Journal of Modern Applied Statistical Methods

Volume 4 | Issue 2 Article 32

11-1-2005

# Vol. 4, No. 2 (Full Issue)

JMASM Editors

# The easy way to find open access journals

DOAJ DIRECTORY OF OPEN ACCESS JOURNALS

www.doaj.org

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

## Aims

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

## Scope

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

## In DOAJ browse by subject

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

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

Funded by
www.soros.org

Hosted by
LUND UNIVERSITY
www.lu.se

Journal of Modern Applied Statistical Methods
November, 2005, Vol. 4, No. 2, 353- 628


1538 – 9472/05/$95.00

# Journal Of Modern Applied Statistical Methods

Journal of Modern Applied Statistical Methods
November, 2005, Vol. 4, No. 2, 353-628


1538 – 9472/05/$95.00

# Editorial Board

Journal of Modern Applied Statistical Methods
November, 2005, Vol. 4, No. 2, 353- 628


1538 – 9472/05/$95.00

# Journal Of Modern Applied Statistical Methods

## *Invited Articles*



## *Regular Articles*

Journal of Modern Applied Statistical Methods
November, 2005, Vol. 4, No. 2, 353-628

1538 – 9472/05/$95.00

Journal of Modern Applied Statistical Methods
November, 2005, Vol. 4, No. 2, 353- 628


1538 – 9472/05/$95.00



*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.

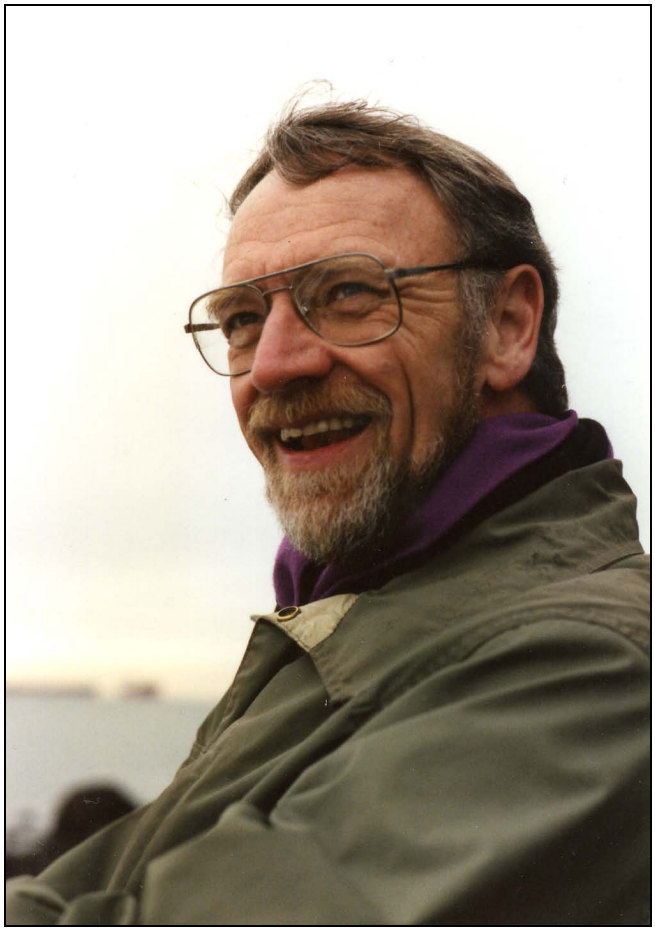


Credits: Photograph of the late Dr. Clifford E. Lunneborg courtesy of Mrs. Pat Lunneborg. The memoriam was written by Dr. Phillip I. Good.

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

Journal of Modern Applied Statistical Methods
November, 2005, Vol. 4, No. 2, 353-628

1538 – 9472/05/$95.00



The ideal man would have all the qualities your mother was always going on about when you were a boy – helpful, considerate, courteous; along with those Tom Sawyer thought important – adventuresome, inspired, and inspiring.

Such was Cliff Lunneborg.
1932 – 2005

Journal of Modern Applied Statistical Methods
November, 2005, Vol. 4, No. 2, 353-371


1538 – 9472/05/$95.00

*INVITED ARTICLES*

# Robust Confidence Intervals for Effect Size in the Two-Group Case

H. J. Keselman
University of Manitoba

James Algina
University of Florida

Katherine Fradette
University of Manitoba

The probability coverage of intervals involving robust estimates of effect size based on seven procedures was compared for asymmetrically trimming data in an independent two-groups design, and a method that symmetrically trims the data. Four conditions were varied: (a) percentage of trimming, (b) type of nonnormal population distribution, (c) population effect size, and (d) sample size. Results indicated that coverage probabilities were generally well controlled under the conditions of nonnormality. The symmetric trimming method provided excellent probability coverage. Recommendations are provided.

Key words: Robust Intervals, effect size statistics, symmetric and asymmetric trimmed means, nonnormality

## Introduction

Journal editorial policies in medicine and psychology encourage researchers to supplement significance testing by reporting confidence intervals (CIs) as well as effect size (ES) statistics. As Fidler, Thomason, Cumming, Finch, and Leeman (2004) note, this movement started in medicine as early as the 1980s (see Rothman 1975, 1978a, 1978b). In psychology, in the past 15 years or so, there has been renewed emphasis on reporting ESs because of editorial policies requiring ESs (e.g., Murphy, 1997; Thompson, 1994) and official support for the practice. According to *The Publication Manual of the American Psychological Association* (2001), "it is almost always necessary to include some index of ES or strength of relationship in your Results section." (p. 25). The practice of reporting ESs has also received support from the APA Task Force on Statistical Inference (Wilkinson and the Task Force on Statistical Inference, 1999). An interest in reporting CIs for ESs has accompanied the emphasis on ESs. Cumming and Finch (2001), for example, presented a primer of CIs for ESs. The purpose of this article is to bring to the attention of researchers in medicine and psychology, and other interested researchers, who set CIs around an ES parameter, a better approach than currently adopted methods.

Algina and Keselman (2003) and Algina, Keselman and Penfield (2005) investigated two two-group ES statistics, looking, in particular, at the confidence coefficient of two intervals associated with each. One of the ES statistics was Cohen's (1965) standardized mean difference statistic

$$d = \frac{\bar{Y}_2 - \bar{Y}_1}{S},$$

H. J. Keselman is Professor of Psychology. Email: kesel@ms.umanitoba.ca. James Algina is Professor of Educational Psychology. Email algina@ufl.edu. Katherine H. Fradette is a doctoral student in the Department of Psychology. Email: umfradet@cc.umanitoba.ca.

where $\bar{Y}_j$ is the mean for the jth level ( $j = 1, 2$ ) of a treatment factor and S is the square root of the pooled variance. The second was

$$d_R = .643\left(\frac{\bar{Y}_{t2} - \bar{Y}_{t1}}{S_W}\right),$$

where $\bar{Y}_{tj}$ denotes the jth 20% trimmed mean, $S_W$ is the square root of the pooled 20% Winsorized variance and .643 is the population 20% Winsorized standard deviation for a standard normal distribution. These authors included .643 in the definition of their robust effect so that the population values of $d_R$ ($\delta_R$) and d ($\delta$) would be equal when data are drawn from normal distributions with equal variances.

However, these authors also pointed out that it is not obligatory to include the .643 multiplier in the definition of $d_R$ and $\delta_R$. Accordingly, the multiplier is excluded in this article. Using each ES statistic, CIs were constructed by using critical values obtained from theory or through a bootstrap method. Algina and Keselman (2003) found that probability coverage for intervals of the usual statistic based on least squares estimators was inaccurate whether or not the interval's critical values were obtained from a theoretical or bootstrap distribution. They also reported that probability coverage was inaccurate when the interval was set around a robust parameter of ES and the critical values for the interval were obtained from a theoretical probability distribution. However, probability coverage was by in large accurate (e.g., .940-.971 for a .95 confidence coefficient) when the interval for the robust parameter of ES was based on critical values obtained through a bootstrap method (see Algina et al., 2005).

Keselman, Wilcox, Lix, Algina and Fradette (in press) found that tests of treatment group equality based on robust estimators performed very well, with respect to Type I error control and power to detect effects in nonnormal heteroscedastic distributions, when adopting robust estimators based on asymmetric trimming of the data. That is, rather than trim a predetermined fixed amount of data from each tail of the empirical distribution, as frequently is recommended in the literature (e.g., 20% from each tail; see Wilcox, 1997; Wilcox & Keselman, 2003), Keselman et al. used nine adaptive procedures that empirically determined the amounts of data that should be trimmed in the right and left tails of each of the nonnormal distributions that they examined in their Monte Carlo investigation. 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.

Based on the two aforementioned studies, it is believed that more accurate confidence coefficients for Algina and Keselman's (2003) and Algina et al.'s (2005) robust parameter of ES could be obtained by adopting the asymmetric trimming procedures enumerated in Keselman et al. (in press). Accordingly, this issue will be investigated in this article.

## Theoretical Background

### ES Statistics and Accompanying CIs

In the two independent-groups paradigm, Cohen's (1965) standardized mean difference statistic, d, is a popular choice for estimating ES. His ES statistic is defined as

$$d = \frac{\bar{Y}_2 - \bar{Y}_1}{S}.$$

Cohen's d estimates

$$\delta = \frac{\mu_2 - \mu_1}{\sigma},$$

where $\mu_j$ is the jth population mean and $\sigma$ is the population standard deviation, assumed to be equal for both groups.

When the scores are independently distributed and are drawn from normal distributions having equal variances, an exact CI for the population ES (i.e., $\delta$) can be constructed by using the noncentral t distribution (see, e.g., Cumming & Finch, 2001 or Steiger & Fouladi, 1997). The noncentral t distribution is

the sampling distribution of the t statistic when $\delta$ is not equal to zero; it has two parameters. The first is the degrees of freedom and equals $N-2$ in the two independent-groups set-up ($[N = n_1 + n_2]$ and the number of observations in a level is denoted by $n_j$). The second parameter is the noncentrality parameter

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

The noncentrality parameter controls the location of the noncentral t distribution. The mean of the noncentral t distribution is $\approx \lambda$ (Hedges, 1981); the accuracy of the approximation improves as N increases.

To find a 95% (for example) CI for $\delta$, one would first use the noncentral t distribution to find a 95% CI for $\lambda$. A CI for $\delta$ can then be obtained by multiplying the limits of the interval for $\lambda$ by $\sqrt{(n_1 + n_2)/n_1 n_2}$. The lower limit of the CI for $\lambda$ is the noncentrality parameter for the noncentral t distribution in which the calculated t statistic

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

is the .975 quantile. The upper limit of the interval for $\lambda$ is the noncentrality parameter for the noncentral t distribution in which the calculated t statistic is the .025 quantile of the distribution (see Steiger & Fouladi, 1997).

The use of the noncentral t distribution is based on the assumption that the data are drawn from normal distributions. If this assumption is not true, there is no guarantee that the actual probability coverage for the interval will match the nominal probability coverage, as was demonstrated by Algina and Keselman (2003). In addition, as noted by Wilcox and Keselman (2003), when data are not normal, the usual population ES can be misleading because the (least squares) means and standard deviations can be affected by skewed data and by outliers. A better strategy, they maintain, is to replace the least squares values by robust estimates, such as trimmed means and Winsorized variances, and, accordingly, estimate a robust population ES.

As an alternative to d, Algina and Keselman (2003) and Algina et al. (2005) (hereafter referred to as A&K) proposed

$$d_R = \left(\frac{\bar{Y}_{t2} - \bar{Y}_{t1}}{S_W}\right).$$

(Remember, the .643 multiplier is not used.)

The robust population ES is

$$\delta_R = \left(\frac{\mu_{t2} - \mu_{t1}}{\sigma_W}\right),$$

where $\mu_{tj}$ is the jth population 20% trimmed mean and $\sigma_W$ is the population analogue of $S_W$. (See appendix 1.)

As Algina and Keselman (2003) and Algina et al. (2005) indicated, an approximately correct CI for $\delta_R$ can also be constructed by using the noncentral t distribution. However, as previously noted, this approach to forming intervals did not provide satisfactory probability coverage when data were obtained from nonnormal distributions. However, Algina et al. did find that probability coverage, under conditions of nonnormality, was generally reasonably good when critical values were obtained through a percentile bootstrap empirical sampling distribution, not from the noncentral t distribution.

## Adaptive Trimming Methods

The theoretical background to the asymmetric trimming methods investigated by Keselman et al. (in press) is now discussed. Based on the work of Hogg (1974, 1982) and others, Reed and Stark (1996) defined seven adaptive location estimators based on measures of tail-length and skewness for a set of n observations. To define these estimators the measures of tail-length and skewness must first be defined. By adopting the notation of Hogg (1974, 1982) and Reed and Stark (1996), based on the ordered values, we let $L_\alpha$ = the mean of

the smallest $[\alpha n]$ observations, where $[\alpha n]$ denote the greatest integer less than $\alpha n$ and $U_\alpha$ = the mean of the largest $[\alpha n]$ observations. When $\alpha = .05$, and, therefore, $L_{(.05)}$ is the mean of the smallest [.05n] observations, $B$ = the mean of the next largest .15n observations, $C$ = the mean of the next largest .30n observations, $D$ = the mean of the next largest .30n observations, and $E$ = the mean of the next largest .15n observations.

*Tail-length measures*. Hogg (1974) defined two measures of tail-length, Q and $Q_1$, where

$$Q = \left(U_{(.05)} - L_{(.05)}\right) \Big/ \left(U_{(.5)} - L_{(.5)}\right) \text{ and}$$

$$Q_1 = \left(U_{(.2)} - L_{(.2)}\right) \Big/ \left(U_{(.5)} - L_{(.5)}\right).$$

Q and $Q_1$ can be used to classify symmetric distributions as light-tailed, medium-tailed or heavy-tailed. Q and $Q_1$ are location free statistics and, moreover, are uncorrelated with location statistics such as trimmed means (Reed & Stark, 1996, p. 12). According to Hogg and Reed and Stark, values of $Q < 2$ imply a light-tailed distribution, $2.0 \le Q \le 2.6$ a medium-tailed distribution, $2.6 < Q \le 3.2$ a heavy-tailed distribution and $Q > 3.2$ a very heavy-tailed distribution. The cutoffs for $Q_1$ are: $Q_1 < 1.81$ (light-tailed), $1.81 \le Q_1 \le 1.87$ (medium-tailed) and $Q_1 > 1.87$ (heavy-tailed).

Hogg (1982) introduced another measure of tail-length:

$$H_3 = \left(U_{(.05)} - L_{(.05)}\right) \Big/ (E - B).$$

With this measure, values of $H_3 < 1.26$ suggest that the tails of the distribution are similar to a uniform distribution, values of 1.26 through 1.76 suggest a normal distribution and values greater than 1.76 suggest the tails are similar to those of a double exponential distribution.

Measures of skewnesss

Reed and Stark (1996) defined four measures of skewness as:

$$Q_2 = \left(U_{(.05)} - T_{(.25)}\right) \Big/ \left(T_{(.25)} - L_{(.05)}\right),$$

$$H_1 = \left(U_{(.05)} - D\right) \Big/ \left(C - L_{(.05)}\right),$$

$$SK_2 = \left(Y_{(1)} - YMD\right) \Big/ \left(YMD - Y_{(n)}\right) \text{ and}$$

$$SK_5 = \left(Y_{(1)} - YM\right) \Big/ \left(YM - Y_{(n)}\right),$$

where YMD is the median, YM is the arithmetic mean, $T_{(.25)}$ is the .25- trimmed mean ($T_\alpha$) given below and $Y_{(1)}$ and $Y_{(n)}$ are, respectively the first and last ordered observations. According to Reed (1998), the $\alpha$-trimmed mean is defined as

$$T_\alpha = \frac{1}{n(1-2\alpha)}\left[\sum_{i=k+1}^{n-k} Y_i + (k - \alpha n)\left(Y_k + Y_{n-k+1}\right)\right].$$

(In this definition a proportion, $\alpha$, has been trimmed from each tail) and the accompanying Winsorized variance $S^2$ is defined as

$$S^2 = \frac{1}{(n-1)(1-2\alpha)^2}$$

$$\left[\sum_{i=k+1}^{n-k} \left(Y_i - T_\alpha\right)^2 + k\left(Y_k - T_\alpha\right)^2 + k\left(Y_{n-k+1} - T_\alpha\right)^2\right]$$

where $k = [\alpha n] + 1$.

Based on the former definitions of tail-length and skewness, Reed and Stark (1996, p. 13) proposed a set of adaptive linear estimators "that have the capability of asymmetric trimming." These authors defined a general scheme for their approach as follows:

1. Set the value for the total amount of trimming from the sample, $\alpha$.

1) Determine the proportion to be trimmed from the lower end of the sample ($\alpha_l$) by the following proportion: $\alpha_l = \alpha\left[UW_X \big/ (UW_X + LW_X)\right]$, where $UW_X$ and $LW_X$ are the numerator and

denominator portions of the previously defined selector statistics (i.e., tail-length and skewness).

2) The upper trimming proportion is then given by $\alpha_u = \alpha - \alpha_l$.

Based on this general schema, Reed and Stark (1996) defined seven hinge estimators, which are trimmed means:

1. HQ $\alpha_l = \alpha\left[UW_Q/(UW_Q + LW_Q)\right]$,
2. $HQ_1$ $\alpha_l = \alpha\left[UW_{Q_1}/(UW_{Q_1} + LW_{Q_1})\right]$,
3. $HH_3$ $\alpha_l = \alpha\left[UW_{H_3}/(UW_{H_3} + LW_{H_3})\right]$,
4. $HQ_2$ $\alpha_l = \alpha\left[UW_{Q_2}/(UW_{Q_2} + LW_{Q_2})\right]$,
5. $HH_1$ $\alpha_l = \alpha\left[UW_{H_1}/(UW_{H_1} + LW_{H_1})\right]$,
6. $HSK_2$ $\alpha_l = \alpha\left[UW_{SK_2}/(UW_{SK_2} + LW_{SK_2})\right]$, and
7. $HSK_5$ $\alpha_l = \alpha\left[UW_{SK_5}/(UW_{SK_5} + LW_{SK_5})\right]$.

Keselman et al. (in press), investigating Type I error rates and power of procedures for testing equality of two trimmed means when variances are not assumed to be equal, examined the Reed and Stark (1996) procedure with various values for $\alpha$ because the literature varies on the amount of recommended (symmetric) trimming. Rosenberger and Gasko (1983) recommended 25% when sample sizes are small, though they thought generally 20% suffices. Wilcox (1997) also recommended 20%, and Mudholkar, Mudholkar and Srivastava (1991) suggested 15%. Ten percent has been considered by Hill and Dixon (1982), Huber (1977), Stigler (1977) and Staudte and Sheather (1990); results reported by Keselman, Wilcox, Othman and Fradette (2002) also support 10% trimming.

Reed and Stark (1996) found, based on a simulation study, that $T_{.10}$, $T_{.15}$, $HSK_2$ and $HSK_5$ were the most efficient estimators when the distribution was symmetric. When the distribution was asymmetric, they found that "**HQ**, $\mathbf{HQ_1}$, $\mathbf{HQ_2}$, $HH_1$, $HSK_2$ and $HSK_5$ [were] consistently among the top four estimators, with $HQ_1$ and $HQ_2$ in the top three" (p. 661).

According to Keselman et al. (in press), one can modify Reed and Stark's (1996) tail-length and skewness measures for the multi-group problem and then apply the modified multi-group measures to the hinge estimators. In particular, they indicated that each of the measures can be modified by taking weighted averages (in a manner analogous to the modifications of tail-length and symmetry measures suggested by Babu, Padmanaban and Puri, 1999) of each numerator and denominator term. For example, for the multi-group problem, where $n_j$ represents the number of observations in each group, $Q_1$ and $Q_2$ can be defined as

$$Q_1 = \left[\sum_j n_j\left(U_{(.2)} - L_{(.2)}\right)\Big/\sum_j n_j\right]\Big/\left[\sum_j n_j\left(U_{(.5)} - L_{(.5)}\right)\Big/\sum_j n_j\right],$$

and

$$Q_2 = \left[\sum_j n_j\left(U_{(.05)} - T_{(.25)}\right)\Big/\sum_j n_j\right]\Big/\left[\sum_j n_j\left(T_{(.25)} - L_{(.05)}\right)\Big/\sum_j n_j\right].$$

The other measures would be similarly modified and these multi-group measures of tail-length and skewness are the measures that are applied to the general scheme proposed by Reed and Stark (1996).

Based on these multi-group tail-length and skewness measures, and their application to the hinge estimators, Keselman et al. (in press) reported that over the 288 empirical values they collected for each method investigated, in which they varied the total percent of data trimmed, sample size, degree of variance heterogeneity, pairing of variances and group sizes and population shape, five methods resulted in exceptionally good control of Type I error rates (HH3, HQ2, HH1, HSK2 and HSK5). With regard to the power to detect nonnull treatment effects, they found that HH3 was uniformly more powerful than the remaining ones.

Robust Estimation

In this study, the methods for constructing CIs for a robust ES, defined by using robust measures of central tendency and variability are investigated. It is important to note that $\alpha$-trimmed means and Winsorized variances can be defined in a number of different ways (Hogg, 1974; Reed, 1998; Keselman et al., in press; Wilcox, 2003). Suppose $n_j$ independent random observations $Y_{1j}, Y_{2j}, \ldots, Y_{n_j j}$ are sampled from population j ($j = 1, 2$). Let $Y_{(1)j} \le Y_{(2)j} \le \cdots \le Y_{(n_j)j}$ represent the ordered observations associated with the jth group. The approach taken by Reed (1998) is based on the work of Hogg (1974). For Hogg, the $\alpha$-trimmed mean is

$$m(\alpha) = (1/h)\sum_{i=g+1}^{n_j-g} Y_{(i)},$$

where $\alpha$ is usually selected so that $g = \lfloor n_j\alpha \rfloor$ and $h = n_j - 2g = n_j - 2\lfloor n_j\alpha \rfloor$. The standard error of $m(\alpha)$ that Hogg suggests is based on the work of Tukey and McLaughlin (1963) and Huber (1970) and, according to these authors, is estimated by

$$S_{m(\alpha)} = \sqrt{SS(\alpha)/h(h-1)},$$

where $SS(\alpha)$ is the Winsorized sum of squares, defined as

$$(g+1)\left[Y_{(g+1)} - m(\alpha)\right]^2 + \left[Y_{(g+2)} - m(\alpha)\right]^2 + \ldots + \left[Y_{(n_j-g-1)} - m(\alpha)\right]^2 + (g+1)\left[Y_{(n_j-g)} - m(\alpha)\right]^2.$$

When allowing for different amounts of trimming in each tail of the distribution, Hogg (1974) defines the trimmed mean as

$$m(\alpha_1, \alpha_1) = (1/h)\sum_{i=g_1+1}^{n_j-g_2} Y_{(i)},$$

where $g_1 = \lfloor n_j\alpha_1 \rfloor$ and $g_2 = \lfloor n_j\alpha_2 \rfloor$ and $h = n_j - g_1 - g_2$. Hogg suggests that the standard deviation of $m(\alpha_1, \alpha_2)$ can be estimated as

$$S_{m(\alpha_1, \alpha_2)} = \sqrt{SS(\alpha_1, \alpha_2)/h(h-1)},$$

where $SS(\alpha_1, \alpha_2)$ can be calculated as

$$\begin{aligned} &(g_1+1)\left[Y_{(g_1+1)} - m(\alpha_1, \alpha_2)\right]^2 \\ &+\left[Y_{(g_1+2)} - m(\alpha_1, \alpha_2)\right]^2 + \ldots \\ &+\left[Y_{(n_j-g_2-1)} - m(\alpha_1, \alpha_2)\right]^2 + \\ &(g_2+1)\left[Y_{(n_j-g_2)} - m(\alpha_1, \alpha_2)\right]^2 \\ &-\frac{\left\{(g_1)\left[Y_{(g_1+1)} - m(\alpha_1, \alpha_2)\right] + \right\}^2}{n_j} \end{aligned}$$

Based on the preceding, our robust estimate of ES for asymmetrically trimmed data is defined as

$$d_R = \frac{m_1(\alpha_1, \alpha_2) - m_2(\alpha_1, \alpha_2)}{\sqrt{\frac{SS_1(\alpha_1, \alpha_2) + SS_2(\alpha_1, \alpha_2)}{N-2}}},$$

where $m_j(\alpha_1, \alpha_2)$ and $SS_j(\alpha_1, \alpha_2)$ are the jth asymmetrically trimmed mean and sum of squares, respectively. (See Appendix 2.)

## Methodology

Probability coverage for seven ES statistics (based on seven hinge estimators: HQ, HQ1, HH3, HQ2, HH1, HSK2, and HSK5) was estimated for all combinations of the following four factors: (a) four values of total trimming, namely 10%, 15%, 20% and 25%, (b) population distribution (four cases from the family of g and h distributions), (c) sample size: $n_1 = n_2 = 20,\ 40,\ 60,\ 80,$ and $100$, and (d) population ES ($PES = \delta_R$) of 0, .2, .5, .8, 1.1, and 1.2. The A&K statistic was also included, where the values of symmetric trimming investigated were 5%, 10%, 15% and 20%.

The data were generated from the family of g and h distributions (Hoaglin, 1985). Specifically, it was chosen to investigate four g and h distributions:

(a) $g = h = 0$, the standard normal distribution ($\gamma_1 = \gamma_2 = 0$),
(b) $g = 0$ and $h = .225$, a long-tailed distribution ($\gamma_1 = 0, \gamma_2 = 154.84$),
(c) $g = .76$ and $h = -.098$, a distribution with skew and kurtosis equal to that for an exponential distribution ($\gamma_1 = 2, \gamma_2 = 6$), and
(d) $g = .225$ and $h = .225$, a long-tailed skewed distribution ($\gamma_1 = 4.90, \gamma_2 = 4673.80$).

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, $Y_{ij} = Z_{ij} \exp\left(\frac{hZ_{ij}^2}{2}\right)$. The $Z_{ij}$ scores were generated by using RANNOR from SAS (1999). In particular, the following method to generate our data was used:

1. The original $Y_{ij}$ data (for both groups) were generated from a desired population distribution (e.g., $g = .225$ and $h = .225$). (NOTE: The original $Y_{i2}$ data are not yet transformed)
2. A bootstrap sample ($Y_{ij}^*$) was obtained from the original sample by sampling $n_1$ observations with replacement from $Y_{i1}$ and $n_2$ observations with replacement from $Y_{i2}$.
3. With the bootstrap data, we determined $\alpha_1$ and $\alpha_2$ for the desired total trimming percentage (e.g., 15%) for each of the seven hinge estimators.
4. The bootstrapped data for group 2 ($Y_{i2}^*$) were then transformed according to $Y_{i2}^* + \sigma_W \times \delta_R$, where $\sigma_W$ depended on the hinge estimator, the total % of trimming, and the population distribution under investigation. For a particular population distribution and total % of trimming, $\sigma_W$ was determined prior to conducting the study. That is, 1,000,000 observations were first generated from the population distribution in question and then the population trimming strategy was determined for each of the hinge estimators under the desired total % of trimming. The $\sigma_W$ values for the seven different hinge estimators were then determined by computing the Winsorized standard deviation of the 1,000,000 observations, using the trimming strategies of each of the estimators.
5. The transformed bootstrap data was then used to compute the trimmed means ($\bar{Y}_{t1}^*$ and $\bar{Y}_{t2}^*$) and the pooled Winsorized standard deviation ($S_W^*$) for each of the 7 different hinge estimator methods, based on the trimming strategies previously determined.

6. For each estimator, the following was computed $d_R^* = \frac{\bar{Y}_{t2}^* - \bar{Y}_{t1}^*}{S_W^*}$.
7. Steps 1 through 6 were repeated 600 times.
8. For each hinge estimator, the 600 bootstrap ES estimates ($d_R^*$) were ranked and the upper and lower limits of the CIs were determined in the following manner. Letting $l = .025B$, rounded to the nearest integer, and $u = B - l$, an estimate of the .025 and .975 quantiles of the distribution of $d_R$ is $d_{R_{(l+1)}}^*$ and $d_{R_{(u)}}^*$.
9. Finally, steps 1 through 8 were repeated 5000 times.

The nominal confidence level for all intervals was .95.

## Results

Table 1 contains average probability coverage rates for the seven hinge estimator methods as well as A&K for setting intervals around the PES for the effects investigated. Bradley's (1978) liberal criterion will be used to judge the robustness of the methods.

Coverage probabilities within the interval .925-.975 are deemed well controlled, while those outside this range are regarded as substantially affected by an investigated effect(s). Values outside the interval will be demarcated with boldface type in the tables. The grand mean coverage probabilities were obtained over 480 conditions and most apparent is that the empirical values are not only contained in Bradley's interval, but, moreover, are actually quite close to the nominal .95 value, with the largest deviation between nominal and empirical values equaling .004. Indeed, the range of empirical values extends from .946 to .949. Similarly, none of the remaining Table 1 values fell outside the Bradley liberal criterion.

Thus, by this standard of robustness, all hinge estimator methods for setting intervals around the robust PES can be regarded as not adversely affected by the effects of percentage of trimming, sample size, PES, and shape of distribution. Indeed, the number of times each of the methods' empirical values fell outside the liberal interval were tabulated and it was found that, over the 3840 estimates (480 conditions X 8 procedures), only 56 were not contained in the interval (less than 1.5% of the values!).

Not surprisingly, 51 of these values occurred when $n = 20$; the remaining five values occurred when $n = 40$. From this tabulation it was also found that, of the hinge estimator procedures, only HSK2 and HSK5 never had a value outside the Bradley interval. However, if the $n = 20$ results are excluded, then HQ, HQ1, and HH3 can be added to this list of procedures that never had a value over the 480 conditions outside the Bradley interval. Also noteworthy is that all 480 of the A&K values were in the Bradley interval.

Nonetheless, one can observe from the tabled values that there are variations in coverage probabilities due to the investigated effects. That is, it appears that coverage probabilities were closer to .95 when the: (a) percentage of total trimming was at least 20% (for A&K the empirical estimates were equal across percentages of symmetric trimming), (b) sample size was at least 80 per group, and (c) nonnormal distribution was not $g = .76$ and $h = -.098$.

Accordingly, exemplars of these empirical coverage probabilities are presented in Tables 2-5, where the four tables are for the four distributions investigated. When $PES = 0$, all empirical coverage probabilities (not tabled) were contained within Bradley's (1978) interval across all sample size and population distributions investigated. In Tables 2-5, 28 of the 1152 empirical values ($\simeq 2.4\%$) were not contained in the .925-.975 interval. Twenty-five of the affected values occurred when data were obtained from the $g = .76$ and $h = -.098$ distribution and when $n = 20$ (Table 4).

The remaining three liberal values also occurred when $n = 20$ but in these instances the data were $g = .225$ and $h = .225$ distributed. One should also notice that empirical values for the A&K procedure were always in Bradley's (1978) interval across the

Table 1. Summary Data for Estimated Coverage Probabilities for Nominal 95% Bootstrap Intervals

| Condition | A&K | HQ | HQ1 | HH3 | HQ2 | HH1 | HSK2 | HSK5 |
|---|---|---|---|---|---|---|---|---|
| Grand Mean | .949 | .947 | .948 | .947 | .947 | .946 | .948 | .948 |
| % Trimming | | | | | | | | |
| 10 | | .943 | .945 | .944 | .944 | .942 | .948 | .948 |
| 15 | | .946 | .949 | .946 | .947 | .946 | .949 | .948 |
| 20 | | .949 | .949 | .948 | .948 | .947 | .948 | .948 |
| 25 | | .949 | .949 | .948 | .949 | .948 | .947 | .948 |
| 5 (Symmetric) | .949 | | | | | | | |
| 10 (Symmetric) | .949 | | | | | | | |
| 15 (Symmetric) | .949 | | | | | | | |
| 20 (Symmetric) | .949 | | | | | | | |
| Sample Size | | | | | | | | |
| 20 | .950 | .939 | .943 | .937 | .938 | .936 | .948 | .949 |
| 40 | .951 | .948 | .950 | .948 | .948 | .946 | .949 | .949 |
| 60 | .946 | .949 | .949 | .949 | .949 | .948 | .947 | .947 |
| 80 | .950 | .950 | .950 | .949 | .950 | .950 | .948 | .948 |
| 100 | .948 | .950 | .949 | .950 | .950 | .950 | .947 | .947 |
| PES | | | | | | | | |
| 0 | .946 | .945 | .945 | .945 | .947 | .946 | .946 | .946 |
| 0.2 | .947 | .946 | .947 | .946 | .948 | .947 | .948 | .948 |
| 0.5 | .949 | .946 | .947 | .946 | .947 | .946 | .947 | .947 |
| 0.8 | .949 | .948 | .949 | .947 | .947 | .946 | .948 | .948 |
| 1.1 | .951 | .949 | .950 | .948 | .948 | .946 | .949 | .949 |
| 1.4 | .953 | .948 | .949 | .947 | .947 | .944 | .949 | .948 |
| Distribution | | | | | | | | |
| g=0/h=0 | .947 | .946 | .946 | .946 | .947 | .947 | .946 | .947 |
| g=0/h=.225 | .951 | .944 | .946 | .944 | .941 | .936 | .946 | .944 |
| g=.76/h=-.098 | .947 | .950 | .950 | .949 | .950 | .950 | .950 | .951 |
| g=.225/h=.225 | .951 | .949 | .950 | .948 | .951 | .951 | .950 | .950 |

*Notes*: Based on definitions of tail-length and skewness, Reed and Stark (1996, p. 13) defined seven hinge estimators that have the capability of asymmetric trimming: HQ, HQ1, HH3, HQ2, HH1, HSK2, HSK5; Sample Size ($n_1 = n_2$); PES-Population Effect Size; $g = X / h = Y$ specifies a particular g and h distribution with specific values of skewness and kurtosis.

Table 2. Estimated Coverage Probabilities for Nominal 95% Bootstrap Intervals ($g = 0$ & $h = 0$).

| PES | n | Trimming | Test | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | A&K | HQ | HQ1 | HH3 | HQ2 | HH1 | HSK2 | HSK5 |
| 0.2 | 20 | 5% | .942 | | | | | | | |
| | | 10% | .943 | .935 | .935 | .935 | .938 | .937 | .939 | .940 |
| | | 15% | .944 | .940 | .941 | .939 | .942 | .942 | .941 | .942 |
| | | 20% | .945 | .942 | .943 | .941 | .944 | .944 | .942 | .942 |
| | | 25% | | .942 | .942 | .942 | .944 | .944 | .942 | .942 |
| | 60 | 5% | .940 | | | | | | | |
| | | 10% | .939 | .945 | .944 | .945 | .945 | .945 | .944 | .944 |
| | | 15% | .940 | .946 | .945 | .945 | .945 | .945 | .945 | .945 |
| | | 20% | .938 | .946 | .945 | .946 | .946 | .946 | .944 | .945 |
| | | 25% | | .945 | .946 | .945 | .946 | .946 | .945 | .946 |
| | 100 | 5% | .948 | | | | | | | |
| | | 10% | .949 | .945 | .944 | .946 | .946 | .946 | .945 | .945 |
| | | 15% | .948 | .947 | .946 | .947 | .947 | .947 | .946 | .945 |
| | | 20% | .947 | .946 | .945 | .945 | .947 | .947 | .946 | .946 |
| | | 25% | | .945 | .945 | .945 | .946 | .946 | .946 | .946 |
| 0.8 | 20 | 5% | .946 | | | | | | | |
| | | 10% | .950 | .939 | .939 | .939 | .940 | .940 | .943 | .944 |
| | | 15% | .951 | .946 | .947 | .943 | .946 | .946 | .946 | .946 |
| | | 20% | .953 | .951 | .951 | .950 | .950 | .949 | .949 | .951 |
| | | 25% | | .949 | .950 | .948 | .952 | .952 | .950 | .952 |
| | 60 | 5% | .943 | | | | | | | |
| | | 10% | .943 | .947 | .949 | .949 | .950 | .950 | .949 | .949 |
| | | 15% | .943 | .949 | .950 | .950 | .949 | .949 | .947 | .947 |
| | | 20% | .947 | .951 | .951 | .950 | .951 | .951 | .950 | .951 |
| | | 25% | | .950 | .949 | .950 | .953 | .953 | .951 | .952 |
| | 100 | 5% | .944 | | | | | | | |
| | | 10% | .944 | .949 | .949 | .949 | .949 | .949 | .949 | .949 |
| | | 15% | .945 | .949 | .949 | .948 | .948 | .948 | .947 | .947 |
| | | 20% | .945 | .950 | .950 | .949 | .949 | .950 | .949 | .949 |
| | | 25% | | .949 | .948 | .948 | .948 | .948 | .947 | .948 |
| 1.4 | 20 | 5% | .943 | | | | | | | |
| | | 10% | .951 | .939 | .939 | .939 | .940 | .940 | .942 | .943 |
| | | 15% | .952 | .946 | .950 | .944 | .947 | .947 | .949 | .949 |
| | | 20% | .954 | .951 | .948 | .952 | .952 | .951 | .954 | .953 |
| | | 25% | | .950 | .951 | .950 | .954 | .953 | .953 | .955 |
| | 60 | 5% | .945 | | | | | | | |
| | | 10% | .946 | .947 | .948 | .947 | .950 | .951 | .948 | .947 |
| | | 15% | .946 | .948 | .947 | .948 | .949 | .949 | .948 | .947 |
| | | 20% | .945 | .951 | .950 | .949 | .948 | .948 | .948 | .948 |
| | | 25% | | .950 | .950 | .949 | .950 | .950 | .950 | .950 |
| | 100 | 5% | .946 | | | | | | | |
| | | 10% | .949 | .948 | .949 | .949 | .949 | .949 | .948 | .948 |
| | | 15% | .949 | .950 | .950 | .950 | .949 | .949 | .948 | .949 |
| | | 20% | .950 | .949 | .951 | .950 | .950 | .950 | .947 | .948 |
| | | 25% | | .949 | .949 | .949 | .949 | .948 | .950 | .950 |

Table 3. Estimated Coverage Probabilities for Nominal 95% Bootstrap Intervals ($g = 0$ & $h = .225$).

| PES | N | Trimming | Test | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | A&K | HQ | HQ1 | HH3 | HQ2 | HH1 | HSK2 | HSK5 |
| 0.2 | 20 | 5% | .944 | | | | | | | |
| | | 10% | .950 | .936 | .937 | .937 | .934 | .933 | .942 | .942 |
| | | 15% | .949 | .935 | .946 | .933 | .943 | .942 | .946 | .947 |
| | | 20% | .946 | .944 | .947 | .943 | .946 | .945 | .946 | .947 |
| | | 25% | | .947 | .947 | .944 | .948 | .947 | .945 | .948 |
| | 60 | 5% | .942 | | | | | | | |
| | | 10% | .943 | .948 | .948 | .948 | .953 | .952 | .948 | .948 |
| | | 15% | .941 | .950 | .950 | .950 | .950 | .951 | .950 | .949 |
| | | 20% | .940 | .948 | .949 | .948 | .949 | .948 | .946 | .946 |
| | | 25% | | .949 | .949 | .948 | .950 | .950 | .945 | .947 |
| | 100 | 5% | .950 | | | | | | | |
| | | 10% | .951 | .951 | .950 | .950 | .949 | .950 | .946 | .947 |
| | | 15% | .950 | .949 | .948 | .949 | .948 | .948 | .948 | .948 |
| | | 20% | .950 | .949 | .948 | .947 | .949 | .950 | .949 | .949 |
| | | 25% | | .948 | .947 | .947 | .949 | .948 | .949 | .946 |
| 0.8 | 20 | 5% | .949 | | | | | | | |
| | | 10% | .959 | .937 | .937 | .937 | .935 | .934 | .946 | .948 |
| | | 15% | .958 | .943 | .953 | .940 | .944 | .943 | .952 | .951 |
| | | 20% | .958 | .952 | .953 | .949 | .950 | .950 | .955 | .955 |
| | | 25% | | .953 | .953 | .952 | .954 | .953 | .955 | .957 |
| | 60 | 5% | .953 | | | | | | | |
| | | 10% | .948 | .949 | .949 | .947 | .952 | .952 | .951 | .951 |
| | | 15% | .946 | .951 | .956 | .951 | .950 | .952 | .953 | .952 |
| | | 20% | .948 | .957 | .952 | .955 | .953 | .953 | .950 | .950 |
| | | 25% | | .954 | .951 | .954 | .953 | .953 | .950 | .952 |
| | 100 | 5% | .950 | | | | | | | |
| | | 10% | .946 | .954 | .955 | .955 | .958 | .959 | .953 | .954 |
| | | 15% | .944 | .955 | .954 | .956 | .953 | .955 | .953 | .953 |
| | | 20% | .947 | .953 | .950 | .953 | .953 | .953 | .951 | .950 |
| | | 25% | | .952 | .951 | .952 | .951 | .951 | .943 | .951 |
| 1.4 | 20 | 5% | .952 | | | | | | | |
| | | 10% | .965 | .934 | .933 | .933 | .929 | .928 | .948 | .947 |
| | | 15% | .963 | .941 | .958 | .938 | .939 | .937 | .954 | .952 |
| | | 20% | .963 | .954 | .946 | .946 | .943 | .942 | .957 | .957 |
| | | 25% | | .950 | .948 | .946 | .949 | .948 | .962 | .958 |
| | 60 | 5% | .960 | | | | | | | |
| | | 10% | .955 | .950 | .947 | .945 | .954 | .951 | .956 | .957 |
| | | 15% | .951 | .949 | .959 | .948 | .950 | .951 | .954 | .954 |
| | | 20% | .949 | .960 | .953 | .957 | .954 | .953 | .952 | .953 |
| | | 25% | | .959 | .953 | .955 | .954 | .954 | .950 | .953 |
| | 100 | 5% | .956 | | | | | | | |
| | | 10% | .955 | .957 | .956 | .956 | .959 | .959 | .954 | .954 |
| | | 15% | .953 | .954 | .951 | .953 | .957 | .957 | .951 | .952 |
| | | 20% | .950 | .956 | .952 | .952 | .954 | .954 | .953 | .953 |
| | | 25% | | .954 | .954 | .954 | .954 | .955 | .935 | .951 |

Table 4. Estimated Coverage Probabilities for Nominal 95% Bootstrap Intervals ($g = .76$ & $h = -.098$).

| PES | N | Trimming | Test | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | A&K | HQ | HQ1 | HH3 | HQ2 | HH1 | HSK2 | HSK5 |
| 0.2 | 20 | 5% | .940 | | | | | | | |
| | | 10% | .946 | .927 | .927 | .927 | .926 | .926 | .943 | .943 |
| | | 15% | .947 | .932 | .941 | .932 | .930 | .929 | .946 | .946 |
| | | 20% | .947 | .941 | .942 | .939 | .935 | .932 | .945 | .946 |
| | | 25% | | .943 | .944 | .942 | .940 | .935 | .945 | .945 |
| | 60 | 5% | .936 | | | | | | | |
| | | 10% | .938 | .944 | .948 | .944 | .944 | .938 | .947 | .948 |
| | | 15% | .938 | .948 | .947 | .949 | .945 | .944 | .946 | .947 |
| | | 20% | .938 | .948 | .949 | .949 | .949 | .946 | .948 | .947 |
| | | 25% | | .947 | .949 | .949 | .948 | .947 | .949 | .949 |
| | 100 | 5% | .948 | | | | | | | |
| | | 10% | .944 | .950 | .949 | .950 | .947 | .946 | .949 | .948 |
| | | 15% | .948 | .949 | .950 | .950 | .949 | .948 | .949 | .949 |
| | | 20% | .949 | .950 | .949 | .948 | .951 | .949 | .948 | .947 |
| | | 25% | | .950 | .948 | .948 | .950 | .949 | .947 | .948 |
| 0.8 | 20 | 5% | .934 | | | | | | | |
| | | 10% | .948 | **.909** | **.914** | **.909** | **.905** | **.895** | .940 | .941 |
| | | 15% | .948 | **.921** | .934 | **.922** | **.912** | **.906** | .948 | .949 |
| | | 20% | .950 | .934 | .939 | .935 | **.921** | **.909** | .948 | .949 |
| | | 25% | | .939 | .942 | .941 | .926 | **.917** | .951 | .948 |
| | 60 | 5% | .949 | | | | | | | |
| | | 10% | .949 | .946 | .947 | .946 | .941 | .933 | .948 | .948 |
| | | 15% | .944 | .948 | .947 | .951 | .946 | .941 | .947 | .947 |
| | | 20% | .944 | .950 | .950 | .951 | .949 | .943 | .945 | .941 |
| | | 25% | | .951 | .951 | .951 | .947 | .947 | .945 | .941 |
| | 100 | 5% | .946 | | | | | | | |
| | | 10% | .948 | .952 | .950 | .951 | .954 | .948 | .946 | .947 |
| | | 15% | .945 | .949 | .949 | .950 | .951 | .952 | .946 | .944 |
| | | 20% | .946 | .948 | .947 | .947 | .947 | .949 | .944 | .936 |
| | | 25% | | .947 | .948 | .946 | .949 | .949 | .941 | .937 |
| 1.4 | 20 | 5% | .929 | | | | | | | |
| | | 10% | .957 | **.903** | **.907** | **.903** | **.892** | **.878** | .942 | .943 |
| | | 15% | .953 | **.912** | .932 | **.913** | **.905** | **.894** | .955 | .954 |
| | | 20% | .956 | .931 | .939 | .931 | **.917** | **.898** | .956 | .952 |
| | | 25% | | .938 | .945 | .938 | .924 | **.911** | .948 | .942 |
| | 60 | 5% | .955 | | | | | | | |
| | | 10% | .953 | .943 | .951 | .942 | .939 | **.921** | .944 | .946 |
| | | 15% | .950 | .952 | .951 | .953 | .944 | .938 | .948 | .943 |
| | | 20% | .949 | .953 | .952 | .953 | .948 | .940 | .944 | .933 |
| | | 25% | | .951 | .954 | .952 | .950 | .946 | .939 | .932 |
| | 100 | 5% | .953 | | | | | | | |
| | | 10% | .952 | .951 | .951 | .949 | .946 | .935 | .953 | .953 |
| | | 15% | .952 | .950 | .950 | .951 | .949 | .945 | .952 | .945 |
| | | 20% | .951 | .950 | .951 | .953 | .952 | .944 | .948 | .932 |
| | | 25% | | .947 | .953 | .950 | .947 | .948 | .936 | .931 |

Table 5. Estimated Coverage Probabilities for Nominal 95% Bootstrap Intervals ($g = .225$ & $h = .225$).

| PES | N | Trimming | Test | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | A&K | HQ | HQ1 | HH3 | HQ2 | HH1 | HSK2 | HSK5 |
| 0.2 | 20 | 5% | .946 | | | | | | | |
| | | 10% | .951 | .929 | .930 | .930 | .932 | .931 | .943 | .944 |
| | | 15% | .950 | .931 | .944 | .930 | .941 | .940 | .946 | .947 |
| | | 20% | .949 | .941 | .946 | .938 | .946 | .944 | .948 | .949 |
| | | 25% | | .947 | .947 | .945 | .949 | .948 | .946 | .947 |
| | 60 | 5% | .944 | | | | | | | |
| | | 10% | .942 | .946 | .946 | .945 | .948 | .948 | .948 | .948 |
| | | 15% | .942 | .947 | .948 | .949 | .951 | .951 | .947 | .948 |
| | | 20% | .939 | .949 | .950 | .950 | .953 | .953 | .947 | .947 |
| | | 25% | | .950 | .950 | .950 | .952 | .952 | .946 | .946 |
| | 100 | 5% | .948 | | | | | | | |
| | | 10% | .950 | .950 | .951 | .952 | .952 | .953 | .947 | .948 |
| | | 15% | .949 | .951 | .948 | .950 | .952 | .952 | .948 | .948 |
| | | 20% | .950 | .950 | .949 | .949 | .950 | .951 | .950 | .950 |
| | | 25% | | .950 | .947 | .948 | .948 | .948 | .950 | .950 |
| 0.8 | 20 | 5% | .950 | | | | | | | |
| | | 10% | .957 | .926 | .928 | .928 | .932 | .931 | .943 | .944 |
| | | 15% | .956 | .934 | .950 | .934 | .944 | .943 | .949 | .951 |
| | | 20% | .956 | .947 | .951 | .942 | .949 | .947 | .953 | .953 |
| | | 25% | | .948 | .948 | .946 | .954 | .952 | .955 | .955 |
| | 60 | 5% | .955 | | | | | | | |
| | | 10% | .949 | .949 | .949 | .947 | .950 | .950 | .951 | .951 |
| | | 15% | .947 | .950 | .955 | .952 | .955 | .957 | .948 | .948 |
| | | 20% | .945 | .957 | .953 | .957 | .954 | .957 | .952 | .952 |
| | | 25% | | .956 | .953 | .955 | .956 | .954 | .953 | .952 |
| | 100 | 5% | .949 | | | | | | | |
| | | 10% | .949 | .954 | .956 | .956 | .956 | .955 | .951 | .951 |
| | | 15% | .946 | .956 | .952 | .954 | .954 | .956 | .950 | .951 |
| | | 20% | .948 | .954 | .951 | .953 | .951 | .954 | .950 | .951 |
| | | 25% | | .951 | .950 | .949 | .951 | .951 | .950 | .950 |
| 1.4 | 20 | 5% | .950 | | | | | | | |
| | | 10% | .965 | **.924** | .926 | .926 | **.924** | **.923** | .946 | .947 |
| | | 15% | .964 | .930 | .955 | .927 | .939 | .940 | .954 | .952 |
| | | 20% | .963 | .950 | .948 | .939 | .946 | .944 | .958 | .955 |
| | | 25% | | .953 | .945 | .943 | .953 | .950 | .957 | .959 |
| | 60 | 5% | .961 | | | | | | | |
| | | 10% | .955 | .949 | .948 | .944 | .951 | .949 | .953 | .953 |
| | | 15% | .952 | .951 | .961 | .949 | .956 | .958 | .952 | .952 |
| | | 20% | .951 | .960 | .958 | .961 | .955 | .958 | .951 | .949 |
| | | 25% | | .963 | .956 | .956 | .957 | .958 | .953 | .951 |
| | 100 | 5% | .958 | | | | | | | |
| | | 10% | .957 | .957 | .957 | .955 | .957 | .958 | .954 | .954 |
| | | 15% | .952 | .957 | .955 | .957 | .956 | .958 | .952 | .953 |
| | | 20% | .952 | .956 | .955 | .956 | .953 | .956 | .953 | .952 |
| | | 25% | | .954 | .954 | .956 | .956 | .956 | .951 | .952 |

Table 6. Ranks

| N | Test | PES=0 | PES=.2 | PES=.5 | PES=.8 | PES=1.1 | PES=1.4 | **Total** |
|---|---|---|---|---|---|---|---|---|
| 20 | HQ | 1 | 2 | 5 | 6 | 3 | 6 | **23** |
| | HQ1 | 5 | 5 | 9 | 8 | 7 | 9 | **43** |
| | HH3 | 0 | 0 | 3 | 3 | 4 | 3 | **13** |
| | HQ2 | 6 | 4 | 8 | 7 | 4 | 5 | **34** |
| | HH1 | 4 | 3 | 7 | 6 | 6 | 5 | **31** |
| | HSK2 | 7 | 8 | 12 | 10 | 10 | 8 | **55** |
| | HSK5 | 12 | 9 | 13 | 10 | 10 | 7 | **61** |
| | **Total** | **35** | **31** | **57** | **50** | **44** | **43** | **260** |
| | | | | | | | | |
| 40 | HQ | 5 | 11 | 10 | 7 | 7 | 8 | **48** |
| | HQ1 | 9 | 15 | 12 | 10 | 11 | 13 | **70** |
| | HH3 | 7 | 13 | 13 | 5 | 10 | 10 | **58** |
| | HQ2 | 8 | 5 | 7 | 9 | 8 | 11 | **48** |
| | HH1 | 9 | 6 | 6 | 5 | 9 | 8 | **43** |
| | HSK2 | 6 | 12 | 15 | 10 | 13 | 11 | **67** |
| | HSK5 | 7 | 12 | 15 | 9 | 9 | 8 | **60** |
| | **Total** | **51** | **74** | **78** | **55** | **67** | **69** | **394** |
| | | | | | | | | |
| 60 | HQ | 14 | 14 | 8 | 12 | 8 | 10 | **66** |
| | HQ1 | 13 | 15 | 12 | 14 | 10 | 11 | **75** |
| | HH3 | 13 | 15 | 9 | 10 | 8 | 6 | **61** |
| | HQ2 | 12 | 14 | 10 | 10 | 9 | 10 | **65** |
| | HH1 | 10 | 13 | 8 | 9 | 11 | 8 | **59** |
| | HSK2 | 9 | 10 | 3 | 14 | 7 | 9 | **52** |
| | HSK5 | 11 | 13 | 4 | 13 | 9 | 8 | **58** |
| | **Total** | **82** | **94** | **54** | **82** | **62** | **62** | **436** |
| | | | | | | | | |
| 80 | HQ | 7 | 12 | 13 | 9 | 10 | 9 | **60** |
| | HQ1 | 3 | 16 | 12 | 11 | 13 | 10 | **65** |
| | HH3 | 8 | 16 | 15 | 11 | 8 | 11 | **69** |
| | HQ2 | 14 | 9 | 8 | 10 | 12 | 14 | **67** |
| | HH1 | 11 | 8 | 6 | 10 | 9 | 9 | **53** |
| | HSK2 | 2 | 16 | 16 | 8 | 12 | 13 | **67** |
| | HSK5 | 4 | 14 | 14 | 9 | 11 | 12 | **64** |
| | **Total** | **49** | **91** | **84** | **68** | **75** | **78** | **445** |
| | | | | | | | | |
| 100 | HQ | 12 | 16 | 12 | 14 | 9 | 9 | **72** |
| | HQ1 | 12 | 14 | 11 | 15 | 13 | 14 | **79** |
| | HH3 | 13 | 14 | 13 | 12 | 10 | 11 | **73** |
| | HQ2 | 16 | 15 | 11 | 12 | 10 | 9 | **73** |
| | HH1 | 16 | 14 | 10 | 11 | 9 | 7 | **67** |
| | HSK2 | 14 | 11 | 1 | 11 | 12 | 11 | **60** |
| | HSK5 | 13 | 11 | 1 | 12 | 13 | 12 | **62** |
| | **Total** | **96** | **95** | **59** | **87** | **76** | **73** | **486** |
| | | | | | | | | |
| | **GT** | **313** | **385** | **332** | **342** | **324** | **325** | **2021** |

Table 7. Total Number of Top Three Rankings for Each Test

| HQ | HQ1 | HH3 | HQ2 | HH1 | HSK2 | HSK5 |
|---|---|---|---|---|---|---|
| 269 | 332 | 274 | 287 | 253 | 301 | 305 |

three tables. (This is expected given the findings we previously enumerated.) One additional point important to mention is that the HSK2 and HSK5 hinge estimators methods as well as the A&K method resulted in well controlled coverage probabilities for the conditions where the affected procedures did not; that is, their coverage probabilities were not affected even though sample size was small ($n_1 = n_2 = 20$) and data were $g = .76$ and $h = -.098$ distributed, for any percentage of total trimming.

Based on the preceding descriptions of our results, it would be difficult to try to pick out the 'best' one, two, or three methods for CIs around the robust PES. Indeed, Table 1 summary results indicate that all empirical values for all procedures were contained in the .925-.975 interval and accordingly, based on these results and the generally robust findings reported in Tables 2-5 (and those not tabled), specific recommendations would be challenging, and perhaps somewhat arbitrary, to make. Nonetheless, applied researchers usually like guidance from quantitative researchers regarding our recommendation of 'best' choice of procedure for their analyses. Accordingly, an even finer examination of our data was made.

In our second phase of analyses, the three hinge estimator methods for setting intervals having coverage probabilities closest to .95 were located; this was done for each combination of sample size, population distribution, total percentage of trimming and PES. Hinge estimator methods having identical empirical coverage probabilities received the same rank (either 1-closest, 2-next closest, or 3-third closest). Preferred ranks were given to deviations that were above .95 as opposed to below .95. Thus, if procedure 'A' resulted in a .951 coverage probability while procedure 'B' had coverage probability of .949, procedure A received the better rank -- the preference was for conservative rather than liberal values. Finally, any value that did not fall into a stringent criterion [($\pm 2\sigma_{1-\alpha}$ for $1-\alpha = .95$) i.e., .945-.955] was excluded from ranking.

Accordingly, in Table 6 the total number of top three rankings as a function of sample size and PES for the seven hinge estimator ES intervals are presented. What one can also see from Table 6 is that: (a) the total number of top three rankings, not surprisingly, increased with the size of sample; for $n_1 = n_2 = 20,\ 40,\ 60,\ 80,$ and $100$, the total number of top three rankings was 260, 394, 436, 445, and 486, respectively; (b) the procedures were most disparate (range=48) from one another in terms of accuracy (i.e., number of top three rankings) when $n_1 = n_2 = 20$ and $40$ and were much more similar to one another when $n_1 = n_2 = 60,\ 80,$ and $100$; and (c) the number of top three rankings increased with PES up until $PES = .2$ and then remained almost the same for $PES = .5$-$1.4$ Finally, the numbers presented in Table 6 and summarized in Table 7 indicate that HQ1 had the greatest number (332) of top three rankings while HSK2 and HSK5 had the second and third most top three rankings (301 and 305, respectively).

## Discussion

Algina and Keselman (2003) and Algina et al. (2005) compared two estimates of ES and associated CIs in an independent two-groups design, in which either least squares or robust estimators were used and where the critical values used in computing the interval were

based on either a theoretical or bootstrap distribution. The procedures were compared under different conditions of nonnormality and for various sample sizes and magnitudes of PES. It was found that probability coverage for the CI was only controlled when the interval used robust estimators (i.e., trimmed means and Winsorized variances) and the critical values of the interval were obtained via a bootstrap empirical distribution. The authors used *a priori* $2\times100\alpha$ % symmetric trimming to remove the biasing effects of skewed data and/or outlying values and only investigated $\alpha = .20$.

In an unrelated study, Keselman et al. (in press) found that tests for treatment group equality based on asymmetrically obtained trimmed means and Winsorized variances, resulted in exceptionally good Type I error control and power to detect effects in nonnormal heterogeneous one-way models. Consequently, it is believed that it would be possible to obtain more accurate probability coverage for intervals of ES in nonnormal models if the ES statistic was based on asymmetrically trimmed data. Accordingly, a Monte Carlo investigation was conducted to probe this hypothesis, varying population shape, magnitude of PES, sample size, and total percentage of trimming.

The results from the investigation clearly suggest that coverage probabilities for robust ES intervals were very well controlled under the conditions of nonnormality that were investigated. That is, only 56 of the 3840 empirical coverage probabilities (less than 1.5% of the values) did not fall within Bradley's (1978) criterion of .925-.975. And, these liberal values (i.e., intervals were too narrow), almost exclusively occurred when sample size was at the minimum value ($n_1 = n_2 = 20$) investigated. However, coverage probabilities, with the exception of two cases, were always within the Bradley interval once sample size reached our medium sample size condition ($n_1 = n_2 = 60$). Thus, based on these findings, any of the hinge estimators for setting a CI around a robust parameter of ES are recommended.

Nonetheless, in the interest of trying to separate the procedures in order to provide a more specific recommendation for researchers intending to set an interval around an ES statistic in a two-groups paradigm, a comparison of the hinge estimator ES intervals with a more stringent criterion was made, a criterion where a procedure would be judged robust if the empirical estimate did not fall outside a .944-.956 interval ($\pm2\sigma_{1-\alpha}$ for $1-\alpha = .95$). Based on this more stringent criterion, the three hinge estimator methods were located having empirical coverage probabilities closest to .95. Specifically, it was found that HQ1, HSK2, and HSK5 had, respectively, the highest number of top three rankings: 332, 301, and 305. Accordingly, from the set of seven hinge estimator ES interval estimation procedures, any one of these three methods are recommended. Keselman et al. (in press) also recommended these three procedures for comparing treatment group trimmed means. Furthermore, the results suggest that, in general, one needs to have group sizes larger than 20 and that one can obtain good coverage with as little as 15% total trimming. The reader should remember however, that the differences between the empirical probabilities among these methods generally occurred in the third decimal place, and therefore, as stated, any of the seven hinge estimator approaches to setting an interval around the PES would be satisfactory, and in particular, much better than the usual approach of setting an interval around the nonrobust PES.

It was also found that *a priori* symmetric trimming provided very accurate probability coverage. All empirical coverage probabilities were within the Bradley (1978) liberal interval. Based on the summary values presented in Table 1, one can also note that the average probabilities are very tightly bunched around the target value of .95. Additionally, it is worth noting that, on average, researchers can obtain a very precise interval when adopting 5% symmetric trimming. Accordingly, the choice between *a priori* fixed trimming and asymmetric trimming methods might rest on ones comfort quotient for fixing the trimming rate prior to an examination of the data versus letting the data determine whether data should be trimmed in each tail of the data distribution and by what amount.

The comments provided by Keselman et al. (in press) regarding the choice of a best method of analysis are echoed. First, it needs to be repeated that no one method will be universally best. It could be that, at times, probability coverage for the classical method (i.e., Cohen's ES statistic) could provide a reasonable CI for ES. And as Wilcox and Keselman (2003) had noted, there is no way of knowing *a priori* which approach will be best. As they recommend, one could compute both approaches, that is, the classical approach and one of the robust methods enumerated in this paper. When the conclusions are the same, one can be comfortable with this common finding, otherwise, a robust approach to setting a CI for ES is recommended.

Keselman et al. noted that researchers should always carefully examine graphs of their data before proceeding with a particular method of analysis. Indeed, as many others have previously noted, a careful examination of outlying values can provide researchers with insights into the phenomenon under investigation.

It is reiterated that the parameter $\delta$ has a serious shortcoming because it is defined by using the usual population mean and standard deviation. These least squares parameters are not robust. While there are several criteria for assessing robustness of a parameter: qualitative robustness, quantitative robustness, and infinitesimal robustness (see Wilcox, 2005, Section 2.1 for a description of these criteria), the general notion is that *a parameter is not robust if a small change in the population distribution can strongly affect the parameter*. It can be shown that the least squares mean and variance are not robust (see, for example, Staudte and Sheather, 1990) when judged by any one of these three criteria. Accordingly, many authors, including us, subscribe to the position that inferences pertaining to robust parameters are more valid than inferences pertaining to the usual least squares parameters when dealing with populations that are nonnormal (e.g., Hampel, Ronchetti, Rousseeuw & Stahel, 1986; Huber, 1981; Staudte & Sheather, 1990; Wilcox & Keselman, 2003).

By itself, Cohen's $\delta$, or any other ES (i.e., $\delta_R$) for that matter, has little value in assessing whether or not a mean difference is large or small. What is required is experience in applying the ES. For example, as part of a review of the power of studies in abnormal and social psychology, Cohen (1962) suggested 0.25, 0.50, and 1.00 as small, medium, and large $\delta$s, respectively. In defense of these values, Cohen argued that the values "were chosen to seem reasonable." (p. 146) and cited three research studies on group differences in IQ research as justification for these guidelines. Cohen was clearly aware of the provisional nature of these guidelines and subsequently (Cohen, 1969) modified the guidelines to 0.20, 0.5, and 0.80, as small, medium, and large $\delta$s, respectively, and again emphasized that he regarded these to be reasonable based on his experience with research in the behavioral sciences. Cohen's guidelines, and his justification for them, illustrate an important point: Understanding of an ES measure will increase through experience with that measure.

## References

Algina, J., & Keselman, H. J. (2003, May). *Confidence intervals for Cohen's effect size.* Paper presented at a conference in honor of H. Swaminathan, University of Massachusetts, Amherst.

Algina, J. Keselman, H. J., & Penfield, R. D. (2005). 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.

American Psychological Association (2001). *Publication manual of the American Psychological Association* (5th ed.). Washington, DC: Author.

Babu, J. G., Padmanabhan A. R., & Puri, M. P. (1999). Robust one-way ANOVA under possibly non-regular conditions. *Biometrical Journal, 41*(1978), 321-339.

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

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology, 65,* 145-153.

Cohen, J. (1965). Some statistical issues in psychological research. In B. B. Wolman (Ed.), *Handbook of clinical psychology* (pp. 95-121). New York: Academic Press.

Cohen, J. (1969). *Statistical power analyses for the behavioral sciences*. New York: Academic Press.

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.

Fidler, F., Thomason, N., Cumming, G., Finch, S., & Leeman, J. (2004). Editors can lead researchers to confidence intervals, but can't make them think. *Psychological Science, 15,* 119-126.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., & Stahel, W. A. (1986). *Robust statistics*. New York: Wiley.

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

Hill, M., & Dixon, W. J. (1982). Robustness in real life: A study of clinical laboratory data. *Biometrics, 38,* 377-396.

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

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

Hogg, R. V. (1982). On adaptive statistical inferences. *Communications in Statistics: Theory and Methods, 11,* 2531-2542.

Huber, P. J. (1970). Studentizing robust estimates. In M. L. Puri (Ed.), *Nonparametric techniques in statistical inference*. London: Cambridge University Press.

Huber, P. J. (1972). Robust statistics: A review. *Annals of Mathematical Statistics, 43,* 1041-1067.

Huber, P. J. (1977). Discussion. *The Annals of Statistics, 5,* 1090-1091.

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

Keselman, H. J., Wilcox, R. R., Lix, L. M., Algina, J., & Fradette, K. H. (in press). Adaptive robust estimation and testing. *British Journal of Mathematical and Statistical Psychology*.

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.

Mudholkar, A., Mudholkar, G. S., & Srivastava, D. K. (1991). A construction and appraisal of pooled trimmed-t statistics. *Communications in Statistics: Theory and Methods, 20,* 1345-135.

Murphy, K. R. (1997). Editorial. *Journal of Applied Psychology, 82,* 3-5.

Reed III, J. F. (1998). Contributions to adaptive estimation. *Journal of Applied Statistics, 25,* 651-669.

Reed III, J. F., & Stark, D. B. (1996). Hinge estimators of location: Robust to asymmetry. *Computer Methods and Programs in Biomedicine, 49,* 11-17.

Rosenberger, J. L., & Gasko, M. (1983). Comparing location estimators: Trimmed means, medians and trimean. In D. Hoaglin, F. Mosteller, & J. Tukey (Eds.), *Understanding robust and exploratory data analysis* (pp. 297-336). New York:.Wiley.

Rothman, K. J. (1975). Computation of exact confidence intervals for the odds ratio. *International Journal of Bio-Medical Computing, 6,* 33-39.

Rothman, K. J. (1978a). A show of confidence. *New England Journal of Medicine, 299,* 1362-1363.

Rothman, K. J. (1978b). Estimation of the confidence limits for the cumulative probability of survival in life table analysis. *Journal of Chronic Disease, 31,* 557-560.

SAS Institute Inc. (1999). *SAS/STAT user's guide, Version 7,* 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 and J. H. Steiger (Eds.), *What if there were no significance tests?* Hillsdale, NJ: Lawrence Erlbaum.

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

Thompson, B. (1994). Guidelines for authors. *Educational and Psychological Measurement, 54*, 837-847.

Tukey, J. W., & McLaughlin, D. H. (1963). Less vulnerable confidence and significance procedures for location based on a single sample: Trimming/Winsorization. *Sankhya, Series A, 25*, 331-352.

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

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

Wilcox, R. R. (2005). *Introduction to robust estimation and hypothesis testing* (2nd Ed.). San Diego: Elsevier 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. and the Task force on Statistical Inference Statistical methods in psychology journals. (1999). *American Psychologist, 54*, 594-604.

## Appendix 1

One question that might be asked about $\delta_R$ is whether it is necessary to multiply

$$\delta_R = \frac{\mu_{t2} - \mu_{t1}}{\sigma_W}$$

by .643 to obtain a robust parameter. The answer is, of course, no. When the multiplier is not used, the difference between the trimmed means is divided by the Winsorized standard deviation. By contrast, when using the multiplier, the difference between the trimmed means is divided by a rescaled Winsorized standard deviation $(\text{i.e., } \sigma_W/.643)$.

The same multiplier would be applied to the sample ES and, as a result, *regardless of whether the multiplier is used, coverage probability is the same.* Therefore, our results have relevance to researchers who prefer to include the multiplier and researchers who prefer to exclude the multiplier. Incorporating the multiplier requires a different value for different levels of trimming. The multipliers for 10%, 15%, and 25% trimming would be $1/\sqrt{.824}$, $1/\sqrt{.734}$, $1/\sqrt{.537}$, respectively.

## Appendix 2

Huber (1972) and Hogg (1974) noted that the best way of conceptualizing the unknown parameter $\theta(\alpha_1,\alpha_1)$ is that it is the population counterpart of $m(\alpha_1,\alpha_1)$. Hogg (1974, p. 920) indicated that in the one-sample case the statistic $[m(\alpha_1,\alpha_2)-\theta(\alpha_1,\alpha_2)]/s_{m(\alpha_1,\alpha_2)}$ has an approximate t-distribution with $h-1$ degrees of freedom if trimming is reasonably symmetric about the mode of a unimodal skewed distribution. Moreover, he noted that, even for fairly skewed situations, the distribution of this statistic will "probably be closer to this approximating distribution than the ratio $[m(\alpha)-\theta]/s_{m(\alpha)}$, which is the statistic based on a symmetrically trimmed mean. (p. 920)".

Journal of Modern Applied Statistical Methods
November 2005, Vol. 4, No. 2, 372-393

1538 – 9472/05/$95.00

# A Single, Powerful, Nonparametric Statistic for Continuous-data Telecommunications Parity Testing

J.D. Opdyke
DataMineIt
Marblehead, MA

Since the enactment of the Telecommunications Act of 1996, extensive expert testimony has justified use of the modified *t* statistic (Brownie et al., 1990) for performing two-sample hypothesis tests comparing Bell companies' CLEC and ILEC performance measurement data (known as parity testing). However, Opdyke (Telecommunications Policy, 2004) demonstrated this statistic to be potentially manipulable and to have literally zero power to detect inferior CLEC service provision under a wide range of relevant data conditions. This article develops a single, nonparametric statistic that is easily implemented (i.e., not computationally intensive) and typically provides dramatic power gains over the modified *t* while simultaneously providing much better Type I error control. The statistic should be useful in a wide range of quality control settings.

Key words: Telecommunications Act, ILEC, CLEC, Location-scale, Mean-variance, Maximum test

## Introduction

The major goal of the Telecommunications Act of 1996, the most sweeping communications-related public policy to be enacted by Congress in over half a century (since the Telecom Act of 1934 – see http://www.fcc.gov/telecom.html) has been to deregulate local telephone service in the United States, making it a fully competitive economic market. To accomplish this, the Act takes a carrot-stick approach: it allows the Bell companies (the incumbent local exchange carriers, or ILECs, now only BellSouth, Qwest, SBC, and Verizon) to enter into the previously deregulated long distance market, something they had been prohibited from doing because of their status as government regulated monopolies. This provides ILECs with the potentially lucrative opportunity to provide one-stop shopping telephone service to their customers, bundling all of their clients' telecommunications needs into a single package from a single service provider.

In return for this carrot, the Act's stick requires that the ILECs first must do two things: (a) allow their competitors (competitive local exchange carriers, or CLECs, the large long distance telephone companies like Sprint, as well as numerous smaller companies) access to and use of their networks, in some cases to resell services at discounted wholesale rates, and (b) provide the CLECs' customers with service "at least equal in quality to" the service they provide to their own customers (Telecommunications Act of 1996, Pub. LA. No. 104-104, 110 Stat. 56 (1996), at §251 (c) (2) (C); and see §251 (c) (2) (B) for the 14 point "COMPETITIVE CHECKLIST" of conditions that ILECs must satisfy to meet the at-least-equal

J.D. Opdyke is President of DataMineIt, a statistical data mining consultancy specializing in the banking and credit, telecommunications, retail and pricing, and advertising and marketing sectors (JDOpdyke@DataMineIt.com, www.DataMineIt.com). I owe special thanks to Geri S. Costanza, M.S., for numerous and valuable insightful discussions.

service provision standard). This at-least-equal service provision is the necessary enforcement mechanism for ensuring that network access (a) occurs in a meaningful way that truly promotes the goal of market competition.

To explain by way of example, if it takes a week on average for a CLEC customer to have a line installed or repaired by the ILEC, but only a day on average for an ILEC customer to receive the same service, no customers would ever switch from the ILEC to any of the CLECs, and markets could never become competitive. The mechanism for properly enforcing the at-least-equal service provision depends on the appropriate utilization of the extensive operations support services (OSS) performance measurement data that ILECs record when providing service to both CLEC and ILEC customers (e.g., how fast is a phone line installed; how fast is a line repaired; how often are repairs made within a certain number of days or by a preset due date, etc.). This utilization has taken the form of monthly statistical parity testing – applying statistical tests to the monthly CLEC and ILEC service data to compare the two groups and make sure that service is, in fact, at least equal for CLEC customers (i.e., in parity).

The specific statistical tests used in OSS parity testing depend on a number of factors, and foremost among these are the hypotheses being tested. The appropriate null and alternate hypotheses for OSS parity testing are listed below (1), in terms of both average service (the mean) and the variability of the service provided (the variance) (see Opdyke, 2004, p. 3-4, for a detailed explanation of why precisely these hypotheses are required in this setting).

$$\begin{aligned} &\text{Ho: } \mu_C \le \mu_I \text{ AND } \sigma_C^2 \le \sigma_I^2 \\ &\qquad\qquad \text{vs.} \\ &\text{Ha: } \mu_C > \mu_I \text{ OR } \sigma_C^2 > \sigma_I^2 \end{aligned} \qquad (1)$$

A statistical test of this pair of joint hypotheses will determine, with a specified level of certainty, whether service to CLEC customers takes no longer *on average* than service to ILEC customers (i.e., $\mu_C \le \mu_I$), and whether the variability of this service is no larger than that characterizing the service provided to ILEC customers (i.e., $\sigma_C^2 \le \sigma_I^2$) (see the FCC's Notice of Proposed Rulemaking, 04/16/98, APPENDIX B, p.B2, for some of the early impetus for testing both means and variances). If the statistical test determines, with a specified level of certainty, that both of these conditions hold, service is deemed to be at least equal, or in parity. If either condition is determined, with a specified level of certainty, to be violated, then service is considered out of parity, or in disparity.

Findings of disparity carry consequences for the ILEC(s) in the form of fines paid to the CLECs, and sometimes to the relevant state(s). These fines, or remedies, can be large (US$ millions), and extensive and/or prolonged findings of disparity can lead to revocation of an ILEC's approval to provide long distance service. Therefore the choice of appropriate, if not the best statistics for OSS parity testing is very important, not only for the individual firms involved, but also for the entire industry. And of course, the best statistics simply are those that, under a classical Neyman-Pearson hypothesis-testing paradigm, are most powerful under the widest range of relevant data conditions, given robust and reasonable Type I error control.

In addition to the hypotheses being tested, the type of data being compared determines which statistical tests can and should be used. Telecommunications OSS performance metrics contain three types of data, and each is listed below with an example of a corresponding performance metric:

- *binary data* – the percentage of repairs completed on time, or within a certain number of days
- *count data* – the number of troubles on a telephone line within a specified time period
- *continuous data* – the average time it takes to install a phone line

For continuous data metrics, the modified *t* (Brownie et al., 1990) has been supported in extensive expert testimony proffered by both CLECs and ILECs, as well as in Opinions and Rulings by various regulatory bodies, as an appropriate statistic to test the relevant joint hypotheses above (see Opdyke, 2004, for extensive citations; all but one of the four major ILEC performance and remedy plans nationwide utilizes the modified *t* as a primary test statistic).

$$t_{\text{mod}} = \frac{\left(\bar{X}_C - \bar{X}_I\right) - \left(\mu_C - \mu_I\right)}{s_I \sqrt{\frac{1}{n_I} + \frac{1}{n_C}}} \quad (2)$$

where

$$s_I = \sqrt{\frac{\sum_{i=1}^{n_I}\left(X_{I_i} - \bar{X}_I\right)^2}{\left(n_I - 1\right)}}, \quad \bar{X}_I = \frac{\sum_{i=1}^{n_I} X_i}{n_I}, \quad \bar{X}_C = \frac{\sum_{i=1}^{n_C} X_i}{n_C},$$

and degrees of freedom ($df$) = $n_I - 1$.

However, Opdyke (2004) demonstrated, via an extensive simulation study and an analytic derivation, that because the modified *t* follows neither the standard normal nor the student's *t* distribution as previously surmised in seven years' of expert testimony (see Opdyke, 2004, for extensive citations), it *potentially* remains vulnerable to what has been termed gaming – intentional manipulation of its score to effectively mask disparity. But far more importantly, the modified *t* also was shown to be virtually powerless to detect inferior CLEC service provision under a wide range of relevant data conditions (i.e., larger CLEC variability under equal or better average service).

Instead, Opdyke (2004) proposed the collective use of several other easily-implemented statistical procedures that typically provide dramatic power gains over the modified *t*. Selection of a specific statistic among those proposed depends on the relative sizes of the two samples being compared, and on whether the particular performance metric being tested is long-tailed or short-tailed (this is the distributional characteristic known as kurtosis). Years of OSS data now exist since the Act was passed to establish such distributional characteristics as population parameters, not as unknowns requiring an additional statistical test. However, even though the FCC itself identified "data distribution, sample size and other characteristics inherent in the data" (FCC NPRM, 11/08/01, p. 37) as factors relevant to the choice of the statistical tests used in parity testing, one expressed concern regarding Opdyke's (2004) approach is that the potential use of different statistics for different performance metrics (and sample sizes) is somehow too complex for implementation in parity testing.

This article addresses this concern by building on the results and recommendations of Opdyke (2004) to develop a single, nonparametric, and generally powerful statistic for use with all continuous–data performance metrics. As shown below, the proposed statistic 1) maintains reasonable Type I error control; 2) is always either nearly as powerful as Opdyke's (2004) multiple procedures, or almost as often, even more powerful; 3) typically provides dramatic power gains over the modified *t*; 4) is easily implemented and not computationally intensive; and 5) should be widely applicable and useful in other quality control settings as well.

## Methodology

### Previously Developed Alternatives to the modified *t*

Under the data conditions relevant to OSS parity testing, Opdyke (2004) found that conditional statistical procedures combining either O'Brien's (1988) generalized *t* test (OBt) or his generalized rank sum test (OBG) with either of two straightforward tests of variances (Shoemaker's, 2003, $F_1$ test, or the modified Levene test of Brown and Forsythe, 1974) were by far the most powerful procedures of the over twenty statistics that were studied. Their combined use is conditioned on the relative sizes of the two sample means, as shown below:

Table 1. Conditional Statistical Procedures, Opdyke (2004)

| Conditional statistical procedure | if $\bar{X}_C > \bar{X}_I$, use… | If $\bar{X}_C \leq \bar{X}_I$ or OB fails to reject Ho:, use… |
|---|---|---|
| OBtShoe | OBt | Shoemaker's $F_1$ |
| OBtLev | OBt | modified Levene |
| OBGShoe | OBG | Shoemaker's $F_1$ |
| OBGLev | OBG | modified Levene |

(*Note*: see Appendix for the calculation of these statistics)

Conditioning on the sample means as shown in Table 1 inflates the size of these tests, so an ad hoc p-value adjustment of p-value = (5/3 * p-value) was used to maintain Type I error control (see Opdyke, 2004, for details). Even after such an adjustment, these tests maintain reasonable, if not impressive power under normal and short-tailed (uniform) data, and somewhat less power under

long-tailed (double exponential) data, although still far more power than the modified *t* under most of these conditions (Opdyke, 2004, p. 20-26).

The conditions under which each of these four tests is most powerful and should be used are summarized in Table 2 below. Notably skewed data, however, first should be transformed, as required by one of the largest state PUCs and strongly endorsed by another of the largest state PUCs (CPUC Interim Opinion, 2001, Appendix J; CPUC Opinion (2002), Appendix J, Exhibit 3 p.2-3; Before the Texas PUC – SBC Testimony, Dysart & Jarosz, 2004; and for optional use with some metrics, SBC Comments, 2002, p.48, 56).

Unfortunately, all of the statistics examined for or used in OSS parity testing suffer from sometimes severe erosions in power under skewness (see Opdyke, 2004, for relevant simulation results; The California Public Utilities Commission also addresses this issue – CPUC Interim Opinion, 2001, p. 112-115, 136, 142, 145, & Appendix J, and CPUC Opinion, 2002, p. 74, 84, & Appendix J). Because these metrics are widely cited as being lognormal (which is typically highly skewed – see CPUC Interim Opinion, 2001, Appendix J, and MCI Worldcom's Performance Assurance Plan: The SiMPL Plan, by George S. Ford, Ph.D., p.5), a logarithmic transformation toward symmetry should provide at least some needed power to detect disparity without, in all practicality, causing distortions in the comparison of CLEC and ILEC service provision.

Table 2. Conditional Statistical Procedures, Opdyke (2004)

| Sample sizes | | Normal & Short-tailed | Long-tailed | Skewed |
|---|---|---|---|---|
| | | OBt | OBG | |
| Bal. | Shoe | OBtShoe | OBGShoe | Transform |
| Unbal. | Lev | OBtLev | OBGLev | Transform |

Once transformed (if necessary), the performance metric is tested with one of the four combined procedures listed in Table 2. This is clear-cut if the sample sizes and distributional characteristics of the metrics being tested unambiguously fall neatly into these four cells (for example, if a metric is at least as short-tailed as the normal distribution, kurtosis = 3, and has very unbalanced sample sizes, use OBtLev).

However, further simulations that parallel those of Opdyke (2004) are required to determine the tipping points defining exactly when to use each of these four statistics. Although these tipping point simulations would be straightforward to perform, one expressed concern about the use of Table 2 is that, the FCC's advisory comment notwithstanding, having to (potentially) use different tests under different sample size and data conditions is somehow too complex for the implementation of parity testing. Although implementing Table 2 is far less complicated than at least one of the four major OSS performance and remedy plans (the BellSouth 'truncated Z' plan, which one FCC economist only half-jokingly refers to as "the balanced averaged disaggregated truncated adjusted modified Z plan", Shiman, 2002, p.283), it unarguably would be preferable if, all else equal (or close), one statistic could accomplish what the conditional use of the multiple statistics in Table 2 does. This is the motivation for this paper, and the development of the statistic presented below.

## A Single Statistic for Continuous-data Parity Testing

Maximum tests – statistics whose scores (p-values) are the maximum (minimum) of two or more other statistics – have been devised and studied in a number of settings in the statistics literature with very favorable results. Neuhäuser et al. (2004) favorably compares a maximum test for the non-parametric two-sample location problem to multiple adaptive tests, finding the former to be most powerful under the widest range of data conditions.

Blair (2002) constructed a maximum test of location that is shown to be only slightly less powerful than each of its constituent tests under their respective ideal data conditions, but notably more powerful than each under their respective non-ideal data conditions (for additional studies using maximum tests, see Fleming & Harrington, 1991, Freidlin & Gastwirth, 2000a, 2000b, Freidlin et al., 2002, Lee, 1996, Ryan et al., 1999, Tarone, 1981, Weichert & Hothorn, 2002, Willan, 1988, & Yang et al., 2005). These findings demonstrate the general purpose of maximum tests – to trade-off minor power losses under ideal data

conditions for a more robust statistic with larger power gains across a wider range of possible (and usually unknown) data distributions.

Although the relevant characteristic of the distributions of continuous-data OSS performance metrics is, for all intents and purposes, known because so many years of data now exist to establish the kurtosis as a population parameter and not a statistical estimate based on samples, a maximum test still could be useful here for several reasons: 1) using only one statistical test unarguably would be more straightforward to implement than (potentially) relying on the four statistics in Table 2 and choosing between them based on a matrix of sample sizes and performance metric kurtoses; 2) the expected power losses compared to Opdyke's (2004) individual tests may be small or negligible; and 3) under some conditions, depending on the constituent tests used, the maximum statistic may be even more powerful than those tests recommended in Opdyke (2004) and shown in Table 2.

To construct a maximum test here, it must be recognized that maximum tests are conditional statistical procedures, and the additional variance introduced by such conditioning will inflate the test's size over that of its constituent statistics (and if left unadjusted, probably over the nominal level of the test as shown in Blair, 2002). But the constituent statistics in Table 2 are already conditional statistical procedures. Consequently, the ad hoc p-value adjustment used below for the purpose of maintaining validity must be large enough to take this double conditioning into account (this actually is triple conditioning because O'Brien's tests themselves are conditional statistical procedures). The adjustment is simply a multiplication of the p-values by constant factors ($\beta$'s). The p-value of the maximum test – OBMax – is defined in (2):

$$p_{OBMax} = \min \begin{pmatrix} p_{OBtShoe} \cdot \beta_{OBtShoe}, \\ p_{OBtLev} \cdot \beta_{OBtLev}, \\ p_{OBGShoe} \cdot \beta_{OBGShoe}, \\ p_{OBGLev} \cdot \beta_{OBGLev}, \\ p_{tsv} \cdot \beta_{tsv}, \\ 1.0 \end{pmatrix} \quad (3)$$

where

$$\beta_{OBtShoe} = \beta_{OBtLev} = \beta_{OBGShoe} = \beta_{OBGLev} = 2.8,$$

and $\beta_{tsv} = 1.8$, and $p_{tsv}$ is the $p$-value corresponding to the separate-variance $t$ test with Satterthwaite's (1946) degrees of freedom (see Appendix for corresponding formulae). Under the relevant data conditions, the behavior of OBMax is compared to that of its constituent tests and the modified $t$ test in the simulation study described below. It is also compared with two other maximum tests – OBMax3 and TVMax – as defined in (2) and (3) below (TVMax for $t$ test, Variance tests, and Maximum test).

$$p_{OBMax3} = \min \begin{pmatrix} p_{OBtLev} \cdot \beta_{OBtLev}, \\ p_{OBtShoe} \cdot \beta_{OBtShoe}, \\ p_{tsv} \cdot \beta_{tsv}, \\ 1.0 \end{pmatrix} \quad (4)$$

where $\beta_{OBtLev} = \beta_{OBtShoe} = 3.0,$ and $\beta_{tsv} = 1.6$

$$p_{TVMax} = \min \begin{pmatrix} p_{modLev} \cdot \beta_{modLev}, \\ p_{ShoeF_1} \cdot \beta_{ShoeF_1}, \\ p_{tsv} \cdot \beta_{tsv}, \\ 1.0 \end{pmatrix} \quad (5)$$

where $\beta_{modLev} = \beta_{ShoeF1} = 3.0,$ and $\beta_{tsv} = 1.6$

Although preferable to ad hoc adjustments based on simulations, analytic derivation of the asymptotic distribution of OBMax, and maximum tests in general, is non-trivial, as Yang et al. (2005) show under even stronger distributional assumptions than can be made with respect to the Table 1 statistics. Derivation of the asymptotic distribution of OBMax is the topic of continuing research (Opdyke, 2005).

Level and Power Simulation Study

The level and power simulations in this article parallel those conducted in Opdyke (2004). Eleven tests were studied: each of the four conditional statistical procedures listed in Table 1 – OBtShoe, OBtLev, OBGShoe, and OBGLev; the separate-variance $t$ test (with Satterthwaite's, 1946, degrees of freedom – $df$) (tsv); the modified $t$ test (with $df = n_1 - 1$, as in Brownie et al., 1990, Comments of SBC, 2002, p.57, and CPUC Opinion, 2001,

Appendix C, p. 2.) (tmod); OBMax as defined above in (1); OBMax3 and TVMax as defined above in (2) and (3), respectively; and two tests of stochastic dominance described below. All of the conditional statistics using O'Brien's (1988) tests are referenced to the *F* distribution, rather than Blair's (1991) critical values, even though doing so would normally violate the nominal level of the test under some conditions, because the p-value adjustment used here explicitly takes this size inflation into account (see Opdyke, 2004, 2005, for further details).

The data was generated from the normal, uniform, double exponential, and lognormal distributions for four different pairs of sample sizes ($n_C = n_I = 30$; $n_C = 30$ & $n_I = 300$; $n_C = 30$ & $n_I = 3000$; and $n_C = n_I = 300$), seven different variance ratios ($\sigma_C^2 / \sigma_I^2 =$ 0.50, 0.75, 1.00, 1.25, 1.50, 1.75, 2.00), and seven different location shifts

$$\left( \begin{array}{c} \mu_C = \mu_I - 2\sigma_I,\ \mu_I - \sigma_I,\ \mu_I - 0.5\sigma_I,\ \mu_I,\ \mu_I + 0.5\sigma_I, \\ \mu_I + \sigma_I,\ \mu_I + 2\sigma_I \end{array} \right),$$

making 784 scenarios. N = 20,000 simulations were run for each scenario, except for scenarios with $n_C = 30$ & $n_I = 3000$, which used N = 5,000.

The normal distribution was chosen as a universal basis for comparison; the uniform and double exponential distributions were chosen as examples of short-tailed and long-tailed distributions, respectively, to examine the possible effects of kurtosis on the tests; and the lognormal distribution was chosen to examine the possible effects of skewness on the tests, and because continuous data OSS performance metrics have been cited widely as often being approximately lognormal. $n_C = n_I = 30$ was chosen because many performance and remedy plans require or allow for the use of permutation tests if at least one of the two samples has less than 30 observations (see The Qwest Performance Assurance Plan, Revised 11/22/2000, p.4-5; SBC Comments, 2002, p. 55, and 13 state Performance Remedy Plans – Attachment 17, p.4-5; and Performance Assurance Plan – Verizon New York Inc., Redlined Version January 2003, Appendix D, p.3-4.), and $n_C = n_I =$ 300 was chosen to examine rates of convergence under equal sample sizes (Pesarin's, 2000, combined permutation test, however, appears to have greater power for the relevant joint hypotheses here than the naïve Monte Carlo permutation test currently implemented by these performance and remedy plans, and at least two companies produce preprogrammed software that automatically performs this test – DataMineIt, http://www.DataMineIt.com, and Methodologica, http://www.methodologica.it/npctest.html).

The extremely unbalanced sample size pairs of $n_C = 30$ & $n_I = 300$ and $n_C = 30$ & $n_I = 3000$ were chosen because such large sample size ratios actually are not uncommon in OSS performance metric data. Also, the number of ILEC phone lines and customers typically dwarf those corresponding to most individual CLECs. Thus, it is important to test the behavior of these statistics under these extreme conditions, even though most simulation studies would focus on smaller and/or more balanced sample sizes. $n_C$ is very rarely, if ever, larger than $n_I$ and thus, only cases involving ($n_I$ / $n_C$) $\geq 1.0$ were examined in this study (Opdyke, 2005, examines $n_I < n_C$ also). Two nominal levels were used for all the simulations: $\alpha = 0.05$ and $\alpha =$ 0.10, bringing the total number of scenarios to 1,568. These two levels bracket the vast majority of the levels used in OSS parity testing. (SBC Comments, 2002, p.49-52; CPUC Opinion, 2002, Appendix J, Exhibit 3, p.4; and Performance Assurance Plan – Verizon New York Inc., Redlined Version January 2003, Appendix D, p.1).

Two other tests also were included in the simulations: Rosenbaum's (1954) test, which counts the number of observations in one sample beyond the maximum of the other as a test of $H_o$: $F(x) \equiv G(x)$ against the alternative of stochastic dominance; and the (one-sided) Kolmogorov-Smirnov statistic (using Goodman's, 1954, Chi-square approximation – see Siegel & Castellan, 1988, p.148), for a non-parametric test of $H_o$: $F(x) \equiv G(x)$ against general (one-sided) alternatives. Although neither is designed specifically to test the joint hypotheses relevant to the OSS parity testing setting, and thus may have less power, they are included for several reasons: (1) as a basis for comparison to the other tests; (2) because researchers often turn to these types of tests when confronted with the joint hypotheses relevant to the parity testing context and examined in this simulation study; and (3) because the Kolmogorov-Smirnov statistic has been described as being "able to detect not only differences in average but differences in dispersion between the two samples as well." (Matlack, 1980, p. 359).

## Results

This simulation study generated 11 x 1,568 = 17,248 level and power results, all of which are available from the author upon request in a Microsoft Excel® workbook (along with a SAS/GRAPH® program for convenient visualization). The key results are summarized in the tables and selected graphs below.

Under symmetry, the p-value adjustments used in OBMax as defined in (3) provide reasonable Type I error control for the relevant range of test levels; as shown in Table 3, violations of the nominal level are modest in size and infrequent (14 of 288 symmetric-data null hypothesis scenarios; violations occur if the observed level is equal to or greater than the one-tailed 95% critical value of the simulation, based on the common Wald approximation of the binomial distribution to the normal distribution, which is very accurate for such large numbers of simulations and $\alpha \geq 0.05$ – see Evans et al., 1993, p. 39, and Cochran, 1977, p. 58).

Even better level control is possible by increasing the adjustment factors – say, by increasing the OB $\beta$'s from 2.8 to 3.0 – but the price paid for this is a loss of power. The adjustment factors used – 2.8 for the OB tests and 1.8 for the separate-variance *t* test – are reasonable as they produce relatively minor level violations, and relatively minor power losses when OBMax is compared to its constituent tests. However, nearly as often as not, OBMax actually provides power *gains* over the conditional use of the Table 2 statistics (graphs of these comparisons are available from the author upon request). OBMax's largest power loss is only slightly over 0.10, and these minor power losses typically occur under simultaneously small CLEC samples, large CLEC variance increases, and *decreases* in the CLEC mean (relative to the ILEC mean).

Its largest power gain, however, exceeds 0.2, and these power gains occur under simultaneously small CLEC samples, typically equal or smaller CLEC variances, and small *increases* in the CLEC mean. The reason for this increased sensitivity to detect small location shifts is the inclusion of the separate-variance *t* test among the constituent tests of OBMax. Including this test mitigates power losses in the one fairly narrow range of conditions where the modified *t* test has a relatively slight, but still noticeable power advantage over the Table 2 constituent tests: for normal and short-tailed data, under simultaneously small CLEC samples, typically equal or smaller CLEC variances, and small increases in the CLEC mean. Including the separate-variance *t* test as a constituent test of OBMax shrinks this loss of power relative to the modified *t* (under only these fairly narrow conditions) typically by a factor of one half, so that the largest power loss remains less than 0.1 (Figure 3).

Far more important to note, however, is that under all other data conditions the power of OBMax is never less than that of the modified *t*, and typically dramatically larger (sometimes a gain of 1.0! - see Figures 3, 4, and 6). The power differences between OBMax and the modified *t* that are shown in Figure 3 are summarized in Table 4 below, although the Figures more accurately and thoroughly convey the story. Figures 5 and 6 show how dramatically OBMax dominates the modified *t* as sample sizes increase. This demonstration of the reasonable power of OBMax, under all symmetric alternatives, should dispel a) expressed concerns in this setting regarding the lack of power of composite tests of location and scale (Mallows, 2002, p. 260); b) admittedly premature conclusions in this setting about the lack of power of relevant rank-based tests (Mallows, 2002, p. 260), which is what the OBG tests are; and c) findings of less (and concerns of too little) power in this setting under unbalanced sample sizes (Gastwirth & Miao, 2002, p. 273).

Table 3. Symmetric Data Level Violations of OBMax

| $\sigma_C^2$ | $\mu_C$ | Sample sizes | Distribution | Nominal level of test (α) | Actual size |
|---|---|---|---|---|---|
| $\sigma_I^2$ | $\mu_I$ | $n_C = n_I = 30$ | Normal | 0.05 | 0.0578 |
| $\sigma_I^2$ | $\mu_I$ | $n_C = 30,\ n_I = 3000$ | Normal | 0.05 | 0.0532 |
| $\sigma_I^2$ | $\mu_I$ | $n_C = n_I = 300$ | Normal | 0.05 | 0.0561 |
| $\sigma_I^2$ | $\mu_I$ | $n_C = 300,\ n_I = 300$ | Uniform | 0.05 | 0.0546 |
| $\sigma_I^2$ | $\mu_I$ | $n_C = n_I = 30$ | Double exponential | 0.05 | 0.0574 |
| $\sigma_I^2$ | $\mu_I$ | $n_C = 30,\ n_I = 300$ | Double exponential | 0.05 | 0.0538 |
| $\sigma_I^2$ | $\mu_I$ | $n_C = 30,\ n_I = 3000$ | Double exponential | 0.05 | 0.0556 |
| $\sigma_I^2$ | $\mu_I$ | $n_C = n_I = 300$ | Double exponential | 0.05 | 0.0596 |
| $\sigma_I^2$ | $\mu_I$ | $n_C = n_I = 30$ | Normal | 0.10 | 0.1115 |
| $\sigma_I^2$ | $\mu_I$ | $n_C = n_I = 300$ | Normal | 0.10 | 0.1073 |
| $\sigma_I^2$ | $\mu_I$ | $n_C = n_I = 30$ | Uniform | 0.10 | 0.1048 |
| $\sigma_I^2$ | $\mu_I$ | $n_C = n_I = 300$ | Uniform | 0.10 | 0.1044 |
| $\sigma_I^2$ | $\mu_I$ | $n_C = n_I = 30$ | Double exponential | 0.10 | 0.1116 |
| $\sigma_I^2$ | $\mu_I$ | $n_C = n_I = 300$ | Double exponential | 0.10 | 0.1095 |

Not surprisingly, OBMax is very similar to OBMax3 and TVMax in terms of both Type I error control and power, except that, under small CLEC and large ILEC samples, OBMax has greater power than TVMax to detect slight CLEC location shifts, especially under leptokurtotic data (the largest power advantages are about 0.08, 0.10, and 0.14 for uniform, normal, and double exponential data, respectively). OBMax3 is more powerful than TVMax, exhibiting the same slight power loss compared to OBMax only under leptokurtotic data (where the largest loss is only about 0.08). Because OBMax is unambiguously more powerful, it is recommended over the other two tests under symmetry. Under asymmetry, however, OBMax violates the nominal level of the test under a specific combination of conditions, for which the OBG rank tests perform poorly (a. large and equal sample sizes; b. equal means; *and* c. a much smaller CLEC variance). Therefore if skewed data is not or cannot be reliably transformed toward symmetry for some reason, OBMax3 is one good alternative to OBMax. OBMax3 has slightly less power, but it always maintains validity, even under skewed data. In fact, it maintains validity far better than does the modified *t* under skewed data.

However, an even better alternative appears to be OBMax2, as presented in the preliminary results of Opdyke (2005). OBMax2 = OBMax3 if a) $s_C^2 \le s_I^2$, b) $\bar{X}_c \le \left(\bar{X}_I + 0.5 s_I\right)$, *and* c) the null hypothesis of symmetry is rejected by the test of D'Agostino et al. (1990) at α = 0.01; otherwise, OBMax2 = OBMax. OBMax2 maintains most of the power gains of OBMax over OBMax3, while also maintaining validity very well under skewed data – again, far better than does the modified *t*, as shown in Table 5 below (note that when $n_C > n_I$, which rarely if ever occurs with OSS data, all $\beta$'s for OBMax2 utilize an additional adjustment: $\beta_X = \beta_X + \min\left[2.5,\ \log_{2.7}\left(n_C / n_I\right)\right]$ – see Opdyke, 2005, for further details).

Figure 1. OBMax rejection rate: Empirical Level and Power (α = 0.05)

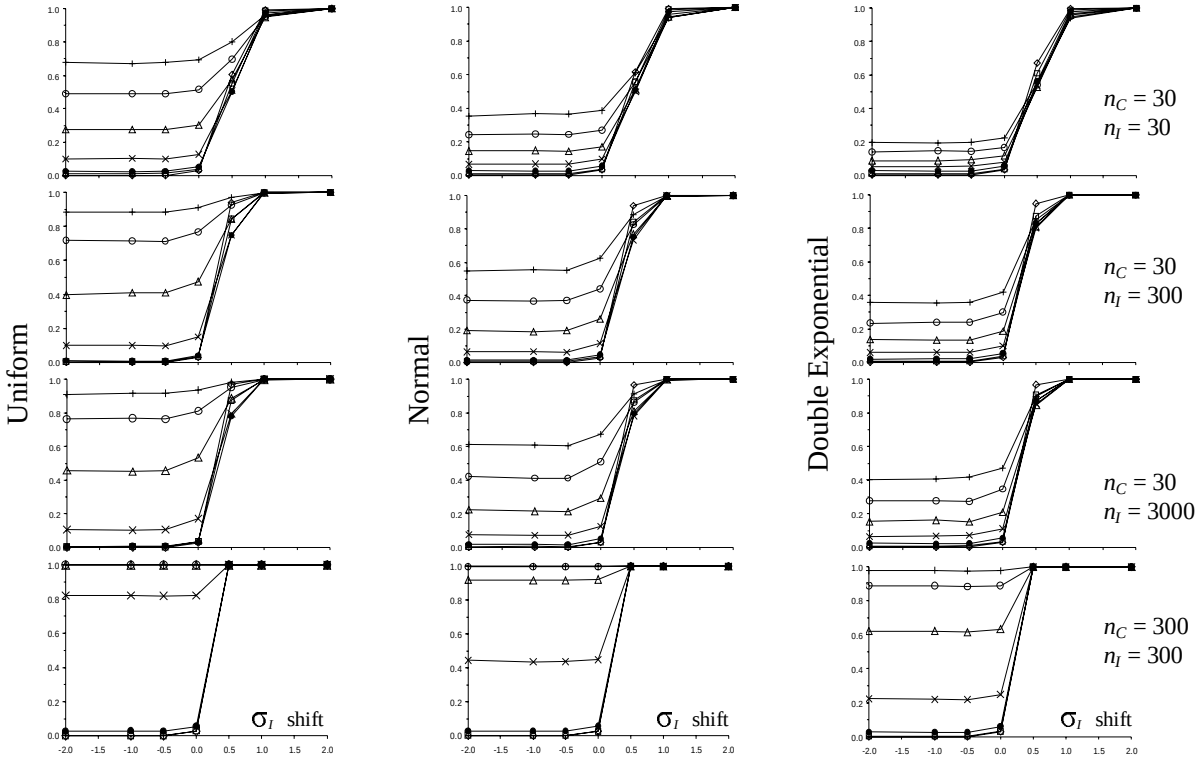



Figure 2. modified *t* rejection rate: Empirical Level and Power (α = 0.05)

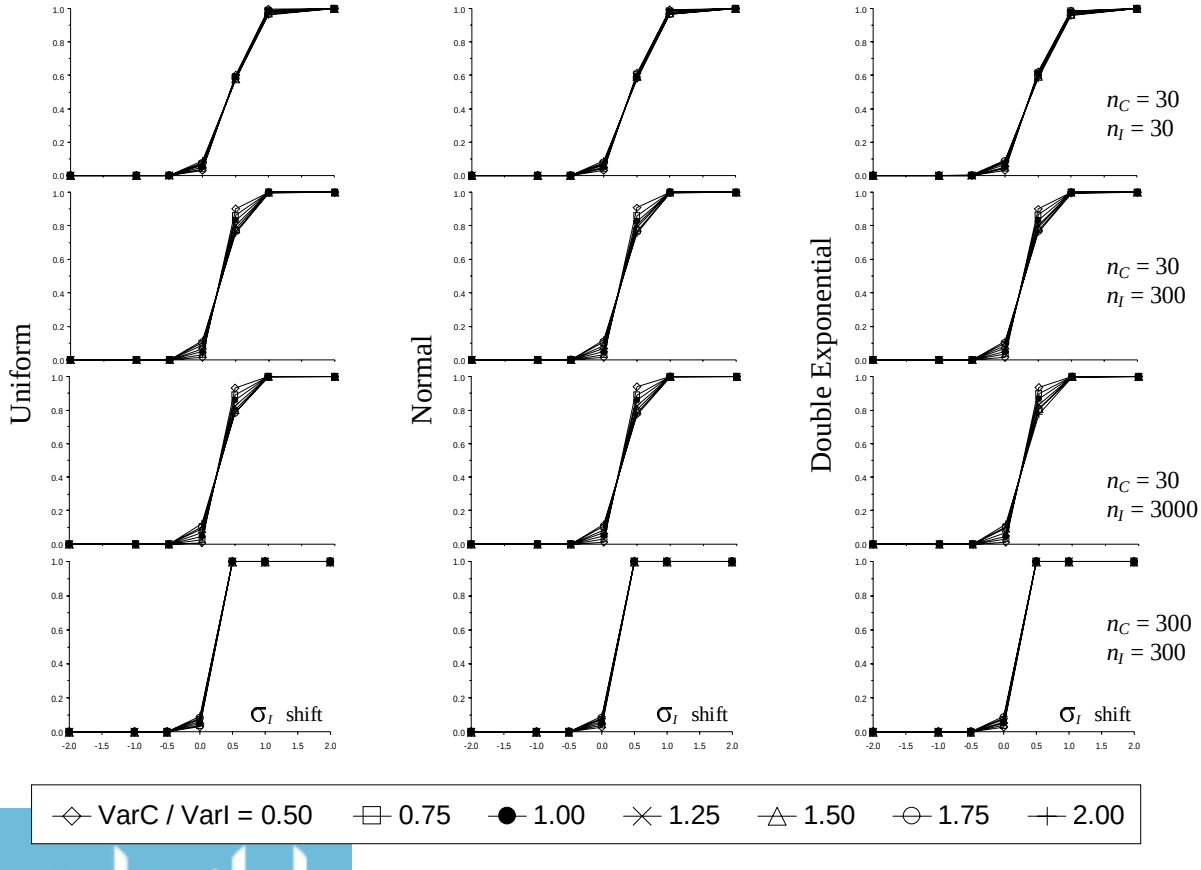


Figure 3. OBMax Power minus modified *t* Power ($\alpha = 0.05$)

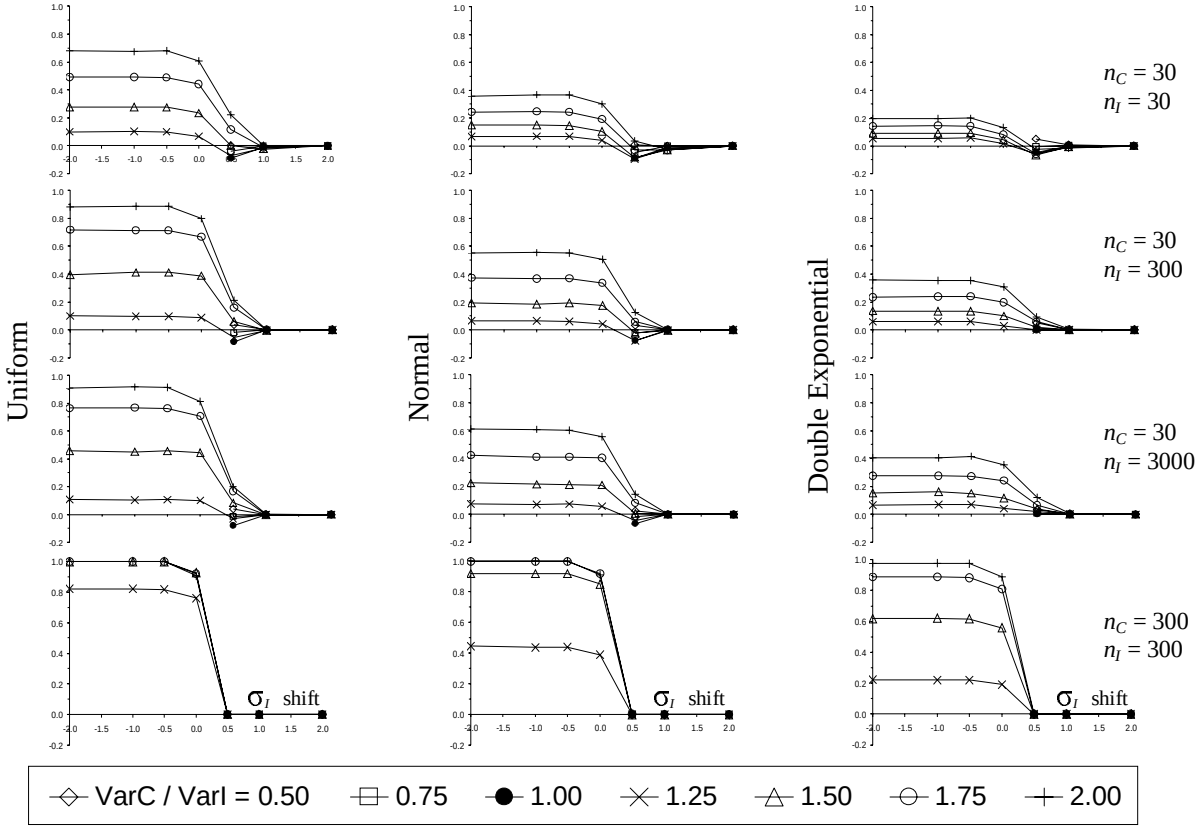



Figure 4. All Alternate Hypothesis Simulations with a Power Difference (309 of 444): OBMax Power minus modified *t* Power ($\alpha = 0.05$)

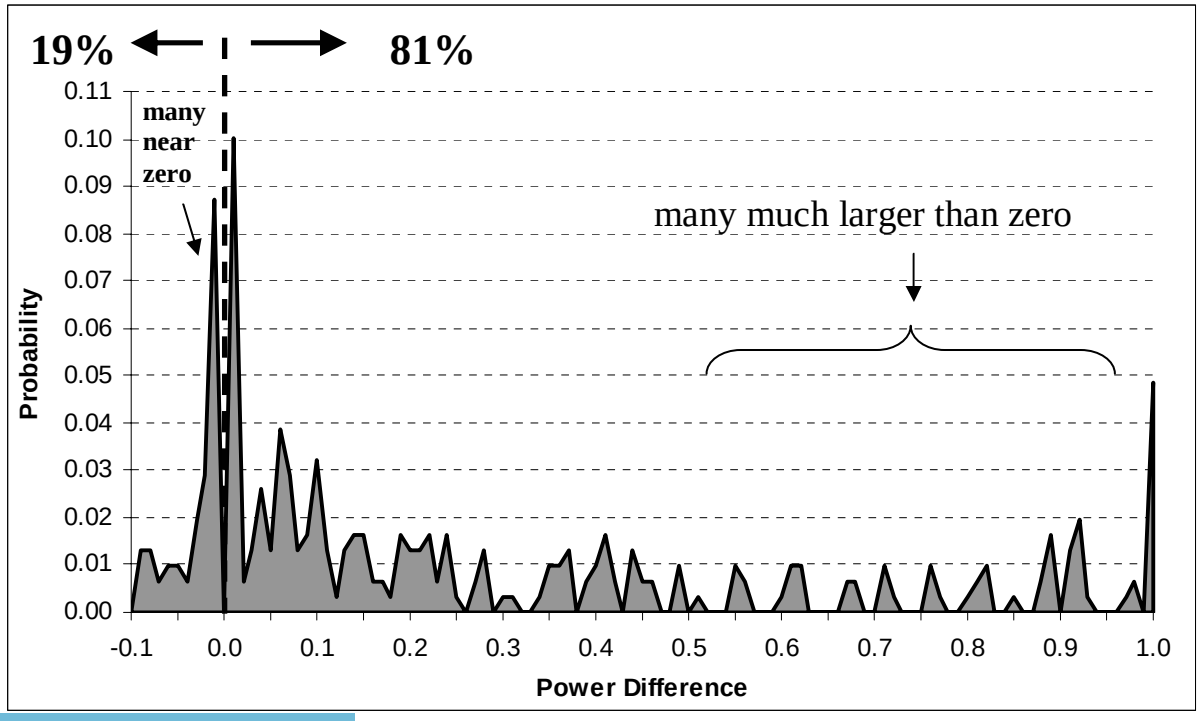


Figure 5. Alternate Hypothesis Simulations of $n_C = n_I = 30$ with a Power Difference (90 of 111): OBMax Power minus modified *t* Power ($\alpha$=0.05)

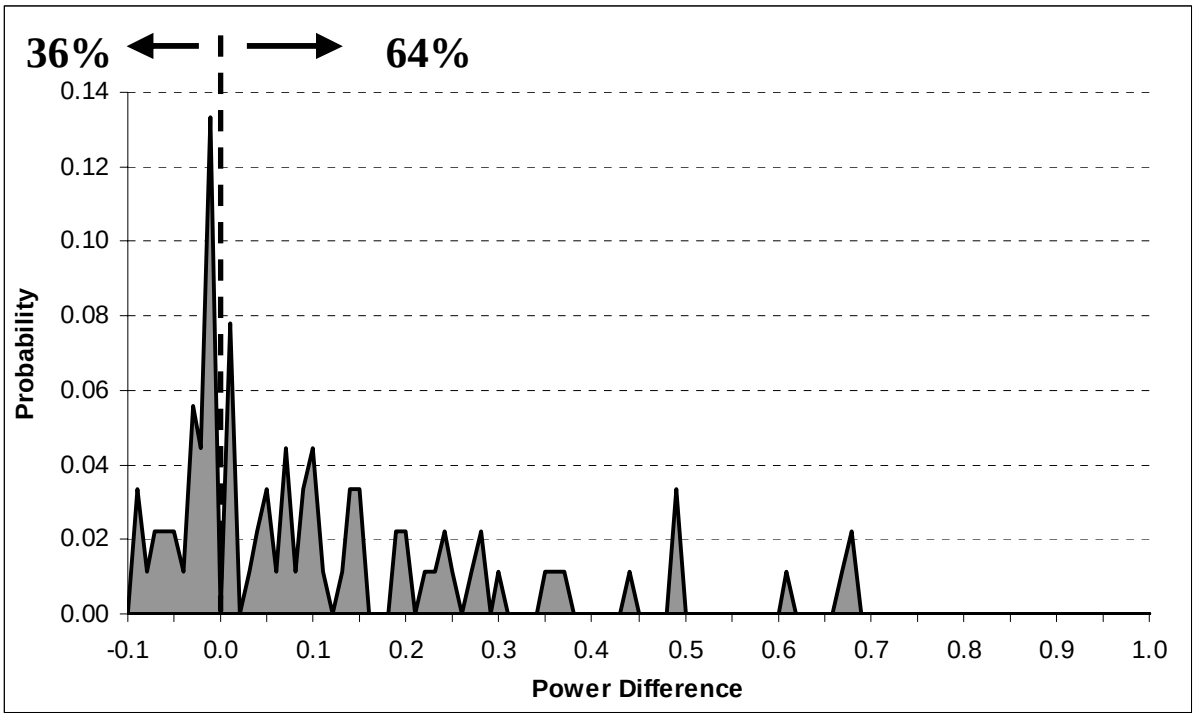



Figure 6. Alternate Hypothesis Simulations of $n_C = n_I = 300$ with a Power Difference (52 of 111): OBMax Power minus modified *t* Power ($\alpha$=0.05)

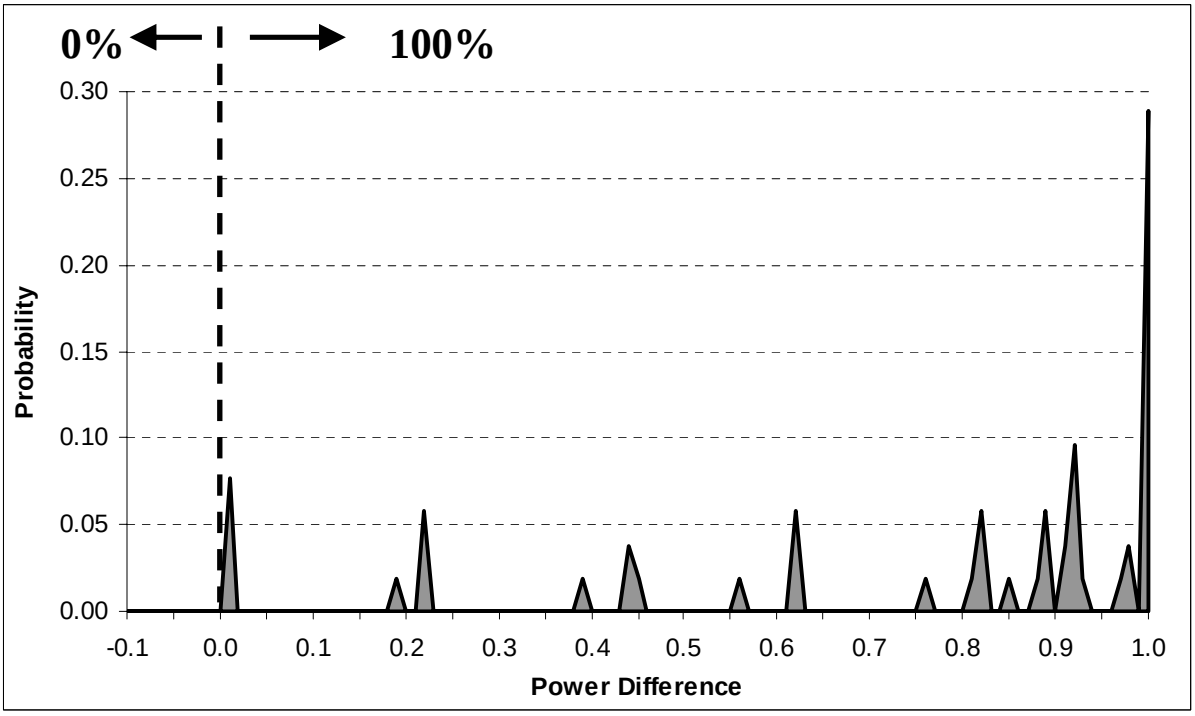


Table 4. modified *t* vs. OBMax: Dominant Test, and Corresponding Power Gains Under Symmetry (α = 0.05) by Magnitude of Mean Difference and Variance Difference

| $\sigma^2 / \mu$ | $\mu_C > \mu_I$ (small difference) | | $\mu_C > \mu_I$ (large difference) | $\mu_C \leq \mu_I$ |
|---|---|---|---|---|
| | Small $n_C$ ( = 30) | Large $n_C$ | | |
| $\sigma_C^2 > \sigma_I^2$ | **Usually OBMax**<br>Max = 0.223<br>Mean = 0.038<br>Median = 0.028 | **EQUAL** | **EQUAL** | **Always OBMax**<br>Max = 1.000<br>Mean = 0.431<br>Median = 0.361 |
| $\sigma_C^2 \leq \sigma_I^2$ | **Usually tmod**<br>Max = 0.051<br>Mean = 0.015<br>Median = 0.006 | **EQUAL** | **EQUAL** | **Ho:** |

OBMax vs. the modified *t*: Where does it matter in terms of remedies?

