DigitalCommons @WayneState

Journal of Modern Applied Statistical Methods

Volume 14 | Issue 1 Article 1

5-1-2015

Vol. 14, No. 1 (Full Issue)

JMASM Editors

Recommended Citation

Editors, JMASM (2015) "Vol. 14, No. 1 (Full Issue)," *Journal of Modern Applied Statistical Methods*: Vol. 14 : Iss. 1 , Article 1.
DOI: 10.22237/jmasm/1430452800
Available at: http://digitalcommons.wayne.edu/jmasm/vol14/iss1/1

בס"ד

Journal of Modern Applied Statistical Methods
May 2015, Vol. 14, No. 1, ii-iv

ISSN 1538 − 9472

Journal of **Modern Applied Statistical Methods**

JMASM (ISSN 1538−9472, http://digitalcommons.wayne.edu/jmasm) is an independent, open access electronic journal, published biannually in May and November by JMASM Inc. (PO Box 48023, Oak Park, MI, 48237) in collaboration with the Wayne State University Library System. *JMASM* seeks to publish (1) new statistical tests or procedures, or the comparison of existing statistical tests or procedures, using computer-intensive Monte Carlo, bootstrap, jackknife, or resampling methods, (2) the study of nonparametric, robust, permutation, exact, and approximate randomization methods, and (3) applications of computer programming, preferably in Fortran (all other programming environments are welcome), related to statistical algorithms, pseudo- random number generators, simulation techniques, and self-contained executable code to carry out new or interesting statistical methods.

Journal correspondence (other than manuscript submissions) and requests for advertising may be forwarded to ea@jmasm.com. See back matter for instructions for authors.

ISSN 1538 − 9472

Journal of Modern Applied Statistical Methods

Vol. 14, No. 1

May 2015

Table of Contents

Invited Articles

Invited Debate

Regular Articles

ISSN 1538 − 9472

Statistical Software Applications and Review

Algorithms and Code

Journal of Modern Applied Statistical Methods
May 2015, Vol. 14, No. 1, 2-11.

ISSN 1538 − 9472

Invited Article
Estimating the Strength of an Association Based on a Robust Smoother

Rand Wilcox
University of Southern California
Los Angeles, CA

It is known that the more obvious parametric approaches to fitting a regression line to data are often not flexible enough to provide an adequate approximation of the true regression line. Many nonparametric regression estimators, often called smoothers, have been derived that are aimed at dealing with this problem. The paper deals with the issue of estimating the strength of an association based on the fit obtained by a robust smoother. A simple approach, already known, is to estimate explanatory power in a fairly obvious manner. This approach has been found to perform reasonably well when using the smoother LOESS. But when using a running interval, which provides a simple way of using any robust measure of location, the method performs poorly, even with a reasonably large sample size. The paper suggests an alternative estimation method that performs much better in simulations.

Keywords: Running interval smoother, explanatory power, cross-validation, Well Elderly 2 Study

Introduction

Consider a situation where the conditional measure of location of some random variable *Y*, given *X*, is given by

$$M\left(Y \mid X\right) = g\left(X\right) \tag{1}$$

where *g*(*X*) some unknown function. As is evident, a common strategy is to assume $g(X) = \beta_0 + \beta_1 X$, where β_0 and β_1 are unknown parameters that are typically estimated using ordinary least squares (OLS) regression with the goal of estimating the conditional mean of *Y* given *X*. There are, however, well known

Dr. Wilcox is Professor of Psychology at the University of Southern California. Email him at rwilcox@usc.edu.

concerns with this approach. First, it is often the case that assuming a straight regression line is unsatisfactory, which has led to the derivation of many nonparametric regression estimators, often called smoothers (e.g., Efromovich, 1999; Eubank, 1999; Fan & Gijbels, 1996; Fox, 2001; Green & Silverman, 1993; Gyöfri, et al., 2002; Härdle, 1990; Hastie & Tibshirani, 1990). Of course, some parametric model might be used to deal with any curvature, but often the more obvious strategies (e.g., include a quadratic term) are not flexible enough in terms of giving a reasonably accurate approximation of the true regression line.