As shown in Figures 3-6 above, OBMax often provides dramatic power gains over the modified *t*, making it much more effective at identifying disparity when it truly exists. A very important point to note here is that the narrow conditions under which the modified *t* has a slight power advantage – small sample sizes and small location shifts (and a typically smaller or equal CLEC variance) – are exactly those that are the least important in terms of the size of the resulting remedies. Under most performance and remedy plans, the formulae for calculating remedies are proportionate functions of the number of lines or customers affected, as well as the magnitude of the degree to which service is out of parity (i.e., how much worse CLEC service is relative to ILEC service). Small sample sizes, and small deviations from parity, together imply the smallest remedies. Small power losses under these conditions (always less than 0.1 under symmetry, and no more than 0.2 under asymmetry when using OBMax2) will result in missed remedies that should be quite small, and perhaps even negligible, relative to overall remedies.

In contrast, under all other conditions of disparity, where both sample sizes and deviations from parity are much larger, the typically dramatic power gains of OBMax over the modified *t* will translate into much larger remedies that the modified *t* will fail to identify. The relative (if not absolute) size of these remedies missed by the modified *t* will dwarf any missed by OBMax when both sample sizes and location shifts are small. Thus, not only are the power gains of OBMax over the modified *t* much larger and more common than the losses, but also much more important in terms of the magnitude of the remedies that should be identified by the statistical test used. Consequently, from both a statistical and remedy-impact perspective, OBMax is dramatically better than the modified *t* at identifying disparate service provision to CLEC customers, and thus, is far more effectively used in parity testing to enforce the at-least-equal service provision of the Act. This makes OBMax is a better tool for achieving the Act's major objective: moving local telephone service from regulation to full competition and, once achieved, preventing backsliding to disparity into the future.

In other quality control settings, too, OBMax should be useful and widely applicable as discussed below, but the questions of how, and how much, the use of OBMax matters in OSS parity testing are examined next.

Table 5. Worst Level Violations of modified *t* vs OBMax2 Under Asymmetry (Opdyke, 2005)

| Statistic | $\sigma_C^2$ | $\mu_C$ | $n_C$ | $n_I$ | Distribution | α | Actual Size | Violation |
|---|---|---|---|---|---|---|---|---|
| OBMax2 | $\sigma_I^2$ | $\mu_I - \sigma_I$ | 300 | 30 | Exponential | 0.05 | 0.0553 | 0.0053 |
| OBMax2 | $\sigma_I^2$ | $\mu_I - 2\sigma_I$ | 300 | 30 | Exponential | 0.05 | 0.0566 | 0.0066 |
| OBMax2 | $\sigma_I^2$ | $\mu_I$ | 300 | 30 | Exponential | 0.05 | 0.0665 | 0.0165 |
| OBMax2 | $0.75\sigma_I^2$ | $\mu_I$ | 300 | 30 | Lognormal | 0.05 | 0.0581 | 0.0081 |
| OBMax2 | $\sigma_I^2$ | $\mu_I$ | 300 | 30 | Lognormal | 0.05 | 0.0623 | 0.0123 |
| OBMax2 | $\sigma_I^2$ | $\mu_I$ | 300 | 30 | Exponential | 0.10 | 0.1053 | 0.0053 |
| OBMax2 | $\sigma_I^2$ | $\mu_I$ | 300 | 30 | Lognormal | 0.10 | 0.1073 | 0.0073 |
| modt | $\sigma_I^2$ | $\mu_I$ | 30 | 30 | Lognormal | 0.05 | 0.0992 | 0.0492 |
| modt | $\sigma_I^2$ | $\mu_I$ | 300 | 30 | Exponential | 0.05 | 0.1003 | 0.0503 |
| modt | $0.50\sigma_I^2$ | $\mu_I$ | 300 | 30 | Lognormal | 0.05 | 0.1034 | 0.0534 |
| modt | $\sigma_I^2$ | $\mu_I$ | 300 | 30 | Lognormal | 0.05 | 0.1082 | 0.0582 |
| modt | $0.75\sigma_I^2$ | $\mu_I$ | 300 | 30 | Lognormal | 0.05 | 0.1089 | 0.0589 |
| modt | $\sigma_I^2$ | $\mu_I$ | 30 | 30 | Lognormal | 0.10 | 0.1451 | 0.0451 |
| modt | $\sigma_I^2$ | $\mu_I$ | 300 | 30 | Exponential | 0.10 | 0.1477 | 0.0477 |
| modt | $0.50\sigma_I^2$ | $\mu_I$ | 300 | 30 | Lognormal | 0.10 | 0.1544 | 0.0544 |
| modt | $0.75\sigma_I^2$ | $\mu_I$ | 300 | 30 | Lognormal | 0.10 | 0.1630 | 0.0630 |
| modt | $\sigma_I^2$ | $\mu_I$ | 300 | 30 | Lognormal | 0.10 | 0.1649 | 0.0649 |

OBMax vs. the modified *t*: How Does It Matter, and How to Decide?

The Act was designed so that, with respect to enforcing the central requirement of at-least-equal service provision to CLEC customers, everything hinges on the performance metric data, and the inferences made about it based on statistical tests. The consequences of OSS parity testing results that indicate disparity undeniably can be large, in terms of both remedies paid by ILECs to CLECs and, in the case of backsliding or prolonged and extensive disparity, the possible revocation of an ILEC's long-distance approval (which carries even larger, long-term financial consequences for both ILECs and CLECs).

Although not all performance metrics have statistical tests applied to them (a minority are comparisons of CLEC service against a fixed benchmark), and continuous data metrics are only a subset of all those subject to statistical parity testing, they still include some of the biggest metrics – i.e., those containing the most data reflecting the largest numbers of customers and phone lines (e.g., average time-to-install). Therefore, a statistic used to test these metrics that fails to identify actual disparity under a wide range of conditions not only distorts the simple and crucial incentive structure clearly and explicitly intended by the Act, but also misses sizeable remedies that would have been identified by a more powerful statistic – in this case, OBMax (or OBMax2).

Therefore, given the results of this study comparing OBMax to the modified *t*, one might ask when using actual OSS data, what is the magnitude of this distortion caused by the modified *t*? How much does it matter in terms of remedies, which is the bottom line in this setting? Although it is possible to approximately answer this question empirically, and the answer could very well be a sizeable amount, it is actually the wrong question to ask here for several reasons. First, it can never be known absolutely whether service provision to CLEC customers is truly inferior because only monthly samples are being considered, not entire populations. It could be, due

to random variation, that CLEC service is not really inferior, but that the given samples make it appear so (in statistical parlance, this is a Type I error). The reverse also can occur (a Type II error). What statistical tests provides is a scientific basis for making an inference, based on the samples that merely represent the true underlying service levels, with a specified degree of certainty (for example, if $\alpha = 0.05$, one can be $[1 - \alpha] = 95\%$ certain that an inference of parity is correct).

This guess or hypothesis about whether service is or is not in parity is the best that can be done, so a researcher can never evaluate the statistical properties of competing tests based (solely) on real data samples. The researcher must know the true answer in the data ahead of time, which is only possible with simulated data (as used in this study), and then see which statistic gets it right most often under the widest range of relevant data conditions. Then it will be known that, if applied to actual data samples that are based on truly disparate service levels, a statistic that is proven to be more powerful under well-constructed simulations will be more powerful under actual data and correctly detect the disparity more often.

That said, a general idea may be obtained as to how much remedies will be affected when using OBMax vs. the modified *t* by applying each to, say, six months of actual data and comparing the resulting remedies (such a comparison obviously would need to be based on identical remedy formulae, with distance-beyond-parity directly or indirectly based on p-values and $\alpha$; if Z-scores are familiar or in current use, then the inverse standard normal function can be used, e.g., $\Phi$(p-value) – $\Phi(\alpha)$ = distance beyond parity). If there are much larger remedies resulting from the use of OBMax, then it will be known that its greater power is driving this result.

However, even if no appreciable difference in remedies is observed (which would be surprising), the question 'How much are remedies actually affected?' is not the key question that needs to be answered because it ignores the important issue of a deterrent effect. If no appreciable difference in remedies is observed, that just means that scenarios under which OBMax is more powerful are not exhibited in the data being examined. But there is no telling that these types of inferior service scenarios will not crop up in the future (or have not cropped up at different times in the past). Because the modified *t* will definitely miss them if they do crop up, why would it ever be used over the more powerful statistic, OBMax? The answer is, it should not, and under a scientifically responsible implementation of applied statistics, it would not.

Thus, in evaluating which statistic to use for OSS parity testing and considering the remedy-impact of using OBMax instead of the modified *t*, the driving question is not, How much will actual remedies differ under OBMax vs. the modified *t*? (although the answer to this probably is noticeably, if not a great deal.); instead, the relevant question is, Under conditions that we know to be disparate, which statistic has greater power to correctly identify the disparity? This question cannot be answered by using actual data and comparing the remedies resulting from the use of each of these two statistics (although this comparison may be interesting), but rather, by the simulation study conducted in this paper. And the answer this study provides is that OBMax does have more power under a wider range of relevant data conditions, and these power gains are often dramatic. The general applicability of OBMax in other settings is discussed briefly below.

### General Utility of OBMax (OBMax2)

OBMax, and OBMax2, are useful in any context where one-sided tests of the first two moments are the primary or exclusive concern, and the researcher needs to test for effects in *either or both* moments (in other words, when the researcher needs to test (1) above). For these joint hypotheses, just as shown in Opdyke (2004) for OBMax's constituent tests, OBMax outperforms a test of stochastic dominance and a widely-used nonparametric distributional test against general alternatives. The Rosenbaum (1954) statistic maintains validity, but generally has much less power than OBMax, especially if the CLEC mean is smaller than the ILEC mean, when it often has absolutely no power to detect a larger CLEC variance (which is consistent with its design). The latter finding also holds for the one-sided Kolmogorov-Smirnov statistic which, although occasionally more powerful than OBMax, often severely violates the nominal level when means are identical but the CLEC variance is *smaller* (which is consistent with *its* design, if not the

relevant joint hypotheses examined here). Thus, OBMax is far superior to statistical tests that many researchers commonly turn to, at least initially, when faced with testing the joint hypotheses of (1) above. Among the settings in which these hypotheses are central is, of course, OSS parity testing; possibly the network access rules aimed at similar telecom deregulation efforts in other countries (Ure, 2003, p. 42-43); possibly the open access energy transmission regulations established by the Federal Energy Regulatory Commission (Gastwirth & Miao, 2002, p. 278); and numerous industrial settings with the need to address the quality control issues of accuracy and/or precision in manufacturing and other processes (Opdyke, 2005). Some important issues warranting further inquiry are listed below.

Further Research

Most of the points below are listed in Opdyke (2004) and remain important issues for further inquiry in this setting.

- In regulatory telecommunications, almost always $n_{CLEC} << n_{ILEC}$, so scenarios of $n_{CLEC} > n_{ILEC}$ were not studied in this paper. However, they are addressed in the further development of OBMax2 in Opdyke (2005).
- Although typically much more powerful than the modified *t*, even under skewed data, OBMax2 still has low power under asymmetry, and exploring ways to increase it is worthy of further study (Opdyke, 2005).
- Although the nominal test levels examined in this study ($\alpha = 0.05$ and $\alpha = 0.10$) bracket the vast majority of the test levels used in telecommunications OSS parity testing, (SBC Comments, 2002, p.49-52; CPUC Opinion, 2002, Appendix J, Exhibit 3, p.4; and Performance Assurance Plan – Verizon New York Inc., Redlined Version January 2003, Appendix D, p.1) other settings may require very different nominal levels (e.g., $\alpha = 0.20$ or $\alpha = 0.01$). Generalizing from the findings of this study to such conditions would not be advisable without further simulation.
- The one major exception to the above point regarding nominal test levels is the BellSouth performance and remedy plan. As previously mentioned, instead of solely using the modified *t* for continuous data performance metrics, this plan relies primarily on a statistic dubbed the truncated *Z* for which a balancing critical value is used as the nominal level of the hypothesis test. This critical value purports to balance or equalize the probability of Type I and Type II error (i.e., incorrect inferences of disparity and parity, respectively). This statistic, however, may remain insensitive to, i.e., have little power to detect, larger CLEC variance for two reasons: first, the formula used to determine the balancing critical value is admittedly essentially unaffected by differences in variances (BellSouth Comments, 2002, Attachment 2 (Part 4), Exhibit No. EJM-1, Appendix C, p.C-9); second, the statistical test scores that are truncated and combined to obtain the truncated *Z* score are simply scores of modified *t* tests adjusted for skewness (BellSouth Comments, 2002, Attachment 2 (Part 3), Exhibit No. EJM-1, Appendix A, p.A-5, with correction from Attachment 2 (Part 2), Appendix D – Technical Description, p. 37). It is not at all clear that a combined statistic based on such truncated *t*-scores has much or any power to detect differences in variances, and a thorough simulation study like the one completed in this paper would be useful to allay or confirm these suspicions.
- Although not the focus of this study, some performance and remedy plans use the general form of the modified *t* statistic as the basis for modifications to statistical tests designed for binary data, like that based on the common Wald approximation to the normal distribution (Comments of SBC, 2002, p. 59). In light of Opdyke's (2004) findings, and all of the problems inherent in using the modified *t* statistic with continuous data performance metrics, such modifications should be viewed with skepticism until subjected to careful analytic scrutiny and empirical simulation. No objections to using the modified *t* for continuous data OSS parity testing were raised. Mulrow (2002) raised no objection to using the modified *t* for continuous data OSS parity testing, although concern was expressed about making modified *t*–like changes to the Wald approximation test for binary data: "This does not seem right to me" (p.280). Instead of this

test, Mulrow (2002) advocated the use of Fisher's exact test. It is a viable and easily implemented alternative already in wide usage in OSS parity testing, although sometimes only for small(er) samples (SBC Performance Remedy Plan – Attachment 17, p. 3). Yet, it can be used for large samples as well because, even as a conditional exact test, it can be implemented very quickly with modern statistical software packages (e.g., SAS®). Agresti and Caffo (2000) provided a simple and effective, although not exact test for both small and large samples, and even better (more powerful), if slightly more complex alternatives, are the unconditional exact tests of Berger and Boos (1994) (available at http://www4.stat.ncsu.edu/~berger/tables.html) and Skipka et al. (2004) (Berger, 1996; Kopit & Berger, 1998). These all are carefully studied and well designed tests for binary data: there is no need to turn to unverified methods of questionable utility in this setting.

- Although not the focus of this study, some performance and remedy plans rely on a normal approximation *Z*-test for comparing CLEC and ILEC sample rates from count data performance metrics, even when those rates are very small (e.g., trouble report rate) and almost certainly highly non-normal (SBC Performance Remedy Plan – Attachment 17, p.3-4; Ameritech Michigan – Performance Remedy Plan – Attachment A, p. 2; and SBC Performance Remedy Plan – Version 3.0 SBC/SNET FCC 20 Business Rules – Attachment A-3, p.A-88). Yet, powerful and easily-implemented tests for comparing two Poisson means have been developed, and may be far superior statistically for such comparisons (Krishnamoorthy & Thomson, 2004). Examination of these metrics' distributions, and a straightforward simulation study, would adequately address this question.

Unheeded Warnings

As mentioned in Opdyke (2004), it is important to note that not everyone has supported the use of the modified *t* in this (and other) settings, although dissension has been conspicuously rare in the OSS parity testing arena. O'Brien (1993), in his discussion of Blair & Sawilowsky's (1993) empirical study unfavorably comparing the modified *t* to O'Brien's (1988) OBt and OBG statistics, points out that the Type I error rates of the modified *t* statistic will severely violate the nominal level of the test under a variety of conditions. Within the parity testing arena, over five years ago GTE voiced a lone, cautionary, and seemingly prescient dissent, given the findings of this current study, regarding use of the modified *t* in OSS parity testing:

> The modified *Z*-test [*t* test] should not be used since it follows no standard formulation of the test statistic. In the absence of a rigorous derivation, its sampling properties and maintained hypotheses are unknown. It has been asserted that the modified *Z*-test [*t* test] is a joint test of the equality of the means and variances of the two distributions; however no rigorous derivation has been provided. … It would clearly be foolish to accept a new and unknown test statistic without further documentation and consideration. (COMMENTS OF GTE, Before the Michigan Public Service Comm., 11/20/98, Attachment B, p.15-16)

(Opdyke, 2004, has since provided an analytic derivation of the asymptotic distribution of the modified *t*: as stated previously, it is *not* standard normal or student's *t* distributed, although it has been described as such in the expert testimony of Dysart & Jarosz, 2004 which, on pages 27-29, egregiously misquotes the derivation and major findings of Opdyke, 2004.)

Meanwhile, others have hedged their bets. While being deposed as an expert witness for AT&T and other CLECs, Dr. Gerald Ford was asked:

> DO YOU BELIEVE THE MODIFIED Z-TEST SHOULD BE REPLACED WITH THESE PROPOSED ALTERNATIVES?
>
> No. The development of the particulars of the performance plan took many months of hard work by some very smart people. It was only after considerable analysis and debate that the Modified Z-test [modified *t* test] was selected as the best test statistic for the performance plan. …I see no reason to alter the test procedures of the existing plan without strong

> evidence that the other tests represent an improvement.
>
> SO YOU BELIEVE THE MODIFIED Z-TEST [modified *t* test] SHOULD BE USED?
>
> Yes, at least until some strong evidence is provided to indicate an alternative test is preferred. (Before the Texas PUC, Rebuttal Testimony of Dr. Gerald Ford for the CLEC Coalition, 08/23/04, p.36)

The goal of this article, with its development of a single, nonparametric, yet generally powerful statistic for continuous-data OSS parity testing, has been to provide the "further documentation and consideration" implicitly requested by GTE (1998), as well as the "strong evidence" of "an improvement" over the modified *t* that Ford (2004) implicitly requested much more recently.

## Conclusion

As summarized in Opdyke (2004), under the Telecommunications Act of 1996, ILECs are required to provide CLEC customers with local telephone service "at least equal in quality to" that which they provide to their own customers if they are to be allowed into the long distance telephone market (Telecommunications Act of 1996, Pub. LA. No. 104-104, 110 Stat. 56 (1996), at §251 (c) (2) (C)). The goal of this carrot-stick approach – the carrot being the potentially lucrative long distance market, and the stick being this requirement of at-least-equal service provision – is to promote competition in the newly deregulated local telephone markets. Implementing and enforcing the at-least-equal service provision requirement has taken the form of OSS parity testing – statistically testing the service data represented in thousands of operations support services performance metrics to ensure that the service provided to CLEC customers is, in fact, at least equal.

Results from these statistical tests indicating average service and/or service variability that is *not* at least equal, i.e., findings of disparity, typically require an ILEC to pay fines (sometimes US$ millions) to the CLEC(s), and sometimes to the state(s); disparity that is consistent and widespread over time (i.e., backsliding) can serve as cause for the revocation of an ILEC's approval to provide long distance service. These stakes are high, not only for individual firms but also for the entire industry, so choosing the correct, if not the best statistics to use in OSS parity testing is a very important decision.

To date, the modified *t* statistic (Brownie et al., 1990) has been approved and used in OSS parity testing across the country. It is used on continuous-data performance metrics as a test of whether average service and/or service variability are at least equal for CLEC customers compared to their ILEC counterparts. However, Opdyke (2004) demonstrated that the modified *t* is an ineffective and misleading choice for this purpose in this setting. It remains *potentially* vulnerable to gaming – intentional manipulation of its score to mask disparity – but far more importantly, it remains absolutely powerless to detect inferior CLEC service provision under a wide range of relevant data conditions. Opdyke (2004) proposed the use of several other easily implemented conditional statistical procedures that are not vulnerable to gaming and typically provide dramatic power gains over the modified *t*. The selection of which among them to use, however, depends on the relative sizes of the two data samples and a distributional characteristic (the kurtosis) of the specific performance metric being tested. Although this is arguably straightforward, a single test that could accomplish the same thing would be preferable, and the development of such a statistic is the motivation for this article.

In this article, an easily-implemented maximum test – OBMax – was developed based on the multiple statistics proposed by Opdyke (2004). OBMax maintains reasonable Type I error control and is always either nearly as powerful as its constituent tests, or almost as often as not, even more powerful. More importantly, it typically provides dramatic power gains over the modified *t*. The one set of narrow conditions under which the modified *t* has a slight power advantage (always less than 0.1 under symmetry) are exactly those under which consequent fines or remedies imposed on ILECs will be the smallest – small CLEC sample sizes and small location shifts (and equal or close-to-equal variances).

In contrast, the typically dramatic power gains of OBMax over the modified *t* under most other conditions of disparity (sometimes gains of even 1.0!) translate into the appropriate identification of

vastly larger amounts of remedies that the modified *t* will miss. From both a statistical and remedy-impact perspective, therefore, OBMax is superior at detecting disparity, and thus, at enforcing the at-least-equal service provision of the Telecommunications Act of 1996. It consequently is an unambiguously better statistic than the modified *t* for use in OSS parity testing to achieve the major objective of the Act: the movement of local telephone service from regulation to full market competition.

## References


Agresti, A., & Caffo, B. (2000), Simple and effective confidence intervals for proportions and differences of proportions result from adding two successes and two failures, *The American Statistician, Vol. 54*, No. 4, 280-288.

Before the Federal Communications Commission, Comments of BellSouth, CC Docket No. 01-318, ATTACHMENT 2, January 22, 2002. [available at http://gullfoss2.fcc.gov/prod/ecfs/retrieve.cgi?native_or_pdf=pdf&id_document=6512980049 through 6512980052]

Before the Federal Communications Commission, In the Matter of Performance Measurements and Reporting Requirements of Operations Support Systems, Interconnection, and Operator Services and Directory Assistance, CC Docket No. 98-56 RM-9101, Motion to Accept Late-Files Documents of U S WEST Communications, Inc., APPENDIX A: Comments of Michael Carnall on Statistical Issues of Detecting Differences in Service Quality, On Behalf of U S WEST Communcations, June 1, 1998. [available at http://gullfoss2.fcc.gov/prod/ecfs/retrieve.cgi?native_or_pdf=pdf&id_document=2078390002 and 2078390003]

Before the Federal Communications Commission, 07/09/03 filing on behalf of SBC Communications, Inc.: Performance Remedy Plan – SBC – Version 3.0 SBC/SNET FCC 20 Business Rules – Attachment A-3: Calculation of Parity and Benchmark Performance and Voluntary Payments. [available at http://gullfoss2.fcc.gov/prod/ecfs/retrieve.cgi?native_or_pdf=pdf&id_document=6514285147]

Before the Federal Communications Commission, in the Matter of Performance Measurements and Standards for Unbundled Network Elements and Interconnection; Performance Measurements and Reporting Requirements for Operations Support Systems, Interconnection, and Operator Services and Directory Assistance; Deployment of Wireline Services Offerings Advanced Telecommunications Capability; Petition of Association for Local Telecommunications Services for Declaratory Ruling; CC Docket No. 01-318; CC Docket No. 98-56; CC Docket No. 98-147; and CC Docket Nos. 98-147, 96-98, and 98-141; Comments of SBC Communications, Inc., January 23, 2002. [available at http://gullfoss2.fcc.gov/prod/ecfs/retrieve.cgi?native_or_pdf=pdf&id_document=6512980270 and 6512980271]

Before the Commonwealth of Massachusetts Department of Telecommunications and Energy, D.T.E. 99-271, MCI Worldcom's Proposed Performance Assurance Plan for Bell Atlantic-Massachusetts, Appendix D – Simplified Measurement of Performance and Liability: The SiMPL Plan, by George S. Ford on Behalf of MCI Worldcom. [available at http://www.state.ma.us/dpu/telecom/99-271/pap/MCIW/appendix_D.pdf]

Before the Michigan Public Service Commission, Case No. U-11830, COMMENTS OF GTE, November 20, 1998, ATTACHMENT B: Statistical Method for OSS Performance Measures, GTE White Paper.

Before the Michigan Public Service Commission, Case No. U-13848, 09/04/03 filing on behalf of SBC Michigan: Ameritech Michigan – Performance Remedy Plan – Attachment A. [available at http://efile.mpsc.cis.state.mi.us/efile/docs/13848/0004.pdf]

Before the Public Utilities Commission of the State of California, Order Instituting Rulemaking on the Commission's Own Motion into Monitoring Performance of Operations Support Systems, Decision 01-01-037, January 18, 2001. – Rulemaking 97-10-016 (Filed October 9, 1997), and Order Instituting Investigation on the Commission's Own Motion into Monitoring Performance of Operations Support Systems. – Investigation 97-10-017 (Filed October 9, 1997): INTERIM OPINION ON PERFORMANCE INCENTIVES, Decision 01-01-037, January 18, 2001. [available at http://www.cpuc.ca.gov/word_pdf/FINAL_DECISION/11842.pdf]

Before the Public Utilities Commission of the State of California, Order Instituting Rulemaking on the Commission's Own Motion into Monitoring Performance of Operations Support Systems, Decision 01-01-037, March 6, 2002. – Rulemaking 97-10-016 (Filed October 9, 1997), and Order Instituting Investigation on the Commission's Own Motion into Monitoring Performance of Operations Support Systems. – Investigation 97-10-017 (Filed October 9, 1997): OPINION ON THE PERFORMANCE INCENTIVES PLAN FOR PACIFIC BELL TELEPHONE COMPANY, Decision 02-03-023, March 6, 2002, APPENDIX J. [available at http://www.cpuc.ca.gov/published/final_decision/13927.htm and http://www.cpuc.ca.gov/PUBLISHED/FINAL_DECISION/13928.htm]

Before the Public Utilities Commission of Texas, Docket No. 28821, SBC Texas' Joint Direct Testimony of William R. Dysart and Dorota Jarosz, July 19, 2004. – available at http://interchange.puc.state.tx.us/WebApp/Interchange/application/dbapps/billings/pgDailySearch.asp

Before the Public Utilities Commission of Texas, Docket No. 28821, Rebuttal Testimony of George S. Ford, Ph.D., on behalf of the CLEC Coalition, August 23, 2004. – available at http://interchange.puc.state.tx.us/WebApp/Interchange/application/dbapps/billings/pgDailySearch.asp

Berger, R.L. (1996), More powerful tests from confidence interval p values, *The American Statistician, Vol. 50*, 314-317.

Berger, R.L., & Boos, D.D. (1994), P Values maximized over a confidence set for the nuisance parameters, *Journal of the American Statistical Association, Vol. 89*, 1012-1016. [available at http://www4.stat.ncsu.edu/~berger/tables.html]

Blair, R.C. (2002), Combining two nonparametric tests of location, *Journal of Modern Applied Statistical Methods*, Vol. 1, No. 1, 13-18.

Blair, R.C. (1991), New critical values for the generalized t and generalized rank-sum procedures, *Communications in Statistics, Vol. 20*, 981-994.

Blair, R.C. & Sawilowsky, S. (1993) Comparison of two tests useful in situations where treatment is expected to increase variability relative to controls, *Statistics in Medicine, Vol. 12*, 2233-2243.

Brown, M. B. & Forsythe, A. B. (1974), Robust tests for the equality of variances, *Journal of the American Statistical Association, Vol. 69*, 364-367.

Brownie, C., Boos, D. D. & Hughes-Oliver, J. (1990), Modifying the t and ANOVA F tests when treatment is expected to increase variability relative to controls, *Biometrics, Vol. 46*, 259-266.

Cochran, W. (1977), *Sampling techniques*, 3rd ed., New York: John Wiley & Sons.

D'Agostino, R.B., A. Belanger, and R.B. D'Agostino, Jr. (1990), A suggestion for using powerful and informative tests of normality, *The American Statistician, 44*: 316-321.

Evans, M., Hastings, N., & Peacock, B. (1993), *Statistical distributions*, 2nd ed., New York: John Wiley & Sons.

Federal Communications Commission, Notice of Proposed Rulemaking, CC Docket No. 98-56, RM-9101, FCC 98-72, APPENDIX B, Adopted April 16, 1998. [available at http://gullfoss2.fcc.gov/prod/ecfs/retrieve.cgi?native_or_pdf=pdf&id_document=2060360001 to 0004]

Federal Communications Commission, Notice of Proposed Rulemaking, CC Docket No. 98-56, FCC 01-331, APPENDIX B, Adopted November 8, 2001. [available at http://gullfoss2.fcc.gov/prod/ecfs/retrieve.cgi?native_or_pdf=pdf&id_document=6512974611 & 4612]

Fleming, T. R. & Harrington, D. P. (1991) *Counting processes and survival analysis*. Wiley, New York.

Freidlin, B. & Gastwirth, J. (2000a) Changepoint tests designed for the analysis of hiring data arising in employment discrimination cases, *Journal of Business and Economic Statistics, Vol. 18*, No. 3, 315-322.

Freidlin, B. & Gastwirth, J. (2000b) On power and efficiency robust linkage tests for affected sibs, *Annals of Human Genetics, 64*, 443-453.

Freidlin, B., Zheng, G., Zhaohai, L., & Gastwirth, J. (2002) Trend tests for case-control studies of genetic markers: power, sample size and robustness, *Human Heredity, Vol. 53*, No. 3, 146-152.

Gastwirth, J. L., & Miao, W. (2002), Comment, *Statistical Science, Vol. 17*, No. 3, 271-276.

Goodman, L.A. (1954), Kolmogorov-Smirnov tests for psychological research, *Psychological Bulletin, 51*, 160-168.

Kopit, Justin S., and Berger, R.L. (1998), A more powerful exact test for a practical difference between binomial proportions, *American Statistical Association Proceedings, Biopharmaceutical Section*, 251-256.

Krishnamoorthy, K., & Thomson, J. (2004), A more powerful test for comparing two poisson means, *Journal of Statistical Planning and Inference, Vol. 119*, Issue 1, 23-35.

Lee, W. J., (1996) some versatile tests based on the simultaneous use of weighted log-rank statistics, *Biometrics, 52*, 721-725.

Levene, H. (1960), Robust tests for equality of variances, in *Contribution to probability and statistics: essays in honor of harold hotelling*, I. Olkin et al., eds., Stanford University Press, Palo Alto, 278-292.

Local Competition User's Group ("LCUG" – Membership: AT&T, Sprint, MCI, LCI, WorldCom), Statistical tests for local service parity, February 6, 1998, Version 1.0.

Mallows, C. (2002), Parity: Implementing the Telecommunications Act of 1996, *Statistical Science, Vol. 17*, No. 3, 256-285.

Matlack, W.F., (1980), *Statistics for public policy and management*, Belmont, CA: Duxbury Press.

Mulrow, E. (2002), Comment, *Statistical Science, Vol. 17*, No. 3, 276-281.

Neuhäuser, M. Büning, H., & Hothorn, L. A. (2004), Maximum test versus adaptive tests for the two-sample location problem, *Journal of Applied Statistics, Vol. 31*, No. 2, 215-227.

O'Brien, P.C. (1988), Comparing two samples: extensions of the t, rank-sum, and log-rank tests, *Journal of the American Statistical Association, Vol. 83*, 52-61.

O'Brien, P.C. (1993), Discussion, *Statistics in Medicine, Vol. 12*, 2245-2246.

Opdyke, J.D. (2004), Misuse of the 'modified' *t* statistic in regulatory telecommunications, *Telecommunications Policy, Vol.28*, 821-866.

Opdyke, J.D. (2006), A nonparametric statistic for joint mean-variance quality control, *American Statistical Association Proceedings - 2005, Section on Quality and Productivity*, forthcoming.

Performance Assurance Plan – Bell-Atlantic, New York, (filed with New York Public Service Commission 04/07/2000). [available at http://www.fcc.gov/telecom.html]

Performance Remedy Plan – SBC, 13 states, Attachment 17. [available at http://www.nrri.ohio-state.edu/oss/Post271/Post271/Texas/performance%20agreement.pdf]

Performance Assurance Plan – Verizon New York Inc., Redlined Version January 2003. [available at http://www.dps.state.ny.us/ny2003pap_redline.PDF and http://www.dps.state.ny.us/nyappndx_a_to_f_h2003pap.PDF]

Pesarin, F. (2001), *Multivariate permutation tests with applications in biostatistics*, John Wiley & Sons, Ltd., New York.

Rosenbaum, S. (1954), Tables for a Nonparametric Test of Location, *Annals of Mathematical Statistics, 25*, 146-50.

Ryan, L.M., Freidlin, B., Podgor, M.J., & Gastwirth, J.L., (1999), Efficiency robust tests for survival or ordered categorical data, *Biometrics, 55*, No. 3, 883-886.

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

Shiman, D. (2002), Comment, *Statistical Science, Vol. 17*, No. 3, 281-284.

Siegel, S. & Castellan, N. John, (1988), *Nonparametric Statistics for the Behavioral Sciences*, 2nd ed., New York: McGraw-Hill.

Shoemaker, L. H. (2003), Fixing the F test for equal variances, *The American Statistician, Vol. 57*, No. 2, 105-114.

Skipka, G., Munk, A., & Freitag, G. (2004), Unconditional exact tests for the difference of binomial probabilities – contrasted and compared, *Computational Statistics and Data Analysis, Vol.47*, No.4, 757-774.

Tarone, R.E., (1981) On the distribution of the maximum of the log-rank statistic and the modified wilcoxon statistic, *Biometrics*, 37, 79-85.

Telecommunications Act of 1996, Pub. LA. No. 104-104, 110 Stat. 56 (1996). [available at http://www.fcc.gov/telecom.html]

The Qwest Performance Assurance Plan, Revised 11/22/2000 [available at http://www.nrri.ohio-state.edu/oss/Post271 /Post271/Qwest_11-22-00_Red-lined_PAP.pdf]

Verizon Performance Assurance Plan, Redlined Version of Current PAP Showing Proposed 2003 Modifications, APPENDIX D. [available at http://www.dps.state.ny.us/ Case_99C0949.htm]

Ure, J. (2003), Competition in the local loop: unbundling or unbungling? *Info: The Journal of Policy, Regulation, and Strategy for Telecommunications, Information and Media*, Vol. 5, No. 5, 38-46.

Weichert, M. & Hothorn, L.A. (2002) Robust hybrid tests for the two-sample location problem, *Communications in Statistics – Simulation and Computation, Vol. 31*, 175-187.

Willan, A.R. (1988) Using the maximum test statistic in the two-period crossover clinical trial, *Biometrics, Vol. 44*, No. 1, 211-218.

Yang, Song, Li Hsu, & Lueping Zhao (2005), Combining asymptotically normal tests: case studies in comparison of two groups, *Journal of Statistical Planning and Inference, Vol. 133*, Issue 1, 139-158.

Zar, J.H., (1999), *Biostatistical analysis*, 4th ed., Upper Saddle River, NJ: Prentice-Hall.

## Appendix

*OBt and OBG*: O'Brien's OBt test involves running the following ordinary least squares regression on pooled data including both samples:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i , \qquad (6)$$

where $y$ is a dummy variable indicating inclusion in the CLEC sample, and $x$ is the performance metric variable. If the parameter on the quadratic term ($\beta_2$) is (positively) statistically significant at the 0.25 level, use the critical value of the overall equation to reject or fail to reject the null hypothesis; if it is not, use the critical value of the overall equation of the following ordinary least squares regression instead:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad (7)$$

O'Brien's OBG test is identical to the OBt test except that the pooled-sample ranks of $x$ are used in the regressions instead of the $x$ data values themselves.

*Modified Levene test*: The modified Levene test requires a simple data transformation: take the absolute value of each data point's deviation from its respective sample median (as per Brown and Forsythe, 1974), and then calculate the usual one-way ANOVA statistic using these transformed values (as per Levene, 1960). The resulting statistic (8) is referenced to the $F$ distribution as usual.

Let $z_{ij} = \left| x_{ij} - \tilde{x}_i \right|$ where $\tilde{x}_i$ is sample $i$'s median (8)

$$W_o = \frac{\sum_i n_i \left( \bar{z}_{i\cdot} - \bar{z}_{\cdot\cdot} \right)^2 \Big/ (g-1)}{\sum_i \sum_j \left( z_{ij} - \bar{z}_{i\cdot} \right)^2 \Big/ \sum_i (n_i - 1)} \sim F_{(g-1),\sum_i (n_i - 1)}$$

where $\bar{z}_i = \sum z_{ij} / n_i$ and $\bar{z}_{\cdot\cdot} = \sum \sum z_{ij} / n_i$

However, because this test is designed as a two-tailed test, and the hypotheses being tested in this setting are one-tailed, the p-value resulting from this test, when used conditionally with O'Brien's tests as in Table 1, must be subtracted from 1.0 if the CLEC sample variance is less than the ILEC sample variance. Or, if one does not need to calculate a p-value that is be known to be larger than α (as when the CLEC sample variance is smaller), the calculation simply can be skipped.

*Shoemaker's $F_1$ test*: Shoemaker's $F_1$ test is simply the usual ratio of sample variances referenced to the $F$ distribution, but using different degrees of freedom:

$$s_C^2 / s_I^2 \sim F_{df_C, df_I} \qquad (9)$$

where $$df_i = 2n_i \Big/ \left( \frac{\hat{\mu}_4}{\hat{\sigma}^4} - \frac{n_i - 1}{n_i - 3} \right)$$

where i = C, I corresponds to the two samples, and $\mu_4$ and $\sigma^4$ are estimated from the two samples when pooled:

$$\hat{\mu}_4 = \sum_{i=1}^{2} \sum_{j=1}^{n_i} \left( x_{ij} - \bar{x}_i \right)^4 \Big/ (n_1 + n_2) \qquad (10)$$

$$\hat{\sigma}^4 = \left[ \left( (n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 \right) \Big/ (n_1 + n_2) \right]^2 \quad (11)$$

Shoemaker (2003) notes that the biased estimate for $\sigma^4$ is used for improved accuracy.

*Separate-variance t test*: The separate-variance *t* test, also known as the Welch or Behrens-Fisher *t* test, is presented below:

$$t_{sv} = \frac{(\bar{X}_C - \bar{X}_I) - (\mu_C - \mu_I)}{\sqrt{\frac{s_I^2}{n_I} + \frac{s_C^2}{n_C}}} \quad (12)$$

where $s_I^2 = \frac{\sum_{i=1}^{n_I} (X_{I_i} - \bar{X}_I)^2}{(n_I - 1)}$, $s_C^2 = \frac{\sum_{i=1}^{n_C} (X_{C_i} - \bar{X}_C)^2}{(n_C - 1)}$,

$\bar{X}_I = \frac{\sum_{i=1}^{n_I} X_i}{n_I}$, and $\bar{X}_C = \frac{\sum_{i=1}^{n_C} X_i}{n_C}$

Satterwaith's (1946) degrees of freedom for $t_{sv}$ is:

$$df = \frac{\left( \frac{s_I^2}{n_I} + \frac{s_C^2}{n_C} \right)^2}{\frac{\left( \frac{s_I^2}{n_I} \right)^2}{(n_I - 1)} + \frac{\left( \frac{s_C^2}{n_C} \right)^2}{(n_C - 1)}} \quad (13)$$

If *df* is not an integer, it should be rounded down to the next smallest integer (Zar, 1999, p. 129)

*Test of D'Agostino et al. (1990)*: The test of D'Agostino et al. (1990) is calculated as follows:

$$g_1 = \frac{k_3}{s^3} = \frac{\frac{n \sum (X_i - \bar{X})^3}{(n-1)(n-2)}}{\sqrt{\left(s^2\right)^3}}, \qquad \sqrt{b_1} = \frac{(n-2)g_1}{\sqrt{n(n-1)}}$$

$$A = \sqrt{b_1} \sqrt{\frac{(n+1)(n+3)}{6(n-2)}} \quad (14)$$

$$B = \frac{3\left(n^2 + 27n - 70\right)(n+1)(n+3)}{(n-2)(n+5)(n+7)(n+9)}$$

$$C = \sqrt{2(B-1)} - 1 \ , \qquad D = \sqrt{C} \ , \qquad E = \frac{1}{\sqrt{\ln D}}$$

$$F = \frac{A}{\sqrt{\frac{2}{C-1}}} \ , \qquad Z_{g_1} = E \ln\left( F + \sqrt{F^2 + 1} \right) \sim \phi(0,1)$$

(~ standard normal)

For one-tailed testing of skewness to the left, check $\Pr\left(Z \le Z_{g_1}\right)$; for skewness to the right, check $\Pr\left(Z \ge Z_{g_1}\right)$. See Zar (1999), p. 115-116, for further details.

Journal of Modern Applied Statistical Methods
November, 2005, Vol. 4, No. 2, 394- 401


1538 – 9472/05/$95.00

*Regular Articles*

# Estimation of Process Variances in Robust Parameter Designs

T. K. Mak Fassil Nebebe

Concordia University
Montreal, Quebec, Canada

The modeling of variation through interactions is appealing in crossed array design as it leads to greater robustness to certain type of model misspecification. As an alternative to signal-to-noise analysis, a new, systematic method based on Taguchi type crossed array design is given. It is shown in this article that when fractional factorial design is used for the outer array, the crossed array design is not robust to the presence of noise-noise interactions and a method of rectifying the problem is suggested.

Keywords: Inner and outer arrays, interactions, off-line quality control, orthogonal polynomial, PerMIA, Taguchi experiment.

## Introduction

Robust design has been widely used in industry to improve productivity and achieve higher quality at a lower cost. The main idea in robust design is to develop product and process designs that can deliver at a minimal cost units of target performance which are usable or functional with maintained quality under all intended operating conditions.

T. K. Mak is a Professor in the Department of Decision Sciences & M.I.S. His research interests are in survey design and analysis, measurement error models, and statistical quality control. He has served as an Associate Editor of the *Canadian Journal of Statistics* and was an elected member of ISI. F. Nebebe is a Associate Professor in the Department of Decision Sciences & M.I.S. His research interests are in Bayes and empirical Bayes methods. He is an International Advisory Board member of *SINET*: *Ethiopian Journal of Science*, and the *Journal of the Ethiopian Statistical Association.*

Thus, one major approach in robust design is to reduce variation in the quality characteristic without actually eliminating the causes of variation (the noise factors). Instead of replacing some components with more expensive ones to achieve smaller variation from target, robust design methodology seeks combinations of levels of factors affecting the quality characteristics that are least sensitive to environmental changes in production or operating conditions. This adjustment to the optimal levels are usually less expensive and are achieved through parameter design.

In parameter design, techniques of design of experiments are widely used to obtain data for a number of experimental runs corresponding to different combinations of the factors. An analysis of the resulting data is performed to approximate the optimal combination yielding the smallest variation from the target. In these regards, Taguchi-type experiments consisting of crossed arrays are sometimes performed, and the experimental data are analyzed using signal to noise ratio as a performance measure. A factor affecting response or product characteristic can be classified as a control factor or a noise factor (internal or external). Control factors are factors the levels or values of which are controllable during production. In contrast, the levels of the noise factors are expensive to control in

production or uncontrollable during use in the lifetime of the product. However, for the purpose of assessing their effects on the quality characteristics, the levels of the noise factors may also be controlled in the experimental runs in parameter design. In crossed array designs, each treatment combination of the control factors considered appears with every member in a set of treatment combinations of noise factors.

Taguchi's crossed array design and the signal-to-noise ratio analysis were criticized in the literature (Box, 1988). Some major difficulties in Taguchi's approach are summarized in Barreau et al. (1999). Crossed array design generally calls for a larger number of experimental runs which may be deemed unnecessary when some of the interactions may be safely assumed to be zero (Shoemaker et al., 1991). Furthermore, the use of signal-to-noise ratio may not always be appropriate as a performance measure to be minimized (Box, 1988), and modeling directly the signal to noise ration as the response in ANOVA is generally not intuitive and problematic. As an alternative design, the use of combined arrays has been suggested in the literature (Welch et al., 1990; Shoemaker et al., 1991).

In combined array design, both the control and noise factors are integrated into the same array, resulting in less number of experimental runs. The resulting data are then analyzed differently, with the control factors affecting variance through their interactions with the noise factors (O'Donnell and Vining, 1997; Myers, 1997). Engel and Huele (1996) used a generalized linear modeling approach to analyze combined array designs.

It is interesting to note that similar approach of modeling through interactions between the control and the noise factors is in fact more appropriate for crossed array designs (Barreau, et al., 1999). Despite some of its major drawbacks, Taguchi's approach is still embraced by many practitioners, largely because of its conceptual simplicity and easier implementation that requires less sophisticated analytical tools. Furthermore, the combined array methodology, though more economical, is less robust than the crossed array design to model misspecification especially when certain significant interactions among control factors are accidentally omitted in the design and analysis.

The number of experimental runs required in a crossed array design can be substantially reduced by employing fractional factorial designs for the inner (involving control factors) and outer array (involving noise factors). Barreau, et al. (1999) examined the role of interactions between control and noise factors in a Taguchi type experiment. These approaches of design and analysis have the advantages of being more economical, and yet are capable of retaining the benefits of having crossed inner and outer arrays.

The use of interaction analysis also throws light on how the noise variables affect the response, and provides a more natural analysis than a direct modeling of the signal-to-noise ratio as a response variable. Design of resolution III can be used for the inner array without any adverse effects on the study of variation or performance measure even if some interactions exist between control factors. However, complication arises when two factor interactions exist between noise factors. Such interactions do not appear in the true unknown objective function to be minimized for finding optimal levels, but it is shown in this paper that they can seriously bias the estimation of this objective function.

It is suggested that this potential bias be corrected based on a small confirmatory experiment. It is also proposed to use orthogonal polynomials in the analysis to facilitate the identification of adjustment variables, variables that only affect variation through the mean function. It is well known that the use of adjustment variables greatly simplifies the process of minimizing variation while having the mean on target. Furthermore, the use of orthogonal polynomials when some variables are quantitative allows one to better relate the analysis to response surface methodology and to obtain interpolated values for improved results in variance minimization.

## Methodology

In this section, an outline of a systematic approach for analyzing data from a crossed array design is given. The details are best explained by

a practical example, which will be left to the next section. Let $y$ be the response variable representing a certain product characteristic. Suppose there are $c$ control variables each with $k_c$ levels, and $n$ noise variables each has $k_n$ levels. For the ease of discussion, all the control and noise variables are assumed to be quantitative, but the necessary modifications when there are both quantitative and qualitative variables will be demonstrated with a real example in the next section.

Suppose that there are $N_c$ treatment combinations in the inner array, which is an orthogonal resolution III main effect plan. Similarly, there are $N_n$ treatment combinations in the outer array, which is an orthogonal resolution III main effect plan. Assume all interactions involving three or more factors (both control and noise factors) are non-significant. For the $i^{th}$ control factor $x_i$, there are $k_c$ levels corresponding to $k_c$ numeric coded values. Denote the set of the $k_c$ numeric coded values by $W$. Let $u_1(x),...,u_{k_c-1}(x)$ be orthogonal polynomials where $u_j(x)$ is a polynomial of degree $j$ such that $\sum_{x_i \in W} u_j(x_i) = 0$, $\sum_{x_i \in W} u_j(x_i)u_{j'}(x_i) = 0$, for all $j$ and $j \neq j'$.

The $n$ noise factors $z_1,...,z_n$ are random variables assumed to be independent and, without loss of generality, to have mean 0 and standard deviation 1. Thus if all the two factor control-control and noise-noise interactions are suppressed, a linear model for the response $y$ conditional also on $z_1,...,z_n$ can be formulated as:

$$\begin{aligned} y &= f(x_1,...,x_c,z_1,...,z_c) + e \\ &= \mu + \sum_{i=1}^{c} \alpha_i^T u(x_i) + \sum_{i'=1}^{n} \gamma_{i'} z_{i'} \\ &\quad + \sum_{i'=1}^{n}\sum_{i=1}^{c} \beta_{i'i}^T u(x_i) z_{i'} + e, \end{aligned}$$

where $\alpha_i$ is a $k_c \times 1$ vector, $\gamma_{i'}$ is a scalar, $\beta_{i'i}$ is a $k_c \times 1$ vector of unknown coefficients, and $u(x) = (u_1(x),...,u_{k_c-1}(x))^T$. Here the error term $e$ has mean 0 and constant variance $\sigma_e^2$. Thus for given $x_1,...,x_c$, treating $z_1,...,z_n$ as random, the variance of $y$ is therefore

$$\sigma^2(x_1,...,x_c) = \sum_{i'=1}^{n} V_{i'}^2 + \sigma_e^2 \qquad (1)$$

where

$$V_{i'} = (\gamma_{i'} + \sum_{i=1}^{c} \beta_{ii'}^T u(x_i))^{\cdot}$$

Thus to estimate the unknown $\alpha_i$, $\gamma_{i'}$ and $\beta_{ii'}$ can be estimated by the least squares estimators $\hat{\alpha}_i$, $\hat{\gamma}_{i'}$ and $\hat{\beta}_{ii'}$ using data collected from a crossed array design where the outer array is an orthogonal Resolution III main effect plan with each noise factors set at two levels -1 and +1 (corresponding to $\pm 1$ standard deviation). The optimal solution for achieving smallest variation is obtained by minimizing the objective function (1). To obtain an approximate solution for smallest variation, one can minimize with respect to $x_1,...,x_c$, the estimated objective function:

$$\begin{aligned} \hat{h}(x_1,...,x_c) &= \sum_{i'=1}^{n} \hat{V}_{i'}^2 \\ &= (\hat{\gamma}_{i'} + \sum_{i=1}^{c} \hat{\beta}_{ii'}^T u(x_i)). \end{aligned}$$

How is this variance minimization procedure affected if some or all of the two factor noise-noise interactions are in fact non-negligible? It is not difficult to see that in such cases, for given $x_1,...,x_c$ the variance of $y$ differs from (2.1) by a positive term that does not involve $x_1,...,x_c$. Thus one might want to minimize the same function $\hat{h}(x_1,...,x_c)$. However, because the main effects in the outer

array are aliased with certain two factor noise-noise interactions, the estimator $\hat{\gamma}_{i'}$ no longer estimates $\gamma_{i'}$ alone, but the sum of $\gamma_{i'}$ and the effects of the two factor noise-noise interactions in the same alias set. Thus it is not appropriate to minimize directly $\hat{h}(x_1,...,x_c)$ without adjustment. It is proposed here that a follow up $2^n$ factorial (or a faction of $2^n$) experiment of the *n* noise factors be performed to estimate all the two factor noise-noise interactions independently. The estimates obtained are used to correct for bias of the estimated coefficients in the function $\hat{h}(x_1,...,x_c)$. This procedure will be illustrated with the example in next Section.

If for a control factor $x_i$, the vector $\beta_{ii'} = 0$ for all $i' = 1,...,n$, then $x_i$ does not appear in the objective function and the optimal solution does not depend on $x_i$. This kind of control factor is called adjustment factor. Their existence greatly simplifies the procedure of minimizing variance while the mean is made on target, as the variation can first be minimized using the non-adjustment control variables, and then the values of the adjustment variable is set to give the targeted mean value. The identification of adjustment variables can be done by examining the magnitudes of the two factor control-noise interactions using graphical technique such as the half normal probability plot (Box, 1988).

With the present formulation through orthogonal polynomials, one can also examine the sum of squares of the orthogonal contrasts corresponding to these interactions. It is also suggested that the effects of the interactions of each control variable with the noise variables on the results of variance minimization be studied for this purpose.

These approaches will also be illustrated with an example in the next section. If the constant variance in the assumed model is violated, one might have to transform the response variable to attain approximate homogeneity of variances. As explained in Box (1988), the minimization of variance in the transformed metric can be seen as approximately minimizing a performance measure independent of the mean (PerMIA).

## Results

The new methods are outlined to re-analyze the data from a crossed array design, studied by Vandenbrande (2000), using signal-to-noise ratio. The data involve a car body paint spray process in which it is required to spray paint on a plate evenly to a desirable width. Although the surface has to be adequately covered, overspray would result in unnecessarily higher cost in paint as well as causing quality problems on other part of the car body. The response measurement *y* is the width of the paint pattern.

There are four control variables: type of gun $x_1$ (a qualitative variable with values 1, 2 and 3 representing three different guns), paint flow $x_2$, paint airflow $x_3$ and atomizing airflow $x_4$. The last three variables are quantitative and each is set at 3 levels (low, medium and high) which we take to be equally spaced and coded as -1, 0, +1. There are three noise factors: color $z_1$, input air pressure $z_2$, and paint viscosity $z_3$. Each of the three noise factors has two levels: -1 and +1. A Taguchi type of crossed array experiment is performed using the $L_9$ and $L_4$ orthogonal arrays for, respectively, the inner and outer arrays, as displayed in Table 1.

There are therefore 36 experimental runs, determined by crossing the 4 treatment combinations in the outer array with each of the 9 treatment combinations in the inner array. The observed data are given in (Vandenbrande, 1998, 1999).

The first step in the analysis involves defining indicator variables for any qualitative control variables and finding orthogonal polynomials for the quantitative control variables. Here, only type of gun is qualitative and we define $x_{11}$ to be equal to 1 for type 1 and 0 otherwise, $x_{12}$ equal to 1 for type 2 and 0 otherwise. The linear and quadratic orthogonal polynomials used for $x_2$, $x_3$ and $x_4$ are $u_1(x)=x$, $u_2(x)=2-3x^2$.

The coefficients of the linear contrast corresponding to x =-1, 0, +1, are $u_1(x)$=-1, 0, +1, and that of the quadratic contrast corresponding to x =-1, 0 ,+1, are $u_2(x)$=-1,2,-1.

Table 1. Inner and outer array layout

| Inner Array | | | |
|---|---|---|---|
| $x_1$ | $x_2$ | $x_3$ | $x_4$ |
| 1 | 0 | 0 | 0 |
| 1 | 1 | 1 | 1 |
| 1 | -1 | -1 | -1 |
| 2 | -1 | 0 | 1 |
| 2 | 0 | 1 | -1 |
| 2 | 1 | -1 | 0 |
| 3 | -1 | 1 | 0 |
| 3 | 0 | -1 | 1 |
| 3 | 1 | 0 | -1 |
| Outer array | | | |
| $z_1$ | -1 | 1 | 1 |
| $z_2$ | -1 | 1 | -1 |
| $z_3$ | -1 | -1 | 1 |

Our model, suppressing two factor control-control, noise-noise as well as higher order interactions is therefore:

$$y = \mu + (\alpha_{11}x_{11} + \alpha_{12}x_{12}) + \sum_{i=2}^{4}(\alpha_{i1}u_1(x_i) + \alpha_{i2}u_2(x_i)) + \sum_{i'=1}^{3}\gamma_{i'}z_{i'} + \sum_{i'=1}^{3}(\beta_{1i'1}x_{11}z_{i'} + \beta_{1i'2}x_{12}z_{i'}) + \sum_{i=2}^{4}\sum_{i'=1}^{3}(\beta_{ii'1}u_1(x_i)z_{i'} + \beta_{ii'2}u_2(x_i)z_{i'}) + e \quad (2)$$

The least squares estimates of $\alpha_{ij}$, $\gamma_{i'}$ and $\beta_{ii'j}$, $i = 1,...4$, $i' = 1,2,3$, $j = 1,2$, and the broken down sum of squares for each degree of freedom are given in Table 2.

In the second step, one may proceed if desirable to identify adjustment variables which do not interact with any of the noise variables. Specifically, we look for quantitative adjustment variables as these variables can be used to make continuous adjustment of the mean to the target value. By looking at the sum of squares (SS) corresponding to the orthogonal contrasts $u(x_i)z_{i'}$, it is seen that the control factor paint flow $x_2$ has small SS of interactions with all three noise factors. This suggests that using $x_2$ as an adjustment variable and drop it from the variance function (1). The effect of excluding $x_2$ from the study of variance will be examined later.

In step 3, minimize the estimated objective function $\hat{h}$ defined in Section 2, or equivalently, the estimated variance function of $y$ given $x_1, x_3$ and $x_4$. In principle, the mean and variance (treating $z_1, z_2, z_3$ as random along with $e$) of $y$ given $x_1, x_2, x_3$ and $x_4$ can be estimated based on the analytical expression for the mean and variance derived from (3.1). However, an equivalent but more intuitive and easily programmable procedure is to calculate the mean and variance based on generated pseudo observations.

To generate these pseudo observations, we first set a new variable $z_4$ to two levels at -1 and +1 as other noise factors. Also let $\hat{\gamma}_4 = \sqrt{MSE}$. The pseudo observations are generated using (3.1) with the least square estimates replacing the unknown coefficients and also the error $e$ by $\hat{\gamma}_4 z_4$. Here, the $z_i$, i=1,…, 4 can be -1 or +1, yielding a total of $2^4$ pseudo observations. The conditional mean and variance of $y$ given $x_1, x_2, x_3$ and $x_4$ can then be estimated by the usual mean and variance of the pseudo observations (with $2^4$ as the divisor in calculating variance). This procedure is justified as it is equivalent to using Gaussian Quadrature to evaluate the first two moments, and the two point Gaussian Quadrature is known to yield exact integral for polynomial of degree 3.

The added advantage of using the approach of pseudo observations is that it can be readily applied to evaluate any expected loss function $L(y)$, not just the quadratic loss function, by calculating the mean loss at the values of the pseudo observations. This can be particularly helpful if an analytical expression for the expected loss is difficult to obtain.

Table 2. Estimates and sum of squares:

$\hat{y}$ = 39.6 + 1.02 $x_{11}$ - 2.57 $x_{12}$ + 3.84 $u_1(x_2)$+0.604 $u_2(x_2)$ + 3.64 $u_1(x_3)$-1.69 $u_2(x_3)$ -2.99 $u_1(x_4)$ +1.37 $u_2(x_4)$ -3.63 $z_1$+ 0.308 $z_2$ - 0.0417 $z_3$ + 3.48 $x_{11}$ $z_1$ + 2.58 $x_{12}$ $z_1$ + 0.550 $x_{11}$ $z_2$ - 0.0500 $x_{12}$ $z_2$ - 1.15 $x_{11}$ $z_3$ + 0.233 $x_{12}$ $z_3$ - 0.0125 $u_1(x_2)$ $z_1$ + 0.0931 $u_2(x_2)$ $z_1$+0.438 $u_1(x_2)z_2$ +0.121 $u_2(x_2)$ $z_2$-0.221 $u_1(x_2)z_3$+0.290 $u_2(x_2)$ $z_3$ -1.46 $u_1(x_3)$ $z_1$-0.253 $u_2(x_3)$ $z_1$-0.550 $u_1(x_3)$ $z_2$+0.717 $u_2(x_3)$ $z_2$ 0.783 $u_1(x_3)$ $z_3$ - 0.889 $u_2(x_3)$ $z_3$+1.73 $u_1(x_4)$ $z_1$-0.519 $u_2(x_4)$ $z_1$-1.08 $u_1(x_4)$ $z_2$-0.717 $u_2(x_4)$ $z_2$ + 0.850 $u_1(x_4)$ $z_3$ + 0.369 $u_2(x_4)$ $z_3$.

| Control factor $x_2$ | | Control factor $x_3$ | | Control factor $x_4$ | |
|---|---|---|---|---|---|
| Effects | Sum of squares | Effects | Sum of squares | Effects | Sum of squares |
| $u_1(x_2)$ $z_1$ | 0.004 | $u_1(x_3)$ $z_1$ | 51.042 | $u_1(x_4)$ $z_1$ | 72.107 |
| $u_2(x_2)$ $z_1$ | 0 .623 | $u_2(x_3)$ $z_1$ | 4.601 | $u_2(x_4)$ $z_1$ | 19.427 |
| $u_1(x_2)z_2$ | 4.594 | $u_1(x_3)$ $z_2$ | 7.260 | $u_1(x_4)$ $z_2$ | 27.735 |
| $u_2(x_2)$ $z_2$ | 1.051 | $u_2(x_3)$ $z_2$ | 36.980 | $u_2(x_4)$ $z_2$ | 36.980 |
| $u_1(x_2)z_3$ | 1.170 | $u_1(x_3)$ $z_3$ | 14.727 | $u_1(x_4)$ $z_3$ | 17.340 |
| $u_2(x_2)$ $z_3$ | 6.067 | $u_2(x_3)$ $z_3$ | 56.889 | $u_2(x_4)$ $z_3$ | 9.827 |

Table 3 gives the estimated standard deviation (column (1)) for all 27 treatment combinations of $x_1, x_3$, and $x_4$. The combination $x_1 = 3,\ x_3 = -1,\ x_4 = 1,$ yields the smallest value of standard deviation of 1.6. However, because of practical consideration, high atomizing air must be combined with somewhat higher fan air.

One might consider the next best combination at $x_1 = 1,\ x_3 = -1,\ x_4 = 0,$ with an estimated standard deviation of 1.8. The use of orthogonal polynomials allows interpolation to obtain improved results at $x_1 = 1,\ x_3 = -1.1,$ $x_4 = -0.4,$ yielding a smaller standard deviation of 1.6. The last few columns of Table 3 give the mean and standard deviation for each of $x_2$ = -1, 0, +1 when $x_2$ is also included in the variance analysis. The difference in standard deviations from column (1) is minimal.

Furthermore, if a target mean of 45 is desired, then $x_2$ should be set around $x_2$ = 1. As pointed out in the last section, the procedure of minimizing variance can be adversely affected if some of the two factor noise-noise interactions are non-zero. Thus we suggest, as a safeguard against this potential problem by assessing these interactions with small number of additional experimental runs. In the present example, each

Table 3. Means and standard deviations

| | | | | | $x_2 = -1$ | | $x_2 = 0$ | | $x_2 = +1$ | |
|---|---|---|---|---|---|---|---|---|---|---|
| $x_1$ | $x_3$ | $x_4$ | (1) | (2) | mean | SD | mean | SD | mean | SD |
| 1 | -1 | -1 | 3.7 | 3.5 | 35.8 | 3.1 | 41.5 | 3.3 | 43.5 | 4.0 |
| 1 | -1 | 0 | 1.8 | 3.0 | 36.9 | 1.4 | 42.6 | 0.9 | 44.6 | 1.0 |
| 1 | -1 | 1 | 4.2 | 4.7 | 29.8 | 3.8 | 35.5 | 4.0 | 37.5 | 3.9 |
| 1 | 0 | -1 | 6.3 | 5.6 | 34.4 | 5.9 | 40.0 | 5.9 | 42.1 | 6.7 |
| 1 | 0 | 0 | 3.3 | 3.1 | 35.5 | 2.9 | 41.2 | 2.5 | 43.2 | 3.5 |
| 1 | 0 | 1 | 3.8 | 3.5 | 28.4 | 3.3 | 34.1 | 3.4 | 36.1 | 4.0 |
| 1 | 1 | -1 | 3.4 | 3.9 | 43.1 | 2.9 | 48.8 | 2.9 | 50.8 | 3.4 |
| 1 | 1 | 0 | 3.6 | 4.9 | 44.2 | 3.6 | 49.9 | 3.3 | 51.9 | 3.0 |
| 1 | 1 | 1 | 2.1 | 3.9 | 37.1 | 1.8 | 42.8 | 2.0 | 44.8 | 1.0 |
| 2 | -1 | -1 | 2.6 | 3.2 | 32.2 | 1.8 | 37.9 | 2.2 | 39.9 | 2.7 |
| 2 | -1 | 0 | 2.3 | 4.0 | 33.4 | 2.2 | 39.0 | 2.0 | 41.0 | 1.2 |
| 2 | -1 | 1 | 3.4 | 4.6 | 26.3 | 3.0 | 31.9 | 3.4 | 33.9 | 2.8 |
| 2 | 0 | -1 | 5.5 | 5.1 | 30.8 | 5.0 | 36.5 | 5.0 | 38.5 | 5.8 |
| 2 | 0 | 0 | 3.1 | 3.7 | 31.9 | 2.8 | 37.6 | 2.5 | 39.6 | 3.1 |
| 2 | 0 | 1 | 2.4 | 2.9 | 24.8 | 1.5 | 30.5 | 1.9 | 32.5 | 2.4 |
| 2 | 1 | -1 | 3.9 | 5.0 | 39.5 | 3.6 | 45.2 | 3.6 | 47.2 | 3.7 |
| 2 | 1 | 0 | 5.1 | 6.5 | 40.6 | 5.2 | 46.3 | 5.0 | 48.3 | 4.5 |
| 2 | 1 | 1 | 3.1 | 5.0 | 33.5 | 3.0 | 39.2 | 3.1 | 41.2 | 2.1 |
| 3 | -1 | -1 | 4.1 | 4.4 | 34.8 | 3.7 | 40.5 | 3.7 | 42.5 | 4.3 |
| 3 | -1 | 0 | 3.6 | 4.8 | 35.9 | 3.6 | 41.6 | 3.3 | 43.6 | 3.2 |
| 3 | -1 | 1 | 1.6 | 3.4 | 28.8 | 0.9 | 34.5 | 1.2 | 36.5 | 0.3 |
| 3 | 0 | -1 | 7.2 | 6.9 | 33.4 | 6.9 | 39.0 | 6.8 | 41.1 | 7.6 |
| 3 | 0 | 0 | 5.5 | 5.8 | 34.5 | 5.4 | 40.1 | 5.0 | 42.2 | 5.5 |
| 3 | 0 | 1 | 3.1 | 3.4 | 27.4 | 2.5 | 33.0 | 2.5 | 35.1 | 3.2 |
| 3 | 1 | -1 | 6.3 | 6.9 | 42.1 | 6.1 | 47.7 | 6.0 | 49.8 | 6.3 |
| 3 | 1 | 0 | 6.9 | 8.0 | 43.2 | 7.0 | 48.9 | 6.8 | 50.9 | 6.6 |
| 3 | 1 | 1 | 3.9 | 5.5 | 36.1 | 3.8 | 41.8 | 3.8 | 43.8 | 3.3 |

main effect in the outer array is aliased with the interaction between the remaining two noise factors. For instance, the coefficient $\hat{\gamma}_3$ of the noise factor "viscosity" is small, but since $z_3$ is aliased with $z_1z_2$, it actually estimates the sum of $\gamma_3 + \gamma_{12}$, where $\gamma_{12}$ is the coefficient of $z_1z_2$.

In the last step, we propose to have a $2^2$ factorial (or a factional factorial so that the interactions suspected to be significant are estimable) of the noise factors conducted at the solution obtained in step 3, i.e. $x_1 = 1$, $x_3 = -1.1$, $x_4 = -0.4$. To estimate $\gamma_{12}$, first subject the fitted value based on (3.1) from each of the $y$ values from the new experiment and estimate $\gamma_{12}$ by the slope of the regression of the adjusted $y$ on $z_1z_2 - z_3$.

As an illustrative example, suppose an estimate $\hat{\gamma}_{12} = -1.855$ is obtained. Then the coefficient $\gamma_3$ can be re-estimated as -0.042-(-1.855) = 1.813. Column (2) of Table 3 now gives the standard deviations based on the new model (model (2) together with the additional term $\gamma_{12}z_1z_2$). The results are markedly different from column (1), and the smallest value no longer occurs at $x_1 = 3$, $x_3 = -1$, $x_4 = 1$, suggesting that such adjustment might be necessary.

## Conclusion

We have suggested in this article a systematic approach in analyzing crossed array designs, where fractional factorial design may be employed in the outer array. This kind of

designs is still popular because of its simplicity and its greater robustness than combined array designs to certain type of model mis-specification. It is however demonstrated that non-ignorable noise-noise interactions may still create problems with the crossed array design. A method of rectifying these difficulties is proposed, but the problem of finding cost effective follow up design to complement the original design is worth studying.

Our approach also assumes the constant variance assumption conditional on values of both the control and noise factors. If this assumption is violated, the response variable may have to be transformed to attain constant variances before the suggested analysis can be carried out.

Alternatively, the use of generalized linear model (Nelder and Lee, 1991) or the approach of Engel (1982) may also be appropriate. The choice of an appropriate transformation may be facilitated using the graphical plot of Box (1988), or the analysis of Chan and Mak (1997). However, even if the quadratic loss function is used in the original metric, the induced loss function in the transformed scale is no longer quadratic. In this case, the expected loss can be approximated using the idea of pseudo observations. This approach is equivalent to using Gaussian Quadrature to carry out the integration in computing the expected loss. As is well known the approximation can be improved by using more data points for the noise factors in generating the pseudo observations. Details will not be given here.

## References

Box, G.E.P. (1988). Signal-to-noise ratios, performance criteria and transformation. *Technometrics*, *30,* 1-31.

Barreau, A., Chassagnon, R., Kobi, & A., Seibilia, B. (1999). Taguchi's parameter design: an improved alternative approach, *53rd Annual quality congress proceedings*, Milwaukee, WI: ASQ., 400-404.

Chan, L.K., & Mak, T.K. (1995). A regression approach for discovering small variation around a target. *Applied Statistics*, *44*, 369-377.

Engel, J. (1992). Modeling variation in industrial experiments. *Applied Statistics*, *41*, 579-593.

Engel, J., & Huele, A. F. (1996). A generalized linear modeling approach to robust design. *Technometrics, 38***,** 365-373.Myers, R. H., Kim, Y., & Griffiths, K. L. (1997). Response surface methods and the use of noise variables. *Journal of Quality Technology, 29*, 429-440.

Nelder, J. A., & Lee, Y. (1991). Generalized linear models for the analysis of Taguchi-type experiments. *Applied Stochastic Models and Data analysis*, *7*, 107-120.

O'Donnell, E. M., & Vining, G. G. (1997). Mean squared error of prediction approach to the analysis of a combined array. *Journal of Applied Statistics*, *24*. 733-746.

Shoemaker, A. C., Tsui, K. L., & Wu, C. F. (1991). Economical experimentation methods for robust design. *Technometrics,33,* 415-427.

Vandenbrande, W. (2000). Make love, not war: combining DOE and Taguchi, *54th Annual quality congress proceedings*, Milwaukee, WI: ASQ., 450-456.

Vandenbrande, W. (1998). SPC in paint application: Mission impossible, *52nd Annual quality congress proceedings*, Milwaukee, WI: ASQ., 708-715.

Welch, W. J., Yu, T. K., Kang, S. M., & Sacks, J. (1990). Computer experiments for quality control by parameter design. *Journal of Quality Technology, 22,* 15-22.

Journal of Modern Applied Statistical Methods
November, 2005, Vol. 4, No. 2, 402-407

1538 – 9472/05/$95.00

# Testing Normality Against The Laplace Distribution

Taisuke Otsu
Cowles Foundation
Yale University

Some normality test statistics are proposed by testing non-nested hypotheses of the normal distribution and the Laplace distribution. If the null hypothesis is normal, the proposed non-nested tests are asymptotically equivalent to Geary's (1935) normality test. The proposed test statistics are compared by the method of approximate slopes and Monte Carlo experiments.

Key words: Normality test; non-nested hypothesis; Cox test; Atkinson test

## Introduction

In statistical analysis, many models and methods rely upon the assumption of normality, which should be examined by some adequate tests. However, in several data (e.g. economic and financial data), the existence of outliers is much frequent, and the observations or disturbances may have some leptokurtic distributions, where the kurtosis is larger than three. In order to detect such leptokurtic non-normal distributions, we apply the method of non-nested testing which has high sensitivity (power) for an explicit alternative hypothesis.

Based on Cox (1961, 1962) and Atkinson (1970), it this article non-nested test statistics between the normal distribution and the Laplace (or double-exponential) distribution, which is a typical leptokurtic distribution are proposed. All of the proposed test statistics are asymptotically normal. When the null hypothesis is normal, these test statistics are asymptotically equivalent to Geary's (1935) normality test statistic.

In the context of regression models, the maximum likelihood estimator with the Laplace distribution error is the least absolute deviation (LAD) estimator. Therefore, these test statistics are also useful to decide whether the LAD regression or the conventional OLS regression should be applied.

By applying Pesaran's (1987) strict definition of non-nested hypotheses, we find that the normal distribution and the Laplace distribution are globally non-nested, and that the power analysis using Pitman-type local alternatives is not available. Therefore, these non-nested test statistics are compared by the method of approximate slope (or Bahadur efficiency) developed by Bahadur (1960, 1967). Furthermore, Monte Carlo simulations are carried out to compare the small sample properties of the proposed tests and other conventional normality tests. Simulation results indicate that these tests show reasonable performances in terms of the size and power.

Taisuke Otsu is an Assistant Professor of Economics. His research interests are in empirical likelihood, nonparametric and semiparametric methods, and microeconometrics. Contact him at 30 Hillhouse Ave., Rm. 36,Department of Economics, Yale University,P.O. Box 208281, New Haven, CT 06520-8281, or taisuke.otsu@yale.edu.

## Non-nested Test Statistics

Throughout this article, demeaned observations are considered, i.e., the mean is assumed to be zero. Let $Y = (Y_1, \ldots, Y_n)$ be independently and identically distributed (iid)

random variables. Consider the following non-nested hypotheses:

$$H_f : f(y;\alpha) = \frac{1}{\sqrt{2\pi\alpha}} \exp\left[ -\frac{y^2}{2\alpha} \right], \quad (1)$$

$$H_g : g(y;\beta) = \frac{1}{2\beta} \exp\left[ -\frac{|y|}{\beta} \right], \quad (2)$$

where $H_f$ is the normal distribution with zero mean, and $H_g$ is the Laplace distribution with zero mean. $H_f$ and $H_g$ belong to separate parametric families and are called non-nested hypotheses. In order to test non-nested hypotheses, Cox (1961, 1962) proposed a testing procedure based on a modified likelihood ratio. When $H_f$ is the null hypothesis and $H_g$ is the alternative hypothesis, the Cox test statistic is written as

$$T_f = L_f(\hat{\alpha}) - L_g(\hat{\beta}) - \mathrm{E}_{\hat{\alpha}}(L_f(\alpha) - L_g(\beta_\alpha)), \quad (3)$$

where $L_f(\alpha) = \sum_{i=1}^{n} \log f(y_i;\alpha)$ and $L_g(\beta) = \sum_{i=1}^{n} \log g(y_i;\beta)$ denotes the log likelihood functions of the hypotheses $H_f$ and $H_g$, respectively, $\hat{\alpha}$ and $\hat{\beta}$ denote the maximum likelihood estimators under $H_f$ and $H_g$, respectively, $\mathrm{E}_{\hat{\alpha}}(\cdot)$ is the expected value under $H_f$ when $\alpha$ takes the value $\hat{\alpha}$, and $\beta_\alpha = \mathrm{plim}_\alpha \hat{\beta}$ is the probability limit of $\hat{\beta}$ under $H_f$ as $n \to \infty$. Define

$$F_i = \log f(Y_i;\alpha),\ G_i = \log g(Y_i;\beta_\alpha),$$
$$F_{\alpha i} = \frac{\partial \log f(Y_i;\alpha)}{\partial \alpha}. \quad (4)$$

Cox (1961, 1962) showed that $T_f$ is asymptotically normal with zero mean and variance

$$\mathrm{V}_\alpha(T_f) = n\left[ \mathrm{V}_\alpha(F_i - G_i) - \frac{\mathrm{C}_\alpha^2(F_i - G_i, F_{\alpha i})}{\mathrm{V}_\alpha(F_{\alpha i})} \right], \quad (5)$$

where $V_\alpha(\cdot)$ and $\mathrm{C}_\alpha(\cdot,\cdot)$ denote the variance and the covariance under $H_f$, respectively.

In the same manner, set the Laplace distribution $H_g$ as the null hypothesis and set the normal distribution $H_f$ as the alternative hypothesis. In this case, the Cox test statistic $T_g$ is written as

$$T_g = L_g(\hat{\beta}) - L_f(\hat{\alpha}) - \mathrm{E}_{\hat{\beta}}(L_g(\hat{\beta}) - L_f(\alpha_\beta)), \quad (6)$$

where $\mathrm{E}_{\hat{\beta}}(\cdot)$ is the expected value under $H_g$ when $\beta$ takes the value $\hat{\beta}$, and $\alpha_\beta = \mathrm{plim}_\beta \hat{\alpha}$ is the probability limit of $\hat{\alpha}$ under $H_g$ as $n \to \infty$. $T_g$ is also asymptotically normal with zero mean and variance $\mathrm{V}_\beta(T_g)$, which is defined in the same manner as (4). If $\mathrm{V}_\alpha(T_f)$ and $\mathrm{V}_\beta(T_g)$ are consistently estimated by $\mathrm{V}_{\hat{\alpha}}(T_f)$ and $\mathrm{V}_{\hat{\beta}}(T_g)$, respectively,

$$N_f = T_f / \sqrt{\mathrm{V}_{\hat{\alpha}}(T_f)},\ N_g = T_g / \sqrt{\mathrm{V}_{\hat{\beta}}(T_g)} \quad (7)$$

can be used as test statistics which follow the standard normal limiting distribution.

In setup (1) and (2), obtain

$$\hat{\alpha} = \textstyle\sum_i Y_i^2 / n,\ \hat{\beta} = \sum_i |Y_i| / n, \quad (8)$$

$$\beta_\alpha = \mathrm{plim}_\alpha \hat{\beta} = \mathrm{E}_\alpha(|Y_i|) = \sqrt{2\alpha/\pi},$$
$$\alpha_\beta = \mathrm{plim}_\beta \hat{\alpha} = \mathrm{E}_\beta(Y_i^2) = 2\beta^2. \quad (9)$$

Therefore, when the null hypothesis is normal and the alternative hypothesis is Laplace, the Cox test statistic is

$$T_f = n\log\left(\frac{\hat{\beta}}{\beta_{\hat{\alpha}}}\right) = n\log\left(\sqrt{\frac{\pi}{2}}\frac{\hat{\beta}}{\sqrt{\hat{\alpha}}}\right), \qquad (10)$$

with the asymptotic variance $\mathrm{V}_{\alpha}(T_f) = \frac{\pi}{2} - \frac{3}{2}$. On the other hand, when the null hypothesis is Laplace and the alternative hypothesis is normal, the Cox test statistic is

$$T_g = \frac{n}{2}\log\left(\frac{\hat{\alpha}}{\alpha_{\hat{\beta}}}\right) = \frac{n}{2}\log\left(\frac{\hat{\alpha}}{2\beta^2}\right), \qquad (11)$$

with the asymptotic variance $\mathrm{V}_{\beta}(T_g) = \frac{1}{4}$.

Next, derive Atkinson's (1970) test. The Atkinson test procedure is derived from the comprehensive probability density function (pdf), which includes $f(y;\alpha)$ and $g(y;\beta)$ as special cases. When $H_f$ is the null hypothesis and $H_g$ is the alternative hypothesis, the Atkinson test statistic is written as

$$TA_f = L_f(\hat{\alpha}) - L_g(\beta_{\hat{\alpha}}) - \mathrm{E}_{\hat{\alpha}}(L_f(\alpha) - L_g(\beta_{\alpha})). \qquad (12)$$

Comparing (3) and (12), the difference between $T_f$ and $TA_f$ is their second terms. Because the Atkinson test $TA_f$ and the Cox test $T_f$ are asymptotically equivalent under $H_f$, the asymptotic variance of $TA_f$ is same as (5) (see Pereira, 1977). Analogous results are obtained for the case where $H_g$ is the null hypothesis and $H_f$ is the alternative hypothesis. In order to conduct the Atkinson test, we can use

$$NA_f = TA_f/\sqrt{\mathrm{V}_{\hat{\alpha}}(T_f)}, \; NA_g = TA_g/\sqrt{\mathrm{V}_{\hat{\beta}}(T_g)} \qquad (13)$$

as test statistics which follow the standard normal limiting distribution. When the null hypothesis is normal and the alternative hypothesis is Laplace, the Atkinson test statistic is:

$$TA_f = n\left(\frac{\hat{\beta}}{\beta_{\hat{\alpha}}} - 1\right) = n\left(\sqrt{\frac{\pi}{2}}\frac{\hat{\beta}}{\sqrt{\hat{\alpha}}} - 1\right), \qquad (14)$$

and when the null hypothesis is Laplace and the alternative hypothesis is normal, the Atkinson test statistic is

$$TA_g = \frac{n}{2}\left(\frac{\hat{\alpha}}{\alpha_{\hat{\beta}}} - 1\right) = \frac{n}{2}\left(\frac{\hat{\alpha}}{2\beta^2} - 1\right). \qquad (15)$$

Because the computation of our non-nested test statistics (i.e., $N_f$, $N_g$, $NA_f$, and $NA_g$) needs only $\hat{\alpha}$ and $\hat{\beta}$, their implementation is quite easy.

$T_f$ and $TA_f$ are related to another normality test suggested by Geary (1935). The Geary test statistic is written as

$$G = \frac{\sum_i |Y_i|}{\sqrt{n\sum_i Y_i^2}} = \frac{\hat{\beta}}{\sqrt{\hat{\alpha}}}, \qquad (16)$$

From (10) and (14), the relationships among $G$, $T_f$, and $TA_f$ are

$$T_f = n\log\left(\sqrt{\frac{\pi}{2}}G\right), \; TA_f = n\left(\sqrt{\frac{\pi}{2}}G - 1\right). \qquad (17)$$

Therefore, if the standardized test statistics is compared, it can be shown that under $H_f$ the Cox test and the Atkinson test are asymptotically equivalent to the Geary test.

Power Comparison

This section considers theoretical properties of the proposed non-nested tests. We first investigate the consistency of the Cox test and the Atkinson test. Pereira (1977) showed that the Cox test is always consistent, but the Atkinson test is not always consistent. From (14) and (15):

$$\text{plim}_{\beta} n^{-1} TA_f = \sqrt{\pi}/2 - 1 \approx -0.1138, \quad (18)$$

$$\text{plim}_{\alpha} n^{-1} TA_g = (1/2)(\pi/4 - 1) \approx -0.1073. \quad (19)$$

Because both $TA_f$ and $TA_g$ converge to non-zero constants, the Atkinson test is consistent in our particular setup.

Using Pesaran's (1987) strict definition of the non-nested hypotheses, which is based upon the Kullback-Leibler information criterion (KLIC), next examine the relationship between the normal distribution ($H_f$) and the Laplace distribution ($H_g$). The KLIC for the pdf $f(y;\alpha)$ against the pdf $g(y;\beta)$ is defined as

$$I_{fg}(\alpha,\beta) = E_{\alpha}(\log f(y;\alpha) - \log g(y;\beta)). \quad (20)$$

Assume that $I_{fg}(\alpha,\beta)$ has a unique minimum at $\beta_*(\alpha)$. Pesaran (1987) defined the closeness of $H_g$ to $H_f$ as

$$C_{fg}(\alpha) = I_{fg}(\alpha, \beta_*(\alpha)). \quad (21)$$

Similarly, define the KLIC for $g(y;\beta)$ against $f(y;\alpha)$ (denote $I_{gf}(\beta,\alpha)$) and the closeness of $H_f$ to $H_g$ (denote $C_{gf}(\beta)$). Using $C_{fg}(\alpha)$ and $C_{gf}(\beta)$, Pesaran (1987) classified the relationship between two hypotheses into three categories, i.e., nested, globally non-nested, and partially non-nested. In the case of (1) and (2), $I_{fg}(\alpha,\beta)$ and $I_{gf}(\beta,\alpha)$ are written as

$$I_{fg}(\alpha,\beta) = -\frac{1}{2}\log(2\pi\alpha) + \log(2\beta) + \frac{1}{\beta}\sqrt{\frac{2\alpha}{\pi}} - \frac{1}{2}, \quad (22)$$

$$I_{gf}(\beta,\alpha) = \frac{1}{2}\log(2\pi\alpha) - \log(2\beta) + \frac{\beta^2}{\alpha} - 1. \quad (23)$$

Because $\beta_*(\alpha) = \sqrt{2\alpha/\pi}$ and $\alpha_*(\beta) = 2\beta^2$,

$$C_{fg}(\alpha) = \log\left(\frac{2}{\pi}\right) + \frac{1}{2} \approx 0.04842, \quad (24)$$

$$C_{gf}(\beta) = \log(\sqrt{\pi}) - \frac{1}{2} \approx 0.07236. \quad (25)$$

Because both $C_{fg}(\alpha)$ and $C_{gf}(\beta)$ are nonzero constants, $H_f$ and $H_g$ are globally non-nested and the power analysis using a local alternative is not available (see Pesaran (1987)).

Because the Pitman-type power analysis cannot be applied, compare the Cox test and the Atkinson test by the method of approximate slopes developed by Bahadur (1960, 1967). The method of approximate slopes compares the convergence rates of the significance levels of tests (to zero) under some fixed alternative hypothesis with some fixed power.

Thus, approximate slopes are useful to analyze the power properties of tests under globally non-nested hypotheses. Let $\tilde{\alpha}_n$ be the asymptotic significance level of some test with a given sample size $n$. The approximate slope is defined as $\lim(-2n^{-1}\log\tilde{\alpha}_n)$. If a test $T_1$ has a greater approximate slope than another test $T_2$, we call that $T_1$ is Bahadur efficient relative to $T_2$. Pesaran (1984) showed that the approximate slopes of the Cox test and the Atkinson test are given by $\text{plim}_{\beta}(n^{-1}N_f^2)$ and $\text{plim}_{\beta}(n^{-1}NA_f^2)$, respectively. Therefore, from (10), (11), (14), and (15),

$$\text{plim}_{\beta} n^{-1} N_f^2 = \frac{\left(\log\left(\frac{\sqrt{\pi}}{2}\right)\right)^2}{\frac{\pi}{2} - \frac{3}{2}} \approx 0.2061, \quad (26)$$

$$\text{plim}_{\beta} n^{-1} NA_f^2 = \frac{\left(\frac{\sqrt{\pi}}{2} - 1\right)^2}{\frac{\pi}{2} - \frac{3}{2}} \approx 0.1828, \quad (27)$$

$$\text{plim}_{\alpha} n^{-1} N_g^2 = \left(\log\left(\frac{\pi}{4}\right)\right)^2 \approx 0.05835, \quad (28)$$

Table 1. Finite sample rejection frequencies of the null hypothesis at the one side 5% level

| DGP | $n$ | $T_f$ | $T_g$ | $TA_f$ | $TA_g$ | BS | SW | DA | AD |
|---|---|---|---|---|---|---|---|---|---|
| | 20 | 0.0429 | 0.1812 | 0.0368 | 0.0239 | 0.0234 | 0.0469 | 0.0526 | 0.0512 |
| Normal | 50 | 0.0451 | 0.6167 | 0.0410 | 0.4438 | 0.0353 | 0.0494 | 0.0488 | 0.0509 |
| | 100 | 0.0498 | 0.9291 | 0.0469 | 0.8875 | 0.0434 | 0.0484 | 0.0525 | 0.0522 |
| | 20 | 0.3427 | 0.0311 | 0.3012 | 0.0014 | 0.2118 | 0.2498 | 0.3556 | 0.2663 |
| Laplace | 50 | 0.7072 | 0.0418 | 0.6945 | 0.0190 | 0.5107 | 0.4105 | 0.6927 | 0.5498 |
| | 100 | 0.9377 | 0.0460 | 0.9339 | 0.0254 | 0.7783 | 0.5386 | 0.9175 | 0.8265 |
| | 20 | 0.1184 | 0.0995 | 0.1066 | 0.0108 | 0.0931 | 0.1102 | 0.1497 | 0.1052 |
| Logistic | 50 | 0.2549 | 0.2859 | 0.2428 | 0.1678 | 0.2313 | 0.1459 | 0.2984 | 0.1682 |
| | 100 | 0.4072 | 0.5356 | 0.3957 | 0.4512 | 0.3673 | 0.1289 | 0.4531 | 0.2367 |

$$\text{plim}_{\alpha} n^{-1} NA_g^2 = \left(\frac{\pi}{4} - 1\right)^2 \approx 0.04605. \qquad (29)$$

In both cases (i.e., the null is normal, and the null is Laplace), the Cox test is Bahadur efficient relative to the Atkinson test. Thus, the Cox test has better global power property than the Atkinson test.

## Results

In order to analyze the finite sample properties of the proposed tests, we conduct Monte Carlo simulation. In addition to the non-nested test statistics in (10), (11), (14), and (15), consider the normality tests by Bowman and Shenton (1975) (BS), Shapiro and Wilk (1965) (SW), D'Agostino (1971) (DA) and Anderson and Darling (1954) (AD), which is a modified Kolmogorov-Smirnov test, as alternative tests.

As the data generating process (DGP), employ the standard normal, standard Laplace, and standard logistic distribution. The sample sizes are set as $n = (20, 50, 100)$. The number of replications is 10000.

Table 1 shows finite sample rejection frequencies of the null hypothesis at the 5% level. From this table, the following may be seen. First, the Cox test $T_f$ with the normal null hypothesis demonstrates better performances than the Atkinson test $TA_f$ in terms of the size accuracy and power. This power superiority of $T_f$ is consistent with the relative Bahadur efficiency of $T_f$. Second, comparing to the other normality tests, $T_f$ has the highest power when the DGP is the standard Laplace distribution. Also $T_f$ is second best when the DGP is the

logistic distribution. Third, the Atkinson test $TA_g$ with the Laplace null hypothesis shows enough power when the DGP is the standard normal distribution. Note that $T_g$ and $TA_g$ can provide additional information, which cannot be obtained by the conventional normality tests based on the normal null hypothesis.

## Conclusion

By applying the Cox and Atkinson test, we propose the non-nested test statistics of the normal and the Laplace distribution. The proposed test statistics proposed are asymptotically normal, and are easily computed. Approximate slopes show that the Cox test has better power properties than the Atkinson test. In simulation, the Cox test with the normal null hypothesis shows higher power for leptokurtic distributions comparing to the other normality tests. The Atkinson test with the Laplace null hypothesis is also useful to analyze distributional forms of data.

## References

Anderson, T. W. & D. A. Darling (1954). A test of goodness-of-fit. *Journal of the American Statistical Association 49*, 765-769.

Atkinson, A. C. (1970). A method for discriminating between models. *Journal of the Royal Statistical Society B32*, 323-353.

Bahadur, R. R. (1960). Stochastic comparison of tests. *Annals of Mathematical Statistics 31*, 276-295.

Bahadur, R. R. (1967). Rates of convergence of estimates and test statistics, *Annals of Mathematical Statistics 38*, 303-324.

Bowman, K.O. & Shenton, B. R. (1975). Omnibus test contours for departures from normality based on $\sqrt{b_1}$ and $b_2$. *Biometrika, 62*, 243-50.

Cox, D. R. (1961). Tests of separate families of hypotheses. In: *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability, Vol. 1*. University of California Press, Berkeley, p. 105-123.

Cox, D. R. (1962). Further results on tests of separate families of hypotheses. *Journal of the Royal Statistical Society B24*, 406-424.

D'Agostino, R. B. (1971). An omnibus test of normality for moderate and large size samples. *Biometrika 58*, 341-348.

Geary, R. C. (1935). The ratio of the mean deviation to the standard deviation as a test of normality. *Biometrika 27*, 310-332.

McAleer, M. (1986). Specification tests for separate models: a survey, Ch. 9 in King, M.L. & Giles, D.E.A. (ed.), *Specification analysis in the linear model*, London: Routledge and Kegan Paul, p. 146-196.

Pereira, B. B. (1977). A note on the consistency and on the finite sample comparisons of some tests of separate families of hypotheses. *Biometrika 64*, 109-113.

Pesaran, M. H. (1984). Asymptotic power comparisons of tests of separate families by Bahadur's approach. *Biometrika 71*, 245-252.

Pesaran, M. H. (1987). Global and partial non-nested hypotheses and asymptotic local power. *Econometric Theory 3*, 69-97.

Shapiro, S. S. & Wilk, M. B. (1965). An analysis of variance test for normality. *Biometrika, 52*, 591-611.

Journal of Modern Applied Statistical Methods
November, 2005, Vol. 4, No. 2, 408-424

1538 – 9472/05/$95.00

# A Discretized Approach to Flexibly Fit Generalized Lambda Distributions to Data

Steve Su
Epi-stat Division, George Institute for International Health
Sydney, New South Wales, Australia

This article presents a flexible approach to fit statistical distribution to data. It optimizes the bin-width of data histogram to find a suitable generalized lambda distribution. In addition to the default optimization, this approach provides additional flexibility akin to the concepts of loess and kernel smoothing, which allow the users to determine the amount of details they would like to smooth over the data. The approach presented in this article will allow users to visually compare and choose the parameters of generalized lambda distribution that best suit their purposes of study.

Key words: generalized lambda distributions, quantile distributions, fitting distributions to data

## Introduction

An essential problem in data analysis is to find a probability distribution that will adequately fit the empirical data. Considerable literature exists in this area, ranging from the parametric work of generalized lambda distribution (Ramberg & Schmeriser, 1974; Ramberg, Tadikamalla, Dudewicz & Mykytka, 1979; Ozturk & Dale, 1985; Freimer, Mudholkar, Kollia, & Lin, 1988; Okur, 1988; King & MacGillivray, 1999; Karian & Dudewicz, 2000; Lakhany & Massuer, 2000) to nonparametric work of kernel density estimation (Silverman, 1985). In spite of these works, no current work exists on allowing a range of possible generalized lambda distribution (GλD) fits to data, pending on users' desire to suppress or accentuate certain features of the data based on prior knowledge of the distribution. This is important when a particular method fails to provide a fit that highlights the essential features of the data exhibited and known by the analyst. In these situations, it will often be preferable to explore other plausible GλDs.

Steve Yu Shuo Su is a Research Fellow at the Epi-stat Division of the George Institute, affiliated with the University of Sydney. His research interests are in applied statistical methods in business and epidemiology. Email: ssu@thegeorgeinstitute.org.

This article proposes an extension of the existing fitting method using GλD which offers more flexibility and in many cases can highlight features of the data not considered by the King and MacGillivray (1999)'s starship method. Instead of optimizing using goodness of fit method, this article suggests an alternative approach which is to optimize based on the number of classes or bins of the data. The number of bins of the data can be determined by the user, offering flexibility to suppress or highlight details, much like the concept of smoothing a data set using different weights in loess or kernel smoothing. This is a valuable tool in practice because the real distribution of the data set is almost never known and the methods developed in this article can be used to conduct sensitivity analysis to assess the effects of using different yet plausible distributions.

The principal emphasis in this article is to allow the user to fit a wide range of different distributions to data set rather than to satisfy the goodness of fit statistics. Also, the exclusive use of goodness of fit statistics in the fitting of distribution to data as was done in previous works (King & MacGillivray, 1999; Lakhany & Massuer, 2000) does not guarantee the resulting distribution fit will satisfy the goodness of fit, but merely tries to maximize it. The beauty of the approach in this article is that it allows the data to be represented in different angles. This is important because unlike theoretical simulated data, real life data is often messy. Very often,

real life data does not have a nice continuous range of values one can get from theoretical simulations. Due to this imperfection, it is often desirable to have an alternative data fitting method that could provide alternative fits beyond the traditional goodness of fit methods. This will give the user a possible range of distribution fits that could arise from the data set and this can lead to valuable sensitivity analysis on the impact of different distributions. The use of goodness of fit criteria could also enhance the credibility of fit under different fits but should not discredit it. This is because it is only possible to test the goodness of fit of one realization of the real life data from its underlying distribution, which may or may not be representative.

The article begins with a literature review on the existing methods of GλD parameters estimation, which progressively result in the development of this new method. Results of the application of the new methods on real life data are then presented and the article concludes with a discussion on the shortcomings of this new method.

Review of Literature

This literature review begins with the basic theory of GλD and discusses some of the fitting methods reported in literature. The literature review then presents two methods that appear to give promising results. These two methods are extended and discussed in the method section.

The Ramberg-Schmeiser (1974) (RS) GλD is an extension of Tukey's lambda distribution (Hastings, Mosteller, Tukey, & C 1947). It is defined by its inverse distribution function:

$$F^{-1}(u) = \lambda_1 + \frac{u^{\lambda_3} - (1-u)^{\lambda_4}}{\lambda_2} \qquad (1)$$

In Expression (1), $0 \le u \le 1$, $\lambda_2 \neq 0$ and $\lambda_1$ ,$\lambda_2$, $\lambda_3$, $\lambda_4$ are respectively the location, scale, skewness and kurtosis parameters of generalized lambda distribution GλD($\lambda_1$ ,$\lambda_2$, $\lambda_3$, $\lambda_4$). In particular, Karian, Dudewicz and MacDonald (1996) noted that GλD is defined if and only if:

$$\frac{\lambda_2}{\lambda_3 u^{\lambda_3 - 1} + \lambda_4 (1-u)^{\lambda_4 - 1}} \ge 0 \qquad (2)$$
$$u \in [0,1]$$

Another distribution known as FMKL GλD also exists, due to the work of Freimer Mudholkar, Kollia and Lin (1988). This distribution is slightly different to RS GλD and they overlap when $\lambda_3=\lambda_4$. The FMKL GλD can be written as:

$$F^{-1}(u) = \lambda_1 + \frac{\frac{u^{\lambda_3} - 1}{\lambda_3} - \frac{(1-u)^{\lambda_4} - 1}{\lambda_4}}{\lambda_2} \qquad (3)$$

Under Expression (3), $0 \le u \le 1$, and $\lambda_1$, $\lambda_2$, $\lambda_3$, $\lambda_4$ are consistent with the interpretations in RS GλD, namely $\lambda_1$ ,$\lambda_2$ are the location and scale parameters and $\lambda_3$, $\lambda_4$ are the shape parameters. In particular, if $\lambda_3=\lambda_4=0$, both RS and FMKL GλD have:

$$F^{-1}(u) = \lambda_1 + \frac{\ln(u) - \ln(1-u)}{\lambda_2} \qquad (4)$$

The fundamental motivation for the development of FMKL GλD is that the distribution is proper over all $\lambda_3$ and $\lambda_4$ (Freimer, Mudholkar, Kollia, & Lin, 1988). This adds convenience to users who wish to program this function as there are fewer restrictions on the values of $\lambda_3$ and $\lambda_4$. The only restriction on FMKL GλD is $\lambda_2>0$.

The extensive use of FMKL GλD is reported in Freimer et al (1988). Due to the wide range of shapes GλD possesses, for example: U shaped, bell shaped, triangular, and exponentially shaped distributions and its simplicity, it has been used in Monte Carlo simulations (Hogben, 1963), the modeling of empirical distributions (Ramberg, Tadikamalla, Dudewicz, & Mykytka, 1979; Okur, 1988), and in the sensitivity analysis of robust statistical methods (Shapiro, Wilk, & Chen, 1968). Other

research works on GλD concentrate on estimating the parameters of the GλD from empirical data and these are discussed below.
In any optimization problem, it is necessary to:

1. Find suitable initial values, and
2. Choose the appropriate optimization scheme.

Perhaps the most common approach has been to use method of moments to estimate the parameters of GλD as demonstrated in Ramberg et al (1979) and Karian and Dudewicz (1996, 2000). These works covered only the RS GλD and often use tables based on the third and fourth moments or percentiles of the data to find suitable initial values. The appropriate optimization scheme involves finding a GλD with parameters $\lambda_1$, $\lambda_2$, $\lambda_3$, $\lambda_4$ that matches closely with the first four moments of the empirical data. This is done numerically through either the Nelder-Simplex (Nelder & Mead, 1965) algorithm as in the work of Ramberg, et al. (1979) or the Newton-Raphson algorithm or tabulated values (Karian & Dudewicz, 2000). Karian and Dudewicz (1996) also discussed the use of the generalized beta distribution to model the distributions that were not covered by the original RS GλD. In Karian and Dudewicz (2000), an alternative method is also demonstrated which matches the RS GλD with the parameters $\lambda_1$, $\lambda_2$, $\lambda_3$, $\lambda_4$ based on the first four percentiles of the data set. This is a variation on the same theme of the matching of moment method but one in which Karian and Dudewicz (2000) reported can produce better fits than in the case with other methods of moment matching under RS GλD.

In a different line of work, Ozturk and Dale (1985) used a version of least squares estimation to find the parameters of RS GλD. They derived the squared distance between empirical data points with the expected values of the order statistics, and numerically minimized this measure using Nelder-Simplex method to derive parameter estimates for the RS GλD.

The literature recognizes that matching the first four moments or using the "least squares" method by Ozturk and Dale (1985) does not necessarily produce a good fit to the data (Karian & Dudewicz, 2000; Lakhany & Massuer, 2000). This is due to different parameters of the GλD can results in the similar first four moments. For example, in the case of the least squares method by Ozturk and Dale (1985), the goal of minimizing the squared distance between empirical data points with the expected values of the order statistics of GλD does not necessarily coincide with the formal goodness of fit objective such as the Kolmogorov-Smirnov Goodness-of-Fit Test.

It is precisely the need to assess the resulting fit with the goodness of fit objective that King and MacGillivray (1999) used the starship methods. In the starship method, grid points comprising of $\lambda_1$, $\lambda_2$, $\lambda_3$, $\lambda_4$ aimed at covering a wide range of GλD, calculated from the sample quantiles. Then, for each of the grid points the theoretical GλD was transformed into uniform distribution and goodness of fit statistics like Anderson-Darling test statistics or Kolmogorov-Smirnov test statistics were calculated. The set of grid points with the lowest Anderson-Darling statistics was then being chosen as the initial values for optimization, usually through the Nelder-Simplex algorithm. The resulting values from the optimization scheme are the parameter estimates of the GλD, given by starship method.

Lakhany and Mausser (2000) suggested a variation of using re-sampling method combined with the method of moments and a goodness of fit test via the FMKL GλD. They first generated initial values for the method of moment matching via quasi random number generator (i.e., the Sobol sequence generator (Bratley & Fox, 1988)), and then found the set of values $\lambda_1$, $\lambda_2$, $\lambda_3$, $\lambda_4$ that matched optimally (through the Nelder-Simplex algorithm) with the first four moments from the data. This set of values was then evaluated through a goodness of test statistic such as adjusted Kolmogorov-Smirnov test statistics. Under this method, any solution that results in a p-value > 0.05 is accepted. Lakhany and Mausser (2000) commented that this method is much more efficient time-wise than the starship method developed by King and MacGillivray (1999) and allows for automatic restarts from different initial values to help to find a distribution that will adequately fit the data. The use of p-values in the optimization scheme, however, can be

somewhat problematic. The deficiency of p-values is well known, since failure to reject does not mean the hypothesis is true since it may be that the sample size is too small to be able to detect differences between the empirical and fitted data. Conversely, rejection of the hypothesis does not mean the fitted model is inappropriate, as the user may have a different purpose to fitting the data other than to satisfy the goodness of fit criteria.

An important improvement of Lakhany and Mausser (2000)'s approach is the flexibility of fits it offers to the users. As different initial values are chosen, different results can be obtained. However, this flexibility is rather limited as the users have no real control over the amount of smoothing they would like to achieve.

The current literature does not appear to cover a comparison of the method of percentiles from Karian and Dudewicz (2000) with the other methods like starship by King and MacGillivray (1999), nor with the automatic re-sampling methods of Lakhany and Massuer (2000). The method below will consider both the method of percentiles under RS GλD and the method of moments under the FMKL GλD. The rationale is that the existing literature appears to recommend these two methods hence these methods are chosen for extension to offer greater flexibility of fit than the methods previously reported.

A detailed discussion of the method of percentiles using the RS GλD and the method of moments using FMKL GλD is outlined below.

Method of percentiles using the RS GλD:

The following is obtained directly from Karian and Dudewicz (2000). For a given data set X with values $x_1$, $x_2$, $x_n$, the p-th percentile defined by Karian and Dudewicz (2000) is $\hat{\pi}_p = y_r + k(y_{r+1} + y_r)$, where Y= $y_1$, $y_2$,… $y_n$ are sorted values of X in ascending order and r is the truncated value of (n+1)×p with k being (n+1)×p-r.

Instead of using the first four moments, the following statistics are used:

$$\begin{aligned}
\hat{\rho}_1 &= \hat{\pi}_{0.5} \\
\hat{\rho}_2 &= \hat{\pi}_{1-v} - \hat{\pi}_v \\
\hat{\rho}_3 &= \frac{\hat{\pi}_{0.5} - \hat{\pi}_v}{\hat{\pi}_{1-v} - \hat{\pi}_{0.5}} \\
\hat{\rho}_4 &= \frac{\hat{\pi}_{0.75} - \hat{\pi}_{0.25}}{\hat{\rho}_2}
\end{aligned} \tag{5}$$

where v is an arbitrary number from 0 to 0.25.

The relationship between the theoretical $\rho_1$, $\rho_2$, $\rho_3$, $\rho_4$ and $\lambda_1$, $\lambda_2$, $\lambda_3$, $\lambda_4$ in the RS GλD is as follows:

$$\begin{aligned}
\rho_1 &= F^{-1}(0.5) = \lambda_1 + \frac{0.5^{\lambda_3} - 0.5^{\lambda_4}}{\lambda_2} \\
\rho_2 &= F^{-1}(1-v) - F^{-1}(v) = \frac{(1-v)^{\lambda_3} - v^{\lambda_4} + (1-v)^{\lambda_4} - v^{\lambda_3}}{\lambda_2} \\
\rho_3 &= \frac{F^{-1}(0.5) - F^{-1}(v)}{F^{-1}(1-v) - F^{-1}(0.5)} = \frac{(1-v)^{\lambda_4} - v^{\lambda_3} + (0.5)^{\lambda_3} - (0.5)^{\lambda_4}}{(1-v)^{\lambda_3} - v^{\lambda_4} + (0.5)^{\lambda_4} - (0.5)^{\lambda_3}} \\
\rho_4 &= \frac{F^{-1}(0.75) - F^{-1}(0.5)}{\rho_2} = \frac{(0.75)^{\lambda_3} - (0.25)^{\lambda_4} + (0.75)^{\lambda_4} - (0.25)^{\lambda_3}}{\rho_2}
\end{aligned} \tag{6}$$

The condition $-\infty < \rho_1 < \infty, \rho_2 \geq 0, \rho_3 \geq 0, \rho_4 \in [0,1]$ must also be true, which is a direct consequence of the definition of $\rho_1$, $\rho_2$, $\rho_3$, $\rho_4$. In Karian and Dudewicz (2000), a fit for the GλD is found by solving Expression (7) through the use of tables. This can also be solved this numerically via Newton-Raphson method.

$$\left|\hat{\rho}_3 - \rho_3\right| <= 10^{-6}, \left|\hat{\rho}_4 - \rho_4\right| <= 10^{-6} \tag{7}$$

In the extended method described below, however, the following minimization scheme in Expression (8) is used. Once $\lambda_3$, $\lambda_4$ are obtained, $\lambda_1$, $\lambda_2$ can be obtained directly via Expression (6).

$$\sqrt{\left(\hat{\rho}_3-\rho_3\right)^2+\left(\hat{\rho}_4-\rho_4\right)^2} \qquad (8)$$

Method of Moments under the FMKL GλD:

In an alternative approach, Lakhany and Mausser (2000) used the method of moments for the FMKL GλD. The following are extracts from Lakhany and Mausser (2000):

For a given data set X with values $x_1$, $x_2$,… $x_n$, the i-th moment $\alpha_i$ is defined in Expression (9).

$$\begin{aligned}
\hat{\alpha}_1 &= \frac{\sum_{i=1}^{n} x_i}{n} \\
\hat{\alpha}_2 &= \frac{\sum_{i=1}^{n} (x_i-\hat{\alpha}_1)^2}{n} \\
\hat{\alpha}_3 &= \frac{\sum_{i=1}^{n} (x_i-\hat{\alpha}_1)^3}{n(\hat{\alpha}_2)^{1.5}} \\
\hat{\alpha}_4 &= \frac{\sum_{i=1}^{n} (x_i-\hat{\alpha}_1)^4}{n(\hat{\alpha}_2)^{2}}
\end{aligned} \qquad (9)$$

Putting a= $\frac{1}{\lambda_2}$ and b= $\lambda_1-\frac{1}{\lambda_1\lambda_2}+\frac{1}{\lambda_2\lambda_4}$, with Y=(X-b)/a, using $E(X^k)=\int_0^1 \left(F^{-1}(u)\right)^k du$ and binomial expansion gives Expression (10).

$$\begin{aligned}
s_k &= E(Y^k) \\
s_k &= \int_0^1 \left(\frac{u^{\lambda_3}}{\lambda_3}-\frac{(1-u)^{\lambda_4}}{\lambda_4}\right)du \\
s_k &= \int_0^1 \sum_{j=o}^{k} \binom{k}{j}(-1)^j \left(\frac{u^{\lambda_3(k-j)}}{\lambda_3^{\,k-j}}-\frac{(1-u)^{\lambda_4 j}}{\lambda_4^{\,j}}\right)du \\
s_k &= \sum_{i=o}^{k} \binom{k}{j}\frac{(-1)^j}{\lambda_3^{\,k-j}\lambda_4^{\,j}}\beta(\lambda_3(k-j)+1,\lambda_4 j+1)
\end{aligned} \qquad (10)$$

In Expression (10), β(*) denotes beta function. Note that both arguments of the beta function must be positive, implying that min($\lambda_3$, $\lambda_4$) > -1/*k* if the distribution is to have finite k-th moments. The k-th central moment (except for the first which is the mean) of the distribution $F^{-1}(u)$ denoted as $\mu_k$ are hence given in Expression (11).

$$\begin{aligned}
\mu_1 &= \frac{1}{\lambda_2}(s_1)-\frac{1}{\lambda_2\lambda_3}+\frac{1}{\lambda_2\lambda_4} \\
\mu_2 &= \frac{1}{\lambda_2}(s_2-s_1^2) \\
\mu_3 &= \frac{1}{\lambda_2^3}(s_3-3s_1s_2+2s_1^3) \\
\mu_4 &= \frac{1}{\lambda_2^4}(s_4-4s_1s_3+6s_1^2s_2-3s_1^4)
\end{aligned} \qquad (11)$$

The theoretical $\alpha_3$ and $\alpha_4$ are given in Expression (12).

$$\begin{aligned}
\alpha_3 &= \frac{s_3-3s_1s_2+2s_1^3}{(s_2-s_1)^{\frac{3}{2}}} \\
\alpha_4 &= \frac{s_4-4s_1s_3+6s_1^2s_2-3s_1^4}{(s_2-s_1)^2}
\end{aligned} \qquad (12)$$

The same methodology now follows as from Lakhany and Mausser (2000). They propose to find $\lambda_3$, $\lambda_4$ by minimizing Expression (13), where $\hat{\alpha}_3$ and $\hat{\alpha}_4$ are sample values using sample moments.

$$\sqrt{\left(\hat{\alpha}_3-\alpha_3\right)^2+\left(\hat{\alpha}_4-\alpha_4\right)^2} \tag{13}$$

Once $\lambda_3$, $\lambda_4$ is determined it is possible to find $\lambda_1$, $\lambda_2$ as shown in Expression (14).

$$\lambda_2=\frac{\sqrt{\left(s_2-s_1^2\right)}}{\hat{\alpha}_2}$$
$$\lambda_1=\hat{\alpha}_1+\frac{1}{\lambda_2}\left(\frac{1}{\lambda_3+1}-\frac{1}{\lambda_4+1}\right) \tag{14}$$

Extension of previous methodology

The principle underlying earlier methods (King & MacGillivray, 1999; Lakhany & Massuer, 2000) is to use goodness of fit as a criteria to determine whether the resulting GλD fits the data adequately. However this, as will be demonstrated later, does not give the potential for a wide range of different plausible distribution fits to data.

The new method described here uses the percentile method from Karian and Dudewicz (2000) and the method of moments with the FMKL GλD. It also uses quasi random numbers to find initial values, but the optimization can be based on the number of classes or bins the user specifies. This optimization scheme allows users to suppress or accentuate part of the distribution as desired, a feature that is not explicitly considered in other methods. The range of initial values should be chosen based on the shape of the distribution shown by the histogram, or they maybe left unspecified with a default set of values chosen.

A full description of the algorithm is provided below:

1. Specify a range of initial values for $\lambda_3$, $\lambda_4$, and the number of initial values to be selected. Here, the $\lambda_3$, $\lambda_4$ are set by default to range from -1.5 to 1.5 for the RS GλD percentile method and -0.25 to 1.5 for the FMKL GλD method of moment. These default values are from author's clinical experiences and appear to work well in most situations. It is possible to change these initial values if desired.

The quasi random generator used is based on the work of Hong and Hickernell (http://www.mcqmc.org/Software.html) and the scrambling method of Owen (1995) and Faure and Tezuka (2000). This code is available from the beta resample library in Splus 6.0 and scrambling methods are applied so that the numbers generated fills uniformly onto the $\lambda_3$, $\lambda_4$ two dimensional space. To increase the speed, it is possible to set the initial values where $\lambda_3$= $\lambda_4$. This appears to work well in many situations. By default, 100 of such initial values are chosen in this case and used in step 2.

2. Evaluate $\lambda_1$, $\lambda_2$ for each of the initial values $\lambda_3$, $\lambda_4$. Remove all the set of values that do not:
    a. Result in a legal parameterization of GλD.
    b. Span the entire region of the data set.

From these sets of initial points, find the values of $\lambda_3$, $\lambda_4$ that matches closely with the data. This is to generate a set of initial values that produce the lowest values in Expression (8) and Expression (13), to be used as initial values in the optimization process.

3. Sort the sample data in ascending order, and divide the data set into evenly spaced classes with bin edges that span

the data set. Calculate the proportion of the sample out of the total sample in each class. Hence Table 1 maybe constructed:

Table 1 Calculating proportion of data in each class

| Classes | 1.5-2 | 2-2.5 | 2.5-3 | 3-3.5 | Sum |
|---|---|---|---|---|---|
| Proportion of data | 0.1 | 0.6 | 0.2 | 0.1 | 1 |

Table 1 shows four classes, with the proportion of the data set falling in each class in the second column. Let the proportion of data in each class be denoted $d_i$ for i=1,2,3..n classes and the proportion of data from the theoretical GλD be the vector $t_i$ for i=1,2,3…n classes. The quantity that one wants to minimize is:

$$\sum_{i=1}^{n} d_i (d_i - t_i)^2 \qquad (15)$$

Expression (15) is the weighted squared deviation of the theoretical proportions with the actual data proportions. This is weighted so that the data with higher proportions are given priority in the minimization scheme. The resulting fit will then be more likely to capture the majority of the data. The weighting factor $d_i$ can be removed if desired. In addition, this optimization scheme also rejects estimations that do not span the entire data set.

The number of classes, n, can be solely determined by the user, or determined by the formula devised by this article (discussed below), or via previous literature works as in Sturges, Scott (1979; 1992) or Freedman and Diaconis (1981).

Sturges' formula is based a bin width of:

$$\text{range(data)}/(\log_2 \text{m} + 1) \qquad (16)$$

This strategy often results the bin width being too wide as reported in Venables and Ripley (2002), and has the disadvantage that "outliers may inflate the range and increase the bin width in the centre of the distribution."

Hyndman (1995) also argued that the use of Sturges' formula should be avoided since there is no sound statistical backing to its derivation.

Scott (1979) used $3.5\hat{\sigma} m^{-1/3}$, although Freedman & Diaconis (1981) proposed $2Rm^{-1/3}$, where R is the inter-quartile range and $\hat{\sigma}$ is the estimated standard deviation from the data, and m is the number of observations in the data. Freedman & Diaconis's (1981) use of inter-quartile range is more robust against outliers and tends to choose smaller bins than the formula by Scott (1979). More complicated rules are also available in Scott (1992) but they are not discussed here.

The methods developed in this article calculate the default number of classes to be optimized over as the one that gives ζ: the minimal squared error between the first two moments of the categorized data with the actual. For example, in the context of Table 1, the first two moments of the categorized data can be calculated using the following table, which takes the mid point of the class intervals and treat the data as discrete. The mean and variance of data shown in Table 2 are 2.4 and 0.1525 respectively; this is then compared with the actual mean and variance of the continuous data with the squared error subsequently calculated. The number of classes chosen for optimization would be the one with minimal squared error or ζ. It is possible to choose any other number of classes such as the formula in Scott (1979) and Freedman & Diaconis (1981).

Table 2 Calculating mean and variance from Table 1

| Observation | 1.75 | 2.25 | 2.75 | 3.25 | Sum |
|---|---|---|---|---|---|
| Proportion of data | 0.1 | 0.6 | 0.2 | 0.1 | 1 |

The philosophy for this approach is to choose the number of classes that best represents the first two moments of the data, so that the distribution fitted would resemble more or less an accurate representation of the data set.

Although formulas for determining the optimal bin width for the histograms interval do exist, users can exercise their judgments by choosing the number of classes. Generally

speaking, higher number of classes will result in details of the distribution being accentuated, while lower number of classes will tend to suppress details of the distribution.

4. The optimal result can be obtained via the Nelder-Mead Simplex algorithm or another suitable numerical optimization algorithm. It is advisable to re-use the initial values in the optimization process to ensure the result obtained is a global minimum rather than a local minimum. Steps 1 to 3 may be repeated if necessary, where the number of classes and the range of initial values can be adjusted until the results are deemed adequate. The final fitting result can be examined by plotting the result on the histogram with the fitted line as well as testing the goodness of fit using the Kolmogorov-Smirnov (KS) test.

## Results

The analysis below is divided into two parts. The first part is a theoretical comparison between data fitting methods with well known statistical distributions. A two sample KS test is carried out by sampling 100 points from the theoretical and fitted distributions and the number of times the p-value exceeds 0.05 is recorded over 1000 times. This will give the user an independent measure as to the adequacy of fits beyond a visual comparison. The second part shows the fitting method over some real life data, and the goodness of fit test is carried out on the comparison between sampling 90% of the real life data with the fitted data using two sample KS test over 1000 runs.

This is also known as the Monte Carlo KS test in this article. It is worth cautioning that the use of goodness of fit as a measure for quality of fit would bias methods that seek to maximize goodness of fit. In fact, it is a circular logic. The use of goodness of fit to assess the quality of fits used in this article will not suffer from this problem, but it needs to bear in mind that the objective of fit in this article was not to maximize the goodness of fit, and so it may not always be as high as starship method (STAR) which uses standard statistical goodness of fit such as Kolmogorov-Smirnov and Anderson Darling test statistics in its data fitting algorithm.

The following compares between the revised percentile method of the RS GλD (RPRS), the revised method of moment under the FMKL GλD (RMFMKL) and the STAR method. Previous literature such as King and MacGillivray (1999), Lakhany and Mausser (2000), and Karian and Dudewicz (2000) have already covered comparisons between the starship methods, the GλD under the RS and FMKL GλD using the method of moments and percentiles as well as the least square method used by Ozturk (1985); hence these will not be repeated here.

### Commentary

The modified methods RPRS and RMFMKL are perhaps not appropriately termed as the percentiles and method of moments are not used in the optimization step but only for choosing the initial values for the optimization process. However, the differences in the two methods highlight the fact that the choices of initial values and type of GλD are important in the outcome of these extended methods, since different results are obtained even though both methods undergo the same optimization scheme.

### Comparison with Theoretical Distributions

Figure 1 and Table 3 show the resulting fits of RPRS, RMFMKL and STAR on well known statistical distributions. Using the default fitting method described above, RPRS and RMFMKL are very close to the actual distribution in Figure 1. This result is further confirmed in Table 3, where more than 90% of the time, the Monte Carlo KS test will indicate there is no difference between the fitted and actual distributions.

The real interest of the method of this article is not in the fitting of theoretical distributions. In the theoretical simulation it is possible to compare between the actual and approximate distributions, but not so in practice. It is precisely the reason that one does not know the real underlying distribution of real life data, one needs a flexible fitting method that could allow us to assess different distribution fits and

the stability of distribution fits under different data representations by the histogram.

The following real life examples will compare different cases where different methods work well under different situations. It will also use the Monte Carlo KS tests results to demonstrate the quality of fit under the goodness of fit objective.

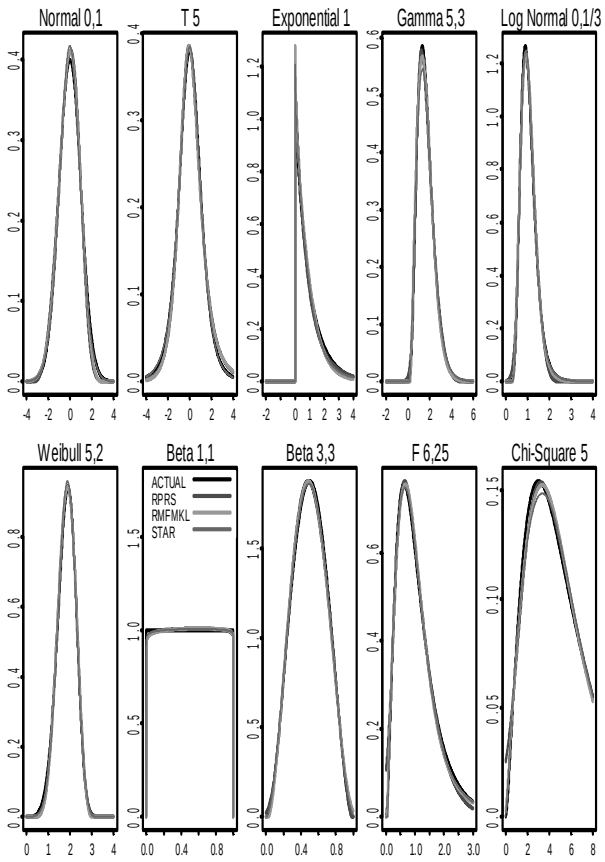



Figure 1: Demonstrating the distribution fits of well known statistical distributions.

Table 3: Monte Carlo KS goodness of fit tests results over 1000 runs. A value close to 1000 indicates high level of confidence of a good fit.

| Distribution | RPRS | RMFMKL | STAR |
|---|---|---|---|
| normal(0,1) | 941 | 966 | 955 |
| student(5) | 943 | 940 | 960 |
| exp(1) | 945 | 905 | 944 |
| gamma(5,3) | 957 | 960 | 961 |
| lognormal(0, | 967 | 977 | 969 |
| weibull(5,2) | 964 | 968 | 952 |
| beta(1,1) | 970 | 963 | 970 |
| beta(3,3) | 966 | 966 | 959 |
| f(6,25) | 939 | 964 | 961 |
| chisq(5) | 962 | 966 | 958 |

Dataset used

The datasets used in here were supplied by research works of Sabri Hassan and Victoria Clout at School of Accountancy in Queensland University of Technology, Australia. The dataset by Sabri Hassan is based on 44 Australian extractive industries firms, listed on the ASX (Australian Stock Exchange) from 1998 to 2001. The dataset used is based on the mean value of each individual company over four years. Market to Book values (sh.mtb), transparency (sh.transp), and profit (sh.profit) variables were extracted and used in this demonstration. There are 176 observations in this data set and the goodness of fit test below will sample 160 observations from this data set and the fitted distribution.

Victoria Clout's data consisted of 361 US firms, listed on the S&P500. The selection requirements were December year-end firms for the 1977 to 1995 period. Similarly, the data used is based on the mean values for each company over the 12 years period. Market to Book ratio (vc.mbr), Ratio of cash and marketable securities over current assets (vc.flex), return on assets (vc.roa) were used in this demonstration. There are 143 observations in this data set and the goodness of fit test below will sample 130 observations from this data set and the fitted distribution.

In addition to financial data, geological data (faithful) on the duration of 272 eruptions

from the Old Faithful geyser in Yellowstone National Park (Hardle, 1991) was also used.

The following examples are designed to demonstrate the flexibility the new methods which can fit alternative, convincing distributions other than suggested by the starship method. It also designed to offer a balanced view on some of the possible deficiencies of this method in relation to satisfying the goodness of fit tests.

Figure 2 is an example of graphical over-fitting by the STAR method, and how the use of default settings described in this article appears to give a more adequate fit. The number of classes to be optimized over is 12, using the default calculations. The histogram shown in Figure 2 is plotted using 100 classes. Using the Monte Carlo KS test, the results are 0, 7 and 732 for RPRS, RMFMKL and STAR respectively. This suggests that STAR is the best fit among the three under the Monte Carlo KS test. It is however possible to improve the Monte Carlo KS test of the RPRS fit by increasing the number of classes to be fitted.

*Example 1: sh.mtb*

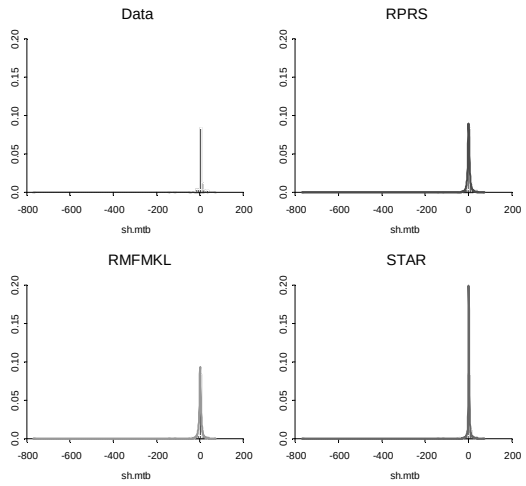



Figure 2: Fitting of sh.mtb data using RPRS, RMFMKL and STAR methods. The extreme scale is due to an extreme outlier, which is retained for illustrative purposes. For example, a certain process may have a huge loss with a very small probability, but it is nevertheless important to model that scenario.

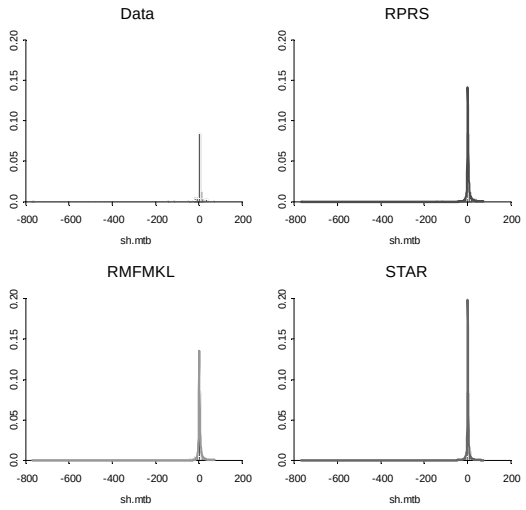
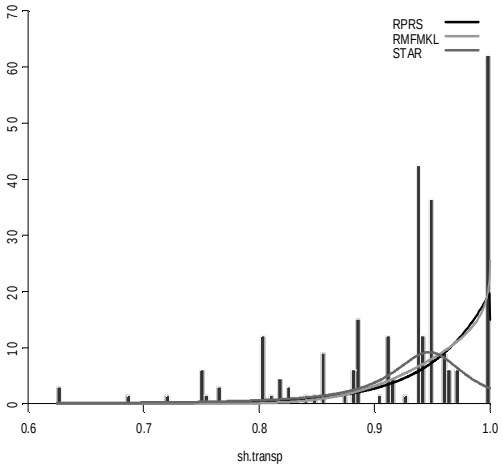



Figure 3: Fitting of sh.mtb data using RPRS, RMFMKL and STAR methods using 150 classes. This shows how it is possible to fit using different histogram bin widths to improve the goodness of fit.

Figure 3 shows the result of such fit graphically and the Monte Carlo KS results are 585, 561 and 749 for RPRS, RMFMKL and STAR. A real strength of the method developed in this article is that it gives a range of plausible fits which the goodness of fit could be assessed objectively. For example, it can be considered that the results in Figure 2 are less likely to be the real representation of the data than Figure 3.

Example 2: sh.transp, alternatives suggested by RPRS, RMFMKL:



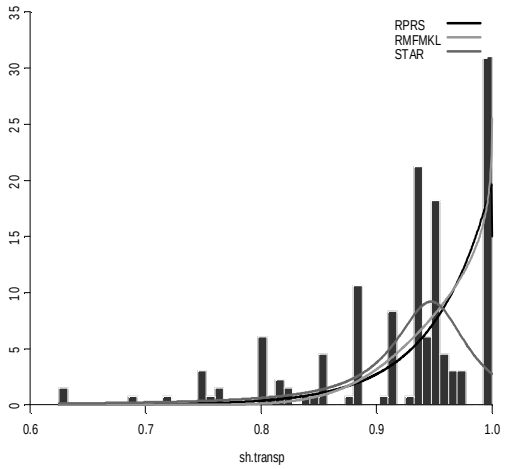



Figure 4: Figures showing fitting of sh.transp data using RPRS, RMFMKL and STAR, the first histogram uses 100 classes while the second histogram uses 50 classes.

The graphs in Figure 4 show two histograms with 100 and 50 classes with the default optimization classes to be optimized over being 31. STAR failed to capture the upward trend of the data. If it is desirable to reach the peak of the histogram data with 100 classes, it is possible to refit RPRS and RMFMKL over 100 classes, resulting in Figure 5. Using 50 or 100 classes will result in Monte Carlo KS test results of 0, 0, and 300 for RPRS, RMFMKL and STAR.

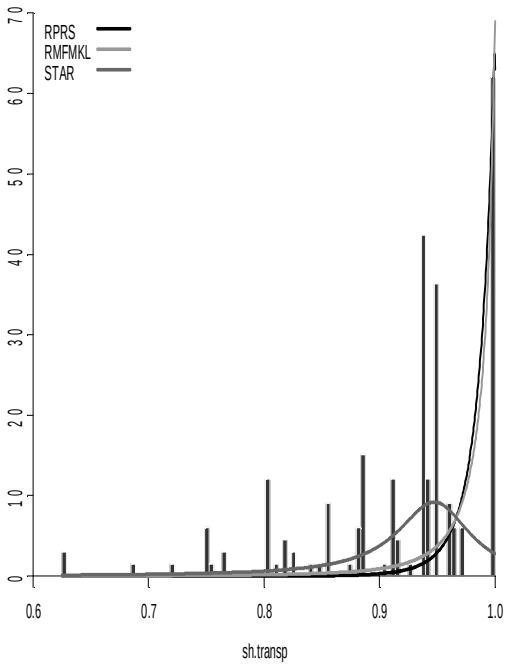



Figure 5: Figure showing alternative fitting of sh.transp sh.transp by RPRS and RMFMKL using 100 histogram classes.

This suggests that none of the methods appear to work well in this case, as STAR although the best out of the three in the Monte Carlo KS test, only really can be said to represent the data 3 times out of 10. In situation like this, where none of the method appears to work well, it is useful to explore other plausible fits and conduct sensitivity analysis to examine the impact on a particular analysis using different distributions.

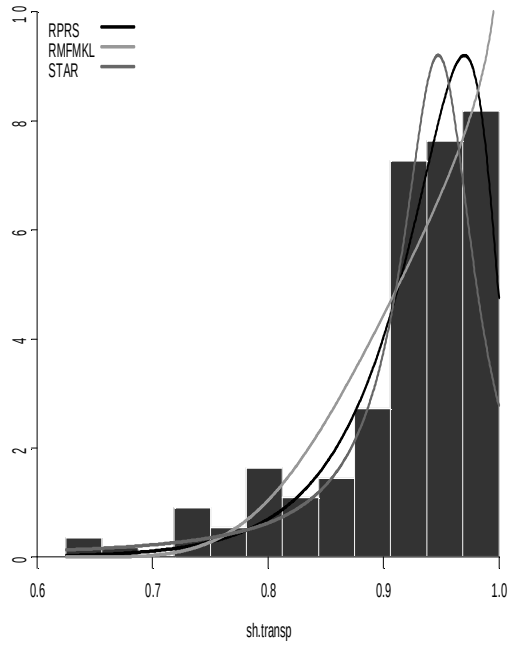



Figure 6: Figure showing alternative fitting of sh.transp using 12 histogram classes.

Figure 6 shows how STAR captured a different representation of the dataset; by manually adjusting the classes of histograms to 12, the fit by STAR appears to be more plausible. Alternative fits by RPRS and RMFMKL using 12 classes appears to represent the data well. This example highlights the importance of allowing alternative methods, since they can give different and possibly valid representations to the same data set. The Monte Carlo KS test results are 23, 2 and 290 for RPRS, RMFMKL and STAR. It also shows the flexibility of RPRS and RMFMKL which can give different fits to the data set depending on the number of classes specified. An additional analysis showing the effect of changing number of classes from 5 to 55 and the corresponding RPRS and RMFMKL fits is shown in Figure 7. All the Monte Carlo KS test results under each of the class suggest 0, 0 and 300 for RPRS, RMFMKL and STAR respectively. The graphs show how different fits may be obtained by varying the number of classes and it is possible these may not change the result of the Monte Carlo KS tests at all. The sharp spike exhibited in Figure 7 for 15 classes is characteristic of RPRS fits, as will be shown in more examples below.

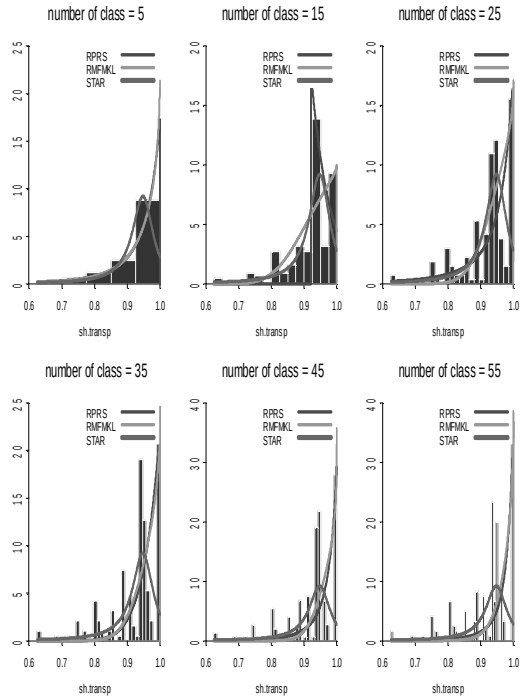



Figure 7: Figure showing alternative fitting of sh.transp using different histogram classes.

Example 3: vc.leverage, similar results:

This example shows that consistent results can often be obtained between different methods. RPRS and RMFMKL used 89 classes by default calculations in this case. The result is shown in Figure 8 below with the histogram exhibiting 100 classes. The Monte Carlo KS tests suggest 882,887 and 945 for RPRS, RMFMKL and STAR respectively. It is normally the case that STAR has somewhat higher goodness of fit score, owing to its fitting objective.

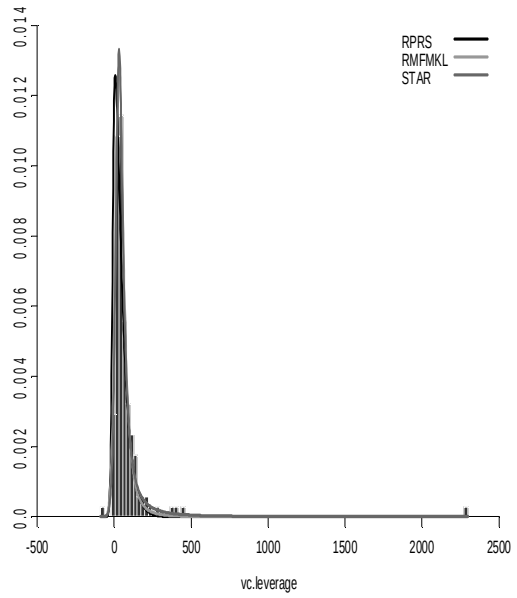



Figure 8: Figure showing fitting of vc.roa data using RPRS, RMFMKL and STAR. All methods give similar results.

Example 4: vc.mbr

RPRS and RMFMKL used 20 classes by default calculations in this optimization scheme. Figure 9 shows a histogram with 100 classes, and all methods give different representations to the dataset. They are all valid representations as suggested by Monte Carlo KS tests, with 929, 887 and 934 for RPRS, RMFMKL and STAR. A striking feature is that RPRS is similar to RMFMKL and they appear to capture the peak of data better than the STAR method. An additional analysis showing the effect of changing number of classes from 5 to 55 and the corresponding RPRS and RMFMKL fits is shown in Figure 10. This example shows how plausible fits can be gauged by using the method described in this article. Table 4 shows the resulting Monte Carlo KS tests for different number of classes and it can be used to as a rough guide as to how credible certain fits are to the data set. For example, for RMFMKL, the most plausible fits are from classes of 15 and 35. This example at Table 4 also shows that the method developed in this article can be as good as STAR method, in addition to offering flexibility to provide convincing fits.

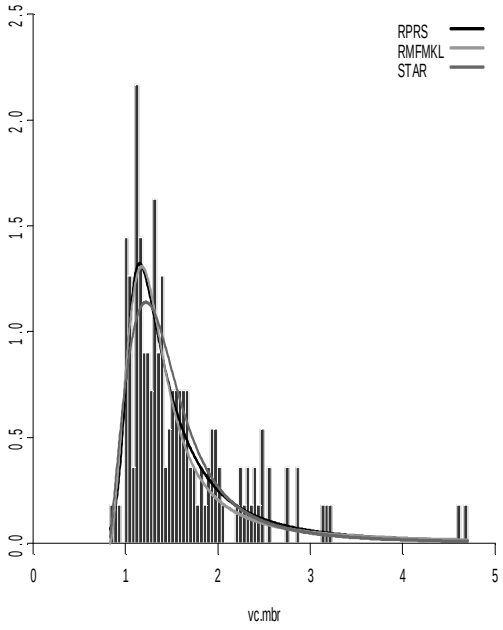



Figure 9: Figure showing fitting of vc.mbr data using RPRS, RMFMKL and STAR. RPRS and RMFMKL appear to represent the peak of the data better than STAR.

Table 4: Monte Carlo KS test for vc.mbr over different number of classes

| Classes<br>Method | 5 | 15 | 25 | 35 | 45 | 55 |
|---|---|---|---|---|---|---|
| RPRS | 481 | 940 | 933 | 905 | 908 | 873 |
| RMFMKL | 354 | 929 | 713 | 932 | 812 | 778 |
| STAR | 932 | 930 | 923 | 917 | 942 | 925 |

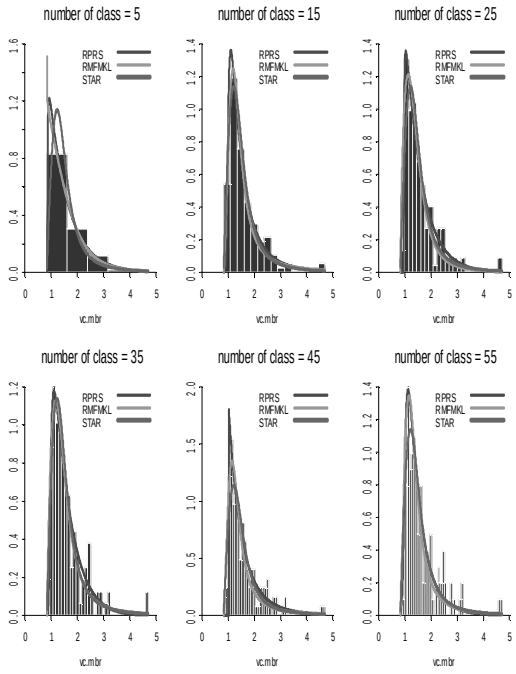



**Figure 10: Figure showing alternative fitting of vc.mbr using different histogram classes.**

*Example 5: faithful, bimodal data, splitting fits by STAR, RPS and RMFMKL:*

This last example shows cases where it may be difficult to fit the data adequately when one encounters a bimodal shaped data. In such cases, the data can be divided into two with two different distributions fitted on each side. Problem can arise when the end points do not match as appeared to be possible with the STAR method in this case. However, as shown in Figure 11, this can be easily corrected for example, by setting the optimization scheme to only include distributions that have maximum values less or equal to 3 for the distribution on the left hand side, and the distribution to have minimum values bigger or equal to 3 on the right hand side.

The original default number of classes was 52 on the RHS of Figure 11 and it does not satisfy the Monte Carlo KS test well, with 614 and 187 for RPRS and RMFMKL. Instead of using the default class calculation, the number of classes was manually adjusted to 20 and this result in Monte Carlo KS test of 855, 873 and 890 for RPRS, RMFMKL and STAR. On the LHS the default setting of 15 classes satisfy the Monte Carlo KS test well, resulting in 921, 927 and 917 for RPRS, RMFMKL and STAR and very similar fits. Figure 11 shows three plausible alternative fits and it is possible some data set may require a mixture of RS and FMKL GλD. The alternative fit by KDE is also provided in Figure 12 for comparison purposes. Figure 12 shows two different fits using KDE. However, the KDE fit, in an attempt to reach the more extreme points of the histogram became less smooth. This rugged appearance will not occur from using generalized lambda distributions.

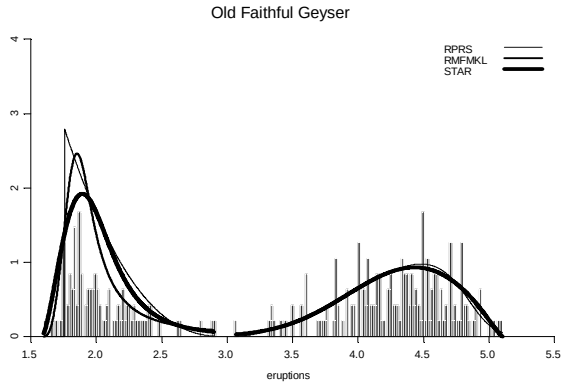



Figure 11: Figure showing fitting of eruptions data using RPRS, RMFMKL and STAR and the use of splitting techniques in fitting bi-modal shaped data. The values below 3 are fitted first and the values above 3 are fitted later.

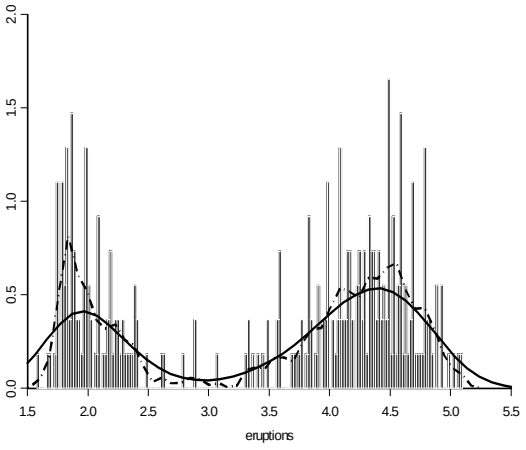



Figure 12: Graph showing two different KDE fits for the eruptions data.

Application of fitting distributions to data using GλD, and a comparison to Kernel Density Estimation method

The use of RPRS or RMFMKL can help users to model a wide variety of distributions as well as acting as a smoothing device with the flexibility of increasing or decreasing levels of details of the data. Another method that allows for density estimation is Kernel Density Estimation (KDE) (Silverman 1985). This is a nonparametric method of estimating the distribution of the data and can often result in a rather rugged appearance compared to the smooth fits from using GλD. Another advantage of using GλD is that the parametric form of the function is known. Consequently, mathematical analysis on the function is possible. In considering re-sampling from the modeled distributions for simulation purposes, both KDE and GλD could be used.

Simulation from KDE and GλD

Simulation from KDE is a simple exercise. KDE calculations give k sets of $(x_1,y_1)\ldots (x_k,y_k)$ co-ordinates which span the distribution of the data. For each consecutive set of points, the area under the line is a trapezium. Let this area be $t_1, t_2,\ldots t_{k-1}$.

Assume one want to sample n numbers from the KDE distribution. For each of the interval i=1,2,3,… k-1, calculate $nt_i$, and generate $nt_i$ numbers from a uniform distribution on the interval, repeating the process for all k-1 intervals.

Simulation from GλD simply requires generating n uniform distribution over [0, 1] and substituting the result into Expression (1) for the RS GλD and Expressions (3) for the FMKL GλD.

Shortcomings of the RPRS AND RMFMKL

All methodologies have their shortcomings, and the method devised here is no exception. The design of the RPRS and RMFMKL can suffer from the following deficiencies.

1. Different results in different runs for the same settings. RPRS and RMFMKL is based on re-sampling methods over the specified range of initial values, hence different runs will result in different initial values being chosen. This is the reason sampling is based on scrambled quasi random sampling (Owen 1995; Hong & Hickernell, 2002) available from the Splus beta resample library, so that the values span evenly throughout the ranges each time. In most cases there are no dramatic changes between each run; however situations do occur when the one run results in a better fit than other runs. This problem can be minimized by increasing the number of values to be sampled in the region. For example, if one million points were chosen over the span of [-1.5, 1.5] then dramatic changes in the result between different runs would be less likely.

2. Optimization method converges falsely or do not converge. This is a problem associated with all numerical optimization schemes, rather than related to this method directly. The program written for RPRS and RMFMKL allows for the quasi-Newton method, conjugate gradients method (Fletcher & Reeves, 1964), the Nelder-Mead algorithm (Nelder & Mead, 1965) and SANN (Belisle, 1992). Hence if one optimization method fails, the other methods can be used instead. So far the use of Nelder-Mead algorithm has proven to be effective in the cases examined here and no case of non convergence have occurred in the application of this optimization procedure.

3. Subjective choice of the number of classes required. Considerable difficulties can arise when choosing number of classes for optimization. While this flexibility is intended, it also may allow data analysts to manipulate the results and choose a method that appears to suit their needs, rather than one that is the most representative of the data. This deficiency does not affect the starship method, which only allows one optimal output based on the goodness of fit measure.

## Conclusion

The exposition in the result section shows the methods developed in this article can offer good alternatives of fitting distribution to data in terms of satisfying Monte Carlo KS tests. While the use of RPRS and RMFMKL offers great flexibility, it also offers rooms for subjective bias in selecting the adequate fit. The use of goodness of fit statistics, however, can help the user to determine the likelihood of a certain distribution fit in the absence of expert knowledge of the underlying data set.

In some situations, where the goodness of fit statistics cannot be adequately satisfied the user could use the methods developed in this article to conduct sensitivity analysis on the impact of results using different distributions. Lastly, improvement on the current RPRS and RMFMKL is also possible by at least two ways, by either improving the optimization algorithm or set an algorithm to quickly find plausible initial values.

## References


Belisle, C. J. P. (1992). Convergence theorems for a class of simulated annealing algorithms on R^d. *Journal of Applied Probability, 29*, 885-895.

Bratley, P., & Fox, B. (1988). Algorithm 659, Implementing Sobol's Quasirandom Sequence Generator. *ACM Transactions on Mathematical Software, 14*(1), 88-100.

Faure, H., & Tezuka, S. (2000). Another random scrambling of digital (t,s)-sequences. *MCQMC 200*0, Hong Kong: Springer-Verlag.

Fletcher, R., & Reeves, C. M. (1964). Function minimization by conjugate gradients. *Computer Journal, 7*, 148-154.

Freedman, D., & Diaconis, P. (1981). On the histogram as a density estimator: L_2 theory. *Zeitschrift für Ahrscheinlichkeitstheorie und Verwandte Gebiete, 57*, 453–476.

Freimer, M., Mudholkar, G., Kollia, G., & Lin, C. (1988). A Study of the generalised Tukey lambda family. *Communications in Statistics- Theory and Methods, 17*, 3547-3567.

Hardle, W. (1991). *Smoothing Techniques with Implementation in S*. New York: Springer.

Hastings, J. C., Mosteller, F., Tukey, J. W., & C, W. (1947). Low moments for small samples: A comparative study of order statstics. *The Annals of Statistics, 18*, 413-426.

Hogben, D. (1963). *Some Properties of Tukey's Test for Non-Additivity*. NJ: Rutgers, The State University of New Jersey.

Hong, H. S. & Hickernell, F. J. (2002). *Implementing scrambled digital sequences*. Unpublished.

Karian, Z., & Dudewicz, E. (2000). *Fitting statistical distributions: The generalized lambda distribution and generalized bootstrap methods*. New York: Chapman & Hall.

Karian, Z., Dudewicz, E., & McDonald, P. (1996). The extended generalized lambda distribution systems for fitting distributions to data: History, completion of theory, tables, applications, the "final word" on moment fits. *Communications in Statistic: Computation and Simulation, 25*(3), 611-642.

King, R., & MacGillivray, H. (1999). A starship estimation method for the generalised lambda distributions. *Australia and New Zealand Journal of Statistics, 41*(3), 353-374.

Lakhany, A., Massuer, H. (2000). Estimating the parameters of the generalised lambda distribution. *Algo Research Quarterly, 3*(3), 47-58.

Nelder, J. A., & Mead, R. (1965). A simplex algorithm for function minimization. *Computer Journal, 7*, 308-313.

Okur, M. (1988). On fitting the generalised lambda distribution to air pollution data. *Atmospheric Environment, 22*, 2569-2572.

Owen, A. (1995). Randomly permuted (t,m,s)-nets and (t,s)-sequences. In (H. Niederreiter& P. J. Shiue, Eds.) *Monte Carlo and Quasi-Monte Carlo Methods in Scientific Computing*, NY: Springer-Verlag. *106***,** 299-317.

Ozturk, A., & Dale, R. (1985). Least squares estimation of the parameters of the generalised lambda distribution. *Technometrics, 27*, 8-84.

Ramberg, J., & Schmeriser, B. (1974). An approximate method for generating asymmetric random variables. *Communications of the Association for Computing Machinery, 17*, 78-82.

Ramberg, J., Tadikamalla, P., Dudewicz, E., Mykytka, E. (1979). A probability distribution and its uses in fitting the data. *Technometrics, 21*, 201-214.

Scott, D. W. (1979). On optimal and data-based histograms. *Biometrika, 66*, 605–610.

Scott, D. W. (1992). *Multivariate density estimation. theory, practice, and visualization.* Indianapolis, IN: Wiley.

Shapiro, S., Wilk, M., & Chen, J. H. (1968). A Comparative Study of Various Tests of Normality. *Journal of American Statistical Association, 63*, 1343-1372.

Silverman, B. W. (1985). *Density estimation for statistics and data analysis*. London, Chapman & Hall.

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S-PLUS*. NY:Springer.

Journal of Modern Applied Statistical Methods
November, 2005, Vol. 4, No. 2, 425-433


1538 – 9472/05/$95.00

# Testing Goodness Of Fit Of The Geometric Distribution: An Application To Human Fecundability Data

Sudhir R. Paul
Department of Mathematics and Statistics
University of Windsor

A measure of reproduction in human fecundability studies is the number of menstrual cycles required to achieve pregnancy which is assumed to follow a geometric distribution with parameter p. Tests of heterogeneity in the fecundability data through goodness of fit tests of the geometric distribution are developed, along with a likelihood ratio test statistic and a score test statistic. Simulations show both are liberal, and empirical level of the likelihood ratio statistic is larger than that of the score test statistic. A power comparison shows that the likelihood ratio test has a power advantage. A bootstrap p-value procedure using the likelihood ratio statistic is proposed.

Key words: Beta-geometric distribution; bootstrap p-value; fecundability data; geometric distribution; likelihood ratio test; score test.

## Introduction

The geometric distribution is important in many real life data analyzes. For example, in fecundability studies (Weinberg & Gladen, 1986), the number of cycles required to achieve pregnancy would be distributed as a geometric distribution with parameter p. However, in real life data situations, the actual variation of the data may exceed that of the geometric distribution, as the parameter p may not remain constant in the course of the experiment. It is then useful to assume that the parameter p varies from observation to observation. One can assume one of many continuous distributions for p in the parameter space 0<p<1. But, the most convenient and most sensible distribution for p is the beta distribution, because it is the natural conjugate prior distribution in the Bayesian sense.

Paul R. Sudhir is University Professor, Professor of Statistics, and Chair, Graduate Studies. His research interests are in correlated data, frailty models, generalized linear models, categorical data analysis, zero-inflated and over-dispersed count data regression models, & dose-response modeling. Email him at smjp@uwindsor.ca.

It also produces a convenient mixed distribution, namely, the beta-geometric distribution. The parameters of this mixed distribution have practical interpretation. In some other analogous applications, such as in Toxicology, the beta-binomial distribution arises as a beta mixture of the binomial distribution (Weil, 1970; Williams, 1975; Crowder, 1978; Otake & Prentice, 1984).

It is assumed that $Y \mid p \sim$ geometric distribution. Let $q = 1 - p$. Then, the probability function of Y is $P(Y = y \mid q) = q^{y-1} p.$

In human reproduction the random variable Y may be the number of menstrual cycles required for conception in which the parameter p may be interpreted as the pre-cycle conception probability or a measure of fecundability (Weinberg & Gladen, 1986). It is assumed that the parameter p is fixed for a given couple, but across couples it varies according to some unspecified underlying distribution which is assumed to be beta with probability density function given by

$$f(p \mid \alpha, \beta) = \frac{p^{\alpha-1}(1-p)^{\beta-1}}{B(\alpha, \beta)}, 0 < p < 1,$$

where

$$\mathrm{B}(\alpha,\beta)=\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}$$

is the beta function and where $\Gamma(\mathrm{a})$ is the gamma function:

$$\Gamma(\mathrm{a})=\int_0^{\infty}\mathrm{x}^{\mathrm{a}-1}\mathrm{e}^{-\mathrm{x}}\mathrm{dx}.$$

The mean and variance of the beta random variable p are $\mu=\frac{\alpha}{\alpha+\beta}$ and $\sigma^2=\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$ respectively. The marginal distribution of Y, then, is

$$\begin{aligned}P(Y=y)&=\int_0^1 P(Y=y\mid p)\ f(p\mid\alpha,\beta)dx\\&=\frac{1}{B(\alpha,\beta)}\int_0^1 p^{\alpha}(1-p)^{y+\beta-2}dx\\&=\frac{B(\alpha+1,y+\beta-1)}{B(\alpha,\beta)}.\end{aligned}$$

This distribution is known as the beta-geometric distribution. In the human reproduction literature P(Y=y) is the probability that conception occurs at y for a randomly selected couple. The beta-geometric distribution can be written in terms of the parameter $\pi=\alpha/(\alpha+\beta)$ and $\theta=1/(\alpha+\beta)$, where p is interpreted as the mean parameter and $\theta$ as the shape parameter (Weinberg & Gladen, 1986), which is given in what follows.

$$\mathrm{P}(\mathrm{Y}=\mathrm{y}\mid\mathrm{n})=\frac{\pi\prod_{\mathrm{r}=0}^{\mathrm{y}-2}\{(1-\pi)+\mathrm{r}\theta\}}{\prod_{\mathrm{r}=0}^{\mathrm{y}-1}\{1+\mathrm{r}\theta\}}.$$

The distribution has mean $\frac{1-\theta}{\pi-\theta}$ and variance $\frac{\pi(1-\pi)(1-\theta)}{(\pi-\theta)^2(\pi-2\theta)}$. Obviously, $\theta=0$ corresponds to the geometric distribution with mean $\frac{1}{\mathrm{p}}$ and variance $\frac{1-\mathrm{p}}{\mathrm{p}^2}$.

The purpose of this article is to develop tests of goodness of fit of the geometric distribution against the beta-geometric distribution. A score test and a likelihood ratio test are developed. The score test (Rao, 1947) is a special case of the more general C($\alpha$) test (Neyman, 1959) in which the nuisance parameters are replaced by their maximum likelihood estimates which are $\sqrt{\mathrm{N}}$ consistent estimates (N=number of observations used in estimating the parameters) . The score or the C($\alpha$) class of tests (i) often maintain, at least approximately, a preassigned level of significance (Bartoo & Puri, 1967), (ii) require estimates of the parameters only under the null hypothesis, and (iii) often produce statistics which are simple to calculate.

These tests are robust in the sense that their optimality remain true whatever the form of the distribution assumed for the data under the alternative hypothesis - a property called robustness of optimality by Neyman and Scott (1966). The C($\alpha$) test has been shown by many authors to be asymptotically equivalent to the likelihood ratio test and to the Wald test (Moran, 1970; Cox & Hinkley, 1974). Potential drawbacks to the use of the likelihood ratio and Wald tests include the fact that both require estimates of the parameters under the alternative hypotheses and often show liberal or conservative behaviour. Examples of this may be found in Barnwal & Paul (1988), Paul (1989), Paul (1996), Paul & Banerjee (1998), and Paul and Islam (1995).

In the present context, although the score test statistic has a very simple form, both the score test and the likelihood ratio test have been found, by simulation, to be liberal. A power comparison, using the empirical quantiles derived from the corresponding size simulation to ensure that each test had approximately the nominal size, has been conducted. This comparison shows that the likelihood ratio test has power advantage over the score test. A

bootstrap likelihood ratio test is therefore proposed to test the fit of a geometric model against the over-dispersed geometric model. The bootstrap likelihood ratio test provides approximately correct p-value (Davison & Hinkley, 1998). McLachlan (1987) uses the bootstrap likelihood ratio test to test for the number of components in mixture of normal distributions. McLachlan notes that the bootstrap and the true null distribution of the likelihood ratio statistics are the same. The bootstrap likelihood ratio test was also used by others in similar contexts (Aitkin, Anderson & Hinde, 1981; Karlis & Xekalaki, 1999).

For the situation in which the data are found to be heterogeneous, maximum likelihood estimates of the parameters of the beta-geometric distribution and the elements of the exact Fisher information matrix are obtained. Two sets of data including one on human fecundability study from Weinberg & Gladen (1986) are analyzed.

## Tests of Goodness of Fit

### Estimation of the Parameters

Suppose data are available on n individuals as $y_i, i = 1, \cdots, n$. The maximum likelihood estimate of the parameter p of the geometric distribution is $\hat{p} = 1/\bar{y}$, where $\bar{y} = \sum_{i=1}^{n} y_i / n$. The likelihood function for

the data based on the beta-geometric distribution is given as

$$L = \pi^n \prod_{i=1}^{n} \frac{\prod_{r=1}^{y_i - 1} \{1 - \pi + (r-1)\theta\}}{\prod_{r=1}^{y_i} \{1 + (r-1)\theta\}}$$

and the corresponding log-likelihood can be written as

$$l = n \log(\pi) + \sum_{i=1}^{n} \left[ \sum_{r=1}^{y_i - 1} \log\{1 - \pi + (r-1)\theta\} - \sum_{r=1}^{y_i} \log\{1 + (r-1)\theta\} \right].$$

The maximum likelihood estimates $\hat{\pi}$ and $\hat{\theta}$ of the parameters $\pi$ and $\theta$ are obtained by solving the maximum likelihood estimating equations $\frac{\partial l}{\partial \pi} = 0$ and $\frac{\partial l}{\partial \theta} = 0$ simultaneously. That is, by solving

$$\frac{n}{\pi} - \sum_{i=1}^{n} \left\{ \sum_{r=1}^{y_i - 1} \frac{1}{1 - \pi + (r-1)\theta} \right\} = 0$$

and

$$\sum_{i=1}^{n} \left\{ \sum_{r=1}^{y_i - 1} \frac{r-1}{1 - \pi + (r-1)\theta} - \sum_{r=1}^{y_i} \frac{r-1}{1 + (r-1)\theta} \right\} = 0,$$

simultaneously subject to the constraints $0 < p < 1$ and $\theta > 0$. Note that there is no closed form solution for these equations. So these equations are to be solved using a numerical procedure such as the Newton-Raphson method or a numerical subroutine, such as the IMSL subroutine ZBRENT or NEQNF.

### The Likelihood Ratio Test

The maximized log-likelihood under the geometric distribution is

$$l_0 = n \log(\hat{p}) + n(\bar{y} - 1) \log(1 - \hat{p}) \qquad (1)$$

and that under the beta-geometric distribution is

$$l_1 = n \log(\hat{\pi}) + \sum_{i=1}^{n} \left[ \sum_{r=1}^{y_i - 1} \log\{1 - \hat{\pi} + (r-1)\hat{\theta}\} - \sum_{r=1}^{y_i} \log\{1 + (r-1)\hat{\theta}\} \right].$$

Then, the likelihood ratio statistic to test for $H_0 : \theta = 0$ against $H_A : \theta > 0$ is $LR = 2(l_1 - l_0)$. Under standard conditions, the asymptotic null distribution of this likelihood ratio statistic would be chi-square with 1 degree of freedom. However, since the parameter $\theta$ is necessarily nonnegative, there is

a boundary problem and the regular asymptotic likelihood theory breaks down in this situation. In the course of a general discussion of asymptotic properties of likelihood procedures when some of the parameters are on the boundary, Self & Liang (1987) derive a representation for the asymptotic distribution of the likelihood ratio statistic. Since the parameter value under $H_0$ is on the boundary of the parameter space it can be easily seen from the results of Self & Liang (1987) that the correct distribution of the LR test is a 50:50 mixture of zero and chi-square with 1 degree of freedom provided $0 < p < 1$.

The Score Test

Define

$$S = \frac{\partial l}{\partial \theta}|_{\theta=0},$$

$$I_{\pi\pi} = E\left(\frac{\partial^2 l}{\partial \pi^2}|_{\theta=0}\right), I_{\pi\pi} = E\left(\frac{\partial^2 l}{\partial \pi \partial \phi}|_{\theta=0}\right),$$

and

$$I_{\theta\theta} = E\left(\frac{\partial^2 l}{\partial \theta^2}|_{\theta=0}\right).$$

Then, a score test statistic for testing $H_0 : \theta = 0$ against $H_A : \theta > 0$ is given by $Z = S/\sqrt{(I_{\theta\theta} - I_{\pi\phi}^2 / I_{\pi\pi})}$. If the nuisance parameter $\pi$ is replaced by its maximum likelihood estimate under the null hypothesis, then, asymptotically, as $n \to \infty$, the distribution of Z is standard normal. Note, under the null hypothesis $\pi$ becomes $p$. Then, the following is obtained

$$S = \frac{1}{1-p}\sum_{i=1}^{n}\sum_{r=1}^{y_i - 1}(r-1) - \sum_{i=1}^{n}\sum_{r=1}^{y_i}(r-1) = 0, \qquad (2)$$

$I_{\pi\pi} = n/\{p^2(1-p)\}, I_{\pi\phi} = -n/p^2$, and

$$I_{\phi\phi} = \frac{n\{2 - 5p + p^2(4-p) - (p-1)(p-2)(1-p)^2\}}{\{p^3(1-p)^2\}}$$

It can be shown that $Var(S) = I_{\phi\phi} - I_{\pi\phi}^2 / I_{\pi\pi} = n/p^2$. Thus, the score test statistic for testing $H_0 : \theta = 0$ against $H_A : \theta > 0$ is given by $Z = S/\sqrt{(n/p^2)}$. If $p$ is replaced by $\hat{p}$, where $\hat{p}$ is the maximum likelihood estimate of the parameter p of the geometric distribution, in Z, then, under the null hypothesis $H_0 : \theta = 0$, the statistic Z will have an asymptotic standard normal distribution. Since this is a one-sided test the null hypothesis is rejected at $100(1-\alpha)\%$ level of significance if $Z > z_\alpha$, where, $z_t$ is the $100(1-t)\%$ point of the standard normal distribution.

Simulations

A simulation experiment was conducted to study size properties of the likelihood ratio statistic LR and the score test statistic Z. Data have been generated from the geometric distribution with values of the geometric parameter $p = .1, .2, .3, .4, .5$, sample sizes, $n = 10,\ 20,\ 50$, and $\alpha = .05, .10$. Each simulation experiment was based on 5000 replications. Empirical size values are given in Table 1.

Table 1: Empirical sizes, in percent, for $H_0$ of score test statistics Z and the likelihood ratio statistic LR

| | | | *p* | | | | |
|---|---|---|---|---|---|---|---|
| *n* | $\alpha$ | Statistics | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 |
| 10 | 0.05 | Z | 8.0 | 6.9 | 7.2 | 6.5 | 6.6 |
| | | LR | 12.0 | 10.6 | 10.6 | 10.5 | 10.0 |
| | | LR1 | 12.0 | 10.6 | 10.6 | 10.5 | 10.0 |
| 20 | | Z | 11.2 | 10.2 | 10.2 | 11.5 | 13.3 |
| | | LR | 13.0 | 11.4 | 11.4 | 12.7 | 15.0 |
| | | LR1 | 12.0 | 10.6 | 10.6 | 10.5 | 10.0 |
| 50 | | Z | 13.3 | 12.6 | 12.3 | 13.8 | 16.4 |
| | | LR | 14.2 | 13.3 | 13.2 | 14.6 | 16.6 |
| | | LR1 | 12.0 | 10.6 | 10.6 | 10.5 | 10.0 |
| 100 | | Z | 13.3 | 12.6 | 12.3 | 13.8 | 16.4 |
| | | LR | 14.2 | 13.3 | 13.2 | 14.6 | 16.6 |
| | | LR1 | 12.0 | 10.6 | 10.6 | 10.5 | 10.0 |
| 500 | | Z | 13.3 | 12.6 | 12.3 | 13.8 | 16.4 |
| | | LR | 14.2 | 13.3 | 13.2 | 14.6 | 16.6 |
| | | LR1 | 12.0 | 10.6 | 10.6 | 10.5 | 10.0 |
| 10 | 0.10 | Z | 14.0 | 12.6 | 12.4 | 12.8 | 12.8 |
| | | LR | 19.0 | 17.1 | 16.6 | 18.0 | 18.3 |
| | | LR1 | 12.0 | 10.6 | 10.6 | 10.5 | 10.0 |
| 20 | | Z | 17.9 | 16.7 | 16.6 | 17.9 | 21.8 |
| | | LR | 20.0 | 18.2 | 18.2 | 19.7 | 23.0 |
| | | LR1 | 12.0 | 10.6 | 10.6 | 10.5 | 10.0 |
| 50 | | Z | 21.6 | 20.2 | 19.9 | 21.9 | 25.5 |
| | | LR | 21.2 | 20.6 | 20.0 | 22.5 | 25.6 |
| | | LR1 | 12.0 | 10.6 | 10.6 | 10.5 | 10.0 |
| 100 | | Z | 13.3 | 12.6 | 12.3 | 13.8 | 16.4 |
| | | LR | 14.2 | 13.3 | 13.2 | 14.6 | 16.6 |
| | | LR1 | 12.0 | 10.6 | 10.6 | 10.5 | 10.0 |
| 500 | | Z | 13.3 | 12.6 | 12.3 | 13.8 | 16.4 |
| | | LR | 14.2 | 13.3 | 13.2 | 14.6 | 16.6 |
| | | LR1 | 12.0 | 10.6 | 10.6 | 10.5 | 10.0 |

Table 2: Empirical powers, in percent, for $H_0$ , at $\alpha = 0.05$, of score test statistics Z and the likelihood ratio statistic LR. The extra-geometric variation is .01(.05)(.1)

| | | *p* | | |
|---|---|---|---|---|
| *n* | Statistics | 0.1 | 0.3 | 0.5 |
| 10 | Z | 6(32)(67) | 8(20)(39) | 5(9)(15) |
| | LR | 7(38)(82) | 9(25)(52) | 5(10)(18) |
| 20 | Z | 11(53)(88) | 22(49)(70) | 10(19)(39) |
| | LR | 10(57)(96) | 25(64)(86) | 12(24)(46) |
| 50 | Z | 15(81)(99) | 53(93)(97) | 8(38)(70) |
| | LR | 16(84)(99) | 54(97)(99) | 13(44)(81) |

From Table 1 it is seen that both the score test statistic and the likelihood ratio statistic are liberal. Empirical level of the likelihood ratio statistic is larger than that of the score test statistic. Also, empirical level increases as the sample size increases. A mean-variance correction of the score test statistic using Taylor series expansion (Paul, 1996) produces empirical levels that are too small compared with the nominal levels.

A power comparison of the two statistics was also conducted. The empirical 95% quantiles derived from the corresponding size simulation have been used to ensure that each test had approximately the nominal size of 0.05. Empirical quantiles were calculated based on 20,000 replications and empirical power calculations were based on 1000 replications. Empirical power values are given in Table 2. The likelihood ratio statistic, in general, shows power advantage, over the score test.

The Bootstrap Goodness of Fit Test

As seen from the simulation results in Section 3, both the likelihood ratio test and the test based on the score test statistic are liberal. However, the likelihood ratio test has some power advantage over the score test. So, following Davison & Hinkley (1997), a bootstrap test of the null hypothesis $H_0 : \theta = 0$ against $H_A : \theta > 0$ is proposed. The bootstrap likelihood ratio test procedure proceeds according to the following steps:

Step 1. Obtain $\hat{p}$ of the parameter $p$ of the geometric distribution from the data. Calculate the value of the likelihood ratio statistic LR, say $LR_0$, from the data.

Step 2. Generate n observations from the fitted null distribution, i.e., the geometric distribution with parameter $p = \hat{p}$ and calculate the likelihood ratio statistic $LR_0^*$.

Step 3. Repeat step 2 B times obtaining B values of the bootstrap likelihood ratio statistic, say, $LR_0^{(b)}$, b=1,2,...,B.

Step 4. Estimate the bootstrap p-value by

$$\hat{p}_{boot} = \frac{1+\#\{LR^{(b)*}_0 \geq LR_0\}}{B+1}.$$

This gives the level at which to reject or not to reject $H_0$. A typical value of B is 1000.

Elements of the Expected Fisher Information Matrix of the Beta-geometric Distribution

In this section, the elements of the expected Fisher Information matrix for the estimates of the parameters of the beta-geometric distribution are derived. The calculations are quite involved, so the details were omitted. The exact expressions are given in what follows.

$$I_{11}=E\left(\frac{-\partial^2 l}{\partial\pi^2}\right)=n/\pi^2+n\sum_{r=2}^{\infty}\frac{P(Y\geq r)}{\{(1-\pi+(r-2)\theta\}^2},$$

$$I_{12}=E\left(\frac{-\partial^2 l}{\partial\pi\partial\phi}\right)=-n\sum_{r=3}^{\infty}\frac{(r-2)P(Y\geq r)}{\{(1-\pi+(r-2)\theta\}^2},$$

and

$$I_{22}=E\left(\frac{-\partial^2 l}{\partial\phi^2}\right)=n\left(\sum_{r=3}^{\infty}\frac{(r-2)^2P(Y\geq r)}{\{(1-\pi+(r-2)\theta\}^2}-\sum_{r=2}^{\infty}\frac{(r-1)^2P(Y\geq r)}{\{1+(r-2)\theta\}^2}\right).$$

Calculations of the above terms do not pose any difficulty if $\infty$ in the upper limit of the summation is replaced by a sufficiently large number, say, 5000. Thus, the estimated variance of $\hat{\pi}$ and $\hat{\theta}$ are

$$\text{var}(\hat{\pi})=\frac{\hat{I}_{22}}{(\hat{I}_{11}\hat{I}_{22}-\hat{I}_{12}{}^2)}$$

and

$$\text{var}(\hat{\theta})=\frac{\hat{I}_{11}}{(\hat{I}_{11}\hat{I}_{22}-\hat{I}_{12}{}^2)}$$

respectively, where $\hat{I}_{11},\hat{I}_{12},\text{and}\hat{I}_{22}$ are estimates of $I_{11},I_{12},\text{and}I_{22}$ respectively obtained by replacing the parameter p by its maximum likelihood estimate.

Examples

Example 1: The data, given in the Table 3 from Weinberg & Gladden (1986), refer to times, taken by couples that were attempting to conceive, until pregnancy results.

Table 3: Data from Weinberg and Gladen (1986) on the number of menstrual cycles to pregnancy

| Cycles | Number of Women |
|---|---|
| 1 | 227 |
| 2 | 123 |
| 3 | 72 |
| 4 | 42 |
| 5 | 21 |
| 6 | 31 |
| 7 | 11 |
| 8 | 14 |
| 9 | 6 |
| 10 | 4 |
| 11 | 7 |
| 12 | 28 |

The data were obtained retrospectively, starting from a pregnancy in each case. Weinberg & Gladen (1986) analyzed fecundability data for a total of 586 women, contributing a total of 1844 cycles. See Weinberg & Gladen (1986) for more details regarding the data. For these data, the data for 12 or more cycles has been combined.

An estimate of the parameter p of the geometric distribution for these data is $\hat{p}=.3177874$. An estimate of the variance is $(1-\hat{p})/\hat{p}^2=$ 6.76. The observed variance, however, is 8.68 which is much larger than the variance predicted by the geometric distribution. This indicates that an over-dispersed geometric distribution may fit the data better than the geometric distribution. Now, the value of the likelihood ratio statistic is LR=14.97 with a p-value (using the 50:50 mixture of 0 and chi-square with 1 degree of freedom)=0.00000006 and the bootstrap p-value is 0.002. In calculating the bootstrap p-value B=500 have been used. The data shows very strong evidence in favor of

the beta-geometric distribution. Note that in this example the p-value of the likelihood ratio statistic is much smaller than the corresponding bootstrap p-value. This is in line with the simulation results in Section 3 that the likelihood ratio test is liberal.

The maximum likelihood estimates of the parameters $\pi$ and $\theta$ of the beta-geometric distribution are $\hat{\pi} = 0.36596$ and $\hat{\theta}$ =0.0745 and the standard errors of the estimates $\hat{\pi}$ and $\hat{\theta}$ are .0162 and .0204 respectively.

Example 2: In example 1 the data produce a value of 14.97 for the likelihood ratio statistic. This is rather large and therefore it is not surprising that both the ordinary likelihood ratio test and the bootstrap likelihood ratio test provide same conclusion. Moreover, the observed variance is about 28% larger than what is predicted by the geometric distribution. Thus, the data given in Table 4 was produced; it was obtained by modifying the data set in Table 3.

Table 4: Modified data of Table 3 on the number of menstrual cycles to pregnancy

| Cycles | Number of Women |
|---|---|
| 1 | 180 |
| 2 | 123 |
| 3 | 72 |
| 4 | 42 |
| 5 | 21 |
| 6 | 31 |
| 7 | 11 |
| 8 | 14 |
| 9 | 6 |
| 10 | 4 |
| 11 | 7 |
| 12 | 18 |

For these data an estimate of the variance predicted by the geometric distribution is $(1-\hat{p})/\hat{p}^2 =$ 6.88 and the corresponding observed variance is 7.72. These two variances are much closer than the two corresponding variances for the data in Table 3. This indicates that the geometric distribution might fit these data well. For these data the value of the likelihood ratio statistic is LR=2.51 with a p-value (using the 50:50 mixture of 0 and chi-square with 1 degree of freedom) = 0.025 and the bootstrap p-value is 0.14. For these data, the bootstrap likelihood ratio procedure shows that the geometric distribution fits the data well at 5% level of significance which is contradicted by the ordinary likelihood ratio test. The reason for this is that the likelihood ratio test is liberal.

## References


Aitkin, M., Anderson, D., & Hinde, J. (1999). Statistical modelling of data on teaching styles. *Journal of the Royal Statistical Society, A, 144*, 419-461.

Barnwal, R. K. & Paul, S. R. (1988). Analysis of one-way layout of count data with negative binomial variation. *Biometrika 75*, 215-222.

Bartoo, J. B. & Puri, P. S. (1967). On optimal asymptotic tests of composite statistical Hypothesis. *The Annals of Mathematical Statistics, 38*, 1845-52.

Crowder, M. J. (1978). Beta-binomial ANOVA for proportions. *Applied Statistics, 27*, 34-37.

Cox, D. R. & Hinkley, D. V. (1974). *Theoretical Statistics*. London: Chapman and Hall.

Davison, A. C. & Hinkley, D. V. (1997). *Bootstrap Methods and Their Application*. Cambridge University Press.

International Mathematical and Statistical Libraries (1994). *IMSL Manual*. The Numerical Solution Source, Houston, Texas.

Karlis, D. & Xekalaki, E (1999). On testing for the number of components in a mixed Poisson model. *Annals of the Institute of Statistical Mathematics, 51*, 149-162.

Otake, M. & Prentice, R. L. (1984). The analysis of chromosomally aberrant cells based on beta-binomial distribution. *Rediation Research, 98*, 456-470.

McLachlan, G. J. (1987). On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture. *Applied Statistics, 36*, 318-324.

Moran, P. A. P. (1970). On asymptotically optimal tests of composite hypothesis. *Biometrika, 57*, 47-55.

Neyman, J. (1959). *Optimal asymptotic tests for composite hypothesis*. In Probability and Statistics: The Harold Cramer Volume, U. Grenander (ed). New York: Wiley.

Neyman, J. & Scott, E. L. (1966). On the use of C $(\alpha)$ optimal tests of composite hypotheses. *Bulletin of the International Statistical Institute, 41*, 477-497.

Paul, S. R. (1989). Test for the equality of several correlation coefficients. *The Canadian Journal of Statistics, 93*, 217-227.

Paul, S. R. (1996). Score tests for intraclass correlation in familial data. *Biometrics 52*, 955-963.

Paul, S. R. & Banerjee, T. (1998). Analysis of two-way layout of count data involving multiple counts in each cell. *Journal of the American Statistical Association, 93*, 1419-1429.

Paul, S. R. & Islam, A. S. (1995). Analysis of proportions based on parametric and semi-parametric models. *Biometrics, 51*, 1400-1410.

Rao, C. R. (1947). Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation. *Proceedings of the Cambridge Philosophical Society, 44*, 50-57.

Self, S. G. & Liang, K. Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association, 82*, 605-610.

Weil, C. S. (1970). Selection of valid number of sampling units and a consideration of their combination in toxicological studies involving reproduction, teratogenesis or carcinogenesis reproduction, teratogenesis. *Food and Cosmetic Toxicology, 8*, 177-182.

Weinberg, P. & Gladen, B. C. (1986). The Beta-geometric distribution applied to comparative fecundability studies. *Biometrics, 42*, 547-560.

Williams, D. A. (1975). Analysis of binary responses from toxicological experiments involving reproduction and teratogenicity. *Biometrics, 31*, 949-952.

Journal of Modern Applied Statistical Methods
November, 2005, Vol. 4, No. 2, 434-445

1538 – 9472/05/$95.00

# Sample Size Calculation and Power Analysis of Time-Averaged Difference

Honghu Liu
David Geffen School of Medicine
UCLA

Tongtong Wu
Department of Biostatistics
UCLA

Little research has been done on sample size and power analysis under repeated measures design. With detailed derivation, we have shown sample size calculation and power analysis equations for time-averaged difference to allow unequal sample sizes between two groups for both continuous and binary measures and explored the relative importance of number of unique subjects and number of repeated measurements within each subject on statistical power through simulation.

Key words: sample size calculation; power analysis; repeated measures design; time-averaged difference

## Introduction

Sample size calculation and power analysis are essentials of a statistical design in studies. As statistical significance is likely the desired results of investigators, proper sample size and sufficient statistical power are of primary importance of a study design (Cohen, 1988). Although a larger sample size yields higher power, one cannot have as large a sample size as one wants, since sample subjects are not free and the resources to recruit subjects are always limited. As a result, a good statistical design that can estimate the needed sample size to detect a desired effect size with sufficient power will be critical for the success of a study.

Some research has been done for sample size calculation and power analysis regarding different designs with cross-sectional data, such as difference between correlations, sign-test (Dixon & Massey, 1969), difference between means with two group t-test or analysis of variance (ANOVA) (Machin, Campbell, Fayers, & Pinol, 1997), contingence tables (Agresti, 1996), difference of proportions between two groups, F-test (Scheffé, 1959), multiple regressions and logistic regressions (Whittemore, 1981; Hsieh et al., 1998).

However, little research has been done about sample size calculation and power analysis with repeated measures design, especially for unbalanced designs, which is widely used in biological, medical, health services research and other fields. For example, in research for diseases with low incidence and prevalence; designs where the non-diseased group is much larger than the diseased group to ensure a sufficient large sample size for multivariate modeling.

Unbalanced repeated measures situations also emerge in cluster randomized trials (Eldridge et al., 2001). Diggle et al. (1994) proposed a basic sample size calculation formula for time-averaged difference (TAD) with both continuous and binary outcome measures for the situation only with equal sample size in each group. Fitzmaurice et al. (2004) proposed a two-stage approach for sample size and power analyses of change in mean response over time for both continuous and binary outcomes.

Statistical software and routines have made sample size calculation and power analysis process much easier and flexible for researchers. With statistical software, one can efficiently

Dr. Honghu Liu Professor of Medicine in the Division of General Internal Medicine & Health Services Research of the David Geffen School of Medicine at UCLA. 911 Broxton Plaza, Los Angeles, CA 90095-1736. Email: hhliu@ucla.edu. Tongtong Wu is a Ph.D. Candidate in the Department of Biostatistics in the UCLA School of Public Health. 911 Broxton Plaza, Los Angeles, CA, 90095-1736. Email: tongtong@ucla.edu

examine designs with different parameters and select the best design to fit the need of a research project. Currently, there are many types of statistical software that can conduct sample size and power analyses. These include the general purpose software which contain power analysis routines such as: NCSS (NCSS, 2002), SPSS (SPSS Inc., 1999), and STATA (STATA Press, 2003); general purpose software that can be used to calculate power (i.e., contain non-central distribution or simulation purpose) such as: SAS (SAS Institute Inc., 1999), S-Plus (MathSoft, 1999), and XLISP-STAT (Wiley, 1990); and stand-alone power analysis software such as: NCSS-PASS 2002 (NCSS, 2002), nQuery advisor (Statistical Solutions, 2000), and PowerPack (Length, 1987). A comprehensive list of sample size and power analysis software can be found at http://www.insp.mx/dinf/stat_list.html.

Although a lot of software can conduct sample size and power analyses, they are basically all for data with different cross-sectional designs. The only software that can conduct sample size and power analyses with repeated measures design is NCC-PASS 2002, which handles power analysis for repeated measures ANOVA design. There is, however, no software available for TAD with repeated measures design.

In this article, a formula has been developed for sample size calculation and power analysis of TAD for both continuous and binary measures to allow unequal sample size between groups. In addition, the relative impact and equivalence of number of subjects and the number of repeated measures from each subject on statistical power was examined. Finally, a unique statistical software for conducting sample size and power analysis for TAD was created.

## Methodology

### Sample size Calculation and Power Analysis

Sample size calculation and power analysis are usually done through statistical testing of the difference under a specific design when the null or alternative hypothesis is true. Although there are many factors that influence sample size and power of a design, the essential factors that have direct impact on sample size and statistical power are type I error ($H_0$ may be rejected when it is true and its probability is denoted by $\alpha$), type II error ($H_0$ may be accepted when it is false and its probability is denoted by $\beta$), effect size (difference to be tested and it is usually denoted by $\Delta$) and variation of the outcome measure of each group (for example, standard deviation $\sigma$). Sample size and power are functions of these factors. Sample size and power analysis formulas link all of them together. For example, the sample size calculation formula for a two group mean comparison can be written as a function of the above factors:

$$n_2 = ((z_{1-\beta} + z_{1-\alpha/2})/(\Delta/S))^2/(1+1/r),$$

where $n_2$ is the sample size for group2, $S$ is the common standard deviation of the two groups, $r$ $0 < r \le 1$ is a parameter that controls the ratio between the sample sizes of group 1 and group 2 (i.e., $n_1 = n_2/r$). $z_{1-\beta}$ is the normal deviate for the desired power, $z_{1-\alpha/2}$ is the normal deviate for the significance level (two-sided test) and $\Delta$ is the difference to be detected.

For given levels of a type I error, a type II error and an effect size, sample size and statistical power are positively related: the larger the sample size, the higher the statistical power. Type I error is negatively related to sample size: the smaller Type I error, the larger sample size that is required to detect the effect size for a given statistical power. The larger type II error, the smaller power and thus one will need smaller sample size to detect a given effect size.

### Repeated Measures Design

### Time-Averaged Difference (TAD)

In many biomedical or clinical studies, researchers use the experimental design that takes multiple measurements on the same subjects over time or under different conditions. By using this kind of repeated measures design, treatment effects can be measured on "units" that are similar and precision can be determined by variation within same subject. Although the analyses become more complicated because

measurements from the same individual are no longer independent, the repeated measures design can avoid the bias from a single snapshot and is very popular in biological and medical research.

Suppose there are two groups, group 1 and group 2, and one would like to compare the means of an outcome, which could vary from time to time or under different situations between the two groups. With cross-sectional design, one will directly compare the means of the outcome between the groups with one single measure from each subject, which may not reflect the true value of the individual.

For example, it is known that an individual's blood pressure is sensitive to many temporary factors, such as mood, the amount of time slept the night before and the degree of physical exercise/movement right before taking the measurement. This is why the mean blood pressure of a patient is always examined from multiple measurements to determine his/her true blood pressure level. If only a single blood measurement is taken from each individual, then comparing mean blood pressure between two groups could be invalid as there is large variation among the individual measures for a given patient. To increase precision, the best way to conduct this is to obtain multiple measurements from each individual and to compare the time-averaged difference between the two groups (Diggle, 1994).

Notations

Suppose that there is a measurement for each individual $y_{g(ij)}$, where $g = 1,2$ indicating which group, $i = 1,...,m_k$ (with $k = 1,2$) indicating the number of individuals in each group, and $j = 1,...,n$ indicating the number of repeated measures from each individual subject. Then TAD will be defined as:

$$d = ((\sum_{i=1}^{m_1}\sum_{j=1}^{n} y_{1(ij)})/n*m_1)) - ((\sum_{i=1}^{m_2}\sum_{j=1}^{n} y_{2(ij)})/n*m_2).$$

The following notations will be used to define the different quantities used in sample size calculation and power analysis for TAD:

1. $\alpha$: Type I error rate
2. $\beta$: Type II error rate
3. $d$: Smallest meaningful TAD difference to be detected
4. $\sigma$: Measurement deviation (assume to be equal for the two groups)
5. $n$: Number of repeated observations per subject
6. $\rho$: Correlation between measures within an individual
7. $m_1, m_2$: Number of subjects in group 1 and group 2, respectively
8. $M = m_1 + m_2$: Total number of subjects in the design
9. $\pi = m_1 / M$: Proportion of number of subjects within group 1 ($\pi = 0.5$ gives equal sample size. $m_1 = \pi M, m_2 = (1-\pi)M$ )

Using the above notations, the next two sections will derive the sample size calculation formula for TAD between two groups with the flexibility of possible unequal sample size from each group for continuous and binary measures, respectively.

Continuous responses

Consider the problem of comparing the time-averaged difference of a continuous response between two groups. Supposed the model is of the following form:

$$Y_{ij} = \beta_0 + \beta_1 x + \varepsilon_{ij}, \quad i = 1,\cdots,M;\ j = 1,\cdots n$$

where $x$ indicates the treatment assignment, $x = 1$ for group 1 and $x = 0$ for group 2. To test if the time-averaged difference is zero is equivalent to test $H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$. Without showing details of derivation, Diggle et al. (1994) have shown the sample size in the situation when group 1 and group 2 have the same sample size. With step by step derivation, here it is shown generally to the cases that the sample sizes of two groups could be unequal. Assume that the within subject correlation

$$Corr(y_{ij}, y_{ik}) = \rho \quad \text{for any} \quad j \neq k$$

and

$$Var\ (y_{ij}) = \sigma^2 .$$

Without lost generality, it is assumed that the smallest meaningful difference $d > 0$, and let the power of the test be $1-\beta$. Under $H_0$:

$$z = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)} \rightarrow N(0,1)$$

The above model can be written in matrix form:

$$Y_i = X'\beta + \varepsilon$$

where

$$X_i = \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & 1 \end{pmatrix} \text{ for group 1}$$

or

$$X_i = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \end{pmatrix} \text{ for group 2}$$

and

$$Y_i = \begin{pmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{in} \end{pmatrix}$$

The variance-covariance matrix (compound symmetry) can be written as

$$\Sigma = \sigma^2 \begin{bmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{bmatrix}.$$

The estimates of regression coefficients of such a model are

$$\hat{\beta} = \left( \sum_i X_i' \Sigma^{-1} X_i \right)^{-1} \left( \sum_i X_i' \Sigma^{-1} Y_i \right),$$

and the estimates of variance estimate are

$$\begin{aligned} \text{var}(\hat{\beta}) &= \sigma^2 \left( \sum_i X_i' \Sigma^{-1} X_i \right)^{-1} \\ &= \frac{\sigma^2[1+(n-1)\rho]}{n[(m_1+m_2)m_2 - m_1^2]} \begin{bmatrix} m_2 & -m_1 \\ -m_1 & m_1+m_2 \end{bmatrix} \end{aligned}$$

By definition, it is known that

$$\begin{aligned} \text{Power} &= 1-\beta \\ &= \Pr(\text{rejecting } H_0 \mid H_1) = \Pr(|z| > z_{1-\alpha/2} \mid H_1) \end{aligned}$$

so,

$$\begin{aligned} &\text{Power} \\ &= \Pr\left( \left| \frac{\hat{\beta}_1}{se(\hat{\beta}_1)} \right| > z_{1-\alpha/2} \mid H_1 \right) \\ &= \Pr\left( \frac{\hat{\beta}_1}{se(\hat{\beta}_1)} > z_{1-\alpha/2} \mid H_1 \right) + \Pr\left( \frac{\hat{\beta}_1}{se(\hat{\beta}_1)} < -z_{1-\alpha/2} \mid H_1 \right) \\ &\approx \Pr\left( \frac{\hat{\beta}_1}{se(\hat{\beta}_1)} > z_{1-\alpha/2} \mid H_1 \right) \end{aligned}$$

it is assumed that $d > 0$, therefore, the second term can be ingored

$$= \Pr\left( \frac{\hat{\beta}_1 - d}{se(\hat{\beta}_1)} > z_{1-\alpha/2} - \frac{d}{se(\hat{\beta}_1)} \mid H_1 \right)$$

Therefore,

$$-z_{1-\beta} = z_{1-\alpha/2} - \frac{d}{se(\hat{\beta}_1)},$$

or

$$(z_{1-\alpha/2} + z_{1-\beta})^2 = \frac{d^2}{\mathrm{var}(\hat{\beta}_1)} = \frac{n[(m_1+m_2)m_2 - m_1^2]d^2}{\sigma^2[1+(n-1)\rho](m_1+m_2)}$$

In other words, given power $1-\beta$, the total sample size needed to detect the smallest meaningful difference $d > 0$ is

$$M = \frac{(z_{1-\alpha/2} + z_{1-\beta})^2[1+(n-1)\rho]s^2}{n(1-\pi-\pi^2)d^2}, \qquad (1)$$

where $s$ is the estimate of standard deviation. When $m_1 = m_2 = m$, the above formula becomes the same as that shown in Diggle et al. (1994) for balanced design:

$$m = \frac{2(z_{1-\alpha/2} + z_{1-\beta})^2[1+(n-1)\rho]s^2}{nd^2}. \qquad (2)$$

Given sample size,

$$z_{1-\beta} = -z_{1-\alpha/2} + \frac{d}{se(\hat{\beta}_1)} = -z_{1-\alpha/2} + \frac{\sqrt{nM(1-\pi-\pi^2)}\cdot d}{\sqrt{1+(n-1)\rho}\cdot s}$$

Therefore, the power of the test can be written as:

$$\text{Power} = 1-\beta = 1-\Phi\left(z_{1-\alpha/2} - \frac{\sqrt{nM(1-\pi-\pi^2)}\cdot d}{\sqrt{1+(n-1)\rho}\cdot s}\right) \qquad (3)$$

Binary responses

Suppose a binary response variable is to be compared between group 1and group 2. Assume

$$\Pr(Y_{ij}=1) = \begin{cases} p_1 \text{ in group 1} \\ p_2 \text{ in group 2} \end{cases}$$

To test if the proportions of responses being 1of the two groups are equal, the following model is considered

$$E(Y_{ij} \mid x_{ij}) = \Pr(Y_{ij}=1 \mid x_{ij}) = \beta_0 + \beta_1 x_{ij},$$
$$i = 1, \cdots, M;\ j = 1, \cdots n$$

where $x$ indicates the treatment assignment, $x=1$ for group 1 and $x=0$ for group 2. this test will be equivalent to test $H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$. Without showing the details, Diggle et al (1994) have shown the sample size in the situation when group 1 and group 2 have the same sample size. With step by step derivation, here it is generalized to the case that the sample size could be different between the two groups.

Suppose $d = p_1 - p_2 > 0$ and the power of the test is $1-\beta$. Under $H_0$, the estimate of $\sigma^2$ is

$$\hat{\sigma}_0^2 = \frac{m_1 p_1 + m_2 p_2}{m_1+m_2}\cdot\left(1 - \frac{m_1 p_1 + m_2 p_2}{m_1+m_2}\right) = \frac{(m_1 p_1 + m_2 p_2)(m_1 q_1 + m_2 q_2)}{m_1+m_2}$$

where $q_1 = 1-p_1$ and $q_2 = 1-p_2$. Under $H_1$, the estimate of $\sigma^2$ is

$$\hat{\sigma}_1^2 = \frac{m_1}{m_1+m_2}p_1 q_1 + \frac{m_2}{m_1+m_2}p_2 q_2 = \frac{m_1 p_1 q_1 + m_2 p_2 q_2}{m_1+m_2}$$

The variance estimator of $\hat{\beta}_1$ is

$$\text{var}(\hat{\beta}_1) = \frac{\sigma^2 (m_1 + m_2)[1 + (n-1)\rho]}{n[(m_1 + m_2)m_2 - m_1^2]},$$

and it is denoted as $\hat{\sigma}_{\hat{\beta}_1, H_0}$ when replacing $\sigma^2$ by $\hat{\sigma}_0^2$, and $\hat{\sigma}_{\hat{\beta}_1, H_1}$ when replacing $\sigma^2$ by $\hat{\sigma}_1^2$.

The power of the test is:

$$\begin{aligned}
&\text{Power} \\
&= \Pr\left(\left|\frac{\hat{\beta}_1}{se(\hat{\beta}_1)}\right| > z_{1-\alpha/2} \mid H_1\right) \\
&\approx \Pr\left(\frac{\hat{\beta}_1}{se(\hat{\beta}_1)} > z_{1-\alpha/2} \mid H_1\right) \quad \text{because we assume } d > 0 \\
&= \Pr\left(\frac{\hat{\beta}_1 - d}{\hat{\sigma}_{\hat{\beta}_1, H_0}} > z_{1-\alpha/2} - \frac{d}{\hat{\sigma}_{\hat{\beta}_1, H_0}} \mid H_1\right) \\
&= \Pr\left(\frac{\hat{\beta}_1 - d}{\hat{\sigma}_{\hat{\beta}_1, H_1}} \cdot \frac{\hat{\sigma}_{\hat{\beta}_1, H_1}}{\hat{\sigma}_{\hat{\beta}_1, H_0}} > z_{1-\alpha/2} - \frac{d}{\hat{\sigma}_{\hat{\beta}_1, H_0}} \mid H_1\right) \\
&= \Pr\left(\frac{\hat{\beta}_1 - d}{\hat{\sigma}_{\hat{\beta}_1, H_1}} > \frac{\hat{\sigma}_{\hat{\beta}_1, H_0}}{\hat{\sigma}_{\hat{\beta}_1, H_1}} \cdot z_{1-\alpha/2} - \frac{d}{\hat{\sigma}_{\hat{\beta}_1, H_1}} \mid H_1\right)
\end{aligned}$$

Therefore,

$$-z_{1-\beta} = \frac{\hat{\sigma}_{\hat{\beta}_1, H_0}}{\hat{\sigma}_{\hat{\beta}_1, H_1}} \cdot z_{1-\alpha/2} - \frac{d}{\hat{\sigma}_{\hat{\beta}_1, H_1}},$$

Or

$$\left(\frac{\hat{\sigma}_{\hat{\beta}_1, H_0}}{\hat{\sigma}_{\hat{\beta}_1, H_1}} \cdot z_{1-\alpha/2} + z_{1-\beta}\right)^2 = \frac{d^2}{\hat{\sigma}^2_{\hat{\beta}_1, H}}$$

i.e.,

$$\left(\sqrt{\frac{(m_1 p_1 + m_2 p_2)(m_1 q_1 + m_2 q_2)}{m_1 p_1 q_1 + m_2 p_2 q_2}} \cdot z_{1-\alpha/2} + z_{1-\beta}\right)^2 = \frac{nM(1-\pi-\pi^2)d^2}{[1+(n-1)\rho][\pi p_1 q_1 + (1-\pi) p_2 q_2]}$$

In other words, given power $1-\beta$, the total sample size needed to detect the smallest meaningful difference $d > 0$ is

$$M = \frac{\left(\sqrt{\frac{(m_1 p_1 + m_2 p_2)(m_1 q_1 + m_2 q_2)}{m_1 p_1 q_1 + m_2 p_2 q_2}} \cdot z_{1-\alpha/2} + z_{1-\beta}\right)^2 [1+(n-1)\rho][\pi p_1 q_1 + (1-\pi) p_2 q_2]}{n(1-\pi-\pi^2)d^2} \tag{4}$$

When $m_1 = m_2$, the above formula is the same as shown in Diggle et al. (1994) for balanced design. Given sample size, the power of the test can be calculated using the following equation:

$$\text{Power} = 1-\beta = \Phi\left(\frac{\hat{\sigma}_{\hat{\beta}_1, H_0}}{\hat{\sigma}_{\hat{\beta}_1, H_1}} \cdot z_{1-\alpha/2} - \frac{d}{\hat{\sigma}_{\hat{\beta}_1, H_1}}\right) \tag{5}$$

The Relative Impact of Number of Subjects and Number of Repeated Measures on Power

As the cost and the amount of effort to recruit subjects or to increase the number of repeated measurements for each participant is often different, it will be useful for investigators to know the relative impact of number of subjects and number of repeated measures on statistical power for testing TAD. The relative importance of the total number of subjects M and number of repeated measures n, which have nonlinear effects on the power, is now investigated. For easy derivation, let's examine the situation of continuous measure.

First, if the within subject correlation is $\rho = 0$, then it can be seen that the number of subjects *M* and number of repeated measures *n* will have exactly the same impact on statistical

power. Using formula (3) and plugging in $\rho = 0$, the power then becomes:

$$\text{Power} = 1-\beta = 1-\Phi\left(z_{1-\alpha/2} - \frac{\sqrt{nM(1-\pi-\pi^2)}\cdot d}{s}\right) \quad (6)$$

It can be explained that when $\rho = 0$ all the observations are independent and thus there is no distinction between the repeated measurements and different subjects. Second, when $\rho = 1$, the number of repeated measures has no more impact on power because it just repeats the same observations over again. This can be seen by plugging in $\rho = 1$ in formula (3):

$$\text{Power} = 1-\beta = 1-\Phi\left(z_{1-\alpha/2} - \frac{\sqrt{M(1-\pi-\pi^2)}\cdot d}{s}\right) \quad (7)$$

To examine the impacts of M and n on the power when $0 < \rho < 1$, the amounts that need to be increased on M and n to achieve the same power are calculated. With other factors fixed and for a given n and M, how much does n need to be increased to achieve the same impact on power when increasing *M* by 1? Recall the power function is

$$\text{Power} = 1-\beta = 1-\Phi\left(z_{1-\alpha/2} - \frac{\sqrt{nM(1-\pi-\pi^2)}\cdot d}{\sqrt{1+(n-1)\rho}\cdot s}\right)$$

.

With other factors fixed, all that is required is to make the term,

$$\frac{nM}{1+(n-1)\rho},$$

a constant to achieve the same power. Let $n'$ be the new $n$ that will have the same impact on power as M increased by 1. Then the following equation can be solved

$$\frac{n(M+1)}{1+(n-1)\rho} = \frac{n'M}{1+(n'-1)\rho},$$

and the following equation is obtained:

$$n' = \frac{n(M+1)(1-\rho)}{M-(M+n)\rho} \quad (8)$$

Thus increasing *n* by the amount,

$$n' - n = \frac{n(1-\rho+n\rho)}{M-(M+n)\rho} \quad (9)$$

is the same as increasing M by 1. This amount of increment depends on M, n and $\rho$. For example, if $\rho = 0.5$, then n needs to increase by $n(1+n)/(M-n)$; if $\rho = 0.05$ n needs to increase by $n(0.95+0.05n)/(0.95M-0.05n)$ in order to have the same impact on power as M increased by 1.

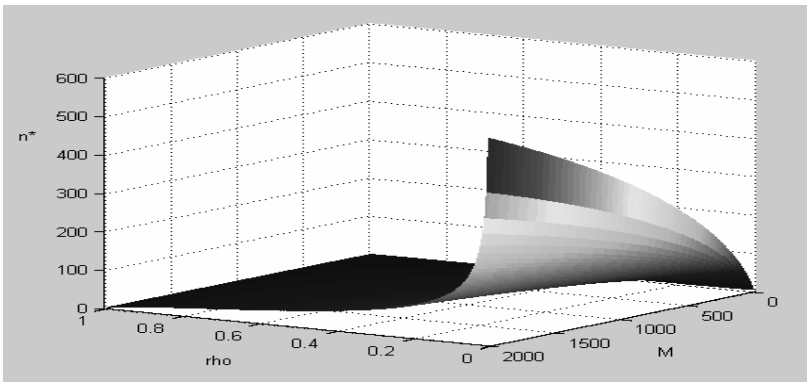
To examine which variable, M or n, has a larger impact on the power, it is required that one checks which variable needs to increase more to get the same power. The larger amount that needs to increase, the lower impact the variable has on statistical power. Set (9) equal to 1 and obtain the following equation.

$$\rho n^2 + n - (1-\rho)M = 0 \quad (10)$$

This is a quadratic function of n, and thus it has two roots

$$n^* = \frac{-1 \pm \sqrt{1+4\rho(1-\rho)M}}{2\rho} \quad (11)$$

Figure 1. The Relationship of n*, $\rho$ and M.




Because n is always greater than 0, the positive root is taken. To say that the amount (9) is greater than 1, is equivalent to stating that equation (10) is greater than 0, or n is greater than n*, the root of (10). In other words, the impact of n on power is smaller than the impact of M when n is greater than n*. Based on (11), one can see that n* depends on both M and $\rho$ nonlinearly. Figure 1 below shows the non-linear relationship among M, n and $\rho$.

This 3-D figure reveals that the threshold n* will increase when M increases but for a same M value, the threshold will be larger when $\rho$ smaller. Figure 2 and Figure 3 are special slides of the 3-D figure of Figure 1. Figure 2 shows the relationship between the threshold n* and $\rho$ for M=300 and Figure 3 shows the relationship between the threshold n* and M for $\rho$ =0.4.

Figure 2. The Relationship of n* and $\rho$, with M = 300 fixed.

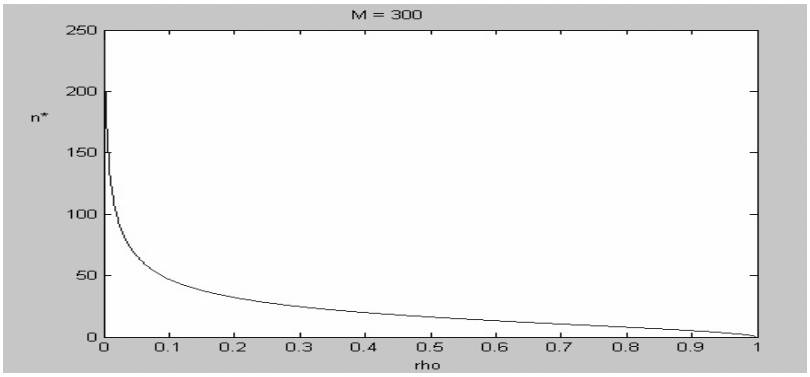
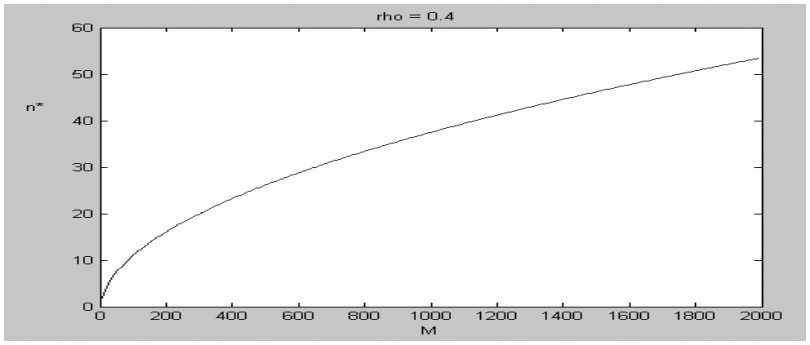



Figure 3. The Relationship of n* and M, with $\rho$ = 0.4 fixed.




%SP_TAD Software, Syntax and Parameters

To implement the algorithm for calculating the sample sizes or power for time-averaged difference, we have written a statistical macro procedure %SP_TAD, where SP stands for sample and power, TAD stands for time averaged difference in SAS/MACRO.

The syntax of the macro is simple and straightforward. To use this macro, one simply needs to invoke the macro with specific values for the parameters required. Here is the list of parameters that need to be specified:

(1) type-------continuous (=1) or binary (=2) responses. This sets up the tone of the type of the outcome measure to be analyzed. The following parameters of (2) to (9) must be provided for continuous responses:
(2) alpha----Type I error rate
(3) beta----- Type II error rate
(4) d--------Smallest meaningful difference to be detected
(5) sigma----Measurement deviation (for continuous responses)
(6) n--------Number of repeated observations per subject
(7) rho------Correlation among each subject
(8) pi--------Proportion of number of subjects within group 1
(9) M--------Total number subjects

For binary outcome, sigma is not needed. Instead, two more parameters need to be provided:

(10) pa-------Pr(Y_ij=1) in group 1
(11) pb-------Pr(Y_ij=1) in group 2

To run the macro, one needs simply to issue:

%sp_tad(type=, alpha=, beta=, d=, sigma=, n=, rho=, pi=, pa=, pb=, M=);

where pa and pb should be left as blank for continuous outcome, and sigma should be left blank for binary outcome. Beta and M should not be provided at the same time. To calculate required sample size, beta must be provided. To calculate power, M must be provided. Type is 1 or 2, where 1 stands for continuous responses and 2 stands for binary responses. The software code is available upon request from the author.

Application

Repeated measures design has wide applications in social, biological, medical and health service research. To avoid possible bias from snapshot of data collection at one time point and to reduce the cost of collecting data from different subjects, repeated measures data are often collected. Through a real example, this section demonstrates the input, output and the functionality of the %SP_TAD software and how the procedure works with continuous outcome measures. For binary outcome measures, the process will be similar.

For continuous measures, an example of a patient's diastolic blood pressure between a treatment and control group is examined (generally, diastolic blood pressure below 85 is considered "normal"). The level of a person's blood pressure could be affected by many temporary factors, such as the amount of time that the person slept last night, the person's mood, physical activity right before taking blood pressure measurement, etc. Thus, a one time snapshot of blood pressure will likely not be accurate. To accurately estimate the level of blood pressure of a patient or a group of patients, means of multiple measurements of blood pressure from a patient are usually used.

Suppose that a design is required to examine the difference of diastolic blood pressure between the treatment and control groups. To avoid bias from one time snapshot, five repeated measures of blood readings were taken from each patient within a week (one reading each day). Based on previous studies, intra-class correlation at the level of 0.4, type I error 0.05 and type II error 0.15 and a common standard deviation of 15 was used. Assume that a difference in mean blood pressure as small as 10 points between the treatment and control groups is desired. Since the treatment is more expensive than the control and more controls than treatment participants is desired, with a ratio of 3:2. Using these parameters, the calculation with the following syntax can be established:

%***sp_tad***(type=**1**, alpha=**0.05**, beta=**0.15**, d=**10**, sigma=**15**, n=**5**, rho=**0.4**, pi=**0.6**, pa=, pb=, M=);

Execute the procedure and the answer is 158 in treatment group and 105 in control group. Assume that the control group had a mean diastolic blood pressure 88. Then, the given sample size of 158 in the treatment group and 105 in the control group with 5 repeated measurements from each patient will allow one to detect a mean diastolic blood pressure of the treatment as low as 78.

For the same question, assume 158 patients in treatment group and 105 patients in the control group with 5 repeated measures of blood pressure. With a type I error 0.05, what kind of power will is needed to detect a difference in mean blood pressure of as small as 10 points? Using the same procedure, these parameters can be instituted and the macro with the following syntax can be executed:

%***sp_tad***(type=**1**, alpha=**0.05**, beta=, d=**10**, sigma=**15**, n=**5**, rho=**0.4**, pi=**0.6**, pa=, pb=, M=263);

The answer for power will be 85%.

## Conclusion

Time-averaged difference of repeated measures data has wide applications in many fields of research. TAD provides the opportunity to examine the difference in means between groups with higher precision using repeated measurements from each subject. This article deals with sample size and power analyses issues for time-averaged difference of repeated measures design. It presents the details of derivation of the general sample size calculation and power analysis formula for TAD with unequal sample size between two groups. Allowing unequal sample size will enable researchers to have the opportunity to choose an unbalanced design so that smaller number of

subjects could be used for the group that is either more expensive, hard to recruit or with limited number of available subjects.

Repeated measures data points also arise from cluster randomized trials, where it typically has repeated individuals within randomized clusters. There is growing literature on the topic starting with initial work involving balanced equally sized groups, but is now extending to more complex situations, of which unequal group sizes is also a possible scenario (Eldridge, 2001).

Repeated measures data has two dimensions of sample sizes: the number of different individuals and the number of repeated measurements from each individual. As shown in the article, because data from different individuals are independent, the number of different subjects seems to have a larger effect on power than the number of repeated measurements from the same subject. However, there is a threshold of the number of repeated measures, which will yield a larger impact by increasing the number of repeated measures than by increasing the number of subjects on statistical power. However, increasing the number of subjects by 1 means to increase the number of observations by n (the new subject gets n repeated measurements as others) and increasing the number of repeated measures by 1 means to increase the number of observations by M (every subject increases one repeated measurement). Thus, when $\rho$ is very small (i.e. about zero), one will need a larger n to exceed n*, the threshold, in order to have a larger impact of increment of n than M on power.

In most of the situations, n is not large and much smaller than M, thus likely M will have larger impact than n. For the two extreme cases where $\rho = 0$ or $\rho = 1$, the impact of the increase of the number of repeated measures will be the same as the increase of the number of individuals in each group ( $\rho = 0$ ) or there will be no impact of increasing the number of repeated measures ( $\rho = 1$ ) on power.

The software created is easy to use and can handle both continuous outcome measure and dichotomous outcome measure by issuing a value of "1" or "0" for the parameter "type". For the same software, one can also estimate the underlying statistical power for a given sample size with a given type I error, type II error, variation and effect size.

## References


Agresti, A. (1996). *An introduction tocategorical data analysis*. Wiley: New York.

Cohen, J. (1988). *Statistical poweranalysis for the behavioral sciences*. Second edition. Lawrence Erlbaum Associates: Hove and London.

Diggle, P. J., Liang, K. Y., & Zegger,S. L. (1994). *Analysis of longitudinal data*. Oxford University Press: Oxford.

Dixson, W. J., Massey, F. J. (1969).*Introduction to statistical analysis*. McGraw-Hill: New York.

Elashoff, J. D. (2000). *nQuery advisor* (Version 4.0.). Statistical Solutions: Cork, Ireland.

Lenth, R. V. (1987). *"PowerPack," Software for IBM PCs and compatibles. Provides an interactive environment for power and sample-size calculations and graphics.*

Eldridge, S., Cryer, C., Defer, G., & Underwood, M. (2001). Sample size calculation for intervention trials in primary care randomizing by primary care group: an empirical illustration from one proposed intervention trial. *Statistics in Medicine 20*(3), 367-376.

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

Hsieh, F. Y., Block, D. A., & Larsen, M. D. (1998). A simple method of sample size calculation for linear and logistic regression. *Statistics in Medicine, 7*, 1623-1634.

Machin, D., Campbell, M., Fayers, P., &Pinol, A. (1997). *Sample size tables for clinical lstudies* (2nd ed.). London: Blackwell Science.

NCSS Statistical Software, NCSS: Kaysville, Utah, 2002.

SAS/IML, User's Guide, Version 8. SAS Institute Inc: Cary , NC, 1999.

SAS/STAT, User's Guide, Version 9. SAS Institute Inc: Cary, NC, 1999.

Scheffé, H. (1959). *The analysis ofvariance*. Wiley: New York.

S-PLUS 2000 User's Guide, MathSoft Data Analysis Products Division: Seattle, WA, 1999.

SPSS Base 10.0 for Windows User's Guide. SPSS Inc.: Chicago IL, 1999.

STATA, Version 8. STATA Press:Texas. 2003.

Tierney, L. (1990). *Lisp-Stat, an object-oriented environment for dynamic graphics*. Wiley: New York.

Whittemore, A. (1981). Sample size forlogistic regression with small response probability. *Journal of the American Statistical Association, 76,* 27-32.

Journal of Modern Applied Statistical Methods
November, 2005, Vol. 4, No. 2, 446-459


1538 – 9472/05/$95.00

# Type I Error Of Four Pairwise Mean Comparison Procedures Conducted As Protected And Unprotected Tests

J. Jackson Barnette
Department of Biostatistics
University of Alabama at Birmingham

James E. McLean
Program of Educational Research
University of Alabama, Tuscaloosa

Type I error control accuracy of four commonly used pairwise mean comparison procedures, conducted as protected or unprotected tests, is examined. If error control philosophy is experimentwise, Tukey's HSD, as an unprotected test, is most accurate and if philosophy is per-experiment, Dunn-Bonferroni, conducted as an unprotected test, is most accurate.

Key words: Type I error control, experimentwise vs. per-experiment error, protected vs. unprotected tests, pairwise comparisons, Tukey's HSD, Dunn-Bonferroni, Dunn-Sidak, Holm's sequentially rejective

## Introduction

Whenever a researcher has more than two comparisons to test, control of the Type I error-rate becomes a concern. Soon after Fisher developed the process of analysis of variance (ANOVA), he recognized the potential problem of the error-rate becoming inflated when multiple *t* tests were performed on three or more groups.

He discussed this problem in the 1935 edition of his famous book, *The Design of Experiments*. His recommendation of using a more stringent alpha when performing his Least Significant Difference Procedure (LSD) is based on this concern. However, researchers still criticized the LSD as providing inadequate control of Type I error. This early recognition of the problem has resulted in hundreds of multiple comparison procedures being developed over the years.

J. Jackson Barnette is Senior Associate Dean for Academic Affairs and Professor of Biostatistics in the School of Public Health at the University of Alabama at Birmingham. Email: barnette@uab.edu. James E. McLean is University Professor and Dean in the College of Education at The University of Alabama, Tuscaloosa.

The earliest example of what is now known as a multiple comparison procedure could be found in 1929, when Working and Hotelling applied simultaneous confidence intervals to regression lines. The Fisher (1935) reference cited earlier was the first application to the process of ANOVA. The Type I error-rate control problem was also referred to by Pearson and Sekar in 1936 and Newman in 1939. Newman described a multiple comparison test that used the "Studentized Range Statistic." It is said that his work was prompted by a discussion he had with Student. Years later, Keuls published an updated version of the procedure (1952) using the Studentized range. That multiple comparison procedure is now known as the Student-Newman-Keuls procedure.

Most studies of Type I error rates for follow-up of pairwise mean differences have been based on what is referred to as experimentwise or familywise error control philosophies. These terms were more extensively described by Ryan (1959) and Miller (1966). Experimentwise (EW) Type I error relates to finding at least one significant difference by chance for the specified alpha level. In these cases, the only difference of concern is the largest mean difference. Experimentwise Type I error control ignores the possibility of multiple Type I errors in the same experiment. The pairwise mean differences for

those other than the largest mean difference are not considered. Type I error control is such that not all possible Type I errors are evaluated. In these cases, many procedures such as Tukey's HSD are considered to have conservative Type I error control since the actual probabilities of finding at least one Type I error are lower than the nominal alpha level.

Per-experiment (PE) Type I error control considers all the possible Type I errors that can occur in a given experiment. Thus, more than one Type I error per experiment is possible and reasonably likely to occur if there is an experimentwise Type I error on the highest mean difference. Klockars & Hancock (1994) pointed out the importance and risks associated with this distinction. They found, using a Monte Carlo simulation, that there was a difference of .0132 in the per-experiment and experimentwise Type I error rates for Tukey's HSD when alpha was set at .05. This discussion was expanded in their 1996 review titled "The Quest for α" (Hancock & Klockars). Thus, when one has exact control of Type I error in the experimentwise situation, the per-experiment Type I error probability is higher. One of the purposes of this research was to examine how much of a difference there may be between experimentwise and per-experiment Type I error rates for four of the most commonly used pairwise multiple comparison procedures when used with alpha levels of .10, .05, and .01, and to determine the relative influence on this difference of number of groups and number of subjects per group. While most Type I error research is based on an experimentwise mode, the per-experiment Type I error is more consistent with the reality of pairwise hypothesis testing. It considers not only the largest mean difference subjected to error control, but all the pairwise differences.

There seems to be an inconsistency of logic when comparing the power of various methods and manners of Type I error control. When it is stated that the Student-Newman-Keuls is more powerful than Tukey's HSD or Holm's procedure is more powerful than Dunn-Bonferroni; the notion is that one method leads to more rejections of partial null hypotheses. However, if one considers the notion of experimentwise Type I error (the largest pairwise difference or more being rejected), then SNK and HSD have the same power and Dunn-Bonferroni and Holm have the same power. Differences in power only come when considering pairwise differences that are found beyond the k number of means steps. Thus, should not error rate take into account the possible false rejections in the entire structure of mean differences, not just the largest one? Per-experiment Type I error control is more consistent with actual pairwise hypothesis decision-making.

Four multiple comparison procedures were selected for this research: Dunn-Bonferroni, Dunn-Sidak, Holm's sequentially rejective, and Tukey's HSD. Based on a review of current literature and commonly used statistical texts, it was concluded that these are among the most frequently used pairwise procedures and represent a variety of approaches to control for Type I error. Since the names of these procedures tend to vary slightly in texts, statistical software, and in the literature, each is described briefly below:

The Dunn-Bonferroni procedure uses the Bonferroni inequality ($\alpha_{PE} \leq \Sigma\alpha_{PC}$) as authority to divide equally the total a priori error among the number of tests to be completed, often following the application of the Fisher LSD procedure. The LSD procedure is equivalent to conducting all pairwise comparisons using independent *t* tests with the $MS_{error}$ as the common pooled variance estimate (Kirk, 1982). An example of the application of the Dunn-Bonferroni would be identifying the a priori α as .05 where tests are required to compare means of five groups using 10 comparisons, running each individual test at the .05/10= .005 level (Hays, 1988). Sidak's modification of the Dunn-Bonferroni procedure, referred to as the Dunn-Sidak procedure substituted the multiplicative computation of the exact error-rate, $\alpha_{PE} = 1 - (1 - \alpha_{PC})^c$ where c is the number of comparisons for the Bonferroni Inequality ($\alpha_{PE} \leq \Sigma\alpha_{PC}$), otherwise following the same procedures (Kirk, 1982).

A procedure proposed by Holm in 1979, Holm's Sequentially Rejective procedure is also referred to as the Sequentially Rejective Bonferroni procedure. Assuming a maximum of

c comparisons to be performed, the first null hypothesis is tested at the $\alpha/c$ level. If the test is significant, the second null hypothesis is tested at the $\alpha/(c - 1)$ level. If this is significant, the testing continues in a similar manner until all c tests have been completed or until a nonsignificant test is run. The testing stops when the first nonsignificant test is encountered (Hancock & Klockars, 1996).

Tukey's Honestly Significant Difference procedure (HSD) was presented originally in a non-published paper by Tukey in 1953. Its popularity has grown to the point where it is, possibly, the most widely used multiple comparison procedure. The HSD is based on the Studentized Range Statistic originally derived by Gossett (a.k.a., Student) (1907-1938). This statistic, unlike the *t* statistic, takes into account the number of means being compared, adjusting for the total number of tests to make all pairwise comparisons (Kennedy & Bush, 1985).

Many researchers follow the practice of conducting post-hoc pairwise multiple comparisons only after a significant omnibus *F* test. Protected tests are conducted only after a significant omnibus *F* test, while unprotected tests are conducted without regard to the significance of the omnibus *F* test. Many common statistical texts either recommend or imply the use of a protected test for all post-hoc multiple comparison procedures (e.g., Hays, 1988; Kennedy & Bush, 1985; Kirk, 1982; Maxwell & Delaney, 1990). While these texts provide a logical basis for this, and excellent reviews of multiple comparison procedures are available (e.g., Hancock & Klockars, 1996; Toothaker, 1993), little empirical evidence is presented, either analytically or empirically, to justify this practice.

The research questions addressed in this research are:

1. Which of these four multiple comparison procedures has the most accurate control of Type I error across the three alpha conditions?

2. Does error control accuracy differ when tests are conducted as protected or unprotected tests?

3. Do methods differ relative to accuracy when conducted as experimentwise vs. per-experiment control?

## Methodology

Monte Carlo methods were used to generate the data for this research. All data comprising the groups whose means were compared were generated from a random normal deviate routine, which was incorporated into a larger compiled QBASIC program that conducted all needed computations. The program was written by the senior author. All sampling and computation, conducted with double-precision, routines were verified using SAS® programs. Final analysis of the summary statistics and correlations was conducted using SAS®.

Several sample size and number of groups arrangements were selected to give a range of low, moderate, and large case situations. The numbers of groups were: 3, 4, 5, 6, 8, and 10 and the sample sizes for each group were: 5, 10, 15, 20, 30, 60, and 100, which when crossed gave 42 experimental conditions. This was replicated for three nominal alphas of .10, .05, and .01. The approach used was to determine what number of replications would be needed to provide an expected .95 confidence interval of +/- .001 around the nominal alpha.

This is an approach to examination of how well observed Type I error proportions are reasonable estimates of a standard nominal alpha. In other words, if alpha is the standard, what proportion of the estimates of actual Type I error proportions can be considered accurate, as evidenced by them being within the expected .95 confidence interval around nominal alpha?

This was based on the assumption that errors would be normally distributed around the binomial proportion represented by nominal alpha. Thus, when alpha was .10, 345742 replications were needed to have a .95 confidence interval of +/- .001 or between .099 and .101. When alpha was .05, 182475 replications were needed to have a .95 confidence interval of +/- .001 or between .049 and .051 and when alpha was .01, 38032 replications were needed to have a .95 confidence interval of +/- .001 or between .009

and .011. Observed Type I error proportions falling into the respective .95 confidence intervals are considered to be accurate estimates of the expected Type I error rate.

Within each nominal alpha/sample size/number of groups configuration, the number of ANOVA replications were generated. Each replication involved drawing of elements of the sample from a distribution of normal deviates, computation of sample means, and the omnibus *F* test. Error rates were determined for protected and unprotected tests for each of the four multiple comparison procedures. While Dunn-Bonferroni, Dunn-Sidak, and HSD use only one critical value for all differences, the pairwise differences were recorded in a hierarchical fashion to determine pairwise differences significant at each of the numbers of steps between means from k down to 2. This approach permitted determination of experimentwise Type I error (at least one Type I error per experiment) or a Type I error for the largest mean difference, and per-experiment Type I errors or the total number of Type I errors observed regardless of where they are in the stepwise structure.

Summary statistics were computed for each alpha level for experimentwise and per-experiment conditions including: the mean proportion of Type I errors, standard deviation of the proportion of Type I errors, and the percentage of those proportions falling in the three regions associated with the .95 confidence interval of nominal alpha +/- 0.001.Additional analysis included computation of differences between per-experiment proportions and experimentwise proportions (PE-EW).

Preliminary analyses were run using the Monte Carlo program to test its accuracy. First, 500,000 standard normal scores (*z* scores) were generated and the statistics for the distribution were computed. This resulted in a mean = -.00096, variance = 1.0013, skewness = .00056, kurtosis = .00067, and the Wilk-Shapiro D = .000734 (nonsignificant). Thus, we concluded that the program generates reasonable normal distributions. Second, 900,000 cases were computed with k ranging from 2 to 10 and n ranging from 5 to 100 with no differences between the group means. In each case, the proportions of significant *F* statistics were computed corresponding to preset alphas of .25, .10, .05, .01, .001, and .0001. The resulting proportions of rejected null hypotheses were .24989, .10106, .05071, .01022, .001004, and .000103 respectively. These results support the accuracy of the Monte Carlo program.

## Results

The first research question is: Which of these four multiple comparison procedures has the most accurate control of Type I error across the three alpha conditions? The results for each of the three alpha conditions are presented in Tables 1 through 3 and Figures 1 through 3. Table 1 and Figure 1 present results when nominal alpha is set at .10, Table 2 and Figure 2 present results when nominal alpha is set at .05, and Table 3 and Figure 3 present results when nominal alpha is set at .01.

When alpha is set at .10, if the Type I error rate philosophy is experimentwise, the most accurate of these four procedures is clearly Tukey's HSD, conducted as an unprotected test, with a mean observed Type I error rate of .09940 and with 78.6% of the observed Type I errors being in the range of .099 to .101. The HSD conducted as a protected test with an experimentwise control philosophy had a mean of .08134, somewhat conservative. All of the other procedures conducted, based on the experimentwise philosophy are conservative procedures with mean Type I error rates in the range of .07239 to .07535 when conducted as unprotected tests and .06695 to .06885 when conducted as protected tests.

If the Type I error control philosophy is per-experiment, the most accurate procedure is clearly the Dunn-Bonferroni, conducted as an unprotected test with a mean observed Type I error rate of .10011 and 85.7% of the observed Type I errors in the range of .099 to .101. When the philosophy is per-experiment and conducted as unprotected tests, the other three methods tend to be liberal with the mean error rate for the Dunn-Sidak at .10481 and the Holm procedure at .10582. Tukey's HSD was very liberal in this situation with a mean error rate of .14579. When conducted as protected tests, HSD was slightly liberal with a mean error of .12741 and the other three methods were reasonably accurate with mean errors of .09466 for the Dunn-Bonferroni,

.09834 for the Dunn-Sidak, and .10036 for Holm's procedure.

When nominal alpha was set at .05, the results were very similar. If the Type I error rate philosophy is experimentwise, the most accurate of these four procedures is clearly Tukey's HSD, conducted as an unprotected test, with a mean observed Type I error rate of .04993 and with 97.6% of the observed Type I errors being in the range of .049 to .051. The HSD conducted as a protected test with an experimentwise control philosophy had a mean of .03865, somewhat conservative. All of the other procedures conducted, based on the experimentwise philosophy are conservative procedures with mean Type I error rates in the range of .03864 to .03943 when conducted as unprotected tests and .03352 to .03395 when conducted as protected tests.

If the Type I error control philosophy is per-experiment, the most accurate procedure is clearly the Dunn-Bonferroni, conducted as an unprotected test with a mean observed Type I error rate of .04998 and 92.9% of the observed Type I errors in the range of .049 to .051. When the philosophy is per-experiment and conducted as unprotected tests, the other three methods tend to be liberal with the mean error rate for the Dunn-Sidak at .05110 and the Holm procedure at .05208. Tukey's HSD was very liberal in this situation with a mean error rate of .06674. When conducted as protected tests, HSD was slightly liberal with a mean error of .05531 and the other three methods were slightly conservative with mean errors of .04483 for the Dunn-Bonferroni, .04560 for the Dunn-Sidak, and .04696 for Holm's procedure.

When nominal alpha was set at .01, the patterns of results were very similar to the .10 and .05 nominal alpha conditions. If the Type I error rate philosophy is experimentwise, the most accurate of these four procedures is clearly Tukey's HSD, conducted as an unprotected test, with a mean observed Type I error rate of .01002 and with 100.0% of the observed Type I errors being in the range of .009 to .011. The HSD conducted as a protected test with an experimentwise control philosophy had a mean of .00702, somewhat conservative. All of the other procedures conducted, based on the experimentwise philosophy are conservative procedures with mean Type I error rates in the range of .00860 to .00865 when conducted as unprotected tests and .00647 to .00649 when conducted as protected tests. If the Type I error control philosophy is per-experiment, the most accurate procedure is clearly the Dunn-Bonferroni, conducted as an unprotected test with a mean observed Type I error rate of .01003 and 97.6% of the observed Type I errors in the range of .009 to .011.

When the philosophy is per-experiment and conducted as unprotected tests, the Dunn-Sidak outcome is very close to the Dunn-Bonferroni outcome with a mean error rate of .01007 and 92.9% of the observed errors in the .009 to .011 range. The other two methods tend to be liberal with the mean error rate for the Holm procedure at .01026 and Tukey's HSD with a mean error rate of .01181. When conducted as protected tests, all four methods were conservative with Tukey's HSD slightly less conservative with a mean error rate of .00878. The other three methods were slightly more conservative with mean errors of .00790 for the Dunn-Bonferroni, .00793 for the Dunn-Sidak, and .00814 for Holm's procedure.

In summary, relative to research question 1 (Which of these four multiple comparison procedures has the most accurate control of Type I error across the three alpha conditions?), if the most accurate control of per-experiment Type I error is desired, the Dunn-Bonferroni, conducted as an unprotected test, is the most accurate across all three levels of alpha. It consistently provides a mean Type I error rate closest to nominal alpha, has the lowest variance, and captures the highest proportion of observed Type I errors in the expected +/- .001 interval. Although the Dunn-Sidak and Holm provide values that are reasonably close, they tend to be slightly more liberal and less accurate, particularly with higher nominal alpha. As alpha decreases, both the Dunn-Sidak and Holm approach the level of accuracy of the Dunn-Bonferroni. Tukey's HSD is liberal as an unprotected test in control of per-experiment Type I error, although this decreases as alpha decreases. If the error control philosophy is experimentwise, Tukey's HSD is the most accurate, conducted as an unprotected test. It has a mean error closest to nominal alpha, the lowest

variance, and the highest proportion of observed Type I errors in the expected +/- .001 interval. When alpha is .10, HSD is slightly less accurate than when alpha is .05 or .01. The other three methods are conservative, with the Dunn-Sidak being slightly less conservative compared with Dunn-Bonferroni and Holm.

The second research question is: Does error control accuracy differ when tests are conducted as protected or unprotected tests? If the interest is in using any of these methods as a protected test, a practice not generally supported by these data, the HSD provides the most accurate control of experimentwise Type I error although it is very conservative at all alpha levels. The other three methods are very conservative in control of experimentwise Type I error. If per-experiment control of Type I error is the philosophy, HSD is liberal when alpha is .10 or .05 but becomes more accurate, even somewhat conservative, when alpha is .01. Of the remaining three, Holm's procedure tends to be more accurate across the three alpha levels. It is clear and expected that unprotected tests are more powerful than protected tests.

The third research question is: Do methods differ relative to accuracy when conducted as experimentwise vs. per-experiment control? It seems pretty clear that the results vary a great deal depending on the Type I error control philosophy. By the very nature of these philosophies, there will be a higher proportion of Type I errors in the per-experiment condition compared with the experimentwise condition. In every case, across alpha levels and for both protected and unprotected tests, the lowest difference between these rates was for the Dunn-Bonferroni, followed relatively closely by the Dunn-Sidak, Holm's procedure has next highest, and the highest difference was for the HSD. Thus, the issue is more a concern if one is using the HSD as compared with the other three methods.

## Conclusion

These results provide insights on two major controversies. One is the need for a significant omnibus F test as the gateway for conducting pairwise follow-ups (i. e., the protected test). Is it not possible, as Hancock & Klockars (1996) pointed out, that this requirement overprotects against finding pairwise differences? These results certainly support that claim, particularly when experimentwise Type I error is the control philosophy. Protected tests were more conservative in every case. It can clearly be concluded that none of these four tests should be used as protected tests when experimentwise error control is used. If per-experiment error control is desired, only the Holm procedure with alpha of .10 was more accurate as a protected test than as an unprotected test. However, that accuracy difference was lower when alpha was .05 or .01.

The other controversy is the use of experimentwise vs. per-experiment Type I error control. Clearly there is a difference in the error rates of these philosophies. The authors of this article contend that per-experiment mode is closest to the realities of pairwise hypothesis testing, because more than just the largest pairwise difference is of interest and all pairwise comparisons are tested. The conventional wisdom, based on experimentwise Type I error control, is that the Dunn-Bonferroni is very conservative and that the HSD is conservative, but less so.

The HSD is often recommended because it is conservative, yet provides reasonable power for finding significant differences; but this relates to experimentwise control and a protected test. Yet, arguments could be made that the HSD gets its power from a higher-than-nominal alpha level. In this research, when HSD is used as a protected test with alpha of .10 or .05, the actual per-experiment Type I error rates are .12741 and .05531 respectively and actual experimentwise Type I error rates were much lower at .08134 and .03865. Thus, the operational alpha level is not the nominal level, but a higher level.

If one is truly interested in maintaining an accurate level of control of Type I error, then methods which are shown to provide accurate actual controls should be used, and the power available can be determined by other comparison conditions: sample size, effect size, number of groups, and error variance. This research indicates that Tukey's HSD, conducted as an unprotected test, is the most accurate control of experimentwise Type I error. If it is

desired that accurate, as advertised, control of per-experiment Type I error be the primary criterion, there is one method that seems to provide that regardless of alpha level and that is the Dunn-Bonferroni conducted as an unprotected test.

These findings are not consistent with common wisdom or with recommendations found or implied in most statistics texts. However, it is hoped that this research influences others to replicate this work, possibly using other methods. Only when one is willing to question our current practice can one be able to improve on it.

Additional study of the discrepancy between experimentwise and per-experiment Type I errors is needed. Determining the importance of this discrepancy is required. The current study did not consider the case of unequal sample sizes or heterogenous variances. Is it the same under conditions of unequal sample sizes and/or variances? While it might be useful to include other procedures such as the Student-Newman-Keuls, Scheffé, and modifications of Holm's procedure, it is believed that it is unlikely that any of these methods will fare better as methods of Type I error control than Tukey's HSD when experimentwise is the control philosophy, or the Dunn-Bonferroni when per-experiment is the control philosophy and unprotected tests are used.