Another concern with least squares regression, as well as the bulk of the smoothers that have been derived, is that they are designed to estimate the conditional mean of *Y*, one concern being that the population mean is not robust in the general sense summarized, for example, by Hampel et al., (1986), Huber and Ronchetti (2009), Staudte and Sheather (1990). (The population mean has an unbounded influence function and its breakdown point is zero.) A related concern is that even a single outlier can highly influence the sample mean, which in turn can give a distorted view of the typical value of *Y* given *X*. Cleveland (1979) derived a smoother (generally known as LOESS) aimed at estimating the conditional mean of *Y* and suggested how it might be modified to handle outliers among the dependent variable. Another robust approach is the running interval smoother in Wilcox (2012). It is more flexible than LOESS in the sense that virtually any robust measure of location can be used. For example, it is easily applied when the goal is to estimate the conditional median, trimmed mean or M-estimator of *Y*. It also can be used to estimate any quantile of interest.

A fundamental goal is estimating the strength of an association given a fit to data. An approach when using any smoother is to use some robust version of explanatory power (e.g., Wilcox, 2012). Explanatory power is

$$\xi^2 = \frac{\tau^2\left(\hat{Y}\right)}{\tau^2\left(Y\right)'}$$

where τ^2 is some measure of variation and $\hat{Y}$ is the predicted value of *Y* based on some fit to the data. The square root of explanatory power is called the explanatory strength of the association. To put ξ^2 in perspective, if $\hat{Y}$ is based on the OLS regression line and τ^2 is taken to be the usual variance, ξ^2 reduces to R^2, the usual coefficient of determination.

Estimating explanatory power would seem to be straightforward. Given a random sample (X_i, Y_i), $i = 1, \cdots n$, let $\hat{Y}_i$ be the predicted value of Y given that $X = X_i$. Let $\hat{\tau}^2\left(\hat{Y}\right)$ be an estimate of $\tau^2(\hat{Y})$ based on $\hat{Y}_1, \cdots, \hat{Y}_n$ and let $\hat{\tau}^2\left(Y\right)$ be an estimate of $\tau^2(Y)$ based on $Y_1, \cdots, Y_n$. The an estimate of explanatory power is simply

$$\hat{\xi}^2 = \frac{\hat{\tau}^2\left(\hat{Y}\right)}{\hat{\tau}^2\left(Y\right)} \qquad (2)$$

This approach seems to perform reasonably well when using LOESS, but when using the running interval smoother, it performs poorly: it can be severely biased (Wilcox, 2008). The goal in this paper is to suggest another estimation method that gives substantially better results.

The next section describes the details of the proposed estimation method. The following section reports simulation results comparing the new estimator to the estimator studied in Wilcox (2008). The final section illustrates the new method using data from the Well Elderly 2 study.

The Proposed Method

The measure of location used here is a 20% trimmed mean. For $Y_1, \cdots, Y_n$ the sample 20% trimmed mean is

$$\frac{1}{n-2g}\sum_{i=g+1}^{n-g} Y_{(i)}$$

where $g = .2n$ rounded down to the nearest integer and $Y_{(1)} \leq \cdots \leq Y_{(n)}$ are the values $Y_1, \cdots, Y_n$ written in ascending order. The 20% trimmed mean has nearly the same efficiency as the mean under normality, but it continues to have high efficiency, relative to the usual sample mean, when sampling from heavy-tailed distributions.

The measure of variation that is used is the 20% Winsorized variance. For $i = 1, \cdots, g$, let $W_i = Y_{(g+1)}$. For $i = g + 1, \cdots, n - g$, let $W_i = Y_{(i)}$ and for $i = n - g + 1, \cdots, n$ let $W_i = Y_{n-g}$. Then the Winsorized variance is just the usual sample variance based on the Winsorized values $W_1, \cdots, W_n$.