Table 1. Observed Per-Experiment and Experimentwise Type I Error Rates for Selected Multiple Comparison Procedures when Conducted as Protected and Unprotected Tests with Alpha= .10

| | | Protected Test | | | Unprotected Test | | |
|---|---|---|---|---|---|---|---|
| | | Per-Experiment Error (PE) | Experiment-wise Error (EW) | PE - EW Difference | Per-Experiment Error (PE) | Experiment-wise Error (EW) | PE - EW Difference |
| Dunn-Bonferroni | M | .09466 | .06695 | .02771 | .10011 | .07239 | .02772 |
| | M – α | –.00534 | –.03305 | | +.00011 | –.02767 | |
| | SD | .00427 | .00962 | | .00075 | .00626 | |
| | % in α +/-.001 | 19.0 | 0 | | 85.7 | 0 | |
| Dunn-Sidak | M | .09834 | .06885 | .02949 | .10481 | .07535 | .02946 |
| | M – α | –.00166 | –.03115 | | +.00481 | –.02465 | |
| | SD | .00401 | .00972 | | .00093 | .00625 | |
| | % in α +/-.001 | 19.0 | 0 | | 0 | 0 | |
| Holm | M | .10036 | .06695 | .03341 | .10582 | .07239 | .03343 |
| | M – α | +.00036 | –.03305 | | +.00582 | –.02761 | |
| | SD | .00739 | .00962 | | .00346 | .00626 | |
| | % in α +/-.001 | 2.4 | 0 | | 7.1 | 0 | |
| HSD | M | .12741 | .08134 | .04607 | .14579 | .09940 | .04639 |
| | M – α | +.02741 | –.01866 | | +.04579 | –.00060 | |
| | SD | .00906 | .00755 | | .01472 | .00102 | |
| | % in α +/-.001 | 0 | 0 | | 0 | 78.6 | |

**Figure 1**
**Accuracy of Type I Error Control with Experimentwise and Per-Experiment Control Conducted as Protected and Unprotected Tests when Nominal Alpha= .10 and % in .10 +/- 0.001**

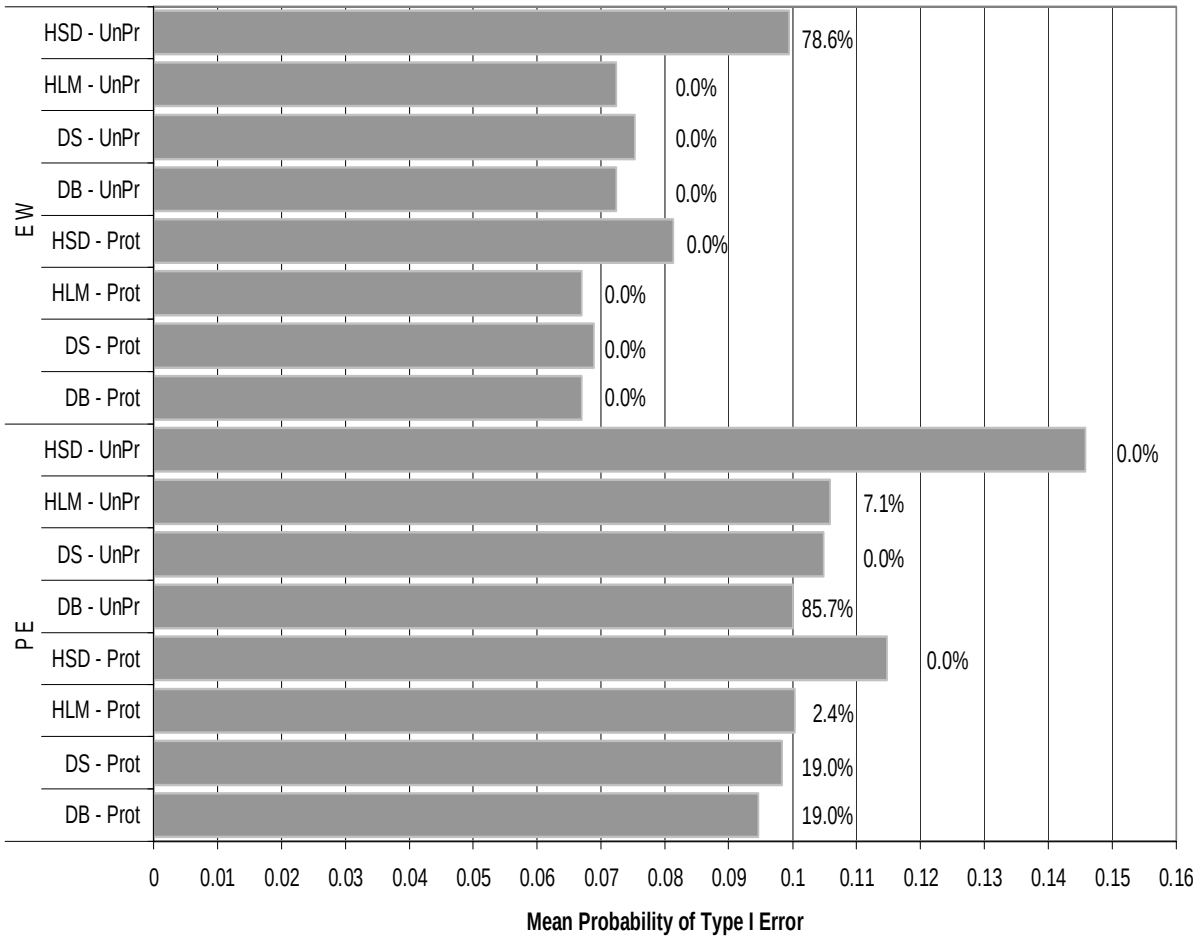


Table 2. Observed Per-Experiment and Experimentwise Type I Error Rates for Selected Multiple Comparison Procedures when Conducted as Protected and Unprotected Tests with Alpha= .05

| | | Protected Test | | | Unprotected Test | | |
|---|---|---|---|---|---|---|---|
| | | Per-Experiment Error (PE) | Experiment-wise Error (EW) | PE - EW Difference | Per-Experiment Error (PE) | Experiment-wise Error (EW) | PE - EW Difference |
| Dunn-Bonferroni | M | .04483 | .03352 | .01113 | .04998 | .03864 | .01134 |
| | M – α | –.00517 | –.01648 | | –.00002 | –.01136 | |
| | SD | .00315 | .00534 | | .00054 | .00294 | |
| | % in α +/-.001 | 7.1 | 0 | | 92.9 | 0 | |
| Dunn-Sidak | M | .04560 | .03395 | .01165 | .05110 | .03943 | .01167 |
| | M – α | –.00440 | –.00405 | | +.00110 | –.01057 | |
| | SD | .00308 | .00536 | | .00052 | .00291 | |
| | % in α +/-.001 | 16.7 | 0 | | 50.0 | 0 | |
| Holm | M | .04696 | .03352 | .01344 | .05208 | .03864 | .01344 |
| | M – α | –.00304 | –.01648 | | +.00208 | –.01136 | |
| | SD | .00433 | .00535 | | .00146 | .00294 | |
| | % in α +/-.001 | 19.0 | 0 | | 33.3 | 0 | |
| HSD | M | .05531 | .03865 | .01666 | .06674 | .04993 | .01681 |
| | M – α | +.00531 | –.01135 | | +.01674 | –.00007 | |
| | SD | .00324 | .00458 | | .00541 | .00048 | |
| | % in α +/-.001 | 2.4 | 0 | | 0 | 97.6 | |

**Figure 2**
**Accuracy of Type I Error Control with Experimentwise and Per-Experiment Control Conducted as Protected and Unprotected Tests when Nominal Alpha= .05 and % in .05 +/- 0.001**

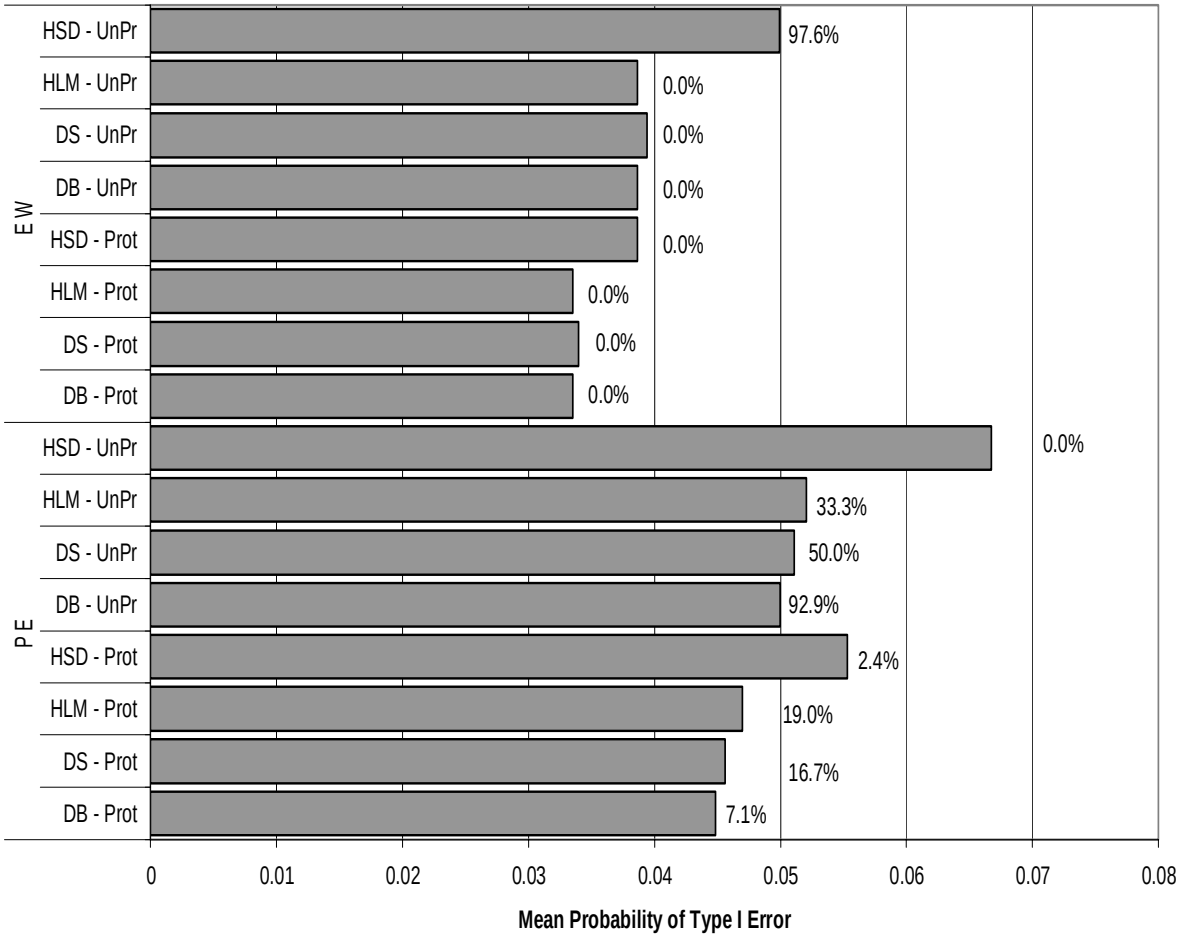
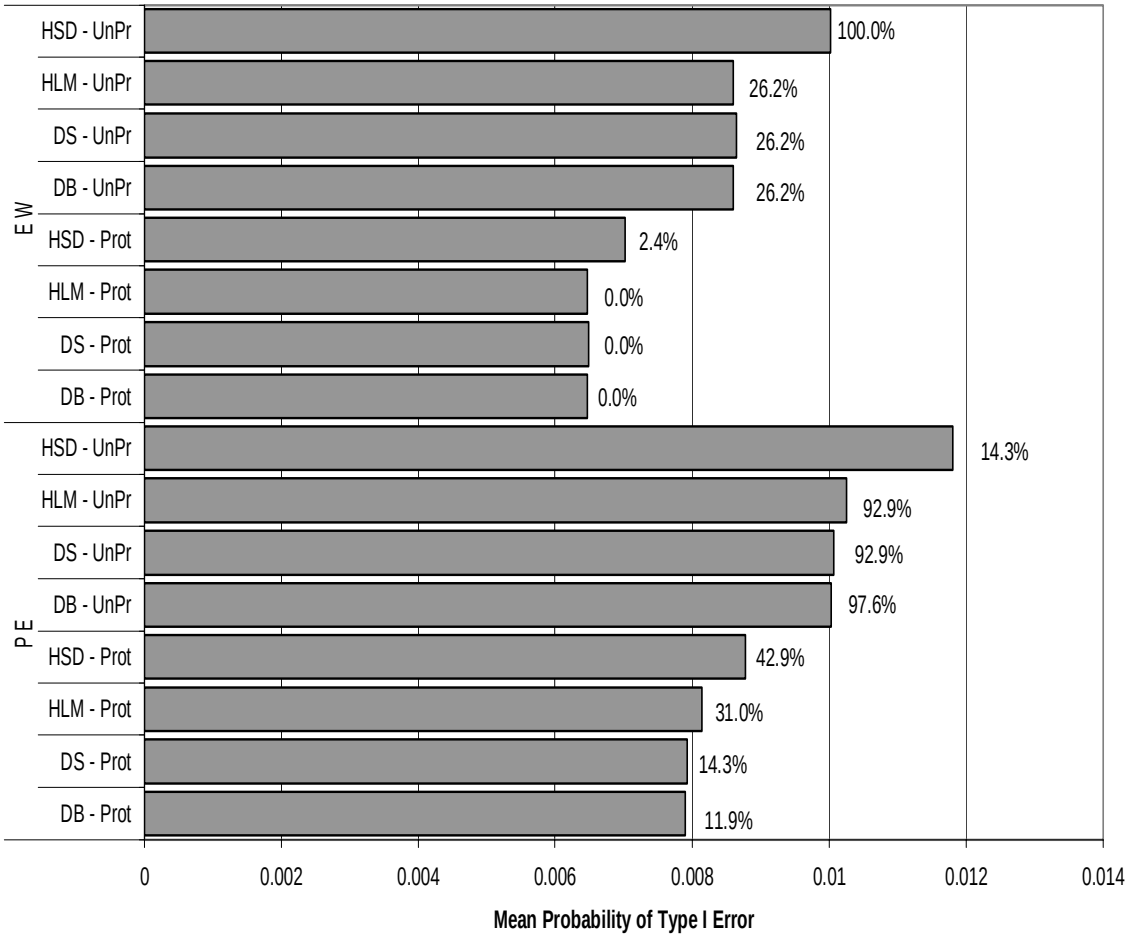


Table 3. Observed Per-Experiment and Experimentwise Type I Error Rates for Selected Multiple Comparison Procedures when Conducted as Protected and Unprotected Tests with Alpha= .01

| | | Protected Test | | | Unprotected Test | | |
|---|---|---|---|---|---|---|---|
| | | Per-Experiment Error (PE) | Experiment-wise Error (EW) | PE - EW Difference | Per-Experiment Error (PE) | Experiment-wise Error (EW) | PE - EW Difference |
| Dunn-Bonferroni | M | .00790 | .00647 | .00143 | .01003 | .00860 | .00143 |
| | M – α | –.00210 | –.00353 | | +.00003 | –.00140 | |
| | SD | .00103 | .00123 | | .00048 | .00059 | |
| | % in α +/-.001 | 11.9 | 0 | | 97.6 | 26.2 | |
| Dunn-Sidak | M | .00793 | .00649 | .00144 | .01007 | .00865 | .00142 |
| | M – α | –.00207 | –.00351 | | +.00007 | –.00135 | |
| | SD | .00103 | .00122 | | .00049 | .00058 | |
| | % in α +/-.001 | 14.3 | 0 | | 92.9 | 26.2 | |
| Holm | M | .00814 | .00647 | .00167 | .01026 | .00860 | .00166 |
| | M – α | –.00186 | –.00353 | | +.00026 | –.00140 | |
| | SD | .00119 | .00123 | | .00054 | .00059 | |
| | % in α +/-.001 | 31.0 | 0 | | 92.9 | 26.2 | |
| HSD | M | 00878 | .00702 | .00176 | .01181 | .01002 | .00179 |
| | M – α | –.00122 | –.00298 | | +.00181 | +.00002 | |
| | SD | .00097 | .00116 | | .00080 | .00043 | |
| | % in α +/-.001 | 42.9 | 2.4 | | 14.3 | 100.0 | |

**Figure 3**
**Accuracy of Type I Error Control with Experimentwise and Per-Experiment Control Conducted as Protected and Unprotected Tests when Nominal Alpha= .01 and % in .01 +/- 0.001**




References

Fisher, R. A. (1935, 1960). *The design of experiments*, 7th ed. London: Oliver & Boyd; New York: Hafner.

Gossett, W. S. (1907-1938) (1943). *Student's collected papers*. (E. S. Pearson & Wishart, J., editors). London: University Press, Biometrika Office.

Hancock, G. R. & Klockars, A. J. (1996). The quest for α: Developments in multiple comparison procedures in the quarter century since Games (1971), *Review of Educational Research, 66*, 269-306.

Hays, W. L. (1988). *Statistics* (4th ed.). New York: Holt, Rinehart, and Winston, Inc.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics, 6*, 65-70.

Kennedy, J. J. & Bush, A. J. (1985). *An introduction to the design and analysis of experiments in behavioral research*. Lanham, MD: University Press of America, Inc.

Kirk, R. E. (1982). *Experimental design: Procedures for the behavioral sciences*. (2nd ed.). Belmont, CA: Brooks Cole.

Klockars, A. J. & Hancock, G. R. (1994). Per-experiment error rates: The hidden costs of several multiple comparison procedures. *Educational and Psychological Measurement, 54* (2), 292-298.

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

Maxwell, S. E. & Delaney, H. D. (1990). *Designing experiments and analyzing data: A model comparison perspective*. Belmont, CA: Wadsworth Publishing Company.

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

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.

Pearson, E. S. & Sekar, C. (1936). The efficiency of statistical tools and a criterion for the rejection of outlying observations. *Biometrika, 28*, 308-320.

Ryan, T. A. (1959). Multiple comparisons in psychological research. *Psychological Bulletin, 56*, 26-47.

Toothaker, L. E. (1993). *Multiple comparison procedures*. Sage University Paper series on Quantitative Applications in the Social Sciences, 07-089. Newbury Park, CA: Sage.

Tukey, J. W. (1953). *The problem of multiple comparisons.* Unpublished manuscript, Princeton University.

Working, H. & Hotelling, H. (1929). Application of the theory of error to the interpretation of trends. *Journal of the American Statistical Association, 35*, 73-85.

Journal of Modern Applied Statistical Methods
November, 2005, Vol. 4, No. 2, 460-472


1538 – 9472/05/$95.00

# Testing For Aptitude-Treatment Interactions In Analysis Of Covariance And Randomized Block Designs Under Assumption Violations

Tim Moses
Educational Testing Service
Princeton, NJ

Alan Klockars
University of Washington

This study compared the robustness of two analysis strategies designed to detect Aptitude-Treatment Interactions to two of their similarly-held assumptions, normality and residual variance homogeneity. The analysis strategies were the test of slope differences in analysis of covariance and the test of the Block-by-Treatment interaction in randomized block analysis of variance. With equal sample sizes in the treatment groups the results showed that residual variance heterogeneity has little effect on either strategy but nonnormality makes the test of slope differences liberal and the test of the Block-by-Treatment interaction conservative. With unequal sample sizes in the treatment groups the often-reported sample size-variance heterogeneity pairing is problematic for both strategies. The findings suggest that the randomized block strategy can be characterized as an overly-conservative alternative to the test of slope differences with respect to robustness.

Key words: Aptitude-treatment interactions, analysis of covariance, randomized block, nonnormality, variance heterogeneity, robustness

## Introduction

One of the important issues in education is identifying when the effect of an instructional strategy depends on some individual difference variable (X) of the student. In their seminal work, Cronbach and Snow (1977) called these effects Aptitude-Treatment Interactions (ATIs). Two analysis approaches for identifying the presence of ATIs differ in terms of how they deal with an originally continuous X.

Tim Moses is an Associate Measurement Statistician at Educational Testing Service. He works primarily on the Advanced Placement Program. Tim completed his PhD in Educational Psychology at the University of Washington. Email him at tmoses@ets.org. Alan Klockars is Professor of Educational Psychology at the University of Washington. His research concerns multiple comparisons and, more recently, methods of conducting ATI research. Email him at klockars@u.washington.edu.

The first is a randomized block analysis of variance approach in which X is stratified into mutually exclusive subsets (Blocks). The second is a regression-based analysis of covariance approach in which the observed continuum of X is used. The question of interest is assessed with a test of the Block-by-Treatment interaction in the randomized block approach and a test of the homogeneity of regression coefficients in the analysis of covariance approach.

The randomized block and the analysis of covariance approaches have been compared in terms of relative power and apparent popularity. When their assumptions are met, both approaches control Type I error to an acceptable level, while the analysis of covariance strategy has superior power (Klockars & Beretvas, 2001; Cronbach & Snow, 1977; Pedhazur, 1997; Aiken & West, 1991). The power advantage is greatest when the randomized block strategy is based on a large number of blocks. In terms of popularity and familiarity for researchers, the randomized block strategy seems to have the advantage (Klockars & Beretvas, 2001; Keselman, Huberty, Lix, Olejnik, Cribbie, Donahue, Kowalchuk,

Lowman, Petoskey, Keselman, & Levin, 1998; Maxwell, O'Callaghan, & Delaney, 1993). The purpose of the current study is to compare the two strategies in terms of a different criterion, their relative robustness to violations of assumptions about the normality and between-group variance homogeneity of the errors.

The two strategies make similar assumptions about the normality and variance homogeneity of the errors, but define error differently. In the randomized block design error is defined as the deviation of the scores from the mean of the Block-Treatment group. This mean reflects the outcome measure (Y) for all individuals in a treatment group who are categorized into the same block based on their X values. The error variance for the randomized block design is called the Subject/Block-by-Treatment Mean Square or S/BT. In analysis of covariance, error is defined as the difference between the Y scores and the predicted value based on the X value of the subject. The predicted value is from the best fitting least squares line for the treatment group. The error variance for analysis of covariance is called the adjusted subject Mean Square or the residual variance.

Research has considered the effects of nonnormality and variance heterogeneity on the robustness of the two strategies, but most of this work has been on the analysis of covariance strategy. None of this work has specifically compared the robustness of the two analysis strategies under the same assumption violations. This research suggests that the two assumption violations have different effects on the robustness of the analysis of covariance and randomized block strategies.

Nonnormality seems to have a stronger impact on the robustness of the analysis of covariance strategy than on the robustness of the randomized block strategy. The analysis of covariance strategy becomes liberal when the error distribution is heavy-tailed and conservative when it is light-tailed (Conover & Iman, 1982; Headrick & Sawilowsky, 2000; Klockars & Moses, 2002). The randomized block strategy is mildly affected by all but the most extreme conditions of nonnormality (Milligan, Wong & Thompson, 1987; Keselman, Carriere, & Lix, 1995).

The effect of variance heterogeneity on robustness depends on whether group sample sizes are equal. With equal sample sizes, variance heterogeneity has a negligible effect on the robustness of the analysis of covariance strategy (Dretzke, Levin & Serlin, 1982; Overton, 2001) and sometimes a negligible (Milligan, Wong & Thompson, 1987) or other times a liberal (Harwell, Rubinstein, Hayes & Olds, 1992) effect on the randomized block strategy. With unequal sample sizes, variance heterogeneity influences the robustness of the two strategies in the same way: when the group with the largest sample size has the smallest error variance (inverse pairing) both strategies are liberal, and when the group with the largest sample size has the largest error variance (direct pairing) both strategies are conservative. The current study considers the variance heterogeneity effect for equal and unequal sample sizes.

Finally, the effect of combined nonnormality and variance heterogeneity is interactive for the analysis of covariance strategy and additive for the randomized block strategy. For the analysis of covariance strategy, the two assumption violations slightly correct for each other (Deshon & Alexander, 1996). For the randomized block strategy, the two assumption violations are not interactive so that robustness depends mostly on the extent of variance heterogeneity (Keselman, et al., 1995; Harwell, et al., 1992).

It is difficult to recommend either analysis of covariance or randomized block as the more robust strategy when the errors are nonnormal and heterogeneous. Comparisons of the two strategies have focused on power when their assumptions are met and their popularity among researchers. The research that has evaluated the impact of the assumption violations on robustness has not directly compared the robustness of the two strategies. The current study was motivated by these concerns. The major questions are 1) for combinations of nonnormality and variance heterogeneity, which strategy is more robust? and 2) how will the relative robustness of these two strategies compare to what is known about their relative power?

## Methodology

A Monte Carlo simulation study was conducted to investigate the relative robustness of the ATI analysis strategies. The null hypothesis of no ATI was true in all conditions. Empirical Type I error rates based on 10,000 iterations were generated for each condition. These empirical Type I error rates were then compared to the nominal Type I error rate of .05. Two treatment groups were used throughout the study. The following conditions were considered.

### Analysis strategies

The standard analysis of covariance test of regression slope heterogeneity (Slopes) and the randomized block Block-by-Treatment Interaction analyses were compared. The randomized block strategy was evaluated using two (RB2) and four (RB4) blocks of X using median and quartile splits of the X variable based on the total sample. While the creation of the X blocks using of the total sample can create slightly unequal sample sizes even though the treatment group sizes are intended to be equal, the use of the total sample was preferred over the excessively liberal strategy of creating the X blocks within each separate treatment group (Myers & Well, 1995).

### Assignment strategies

Two major strategies for assigning subjects to treatment conditions in randomized block and analysis of covariance are random assignment and assignment that utilizes subjects' X scores (Lomax, 2001; Myers & Well, 1995). When subjects are randomly assigned to treatments without regard for X, the randomized block strategy creates X blocks after treatments are administered (post hoc blocking). When subjects are assigned to treatments based on their X score, the randomized block strategy first creates the desired number of blocks in the total sample and then randomly assigns equal numbers of subjects to each of the treatments from each of the blocks. The approach of assigning subjects to treatments based on X and using the analysis of covariance is called systematic assignment (Dalton & Overall, 1977), meaning that subjects are first ranked on X and then assigned to treatments in a systematic pattern (i.e. 12211221…).

The consideration of analysis and assignment strategy resulted in six strategies to be investigated: analysis of covariance with random assignment, analysis of covariance with systematic assignment, RB2 and RB4 with random assignment (post hoc blocking) and RB2 and RB4 with assignment from the blocks.

### Normality

Three shapes were used for X and the errors of Y, including a normal shape (skew=0, kurtosis=0), a skewed and heavy-tailed shape (skew=1, kurtosis=10) and an extremely skewed and heavy-tailed shape (skew=3, kurtosis=50). The shapes were generated with Fleishman's (1978) method (described below).

### Variance Heterogeneity

Between-group variance heterogeneity was created to obtain a specified residual variance ratio of the treatment groups' residual variances based on the groups' deviations from their own regression lines. The variance heterogeneity considered in this study corresponds to how variance heterogeneity occurs in observed datasets (Oswald, Saad, & Sackett, 2000), meaning that groups differed more on their X-Y correlations and Y variances than on their X variances. The three considered residual variance ratios for the groups were 1/1, 3/1 and 15/1. For the conditions of unequal sample size, the residual variances were directly and inversely paired with the treatment group sample sizes.

To assess the correspondence of the considered levels of residual variance heterogeneity from treatment group regression lines to levels of variance heterogeneity from Block-by-Treatment Y means, Tables 1 and 2 give the ratios of the largest-to-smallest variances for the Block-by-Treatment cells of the RB2 and RB4 designs for all levels of assumption violations considered in this study. As analytical methods for deriving Y variances after forming categories on a correlated X variable are valid only for symmetric distributions (Maxwell & Delaney, 1993), the approach taken to produce the ratios in Tables 1 and 2 was simply to generate each distribution

and residual variance heterogeneity combination in a total sample of 100,000 observations and then compute Y variances for the randomized block designs based on random assignment to treatment conditions (note that the variance ratios based on assignment from the X blocks are almost exactly equal).

Data were simulated so that the correlation was either .3 or .7 for one group. For the second group, the correlation was somewhat different from .3 or .7 so that, combined with a different Y variance, this second group's slope was equal the first group's slope while a desired level of variance heterogeneity was obtained.

Sample Size

Forty or eighty subjects per treatment group were used. The conditions of unequal sample size used forty subjects in one group and eighty in the other.

Data Generation Method

The following data generation method was used to create X and Y variables of desired distributions, variances and correlations while allowing for different assignment strategies to the treatment conditions.

1) N values of one standard normal variate, Z, were generated, where N was the total sample size based on two treatment groups that were intended to be of equal sample size.

2) X was created as a transformation of Z using Fleishman's (1978) method for generating nonnormal variables:

$$X = a + bZ + cZ^2 + dZ^3 \qquad (1)$$

The constants (a, b, c, and d) determined the first (mean), second (variance), third (skew) and fourth (kurtosis) moments of X. The values of the constants were derived to obtain the three distributions of interest in this study, where each distribution had a mean and variance of 0 and 1, respectively. The constants and resulting distributions are listed in Table 3.

3) An error variable for Y (E) was generated exactly as X was in steps 1 and 2. E had the same distribution as X.

4) Equal numbers of Xs and Es were randomly assigned to treatment groups 1 and 2. Depending on the particular strategy being studied, this involved either random assignment from the total available dataset (analysis of covariance and randomized block with post hoc blocking), random assignment from blocks of X (randomized block with assignment from the X blocks) or systematic assignment of the ranked X values to treatment groups (analysis of covariance with systematic assignment). The assignment strategies were the same in the unequal sample size conditions as in the equal sample size conditions, but after assignment one treatment group's sample size was reduced by ½, approximating an experimental study with massive loss of subjects from one of the two treatment groups.

5) Y was created as a function of X and E:

$$Y = \sigma_{Yk}[\rho_k X + (1 - \rho_k^2)^{.5} E] \qquad (2),$$

where $\rho_k$ was the desired X-Y correlation and $\sigma_{Yk}$ is the desired standard deviation of Y for treatment group k. The values $\rho_k$ and $\sigma_{Yk}$ were determined for both treatment groups such that the two groups had the desired residual variance ratio and the null hypothesis of no slope differences was true. The values used are summarized in Table 4.

Table 1 Simulated ratios of largest-to-smallest Y variances in the Block-by-Treatment cells of the randomized block designs (XY correlation = .3, N=100,000).

| Distribution of X and E | | ***Residual Variance Ratio*** | | | | | |
|---|---|---|---|---|---|---|---|
| | | 1/1 | | 3/1 | | 15/1 | |
| Skew | Kurtosis | RB2 | RB4 | RB2 | RB4 | RB2 | RB4 |
| 0 | 0 | 1.0/1 | 1.1/1 | 2.9/1 | 3.1/1 | 14.5/1 | 15.4/1 |
| 1 | 10 | 1.0/1 | 1.1/1 | 3.0/1 | 3.2/1 | 14.9/1 | 16.3/1 |
| 3 | 50 | 1.1/1 | 1.3/1 | 3.0/1 | 3.4/1 | 14.3/1 | 15.3/1 |

Table 2 Simulated ratios of largest-to-smallest Y variances in the Block-by-Treatment cells of the randomized block designs (XY correlation = .7, N=100,000).

| Distribution of X and E | | ***Residual Variance Ratio*** | | | | | |
|---|---|---|---|---|---|---|---|
| | | 1/1 | | 3/1 | | 15/1 | |
| Skew | Kurtosis | RB2 | RB4 | RB2 | RB4 | RB2 | RB4 |
| 0 | 0 | 1.0/1 | 1.2/1 | 2.5/1 | 3.2/1 | 11.6/1 | 15.1/1 |
| 1 | 10 | 1.3/1 | 1.9/1 | 2.7/1 | 3.8/1 | 11.7/1 | 16.9/1 |
| 3 | 50 | 1.8/1 | 3.4/1 | 2.8/1 | 4.8/1 | 10.7/1 | 17.5/1 |

Table 3 Fleishman constants used to generate the variables

| Skew | Kurtosis | a | b | c (=-a) | d |
|---|---|---|---|---|---|
| 0 | 0 | 0 | 1 | 0 | 0 |
| 1 | 10 | -.08772 | .56426 | .08772 | .12621 |
| 3 | 50 | -.17038 | -.04789 | .17038 | .26005 |

Table 4 Correlations and standard deviations used to create levels of residual variance heterogeneity.

| Residual Variance Ratio | $\rho_k$ for Group 1 | $\sigma_{Yk}$ for Group 1 | $\rho_k$ for Group 2 | $\sigma_{Yk}$ for Group 2 |
|---|---|---|---|---|
| | | Low X-Y Relationship | | |
| 1/1 | 0.3 | 1 | 0.3 | 1 |
| 1/3 | 0.3 | 1 | 0.171871 | 1.679143 |
| 1/15 | 0.3 | 1 | 0.080933 | 3.706751 |
| | | High X-Y Relationship | | |
| 1/1 | 0.7 | 1 | 0.7 | 1 |
| 1/3 | 0.7 | 1 | 0.492773 | 1.421127 |
| 1/15 | 0.7 | 1 | 0.24535 | 2.853069 |

Programming

The programming for this study was done in SAS, using the CALL RANNOR (SAS Institute Inc., 1999a) routine for creating standard normal deviates and the PROC GLM (SAS Institute Inc., 1999b) function with Type III Sums of Squares for implementing the analysis strategies.

Assessing the Type I Error Rates

To identify the conditions with the strongest influence on Type I error, ANOVAs of the six manipulated variables and their two, three, four, five and six-way interactions were used. These ANOVAs were conducted separately for the equal and unequal sample size conditions. For equal sample sizes, the six independent variables (and their number of levels) were analysis strategy (3), assignment strategy (2), nonnormality (3), residual variance ratio (3), sample size (2) and overall X-Y correlation (2). For unequal sample sizes, the six independent variables (and their number of levels) were analysis strategy (3), assignment strategy (2), nonnormality (3), residual variance ratio (3), sample size-residual variance pairing (direct or inverse, 2) and overall X-Y correlation (2). Due to the stability of the empirical error rates, the two ANOVAs captured 100% of the variation in Type I error. Representative tables that illustrated the most important effects from the ANOVAs are also provided. The Type I error rates in these tables were considered as meaningfully different from the nominal .05 rate based on the criterion of +/- 2 standard errors range (.046-.054). Note that the +/- 2 standard error range is almost identical to Bradley's (1978) conservative range (.045-.055).

## Results

Equal Sample Sizes

Table 5 presents the ten effects with the largest mean squares from the ANOVA of the error rates for equal sample sizes in the treatment groups. These ten effects accounted for 84.6% of the variation in Type I error rates. The two strongest effects were the analysis strategy and the analysis*normality interaction, accounting for 72.3% of the variation in Type I error. The assignment strategy's main effect and interactions with analysis, analysis*normality were also visible, but to a much smaller extent. Residual variance heterogeneity, XY correlation and sample size had small main effects.

Tables 6 and 7 illustrate the results of Type I error effects for equal treatment group sample sizes. These tables present the empirical Type I error rates for three analysis strategies across normality and residual variance heterogeneity ratios for the treatment group sample sizes of 40 and the overall XY correlation of .3. Table 6 shows the results for random assignment to treatment conditions. Table 7 shows the results when X was used to assign subjects to treatment conditions.

Table 5 The Ten Effects with the Largest Mean Squares, Equal Sample Sizes

| Source | Sum of Squares (multiplied by 1,000) | df | Mean Square (multiplied by 1,000) |
|---|---|---|---|
| Analysis | 5.644 | 2 | 2.822 |
| Analysis*Normality | 5.350 | 4 | 1.338 |
| Analysis*Assignment | .456 | 2 | .228 |
| Analysis*N | .342 | 2 | .171 |
| Correlation | .148 | 1 | .148 |
| Assignment | .117 | 1 | .117 |
| N | .115 | 1 | .115 |
| ResVarHet | .204 | 2 | .102 |
| Analysis*Normality*Assignment | .335 | 4 | .084 |
| Correlation*Normality | .143 | 2 | .072 |

Table 6 Type I Error Rates for Treatment Groups of 40, an XY correlation of .3, and Random Assignment to Treatment Conditions.

| Distribution of X and E | | ***Residual Variance Ratio*** | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | 1/1 | | | 3/1 | | | 15/1 | | |
| Skew | Kurtosis | Slopes | RB2 | RB4 | Slopes | RB2 | RB4 | Slopes | RB2 | RB4 |
| 0 | 0 | .047 | .048 | .052 | .046 | .046 | .051 | .051 | .051 | .054 |
| 1 | 10 | .054 | .046 | .051 | .054 | .045* | .051 | .055* | .052 | .056* |
| 3 | 50 | .068* | .044* | .044* | .058* | .042* | .042* | .066* | .036* | .038* |

* Outside the +/- 2 standard error range (.046 to .054).

Table 7 Type I Error Rates for Treatment Groups of 40, an XY correlation of .3, and Assignment to Treatment Conditions Utilizing X.

| Distribution of X and E | | ***Residual Variance Ratio*** | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | 1/1 | | | 3/1 | | | 15/1 | | |
| Skew | Kurtosis | Slopes | RB2 | RB4 | Slopes | RB2 | RB4 | Slopes | RB2 | RB4 |
| 0 | 0 | .050 | .050 | .051 | .052 | .050 | .051 | .053 | .052 | .056* |
| 1 | 10 | .056* | .046 | .043* | .061* | .050 | .045* | .071* | .053 | .051 |
| 3 | 50 | .069* | .041* | .034* | .076* | .040* | .034* | .088* | .039* | .033* |

* Outside the +/- 2 standard error range (.046 to .054).

The most visible effect shown in Tables 6 and 7 is the effect of nonnormality on the analysis strategies. For the analysis of covariance strategy, increased nonnormality made Type I error liberal. For the randomized block strategies, increased nonnormality made Type I error conservative. The effect of nonnormality on the strategies was slightly larger when assignment to treatments used X (Table 7) than when assignment to treatments was random (Table 6). The effect of residual variance heterogeneity was very small when subjects are randomly assigned to treatments (Table 6), though RB4 was significantly liberal in two of the four sample size-correlation conditions where residual variance heterogeneity was most extreme. When subjects were assigned to treatments based on X, residual variance heterogeneity seemed to increase the liberalness of the analysis of covariance test when there was nonnormality. The results shown in Tables 6 and 7 were similar for the higher sample size and XY correlation.

Unequal Sample Sizes

Table 8 presents the ten effects with the largest mean squares from the ANOVA of the error rates for unequal sample sizes in the treatment groups. The mean squares were much larger when sample sizes were unequal, indicating that variations in Type I error are much greater for unequal sample sizes than for equal sample sizes. The ten effects in Table 8 accounted for 98.9% of the variation in Type I error rates. The two strongest effects were the residual variance-sample size pairing (direct or inverse) and this pairing in interaction with the levels of residual variance heterogeneity, 80.5% of the variation in Type I error. Many of the remaining ten effects in Table 8 also involved interactions with the residual variance-sample size pairing and the levels of residual variance heterogeneity. The main effects and interactions with analysis strategy accounted for less than 8% of total variability in Type I error, suggesting small but visible differences in the robustness of the three analysis strategies. The effects of assignment strategy, overall XY

correlation, sample size and normality effects were very small when group sample sizes were unequal.

Tables 9 and 10 illustrate the effects of directly-paired sample sizes and residual variance ratios where the overall XY correlation was .3 and the assignment strategy was either random (Table 9) or based on X (Table 10). With equal residual variances (a residual variance ratio of 1/1), the slope test became liberal, RB2 became conservative and RB4 was not seriously affected. With residual variance heterogeneity, all Type I error rates became extremely conservative. The most conservative strategy was RB4. The RB2 and the analysis of covariance strategies had similar Type I error rates when distributions were normal. The combination of nonnormality and residual variance heterogeneity was visibly interactive for the analysis of covariance strategy, which became slightly less conservative as distributions became more nonnormal. In contrast, the effect of nonnormality was very small for RB2 and RB4. The error rates in Tables 9 and 10 are similar, suggesting that the assignment strategy used makes little difference when sample sizes are unequal.

Tables 11 and 12 illustrate the effects of inversely-paired sample sizes and residual variances. With no residual variance heterogeneity, nonnormality made the analysis of covariance test liberal, RB2 conservative, and had little effect on RB4. As residual variances became different all three analysis strategies became liberal, where the randomized block strategy based on four blocks (RB4) was the most liberal and the analysis of covariance and RB2 strategies had similarly-liberal Type I error rates. The combination of nonnormality and residual variance heterogeneity made all three strategies slightly less liberal than residual variance heterogeneity with normality. The error rates in Tables 11 and 12 are very similar, suggesting that assignment strategy makes little difference when sample sizes are unequal (like the results of direct pairing).

Table 8 The Ten Effects with the Largest Mean Squares, Unequal Sample Sizes

| Source | Sum of Squares (multiplied by 1,000) | df | Mean Square (multiplied by 1,000) |
|---|---|---|---|
| Pairing | 340.380 | 1 | 340.380 |
| Pairing*ResVarHet | 230.011 | 2 | 115.006 |
| ResVarHet | 55.485 | 2 | 27.743 |
| Analysis*Pairing | 23.954 | 2 | 11.977 |
| Analysis | 13.601 | 2 | 6.800 |
| Analysis*Pairing*ResVarHet | 18.513 | 4 | 4.628 |
| Analysis*ResVarHet | 11.645 | 4 | 2.911 |
| Pairing*Normality | .447 | 2 | 2.236 |
| Pairing*Correlation | .622 | 1 | .622 |
| Pairing*ResVarHet*Normality | 2.362 | 4 | .591 |

Table 9 Type I Error Rates for the Direct Pairing of Sample Size (80, 40) and Residual Variance, an XY correlation of .3, and Random Assignment to Treatment Conditions.

| Distribution of X and E | | ***Residual Variance Ratio*** | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | 1/1 | | | 3/1 | | | 15/1 | | |
| Skew | Kurtosis | Slopes | RB2 | RB4 | Slopes | RB2 | RB4 | Slopes | RB2 | RB4 |
| 0 | 0 | .050 | .050 | .050 | .021* | .021* | .012* | .008* | .008* | .003* |
| 1 | 10 | .050 | .049 | .051 | .025* | .022* | .015* | .015* | .006* | .003* |
| 3 | 50 | .060* | .045* | .050 | .040* | .020* | .016* | .026* | .006* | .002* |

* Outside the +/- 2 standard error range (.046 to .054).

Table 10 Type I Error Rates for the Direct Pairing of Sample Size (80, 40) and Residual Variance, an XY correlation of .3, and Assignment to Treatment Conditions Utilizing X.

| Distribution of X and E | | ***Residual Variance Ratio*** | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | 1/1 | | | 3/1 | | | 15/1 | | |
| Skew | Kurtosis | Slopes | RB2 | RB4 | Slopes | RB2 | RB4 | Slopes | RB2 | RB4 |
| 0 | 0 | .046 | .049 | .051 | .023* | .019* | .012* | .009* | .008* | .004* |
| 1 | 10 | .050 | .047 | .051 | .030* | .020* | .013* | .014* | .008* | .003* |
| 3 | 50 | .062* | .045* | .052 | .042* | .022* | .017* | .032* | .006* | .002* |

* Outside the +/- 2 standard error range (.046 to .054).

Table 11 Type I Error Rates for the Inverse Pairing of Sample Size (40, 80) and Residual Variance, an XY correlation of .3, and Random Assignment to Treatment Conditions.

| Distribution of X and E | | ***Residual Variance Ratio*** | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | 1/1 | | | 3/1 | | | 15/1 | | |
| Skew | Kurtosis | Slopes | RB2 | RB4 | Slopes | RB2 | RB4 | Slopes | RB2 | RB4 |
| 0 | 0 | .049 | .053 | .050 | .099* | .097* | .138* | .149* | .149* | .245* |
| 1 | 10 | .049 | .045* | .052 | .097* | .094* | .128* | .143* | .147* | .238* |
| 3 | 50 | .060* | .043* | .050 | .092* | .085* | .114* | .114* | .138* | .210* |

* Outside the +/- 2 standard error range (.046 to .054).

Table 12 Type I Error Rates for the Inverse Pairing of Sample Size (40, 80) and Residual Variance, an XY correlation of .3, and Assignment to Treatment Conditions Utilizing X.

| Distribution of X and E | | ***Residual Variance Ratio*** | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | 1/1 | | | 3/1 | | | 15/1 | | |
| Skew | Kurtosis | Slopes | RB2 | RB4 | Slopes | RB2 | RB4 | Slopes | RB2 | RB4 |
| 0 | 0 | .049 | .048 | .052 | .102* | .099* | .142* | .160* | .152* | .248* |
| 1 | 10 | .054 | .047 | .050 | .097* | .100* | .127* | .147* | .153* | .240* |
| 3 | 50 | .061* | .048 | .052 | .092* | .081* | .111* | .131* | .145* | .215* |

* Outside the +/- 2 standard error range (.046 to .054).

## Conclusion

The purpose of the current study was to compare the robustness of two standard analysis strategies for detecting Aptitude-Treatment Interactions when two of their commonly-held assumptions were violated (nonnormal distributions and heterogeneous variances). The two strategies were the test for slope heterogeneity in analysis of covariance and the test of the Block-by-Treatment Interaction in randomized block analysis of variance. In addition, the strategies were evaluated based on two different assignment strategies, random assignment and assignment that utilized X.

The findings supported and extended the findings of previous studies that considered either the randomized block strategy (Milligan, Wong & Thompson, 1987; Keselman, Carrier & Lix, 1995; Harwell, Rubinstein, Hayes & Olds, 1992) or the analysis of covariance strategy (Conovar & Iman, 1982; Headrick & Sawilowsky, 2000; Klockars & Moses, 2002; Dretzke, Levin & Serlin, 1982; Overton, 2001; Deshon & Alexander, 1996; Conerly & Mansfield, 1988) separately. With equal sample sizes, the effect of nonnormality was much stronger than the effect of residual variance heterogeneity, causing the analysis of covariance strategy to get significantly liberal and the randomized block strategy to get significantly conservative. The effect of nonnormality was stronger when assignment to treatment groups was based on X than when assignment was random. With unequal sample sizes, the effect of residual variance heterogeneity was much stronger than the effect of nonnormality, causing the analysis strategies to get significantly conservative when residual variances were directly paired with sample sizes and liberal when residual variances were inversely paired with sample sizes. For unequal sample sizes the assignment strategy did not matter. Finally, for unequal sample sizes the combination of nonnormality and heterogeneous residual variances was interactive for the analysis of covariance strategy and slightly additive for the randomized block strategy. These findings suggest how the issue of robustness can contribute to several years of discussion on the relative merits of the randomized block and analysis of covariance strategies (Cox, 1957; Feldt, 1958; Cronbach & Snow, 1977; Aiken & West, 1991; Pedhazur, 1997; Lomax, 2001; Myers & Well, 1995; Klockars & Beretvas, 2001).

The magnitude of the effects of assumption violations on the robustness of the analysis strategies for equal sample sizes was somewhat smaller than expected. While heavy-tailed distributions did inflate the Type I error for the slope test, the inflation was rather small (up to about .09) given the extremely nonnormal distributions used. Two factors that kept Type I error from fluctuating too widely for extreme nonnormality were the assignment strategies, which made the treatment groups similar in the X distributions and therefore spread the extreme observations fairly evenly across the groups, and the use of a data generation method that created Y's nonnormality rather indirectly through adding nonnormality to X and E. Consistent with previous studies that used a similar data generation method (Conover & Iman, 1982; Luh & Gou, 2000), nonnormality has to be extreme and fairly unrealistic (Micceri, 1989) in order to see its effects on robustness with this data generation method.

The small effect of variance heterogeneity for the randomized block strategy with two blocks and equal sample sizes was surprising given the many studies that discuss the strong influence variance heterogeneity has on standard tests of means (Lix, Keselman, & Keselman, 1996) and interactions (Harwell, Rubinstein, Hayes, & Olds, 1992). However, many studies of the variance heterogeneity assumption focus much more on unequal sample sizes than on equal sample sizes (e.g. Milligan, Wong & Thompson, 1987; Keselman, Carriere & Lix, 1995), giving the impression that unequal sample sizes almost always accompany variance heterogeneity. For example, Milligan et al's study focuses almost completely on the effect of variance heterogeneity and unequal sample sizes, giving only a very quick mention of finding a negligible effect of heterogeneous variances when sample sizes were equal (p. 469). It is possible that the variance heterogeneity created from given levels of residual variance heterogeneity (Tables 1 and 2) was not large enough to impact the randomized

block strategy with two blocks and equal sample sizes. In contrast to the randomized block strategy with two blocks, the randomized block strategy with four blocks resulted in greater levels of variance heterogeneity and did get liberal even when sample sizes were equal.

The explanations of the effects of the assumption violations on the analysis strategies are fairly well-known. Nonnormality makes treatment group slope estimates differ because of high-leverage observations that are extreme on both X and Y, resulting in inflated numerators of the F ratio. In addition, the standard errors of the slopes are smaller than they should be because the denominators of these standard errors use the sum of squares of X, which gets large as observations get more extreme. As the XY correlation increases, so does nonnormality's liberal effect on the test of slopes. For randomized block's tests of means, nonnormal Y's inflate standard deviations and standard errors, resulting in conservative tests. Nonnormal distributions can also affect mean estimates as well. In general, nonnormality has a stronger influence on sums of squares (standard deviations and standard errors) and sums of products (covariances) than it does on sums of raw data (means).

The effects of heterogeneous variances for equal and unequal sample sizes are also straightforward. The randomized block and analysis of covariance F tests use denominators that pool within-group variability across the groups. When sample sizes are equal, this pooling reasonably weights each group's variance equally. When sample sizes are unequal, the variance of the larger group gets weighted more heavily than that of the smaller group, which can over or underestimate random error and lead to conservative or liberal tests, respectively.

Given the effects of the assumption violations on the standard analysis strategies, many alternative strategies have been proposed. In fact, this study was motivated by a view of the randomized block strategy as an alternative strategy to the analyses of covariance strategy that might be more robust to nonnormal distributions. Other alternatives to the slope test include parametric alternative tests for heterogeneous residual variances (Deshon & Alexander, 1996; Overton, 2001; Dretzke, Levin & Serlin, 1982), ranking strategies for nonnormality (Conover & Iman, 1982; Headrick & Sawilowsky, 2000; Klockars & Moses, 2002), and combinations of strategies designed for addressing combinations of assumption violations (Luh & Guo, 2000, 2002). Given researchers' noted tendency to favor more familiar analysis strategies, the randomized block strategy was a practically-important method to evaluate. The findings of this study show that the randomized block strategy suffers from its own problems with respect to robustness. Given its relatively low power (Klockars & Beretvas, 2001) the randomized block strategy is probably best viewed as an overly conservative alternative to the slope strategy, along the same lines as ranked analysis of covariance. The low power of the randomized block test makes its recommendation difficult, especially given the complaints of low power in interaction studies (Aguinis & Pierce, 1998).

One interesting extension of this study would be to evaluate applications of alternative strategies that can address assumption violations within both the randomized block framework and the analysis of covariance framework. A combination of approaches like trimming/winsorizing observations or trimming test statistics for nonnormality and using a parametric alternative test that does not pool treatment group variances for variance heterogeneity has been shown to be effective for improving the robustness and power of tests of means (Keselman, Wilcox, Othman, Fradette, 2002; Luh & Guo, 1999; Keselman, Othman, Wilcox & Fradette, 2004). Some of these combinations of alternative strategies are applicable to tests of interactions. Along these same lines, some ways to trim observations and test statistics for nonnormality and also to use similar parametric alternative tests for heterogeneous residual variances have been considered for the analysis of covariance slope test (Luh & Guo, 2000, 2002). The relative effectiveness of these combinations of alternative strategies for analysis of covariance and randomized block strategies under the same degrees of assumption violations would be interesting to evaluate.

References

Aguinis, H., & Pierce, C. A. (1998). Heterogeneity of error variance and the assessment of moderating effects of categorical variables: A conceptual review. *Organizational Research Methods, 1,* 296-314.

Aiken, L. S. & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions.* Newbury Park, CA: Sage.

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

Conerly, M. D. & Mansfield, E. R. (1988). An approximate test for comparing heteroscedastic regression models. *Journal of the American Statistical Association, 83*(403), 811-817.

Conover, W. J. & Iman, R. L. (1982). Analysis of covariance using the rank transformation. *Biometrics, 38,* 715-724.

Cox, D. R. (1957). The use of a concomitant variable in selecting an experimental design. *Biometrika, 44,* 150-158.

Cronbach, L. J. & Snow, R. E. (1977). *Aptitude and instructional methods.* Irvington, New York.

Dalton, S., & Overall, J. C. (1977). Nonrandom assignment in ANCOVA: The alternative ranks design. *The Journal of Experimental Education, 46,* 58-62.

DeShon, R. P.; Alexander, R. A. (1996). Alternative procedures for testing regression slope homogeneity when group error variances are unequal. *Psychological Methods, 1*(3), 261-277.

Dretzke, B. J.; Levin, J. R.; Serlin, R. C. (1982). Testing for regression homogeneity under variance heterogeneity. *Psychological Bulletin, 91,* 376-383.

Feldt, L. S. (1958). A comparison of the precision of three experimental designs employing a concomitant variable. P*sychometrika, 23,* 335-353.

Fleishman, A. I. (1978). A method for simulating non-normal distributions. *Psychometrika, 43*(4), 521-532.

Harwell, M. R., Rubinstein, E. N., Hayes, W. S., & Olds, C. C. (1992). Summarizing monte carlo results in methodological research: The one- and two-factor fixed effects ANOVA cases. *Journal of Educational Statistics, 17(4),* 315-339.

Headrick, T. C.; Sawilowsky, S. S. (2000). Properties of the rank transformation in factorial analysis of covariance. *Communications in Statistics- Simulation and Computation, 29,* 1059-1087.

Keselman, H. J., Carriere, K. C., & Lix, L. M. (1995). Robust and powerful nonorthogonal analyses. *Psychometrika, 60(3),* 395-418.

Keselman, H. J., Huberty, C. J., Lix, L. M., Olejnik, S., Cribbie, R. A., Donahue, B., Kowalchuk, R. K., Lowman, L. L., Petoskey, M. D., Keselman, J. C., & Levin, J. R. Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA and ANCOVA analyses. *Review of Educational Research, 68*(3), 350-386.

Keselman, H. J., Othman, A. R., Wilcox, R. R. & Fradette, K. (2004). The new and improved two-sample t test. *Psychological Science, 15*(1), 47-51.

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*(2), 288-309.

Klockars, A. J., & Beretvas, S. N. (2001). Analysis of covariance and randomized block design with heterogeneous slopes. *Journal of Experimental Education, 69*(4), 393-410.

Klockars, A. & Moses, T. (2002). Type I error rates for rank-based tests of homogeneity of slopes. *Journal of Modern Applied Statistical Methods, 1*(2), 452-460.

Lix, L. M., Keselman, J. C., & Keselman, H. J. (1996). Consequences of assumption violations revisited: A quantitative review of alternatives to the one-way analysis of variance F test. *Review of Educational Research, 66*(4), 579-619.

Lomax, R. G. (2001). *Statistical Concepts: A second course for education and the behavioral sciences (2nd edition).* Lawrence Erlbaum Associates: Mahwah, New Jersey.

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

Luh, W., & Guo, J. (2000). Approximate transformation trimmed mean methods to the test of simple linear regression slope equality. *Journal of Applied Statistics, 27*(7), 843-857.

Luh, W., & Guo, J. (2002). Using Johnson's transformation with approximate test statistics for the simple regression slope homogeneity. *The Journal of Experimental Education, 71*(1), 69-81.

Maxwell, S. E., O'Callaghan, M. F., & Delaney, H. D. (1993). Analysis of covariance. In L. K. Edwards (Ed.), *Applied analysis of variance in behavioral science.* New York: Marcel Dekker.

Maxwell, S. E. & Delaney, H. D. (1993). Bivariate median splits and spurious statistical significance. *Psychological Bulletin, 113*(1), 181-190.

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

Milligan, G. W., Wong, D. S., & Thompson, P. A. (1987). Robustness properties of nonnorthogonal analysis of variance. *Psychological Bulletin, 101*(3), 464-470.

Myers, J. L. & Well, A. D. (1995). *Research design and statistical analysis.* Lawrence Erlbaum Associates: Hillsdale, New Jersey.

Oswald, F. L., Saad, S.; Sackett, P. R. (2000). The homogeneity assumption in differential prediction analysis: Does it really matter? *Journal of Applied Psychology, 85*(4), 536-541.

Overton, R. C. (2001). Moderated multiple regression for interactions involving categorical variables: A statistical control for heterogeneous variance across two groups. *Psychological Methods, 6*(3), 218-233.

Pedhazur, E. J. (1997). *Multiple regression in behavioral research.* Fort Worth, Texas.

SAS Institute Inc. (1999a). *SAS language reference: Dictionary, version 8.* Cary, NC: Author.

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

Journal of Modern Applied Statistical Methods
November, 2005, Vol. 4, No. 2, 473- 481

1538 – 9472/05/$95.00

# Quasi-Maximum Likelihood Estimation For Latent Variable Models With Mixed Continuous And Polytomous Data

Jens C. Eickhoff
Department of Biostatistics & Medical Informatics
University of Wisconsin – Madison

Latent variable modeling is a multivariate technique commonly used in the social and behavioral sciences. The models used in such analysis relate all observed variables to latent common factors. In many situations, however, some outcome variables are in polytomous form while other outcomes are measured on a continuous scale. Maximum likelihood estimation for latent variable models with mixed polytomous and continuous outcomes is computationally intensive and may become difficult to implement in many applications. In this article, a computationally practical, yet efficient, Quasi-Maximum Likelihood approach for latent variable models with mixed continuous and polytomous variables is proposed. Asymptotic properties of the estimator are discussed. Simulation studies are conducted to examine the empirical behavior and to compare it with existing methods.

Key words: multivariate analysis, polytomous outcome variables, Quasi-ML estimation.

## Introduction

The problem of analyzing concepts or variables which are not directly observable and can only be measured through related indicators arises frequently in practice. In these situations, latent variable modeling provides a useful statistical technique. Statistical methods for analyzing covariances and other relationships between latent and observed variables were historically originated in psychometrics in the form of factor analysis which has later been extended to the more general structural equation analysis (Bentler, 1995; Bollen, 1989; Jöreskog and Sörbom, 1996). Today, latent variable models are extensively used in the behavioral and social sciences.

Most latent variable models are based on the assumption that the observed variables are continuous with a multivariate normal distribution. However, in many studies where data are obtained based on questionnaires, some or all observed outcome variables are typically in polytomous form. For example, data are frequently collected based on questionnaires with Likert scales (e.g., ``disagree", ``neutral", ``agree") responses. Because of its importance in many applications, there has been much attention in latent variable modeling involving polytomous outcomes and it remains an active area of research.

Bock and Lieberman (1970) considered a maximum likelihood method for factor analysis models with dichotomous outcome variables and only one factor. However, direct maximum likelihood analysis for models involving higher dimensional latent variables becomes computationally impractical because it requires maximization over multiple intractable integrals. This led to the development of multi-stage weighted least square estimation based on limited first and second-order sampling using polychoric and polyserial correlations (Muthén, 1984; Lee & Poon, 1987). Multi-stage weighted least squares (WLS) estimation procedures for structural equation models with polytomous outcome variables have been implemented in popular psychometrical software packages including LISCOMP (Muthén, 1987), EQS (Bentler, 1995), LISREL/PRELIS (Jöreskog &

Jens C. Eickhoff is an Associate Scientist in the Department of Biostatistics & Medical Informatics. He obtained his Ph. D. from Iowa State University. Email him at E-mail: eickhoff@biostat.wisc.edu.

Sörbom, 1996), and Mplus (Muthén & Muthén, 1998). These procedures, however, can experience problems of numerical instability, bias, non-convergence, and non-positive definiteness of weight matrices in situations of small sample sizes but large number of outcome variables (Reboussin & Liang, 1998). Sammel & Ryan (1997) and Shi & Lee (2000) used a Monte Carlo EM algorithm to perform maximum likelihood estimation in latent variables models with mixed discrete and continuous outcome variables. These procedures are computationally intensive as each E-step is approximated by Monte Carlo integration and no closed-form expressions are available in the M-steps. Moreover, many iterations are typically required to achieve convergence.

In this article, a computationally practical, yet efficient, Quasi-ML estimation procedure is proposed for factor analysis and structural equation models with mixed continuous and polytomous outcome variables. Asymptotic properties and standard error estimation are discussed. The Quasi-ML estimation can be easily implemented and does not require intensive computations. Simulation studies indicate that the proposed Quasi-ML estimator is substantially more efficient than traditional multi-stage WLS estimators, especially for models where the number of continuous outcome variables exceeds the number of polytomous outcomes.

This article is organized as follows. In the Methodology section, the general model and motivation for the proposed approach, as well as the Quasi-ML estimation procedure and the computation of asymptotic standard errors are described. The results of a simulation study, where the performance of the proposed Quasi-ML estimation is compared with traditional multi-stage weighted least square estimation techniques, is presented in the Results section. Finally, a brief conclusion is given in the last section.

## Methodology

Consider a multivariate mixed-type variable situation with $p_1$ continuous and $p_2$ polytomous outcome variables and $n$ observations. Let $y_i = (y_{1i}, \cdots, y_{p_1 i})'$ denote the set of continuous outcome variables and $z_i = (z_{1i}, \cdots, z_{p_2 i})'$ denote the set of polytomous outcome variables, each with $c(k)$ categories $(k = 1, \cdots, p_2)$, measured on the $i^{th}$ individual. To motivate the model, assume that the set of continuous and polytomous outcome variables can be explained by a smaller number of $q\ (q < p_1 + p_2)$ unobserved latent variables $f_i = (f_{1i}, \cdots, f_{qi})'$. For ease of notation, a measurement or confirmatory factor analysis model is considered as follows. The notation can be easily extended to utilize the more general structural equation model framework. The standard linear measurement model for the continuous outcome variables for the $i^{th}$ observation can be expressed as

$$y_i = \mu + \mathbf{\Lambda}\, f_i + \varepsilon_i, \quad i = 1, \cdots, n, \qquad (1)$$

where $\varepsilon_i$ is a vector of measurement errors and the parameters $\mu$ and $\mathbf{\Lambda}$ contain some restricted elements. It is assumed that

$$f_i \sim N(\mu_f, \mathbf{\Sigma}_f),$$
$$\varepsilon_i \sim N(0, \mathbf{\Psi}),$$

where the elements of $\mu_f, \mathbf{\Sigma}_f$, and $\mathbf{\Psi}$ are unrestricted, free parameters. Furthermore, it is assumed that, conditional on $f_i$, the elements of $y_i$ are independent, i.e., $\mathbf{\Psi}$ is set to be a diagonal matrix. Likewise, for the polytomous outcome variables, it is assumed that conditional on $f_i$, the elements of $z_i$ are independent and that each $z_{ki}$, $(k = 1, \cdots, p_2)$ relates to the latent variables through a probit response probability function, i.e.,

$$P(z_{ki} \le c_j \mid f_i) = \mathbf{\Phi}(\alpha_{k_j} + \beta_k' f_i), \qquad (2)$$

for category $c_j$, $j = 1, \cdots, c(k) - 1$ and $\alpha_{k_1} < \cdots < \alpha_{k_{c(k)-1}}$. The intercept and slope parameters, $\alpha_{k_j}$ and $\beta_k$, describe the

measurement properties of the $k^{th}$ polytomous outcome variable.

The model described by (1) and (2) contains the factor indeterminacy inherent in this type of latent variable models. That is, the same model can be expressed using transformed parameters and factors. To remove this indeterminacy, the following standard identification form (Wall & Amemiya, 2000) for sub-model (1) is used,

$$y_i = \begin{pmatrix} 0 \\ \mu_y \end{pmatrix} + \begin{pmatrix} I_q \\ \Lambda_y \end{pmatrix} f_i + \varepsilon_i, \quad i = 1, \cdots, n,$$

where $\mu_y$ is a $(p_1 - q)\times 1$ vector and $\Lambda_y$ is a $(p_1 - q)\times q$ matrix with unrestricted parameters. If $q > p_1$, additional measurement parameters in sub-model (2) are restricted. Note that this is an interpretable and meaningful identification parameterization which allows for assessing latent variable characteristics because parameters corresponding to the latent variables, i.e., $\mu_f$ and $\Sigma_f$, remain unrestricted. This is particularly useful in multi-group analysis situations where the main interest lies in the comparison of latent variable characteristics between different sampling groups, e.g., sex, gender, etc.

Quasi-Maximum Likelihood Estimation

Let $\boldsymbol{Y} = (y_1, \cdots, y_n)$ and $\boldsymbol{Z} = (z_1, \cdots, z_n)$ denote the observed data matrices from a random sample of the underlying population. Furthermore, denote the model parameters as,

$$\alpha = (\alpha_{1_1}, \cdots, \alpha_{1_{c(1)-1}}, \cdots, \alpha_{p_{2_1}}, \cdots, \alpha_{p_{2c(p2)-1}})',$$
$$\beta = (\beta_1', \cdots, \beta_{p_2}')',$$

and

$$\theta_y = \left(\mu_y', (\text{vec}\,\Lambda_y)', (\text{vec}\,\Psi)'\right)',$$
$$\theta_z = \left(\alpha', (\text{vec}\,\beta)'\right)',$$
$$\theta_f = \left(\mu_f', (\text{vec}\,\Sigma_f)'\right)'.$$

The log-likelihood function based on the observed data is given by

$$l(\theta_y, \theta_z, \theta_f \mid \text{Y, Z}) = \log p(\text{Y}; \theta_y, \theta_f) + \log p(\text{Z} \mid \text{Y}; \theta_z, \theta_f). \quad (3)$$

Because $\log p(\boldsymbol{Z} \mid \boldsymbol{Y}; \theta_z, \theta_f)$ involves multiple integration which cannot be evaluated in closed form, direct maximization of this log-likelihood function is impractical. Various approaches have been proposed to overcome this computational burden. Sammel & Ryan (1997) and Shi & Lee (2000) proposed utilizing a Monte Carlo EM estimation approach. However, the EM algorithm is known to be slow and may require many iterations to achieve convergence. Moreover, the M-step in these approaches requires iterative procedures which might be time consuming, especially in models involving many polytomous outcomes.

The Quasi-ML approach (Besag, 1975) has become a popular tool in situations where the true likelihood function is computationally intractable but can be approximated by a function that is easier is evaluate. Quasi-ML methods may not always yield efficient estimators but they are usually consistent as long as the first derivatives of the quasi likelihood function has mean 0 at the true parameter values (Le Cessie & Houwelingen, 1994). In the following, a Quasi-ML approach is proposed where the second term of the right hand side of the log-likelihood function in (3) is approximated by a function which is computationally easy to evaluate. Specifically, the Quasi- log-likelihood for the $i^{th}$ observation is expressed as

$$l_i^p = \log p(y_i; \theta_y, \theta_f) + \sum_{k=1}^{p_2} p(z_{ki} \mid y_i; \theta_z, \theta_f),$$

where $p(y_i; \theta_y, \theta_f)$ is a multivariate normal density function with mean

$$\mu(\theta_y, \theta_f) = (\mu_f', (\mu_y + \Lambda_y \mu_f)')'$$

and covariance matrix

$$\boldsymbol{\Sigma}(\theta_y, \theta_f) = \begin{pmatrix} I_q \\ \boldsymbol{\Lambda}_y \end{pmatrix} \boldsymbol{\Sigma}_f \begin{pmatrix} I_q & \boldsymbol{\Lambda}_y \end{pmatrix} + \boldsymbol{\Psi}.$$

Standard evaluation of the conditional distribution, $z_{ki} \mid y_i$, leads to

$$P(z_{ki} \le c_j \mid y_i; \theta_y, \theta_f) = \boldsymbol{\Phi}\left( \frac{\alpha_{k_j} + \beta_k' \mu_{f_i \mid y_i}}{\sqrt{1 + \beta_k' \boldsymbol{\Sigma}_{f_i \mid y_i} \beta_k}} \right),$$

where $1 \le k \le c(k) - 1$ and

$$\mu_{f_i \mid y_i} = \mu_f + \boldsymbol{\Sigma}_f \begin{pmatrix} I_q & \boldsymbol{\Lambda}_y \end{pmatrix} \left( \begin{pmatrix} I_q \\ \boldsymbol{\Lambda}_y \end{pmatrix} \boldsymbol{\Sigma}_f \begin{pmatrix} I_q \\ \boldsymbol{\Lambda}_y \end{pmatrix}' \right)^{-1} \left( y_i - \begin{pmatrix} \mu_f \\ \mu_y - \boldsymbol{\Lambda} \mu_f \end{pmatrix} \right),$$

$$\boldsymbol{\Sigma}_{f_i \mid y_i} = \boldsymbol{\Sigma}_f - \boldsymbol{\Sigma}_f \begin{pmatrix} I_q & \boldsymbol{\Lambda}_y \end{pmatrix} \left( \begin{pmatrix} I_q \\ \boldsymbol{\Lambda}_y \end{pmatrix} \boldsymbol{\Sigma}_f \begin{pmatrix} I_q \\ \boldsymbol{\Lambda}_y \end{pmatrix}' \right)^{-1} \begin{pmatrix} I_q \\ \boldsymbol{\Lambda}_y \end{pmatrix} \boldsymbol{\Sigma}_f.$$

The total Quasi log-likelihood is then the sum of the $l_i^p$'s, i.e.,

$$\begin{aligned} l^p &= \sum_{i=1}^{n} l_i^p = \sum_{i=1}^{n} \log p(y_i; \theta_y, \theta_f) + \sum_{i=1}^{n} \sum_{k=1}^{p} p(z_{ki} \mid y_i; \theta_z, \theta_f) \\ &\propto -\frac{n}{2} \left( \log \left| \Sigma(\theta_y, \theta_f) \right| + \frac{n-1}{n} tr(\mathrm{S}_y \Sigma^{-1}(\theta_y, \theta_f)) + \right. \\ &\quad \left. \left( \bar{y} - \mu(\theta_y, \theta_f) \right)' \Sigma^{-1}(\theta_y, \theta_f) \left( \bar{y} - \mu(\theta_y, \theta_f) \right) \right) \\ &\quad + \sum_{i=1}^{n} \sum_{k=1}^{p} p(z_{ki} \mid y_i; \theta_z, \theta_f), \end{aligned} \tag{4}$$

where $\bar{y}$ is the sample mean, and $\boldsymbol{S}_y$ is the empirical covariance matrix of $y_i = (y_{1i}, \cdots, y_{p_i i})'$. Note that for a model with several continuous outcomes but only one polytomous outcome variable, the Quasi-log-likelihood function (4) is identical with the log-likelihood function (3).

The Quasi-ML estimator $(\hat{\theta}_y, \hat{\theta}_z, \hat{\theta}_f)$ is obtained by solving

$$\begin{aligned} &S(\theta_y, \theta_z, \theta_f) \\ &= \sum_{i=1}^{n} s_i(\theta_y, \theta_z, \theta_f) = \sum_{i=1}^{n} \frac{\partial l_i^p(\theta_y, \theta_z, \theta_f)}{\partial(\theta_y, \theta_z, \theta_f)} = 0. \end{aligned} \tag{5}$$

Explicit solutions for solving (5) are not available and therefore an iterative procedure is required. Because the number of parameters in (4) is usually relatively large, a derivative free optimization procedure as the Nelder-Mead simplex algorithm may not be computationally efficient. On the other hand, using an efficient optimization procedure such as the Newton-Raphson algorithm requires evaluation the first partial derivatives and the Hessian matrix which might be, due to the complexity of the objective function in (4), a tedious task. A good compromise is using a quasi Newton-Raphson algorithm with numerical derivatives which is easy to implement and numerically stable.

Standard Errors

For the computation of confidence intervals for the Quasi-ML parameter estimates, standard error estimates are required. A sandwich estimator can be used to estimate standard errors of Quasi-ML parameter estimates. It follows from the delta theorem that, under mild regularity conditions (see, e.g., Stuart and Ord, 1991), the distribution of $\sqrt{n}(\hat{\theta}_y - \theta_y, \hat{\theta}_z - \theta_z, \hat{\theta}_f - \theta_f)'$ converges to a $N(0, \boldsymbol{\Delta})$ distribution with

$$\boldsymbol{\Delta} = n \boldsymbol{I}^{-1} \boldsymbol{D} \boldsymbol{I}^{-1},$$

where

$$\begin{aligned} \boldsymbol{D} &= \mathrm{cov}\left( S(\theta_y, \theta_z, \theta_f) \right), \\ \boldsymbol{I} &= E\left( S(\theta_y, \theta_z, \theta_f) \right). \end{aligned}$$

Estimates of $\boldsymbol{D}$ and $\boldsymbol{I}$ can be obtained by computing

$$\hat{\boldsymbol{D}} = \sum_{i=1}^{n} s_i(\hat{\theta}_y, \hat{\theta}_z, \hat{\theta}_f) \left( s_i(\hat{\theta}_y, \hat{\theta}_z, \hat{\theta}_f) \right)' \tag{6}$$

and

$$\hat{\boldsymbol{I}} = -\sum_{i=1}^{n} \frac{\partial\, s_i(\hat{\theta}_y, \hat{\theta}_z, \hat{\theta}_f)}{\partial\,(\theta_y, \theta_z, \theta_f)'}. \tag{7}$$

Expressions (6) and (7) can be obtained using the numerical first and second order derivatives in the last iteration step of the quasi Newton-Raphson algorithm used to solve (5).

Starting Values

As the quasi Newton-Raphson algorithm used to solve (5) is an iterative procedure, starting values for the model parameters are required. One way to obtain starting values is to treat the sub-models (1) and (2) separately. Specifically, starting values for the parameters corresponding to sub-model (1) can be computed using standard estimation procedures for fitting latent variable models with continuous outcomes (Bollen, 1989). These estimates can be used to estimate factor scores, i.e.

$$\tilde{f}_i = \left(\begin{pmatrix} I_q \\ \tilde{\boldsymbol{\Lambda}}_y \end{pmatrix}' \tilde{\boldsymbol{\Psi}}^{-1} \begin{pmatrix} I_q \\ \tilde{\boldsymbol{\Lambda}}_y \end{pmatrix}\right)^{-1} \begin{pmatrix} I_q \\ \tilde{\boldsymbol{\Lambda}}_y \end{pmatrix}' \tilde{\boldsymbol{\Psi}}^{-1} \left(y_i - \begin{pmatrix} 0 \\ \tilde{\mu}_y \end{pmatrix}\right),$$

where $\tilde{\boldsymbol{\Lambda}}_y, \tilde{\boldsymbol{\Psi}}$, and $\tilde{\mu}_y$ are parameter estimates obtained using standard estimation procedures for latent variables models with continuous outcomes. The latent variable $f_i$ of sub-model (2) can then be replaced by the factor scores $\tilde{f}_i$ and standard probit regression can be performed to obtain starting values for $\theta_z$.

## Results

The purpose of this simulation study is to compare the performance of the proposed Quasi-ML estimation approach with the traditional multi-stage WLS estimation approach which is currently considered the gold standard of fitting mixed latent variable models with continuous and polytomous outcomes. In the following, a confirmatory factor analysis model models with three continuous outcome variables and various numbers of polytomous outcome variables are considered. It is assumed that each polytomous outcome variable has three categories. Sub-model (1) is given by

$$\begin{pmatrix} y_{1i} \\ y_{2i} \\ y_{3i} \end{pmatrix} = \begin{pmatrix} 0 \\ \mu_{y_1} \\ \mu_{y_2} \end{pmatrix} + \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ \lambda_1 & \lambda_2 \end{pmatrix} \begin{pmatrix} f_{1i} \\ f_{2i} \end{pmatrix} + \begin{pmatrix} \varepsilon_{1i} \\ \varepsilon_{2i} \\ \varepsilon_{3i} \end{pmatrix},$$

where

$$\begin{pmatrix} f_{1i} \\ f_{2i} \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_{f_1} \\ \mu_{f_2} \end{pmatrix}, \begin{pmatrix} \sigma^2_{f_1} & \sigma_{f_1,f_2} \\ \sigma_{f_1,f_2} & \sigma^2_{f_2} \end{pmatrix}\right)$$

$i = 1, \cdots n$, and $\varepsilon_{ki}$, $k = 1,2,3$, are iid with $N(0, \psi^2)$ distribution. The parameters $\mu_{y_2}, \mu_{y_2}, \lambda_1, \lambda_2, \sigma^2_{f_1}, \sigma_{f_1,f_2}, \sigma^2_{f_2}$, and $\psi^2$ are unrestricted parameters with the true values $\mu_{y_2} = \mu_{y_3} = 1$, $\lambda_1 = \lambda_2 = 0.8$, $\sigma^2_{f_1} = \sigma^2_{f_2} = 1$, $\sigma_{f_1,f_2} = 0.5$, and $\psi^2 = 0.36$.

Sub-model (2), which corresponds to the polytomous outcome variables, each with three categories, is given by,

$$P(z_{ki} = c_j \mid f_1, f_2) = \begin{cases} \Phi(\alpha_{k_1} + \beta_{k1} f_1 + \beta_{k2} f_2), & j = 1 \\ \Phi(\alpha_{k_2} + \beta_{k1} f_1 + \beta_{k2} f_2) - \Phi(\alpha_{k_1} + \beta_{k1} f_1 + \beta_{k2} f_2), & j = 2 \end{cases},$$

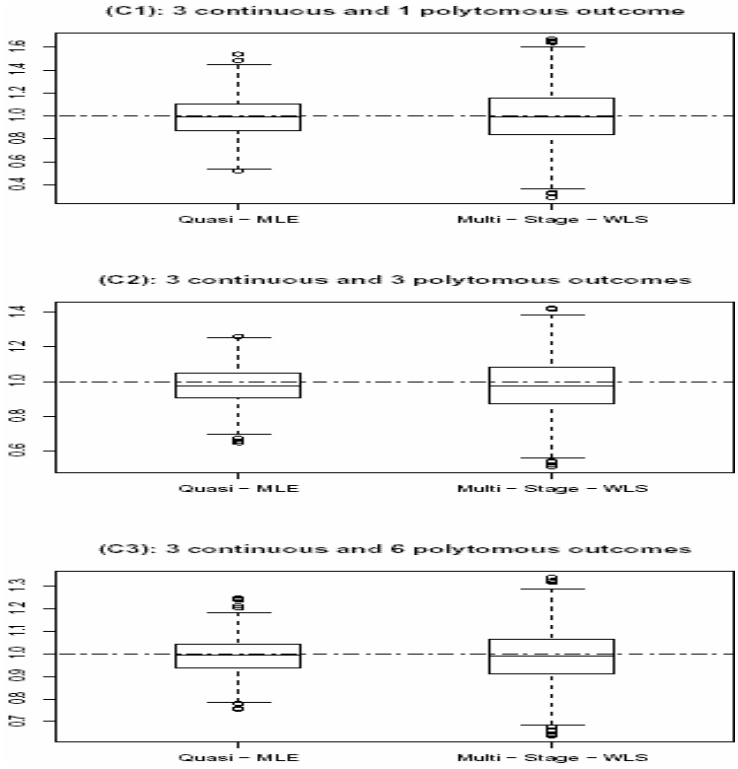
where $\alpha_{k_1}, \alpha_{k_2}, \beta_{k1}$, and $\beta_{k2}$ are unrestricted parameters with true values $\alpha_{k_1} = 0.8$, $\alpha_{k_2} = 1.6$, $\beta_{k1} = 0.6$, and $\beta_{k2} = -0.6$. To facilitate generalization of the simulation results, the following three conditions on the number of polytomous outcome variables in the confirmatory factor models are considered:

(C1): Number of polytomous outcomes: 1

(C2): Number of polytomous outcomes: 3

(C3): Number of polytomous outcomes: 6

Figure 1: Boxpots for Quasi-ML and Multi-Stage WLS Estimators of $\sigma^2_{f_2}$ under Experimental Conditions (C1) – (C3) ($n = 500$)




Note that under experimental condition (C1), the Quasi-ML estimates are equivalent to the ML estimates. In order to compare the Quasi-ML estimation approach with the multi-stage WLS estimation approach, the model part corresponding to the polytomous outcome variables is first re-parameterized to the threshold model. This can be achieved by standardizing the intercept parameters $\alpha_{k_1}, \alpha_{k_2}$ to $\alpha^*_{k_1} = \alpha_{k_1} / \sqrt{1 - \beta' \boldsymbol{\Sigma}_f \beta} = 1$, $\alpha^*_{k_2} = \alpha_{k_2} / \sqrt{1 - \beta' \boldsymbol{\Sigma}_f \beta} = 2$, and the slope parameters $\beta_{k1}, \beta_{k2}$ to $\beta^*_{k1} = \beta_{k1} / \sqrt{1 - \beta' \boldsymbol{\Sigma}_f \beta} = 0.75$ and $\beta^*_{k2} = \beta_{k2} / \sqrt{1 - \beta' \boldsymbol{\Sigma}_f \beta} = -0.75$, respectively.

The computation of the multi-stage WLS procedure was performed by using LISREL 8 and PRELIS 2. The Quasi-ML estimates were computed using R version 1.8.1.

The sample sizes considered were $n = 100$, $n = 500$, and $n = 1{,}000$. For each $n$ and experimental condition (C1), (C2), and (C3), 1,000 simulations on samples were generated. The starting values for the Quasi-ML approach were computed as described in the previous section. Non-convergence was experienced in some cases for the multi-stage WLS approach when $n = 100$, especially for the model with 3 continuous and 6 polytomous outcomes (C3). For $n = 500$, the multi-stage WLS estimation procedure became numerically more stable. There were no convergence difficulties experienced for the Quasi-ML estimation for all sample sizes.

Figure 1 presents boxplots for the two estimators of the variance parameter $\sigma^2_{f_2}$ when $n = 500$, depicting the empirical distribution around the true parameter value $\sigma^2_{f_2} = 1.0$ under

Table 1: Empirical Bias and Root Mean Squared Error for Quasi-ML and Multi-Stage WLS Estimators for $\sigma^2_{f_2}$ under Experimental Conditions (C1) – (C3)

| Experimental Condition | *n* | | Quasi-MLE | Multi-Stage WLS |
|---|---|---|---|---|
| (C1) | 100 | Bias | 0.044 | 0.054 |
| | | RMSE | 0.142 | 0.220 |
| | 500 | Bias | 0.016 | 0.015 |
| | | RMSE | 0.090 | 0.156 |
| | 1,000 | Bias | 0.010 | 0.008 |
| | | RMSE | 0.052 | 0.120 |
| (C2) | 100 | Bias | -0.010 | -0.012 |
| | | RMSE | 0.166 | 0.238 |
| | 500 | Bias | 0.026 | 0.023 |
| | | RMSE | 0.110 | 0.165 |
| | 1,000 | Bias | -0.009 | 0.011 |
| | | RMSE | 0.079 | 0.118 |
| (C3) | 100 | Bias | -0.081 | 0.022 |
| | | RMSE | 0.199 | 0.244 |
| | 500 | Bias | 0.009 | -0.007 |
| | | RMSE | 0.131 | 0.155 |
| | 1,000 | Bias | 0.003 | -0.001 |
| | | RMSE | 0.102 | 0.129 |

experimental conditions (C1) – (C3). The general pattern given in Figure 1 can also been seen in boxplots for the other parameters and sample sizes. Table 1 gives the empirical bias and root mean squared error (RMSE) of the two estimators for the latent variable covariance parameters $\sigma^2_{f_1}$, $\sigma_{f_2,f_2}$, and $\sigma^2_{f_2}$. The cases where the multi-stage WLS estimator didn't converge were excluded when computing the empirical bias and RMSE.

The results indicate that the Quasi-ML estimator and the multi-stage WLS estimator are both unbiased for all coefficients and sample sizes. Under experimental conditions (C1) and (C2), the Quasi-ML estimate exhibit considerable less variability than the multi-stage WLS estimates. As the number of polytomous outcome variables increases this difference in RMSE between the two estimators becomes smaller. However, even under experimental condition (C3) (3 continuous and 6 polytomous outcomes), the Quasi-ML estimates still exhibit slightly less variability than the multi-stage WLS estimates.

Table 2 presents the empirical coverage probabilities of the nominal 95% confidence intervals for the Quasi-ML estimates of the latent variable covariance parameters $\sigma^2_{f_1}$, $\sigma_{f_2,f_2}$, and $\sigma^2_{f_2}$. The intervals were obtained by taking an estimate $\pm 1.96$ times the corresponding estimated standard error. For all sample sizes, the constructed intervals give an empirical coverage close to the nominal level. Similar results were obtained for the other model parameters. Overall, the results indicate that the Quasi-ML standard errors can be used for valid statistical inference on the model parameters.

## Conclusion

Multivariate polytomous data are common in psychosocial research. Consequently, there has been recently an increased interest in latent

Table 2: Empirical Coverage Probabilities for Quasi-ML estimates of Nominal 95% Confidence Intervals for Latent Variable Covariance Parameters

| $n$ | $\sigma^2_{f_1}$ | $\sigma_{f_2,f_2}$ | $\sigma^2_{f_2}$ |
|---|---|---|---|
| 100 | 91.2% | 90.1% | 90.9% |
| 500 | 92.8% | 91.3% | 92.6% |
| 1,000 | 94.0% | 92.9% | 93.9% |

variable modeling involving polytomous outcome variables.

The parameter estimation of these types of models is computationally challenging. Traditional estimation techniques include multi-stage WLS procedures. However, it has been demonstrated that multi-stage WLS procedures can experience serious numerical problems, especially in situations of low prevalence, small sample sizes, or when fitting models with a large number of outcome variables.

Maximum likelihood estimation procedures have been proposed utilizing various types of EM algorithms (Sammel & Ryan, 1997; Shi & Lee, 2000). These procedures are numerically stable, yet computationally very intensive. In this article, a Quasi-ML method is proposed for parameter estimation of latent variable models with mixed continuous and polytomous variables. The procedure is computationally practical and can be easily implemented into standard statistical software (e.g., R, Splus, etc).

Simulation studies indicate that the proposed Quasi-ML estimator tends to be more efficient than traditional multi-stage WLS estimator, especially for models where the number of polytomous outcome variables is smaller than the number of continuous outcome variables. The Quasi-ML estimation of standard errors showed no substantial bias which warrants the performance of valid statistical inference. In summary, the proposed Quasi-ML estimation procedure appears to be efficient, computationally feasible, and a practical approach for latent variable models involving both continuous and polytomous outcomes.

References

Bentler, P. M. (1995). *EQS: Structural Equation Program Manual*. Los Angeles: BMDP Statistical Software.

Besag, J. (1975). The statistical analysis of non-lattice data. *Statistician, 24*, 179-195.

Bock, R. D. & Lieberman, M. (1970). Fitting a response model for n dichotomously scored items. *Psychometrika, 35*, 179-197.

Bollen, K. A. (1989). *Structural equations with latent variables.* New York: Wiley.

Jöreskog, K. & Sörbom, D. (1996). *LISREL 8: User's Reference Guide*. Chicago: Scientific Software International.

Le Cessie, S. & *Van Houwelingen, J. C. (1994). Logistic regression for binary correlated data. Applied Statistics, 43*, 95-108.

Lee, S. Y. & Poon, W. Y. (1987). Two-step estimation of multivariate polychoric correlations. *Communications in Statistics: Theory and Methods, 16*, 307-320.

Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical and continuous latent variable indicators. *Psychometrika, 49*, 115-132.

Muthén, B. (1987) *LISCOMP, Analysis of linear structural equations with a comprehensive measurement model. Theoretical integration and uses's guide.* Mooresville, IN: Scientific Software.

Muthén, B. & Muthén, L. (1998). *Mplus User's Guide*. Los Angeles. CA: Muthen & Muthen.

Reboussin, B. A. & Liang, K. Y. (1998). An estimating equations approach for the LISCOMP model. *Psychometrika, 63*, 165-182.

Sammel, M. D. & Ryan, L. M. (1997). Latent variable models with mixed discrete and continuous outcomes. *Journal of the Royal Statistical Society B, 59*, 667-678.

Shi, J. Q. & Lee, S. Y. (2000). Latent variable models with mixed continuous and polytomous data. *Journal of the Royal Statistical Society B, 62*, 77-87.

Stuart, A. & Ord, J.K. (1991). *Kendall's advanced theory of statistics*. London: Arnold.

Wall, M. M. & Amemiya, Y. (2000). Estimation for polynomial structural equation models. *Journal of the American Statistical Association, 95*, 929-940.

Journal of Modern Applied Statistical Methods
November, 2005, Vol. 4, No. 2, 482-499

1538 – 9472/05/$95.00

# A Bayesian Subset Analysis Of Sensory Evaluation Data

Balgobin Nandram
Department of Mathematical Sciences
Worcester Polytechnic Institute

In social sciences it is easy to carry out sensory experiments using say a *J*-point hedonic scale. One major problem with the *J*-point hedonic scale is that a conversion from the category scales to numeric scores might not be sensible because the panelists generally view increments on the hedonic scale as psychologically unequal. In the current problem several products are rated by a set of panelists on the *J*-point hedonic scale. One objective is to select the best subset of products and to assess the quality of the products by estimating the mean and standard deviation response for the selected products. A priori information about which subset is the best is incorporated, and a stochastic ordering is modified to select the best subset of the products. The method introduced in this article is sampling based, and it uses Monte Carlo integration with rejection sampling. The methodology is applied to select the best set of entrees in a military ration, and then to estimate the probability of at least a neutral response for the judged best entrees. A comparison is made with the method, which converts the category scales to numeric scores.

Key words: Bayes factor; composition method; stochastic ordering; rejection sampling.

## Introduction

Consider the problem of selecting the best subset of a number of multinomial populations with ordinal categories. This can be accomplished by first converting the nominal data to numeric scores, and then a standard multiple comparison procedure can be performed on these scores. However, this procedure can go badly wrong when the conversion is made. It is, therefore, the purpose of this article to describe a straightforward method based on a stochastic ordering of the multinomial populations for selecting the best subset of populations and then to estimate parameters used to assess the quality of the best subset without conversion of the nominal data. A Bayesian approach is preferred because it is natural to incorporate a priori information about which subset is the best.

In sensory evaluation of food acceptability, judges are asked to rate several products on the 9-point scale with qualitative responses ranging from “dislike extremely” to “neither like nor dislike” to “like extremely” on an ordinal scale. Usually in the analysis these nominal values are converted to scores ranging from 1 to 9 where an attempt is made to associate “dislike extremely” with 1, “neither like nor dislike” with 5, “like extremely” with 9, and intermediate nominal values have graduated meanings. The use of scores has several disadvantages, which weaken the interpretation that can be placed on the analysis of sensory evaluation data.

First, the intervals between categories are psychologically unequal. Second, judges tend to avoid the use of extreme categories by grouping judgments into the center of the scale, and sometimes avoiding even “neither like nor dislike” response. Third, scale values have no numerical relationship. Thus, it is difficult to make conclusions concerning ratios of acceptability of the food products when

Balgobin Nandram is a Professor of Statistics, and a fellow of the American Statistical Association. His research interests are in survey methodology, Bayesian statistics, categorical data analysis, computational statistics and simulation, health, industrial and environmental statistics, and statistical education. Email him at balnan@WPI.EDU.

qualitative responses are converted to quantitative responses.

Newel (1982) applied the method of McCullagh (1980) to analyze sensory data and was able to overcome some of the advantages in using scores. This method for ordinal data treats the response categories as contiguous intervals on a continuous scale with unknown cutpoints $\theta_1, \ldots, \theta_{J-1}$, where for the $J$-point scale $J = 9$. Inherent in these models is the stochastic ordering with the use of scores unnecessary. Let $\pi_{ij}$ denote the probability of the $j^{th}$ response in the $i^{th}$ population, and $\gamma_{ij} = \sum_{s=1}^{j} p_{is}$ be the cumulative probability of the $i^{th}$ population. Then Newel (1982) entertained a model of the form

$$\log\{\gamma_{ij}/(1-\gamma_{ij})\} = (\theta_j - \beta_i)/\tau_i,\ i = 1, \ldots, I,$$
$$j = 1, \ldots, J-1,$$

where $\beta_i$ and $\tau_i$ are relative measures of location and spread respectively of the $i^{th}$ population. This model incorporates the location of the ratings and the consistency of the panelists' responses directly.

Such a model is usually fitted using nonlinear iteratively reweighted least squares; see, for example, Green (1985). While this is an attractive model, besides the cell probabilities, it introduces $2I + J$ new parameters. Moreover, while one can choose the best population as the one with the largest $\beta_i$, and perhaps the smallest $\tau_i$, this modeling does not address the problem of selecting the best population directly, and in fact, it is difficult to assess the uncertainty in selecting the best population. Also as the analysis relies heavily on asymptotic theory, with sparse data this approach will provide poor estimates for the cutpoints $\theta_j$, and hence the other parameters. A more appropriate method is associated with ranking and selection.

Recent Bayesian work on selection and ranking includes the approach of Morris and Christiansen (1996). They used a simple two-level Bayes empirical Bayes model to select the best mean. They generated samples from the product normal posterior distribution of the means, and obtained posterior probabilities that each of the means is the largest. Goldstein and Spiegelhalter (1996) described statistical issues in ranking institutions in the areas of health and education based on outcome data by using certain performance indicators. They obtained interval estimates of the ranks of these indicators for the different institutions, using both Bayesian and non-Bayesian methods. Similar to Morris and Christiansen (1996), Goldstein and Spiegelhalter (1996) did not incorporate uncertainty directly about the ranks of the performance indicators. Moreover, these authors did not consider the ranking of several multinomial populations nor did they consider sensory evaluation data. However, the sampling-based approach of these authors is closest in spirit to the work in this article.

In fact, Nandram (1997) obtained the best multinomial population (not best subset) among a set of populations, converting the nominal data on the hedonic scale to numeric scores. A number of independent nonidentical multinomial populations with the same ordinal categories are considered. This approach is different from that in the ranking and selection literature because it incorporates the prior belief about which population is the best by assigning a nonzero probability to the event that any population could be the best population (Nandram, 1997). The simple tree order (see Robertson, Wright and Dykstra, 1988) is used to obtain the most probable population under a variation of the stochastic ordering. Consider two discrete random variables, $P$ and $Q$, which take the same values $a_j$ (increasing in $j$) with probabilities $p_j$ and $q_j$ respectively, $j = 1, \ldots, J - 1$, where

$$\sum_{j=1}^{J} p_i = \sum_{j=1}^{J} q_i = 1.$$

then

$$P \overset{st}{\geq} Q$$

if, and only if,

$$\sum_{i=1}^{s} p_i \le \sum_{i=1}^{s} q_i, \quad s = 1, \ldots, J-1. \qquad (1)$$

This is the situation for two multinomial populations which are stochastically ordered (*P* stochastically greater than *Q*) with the same ordered categories; see, for example, Sampson and Whitaker (1989). This stochastic ordering is modified to obtain a criterion which will be used to select the best population or best subset of populations without using the values $a_j$ on the ordinal scale.

The Bayesian analysis is pertinent as there is useful information about which is the best product. In the non-Bayesian approach, it is difficult to express uncertainty about which population is the best. Moreover, as the non-Bayesian methods do not express uncertainty about the best population, estimation after selection becomes a delicate and tricky issue. In the Bayesian method the parameters can be estimated in a straightforward manner by mixing with appropriate weights (posterior probabilities); see Nandram (1997).

The objective is to select the best population (or subset) among a number of multinomial populations, whose cell counts arise from sensory evaluation, and to show how to estimate the parameters of the selected population. The method is sampling based, and it uses Monte Carlo integration which is accommodated by rejection sampling. A methodology is described, and it is shown how to compute efficiently the relevant quantities. Next, the sensory data obtained from the Natick food experiment is described and the methodology is applied to select the best entree. Finally, there are conclusions.

## Methodology

The objective is to develop a method to judge the best multinomial population or the best subset of multinomial populations without converting the ordinal categories to numeric scores by modifying the stochastic ordering. Estimation is performed to make inference about the quality of product. In general, it is assumed that there are *I* multinomial populations, and the best subset of size $\ell < I$ subsets is to be selected. There are $T = I!/\ell!(I-\ell)!$ distinct subsets of size $\ell$ which are denoted by $I_t$, $t = 1 \ldots, T$. For example, with $I = 3$, $\ell = 2$, the set of all products is {1, 2, 3}, $T = 3$, and the subsets are $I_1$ = {1, 2}, $I_2$ = {1, 3} and $I_3$ = {2, 3}. The primary objective is to select the best subset among the $I_t$.

### Model

*I* multinomial populations with *J* categories are considered. For the $i^{th}$ population, the counts, denoted by $\underset{\sim}{n}_i = (n_{i1}, \ldots, n_{iJ})'$, $i = 1, \ldots, I$, are taken. In many applications it is reasonable to assume that the $\underset{\sim}{n}_i$ have independent multinomial distributions with probabilities $\underset{\sim}{p}_i = (p_{i1}, \ldots, p_{iJ})'$, $\sum_{j=1}^{J} p_{ij} = 1$.

Letting

$$\underset{\sim}{p} = (\underset{\sim}{p}'_1, \ldots, \underset{\sim}{p}'_I)',$$

the joint likelihood is

$$l(\underset{\sim}{p} \mid \underset{\sim}{n}) \propto \prod_{i=1}^{I} \prod_{j=1}^{J} p_{ij}^{n_{ij}}. \qquad (2)$$

A priori, without any order restriction on the $p_{ij}$, we take independent Dirichlet distributions for the $p_i$,

$$\pi(\underset{\sim}{p}) = \prod_{i=1}^{I} \frac{\prod_{j=1}^{J} p_{ij}^{\alpha_{ij}-1}}{D(\underset{\sim}{\alpha}_i)}, \qquad (3)$$

where the $\underset{\sim}{\alpha}_i = (\alpha_{i1}, \ldots, \alpha_{iJ})'$ and $\alpha_{ij}$ are fixed quantities to be specified. Note that in (3) $D(\underset{\sim}{a}) = \{\prod_{j=1}^{J} \Gamma(a_j)\}\{\Gamma(\sum_{j=1}^{J} a_j)\}^{-1}$ and $\Gamma(\cdot)$ is the gamma function. In (3), $\alpha_{ij} = ½$ is taken for three reasons. First, it is difficult to elicit information about $\alpha_{ij}$ even though they can be interpreted as cell counts in a prior sensory evaluation. Second, one does not want to model

similarity among the different products as it is believed that a priori some of them are better than others. Third, it simplifies the computation a lot if the $\alpha_{ij}$ are taken known, rather than if an assumption is made about their distributions a priori. Thus, to ensure the maximum heterogeneity (no preference) Jeffreys' reference prior is used (i.e., $\alpha_{ij} = ½$), a proper density in this application. In classical statistics, this is equivalent to adding a ½ to the cell counts; a recommendation usually made for sparse categorical tables. Rather, prior information will be inputted through the belief about which is the best product.

Criteria for Selection

One criterion that can be used is based on the random variable $X_i$ representing values on the hedonic scale. That is, letting $a_j$ denote the values on the ordinal scale,

$$Pr(X_i = a_j \,|\, \underset{\sim}{p}_i) = p_{ij}, \quad j = 1,\ldots,J, \quad i = 1,\ldots,I$$

and the mean of $X_i$ is denoted by $\mu_i = \sum_{j=1}^{J} a_j p_{ij}$ .

First, to introduce the general criterion, suppose a single population is selected; let *b* denote the selected population. The best (selected) population is defined as the one for which

$$\mu_b \geq \max\{\mu_i, i = 1,\ldots,I\}. \qquad (4)$$

That is, the population with the largest mean is selected. Thus, the best population is defined by using the simple tree order; see Robertson, Wright and Dykstra (1988). Such an order restriction arises naturally in many situations. For example, if an investigator wishes to compare several treatments with a new one, the prior information that the new treatment mean is at least as large as the others might be entertained. Because of its simplicity, (4) is popular.

Nandram (1997) used criteria based on the mean, standard deviation and coefficient of variation of the $X_i$ to obtain the best multinomial population (not best subset) among a set of populations. However, he used the scores on the hedonic scale to construct these criteria.

For subset selection, let $I_b$ denote the set containing the $\ell$ best populations. (Note that $I_b$ is a proper nonempty subset of the set of $I$ products.) Then, based on the means, the (best) selected set of populations $I_b$ is defined as the one for which

$$\min\{\mu_i; i \in I_b\} \geq \max\{\mu_i; i \notin I_b\}. \qquad (5)$$

Note that (4) is a special case of (5), and (5) can be viewed as an extension of the simple tree order.

Unfortunately, the method of subset selection based on the mean, uses the category scales. The $a_j$ are almost always unknown and are usually taken as $a_j = j$, $j = 1,\ldots, J$. The thesis is that this is inaccurate, and an alternative solution based on a modification of the stochastic ordering is sought. However, the method of subset selection based on the mean will be used for comparison with the method which does not use the category scales.

A single criterion based on a version of the stochastic ordering is obtained, but first, an explanation for why the stochastic ordering cannot be used directly is provided. For simplicity, consider selecting the best population. Let $A_{is} = \{p : \sum_{j=1}^{s} p_{ij} \leq \max(\sum_{j=1}^{s} p_{tj}, t = 1,\ldots,I, t \neq i)\}$, $s = 1,\ldots,J-1$, and $S_i = \bigcap_{j=1}^{J-1} A_{ij}$ . Then for each $j$ the $A_{ij}$ are mutually exclusive, $\sum_{i=1}^{I} P(A_{ij}) = 1$, and $P(S_i) \leq \min\{P(A_{ij}), j = 1,\ldots,J-1\}$ . As the $P(A_{ij})$ are different for each *i*, for some choice of *s* and some *i*, $P(A_{is}) > \min\{P(A_{ij}), j = 1,\ldots,J-1\}$ .

Then, $\sum_{i=1}^{I} P(S_i) < \sum_{i=1}^{I} P(A_{is}) = 1$. That is, while the $S_i$ are mutually exclusive, they are not exhaustive. In fact, $P(S_i)$ is not the probability that the $i^{th}$ population is the best; the $P(S_i)$ could be extremely small and $\sum_{i=1}^{I} P(S_i) << 1$. Thus, for each j $\{A_{ij}, I = 1,\ldots,I\}$ will be used as a partition to identify the best population or subset.

Letting

$$\Delta_{ik} = \sum_{j=k}^{J} p_{ij}, k = 2,\ldots,J,\ i = 1,\ldots,I, \qquad (6)$$

these $\Delta_{ik}$ are measures of the quality of the $i^{th}$ product. Note that $\Delta_{ik}$ is the probability of getting at least response $k$ on the ordinal scale (e.g., $\Delta_{i,\ \frac{J+1}{2}}$ is the probability of getting at least a neutral response). To express uncertainty about the best subset of populations, let $B$ denote the random variable indicating the best population and $\kappa$ denote exclusively the measure of quality which is used. Let $A_{tk} = \{p : \min\{\Delta_{ik}, i \in I_t\} \geq \max\ \{\Delta_{ik}, i \notin I_t\},\ \ t = 1,...,T, k = 2,...,J$, and $S_{t2} = A_{t2} = S_{tk} = A_{ts} - \bigcup_{j=2}^{s-1} A_{tj},\ \ s = 3,\ldots,J$. Then, $\kappa = k$ if $\underset{\sim}{p} \in A_{tk},\ \ k = 2,\ldots,J$ is defined (However, note that $\kappa$ is a nuisance parameter.). The criterion based on $S_{bk}$ is defined as the modified stochastic ordering (MSO) criterion. Then,

$$Pr(B = b, \kappa = k) = \omega_{bk},\ b = 1,\ldots,T,$$
$$k = 2,\ldots,J,\ \sum_{b=1}^{I}\sum_{k=1}^{J} \omega_{bk} = 1, \qquad (7)$$

where the $\omega_{bk}$ are to be specified. Letting $\lambda_b = \sum_{k=1}^{J-1} \omega_{bk}$, a priori the best population is the $b^{th}$ population for which $\lambda_b = \max\{\lambda_t, t = 1,\ldots,T\}$. The $\lambda_b$ are to be updated using the data.

Incorporating prior information about which is the best entree through the $\omega_{bk}$ rather than the $\alpha_{ij}$ is preferred. It should be noted that it is conceptually simple and convenient to use the random variables $B$ and $\kappa$ to model uncertainty about which is the best entree. On the other hand, it is much more difficult to add information about which is the best entree through the $\alpha_{ij}$. However, unless the $\alpha_{ij}$ are all equal, their specification will give latent information about which is the best entree, but this information is difficult to discern.

In addition, if there is a reluctance to specify the $\alpha_{ij}$, then in the Bayesian paradigm they are random variables, and the problem of selection and estimation becomes extremely difficult, especially if one wants to incorporate uncertainty about which is the best population.

For the criterion given by (5) based on the mean, $k = 1$ will be taken and define $S_{b1} = \{p : \min\{\mu_i, i \notin I_b\} \geq \max\ (\mu_i, i \notin I_b\}, b = 1,\ldots,T$. The criterion based on $S_{b1}$ will be called the mean response ordering (MRO) criterion.

Then the prior distribution on $\underset{\sim}{p}$ in (3) becomes

$$\pi\left(\underset{\sim}{p} \middle| B = b, \kappa\right) = \begin{cases} c_{bk}(\underset{\sim}{\alpha}) \prod_{i=1}^{I} \dfrac{\prod_{j=1}^{J} p_{ij}^{\alpha_{ij}-1}}{D(\underset{\sim}{\alpha_i})}, & \underset{\sim}{p} \in S_{bk}, \\ 0, & \textit{otherwise}, \end{cases} \qquad (8)$$

where

$$\underset{\sim}{\alpha} = (\underset{\sim}{\alpha'}_1, \ldots, \underset{\sim}{\alpha'}_I)'$$

and

$$c_{bk}(\underset{\sim}{\alpha})^{-1} = \int_{S_{bk}} \prod_{i=1}^{I} \frac{\prod_{j=1}^{J} p_{ij}^{\alpha_{ij}-1}}{D(\underset{\sim}{\alpha_i})} \underset{\sim}{dp},\ b = 1,\ldots,I,\ k = 2,\ldots,J.$$

Note that

$$c_{bk}(\underset{\sim}{\alpha})^{-1} = \Pr(\underset{\sim}{p} \in S_{bk}), b = 1,\ldots,T,\ k = 2,\ldots,J.$$

These quantities are to be updated by the data, and are to be used to update the $\omega_{bk}$ which, in turn, are to be used to judge the best product or set of products.

Bayesian Selection and Estimation

Now, it is shown how to use the data to judge the best subset, and then to make inference about the best set of populations.

Let

$$n'_{ij} = n_{ij} + \alpha_{ij},\ \underset{\sim}{n}'_i = (n'_{i1},\ n'_{i2},\ \ldots,\ n'_{iJ})'$$

and

$$\underset{\sim}{n}' = \{n'_{ij} : i = 1,\ldots,I;\ j = 1,\ldots,J)\}\ .$$

Using Bayes' theorem, the joint posterior distribution of $\underset{\sim}{p}$, $B$ and $\kappa$ is

$$\begin{aligned} &f(\underset{\sim}{p}, B=b, \kappa=k \mid \underset{\sim}{n}) \\ &= f(\underset{\sim}{p} \mid \underset{\sim}{n}, B=b, \kappa=k) P(B=b, \kappa=k \mid \underset{\sim}{n}) \end{aligned} \tag{9}$$

where

$$f(\underset{\sim}{p} \mid \underset{\sim}{n},\ B=b,\ \kappa=k)$$

and

$$P(B=b,\ \kappa=k \mid \underset{\sim}{n})$$

are to be described. First,

$$f\left(\underset{\sim}{p} \mid \underset{\sim}{n}, B=b,\ \kappa=k\right) = \begin{cases} c_{bk}(\underset{\sim}{\alpha}) \prod_{i=1}^{I} \dfrac{\prod_{j=1}^{J} p_{ij}^{n'_{ij}-1}}{D(\underset{\sim}{n}'_i)}, & \underset{\sim}{p} \in S_{bk} \\ 0, & \textit{otherwise}, \end{cases} \tag{10}$$

where

$$c_{bk}(\underset{\sim}{n}')^{-1} = \int_{S_{bk}} \prod_{i=1}^{I} \frac{\prod_{j=1}^{J} p_{ij}^{n'_{ij}-1}}{D(\underset{\sim}{n}'_i)}\ d\underset{\sim}{p},$$

*b=1,…,T, k=2,…,J.*

For convenience, letting $\bar{S}_{bk}$ be the complement of $S_{bk}$,

The following is defined as,

$$\begin{aligned} &\bar{c}_{bk}(\underset{\sim}{n}')^{-1} \\ &= 1 - c_{bk}(\underset{\sim}{n}')^{-1} \int_{\bar{S}_{bk}} \prod_{i=1}^{I} \frac{\prod_{j=1}^{J} p_{ij}^{n'_{ij}-1}}{D(\underset{\sim}{n}'_i)}\ d\underset{\sim}{p}. \end{aligned}$$

Second, letting

$$r_{bk}(\underset{\sim}{n}') = c_{bk}(\underset{\sim}{\alpha}) c_{bk}(\underset{\sim}{n}')^{-1},$$

$$b = 1,\ldots,T,\ k = 2,\ldots,J,$$

$$(B=b,\ \kappa=k \mid \underset{\sim}{n}) = \hat{\omega}_{bk} = \omega_{bk} r_{bk}(\underset{\sim}{n}')$$

$$Pr\ \left\{ \sum_{t=1}^{T} \sum_{j=2}^{J} \omega_{tj} r_{tj}(\underset{\sim}{n}') \right\}^{-1}. \tag{11}$$

Letting

$$\hat{\lambda}_b = \sum_{j=2}^{J} \hat{\omega}_{bj}, \tag{12}$$

in (11), a posteriori the best subset is the $b^{th}$ subset for which $\hat{\lambda}_b = \max(\hat{\lambda}_t, t = 1,\ldots,T)$.

Consider testing $H_o$: $b^{th}$ subset is the best versus $h_1$: $b^{th}$ subset is not the best where $Pr(H_0) = \lambda_b = 1 - Pr(H_1)$. Then the Bayes factor, $B_f$, for testing $H_0$ versus $H_1$ is

$$B_f = \{\hat{\lambda}_b / (1-\hat{\lambda}_b)\} \{\lambda_b / (1-\lambda_b)\}^{-1}.$$

Letting

$$c_b^*(\underset{\sim}{\alpha})^{-1} = \sum_{j=2}^{J} c_{bj}(\underset{\sim}{\alpha})^{-1}$$

and

$$c_b^*(\underset{\sim}{n}')^{-1} = \sum_{j=2}^{J} c_{bj}(\underset{\sim}{n}')^{-1},$$

it follows easily from (11) and (12) that the Bayes factor is also given by

$$B_f = \{c_b^*(\underset{\sim}{\alpha})c_b^*(\underset{\sim}{n}')^{-1} - c_b^*(\underset{\sim}{n}')^{-1}\}\{1 - c_b^*(\underset{\sim}{n}')^{-1}\}^{-1}$$

$$\approx c_b^*(\underset{\sim}{\alpha})c_b^*(\underset{\sim}{n}')^{-1} \approx$$

$$(J-1)^{-1}\sum_{j=2}^{J} c_{bj}(\underset{\sim}{\alpha})c_{bj}(\underset{\sim}{n}')^{-1} = \bar{r}_b(\underset{\sim}{n}') \ . \qquad (13)$$

In (13) the first approximation follows because in many examples $c_b^*(\underset{\sim}{n}') >> 1$. This is true when there is a large number of subsets as in our application. Also in (13) the second approximation follows if the $c_{bk}(\underset{\sim}{\alpha})$ are approximately constant which is the case with a uniform prior on $B$ and κ. Note that $\bar{r}_b(\underset{\sim}{n}')$ is the average of the $\bar{r}_{bk}(\underset{\sim}{n}')$ in (11). Thus, it is interesting to observe that one might interpret $\bar{r}_b(\underset{\sim}{n}')$ as the Bayes factor, which, in turn, can be interpreted as the odds for $H_0$ provided by the data. For a review of the literature on the Bayes factor and its interpretation see Kass and Raftery (1995).

Inference proceeds by first picking with uncertainty the best subset (i.e., the subset with the largest $\hat{\lambda}_b$). Whether the frequentist method or the Bayesian method is used, the statistician will be uncertain about which is the best subset of populations. However, in the Bayesian method, as presented here the statistician can incorporate uncertainty about the best population, and this is attractive because by (11) the uncertainty about the best population a posteriori can be quantified. In addition, a posteriori inference about the parameters of the judged best population is obtained by using the posterior distribution

$$\pi\left(\underset{\sim}{p}_b \middle| \underset{\sim}{n}\right) = \sum_{t=1}^{T} \hat{\lambda}_t \pi(\underset{\sim}{p}_b \middle| \underset{\sim}{n},\ B = t) \ . \qquad (14)$$

The elegance in the current approach is contained in (14), as the weakness in the classical approach, is that after the best population is obtained the methods usually proceed as though it is known with certainty which is the best population.

The expression in (14) can be simplified. For

$$\pi\left(\underset{\sim}{p}_b \middle| \underset{\sim}{n}\right) = \hat{\lambda}_b \pi(\underset{\sim}{p}_b \middle| \underset{\sim}{n},\ \underset{\sim}{p} \in S_b) + (1-\hat{\lambda}_b)\pi(\underset{\sim}{p}_b \middle| \underset{\sim}{n},\ \underset{\sim}{p} \in \bar{S}_b) \ , \qquad (15)$$

where

$$\pi(\underset{\sim}{p}_b \middle| \underset{\sim}{n}, \underset{\sim}{p} \in \bar{S}_b) = \begin{cases} c_b^*(\underset{\sim}{n}')\prod_{i=1}^{I} \dfrac{\prod_{j=1}^{J} p_{ij}^{n_{ij}'-1}}{D(\underset{\sim}{n}_i')}, & \underset{\sim}{p} \in \bar{S}_b \\ 0, & \text{otherwise}, \end{cases}$$

$S_b = \bigcup_{K=2}^{J} S_{bk}$, $\bar{S}_b$ is the component of $S_b$, and $\bar{c}_b^*(\underset{\sim}{n}')^{-1} = 1 - c_b^*(\underset{\sim}{n}')^{-1}$.

When the criterion based on the mean is used, the following is taken

$$\beta_i = \sum_{j=1}^{J} j p_{ij}$$

and

$$\tau_i = \left\{\sum_{j=1}^{J} p_{ij}(j-\beta_i)^2\right\}^{1/2}, \ i = 1, \ldots, I.$$

When the criterion based on the modified stochastic ordering is used, the following is taken

$$\ln\{\gamma_{ij}/(1-\gamma_{ij})\} = (\theta_j - \beta_i)/\tau_i, \ i = 1,\ldots,I, \ j = 1,\ldots,J-1,$$

where

$$\gamma_{ij} = \sum_{s=1}^{j} p_{is} \text{ and } \theta_1 < \theta_2 < \ldots < \theta_{J-1}$$

are the unknown cutpoints. A posteriori inference about $\beta_i$ and $\tau_i$ can be obtained by using (15). Inference is made about the population means $\beta_i$ and standard deviations $\tau_i$, $i=1,\ldots,I$.

Computations

In this section, a description of how to compute $\hat{\lambda}_b$ in (12) and $\pi\left(\underset{\sim}{p}_b \middle| \underset{\sim}{n}\right)$ in (15) is provided.

First, consider $\hat{\lambda}_b$. Although it is more accurate to compute $\bar{r}_{bk}(\underset{\sim}{n}')$ directly rather than first computing $c_{bk}(\underset{\sim}{\alpha})$ and $c_{bk}(\underset{\sim}{n}')$ separately, a simple method is proposed which first obtains $c_{bk}(\underset{\sim}{\alpha})$ and $c_{bk}(\underset{\sim}{n}')$. How to obtain $c_{bk}(\underset{\sim}{n}')$, or $\bar{c}_{bk}(\underset{\sim}{n}')$ is described, for which the simple method suggested by Nandram, Sedransk and Smith (1997) is used. The problem of estimating $\bar{r}_{bk}(\underset{\sim}{n}')$ directly is a special case of the more general problem associated with estimating the ratio of two normalization constants; see, for example, Meng and Wong (1996) and Chen and Shao (1997) who used Markov chain Monte Carlo methods. (These refinements are unnecessary in this application.) Denoting the joint unrestricted posterior distribution of $\underset{\sim}{p}$ by

$$f''\left(\underset{\sim}{p} \middle| \underset{\sim}{n}\right),$$

therefore,

$$f''(\underset{\sim}{p}|\underset{\sim}{n}) = \begin{cases} \prod_{i=1}^{I} \dfrac{\prod_{j=1}^{J} p_{ij}^{n'_{ij}-1}}{D(\underset{\sim}{n}'_i)} & 0 \le p_{ij} \le 1, \ \sum_{j=1}^{J} p_{ij} = 1 \\ 0, & \text{otherwise}. \end{cases} \tag{16}$$

$N$ independent multivariate samples are selected from the unrestricted product Dirichlet distributions with parameters $\underset{\sim}{n}'_i, i = 1,\ldots,I$ in (16), and find the number $\dfrac{N_{\underset{\sim}{n}'_i}}{N}$ falling inside $S_{bk}$. (Note that $\bar{c}_{bk}(\underset{\sim}{n}')^{-1}$ is estimated by the proportion $1 - T_{\underset{\sim}{n}'}$, falling outside $S_{bk}$.) The Monte Carlo sample size, $N$, is obtained by taking, for example,

$$Pr\left\{\left|c_{bk}(\underset{\sim}{n}')T_{\underset{\sim}{n}'} - 1\right| < .01\right\} = 0.95. \tag{17}$$

For the examples discussed, $N$=10,000 is taken. The computations for $c_{bk}(\underset{\sim}{n}')^{-1}$ or $\bar{c}_{bk}(\underset{\sim}{n}')^{-1}$ are performed for whichever requires smaller Monte Carlo sample size in (17). Estimates of the $c_{bk}(\underset{\sim}{\alpha})$ are obtained in a similar manner. But note that with a uniform prior on $B$ and κ, it is unnecessary to compute $c_{bk}(\underset{\sim}{\alpha})$ since they are all equal. Otherwise, $r_{bk}(\underset{\sim}{n}')$ are obtained by monitoring the estimates of the ratios of $c_{bk}(\underset{\sim}{\alpha})$ and $c_{bk}(\underset{\sim}{n}')$ for convergence. Again 10,000 iterates suffice.

Samples from the posterior distribution of $\underset{\sim}{p}_b$, $\pi\left(\underset{\sim}{p}_b \middle| \underset{\sim}{n}\right)$ in (15), can be obtained by using the composition method (Tanner 1993). First, draw a uniform random variate, $U \sim U(0,1)$. Then if $U \le \hat{\omega}_{bk}$, draw $\underset{\sim}{p}_b$ from $\pi\left(\underset{\sim}{p}_b \middle| \underset{\sim}{n}, \underset{\sim}{p} \in S_{bk}\right)$; otherwise draw $\underset{\sim}{p}_b$ from $\pi\left(\underset{\sim}{p}_b \middle| \underset{\sim}{n}, \underset{\sim}{p} \in \bar{S}_{bk}\right)$. Samples of $\underset{\sim}{p}_b$ from $\pi\left(\underset{\sim}{p}_b \middle| \underset{\sim}{n}, \underset{\sim}{p} \in S_{bk}\right)$ can be obtained simply by drawing $\underset{\sim}{p}_b$ from $f''\left(\underset{\sim}{p} \middle| \underset{\sim}{n}\right)$ and then if $\underset{\sim}{p} \in S_{bk}$, accept it. Similarly, samples from $\pi\left(\underset{\sim}{p}_b \middle| \underset{\sim}{n}, \underset{\sim}{p} \in \bar{S}_{bk}\right)$ are obtained by simply drawing $\underset{\sim}{p}$ from $f''\left(\underset{\sim}{p} \middle| \underset{\sim}{n}\right)$ and then if $\underset{\sim}{p} \in \bar{S}_{bk}$, accept it. However, it is still possible to obtain

samples from $\pi''\left(p_b \middle| n\right)$ more efficiently.

It is not difficult to show that $\lambda_b c_b^*(\alpha) \neq 1$, then $\hat{\lambda}_b c_b^*(n') < 1$ if and only if $(1-\hat{\lambda}_b)\bar{c}_b^*(n') > 1$. Also, it is not difficult to show that if $\hat{\lambda}_b c_b^*(n') < 1$, then

$$\pi\left(p_b \middle| n\right) = \hat{\lambda}_b c_b^*(n')\, f''\left(p_b \middle| n\right) + (1-\hat{\lambda}_b c_b^*(n'))\,\pi\left(p_b \middle| n\right),\ p \in \bar{S}_b) \tag{18}$$

and if

$$(1-\hat{\lambda}_b)\bar{c}_b^*(n') < 1,$$

then

$$\pi\left(p_b \middle| n\right) = (1-\hat{\lambda}_b)\bar{c}_b^*(n')\, f''\left(p_b \middle| n\right) + (1-(1-\hat{\lambda}_b)\bar{c}_b^*(n'))\,\pi\left(p_b \middle| n\right),\ p \in S_b). \tag{19}$$

Note that $f''(p_n \mid n)$ is obtained by marginalization of the posterior distribution $f''(p \mid n)$, in (16). Related arguments are given by Bhattacharya and Nandram (1996). Note that the application $\hat{\lambda}_b$ could be very small and $c_b^*(n')^{-1}$ very close to 1, so that it is very likely that (18) is the choice.

Thus, samples from the posterior distribution $\pi\left(p_b \middle| n\right)$ can be obtained by using the composition method in either (15), (18) or (19). Notice that it is really simple to draw from $f''\left(p_b \middle| n\right)$. In practice, if $\hat{\lambda}_b c_b^*(n')$ is large but less than 1, draws can be made easily from (18), or if $(1-\hat{\lambda}_b)\bar{c}_b^*(n')$ is large but less than 1, draws can be made easily from (19). In the event that $\hat{\lambda}_b c_b^*(n')$ and $\bar{c}_b^*(n')^{-1}$ are small, or $(1-\hat{\lambda}_b)\bar{c}_b^*(n')$ and $c_b^*(n')^{-1}$ are small, one can draw efficiently from (15).

Posterior inference of any function of $p_b$ (e.g., $\Delta_{bk}$) can be obtained by using samples from $\pi''\left(p_b \middle| n\right)$ in a straightforward manner. Noting that $p$ is first drawn from $\pi''\left(p_b \middle| n\right)$, and the components $p_b$ are stripped off, one can take $\hat{p}_b^{(h)}$, $h=1,\ldots,M$ to be $M$ vectors drawn from $\pi''\left(p_b \middle| n\right)$ and $\Delta_{bk}^{(h)} = \sum_{j=k}^{J} p_{bj}^{(h)}$, $h=1,\ldots,M$. Then $E\left(\Delta_{bk} \middle| n\right)$ is estimated by $\bar{\Delta}_{bk} = M^{-1}\sum_{h=1}^{M}\Delta_{bk}^{(h)}$ and $\mathrm{var}\left(\Delta_{bk} \middle| n\right)$ is estimated by $\bar{\Delta}_{bk} = (M-1)^{-1}\sum_{h=1}^{M}(\Delta_{bk}^{(h)} - \bar{\Delta}_{bk})^2$. Note that in these estimation procedures independent samples are used, not dependent samples as in Markov chain Monte Carlo methods.

To make inference about $\beta_i$ and $\tau_i$ a random sample $p^{(1)},\ldots,p^{(M)}$ is first obtained from $\pi''\left(p_b \middle| n\right)$. Then using the criterion based on the mean, the following is computed

$$\beta_i^{(h)} = \sum_{j=1}^{J} j\, p_{ij}^{(h)} \quad \text{and} \quad \tau_i^{(h)} = \left\{\sum_{j=1}^{J} p_{ij}^{(h)}\,(j-\beta_i^{(h)})^2\right\}^{1/2},$$
$$i = 1,\ldots,I,\ h = 1,\ldots,M.$$

For the criterion based on the modified stochastic ordering, nonlinear least squares minimizing is used

$$\sum_{i=1}^{I}\sum_{j=1}^{J-1}\left\{\ln\{\gamma_{ij}^{(h)}/(1-\gamma_{ij}^{(h)})\} - (\theta_j^{(h)} - \beta_i^{(h)})/\tau_i^{(h)}\right\}^2$$

to obtain $\theta_j^{(h)}, \beta_i^{(h)}$ and $\tau_i^{(h)}$, $h=1,\ldots,M$; see appendix A for the appropriate equations. (The iterative procedure converges quickly in less than 5 steps.) Then a posteriori we take

$$\hat{\beta}_i = M^{-1}\sum_{h=1}^{M}\beta_i^{(h)}$$

and

$$\hat{\tau}_i = M^{-1}\sum_{h=1}^{M}\tau_i^{(h)}$$

with corresponding standard deviation given by

$$\left\{(M-1)^{-1}\sum_{h=1}^{M}\left(\beta_i^{(h)}-\hat{\beta}_i\right)^2\right\}^{1/2}$$

and

$$\left\{(M-1)^{-1}\sum_{h=1}^{M}\left(\tau_i^{(h)}-\hat{\tau}_i\right)^2\right\}^{1/2}.$$

Analysis of the Military Data

In this section, the methodology is applied to the Natick Food Experiment. The Meal, Ready-To-Eat (MRE) has twelve meals (menus), each consisting of four to six food items. The system contains 39 distinct foods. Some of these items occur in more than one meal and are regarded as different items in different meals, so the total number of items studied is 52. These items can be classified into five principal types: entrees, pastries, vegetables, fruits and miscellaneous. Chen, Nandram and Ross (1996) analyzed these data to predict shelf lives of the entrees, and they classified the entrees according to whether their shelf lives are short, medium or long.

Meals were purchased through the military supply procedures of the armed-forces procurement system, and the taste testing was carried out at the Natick Laboratories (NLABS). On arrival at NLABS they were inspected for completeness, immediately tested at room temperature (21ºC) and stored at four different temperatures. Those stored at room temperature were withdrawn and tested at 12, 24 36, 48, 60 months' storage.

The meals were opened by test monitors, and each item served to a panel of 36 untrained subjects who judged its acceptability on a 9-point hedonic rating scale. At a session, each consumer evaluated all the items in one meal which consists of four to six items (including an entree) served one at a time in random order with a mouth-rinsing between items.

Each item in the entire meal, which consists of the entree and the other items, was rated on the 9-point hedonic scale by each panelist (Only one storage temperature was tested for that particular meal, and other temperatures for the same meal were judged mostly by other panelists.). The panelists were chosen from a pool of volunteers comprising both military and civilian staff at NLABS. At most, two meals were tested each day, one in the morning session and one in the afternoon. Care was taken so that no panelist was used twice in the same day. Thus, it is not unreasonable to entertain the assumption that the responses across meals and storage temperatures are uncorrelated.

The samples were coded alphabetically when presented to the test-subjects. The items were all served at room temperature as they came from the package, except for the dehydrated items, which were re-hydrated with water at 60ºC before serving. The tests took place in semi-isolated booths at NLABS under standard fluorescent lighting conditions. At any withdrawal period as many as 48 sessions (twelve menus at four temperatures) were required, which means that the tests went up to 5 weeks, and individual panelists were used about ten times during that period. Thus, it is natural to assume that the responses on each item in a meal follow a multinomial distribution, with different distributions for different entrees.

For each of the 23 combinations of time and temperature, there were sensory ratings for each of the 36 panelists, and so the data for each item consisted of 828 scores. The results were studied for 12 entrees: pork sausage (1), ham-chicken loaf (2), beef patty (3), barbecued beef (4), beef stew (5), frankfurters (6), turkey (7),

beef in gravy (8), chicken (9), meat balls (10), ham slices (11) and beef in sauce (12).

Our contact at NLABS suggested, of course with uncertainty, that among the best entrees are 5, 9 and 11. In fact, Chen, Nandram and Ross (1995) found that at room temperature the shelf lives of 5, 9 and 11 are very long (12, 8 and 14 years respectively) making these estimates less useful.

In Table 1 the responses of the 36 panelists for each entree are presented for the entrees withdrawn after 12 months' storage; the last two columns contain the average (avg) and standard deviation (std) of the 36 scores. Here, chicken (entree 9) has the largest average and the smallest standard deviation, and beef stew (entree 5) seems to be a good competitor.

Further, a Bonferroni multiple comparison procedure was performed using the ANOVA procedure of SAS on the raw data. Of course, this procedure assumes that the 36 scores are normally distributed. At 12 months' storage, the procedure indicated no significant differences between the means of the entrees, suggesting that there is no best entree at 12 months' storage. Thus, a procedure which is more sensitive than classical multiple comparison is needed.

## Numerical Results

The data on the sensory evaluation of the twelve entrees withdrawn after twelve months' storage was used. Selection and estimation were studied in turn. The best subset of entrees with $t$ entrees, $t$ = 1,…, 4 were considered. First, a uniform prior on all subsets of size $t$ was considered. That is, $\lambda_b = T^{-1}, b = 1,\ldots,T$ was taken. To make comparisons a much larger prior probability $\lambda_b$ = .25 for a pre-assigned best subset and the remaining probability split equally among the ($T$ - 1) subsets was also studied. To further assess difference between the criteria based on the mean response ordering (MRO) and the modified stochastic ordering (MSO) the observed data was perturbed by replacing each of the last two cell counts by the average of the observed cell counts for the last two cells for each entree.

In Table 2, the posterior probability $\hat{\lambda}_b$ and the Bayes factor *Bf* associated with the presumed best subsets which are {9} {5, 9} ,{ 5, 9, 11} , { 5, 7, 9, 11} by criterion, data and prior weight $\lambda_b$ is presented. For the observed data when uniform prior weight is used, except for the best entree which is {9} when the MRO is used and {11} when the MSO is used, the determined subsets of size 2, 3 and 4 are the same, being exactly the presumed best subsets.

The best subsets with prior $\lambda_b$ = .25 are the same as the presumed best subsets. The posterior probabilities increase as the number of subsets increase for both MRO and MSO, but much more rapidly for the MRO. For the perturbed data, there are substantial differences between the MRO and the MSO with the uniform prior. The posterior probability decreases with the number of subsets for the MRO and less rapidly for the MSO. But in both cases the Bayes factor increases rapidly with the number of subsets, more rapidly for the MRO.

Note that the best subsets of sizes 1, 2, 3, 4 with the MRO are {5}, {5, 9} {5, 9, 11}, {1, 5, 9, 11} respectively as compared with {11}, {9, 11} {5, 9, 10}, {5, 7, 9, 10}. The best subsets with the perturbed data and $\lambda$= .25 are the same as those for the observed data for both the MRO and the MSO. Thus, the two criteria can lead to different judged best subsets. However, if the prior probability on the best subset is substantial, the two criteria provide the same best subsets, the evidence with the MRO is slightly larger than with the MSO.

In Table 3, a sensitivity analysis to investigate misspecifications with the presumed best subsets is presented. A prior probability of $\lambda_b$ = .25 is assigned to the possibly worst subsets {2}, {*2*, 4}, {*2*, 4, 6} and {*2*, 4, 6, 12} with a probability of .75 assigned equally to the remaining $T$ – 1 subsets. Again, the observed and the perturbed data are considered. With the MSO the evidence for the presumed best subsets is very weak, and in fact, the best judged subsets are the ones expected. However, with the MSO the best subsets are the same as assigned for sizes 1, 2, 3 with very weak evidence, and for size 4 the best subset is {5, 7, 9, 10} rather than

{5, 7, 9, 11} as specified by the MRO (Note that the evidence is substantial.). Although the judged best subsets for the perturbed data and the observed data are the same, there are substantial differences between the MRO and the MSO for the perturbed data. The determined subsets are different at every size and interestingly the best subset of size 4 has associated with it fairly large Bayes factors (82.5 versus 29.2). Thus, it is important to specify the correct subset a priori especially if a large prior probability is placed on such a subset. Note that the determined subsets are different for the four scenarios.

Thus, the best subsets of any size are likely to be different for the two criteria, suggesting that it is risky to use the category scales when selecting the best subsets.

Next, consider estimation of the mean response $\beta_i$ and the measure of variability $\tau_i$ for which the posterior mean and standard deviation are obtained. Letting $\delta$ denote either $\beta_i$ or $\tau_i$, we take $\text{AVG}_C = \hat{E}(\delta|\underset{\sim}{n})$ and $\text{STD}_C = \{\hat{\text{var}}(\delta|\underset{\sim}{n})\}^{1/2}$ under criterion based on $C$ (MRO or MSO). Then, consider the ratio $R_{avg} = AVG_{mso}/AVG_{mro}$ and $R_{std} = STD_{mso}/STD_{mro}$.

In Table 4, results are presented for the observed data by prior weight for the modified stochastic ordering (MSO) for subsets of size 4. Columns 3 and 4, and 7 and 8, show there are minor differences between posterior means for $\beta_i$ and $\tau_i$ respectively for $\lambda = T^{-1}$ and $\lambda$=.25. In addition, columns 5 and 9 show minor differences between the point estimates when the MRO and MSO are used. However, columns 6 and 10 show substantial differences between the MRO and MSO. $R_{std}$ under the MSO is at least twice as large under the MRO for the $\beta_i$ and at least one and a half times as large for the $\tau_i$. Note also that there are differences for $R_{std}$ between $\lambda = T^{-1}$ and $\lambda$=.25 (e.g., compare the values for entrees 7 and 10 in column 6). Thus, for estimation when little difference is expected between the posterior means with the MRO and MSO, there are substantial differences between the standard deviations.

In Table 5, ranges are considered for the ratios $R_{avg}$ and $R_{std}$ for subsets of sizes 1-4 $\lambda = T^{-1}$ and $\lambda$=.25 and for the observed data sets and the perturbed data sets for the $\beta_i$ and the $\tau_i$. The ranges for $R_{avg}$ are very similar for both $\beta_i$ and $\tau_i$ for all scenarios (i.e., the posterior means are very similar under MRO and MSO). The standard deviations are much larger under the MSO for $\beta_i$, but not so large for the $\tau_i$, and there is a slight increase in the ranges of $R_{std}$ from $T^{-1}$ to $\lambda$=.25. In addition, as expected, note that there are virtually no differences in estimation for various sizes of the subsets.

Table 1: Panelists' responses for the military sensory evaluation Response Categories

| Entree | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | avg | std |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 2 | 0 | 1 | 5 | 4 | 6 | 8 | 8 | 2 | 6.08 | 2.01 |
| 2 | 0 | 4 | 1 | 7 | 4 | 8 | 6 | 5 | 1 | 5.50 | 1.93 |
| 3 | 2 | 1 | 3 | 7 | 3 | 8 | 8 | 4 | 0 | 5.33 | 1.94 |
| 4 | 0 | 2 | 1 | 3 | 5 | 10 | 8 | 7 | 0 | 6.00 | 1.64 |
| 5 | 0 | 0 | 1 | 3 | 7 | 6 | 8 | 10 | 1 | 6.42 | 1.50 |
| 6 | 0 | 3 | 4 | 7 | 4 | 8 | 8 | 2 | 0 | 5.17 | 1.75 |
| 7 | 0 | 1 | 0 | 5 | 4 | 10 | 10 | 5 | 1 | 6.14 | 1.50 |
| 8 | 1 | 3 | 2 | 3 | 4 | 12 | 7 | 4 | 0 | 5.50 | 1.86 |
| 9 | 0 | 0 | 1 | 5 | 0 | 9 | 14 | 6 | 1 | 6.44 | 1.40 |
| 10 | 0 | 0 | 2 | 5 | 4 | 7 | 11 | 7 | 0 | 6.14 | 1.51 |
| 11 | 2 | 1 | 2 | 1 | 1 | 5 | 17 | 6 | 1 | 6.25 | 1.98 |
| 12 | 2 | 2 | 5 | 3 | 0 | 13 | 6 | 3 | 2 | 5.42 | 2.16 |

*Note*: Meals were withdrawn after twelve months' storage.

Table 2: Posterior probability, Bayes factor and the judged best subset ($b$) of entrees with a prior probability on the presumed best subset by data, criterion and prior weight

| Observed Data | | | | | | Perturbed Data | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| MRO | | | MSO | | | MRO | | | MSO | | |
| $\hat{\lambda}_b$ | $B_f$ | $t_b$ | $\hat{\lambda}_b$ | $B_f$ | $t_b$ | $\hat{\lambda}_b$ | $B_f$ | $t_b$ | $\hat{\lambda}_b$ | $B_f$ | $t_b$ |
| (a) $\lambda_b = T^{-1}$ | | | | | | | | | | | |
| .36 | 5.1 | 9 | .21 | 2.9 | 11 | .32 | 5.2 | 5 | .21 | 2.9 | 11 |
| .72 | 22.7 | 5, 9 | .34 | 4.7 | 5, 9 | .20 | 16.6 | 5, 9 | .10 | 7.5 | 9, 11 |
| .85 | 50.2 | 5, 9, 11 | .59 | 12.8 | 5, 9, 11 | .13 | 31.3 | 5, 9, 11 | .06 | 14.7 | 5, 9, 10 |
| .88 | 64.4 | 5, 7, 9, 11 | .69 | 20.1 | 5, 7, 9, 11 | .11 | 62.0 | 1, 5, 9, 11 | .04 | 22.0 | 5, 7, 9, 10 |
| (b) $\lambda_b = .25$ | | | | | | | | | | | |
| .63 | 5.1 | 9 | .38 | 1.9 | 9 | .59 | 4.3 | 9 | .41 | 2.1 | 9 |
| .88 | 22.7 | 5, 9 | .61 | 4.7 | 5, 9 | .85 | 16.6 | 5, 9 | .62 | 4.9 | 5, 9 |
| .94 | 50.2 | 5, 9, 11 | .81 | 12.8 | 5, 9, 11 | .91 | 31.3 | 5, 9, 11 | .80 | 11.7 | 5, 9, 11 |
| .96 | 64.4 | 5, 7, 9, 11 | .87 | 20.1 | 5, 7, 9, 11 | .94 | 44.9 | 5, 7, 9, 11 | .88 | 21.7 | 5, 7, 9, 11 |

*NOTE*: The presumed best subsets are {9}, {5, 9}, {5, 9,11}, {5, 7, 9, 11}; a probability $\lambda_b$ is assigned to each of these subsets and $(1-\lambda_b)(T-1)^{-1}$ is assigned to each of the remaining $(T-1)$ subsets; mean response ordering (MRO), modified stochastic ordering (MSO)

Table 3: Posterior probability, Bayes factor for the judged best subset ($b$) of entrees under misspecification of the presumed best subset by data, criterion and prior weight

| Observed Data | | | | | Perturbed Data | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Preassigned | | Determined | | | Preassigned | | Determined | | |
| $\hat{\lambda}_b$ | $B_f$ | $\hat{\lambda}_b$ | $B_f$ | $t_b$ | $\hat{\lambda}_b$ | $B_f$ | $\hat{\lambda}_b$ | $B_f$ | $t_b$ |
| (a) Mean Response Ordering (MRO) | | | | | | | | | |
| .24 | 0.3 | .25 | 6.9 | 9 | .10 | 0.3 | .30 | 5.8 | 5 |
| .02 | 0.0 | .25 | 43.8 | 5, 9 | .01 | 0.0 | .20 | 21.8 | 5, 9 |
| .00 | 0.0 | .19 | 100.2 | 5, 9, 11 | .00 | 0.0 | .13 | 41.6 | 5, 9, 11 |
| .00 | 0.0 | .12 | 128.7 | 5, 7, 9, 11 | .00 | 0.0 | .11 | 82.5 | 1, 5, 9, 11 |
| (b) Modified Stochastic Ordering (MSO) | | | | | | | | | |
| .39 | 0.6 | .39 | 0.6 | 4 | .20 | 0.7 | .20 | 0.7 | 4 |
| .25 | 0.3 | .25 | 0.3 | 2, 4 | .16 | 0.6 | .16 | 0.6 | 2, 4 |
| .37 | 0.6 | .37 | 0.6 | 2, 4, 6 | .19 | 0.7 | .19 | 0.7 | 2, 4, 6 |
| .05 | 0.1 | .07 | 69.2 | 5, 7, 9, 10 | .01 | 0.0 | .04 | 29.2 | 5, 7, 9, 10 |

*NOTE*: The presumed worst subsets are {2}, {2, 4}, {2, 4, 6}, {2, 4, 6, 12}; a probability $\lambda_b = .25$ is assigned to each of these subsets and $(1-\lambda_b)(T-1)^{-1}$ is assigned to each of the remaining $(T-1)$ subsets.

Table 4: Posterior mean and standard deviation of μ and τ under MSO, and ratios of posterior means and standard deviations for all entrees based on the judged best four entrees using the observed data by prior weight

| | | μ | | | | τ | | | |
|---|---|---|---|---|---|---|---|---|---|
| λ | Entree | AVG | STD | $R_{avg}$ | $R_{std}$ | AVG | STD | $R_{avg}$ | $R_{std}$ |
| $T^{-1}$ | 1 | 6.52 | 0.77 | 1.09 | 2.38 | 2.47 | 0.41 | 1.20 | 1.90 |
| | 2 | 5.76 | 0.64 | 1.06 | 2.05 | 1.83 | 0.35 | 0.93 | 2.07 |
| | 3 | 4.71 | 0.69 | 0.89 | 2.21 | 1.83 | 0.38 | 0.93 | 2.16 |
| | 4 | 5.82 | 0.68 | 0.99 | 2.46 | 1.49 | 0.31 | 0.86 | 1.60 |
| | 5 | 6.95 | 0.60 | 1.11 | 2.28 | 1.68 | 0.33 | 1.01 | 1.86 |
| | 6 | 4.91 | 0.65 | 0.95 | 2.27 | 1.48 | 0.30 | 0.82 | 2.04 |
| | 7 | 6.49 | 0.60 | 1.08 | 2.28 | 1.69 | 0.31 | 1.03 | 1.64 |
| | 8 | 4.94 | 0.68 | 0.91 | 2.25 | 1.74 | 0.35 | 0.91 | 1.86 |
| | 9 | 6.91 | 0.60 | 1.10 | 2.39 | 1.65 | 0.32 | 1.04 | 1.65 |
| | 10 | 6.11 | 0.67 | 1.02 | 2.56 | 1.42 | 0.29 | 0.85 | 1.76 |
| | 11 | 6.25 | 0.78 | 1.02 | 2.42 | 2.33 | 0.40 | 1.15 | 1.59 |
| | 12 | 5.37 | 0.76 | 1.00 | 2.22 | 2.51 | 0.41 | 1.16 | 2.21 |
| .25 | 1 | 6.28 | 0.71 | 1.08 | 2.80 | 2.47 | 0.41 | 1.18 | 1.99 |
| | 2 | 5.69 | 0.64 | 1.05 | 2.18 | 1.83 | 0.35 | 0.93 | 2.07 |
| | 3 | 4.67 | 0.67 | 0.88 | 2.18 | 1.82 | 0.38 | 0.93 | 2.13 |
| | 4 | 5.65 | 0.65 | 0.98 | 2.85 | 1.51 | 0.32 | 0.85 | 1.73 |
| | 5 | 7.11 | 0.58 | 1.12 | 2.69 | 1.66 | 0.32 | 1.01 | 1.91 |
| | 6 | 4.89 | 0.65 | 0.95 | 2.29 | 1.48 | 0.31 | 0.82 | 2.08 |
| | 7 | 6.73 | 0.56 | 1.09 | 3.00 | 1.68 | 0.31 | 1.06 | 1.80 |
| | 8 | 4.85 | 0.66 | 0.89 | 2.28 | 1.74 | 0.36 | 0.91 | 1.93 |
| | 9 | 7.02 | 0.57 | 1.11 | 2.73 | 1.65 | 0.31 | 1.05 | 1.74 |
| | 10 | 5.90 | 0.64 | 1.01 | 3.09 | 1.43 | 0.31 | 0.84 | 1.82 |
| | 11 | 6.44 | 0.75 | 1.03 | 3.37 | 2.28 | 0.38 | 1.18 | 1.78 |
| | 12 | 5.28 | 0.74 | 0.99 | 2.25 | 2.51 | 0.41 | 1.16 | 2.22 |

Table 5: Ranges of ratios of posterior means and standard deviations of μ and τ based on the judged best subset of sizes 1– 4 by data and prior weight

| | μ | | τ | |
|---|---|---|---|---|
| λ | $R_{avg}$ | $R_{std}$ | $R_{avg}$ | $R_{std}$ |
| (a) Observed data | | | | |
| $T^{-1}$ | 0.89-1.12 | 2.02-2.56 | 0.82-1.21 | 1.54-2.21 |
| .25 | 0.88-1.13 | 2.05-3.37 | 0.82-1.21 | 1.61-2.22 |
| (b) Perturbed data | | | | |
| $T^{-1}$ | 0.92-1.13 | 1.78-2.23 | 0.83-1.21 | 1.68-2.26 |
| .25 | 0.91-1.13 | 1.75-3.17 | 0.83-1.26 | 1.70-2.26 |

## Conclusion

The method for how to obtain the best subset of a set of multinomial populations and how to estimate the parameters of any of the selected population has been shown. In addition, it has been shown that the judged best subset can be different under the modified stochastic ordering and the mean response ordering. The methodology applies generally to many sensory data problems when a nonparametric approach might be desirable and when there are small cell counts. For an alternative nonparametric Bayesian approach to estimate several similar multinomial populations see Quintana (1998). He used a Dirichlet process prior to obtain a more robust specification of exchangeability. The method to obtain the best subset of entrees that was outlined in this article is much simpler.

Specifically, five tasks were accomplished. First, a more formal framework for selection than Morris and Christiansen (1996) and Goldstein and Spiegelhalter (1996) has been obtained. The main feature of the estimation method is that it weighs the different subsets according to which one is believed to be best. As there is a joint posterior distribution of the best population and its parameters, estimation proceeds in a simple manner. Second, most non-Bayesian procedures in ranking and selection, use the normality assumption. A normal approximation was not used in this analysis; instead work was done directly with the multinomial assumption. Third, work was done with all the categories in the multinomial table (i.e., collapsing to remove sparseness has not been done). Fourth, this method is sampling based, facilitating a complete probabilistic analysis of the best subset of multinomial populations. Fifth, the method for how to estimate the average response score and standard deviation for each food without actually using the numeric scores has been shown.

With respect to the application discussed, future work will address more complicated issues associated with different storage temperatures, and the other items including the entrees in each meal. It will be useful to obtain the best subset at all temperatures for all rated items in each food. More generally, a number of items is usually rated in accordance with a number of different characteristics. Then, one might wish to find the best subset of items when all the characteristics are taken simultaneously.

## References

Bhattacharya, B. & Nandram, B. (1996). Bayesian inference for multinomial populations under stochastic ordering. *Journal of Statistical Computation and Simulation*, *54*, 145- 163.

Chen, M-H., Nandram, B. & Ross, E.W. (1996). Bayesian prediction of the shelf-life of a military ration with sensory data. *Journal of Agricultural, Biological and Environmental Statistics*, *1*, 377-392.

Chen, M-H. & Shao, Q-M. (1997). Estimating ratios of normalizing constants for densities with different dimensions. *Statistica Sinica*, *7*, 607-630.

Goldstein, H. & Spiegelhalter, D. J. (1996). League tables and their limitations: Statistical issues in comparisons of institutional performance. *Journal of the Royal Statistical Society, A 159*, 384-443.

Green, P.G. (1984). Iteratively reweighted least squares for the maximum likelihood estimation, and some robust and resistant alternatives (with discussion). *Journal of the Royal Statistical Society, B 46*, 149-192.

Kass, R. E. & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association, 90*, 773-795.

McCullagh, P. (1980). Regression models for ordinal data (with discussion). *Journal of theRoyal Statistical Society, B 42*, (with discussions), 109-142.

Meng, X. L. & Wong, W. H. (1996). Simulating ratios of normalizing constants via a simple identity: A theoretical exploration. *Statistica Sinica*, *6*, 831-860.

Morris, C. N. & Christiansen, C. L. (1996). Hierarchical models for ranking and for identifying extremes, with applications. In *Bayesian Statistics V, eds.:* J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, Oxford University Press, 277-296.

Nandram, B. (1997). Bayesian inference for the best ordinal multinomial population. In *Case Studies in Bayesian Statistics, eds.: C. Gatsonis, J. S. Hodges, R. E. Kass and N. D. Singpurwalla, Volume III, Springer-Verlag*, 399-418.

Nandram, B., Sedransk, J., & Smith, S. J. (1997). Order restricted bayesian estimation of the age composition of a population of Atlantic cod. *Journal of the American Statistical Association, 92*, 33-40.

Newell, G. J. (1982). Use of linear logistic model for the analysis of sensory evaluation data. *Journal of Food Science, 47*, 818-820.

Quintana, F. A. (1998). Nonparametric Bayesian analysis for assessing homogeneity in $k \times \ell$ contingency tables with fixed right margin totals. *Journal of the American Statistical Association, 93*, 1140-1149.

Robertson, T., Wright, F. T. & Dykstra, R. L. (1988). *Order restricted statistical inference*. John Wiley & Sons, Inc., New York.

Sampson, A. R. & Whitaker, L. R. (1989). Estimation of multivariate distributions under stochastic ordering. *Journal of the American Statistical Association, 84*, 541-548.

Tanner, M. A. (1993). *Tools for statistical inference: methods for exploration of posterior distributions and likelihood functions*. Springer-Verlag, New York.

## Appendix A

For the iterative nonlinear least squares, one would take

$$\Delta_{ij} = \ln\{\gamma_{ij}/(1-\gamma_{ij})\} = (\theta_j - \beta_i)/\tau_i$$

where

$$\theta_1 < \theta_2 < \ldots < \theta_{J-1}, \gamma_{ij}$$
$$= \sum_{p=1}^{j} p_{ip},\ i = 1, 2, \ldots, I,\ j = 1, 2, \ldots, J-1.$$

Let

$$\bar{\theta} = (J-1)^{-1} \sum_{j=1}^{J-1} \theta_j,\ \bar{\Delta}_i = (J-1)^{-1} \sum_{j=1}^{J-1} \Delta_{ij},$$

$$\omega_{ij} = \left\{ \sum_{j=1}^{J-1} (\theta_j - \beta_i)/\Delta_{ij} \right\}^{-1} \{(\theta_j - \beta_i)/\Delta_{ij}\},$$

$$i = 1, 2, \ldots, I,\ j = 1, 2, \ldots, J-1.$$

Then, the normal equations, obtained by minimizing

$$\sum_{i=1}^{I} \sum_{j=1}^{J-1} \{\Delta_{ij} - (\theta_j - \beta_i)/\tau_i\}^2$$

over $\theta_j$, $\beta_i$, and $\tau_i$, are

$$\theta_j = \left( \sum_{i=1}^{I} \tau_i^{-2} \right)^{-1} \sum_{i=1}^{I} \tau_i^{-2} (\tau_i \Delta_{ij} + \beta_i) \quad j = 1, 2, \ldots, J-1, \tag{A.1}$$

$$\beta_i = \bar{\theta} - \tau_i \bar{\Delta}_i \qquad i = 1, 2, \ldots, I, \tag{A.2}$$

$$\tau_i = \sum_{j=1}^{J-1} \omega_{ij} (\theta_j - \beta_i) \Delta_{ij}^{-1}. \tag{A.3}$$

Letting

$$\hat{p}_{ij} = n_{ij} / n_i., \; \Delta_{ij}^{*} = \ln \left\{ \frac{(\hat{p}_{ij} + 1/2n_i.)}{(1 - \hat{p}_{ij} + 1/2n_i.)} \right\}$$

with

and starting values are obtained by taking

$$n_i. = \sum_{j=1}^{J} n_{ij,} \text{ for } i=1, 2,\ldots, I, j=1, 2,\ldots, J-1,$$

$$\beta_i = \sum_{j=1}^{J} j \hat{p}_{ij}, \quad \tau_i = \left\{ \sum_{j=1}^{J} \hat{p}_{ij} (j - \beta_i)^2 \right\}^{1/2},$$

$$\theta_j = \left( \sum_{i=1}^{I} \tau_i^{-2} \right)^{-1} \sum_{j=1}^{I} \tau_i^{-2} (\tau_i \Delta_{ij}^{*} + \beta_i).$$

Starting with a random sample $\underset{\sim}{p}^{(1)}, \underset{\sim}{p}^{(2)}, \ldots, \underset{\sim}{p}^{(M)}$, taking $\Delta_{ij}^{(h)} = \ln\{\gamma_{ij}^{(h)} / (1 - \gamma_{ij}^{(h)})\}$ and solving the normal equations (A.1), (A.2), (A.3), samples $\theta_j^{(h)}$, $\beta_i^{(h)}$, and $\tau_i^{(h)}$, $h=1, 2,\ldots, M$ are obtained from their empirical posterior distributions.

Journal of Modern Applied Statistical Methods
November, 2005, Vol. 4, No. 2, 500-508

1538 – 9472/05/$95.00

# An Estimator Of Intervention Effect On Disease Severity

David Siev
USDA Center for Veterinary Biologics

When a medical intervention prevents a dichotomous outcome, the size of its effect is often estimated with the prevented fraction. Some interventions may reduce the severity of an outcome without entirely preventing it. To quantify the effect of a severity-moderating intervention, a measure termed the mitigated fraction (*MF*) is proposed. *MF* has broad applicability, because it measures the overlap of two empirical distributions based on their stochastic ordering. It is also useful in the specific context of medical interventions, because it shares certain structural and functional features with the prevented fraction. The two measures may be applied together in a single semiparametric model with components for outcome prevention and for severity conditional on the presence of the outcome.

Key words: mitigated fraction, prevented fraction, vaccine efficacy

## Introduction

When a medical intervention is intended to prevent a dichotomous outcome, such as the presence or absence of disease, an estimator known as the prevented fraction (*PF*) is commonly used to measure its effect. Vaccine efficacy, for example, is often estimated using some form of prevented fraction. Some interventions are, however, intended to reduce disease severity without entirely preventing disease. It would be valuable to have an estimator that is broadly applicable for evaluating vaccine efficacy in reducing disease severity (Mehrotra, 2004). An estimator that has proved useful in animal vaccine studies is the mitigated fraction (*MF*). The mitigated fraction is a new incarnation of an old statistic with a number of salient attributes. It is both analogous in function and homologous in structure to the prevented fraction.

David Siev acknowledges helpful comments of many colleagues, particularly B. Fergen, P. Dixon, T. Katz, D. Sweeney, J. Zimmerman. Email him at David.Siev@aphis.usda.gov.

For vaccination, *PF* is the relative decrease in the probability a vaccinate will become a case, while *MF* is the relative increase in the probability that a vaccinate's disease will be less severe than a nonvaccinate's disease. This article shows its origin, describes some of its features, and illustrates how *PF* and *MF* may be components of a nested model.

## Example

A swine respiratory disease vaccine study included groups of pigs treated with either vaccine or placebo. All subjects were exposed to the pathogen and subsequently sacrificed. At postmortem examination, the extent of gross lesions in the lungs of each subject was estimated by visual approximation. Two observers independently sketched on a grid the dorsal and ventral surfaces of each of the seven lung lobes. The fraction of each lobe was taken as the average of the two surfaces and two observers. The lobe fractions were weighted (by their standard relative mass) and summed to arrive at the fraction of the lungs consisting of gross lesions. They are shown in Figure 1.

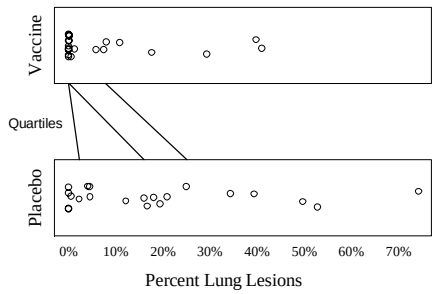



Figure 1. Fraction of lungs consisting of gross lesions. Number of subjects – placebo: 21, vaccine: 22. Points are jitter vertically to aid visualization.

How then should one analyze and summarize the findings of this study? The subjects could be divided into unaffected (0% lesions) and affected (more than 0% lesions). The prevented fraction could then be estimated, using methods for binary data. Important information is lost, however, if one only considers whether the response was present or absent and ignores its severity, particularly because most subjects were affected, and there was a wide range of response.

An approach often seen with this type of data is to calculate the average percent in each group and compare the group averages by their difference or relative difference. Taking averages is not the soundest way to summarize data that are highly skewed and border a boundary of the parameter space. The resulting summary measure also does not illuminate the vaccine's impact on individual subjects, as does *PF*, which is the relative decrease in the probability a vaccinate will become a case. A measure analogous to *PF* is *MF*, the relative increase in the probability that a vaccinate's disease will be less severe than a nonvaccinate's disease. An interesting question is whether to estimate *MF* for the entire set of data, or only for those affected by challenge. That point will be considered further when the example is revisited.

Mitigated Fraction

Prevented fraction has the general form $PF = 1 - p_2/p_1$, where, say, $p_1$ is the expected fraction of nonvaccinates affected by disease, and $p_2$ is the corresponding expectation among vaccinates. As the usual estimator of vaccine effect, *PF* is often simply termed vaccine efficacy (*VE*) in vaccine studies. Besides binomial expectations, *VE* may be constructed from other parameters that are related in some way to the probability of disease transmission (see Table 1 of Halloran et al., 1997, for an overview).

Suppose that all subjects in a vaccine trial become sick, whether vaccinated or not. Rather than looking at the effect of vaccination on the relative probability of contracting the disease, one might now wish to consider the effect of vaccination on the relative probability that the disease is milder. An estimator may be constructed that is both analogous to *PF* in function (summarizing subject probabilities) and homologous to *PF* in structure (difference relative to nonintervention).

To highlight these features, it is called the mitigated fraction (*MF*). That is, $MF = 1 - t_2/t_0$ where $t_2$ is the estimated probability that a vaccinate's disease is more severe than that of a nonvaccinate, and $t_0$ is the probability of greater severity in the absence of vaccination. *MF* may range from -1 to 1, unlike *PF*, which can take any real value no greater than 1. The difference in their ranges is related to the fact that the constituent probabilities in *MF* are relative (more or less severe than the other treatment group), while those in *PF* are not (presence or absence of disease). In practice, if a vaccine does not actually cause disease, both *MF* and *PF* will take values from 0 to 1.

If disease severity can be graded by some continuous measure or discrete assessment in a way that results in unambiguous ranks, the mitigated fraction is estimated by

$$MF = \{2W_1 - n_1(1 + n_1 + n_2)\}/n_1 n_2$$

where *W* is the familiar Wilcoxon rank sum statistic, *n* is the number of subjects in a group, and the subscripts are 1 for nonvaccinates and 2 for vaccinates.

Background

A general problem is how to distinguish between samples of two populations in some quantifiable way that avoids all parametric assumptions. A useful approach is to consider the stochastic ordering of the two empirical distributions. Figure 2 illustrates two estimators that do so,

$$T_i = \text{Prob}(Y_i > Y_j) + \tfrac{1}{2}\text{Prob}(Y_i = Y_j) .$$

For continuous random variables Prob($Y_i = Y_j$) = 0, of course, and the second term is omitted from the figure label for simplicity, but without loss of generality. If two distributions are stochastically identical, the probability that a realization from one of them is greater or lower than a realization from the other is one half. Consequently, $\theta_i$ rescales $T_i$ to range from –1 to 1, with 0 corresponding to the null probability, ½.

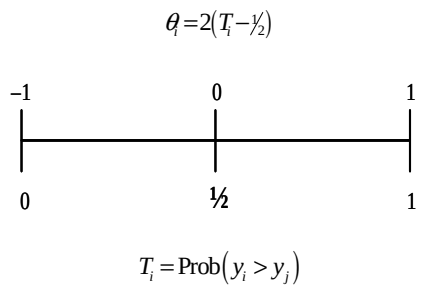



Figure 2. Because $T_i$ and $T_j$ are complementary probabilities, summing to one and equidistant from ½, $\theta_i$ may be reformulated as

$$\begin{aligned}\theta_i &= T_i - T_j \\ &= P(Y_j < Y_i) - P(Y_j > Y_i)\end{aligned}$$

In other words, $\theta_i$ is a measure of the overlap between the two distributions based on their stochastic ordering. A general measure of the overlap of two distributions is simply $\theta$, the absolute value of either $\theta_i$. $\theta = |\theta_i| = 2\left(T - \tfrac{1}{2}\right)$, where

$$T = \sup\{\text{Prob}(y_1 > y_2), \text{Prob}(y_1 < y_2)\} .$$

$\theta$ is used when comparing distributions that have no particular relative ordering. $\theta_i$, on the other hand, is useful when the distributions arise in a particular setting that establishes an ordered relationship. For example, population 2 may be manifesting the effect of a medical intervention that is being compared to population 1, representing placebo treatment.

These estimators are generalizations of known statistics. For example, mean ridits (Bross, 1958) are $T_i$, and Somers' *d* statistics (Somers, 1962) are $\theta_i$. (Vigderhous (1979) noted the connection between ridits and Somers' *d*). Somers' *d* was conceived as a measure of association between two ordinal variables, in contrast to ridit analysis, which was designed to compare the distributions of an ordinal variable in each of two distinct populations. Here, they are generalized to encompass data of all types that are not necessarily categorical and may arise from independent or correlated distributions. This general approach has been advocated by other authors (Wolf & Hogg, 1971).

It is well known that an estimate of *T* may be recovered from the Wilcoxon-Mann-Whitney statistic (Wolf & Hogg, 1971, equation 1). That may be done as follows.

$$T_i = \frac{U_i}{n_i n_j} = \frac{W_i - n_i(n_i+1)/2}{n_i n_j}$$

where

$W_i$ = sum of the ranks in group *i* (the Wilcoxon rank sum statistic), and $U_i$ = number of times a $y_{jk}$ precedes a $y_{ih}$ (the Mann-Whitney U statistic), i.e.,

$$U_i = \sum_{k=1}^{n_j}\sum_{h=1}^{n_i} \text{H}(y_{jk}, y_{ih}) ,$$

where

$$\text{H}(a,b) = 1 \text{ if } a<b;\ 0 \text{ if } a>b;\ \text{and } \tfrac{1}{2} \text{ if } a=b ,$$

and $y_{ih}$ is the response of subject *h* ($h$ = 1 ... $n_i$) in group *i* ($i$ = 1, 2).

Substituting $\theta_i = 2(T_i - ½)$ gives

$$\theta_i = \left\{2W_i - n_i(1 + n_i + n_j)\right\} \Big/ n_i n_j$$

Stratified Design

To estimate $\theta$ from stratified data use $T_i = \sum_r U_{ir} \Big/ \sum_r n_{ir} n_{jr}$, where *r* indexes the strata. For matched pairs, this reduces to a simple binomial fraction $T_i = \sum_r \mathrm{I}(y_{jr} < y_{ir}) \Big/ R$, where *R* is the number of pairs and $\mathrm{I}(\bullet)$ is the indicator function. In that case, interval estimation can proceed by familiar methods for binomial fractions.

Subject Components

*MF* may be decomposed into the contribution of individual subjects. The component for a vaccinated subject *j* is $s_j = \frac{2}{n_1} \sum_{k=1}^{n_1} \mathrm{H}(y_{2j}, y_{1k}) - 1$, which is its contribution to $MF = \frac{1}{n_2} \sum_{j=1}^{n_2} s_j$. *MF* is thus the mean of the individual subject components.

Confidence Intervals

Confidence intervals using normal approximations can be derived from the asymptotic variance for *W* or the asymptotic variance for Somers' *d* provided by popular software packages. Such intervals depend on assumptions are preferably avoided and may even contain inadmissable values. An alternative is to calculate confidence intervals for *MF* by one of the bootstrap methods (Efron & Tibshirani, 1993); this is an area of ongoing investigation.

Graphical Representation (Example)

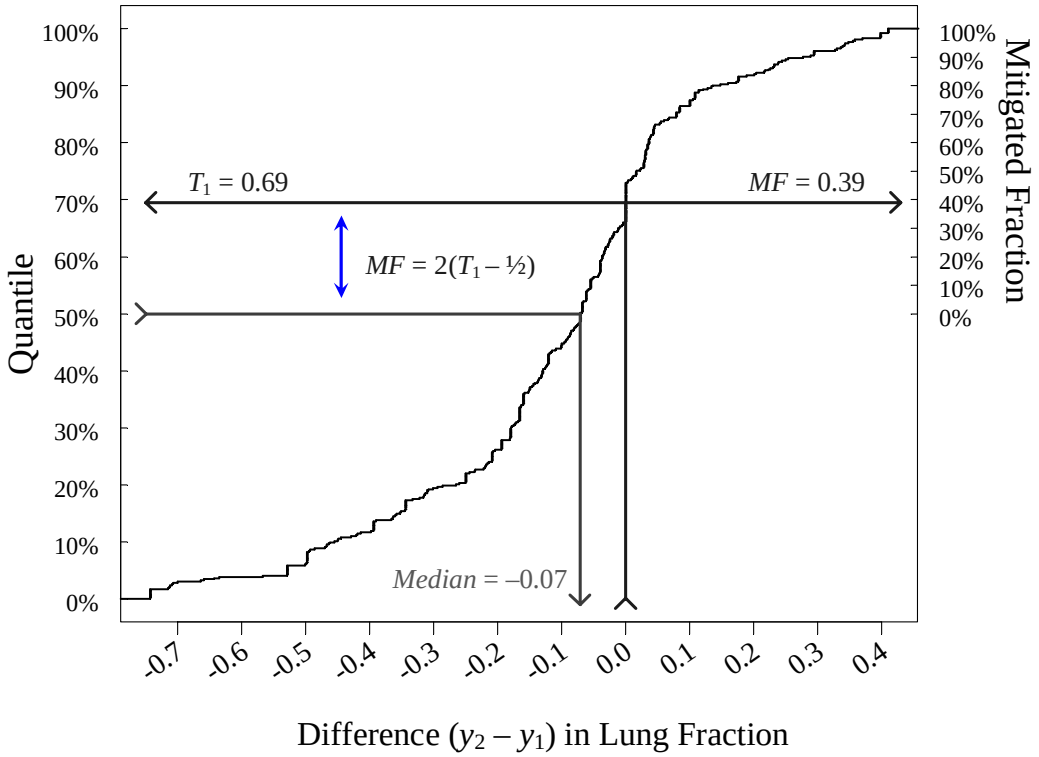
Figure 3 shows the empirical cumulative distribution function of the difference distribution, $F(Y_2 - Y_1)$, obtained from taking all pairwise differences between the groups in our example: $d_{ij} = y_{2i} - y_{1j}$, where $i = 1,...,n_2$ and $j = 1,...,n_1$. The arrow leading from the 50% quantile indicates the median difference (the Hodges-Lehmann estimator), which gives some idea of the amount of shift between the two distributions. The quantile corresponding to a difference of zero is the probability that a vaccinate's disease is less severe than that of a nonvaccinate ($T_1$). Rescaling the difference between $T_1$ and the median gives *MF*, shown in the right hand *y* axis. *MF* is thus a rescaled quantile of the difference distribution.

In contrast to the median difference, which is in the original units of measurement on the abscissa (*x* axis), *MF* reflects probabilities on the ordinate (*y* axis). In this example, $T_1 = 0.69$ means that 69% of the nonvaccinates are expected to be more severely affected than the vaccinates, $MF = 2(T_1 - ½) = 0.39$, (*95% bootstrap CI:* 0.06 to 0.68). The vaccine benefited an estimated 39% of the 50% of vaccinates who, in the absence of vaccination, would have been more severely affected than nonvaccinates.

Interpretation and application of *MF*

*MF* is the increase due to vaccination of the probability that a vaccinate's disease will be less severe than a nonvaccinate's disease, relative to the probability that it would have been less severe had the individual not been vaccinated. It is important to avoid direct comparison between *PF* and *MF*, which have somewhat different implications. Many of the usual estimators of vaccine efficacy are concerned with the prevention of outcomes that

Figure 3. Empirical difference distribution showing *MF* as a rescaled quantile.




are links in the chain of disease transmission, such as infection or infectivity, and in this respect *MF* is not like them. *PF* also relies on explicit case definitions, while *MF* is intended for situations where disease severity need only be clearly graded.

*MF* is analogous to *PF* in that it is based on estimated subject probabilities. Some relative difference measures that attempt to mimic *PF* in formulation may not necessarily have an analogous implication and should be interpreted cautiously. For example, a formulation that is often used to emulate *PF* is the relative difference of means $((\bar{y}_1 - \bar{y}_2)/\bar{y}_1)$. This is, at best, a comparison of population averages rather than subject distribution. It is rarely appropriate as the sole assessment of vaccine efficacy when the outcome is continuous rather than dichotomous (and it is particularly misleading when the data may not have arisen from a location-scale distribution). Although such estimators may be devised to emulate the configuration of *PF*, they fail to capture a similar meaning, since what is important about the constituent parameters in *PF* is not that they are means but that they are category probabilities. In this respect, *MF* is an estimator that is analogous to *PF*.

The use of mean based estimators may also arise from an understandable desire to quantify the amount of severity reduction. Unfortunately, such estimators are sensitive to the form and scale of the response measurement, which may vary substantially between similar studies. *MF*, on the other hand, is invariant to order-preserving transformations of the data. The price for such invariance is that *MF* gives no information about the magnitude of disease severity reduction, and a large value of *MF* may result from a small but highly probable reduction

in severity. That is why it is a good idea to accompany *MF* with an estimator in the original units of measurement, such as the empirical quartiles illustrated in Figure 1.

*MF* may also be estimated under a range of parametric assumptions, thereby offering a common approach to studies of various types. The example illustrates its most general application, where there are no assumptions other than that the data are legitimately ranked. *MF* could just as readily be estimated from ordinal categories or continuous data. With categorical data, the estimator based on *W* corresponds to the ridit estimator. In parametric analyses, the probabilities are obtained from the estimated cumulative distribution functions. For example, the frequency table shows the number of subjects of a drug trial in categories of increasing disease severity. (The data are a subset of those analyzed by Poon (2004).) By the formula, estimated *MF* = 0.08 (*95% bootstrap CI:* -0.07, 0.23). By Poon's latent normal model, estimated *MF* = 0.10 (*95% profile likelihood CI:* -0.11, 0.30). Regardless how the probabilities are estimated, the meaning of *MF* remains the same.

| | *increasing disease severity* → | | | | |
|---|---|---|---|---|---|
| *placebo* | 2 | 22 | 54 | 29 | 3 |
| *drug* | 4 | 23 | 45 | 22 | 2 |

## Conditional *MF* in Nested Models

### Nested Model 1

Consider a model with a component for the presence or absence of disease and a component for disease severity among only those who become sick. Suppose resistance to the pathogen is dichotomous, while the immune response to vaccination among those susceptible to challenge follows some discrete or continuous distribution. Such a model may be formulated

$$f(y) = \pi^{d}\left[(1-\pi)\, f(y \mid y>0)\right]^{1-d},$$

where $d = \mathrm{I}(y=0)$ (i.e. $d$ is an indicator taking the value 1 if y=0 and 0 otherwise) and $\pi = E(d)$, its expectation. The likelihood is then factored into a Bernoulli likelihood and a conditionally independent part which contributes to the total only for responders. This is a nested model with conditionally independent components. Since participation in the second part is conditional on crossing the hurdle of the first part, this type of nested model is sometimes termed a hurdle model (Mullahy, 1986).

If $f(y \mid y>0)$ were completely specified, say as a beta density, maximum likelihood estimation could be used to assess how the treatment groups differed with respect to prevention, conditional severity, or both. If complete specification is not warranted, *PF* may be estimated from the first part and $MF_C$, the conditional mitigated fraction among those affected, from the second part. To do so, let

$$p_i = 1-\pi_i$$

and

$$T_i^C = T_i \mid y_i>0,\, y_j>0\,.$$

Then,

$$PF = 1-p_2/p_1 \text{ and } MF_C = 2T_1^C - 1\,.$$

The conditionally independent nature of the nested components distinguishes the nested model from more complex mixture models. For example, continuous data with many zeros would, in some cases, be analyzed with a zero-inflated model. In contrast to a nested model, the nonresponse portion of a zero-inflated model describes a latent mixture of two populations, one which may be incapable of response and another capable of response but with response zero according to distribution $f_Y(y)$, leading to the formulation

$$f(y) = \left\{\lambda + (1-\lambda) f_Y(0)\right\}^{d}\left[(1-\lambda) f_Y(y \mid y>0)\right]^{1-d},$$

where $\lambda$ is the population mixture parameter.

An example of a nested model for categorized data is the well-known continuation-ratio factorization of the multinomial likelihood

into conditionally independent binomial components. It may be parameterized $L(\underline{\pi}) \propto \prod_{j=1}^{J} \delta_j^{y_j}(1-\delta_j)^{n-r_j}$, where, for the *j*th of *J* categories, $y_j$ is the category count, $\pi_j$ is the category probability, $r_j = \sum_{k=1}^{j} y_k$ is the cumulative category count, and $n = \sum_{j=1}^{J} y_j$ is the total.

The continuation ratios are $\delta_j = \pi_j / \sum_{k=j}^{J} \pi_k$, the probability of being in category *j* given not in any previous category. Continuation-ratio models are useful for tabulated health events that occur in a natural sequence. For example, the impact of a pathogen on reproductive health may be seen by the presence of normal conception, gestation, parturition, and neonatal vigor, and a subject's inclusion at any stage depends on successfully passing the previous stage. Continuation-ratio models may also be applied to ordinal categories, such as disease severity, if they are similarly considered to be nested. In some situations they may offer an alternative to the more common cumulative probability models.

Suppose disease is categorized as absent, mild, moderate, and severe, and the counts for the two groups are arrayed in a 4 x 2 contingency table. *MF* could be estimated from the entire table, or separate estimates could be obtained for *PF* and $MF_C$. *PF* would be estimated from the 2 x 2 table collapsing over categories 2 through 4, while $MF_C$ would be estimated from the 3 x 2 table that excludes the first category. A similar rationale could be applied to ranked data if each rank were thought to represent a discrete category.

Implications of Nested Model

What are the implications of the nested model for prevention and conditional severity? Suppose all nonvaccinates are sick while some vaccinates are unaffected ($p_1 = 1$, $p_2 < 1$), and disease severity is reduced among the vaccinates. *MF* is then a simple function of its components: $MF = 1 - (1 - MF_C)(1 - PF)$. Otherwise, in most practical situations where the vaccine both prevents disease ($PF > 0$) and reduces its severity among those affected ($MF_C > 0$), the relationship would be $MF < 1 - (1 - MF_C)(1 - PF)$. If the vaccine reduces disease severity among the affected but has no effect on disease prevention, although resistant individuals are found among both nonvaccinates and vaccinates ($p_1 = p_2 < 1$), the inequality reduces to $MF < MF_C$. In both latter situations, $MF_C$ and *PF* provide illuminating information and may be examined separately from *MF*. On the other hand, in the unlikely but not impossible case that the vaccine were to prevent disease but increase severity among affected vaccinates ($MF_C < 0$), *MF* could be a useful summary which balances the benefit of prevention against the detriment of increased severity.

Nested Model 2

Nested models may also be constructed when the first component is at the end, rather than the beginning, of the disease process. For example, suppose participation in the evaluation of disease severity depends on whether or not a subject survives. The model would then be

$$f(y) = \left[ f(y \mid x = 0)\, \pi \right]^{x} (1 - \pi)^{1-x},$$

where each observation consists of the pair $\{y, x\}$, *y* is the measurement of disease severity, and *x* takes the values 0 if the subject has died and 1 otherwise.

Implications of Nested Model 2

What are the implications of the nested model for severity given that a terminal outcome has not occurred? Suppose a subject dies. Is its prior disease severity relevant? There are several possibilities. For example, in an established clinical model where the severity of gross lesions predicts a possibly fatal disease, it may be valid to include the observations of all subjects, surviving or not, to assess disease severity. On the other hand, there may be no clear association between the observation and disease. Acute death may occur in response to pathogen challenge without any clinical signs at all. Retaining the observations of the dead

subjects when the severity measure is unrelated to a primary clinical outcome perpetuates an incoherent clinical model. In such cases, rank based methods are sometimes applied after assigning the dead subjects a common value greater than the maximum value of the surviving subjects. This approach treats death as simply the severest manifestation of disease, ignoring the qualitative difference between death and survival. A third position is that death is a critical event, but the prior disease severity of dead subjects is of no practical interest, leading us to exclude them from the evaluation of disease severity, but including all subjects when considering mortality. Since participation in disease severity evaluation is conditional on survival, a nested model may be constructed in which each observation consists of the pair $\{y, x\}$, where $x$ indicates whether or not the subject has died, and $y$ is the measurement of disease severity (nested model 2).

Example revisited

In the swine vaccine example, an estimate of the mitigated fraction is $MF = 0.39$ (*95% bootstrap CI:* 0.06 to 0.68). (The asymptotic approximation is 0.07, 0.71.) A number of subjects in the study did not succumb at all to pathogen challenge. Suppose resistance to the pathogen is dichotomous, while the immune response to vaccination among those susceptible to challenge follows some continuous distribution. The dichotomous response may be described by *PF*, and the continuous response by $MF_C$ , the conditional mitigated fraction among those affected. *PF* and $MF_C$ would be derived from the conditionally independent components of a hurdle model (nested model 1).

The value of nested models is that they allow simultaneous inference on two components that are conditionally independent. In the example, one would estimate *PF* by categorizing all observations as disease positive if the pathological lung fraction is greater than zero and disease negative otherwise. $MF_C$ is then estimated using only the nonzero observations. Taking that approach, point and interval estimates are $PF = 0.21$ (-0.15, 0.49), and $MF_C = 0.42$ (0.01, 0.49). Apparently, the study is insufficient for conclusive inference on either one alone.

## Conclusion

Although it is easily calculated from the Wilcoxon statistic, *MF* is aimed at estimation rather than hypothesis testing. Consequently, it helps focus attention on the clinical relevance of the outcome. Nonparametric tests are sometimes abused by those who seem to think that avoiding certain parametric assumptions also eliminates the need for forethought in study design. Care is particularly needed when observations are recorded in the form of derived ratings such as complex scoring schemes which, unlike simple grading scales, often do not preserve a clear correspondence of score with disease severity. Unless one is confident in the scores' validity when ranked, the methods shown here should not be used. Nonparametric analysis will not salvage a poorly designed scoring scheme.

Estimation requires an outcome that is quantitatively meaningful as well as clinically relevant. The study protocol should explicitly specify the outcome variable and describe how it will be recorded. Outcome specification should also aim to highlight the random structure of the data rather than conceal or ignore it by appeal to rank based methods.

For this reason, the use of nonparametric techniques in pivotal confirmatory studies has been discouraged (e.g. Longford and Nelder, 1999). Critics point out that reliance on nonparametric methods may simply postpone the search for a suitable scale of measurement and clarification of its stochastic nature, which are prerequisites for planning a study able to yield informative estimates of the size and uncertainty of relevant effects. Full distributional specification of a germane response variable is certainly ideal. Nevertheless, the basis of *MF* on ranks gives it the very qualities that are valuable in certain types of studies, particularly where a measure based on subject probabilities is preferable to an alternative measure formed from averages.

Because the mitigated fraction is comparable in structure and function to the prevented fraction, it is a useful method of estimating the benefit of an intervention that

reduces disease severity. Like *PF*, *MF* evaluates the intervention's effect by the probability a subject will benefit from the intervention. For this reason, $MF_C$ and *PF* may illuminate different aspects of the same intervention when they are components of a nested model, and *MF* may be useful in comparisons between studies. For example, animal vaccine studies typically entail challenging all subjects with the virulent pathogen. The response to challenge often varies in magnitude between studies, and, when the response is an uncategorized measure of disease severity, the relative difference between mean group responses often varies, as well. While it is difficult to completely standardize the evaluation of such studies, *MF* estimates the probability of a beneficial response to vaccination, offering a way to assess the degree of vaccine effect at different times or locations.

## References

Bross, I. D. J. (1958). How to use ridit analysis. *Biometrics, 14,* 18–38.

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

Halloran, M. E., Struchiner, C. J., & Longini, I. M. (1997). Study designs for evaluating different efficacy and effectiveness aspects of vaccines. *American Journal of Epidemiology, 146,* 789–803.

Longford, N. T. & Nelder, J. A. (1999). Statistics versus statistical science in the regulatory process. *Statistics in Medicine, 18,* 2311-2320.

Mehrotra, D. V. (2004). Vaccine clinical trials: A statistical primer. *Biopharmaceutical Report, 12*(1), 1-7.

Mullahy, J. (1986). Specification and testing of some modified count data models. *Journal of Econometrics, 33,* 341–365.

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

Vigderhous, G. (1979). Equivalence between ordinal measures of association and tests of significant differences between samples. *Quality and Quantity, 13,* 187–201.

Wolfe, D. A. & Hogg, R. V. (1971). On constructing statistics and reporting data. *American Statistician, 25*(4), 27-30

Journal of Modern Applied Statistical Methods
November, 2005, Vol. 4, No.2, 509-513

1538 – 9472/05/$95.00

# Estimating The Slope Of Simple Linear Regression In The Presence Of Outliers

Mohammed Al-Haj Ebrahem  Amjad D. Al-Nasser
Department of Statistics, Faculty of Science, Yarmouk University
Irbid, Jordan

In this article, an estimation procedure to simple linear regression in the presence of outliers is proposed. The performance of the proposed estimator, the AM estimator, is compared with other traditional estimators: least squares, Theil type repeated median, and geometric mean. A numerical example is given to illustrate the proposed estimator. Simulation results indicate that the proposed estimator is accurate and has a high precision in the presence of outliers.

Key words: Least squares, geometric mean, Theil-type estimators, simple linear regression, outliers

## Introduction

Regression analysis was first developed by Sir Francis Galton in the later part of the 19th century. Galton had studied the relation between heights of parents and children and noted that the heights of children of both tall and short parents appeared to revert or regress to the mean of the group. Galton developed a mathematical description of this tendency, the precursor of today's regression models (Neter, et. al., 1996).

Consider the simple linear regression model:

$$y_i = \alpha + \beta x_i + \varepsilon_i \text{ , i = 1,2,…,n} \qquad (1)$$

where $y_i$ is the response variable in the ith trial, α (intercept) and β(slope) are parameters. $X_i$ is a known constant, namely; the value of the predictor variable in the ith trial. $\varepsilon_i$ is a random error term with mean zero and variance $\sigma^2$.

Mohammed Al-Haj is an Assistant Professor in the Department of Statistics. His research interests are is in reliability, accelerated life testing, and non-parametric regression models. E-mail: m_hassanb@hotmail.com. Email Amjad D. Al-Nasser at amjadn@yu.edu.jo.

Most of the methods used in the literature to estimate the model parameters are based on the normality assumption. However, in some situations it is unreliable to use the normality assumption to identify the model; instead one may use non-parametric estimation approach. Moreover, if the data contains outlier observations, then robust methods are needed to polish the effect of the outliers. More details can be found in Montgomery and Peck (1992), Rousseeuw and Leroy (1987), Davies (1993), Fernandez (1997), and Olive (2005). A new non-parametric procedure is proposed in order to estimate the slope of model (1).

## Estimation Methods for Simple Linear Regression Model

The various estimators that have been suggested for the slope are as follows:

### (1) Method of Least Squares (LS)

The least square criterion requires that one consider the sum of n squared deviations; this criterion is denoted by Q

$$Q = \sum_{i=1}^{n} \left( y_i - \alpha - \beta x_i \right)^2$$

According to the method of least squares, the estimates of α (intercept) and β(slope) are those values $\hat{\alpha}_{ls}$, $\hat{\beta}_{ls}$ respectively, that minimize the criterion Q for the given sample observations

$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, using the analytical approach it can be shown that the estimate values of α (intercept) and β(slope) are

$$\hat{\beta}_{ls} = \frac{Sxy}{Sxx} \text{ and } \hat{\alpha}_{ls} = \bar{y} - \hat{\beta}_{ls}\bar{x}$$

where

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 \ ,$$

$$S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}), \ \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

and

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

Note that $\hat{\beta}_{ls}$ is unbiased estimator of β. However, regression outliers (either in x or in y) pose a serious threat to least squares analysis.

(2) The Geometric Mean Functional Relationship (GM)

This estimator was proposed by Dent (1935). This estimator has been widely used, especially in fisher's researches:

$$\hat{\beta}_{GM} = Sign(Cov(x,y)) * \left( \frac{\sum (y_i - \bar{y})^2}{\sum (x_i - \bar{x})^2} \right)^{1/2}$$

It can be noted that this estimator is symmetric in x and y. Where Cov(x,y) is the covariance of x and y. $\hat{\beta}_T = median(B_{ij})$

(3) Repeated Median Theil-Type Method (T)

Theil (1950) proposed this method. The data are ordered either to the x variable or the y variable. Find all possible pairs of observations, assuming that all $x_i$'s are distinct,

$$B_{ij} = \frac{y_{[j]} - y_{[i]}}{x_{(j)} - x_{(i)}}; \ i = 1,2,\ldots, j-1 \ , \ j = 2,3,\ldots,n$$

which yields $\binom{n}{2}$ slope values, then where m can be chosen to be the maximum divisor of n such that $m \le r$. For example, when n = 20 then m = 4 and r =5 are selected.

(4) Proposed Method (AM)

This method consists of ordering the observed pairs $(x_i, y_i)$'s, i = 1,2,…,n; by the magnitude of $x_i$'s , assuming that all $x_i$'s are distinct, then divide the observation into some groups and find all possible paired slopes. The procedure can be described as follows:

a) Arrange the observations in ascending order on the basis of the values of $x_i$; i.e., $x_{(1)} \le x_{(2)} \le \ldots \le x_{(n)}$ and the associated $y_{[1]}, y_{[2]}, \ldots, y_{[n]}$ of the original data are taken; then the new pairs will be $(x_{(i)}, y_{[i]})$

b) Divide the data into m-subgroup each of size r such that m*r = n; then the sample can be rewritten in the form in Figure 1 on the following page.

c) Find all possible paired slopes

$$\left\{ b(k)_{ij} = \frac{y_{[j]} - y_{[i]}}{x_{(j)} - x_{(i)}}; \ i = 1,2\ldots, j-1; \ j = 2,3,\ldots,r \right\};$$
$$k = 1,2,\ldots,m$$

d) Then the estimated value of the slope can be defined as follows:

$$\hat{\beta}_{AM} = \underset{k}{Median}\left\{ b(k)_{ij}, \ i = 1,2\ldots, j-1; \ j = 2,3,\ldots,r \right\};$$
$$k = 1,2,\ldots,m$$

Note that the suggested estimator is in the form of Theil's estimator with $m\binom{r}{2}$ paired slopes to be evaluated. If the sample size n is a prime number, then the estimates leads exactly to the repeated median Theil type estimator.

$$\begin{array}{cccc} (x_{(1)}, y_{[1]}) & (x_{(2)}, y_{[2]}) & \cdots & (x_{(r)}, y_{[r]}) \\ (x_{(r+1)}, y_{[r+1]}) & (x_{(r+2)}, y_{[r+2]}) & \cdots & (x_{(2r)}, y_{[2r]}) \\ \vdots & \vdots & \vdots & \vdots \\ (x_{((m-1)*r+1)}, y_{[(m-1)*r+1]}) & \cdots & \cdots & (x_{(mr)}, y_{[mr]}) \end{array}$$

Figure 1

However the advantage of the proposed one is in abstracting the number of paired slopes to be evaluated, for example when n = 100, 4950 paired slopes are needed to be evaluated by using T method. By using the suggested method (AM), where r = m = 10, only 450 paired slopes are needed, which is a good advantage for this method.

Numerical Example

In order to compare various estimation methods, the so-called Pilot-Plant data from Daniel and Wood (1971) is considered. The observed (y) corresponds to acid content determined by titration and the observed (x) is the organic acid content determined by extraction and weighing. Moreover, Rousseeuw and Leroy (1987) analyzed this data further by assuming that one of the observations is wrongly recorded, i.e. the x-value of the sixth observation might have been wrongly recorded as 370 instead of 37. Based on the data which consist of 20 observations, and for the fact the x's data point should be distinct, $x_{20}$ is substituted to be 168 instead of 167. The various estimated slopes yielded the results as shown in Table.1.

In this example, for the proposed method, the original sample is divided into 4 sub-samples, each of size 5. The results showed that traditional LS and GM methods have been strongly affected by the single outliers. On the other hand, AM and T are hardly affected by the wild observation.

Simulation Study

To illustrate the performance of the proposed method in the presence of outliers, a simulation study was carried out as follows: it begins by generating 100 observations according to the model; $y_i = 1 + x_i + \varepsilon_i$, where $x_i = 10\frac{i}{n}$ and $\varepsilon_i \sim N(0,1)$. Then, the data is contaminated; at each step a certain percentage of the observations are deleted and replaced with outliers' observations. The contaminated data point was generated according to the given relationship where $\varepsilon_i \sim N(20,25)$. Table.2 presents the values of the estimated slopes:

The properties of these methods were investigated further by looking at the mean square of error (MSE) in 10000 trials. For each 10000 trials, samples of size 20 and 50 were generated, the simulation results are represented in Table.3.

Table.1 The slope estimates using different methods for Pilot-Plant data

| Slope | $x_6 = 370$ | $x_6 = 37$ |
|---|---|---|
| Least Squares (LS) | 0.0808 | 0.3211 |
| Geometric Mean (GM) | 0.2148 | 0.3220 |
| Theil (T) | 0.3170 | 0.3194 |
| Proposed method (AM) | 0.3273 | 0.3480 |

Table.2. Slope Estimates with n= 100 and $\beta$ =1

| Contamination (%) | LS | GM | T | AM |
|---|---|---|---|---|
| 0 | 0.9977 | 1.0590 | 0.9906 | 0.8491 |
| 10 | -0.1176 | -1.9339 | 0.8585 | 0.7911 |
| 20 | -0.9760 | -2.4261 | 0.6003 | 0.7675 |
| 30 | -1.6041 | -2.7429 | -.05473 | 0.7574 |
| 40 | -1.9215 | -2.7781 | -1.4783 | 0.5783 |
| 50 | -2.0421 | -2.8190 | -1.7236 | 0.5214 |

Table.3. MSE of the Slope in the presence of outliers

| Contamination (%) | Sample Size / Slope | 20 | 50 |
|---|---|---|---|
| 0 | LS | 6.0016E-03 | 2.3847E-03 |
| | GM | 8.4800E-03 | 5.4053E-03 |
| | T | 6.5697E-03 | 2.5118E-03 |
| | AM | 1.2690E-01 | 7.1048E-02 |
| | | | |
| 10 | LS | 1.2115E+00 | 1.1850E+00 |
| | GM | 6.1172E+00 | 6.5467E+00 |
| | T | 2.7433E-02 | 2.1701E-02 |
| | AM | 2.7372E-01 | 1.9499E-01 |
| | | | |
| 20 | LS | 3.7599E+00 | 3.7167E+00 |
| | GM | 1.1129E+01 | 1.1212E+01 |
| | T | 1.8782E-01 | 1.7369E-01 |
| | AM | 2.3882E-01 | 1.0105E-01 |
| | | | |
| 30 | LS | 6.4511E+00 | 6.3880E+00 |
| | GM | 1.3218E+01 | 1.3285E+01 |
| | T | 2.4676E+00 | 2.2527E+00 |
| | AM | 3.2630E-01 | 3.0625E-01 |
| | | | |
| 40 | LS | 8.4146E+00 | 8.3348E+00 |
| | GM | 1.4609E+01 | 1.4647E+01 |
| | T | 5.8036E+00 | 5.6501E+00 |
| | AM | 2.1543E-01 | 1.5468E-01 |
| | | | |
| 50 | LS | 9.12418E+00 | 9.04105E+00 |
| | GM | 1.52952E+01 | 1.53539E+01 |
| | T | 7.13609E+00 | 7.00981E+00 |
| | AM | **5.62811E-01** | 3.85401E-01 |

## Conclusion

Our simulation results from Table.3 indicate that, in terms of MSE the performance of the four estimators in the absences of outliers are comparable. However, as the degree of contamination increases LS and GM methods became very sensitive to the presence of outliers. Theil-Type estimator (T), clearly affected with the outliers when the contamination became 30% or more. It is very clear that the proposed estimator (AM) is very robust in the presence of outliers. As a conclusion, the AM estimator can be consider as a good alternative to the traditional methods because it is able to produce satisfactory results even in the presence of a large amount of outliers.

## References

Daniel, C. & Wood, F. S. (1975). *Fitting equation to data*. John Wiley: New York.

Davies, P. L. (1993). Aspects of robust linear regression. *The Annals of Statistics, 21*, 1843-1899.

Dent, B. (1935). On observations of points connected by a linear relation. *Proceedings of the Physical Society, 47*, 92-106.

Fernandez, G. C. J. (1997). *Detection of model specification, outlier and multicollinearity in multiple regression using Partial regression/Residual plots.* SAS Users Group International Conference. San Diego, CA.

Neter, J., Kutner, M. H., Nachtsheim, C. J., & Wasserman, W. (1996). *Applied linear statistical models*. McGraw-Hill Inc.

Montgomery, D. C. & Peck, E. A. (1992). *Introduction to linear regression analysis (2nd ed.)*. John Wiley: New York.

Olive, D. J. (2005). Two simple resistant regression estimators. *Computational Statistics and Data Analysis*, To Appear.

Rousseeuw, P. J. & Leroy, A. (1987). *Robust regression outlier detection*. John Wiley: New York.

Theil, H. (1950). A rank-invariant methods of linear and polynomial regression analysis. *Indagationes Mathematicae, 12*(85), 173.

Journal of Modern Applied Statistical Methods
November, 2005, Vol. 4, No. 2, 514-521

1538 – 9472/05/$95.00

# Comparison Of Statistical Tests In Logistic Regression: The Case Of Hypernatreamia

Stylianos Katsaragakis
University of Athens

Christos Koukouvinos Stella Stylianou Eleni-Maria Theodoraki
National Technical University of Athens

The logistic regression has become an integral component of any medical data analysis concerning binary responses. The main issue rising after the adaptation of the final model is its goodness-of-fit. The fit of the model is assessed via the overall measures and summary statistics and comparing them in the case of hypernateamia.

Key words: Logistic regression, goodness-of-fit, covariates

## Introduction

The use of overall summary measures of goodness-of-fit has become an important and easily performed step in building logistic regression models. Pearson chi-square sum-of-squares statistics and the Score test are recommended due to their superior power in the simulations, but one must keep in mind that in small sample cases there is lack of detecting subtle deviations from the model (Hosmer, 1997). When it comes to sparse data, a non-significant result of a goodness-of-fit test does not tell that the model is correct, it just tells that the lack-of-fit is not large enough for the model to be rejected (Kuss, 2002).

In general, there are two different approaches to assessing goodness-of-fit in logistic regression models (e.g., Cook, 1979; Pregibon, 1981). The first one, residual analysis, investigates the model on the level of individuals and looks for those observations which are not adequately described by the model or which are highly influential on the model fit. The second approach seeks to combine the information on the amount of lack-of-fit in a single number. Statistical tests, so-called goodness-of-fit tests, are then calculated to judge if this lack-of-fit is significant or due to random chance and can be distinguished to specific and global. Global tests do not evaluate specific alternatives, rather test unspecific hypotheses of the form 'the model fits' versus the alternative 'the model does not fit'.

The goal is to investigate the choice of statistic test for assessing the coefficients of parameters as well as the goodness of fit by examining the medical disorder called hypernatreamia. For this purpose, three well known statistic tests will be used: the Likelihood Ratio statistic (LR), the Wald test (W) and the Score test (Scr) (Hosmer, 1989), although some authors warn that for large coefficients, standard error is inflated, lowering the Wald statistic (chi-square) value (Hosmer, 1989) and the likelihood-ratio test is more reliable for small sample sizes than the Wald test (Argesti, 1996). Methods for checking goodness-of-fit, are less developed, which may be due to the relative youth and enhanced mathematical complexity of the logistic regression model compared to, for example, the linear regression model (e.g., Bendel, 1977; Cook, 1977).

The study includes 314 patients treated at the Surgery Intensive Care Unit of a central hospital in Athens during 1996 - 2003. All data have been extracted from the Central Data Base of the Unit in which are recorded all demographic information (ID, age, sex, disease,

Stylianos Katsaragakis is an Associate Professor in the First Propedeutic Clinic of Surgery, Ippokratio Hospital, Athens, Greece. Christos Koukouvinos is a Professor in the Department of Mathematics, Zografou 15773, Athens, Greece. Email: ckoukouv@math.ntua.gr. Dr. Stella Stylianou is at the Department of Mathematics, Zografou 15773, Athens, Greece. Eleni-Maria Theodoraki is a postgraduate student in Biostatistics.

APACHE II score), daily biochemical indication and medical treatment and mortality. These patients have been chosen, excluding some from the 364 recorded, due to their staying in the ICU less than 3 days, which is thought to be a cutpoint for the ones who enter only for after surgery treatment. In addition, the patients under examination have not been transported to other hospital in order to be aware of the final condition of their health.

To compare the groups of patients having expressed the disorder hypernatreamia, with a control group, there were 35 patients from the first one with at least one indication of the electrolyte Na >147mmol/l during their staying in the ICU and 279 from the second group. With the aim of studying their behaviour, possible risk factors, sepsis criteria, Apache II score, medical treatment and mortality were examined.

In this article, the case of hypernatreamia with a multiple logistic regression model is considered.

The Logistic Regression Model

Logistic regression is part of generalized linear models (McCullagh, 1983), which allows one to predict a discrete outcome, from a set of variables that may be continuous, discrete, dichotomous, or a mix of any of these. Dichotomous (binary) outcome is the most common situation in biology and epidemiology, standing for the presence or absence of a disease, success or failure etc. Although discriminant analysis may also predict group membership (e.g., Costanza, 1979; Efron, 1975), it can be used only with two groups, so in the cases of categorical, or a mix of continuous and categorical covariates, logistic regression is preferred (e.g., Cook, 1979; Fleiss, 1979; Furnival, 1974; Mickey, 1989).

What seems to distinguish logistic regression to linear is conditional mean $E(Y/x)$, the mean value of the outcome variable, given the value of the independent variable. In linear regression, it is assumed that this mean may be expressed as an equation linear in x, which implies that $E(Y/x)$ may take any value as x ranges between $-\infty$ and $+\infty$, but with dichotomous data conditional mean must be greater than or equal to zero and less than or greater to one. The second important difference concerns the conditional distribution of the outcome variable. In the linear regression model, it is assumed that an observation of the outcome variable may be expressed as $y = E(Y/x) + \varepsilon$, where the error ε follows a normal distribution [$\varepsilon \sim N(\mu, \sigma^2)$], whereas in logistic ε follows the binomial one.

Logistic regression makes no assumption about the distribution of the independent or predictor variables, that is they do not have to be normally distributed (Lawless, 1978), linearly related or of equal variance within each group so the relationship between the predictor and response variables is not a linear function.

Let $f(x) = P(Y = 1/\vec{x})$, where the vector

$$\vec{x} = (x_1, x_2, \ldots x_p)$$

denotes a collection of $p$ covariates. Then the logistic regression function, in form of the logit transformation

$$g(\vec{x}) = \ln\left[\frac{f(x)}{1-f(x)}\right] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_p x_p$$

is:

$$f(x) = \frac{e^{g(\vec{x})}}{1+e^{g(\vec{x})}}.$$

During model creation, variables can be entered into the model in the order specified by the researcher or logistic regression can test the fit of the model after each coefficient is added or deleted, called stepwise regression. Stepwise regression is used in the exploratory phase of research but it is not recommended for theory testing. Forward variable selection enters the variables in the block one at a time based on entry criteria and backward stepwise regression appears to be a preferred method of exploratory analysis, where the analysis begins with a full or saturated model and variables are eliminated from the model in an iterative process.

Backward selection is sometimes less successful than forward or stepwise selection because the full model fit in the first step is the

model most likely to result in a complete or quasi-complete separation of response values. The fit of the model is tested after the elimination of each variable to ensure that the model still adequately fits the data. When no more variables can be eliminated from the model, the analysis has been completed. The process by which coefficients are tested for significance for inclusion or elimination from the model involves several different techniques (e.g., Bendel, 1977; Costanza, 1979). Some of these tests are described in the next section.

Assessment of the Coefficients of the Model

A Wald test is used to test the statistical significance of each coefficient $\beta_i$ in the model. A Wald test calculates a *z* statistic, which is:

$$z = \frac{\beta_i}{SE(\beta_i)}.$$

This z value is then squared, yielding a Wald statistic with a chi-square distribution with p+1 degrees of freedom, where p is the number of covariates. The likelihood-ratio test uses the ratio of the maximized value of the likelihood function for the saturated model ($L_1$) over the maximized value of the likelihood function for the current model ($L_0$). The likelihood-ratio test statistic equals:

$$-2\log\left(\frac{L_0}{L_1}\right) = -2[\log(L_0) - \log(L_1)] = -2(L_0 - L_1).$$

This log transformation of the likelihood functions yields a chi-squared statistic with p degrees of freedom equal to the number of covariates of the model. This appears to be the recommended test statistic to use, when building a model through backward stepwise elimination.

The score statistic is a quadratic form based on the vector of partial derivatives of the log-likelihood function with respect to the parameters of interest, evaluated at the values postulated by the null hypothesis.

Let

$$L(\beta|Y) = \prod_{i\in S} P_i^{w_i Y_i}(1-P_i)^{w_i(1-Y_i)} - \prod_{i\in S}\left(\frac{P_i}{1-P_i}\right)^{w_i Y_i}(1-P_i)^{w_i}$$

be the weighted likelihood function and

$$\log_e L(\beta|Y) = \sum_{i\in S}\left\{w_i \log_e\left(\frac{P_i}{1-P_i}\right) + w_i \log_e(1-P_i)\right\}$$
$$= \sum_{i\in S} w_i Y_i X_i^T \beta - \sum_{i\in} w_i \log_e(1+e^{X_i^T\beta})$$

be the log likelihood function. Then, the (p + 1) x 1 score vector, S(β), is given by

$$S(\beta) = \frac{\partial}{\partial\beta}\log_e L(\beta|Y) = \sum_{i\in S} w_i X_i^T (Y_i - P_i)$$

Testing the Fit of the Model

For a particular covariate pattern, the Pearson residual is defined as follows:

$$r(y_i, \hat{\pi}_j) = \frac{(y_j - m_j\hat{\pi}_j)}{\sqrt{m_j\hat{\pi}_j(1-\hat{\pi}_j)}}$$

The summary statistic based on these residuals is the Pearson chi-square statistic

$$X^2 = \sum_{j=1}^{J} r(y_j, \hat{\pi}_j)^2$$

and the deviance residual:

$$d(y_j, \hat{\pi}_j) = \pm\left\{2\left[y_j \ln\left(\frac{y_j}{m_j\hat{\pi}_j}\right) + (m_j - y_j)\ln\left(\frac{(m_j - y_j)}{m_j(1-\hat{\pi}_j)}\right)\right]\right\}^{1/2}$$

The distribution of the statistics $X^2$ and D under the assumption that the fitted model is correct in all aspects is supposed to be chi-square with degrees of freedom equal to J-p-1.

The Hosmer-Lemeshow goodness-of-fit statistic is obtained by calculating the Pearson chi-square statistic from the $2\times g$ table of observed and expected frequencies, where $g$ is the number of groups. The statistic is written as:

$$\chi^2_{HL} = \sum_{i=1}^{g} \frac{\left(O_i - N_i\overline{\pi}_i\right)^2}{N_i\overline{\pi}_i\left(1-\overline{\pi}_i\right)}$$

where $N_i$ is the total frequency of subjects in the $i$th group, $O_i$ is the total frequency of event outcomes in the $i$th group, and $\overline{\pi}_i$ is the average estimated probability of an event outcome for the $i$th group. The Hosmer-Lemeshow statistic is then compared to a chi-square distribution with ($g$-$n$) degrees of freedom, where the value of $n$ can be specified in the lackfit option in the model statement. The default is $n$=2. Large values of $X^2_{HL}$ (and small $p$-values) indicate a lack of fit of the model.

Comparison of the Coefficients-Results

The data set used to compare the statistical tests contains 24 covariates for each of the two groups of patients under examination (hypernatreamic-control patients). At a brief description it is observed that both groups have statistically comparable ages ($t_{290,\ 0.025}$=-0.753, p=0.452), the sepsis score ($X^2_4(0.05)$=6.979, p=0.137) as well as the **A**cute **P**hysiology **A**nd **C**hronic **H**ealth **E**valuation, ($X^2_1(0.05)_{Kruskall\ \ Wailes}$ = 1.174, p = 0.279), which both estimate the condition of health of each patient at his entrance in the ICU, does not seem to differentiate between two groups.

It is of interest now to explore the relationship between the covariates and the presence or absence of hypernatreamia. Using a univariate model containing the intercept and every time the variable of interest, it seems to exist strong relationships with the binary outcome indicating that patients with high values of Na differentiate from the control group. But can this univariate result be used to confirm, for example, that hypernateamia is associated with mortality - taking under consideration all possible risk factors? That is one of the questions generated and concerns a set of covariates that can be partly answered with a multivariable logistic regression analysis.

For this purpose, variables are included in the model that has been shown to be associated with hypernatreamia. Covariates of interest included age, gender, evaluation of the stage of the patients condition (APACHE, sepsis score), resuscitation fluids and antibiotics containing Na. The multivariate logistic regression model also included the interactions of plasma (FFP) with the antibiotics containing furosemide, teicoplanin and humanxlasix to examine if their combination is mischievous, that is they lead to hypernatreamia.

The analysis was conducted with the SAS program and the method used for the binary model was the full one. 31 observations were deleted due to missing values for the explanatory variables so the number of observations that finally contributed to the analysis was 283 (30 patients who expressed the disorder and 253 control patients). The importance of a variable is defined in terms of a measure of the statistical significance of the coefficient of the model ($p<0.05$), which denotes the fixed decision rule for the inclusion of variables at the procedure used. However there seems to be an indication of the influential role for some covariates ($p<0.10$) that needs to be taken under consideration and are therefore illustrated.

The results for the logistic regression model to be assessed are presented in table 1. Initially the model contained all the possible interaction factors, which have already been discussed, with no statistically significant results; therefore only the main effects were used. With the exception of the design variable sepsis, there is clear evidence that each of the variables has some association with the outcome. This observation is based on an inspection of the 95% Wald confidence interval estimates which, either do not contain 1 or just barely do. At this point, a decision concerning the variable age had to be made, as it is known to be a biologically important variable, yet is not statistically significant in this model. For this reason the covariate's estimate and the Wald test's value at the Analysis of Maximum Likelihood Estimates table were included. In search of a confounding effect, it was found that

Table 1: Analysis of Maximum Likelihood Estimates

| Parameter | DF | Estimate | Standard Error | Wald Chi-Square | Pr>ChiSq |
|---|---|---|---|---|---|
| Intercept | 1 | -52.186 | 353.700 | 0.022 | 0.883 |
| APACHE | 1 | 0.121 | 0.073 | 2.748 | 0.097 |
| daysofst | 1 | 2.356 | 0.624 | 14.245 | 0.000 |
| age | 1 | 0.035 | 0.037 | 0.884 | 0.347 |
| qfurosemide | 1 | -0.145 | 0.050 | 8.462 | 0.004 |
| qffp | 1 | -0.590 | 0.253 | 5.427 | 0.020 |
| qimipeneme | 1 | 0.844 | 0.292 | 8.386 | 0.004 |
| qteicoplanin | 1 | 1.024 | 0.527 | 3.776 | 0.052 |
| qsod. hlopideamp 15% | 1 | -0.389 | 0.109 | 12.877 | 0.000 |
| sex (0) | 1 | 1.177 | 0.597 | 3.887 | 0.049 |
| death (0) | 1 | -3.782 | 1.068 | 12.549 | 0.000 |
| sepsis (0) | 1 | 15.483 | 8.240 | 3.531 | 0.060 |
| sepsis (1) | 1 | 14.758 | 8.298 | 3.163 | 0.075 |
| sepsis (2) | 1 | 12.958 | 7.949 | 2.658 | 0.103 |
| sepsis (3) | 1 | 15.469 | 8.276 | 3.494 | 0.062 |
| ffp (0) | 1 | -1.099 | 0.630 | 3.043 | 0.081 |
| imipeneme (0) | 1 | -3.514 | 1.646 | 4.559 | 0.033 |
| teicoplanin (0) | 1 | -16.705 | 6.381 | 6.854 | 0.000 |

Table 2: Odds Ratio Estimates

| Effect | Point Estimate | 95% Wald Confidence Limits | |
|---|---|---|---|
| APACHE | 0.886 | 0.767 | 1.022 |
| daysofst | 0.095 | 0.028 | 0.322 |
| age | 1.035 | 0.963 | 1.114 |
| qfurosemide | 1.156 | 1.049 | 1.275 |
| qffp | 1.804 | 1.098 | 2.963 |
| qimipeneme | 0.430 | 0.243 | 0.761 |
| qsod. Chlopideamp 15% | 0.359 | 0.128 | 1.009 |
| sex (0 vs 1) | 0.095 | 0.009 | 0.986 |
| death (0 vs 1) | >999.999 | 29.340 | >999.999 |
| sepsis (0 vs 4) | <0.001 | <0.001 | 290.589 |
| sepsis (1 vs 4) | <0.001 | <0.001 | 689.112 |
| sepsis (2 vs 4) | <0.001 | <0.001 | >999.999 |
| sepsis (3 vs 4) | <0.001 | <0.001 | 337.138 |
| ffp (0 vs 1) | 9.006 | 0.762 | 106.412 |
| imipeneme (0 vs 1) | <0.001 | <0.001 | 0.562 |
| teicoplanin (0 vs 1) | <0.001 | <0.001 | <0.001 |

the absence of age indeed acts as a confounder changing remarkably the significance status of the model. Assessing the reduced model for that case, the LR and Score Tests

$$(X^2_{26}(0.05)_{(LR)(f-age)} = 126.486,$$

$$X^2_{26}(0.05)_{(Scr)(f-age)} = 123.824,\ p<0.0001)$$

agrees with the saturated one

$$(X^2_{27}(0.05)_{(LR)f} = 141.465,\ X^2_{27}(0.05)_{(Scr)f} = 120.634,\ p<0.0001)$$

and there is a small change

$$(X^2_{277}(0.05)_{(Pearson)} = 217.715\ (p=0.997),$$

$$X^2_8(0.05)(HL) = 3.322,\ (p=0.913)$$

in the Pearson and Hosmer-Lemeshow goodness-of-fit tests

$$(X^2_{255}(0.05)_{(Pearson)} = 128.107\ (p=1.000),$$

$$X^2_8(0.05)(HL) = 2.333,\ p=0.969)$$

reflecting the reduction of effectiveness in describing the outcome due to the absence of age.

Examining the results, it was also observed that the estimated coefficients for a set of variables in the model changed significantly when gender was deleted. Hence, there is clear evidence of a confounding effect due to gender describing that it is associated with both the outcome variable of interest, hypernatreamia, and the risk factors. Comparing the LR and Score tests of that model with the full one, it was found that although the LR and Score tests don't seem to denote that the absence of the variable produces an alteration in the model

$$(X^2_{26}(0.05)_{(LR)(f-gender)} = 136.777,$$

$$X^2_{26}(0.05)_{(Scr)\,gender} = 120.05,$$

$$p<0.0001,\ X^2_{27}(0.05)_{(LR)f} = 141.465,$$

$$X^2_{27}(0.05)_{(Scr)f} = 120.634, p<0.0001),$$

the goodness-of-fit statistics seem to ascertain a small one

$$(X^2_{256}(0.05)_{(Pearson)(f-gender)} = 194.389$$

$$(p=0.998),\ X^2_8(0.05)_{(HL)(f-gender)} = 2.127$$

$$(p=0.977),\ X^2_{255}(0.05)_{(Pearson)f} = 128.107$$

$$(p=1.000),\ X^2_8(0.05)_{(HL)f} = 2.334 = 0.969).$$

The confounding status of sepsis score has also been examined, confirming that it is interactively associated with both the disorder and the covariates. The results of the comparison are very interesting since the absence of the polytomous covariate sepsis score produces remarkable changes to the model fit. In specific, although the saturated model seems to fit well, the null hypothesis for the reduced model is rejected

$$(X^2_{259}(0.05)_{(Pearson)(f-sepsis)} = 591.935$$

$$(p<0.001),\ X^2_8(H-L)_f = 20.167\ (p=0.0097)).$$

Considering that the overall goal is to obtain the best fitting model while minimizing the number of parameters, the next step is to fit a reduced model containing only those variables thought to be significant, and compare it to the full model containing all the variables. The results fitting a model with intercepts only and for fitting a model with intercepts and explanatory variables, show that the overall statistic tests reject the global null hypothesis

BETA=0 in the case of both the reduced and the full model.

$( X^2_7(0.05)_{(LR)_r}$ =65.395, $X^2_7(0.05)_{(Scr)_r}$ =94.37

7,p<0.0001) $X^2_{27}(0.05)_{(LR)_f}$ =141.465,

$X^2_{27}(0.05)_{(Scr)_f}$ =120.634, p<0.0001).

However examining the Pearson and Hosmer-Lemeshow statistics

$( X^2_8(0.05)_{(HL)}$ =17.756 (p=0.023),

$X^2_{278}(0.05)_{(Pearson)}$ =1316.375 (p<0.0001)

a remarkable change demonstrating a better fit of the full model is observed

$( X^2_8(0.05)_{(HL)}$ =128.107 p=1.000,

$X^2_{278}(0.05)_{(Pearson)}$ =2.333, p=0.969)).

During model assessment, it was observed that deviance does not seem to alter

$( X^2_{255}(0.05)_{(Deviance)_f} = 49.891$

(p=1.000), $X^2_{277}(0.05)_{(Deviance)_{(f-age)}} = 78.103$(p

=1.000), $X^2_{256}(0.05)_{(Deviance)_{(f-gender)}} = 54.58$

(p=1.000)),

placing all models containing confounders or other reduced models in the same goodness-of-fit status with the full model. That happens even in the last case of the confounding of sepsis score when Pearson and Hosmer-Lemeshow tests agree in rejecting the goodness-of-fit but deviance fails to identify such alteration

$( X^2_{255}(0.05)_{(Deviance)_{(f-sepsis)}} = 88.531$, p=1.000).

The estimated coefficients and odds ratio show that women are 10.6 times more likely to express the disorder (p<0.05) than men, mortality increases to hypernatriemic patients (p<0.01) and the ones with sepsis score 4 are much less likely to get hypernatreamic compared to any of the other 3 sepsis levels (0, 1, 2, 3). In the case of the design variables of sepsis, although between levels 2 and 4 there seems to be a marginal relationship at the 10% level **(**p=0.103), the variable was included because the W statistics for all relative coefficients exceed 2 (Hosmer & Lemeshow, 1989).

There is great interest to the influential part that the antibiotics and resuscitation fluids containing Na, play during patients treatment in ICU. Especially, patients that were treated intravenously with furosemide increased the risk of getting hypernatriemic 15% every time they accepted 20mg as long as getting FFP they increased the risk 9 times from those who didn't (an increase of 1 point led to a 80% increase of risk).

## Conclusion

During or after model creation, there seems to be efficiency and applicability of the proposed Wald Test, Likelihood Ratio Test, and Score test, because they agree in refining the significance of the coefficients. Our comparison of the proposed goodness-of-fit statistics Pearson chi-square and Hosmer-Lemeshow, showed small deviations between them at the omission of important confounders, but both are much more powerful from deviance in detecting the fit of the model. That leads to an important association between the behaviour of the logistic regression model through the application of different assessment statistics, in representing best the biological mechanism, hence correctly logistic regression is a significant tool in any medical data analysis of an ordinal response model with both categorical and continuous covariates.

References

Argesti, A. (1996). *An introduction to categorical data analysis*. Wiley.

Bendel, R. B., & Afifi, A. (1977). Comparison of stopping rules in forward regression. *Journal of the American Statistical Association, 72*, 46-53.

Cook, R. D. (1977). Detection of influential observations in linear regression. *Technometrics*, *19*, 15-18.

Cook, R. D. (1979). Influential observations in linear regression. *Journal of the American Statistical Association*, *74*, 169-174.

Costanza, M. C., & Afifi, A. (1979). Comparison of stopping rules in forward stepwise discriminant analysis. *Journal of the American Statistical Association*, *74*, 777-785.

Efron, B. (1975). The efficiency of logistic regression compared to normal discriminant function analysis. *Journal of the American Statistical Association*, *70*, 892-898.

Fleiss, J. (1979). Confidence intervals for the odds ratio in case control studies: State of the art. *Journal of Chronic Diseases*, *32*, 69-77.

Furnival, G. M., & Wilson, R. W. (1974). Regression by leaps and bounds. *Technometrics*, *16,* 499-511.

Hauck, W. W. (1985). A comparison of the logistic risk function and the proportional hazards model in prospective epidemiologic studies. *Journal of Chronic Diseases*, *38*, 125-126.

Hosmer, D. W., Hosmer, T., Cessie , S. Le, & Lemeshow, S. (1997). A comparison of goodness-of-fit tests for logistic regression model. *Statistics in Medicine*, *16*, 965-980.

Hosmer, D. W., & Lemeshow, S. (1989). *Applied logistic regression*, Wiley.

Kuss, O. (2002). Global goodness-of-fit tests in logistic regression with sparse data. *Statistics in Medicine*, *21*, 3789-3801.

Lawless, J. F., & Singhal, K. (1978). Efficient screening of non-normal regression models. *Biometrics*, *34*, 318-327.

McCullagh, P., & Nelder, J. A. (1983). *Generalized linear models*. Chapman Hall: London.

Mickey, J., & Greenland, S. (1989). A study of the impact of confounder - selection criteria on effect estimation. *American Journal of Epidemiology*, *129*, 125-137.

Pregibon, D. (1981). Logistic regression diagnostics. *Annals of Statistics*, *9*, 705-724.

Pulkstenis, E., & Robinson, T. J. (2002). Two goodness of fit tests for logistic regression models with continuous variables. *Statistics in Medicine*, *21*, 79-83.

Journal of Modern Applied Statistical Methods
November, 2005, Vol. 4, No. 2, 522-527

1538 – 9472/05/$95.00

# Simulation Procedure In Periodic Cancer Screening Trials

Dongfeng Wu Xiaoqin Wu
Department of Mathematics and Statistics
Mississippi State University

Ioana Banicescu
Department of Computer Science and Engineering
Mississippi State University

Ricolindo L. Cariño
Center for Computational Sciences ERC
Mississippi State University

A general simulation procedure is described to validate model fitting algorithms for complex likelihood functions that are utilized in periodic cancer screening trials. Although screening programs have existed for a few decades, there are still many unsolved problems, such as how age or hormone affects the screening sensitivity, the sojourn time in the preclinical state, and the transition probability from disease-free state to the preclinical state. Simulations are needed to check reliability or validity of the likelihood function combined with the associated effect functions. One bottleneck in the simulation procedure is the very time consuming calculations of the maximum likelihood estimates (MLE) from generated data. A practical procedure is presented, along with results for when both sensitivity and transition probability into the preclinical state are age-dependent. The procedure is also suitable for other applications.

Key words: periodic screening, breast cancer, early detection, sensitivity, sojourn time, transition probability, mammogram, clinical breast examination, incidence

## Introduction

According to a recent report of the National Institute of Health (NIH 2000), breast cancer is the most common form of cancer among women in the United States and the second leading cause of cancer deaths among women. One of the procedures to manage the disease is periodic cancer screening, which has been utilized for a few decades. The motivation for screening is to detect the disease early even before clinical symptoms come up. The benefit for early detection is obvious. People in whom cancer is detected earlier usually have a better prognosis. Early treatments hopefully will lead to more cure and prolonged survival of cancer patients.

In a screening program, a large group of asymptomatic individuals are enrolled in the program to detect the presence of a specific disease. The natural history of the disease for an individual is assumed to follow a progressive stochastic model, which consists of three states, denoted by $S_0 \to S_p \to S_c$, corresponding, respectively, to the disease-free state; the preclinical disease state, in which an asymptomatic individual unknowingly has disease that the screening exam can detect; and the clinical state when the disease manifests itself in clinical symptoms. The screening sensitivity is the probability that the screening exam is positive, given that the individual is in the preclinical stage. The sojourn time refers to the time beginning when the disease first

Dongfeng Wu is an Assistant Professor, with research interests in cancer screening probability modeling and inference. She is on the Editorial Board of *JMASM*. Xiaoqin Wu is a PhD candidate. His research interest is in PDE modoling and statistical modeling. Ioana Banicescu is an Associate Professor, with research interests in parallel algorithms, scientific computing, scheduling theory, optimization and prediction. Ricolindo L. Carino received his Ph.D. from La Trobe University, and is a member of the research faculty. His main interest is parallel computing for scientific applications.

develops until the manifestation of clinical symptoms, that is $(S_c - S_p)$. The transition probability into the preclinical stage is the probability density function of making transition from the disease-free to the preclinical state. Knowledge of the sensitivity of the screening modality is necessary for evaluating the predictive performance of a screening exam. The screening sensitivity may depend on a variety of factors, including age, position, location and size of the tumor, and the experience of the radiologist, etc. For example, recent studies indicate that the sensitivity of mammography increases with age at diagnosis (Shapiro, et. al., 1988; Miller, et. al., 1992a, 1992b), attributable to the fact that breast tissue tends to be more dense and fibrous in younger women, and more soft and fatty in older women (Kerlikowske, et. al., 1996).

There is great interest in determining the properties of the sensitivity, the sojourn time distribution and the transition probability density function into the preclinical state. Much work has been done in this area (Shen & Zelen, 1999; Shen, et. al., 2001; Wu, et. al., 2005). The research is still ongoing because many researchers are trying to explore how age or hormone changes may affect the sensitivity, the sojourn time, and the transition probability. One of the common features in the research is to derive the correct likelihood function and to propose correct age effect (or hormone effect) functions based on the stochastic model and the screening data. However, it is imperative to validate the reliability of the likelihood function and the associated effect functions before these can be applied to real data. This validation may be accomplished through simulation, which has become an acceptable procedure to check that the model fitting and the complex algorithms work well with this complicated likelihood.

The remainder of the article is organized as follows. A generalized stochastic model and its likelihood function in a periodic cancer screening program is introduced, as well as the age-dependent sensitivity and transition probability density. The simulation procedure, the corresponding algorithm and results of applying it to a sample scenario are then presented. It will conclude with a discussion of the results of the research.

## The Model

Consider a cohort of initially asymptomatic individuals who enroll in a screening program. The sensitivity is denoted by $\beta(t)$, where $t$ is the individual's age at the screening exam. Define $w(t)dt$ as the probability of a transition from $S_0$ to $S_p$ during $(t, t+dt)$. Let $q(t)$ be the probability density function of the sojourn time in $S_p$. Finally, let $Q(z) = \int_z^\infty q(x)dx$, that is, $Q(z)$ is the survivor function of the sojourn time in the preclinical state $S_p$. Throughout this article, the time variable $t$ represents the participating individual's age. If random variables $T$ and $S$ are the duration times in $S_0$ and $S_p$ respectively, then an individual will enter the clinical state $S_c$ at age $T+S$, the probability density function of $T+S$ is

$$I(t) = \int_0^t w(x)q(t-x)dx,$$

which is the observable incidence of clinical cases.

Consider a cohort of women in the study group who are all aged $t_0$ at study entry, and a protocol calls for $K$ ordered screening examinations occur at ages $t_0 < t_1 < \cdots < t_{K-1}$, where $t_i = t_0 + i$ for annual screening exams. Define the $i$-th screening interval as the time interval between the $i$-th and the $(i+1)$-th screening exams $(t_{i-1}, t_i)$, $i=1,2,\ldots, K-1$. The $i$-th generation of individuals consists of those who enter $S_p$ during this interval. The $0$-th generation includes all who enter $S_p$ before the initial screening exam; let $t_{-1} \equiv 0$.

For each screening exam, let $n_{i,t_0}$ be the total number of individuals in this cohort examined at the $i$-th screening; $s_{i,t_0}$ is the number of cases detected at the $i$-th screening exam; and $r_{i,t_0}$ is the number of cases diagnosed in the clinical state $S_c$ within the interval $(t_{i-1}, t_i)$. The latter cases are called interval cases.

Let $D_{k,t_0}$ be the probability that an individual will be diagnosed at the *k*-th scheduled exam (at which her age is $t_{k-1} = t_0 + k - 1$) given that she is already in the preclinical state. Let $I_{k,t_0}$ be the probability of being incident in the *k*-th screening interval. In Wu, et. al., 2005, these two probabilities were derived as:

$$D_{k,t_0} = \beta(t_{k-1})\{\sum_{i=0}^{k-2}[1-\beta(t_i)]\cdots[1-\beta(t_{k-2})]$$
$$\int_{t_{i-1}}^{t_i} w(x)Q(t_{k-1}-x)dx + \int_{t_{k-2}}^{t_{k-1}} w(x)Q(t_{k-1}-x)dx\}.$$

$$I_{k,t_0} = \sum_{i=0}^{k-1}[1-\beta(t_i)]\cdots[1-\beta(t_{k-1})]$$
$$\int_{t_{i-1}}^{t_i} w(x)[Q(t_{k-1}-x)-Q(t_k-x)]dx$$
$$+\int_{t_{k-1}}^{t_k} w(x)[1-Q(t_k-x)]dx.$$

The likelihood function for this cohort of women is

$$L(\cdot\,|\,t_0) = \prod_{k=1}^{K} D_{k,t_0}^{s_{k,t_0}} I_{k,t_0}^{r_{k,t_0}} (1-D_{k,t_0}^{s_{k,t_0}} - I_{k,t_0}^{r_{k,t_0}})^{n_{k,t_0}-s_{k,t_0}-r_{k,t_0}} \quad (1)$$

The full likelihood for the study group across all ages is

$$L = \prod_{t_0}\prod_{k=1}^{K} D_{k,t_0}^{s_{k,t_0}} I_{k,t_0}^{r_{k,t_0}} (1-D_{k,t_0}^{s_{k,t_0}} - I_{k,t_0}^{r_{k,t_0}})^{n_{k,t_0}-s_{k,t_0}-r_{k,t_0}} \quad (2)$$

The age effect was modeled in the sensitivity and the transition probability simultaneously in the following way. The sensitivity $\beta$ is associated with age *t* by a logistic link,

$$\beta(t) = \frac{1}{1+\exp(-b_0 - b_1 * (t-\bar{t}))},$$

Where $\bar{t}$ is the average age at entry in the whole study group. If $b_1 > 0, \beta(t)$ will be a monotone increasing function of age *t*.

The transition probability density function *w(t)* is the instantaneous probability of a transition from $S_0$ to $S_p$. The integral $\int_0^\infty w(t)dt$ represents a lifetime risk for a healthy female to transit into the preclinical state. According to the NCI's SEER database (Ries et al. 2002), a woman's lifetime risk of being diagnosed with breast cancer is 15.7%, which is less than a women's lifetime risk of entering the preclinical disease state. Hence, 20% was chosen as a reasonable upper bound. The following was chosen

$$w(t) = \frac{0.2}{\sigma t\sqrt{2\pi}}\exp\{-\frac{(\log t-\mu)^2}{2\sigma^2}\},$$

which is the pdf of lognormal($\mu$, $\sigma^2$) multiplied by 20%. That is, w(t) is a sub-density function, where $\mu$ and $\sigma^2$ are parameters to be estimated.

The loglogistic distribution was adopted to model the sojourn time in the preclinical state,

$$q(x) = \frac{\kappa x^{\kappa-1}\rho^{\kappa}}{[1+(x\rho)^{\kappa}]^2}, x > 0,$$

where *x* is the sojourn time, and $\kappa$ and $\rho$ are positive parameters, represent the scale and location in the loglogistic family. An advantage of this family over the exponential is that it has two parameters and is more robust in the tails. Another advantage of this family is that its relatively simple form achieved for the survivor function and the hazard function. Its first moment can be calculated directly from

$$EX = \frac{\pi}{\rho\kappa}\csc\left(\frac{\pi}{\kappa}\right).$$

For the *r*-th moment to exist, $\kappa > r$ is needed. For justifications on how these age effect functions are chosen, see Wu et. al., 2005.

Simulation Procedure and Results

The purpose of the simulation is to check the reliability of the likelihood function as

screening sensitivity and transition probability are both age-independent. The key steps were summarized in the non-routine simulation study here. In fact, based on the steps here, one can explore other possible associated functions between age and sensitivity, age and transition density, age and sojourn time, etc.

In the proposed model, there are six unknown parameters, that is, $\theta = (b_0, b_1, \mu, \sigma^2, \kappa, \rho)$. Theoretically the parameters have a domain of either $(-\infty, \infty)$ or $(0, \infty)$. The practical meaning of these parameters will limit them to a finite range. The range for each of them was identified as: $0 < b_0 < 5$, $-0.2 < b_1 < 0.2$, $3.5 < \mu < 4.5$, $0 < \sigma^2 < 1$, $0.1 < \rho < 2.0$, and $1 < \kappa < 5$. For justifications of these ranges, see Wu, et. al., 2005.

This simulation consisted of two stages. First, age-dependent screening data based on input values of $\theta = (b_0, b_1, \mu, \sigma^2, \kappa, \rho)$ were generated, assuming that initially there are about 100,000 individuals in each age group from age 40 to 64 who will take part in the periodic screening exams. For the input values of $\theta$, the values for $b_0, b_1, \mu, \sigma^2, \kappa$ and $\rho$ was randomly chosen from the valid range above. Second, the MLE $\hat{\theta}$ was computed from our likelihood function using the simulated data. This procedure was repeated n = 1,000 times, then the sample mean and the sample standard deviation of the MLE were collected, and were compared with the input values of $\theta$. If the MLE is close to the true input value of $\theta$, then our likelihood function and the age- dependent functions work well in the modeling.

Here are more details in Step 1: Suppose there are M= 100,000 women who were born in the same year, and who will take part in the screening exam at age $t_0$. Their duration time spent in the disease-free state ($S_0$) and in the preclinical state ($S_p$) can be generated by the density functions *w(t)* and *q(t)* correspondingly. Since w(t) is a sub-density function, it is not obvious how to generate random variables directly from its density. The number of incident cases from disease-free into preclinical state age by age will be generated, using the probability *w(t)dt* which is binomially distributed. Then, for women in the preclinical state at age *t*, their incident time can be generated uniformly in *(t, t+1)*. See Appendix for programming details.

For details in Step 2: The log likelihood function can be implemented in C language. Then, taking the negative value of the log likelihood and calling the S-PLUS routine "nlminb" will provide a local minimum. This local minimum corresponds to a local maximum in the log likelihood. However, computer software has not been found that can find the global minimum (maximum) for a general function. To overcome this problem, the initial point of $\theta$ was chosen randomly and the procedure was repeated 5 times for each simulated data and find the global maximum.

The simulation programming code, written in C++ and S-PLUS, is attached in the Appendix. It runs well in a PC environment. Eight simulation results are listed in Table 1. For each true value of $\theta$, the sample mean and sample standard error (S.E.) of the MLE of $\theta$ from 1000 simulations are listed. The consistency between the sample mean of the MLE and the input parameters is clearly shown.

## Conclusion

The purpose of this article is to provide a simulation procedure in periodic cancer screening trials, with the computer programming code in C++ and S-PLUS. A practical issue encountered in the simulation is that it is very time consuming when MLE was calculated from the simulated data. The procedure for each MLE calculation usually takes about 20 minutes if the code is written in S-PLUS, making it impractical to repeat the procedure for 1000 times. To decrease the computation time, the likelihood part was implemented in C++, which resulted in the whole 1000 simulation procedure finishing in two or three days. The simulation and programming code can be slightly modified to fit other age effect or hormone effect models as well. Hopefully this will help other researchers in this area to carry out their simulation studies.

Table 1. Summary of the simulation results for the six parameters

| | $b_0$ | $b_1$ | $\mu$ | $\sigma^2$ | $\kappa$ | $\rho$ |
|---|---|---|---|---|---|---|
| True value | 2.07 | -0.05 | 4.05 | 0.80 | 4.54 | 0.70 |
| MLE estimate | 2.073 | -0.051 | 4.053 | 0.799 | 4.525 | 0.698 |
| S.E. of MLE | 0.112 | 0.006 | 0.042 | 0.018 | 0.245 | 0.016 |
| True value | 0.91 | -0.07 | 4.24 | 0.51 | 3.01 | 0.74 |
| MLE estimate | 0.879 | -0.069 | 4.242 | 0.510 | 3.046 | 0.730 |
| S.E. of MLE | 0.093 | 0.004 | 0.019 | 0.015 | 0.150 | 0.029 |
| True value | 2.72 | -0.12 | 3.65 | 0.55 | 3.73 | 0.65 |
| MLE estimate | 2.714 | -0.120 | 3.652 | 0.551 | 3.750 | 0.647 |
| S.E. of MLE | 0.157 | 0.011 | 0.021 | 0.018 | 0.133 | 0.012 |
| True value | 3.14 | 0.12 | 4.42 | 0.86 | 1.16 | 1.23 |
| MLE estimate | 3.169 | 0.123 | 4.420 | 0.861 | 1.161 | 1.223 |
| S.E. of MLE | 0.308 | 0.029 | 0.024 | 0.034 | 0.015 | 0.025 |
| True value | 0.47 | -0.17 | 3.59 | 0.15 | 1.67 | 0.76 |
| MLE estimate | 0.475 | -0.170 | 3.591 | 0.150 | 1.667 | 0.752 |
| S.E. of MLE | 0.053 | 0.004 | 0.005 | 0.004 | 0.023 | 0.018 |
| True value | 1.64 | 0.02 | 3.93 | 0.08 | 2.37 | 1.05 |
| MLE estimate | 1.612 | 0.022 | 3.930 | 0.080 | 2.377 | 1.037 |
| S.E. of MLE | 0.150 | 0.004 | 0.003 | 0.001 | 0.054 | 0.037 |
| True value | 2.81 | 0.19 | 4.03 | 0.67 | 3.07 | 0.82 |
| MLE estimate | 2.710 | 0.181 | 4.029 | 0.670 | 3.094 | 0.812 |
| S.E. of MLE | 0.137 | 0.013 | 0.033 | 0.014 | 0.083 | 0.012 |
| True value | 3.74 | -0.04 | 4.36 | 0.72 | 2.74 | 0.81 |
| MLE estimate | 3.650 | -0.039 | 4.361 | 0.721 | 2.762 | 0.801 |
| S.E. of MLE | 0.538 | 0.030 | 0.024 | 0.027 | 0.075 | 0.021 |

For more details on how to combine C++ and S-PLUS code, see S-PLUS manual. Current efforts are in transporting this procedure to run on a cluster of Linux workstations. If this effort is successful, the simulation time will be shortened to a few hours.

References

Chen, T. H. H., Kuo, H. S., Yen, M. F., Lai, M. S., Tabar, L. & Duffy, S. W.(2000). Estimation of sojourn time in chronic disease screening without data on interval cases. *Biometrics 56*, 167-172.

Cox, D. R. & Oakes, D. (1984). *Analysis of survival data*. Chapman & Hall/CRC.

Day, N. E. & Walter, S. (1984). Simplified models of screening for chronic disease: Estimation procedures from mass screening programs. *Biometrics 40*, 1-13.

Eddy, D. M. (1980). *Screening for cancer: Theory, analysis, and design*. Englewood Cliffs, NJ: Prentice Hall.

Kerlikowske, K., Grady, D., Barclay, J., Sickles, E. A., & Ernster, V. (1996). Effect of age, breast density, and family history on the sensitivity of first screening mammography. *Journal of the American Medical Association 276*, 33-38.

Lee, S. J. & Zelen, M. (1998). Scheduling periodic examinations for the early detection of disease: Applications to breast cancer. *Journal of the American Statistical Association 93*, 1271-1281.

Miller, A. B., Baines C. J., To, T. & Wall, C. (1992a). Canadian national breast screening study: 1. Breast cancer detection and death rates among women aged 40 to 49 years. *Canadian Medical Association Journal 147*(10), 1459-76.

Miller, A. B., Baines, C.J., To, T. & Wall, C. (1992b). Canadian national breast screening study: 2. Breast cancer detection and death rates among women aged 50 to 59 years. *Canadian Medical Association Journal 147*(10), 1477-88.