The running-interval smoother is applied as follows. For some constant f, declare x to be close to X_i if

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

where $MADN = MAD/.6745$, MAD is the median of the values.

$|X_1 - M|, \cdots, |X_n - M|$ and M is the usual sample median of the X_i values. Let $N(X_i) = \{ j{:}|X_j - X_i| \le f \times MADN \}$. That is, $N(X_i)$ indexes the set of all X_j values that are close to X_i. Then $M(Y \mid X_i)$ is taken to be some measure of location based on all Y_j values such that $j \in N\left(X_i\right)$ and here, a 20% trimmed mean is used. It appears that often a good choice for the span, f, is $f = 1$ (e.g., Wilcox, 2012) and this value is used here.

Method M1

Letting $\hat{Y}_i = M\left(Y \mid X_i\right)$ based on the running interval smoother just described, method M1 consists of simply computing (2) using the Winsorized variance.

Method M2

Method M2 differs from method M1 in two fundamental ways. First, $\hat{Y}_i$ is based on a leave-one-out cross-validation approach in conjunction with the running interval smoother. That is, $\hat{Y}_i$ in method M1 is replaced by $\breve{Y}_i = M\left(Y \mid X_i\right)$, which is based on $(X_1,Y_1), \cdots, (X_n,Y_n)$, ignoring the point (X_i,Y_i) rather than using all n points. For notational convenience, let T_i be the trimmed mean of $Y_1, \cdots, Y_n$, excluding Y_i. The other difference, compared to method M1, is that the estimate of explanatory power is replaced by

$$\breve{\xi}^2 = \frac{\tau^2\left(T_1,\cdots,T_n\right) - \tau^2\left(\breve{Y}_1,\cdots,\breve{Y}_n\right)}{\tau^2\left(T_1,\cdots,T_n\right)} \qquad (3)$$

Note that (3) mimics a standard way of writing the coefficient of determination. That is, it reflects the proportion of variation accounted for by the dependent variable and the fit obtained by the running interval smoother.

Simulation Results

Simulations were used to compare the bias and mean squared error of methods M1 and M2 when estimating ξ. For the first set of simulations data were generated from the model $Y = \frac{1}{\sqrt{3}} X + e$. The true value of ξ^2 was determined by noting that $\xi^2 = \tau_x^2 / \left(\tau_x^2 + \tau_e^2\right)$, in which case the explanatory strength of the association is $\xi = .5$. The sample size is taken to be 50. Both X and e were taken to have one of four g-and-h distributions, which contain the standard normal distribution as a special case. More precisely, if Z has a standard normal distribution, then

$$
\begin{aligned}
W &= \frac{\exp(gZ)-1}{g}\exp\left(h\frac{Z^2}{2}\right), \text{if } g > 0 \\
&= Z\exp\left(h\frac{Z^2}{2}\right), \text{ if } g = 0
\end{aligned}
$$

has a g-and-h distribution where g and h are parameters that determine the first four moments. The four distributions used here were the standard normal ($g = h = 0$), a symmetric heavy-tailed distribution ($h = 0.2$, $g = 0.0$), an asymmetric distribution with relatively light tails ($h = 0.0$, $g = 0.2$), and an asymmetric distribution with heavy tails ($g = h = 0.2$). Table 1 shows the skewness (κ_1) and kurtosis (κ_2) for each distribution. More properties of the g-and-h distribution are summarized by Hoaglin (1985).

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

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

Let $\hat{\xi}_1$ and $\hat{\xi}_2$ be the estimates of ξ based on methods M1 and M2, respectively. Bias was measured with $E\left(\hat{\xi}_j - \xi\right)$, $j = 1, 2$. To add perspective, bias also was measured with the median difference. The accuracy of the estimators was also measured with mean squared error, $E\left(\hat{\xi}_j - \xi\right)^2$, as well as the median squared error.

Table 2 shows the estimated bias when $n = 100$ and $Y = \beta X + e$ for three choices of the slope: 0, .5 and 1. As can be seen, generally M2 is less biased, and in various situations substantially so despite the reasonably large sample size. Note that the bias associated with M1 can be quite severe, the estimates being approximately −.2 in some cases.