National Institute of Health (2000). NIH Publication No. 00-1556, 12/12/2000.

Shapiro, S., Venet, W., Strax, P. & Venet L. (1988). *Periodic screening for breast cancer. The Health Insurance Plan Project and its Sequelae, 1963-1986.* Baltimore: The Johns Hopkins University Press.

Shen, Y., Wu, D. & Zelen, M. (2001). Testing the independence of two diagnostic tests. *Biometrics 57*, 1009-1017.

Shen, Y. & Zelen, M. (1999). Parametric estimation procedures for screening programmes: Stable and nonstable disease models for multimodality case finding. *Biometrika 86*, 503-515.

Straatman, H., Peer, P. G. M. & Verbeek, A. L. M. (1997). Estimating lead time and sensitivity in a screening program without estimating the incidence in the screened group. *Biometrics 53*, 217-229.

Walter, S. D. & Day, N. E. (1983). Estimation of the duration of a preclinical disease state using screening data. *American Journal of Epidemiology 118*, 856-86.

Wu, D., Rosner, G. & Broemeling, L. (2005). MLE and bayesian inference of age-dependent sensitivity and transition probability in periodic screening. *Biometrics, 61*, 1056-1063.

Zelen, M. (1993). Optimal scheduling of examinations for the early detection of disease. *Biometrika 80*, 279-93.

Journal of Modern Applied Statistical Methods
November, 2005, Vol. 4, No. 2, 528-537


1538 – 9472/05/$95.00

# Selection Of Independent Binary Features Using Probabilities: An Example From Veterinary Medicine

Ludmila I. Kuncheva  Zoë S.J. Hoare
School of Informatics
University of Wales, Bangor, UK

Peter D. Cockcroft
Department of Clinical Veterinary Medicine
University of Cambridge, UK

Supervised classification into $c$ mutually exclusive classes based on $n$ binary features is considered. The only information available is an $n \times c$ table with probabilities. Knowing that the best $d$ features are not the $d$ best, simulations were run for 4 feature selection methods and an application to diagnosing BSE in cattle and Scrapie in sheep is presented.

Key words: Feature selection, classification, independent features, binary features, veterinary medicine.

## Introduction

Consider the differential diagnosis of BSE in cattle based on the probabilistic description of BSE and 56 alternative diseases with similar symptoms. There are many possible disease-related signs that may be observed as present/absent on an animal. For example, over 240 signs related to BSE and the 56 other diagnoses can be listed (Brightling et al., 1996; White, 1984). To build a diagnostic system, a data set is needed with observations for a number of cattle with their verified diagnoses. In the lack of such a data set, one must rely on estimates of the individual class-conditional probabilities that sign $x_i$ is present, given disease $\omega_j$, where $i \in \{1,2,...,n\}$ and $j \in \{1,2,...,c\}$.

Ludmila Kuncheva is a Senior Lecturer. Her interests include pattern recognition and classifier ensembles. Email her at: l.i.kuncheva@bangor.ac.uk. Zoe Hoare studied mathematics at the University of Wales, Bangor, UK. She is a doctoral student in the area of pattern recognition and its application to veterinary medicine. Email her at z.s.hoare@bangor.ac.uk. Peter Cockroft is a Senior Lecturer, member of the Royal College of Veterinary Surgeons (RCVS), and a holder of the RCVS's Diploma in Cattle Health and Production. Email him at pdc24@hermes.cam.ac.uk.

The information available in this problem is organized as shown in Table 1.

Table 1. Class-conditional probabilities for the individual features (the only information available)

| | $\omega_1$ | ... $\omega_i$ ... | $\omega_c$ |
|---|---|---|---|
| $x_1$ | | ... | |
| ... | | | |
| $x_k$ | ... | $P(x_k = 1 \mid \omega_i)$ | ... |
| ... | | | |
| $x_n$ | | ... | |

It is unrealistic to expect that a system based on these probabilities will fare well in practice because no relationship between the diagnostic signs (features) has been taken into account. In an ideal scenario, a data set will be collected using all features and the relationships between the features will be estimated from it. In reality, measuring only a small number of relevant features may be feasible.

The goal is to select $d$ features ($d < n$), which form a subset with the smallest classification error. Denote by $\mathbf{x}$ the binary vector with the $n$ features. The features are assumed to be conditionally independent, that is,

$$P(\mathbf{x} \mid \omega_j) = \prod_{i=1}^{n} P(x_i \mid \omega_j) \qquad (1)$$

The assumption of independence is enforced upon this study because only (some estimates of) the individual class-conditional probabilities are available. Pattern recognition literature in the 1970s abounds with analyses of the case of independent binary features. Perhaps the most curious result is due to Toussaint (1971). If there are three independent binary features, the best combination of two features may not include the single best feature. Thus, the most desirable selection criterion – the probability of error – will not guarantee the optimal solution if applied in a stepwise manner as in stepwise linear regression.

In this article, four procedures for selecting a subset of features are examined and the results are compared with those obtained with the whole feature set. The feature selection methods are illustrated on two problems taken from veterinary medicine: differential diagnosis of BSE in cattle and Scrapie in sheep.

## Methodology

Feature selection is one of the oldest topics in pattern recognition and machine learning (Stearns, 1976; Van Campenhout, 1982; Jain and Chandrasekaran, 1982; Patrick, 1972). Surveys on more recent state-of-the-art and comparisons between feature selection procedures can be found in (Dash & Liu, 1997; Blum & Langley, 1997; Jain & Zongker, 1997; Aha & Bankert, 1995).

### Evaluation of the Feature Subsets

The most intuitive measure of quality of a feature subset is the error of a classifier built on these features. In theory, one can calculate the error under the assumption that the probabilities are equal to their expert estimates. The optimal classifier for independent features is the Naïve Bayes classifier. Denote by $P_j$ the prior probability for class $\omega_j$. Let $\mathbf{x} = [x_1,\ldots,x_n]^T$ be a binary vector to be labeled into one of the $c$ mutually exclusive classes. A discriminant function is calculated for each class,

$$\mu_j(\mathbf{x}) = P_j P(\mathbf{x} \mid \omega_j) = P_j \prod_{i=1}^{n} P(x_i \mid \omega_j), \quad j = 1,\ldots,c \tag{2}$$

$\mathbf{x}$ is labeled in the class with the largest discriminant value. There are $2^d$ possible binary vectors $\mathbf{x}$ for a candidate subset $S$ with $d$ features. The (probability for the) minimum classification error for the subset can be calculated as

$$P_e = \sum_{\mathbf{x}} P(\mathbf{x}, \text{error}) = 1 - \sum_{\mathbf{x}} \max_j \left[ P_j \prod_{i \in S} P(x_i \mid \omega_j) \right] \tag{3}$$

Equation (3) shows the difficulty in calculating the error for large $d$. Every $\mathbf{x}$ must be visited to decide which class label to assign to it. There are indirect criteria related to the error which may be faster to calculate, but direct calculation of the error in some form is preferable (Dash & Liu, 1997). Monte Carlo simulations were chosen for estimating the error of the selected feature subset. The probabilities for each class were available and it was therefore possible to generate randomly a sample from each class with $n$ independent features. Using the selected feature subset, the Naïve Bayes classifier was applied for the objects in this sample.

### The Single-Best Method (SB)

It is known that the individually best $d$ features do not necessarily form the best subset of $d$ features (Toussaint, 1971). Nonetheless, the method is quick and sometimes surprisingly efficient. The error for each feature is calculated separately using (3) (note that there are only two possible $\mathbf{x}$'s for each feature: present or absent), the errors are sorted in ascending order and the top $d$ features are retained. In this method, one can pick a desired value for $d$.

The complexity of a feature selection algorithm is typically measured by the number of calculations of the classification error needed to select $d$ out of $n$ features. Thus the single-best method needs just $n$ evaluations regardless of the number $d$.

Sequential Forward Selection (SFS)

This is the method traditionally used in stepwise regression. To start, there is an empty set, *S*, of chosen features. Each feature must be evaluated separately as in the single-best method and the best individual feature is placed in *S*. At the next step, all pairs of features which contain the feature selected already and one other feature are evaluated. The pair with the smallest error is retained as *S*. Then, one must check all triples of features, and so on, until the desired cardinality *d* of *S* is reached. This procedure does not guarantee finding the optimal set of *d* features even in this simple case of independent binary features. The reason for this can be explained again with the Toussaint's counter example: the best set of two does not necessarily contain the single best feature.

Below, an example illustrating both the non-optimality of the sequential feature selection (SFS) and the calculation of the error though equation (3) is shown.

Consider three features, $x_1$, $x_2$, and $x_3$, and two classes, $\Omega = \{\omega_1, \omega_2\}$. The non-traditional data considered in this study is given in the form of probability estimates $P(x_i = 1 \mid \omega_j)$, as shown in Table 2.

Table 2. An example of a set of probabilities for 3 features and 2 classes

| | $\omega_1$ | $\omega_2$ |
|---|---|---|
| $x_1$ | 0.1 | 0.5 |
| $x_2$ | 0.6 | 0.1 |
| $x_3$ | 0.8 | 0.4 |

Denote $a = P(x_k = 1 \mid \omega_1)$ and $b = P(x_k = 1 \mid \omega_2)$ for some $x_k$. Assuming equal prior probabilities for the two classes, the probability of correct classification for feature $x_k$ is

$$P(k) = 1/2\{\max(a,b) + \max(1-a, 1-b)\} \quad (4)$$

Using (4), the individual errors for the features are $\varepsilon_1 = 1 - ½ [\max(.1,.5) + \max(.9,.5)] = 0.30$, $\varepsilon_2 = 0.25$ , and $\varepsilon_2 = 0.30$ . Consider a pair of features, $(x_k, x_j)$, and denote the probabilities for $x_j$ as $p = P(x_j = 1 \mid \omega_1)$ and $q = P(x_j = 1 \mid \omega_2)$. Substituting again in equation (3), the probability of correct classification for the pair of features is

$$\begin{aligned} P(k,j) = 1/2\{&\max(a\,p, bq) \\ &+ \max[(1-a)\,p, (1-b)q] \\ &+ \max[a(1-p), b(1-q)] \\ &+ \max[(1-a)(1-p), (1-b)(1-q)]\} \end{aligned} \quad (5)$$

The errors for the three pairs of features for the example in Table 2 are

$\varepsilon_{12}$ = 1–½(max(.1×.6,.5×.1)
+ max(.9×.6,.5×.1)
+ max(.1×.4,.5×.9)
+ max(.9×.4,.5×.9))
= 0.25,

$\varepsilon_{13}$ = 0.24, and $\varepsilon_{23}$ = 0.25.

As $\varepsilon_{13}$ is the smallest pair-wise error, and $\varepsilon_2$ is the smallest individual error, the best pair of independent features, $(x_1, x_3)$, does not include the single best feature $x_2$.

SFS is probably the most widely used procedure because it has both reasonable error and reasonable complexity for "traditional" data sets (Aha & Bankert, 1995; Jain & Zongker, 1997).

At the first step, SFS evaluates all *n* features, at the second step, *n*-1 evaluations are needed as there are *n*-1 possible pairs. For selecting *d* features, SFS needs the following number of evaluations of the error

$$\sum_{i=0}^{d-1} (n-i) \quad (6)$$

However, the complexity calculation is not that simple when the features from probabilistic data

as shown in Table 1 are selected. For the calculation of the theoretical error, the algorithm has to visit every **x** in the possible feature space, find out which is the maximum discriminant function, and add the contribution of the error

for **x** based on the class label decision. The fact that the features are treated as independent does not make the task any easier. The complexity of SFS will depend heavily on the number of features in the evaluated subsets.

Complexity of feature selection algorithms for probabilistic data can be evaluated by the total number of **x**'s visited in the process of selecting $d$ out of the $n$ features. The complexity for the single-best method is $C_{SB} = 2n$, and for the SFS, $C_{SFS} = \sum_{i=0}^{d-1}(n-i)2^{i+1}$.

Class-Pairs Feature Selection (CP)

Ji and Bang (2000) proposed the following feature selection method. A single feature is selected for each pair of classes.

Table 3 shows the data structure used by the algorithm, where $C_{ij}$ = class pairs, ( $i \neq j$ ), $x_k$ = $k$-th feature, ($k$ = 1,..,$n$), $P_{ij}(k)$ = discriminatory power of feature $k$ for $C_{ij}$. Using (4), the values of $P_{ij}(k)$ are calculated as the probability of correct classification between classes $\omega_i$ and $\omega_j$ for feature $x_k$.

Table 3. The table for the class-pairs method for feature selection (Ji and Bang, 2000).

| | $C_{ij}$ | | |
|---|---|---|---|
| | … | | |
| $x_k$ | … | $P_{ij}(k)$ … | $T_k$ |
| | … | | |
| | $E_{ij}$ | | |

The following values are then calculated

- $E_{ij} = \sum_k P_{ij}(k)$, the relative ease of classifying the pair $C_{ij}$, and
- $T_k = \sum_{ij} P_{ij}(k)$, the relative discriminatory power of feature $x_k$.

The algorithm begins with an empty set of features. The class pair that is the hardest to discriminate (has the smallest $E_{ij}$) is identified from the table. The feature with the highest discriminatory power for this pair is added to the subset, if not already selected. If more than one feature has the highest $P_{ij}(k)$ in the chosen column, then the feature with the highest value of $T_k$ is selected. The hardest pair is removed from the table and the process continues with the next hardest pair of classes (Note that the classes are not removed altogether, only the column of the table is removed.). The process stops once all class pairs have been covered.

The maximum number of features this method will select is $\max\{(c(c\text{-}1)/2, n\}$. However, Ji and Bang (2000) claim that the number selected will be much less than either of these. This method may also be restricted at any point to pick only $d$ features. The complexity of the class-pair method (measured again by the total number of **x**'s visited) is $C_{CP} = c\ (c\text{-}1)\ n$. This calculation reflects only the preparation phase (setting up Table 3), and does not take into account the actual procedure which constructs the feature subset.

Feature-Pairs Feature Selection (FP)

The selection methods considered above are either overly simplistic but scale well with $n$, $c$, and $d$ (single-best) or they are computationally demanding but more accurate (SFS). Optimality of the selected feature subset is not guaranteed in any case. The class-pairs method is one possible method that scales well and may be accurate. Here, another method is proposed for feature selection from probabilities, called feature-pairs method.

The process is started with an empty set of features. All pairs of features are evaluated and the best pair is added to the set. While the desired number of features is not reached, add the features from the next best pair which are not already among the selected features. Suppose that $d$-1 features are already in the set, and there

is a pair of features such that neither of the two members of the pair is in the set. One may either take both features and exit with $d$+1 features or randomly select one member of the pair to make up the total of $d$ features in the set. The complexity of the feature-pairs method (using the number of visited **x**'s) is $C_{FP} = n\,(n\text{-}1)$.

All four methods are based on a true calculation of the classification error plus some heuristic about how one forms the feature subset. The experimental results in the next section help to evaluate the assets and drawbacks of the four methods.

## Results

### A Small-Scale Simulation Study

To include SFS in the comparisons, a relatively small example with $n$ = 20 features was chosen and the number classes, $c$, was varied from 3 to 10. The number of selected features, $d$, was varied from 2 to 10.

For each $c$, 50 random matrices of size $20\times c$ were generated from uniform random distribution. Each matrix represented the probabilities for the features and classes as shown in Table 1. For each such matrix and each $d$, the four feature selection algorithms were applied and the best subset of size $d$ was found.

To evaluate the selected subsets, a traditional data set was generated randomly for every pair ($c$,$d$). One hundred data points were generated from the distribution of each class and the Naïve Bayes classifier was used to label these points. The error was estimated as the percent mismatch with the true class label.

An example of the simulation algorithm is given below. Consider the problem presented in Table 2. Suppose that Method $X$ picked features ($x_1$, $x_3$). Set a misclassification counter to 0. The steps below are repeated 100 times for each class.

(Step 1) Generate a data point from class $\omega_1$. To do this, pick a vector of 3 random numbers, one for each feature, e.g. $[0.2736, 0.9241, 0.7102]^T$. Compare this vector with the first column of Table 2 (corresponding to $\omega_1$). If the generated number for $x_i$ is smaller than the corresponding probability in the table, set $x_i$ to 1; else set $x_i$ to 0. For this example, the generated data point is **x** = [0, 0, 1].

(Step 2) Classify the data point using Naïve Bayes and only the chosen features. For this example ($x\_1$=0, $x\_3$=1), the two discriminant functions for **x** are

$$\mu_1(\mathbf{x}) = 1/2\,(0.9\times 0.8) = 0.36$$
$$\mu_2(\mathbf{x}) = 1/2\,(0.5\times 0.4) = 0.10$$

(Step 3) Choose a class label by the maximum discriminant function and note whether there is a mismatch with the class label whose distribution is currently being used. In the example, label $\omega_1$ is chosen so the misclassification counter remains unchanged.

Figure 1 shows the probability of error versus the number of selected features, $d$, for $c$ = 10 classes. Each point on the figure is the average error over the 50 random matrices.

As expected, SFS gives the lowest error. The single-best and the feature-pairs methods are approximately the same with a slight preference to feature-pairs, and the class-pairs method is the worst. For $d$=2 selected features, SFS is the second best method because feature pairs selects the true best pair features.

Figure 1. Probability of error versus the number of selected features ($n$=20, $c$=10).

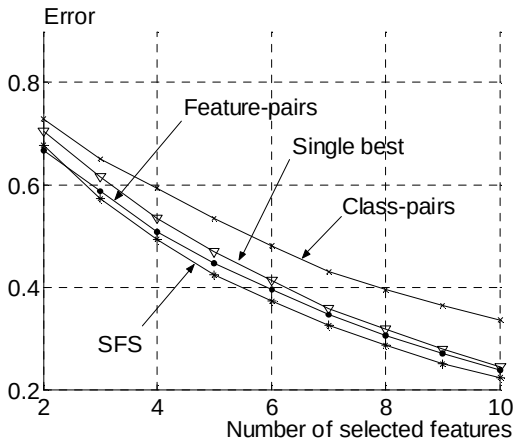



Table 4 gives the classification error averaged across the 50 random matrices of probabilities for 2 and 10 selected features (out of 20), for $c$ = 3,…, 10 classes.

Table 4. Classification error (in %) with 2 and 10 features for $c$ = 3,…, 10 classes. CP stands for class-pairs method, SB for the single-best method and FP for the feature-pairs method.

(a)

$d$ = 2 selected features

| $c$ | CP | SFS | SB | FP |
|---|---|---|---|---|
| 3 | 21.2 | 17.9 | 22.7 | 16.8 |
| 4 | 40.1 | 31.7 | 36.1 | 30.3 |
| 5 | 49.6 | 42.9 | 47.2 | 41.1 |
| 6 | 57.9 | 51.0 | 54.2 | 49.4 |
| 7 | 62.6 | 56.2 | 60.3 | 54.3 |
| 8 | 67.5 | 61.3 | 64.3 | 59.4 |
| 9 | 70.2 | 65.1 | 67.8 | 63.8 |
| 10 | 72.8 | 67.8 | 70.6 | 66.8 |

(b)

$d$ = 10 selected features

| $c$ | CP | SFS | SB | FP |
|---|---|---|---|---|
| 3 | 14.4 | 4.2 | 4.4 | 4.5 |
| 4 | 16.8 | 7.3 | 7.9 | 8.0 |
| 5 | 16.1 | 9.8 | 10.8 | 11.2 |
| 6 | 21.2 | 13.7 | 15.0 | 15.1 |
| 7 | 25.0 | 15.5 | 17.2 | 17.3 |
| 8 | 29.1 | 18.4 | 20.4 | 19.8 |
| 9 | 31.2 | 20.8 | 23.0 | 22.8 |
| 10 | 33.6 | 22.3 | 24.3 | 23.9 |

The results in Table 4 confirm the superiority of SFS for more than 2 features and it also shows that the class-pairs method gives the largest error. There is an interesting turn about the single-best and feature-pairs methods. For small number of classes (3 to 7) SB was slightly better whereas for larger number of classes (8 to 10) FP was the better of the two methods. This behavior is an indication that for larger scale problems FP may be the more accurate method.

A Larger-Scale Simulation Study

SFS was excluded from this experiment because of its large computational time. The same experiments, as in the previous section, were run with a total number of features $n$ = 100 and number of classes $c$ = 50. The number of selected features was d ∈ {5, 10, 15,…, 50}. Figure 2 shows the error versus the number of selected features for SB, CP and FP. The curves are close together but the errors for all $d$ are related as $E_{FP} < E_{SB} < E_{CP}$. The differences between $E_{FP}$ and $E_{SB}$ are not statistically significant.

Figure 2. Probability of error versus the number of selected features ($n$ = 100, $c$ = 50).

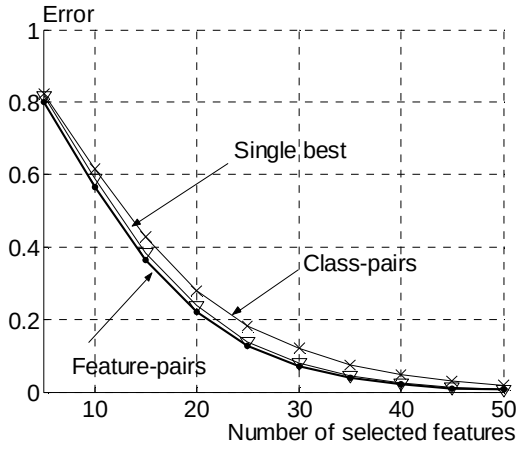
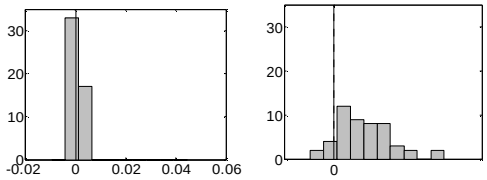



Figure 3 shows the histogram of the 50 differences $E_{SB} - E_{FP}$ for 50 and 25 selected features. For 50 features, $E_{SB} - E_{FP}$ was positive in 64% of the runs, the same in 6% of the runs and negative in 30% of the runs. For 25 selected features, $E_{SB} - E_{FP}$ was positive in 94% of the runs and negative in 6% of the runs. This suggests that there may be optimal ratios $c$:$d$:$n$ for which FP is distinctly better than SB.

Figure 3. Histograms of the 50 differences $E_{SB} - E_{FP}$ for $d$ = 50 selected features (a) and $d$ = 25 selected features (b).




(a) (b)

The computational time ratio for the three methods was approximately $C_{SB}$:$C_{CP}$:$C_{FP}$ = 1:8:23.

The above simulations do not assume any relationship between the classes. The matrices are generated uniformly which means that the correlations between the columns will be

close to 0, as will be the correlations between the rows. In real problems, the class profiles will rarely be uncorrelated. Below, the four methods are explored on two real diagnostic problems where only probabilistic data is available.

An Application to Diagnosis of BSE in cattle and Scrapie in Sheep

The above feature selection methods were applied for selecting diagnostic signs in two problems coming from veterinary medicine.

BSE and Scrapie are fatal neurodegenerative diseases. Both are notifiable diseases which have no known cure. There is currently no ante-mortem test for the diseases that can be used routinely in the field. Notifiable diseases have a major impact on human health, welfare and economics. There was a BSE epidemic in Britain in the 90's and with the first ever BSE case diagnosed in the USA at the end of 2003, the problem of these diseases is global. Therefore, the recognition of the clinical presentations of the two diseases and the need to differentiate them from other diseases is important. In veterinary medicine, prevalence of disease, the conditional dependencies of clinical signs, and the sign frequencies within diseases are rarely, if ever available; demonstrating the need to work with probability data.

Table 5 shows the results from the feature selection experiments with the BSE data. SFS was applied to select 10 of the 242 features and simulated data from the distributions of the 57 classes. The three selection methods SB, CP, and FP, which have lower capacity than SFS were run for $d$ = 10 features too. The first 4 rows in Table 4 show the classification error for $d$ = 10.

Next, the class-pairs method was run letting it stop when all class pairs have been accounted for. CP selected a total of 58 features. Leaving SFS aside, the other two low-complexity methods were run for 58 features. The classification error is displayed in rows 5-8 in Table 5. Finally, the error with using all features was estimated as a tight lower bound on the classification error.

Table 5. Results from feature selection on the BSE probabilities.

| Method ($d$) | Error |
|---|---|
| SFS (10) | 0.4258 |
| SB (10) | 0.6432 |
| CP (10) | 0.5865 |
| FP (10) | 0.5482 |
| CP (58) | 0.0172 |
| SB (58) | 0.0309 |
| FP (58) | 0.0256 |
| ALL (242) | 0.0049 |

The results show that the closest rival to SFS for small number of features is the FP method proposed here. Contrary to the results in the previous section though, CP is better than SB. This shows that in real-life problems when there is dependency between the classes, CP may be a better solution than SB. When run all the way, CP provides the smallest classification error of the three low complexity methods followed by FP and then SB.

Note the large differences between the error probabilities for small number of features. These differences strongly suggest that SFS should be applied as long as the computation time is acceptable. To illustrate the differences between the selected sets of features, Table 6 shows the signs selected by SFS (a) and SB (b) in the order they entered the set.

The same pattern of experiments was repeated for the data containing the probabilities for Scrapie and 62 alternative diseases. Twelve features were selected by SFS. The 3 lower-complexity methods were run for $d$ = 12. The errors are shown in Table 7. The class-pairs method (CP) was run again until all class pairs were covered. The number of selected features was 77. SB and FP were then run for the same number of features. Table 7 ranks the feature selection methods exactly in the same way as Table 5. Again, the discrepancies with the simulation study in the previous sub-section can be attributed to the fact that the classes here are not independent. The CP method manages to capture some dependency between the classes and, if run all the way, it selects better subsets of features than SB and FP. Table 8 mirrors table 6 by showing the signs selected for diagnosing Scrapie and the 63 alternative diseases.

Table 6. Signs selected by SFS and SB for diagnosing BSE and 56 other diseases in cattle

| (a) Signs selected by SFS |
|---|
| Gait abnormal, unspecified |
| Circling in one direction |
| Hypo-responsive to external stimuli |
| Milk yield less than normal (individual) |
| Rumen rate nil, (0 per 2min) |
| Eye menace response absent |
| Hyper-responsive to external stimuli |
| Dyspoena, unspecified |
| Posture recumbency |
| Temperature >39.5 degrees Celsius |

| (b) Signs selected by SB |
|---|
| Gait abnormal, unspecified |
| Dyspoena, unspecified |
| Dyspoena, rate increased shallow |
| Diarrhoea, unspecified |
| Gait uncoordinated\exaggerated |
| Rumen rate slow (1 per 2min) |
| Diarrhoea, acute, profuse |
| Circling in one direction |
| Gait stiff |
| Head rotated, tilted or deviated |

Table 7. Results from feature selection on the Scrapie probabilities.

| Method ($d$) | Error |
|---|---|
| SFS (12) | 0.5975 |
| SB (12) | 0.7635 |
| CP (12) | 0.6930 |
| FP (12) | 0.6610 |
| CP (77) | 0.0625 |
| SB (77) | 0.0992 |
| FP (77) | 0.0649 |
| ALL (285) | 0.0252 |

Table 8. Signs selected by SFS and SB for diagnosing Scrapie and 63 other diseases in sheep

| (a) Signs selected by SFS |
|---|
| Foul odour skin |
| Mastitis |
| Exercise intolerance |
| Paraparesis |
| Weight Loss |
| Generalized weakness |
| Anorexia |
| Generalized lameness or stiffness |
| Ataxia |
| Underweight, thin etc |
| Dullness |
| Reluctant to move |

| (b) Signs selected by SB |
|---|
| Foul odour skin |
| Mastitis |
| Matted \ dirty wool \ hair |
| Moist skin\wool \hair |
| Skin necrosis |
| Exercise intolerance |
| Hyperkeratosis |
| Lymphadenopathy |
| Alopecia |
| Pruritus |
| Weight loss |
| Dullness |

## Conclusion

The problem of selecting a subset of $n$ binary features to discriminate between $c$ mutually exclusive classes was explored. The information available here is in the form of an $n \times c$ table with class-conditional probabilities for the $n$ binary features, i.e., $P(x_i=1|\omega_j)$, $i = 1,\ldots,n$, $j = 1,\ldots,c$. Selecting the best subset of features seems easy because all the probabilistic

information is available and the features are assumed to be independent. The difficulty comes from the complexity of the evaluation of the theoretical classification error for a subset of features.

An easy way out would be to generate a sample and run it through the Naïve Bayes classifier using only the features in the subset. Three methods were applied from the literature (SFS, SB and CP) and a method was proposed based on features pairs (FP) for feature selection using probabilities. It was found that SFS was the most accurate but also the most computationally demanding of the four methods. The simulation experiments with generated random distributions suggested that CP was inferior to SB and FP, but did not favor strongly any of SB or FP. The experiments with two real data matrices from veterinary medicine demonstrated that CP is also a valuable method when larger subsets of features are acceptable. FP was found to be the best alternative to SFS for small and medium subsets.

There are at least two caveats that need to be mentioned. First, features are rarely independent in real life problems. By assuming independence, one runs the risk of missing an important feature which does not have a reasonable predictive value on its own, but is highly important in combination with others. However, in the absence of any further information, the independence assumption is the only option. Second, the estimates of the probabilities given as the information to work upon (Table 1) might not be very close to the true probabilities. A sensitivity study can be run by perturbing the probability estimates and observing how the selected feature subset changes.

The acid test for the quality of the selected subset of features would be the error on real data. However, the aim of this study is a preliminary feature selection so that a real data set can be collected using these features. Therefore, at this stage, a reasonably large feature set should be provided. The hope is that highly discriminative combinations of features will be discovered within using systematically collected data.

References

Aha, D.W. & Bankert, R. L. (1995). A comparative evaluation of sequential feature selection algorithms. *In Proc. 5th International Workshop on AI and Statistics*, pages 1-7, Ft Lauderdale, FL.

Blum, A. & Langley, P. (1997). Selection of relevant features and examples in machine learning. *Artificial Intelligence, 97*, 245-271.

Brightling, P., Larcombe, M. T., Blood, D. C., & Kennedy, P. C. (1996). Development and the use of Bovid-3, an expert system for veterinarians involved in diagnosis, treatment and prevention of diseases of cattle. In *XIX World Buitrics Congress Proceedings, 2*, pages 528-532.

Dash, M., & Liu, H. (1997). Feature selection for classification. *Intelligent Data Analysis, 1*, 131-156.

Domingos, P. & Pazzani, M. (1997). On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning, 29*, 103-130.

Jain, A. K., & Zongker, D. (1997). Feature selection: evaluation, application and small sample performance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *19*, 153-158.

Jain, A. K., & Chandrasekaran, B. (1982). Dimensionality and sample size considerations in pattern recognition practice. In Krishnaiah, P.R. & Kanal, L. N., editors, *Handbook of Statistics*, pages 835--855. North Holland.

Ji, H. & Bang, S. Y. (2000). Feature selection for multiclass classification using pairwise discriminatory measure and covering concept. *Electronics Letters, 36(6)*,524-525

MAFF (2000). Animal Health 2000. MAFF, London.

Patrick, E. (1972). *Fundamentals of Pattern Recognition*. Prentice-Hall, Inc., Englewood Cliffs, N.J.

Stearns, S. (1976). On selecting features for pattern classifiers. *In Proc 3-d International Conference on Pattern Recognition*, pages 71-75, Coronado, CA.

Tax. D., & Duin. R. (2002). Using two-class classifiers for multi-class classification. In *Proc. 16th International Conference on Pattern Recognition, 2,* 124-127.

Toussaint, G. T. (1971). Note on optimal selection of independent binary-valued features for pattern recognition. *IEEE Transactions on Information Theory, 17,* 618.

Van Campenhout, J. M. (1982). Topics in measurement selection. In Krishnaiah, P. R. & Kanal, L.N., editors, *Handbook of Statistics*, pages 793-803. North Holland.

White, M.E. (1984). Consultant: computer-assisted differential diagnosis, *Veterinary Computing, 2,* 9-12.

Journal of Modern Applied Statistical Methods
November, 2005, Vol. 4, No. 2, 538-544


1538 – 9472/05/$95.00

# Kim And Warde's Mixed Randomized Response Technique For Complex Surveys

Amitava Saha
Directorate General of Mines Safety
India

The randomized response (RR) technique introduced by Warner (1965) was found to be an effective method for reducing answer bias and ensuring better respondent cooperation in estimating the proportion of people in a community bearing a sensitive attribute. Chaudhuri (2001a, 2001b, 2002, 2003) extended Warner's method and several other well-known RR devices to complex surveys adopting a varying probability sampling design. Kim and Warde (2004) proposed an RR model assuming that the sample is selected with simple random sampling (SRS) with replacement (SRSWR). Here, the method of estimation is presented when sample is chosen with varying selection probabilities and Kim and Warde's RR procedure is applied for estimating a sensitive proportion. Also illustrated is a numerical example that unequal probability sampling performs better than SRS.

Key words: Answer bias; randomized response; sensitive attribute; simple random sampling; varying probability sampling

## Introduction

Warner (1965) proposed a method called randomized response (RR) to ensure better respondent cooperation and honest responses in surveys involving collection of information on certain sensitive attributes. It has been found that Warner's technique is capable of reducing answer bias and refusals considerably in surveys where a question of sensitive nature is involved. This method has been studied extensively and as a consequence, numerous modifications of it as well as several other methods have emerged in the literature of RR. Among many others, Horvitz et al. (1967), Greenberg et al. (1969), Kuk (1990), Christofides (2003), Mangat and Singh (1990) made notable contributions.

Most of the works cited here have been done assuming that the sample is selected with simple random sampling (SRS) with replacement (SRSWR). But in practice, in the socio-economic surveys, the respondents are usually selected with varying probability sampling. Thus, to meet the demand of the social surveys, Chaudhuri (2001a, 2001b, 2002, 2004) extended some of the RR procedures to complex survey situations.

Most of the works cited here have been done assuming that the sample is selected with simple random sampling (SRS) with replacement (SRSWR). But in practice, in the socio-economic surveys, the respondents are usually selected with varying probability sampling. Thus, to meet the demand of the social surveys, Chaudhuri (2001a, 2001b, 2002, 2004) extended some of the RR procedures to complex survey situations.

Kim and Warde (2005) proposed a mixed RR model in an attempt to improve Moors (1971) model after taking due consideration of the inherent privacy problem of Moors (1971) RR device. They have also discussed how their method may be applied when stratified sampling design is used. But the entire development of Kim and Warde (2005) is based on the assumption that the sample is selected with SRSWR. Since in large-scale sample surveys equal probability sampling is rarely used, necessary modifications need to be developed for adopting this method to complex sample surveys where varying probability sampling designs are often used. Here, Kim and

Contact information for Amitava Saha is Dhanbad, Jharkhand – 826001, India. E-Mail: saha_amitava@hotmail.com

Warde's (2005) procedure is presented when a varying probability sampling design is adopted rather than SRSWR. As well, a numerical illustration of the performance of the extended procedure under varying and equal probability sampling is presented.

Kim and Warde's (2005) Device in Complex Surveys

Kim and Warde's (2005) method for complex surveys is described in section 2. A numerical study for comparing the relative performances is reported in section 3.

Let $U=(1,\cdots\cdots,i,\cdots\cdots N)$ be a finite population of $N$ individuals and $y_i$ be the value of a variable of interest, say, $y$ on the $i$th individual such that $y_i = 1$ if $i$ bears a sensitive attribute $A$ = 0 if $i$ bears the complementary attribute $A^C$. The problem is to estimate the proportion of people in $U$ bearing the character $A$, i.e., $\pi_A=\left(\sum_{i=1}^{N} y_i\right)\Big/N=Y/N$ where $Y=\sum_{i=1}^{N} y_i$ on choosing a sample, say, $s$ of size $n$ from $U$ according to any arbitrary sampling design $p$.

It is also assumed that $x_i$ be the value of a variable $x$ on the $i$th individual in $U$ such that $x_i = 1$ if $i$ bears a non-sensitive attribute $B$ = 0 if $i$ bears $B^C$, the complement of $B$. Kim and Warde (2004) proposed a method for estimating $\pi_A$ when a sample of size $n$ is drawn from $U$ by SRSWR. However, in this article it is assumed that instead of selecting the individuals by SRSWR only, they are chosen following any arbitrary sampling design $p$.

In Kim and Warde's (2005) device every sampled person is requested to answer a direct question about his/her possession of a non-stigmatizing or innocuous character, say, $B$ and on receiving a '*yes*' reply to this non-sensitive question the individual is instructed to use an RR device $R_1$ where a pack of cards marked $A$ and $B$ in proportions $p_1:(1-p_1), 0<p_1<1$ is kept. The respondent is then requested to draw a card at random from this pack, unnoticed by the interviewer and to report the true value of $y$ or $x$ according as $A$-marked or $B$-marked card is drawn. If a respondent answers '*no*' to the initial direct question, he/she is requested to go to another RR device, $R_2$, in which there is another pack of cards marked $A$ and $A^C$ in proportions $p_2:(1-p_2), 0<p_2<1, p_2\neq 1/2$. The respondent is then instructed to choose a card randomly from this pack and to report the true value of $y$, i.e., either '1' or '0', if there is a match (mismatch) between his/her true $y$ character and the card type drawn. Here, it is assumed that the sensitive and the innocuous questions are unrelated and also that the RR devices $R_1$ and $R_2$ are independent.

Suppose that out of the $n$ selected persons $n_1$ reply '*yes*' to the direct question and the remaining $n_2 = n - n_1$ persons provided a '*no*' answer to it. Now, the following are defined:

$I_i$ = 1 if the $i$th selected individual bears the sensitive character and draws an A – marked card or if the individual bears the non-sensitive character and chooses a B – marked card

= 0 else on using $R_1$.

Then $P(I_i = y_i) = p_1$ and $P(I_i = x_i) = 1 - p_1$ and writing $E_R, V_R$ as the expectation and variance operators with respect to any arbitrary RR device it is easy to check that,

$$E_R(I_i) = p_1 y_i + (1-p_1)x_i$$

$$= p_1 y_i + (1-p_1).$$

This is because a respondent using the device $R_1$ has already responded '*yes*' to the initial direct innocuous question. Thus, it follows that for

$$r_i = [I_i - (1-p_1)]/p_1, 0<p_1<1,\ E_R(r_i) = y_i,\ \forall i \in U$$

and

$$V_R(r_i) = \frac{V_R(I_i)}{p_1^{\,2}} = \frac{(1-p)(1-y_i)^2}{p_1} = V_{1i}.$$

It may be seen that $r_i$ is an unbiased estimator for $y_i$ and also an unbiased estimator for $V_{1i}$ is given by $v_{1i} = \frac{(1-p)(1-r_i)^2}{p_1}$. Further,

let $J_i$ = 1 if $i$th selected individual bears the sensitive attribute $A$ and draws an $A$-marked card = 0 else, on applying $R_2$. Then,

$$P(J_i = y_i) = p_2 \text{ and } P(J_i = 1 - y_i) = 1 - p_2$$

and

$$E_R(J_i) = p_2 y_i + (1-p_2)(1-y_i) = (2p_2 - 1)y_i + (1-p_2),$$
$$V_R(J_i) = p_2(1-p_2).$$

For $u_i = [J_i - (1-p_2)]/(2p_2 - 1)$, $p_2 \neq 1/2$, there is $E_R(u_i) = y_i$, $\forall\, i \in U$ and $V_R(u_i) = \dfrac{p_2(1-p_2)}{(2p_2-1)^2} = V_{2i}$, say. Thus, $u_i$ is also unbiased for $y_i$ and an unbiased estimator of $V_{2i}$ is given by $v_{2i} = V_{2i}$.

Let $s_1$ and $s_2$ be respectively the sets of sampled individuals offering '*yes*' and '*no*' responses to the initial direct innocuous question such that $s_1 \cup s_2 = s$ and write $E_p, V_p$ respectively to denote the operators for expectation and variance with respect to the probability design $p$. Suppose that $t_k = \sum_{i=1}^{N} b_{s_k i} I_{s_k i} y_i$ where $I_{s_k i} = 1\,(0)$, if $i \in s_k\,(\notin s_k), k = 1, 2$ and $b_{s_k i}$'s are constants free of $\underline{Y} = (y_1, \ldots\ldots\ldots, y_N)$ such that $E_p(b_{s_k i} I_{s_k i}) = 1,\ \forall\, i \in U$ be a homogeneous linear unbiased estimator for $Y = \sum_{i=1}^{N} y_i$. The following is written as:

$$V_p(t_k) = \sum_{i=1}^{N} y_i^2 c_{ki} + \sum_{i \neq j} y_i y_j c_{kij}$$

where

$$c_{ki} = E_p(b_{s_k i}^2 I_{s_k i}) - 1$$

and

$$c_{kij} = E_p(b_{s_k i} I_{s_k i} - 1)(b_{s_k j} I_{s_k j} - 1)$$

and an unbiased estimator of $V_p(t_k), k = 1, 2$ as

$$v_p(t_k) = \sum_{i=1}^{N} y_i^2 c_{s_k i} I_{s_k i} + \sum_{i \neq j} y_i y_j c_{s_k ij} I_{s_k ij}$$

where $I_{s_k ij} = I_{s_k i} I_{s_k j}$ and $c_{s_k i}$, $c_{s_k ij}$ are $\underline{Y}$-free constants satisfying $E_p(c_{s_k i} I_{s_k i}) = c_{ki}$ and

$$E_p(c_{s_k ij} I_{s_k ij}) = c_{kij},\ k = 1, 2.$$

Because $y_i$'s are unascertainable, two unbiased estimators for $Y$ based on $s_1$ and $s_2$ are obtained

$$e_1 = \sum_{i \in s_1} b_{s_1 i} I_{s_1 i} r_i$$

and

$$e_2 = \sum_{i \in s_1} b_{s_2 i} I_{s_2 i} u_i$$

and accordingly, two unbiased estimators for $\pi_A = Y/N$ are given by

$$\overline{e_1} = e_1/N \quad \text{and} \quad \overline{e_2} = e_2/N.$$

Now, following Raj (1968) and Rao (1975), two unbiased estimators for $V(e_1)$ and $V(e_2)$ are obtained as:

$$v_1(e_1) = v_p(t_1)\big|_{\underline{Y} = \underline{R}} + \sum_{i=1}^{N} b_{s_1 i} I_{s_1 i} v_{1i}$$

$$v_2(e_1) = v_p(t_2)\big|_{\underline{Y} = \underline{R}} + \sum_{i=1}^{N} (b_{s_1 i}^2 - c_{s_1 i}) I_{s_1 i} v_{1i}$$

$$v_1(e_2) = v_p(t_2)\big|_{\underline{Y} = \underline{R}} + \sum_{i=1}^{N} b_{s_2 i} I_{s_2 i} V_{2i}$$

$$v_2(e_2) = v_p(t_2)\big|_{\underline{Y} = \underline{R}} + \sum_{i=1}^{N} (b_{s_2 i}^2 - c_{s_2 i}) I_{s_2 i} V_{2i}.$$

Since both $e_1$ and $e_2$ are unbiased estimators for $Y$, an unbiased estimator of $Y$ based on $e_1$ and $e_2$ is given by

$$e = \frac{n_1}{n} e_1 + \frac{n_2}{n} e_2$$

and

$$V(e) = \left(\frac{n_1}{n}\right)^2 V(e_1) + \left(\frac{n_2}{n}\right)^2 V(e_2)$$

$$= \left(\frac{n_1}{n}\right)^2 V(e_1) + \left(1 - \frac{n_1}{n}\right)^2 V(e_2).$$

Also, an unbiased estimator of $\pi_A$ is given by $\hat{\pi}_A = \frac{n_1}{n}\bar{e}_1 + \frac{n_2}{n}\bar{e}_2$. Again, as the two RR devices are independent, unbiased variance estimators for $V(e)$ are derived as

$$v_1(e) = \left(\frac{n_1}{n}\right)^2 v_1(e_1) + \left(\frac{n_2}{n}\right)^2 v_1(e_2)$$

$$v_2(e) = \left(\frac{n_1}{n}\right)^2 v_2(e_1) + \left(\frac{n_2}{n}\right)^2 v_2(e_2)$$

and similarly, the unbiased estimators for $V(\hat{\pi}_A)$ are given by

$$v_1(\hat{\pi}_A) = \left(\frac{n_1}{n}\right)^2 v_1(\bar{e}_1) + \left(\frac{n_2}{n}\right)^2 v_1(\bar{e}_2)$$

$$v_2(\hat{\pi}_A) = \left(\frac{n_1}{n}\right)^2 v_2(\bar{e}_1) + \left(\frac{n_2}{n}\right)^2 v_2(\bar{e}_2).$$

A Numerical Example

Artificial data relating to a community of $N$ = 129 individuals is considered. As well, the problem of estimating the proportion of individuals evading income tax during the last financial year in the said community on choosing a sample of $n$ = 37 individuals is considered. The individuals from this population were selected according to three different sampling schemes, namely, simple random sampling with replacement (SRSWR), simple random sampling without replacement (SRSWOR) and Rao-Hartley-Cochran (RHC, 1962) sampling scheme as a representative of varying probability sampling.

Here, $y_i$ = 1(0) is defined if the $i$th individual evades (does not evade) income tax during the last financial year and $x_i$ = 1(0) if the $i$th individual prefers (does not prefer) football to basketball. The amount of expenditure incurred in a particular month in the household to which an individual belongs to is considered as the size-measure for selection of the individuals by RHC sampling strategy.

In the RHC scheme, first the population of $N$ units is randomly divided into $n$ random groups, the $i$th group having $N_i$ units such that $\sum_n N_i = N$, where $\sum_n$ denotes the sum over the $n$ random groups. Then, denoting $A_i = a_{i_1} + \cdots\cdots\cdots + a_{i_{N_i}}$ as the sum of the normed size-measures $a_i$'s for the units belonging to the $i$th group, one unit is chosen from the $i$th group with a probability proportional to $A_i$ divided by it's $a$-value. This process is repeated for all the $n$ groups. Now, writing for simplicity $(y_i, a_i)$ as the $(y,a)$-value for the unit selected from the $i$th group, an unbiased estimator for $Y$ is given by

$$t = \sum_n (A_i / a_i)\, y_i$$

along with an unbiased variance estimator for $V(t)$ as

$$v(t) = B\sum_n A_i \left(\frac{y_i}{a_i} - t\right)^2$$

where

$$B = \left(\sum_n N_i^2 - N\right) \Big/ \left(N^2 - \sum_n N_i^2\right).$$

Here, $y_i$'s are unknown and so are to be estimated. Suppose that $w_i$ be an unbiased estimator for $y_i$ and $v_i$ be an unbiased estimator for $V_R(w_i)$. Then, one may employ the unbiased estimator

$$t = \sum_n (A_i / a_i)\, w_i$$

for estimating $Y$ and an unbiased variance estimator of $V(e)$, following Chaudhuri, Adhikary and Dihidar (2000) is given by

$$v(e) = v(t)\big|_{\underline{Y}=\underline{W}} + \sum_{i=1}^{N} b_{si} I_{si} v_i$$

where $\underline{W} = (w_1, \cdots\cdots\cdots, w_N)$. Let $e$ be any point estimator for the parameter $\theta$ and $v(e)$ be an unbiased estimator of $V(e)$. Then, assuming $\delta = (e-\theta)/\sqrt{v(e)}$ to be a standard normal deviate, the following two criteria are considered:

Table 1: Comparative performances of alternative procedures

| $p_1$ | $p_2$ | RHC | | | SRSWOR | | | SRSWR | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | $\hat{\pi}_A$ | CV | Length of CI | $\hat{\pi}_A$ | CV | Length of CI | $\hat{\pi}_A$ | CV | Length of CI |
| **$n_1$ = 30** | | | | | | | | | | |
| 0.98 | 0.47 | 0.65 | 11.4 | 0.366 | 0.40 | 16.9 | 0.264 | 0.59 | 18.5 | 0.265 |
| 0.92 | 0.48 | 0.74 | 15.0 | 0.397 | 0.37 | 17.4 | 0.281 | 0.46 | 18.9 | 0.313 |
| 0.93 | 0.76 | 0.68 | 14.9 | 0.475 | 0.32 | 17.3 | 0.276 | 0.40 | 18.1 | 0.315 |
| 0.81 | 0.84 | 0.85 | 17.9 | 0.466 | 0.34 | 21.6 | 0.319 | 0.34 | 24.9 | 0.362 |
| 0.89 | 0.68 | 0.65 | 16.4 | 0.491 | 0.32 | 19.4 | 0.290 | 0.42 | 22.1 | 0.327 |
| **$n_1$ = 25** | | | | | | | | | | |
| 0.98 | 0.47 | 0.44 | 13.9 | 0.362 | 0.48 | 15.8 | 0.222 | 0.43 | 18.7 | 0.264 |
| 0.92 | 0.48 | 0.43 | 17.1 | 0.351 | 0.41 | 19.7 | 0.253 | 0.44 | 20.8 | 0.273 |
| 0.93 | 0.76 | 0.41 | 17.5 | 0.345 | 0.47 | 19.7 | 0.234 | 0.41 | 23.1 | 0.278 |
| 0.81 | 0.84 | 0.49 | 19.7 | 0.375 | 0.39 | 23.9 | 0.294 | 0.38 | 26.8 | 0.332 |
| 0.89 | 0.68 | 0.43 | 18.2 | 0.379 | 0.37 | 20.1 | 0.267 | 0.36 | 22.2 | 0.297 |
| **$n_1$ = 20** | | | | | | | | | | |
| 0.98 | 0.47 | 0.33 | 15.1 | 0.282 | 0.35 | 18.9 | 0.217 | 0.32 | 20.3 | 0.242 |
| 0.92 | 0.48 | 0.39 | 18.6 | 0.229 | 0.39 | 21.2 | 0.210 | 0.32 | 23.7 | 0.258 |
| 0.93 | 0.76 | 0.32 | 19.4 | 0.260 | 0.31 | 22.6 | 0.235 | 0.30 | 24.6 | 0.260 |
| 0.81 | 0.84 | 0.29 | 21.7 | 0.206 | 0.24 | 24.1 | 0.275 | 0.24 | 27.6 | 0.297 |
| 0.89 | 0.68 | 0.27 | 21.6 | 0.257 | 0.36 | 24.2 | 0.230 | 0.30 | 26.8 | 0.267 |
| **$n_1$ = 15** | | | | | | | | | | |
| 0.98 | 0.47 | 0.27 | 17.8 | 0.193 | 0.27 | 20.7 | 0.192 | 0.27 | 23.4 | 0.204 |
| 0.92 | 0.48 | 0.28 | 20.7 | 0.237 | 0.20 | 24.7 | 0.217 | 0.26 | 27.4 | 0.217 |
| 0.93 | 0.76 | 0.25 | 21.9 | 0.178 | 0.32 | 25.1 | 0.172 | 0.24 | 27.7 | 0.227 |
| 0.81 | 0.84 | 0.20 | 23.2 | 0.162 | 0.17 | 27.5 | 0.246 | 0.17 | 29.7 | 0.261 |
| 0.89 | 0.68 | 0.23 | 23.6 | 0.240 | 0.28 | 26.2 | 0.198 | 0.28 | 28.4 | 0.210 |

(i) the coefficient of variation (CV) defined as $CV = \left(\sqrt{v(e)}/e\right)\times 100$; and

(ii) the length of the confidence intervals (CI's) $\left(e-1.96\sqrt{v(e)}, e+1.96\sqrt{v(e)}\right)$ given by $2\times 1.96\sqrt{v(e)}$

for comparing the relative performances of the alternative sampling procedures.

For the artificial population $\pi_A = 0.6202$. Table 1 outlines the performances of the alternative estimators for different choices of $n_1$, $p_1$ and $p_2$.

## Conclusion

Irrespective of the values of $n_1$, SRSWOR performs better than SRSWR in terms of the two criteria for comparison considered here and the RHC scheme turns out to be the best sampling scheme in terms of the criterion CV. As the values of $n_1$, i.e. the number of individuals replying '*yes*' to the initial direct question increases, improvement in the efficiency level of the estimator is observed for all three sampling designs.

This implies that for producing efficient estimators by applying the method discussed above, one has to choose the direct innocuous question judiciously so that more numbers of interviewees answer '*yes*' to the initial direct question. Thus, the extended method of estimation as discussed here may be effectively used in complex sample surveys for collection of information on sensitive attributes.

## References


Chaudhuri (2004). Christofides' randomized response technique in complex sample surveys. *Metrika*,*60*(3), 23-228.

Chaudhuri, A. (2002). Estimating sensitive proportions from randomized responses in unequal probability sampling. *Calcutta Statistical Association Bulletin, 52*, (205-208), 315-322.

Chaudhuri, A. (2001a). Using randomized response from a complex survey to estimate a sensitive proportion in a dichotomous finite population. J*ournal of Statistical Planning & Inference, 94*, 37 - 42.

Chaudhuri, A. (2001b). Estimating sensitive proportions from unequal probability sample using randomized responses. *Pakistan Journal of Statistics*, *17*(3), 259 - 270.

Chaudhuri, A., Adhikary, A.K. and Dihidar, S. (2000). Mean square error estimation in multi-stage sampling. *Metrika, 52*(2), 115-131.

Chaudhuri, A. & Mukerjee, R. (1988). *Randomized response: Theory and techniques*. Marcel Dekker Inc. N.Y.

Christofides, T. C. (2003). A generalized randomized response technique. *Metrika, 57*, 195 – 200.

Greenberg, B. G., Abul-Ela, Simmons, W. R. & Horvitz, D. G. (1969). The unrelated question randomized response model: theoritical framework. *Journal of the American Statistical Association, 64*, 520-539.

Horvitz, D. G., Shah, B. V., & Simmons, W. R. (1967). The unrelated question randomized response model. *Proceedings of the Social Statistics Section of the American Statistical Association.* 65-72.

Kim, Jong-Min & Warde, D. W. (2005). A mixed randomized response model. *Journal of Statistical Planning and Inference, 133*(1), 211-221.

Kim, Jong-Min & Warde, D. W. (2004). A stratified Warner's randomized response model. *Journal of Statistical Planning and Inference 120*, 155-165.

Kuk, A.Y.C. (1990). Asking sensitive question indirectly. *Biometrika, 77*, 436-438.

Moors, J. J. A. (1971). Optimization of the unrelated question randomized response model. *Journal of the American Statistical Association. 66*, 627-629.

Raj, D. (1968). *Sampling Theory*. Mc-graw Hill. N.Y.

Rao, J. N. K (1975). Unbiased variance estimation for multi-stage designs. Sankhya C, 37, 133-139.

Rao, J. N. K., Hartley, H. O., & Cochran, W. G. (1962). On a simple procedure of unequal probability sampling without replacement. *Journal of the Royal Statistical Society. B, 24*, 482-491.

Warner, S. L. (1965). Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association. 60*, 63-69.

Journal of Modern Applied Statistical Methods
November, 2005, Vol. 4, No. 2, 545-552

1538 – 9472/05/$95.00

# Nonparametric Pooling And Testing Of Preference Ratings For Full-Profile Conjoint Analysis Experiments

Rosa Arboretti G. Marco Marozzi
University of Ferrara

Luigi Salmaso
University of Padua

The problem of pooling customer preference ratings within a conjoint analysis experiment has been addressed. A method based on the nonparametric combination of rankings has been proposed to compete with the usual method based on the arithmetic mean. This method is nonparametric with respect to the underlying dependence structure and so no dependence model must be assumed. The two methods have been compared using Spearman's rank correlation coefficient and related test. Moreover, a further nonparametric testing method has been considered and proposed; this method takes both correlation and distance between ranks into account. By means of a simulation study it has been shown that the NPC Ranking method performs better than the arithmetic mean.

Key words: conjoint analysis, nonparametric inference, nonparametric combination, ranking.

## Introduction

In recent years, there has been a growing level of competitiveness in the offer of products. From a company point of view, one of the conditions of competitive success is a product's high level of correspondence to the varying requirements of the customer (Porter, 1998). Indeed, successful companies invest considerable resources and skills into planning and designing their products in order to incorporate the various requirements of customers into the product itself. The most competitive companies are currently those which use approaches and instruments designed to capture the so-called voice of customer (VOC). In order to do so, companies describe the product idea in terms which the customer can actually perceive. After its definition, the newly developed concept is tested by means of surveys in the field which aim to highlight which characteristics are most important to the customer and what his/her true intentions are in terms of purchasing/fruition. In this way, it is possible to modify the product concept before fully implementing it, in order to maximize adherence to the needs and expectations of potential customers by identifying specific segments of customers. The methods used are generally based on Conjoint Analysis (Dolan, 1993; Gustafsson, Herrmann, & Huber, 2001).

The term Conjoint Analysis refers to a set of predominantly statistical methodologies which aim to study customer choice models starting with opinions and preferences expressed by customers on various profiles of a product which is going to be developed. Even recent literature on such methodologies is rather fragmented and presents some critical elements, both in terms of the procedure for the definition of the survey design and in terms of the subsequent statistical analysis of gathered data (Gustafsson et al., 2001; Green, Krieger, & Wind, 2001). In particular, it should be noted that the arithmetic mean (whether weighted or

Rosa Arboretti Giancristofaro is an Assistant Professor of Statistics. M. Marozzi is a research fellow. L. Salmaso is an Associate Professor of Statistics. This research was conducted at Center for Modeling, Computing and Statistics of the University of Ferrara (CMCS-UNIFE, http://cmcs.unife.it). The main research interests of the authors are nonparametric statistical methods for marketing, business and finance. Send correspondence to M. Marozzi at m.marozzi@economia.unife.it.

not) is mainly used for pooling preference ratings.

One problem that may arise when customer preference ratings are averaged is the so-called majority fallacy (Moore, 1980). This problem occurs when the item chosen by the average customer is not the item chosen most often. For example, if half of the people like large cars and the other half like small ones, the average person would like medium-sized cars, even if no real person wants one. In this article, the problem of pooling preference ratings is addressed. In particular, the Nonparametric Combination of Rankings method (NPC Ranking; Lago & Pesarin, 2000; Arboretti, 2003) is used and extended. A simulation study is performed to show that the NPC Ranking method performs generally better than the arithmetic mean. To this end, Spearman's rank correlation coefficient is considered and a new nonparametric test $T_p$ for ranking comparison is proposed. Furthermore, to study the power of Spearman's $T_s$ and $T_p$ test in detecting ranking shifts, a further simulation study is performed.

## The pooling of preference ratings using the NPC Ranking methodology

In developing a new product/service a company may take $K \geq 2$ attributes (factors) with $P_1, P_2, \ldots, P_K$ values (levels) into consideration. Let $M = \prod_{k=1}^{K} P_k$ be the number of possible combinations of levels (treatments). For each treatment (product/service profile) a hypothetical dummy variable is defined as $d_{mkp}=1$, if the level of factor $k$ is $p$ for treatment $m$, otherwise $d_{mkp}=0$. It is assumed that customers assess the overall utility (worth) of a product/service by combining the separate utility value of each attribute. The additive model for total worth of profile $m$ is therefore:

$$Y_m = \sum_{k=1}^{K} \sum_{p=1}^{P_K} v_{kp} d_{mkp} + \varepsilon_m, \; m=1, \ldots, M,$$

where the coefficient $v_{kp}$ denotes the part-worth for level $p$ of factor $k$ and $\varepsilon_1, \ldots, \varepsilon_m$ are iid random residuals with 0 mean and $\sigma^2$ variance.

The full-profile method of treatment presentation is considered. Each treatment is described on a profile card. Let us consider $n$ customers who are asked to rate each of $M$ profiles on a scale of 1 to 10. The problem of how to obtain this ranking, i.e. how to pool customer preferences, is addressed in the article. Let $X_{mi}$ be the rate of profile $m$ given by customer $i$ ($i$=1, ..., $n$). Of course, if $X_{mi} > X_{m'i}$, then customer $i$ rates profile $m$ better than profile $m'$. In the literature this problem is solved by averaging customer ratings $\overline{X}_m = \frac{1}{n}\sum_{i=1}^{n} X_{im}$, $m$=1, ..., $M$, and profile $\tilde{m}$ such that $\overline{X}_{\tilde{m}} = \max\left(\overline{X}_1, \ldots, \overline{X}_M\right)$ is then the best profile ${}_A R_{\tilde{m}} = M$ (first rank position), profile $\hat{m}$ such that $\overline{X}_{\hat{m}} = \max_{\{i=1,\ldots,M, m\neq\tilde{m}\}}\left(\overline{X}_1, \ldots, \overline{X}_M\right)$ is the profile with the second rank position ${}_A R_{\hat{m}} = M-1$, and so on. For simplicity's sake, it is assumed that there are no ties in ranking positions.

An alternative way to pool preferences is based on the NPC ranking method (Lago & Pesarin, 2000). The procedure consists of three steps. In the first step, a score for profile $m$ is computed as follows:

$$\lambda_{mi} = \frac{\#\left(X_{mi} \geq X_{m'i}\right) + 0.5}{M+1},$$

where $\#\left(X_{mi} \geq X_{m'i}\right)$ indicates the rank transformation of $X_{mi}$. This step is repeated for each customer $i$ and profile $m$. With respect to relative rank transformation $\#\left(X_{mi} \geq X_{m'i}\right)/M$ of $X_{mi}$, 0.5 and 1 have been added respectively to the numerator and the denominator to obtain $\lambda_{mi}$ varying in the open interval (0, 1). The reason for such corrections is merely computational, in order to avoid numerical problems with logarithmic transformations later on. Note that the scores $\lambda_{mi}$ are one-to-one increasingly related with the ranks $\#\left(X_{mi} \geq X_{m'i}\right)$. By considering $\lambda_{mi}$s after the first step, it is straightforward to obtain a (partial) ranking of the $M$ profiles for each customer, but it is the global profile rank that is of interest.

In the second step, the scores that customers have assigned to profile $m$ are combined as follows:

$$C_m = -\sum_{i=1}^{n} \ln(1-\lambda_{mi}).$$

This step is repeated for the remaining $M$-1 profiles and it performs a nonparametric combination of customers' scores. In the last step, the (global) ranking for profile $m$ is computed as ${}_B R_m = \#(C_m \geq C_{m'})$. Of course profile $\tilde{m}$ with ${}_B R_{\tilde{m}} = M$ is the first rank position profile, $\hat{m}$ with ${}_B R_{\hat{m}} = M-1$ is the second one, and so on.

It should be noted that Fisher's omnibus combining function is used in the second step. Other possible combining functions are Liptak's $\sum_{i=1}^{n} \Phi^{-1}(\lambda_{mi})$, where $\Phi$ is the cumulative distribution function of a standard normal distribution, Tippett's $\max_{i\in\{1,\ldots,n\}}(\lambda_{mi})$, the logistic function $\sum_{i=1}^{n} \ln\left(\frac{\lambda_{mi}}{1-\lambda_{mi}}\right)$ and the additive function $\sum_{i=1}^{n} \lambda_{mi}$ (Lago & Pesarin, 2000). These combining functions (say $\psi$) satisfy three properties:

(i) $\psi$ is continuous in all $\lambda_{mi}$ arguments;
(ii) $\psi$ is non-decreasing in each $\lambda_{mi}$ argument: $\psi(\ldots,\lambda_{mi},\ldots) \geq \psi(\ldots,\lambda'_{mi},\ldots)$ if $0 < \lambda'_{mi} < \lambda_{mi} < 1$ for whatever $i \in \{1, \ldots, n\}$;
(iii) $\psi$ is symmetric with respect to permutations of the arguments: if $u_1, \ldots, u_n$ is a permutation of 1, …, $n$ then $\psi(\lambda_{m1},\ldots,\lambda_{mn}) \geq \psi(\lambda_{mu_1},\ldots,\lambda_{mu_n})$.

It should also be noted that a central feature of NPC Ranking is the possibility of assigning different degrees of importance to different types of customers. If the company developing the new product/service is more interested in a certain group of customers, it can assign them a weight of $0.5<w<1$ (and weight 1-$w$ to the remaining ones). This weighted approach is taken into account in step two of the procedure by computing $-\sum_{i=1}^{n} w_i \ln(1-\lambda_{mi})$ instead of $-\sum_{i=1}^{n} \ln(1-\lambda_{mi})$, where $w_i=w$ if customer $i$ belongs to the group of interest and $w_i=1-w$ if he does not. It is straightforward to consider more than two weights.

A comparison of preference pooling methods: Spearman's $I_s$ and $I_p$ indicators

To show that NPC Ranking generally performs better than the arithmetic mean in pooling preference ratings, a new indicator $I_p$ is presented and Spearman's rank correlation coefficient is also considered. Spearman's well-known correlation coefficient is defined as:

$$I_s = \frac{3\sum_{m=1}^{M} (R_m - \pi_m)^2}{M\left(M^2-1\right)},$$

where $R_m$ is the observed rank for profile $m$ and $\pi_m$ is the reference rank. $I_s$ takes values in [0, 1] and small values of $I_s$ are associated with similar values of $R_m$ and $\pi_m$. Another indicator is considered:

$$I_p = \sum_{m<m'} [k_{mm'}(1 + l_{mm'} + h_{mm'})],$$

where $K_{mm'} = 1$ when $(\pi_m - \pi_{m'})(R_m - R_{m'}) < 0$ otherwise $K_{mm'} = 0$, $l_{mm'} = |\pi_m - \pi_{m'}| - 1$ and $h_{mm'} = |R_m - R_{m'}| - 1$. $K_{mm'}$ takes into account whether or not the observed and reference rankings are coherent (i.e. positive correlated), $l_{mm'}$ ($h_{mm'}$) and it takes into account how far observed (reference) ranks are from each other. Values of $I_p$ close to 0 indicate that the observed ranking is very similar to the reference ranking. It is straightforward to show that

$$0 \leq I_p \leq \left[\frac{1}{6} M(M-1)(2M-1)\right]$$

and so

$$\frac{6\sum_{m<m'} [k_{mm'}(1 + l_{mm'} + h_{mm'})]}{M(M-1)(2M-1)}$$ takes values in [0, 1].

A simulation study has been performed. More precisely, a conjoint analysis experiment with three factors (I, II and III) each with two levels (+ and -) is considered. There are $2^3$=8 different profiles. It is assumed that the true profile ranking (reference ranking) is known. Consider table 1, where profile 8 is the best and profile 1 is the worst. Assume the eight profiles are presented to five customers.

Table 1 Reference ranking of profiles

| Profile | Factors | | | Preference Rating |
|---|---|---|---|---|
| | I | II | III | |
| 1 | - | - | - | 1 |
| 2 | - | - | + | 2 |
| 3 | - | + | - | 3 |
| 4 | - | + | + | 4 |
| 5 | + | - | - | 5 |
| 6 | + | - | + | 6 |
| 7 | + | + | - | 7 |
| 8 | + | + | + | 8 |

Customer profile ratings are simulated by adding to the reference ranking a random error taken from continuous distributions such as normal N(0,1), exponential exp(1), uniform U(0,1) and Cauchy Cau(0,1), and from discrete distributions such as binomial Bi(8,0.5) and Poisson $P(1)$: $Y_{mi}=\pi_m+\varepsilon_{mi}$, where $Y_{mi}$ is the rate of profile $m$ for customer $i$, $\mu_m$ is the reference rank/rate of profile $m$ ($\pi_m=m$) and $\varepsilon_{mi}$ is the random error denoting the distance between $Y_{mi}$ and the reference value. $[Y_{mi}]$, $m$=1,…,8 and $i$=1,…,5 is a 8x5 matrix of real numbers. By computing the arithmetic mean or applying the NPC Ranking, two 8x1 vectors of ranks ${}_A\underline{R}$ or ${}_B\underline{R}$ are obtained. 1000 matrixes are randomly generated and 1000 pairs of vectors are then computed. Let ${}_A\underline{R}^{(c)}$ and ${}_B\underline{R}^{(c)}$ indicate the vector of ranks obtained by using the arithmetic mean and the NPC Ranking for simulation $c$($c$=1,…,1000). Let $\underline{\pi}'=(1,2,...,8)$. In order to establish which of the two methods is better, Spearman's $I_s$ and $I_p$ indicators are computed.

More precisely, the two methods are compared using the $I_p$ indicator by computing

$$ {}_{AB}Q_p'=\#\left(I_p\left({}_B\underline{R}^{(c)},\underline{\pi}\right)\le I_p\left({}_A\underline{R}^{(c)},\underline{\pi}\right)\right)/1000, $$

the proportion of simulations in which $I_p\left({}_B\underline{R}^{(c)},\underline{\pi}\right)$ is less than or equal to $I_p\left({}_A\underline{R}^{(c)},\underline{\pi}\right)$. If this proportion is greater than ${}_{AB}Q_p''=\#\left(I_p\left({}_A\underline{R}^{(c)},\underline{\pi}\right)\le I_p\left({}_B\underline{R}^{(c)},\underline{\pi}\right)\right)/1000$, then the NPC Ranking method is preferable because rankings obtained using this method are more similar to the reference ranking than those obtained using the arithmetic mean. It is worth noting that ${}_{AB}Q_p'+{}_{AB}Q_p''>1$ because the equalities are counted both in ${}_{AB}Q_p'$ and ${}_{AB}Q_p''$. A similar comparison is performed by considering the $I_s$ indicator and computing ${}_{AB}Q_s'=\#\left(I_s\left({}_B\underline{R}^{(c)},\underline{\pi}\right)\le I_s\left({}_A\underline{R}^{(c)},\underline{\pi}\right)\right)/1000$ and ${}_{AB}Q_s''=\#\left(I_s\left({}_A\underline{R}^{(c)},\underline{\pi}\right)\le I_s\left({}_B\underline{R}^{(c)},\underline{\pi}\right)\right)/1000$. It is also of some interest to compare $I_p$ and $I_s$ indicators themselves. To this end, ${}_{ps}Q_A'$, ${}_{ps}Q_A''$, ${}_{ps}Q_B'$ and ${}_{ps}Q_B''$ are computed as follows:

$$ {}_{ps}Q_A'=\#\left(I_p\left({}_A\underline{R}^{(c)},\underline{\pi}\right)\le I_s\left({}_A\underline{R}^{(c)},\underline{\pi}\right)\right)/1000, $$
$$ {}_{ps}Q_A''=\#\left(I_s\left({}_A\underline{R}^{(c)},\underline{\pi}\right)\le I_p\left({}_A\underline{R}^{(c)},\underline{\pi}\right)\right)/1000 \text{ and} $$
$$ {}_{ps}Q_B'=\#\left(I_p\left({}_B\underline{R}^{(c)},\underline{\pi}\right)\le I_s\left({}_B\underline{R}^{(c)},\underline{\pi}\right)\right)/1000, $$
$$ {}_{ps}Q_B''=\#\left(I_s\left({}_B\underline{R}^{(c)},\underline{\pi}\right)\le I_p\left({}_B\underline{R}^{(c)},\underline{\pi}\right)\right)/1000. $$

If ${}_{ps}Q_A'\ge{}_{ps}Q_A''$ then $I_p$ is better than $I_s$ when the average method is used. If ${}_{ps}Q_B'\ge{}_{ps}Q_B''$ then $I_p$ is better than $I_s$ when the NPC Ranking method is used.

Table 2 Simulation results

| Distribution | ${}_{AB}Q_p'$ | ${}_{AB}Q_p''$ | ${}_{AB}Q_s'$ | ${}_{AB}Q_s''$ | ${}_{ps}Q_A'$ | ${}_{ps}Q_A''$ | ${}_{ps}Q_B'$ | ${}_{ps}Q_B''$ |
|---|---|---|---|---|---|---|---|---|
| Normal | 0.531 | 0.771 | 0.526 | 0.772 | 1.000 | 0.076 | 1.000 | 0.125 |
| Exponential | 0.650 | 0.447 | 0.592 | 0.461 | 0.996 | 0.015 | 0.991 | 0.021 |
| Uniform | 0.441 | 0.757 | 0.439 | 0.758 | 1.000 | 0.109 | 1.000 | 0.017 |
| Cauchy | 0.649 | 0.378 | 0.655 | 0.385 | 0.771 | 0.296 | 0.662 | 0.412 |
| Binomial | 0.559 | 0.487 | 0.600 | 0.436 | 0.844 | 0.196 | 0.906 | 0.112 |
| Poisson | 0.534 | 0.528 | 0.592 | 0.461 | 0.936 | 0.111 | 0.961 | 0.005 |

As reported in table 2, NPC Ranking is better than the arithmetic mean for Exponential, Cauchy, Binomial and Poisson distributions, using both $I_p$ and $I_s$ indicators. Only for normal and uniform distributions the arithmetic mean (as can be expected) is better than NPC Ranking. As regards indicator comparisons, $I_p$ is clearly better than $I_s$ when the arithmetic mean is used as well as when NPC Ranking is used, because ${}_{ps}Q_A'$ and ${}_{ps}Q_B'$ are greater than ${}_{ps}Q_A''$ and ${}_{ps}Q_B''$ respectively, for all considered distributions.

In order to obtain further insight into $I_p$ and $I_s$ indicator comparison, instead of reference ranking $\underline{\pi}'=(1,2,3,4,5,6,7,8)$, ranking $\underline{\gamma}'=(1,2,3,6,4,5,7,8)$ has been considered in Monte Carlo simulations. The reference ranking is still $\underline{\pi}$, but now random errors $\varepsilon_{mi}$ are added to $\underline{\gamma}$ and not to $\underline{\pi}$. The power simulation study is set out as follows: indicators $I_s$ and $I_p$ are considered as test statistics within a permutation framework, i.e.:

$$T_p = \#\left(I_p^* \geq I_p^{obs}\right)/B$$

and

$$T_s = \#\left(I_s^* \geq I_s^{obs}\right)/B,$$

where $I_p^*$ and $I_s^*$ are obtained by a random permutation of the observed ranking, $I_p^{obs}$ and $I_s^{obs}$ are the values of indicators $I_s$ and $I_p$ calculated by comparing the observed ranking with the reference ranking, and $B$ is the number of all possible permutations in a $2^3$ factorial design (i.e. 8!=40320 permutations).

Tables 3-5 report the results of the simulation study when errors are normal N(0,1), uniform U(0,1), exponential exp(1), Cauchy Cau(0,1), binomial Bi(8,0.5) and Poisson $P(1)$. $T_{sA}$ and $T_{sB}$ ($T_{pA}$ and $T_{pB}$) indicate that the test statistic used is in both cases $T_s$ ($T_p$); although the global ranking is obtained either using the arithmetic mean (indicated by the subscript A) or the NPC method (indicated by the subscript B). Simulation results show that a global ranking obtained using the arithmetic mean allows both test statistics $T_s$ and $T_p$ to gain more power than when the global ranking is obtained using the NPC method, when the underlying error distribution is either normal or uniform. When the error distribution is binomial and Poisson, the power is very similar between the two global ranking procedures.

On the contrary, the power is greater for both $T_s$ and $T_p$ when the global ranking is obtained using the NPC method when the underlying error distribution is exponential or Cauchy. However, it is important to emphasize that both $T_s$ and $T_p$ tests are unbiased, because they indicate that the ranking under $H_1$ is different with respect to the reference ranking, even when the nominal significance level is very small. Moreover, they are consistent tests (for more details see e.g. Pesarin 2001)

## Conclusion

The problem of pooling customer preference ratings within a conjoint analysis experiment has

Table 3 Estimated power, normal and uniform error distributions

| | normal | | | | uniform | | | |
|---|---|---|---|---|---|---|---|---|
| $\alpha$ | $T_{sA}$ | $T_{sB}$ | $T_{pA}$ | $T_{pB}$ | $T_{sA}$ | $T_{sB}$ | $T_{pA}$ | $T_{pB}$ |
| 0.010 | 0.030 | 0.032 | 0.032 | 0.038 | 0.163 | 0.135 | 0.163 | 0.135 |
| 0.025 | 0.196 | 0.212 | 0.206 | 0.220 | 0.464 | 0.384 | 0.493 | 0.401 |
| 0.050 | 0.716 | 0.634 | 0.728 | 0.672 | 0.776 | 0.656 | 0.795 | 0.672 |
| 0.075 | 0.896 | 0.838 | 0.896 | 0.836 | 0.907 | 0.811 | 0.907 | 0.813 |
| 0.100 | 0.956 | 0.886 | 0.958 | 0.888 | 0.943 | 0.870 | 0.948 | 0.878 |
| 0.200 | 1.000 | 0.996 | 1.000 | 0.996 | 0.996 | 0.981 | 0.997 | 0.986 |
| 0.300 | 1.000 | 1.000 | 1.000 | 1.000 | 0.999 | 0.994 | 0.999 | 0.995 |
| 0.400 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 0.998 | 1.000 | 0.998 |
| 0.500 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| 0.600 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| 0.700 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| 0.800 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| 0.900 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |

Table 4 Estimated power, exponential and Cauchy error distributions

| | exponential | | | | Cauchy | | | |
|---|---|---|---|---|---|---|---|---|
| $\alpha$ | $T_{sA}$ | $T_{sB}$ | $T_{pA}$ | $T_{pB}$ | $T_{sA}$ | $T_{sB}$ | $T_{pA}$ | $T_{pB}$ |
| 0.010 | 0.017 | 0.020 | 0.017 | 0.020 | 0.054 | 0.139 | 0.054 | 0.139 |
| 0.025 | 0.051 | 0.053 | 0.062 | 0.059 | 0.224 | 0.392 | 0.238 | 0.413 |
| 0.050 | 0.137 | 0.150 | 0.147 | 0.160 | 0.419 | 0.649 | 0.433 | 0.673 |
| 0.075 | 0.218 | 0.220 | 0.220 | 0.221 | 0.537 | 0.805 | 0.538 | 0.806 |
| 0.100 | 0.286 | 0.272 | 0.304 | 0.279 | 0.590 | 0.875 | 0.595 | 0.878 |
| 0.200 | 0.525 | 0.480 | 0.558 | 0.504 | 0.738 | 0.968 | 0.743 | 0.975 |
| 0.300 | 0.652 | 0.625 | 0.675 | 0.647 | 0.819 | 0.990 | 0.823 | 0.991 |
| 0.400 | 0.772 | 0.751 | 0.774 | 0.755 | 0.892 | 0.996 | 0.889 | 0.996 |
| 0.500 | 0.841 | 0.830 | 0.842 | 0.831 | 0.929 | 0.999 | 0.929 | 0.998 |
| 0.600 | 0.898 | 0.883 | 0.902 | 0.895 | 0.955 | 1.000 | 0.959 | 1.000 |
| 0.700 | 0.936 | 0.928 | 0.937 | 0.928 | 0.976 | 1.000 | 0.976 | 1.000 |
| 0.800 | 0.963 | 0.962 | 0.964 | 0.964 | 0.990 | 1.000 | 0.989 | 1.000 |
| 0.900 | 0.986 | 0.984 | 0.987 | 0.984 | 0.996 | 1.000 | 0.996 | 1.000 |
| 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |

Table 5 Estimated power binomial and Poisson error distributions

| | binomial | | | | Poisson | | | |
|---|---|---|---|---|---|---|---|---|
| $\alpha$ | $T_{sA}$ | $T_{sB}$ | $T_{pA}$ | $T_{pB}$ | $T_{sA}$ | $T_{sB}$ | $T_{pA}$ | $T_{pB}$ |
| 0.010 | 0.492 | 0.408 | 0.506 | 0.409 | 0.012 | 0.020 | 0.018 | 0.020 |
| 0.025 | 0.924 | 0.860 | 0.933 | 0.869 | 0.050 | 0.065 | 0.075 | 0.068 |
| 0.050 | 0.994 | 0.985 | 0.998 | 0.987 | 0.153 | 0.159 | 0.196 | 0.180 |
| 0.075 | 1.000 | 0.998 | 1.000 | 0.999 | 0.244 | 0.264 | 0.273 | 0.266 |
| 0.100 | 1.000 | 0.999 | 1.000 | 0.999 | 0.312 | 0.329 | 0.352 | 0.350 |
| 0.200 | 1.000 | 1.000 | 1.000 | 1.000 | 0.521 | 0.559 | 0.574 | 0.582 |
| 0.300 | 1.000 | 1.000 | 1.000 | 1.000 | 0.656 | 0.681 | 0.699 | 0.696 |
| 0.400 | 1.000 | 1.000 | 1.000 | 1.000 | 0.759 | 0.785 | 0.790 | 0.788 |
| 0.500 | 1.000 | 1.000 | 1.000 | 1.000 | 0.833 | 0.857 | 0.855 | 0.855 |
| 0.600 | 1.000 | 1.000 | 1.000 | 1.000 | 0.902 | 0.897 | 0.923 | 0.908 |
| 0.700 | 1.000 | 1.000 | 1.000 | 1.000 | 0.942 | 0.936 | 0.947 | 0.936 |
| 0.800 | 1.000 | 1.000 | 1.000 | 1.000 | 0.967 | 0.962 | 0.969 | 0.966 |
| 0.900 | 1.000 | 1.000 | 1.000 | 1.000 | 0.985 | 0.993 | 0.987 | 0.994 |
| 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |

been addressed. A nonparametric method based on the nonparametric combination of rankings has been proposed to compete with the traditional method based on the arithmetic mean. In order to compare these two methods, Spearman's rank correlation coefficient has been considered. Moreover, a further nonparametric method has been considered and proposed. This method takes both correlation and distance between ranks into account. By means of a simulation study, it has been shown that the NPC Ranking method performs better than the arithmetic mean.

The NPC Ranking procedure requires only one assumption in terms of variables, i.e. the inequality $X_{mi} \geq X_{m'i}$ means that customer *i* rates profile *m* better than profile *m'*. It should also be noted that a central feature of NPC Ranking is the possibility of assigning different degrees of importance to different types of customers.

Fisher's omnibus combining function has been used. Other combining functions, such as Liptak's, Tippett's, the logistic and additive functions may also be used (for more details see Lago & Pesarin, 2000).

A power simulation study showed that permutation tests based on $I_s$ and $I_p$ statistics clearly indicate that the ranking under $H_1$ is different with respect to the reference ranking, even when the nominal significance level, chosen for the comparison, is very small.

Within a conjoint analysis experiment, practitioners should take the NPC Ranking method into account for the pooling of customer preference ratings. A computer program to perform the analysis is available at the website http://cmcs.unife.it.

## References

Giancristofaro, R. A. (2003). A new conjoint analysis procedure with application to marketing research. *Communications in Statistics - Theory and Methods, 32*, 2271-2286.

Dolan, R. J. (1993). *Managing the new product development process: Cases and notes*. Addison-Wesley: Reading, MA.

Green, P. E., Krieger, A. M., & Wind, Y. (2001). Thirty years of conjoint analysis: Reflections and prospects. *Interfaces, 31*(3), S56-S73.

Gustafsson, A., Herrmann, A., & Huber, F. (2001). *Conjoint measurement: Methods and applications.* Springer: Berlin.

Lago, A., & Pesarin, F. (2000). Nonparametric combination of dependent rankings with application to the quality assessment of industrial products. *Metron, LVIII,* 39-52.

Moore, W. L. (1980). Levels of aggregation in conjoint analysis: An empirical comparison. *Journal of Marketing Research, 11*, 516-523.

Porter, M. E. (1998). *Competitive strategy: Techniques for analyzing industries and competitors.* Free Press: New York.

Journal of Modern Applied Statistical Methods
November, 2005, Vol. 4, No. 2, 553-566

1538 – 9472/05/$95.00

# Statistical Model And Estimation Of The Optimum Price For A Chain Of Price Setting Firms

Chengjie Xiong
Division of Biostatistics
Washington University in St. Louis

Kejun Zhu
College of Management
China University of Geosciences (Wuhan)

A stochastic approach is used to model the economics of a chain of price setting firms. It is assumed that these firms have fixed capacities in their products, but random demands for their products. The optimum price, the optimum revenue, and the expected marginal revenue at a given price are investigated. The method of maximum likelihood is used to provide both point and confidence interval estimates. The coverage probabilities of confidence interval estimates based on a simulation study are presented.

Key words: Asymptotic confidence interval; capacity; gamma distribution; marginal revenue; maximum likelihood estimate (MLE); optimum revenue; Poisson distribution.

## Introduction

Fixed capacity is very common in businesses. For example, an established hotel must operate with a fixed number of rooms; and an established restaurant has a fixed number of seats. While the capacity is fixed for many firms, the demand for their products is uncertain. By their very nature, the hotel and the restaurant cannot respond to the uncertain demand by inventory adjustments, nor for that matter, by using high priced resources to temporarily increase production when demand is high. The most important goal for these firms is to choose a price that maximizes their expected profits under random demand for a fixed capacity. Many authors have studied the problem of firm decision making when demand for the product is uncertain. Epstein (1978) and Turnovsky (1973), provided the classic approach to the problem. Scott, Highfill, and Sattler (1988) and Balvers and Miller (1992) studied several production side questions such as the derived factor demand with capacity constraints. Flacco and Kroetch (1986) and Booth (1990), investigated the production levels and/or inventory adjustments in the decision making.

In this article, it is assumed that these firms operate as monopolies and are risk neutral. It is also assumed that capacity is a strict upper bound on the provision of service and must be set before the demand is arriving. With these same assumptions, Scott, Sattler, and Highfill (1995) studied the optimum price for a single firm when the demand is random. Highfill, Quigg, Sattler, and Scott (2000) investigated the problem of capacity decision for a single firm when the product demand is uncertain. Here, a chain of price setting firms with random demands are considered and the optimum price and its estimation applicable to a population of firms is studied. There are two levels of uncertainty in the demand side now: one is the demand uncertainty for any given firm in the chain, the other is the demand uncertainty from firm to firm in the chain. Therefore, two statistical models are needed to model the demand at two different stages, one for a given

Financial support for this study was provided in part to Kejun Zhu by the National Natural Science Foundation, Grant #70273044, of the People's Republic of China. The funding agreement ensured the authors' independence in designing the study, interpreting the data, writing and publishing the report. Address correspondence to the first author at: chengjie@wubios.wustl.edu.

firm and the other for across the firms in the chain.

A simple example of the kind of problem under consideration in this article is a chain of hotels which operates with the number of rooms as the strict upper bound for the service. The variability in demand will cause the hotels to experience excess capacity and excess demand. Both excesses will depend on the capacity of the firms and the probability distribution for the demand. The question answered in this article is, for a randomly selected hotel in the chain, how the price should be set and estimated so that the maximum profit can be achieved.

In the following section, the statistical model is proposed and the optimum price is studied by assuming that all the parameters are known in the model. Also, the effect of capacity on the optimum price is considered. Next, the estimation for the model parameters is provided and asymptotic confidence intervals for the optimum price, the optimum revenue, and the expected marginal revenue at a given price are presented.

It is convenient to use a chain of hotels as the economic reference of a chain of firms in this article. The results in this article apply to all businesses where capacity is a strict upper bound on the provision of service and the demand is random.

## The Model and the Optimum Price

For a given hotel $H$ in a chain of hotels, let $Y|_H$ be the number of people to rent a room. The uncertain number of people to rent a room is treated as a standard queuing problem with the quantity demanded a random variable distributed as Poisson whose mean is $\lambda_H$, i.e.,

$$P(Y|_H = y) = e^{-\lambda_H}\frac{\lambda_H^y}{y!},$$

for $y$=0,1,2,....

In order to model the demand variability from hotel to hotel in the chain, it is assumed that the population of demand mean $\lambda_H$ of $Y|_H$ from the hotels follows a Gamma distribution with index $\alpha$>0 and scale parameter $\theta$>0, i.e., $\lambda_H$ is distributed according to the probability density function

$$f(\lambda) = \frac{1}{\Gamma(\alpha)\theta^a}\lambda^{a-1}e^{-\frac{\lambda}{\theta}}.$$

It is also assumed that $\alpha$ is independent of price and $\theta$ is linearly and inversely related to unit price $p$, i.e.,

$$\theta = a + bp,$$

where $a$>0 and $b$<0 are two constants.

Let $Y$ denote the number of people to rent a room from a hotel randomly sampled from the chain. The probability distribution of $Y$ is then given by

$$P(Y=y) = \int_0^\infty e^{-\lambda}\frac{\lambda^y}{y!}f(\lambda)d\lambda$$
$$= \frac{\alpha(\alpha+1)...(\alpha+y-1)}{y!(1+\theta)^\alpha}\left(\frac{\theta}{\theta+1}\right)^y,$$

for $y$=0,1,2,....

The distribution of $Y$ is the well known negative binomial distribution when $\alpha$ is a positive integer. The index parameter $\alpha$ in the model allows for the flexibility to choose different densities in the Gamma family to model the demand variability across the hotels. Let $P_\alpha$ denote the probability of events instead of just $P$ to indicate the dependence of the probabilities on the parameter $\alpha$. Notice that $EY = E(EY|_H) = E\lambda_H = \alpha\theta = \alpha a + \alpha bp$, $a$>0, $b$<0. The expected number of people to rent a room from this randomly selected hotel in the chain is also linearly and inversely related to price $p$.

Suppose that $c$ is the capacity number of rooms in the hotel. Let $X$ be the unit sales of the hotel. Then

$$X = \begin{cases} Y, & Y \le c \\ c, & Y > c \end{cases}.$$

Therefore

$$P_a(X = x) = \begin{cases} P(Y = x), & x < c \\ 1 - \sum_{x=0}^{c-1} P(Y = x), & x = c \end{cases} = \begin{cases} \frac{\alpha(\alpha+1)\ldots(\alpha+x-1)}{x!(1+\theta)^{\alpha}} \left(\frac{\theta}{\theta+1}\right)^x, & x < c \\ 1 - \sum_{k=0}^{c-1} \frac{\alpha(\alpha+1)\ldots(\alpha+k-1)}{k!(1+\theta)^{\alpha}} \left(\frac{\theta}{\theta+1}\right)^k, & x = c. \end{cases}$$

When the demand is random and the capacity is fixed, there are positive probabilities that excess demand (denoted by *ED*) and excess capacity (denoted by *EC*) occurs. It is straightforward to find the probability of excess demand and the probability of excess capacity as

$$P_\alpha(ED) = \sum_{x=c+1}^{\infty} \frac{\alpha(\alpha+1)\ldots(\alpha+x-1)}{x!(1+\theta)^{\alpha}} \left(\frac{\theta}{\theta+1}\right)^x$$

and

$$P_\alpha(EC) = \sum_{x=0}^{c-1} \frac{\alpha(\alpha+1)\ldots(\alpha+x-1)}{x!(1+\theta)^{\alpha}} \left(\frac{\theta}{\theta+1}\right)^x,$$

respectively. Two integral representations of these probabilities and their derivatives are given, which will be used later in the article:

$$P_\alpha(ED) = \frac{1}{B(\alpha, c+1)} \theta^{c+1} \int_1^{\infty} \frac{t^c}{(1+t\theta)^{\alpha+c+1}} dt \qquad (1)$$

$$P_\alpha(EC) = \frac{1}{B(\alpha, c)} \theta^{c} \int_1^{\infty} \frac{t^{c-1}}{(1+t\theta)^{\alpha+c}} dt \qquad (2)$$

$$\frac{dP_\alpha(ED)}{d\theta} = \frac{\theta^c}{B(\alpha, c+1)(1+\theta)^{\alpha+c+1}}, \qquad (3)$$

$$\frac{dP_\alpha(EC)}{d\theta} = -\frac{\theta^{c-1}}{B(\alpha, c)(1+\theta)^{\alpha+c}}, \qquad (4)$$

where $B(\alpha,\beta)=\Gamma(\alpha)\Gamma(\beta)/\Gamma(\alpha+\beta)$ is the Beta function. (1) and (2) can be obtained by using equation (6) and (7) from Highfill, Quigg, Sattler and Scott (2000) and applying Fubini's Theorem for the exchange of integrals. (3) and (4) can be obtained by directly taking derivatives from (1) and (2), respectively. Combining (3) and (4) further gives

$$\theta\alpha \frac{dP_{\alpha+1}(EC)}{dp} + c\frac{dP_\alpha(ED)}{dp} = 0. \qquad (5)$$

The expected unit sales of the hotel is then

$$EX = \sum_{x=0}^{c} x \frac{\alpha(\alpha+1)\ldots(\alpha+x-1)}{x!(1+\theta)^{\alpha}} \left(\frac{\theta}{\theta+1}\right)^x + cP_\alpha(ED) = \theta\alpha P_{\alpha+1}(EC) + cP_\alpha(ED).$$

Therefore, the expected unit sales of the hotel contain two parts, one is the expected demand $\theta\alpha$ multiplied by the probability of excess capacity at index $\alpha+1$, and the other is the capacity $c$ multiplied by the probability of excess demand.

For any hotel in the chain, the problem is to find the optimum price that maximizes the expected profit based on the fixed capacity. It is assumed that these hotels have a constant non-stochastic marginal cost function. Therefore, as pointed out by Highfill, Quigg, Sattler and Scott (2000), the constant can be set at zero since the analysis is not materially affected by the choice of this constant (i.e., one can concentrate on the expected revenue). Let $R$ be the revenue for a randomly selected hotel, i.e., $R=Xp$. The expected revenue is

$$ER = p\theta\alpha P_{\alpha+1}(EC) + pcP_\alpha(ED).$$

Therefore, the expected revenue of the hotel contains two parts too, one is the expected revenue for all demand multiplied by the probability of excess capacity at index $\alpha+1$, the other is the revenue at capacity multiplied by the probability of excess demand. The following theorem gives the optimum price which maximizes the expected revenue.

Theorem 1

The optimum price $p^*$ is the unique solution to the equation:

$$(\theta\alpha + pb\alpha)P_{\alpha+1}(EC) + cP_{\alpha}(ED) = 0. \quad (6)$$

In addition, $p^* > -\frac{a}{2b}$ and $\lim_{c\to\infty} p^* = -\frac{a}{2b}$. Refer to the Appendix for the proof.

Let $\theta^* = a + bp^*$. Denote the optimum expected revenue, the probability of excess capacity and the probability of excess demand at optimum price $p^*$ by $ER^*$, $P_{\alpha}(EC^*)$ and $P_{\alpha}(ED^*)$, respectively. Recall that the expected demand is $EY=\alpha(a+bp)$ and the expected unit sales is $EX = \theta\alpha P_{\alpha+1}(EC) + cP_{\alpha}(ED)$. It is always true that $EX<EY$, because $X<Y$. If the capacity is hypothetically infinity, then $X= Y$ and $ER=p\alpha(a+bp)$. Therefore $ER$ attains the maximum $-\alpha a^2/(4b)$ when price $p = -a/(2b)$. Theorem 1 indicates that in real world business applications where the capacity $c$ is always a finite number, the optimum price for the hotel is always larger than that in the limiting capacity situation, and the optimum revenue for the hotel is always smaller than that in the limiting capacity situation. But, as the capacity increases, the optimum price and the optimum revenue approach their limiting values respectively.

Scott, Sattler and Highfill (1995) defined the expected marginal revenue ($EMR$) as $EMR = dER/dEX$. The expected marginal revenue measures the change in expected revenue for a given change in expected unit sales. Notice that $dEX/dp = b\alpha P_{\alpha+1}(EC)$.

Therefore,

$$EMR = \frac{dER}{dp} / \frac{dEX}{dp}$$

$$= \frac{(\theta\alpha + pb\alpha)P_{\alpha+1}(EC) + cP_{\alpha}(ED)}{b\alpha P_{\alpha+1}(EC)}$$

$$= 2p + \frac{a}{b} + \frac{cP_{\alpha}(ED)}{b\alpha P_{\alpha+1}(EC)}.$$

As the capacity approaches infinity, $P_{\alpha}(ED)$ approaches 0 and $P_{\alpha+1}(EC)$ approaches 1. Therefore, the expected marginal revenue approaches the standard marginal revenue under linear demand.

In order to understand the dependence of $p^*$ on capacity $c$, the effect of an additional unit of capacity on the optimum price $p^*$ is analyzed. Suppose that the hotel capacity is increased from $c$ to $c$+1. Assume that the optimum price is changed from $p^*$ to $p^* + \Delta p^*$ and the optimum expected revenue is changed from $ER^*$ to $ER^* + \Delta ER^*$ accordingly. The following theorem presents the effect of an additional unit of capacity on $p^*$ and $ER^*$.

Theorem 2

(1) There exists a constant $C$ depending only on $a$ and $\alpha$ such that if $c>C$ then $\Delta p^* < 0$. In addition, $\lim_{c\to\infty} \Delta p^* = 0$.

(2) $\Delta ER^* > 0$ for every $c\geq 1$. In addition, $\lim_{c\to\infty} \Delta ER^* = 0$. Refer to the appendix for the proof.

Theorem 2 indicates that the optimum price will decrease after the capacity increases to a certain level, but the drop in optimum price for each unit increase of capacity approaches 0 when the capacity approaches infinity. On the other hand, when the capacity increases, there is always a positive probability that the extra unit will be taken by customers. Therefore, the optimum expected revenue will always increase. But the increase in the optimum revenue for each unit increase of capacity also approaches 0 when the capacity approaches infinity.

Estimation and Inference

In the previous section, the optimum price and optimum revenue were discussed when all model parameters are assumed known. In this section, it is first assumed that the index parameter $\alpha$ is known in the model and the

estimate of the unknown parameters $a$ and $b$ is discussed using data collected from the hotels in the chain. Suppose that hotels operate independently and $n$ hotels in the chain have been observed, resulting in the data $(p_i, c_i, x_i, \delta_i)$, $i$=1,2,...,$n$, where $p_i, c_i, x_i$ are the price, the capacity, and the unit sales of the $i$-th hotel, respectively, and

$$\delta_i = \begin{cases} 1, & y_i \le c_i \\ 0, & y_i > c_i \end{cases},$$

where $y_i$ is the demand of the $i$-th hotel. The maximum likelihood estimators for $a$ and $b$ maximize the likelihood function:

$$L(\alpha, a, b) \propto \prod_{i=1}^{n} \left\{ \frac{\theta_i^{\delta_i x_i}}{(\theta_i+1)^{\delta_i(x_i+\alpha)}} [P_\alpha(ED)_i]^{1-\delta_i} \right\},$$

where

$$P_\alpha(ED)_i = \sum_{x=0}^{c_i+1} \frac{\alpha(\alpha+1)...(\alpha+x-1)}{x!(1+\theta_i)^\alpha} \left(\frac{\theta_i}{\theta_i+1}\right)^x,$$

and

$$\theta_i = a + bp_i, i = 1,2,..., n.$$

Because

$$\ln L \propto \left[ \sum_{i=1}^{n} \delta_i x_i \ln\theta_i - \delta_i(\alpha+x_i)\ln(\theta_i+1) \right] + \sum_{i=1}^{n} \left[ (1-\delta_i) \ln P_\alpha(ED)_i \right],$$

the maximum likelihood estimators of $a$ and $b$ solve the following system of equations:

$$\frac{\partial \ln L}{\partial a} = \sum_{i=1}^{n} \left\{ \frac{\delta_i x_i}{\theta_i} - \frac{\delta_i(\alpha+x_i)}{\theta_i+1} \right\} + \sum_{i=1}^{n} \left\{ \frac{(1-\delta_i)\theta_i^{c_i}}{B(\alpha, c_i+1)(1+\theta_i)^{\alpha+c_i+1} P_\alpha(ED)_i} \right\} = 0,$$

$$\frac{\partial \ln L}{\partial b} = \sum_{i=1}^{n} p_i \left\{ \frac{\delta_i x_i}{\theta_i} - \frac{\delta_i(\alpha+x_i)}{\theta_i+1} \right\} + \sum_{i=1}^{n} p_i \left\{ \frac{(1-\delta_i)\theta_i^{c_i}}{B(\alpha, c_i+1)(1+\theta_i)^{\alpha+c_i+1} P_\alpha(ED)_i} \right\} = 0.$$

It is assumed that there are at least two different prices in the data and $n$ is large enough so that not all $\delta_i$ are 0. Then, the maximum likelihood estimates uniquely exist. However, except for trivial situations, the solutions to the system cannot be found in a close form. But numerical methods as discussed in Press, Flannery, Teukolsky, and Vetterling (1986) such as Newton-Raphson method can be easily implemented to find the solutions. The symbols $\hat{a}$ and $\hat{b}$ are used to denote the maximum likelihood estimator for $a$ and $b$, respectively. Let

$$\Sigma^{-1} = \begin{pmatrix} \sigma_{11}' & \sigma_{12}' \\ \sigma_{12}' & \sigma_{22}' \end{pmatrix} = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix}^{-1},$$

where

$$\sigma_{11} = E\left( -\frac{\partial^2 \ln L}{\partial a^2} \right) = \sum_{i=1}^{n} \left\{ \frac{\alpha P_{\alpha+1}(EC)_i}{\theta_i} - \frac{\alpha[(1-P_\alpha(ED)_i) + \theta_i P_{\alpha+1}(EC)_i]}{(\theta_i+1)^2} \right\} - \sum_{i=1}^{n} \left\{ \frac{\theta_i^{c_i-1}}{B(\alpha, c_i+1)} \left[ \frac{c_i - (\alpha+1)\theta_i}{(1+\theta_i)^{\alpha+c_i+2}} \right] \right\} + \sum_{i=1}^{n} \left\{ \frac{\theta_i^{2c_i}}{B^2(\alpha, c_i+1)(\theta_i+1)^{2(\alpha+c_i+1)} P_\alpha(ED)_i} \right\},$$

$$\sigma_{12} = E\left( -\frac{\partial^2 \ln L}{\partial a \partial b} \right)$$

$$= \sum_{i=1}^{n} \left\{ \frac{p_i \alpha P_{\alpha+1}(EC)_i}{\theta_i} \right\}$$

$$- \sum_{i=1}^{n} \left\{ \frac{p_i \alpha [(1 - P_\alpha(ED)_i) + \theta_i P_{\alpha+1}(EC)_i]}{(\theta_i + 1)^2} \right\}$$

$$- \sum_{i=1}^{n} \left\{ \frac{\theta_i^{c_i - 1} p_i}{B(\alpha, c_i + 1)} \left[ \frac{c_i - (\alpha+1)\theta_i}{(1+\theta_i)^{\alpha + c_i + 2}} \right] \right\}$$

$$+ \sum_{i=1}^{n} \left\{ \frac{\theta_i^{2c_i} p_i}{B^2(\alpha, c_i + 1)(\theta_i + 1)^{2(\alpha + c_i + 1)} P_\alpha(ED)_i} \right\},$$

$$\sigma_{22} = E\left( -\frac{\partial^2 \ln L}{\partial b^2} \right)$$

$$= \sum_{i=1}^{n} \left\{ \frac{p_i^{\,2} \alpha P_{\alpha+1}(EC)_i}{\theta_i} \right\}$$

$$- \sum_{i=1}^{n} \left\{ \left[ \frac{\alpha [(1 - P_\alpha(ED)_i) + \theta_i P_{\alpha+1}(EC)_i]}{p_i^{\,-2} (\theta_i + 1)^2} \right] \right\}$$

$$- \sum_{i=1}^{n} \left\{ \frac{\theta_i^{c_i - 1} p_i^{\,2}}{B(\alpha, c_i + 1)} \left[ \frac{c_i - (\alpha+1)\theta_i}{(1+\theta_i)^{\alpha + c_i + 2}} \right] \right\}$$

$$+ \sum_{i=1}^{n} \left\{ \left[ \frac{\theta_i^{2c_i} p_i^{\,2}}{B^2(\alpha, c_i + 1)(\theta_i + 1)^{2(\alpha + c_i + 1)} P_\alpha(ED)_i} \right] \right\},$$

and

$$P_{\alpha+1}(EC)_i = \sum_{x=0}^{c_i - 1} \frac{(\alpha+1)\ldots(\alpha+x)}{x!(\theta_i + 1)^{1+\alpha}} \left( \frac{\theta_i}{\theta_i + 1} \right)^x.$$

These equations are obtained by using equation (3) in the previous section and

$$E(1 - \delta_i) = P_\alpha(ED)_i,$$
$$E(\delta_i X_i) = \theta_i \alpha P_{\alpha+1}(EC)_i, \quad i = 1,2,\ldots,n.$$

A randomly selected hotel from the chain is considered and the estimate for the optimum price and the optimum revenue for the hotel is given. Also, the expected marginal revenue at a given price $p$ is estimated. Again, it is assumed that $c$ is the capacity of the hotel and similar notations are used. Let $p^* = p^*(a,b)$ be the solution to

$$(\theta\alpha + pb\alpha) P_{\alpha+1}(EC) + cP_\alpha(ED) = 0.$$

A direct application of the chain rule when taking the derivative from both sides of the equation gives

$$\frac{\partial p^*}{\partial a} = \frac{p^* b\theta^{*(c-1)} - B(\alpha+1, c)(1+\theta^*)^{\alpha+1+c} P_{\alpha+1}(EC^*)}{b[2B(\alpha+1, c)(1+\theta^*)^{\alpha+1+c} P_{\alpha+1}(EC^*) - p^* b\theta^{*(c-1)}]} \quad (7)$$

$$\frac{\partial p^*}{\partial b} = \frac{p^*[p^* b\theta^{*(c-1)} - 2B(\alpha+1, c)(1+\theta^*)^{\alpha+1+c} P_{\alpha+1}(EC^*)]}{b\left[2B(\alpha+1, c)(1+\theta^*)^{\alpha+1+c} P_{\alpha+1}(EC^*) - p^* b\theta^{*(c-1)}\right]} \quad (8)$$

Notice that

$$ER^* = p^*[\theta^* \alpha P_{\alpha+1}(EC^*) + cP_\alpha(ED^*)]$$
$$= -p^{*2} b\alpha P_{\alpha+1}(EC^*).$$

Applying the chain rule again,

$$\frac{\partial ER^*}{\partial a}$$

$$= -b\alpha p^* \left[ 2P_{\alpha+1}(EC^*) \frac{\partial p^*}{\partial a} \right]$$

$$+ \left[ \left( 1 + b\frac{\partial p^*}{\partial a} \right) \frac{b\alpha p^{*2} \theta^{*(c-1)}}{B(\alpha+1, c)(1+\theta^*)^{\alpha+1+c}} \right],$$

$$\frac{\partial ER^*}{\partial b}$$

$$= -\alpha p^* \left[ \left( 2b \frac{\partial p^*}{\partial b} + p^* \right) P_{\alpha+1}(EC^*) \right]$$

$$+ \left[ \left( p^* + b \frac{\partial p^*}{\partial b} \right) \frac{\alpha p^{*2}\, b\theta^{*(c-1)}}{B(\alpha+1,c)(1+\theta^*)^{\alpha+1+c}} \right],$$

where $\partial p^*/\partial a$ and $\partial p^*/\partial b$ are given by (7) and (8). Recall that at a given price *p*, the expected marginal revenue

$$EMR = 2p + a/b + cP_\alpha(ED)/[b\alpha P_{\alpha+1}(EC)]$$

is a function of *a* and *b*. Another application of the chain rule yields

$$\frac{\partial EMR}{\partial a} = \frac{1}{b} + \frac{c\theta^{c-1}[B(\alpha+1,c)\theta P_{\alpha+1}(EC) + B(\alpha,c+1)P_\alpha(ED)]}{b\alpha B(\alpha,c+1)B(\alpha+1,c)(1+\theta)^{\alpha+1+c} P_{\alpha+1}(EC)^2}$$

$$\frac{\partial EMR}{\partial b} = -\frac{a}{b^2} - \frac{cP_\alpha(ED)}{b^2\alpha P_{\alpha+1}(EC)} + \frac{cp\theta^{c-1}[B(\alpha+1,c)\theta P_{\alpha+1}(EC) + B(\alpha,c+1)P_\alpha(ED)]}{b\alpha B(\alpha,c+1)B(\alpha+1,c)(1+\theta)^{\alpha+1+c} P_{\alpha+1}(EC)^2}.$$

Let

$$\sigma^2 = \begin{pmatrix} \frac{\partial p^*}{\partial a} & \frac{\partial p^*}{\partial b} \end{pmatrix} \Sigma^{-1} \begin{pmatrix} \frac{\partial p^*}{\partial a} & \frac{\partial p^*}{\partial b} \end{pmatrix}^t,$$

$$\delta^2 = \begin{pmatrix} \frac{\partial ER^*}{\partial a} & \frac{\partial ER^*}{\partial b} \end{pmatrix} \Sigma^{-1} \begin{pmatrix} \frac{\partial ER^*}{\partial a} & \frac{\partial ER^*}{\partial b} \end{pmatrix}^t,$$

and

$$\tau^2 = \begin{pmatrix} \frac{\partial EMR}{\partial a} & \frac{\partial EMR}{\partial b} \end{pmatrix} \Sigma^{-1} \begin{pmatrix} \frac{\partial EMR}{\partial a} & \frac{\partial EMR}{\partial b} \end{pmatrix}^t,$$

where *t* stands for the transpose. Finally, let

$$\hat{\Sigma}^{-1} = \begin{pmatrix} \hat{\sigma}_{11}' & \hat{\sigma}_{12}' \\ \hat{\sigma}_{12}' & \hat{\sigma}_{22}' \end{pmatrix} = \begin{pmatrix} \hat{\sigma}_{11} & \hat{\sigma}_{12} \\ \hat{\sigma}_{12} & \hat{\sigma}_{22} \end{pmatrix}^{-1}$$

be the MLE of $\Sigma^{-1}$. Let $\hat{p}^*$, $\hat{\sigma}^2$, $E\hat{R}^*$, $\hat{\delta}^2$, $E\hat{M}R$, and $\hat{\tau}$ be the MLEs of $p^*, \sigma^2$, $ER^*$, $\delta^2$, *EMR*, and τ, respectively. Since $p^*, \sigma^2$, $ER^*$, $\delta^2$, *EMR*, and τ are functions of *a* and *b*. Their MLEs are obtained by replacing *a* and *b* by $\hat{a}$ and $\hat{b}$ in their functions, respectively.

For $0<\gamma<1$, let *Z* be the standard normal distribution and $z_{\gamma/2}$ be such that $\Pr(Z \geq z_{\gamma/2}) = \gamma/2$. The following theorem gives the confidence interval estimations for *a*, *b*, $p^*$, $ER^*$, and *EMR*.

Theorem 3

If there exist two constants $D_1$ and $D_2$ not dependent on *n* such that $p_i < D_1$, $a + bD_1 > 0$, and $c_i \leq D_2$ for *i*=1,2,...,*n*, then the following statements are correct (refer to the appendix for the proofs):

(1) An asymptotic 100(1-γ)% confidence interval for *a* is $\hat{a} \pm z_{\gamma/2}\sqrt{\hat{\sigma}_{11}'}$,

(2) An asymptotic 100(1-γ)% confidence interval for *b* is $\hat{b} \pm z_{\gamma/2}\sqrt{\hat{\sigma}_{22}'}$,

(3) An asymptotic 100(1-γ)% confidence interval for $p^*$ is $\hat{p}^* \pm z_{\gamma/2}\hat{\sigma}$,

(4) An asymptotic 100(1-γ)% confidence interval for $ER^*$ is $\hat{ER}^* \pm z_{\gamma/2}\hat{\delta}$,

(5) An asymptotic 100(1-γ)% confidence interval for *EMR* at a given price *p* is $\hat{EMR} \pm z_{\gamma/2}\hat{\tau}$.

In the more realistic situation when none of parameter α, *a* and *b* are known, a stepwise procedure to find the maximum likelihood estimators of α, *a* and *b* is proposed. The traditional approach of maximizing a likelihood function is simply by setting the derivative of the likelihood function with respect to each parameter to 0 simultaneously and then solving the system of equations. This approach becomes very complicated in this case because the derivative of the likelihood function with respect to the index parameter α is rather complicated.

It is proposed that the maximum likelihood estimators $(\hat{\alpha}, \hat{a}, \hat{b})$ should be obtained by first using the method described above to get the maximum likelihood estimators $\hat{a}(\alpha)$ and $\hat{b}(\alpha)$ for specified α values, and then combining with a search procedure to obtain $\hat{\alpha}$, the value of α that maximizes $L_{\max}(a) = L(\alpha, \hat{a}(\alpha), \hat{b}(\alpha))$. The simplex search method of Nelder and Mead (1965) has proved successful in many problems, particularly when there are not too many parameters present. Other search procedures such as those of Powell (1964) and Fletcher and Reeves (1964) are also widely used. After the maximum likelihood estimators $(\hat{\alpha}, \hat{a}, \hat{b})$ are obtained, Theorem 3 can still be used to obtain the asymptotic confidence intervals for model parameters when α is replaced by $\hat{\alpha}$. These asymptotic confidence intervals are still valid based on the fact that $\hat{\alpha}$ is a strongly consistent estimator to α.

Notice that all confidence intervals given by Theorem 3 are asymptotic confidence intervals whose coverage probability approaches 100(1-γ)% when the sample size *n* approaches infinity. In order to assess how these confidence intervals perform with a limited sample size, a simulation study was carried out to compare the empirical coverage to the nominal coverage probability for a selected set of sample size *n*. The following values were chosen α=2, *c*=50, *a*=100, *b*=-1. For each selected sample size for *X*, one third of the sample comes from each unit price of *p*=40, 65, 90. For a given unit price *p*, the one third of the sample for *X* are simulated by using the distribution of *X* as given in Section 2.

In order to generate these samples, random samples on the integer set {1, 2,...,51} based on the 51 probabilities of *X* from *X*=0 to *X*=50 as given in Section 2 are first generated using the random number generating function RANTBL from Statistical Analysis System (1999). One is then subtracted from the samples to give the random samples for *X*. Table 1 presents the empirical coverage probability of the true parameter values. Each empirical coverage probability reported by Table 1 is computed from a simulation of 500 independent confidence intervals based on 500 independent samples of *X* for parameters *a*, *b*, $p^*, ER^*$, and *EMR* at *p*=60. The optimum price $p^*$ as the solution to (6) is computed using the Newton-Raphson method. All confidence intervals are computed based on Theorem 3 when the index parameter α is replaced by the maximum likelihood estimator α. The maximum likelihood estimators $(\hat{\alpha}, \hat{a}, \hat{b})$ are obtained by the stepwise procedure described above using the simplex search method of Nelder and Mead (1965) when $L_{\max}(a) = L(\alpha, \hat{a}(\alpha), \hat{b}(\alpha))$ is maximized. All the nominal confidence levels in Table 1 are 95% (γ=5%).

Table 1. Empirical Coverage Probability of Confidence Intervals

| | α=2, *c*=50, *a*=100, *b*=-1 | | | | |
|---|---|---|---|---|---|
| Sample size | *a* | *b* | $p^*$ | $ER^*$ | *EMR* at *p*=60 |
| 18 | 0.910 | 0.912 | 0.924 | 0.896 | 0.906 |
| 24 | 0.924 | 0.940 | 0.970 | 0.936 | 0.924 |
| 30 | 0.962 | 0.932 | 0.938 | 0.936 | 0.944 |
| 36 | 0.932 | 0.944 | 0.938 | 0.960 | 0.952 |
| 42 | 0.960 | 0.964 | 0.942 | 0.928 | 0.958 |
| 48 | 0.932 | 0.972 | 0.938 | 0.942 | 0.946 |
| 60 | 0.962 | 0.946 | 0.932 | 0.952 | 0.958 |
| 75 | 0.972 | 0.948 | 0.946 | 0.958 | 0.946 |
| 90 | 0.958 | 0.940 | 0.960 | 0.946 | 0.958 |
| 105 | 0.954 | 0.952 | 0.944 | 0.952 | 0.964 |
| 120 | 0.956 | 0.946 | 0.948 | 0.954 | 0.952 |
| 150 | 0.958 | 0.952 | 0.944 | 0.948 | 0.950 |
| 180 | 0.946 | 0.952 | 0.954 | 0.954 | 0.946 |
| 240 | 0.944 | 0.958 | 0.944 | 0.954 | 0.946 |
| 300 | 0.948 | 0.956 | 0.958 | 0.944 | 0.956 |

## Conclusion

This article has proposed a two-stage statistical model to model the demand variability from a chain of price setting firms. The demand variability from within a firm is modeled by a Poisson distribution, and the demand variability from across the firms is modeled by a Gamma distribution. It was shown that the optimum price under a capacity constraint decreases after the capacity increases to a certain level. On the other hand, the optimum expected revenue increases when the capacity increases. The article also provides a stepwise procedure to find the maximum likelihood estimates of model parameters. The proposed method does not require taking the derivative of the likelihood function with respect to the index parameter α.

Asymptotic confidence interval estimates are developed for the optimum price, the optimum revenue, and the expected marginal revenue at a given price based on the asymptotic normality for the maximum likelihood estimates. A limited simulation study seems to suggest that a relatively large sample size (>100) is required for the asymptotic confidence intervals to achieve the nominal coverage probability.

## References

Billingsley, P. (1986). *Probability and measure* (2nd ed.). New York: John Wiley & Sons**.** Turnovsky, S. J. (1973). Production flexibility, uncertainty, and the behavior of the competitive firms. *International Economic Review, 14*, 395-412.

Balvers, R. J., Miller, N. C. (1992). Factor demand under conditions of product demand and supply uncertainty. *Economic Inquiry, 30,* 544-555.

Booth, L. A. (1990). Note on adjustment to production uncertainty and the theory of the Firm. *Economic Inquiry, 28,* 616-621.

Cox, D. R. & Hinkley, D. V. (1974). *Theoretical statistics*. London: Chapman & Hall.Scott, R. C., Sattler, E. L., Highfill, J. K. (1995). A hotel capacity utilization model. *The Journal of Economics, 21,* 101-105.

Epstein, L. G. (1978). Production flexibility and behavior of competitive firms under price uncertainty. *Review of Economic Studies, 45,* 251-261.

Flacco, P. R., Kroetch, B. G. (1986). Adjustment to production uncertainty and the theory of the firm. *Economic Inquiry, 24,* 485-495.

Fletcher, R. & Reeves, C. M. (1964). Function minimization by conjugate gradients. *Computer Journal, 7,* 149-154.

Highfill, J. K., Quigg, D., Sattler, E. L., & Scott, R. C. (2000). The capacity decision when product demand is uncertain: A timing approach. *The Journal of Economics, 26, 1,* 71-85.

Nelder, J. A. & Mead, R. A. (1965). Simplex method for function minimization. *Computer Journal, 7,* 380-383.

Powell, M. J. D. (1964). An efficient method for finding the minimum of a function of several variables without calculating derivatives. *Computer Journal, 1,* 155-162.

Press, W. H., Flannery, B. P., Teukolsky, S. A., & Vetterling, W. T. (1986). *Numerical recipes: The art of scientific computing.* Cambridge: Cambridge University Press.

SAS Institute, Inc. (1999). *SAS Language* (Version 8), Cary, NC.

Scott, R. C., Highfill, J. K., Sattler, E. L. (1988). Advantage to a risk neutral firm of flexible resources under demand uncertainty. *Southern Economic Journal, 54,* 934-949.

## Appendix

Proof of Theorem 1

The optimum price $p^*$ maximizes *ER* and therefore solves $dER/dp = 0,$ i.e.,

$$(\theta\alpha + pb\alpha)P_{\alpha+1}(EC) + cP_{\alpha}(ED) + p\theta\alpha\frac{dP_{\alpha+1}(EC)}{dp} + pc\frac{dP_{\alpha}(ED)}{dp} = 0.$$

Thus, using equation (5) in Section 2, it is concluded that $p^*$ satisfies the equation:

$$(\theta\alpha + pb\alpha)P_{\alpha+1}(EC) + cP_{\alpha}(ED) = 0.$$

In addition,

$$\frac{d^2ER}{dp^2} = 2b\alpha P_{\alpha+1}(EC) + pb\alpha\frac{dP_{a+1}(EC)}{dp}$$

is negative by the fact that $b<0$ and equation (4). It then follows that $p^*$ is the unique solution to (6). It is clear that the first term in (6) has to be negative to make (6) hold. Therefore, $p^*$ satisfies $\theta\alpha+pb\alpha<0$, i.e., $p^* > -a/(2b)$. Since $\lim_{c\to\infty} P_{\alpha+1}(EC) = 1,$ it follows from (6) that $\lim_{c\to\infty}(\theta\alpha + pb\alpha) = 0$, i.e., $\lim_{c\to\infty} p^* = -a/(2b)$.

Proof of Theorem 2

(1): For $0 < p < -a/b$ and $\theta=a+bp$, let

$$f(p,c) = \frac{dER}{dp} = (\theta\alpha + pb\alpha)P_{\alpha+1}(EC) + cP_{\alpha}(ED).$$

A direct application of equation (6) gives

$$\frac{c!(1+\theta*)^{\alpha}}{(\alpha+1)...(\alpha+c)}\left(\frac{\theta*+1}{\theta*}\right)^{c} f(p*,c+1)$$

$$=\frac{\theta*\alpha+p*b\alpha}{\theta*+1}-\frac{c\alpha\theta_i}{(c+1)(\theta_i+1)}+I(c),$$

where

$$I(c)=\frac{c!(1+\theta*)^{\alpha}}{(\alpha+1)...(\alpha+c)}\left(\frac{\theta*+1}{\theta*}\right)^{c}$$

$$\sum_{x=c+2}^{\infty}\frac{\alpha(\alpha+1)...(\alpha+x-1)}{x!(1+\theta*)^{\alpha}}\left(\frac{\theta*}{\theta*+1}\right)^{x}.$$

Replacing $c$ by $c+1$ in equation (1) of Section 1, provides the following,

$$I(c)=\theta^{*2}(1+\theta^{*})^{\alpha-1}\frac{\alpha(c+\alpha+1)}{c+1}$$

$$\int_0^1 \frac{[(1+\theta^{*})t]^{c+1}}{(1+\theta^{*}t)^{\alpha+c+2}}dt.$$

For any $1>s>0$,

$$\int_0^1 \frac{[(1+\theta^{*})t]^{c+1}}{(1+\theta^{*}t)^{\alpha+c+2}}dt$$
$$=\int_0^s \frac{[(1+\theta^{*})t]^{c+1}}{(1+\theta^{*}t)^{\alpha+c+2}}dt$$
$$+\int_s^1 \frac{[(1+\theta^{*})t]^{c+1}}{(1+\theta^{*}t)^{\alpha+c+2}}dt.$$

Because

$$\int_0^s \frac{[(1+\theta^{*})t]^{c+1}}{(1+\theta^{*}t)^{\alpha+c+2}}dt<\left[\frac{(1+\theta^{*})s}{1+\theta^{*}s}\right]^{c+1},$$

$\lim_{c\to\infty}\int_0^s \frac{[(1+\theta^{*})t]^{c+1}}{(1+\theta^{*}t)^{\alpha+c+2}}dt=0$ by the fact that $\lim_{c\to\infty}\theta^{*}=\frac{a}{2}$ and $\frac{(1+a/2)s}{1+as/2}<1$.

Because

$$\int_s^1 \frac{[(1+\theta^{*})t]^{c+1}}{(1+\theta^{*}t)^{\alpha+c+2}}dt<1-s,$$

it follows that

$$\lim_{s\to 1^{-}}\int_s^1 \frac{[(1+\theta^{*})t]^{c+1}}{(1+\theta^{*}t)^{\alpha+c+2}}dt=0,$$

where the convergence is uniform on $c$. Thus, $\lim_{c\to\infty}I(c)=0$, which further implies that

$$\lim_{c\to\infty}\left[\frac{\theta^{*}\alpha+p^{*}b\alpha}{1+\theta^{*}}-\frac{c\alpha}{c+1}\left(\frac{\theta^{*}}{\theta^{*}+1}\right)+I(c)\right]$$
$$=-\frac{\alpha a}{a+2}<0.$$

Therefore, there exists a constant $C$ depending on only $a$ and $\alpha$ such that if $c>C$ then $f(p^{*},c+1)<0$. Because $f(p^{*}+\Delta p^{*},c+1)=0$ and $df(p,c+1)/dp<0$, it follows that $f(p,c+1)>0$ when $0<p<p^{*}+\Delta p^{*}$ and $f(p,c+1)<0$ when $p^{*}+\Delta p^{*}<p<-a/b$. Hence $p^{*}>p^{*}+\Delta p^{*}$, i.e., $\Delta p^{*}<0$. $\lim_{c\to\infty}\Delta p^{*}=0$ follows from the fact that $\lim_{c\to\infty}p^{*}=-a/(2b)$.

(2): For $0<p<-a/b$, let $g(p,c)=ER=p[\theta\alpha P_{\alpha+1}(EC)+cP_{\alpha}(ED)]$. Then

$$\Delta ER^{*}=g(p^{*}+\Delta p^{*},c+1)-g(p^{*},c)$$
$$=g(p*+\Delta p*,c+1)-g(p*,c+1)$$
$$+g(p*,c+1)-g(p*,c).$$

$g(p^* + \Delta p^*, c+1) - g(p^*, c+1) > 0$ by the fact that $p^* + \Delta p^*$ maximizes $g(p,c+1)$ over $p$. $\Delta ER^* > 0$ follows from the fact that

$$g(p^*, c+1) - g(p^*, c)$$
$$= p^* \sum_{x=c+1}^{\infty} \frac{\alpha(\alpha+1)...(\alpha+x-1)}{x!(1+\theta^*)^{\alpha}} \left( \frac{\theta^*}{\theta^*+1} \right)^x > 0.$$

Finally, since $ER^* = -p^{*2} b \alpha P_{\alpha+1}(EC^*)$ and $\lim_{c\to\infty} P_{\alpha+1}(EC^*) = 1$, $\lim_{c\to\infty} \Delta ER^* = 0$ follows from the fact that $\lim_{c\to\infty} ER^* = -\alpha a^2/(4b)$.

Proof of Theorem 3

The asymptotic normality is first given for the maximum likelihood estimator $(\hat{a}\ \hat{b})^t$ of $(a\ b)^t$ ($t$=transpose). Notice that the data come from independent but not identically distributed distributions. Cox and Hinkley (1974) pointed out that the asymptotic normality for the MLEs of such distributions requires two conditions: one is a central limit theorem to $(\partial \ln L/\partial a \quad \partial \ln L/\partial b)^t$ with a nonsingular asymptotic distribution, the other is a weak law of large numbers to insure the convergence in probability of

$$-\frac{1}{n}\begin{pmatrix} \frac{\partial^2 \ln L}{\partial a^2} & \frac{\partial^2 \ln L}{\partial a \partial b} \\ \frac{\partial^2 \ln L}{\partial a \partial b} & \frac{\partial^2 \ln L}{\partial b^2} \end{pmatrix} - \frac{1}{n}\Sigma$$

to zero.

To prove a central limit theorem to $(\partial \ln L/\partial a \quad \partial \ln L/\partial b)^t$, one only needs to do so for

$$= \sum_{i=1}^{n} (t_1 + t_2 p_i) \left\{ \frac{\delta_i (X_i - \alpha\theta_i)}{\theta_i(\theta_i+1)} \right\}$$

$$+ \sum_{i=1}^{n} \left\{ \frac{(t_1 + t_2 p_i)(1-\delta_i)\theta_i^{c_i}}{B(\alpha, c_i+1)(1+\theta_i)^{\alpha+c_i+1} P_{\alpha}(ED)_i} \right\}$$

for any choices of $t_1$ and $t_2$. For $i$=1,2,...,$n$, let

$$T_i = (t_1 + t_2 p_i) \left\{ \frac{\delta_i (X_i - \alpha\theta_i)}{\theta_i(\theta_i+1)} \right\}$$

$$+ \frac{(t_1 + t_2 p_i)(1-\delta_i)\theta_i^{c_i}}{B(\alpha, c_i+1)(1+\theta_i)^{\alpha+c_i+1} P_{\alpha}(ED)_i}.$$

It is clear that $ET_i = 0$. Let $\sigma_{T_i}^2 = ET_i^2$. A careful computation using

$$E(\delta_i X_i^2) = \alpha(\alpha+1)\theta_i^2 P_{\alpha+2,c-1}(EC)_i + \alpha\theta_i P_{\alpha+1,c}(ED)_i$$

gives

$$\sigma_{T_i}^2 = \frac{(t_1 + t_2 p_i)^2 \theta_i^{2c_i}}{B(\alpha, c_i+1)^2 (1+\theta_i)^{2(\alpha+c_i+1)} P_{\alpha,c}(ED)_i}$$
$$+ \frac{(t_1 + t_2 p_i)^2}{\theta_i^2(\theta_i+1)^2} \left[ \alpha(\alpha+1)\theta_i^2 P_{\alpha+2,c-1}(EC)_i \right]$$

$$+ \frac{(t_1 + t_2 p_i)^2 \alpha(1-2\alpha\theta_i) P_{\alpha+1,c}(EC)_i}{\theta_i(\theta_i+1)^2}$$

$$+ \frac{(t_1 + t_2 p_i)^2 \alpha^2 \left[1 - P_{\alpha,c}(ED)_i\right]}{(\theta_i+1)^2}..$$

Notice that in the above equation, two indices $\alpha$ and $c$ were used in the notation $P_{\alpha,c}(EC)_i$ to indicate the dependence of the probability on these two parameters. Since, for given $t_1$ and $t_2$, $\sigma_{T_i}^2$ is a positive continuous function of $(\theta_i, c)$ when $0 < a + bD_1 \le \theta_i \le a$ and $1 \le c \le D_2$, $\sigma_{T_i}^2$ has a positive lower bound and a positive upper bound not dependent on $i$. Thus,

$\sigma_n^2 = \sum_{i=1}^{n} \sigma_{T_i}^2$ approaches infinity when $n$ approaches infinity. Notice that $T_i$ is bounded. Therefore, for any $\varepsilon$ >0,

$$\lim_{n\to\infty} \sum_{i=1}^{n} \frac{1}{\sigma_n^2} E\{T_i^2 \chi_{\{T_i \geq \varepsilon\sigma_n\}}\} = 0,$$

where $\chi_{\{T_i \geq \varepsilon\sigma_n\}}$ is the indicator of $\{T_i \geq \varepsilon\sigma_n\}$ (i.e., the Lindeberg condition for $T_i$ holds). This proves the central limit theorem for $(\partial \ln L/\partial a \quad \partial \ln L/\partial b)^t$.

To prove a weak law of large numbers to insure the convergence in probability of

$$-\frac{1}{n}\begin{pmatrix} \frac{\partial^2 \ln L}{\partial a^2} & \frac{\partial^2 \ln L}{\partial a \partial b} \\ \frac{\partial^2 \ln L}{\partial a \partial b} & \frac{\partial^2 \ln L}{\partial b^2} \end{pmatrix} - \frac{1}{n}\Sigma$$

to zero, write $\frac{\partial^2 \ln L}{\partial a^2} = \sum_{i=1}^{n} U_i$, $\frac{\partial^2 \ln L}{\partial a \partial b} = \sum_{i=1}^{n} V_i$, and $\frac{\partial^2 \ln L}{\partial b^2} = \sum_{i=1}^{n} W_i$,

where

$$U_i = \frac{\delta_i(X_i+\alpha)}{(\theta_i+1)^2} - \frac{\delta_i X_i}{\theta_i^2} + \frac{(1-\delta_i)\theta_i^{c_i-1} P_\alpha (ED)_i^{-1}}{B(\alpha, c_i+1)(1+\theta_i)^{\alpha+c_i+1}} \left[\frac{c_i - (\alpha+1)\theta_i}{(1+\theta_i)}\right]$$

$$- \frac{(1-\delta_i)\theta_i^{2c_i}}{B^2(\alpha, c_i+1)(1+\theta_i)^{2(\alpha+c_i+1)} P_\alpha^{\,2} (ED)_i},$$

and $V_i = p_i U_i$, $W_i = p_i^2 U_i$. Since $\sigma_{U_i}^2$, $\sigma_{V_i}^2$ and $\sigma_{W_i}^2$ are all positive continuous functions of $(\theta_i, c)$ when $0 < a + bD_1 \leq \theta_i \leq a$ and $1 \leq c \leq D_2$, they all have positive upper bounds. Because

$$\sigma^2_{\frac{\partial^2 \ln L}{\partial a^2}} = \sum_{i=1}^{n} \sigma^2_{U_i},$$

$$\sigma^2_{\frac{\partial^2 \ln L}{\partial a \partial b}} = \sum_{i=1}^{n} \sigma^2_{V_i},$$

$$\sigma^2_{\frac{\partial^2 \ln L}{\partial b^2}} = \sum_{i=1}^{n} \sigma^2_{W_i},$$

it follows that

$$\lim_{n\to\infty} \frac{\sigma_{\frac{\partial^2 \ln L}{\partial a^2}}}{n} = \lim_{n\to\infty} \frac{\sigma_{\frac{\partial^2 \ln L}{\partial a \partial b}}}{n} = \lim_{n\to\infty} \frac{\sigma_{\frac{\partial^2 \ln L}{\partial b^2}}}{n} = 0.$$

The weak law of large numbers to

$$-\frac{1}{n}\begin{pmatrix} \frac{\partial^2 \ln L}{\partial a^2} & \frac{\partial^2 \ln L}{\partial a \partial b} \\ \frac{\partial^2 \ln L}{\partial a \partial b} & \frac{\partial^2 \ln L}{\partial b^2} \end{pmatrix}$$

follows from Theorem 6.2 of Billingsley (1986). Therefore, as $n\to\infty$,

$$\Sigma^{1/2}\left\{\begin{pmatrix} \hat{a} \\ \hat{b} \end{pmatrix} - \begin{pmatrix} a \\ b \end{pmatrix}\right\} \to \mathbf{N}\left\{\begin{pmatrix} 0 \\ 0 \end{pmatrix}, I_{2\times 2}\right\}$$

in distribution, where $I_{2\times 2}$ is the 2×2 identity matrix, i.e., asymptotically,

$$\begin{pmatrix} \hat{a} \\ \hat{b} \end{pmatrix} - \begin{pmatrix} a \\ b \end{pmatrix}$$

is distributed as $\mathbf{N}\left\{\begin{pmatrix}0\\0\end{pmatrix},\Sigma^{-1}\right\}$. (1) and (2) follow directly from the asymptotic normality of $\hat{a}$ and $\hat{b}$, respectively. (3) follows from the fact that as $n\to\infty$, the MLE of $p^* = p^*(a,b)$ satisfies that

$$\frac{\hat{p}^* - p^*}{\sigma} \to N(0,1)$$

in distribution. (4) follows from the fact that as $n\to\infty$, the MLE of $ER^*$ satisfies that

$$\frac{E\hat{R}^* - ER^*}{\delta} \to N(0,1)$$

in distribution. (5) follows from the fact that as $n\to\infty$, the MLE of *EMR* satisfies that

$$\frac{E\hat{M}R - EMR}{\tau} \to N(0,1)$$

in distribution.

Journal of Modern Applied Statistical Methods
November, 2005, Vol. 4, No. 2, 567-582

1538 – 9472/05/$95.00

# A Nonrigorous Approach Of Incorporating Sensitizing Rules Into Multivariate Control Charts

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

Multivariate control charts are becoming more important in the monitoring of processes in manufacturing industries because the quality of a process is usually determined by several correlated variables (quality characteristics). The most popular multivariate process control procedure is based on the Hotelling control chart. It is used to monitor the mean vector of a process. A nonrigorous approach of using four sensitizing rules is introduced to improve the performance of a conventional Hotelling chart. The use of these rules on a conventional Hotelling chart do not require a transformation of the $T^2$ statistics into normal random variables. Thus, the $T^2$ statistics incorporating these rules can be plotted on the same scale as they are plotted on a Hotelling chart. Numerous SAS and Mathematica programs are given to aid quality control practitioners in implementing these rules in real life problems. The aim of this article is to make the implementation of sensitizing rules appealing and user friendly to practitioners.

Key Words: sensitizing rules; Hotelling; average run length (ARL); in-control; out-of-control (o.o.c.); Markov chain; upper control limit (UCL)

## Introduction

Since its inception (Hotelling, 1947), numerous extensions have been made to the conventional Hotelling $T^2$ chart. Tracy, Young and Mason (1992) discussed an exact method based on the beta distribution for constructing multivariate control limits at the start-up stage. Timm (1996) introduced the use of a single step and stepdown finite intersection test (FIT) to evaluate whether a multivariate process is in-control or out-of-control. Runger (1996) discussed an approach based on projections, which simplifies the construction and understanding of a multivariate Hotelling chart. A comparison of using various estimators of the covariance matrix for the Hotelling chart was made by Sullivan and Woodall (1996).

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

Prins and Mader (1997) provided some interesting discussion on multivariate control charts for subgrouped data and individual observations. Key implementation and interpretation issues as well as assessing the problems that currently exist when using multivariate charts were examined by Mason, Champ, Tracy, Wierda and Young (1997). Aparisi (1997) proposed sampling plans for the multivariate $T^2$ control chart.

Various approaches in the identification of the problematic quality characteristics when the $T^2$ chart signals an o.o.c. are suggested in the literature. These include the works of Doganaksoy, Faltin and Tucker (1991), Holmes and Mergen (1995), Mason, Tracy and Young (1995; 1997), Runger, Alt and Montgomery (1996) and Nedumaran and Pignatiello (1998). Apley and Tsung (2002) investigated and provided guidelines for designing the autoregressive $T^2$ chart in the monitoring of univariate autocorrelated processes. The usefulness of the Hotelling $T^2$ statistic for the monitoring of batch processes in both Phase I and Phase II operations were shown in Mason, Chou and Young (2001). Vargas (2003)

suggested $T^2$ charts based on robust estimators of location and dispersion using minimum volume ellipsoid (MVE) estimators, which are effective in detecting any reasonable number of outliers.

Sensitizing rules are supplementary criteria that are used to increase the sensitivity of a univariate control chart to small process shifts so that assignable causes can be detected quicker (Montgomery, 2001). Nelson (1984) provided a good discussion of some of these rules. Champ and Woodall (1987) studied the ARL performances of a univariate Shewhart chart with various sensitizing rules and found that the use of these rules improve the ability of the chart to detect smaller shifts at the expense of the Type-I error. To overcome this problem, Klein (2000) introduced two alternative schemes to the $\bar{X}$ chart, namely rules 2-of-2 and 2-of-3. The Type-I error of these two rules can be fixed by the user and then their respective limits are determined using a Markov chain approach.

One fundamental requirement of using sensitizing rules on a control chart is that the consecutive statistics plotted on the chart must be normally distributed. This is aside from the independent and identically distributed (i.i.d.) assumption of the sequence of control chart statistics. To meet the normality requirement, Khoo and Quah (2003) and Kooh, Quah, and Low (2004), suggested an approach of transforming the Hotelling statistic into a standard normal random variable prior to the application of different sensitizing rules on a multivariate chart. Their suggestion by means of transformation allows the use of such rules on the Hotelling control chart. Though their suggestion is a useful contribution to multivariate quality control, it has increased the complexity of using a Hotelling chart to a certain extent, which may make the suggested approach less appealing to some practitioners.

The main objective in this article is to solve the above problem by making the incorporation of sensitizing rules into a Hotelling chart user friendly so that quality control practitioners will find such enhancements useful in their work. Unlike the previous works of Khoo and Quah (2003) and Kooh, Quah, and Low (2004), the new approach suggested in this article does not require the transformation of a $T^2$ statistic into a standard normal random variable, hence it is referred to as a nonrigorous approach. Besides ease of implementation, another remarkable advantage of the new approach is that it allows the $T^2$ statistics to be plotted on their original scale on a Hotelling control chart. Thus, the use of the conventional Hotelling chart can still be maintained by drawing additional limits on the chart for the sensitizing rule being implemented.

SAS programs are provided for cases of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ known and unknown, involving both individual measurements and subgrouped data. Now, practitioners can easily compute the limits of each of the four rules by running the SAS programs after entering the desired values of the required parameters.

## The Conventional Hotelling $T^2$ Control Chart

In the monitoring of a multivariate process where the data belong to individual observations and follow a multivariate normal distribution, i.e., $\boldsymbol{X}_i \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, i = 1, 2, …, the following $T^2$ statistics are used (Tracy, Young and Mason, 1992):

$$T_i^2 = (\boldsymbol{X}_i - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{X}_i - \boldsymbol{\mu}), \; i = 1, 2, \ldots . \quad (1)$$

Here, $T_i^2 \sim \chi_p^2$ where $p$ is the number of quality characteristics monitored simultaneously. For the case where both $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are unknown, the equation below which is given in Tracy, Young and Mason (1992) is used:

$$T_f^2 = (\boldsymbol{X}_f - \bar{\boldsymbol{X}}_m)' \boldsymbol{S}_m^{-1} (\boldsymbol{X}_f - \bar{\boldsymbol{X}}_m), \; f = 1, 2, \ldots . \quad (2)$$

It is shown in Tracy, Young and Mason (1992) that the exact distribution of $T_f^2$ is $T_f^2 \sim \frac{p(m-1)(m+1)}{m(m-p)} F_{p,m-p}$, where $p$ is the number of quality characteristics, $m$ is the size of the stable reference sample, $\bar{\boldsymbol{X}}_m$ and $\boldsymbol{S}_m$ are estimates of the mean vector and covariance matrix from a stable reference sample of size $m$ respectively. $\boldsymbol{X}_f$ in equation (2) denotes a future multivariate

normal $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ observation taken at time *f*, so that the state of a process at that time can be determined.

For subgrouped data, the test statistics plotted on the Hotelling $T^2$ chart are

$$T_j^2 = n(\overline{\boldsymbol{X}}_j - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\overline{\boldsymbol{X}}_j - \boldsymbol{\mu}), \quad j = 1, 2, \ldots, \quad (3)$$

where *j* is the subgroup number. It is assumed that the joint probability distribution of the *p* quality characteristics is the *p*-variate normal distribution. In equation (3), $\overline{\boldsymbol{X}}_j$ is a $p \times 1$ vector of sample means for each of the *p* quality characteristics from a sample of size *n*, $\boldsymbol{\mu}$ is a vector of in-control means for each of the *p* quality characteristics and $\boldsymbol{\Sigma}$ is the covariance matrix. It is noted in Montgomery (2001) that $T_j^2 \sim \chi_p^2$. If both $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are unknown, the estimates of these parameters are $\overline{\overline{\boldsymbol{X}}}$ and $\boldsymbol{S}$ respectively. Here, $\overline{\overline{\boldsymbol{X}}}$ and $\boldsymbol{S}$ are the sample grand mean vector and the sample covariance matrix estimated from an in-control preliminary data set whose formulas are given in Montgomery (2001).

There are two phases of control chart usage, namely phase 1 and phase 2. Phase 1 is a stage where the chart is used for establishing control while in phase 2, the chart is used to monitor a future production. It is shown in Montgomery (2001) that in phase 1, $T_j^2 \sim \frac{p(m-1)(n-1)}{mn-m-p+1} F_{p,mn-m-p+1}$ and in phase 2, $T_j^2 \sim \frac{p(m+1)(n-1)}{mn-m-p+1} F_{p,mn-m-p+1}$ where

$$T_j^2 = n(\overline{\boldsymbol{X}}_j - \overline{\overline{\boldsymbol{X}}})' \boldsymbol{S}^{-1} (\overline{\boldsymbol{X}}_j - \overline{\overline{\boldsymbol{X}}}), \quad j = 1, 2, \ldots, (4)$$

Note that the SAS programs given in the next section for the computation of the limits of the $T^2$ chart based on the statistics in equation (4) incorporating the various rules are made for the case involving phase 2.

Implementing Sensitizing Rules on the Conventional Hotelling $T^2$ Control Chart: A Nonrigorous Approach

To apply the sensitizing rules on the conventional Hotelling $T^2$ chart, first one needs to know the distribution of the $T^2$ statistics in equations (1) – (4). If the probability density function of the $T^2$ statistic is represented by *f*(*t*), then the upper control limit (UCL) of the various sensitizing rules can be determined by solving the following integral:

$$\int_{UCL}^{\infty} f(t)\,dt = p_A \,. \quad (5)$$

Here, $p_A$, denotes the probability of a point plotting above the UCL. The following four rules will be considered:

The 2-of-2 Rule ($S_I$)

This rule signals an out-of-control if two successive points plot above the UCL. For this rule, the in-control ARL ($ARL_0$) formula given by Khoo and Quah (2003) is

$$ARL_0 = \frac{1+g}{g^2}, \quad (6)$$

where *g* is the probability of a point falling above the UCL. The following Mathematica 4.0 program can be used to calculate the probability, *g*, based on a fixed $ARL_0$ (denoted by ARL0 in Figure 1) value.

Figure 1. A Mathematica program to compute *g* for rule $S_I$

```
ARL0 =
NSolve[(1+g)/g^2 == ARL0, g]
```

After obtaining the probability, g, equation (5) is used to compute the UCL of this rule. The SAS version 8.02 program in Figure 2 is used to compute the UCL of this rule for the $T^2$ chart based on the $T^2$ statistics in equations (1) and (3).

Figure 2. A SAS program to compute the UCL for the $T^2$ chart based on equations (1) and (3)

```
Data EQ1and3;
p= ;
g= ;
UCL=Cinv(1-g,p);
run;
proc print;
run;
```

In Figure 2, UCL = Cinv (1-g, p), where Cinv (1-g, p) refers to the 1-*g* percentile of the chi-square distribution with *p* degrees of freedom. Here, the user needs to enter the desired values of *g* and *p*, where p refers to the number of quality characteristics. Note that this program can be used by practitioners to compute the UCL of the 2-of-2 rule for the $T^2$ chart of both individual measurements and subgrouped data when the standards $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are both known.

For the case of individual measurements when both $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are unknown and are estimated, the limit (UCL) of this rule for the $T^2$ chart based on the distribution of the $T_f^2$ statistics in equation (2), i.e., $T_f^2 \sim \frac{p(m-1)(m+1)}{m(m-p)} F_{p,m-p}$ is computed using the SAS program given in Figure 3.

Figure 3. A SAS program to compute the UCL for the $T^2$ chart based on equation (2)

```
Data EQ2;
p= ;
m= ;
g= ;
a=p;
b=m-p;
UCL=p*(m-
1)*(m+1)/(m*(m-
p))*Finv(1-g,a,b);
run;
proc print;
run;
```

The program shows UCL = $\frac{p(m-1)(m+1)}{m(m-p)}$ Finv$(1-g, a, b)$, where a = p and b = m–p. Note that "Finv (1–g, a, b)" is the $1-g$ percentile of the $F$ distribution with parameters *a* and *b*. Here, the user needs to enter the values of *p*, *m*, and *g* in the program, where the notation *m* has been defined in the previous section.

Similarly, the limit of this rule for the $T^2$ chart involving subgrouped data when the standard values of both $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are unknown, i.e., the case in equation (4), is calculated using the SAS program in Figure 4. This program deals with the case of monitoring a future production, which is also referred to as phase 2.

Figure 4. A SAS program to compute the UCL for the $T^2$ chart based on equation (4)

```
Data EQ4;
p= ;
m= ;
n= ;
g= ;
a=p;
b=m*n-m-p+1;
UCL=p*(m+1)*(n-
1)/(m*n-m-
p+1)*Finv(1-g,a,b);
run;
proc print;
run;
```

The 2-of-3 Rule ($S_{II}$)

An out-of-control signal is given by this rule if two of three successive points plot above the UCL. For this case, by solving the corresponding linear system given in Khoo and Quah (2003), the $ARL_0$ formula is found to be

$$ARL_0 = \frac{1+2g-g^2}{g^2(2-g)} \qquad (7)$$

where g denotes the probability of a point falling above the UCL. Figure 5 provides a Mathematica 4.0 program for the computation of the probability *g* based on a fixed value of

$ARL_0$.

Figure 5. A Mathematica program to compute $g$ for rule $S_{II}$

```
ARL0 =
NSolve[(1 + 2g - g^2)/(g^2 (2 - g)) == ARL0, g]
```

Equation (5) is used to compute the UCL once the value of $g$ is obtained. The UCL of this rule for the $T^2$ chart based on the $T^2$ statistics in equations (1) and (3) can be computed using the SAS program in Figure 2 while that based on equations (2) and (4) are computed using the SAS programs shown in Figures (3) and (4) respectively.

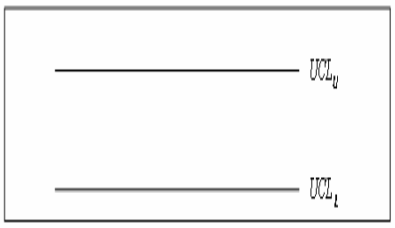
The Combined 1-of-1 and 2-of-2 Rules ($S_{III}$)

These combined rules signal an out-of-control if either a point plots above $UCL_U$ or two successive points plot between $UCL_L$ and $UCL_U$. The $ARL_0$ formula (Khoo, Quah and Low, 2004) is

$$ARL_0 = \frac{1+g}{g^2+h+gh} \qquad (8)$$

where $g$ is the probability that a point falls between $UCL_L$ and $UCL_U$ while $h$ denotes the probability of a point plotting above $UCL_U$. Figure 6 gives a graphical illustration of the limits.

Figure 6. The $UCL_U$ and $UCL_L$ limits for the combined rules




When the $T^2$ statistics are based on the formulas in equations (1) and (3), and for an arbitrary value of $p$, the UCLs of the corresponding conventional $T^2$ charts for these two cases can be computed using the SAS program given in Figure 2. For this case, $g$ is the desired Type-I error of each of the conventional chart. Similarly, the UCLs of the conventional $T^2$ charts based on the $T^2$ statistics in equations (2) and (4) can be obtained using the programs in Figures 3 and 4 respectively. After obtaining the UCL value of the $T^2$ chart for any of the four cases (equations (1), (2), (3) or (4)) of interest, choose a value of $UCL_U$, which is greater than that of the UCL. With this value of $UCL_U$, find $h$, the probability of a point falling above $UCL_U$. $h$ can be computed using the SAS programs in Figures 7, 8 and 9 for cases involving equations (1) and (3), equation (2) and equation (4) respectively.

A brief explanation for the program in Figure 8 will now be given. Because $UCL_U = \frac{p(m-1)(m+1)}{m(m-p)} F_{1-h,p,m-p}$, then the $1-h$ percentile of the $F$ distribution with parameters $p$ and $m-p$ is $F_{1-h,p,m-p} = UCL_U \Big/ \frac{p(m-1)(m+1)}{m(m-p)}$. Note that in Figure 8, $F_{1-h,p,m-p}$ is denoted as Finv. Thus, $h = 1 - P\left(Y < F_{1-h,p,m-p}\right)$ where $Y$ follows an $F$ distribution with parameters $p$ and $m-p$. In Figure 8, this probability is represented by h=**1–**Probf(Finv;a,b). The SAS program in Figure 9 can be explained in a similar manner.

Once the probability, $h$ is obtained, find the probability $g$ using equation (8) based on the $ARL_0$ value, which is chosen earlier. The Mathematica 4.0 program in Figure 10 can be used in this computation. Next, equation (5) is used to compute the limit $UCL_L$ by substituting $p_A$ with $g + h$. The computation of $UCL_L$ can be made using the SAS programs in Figures 11, 12 and 13 for the $T^2$ charts involving equations (1) and (3), equation (2) and equation (4) respectively. The user only needs to enter all the required values in the program which are already known at this stage.

Figure 7. A SAS program to compute *h* for the $T^2$ chart based on equations (1) and (3)

```
Data EQ1and3;
p= ;
UCLu= ;
h=1-Probchi(UCLu,p);
run;
proc print;
run;
```

Figure 8. A SAS program to compute *h* for the $T^2$ chart based on equation (2)

```
Data EQ2;
p= ;
m= ;
UCLu= ;
a=p;
b=m-p;
Finv=UCLu/(p*(m-1)*(m+1)/
(m*(m-p)));
h=1-Probf(Finv;a,b);
run;
proc print;
run;
```

Figure 9. A SAS program to compute *h* for the $T^2$ chart based on equation (4)

```
Data EQ4;
p= ;
m= ;
n= ;
UCLu= ;
a=p;
b=m*n-m-p+1;
Finv=UCLu/(p*(m+1)*(n-1)/(m*n-m-
p+1));
h=1-Probf(Finv;a,b);
run;
proc print;
run;
```

Figure 10. A Mathematica program to compute *g* for rule $S_{III}$

h =
ARL0 =

$$\text{NSolve}\left[\frac{1+g}{g^2+h+gh} == \text{ARL0}, g\right]$$

Figure 11. A SAS program to compute the $UCL_L$ for the $T^2$ chart based on equations (1) and (3)

```
Data EQ1and3;
p= ;
g= ;
h= ;
UCLL=Cinv(1-g-h,p);
run;
proc print;
run;
```

Figure 12. A SAS program to compute the $UCL_L$ for the $T^2$ chart based on equation (2)

```
Data EQ2;
p= ;
m= ;
g= ;
h= ;
a=p;
b=m-p;
UCLL=p*(m-1)*(m+1)/(m*
(m-p))*Finv(1-g-h,a,b);
run;
proc print;
run;
```

Figure 13. A SAS program to compute the $UCL_L$ for the $T^2$ chart based on equation (4)

```
Data EQ4;
p= ;
m= ;
n= ;
g= ;
h= ;
a=p;
b=m*n–m–p+1;
UCLL=p*(m+1)*(n-1)/(m*n-m-
p+1)*Finv(1-g-h,a,b);
run;
proc print;
run;
```

The combined 1-of-1 and 2-of-3 rules ($S_{IV}$)

These combined rules give an out-of-control signal if a point exceeds $UCL_U$, or if two of three consecutive points plot between $UCL_L$ and $UCL_U$ (see Figure 6). Here, the $ARL_0$ formula is (Khoo, Quah and Low, 2004):

$$ARL_0 = \frac{-1+g^2+g(-2+h)}{g^3+2g^2(-1+h)-h+g(-2+h)h}. \quad (9)$$

In equation (9), $g$ is the probability of a point falling between $UCL_L$ and $UCL_U$ and $h$ is the probability that a point plots above the $UCL_U$.

Similar to the previous combined rules, first choose a $UCL_U$ value that is larger than the UCL limit of the conventional $T^2$ chart. The UCL of the conventional chart for the four different cases in equations (1), (2), (3) and (4) based on a desired Type-I error can be easily determined using the same approach discussed for rule $S_{III}$. Based on a chosen value of $UCL_U$, find $h$, the probability of a point plotting above $UCL_U$. $h$ is found from the programs in Figures 7, 8 and 9 for cases involving equations (1) and (3), equation (2) and equation (4) respectively.

After obtaining $h$, find the probability $g$ from equation (9). This is made using the Mathematica 4.0 program in Figure 14. Then, use equation (5) to calculate the limit $UCL_L$ by replacing $p_A$ with $g + h$. $UCL_L$ can be calculated from the SAS programs in Figures 11, 12 and 13 for the $T^2$ charts of equations (1) and (3), equation (2) and equation (4) respectively.

Figure 14. A Mathematica program to compute $g$ for rule $S_{IV}$

```
h =
ARL0 =
NSolve[ (-1 + g^2 + g(-2 + h))/(g^3 + 2g^2(-1 + h) - h + g(-2 + h)h)
== ARL0, g]
```

Performance Evaluation by Means of a Simulation Study

A simulation study is conducted using Statistical Analysis System (SAS) version 8.02 to evaluate the performances of the sensitizing rules discussed in the previous section. The process is assumed to follow a bivariate normal, $N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution. The in-control mean vector is $\boldsymbol{\mu}_0 = (0,0)'$ while the covariance matrix is $\boldsymbol{\Sigma} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$, where $\rho$ is the correlation coefficient between the two quality characteristics. Due to the directionally invariant property of the Hotelling, $T^2$ control chart, the value of $\rho$ ($-1 < \rho < 1$) will not have any influence on the performance of the chart. The chart's performance is only dependent on the magnitude of a shift given by $\lambda$. Hence, $\rho = 0$ is considered in this simulation study. The magnitude of shifts in the mean vector considered are $\lambda \in \{0, 0.25, 0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0, 5.0\}$ for the case of individual observations and $\lambda \in \{0, 0.25, 0.30, 0.40, 0.50, 0.75, 1.00, 1.50, 2.00, 3.00\}$ for the case of subgrouped data where $\lambda^2$ is the noncentrality parameter given by

$$\lambda^2 = (\boldsymbol{\mu}_S - \boldsymbol{\mu}_0)' \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_S - \boldsymbol{\mu}_0). \quad (10)$$

Here, $\boldsymbol{\mu}_S = (\delta,0)'$ represents the off-target mean vector.

Three in-control ARL values are

considered, i.e., 500, 750 and 1000. The $T^2$ statistics in equations (1) for individual observations and (3) for subgrouped data, are considered because this simulation study is conducted with the assumption that the on-target values of both $\boldsymbol{\mu}_0$ and $\boldsymbol{\Sigma}$ are known. The limits of the conventional $T^2$ charts and that based on the sensitizing rules for individual observations and subgrouped data with a sample size, *n*, are similar for the same rule if they have a similar in-control ARL because the charts' statistics follow the same distribution, i.e., $\chi^2_2$. Note that the limits of all the rules are computed using the SAS programs given in the previous section. The values of these limits for the various rules are shown in Tables 1 – 9. For the subgrouped data, samples of sizes $n$ = 5 and 10 are considered. For the combined rules of $S_{III}$ and $S_{IV}$, the $UCL_U$ value of 15 is used for the $T^2$ charts in Tables 1 – 9. Note that $UCL_U$ = 15 is greater than the limits of the conventional $T^2$ charts for all $ARL_0$ values.

The simulation results for the conventional $T^2$ chart together with the limits of the $S_I, S_{II}, S_{III}$ and $S_{IV}$ schemes are shown in Tables 1 – 9 where the first three tables are based on individual observations, the next three tables are based on subgrouped data with sample size, $n$ = 5 and the last three tables are based on subgrouped data with sample size, $n$ = 10. Tables 1, 4 and 7 have an in-control ARL of 1000, Tables 2, 5 and 8 with $ARL_0$ of 750 while the $ARL_0$ value in Tables 3, 6 and 9 is 500.

The results in all the tables show that the 2-of-2 ($S_I$) and 2-of-3 ($S_{II}$) rules outperform the conventional $T^2$ chart in most cases except for very large magnitude of shifts. For the results of the individual observations in Tables 1 – 3, these two sensitizing rules outperform the conventional $T^2$ chart for $0 < \lambda < 3$ and they are only slightly less effective than the latter when $\lambda > 3$. For the results of the subgrouped data in Tables 4 – 9, the performances of these two rules are superior to the $T^2$ chart for $0 < \lambda < 1$. The performances of these two rules are only slightly inferior to the latter for $\lambda > 1$. The combined rules of $S_{III}$ and $S_{IV}$, however, provide excellent results where they improve the performances of the conventional $T^2$ chart for small to moderate magnitude of shifts while maintaining the same sensitivity for large shifts. This is evident from the results in Tables 1 – 9. The results show that the performances of the combined rules of $S_{III}$ and $S_{IV}$ are at par with that of rules $S_I$ and $S_{II}$ for small to moderate magnitude of shifts while slightly outperforming the two latter rules for large shifts.

## Examples of Application

### Example 1

This example deals with a small magnitude of shift in the mean vector involving individual measurements. The first 20 bivariate observations are generated from a bivariate normal, $N_2(\boldsymbol{\mu}_0, \boldsymbol{\Sigma})$ distribution, where $\boldsymbol{\mu}_0 = (0,0)'$ is the on target mean vector and $\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$ is the covariance matrix. These bivariate observations represent the data from an in-control process. For the o.o.c. case which consists of the next 20 observations, the process is assumed to follow a $N_2(\boldsymbol{\mu}_S, \boldsymbol{\Sigma})$ distribution, where $\boldsymbol{\mu}_S = (1,0)'$. Note that all the observations are generated using the SAS program. Because $\boldsymbol{\mu}_0$ and $\boldsymbol{\Sigma}$ are both known, the $T^2$ statistics are computed using equation (1). An in-control ARL of 500 is considered. The values of the $T^2$ statistics and variables $X_1$ and $X_2$ for vector $\mathbf{X} = (X_1, X_2)'$ from observations 1 – 40 are presented in Table 10.

The $T^2$ statistics are plotted on the Hotelling $T^2$ chart whose limit is computed from the conventional rule using the SAS program in Figure 2 to be UCL = 12.4292 because $p$ = 2 and $g = 1/500$. Besides the conventional approach, an additional o.o.c. test considered is that based on the combined 1-of-1 and 2-of-2 rules, a.k.a., rule $S_{III}$. The $UCL_U$ of this rule is set as 15 so that $UCL_U$ > UCL.

Table 1. ARL profiles based on $ARL_0 = 1000$ and $\boldsymbol{\mu}_S = (\delta,0)'$ for individual observations

| $\lambda = \delta$ | Conventional $T^2$ ($UCL$ = 13.8155) | $S_I$ ($UCL$ = 6.87614) | $S_{II}$ ($UCL$ = 7.54488) | $S_{III}$ ($UCL_U = 15$ & $UCL_L = 7.64089$) | $S_{IV}$ ($UCL_U = 15$ & $UCL_L = 8.29725$) |
|---|---|---|---|---|---|
| 0 | 1001.83 | 996.14 | 998.95 | 1002.31 | 999.65 |
| 0.25 | 817.40 | 817.23 | 805.71 | 805.11 | 801.37 |
| 0.5 | 499.37 | 491.04 | 460.87 | 465.98 | 457.55 |
| 1.0 | 146.70 | 120.35 | 106.26 | 115.68 | 109.06 |
| 1.5 | 44.39 | 31.55 | 27.52 | 30.72 | 28.26 |
| 2.0 | 15.83 | 11.12 | 9.84 | 10.82 | 9.97 |
| 2.5 | 6.83 | 5.33 | 4.89 | 4.86 | 4.57 |
| 3.0 | 3.48 | 3.38 | 3.16 | 2.75 | 2.70 |
| 3.5 | 2.11 | 2.55 | 2.46 | 1.91 | 1.91 |
| 4.0 | 1.50 | 2.21 | 2.18 | 1.48 | 1.49 |
| 5.0 | 1.09 | 2.02 | 2.02 | 1.10 | 1.11 |

Table 2. ARL profiles based on $ARL_0 = 750$ and $\boldsymbol{\mu}_S = (\delta,0)'$ for individual observations

| $\lambda = \delta$ | Conventional $T^2$ ($UCL$ = 13.2401) | $S_I$ ($UCL$ = 6.58356) | $S_{II}$ ($UCL$ = 7.24851) | $S_{III}$ ($UCL_U = 15$ & $UCL_L = 7.08929$) | $S_{IV}$ ($UCL_U = 15$ & $UCL_L = 7.74539$) |
|---|---|---|---|---|---|
| 0 | 749.29 | 750.54 | 750.81 | 753.21 | 751.39 |
| 0.25 | 617.53 | 615.78 | 605.18 | 606.72 | 598.38 |
| 0.5 | 384.40 | 375.83 | 360.34 | 357.75 | 353.88 |
| 1.0 | 117.18 | 98.12 | 86.27 | 91.83 | 85.88 |
| 1.5 | 36.85 | 26.87 | 23.53 | 25.51 | 23.52 |
| 2.0 | 13.56 | 9.98 | 9.06 | 9.34 | 8.52 |
| 2.5 | 6.00 | 5.00 | 4.63 | 4.43 | 4.20 |
| 3.0 | 3.20 | 3.24 | 3.07 | 2.61 | 2.57 |
| 3.5 | 2.01 | 2.49 | 2.45 | 1.86 | 1.85 |
| 4.0 | 1.46 | 2.19 | 2.16 | 1.46 | 1.46 |
| 5.0 | 1.07 | 2.02 | 2.01 | 1.10 | 1.11 |

Table 3. ARL profiles based on $ARL_0 = 500$ and $\boldsymbol{\mu}_S = (\delta,0)'$ for individual observations

| $\lambda = \delta$ | Conventional $T^2$ ($UCL$ = 12.4292) | $S_I$ ($UCL$ = 6.16989) | $S_{II}$ ($UCL$ = 6.82846) | $S_{III}$ ($UCL_U = 15$ & $UCL_L = 6.47195$) | $S_{IV}$ ($UCL_U = 15$ & $UCL_L = 7.1244$) |
|---|---|---|---|---|---|
| 0 | 500.59 | 498.26 | 498.02 | 501.13 | 500.24 |
| 0.25 | 419.81 | 414.91 | 416.32 | 407.17 | 410.95 |
| 0.5 | 265.92 | 259.30 | 247.35 | 248.02 | 243.32 |
| 1.0 | 85.71 | 73.99 | 64.97 | 68.82 | 63.23 |
| 1.5 | 28.42 | 21.68 | 19.36 | 20.41 | 19.07 |
| 2.0 | 10.90 | 8.62 | 7.88 | 7.94 | 7.32 |
| 2.5 | 5.06 | 4.53 | 4.23 | 3.96 | 3.80 |
| 3.0 | 2.81 | 3.05 | 2.93 | 2.46 | 2.45 |
| 3.5 | 1.82 | 2.42 | 2.38 | 1.81 | 1.81 |
| 4.0 | 1.36 | 2.16 | 2.13 | 1.45 | 1.44 |
| 5.0 | 1.06 | 2.01 | 2.01 | 1.10 | 1.11 |

Table 4. ARL Profiles based on $ARL_0 = 1000$, $\boldsymbol{\mu}_S = (\delta,0)'$ and $n = 5$

| $\lambda = \delta$ | Conventional $T^2$ ($UCL$ = 13.8155) | $S_I$ ($UCL$ = 6.87614) | $S_{II}$ ($UCL$ = 7.54488) | $S_{III}$ ($UCL_U = 15$ & $UCL_L = 7.64089$) | $S_{IV}$ ($UCL_U = 15$ & $UCL_L = 8.29725$) |
|---|---|---|---|---|---|
| 0 | 999.15 | 999.87 | 999.81 | 1000.25 | 995.78 |
| 0.25 | 436.34 | 427.85 | 401.41 | 402.44 | 391.53 |
| 0.30 | 330.97 | 314.71 | 288.86 | 295.56 | 284.64 |
| 0.40 | 190.53 | 160.42 | 143.78 | 155.69 | 146.70 |
| 0.50 | 108.59 | 84.89 | 74.51 | 81.93 | 78.89 |
| 0.75 | 30.22 | 21.23 | 18.58 | 20.22 | 18.96 |
| 1.00 | 10.47 | 7.72 | 6.86 | 7.17 | 6.77 |
| 1.50 | 2.43 | 2.71 | 2.65 | 2.07 | 2.09 |
| 2.00 | 1.25 | 2.07 | 2.06 | 1.25 | 1.26 |
| 3.00 | 1.00 | 2.00 | 2.00 | 1.00 | 1.00 |

Table 5. ARL Profiles based on $ARL_0 = 750$, $\boldsymbol{\mu}_S = (\delta,0)'$ and $n = 5$

| $\lambda = \delta$ | Conventional $T^2$ ($UCL$ = 13.2401) | $S_I$ ($UCL$ = 6.58356) | $S_{II}$ ($UCL$ = 7.24851) | $S_{III}$ ($UCL_U = 15$ & $UCL_L = 7.08929$) | $S_{IV}$ ($UCL_U = 15$ & $UCL_L = 7.74539$) |
|---|---|---|---|---|---|
| 0 | 747.90 | 754.01 | 748.27 | 747.18 | 754.15 |
| 0.25 | 337.59 | 332.34 | 295.56 | 302.83 | 290.67 |
| 0.30 | 257.82 | 239.66 | 221.70 | 228.97 | 224.04 |
| 0.40 | 151.95 | 129.44 | 114.90 | 121.92 | 113.32 |
| 0.50 | 90.65 | 70.32 | 61.06 | 66.66 | 62.29 |
| 0.75 | 25.54 | 18.36 | 16.75 | 17.41 | 15.88 |
| 1.00 | 9.04 | 6.97 | 6.22 | 6.34 | 5.97 |
| 1.50 | 2.24 | 2.64 | 2.55 | 2.02 | 2.02 |
| 2.00 | 1.20 | 2.06 | 2.06 | 1.24 | 1.24 |
| 3.00 | 1.00 | 2.00 | 2.00 | 1.00 | 1.00 |

Table 6. ARL Profiles based on $ARL_0 = 500$, $\boldsymbol{\mu}_S = (\delta,0)'$ and $n = 5$

| $\lambda = \delta$ | Conventional $T^2$ ($UCL$ = 12.4292) | $S_I$ ($UCL$ = 6.16989) | $S_{II}$ ($UCL$ = 6.82846) | $S_{III}$ ($UCL_U = 15$ & $UCL_L = 6.47195$) | $S_{IV}$ ($UCL_U = 15$ & $UCL_L = 7.1244$) |
|---|---|---|---|---|---|
| 0 | 504.07 | 499.00 | 503.19 | 498.88 | 503.65 |
| 0.25 | 232.42 | 232.34 | 214.33 | 210.87 | 210.45 |
| 0.30 | 182.91 | 175.59 | 154.77 | 156.20 | 153.86 |
| 0.40 | 110.66 | 94.52 | 85.90 | 89.19 | 82.59 |
| 0.50 | 67.85 | 53.89 | 47.81 | 51.51 | 46.06 |
| 0.75 | 20.25 | 15.06 | 13.70 | 14.18 | 13.10 |
| 1.00 | 7.52 | 6.24 | 5.48 | 5.59 | 5.13 |
| 1.50 | 1.97 | 2.57 | 2.47 | 1.96 | 1.90 |
| 2.00 | 1.15 | 2.05 | 2.05 | 1.25 | 1.26 |
| 3.00 | 1.00 | 2.00 | 2.00 | 1.00 | 1.00 |

Table 7. ARL Profiles based on $ARL_0 = 1000$, $\boldsymbol{\mu}_S = (\delta,0)'$ and $n = 10$

| $\lambda = \delta$ | Conventional $T^2$ ($UCL$ = 13.8155) | $S_I$ ($UCL$ = 6.87614) | $S_{II}$ ($UCL$ = 7.54488) | $S_{III}$ ($UCL_U = 15$ & $UCL_L = 7.64089$) | $S_{IV}$ ($UCL_U = 15$ & $UCL_L = 8.29725$) |
|---|---|---|---|---|---|
| 0 | 995.08 | 1004.27 | 1003.12 | 995.48 | 995.58 |
| 0.25 | 255.15 | 215.16 | 203.18 | 218.26 | 188.25 |
| 0.30 | 164.26 | 140.21 | 124.81 | 141.86 | 120.73 |
| 0.40 | 78.44 | 55.18 | 50.65 | 57.51 | 51.57 |
| 0.50 | 38.28 | 25.30 | 23.30 | 24.94 | 23.69 |
| 0.75 | 8.17 | 6.55 | 6.08 | 6.00 | 5.47 |
| 1.00 | 2.94 | 2.99 | 2.91 | 2.45 | 2.45 |
| 1.50 | 1.13 | 2.03 | 2.04 | 1.18 | 1.18 |
| 2.00 | 1.00 | 2.00 | 2.00 | 1.01 | 1.01 |
| 3.00 | 1.00 | 2.00 | 2.00 | 1.00 | 1.00 |

Table 8. ARL Profiles based on $ARL_0 = 750$, $\boldsymbol{\mu}_S = (\delta,0)'$ and $n = 10$

| $\lambda = \delta$ | Conventional $T^2$ ($UCL$ = 13.2401) | $S_I$ ($UCL$ = 6.58356) | $S_{II}$ ($UCL$ = 7.24851) | $S_{III}$ ($UCL_U = 15$ & $UCL_L = 7.08929$) | $S_{IV}$ ($UCL_U = 15$ & $UCL_L = 7.74539$) |
|---|---|---|---|---|---|
| 0 | 750.00 | 750.34 | 751.20 | 750.69 | 747.47 |
| 0.25 | 185.83 | 176.64 | 148.82 | 172.52 | 149.70 |
| 0.30 | 128.30 | 114.54 | 97.52 | 109.93 | 93.93 |
| 0.40 | 61.94 | 48.94 | 44.05 | 46.01 | 42.01 |
| 0.50 | 32.25 | 23.26 | 18.93 | 21.38 | 19.70 |
| 0.75 | 6.98 | 5.98 | 5.31 | 5.14 | 4.81 |
| 1.00 | 2.76 | 2.97 | 2.88 | 2.35 | 2.36 |
| 1.50 | 1.13 | 2.04 | 2.03 | 1.18 | 1.18 |
| 2.00 | 1.00 | 2.00 | 2.00 | 1.01 | 1.01 |
| 3.00 | 1.00 | 2.00 | 2.00 | 1.00 | 1.00 |

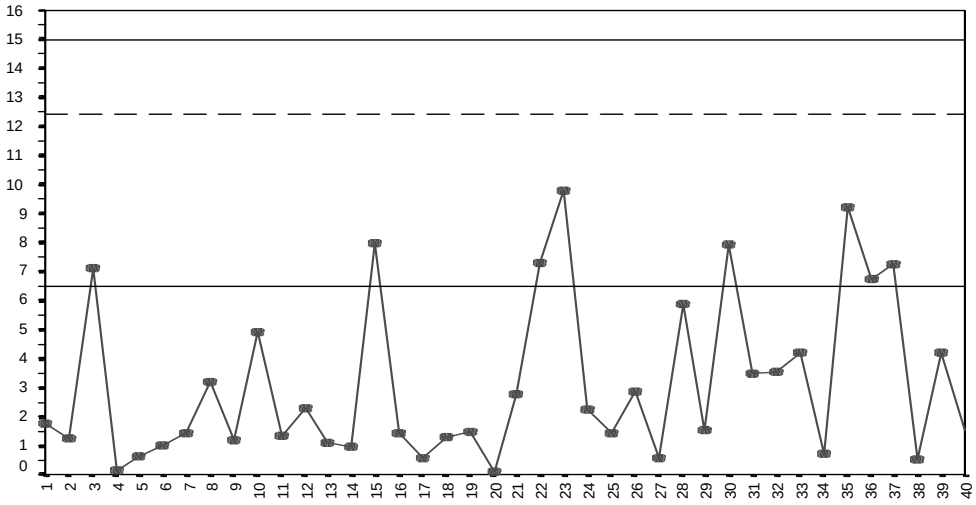
Table 9. ARL Profiles based on $ARL_0 = 500$, $\boldsymbol{\mu}_S = (\delta,0)'$ and $n = 10$

| $\lambda = \delta$ | Conventional $T^2$ ($UCL$ = 12.4292) | $S_I$ ($UCL$ = 6.16989) | $S_{II}$ ($UCL$ = 6.82846) | $S_{III}$ ($UCL_U = 15$ & $UCL_L = 6.47195$) | $S_{IV}$ ($UCL_U = 15$ & $UCL_L = 7.1244$) |
|---|---|---|---|---|---|
| 0 | 505.75 | 502.30 | 501.52 | 505.22 | 499.73 |
| 0.25 | 137.80 | 126.49 | 111.39 | 119.39 | 104.06 |
| 0.30 | 93.79 | 82.06 | 76.72 | 80.06 | 72.60 |
| 0.40 | 45.25 | 34.72 | 34.27 | 35.25 | 32.49 |
| 0.50 | 23.63 | 18.41 | 16.14 | 16.84 | 15.59 |
| 0.75 | 6.16 | 5.35 | 4.76 | 4.46 | 4.33 |
| 1.00 | 2.48 | 2.81 | 2.77 | 2.20 | 2.23 |
| 1.50 | 1.11 | 2.03 | 2.02 | 1.17 | 1.18 |
| 2.00 | 1.00 | 2.00 | 2.00 | 1.01 | 1.01 |
| 3.00 | 1.00 | 2.00 | 2.00 | 1.00 | 1.00 |

Table 10. The Computed $T_i^2$ Statistics for Example 1

| Obs. no., $i$ | $X_1$ | $X_2$ | $T_i^2$ | Obs. no., $i$ | $X_1$ | $X_2$ | $T_i^2$ |
|---|---|---|---|---|---|---|---|
| 1 | –0.344 | –1.286 | 1.774 | 21 | 1.585 | 0.361 | 2.762 |
| 2 | –0.882 | 0.150 | 1.245 | 22 | 2.569 | 2.007 | 7.295 |
| 3 | –1.990 | 0.545 | 7.125 | 23 | 3.045 | 0.909 | 9.772 |
| 4 | –0.343 | –0.067 | 0.132 | 24 | 1.297 | –0.005 | 2.252 |
| 5 | –0.800 | –0.358 | 0.643 | 25 | 1.168 | 0.830 | 1.446 |
| 6 | –0.620 | 0.364 | 0.990 | 26 | 0.595 | –1.080 | 2.884 |
| 7 | –0.004 | –1.041 | 1.440 | 27 | 0.314 | 0.769 | 0.597 |
| 8 | 1.479 | –0.131 | 3.197 | 28 | 1.875 | –0.386 | 5.854 |
| 9 | –1.082 | –0.478 | 1.175 | 29 | 0.393 | –0.823 | 1.540 |
| 10 | 1.549 | –0.602 | 4.927 | 30 | 1.070 | –1.718 | 7.911 |
| 11 | –0.317 | –1.128 | 1.353 | 31 | 1.841 | 1.167 | 3.471 |
| 12 | 0.408 | 1.464 | 2.282 | 32 | 1.868 | 1.100 | 3.525 |
| 13 | 0.639 | 1.037 | 1.094 | 33 | 1.214 | –0.823 | 4.202 |
| 14 | –0.879 | –0.080 | 0.945 | 34 | 0.151 | –0.643 | 0.712 |
| 15 | –2.294 | 0.286 | 7.997 | 35 | 2.046 | –0.917 | 9.202 |
| 16 | 0.060 | 1.066 | 1.434 | 36 | 1.804 | 2.521 | 6.749 |
| 17 | –0.586 | 0.127 | 0.578 | 37 | 0.988 | –1.678 | 7.264 |
| 18 | –0.818 | 0.279 | 1.300 | 38 | –0.344 | –0.718 | 0.516 |
| 19 | –0.600 | 0.610 | 1.464 | 39 | 1.873 | 0.223 | 4.188 |
| 20 | 0.127 | –0.209 | 0.115 | 40 | 0.671 | 1.229 | 1.514 |

Figure 15. The $T^2$ chart with limits of the conventional and $S_{III}$ rules for example 1

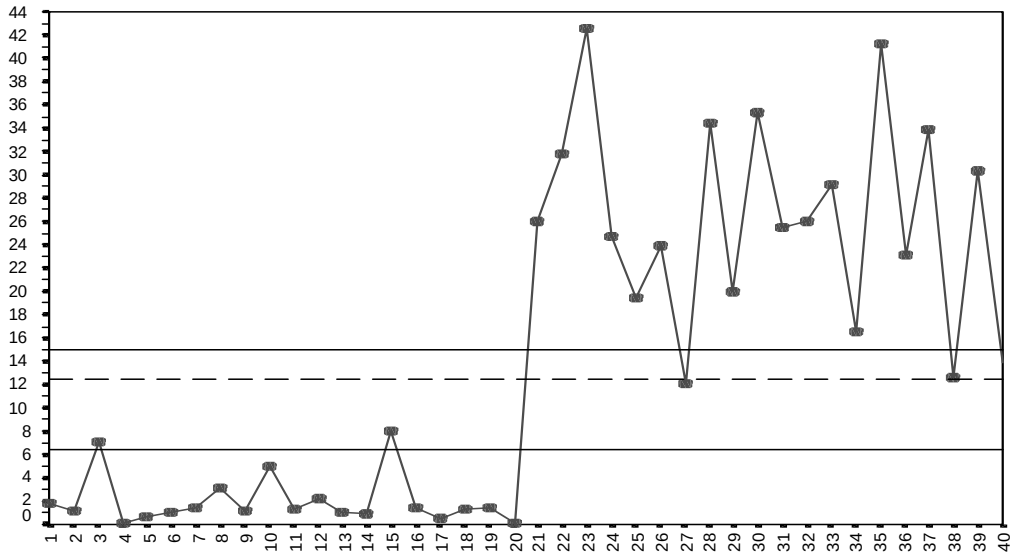



*Note*: The top parallel line is $UCL_u$, the slashed parallel line is UCL, and the lower parallel line is $UCL_l$.

Table 11. The computed $T_i^2$ statistics for example 2

| Obs. no., $i$ | $X_1$ | $X_2$ | $T_i^2$ | Obs. no., $i$ | $X_1$ | $X_2$ | $T_i^2$ |
|---|---|---|---|---|---|---|---|
| 1 | –0.344 | –1.286 | 1.774 | 21 | 4.585 | 0.361 | 26.002 |
| 2 | –0.882 | 0.150 | 1.245 | 22 | 5.569 | 2.007 | 31.820 |
| 3 | –1.990 | 0.545 | 7.125 | 23 | 6.045 | 0.909 | 42.495 |
| 4 | –0.343 | –0.067 | 0.132 | 24 | 4.297 | –0.005 | 24.648 |
| 5 | –0.800 | –0.358 | 0.643 | 25 | 4.168 | 0.830 | 19.473 |
| 6 | –0.620 | 0.364 | 0.990 | 26 | 3.595 | –1.080 | 23.965 |
| 7 | –0.004 | –1.041 | 1.440 | 27 | 3.314 | 0.769 | 12.038 |
| 8 | 1.479 | –0.131 | 3.197 | 28 | 4.875 | –0.386 | 34.401 |
| 9 | –1.082 | –0.478 | 1.175 | 29 | 3.393 | –0.823 | 19.972 |
| 10 | 1.549 | –0.602 | 4.927 | 30 | 4.070 | –1.718 | 35.340 |
| 11 | –0.317 | –1.128 | 1.353 | 31 | 4.841 | 1.167 | 25.532 |
| 12 | 0.408 | 1.464 | 2.282 | 32 | 4.868 | 1.100 | 26.064 |
| 13 | 0.639 | 1.037 | 1.094 | 33 | 4.214 | –0.823 | 29.209 |
| 14 | –0.879 | –0.080 | 0.945 | 34 | 3.151 | –0.643 | 16.498 |
| 15 | –2.294 | 0.286 | 7.997 | 35 | 5.046 | –0.917 | 41.236 |
| 16 | 0.060 | 1.066 | 1.434 | 36 | 4.804 | 2.521 | 23.100 |
| 17 | –0.586 | 0.127 | 0.578 | 37 | 3.988 | –1.678 | 33.879 |
| 18 | –0.818 | 0.279 | 1.300 | 38 | 2.656 | –0.718 | 12.637 |
| 19 | –0.600 | 0.610 | 1.464 | 39 | 4.873 | 0.223 | 30.282 |
| 20 | 0.127 | –0.209 | 0.115 | 40 | 3.671 | 1.229 | 13.966 |

Figure 16. The $T^2$ chart with limits of the conventional and $S_{III}$ rules for example 2




*Note*: The top parallel line is $UCL_u$, the slashed parallel line is UCL, and the lower parallel line is $UCL_l$.

From the SAS programs in Figures 7 and 11, $UCL_L$ is computed to be 6.47195. The $T^2$ statistics are plotted on the $T^2$ chart with limit UCL = 12.4292 on Figure 15. Additional limits which consist of $UCL_U$ = 15 and $UCL_L$ = 6.47195 are drawn on this chart for rule $S_{III}$. Figure 15 shows that the conventional rule fails to detect a shift in the mean vector. The superiority of rule $S_{III}$ is obvious in that it detects the first off-target signal at observation 23.

Example 2

The data in this example, which are generated using the SAS program, involves a shift of a large magnitude in the mean vector. Here, the first 20 bivariate observations are generated from a $N_2(\boldsymbol{\mu}_0, \boldsymbol{\Sigma})$ distribution, where $\boldsymbol{\mu}_0 = (0,0)'$ is the on target mean vector and $\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$ is the covariance matrix. This is followed by generating another 20 bivariate observations from a $N_2(\boldsymbol{\mu}_S, \boldsymbol{\Sigma})$ distribution where $\boldsymbol{\mu}_S = (4,0)'$, to represent the o.o.c. process. The $T^2$ statistics are computed from the formula in eq. (1). The values of the $T^2$ statistics and quality characteristics $X_1$ and $X_2$ for observations 1 – 40 are given in Table 11.

Figure 16 gives the $T^2$ chart, which consists of the $T^2$ statistics plotted on it. The same value of $ARL_0$ considered in Example 1 is used here. The UCL of the conventional $T^2$ chart is computed using the same approach described in Example 1 to be 12.4292. Similar to Example 1, rule $S_{III}$ is also considered. The limits of this rule are obtained using the same approach to be $UCL_U$ = 15 and $UCL_L$ = 6.47195. An o.o.c. signal is detected at observation 21 by both the conventional and $S_{III}$ rules. This example shows that rule $S_{III}$ has the same sensitivity as the conventional rule in the detection of a large magnitude of shift.

## Conclusion

This article provides a nonrigorous approach of implementing sensitizing rules on a Hotelling control chart. The advantage of the approach presented in this article where the $T^2$ statistics do not need to be transformed into normal random variables enable the statistics to be plotted on the original scale so that the incorporation of runs rules can be made on the same conventional chart without having to maintain a separate chart specially designed for plotting the transformed variables which follow a normal distribution. The suggested approach is a remarkable improvement of the earlier works of Khoo and Quah (2003) and Khoo, Quah and Low (2004). The Mathematica and SAS programs provided in this article will certainly serve as useful tools in assisting practitioners in the design and implementation of the various rules.

## References

Apley, D. W. & Tsung, F. (2002). The autoregressive $T^2$ chart for monitoring univariate autocorrelated processes. *Journal of Quality Technology*, *34*, 80 – 96.

Aparisi, F. (1997). Sampling plans for the multivariate $T^2$ control chart. *Quality Engineering, 10,* 141 – 147.

Champ, C. W. & Woodall, W. H. (1987). Exact results for shewhart control charts with supplementary runs rules. *Technometrics, 29,* 393 – 399.

Doganaksoy, N., Faltin, F. W., & Tucker, W. T. (1991). Identification of out of control quality characteristics in a multivariate manufacturing environment. *Communications in Statistics: Theory and Methods, 20,* 2775 – 2790.

Holmes, D. S. & Mergen, A. E. (1995). Identifying the sources for out-of-control signals when the $T^2$ control chart is used. *Quality Engineering, 8,* 137 – 143.

Hotelling, H. (1947). *Multivariate quality control, techniques of statistical analysis*, Eisenhart, Hastay and Wallis (eds.). New York: McGraw-Hill.

Khoo, M. B. C., & Quah, S. H. (2003). Incorporating runs rules into Hotelling $\chi^2$ control charts. *Quality Engineering, 15,* 671 – 675.

Khoo, M. B. C., Quah, S. H., & Low, H. C. (2004). Powerful rules for the Hotelling's $\chi^2$ control chart. *Quality Engineering, 17,* 139 – 149.

Klein, M. (2000). Two alternatives to the shewhart $\bar{X}$ control chart. *Journal of Quality Technology, 32,* 427 – 431.

Mason, R. L., Champ, C. W., Tracy, N. D., Wierda, S. J., & Young, J. C. (1997). Assessment of multivariate process control techniques. *Journal of Quality Technology, 29,* 140 – 143.

Mason, R. L., Chou, Y. M., & Young, J. C. (2001). Applying Hotelling's $T^2$ statistic to batch processes. *Journal of Quality Technology, 33,* 466 – 479.

Mason, R. L., Tracy, N. D., & Young, J. C. (1995). Decomposition of $T^2$ for multivariate control chart interpretation. *Journal of Quality Technology, 27,* 99 – 108.

Mason, R. L., Tracy, N. D., & Young, J. C. (1997). A practical approach for interpreting multivariate $T^2$ control chart signals. *Journal of Quality Technology, 29,* 396 – 406.

Montgomery, D. C. (2001). *Introduction to statistical quality control* (4th ed.). New York: John Wiley and Sons.

Nelson, L. S. (1984). The shewhart control chart: Tests for special causes. *Journal of Quality Technology, 16,* 237 – 239.

Nedumaran, G. & Pignatiello, J. J. (1998). Diagnosing signals from $T^2$ and $\chi^2$ multivariate control charts. *Quality Engineering, 10,* 657 – 667.

Prins, J. & Mader, D. (1997). Multivariate control charts for grouped and individual observations. *Quality Engineering, 10,* 49 – 57.

Runger, G. C. (1996). Projections and the $U^2$ multivariate control chart. *Journal of Quality Technology, 28,* 313 – 319.

Runger, G. C., Alt, F. B., & Montgomery. D. C. (1996). Contributions to a multivariate statistical process control chart signal. *Communications in Statistics: Theory and Methods, 25,* 2203 – 2213.

Sullivan, J. H., & Woodall W. H. (1996). A comparison of multivariate control charts for individual observations. *Journal of Quality Technology, 28,* 398 – 408.

Tracy, N. D., Young, J. C., & Mason, R. L. (1992). Multivariate control charts for individual observations. *Journal of Quality Technology, 24,* 88 – 95.

Timm, N. H. (1996). Multivariate quality control using finite intersection tests. *Journal of Quality Technology, 28,* 233 – 243.

Vargas, J. A. (2003). Robust estimation in multivariate control charts for individual observations. *Journal of Quality Technology, 35,* 367 – 376.

Journal of Modern Applied Statistical Methods
November, 2005, Vol. 4, No. 2, 583-586


1538 – 9472/05/$95.00

# *BRIEF REPORTS*

## Inference on $P(Y < X)$ in a Pareto Distribution

M. Masoom Ali
Department of Mathematical Sciences
Ball State University

Jungsoo Woo
Department of Statistics
Yeungnam University

Inference on the reliability $R = P(Y < X)$ in a Pareto distribution with a known scale parameter is considered. Point estimates and confidence intervals of $R$ are obtained a test of hypothesis is also considered.

Key words: MLE, MSE

### Introduction

A Pareto distribution is given by

$$f(x;\alpha,\beta)=\frac{\alpha}{\beta(1+x/\beta)^{\alpha+1}},\ x>0,\ \alpha,\ \beta>0.$$

Pareto law has been universal and inevitable, regardless of taxation and social and political conditions. More recently, attempts have been made to explain many empirical phenomena using the Pareto distribution (see Moothathu, 1984; Arnold & Press, 1983). Ali, et al, (2005a and 2005b) considered the problem for some other distributions. The probability that a Weibull random variable $Y$ is less than another independent Weibull random variable $X$ was considered by McCool (1991). Baklizi (2003) considered the confidence interval of $P(X < Y)$ in the exponential case with common location.

M.Masoom Ali is George & Frances Ball Distinguished Professor of Statistics. His research interests are in order statistics, Bayesian statistics, statistical inference, and distribution problems. Email him at mali@bsu.edu. Jungsoo Woo is Professor of Statistics. His research interests are in Bayesian statistics, statistical inference and distribution problems.

The problem of estimating and of drawing inferences about the probability that a random variable $Y$ is less than another independent random variable $X$ arise in reliability studies.

When $Y$ represents the random variable of a stress that a device will be subjected to in service and $X$ represents the strength that varies from item to item in the population of devices, then the reliability $R$, i.e., the probability that a randomly selected device functions successfully, is equal to $P(Y < X)$. The same problem also arises in the context of statistical tolerance where Y represents, say, the diameter of a shaft and $X$ the diameter of a bearing that is to be mounted on the shaft. The probability that the bearing fits without interference is the $P(Y < X)$. In biometry, Y represents a patient's remaining years of life if treated with drug A and X represents the patient's remaining years when treated with drug B. If the choice of drug is left to the patient, person's deliberations will center on whether $P(Y < X)$ is less than or greater than 1/2.

In this article, the problem of estimating $P(Y < X)$ in a Pareto distribution with a known scale parameter, including point and interval estimation is considered and also a test of hypothesis.

### Inference on $P(Y < X)$

Let X and Y be independent random variables from Pareto distributions with parameters $(\alpha_x, \beta)$ and $(\alpha_y, \beta)$ respectively.

Then from formula 3.381(4) in Gradshteyn and Ryzhik (1965), the following fact is obtained.

Fact 1:

$$R \equiv P(Y < X) = 1 - \frac{\alpha_x}{\alpha_x + \alpha_y} = \frac{\rho}{1+\rho}$$

is a monotone function of $\rho$, where $\rho \equiv \frac{\alpha_y}{\alpha_x}$.

Proof:

$$R = P(Y < X)$$
$$= 1 - \iint_{0<y<x<\infty} f_X(x;\alpha_x,\beta)\, f_Y(y;\alpha_y,\beta)\,dxdy$$

where $f_X$ is the Pareto density with parameters $(\alpha_x, \beta)$ and $f_Y$ is the Pareto distribution with parameters $(\alpha_y, \beta)$. By formula 3.381(4) in Gradshteyn and Ryzhik (1965), one can integrate and obtain the following.

$$R = P(Y < X) = 1 - \alpha_x \cdot B(1, \alpha_x + \alpha_y),$$

where, $B(a,b)$ is a beta function. Using $B(a,b) = \Gamma(a)\Gamma(b)/\Gamma(a+b), a>0, b>0,$ the above result is obtained.

Because $R$ is a monotone function of $\rho$, inference on $\rho$ is equivalent to inference on $R$. Attention is confined to the parameter $\rho$ (see McCool, 1991). Assume $X_1, X_2, ..., X_m$ and $Y_1, Y_2, ..., Y_n$ are independent random samples from $f_X(x;\alpha_x,\beta_0)$ and $f_Y(y;\alpha_y,\beta_0)$, respectively, where $\beta_0$ is known. From Johnson *et al* (1995), MLE's of $\alpha_x$ and $\alpha_y$ are

$$\hat{\alpha}_x = \frac{m}{\sum_{i=1}^{m} \ln(1 + X_i/\beta_0)}$$

and

$$\hat{\alpha}_y = \frac{n}{\sum_{i=1}^{n} \ln(1 + Y_i/\beta_0)}.$$

The following results in Fact 2 are well-known.

Fact 2: (a) Assume $X_1, X_2, ..., X_m$ be a random sample from a Pareto distribution with parameters $(\alpha_x, \beta_0)$. Then $\sum_{i=1}^{m} \ln(1 + X_i/\beta_0)$ follows a gamma distribution with a shape parameter $m$ and a scale parameter $1/\alpha_x$. (b) If a random variable $X$ follows a gamma distribution with shape $\alpha$ and scale $\beta$ then

$$E(1/X^k) = \frac{\Gamma(\alpha - k)}{\Gamma(\alpha)} \cdot \frac{1}{\beta^k} \text{ if } \alpha > k.$$

From the definition of $\hat{\rho} \equiv \frac{\hat{\sigma}_y}{\hat{\sigma}_x}$, the MLE of $\rho$

$$\text{is } \hat{\rho} = \frac{\hat{\alpha}_y}{\hat{\alpha}_x} = \frac{n}{m} \cdot \frac{\sum_{i=1}^{m} \ln(1 + X_i/\beta_0)}{\sum_{i=1}^{n} \ln(1 + Y_i/\beta_0)}.$$

From Fact 2(a) and (b), one can obtain the following fact.

Fact 3:

$$E(\hat{\rho}) = \frac{n}{n-1}\rho$$

and

$$Var(\hat{\rho}) = \frac{n^2(m+n-1)}{m(n-1)^2(n-2)}\rho^2, n > 2.$$

From Johnson *et al* (1995),

$$\tilde{\alpha}_x = \frac{m-1}{\sum_{i=1}^{m} \ln(1 + X_i/\beta_0)},$$

and

$$\tilde{\alpha}_y = \frac{n-1}{\sum_{i=1}^{n} \ln(1+Y_i/\beta_0)}$$

are UMVUE of $\alpha_x$ and $\alpha_y$, respectively.

Define $\tilde{\rho} = \frac{\tilde{\sigma}_y}{\tilde{\sigma}_x} = \frac{n-1}{m-1} \cdot \frac{\sum_{i=1}^{m} \ln(1+X_i/\beta_0)}{\sum_{i=1}^{n} \ln(1+Y_i/\beta_0)}$.

Then one can obtain the following expectation and variance.

$$E(\tilde{\rho}) = \frac{m}{m-1}$$

and

$$Var(\tilde{\rho}) = \frac{m}{(m-1)^2}\rho^2.$$

Therefore, it is obtained:

Fact 4: $MSE(\hat{\rho}) < MSE(\tilde{\rho})$.

To consider a confidence interval for $\rho$, the following random variables are defined. Let

$$Z \equiv \sum_{i=1}^{m} \ln(1+X_i/\beta_0),$$
$$W \equiv \sum_{i=1}^{n} \ln(1+Y_i/\beta_0)$$

and $U \equiv Z/W$.

By formula 3.381(4) in Gradshteyn and Ryzhik (1965) and the quotient pdf of two independent random variables, the pdf of $U$ is obtained as follows.

$$f_U(u) = \frac{u^{m-1}}{B(m,n)\rho^m}(1+\frac{u}{\rho})^{-m-n}, u>0,$$

where $B(m,n)$ is the Beta function. From the density of $U = Z/W$, one can easily find the distribution of $B \equiv \frac{U}{\rho+U}$.

Fact 5: Let $B \equiv \frac{U}{\rho+U}$. Then, $B$ follows a beta distribution with parameters *m* and *n*. Based on the pivot quantity $B$, a confidence interval of $\rho$ is considered. From the beta distribution function, for a given $0<\alpha<1$, there exists $0<b_\alpha<1$ such that

$$\alpha = \int_0^{b_\alpha} \frac{1}{B(m,n)} x^{m-1}(1-x)^{n-1}dx.$$

Here, for a given $0<\alpha<1$, $b_\alpha$ can be easily evaluated by inverse function of the beta distribution using statistical software. Hence, a $(1-\alpha)100\%$ confidence interval of $\rho$ can be obtained as

$$\left(\frac{m}{n}\cdot\frac{1-b_{1-\alpha/2}}{b_{1-\alpha/2}}\cdot\hat{\rho}, \frac{m}{n}\cdot\frac{1-b_{\alpha/2}}{b_{\alpha/2}}\cdot\hat{\rho}\right)$$

and from the result of Fact 3, its expected length is

$$E(L) = \frac{m}{n-1}\left(\frac{1}{b_{\alpha/2}} - \frac{1}{b_{1-\alpha/2}}\right)\rho.$$

Next, the null hypothesis is tested

$H_0 : \alpha_x = \alpha_y$ against $H_1 : \alpha_x \neq \alpha_y$. Let
$\Theta = \{(\alpha_x, \alpha_y) \mid \alpha_x > 0, \alpha_y > 0\}$, and
$\theta = (\alpha_x, \alpha_y)$.

Then the joint probability density function of $X_1, X_2, ..., X_m, Y_1, Y_2, ..., Y_n$ is

$$L(\theta) = f_\theta(x,y) = \frac{\sigma_x^m \sigma_y^n}{\beta_0^{m+n}} \prod_{i=1}^{m} (1 + x_i/\beta_0)^{-\alpha_x - 1}$$

$$\prod_{i=1}^{n} (1 + y_i/\beta_0)^{-\alpha_y - 1}.$$

Differentiating with respect to $\alpha_x$ and $\alpha_y$, the MLE's are obtained as follows.

$$\hat{\alpha}_x = \frac{m}{\sum_{i=1}^{m} \ln(1 + x_i/\beta_0)}$$

and

$$\hat{\alpha}_y = \frac{n}{\sum_{i=1}^{n} \ln(1 + y_i/\beta_0)}.$$

If $\alpha_x = \alpha_y = \alpha$, then the MLE of $\alpha$ is

$$\hat{\alpha} = \frac{m+n}{\sum_{i=1}^{m} \ln(1 + x_i/\beta_0) + \sum_{i=1}^{n} \ln(1 + y_i/\beta_0)}.$$

From the definition of likelihood ratio test, the likelihood ratio test function is given by

$$\Lambda(x,y) = \left(\frac{m+n}{m}\right)^m \left(\frac{m+n}{n}\right)^n \times \frac{1}{(1+1/U)^m} \frac{1}{(1+U)^n},$$

where

$$U = \frac{\sum_{i=1}^{m} \ln(1 + X_i/\beta_0)}{\sum_{i=1}^{n} \ln(1 + Y_i/\beta_0)}.$$

Therefore, $\Lambda(x,y) < c$ is equivalent to $U < c_1$ or $U > c_2$. Under $H_0 : \alpha_x = \alpha_y$, i.e., $\rho = 1$, from Fact 5, the statistic

$$B_0 = \frac{U}{\rho + U} = \frac{U}{1+U}$$

follows a beta distribution with $m$ and $n$. Because $B_0$ is a monotone increasing function of $U$, so $U < c_1$ or $U > c_2$ is equivalent to $B_0 < b_1$ or $B_0 > b_2$. $b_1$ and $b_2$ can be obtained by inverse function of a beta distribution and using a statistical software.

## References


Ali, M. Masoom & Woo, J. (2005a). Inference on reliability P(Y<X) in a p-dimensional rayleigh distribution. *Mathematical and Computer Modelling - An International Journal*, *42*, 367-373.

Ali, M. Masoom & Woo, J. (2005b). Inference on reliability P(Y<X) in the Levy distribution. *Mathematical and Computer Modelling – An International Journal*, *41*, 965-971.

Abramowitz, M. & Stegun, I. A. (1970). *Handbook of Mathematical Functions*. Dover Publications: New York.

Arnold, B. C. & Press, S. (1983). Bayesian inference for Pareto population. *Journal of Econometrics*, *21*, 287-306.

Baklizi, A. (2003). Confidence intervals for P(X<Y) in the exponential case with common location parameter. *Journal of Modern Applied Statistical Methods*, *2*(2), 341-349.

Gradshteyn, I. S. & Ryzhik, I. M. (1965). *Table of integrals, series and products*. Academic Press: New York.

Johnson, N. L., Kotz, S., & Balakrishnan, N. (1995). *Continuous distributions*. Houghton Mifflin Co.: Boston.

McCool, J. I. (1991). Inference on P(X < Y) in the Weibull case. *Communication in Statistics, Simulations. 20*(1), 129-148.

Moothathu, T. S. K. (1984). Characterizations of power function and Pareto distribution in terms of their Lorenz curves. *Research Reviews*, *3*, 118-123.

Obenhettinger, F. (1974). *Tables of mellin transforms*. Springer-Verlag: New York.

Journal of Modern Applied Statistical Methods
November, 2005, Vol. 4, No. 2, 587-590

1538 – 9472/02/$95.00

# Training Statisticians To Be Alert To The Dangers Of Misapplying Statistical Methods

Vance W. Berger
Biometry Research Group
National Cancer Institute

Statisticians are faced with a variety of challenges. Their ability to cope successfully with these challenges depends, in large part, on the quality of their training. It is not the purpose of this article to present a comprehensive training plan that will overhaul the standard curriculum a statistician might follow under current training regimens (i.e., in a degree program). Rather, the objective is to point out important areas that appear to be under-represented in standard curricula and correspondingly overlooked too often in practice. The hope is that these areas might be better integrated into the training of the next generation of statisticians.

Key words: Assumptions; design-based analysis; exact conditional test; limitations; permutation test.

## Introduction

The ability of statisticians to cope successfully with the wide variety of challenges they face depends, in large part, on the quality of their training. Key components of any training program for statisticians include mathematics, probability theory, statistical inference, and computing. Such classical statistics training would put the statistician in a position to offer solutions to a variety of problems, and defend these solutions. Yet "statistics can be used to form highly technical and even technically correct support for statements which are in fact not true" (Vardeman & Morris, 2003, p. 25). Kimball (1957) described a Type III error as the right answer to the wrong question; earlier Huff (1954) described this phenomenon as a semi-attached figure. It may be overly harsh to use so broad a brush to describe each right answer to a wrong question as an error. Optimal solutions for contrived problems that bear some resemblance to the true problems may also serve as appropriate, if not ideal, solutions for the true problem. On the other hand, an optimal solution to the surrogate problem may not be even a minimally acceptable solution to the true problem.

Vance W. Berger is Mathematical Statistician at the National Cancer Institute. E-mail: vb78c@nih.gov.

Few general rules exist to allow a statistician to be certain that the ideal solution to one problem is actually an appropriate solution to another related problem, so often subject matter knowledge must be used to evaluate a proposed solution to a given problem.

## Unreasonable Assumptions

Many frequently applied statistical methods, including t-tests, linear regression, the analysis of variance (ANOVA), the analysis of covariance (ANCOVA), multivariate ANOVA (MANOVA), and the chi-square test, are based on random sampling and/or normality. In practice, these methods are often used even when neither of these conditions holds. It is also common for methods based on compound symmetry of the variance/covariance matrix, interval scaling of the data, proportional odds or hazards, common variances, or additivity to be used when these conditions do not hold. Statisticians must be concerned with such issues as 1) the evidence for or against each of these conditions holding in a given application and 2) the performance of specific analyses when some or all of these conditions fail to hold. Regarding the first issue, we note the impossibility of demonstrating that certain of these conditions hold in practice.

For example, although a statement such as 'the data are normally distributed' may appear innocuous, this statement simultaneously rules out

every distribution that is not Gaussian, including any distribution with finite support. Also, given the mean and variance, this statement specifies a fixed positive probability of a data point falling in any interval, no matter how far from the largest or smallest observations. As such, this seemingly simple statement actually represents an uncountable number of sub-statements, many of which could not possibly be true. The question is not so much whether the statement is true as it is how well would a procedure derived with the assumption perform without it. This raises the question of what exactly is the true question, when all the assumptions have been stripped away.

If a p-value is required for a between-group comparison, then the true question is 'How likely would it be, if there were no treatment effect, to obtain results as extreme as or more extreme than those which were found'? The answer to this question is a probability, and the relevant probability space is defined based on the observed outcome and all other outcomes that could have occurred given the study design. With random sampling from a normal distribution, the probability space would be based on repeated sampling from a normal distribution. Perhaps a t-test would be used, because it is the optimal solution to the problem of comparing the means of normal populations with equal but unknown variances. But, how well does the t-test perform as an answer for the original question?

To answer this question, the correct answer to the original question must be defined. If there is random allocation but not random sampling, then the platinum standard is an exact design-based permutation test (Tukey, 1993). The frequent assurances that standard statistical methods are robust to violations of their assumptions tend to be based on studies of performance when one assumption at a time is violated. In reality, if an analysis requires assumptions to be valid, then it is vulnerable to the possibility that two of its assumptions may be violated simultaneously. In this case, robustness may be lost (Hunter & May, 1993).

In some cases it may not be possible or feasible to compute an exact p-value. But if the exact p-value is available, as it often is, then the numerical difference between it and the approximate p-value is a better measure of robustness than the usual checks that are made of assumptions. Using this metric, Berger (2000) presented a real data set (specifically, sotalol for reinfarctions) whose assumptions appeared to have been met, yet the exact Smirnov test p-values were 0.0485 (two-sided) and 0.0258 (one-sided), and the approximate p-values were 0.9910 and 0.6823, respectively. This discrepancy can be attributed to the poor approximation of the approximate Smirnov reference distribution to the exact one. That is, the value of the test statistic remains the same whether the exact or approximate test is being used, but the p-value it produces fluctuates wildly as the reference distribution to which it is compared varies.

This is hardly an isolated example, nor is the phenomenon specific to the Smirnov test. Little (1989) presented another real data set, specifically a 2×2 table with cell counts {(170,2);(162,9)}. Each expected cell count is at least 5, so the usual check of the chi-square assumption would be passed, and the chi-square test would tend to be used in practice. Yet at the one-sided 0.025 alpha level the chi-square test would find significance (p=0.0162). and would not even be close to the border, although Fisher's exact test would not reach statistical significance (p=0.0299). Three more examples follow. Using the exact Wilcoxon test, Williams, et al. (2000) demonstrated that compared to routine appointments, open access reduces secondary care costs for inflammatory bowel disease.

Barber and Thompson (2000) unwittingly demonstrated that for this data set, either the normality assumption was sufficiently flawed or the difference in means was sufficiently accompanied by shifts in shape and/or scale that the t-test failed to detect this true difference. Likewise, in a study of the effect of neuromuscular training, Hewett, et al., (1999) used the chi-square test to analyze knee injuries in female athletes. Clancy (2000) commented:

> Because the observed and expected number of knee injuries was less than five in at least one cell, an approximate method is inappropriate. An appropriate method in this instance would have been a Fisher's exact test. Incidentally, use of this exact method demonstrated no statistical significance …, suggesting that the extreme variability present in the

> small sample resulted in an incorrect finding when an approximate method was used. This provides all sports medicine researchers with a potent example of why appropriate statistical analysis is extremely important. (p. 615)

Chaudry, et al. (2002) found p-values of 0.004, 0.016, 0.006, 0.001, and <0.001, using t-tests, for five measures (interest, importance, relevance, validity, believability) of readers' perceptions of papers with and without declaration of competing interests. Jacobs (2003) pointed out that the t-test was applied inappropriately, and, using an exact test, found three of these p-values to be non-significant (interest, p=0.054; importance, p=0.21; relevance, p=0.054). Clearly, assumption-based tests are at times used when they should not be. Bross (1990) stated,

> [T]he user of a statistical method has the responsibility for dealing with the *scientific* question: Are the assumptions valid? In particular, when human health and safety might be jeopardized ..., a statistician has a direct responsibility to protect the public health and safety by following fail-safe principles in dealing with any assumptions. (p. 1216)

Some assumptions are more realistic than others, but if they were known to be true, then they would not be assumptions. As such, one could argue that all things being equal, it is best not to rely on assumptions unless there is a good reason to.

In some cases, there are good statistical methods that require no assumptions at all. For example, design-based between-group permutation tests of the null hypothesis of no difference require no assumptions in randomized clinical trials (Berger, 2000). In other cases, progress can be measured by a reduction, but not elimination, of assumptions. Weerahandi and Berger (1999), for example, derived analyses of growth curves that retain the normality assumption but dropped other assumptions. The use of assumption-minimizing methods, along with the proper respect for uncertainty regarding any assumptions that are made, might be regarded as part and parcel of good statistical practice.

Biased Sampling

Without a reason to suspect systematic bias in the sampling procedure, information about the sample would be used, without adjustment, to draw inferences about the population. This would be optimal in the case of unbiased (perhaps random) sampling. Although it is uncommon for a clinical trial to employ random sampling from the target population, this approach is still used in practice, because the sample is still thought to represent the target population from which it was drawn. Whether or not this is true varies with the situation, but there are cases in which the sampling is biased in a known way. Many randomized clinical trials utilize what is called an open-label run-in phase prior to randomization.

Such a run-in phase is characterized by each patient being exposed to the same treatment. On the basis of their response during this run-in phase, patients are selected for or excluded from the subsequent randomization. Generally, good or bad responders are excluded as the run-in phase used placebo or the active treatment, respectively. But, the treatment used in the run-in phase is then used again as one of the treatments to which patients may be randomized. The effect is over-representation of either active responders or of control non-responders (or, sometimes, both). The advantage for the active treatment group can greatly exaggerate the estimated magnitude of treatment effect (Berger, Rezvani, & Makarewicz, 2003). An optimal analysis should provide a good answer to the question of whether or not treatment A is more effective than treatment B in the sample. But with run-in selection, this optimal answer represents an intentionally distorted answer to the question of whether or not treatment A is more effective than treatment B in the target population.

## Conclusion

It is hoped that the next generation of statistical researchers will work towards deriving better solutions to the important practical questions that need answering. Often, this will involve deriving more powerful assumption-minimizing analyses. We also hope that the next generation of statistical practitioners will appreciate and use these maximally robust procedures more comprehensively. A good step for aspiring statisticians to take now, to help become part of

the solution later, would be to take classes in non-parametric analyses and robust methods, and to develop an interest in the nature of experiments (including limitations) and the way that data sets are generated. It is also useful for one to recognize what it is that (s)he does not know. All too often it is heard that data are used to prove or conclusively demonstrate a hypothesis, when in fact the inference from data analysis is inductive, and not deductive, so proof is not attainable. If, e.g., assumptions were used in an analysis, then the appearance of a treatment effect could be 1) a real treatment effect; 2) a Type I error; or 3) an artifact due to the assumption not being true. A low p-value allows one to probabilistically rule out the second of these explanations, but not the third. Even if the analysis did not explicitly rely on any assumptions, there is still the implicit assumption that an apparent treatment effect cannot be attributed exclusively to a bias. Selection bias, e.g., can create the appearance of a treatment effect where in fact none exists (Berger, 2005).

Even if every known bias can be ruled out, it is still possible that some other bias exists but is yet to be discovered. Hence, there may be any number of explanations for a given observation (such as a data pattern apparently indicative of a treatment effect), and introspection may help anticipate problems not yet identified, and may allow statisticians to perform analyses and design studies that not only gain acceptance in the present, but also stand the test of time in the future (Berger & Matthews, 2005).

## References


Barber, J. A. & Thompson, S. G. (2000). Would have been better to use t-test than Mann-Whitney U-test. *British Medical Journal, 320,* 7251, 1730.

Berger, V. W. (2000). Pros and cons of permutation tests. *Statistics in Medicine, 19,* 1319-1328.

Berger, V. W. (2005). Selection bias and covariate imbalances in randomized clinical trials. Chichester: John Wiley & Sons,

Berger, V. W. & Matthews, J. R. (2005). Conducting today's trials by tomorrow's standards. *Pharmaceutical Statistics, 4*, 155-159.

Berger, V. W., Rezvani, A., & Makarewicz, V. (2003). Direct effect on validity of response run-in selection in clinical trials. *Controlled Clinical Trials, 24*, 2, 156-166.

Bross, I. D. (1990). How to eradicate fraudulent statistical methods: Statisticians must do science, *Biometrics, 46*, 1213-1225.

Chaudhry, S., Schroter, S., Smith, R., & Morris, J. (2002). Does declaration of competing interests affect readers' perceptions? A randomized trial, *British Medical Journal*, *325*, 1391-1392.

Clancy W. G. (2000). Letter to the editor. *American Journal of Sports Medicine*, *28*, 4, 615.

Hewett, T. E., Lindenfeld, T. N., Riccobene, J. V., & Noyes, F. R. (1999). The Effect of Neuromuscular Training on the incidence of knee injury in female athletes. *The American Journal of Sports Medicine, 27*, 6, 699-706.

Huff, D (1954). *How to lie with statistics.* New York: W. W. Norton & Company.

Hunter, M. A. & May, R. B. (1993). Some myths concerning parametric and nonparametric tests. *Canadian Psychology*, *34*, 384-389.

Jacobs, A. (2003). Clarification needed about possible bias and statistical testing. *British Medical Journal USA*, *3*, 93.

Kimball, A. W. (1957). Errors of the third kind in statistical consulting. *Journal of the American Statistical Association,52,* 133-142.

Little, R. J. A. (1989). Testing the equality of two independent binomial proportions. *The American Statistician, 43*, 4, 283-288.

Tukey J. W. (1993). Tightening the clinical trial. *Controlled Clinical Trials, 14*, 266-285.

Vardeman, S. B. & Morris, M. D. (2003). Statistics and ethics: Some advice for young statisticians. *The American Statistician, 57*(1), 21-26.

Weerahandi, S. & Berger, V. W. (1999). Exact inference for growth curves with intraclass correlation structure. *Biometrics, 55*(3), 921-924.

Williams, J. G., Cheung, W. Y., Russell, I. T., Cohen, D. R., Longo, M., and Lervy, B. (2000). Open access follow-up for inflammatory bowel disease: Pragmatic randomized trial and cost effectiveness study. *British Medical Journal, 320*, 544-548.

Journal of Modern Applied Statistical Methods
November, 2005, Vol. 4, No.2, 591-597

1538 – 9472/05/$95.00

# Power of the *t* Test for Normal and Mixed Normal Distributions

Marilyn S. Thompson Samuel B. Green Yi-hsin Chen Shawn Stockford Wen-juo Lo
Division of Psychology in Education
Arizona State University

Previous research suggests that the power of the independent-samples *t* test decreases when population distributions are mixed normal rather than normal, and that robust methods have superior power under these conditions. However, under some conditions, the power for the independent-samples *t* test can be greater when the population distributions for the independent groups are mixed normal rather than normal. The implications of these results are discussed.

Key words: *t* test, mixed normal, power

## Introduction

The accepted belief in modern statistical practice is that the assumption of normality for parametric tests, such as the independent-samples t test and the analysis-of-variance F test, seldom, if ever, holds in practice. In psychology and education, Micceri (1989) offered empirical support for this conclusion. He examined over 400 large-sample data sets that included achievement and psychometric measures and found that they had a variety of shapes (e.g., skewed) and generally could not be described as normal.

For a number of years, violation of the normality assumption was not seen as a serious problem in that a number of studies showed that nonnormality, in and of itself, had a minimal effect on Type I error rate unless sample size is quite small (e.g., Boneau, 1960; Glass, Peckham, & Sanders, 1972; Ramsey, 1980; Rogan & Keselman, 1977).

Marilyn Thompson (m.thompson@asu.edu) and Samuel Green (samgreen@asu.edu) are faculty members in the Measurement, Statistics, and Methodological Studies program in the Division of Psychology in Education at Arizona State University. Yi-hsin Chen, Shawn Stockford, and Wen-juo Lo are graduate students in this program.

More recently, researchers have demonstrated that violation of the normality assumption may, however, have a deleterious effect on the power of parametric tests (e.g., MacDonald, 1999; Lix & Keselman, 1998; Wilcox, 1995). Based on these findings and others concerning violation of the homogeneity of variance assumption, Keselman, Wilcox, and Lix (2003) suggested that the application of standard parametric methods should be greatly restricted, and robust methods requiring minimal distributional assumptions should be used in their place. More specifically, they argued that robust methods, such as those using trimmed means and bootstrapping, are superior in terms of Type I and II error rates across a wide number of conditions encountered in practice.

The mixed normal distribution has been used extensively to illustrate the detrimental effect of nonnormality and specifically outliers on parametric tests and, most frequently, on the independent-samples t test (e.g., MacDonald, 1999; Wilcox, 1997, 2001). Based on these presentations, the independent-samples t test shows a dramatic decrease in power when the population distributions for the two independent groups are mixed normal rather than normal. A small-scale simulation may be used to illustrate the decrease in power found in these studies.

Consider the power of the independent-samples t test with 12 observations in each group under normal and mixed normal conditions. For the normal condition, data are generated from normal distributions with means

of 0 and 3 for first and second groups, respectively. The population variances are held constant across groups at 1. Based on 4000 replications, the empirically determined power is 1.00.

For the mixed normal condition, normal data are generated for each group from primary and secondary subpopulations with probabilities of .80 and .20, respectively. The means of the normal distributions for the primary and secondary subpopulations are identical to those under the normal condition: means of 0 for the first group and means of 3 for the second group. As in the normal condition, the variances for the primary distributions are set to 1 in both groups; however, the variances for the secondary distributions are set to 400 in both groups to simulate outliers. Based on 4000 replications, the empirical power is .21 under the mixed normal condition, much lower than the 1.00 found under the normal condition.

The explanation for these results and ones like them is that the standard error of the difference in means is much larger for the mixed normal distribution than for the normal distribution (e.g., Wilcox, 2001). For this example, the within-group variances increased from 1.00 for the normal condition to 80.80 for the mixed normal condition [i.e., combined across the primary and secondary distributions: .80 (1) + .20 (400) = 80.80], as a function of introducing the secondary distribution with a much larger variance (i.e., 400). Because the within-group variances increased for the mixed normal condition, the standard error of the difference in means increased, and the power decreased.

In the current Monte Carlo study, unexpected results were found when investigating the comparative power of the independent-samples t test under normal and mixed normal conditions. Conditions were included that were similar to those in previous research: the variances for the normal distributions were set equal to the variances of the primary distributions of the mixed normal distributions. In these conditions, the combined variances for the mixed normal distributions were greater due to the larger variances of the secondary distributions. However, different from previous studies, control conditions were included in which normal distributions had variances set equal to the combined variances in the mixed normal conditions. Presumably, the power of the independent-samples t test would be equivalent for the normal and mixed normal conditions if the population variances for the two conditions were equal and, thus, the standard errors of the difference in means were equal. However, the results of this study demonstrate the counterintuitive result that the power may be greater under the mixed normal condition.

## Methodology

Data were generated using the normal pseudorandom number generator available in the IML procedure in SAS 8.2. Fifty-four conditions were created by manipulating four factors: the form of the population distribution, variances of these distributions, sample size, and mean differences.

Form of distributions. Data were generated for two independent groups from populations with normal or mixed normal distributions.

Variance. When the distributions were normal, the variances were equal to 1 for both groups or 80.8 for both groups. When the distributions were mixed normal, the variances for both groups were 1 for the normal distribution with a probability of .80 and 400 for the normal distribution with a probability of .20; therefore, the mixed normal distributions had a combined variance of 80.8.

Sample size. The total sample size (N) consisted of 24, 48, or 96 cases, with an equal number of cases in each of the two independent groups.

Mean differences. To evaluate the Type I error rates of the test statistics, data were generated such that the differences in population means were equal to zero. To assess power, data were generated so that the population mean for one group was zero, and the population mean for the second group was one of five values: 0.5, 1.0, 1.5, 3.0, or 4.5. For mixed normal distributions, the means of the primary and secondary distributions for any one group were always the same.

Data Analysis

Two-tailed independent-samples t tests were conducted using the ttest procedure within SAS 8.2 and evaluated at the .05 level. Four-thousand replications were generated for each of the 54 conditions. Empirical alphas were computed for the conditions in which the means were equivalent. Empirical powers were calculated as proportions of rejections of a false null hypothesis in the correct direction for conditions in which the means differed between groups.

In addition, empirical Type III error rates–proportions of rejections of a false null hypothesis in the wrong direction–were computed. However, Type III error rates were excluded from the discussion because they were strongly inversely related to power and were uniformly very low; Type III error rates were less than .01 for 87% of the conditions and never exceeded .02.

## Results

Empirical Alphas

For the six conditions with normal distributions and equal population means, the empirical alphas were very close to .05, ranging from .046 to .054. These results were expected in that all assumptions of the independent-samples t test were met under these conditions. On the other hand, the empirical alphas were somewhat conservative when the distributions were mixed normal, particularly for smaller sample sizes. The alphas were .025, .042, and .048 with Ns of 24, 48, and 96, respectively. Given these results, any power advantage observed under mixed normal conditions cannot be attributed to inflated alphas.

Empirical Powers

Figure 1 shows the power of the t test as a function of the difference in means and sample size for three population distributions: mixed normal with a variance of 80.8, normal with a variance of 80.8, and normal with a variance of 1.0. As expected, the power was greater for conditions with a normal distribution and a variance of 1 than for conditions with a mixed normal distribution and a variance of 80.8. The differential power was substantial across most sample sizes and mean differences.

The more provocative findings were the power comparisons between the mixed normal and the normal distributions when both distributions had within-group variances of 80.8. For these comparisons, the power tended to be greater when distributions were mixed normal, particularly for the smaller sample sizes (N of 24 or 48). This power differential became larger as the difference in means increased. In contrast, the power differential was minimal for the largest sample size (N = 96).

Exploration of the Power Differential

The results indicate that the power for an independent-samples t test is greater when samples are drawn from mixed normal distributions rather than normal distributions, given both distributions have comparable variances. To better understand these results, it is useful to examine relevant population and sampling distributions.

In Figure 2, three sets of population distributions with means of 0 and 4.5 (and equal variances) are presented: mixed normal distributions with within-group variances of 80.8; normal distributions with within-group variances of 1.0; and normal distributions with within-group variances of 80.8. Examination of these population distributions suggests that some sample distributions from the mixed normal may be more similar to those from the normal with variances of 1.0 than those from the normal with variances of 80.8, particularly for smaller samples. In these samples from mixed normal distributions, there should be a greater likelihood of rejecting the null hypothesis than in samples drawn from the normal distribution with a variance of 80.8. However, sampling distributions are next examined to gain a deeper insight into the differential power of t test under normal and mixed normal conditions.

Table 1 shows the sampling distributions of the *t* statistic, the difference in means, and the pooled within-group variance for 30,000 samples drawn from normal and mixed normal distributions with a difference in means equal to 4.5, within-group variances of 80.8, and *N*s of 24 (with equal sample sizes).

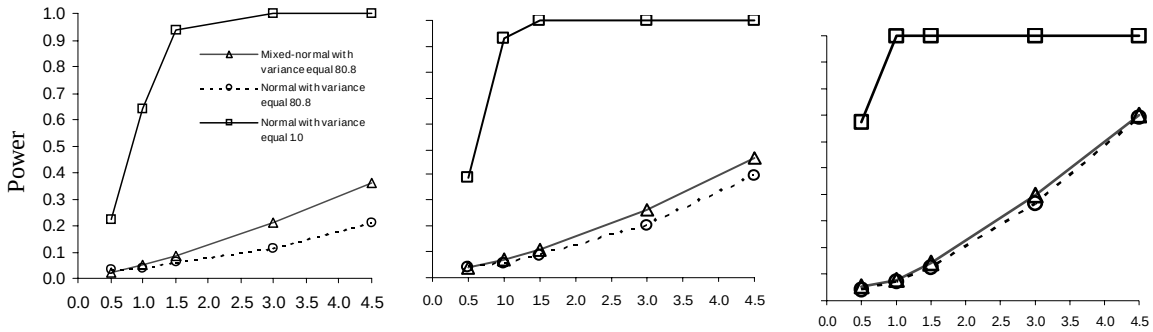



Figure 1. Power of the *t* test as a function of the difference in means and sample size for three population distributions: mixed normal with $\sigma^2 = 80.8$, normal with $\sigma^2 = 80.8$, and normal with $\sigma^2 = 1.0$. From left to right, $N = 24$; $N = 48$; and $N = 96$ (with equal cases across group).

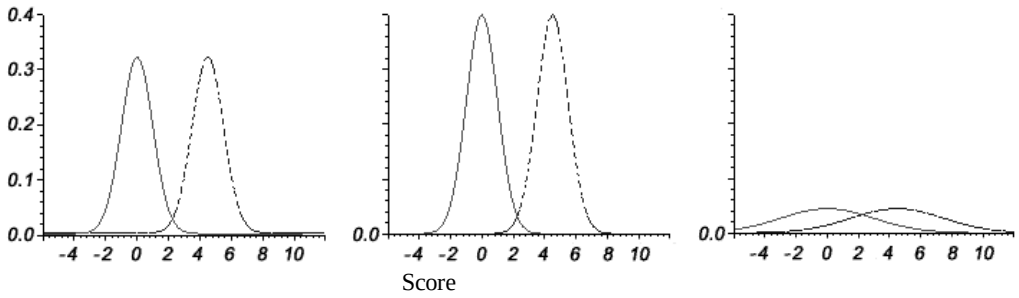



Figure 2. Group population distributions for three conditions where variances are equal across groups and the difference in means is 4.5. From left to right, mixed normal distributions with $\sigma^2 = 80.8$; normal distributions with $\sigma^2 = 1.0$; and normal distributions with$\sigma^2 = 80.8$.

As shown in the first row, the t distribution for the mixed normal condition was quite skewed and thick tailed (i.e., skewness = 2.37 and kurtosis = 11.23) compared to the t distribution for the normal condition (i.e., skewness = 0.19 and kurtosis = 0.37). Given |tcritical(22)| = 2.07, the empirical power of the t test was .34 for the mixed normal distribution, which was considerably larger than the empirical power of .21 for the normal condition.

The t statistic is a function of three quantities: the difference in means, the pooled variance, and sample size–and the latter was held constant. As shown in the second row of Table 1, the sampling distributions for the difference in means were symmetric and quite similar, except that the sampling distribution for the mixed normal was somewhat kurtotic (kurtosis = .45). As presented in the third row of Table 1, the sampling distributions for the pooled variance were very different for the two types of distributions. Although the means of the variances were nearly equal (normal: 80.76; mixed normal: 80.65), the variance of the pooled variance was 6.56 times larger for the mixed normal than for the normal condition. Further, the sampling distribution of the pooled variances was more skewed and had thicker tails for the mixed normal condition compared to the normal

condition (normal condition: skewness = 0.59 and kurtosis = 0.52; mixed normal: skewness = 1.38 and kurtosis = 2.90). Most importantly, a much larger proportion of replications had small variances for the mixed normal distribution than for the normal distribution. For example, approximately 11% of the pooled variances were less than 16 for the mixed normal condition, while none were less than 16 for the normal condition.

A greater percentage of small pooled variances are obtained with the mixed normal in comparison with the normal distribution in that the secondary distribution (with the large population variance of 400) for the mixed normal may have no or minimal effect on the pooled variance in some samples.

For example, some samples may contain no scores from the secondary distribution, and others may contain one score from the secondary distribution, but not an extreme score. The smaller pooled variances produce larger t values and, thus, greater power for the mixed normal distribution in comparison with the normal distribution with the equal population variances.

## Conclusion

The results do not contradict the primary conclusions of previous research on the mixed normal distribution and the independent-samples t test. To the extent that the population distributions have outliers, the power of the t test is diminished. In the context of the mixed normal distribution, the power of the independent-samples t test decreases dramatically as the probability of a secondary distribution with a large variance increases from .00 to .20. In the presence of extreme scores, robust methods such as trimmed means become advantageous.

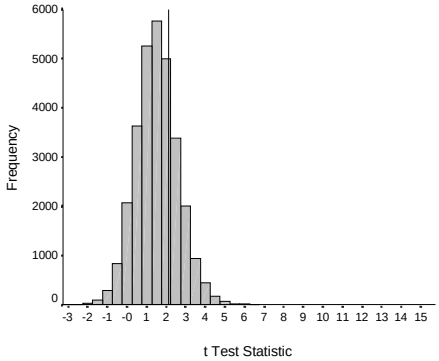
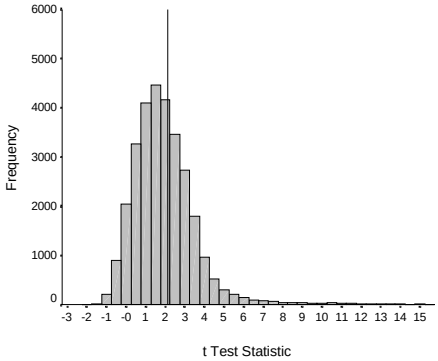
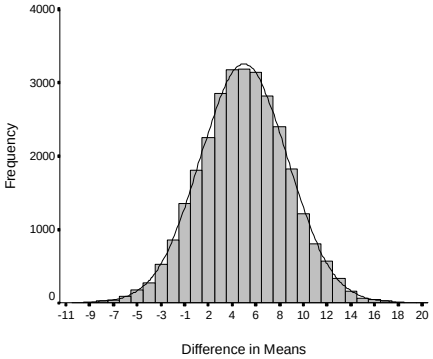
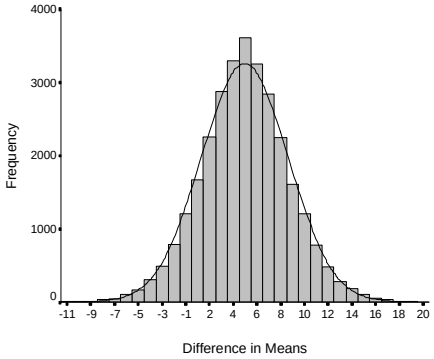
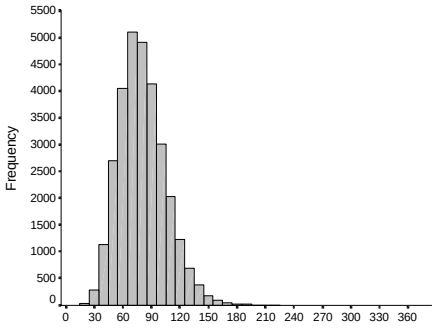
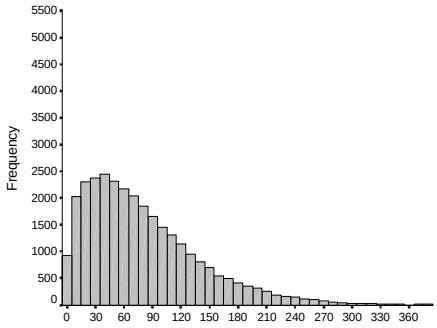
The results, however, contradict the hypothesis that the power of the test for normal and mixed normal conditions would be equal if the within-group variances were held constant or, comparably, if the effect sizes (difference in means divided by the within-group standard deviation) were held constant. Under these conditions, the power, in fact, was greater for the mixed normal distribution in that some samples produce relatively small pooled variance as a function of having few, if any, outliers drawn from the secondary distributions. The superior power was achieved despite the conservative Type I error rate for the mixed normal.

These results support a number of conceptual points. First, care should be used in discussing the diminished power of the independent-samples t test when population distributions are mixed normal rather than normal. An accurate statement is that the independent-samples t test has diminished power with a mixed normal distribution in comparison with the normal distribution to the extent that the secondary normal distribution has a much larger variance than the primary distribution and the probability of the secondary distribution is relatively large.

Second, although the independent-samples t test is the most powerful method for comparing two means if the assumptions, including normality, are met, variations of this statement may not be true. In particular, it is not true that the independent-samples t test has greater power if the population distributions are normal in comparison with other distributions, holding all other conditions constant. As demonstrated in this study, the independent-samples t test can have greater power when the population distributions are mixed normal rather than normal, given the variances of these two types of distributions are held constant.

Third, these results may be used to speculate about trimming strategies for the independent-samples t test. Some samples may include no outliers, even though the population distributions have outliers. For these samples, robust methods relying on trimming lower the likelihood of rejecting the null hypothesis by reducing the effective sample size without decreasing the pooled variance. Adaptive trimming methods–ones that trim based on the outliers present in the sample data–should produce greater power in these circumstances than those that use a fixed proportion of trimming (e.g., trim 20% from both tails of sample distributions). Future Monte Carlo studies are required to investigate whether adaptive trimming methods under these conditions maintain proper control of Type I error while increasing power.

Table 1. Sampling distributions based on independent samples of equal size ($N$ = 24) drawn from two population distributions that are both either normal or mixed normal with a difference in population means of 4.5 and a common population variance of 80.8

| Sampling distribution | Population distributions: Normal | Population distributions: Mixed normal[c] |
|---|---|---|
| *t* test statistic[a] |  |  |
| Difference in means[b] |  |  |
| Pooled variance |  |  |



[a]The vertical reference line indicates the critical value for rejecting the null hypothesis in the correct direction: $t(22)=2.07$.
[b]A normal curve is superimposed on the plots of the difference in means.
[c]The abscissas for the distributions based on the mixed normal were not extended to include all possible values of statistics if the frequencies for intervals including these values were sufficiently small (< .04 % of samples) that they could not be observed on the graphs. The most extreme values not shown were for the pooled variance, with six values being greater than 500.

References

Boneau, C. A. (1962). A comparison of the power of the U and t-tests. Psychological Review, 69, 246-256.

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

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.

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

MacDonald, P. (1999). Power, Type I, and Type III error rates of parametric and nonparametric statistical tests. The Journal of Experimental Education, 67, 367-379.

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

Ramsey, P.H. (1980). Exact type I error rates for robustness of Student's T test with unequal variances. Journal of Educational Statistics, 5, 337-349.

Rogan, J. C., & Keselman, H. J. (1977). Is the ANOVA F-test robust to variance heterogeneity when sample sizes are equal?: An investigation via a coefficient of variation. American Educational Research Journal, 14, 493-498.

Wilcox, R.R. (1995). ANOVA: The practical importance of heteroscedastic method, using trimmed means versus means, and designing simulation studies. British Journal of Mathematical and Statistical Psychology, 48, 99-114.

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

Wilcox, R. R. (2001). Fundamentals of modern statistical methods: Substantially improving power and accuracy. New York: Springer-Verlag.