Table 2. Estimated mean bias and median bias, $Y = \beta X + e$, $n = 100$

g	h	β	*mean bias*		*median bias*	
			M1	***M2***	***M1***	***M2***
0.0	0.0	0.0	.110	.081	.101	.000
0.0	0.2	0.0	.115	.078	.104	.000
0.2	0.0	0.0	.110	.085	.101	.000
0.2	0.2	0.0	.115	.082	.105	.000
0.0	0.0	0.5	-.140	-.099	-.139	-.065
0.0	0.2	0.5	-.178	-.072	-.178	-.035
0.2	0.0	0.5	-.144	-.108	-.142	-.070
0.2	0.2	0.5	-.179	-.081	-.138	-.045
0.0	0.0	1.0	-.132	-.074	-.129	-.057
0.0	0.2	1.0	-.197	-.059	-.197	-.039
0.2	0.0	1.0	-.139	-.077	-.134	-.057
0.2	0.2	1.0	-.201	-.064	-.200	-.047

Table 3 reports the estimated squared error. Method M2 does not dominate. But M1 never offers a striking advantage, while in some situations M2 is substantially better.

Tables 4 and 5 report the estimated bias and squared error loss when $Y = .5X^2 + e$. In terms of bias, the advantage of M2 over M1 is even more striking compared to the results in Table 2. Also, in terms of both the mean and median squared error, all indications are that M2 performs better than M1.

Table 3. Estimated mean squared error (MSE) and median squared error (MEDSE), *Y* = *βX* + *e*, *n* = 100

g	h	β	*MSE*		*MEDSE*	
			M1	***M2***	***M1***	***M2***
0.0	0.0	0.0	.016	.021	.010	.000
0.0	0.2	0.0	.017	.021	.011	.000
0.2	0.0	0.0	.016	.023	.010	.000
0.2	0.2	0.0	.018	.022	.011	.000
0.0	0.0	0.5	.019	.044	.009	.011
0.0	0.2	0.5	.030	.038	.018	.011
0.2	0.0	0.5	.020	.048	.010	.012
0.2	0.2	0.5	.031	.040	.019	.011
0.0	0.0	0.7	.024	.018	.017	.005
0.0	0.2	0.7	.047	.017	.039	.004
0.2	0.0	0.7	.026	.019	.018	.005
0.2	0.2	0.7	.049	.019	.040	.005

Table 4. Estimated mean bias and median bias, *Y* = .5X^2 + *e*, *n* = 100

g	h	*mean bias*		*median bias*	
		M1	***M2***	***M1***	***M2***
0.0	0.0	-.201	-.085	-.208	-.050
0.0	0.2	-.182	-.015	-.191	.025
0.2	0.0	-.203	-.067	-.210	-.036
0.2	0.2	-.182	.004	-.190	.043

Table 5. Estimated mean squared error (MSE) and median squared error (MEDSE), *Y* = .5X^2 + *e*, *n* = 100

g	h	*MSE*		*MEDSE*	
		M1	***M2***	***M1***	***M2***
0.0	0.0	.045	.037	.043	.013
0.0	0.2	.039	.036	.036	.025
0.2	0.0	.046	.036	.044	.012
0.2	0.2	.040	.035	.036	.016

An Illustration

The Well Elderly 2 study (Clark et al., 2012; Jackson et al., 2009) was generally concerned with assessing the efficacy of an intervention strategy aimed at improving the physical and emotional health of older adults. One goal was to determine the association between the cortisol awakening response (CAR) and a measure of depressive symptoms after intervention. CAR is defined to be the change in cortisol concentration that occurs during the first hour after waking from sleep. Extant studies (e.g., Clow et al., 2004; Chida & Steptoe, 2009) indicate that various forms of stress are associated with the CAR.