Journal of Modern Applied Statistical Methods
November, 2005, Vol. 4, No. 2, 598-600

1538 – 9472/05/$95.00

# Misconceptions Leading to Choosing the t Test Over the Wilcoxon Mann-Whitney Test for Shift in Location Parameter

Shlomo S. Sawilowsky
Wayne State University

There exist many misconceptions in choosing the t over the Wilcoxon Rank-Sum test when testing for shift. Examples are given in the following three groups: (1) false statement, (2) true premise, but false conclusion, and (3) true statement irrelevant in choosing between the t test and the Wilcoxon Rank Sum test.

Key words: t test, Wilcoxon Rank-Sum test, robustness, power

## Introduction

For treatment effects modeled as a shift in location parameter, the t test can be decidedly nonrobust to departures from population normality unless certain conditions have been met (Sawilowsky & Blair, 1992). When normality is met or nearly met (which occurs rarely), the t test maintains a very small power advantage over the Wilcoxon Rank Sum / Mann-Whitney U test. When normality is violated, the Wilcoxon Rank Sum Test can be three or four times more powerful than the independent samples t test (Blair, 1980; Blair & Higgins, 1980a, 1980b, 1981; Blair, Higgins, & Smitely, 1980; Sawilowsky & Blair, 1992). The power advantages of the nonparametric test actually increases with sample size for the low to mid-level parts of the t test's power spectrum.

Although the power advantage is not as spectacular as with the independent samples case, the Wilcoxon Signed-Ranks test for two dependent samples nevertheless maintains a considerable power advantage over the dependent samples t test for similar conditions (Blair & Higgins, 1985a, 1985b).

Shlomo S. Sawilowsky is Professor, WSU Distinguished Faculty Fellow, and Careful Data Analyst. Email him at shlomo@wayne.edu.

The dates of the Monte Carlo studies cited above are from 1980 – 1992. Promise for these small sample results was available decades prior on the basis of large sample asymptotic theory. This understanding had even penetrated to the level of a *book review* written in 1968! "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%" (Meeter, 1968).

Thus, sane researchers opt to use the Wilcoxon Rank Sum test when testing for shift in location. Overly cautious researchers, with no justification, opt to perform both the t test and the Wilcoxon Rank Sum test, and accept the Wilcoxon only if it rejects and the t doesn't. (This is a misguided practice, as it leads to an increase in experiment-wise Type I errors.) Pedantic researchers, oblivious to the Monte Carlo results of the past 25 years, and asymptotic results for the past half-century, simply ignore the Wilcoxon Rank Sum test in favor of the t test.

In the course of reviewing articles submitted to the sixteen journals that I have provided ad hoc reviews over the past 15 years, I have compiled a list of constantly recycling reasons given for preferring the t test over the Wilcoxon Rank Sum test when testing for shift in location. They are presented below without expansive commentary, in the hopes that they never again resurface.

The misconceptions are categorized in three groups: (1) false statement, (2) true premise, but false conclusion, and (3) true statement irrelevant in choosing between the t test and the Wilcoxon Rank Sum test.

(1) False Statement

- the Wilcoxon is only for use when the data are originally in the form of ranks
- the Wilcoxon's ranking procedure throws away useful information
- the Wilcoxon is only for use in the presence of outliers
- the Wilcoxon should only be used for small samples
- the t is robust with respect to Type I errors
- the t is more powerful
- if a modern procedure should be used, it should be a permutation test, not the Wilcoxon

(2) True Premise, but False Conclusion

- the Wilcoxon is a test of $f_i(x) = g_i(x)$ (true), so even if it does reject and the t doesn't, it is probably due to some difference other than the mean (e.g., scale) (false)
- the Wilcoxon's underlying assumptions are weaker (true), therefore the hypothesis being tested is less interesting (false)
- in terms of central tendency, the Wilcoxon pertains to the median (true), which is less interesting than the mean (false)
- the t is expandable to the k samples case (true), but the Wilcoxon is not (false)
- the t is expandable to the multivariate case (true), but the Wilcoxon is not (false)
- the t is expandable to the factorial case (true), but the Wilcoxon is not (false)

(3) True Statement Irrelevant in Choosing Between the t and Wilcoxon

- the t is a classical test
- results based on the t have been accumulating for almost a century, permitting direct comparison of results over time
- the t on the ranks is equivalent to the Wilcoxon on the original scores
- the hypotheses being tested for the t and Wilcoxon aren't exactly the same
- the t is the Uniformly Most Powerful Unbiased test under normality
- the t is robust with respect to Type II errors for departures from normality
- for very small sample sizes the t can be conducted at $\alpha = .05$ or .01, but the Wilcoxon cannot because there are no critical values
- at relatively small sample sizes, the Wilcoxon test cannot be conducted at exactly the $\alpha = .05$ or .01 levels due to the discrete nature of the sampling distribution
- even its inventor called the Wilcoxon test a "quick and dirty" or "crude" procedure

## References

Blair, R. C. (1980). *A comparison of the power of the two independent means* t *test to that of theWilcoxon's rank-sum teset for samples of various populations.* Unpublished doctoral dissertation, University of South Florida, Tampa, FL.

Blair, R. C., & Higgins, J. J. (1980a). A comparison of the *t* test and the Wilcoxon statistics when samples are drawn from a certain mixed normal distribution. *Evaluation Review, 4*, 645-656.

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

Blair, R. C., & Higgins, J. J. (1981). A note on the asymptotic relative efficiency of the Wilcoxon rank-sum test relative to the independent means t test under mixtures of two normal distributions. *British Journal of Mathematical and Statistical Psychology*, *31*, 125-128.

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

Meeter, D. (1968). Book Reviews, *Journal of the American Statistical Association*, 62, p. 1505)

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

Journal of Modern Applied Statistical Methods
November, 2005, Vol. 4, No. 2, 601-608

1538 – 9472/05/$95.00

# *Early Scholars*

## Sample Size Selection for Pair-Wise Comparisons Using Information Criteria

Xuemei Pan C. Mitchell Dayton
University of Maryland

This article provides results for rates of correct identifications of paired-comparison information criteria and Tukey HSD as functions of the pattern of mean differences and of sample size. Therefore, the tables provided are useful for selecting sample sizes in real world applications.

Key words: PCIC, sample size, power, information criteria

### Introduction

Model-comparison procedures using information-theoretic criteria such as AIC or BIC provide the basis for attractive alternatives to traditional pairwise comparison procedures such as Tukey HSD tests and its many variations. Known as paired-comparisons information criterion, or PCIC, these methods avoid many of the problems associated with conducting a series of correlated significance tests.

In presenting the theoretical background for PCIC, Dayton (1998) reported a small-scale simulation study that provided some evidence concerning the probability of detecting exactly all true pairwise differences among means from several samples. This is referred to as all-pairs power Ramsey (1978) or as the true-model rate by Cribbie and Keselman (2003). Dayton (1998) found that the all-pairs power for PCIC was found to be generally better than that of HSD. In a much more extensive study of PCIC compared with three step-wise multiple comparison procedures (MCPs), Cribbie and Keselman (2003) reported that "when all population means were not equal… {PCIC}… had significantly higher true-model rates than any of the stepwise MCPs." Similarly, Cribbie (2003) reported a simulation study that compared several conventional multiple comparison procedures with PCIC and concluded that PCIC "…had consistently larger true models rates than did familywise error controlling MCPs."

Information is provided in this article concerning the performance of PCIC with respect to rates of correct identifications of patterns of mean differences as a function of sample size and thus, the results are useful for selecting sample sizes for real world applications. These results supplement the very limited simulation results for minimum sample size requirements for selected power levels provided by Dayton (2003).

### Summary of PCIC

For K independent groups, many popular pairwise-comparison procedures compute test statistics for each of the K(K – 1)/2 unique pairs of means and refer these statistics to an appropriate null distribution. Tukey HSD tests, for example, are based on the studentized range statistic for a span of K means. Thus, K (K – 1)/2 hypotheses of the form $\mu_k = \mu_{k'}$ for $k \neq k'$ are tested. Among the problems with procedures such as this as cited by Dayton (1998) are:

(1) Some arbitrary technique is necessary to control the family-wise type I error rate for the set of correlated pairwise tests;

Xuemei Pan is a Ph D candidate. Her research interests include latent class modeling and model comparison procedures. E-mail her at xpan1@umd.edu. C. Mitchell Dayton is Professor and Chair. His research interests include experimental design and latent class modeling. Email him at cdayton@umd.edu.

(2) The issues of homogeneity of variance and differential sample size pose problems for many paired-comparison procedures;
(3) Intransitive decisions (e.g., outcomes suggesting mean 1 = mean 2, mean 2 = mean 3, but mean 1 < mean 3) are the rule rather than the exception with typical paired comparison procedures since they entail a series of discrete, pairwise significance tests;
(4) There exists a large variety of competing procedures that differ in how type I error is controlled and consequently, in power.

Dayton (1998) proposed using information-theoretic model-selection criteria such as AIC (Akaike, 1973) or BIC (Schwarz, 1978) for selecting the most appropriate ordering of subsets of means for purposes of interpretation. By considering patterns of mean differences, rather than pair-wise differences, the PCIC approach avoids many of the objections raised above. Furthermore, the interpretation of results is facilitated by PCIC to a much greater degree than by conventional pair-wise comparison procedures.

For K independent means, there are a total of $2^{K-1}$ patterns of ordered subsets with equal means within subsets. For example, with three groups for which the means are ranked and labeled 1, 2, 3, the $2^2 = 4$ distinct ordered subsets are {123}, {1,23}, {12,3} and {1,2,3}, where a comma is used to separate subsets that are unequal in mean value. The basic approach in PCIC is to compute AIC (or, BIC) for each ordered subset based on appropriate model assumptions. Then, the preferred model for purposes of interpretation is the one that satisfies a min(AIC), or min(BIC), criterion.

Assuming a given model and distributional form for the data (e.g., normal), AIC is computed as $-2\text{Log}_e(L) + 2p$, where p is the number of independent parameters estimated in calculating the likelihood, L, for the observed data. Typically, the additive term, 2p, is viewed as a penalty that reflects the complexity of the model. Similarly, BIC is computed as $-2\text{Log}_e(L) + \text{Log}_e(N)p$ where N is the total sample size. For a model with T subsets of means, p equals T+1 assuming homogeneity of variance for the K groups (see Dayton, 1998; 2003, for discussion of related models without the assumption of homogeneity). For example, for the pattern {1, 2, 34} there are three ordered subsets of means so the value of T is 4. The four parameters that are estimated are the mean of group 1, the mean of group 2, the combined mean of groups 3 and 4 and the pooled variance across the four groups. It should be noted that in computing the likelihood for the data, maximum-likelihood estimates for variances are biased (e.g., use N in the denominator for computing the pooled variance).

AIC does not directly involve the sample size in its computation and, as noted by Bozdogan (1987), lacks certain properties of asymptotic consistency usually associated with increasing sample sizes. Also, since $\text{Log}_e(N)$ is larger than the penalty coefficient, 2 for AIC when N is greater than seven, AIC and BIC may, and often do, result in different orderings of subsets of means with, predictably, simpler models being favored by BIC, although AIC tends to select more complex models (i.e., models with a greater number of subsets of means).

## Methodology

The main focus of this research was to provide some guidance for selecting sample sizes for comparisons based on information criteria. Power is not only a function of effect size and sample size but also varies in terms of the population pattern of mean differences. In addition for AIC, but not BIC or other asymptotically consistent methods, there are theoretical maximum power levels with respect to certain patterns of mean differences.

In theory, probabilities for selecting models with larger numbers of subsets of means than the true model can be calculated for AIC using results provided by Bozdogan (1987). These calculations provide the upper limits on power that AIC can attain regardless of sample size (as noted above, AIC is not asymptotically consistent). Therefore, when using AIC it is theoretically possible to choose an over-parameterized model even as the sample size

approaches infinity. Model selection criteria which have this property are sometimes called dimension inconsistent. For example with 5 groups, the maximum powers, in theory, for true models with 1 to 5 clusters of means are: .504, .596, .707, .843, and 1.000, respectively. Thus, for one or two clusters of means there is no sample size that will yield all-pairs power of 2/3 for AIC with 5 groups.

For determining minimum sample size requirements, four sets of conditions were considered:

(1) Number of independent groups: k=3, 4, 5 and 6.

(2) Effect size, f, using Cohen's (1969) definition with small (.1), medium (.25) and large (.4) levels for the corresponding one-way ANOVA design with equally-spaced population means.

(3) Power: .50, .67 and .80 representing low, medium and large values.

(4) Patterns of population means: A variety of patterns were examined as shown in the sample size tables below.

Programming in the matrix language, Gauss (Aptech Systems, Inc., 2002), was used to determine minimum sample size requirements for AIC, BIC and HSD. Data were generated by using 1,000 pseudo-random, homoscedastic normal samples of equal sizes with sample sizes starting at 10 per group and incremented by five per group at each iteration. Iterations terminated and the sample size recorded when the specified power (.50, .67 or .80) was attained or, if not attained, when a sample size of 1000 per group was reached.

For AIC and BIC, the proportion of cases for which the selection procedure resulted in selection of the correct data-generating model represents the true-model (or, accuracy) rate. For HSD, pairwise q tests were calculated for all pairs of means and a count was made of the number of correct decision in the sense of identifying the correct pattern (e.g., to be counted as correct for the population pattern {1, 2, 3, 4, 5}, all 10 pairwise differences had to be significant at the .05 level). Note that the simulations only involved equal sample sizes with equal population variances.

## Results

Results for minimum sample sizes are shown in Tables 1, 2 and 3 for effect sizes of .10, .25 and .40, respectively. As expected from prior power studies, HSD often requires considerably larger sample sizes to attain specified power levels than do methods based on information criteria. However, there are substantive differences among the methods for specific cases. The following generalities apply:

(A) When all means are different, AIC requires uniformly much smaller sample sizes than either BIC or Tukey HSD for any number of groups. For example, this superiority of AIC is displayed in Figure 1 that shows minimum sample size requirements for AIC, BIC and Tukey HSD with medium effect size, .25, medium power, .67, and all means different. On the other hand, the minimum sample size requirements for BIC and Tukey HSD are essentially equivalent for this case.

(B) As a rule of thumb, AIC requires smaller minimum samples sizes than BIC or Tukey HSD when the number of clusters of homogeneous means is greater than one-half the number of groups. Occasionally this rule fails since AIC cannot, in theory, attain .67 or .8 power, as noted above.

(C) When the number of clusters of homogeneous means is less than one-half the number of groups, BIC tends to perform better than either AIC or Tukey HSD although this advantage tends to vanish when all group means are equal. On the basis of the poor performance of AIC for the null pattern, it was suggested by Dayton(1998) that an omnibus test be conducted as the first step in any analysis and that additional analyses be contingent on attaining significance with the omnibus test. However, a preliminary omnibus test provides no benefit for the BIC strategy.

(D) For three or more clusters of homogeneous means, those patterns with two or more groups clustered in the center yield higher accuracy rates than when the groups are clustered in the tail for all three methods. For

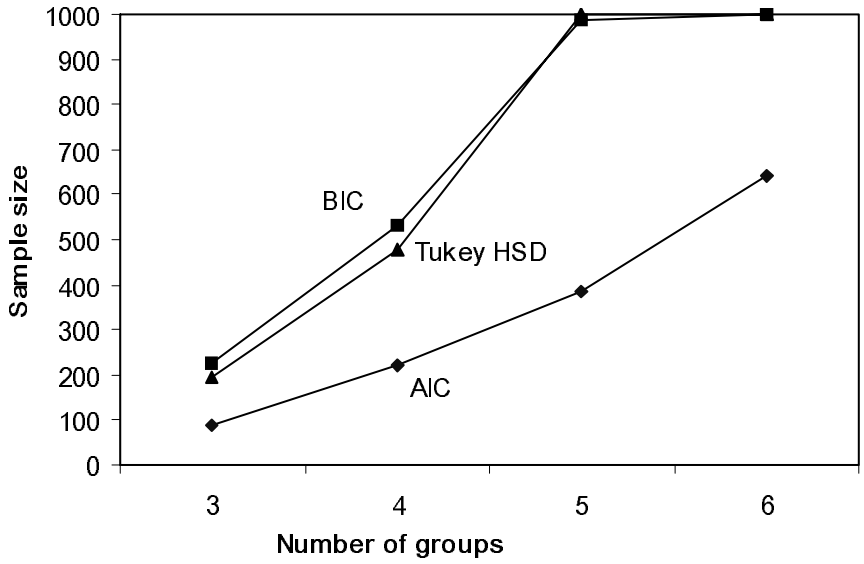



Figure1 All means different Power=. 67

example, with four groups, the pattern {1, 23, 4} has higher accuracy rates than pattern {12, 3, 4} even though both patterns contain three clusters of means. Similarly, for six groups, the five-cluster pattern {1, 2, 34, 5, 6} requires smaller minimum sample size requirements than the five-cluster pattern {12, 3, 4, 5, 6}.

In general, inconsistent performance between the two PCIC methods, AIC and BIC, can be explained by differences in their penalty terms. In general, AIC tends to select more complex models than BIC. Thus, when errors are made, AIC can be viewed as tending to overfit models whereas BIC can be viewed as tending to underfit models.

Table 1. Minimum Sample Size Requirement: Effect Size=0.10

| | AIC | | | BIC | | | Tukey HSD | | |
|---|---|---|---|---|---|---|---|---|---|
| Power | .5 | .67 | .8 | .5 | .67 | .8 | .5 | .67 | .8 |
| Pattern of means | | | | | | | | | |
| Three groups | | | | | | | | | |
| {1,2,3} | 560 | 750 | 985 | M | M | M | M | M | M |
| {12,3} | 100 | 185 | 325 | 225 | 310 | 415 | 345 | 450 | 565 |
| {123} | 10 | 10 | * | 10 | 10 | 10 | 10 | 10 | 10 |
| Four groups | | | | | | | | | |
| {1,2,3,4} | M | M | M | M | M | M | M | M | M |
| {12,3,4} | 615 | 860 | M | M | M | M | M | M | M |
| {1,23,4} | 390 | 550 | 835 | 910 | M | M | M | M | M |
| {123,4} | 110 | 220 | * | 175 | 245 | 325 | 370 | 480 | 580 |
| {1234} | 10 | * | * | 10 | 10 | 10 | 10 | 10 | 10 |
| Five groups | | | | | | | | | |
| {1,2,3,4,5} | M | M | M | M | M | M | M | M | M |
| {12,3,4,5} | M | M | M | M | M | M | M | M | M |
| {12,3,45} | 655 | M | * | M | M | M | M | M | M |
| {1,234,5} | 360 | 595 | * | 665 | 805 | 980 | M | M | M |
| {1234,5} | 105 | * | * | 135 | 205 | 260 | 385 | 495 | 575 |
| {12345} | 10 | * | * | 10 | 10 | 10 | 10 | 10 | 10 |
| Six groups | | | | | | | | | |
| {1,2,3,4,5,6} | M | M | M | M | M | M | M | M | M |
| {12,3,4,5,6} | M | M | M | M | M | M | M | M | M |
| {1,2,34,5,6} | M | M | M | M | M | M | M | M | M |
| {1,2,3,456} | M | M | * | M | M | M | M | M | M |
| {1,2,345,6} | M | M | * | M | M | M | M | M | M |
| {12,34,56} | 515 | * | * | 710 | 930 | M | M | M | M |
| {12,345,6} | 405 | * | * | 465 | 580 | 740 | M | M | M |
| {12345,6} | 160 | * | * | 125 | 170 | 230 | 385 | 465 | 545 |
| {123456} | * | * | * | 10 | 10 | 10 | 10 | 10 | 10 |

* AIC, cannot, in theory attain this power
M Sample size >1000

Table 2. Minimum Sample Size Requirement: Effect Size=0.25

| | AIC | | | BIC | | | Tukey HSD | | |
|---|---|---|---|---|---|---|---|---|---|
| Power | .5 | .67 | .8 | .5 | .67 | .8 | .5 | .67 | .8 |
| Pattern of means | | | | | | | | | |
| Three groups | | | | | | | | | |
| {1,2,3} | 90 | 125 | 160 | 225 | 265 | 325 | 195 | 240 | 285 |
| {12,3} | 20 | 30 | 60 | 30 | 45 | 60 | 60 | 75 | 90 |
| {123} | 10 | 10 | * | 10 | 10 | 10 | 10 | 10 | 10 |
| Four groups | | | | | | | | | |
| {1,2,3,4} | 220 | 275 | 335 | 530 | 640 | 730 | 480 | 575 | 655 |
| {12,3,4} | 100 | 145 | 235 | 210 | 255 | 310 | 250 | 305 | 360 |
| {1,23,4} | 60 | 90 | 125 | 120 | 155 | 185 | 200 | 230 | 280 |
| {123,4} | 20 | 45 | * | 25 | 35 | 50 | 65 | 80 | 95 |
| {1234} | 10 | * | * | 10 | 10 | 10 | 10 | 10 | 10 |
| Five groups | | | | | | | | | |
| {1,2,3,4,5} | 385 | 475 | 565 | 985 | M | M | M | M | M |
| {12,3,4,5} | 240 | 320 | 485 | 520 | 620 | 740 | 585 | 670 | 765 |
| {12,3,45} | 100 | 185 | * | 145 | 200 | 245 | 335 | 395 | 450 |
| {1,234,5} | 55 | 90 | * | 85 | 100 | 130 | 175 | 210 | 240 |
| {1234,5} | 20 | * | * | 20 | 25 | 40 | 60 | 75 | 90 |
| {12345} | 10 | * | * | 10 | 10 | 10 | 10 | 10 | 10 |
| Six groups | | | | | | | | | |
| {1,2,3,4,5,6} | 640 | 760 | 925 | M | M | M | M | M | M |
| {12,3,4,5,6} | 460 | 610 | 880 | M | M | M | M | M | M |
| {1,2,34,5,6} | 310 | 420 | 550 | 670 | 765 | 900 | 885 | M | M |
| {1,2,3,456} | 260 | 420 | * | 545 | 650 | 740 | 680 | 765 | 905 |
| {1,2,345,6} | 170 | 270 | * | 300 | 360 | 430 | 470 | 540 | 625 |
| {12,34,56} | 85 | 250 | * | 105 | 140 | 175 | 360 | 415 | 480 |
| {12,345,6} | 65 | * | * | 65 | 90 | 110 | 260 | 305 | 340 |
| {12345,6} | 30 | * | * | 20 | 25 | 40 | 65 | 75 | 90 |
| {123456} | * | * | * | 10 | 10 | 10 | 10 | 10 | 10 |

* AIC, cannot, in theory attain this power
M Sample size >1000

Table 3. Minimum Sample Size Requirement: Effect Size=0.40

| | AIC | | | BIC | | | Tukey HSD | | |
|---|---|---|---|---|---|---|---|---|---|
| Power | .5 | .67 | .8 | .5 | .67 | .8 | .5 | .67 | .8 |
| Pattern of means | | | | | | | | | |
| Three groups | | | | | | | | | |
| {1,2,3} | 40 | 50 | 60 | 75 | 95 | 115 | 80 | 90 | 120 |
| {12,3} | 10 | 15 | 25 | 10 | 15 | 25 | 25 | 30 | 40 |
| {123} | 10 | 10 | * | 10 | 10 | 10 | 10 | 10 | 10 |
| Four groups | | | | | | | | | |
| {1,2,3,4} | 85 | 105 | 135 | 195 | 230 | 275 | 190 | 220 | 260 |
| {12,3,4} | 40 | 55 | 85 | 75 | 90 | 110 | 105 | 125 | 145 |
| {1,23,4} | 25 | 40 | 60 | 45 | 55 | 70 | 80 | 90 | 105 |
| {123,4} | 10 | 15 | * | 10 | 15 | 20 | 25 | 30 | 35 |
| {1234} | 10 | * | * | 10 | 10 | 10 | 10 | 10 | 10 |
| Five groups | | | | | | | | | |
| {1,2,3,4,5} | 160 | 195 | 235 | 365 | 425 | 475 | 365 | 415 | 480 |
| {12,3,4,5} | 100 | 130 | 190 | 200 | 235 | 290 | 245 | 285 | 315 |
| {12,3,45} | 50 | 70 | * | 60 | 80 | 100 | 140 | 165 | 185 |
| {1,234,5} | 25 | 35 | * | 30 | 40 | 50 | 75 | 90 | 105 |
| {1234,5} | 10 | * | * | 10 | 15 | 15 | 25 | 35 | 40 |
| {12345} | 10 | * | * | 10 | 10 | 10 | 10 | 10 | 10 |
| Six groups | | | | | | | | | |
| {1,2,3,4,5,6} | 250 | 300 | 365 | 580 | 690 | 765 | 600 | 675 | 760 |
| {12,3,4,5,6} | 175 | 235 | 350 | 385 | 455 | 525 | 455 | 515 | 580 |
| {1,2,34,5,6} | 120 | 155 | 240 | 235 | 280 | 330 | 355 | 400 | 450 |
| {1,2,3,456} | 105 | 155 | * | 190 | 235 | 275 | 260 | 305 | 350 |
| {1,2,345,6} | 65 | 105 | * | 105 | 130 | 160 | 190 | 220 | 245 |
| {12,34,56} | 40 | 85 | * | 40 | 50 | 65 | 145 | 165 | 190 |
| {12,345,6} | 30 | * | * | 25 | 35 | 45 | 105 | 115 | 130 |
| {12345,6} | 15 | * | * | 10 | 10 | 20 | 25 | 30 | 40 |
| {123456} | * | * | * | 10 | 10 | 10 | 10 | 10 | 10 |

* AIC, cannot, in theory attain this power
M Sample size >1000

## References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B.N. Petrov & F. Csake (eds.), *Second International Symposium on Information Theory*. Budapest: Akademiai Kiado, 267-281.

Aptech Systems, Inc. (2002). *Gauss for Windows version 4*. Maple Valley, WA

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

Cohen, J. (1969). *Statistical power analysis for the behavioral sciences*. New York: Academic Press.

Cribbie, R. A. & Keselman, H. J. (2003). Pairwise multiple comparisons: A model comparison approach versus stepwise procedures. *British Journal of Mathematical & Statistical Psychology*, *56*, 167-182.

Cribbie, R. A. (2003). Pairwise multiple comparisons: New yardstick, new results. *The Journal of Experimental Education*, *71*, 251-265.

Dayton, C. M. (1998). Information criteria for the paired-comparisons problem. *The American Statistician*, *52*, 144-151.

Dayton, C. M. (2003). Information criteria for pairwise comparisons. *Psychological Methods*, *8*, 61-71.

Ramsey, P. H. (1978). Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika*, *52*, 333-343.

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

Journal of Modern Applied Statistical Methods
November, 2005, Vol. 4, No. 2, 609-620


1538 – 9472/05/$95.00

*JMASM Algorithms and Code*

# JMASM20: Exact Permutation Critical Values For The Kruskal-Wallis One-Way ANOVA

Justice I. Odiase  Sunday M. Ogbonmwan
Department of Mathematics
University of Benin, Nigeria

The exhaustive enumeration of all the permutations of the observations in an experiment is the only possible way of truly constructing exact tests of significance. The permutation paradigm requires no distributional assumptions and works well with values that are normal, almost normal and non-normally distributed. The Kruskal-Wallis test does not require the assumptions that the samples are from normal populations and that the samples have the same standard deviation. In this article, the exact permutation distribution of the Kruskal-Wallis test statistic is generated empirically by actually obtaining all the distinct permutations of an experiment. The tables of exact critical values for the Kruskal-Wallis one-way ANOVA are produced.

Keywords: Permutation test, Kruskal-Wallis test, p-value, permutation algorithm, one-way ANOVA.

## Introduction

Variation is inherent in nature and errors are made occasionally when inferences are drawn from experiments. The risk in decision making cannot be totally eliminated but it can be controlled if correct statistical procedures are employed. The unconditional permutation approach is a statistical procedure that ensures that the probability of a type I error is exactly $\alpha$ and ensures that the resulting distribution of the test statistic is exact (Agresti, 1992; Good, 2000; Pesarin, 2001).

Scheffe (1943) demonstrated that for a general class of problems, the permutation approach is the only possible method of constructing exact tests of significance. It is asymptotically as powerful as the best parametric test (Hoeffding, 1952). In this article, consideration is given to the exhaustive permutation of the ranks of the observations in a single factor multi-sample experiment to arrive at the exact distribution of the Kruskal-Wallis (K-W) test statistic.

The method of obtaining an exact test of significance originated with Fisher (1935). The essential feature is that all the distinct arrangements of the observations are considered, with the proviso that all permutations are equally likely under the null hypothesis. An exact test on the level of significance $\alpha$ is constructed by choosing a proportion, $\alpha$, of the permutation as the critical region.

Statisticians have considered for some decades the possibility of generating exact critical values for the common test statistics that are in use today. This has resulted in the development of several ways such as the exact conditional permutation approach (Fisher, 1935; Agresti, 1992), the Monte Carlo approaches such as the Bootstrap (Efron, 1979; Efron and Tibshirani, 1993), the Bayesian approach (Casella & Robert, 2004), and the likelihood approach (Owen, 1988; Barndorff-Nielsen & Hall, 1988).

J. I. Odiase is a Lecturer in the Department of Mathematics. His areas of research are statistical computing and nonparametric statistics. Email him at justiceodiase@yahoo.com. S. M. Ogbonmwan is an Associate Professor of Statistics, Department of Mathematics, University of Benin, Nigeria. His areas of research are statistical computing and nonparametric statistics. Email him at ogbonmwasmaltra@yahoo.co.uk.

The works of Siegel and Castellan (1989), Conover (1999), Headrick (2003), Bagui & Bagui (2004) are contributions to the quest for exact critical values but the distributions are obtained from either simulation or asymptotic approximations of the distribution of the K-W test statistic. For small samples, $n_i \leq 5$, i = 1(1)p in a p-sample experiment, the null distribution of K-W statistic is not known and a chi-square approximation will not be a good approximation, (see Bagui & Bagui (2004)). The consideration given in this article produces the exact distribution of the K-W test statistic for small samples.

Distribution-free analysis of variance

The single-factor ANOVA model for comparing p populations or treatment means assumes that for i = 1, 2, …, p, a random sample of size n is drawn from a normal population with mean $\mu_i$ and variance $\sigma^2$. The normality assumption is required for the validity of the F test while the validity of the Kruskal-Wallis test for testing equality of the $\mu_i$'s (Kruskal & Wallis, 1952) depends only on the amount by which observed values deviate from their means $\mu_i$'s (random error) having the same continuous distribution.

Given a multisample experiment with

$$X_i = \left(X_{i1}, X_{i2}, ..., X_{in_i}\right)^T, \text{ i = 1(1)p}$$

and

$$\mathbf{X_N} = \left(X_1, X_2, ..., X_p\right),$$

where $N = \sum_{i=1}^{p} n_i$, the total number of observations in the data set. Suppose that one ranks all the N observations from 1 (smallest $X_{ij}$) to N (largest $X_{ij}$), the permutation test procedure presented in this article, computes an empirical estimate of the cumulative distribution of the test statistic T under the null hypothesis. Let the layout of the ranks of the observations $X_{ij}$ be as follows:

$$R_i = \left(r_{i1}, r_{i2}, ..., r_{in_i}\right)^T, \text{ i = 1(1)p.}$$

and

$$\mathbf{R_N} = \left(R_1, R_2, ..., R_p\right), \text{N} = \sum_{i=1}^{p} n_i .$$

Under the null hypothesis, $\mathbf{R_N}$ is composed of N independent and identically distributed random variables and hence conditioned on the observed data set. An exhaustive permutation of the ranks yields

$$\text{M} = \frac{N!}{\prod_{i=1}^{P}\left[(n_i)!\right]}$$

permutations of the N ranks of the variates of p subsets of size $n_i$, i = 1(1)p which are equally likely, each having the conditional probability $M^{-1}$.

When $H_0 : \mu_1 = \mu_2 = ... = \mu_p$ is true, the N observations are assumed to have come from the same distribution, in which case all possible assignments of the rank 1, 2, …, N to the p samples are equally likely and the ranks will be intermingled in these samples. Let $R_{ij}$ denote the rank of the jth observation in the ith treatment $X_{ij}$. Let $R_{i\bullet}$ and $\bar{R}_{i\bullet}$ denote respectively the total and mean of the ranks in the ith treatment. The K-W test statistic is a measure of the extent to which the $\bar{R}_{i\bullet}$'s deviate from their common expected value $\frac{N+1}{2}$, and $H_0$ is rejected if the computed value of the statistic indicates too great a discrepancy between observed and expected rank averages. The K-W test statistic is

$$\text{H} = \frac{12}{N(N+1)}\sum_{i=1}^{p}\frac{R_{i\bullet}^2}{n_i} - 3(N+1).$$

If $H_0$ is rejected when $H \geq c$, then c should be chosen so that the test has level $\alpha$. That is, c should be the upper-tail critical value of the distribution of H when $H_0$ is true. Under $H_0$, each possible assignment can be enumerated, the value of H determined for each one, and the null distribution obtained by

counting the number of times each value of H occurs. When $H_0$ is true, the large-sample approximation is applied if p = 3, $n_i \geq 6$, i = 1(1)3 or $p > 3$, $n_i \geq 5$, i = 1(1)p (Devore, 1982; Rohatgi, 1984). H has approximately a chi-squared distribution with p – 1 degrees of freedom. An approximate level $\alpha$ test is given by: Reject $H_0$ if $H \geq \chi^2_{\alpha,\, p-1}$.

## Methodology

The process of obtaining the permutations starts by choosing the test statistic T and the acceptable significance level $\alpha$. Let $\pi_1$, $\pi_2$, …, $\pi_n$ be a set of all distinct permutations of the ranks of the data set in the experiment. The permutation test procedure is as follows:

1. Rank the observations of the experiment as required by the K-W test.
2. Compute the observed value of the K-W test statistic ($H_1 = t_0$).
3. Obtain a distinct permutation $\pi_i$, of the ranks in Step 1.
4. Compute the K-W test statistic $H_i$ for permutation $\pi_i$ in Step 3, that is, $H_i = H(\pi_i)$.
5. Repeat Steps 3 and 4 for i = 2, 3, …, M.
6. Construct an empirical cumulative distribution for H

$$p_0 = p(H \leq H_i) = \frac{1}{M}\sum_{i=1}^{M}\psi(t_0 - H_i),$$

where

$$\psi(\cdot) = \begin{cases} 1, & \text{if } t_0 \geq H_i \\ 0, & \text{if } t_0 < H_i \end{cases}.$$

7.Under the empirical distribution, if $p_0 \leq \alpha$, reject the null hypothesis.

The complexity in permutation test lies in obtaining all the distinct permutations of the observations in a given experiment. For example, a four-sample experiment with six variates in each sample requires 2,308,743,493,056 permutations. The frequency distribution is constructed for all the distinct occurrences of the test statistic from which the probability distribution of the test statistic is computed.

The number of permutations of the ranks of a two-sample experiment is

$$\sum_{i=0}^{n}\binom{n_1}{i}\binom{n_2}{i},\ n = \min(n_1, n_2),$$

see Odiase & Ogbonmwan (2005) for details.

After obtaining the permutations of the ranks of a two sample experiment, the number of ways to permute the ranks of any $n_3$ of the combined ranks ($n_1 + n_2 + n_3$) of the variates of the three-sample experiment yields

$$\binom{n_1+n_2+n_3}{n_3}\sum_{i=0}^{n}\binom{n_1}{i}\binom{n_2}{i} = \binom{\sum_{k=1}^{3}n_k}{n_3}\sum_{i=0}^{n}\binom{n_1}{i}\binom{n_2}{i}$$

A complete enumeration of the distinct permutations of the ranks of a four-sample experiment yields

$$\binom{\sum_{k=1}^{4}n_k}{n_4}\binom{\sum_{k=1}^{3}n_k}{n_3}\sum_{i=0}^{n}\binom{n_1}{i}\binom{n_2}{i} = \prod_{j=3}^{4}\binom{\sum_{k=1}^{j}n_k}{n_j}\sum_{i=0}^{n}\binom{n_1}{i}\binom{n_2}{i}$$

Continuing in this manner, for p ≥ 3 treatments, the distinct permutations of the ranks of the variates are enumerated through

$$\prod_{j=3}^{p}\binom{\sum_{k=1}^{j}n_k}{n_j}\sum_{i=0}^{n}\binom{n_1}{i}\binom{n_2}{i} = \prod_{j=1}^{p}\binom{\sum_{k=1}^{j}n_k}{n_j}.$$

For the balanced case, $n_1 = n_2 = \ldots = n_p = n$, the number of distinct permutations of the ranks of the variates is $\prod_{j=1}^{p} \binom{jn}{n}$. As an illustration, let

$$R_i = \left(r_{i1}, r_{i2}, \ldots, r_{in_i}\right)^T, \; i = 1(1)p$$

and

$$\mathbf{R_N} = \left(R_1, R_2, \ldots, R_p\right).$$

Consider a three-sample experiment with observations $x_{ij}$, $n_1 = 3$, $n_2 = n_3 = 2$, that is, $\begin{pmatrix} x_{11} & x_{21} & x_{31} \\ x_{12} & x_{22} & x_{32} \\ x_{13} & & \end{pmatrix}$. Assuming there are no ties, the configuration of the ranks of the experiment can be taken as $\begin{pmatrix} r_{11} & r_{21} & r_{31} \\ r_{12} & r_{22} & r_{32} \\ r_{13} & & \end{pmatrix}$. An exhaustive permutation of this experiment yields 210 distinct permutations of the ranks.

First obtain the 6 permutations of the ranks of the 4 variates of the last two treatments, that is,

$$\begin{pmatrix} r_{21} & r_{31} \\ r_{22} & r_{32} \end{pmatrix}, \begin{pmatrix} r_{31} & r_{21} \\ r_{22} & r_{32} \end{pmatrix}, \begin{pmatrix} r_{32} & r_{31} \\ r_{22} & r_{21} \end{pmatrix}, \begin{pmatrix} r_{21} & r_{22} \\ r_{31} & r_{32} \end{pmatrix}, \begin{pmatrix} r_{21} & r_{31} \\ r_{32} & r_{22} \end{pmatrix}, \begin{pmatrix} r_{31} & r_{21} \\ r_{32} & r_{22} \end{pmatrix}.$$

There are 35 ways to permute any 3 ranks of the combined 7 ranks of the variates of the experiment.

$$\begin{pmatrix} r_{11} \\ r_{12} \\ r_{13} \end{pmatrix}, \begin{pmatrix} r_{11} \\ r_{12} \\ r_{21} \end{pmatrix}, \begin{pmatrix} r_{11} \\ r_{21} \\ r_{13} \end{pmatrix}, \begin{pmatrix} r_{21} \\ r_{12} \\ r_{13} \end{pmatrix}, \begin{pmatrix} r_{11} \\ r_{12} \\ r_{22} \end{pmatrix}, \begin{pmatrix} r_{11} \\ r_{22} \\ r_{13} \end{pmatrix},$$
$$\begin{pmatrix} r_{22} \\ r_{12} \\ r_{13} \end{pmatrix}, \begin{pmatrix} r_{11} \\ r_{12} \\ r_{31} \end{pmatrix}, \begin{pmatrix} r_{11} \\ r_{31} \\ r_{13} \end{pmatrix}, \begin{pmatrix} r_{31} \\ r_{12} \\ r_{13} \end{pmatrix}, \begin{pmatrix} r_{11} \\ r_{12} \\ r_{32} \end{pmatrix}, \begin{pmatrix} r_{11} \\ r_{32} \\ r_{13} \end{pmatrix},$$
$$\begin{pmatrix} r_{32} \\ r_{12} \\ r_{13} \end{pmatrix}, \begin{pmatrix} r_{11} \\ r_{21} \\ r_{22} \end{pmatrix}, \begin{pmatrix} r_{21} \\ r_{12} \\ r_{22} \end{pmatrix}, \begin{pmatrix} r_{21} \\ r_{22} \\ r_{13} \end{pmatrix}, \begin{pmatrix} r_{11} \\ r_{21} \\ r_{31} \end{pmatrix}, \begin{pmatrix} r_{21} \\ r_{12} \\ r_{31} \end{pmatrix},$$
$$\begin{pmatrix} r_{21} \\ r_{31} \\ r_{13} \end{pmatrix}, \begin{pmatrix} r_{11} \\ r_{21} \\ r_{32} \end{pmatrix}, \begin{pmatrix} r_{21} \\ r_{12} \\ r_{32} \end{pmatrix}, \begin{pmatrix} r_{21} \\ r_{32} \\ r_{13} \end{pmatrix}, \begin{pmatrix} r_{11} \\ r_{22} \\ r_{31} \end{pmatrix}, \begin{pmatrix} r_{22} \\ r_{12} \\ r_{31} \end{pmatrix},$$
$$\begin{pmatrix} r_{22} \\ r_{31} \\ r_{13} \end{pmatrix}, \begin{pmatrix} r_{11} \\ r_{22} \\ r_{32} \end{pmatrix}, \begin{pmatrix} r_{22} \\ r_{12} \\ r_{32} \end{pmatrix}, \begin{pmatrix} r_{22} \\ r_{32} \\ r_{13} \end{pmatrix}, \begin{pmatrix} r_{11} \\ r_{31} \\ r_{32} \end{pmatrix}, \begin{pmatrix} r_{31} \\ r_{12} \\ r_{32} \end{pmatrix},$$
$$\begin{pmatrix} r_{31} \\ r_{32} \\ r_{13} \end{pmatrix}, \begin{pmatrix} r_{21} \\ r_{22} \\ r_{31} \end{pmatrix}, \begin{pmatrix} r_{21} \\ r_{22} \\ r_{32} \end{pmatrix}, \begin{pmatrix} r_{21} \\ r_{31} \\ r_{32} \end{pmatrix}, \begin{pmatrix} r_{22} \\ r_{31} \\ r_{32} \end{pmatrix}.$$

Each of the 35 ways will combine with the 6 permutations of the remaining 4 ranks of the variates making up the last two treatments in any configuration of the experiment, that is,

$$\binom{7}{3} \sum_{i=0}^{2} \binom{2}{i} \binom{2}{i}.$$

Consider the set of all these 210 permutations, for each one of them, compute the test statistic of interest and hence calculate the probability of the different values of the test statistic based on the number of times each is occurring. When ties occur in the data set, the tied observations are usually assigned the mean of the ranks they would have been assigned if they were distinct. Ties do not pose any problem to the permutation test presented in this article. Assuming no ties, the experiment just presented will have ranks {1, 2, 3, 4, 5, 6, 7} represented

as $\begin{pmatrix} 1 & 4 & 6 \\ 2 & 5 & 7 \\ 3 & & \end{pmatrix}$ and the distinct permutations of these ranks lead to the remaining 209 permutations.

Permutation algorithms

Considering the associated complexity in a complete enumeration of the distinct permutations necessary for the compilation of the distribution of the K-W test statistic, computer algorithms for an exhaustive enumeration are now presented.

The first step in developing permutation algorithm is to formulate an initial configuration of the ranks of the variates of an experiment by taking the trivial configuration given below as:

$$\begin{pmatrix} 1 & n_1+1 & n_1+n_2+1 & \cdots & \sum_{i=1}^{p-1} n_i+1 \\ 2 & : & : & : & \\ 3 & : & : & : & \\ 4 & : & : & : & \\ : & : & : & : & \\ : & & & & \\ n_1 & n_1+n_2 & n_1+n_2+n_3 & \cdots & \sum_{i=1}^{p} n_i \end{pmatrix}$$

Algorithm (PERMUTATION) of Odiase & Ogbonmwan (2005) can handle the permutation of the ranks of the variates in a two-sample experiment. Algorithm 1 in this article generates the distinct permutations of the ranks of the variates of a three-sample experiment and relies on the permutation of the ranks of the variates in a two-sample experiment.

Algorithm 2 calls Algorithm 1 and then generates the distinct permutations of the ranks of the variates of a four-sample experiment. Algorithms 1 and 2 can be extended to take care of the sample sizes under consideration.

## Results

Critical values for the K-W test statistic

The algorithms were implemented in Intel Visual Fortran. Figures 1 – 10 show the small sample distribution of the K-W test statistic for different sample sizes for 3 and 4 samples. The resulting tables of exact critical values as obtained from the exact permutation distribution of the K-W test statistic are presented in Tables 1 and 2.

## Conclusion

Figures 1 and 2 reveal the fact that the chi squared distribution, which is the large sample approximation of the K-W test statistic, will poorly approximate the exact distribution of the K-W test statistic for very small sample sizes. As sample sizes increase, the shape of the chi squared distribution begins to emerge as seen in Figures 3 – 10.

The critical values for a test statistic are usually determined by cutting off the most extreme 100α% of the theoretical frequency distribution of the test statistic, where α is the level of significance, see Siegel and Castellan (1989). The critical values of the K-W test statistic contained in Tables 1 and 2 are obtained from the enumeration of all the distinct permutations of the ranks of the variates in an experiment. These critical values are exact and therefore ensures that the probability of a type I error in decisions arising from the use of the K-W test is exactly α.

```
Algorithm 1 (3 samples)
1: for II10 ← 1, P do
2: for JJ10 ← 1, K(II10) do
3: Y(JJ10, II10) ← Z1(JJ10, II10)
4: Y1(JJ10, II10) ← Z1(JJ10, II10)
5: end for
6: end for
7: Obtain a distinct permutation of ranks in the last two samples
        Exchange one rank
8: for JJ1 ← 1, K(2) do
9: TEMPA ← Y1(JJ1, P - 2)
10: for II1 ← P-1, P do
11: for 110 JJ2 ← 1, K(II1) do
12: Y1(JJ1, P - 2) ← Y1(JJ2, II1)
13: Y1(JJ2, II1) ← TEMPA
14: Obtain a distinct permutation of ranks in the last two samples
15: end for
16: end for
17: end for
        Exchange two ranks
18: for II ← 1, K(2) – 1 do
19: TEMPA1 ← Y1(II, P – 2)
20: for JJ ← II + 1, K(2) do
21: TEMPA2 ← Y1(JJ, P – 2)
22: for LL ← P – 1, P do
23: for II1 ← 1, K(LL) do
24: for LL1 ← LL, P do
25: if LL ← LL1 then
26: TT ← II1 + 1
27: else
28: TT ← 1
29: end if
30: for JJ1 ← TT, K(LL1) do
31: Y1(II, P - 2) ← Y1(II1, LL)
32: Y1(II1, LL) ← TEMPA1
33: Y1(JJ, P - 2) ← Y1(JJ1, LL1)
34: Y1(JJ1, LL1) ← TEMPA2
35: Obtain a distinct permutation of ranks in the last two samples
36: end for
37: end for
38: end for
39: end for
40: end for
41: end for
42: …
        Restore original ranks
43: for II0 ← 1, P do
44: for JJ0 ← 1, K(II0) do
45: Z1(JJ0, II0) ← Z(JJ0, II0)
46: end for
47: end for
```

Algorithm 2 (4 samples)

Generate ranks

1: KK ← 0
2: for I ← 1, P do
3: KK ← KK + K(I-1)
4: for J ← 1, K(I) do
5: Z(J, I) ← KK + J
6: Z1(J, I) ← Z(J, I)
7: Y(J, I) ← Z(J, I)
8: Y1(J, I) ← Y(J, I)
9: X(J, I) ← Z(J, I)
10: X1(J, I) ← X(J, I)
11: end for
12: end for
13: call Algorithm 1
14: for R2 ← 1, P do
15: for R3 ← 1, K do
16: Y(R3,R2) ← Z1(R3,R2)
17: Y1(R3,R2) ← Z1(R3,R2)
18: end for
19: end for

Adjust Algorithm 1 as follows and insert here:

20: Change all the loop variables
21: Change the variable names TEMPA, TEMPA1, TEMPA2, TT, TT1, …
22: Replace Steps 10, 22, … with [Variable name ← P – 2, P]
23: Replace all [P – 2] with [P – 3]
24: Replace [Y1] with [Z1]
25: Replace [Obtain a distinct permutation of ranks in the last two samples] with [Call Algorithm 1]
26: Construct the empirical distribution of H
27: Sort values of H in ascending order of magnitude
28: Construct the CDF for H

Figures 1 – 10: Distribution of Kruskal-Wallis test statistic for different sample sizes

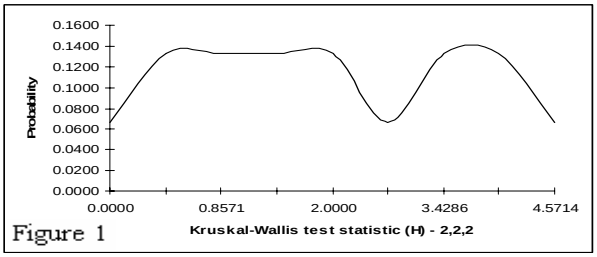


Figure 1

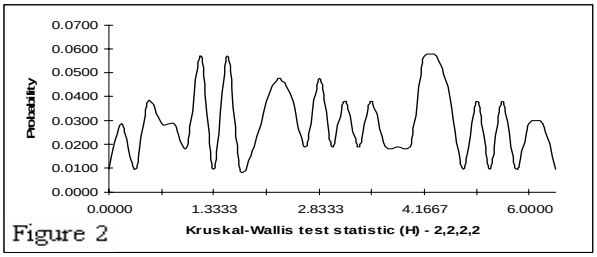


Figure 2

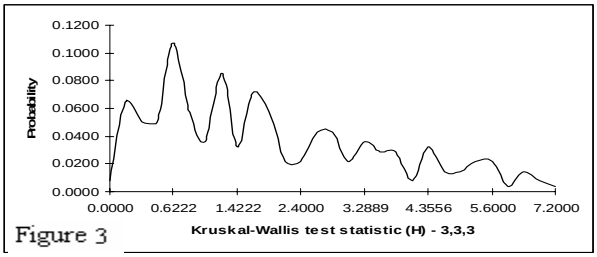


Figure 3

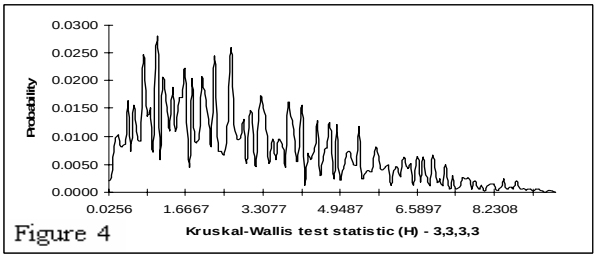


Figure 4

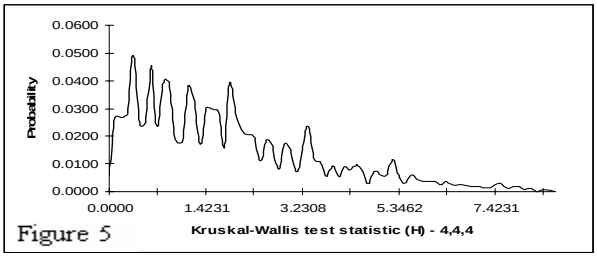


Figure 5

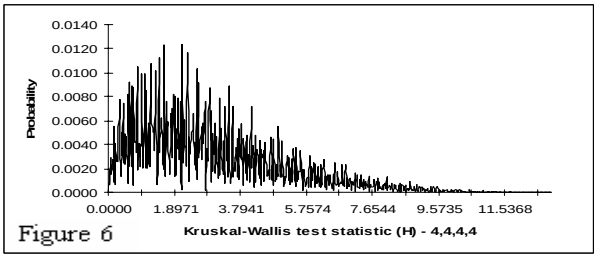


Figure 6

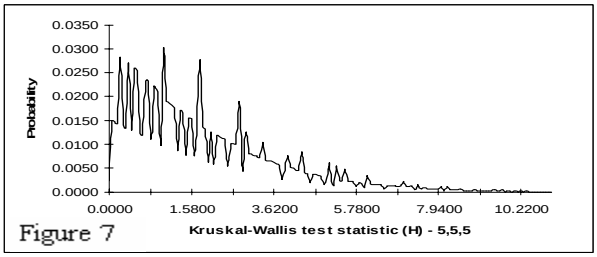


Figure 7

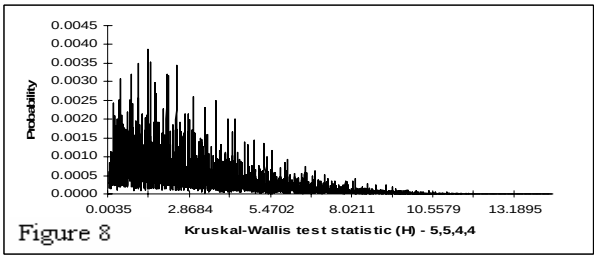


Figure 8

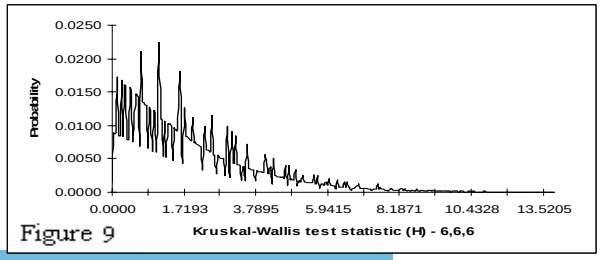


Figure 9

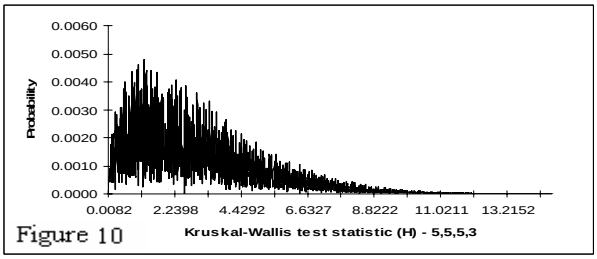


Figure 10

Table 1: Critical values for Kruskal-Wallis test statistic (3 samples)

| *Sample Size* | $H_{0.9000}$ | $H_{0.9500}$ | $H_{0.9750}$ | $H_{0.9900}$ | $H_{0.9950}$ | $H_{0.9975}$ | $H_{0.9990}$ |
|---|---|---|---|---|---|---|---|
| 2,2,1 | 3.6000 | | | | | | |
| 2,2,2 | 3.7143 | 4.5714 | | | | | |
| 3,2,1 | 4.2857 | | | | | | |
| 3,2,2 | 4.4643 | 4.5000 | 5.3571 | | | | |
| 3,3,1 | 4.5714 | | | | | | |
| 3,3,2 | 4.5556 | 5.1389 | 5.5556 | 6.2500 | | | |
| 3,3,3 | 4.6222 | 5.6000 | 5.9556 | 6.4889 | | | |
| 4,2,1 | 4.0179 | 4.8214 | | | | | |
| 4,2,2 | 4.4583 | 5.1250 | 5.3333 | 6.0000 | | | |
| 4,3,1 | 3.8889 | 5.4000 | 5.3889 | | | | |
| 4,3,2 | 4.4444 | 5.4000 | 5.8000 | 6.3000 | 6.4444 | 7.0000 | |
| 4,3,3 | 4.7000 | 5.7273 | 6.0182 | 6.7455 | 7.0000 | 7.3182 | 8.0182 |
| 4,4,1 | 4.0667 | 4.8667 | 6.0000 | 6.1667 | | | |
| 4,4,2 | 4.4455 | 5.2364 | 6.0818 | 6.8727 | 7.0364 | 7.2818 | 7.8545 |
| 4,4,3 | 4.4773 | 5.5758 | 6.3864 | 7.1364 | 7.4773 | 7.8485 | 8.3258 |
| 4,4,4 | 4.5000 | 5.6538 | 6.5769 | 7.5385 | 7.7308 | 8.1154 | 8.7692 |
| 5,1,1 | 3.8571 | | | | | | |
| 5,2,1 | 4.0500 | 4.4500 | 5.2500 | | | | |
| 5,2,2 | 4.2933 | 5.0400 | 5.6933 | 6.1333 | | | |
| 5,3,1 | 3.8400 | 4.8711 | 5.7600 | 6.4000 | | | |
| 5,3,2 | 4.4945 | 5.1055 | 5.9491 | 6.8218 | 6.9491 | 7.1818 | 7.6364 |
| 5,3,3 | 4.4121 | 5.5152 | 6.3030 | 6.9818 | 7.5152 | 7.8788 | 8.2424 |
| 5,4,1 | 3.9600 | 4.8600 | 5.7764 | 6.8400 | 6.9545 | | |
| 5,4,2 | 4.5182 | 5.2682 | 6.0409 | 7.1182 | 7.5682 | 7.8136 | 8.1136 |
| 5,4,3 | 4.5231 | 5.6308 | 6.3949 | 7.3949 | 7.9064 | 8.2564 | 8.6256 |
| 5,4,4 | 4.6187 | 5.6176 | 6.5967 | 7.7440 | 8.1560 | 8.7033 | 9.1286 |
| 5,5,1 | 4.0364 | 4.9091 | 5.7818 | 6.8364 | 7.7455 | 7.7455 | 8.1818 |
| 5,5,2 | 4.5077 | 5.2462 | 6.2308 | 7.2692 | 8.0769 | 8.2923 | 8.6846 |
| 5,5,3 | 4.5363 | 5.6264 | 6.4879 | 7.5429 | 8.2637 | 8.7912 | 9.2835 |
| 5,5,4 | 4.5200 | 5.6429 | 6.6714 | 7.7914 | 8.4629 | 9.0257 | 9.5057 |
| 5,5,5 | 4.5000 | 5.6600 | 6.7200 | 7.9800 | 8.7200 | 9.3800 | 9.9200 |
| 6,1,1 | 4.0833 | | | | | | |
| 6,2,1 | 3.8222 | 4.6222 | 5.4000 | | | | |
| 6,2,2 | 4.4364 | 5.0182 | 5.5273 | 6.5455 | 6.6545 | | |
| 6,3,1 | 3.8182 | 4.8545 | 5.8545 | 6.5818 | | | |
| 6,3,2 | 4.5455 | 5.2273 | 6.0606 | 6.7273 | 7.5000 | 7.5758 | 8.1818 |
| 6,3,3 | 4.5385 | 5.5513 | 6.3846 | 7.1923 | 7.6154 | 8.3205 | 8.6282 |
| 6,4,1 | 3.8636 | 4.9242 | 5.6970 | 7.0833 | 7.5000 | 7.9545 | |
| 6,4,2 | 4.4359 | 5.2628 | 6.1090 | 7.2115 | 7.8205 | 8.3077 | 8.6667 |
| 6,4,3 | 4.5989 | 5.6044 | 6.5000 | 7.4670 | 8.0275 | 8.6538 | 9.1703 |
| 6,4,4 | 4.5238 | 5.6667 | 6.5952 | 7.7238 | 8.3238 | 8.8810 | 9.6286 |
| 6,5,1 | 3.9205 | 4.8359 | 5.8615 | 6.9974 | 8.0667 | 8.4359 | 8.8846 |
| 6,5,2 | 4.4747 | 5.3187 | 6.1890 | 7.2989 | 8.1868 | 8.7473 | 9.1890 |
| 6,5,3 | 4.4971 | 5.6000 | 6.6210 | 7.5600 | 8.2971 | 9.0286 | 9.6686 |
| 6,5,4 | 4.5000 | 5.6558 | 6.7358 | 7.8958 | 8.6400 | 9.2933 | 9.9600 |
| 6,5,5 | 4.5294 | 5.6985 | 6.7809 | 8.0118 | 8.8353 | 9.5809 | 10.2706 |
| 6,6,1 | 3.9780 | 4.8571 | 5.9121 | 7.0659 | 7.9341 | 8.9231 | 9.3077 |
| 6,6,2 | 4.4190 | 5.3524 | 6.1714 | 7.4095 | 8.1524 | 8.9333 | 9.6762 |
| 6,6,3 | 4.5250 | 5.6000 | 6.6833 | 7.6833 | 8.4167 | 9.2250 | 10.1250 |
| 6,6,4 | 4.5184 | 5.7206 | 6.7831 | 7.9890 | 8.7206 | 9.4118 | 10.3419 |
| 6,6,5 | 4.5412 | 5.7516 | 6.8379 | 8.1190 | 8.9817 | 9.7242 | 10.5242 |
| 6,6,6 | 4.5380 | 5.7193 | 6.8772 | 8.1871 | 9.0877 | 9.8713 | 10.8421 |

Table 2: Critical values for Kruskal-Wallis test statistic (4 samples)

| *Sample Size* | $H_{0.9000}$ | $H_{0.9500}$ | $H_{0.9750}$ | $H_{0.9900}$ | $H_{0.9950}$ | $H_{0.9975}$ | $H_{0.9990}$ |
|---|---|---|---|---|---|---|---|
| 2,2,1,1 | 4.7143 | | | | | | |
| 2,2,2,1 | 5.0357 | 5.3571 | 5.6786 | | | | |
| 2,2,2,2 | 5.5000 | 6.0000 | 6.1667 | | | | |
| 3,2,1,1 | 4.8929 | 5.4643 | | | | | |
| 3,2,2,1 | 5.3889 | 5.8056 | 6.0556 | 6.5000 | | | |
| 3,2,2,2 | 5.6444 | 6.2444 | 6.6444 | 7.0000 | 7.1333 | 7.5333 | |
| 3,3,1,1 | 5.2222 | 5.8889 | | | | | |
| 3,3,2,1 | 5.6222 | 6.1556 | 6.5111 | 7.0444 | 7.2000 | 7.4000 | |
| 3,3,2,2 | 5.7273 | 6.4727 | 7.0000 | 7.6364 | 7.7273 | 8.0000 | 8.1273 |
| 3,3,3,1 | 5.5818 | 6.5273 | 6.8909 | 7.3273 | 7.7636 | 8.0545 | 8.3455 |
| 3,3,3,2 | 5.8182 | 6.6818 | 7.4697 | 7.9545 | 8.3182 | 8.5606 | 8.9242 |
| 3,3,3,3 | 5.9744 | 6.8974 | 7.6154 | 8.4359 | 8.7436 | 9.1538 | 9.4615 |
| 4,2,1,1 | 5.2083 | 5.4583 | 6.0833 | | | | |
| 4,2,2,1 | 5.5000 | 6.0000 | 6.5000 | 6.8000 | | | |
| 4,2,2,2 | 5.6727 | 6.4364 | 6.9818 | 7.3091 | 7.8545 | 7.9636 | 8.2909 |
| 4,3,1,1 | 4.9778 | 6.0444 | 6.5667 | 6.7111 | | | |
| 4,3,2,1 | 5.5727 | 6.3000 | 6.9091 | 7.3636 | 7.7273 | 7.8909 | 8.1818 |
| 4,3,2,2 | 5.7121 | 6.6136 | 7.3182 | 7.8485 | 8.2500 | 8.5909 | 8.8939 |
| 4,3,3,1 | 5.6667 | 6.5379 | 7.2727 | 7.7500 | 8.1212 | 8.3561 | 8.8409 |
| 4,3,3,2 | 5.8590 | 6.7821 | 7.5577 | 8.3205 | 8.7179 | 9.0577 | 9.4038 |
| 4,3,3,3 | 6.0000 | 6.9670 | 7.7582 | 8.6538 | 9.2308 | 9.5769 | 10.0000 |
| 4,4,1,1 | 5.1273 | 5.8636 | 6.9273 | 7.5000 | | | |
| 4,4,2,1 | 5.5455 | 6.3636 | 7.1364 | 7.8864 | 8.2273 | 8.5682 | 8.7045 |
| 4,4,2,2 | 5.7692 | 6.6923 | 7.5192 | 8.3077 | 8.6731 | 9.0577 | 9.4423 |
| 4,4,3,1 | 5.6603 | 6.6154 | 7.4808 | 8.2179 | 8.5769 | 8.8654 | 9.2949 |
| 4,4,3,2 | 5.8901 | 6.8626 | 7.7363 | 8.6099 | 9.1538 | 9.4835 | 9.9121 |
| 4,4,3,3 | 6.0048 | 7.0333 | 7.9238 | 8.8667 | 9.4905 | 9.9667 | 10.4619 |
| 4,4,4,1 | 5.6374 | 6.7088 | 7.6319 | 8.5714 | 8.9505 | 9.2473 | 9.7253 |
| 4,4,4,2 | 5.9000 | 6.9429 | 7.8857 | 8.8571 | 9.4714 | 9.9143 | 10.4000 |
| 4,4,4,3 | 6.0292 | 7.1292 | 8.0542 | 9.0667 | 9.7167 | 10.3417 | 10.9000 |
| 4,4,4,4 | 6.0662 | 7.2132 | 8.2059 | 9.2647 | 9.9485 | 10.5662 | 11.3382 |
| 5,2,1,1 | 5.1067 | 5.7600 | 6.0667 | 6.6000 | | | |
| 5,2,2,1 | 5.5309 | 6.0327 | 6.5782 | 7.2000 | 7.4727 | 7.8000 | |
| 5,2,2,2 | 5.6182 | 6.5273 | 7.1545 | 7.6636 | 8.0182 | 8.3818 | 8.6818 |
| 5,3,1,1 | 5.1309 | 6.0036 | 6.8764 | 7.1673 | 7.4000 | | |
| 5,3,2,1 | 5.5030 | 6.3303 | 7.0939 | 7.7455 | 8.1818 | 8.2909 | 8.7273 |
| 5,3,2,2 | 5.7538 | 6.6564 | 7.4641 | 8.1949 | 8.6256 | 8.9333 | 9.4231 |
| 5,3,3,1 | 5.6564 | 6.6000 | 7.4205 | 8.1179 | 8.5282 | 8.8974 | 9.2564 |
| 5,3,3,2 | 5.8571 | 6.8220 | 7.6505 | 8.5912 | 9.0571 | 9.4176 | 9.8549 |
| 5,3,3,3 | 5.9981 | 7.0114 | 7.8267 | 8.8400 | 9.4571 | 9.9067 | 10.4095 |
| 5,4,1,1 | 5.2000 | 6.0182 | 6.8000 | 7.8591 | 8.2000 | 8.2955 | 8.6364 |
| 5,4,2,1 | 5.5615 | 6.4077 | 7.2115 | 8.1692 | 8.5731 | 8.9423 | 9.3231 |
| 5,4,2,2 | 5.7725 | 6.7220 | 7.5989 | 8.4692 | 9.0495 | 9.4451 | 9.8604 |
| 5,4,3,1 | 5.6396 | 6.6813 | 7.5253 | 8.3989 | 8.9802 | 9.3484 | 9.7934 |
| 5,4,3,2 | 5.8933 | 6.9171 | 7.7933 | 8.8000 | 9.3933 | 9.8733 | 10.3543 |
| 5,4,3,3 | 6.0292 | 7.0892 | 7.9892 | 9.0292 | 9.6958 | 10.2892 | 10.8558 |
| 5,4,4,1 | 5.6686 | 6.7429 | 7.6743 | 8.7171 | 9.3029 | 9.6971 | 10.2114 |
| 5,4,4,2 | 5.9400 | 6.9850 | 7.9475 | 9.0000 | 9.6625 | 10.2525 | 10.7875 |
| 5,4,4,3 | 6.0346 | 7.1669 | 8.1346 | 9.2118 | 9.9397 | 10.5574 | 11.2963 |
| 5,4,4,4 | 6.0608 | 7.2569 | 8.2725 | 9.3902 | 10.1373 | 10.8020 | 11.5882 |
| 5,5,1,1 | 5.0923 | 6.0154 | 6.8769 | 8.0769 | 8.6000 | 8.9077 | 9.0923 |

Table 2: Continued

| *Sample Size* | $H_{0.9000}$ | $H_{0.9500}$ | $H_{0.9750}$ | $H_{0.9900}$ | $H_{0.9950}$ | $H_{0.9975}$ | $H_{0.9990}$ |
|---|---|---|---|---|---|---|---|
| 5,5,2,1 | 5.5648 | 6.5341 | 7.2725 | 8.3077 | 9.0198 | 9.4352 | 9.7582 |
| 5,5,2,2 | 5.7943 | 6.7714 | 7.6457 | 8.6286 | 9.2914 | 9.8800 | 10.3429 |
| 5,5,3,1 | 5.6476 | 6.7371 | 7.6286 | 8.5962 | 9.2743 | 9.7619 | 10.2191 |
| 5,5,3,2 | 5.9150 | 6.9417 | 7.8750 | 8.9467 | 9.6350 | 10.1667 | 10.8200 |
| 5,5,3,3 | 6.0118 | 7.1176 | 8.0588 | 9.1882 | 9.9176 | 10.5529 | 11.2353 |
| 5,5,4,1 | 5.6625 | 6.7800 | 7.7625 | 8.8625 | 9.5500 | 10.1025 | 10.5900 |
| 5,5,4,2 | 5.9338 | 7.0279 | 8.0162 | 9.1500 | 9.8868 | 10.5154 | 11.1904 |
| 5,5,4,3 | 6.0523 | 7.2157 | 8.2092 | 9.3562 | 10.1307 | 10.7895 | 11.5739 |
| 5,5,4,4 | 6.0684 | 7.2895 | 8.3421 | 9.5351 | 10.3281 | 11.0228 | 11.8439 |
| 5,5,5,1 | 5.6824 | 6.8294 | 7.8000 | 9.0176 | 9.7588 | 10.3941 | 10.9588 |
| 5,5,5,2 | 5.9451 | 7.0745 | 8.0941 | 9.2863 | 10.0980 | 10.7451 | 11.5137 |
| 5,5,5,3 | 6.0433 | 7.2456 | 8.2889 | 9.4959 | 10.3193 | 10.9930 | 11.8257 |

## References


Agresti, A. (1992). A survey of exact inference for contingency tables, *Statistical Science, 7*, 131-177.

Bagui, S., & Bagui, S. (2004). An algorithm and code for computing exact critical values for the Kruskal-Wallis nonparametric one-way ANOVA. *Journal of Modern Applied Statistical Methods, 3*, 498-503.

Barndorff-Nielsen, O. E., & Hall, P. (1988). On the level-error after Bartlett adjustment of the likelihood ratio statistic. *Biometrika, 75*, 374-378.

Conover, W. J. (1999). *Practical nonparametric statistics*. New York: Wiley.

Devore, J. L. (1982). *Probability and statistics for engineering and the sciences*. California: Brooks/Cole Publishing Company.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics, 7*, 1-26.

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

Fisher, R. A. (1935). *The design of experiments*. Edinburgh: Oliver and Boyd

Good, P. (2000). *Permutation tests: A practical guide to resampling methods for testing hypotheses (2nd ed.)*. New York: Springer Verlag.

Headrick, T. C. (2003). An algorithm for generating exact critical values for the Kruskal-Wallis one-way ANOVA. *Journal of Modern Applied Statistical Methods, 2*, 268-271.

Hoeffding, W. (1952). Large sample power of tests based on permutations of observations. *The Annals of Mathematical Statistics, 23*, 169-192.

Kruskal, W. H., & Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association, 47*, 583-634.

Odiase, J. I., & Ogbonmwan, S. M. (2005) An algorithm for generating unconditional exact permutation distribution for a two-sample experiment. *Journal of Modern Applied Statistical Methods, 4*, 319-332.

Owen, A. B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika, 75*, 237-249.

Pesarin, F. (2001). *Multivariate permutation tests.* New York: Wiley.

Rohatgi, V. K. (1984). *Statistical inference*. New York: John Wiley & Sons.

Scheffe, H. (1943) Statistical inference in the nonparametric case. *The Annals of Mathematical Statistics, 14*, 305-332.

Siegel, S., & Castellan, N. J. (1989). *Nonparametric statistics for the behavioural sciences (3rd ed.)*. New York: McGraw-Hill.

Journal of Modern Applied Statistical Methods
November, 2005, Vol. 4, No. 2, 621-626

1538 – 9472/05/$95.00

# JMASM21: PCIC_SAS: Best Subsets Using Information Criteria

C. Mitchell Dayton Xuemei Pan
Department of Measurement & Statistics
University of Maryland

PCIC_SAS is a SAS program for identifying optimal subsets of means based on independent groups. All possible configurations of ordered subsets of groups are considered and a best model is identified using both the AIC and BIC information criteria. Results for models with homogeneous variances as well as models with heterogeneity of variance in the same pattern as the means are reported.

Key words: PCIC_SAS, information criteria, AIC, BIC, paired-comparisons

## Introduction

Researchers often use analysis of variance (ANOVA) to investigate mean differences among several groups. If the null hypothesis of equality of means is rejected, it is common practice to employ multiple comparison techniques to further study the pattern of differences among the means. For example, Kirk (1995) described 22 multiple comparison procedures including nine pairwise comparisons such as the Tukey honestly significantly different (HSD) procedure and Dunnett's T3 test. Statistical packages often include a variety of competing procedures with, for example, SAS 8.1 allowing the user to choose among 12 distinct methods for pairwise comparisons. Often, these procedures depend upon interpreting multiple significance tests. As detailed in the next section, Dayton (1998, 2003) advocated replacing these procedures by a holistic model selection approach based on information criteria. The purpose of this article is to describe and make available to applied researchers a SAS program, PCIC_SAS, that implements this modern information theoretic approach for comparisons among means.

C. Mitchell Dayton is Professor and Chair of the Department of Measurement, Statistics & Evaluation. His research interests include experimental design and latent class modeling. Email him at cdayton@umd.edu. Xuemei Pan is a Ph D candidate in the Department of Measurement, Statistics & Evaluation. Her research interests include latent class modeling and model comparison procedures. Email her at xpan1@umd.edu. The program mentioned in this article is available at www.education.umd.edu/EDMS/Latent/PCIC.txt

## Application of Information Criteria to the Paired-Comparisons of Means

The widely-used Tukey Honestly Significantly Different (HSD) procedure for K independent group means involves the computation of q statistics for the K(K – 1)/2 different pairs of means and refers these statistics to the appropriate null distribution of the studentized range statistic for a span of K means. Like similar pair-wise comparison procedures, Tukey HSD entails testing K(K – 1)/2 hypotheses of the form $\mu_k = \mu_{k'}$ for $k \neq k'$. Often this is done subsequent to testing the omnibus hypothesis of equality of means (i.e., $\mu_k = \mu$ for k = 1,…,K) using analysis of variance techniques. Theoretically, the omnibus test is not required since the K-range pairwise comparison is an equivalent, although less powerful, test. There are many optional procedures based on modifications to the Tukey procedure or based on related notions using stepwise procedures. See, for example, the Kirk (1995) reference cited above for details of many of these procedures.

Among the problems with pairwise comparison procedures cited by Dayton (1998, 2003) are:

(1) Some arbitrary technique is utilized to control the family-wise type I error rate for the set of correlated pairwise tests;
(2) The issues of homogeneity of variance and differential sample size pose problems for many paired-comparison procedures;
(3) Intransitive decisions (e.g., outcomes suggesting mean 1 = mean 2, mean 2 = mean 3, but mean 1 < mean 3) are the rule rather than the exception with typical paired comparison procedures because they entail a series of discrete, pairwise significance tests;
(4) There exists a large variety of competing procedures that differ in how type I error is controlled and, consequently, in power (e.g., SPSS 11.5 for Windows offers eighteen distinct procedures to choose among).

For K independent groups, there is a total of $2^{K-1}$ patterns of ordered subsets with equal means within subsets. For example, with four groups with means ranked and labeled 1, 2, 3, 4, the $2^3 = 8$ distinct ordered subsets are {1234}, {1,234}, {12,34}, {123,4}, {1,2,34}, {1,23,4}, {12,3,4} and {1,2,3,4}, where a comma is used to separate subsets with unequal means. Dayton (1998, 2003) proposed using model-selection criteria such as the Akaike (1973) AIC statistic for selecting the most appropriate ordering of subsets of means for purposes of interpretation. In particular, this approach avoids many of the objections that can be raised with respect to conventional pairwise comparison procedures. Information criteria such as AIC are based on the logarithm of the likelihood of the data, $Log_e$(likelihood). Sclove (1987) noted that AIC represents a penalized log-likelihood function of the general form:

$$-2Log_eL(likelihood) + a(n)p$$

where a(n) is a function that may depend upon the total sample size, n, and p is the number of independent parameters estimated in fitting the model to the data. Akaike's AIC is equal to

$$-2Log_eL(likelihood) + 2p$$

which does not directly depend upon sample size. Various adaptations of or alternatives to AIC have been suggested that, unlike AIC, are explicitly dependent upon sample size. In particular, the Schwarz (1978) BIC statistic and the Bozdogan (1987) CAIC statistic use penalty terms equal to $Log_e(n)$ and $Log_e(n) + 1$, respectively. As noted by Bozdogan (1987), these latter procedures are asymptotically consistent in the sense that, when the null case is the true model, the probability of selecting the true model approaches one, rather than an arbitrary significance level, as is true for conventional hypothesis testing procedures. It is beyond the scope of this article to discuss the basis for selecting among alternative information criteria. However, these issues are discussed in Dayton (2003).

In practice, AIC (or, BIC) is computed for all competing models that the researcher wishes to compare. Then, from an information theoretic perspective, the model satisfying a min(AIC) (or, min(BIC)) criterion is selected as the best approximating model for the data being analyzed. Note that the min(AIC) (or, min(BIC)) strategy does not suggest that the selected model either fits or does not fit the data but that, among the models being compared, it is, in the information sense, the best choice. If additional models were added to the basis of comparison, a different selection might occur although the previously computed AIC values would not be altered.

The program, PCIC_SAS, computes both the Akaike AIC and the Schwarz BIC statistics for all $2^{K-1}$ distinct ordered subsets. Since the number of ordered subsets can, in practice, become quite large (e.g., 512 for K = 10 groups but 524,288 for K = 20 groups), only the ordered subsets corresponding to the smallest AIC and BIC values, as specified by the user (e.g., 5), are printed out. There is no limit to the number of groups that can be analyzed but, of course, execution time can become relatively

long for large K. In PCIC_SAS, it is assumed that the observations arise from normal densities.

Note, that the log-likelihood is maximized for any given model when variance estimates are computed using the sample size, n, rather than n-1, in the denominator. PCIC_SAS calculates AIC and BIC based on the usual assumption of homogeneity of variance as well as based on a restricted heterogeneous variance model in which it is assumed that there is a unique population variance for each of the distinct subsets of means. For the homogeneous case, the conventional analysis of variance within-groups sum of squares, $SS_w$, is converted to a variance estimate, $SS_w/n$, where n is the total sample size. For the restricted, heterogeneous variance case, an estimated variance for a subset of means can be obtained (a) by pooling the estimates from the separate groups or (b) by computing the sample variance for the combined sample. The latter approach is illustrated in Dayton (1998, 2003) and is the procedure incorporated into PCIC_SAS.

For a model with T subsets of means, the number of independent parameters, p, is equal to T+1 for the homogeneous case and 2T for the restricted heterogeneous case. Because $Log_e(n)$ is greater than 2 for n greater than 7, AIC and BIC may, and often do, result in different orderings of subsets of means with, predictably, simpler models being favored by BIC because of the larger penalty term. In Dayton (1998), results of a limited simulation with AIC and CAIC (the slightly different criterion than BIC with penalty term $Log_e(n+1)p$ suggested by Bozdogan (1987)), it was found that: "Overall…the accuracy of CAIC is always approximately equal to or superior to Tukey HSD but tends to be lower than AIC when there are relatively many clusters of means, especially with smaller sample sizes." For a more extensive simulation providing favorable results for PCIC, see Cribbie and Keselman (2003).

Using the PCIC_SAS Program

PCIC_SAS is written in the SAS programming language. For general-purpose analysis with a major statistical computer package, there is no other program that computes AIC and/or BIC for the models available in PCIC_SAS. For a small number of groups (e.g., 5 or less), it is reasonably easy to program the computations in a spreadsheet as was reported by Dayton (1998). For users of the matrix-language, Gauss (Aptech Systems, 1997), appropriate code that provides input from spreadsheets such as Microsoft Excel is available (Dayton, 2001).

Data for analysis with PCIC_SAS can be in a SAS data base or imported into SAS from a spreadsheet or database program. It is conventional to code the groups with names, or 1, 2, etc., or A, B, etc. but PCIC_SAS rearranges the groups in rank order of means, from smallest to largest, and presents groups in ranked order, 1, 2, etc., in the output. Results are directed to the SAS output screen that can be printed and/or saved.

Example

Summary statistics for five ethic groups, based on a 5% random sample of cases from the NELS88 database, are presented below (see //nces.ed.gov/surveys/nels88/ for information about the longitudinal study of youth). The dependent variable is mathematics achievement on a standardized scale with population mean of about 50 and standard deviation of about 10. The five groups, as documented with the database, are: (1) API (Asian/Pacific Islander), (2) Hispanic, (3) Black-Non-Hispanic, (4) White-Non-Hispanic, and (5) American Indian. In rank order of means from low to high on the output these become: (3) Black-Non-Hispanic, (2) Hispanic, (5) American Indian, (4) White-Non-Hispanic and (1) API. The PCIC_SAS summary table and output for the five smallest values of AIC and BIC are summarized below:

Summary Table - group means in original order

| Obs | race | _FREQ_ | mean | sd | n | varunb | varmle | sum | ss |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 75 | 53.25 | 10.26 | 75.00 | 105.19 | 103.79 | 3993.45 | 7783.89 |
| 2 | 2 | 139 | 47.00 | 8.28 | 139.00 | 68.50 | 68.01 | 6532.98 | 9453.36 |
| 3 | 3 | 153 | 45.63 | 8.37 | 153.00 | 70.09 | 69.63 | 6981.58 | 10654.00 |
| 4 | 4 | 798 | 52.96 | 10.14 | 798.00 | 102.78 | 102.65 | 42258.81 | 81913.54 |
| 5 | 5 | 44 | 47.21 | 7.22 | 44.00 | 52.15 | 50.96 | 2077.40 | 2242.25 |
| | | 1209 | | | | | | 112047.04 | |

Summary Table - group means in rank order

| Obs | race | _FREQ_ | mean | sd | n | varunb | varmle | sum | ss |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 3 | 153 | 45.63 | 8.37 | 153.00 | 70.09 | 69.63 | 6981.58 | 10654.00 |
| 2 | 2 | 139 | 47.00 | 8.28 | 139.00 | 68.50 | 68.01 | 6532.98 | 9453.36 |
| 3 | 5 | 44 | 47.21 | 7.22 | 44.00 | 52.15 | 50.96 | 2077.40 | 2242.25 |
| 4 | 4 | 798 | 52.96 | 10.14 | 798.00 | 102.78 | 102.65 | 42258.81 | 81913.54 |
| 5 | 1 | 75 | 53.25 | 10.26 | 75.00 | 105.19 | 103.79 | 3993.45 | 7783.89 |

AIC and BIC for Homogeneous Case

Rank of AIC, value of AIC and ordered subsets for homogeneous variance case:

AIC_HOMOG

| | | | | | | |
|---|---|---|---|---|---|---|
| 1 | 8914.598 | 1 | 1 | 1 | 2 | 2 |
| 2 | 8914.785 | 1 | 2 | 2 | 3 | 3 |
| 3 | 8916.240 | 1 | 1 | 2 | 3 | 3 |
| 4 | 8916.535 | 1 | 1 | 1 | 2 | 3 |
| 5 | 8916.722 | 1 | 2 | 2 | 3 | 4 |

Rank of BIC, value of BIC and ordered subsets for homogeneous variance case:

BIC_HOMOG

| | | | | | | |
|---|---|---|---|---|---|---|
| 1 | 8929.890 | 1 | 1 | 1 | 2 | 2 |
| 2 | 8935.175 | 1 | 2 | 2 | 3 | 3 |
| 3 | 8936.630 | 1 | 1 | 2 | 3 | 3 |
| 4 | 8936.926 | 1 | 1 | 1 | 2 | 3 |
| 5 | 8942.210 | 1 | 2 | 2 | 3 | 4 |

AIC and BIC for Heterogeneous Case

Rank of AIC, value of AIC and ordered subsets for patterned heterogeneous variance case:

AIC_HETEROG

| | | | | | | |
|---|---|---|---|---|---|---|
| 1 | 8895.898 | 1 | 1 | 1 | 2 | 2 |
| 2 | 8897.075 | 1 | 2 | 2 | 3 | 3 |
| 3 | 8897.724 | 1 | 1 | 2 | 3 | 3 |
| 4 | 8899.729 | 1 | 2 | 3 | 4 | 4 |
| 5 | 8899.838 | 1 | 1 | 1 | 2 | 3 |

Rank of BIC, value of BIC and ordered subsets for patterned heterogeneous variance case:

BIC_HETEROG

| | | | | | | |
|---|---|---|---|---|---|---|
| 1 | 8916.288 | 1 | 1 | 1 | 2 | 2 |
| 2 | 8927.660 | 1 | 2 | 2 | 3 | 3 |
| 3 | 8928.309 | 1 | 1 | 2 | 3 | 3 |
| 4 | 8930.423 | 1 | 1 | 1 | 2 | 3 |
| 5 | 8936.311 | 1 | 1 | 2 | 2 | 2 |

Interpretation

For AIC, all five reported heterogeneous-variance models have smaller values than the best homogeneous-variance model and for BIC this is true for the first three heterogeneous models. Thus, models with variances that differ among subsets of means are favored over homogeneous-variance models. Based on both AIC and BIC, the preferred model is reported as: 1, 1, 1, 2, 2. This suggests that there are two subsets of means comprised of the groups with the three smallest means in one subset and the groups with the two largest means in the second subset. This corresponds to the pattern {Black-Non-Hispanic, Hispanic, American Indian} in the subset with smaller means and {White-Non-Hispanic, API} in the subset with larger means. Note that the conclusion should not be drawn that, for example, the means are equal for the White-Non-Hispanic and API groups but, rather that the data are not sufficiently reliable to permit an ordering within that subset. The variances for the two subsets are not reported but can be easily computed from the output (see Dayton, 1998) and are equal to 67.02 and 102.75, respectively.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csake (eds.), *Second International Symposium on Information Theory*. Budapest: Akademiai Kiado, 267-281.

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

Akaike, H. (1987). Factor analysis and AIC. *Psychometrika, 52*, 317-332.

Aptech Systems, Inc. (1997). GAUSS for Windows NT/95: Version 3.2.32, Maple Valley, WA.

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

Cribbie, R. A. & Keselman, H. J. (2003). A power comparison of pairwise multiple comparison procedures: A model testing approach versus stepwise procedures. *British Journal of Statistical & Mathematical Psychology, 56*, 157-182.

Dayton, C. M. (1998). Information Criteria for the Paired-Comparisons Problem. *American Statistician, 52*, 144-151.

Dayton, C. M. (2001). SUBSET: Best subsets using information criteria. *Journal of Statistical Software, 6*(2).

Dayton, C. M. (2003). Information criteria for pairwise comparisons. *Psychological Methods, 8, 61-7*.

Greeley, A. M., McCready, W. C. & Theisen, G. (1980). *Ethnic drinking subcultures*. New York: Praeger.

Kirk, R. E. (1995). *Experimental design* (3rd ed.). Brooks/Cole.

Kullback, S. & Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics, 22*, 79-86.

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

Sclove, S. (1987). Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika, 52*, 333-343.

Appendix

The theoretical background for AIC derives from information-theoretic concepts originally presented by Kullback and Leibler (1951). The mathematical material presented in this section is supplementary to that presented above and can be skimmed or omitted without any serious loss of understanding of the PCIC technique.

Adapting the notation of Akaike (1973, 1974, 1987) for univariate data, the Kullback-Leibler information for the true distribution, $g_t(x)$, of random variable x, relative to some other distribution, $g_o(x)$, is:

(1) $I(g_t; g_o) = E(\text{Log}_e[g_t(x)]) - E(\text{Log}_e[g_o(x)])$

where all expectations are taken with respect to $g_t(x)$. In statistical applications making use of maximum likelihood estimation, let $\mathbf{x} = \{x_i\}$ be n values of an iid random variable, x, with true density function $g(\cdot \mid \boldsymbol{\theta})$ based on the parameter vector, $\boldsymbol{\theta}$, and let $\boldsymbol{\theta}_x$ be the usual maximum likelihood estimator (MLE) of $\boldsymbol{\theta}$ found by maximizing $g(\mathbf{x} \mid \boldsymbol{\theta})$ over the sample by treating $\boldsymbol{\theta}$ as variable. Assuming p independent parameters, a large-sample result for the distribution of likelihood ratios is:

$$(2) \qquad L_1 = 2\{\mathrm{Log}_e[g(\mathbf{x} \mid \boldsymbol{\theta}_x)] - \mathrm{Log}_e[g(\mathbf{x} \mid \boldsymbol{\theta}_t)]\} = \chi_p^2$$

where $\chi_p^2$ is central chi-square with p degrees of freedom.

Let y be an additional observation from the same distribution as **x**. Akaike (1974) shows that, asymptotically:

$$(3) \qquad L_2 = 2\{E_y\mathrm{Log}_e[g(y \mid \boldsymbol{\theta}_x)] - E_y\mathrm{Log}_e[g(y \mid \boldsymbol{\theta}_t)]\} = -\chi_p^2$$

Then:

$$(4) \qquad E(L_1 - L_2) = 2\mathrm{Log}_e[g(\mathbf{x} \mid \boldsymbol{\theta}_x)] - 2E_y\mathrm{Log}_e[g(y \mid \boldsymbol{\theta}_x)] \approx 2p.$$

Noting that the first term in Equation (1) is constant for any model, Akaike defines the AIC estimator of Kullback-Leibler information as:

$$(5) \qquad \text{Constant} - E_y\mathrm{Log}_e[g(y \mid \boldsymbol{\theta}_x)] \approx -2\mathrm{Log}_e[g(\mathbf{x} \mid \boldsymbol{\theta}_x)] + 2p = \mathrm{AIC}$$

For M different models for the same data, the Akaike min(AIC) procedure involves using Equation (5) to calculate $\mathrm{AIC}_m$, m = 1,…,M, for the models and selecting the model with $\min(\mathrm{AIC}_m)$ as the preferred model. The conventional interpretation of AIC is as an estimate of the loss of precision (or, increase in information) that results when $\boldsymbol{\theta}_x$, the MLE, is substituted for the true parametric value, $\boldsymbol{\theta}_t$, in the likelihood function.

Sclove (1987) notes that AIC represents a penalized log-likelihood function of the general form:

$$(6) \qquad -2\mathrm{Log}_e[g(\mathbf{x} \mid \boldsymbol{\theta}_x)] + a(n)p$$

where a(n) is a function that may depend upon the total sample size, n. Various adaptations of AIC have been suggested that, unlike AIC, make the statistic dependent upon sample size. In particular, the Schwarz (1978) BIC statistic and the Bozdogan (1987) CAIC statistic use penalty terms equal to $\mathrm{Log}_e(n)$ and $\mathrm{Log}_e(n) + 1$, respectively. As noted by Bozdogan (1987), these latter procedures are asymptotically consistent in the sense that, when the null case is the true model, the probability of selecting the true model approaches one, rather than an arbitrary significance level, as is true for conventional hypothesis testing procedures.

Journal of Modern Applied Statistical Methods
November, 2005, Vol. 4, No. 2, 627-628

1538 – 9472/05/$95.00

# Statistical Pronouncements IV

"By sampling, we can learn only about collective properties of populations, not about the properties of individuals" – William G. Cochran, Frederick Mosteller, & John W. Tukey (1954, Principles of sampling, *Journal of the American Statistical Association*, 49, p. 17).

"If a student has not already at least some facility with graphs and logarithms then he is, I believe, ill-advised to start to grapple with the theory of statistics" – Bernard L. Welch, (1954, Book Reviews, *Journal of the American Statistical Association*, 49, p. 378).

"Type, paper, and binding are good" – H. W. Norton (1954, Book Reviews, *Journal of the American Statistical Association*, 49, p. 390).

"It is curious that there are many people who are established scientists, and many others offering to become scientists, who have so little mathematics" – H. W. Norton (1954, Book Reviews, *Journal of the American Statistical Association*, 49, p. 390).

"Choice of subject matter is always an author's prerogative" – J. H. Curtiss, (1954, Book Reviews, *Journal of the American Statistical Association*, 49, p. 401).

"Once upon a time the calculation of the first four moments was an honorable art in statistics" – John W. Tukey (1954, Unsolved problems of experimental statistics, *Journal of the American Statistical Association, 49*, p. 717).

"Why isn't someone writing a book on one- and two-sample techniques? (After all, there is a book being written on the straight line!" – John W. Tukey (ibid, p. 721).

"With this issue, the *Journal* will discontinue publication of random digits" – (1954, RANDOM DIGITS (20,876-21,875), *Journal of the American Statistical Association, 49*, p. 928).

"At a Galton Laboratory tea in 1937, when there were few text books to guide a student in study of statistical methods for research, Fisher remarked that the way to obtain a good one would be for everyone who might feel the urge to try his hand and see which product would survive… The flood is now upon us" – H. Fairfield Smith, (1955, Book Reviews, *Journal of the American Statistical Association*, 50, p. 975).

"If told another elementary text is to be written, my reaction is: Please, not another!" – H. Fairfield Smith, (1955, Book Reviews, *Journal of the American Statistical Association*, 50, p. 979).

"Leonard J. Savage recommended us to read about the foundations of statistics sitting bolt upright on a hard chair, at a desk, and now [Michel] Loève asks us to approach his monumental treatise on the foundations of probability theory "armed permanently with patience, pebble, and reed" – Walter L. Smith, (1955, Book Reviews, *Journal of the American Statistical Association*, 50, p. 986).

"This is by far the largest and best collection of random digits yet" – W. Allen Wallis (1955, Book Reviews, *Journal of the American Statistical Association*, 50, p. 987)

"Good examples in theoretical statistics are not easy to find" – Herman Chernoff (1955, Book Reviews, *Journal of the American Statistical Association*, 50, p. 1334).

"The naive reader is almost certain to form a set of incorrect ideas concerning inference about distribution means. He is likely to feel that the assumption of normality of distribution is about on the same level as the use of a sharp pencil - nice but not exactly necessary" – Leo Katz (1955, Book Reviews, *Journal of the American Statistical Association*, 50, p. 1344).

"The practitioner of statistical inferences must understand much more of his art than he brings to bear on a specific problem; therefore, the 'cookbook' approach cannot succeed" – Leo Katz (ibid, p. 1344).

"I never knew a statistician who thought he knew enough mathematics" – Leonard J. Savage (1955, Book Reviews, *Journal of the American Statistical Association*, 50, p. 1352).

"Nondeductive reasoning is of paramount importance to the statistician" – Leonard J. Savage (ibid, p. 1352).

"The statistician like the scientist has to be concerned primarily with the collection and arrangement of and the reasonable inferences from observed data. Some mathematics will surely help, too much will surely hinder" – Edwin B. Wilson (1955, Book Reviews, *Journal of the American Statistical Association*, 50, p. 1356).

"Using P = .05 [is] all right if understood, but the businessman, the investor, the weather forecaster, the executive, or the card player who waited for that degree of significance would be so out of the game as to be without a livelihood" – Edwin B. Wilson (ibid, p. 1357).

"Science is always provisional and usually approximately, and thus constantly being corrected" – Edwin B. Wilson (ibid, p. 1357).

"Every statistician knows something about stochastic processes, though like me he may be late to learn, and never entirely comfortable with, that awesome sounding name – Leonard J. Savage (1956, Book Reviews, *Journal of the American Statistical Association*, 51, p. 383)

"The best philosophers are often mathematicians" – – I. J. Good (1956, Book Reviews, *Journal of the American Statistical Association*, 51, p. 388).

"Decision theory is not a subject that can be appreciated in all its austere details by a statistician with less than one or two years of experience of real life" – I. J. Good (1956, Book Reviews, *Journal of the American Statistical Association*, 51, p. 388).

"To all of us some of the time and to some of us all of the time it seems that economics fails to make progress as other sciences do" – Robert M. Solow (1956, Book Reviews, *Journal of the American Statistical Association*, 51, p. 398).

"1955 saw the creation of a new Committee to Investigate Statistics as Evidence. The Committee, under the chairmanship of John Tukey, was appointed in response to recommendations based on the fact that many lawyers fail to recognize the validity of statistics as evidence" – American Statistical Association (1956, Report of the board of directors, 1955, *Journal of the American Statistical Association*, 51, p. 424).

"Much oh!ing and ah!ing has been heard in the land about those prodigious giants, the new electronic computing machines" – Thornton C. Fry (1956, The automatic computer in industry, *Journal of the American Statistical Association*, 51, p. 565).

"The theory of decision making, the natural sequel to hypothesis testing, has elevated the notion of risk to an even higher place in the hierarchy of ideas passed on from professor to student" – A. W. Kimball (1957, Errors of the third kind in statistical consulting, *Journal of the American Statistical Association*, 52, p. 133).

"A scientist with ideas frames his hypotheses and wishes to test them" – E. J. G. Pitman (1957, Statistics and science, *Journal of the American Statistical Association*, 52, p. 323).

"Nonparametric methods are needed in many fields, and can be applied in all" – I. Richard Savage (1957, Nonparametric statistics, *Journal of the American Statistical Association*, 52, p. 331).

# DataMineIt℠

## announces

# PermuteIt™ v2.0

## The fastest, most comprehensive and robust permutation test software on the market today.

**Permutation tests increasingly are the statistical method of choice for addressing business questions and research hypotheses across a broad range of industries. Their distribution-free nature maintains test validity where many parametric tests (and even other nonparametric tests), encumbered by restrictive and often inappropriate data assumptions, fail miserably. The computational demands of permutation tests, however, have severely limited other vendors' attempts at providing useable permutation test software for anything but highly stylized situations or small datasets and few tests. PermuteIt™ addresses this unmet need by utilizing a combination of algorithms to perform non-parametric permutation tests very quickly – often more than an order of magnitude faster than widely available commercial alternatives when one sample is large and many tests and/or multiple comparisons are being performed (which is when runtimes matter most). PermuteIt™ can make the difference between making deadlines, or missing them, since data inputs often need to be revised, resent, or recleaned, and one hour of runtime quickly can become 10, 20, or 30 hours.**

**In addition to its speed even when one sample is large, some of the unique and powerful features of PermuteIt™ include:**

- **the availability to the user of a wide range of test statistics for performing permutation tests on continuous, count, & binary data, including: pooled-variance t-test; separate-variance Behrens-Fisher t-test, scale test, and joint tests for scale and location coefficients using nonparametric combination methodology; Brownie et al. "modified" t-test; skew-adjusted "modified" t-test; Cochran-Armitage test; exact inference; Poisson normal-approximate test; Fisher's exact test; Freeman-Tukey Double Arcsine test**

- **extremely fast exact inference (no confidence intervals – just exact p-values) for most count data and high-frequency continuous data, often several orders of magnitude faster than the most widely available commercial alternative**

- **the availability to the user of a wide range of multiple testing procedures, including: Bonferroni, Sidak, Stepdown Bonferroni, Stepdown Sidak, Stepdown Bonferroni and Stepdown Sidak for discrete distributions, Hochberg Stepup, FDR, Dunnett's one-step (for MCC under ANOVA assumptions), Single-step Permutation, Stepdown Permutation, Single-step and Stepdown Permutation for discrete distributions, Permutation-style adjustment of permutation p-values**

- **fast, efficient, and automatic generation of all pairwise comparisons**

- **efficient variance-reduction under conventional Monte Carlo via self-adjusting permutation sampling when confidence intervals contain the user-specified critical value of the test**

- **maximum power, and the shortest confidence intervals, under conventional Monte Carlo via a new sampling optimization technique (see Opdyke, JMASM, Vol. 2, No. 1, May, 2003)**

- **fast permutation-style p-value adjustments for multiple comparisons (the code is designed to provide an additional speed premium for many of these resampling-based multiple testing procedures)**

- **simultaneous permutation testing and permutation-style p-value adjustment, although for relatively few tests at a time (this capability is not even provided as a preprogrammed option with any other software currently on the market)**

**For Telecommunications, Pharmaceuticals, *f*MRI data, Financial Services, Clinical Trials, Insurance, Bioinformatics, and just about any data rich industry where large numbers of distributional null hypotheses need to be tested on samples that are not extremely small and parametric assumptions are either uncertain or inappropriate, PermuteIt™ is the optimal, and only, solution.**

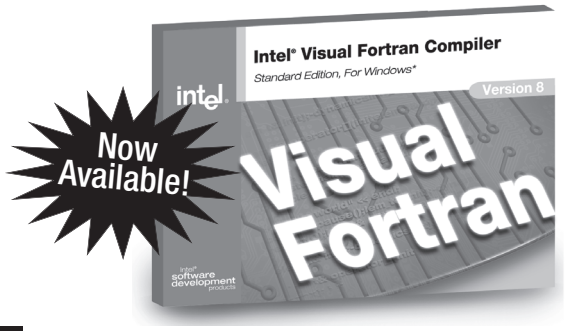
**To learn more about how PermuteIt™ can be used for your enterprise, and to obtain a demo version, please contact its author, J.D. Opdyke, President, DataMineIt℠, at JDOpdyke@DataMineIt.com or www.DataMineIt.com.**

**DataMineIt℠ is a technical consultancy providing statistical data mining, econometric analysis, and data warehousing services and expertise to the industry, consulting, and research sectors. PermuteIt™ is its flagship product.**

# Two Years in the Making...

## Intel® Visual Fortran 8.0

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




### Visual Fortran Timeline

**1997** DEC releases Digital Visual Fortran 5.0

**1998** Compaq acquires DEC and releases DVF 6.0

**1999** Compaq ships CVF 6.1

**2001** Compaq ships CVF 6.6

**2001** Intel acquires CVF engineering team

**2003** Intel releases Intel Visual Fortran 8.0

### Intel Visual Fortran 8.0

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

## Performance

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

## Compatibility

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

## Support

1 year of free product upgrades and Intel Premier Support

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

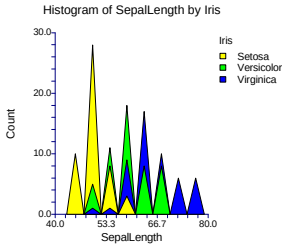
*—Dr. Robert R. Trippi*
*Professor Computational Finance*
*University of California, San Diego*

*FREE trials available at:*
programmersparadise.com/intel

Programmer's Paradise®

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

**NCSS**
**329 North 1000 East**
**Kaysville, Utah 84037**




# Announcing NCSS 2004
## Seventeen New Procedures

**NCSS 2004** is a new edition of our popular statistical **NCSS** package that adds seventeen new procedures.

### New Procedures
Two Independent Proportions
Two Correlated Proportions
One-Sample Binary Diagnostic Tests
Two-Sample Binary Diagnostic Tests
Paired-Sample Binary Diagnostic Tests
Cluster Sample Binary Diagnostic Tests
Meta-Analysis of Proportions
Meta-Analysis of Correlated Proportions
Meta-Analysis of Means
Meta-Analysis of Hazard Ratios
Curve Fitting
Tolerance Intervals
Comparative Histograms
ROC Curves
Elapsed Time Calculator
T-Test from Means and SD's
Hybrid Appraisal (Feedback) Model

### Documentation
The printed, 330-page manual, called ***NCSS User's Guide V***, is available for $29.95. An electronic (pdf) version of the manual is included on the distribution CD and in the Help system.

### Two Proportions
Several new exact and asymptotic techniques were added for hypothesis testing (null, noninferiority, equivalence) and calculating confidence intervals for the difference, ratio, and odds ratio. Designs may be independent or paired. Methods include: Farrington & Manning, Gart & Nam, Conditional & Unconditional Exact, Wilson's Score, Miettinen & Nurminen, and Chen.

### Meta-Analysis
Procedures for combining studies measuring paired proportions, means, independent proportions, and hazard ratios are available. Plots include the forest plot, radial plot, and L'Abbe plot. Both fixed and random effects models are available for combining the results.

### Curve Fitting
This procedure combines several of our curve fitting programs into one module. It adds many new models such as Michaelis-Menten. It analyzes curves from several groups. It compares fitted models across groups using computer-intensive randomization tests. It computes bootstrap confidence intervals.

### Tolerance Intervals
This procedure calculates one and two sided tolerance intervals using both distribution-free (nonparametric) methods and normal distribution (parametric) methods. Tolerance intervals are bounds between which a given percentage of a population falls.

### Comparative Histogram
This procedure displays a comparative histogram created by interspersing or overlaying the individual histograms of two or more groups or variables. This allows the direct comparison of the distributions of several groups.

### Random Number Generator
Matsumoto's Mersenne Twister random number generator (cycle length > 10**6000) has been implemented.

### Binary Diagnostic Tests
Four new procedures provide the specialized analysis necessary for diagnostic testing with binary outcome data. These provide appropriate specificity and sensitivity output. Four experimental designs can be analyzed including independent or paired groups, comparison with a gold standard, and cluster randomized.

### ROC Curves
This procedure generates both binormal and empirical (nonparametric) ROC curves. It computes comparative measures such as the whole, and partial, area under the ROC curve. It provides statistical tests comparing the AUC's and partial AUC's for paired and independent sample designs.

### Hybrid (Feedback) Model
This new edition of our hybrid appraisal model fitting program includes several new optimization methods for calibrating parameters including a new genetic algorithm. Model specification is easier. Binary variables are automatically generated from class variables.

### Statistical Innovations Products
Through a *special arrangement* with Statistical Innovations (S.I.), NCSS customers will receive $100 discounts on:

**Latent GOLD**Ò - latent class modeling
**SI-CHAID**Ò - segmentation trees
**GOLDMineR**Ò - 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 _____

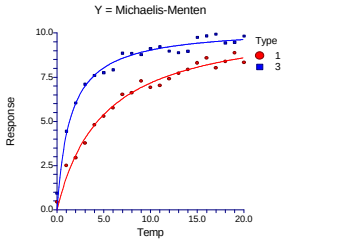
ADDRESS _____

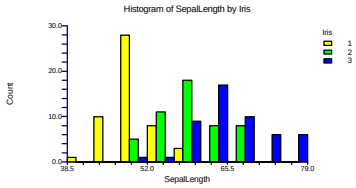
ADDRESS _____

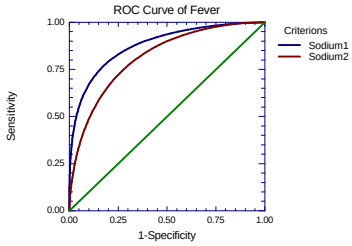
ADDRESS _____

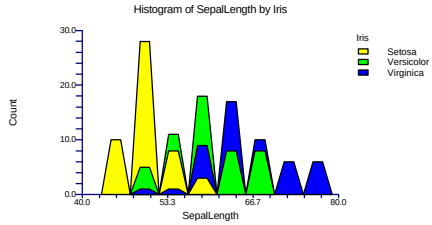
CITY _____ STATE _____

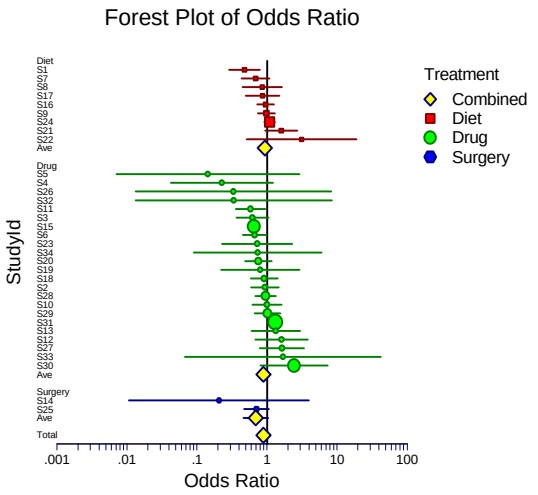
ZIP/POSTAL CODE _____ COUNTRY _____




# 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

### 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

*New Edition in 2004

# PASS 2002

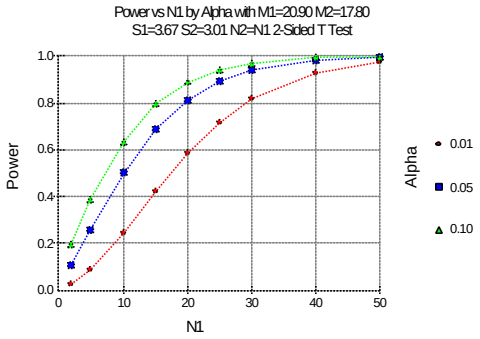
## Power Analysis and Sample Size Software from NCSS

***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

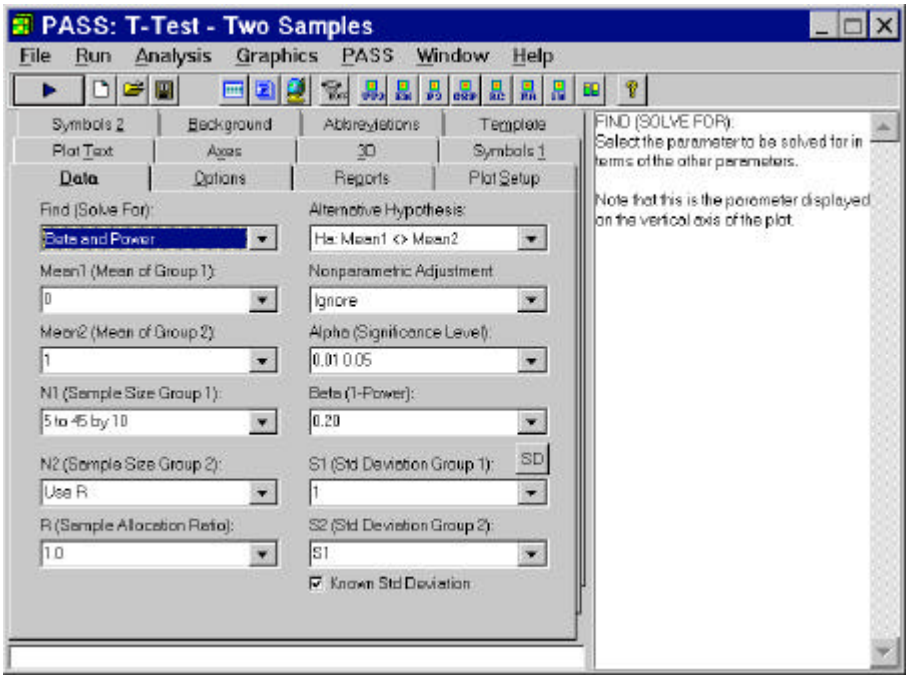
# *PASS 2002* adds power analysis and sample size to your statistical toolbox

**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.

## 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

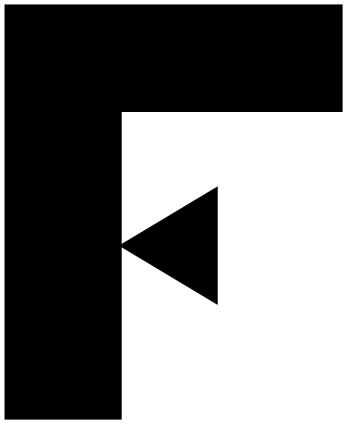
ADDRESS

CITY/STATE/ZIP

COUNTRY (IF OTHER THAN U.S.)

*"Perfection is achieved, not when there is nothing more to add, but when there is nothing left to take away."*

- Antoine de Saint Exupery



F is a carefully crafted subset of the most recent version of Fortran, the world's most powerful numeric language.

Using F has some very significant advantages:

- Programs written in F will compile with any Fortran compiler
- F is easier to use than other popular programming languages
- ***F compilers are free*** and available for Linux, Windows, and Solaris
- Several books on F are available
- F programs may be linked with C, Fortran 95, or older Fortran 77 programs

F retains the modern features of Fortran—modules and data abstraction, for example—but discards older error-prone facilities of Fortran.

It is a safe and portable programming language.

F encourages Module-Oriented Programming.

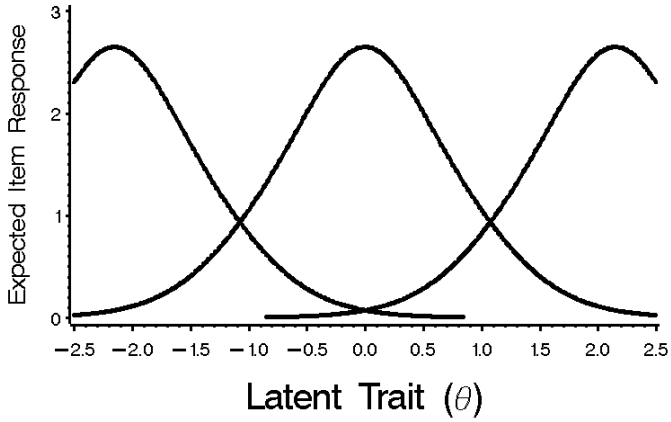
It is ideal for teaching a programming language in science, engineering, mathematics, and finance.

It is ideal for new numerically intensive programs.

The Fortran Company
11155 E. Mountain Gate Place, Tucson, AZ 85749 USA
+1-520-256-1455  +1-520-760-1397 (fax)
http://www.fortran.com  info@fortran.com

# *Introducing GGUM2004*

## *Item Response Theory Models for Unfolding*




The new GGUM2004 software system estimates parameters in a family of item response theory (IRT) models that unfold polytomous responses to questionnaire items. These models assume that persons and items can be jointly represented as locations on a latent unidimensional continuum. A single-peaked, nonmonotonic response function is the key feature that distinguishes unfolding IRT models from traditional, "cumulative" IRT models. This response function suggests that a higher item score is more likely to the extent that an individual is located close to a given item on the underlying continuum. Such single-peaked functions are appropriate in many situations including attitude measurement with Likert or Thurstone scales, and preference measurement with stimulus rating scales. This family of models can also be used to determine the locations of respondents in particular developmental processes that occur in stages.

The GGUM2004 system estimates item parameters using marginal maximum likelihood, and person parameters are estimated using an expected *a posteriori* (EAP) technique. The program allows for up to 100 items with 2-10 response categories per item, and up to 2000 respondents. GGUM2004 is compatible with computers running updated versions of Windows 98 SE, Windows 2000, and Windows XP. The software is accompanied by a detailed technical reference manual and a new Windows user's guide. ***GGUM2004 is free*** and can be downloaded from:

http://www.education.umd.edu/EDMS/tutorials

***GGUM2004 improves upon its predecessor (GGUM2000) in several important ways:***

- It has a user-friendly graphical interface for running commands and displaying output.
- It offers real-time graphics that characterize the performance of a given model.
- It provides new item fit indices with desirable statistical characteristics.
- It allows for missing item responses assuming the data are missing at random.
- It allows the number of response categories to vary across items.
- It estimates model parameters more quickly.

Start putting the power of unfolding IRT models to work in your attitude and preference measurement endeavors. Download your free copy of GGUM2004 today!

WWW.DIGARCH.COM

# DIGITAL EXCAVATOR

## Are you involved in Data Modeling or Data Mining?

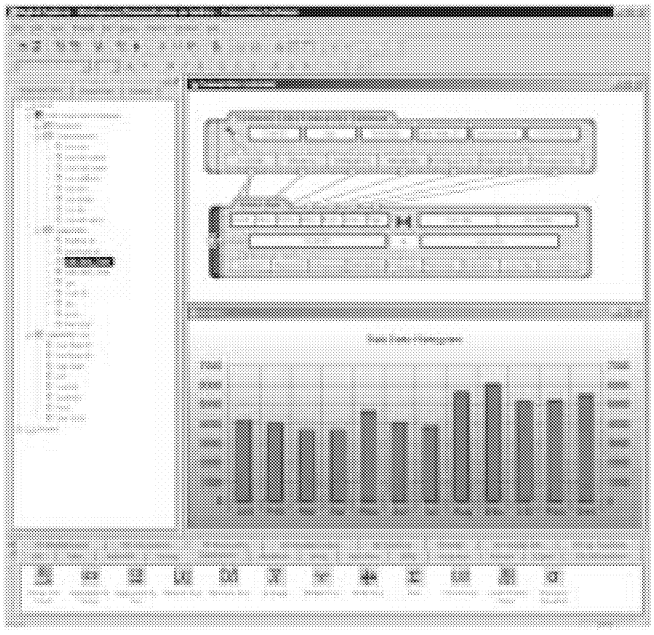
## Are you spending a large percentage of your time dealing with data issues?

If so, you will be happy to know that we have developed a tool that specifically addresses the data prep tasks associated with data modeling and data mining. The tool is called the Digital Excavator from Digital Archaeology (www.digarch.com). Data modelers are well aware of the time-consuming and sometimes frustrating nature of data set-up. In many cases data preparation can represent 60%-80% of the data mining project length. With Digital Archaeology's Digital Excavator, data preparation tasks are streamlined, results are more accurate, and the modeler has more time to focus on finding the appropriate mathematical solution--rather than wasting time with painful data issues. Digital Archaeology's software is intuitive, visual, self-documenting, and deploys what a number of analysts and customers have termed the "most elegant" user interface for data analysis and exploration ever conceived. It's the only tool specifically designed for the data prep tasks of data modeling.

## Visit our website and see for yourself! >>>> www.digarch.com

Functions have been created which perform the following:

- Frequency Distributions
- Categorical Variable Profile
- Continuous Variable Profile
- Histograms
- De-duping
- Find and Replace Missing Values
- Find and Split Out Outliers
- Binning
- Correlation Matrix
- Cross-Tabs
- Panel Variables (Occupancy Map)
- Lag functions
- Decimal Scaling
- Rank and Sample Variables
- Recency, Frequency, Monetary Analysis
- N-Tile Distributions
- Gains Charts
- Many others



DIGITAL ARCHAEOLOGY
A DELANO COMPANY

15721 COLLEGE BOULEVARD
LENEXA, KS 66219
1-877-DIGARCH (344-2724)
WWW.DIGARCH.COM

## 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
November, 2005, Vol. 4, No. 2, 353-628


1538 – 9472/04/$95.00

## 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