Simply using Pearson's correlation yields $r = .07$, which is not significant at the .05 level when using Student's t test ($p = .22$). There are outliers suggesting the use of some robust generalization of Pearson's correlation. The skipped correlation in Wilcox (2012, section 9.4.3) is estimated to be .07. Kendall's tau and Spearman's rho are .038 and .057, respectively. So all of these correlation coefficients fail to detect any association and suggest that any association that might exist is relatively weak. However, a test of the hypothesis that the regression line is straight (using the method in Wilcox, 2012, section 11.6.1) is significant ($p < .001$). Based on method M1, the strength of the association is estimated to be .12 compared to .31 using method M2.

Concluding Remarks

It is not being suggested that better-known correlation coefficients should be abandoned in favor of method M2. If, for example, a correct parametric model has been specified, under normality Pearson's correlation provides a more accurate estimate of the true association in terms of both bias and mean squared error. A difficulty is that no single estimator dominates and the optimal estimator depends in part on the true nature of the association, which of course is unknown. If, for example, a smoother suggests that the regression line is reasonably straight, and if outliers do not appear to be a serious issue, Pearson's correlation seems reasonable. But it can be difficult determining whether some specified parametric model is sufficiently accurate to justify using something other than method M2. In the illustration, for example, the hypothesis of a straight line was rejected. But even if this hypothesis is not rejected, there is the issue of whether the test of the hypothesis that the regression line is straight has enough power to justify assuming a straight line when estimating the strength of the association. Strategies for deciding which estimator to use, or how to resolve any discrepancies among the estimators that are used, are in need of further study.

The running interval smoother can be used when there are two or more independent variables. A few simulations were run with two independent variables yielding results similar to those reported in Tables 2 and 3. But a more extensive investigation is in order.

References

Chida, Y. & Steptoe, A. (2009). Cortisol awakening response and psychosocial factors: A systematic review and meta-analysis. *Biological Psychology, 80*, 265-278.

Clark, F., Jackson, J., Carlson, M., Chou, C.-P., Cherry, B. J., Jordan-Marsh, M., … Azen, S. P. (2012). Effectiveness of a lifestyle intervention in promoting the well-being of independently living older people: results of the Well Elderly 2 Randomised Controlled Trial. *Journal of Epidemiology and Community Health, 66,* 782-790. doi:10.1136/jech.2009.099754

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association, 74*, 829-836.

Clow, A., Thorn, L., Evans, P. & Hucklebridge, F. (2004). The awakening cortisol response: Methodological issues and significance. *Stress, 7*, 29-37.

Efromovich, S. (1999). *Nonparametric curve estimation: Methods theory and applications*. New York: Springer-Verlag.

Eubank, R. L. (1999). *Nonparametric regression and spline smoothing*. New York: Marcel Dekker.

Fan, J. & Gijbels, I. (1996). *Local polynomial modeling and its applications*. Boca Raton, FL: CRC Press.

Fox, J. (2001). *Multiple and generalized nonparametric regression*. Thousand Oaks, CA: Sage.

Green, P. J. & Silverman, B. W. (1993). *Nonparametric regression and generalized linear models: A roughness penalty approach*. Boca Raton, FL: CRC Press.

Györfi, L., Kohler, M., Krzyzk, A. & Walk, H. (2002). *A distribution-free theory of nonparametric regression*. New York: Springer Verlag.

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

Hardle, W. (1990). Applied nonparametric regression. *Econometric Society Monographs No. 19*, Cambridge, UK: Cambridge University Press

Hastie, T. J. & Tibshirani, R. J. (1990). *Generalized additive models*. New York: Chapman and Hall

Hoaglin, D. C. (1985). Summarizing shape numerically: The g-and-h distributions. In D. Hoaglin, F. Mosteller and J. Tukey (Eds.) *Exploring Data Tables, Trends, and Shapes.* (pp. 461-515). New York: Wiley.

Huber P and Ronchetti E. M. (2009). *Robust statistics.* 2nd Ed. New York: Wiley.

Jackson, J., Mandel, D., Blanchard, J., Carlson, M., Cherry, B., Azen, S., … Clark, F. (2009). Confronting challenges in intervention research with ethnically diverse older adults: the USC Well Elderly II trial. *Clinical Trials, 6*, 90-101.

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

Wilcox, R. R. (2008). Estimating explanatory power in a simple regression model via smoothers. *Journal of Modern Applied Statistical Methods*, *7*(2), 368-375. http://digitalcommons.wayne.edu/jmasm/vol7/iss2/2/

Wilcox, R. R. (2012). *Introduction to robust estimation and hypothesis testing*. New York: Elsevier.

Journal of Modern Applied Statistical Methods
May 2015, Vol. 14, No. 1, 12-23.

ISSN 1538 − 9472

Invited Debate Are Per-Family Type I Error Rates Relevant in Social and Behavioral Science?

Andrew V. Frane
University of California Los Angeles
Los Angeles, CA

The familywise Type I error rate is a familiar concept in hypothesis testing, whereas the per‑family Type I error rate is rarely addressed. This article uses Monte Carlo simulations and graphics to make a case for the relevance of the per‑family Type I error rate in research practice and pedagogy.

Keywords: Type I error, multiple comparisons, simultaneous inference

Introduction

The familywise Type I error rate (FWER; Tukey, 1953), which is the probability of making at least one Type I error in a family of hypotheses, is a familiar concept in quantitative research. Much less frequently addressed is the per-family Type I error rate (PFER; Tukey, 1953), which is the number of Type I errors expected to occur in a family of hypotheses (in other words, the sum of probabilities of Type I error for all the hypotheses in the family). The unpopularity of the PFER may stem largely from the fact that it is a stricter standard than the FWER, so controlling it can be more costly in statistical power (potentially increasing the Type II error rate). Given the tremendous pressure on researchers to find statistically significant p-values, any reduction in statistical power is a hard sell. However, as noted by a previous article in this journal (Barnette & McLean, 2005) and by others (Klockars & Hancock, 1994; Ryan, 1959, 1962), it is arguable that the PFER is often more relevant than the FWER in social and behavioral science research. The argument is essentially as follows: Committing multiple Type I errors simultaneously is worse than committing only one, yet unlike the PFER, the FWER does not distinguish between making one Type I

Mr. Frane is a doctoral student of cognitive psychology in the Visual and Multisensory Perception Lab. Email him at avfrane@ucla.edu.

error in a family and making several Type I errors in a family. Moreover, one might reason that because both the maximum FWER and the maximum PFER are equal to α when there is only one comparison, both error rates should remain less than or equal to α when there are multiple comparisons if Type I error is to be considered uninflated.

Readers may debate the comparative merits of the FWER and the PFER. The goal of this article is not to definitively advocate for one standard over the other, but rather to point out that although both error rates have merits, the PFER is almost universally ignored and may deserve more attention. For example, in statistics textbooks for the social and behavioral sciences, there is generally no mention of the PFER even when the FWER is addressed (e.g., Goodwin, 2010; Hinton, 2004; Howell, 2014; Mertler & Vannatta, 2010; Meyers, Gamst, & Guarino, 2006; Sirkin, 2006; Stevens, 2009; Tabachnick & Fidell, 2012; Wetcher-Hendricks, 2011). And although some classic texts on simultaneous inference discuss the PFER (e.g., Hochberg & Tamhane, 1987; Miller, 1966; Tukey, 1953), many newer books on the subject do not (e.g., Dickhaus, 2014; Dmitrienko et al., 2010; Hsu, 1996).

This study briefly describes some popular Type I error rate controlling procedures, distinguishing PFER control from FWER control. Then examples from the applied statistics literature are used to show how widespread disregard of the PFER may be causing confusion. Then Monte Carlo simulations are used to demonstrate that in multivariate contexts the PFER can be substantially inflated even when the FWER is controlled, particularly when outcome variables are correlated.

Controlling the PFER using the Bonferroni procedure

The Bonferroni procedure caps the maximum PFER at α by testing each hypothesis at a nominal alpha level of α / m, where m is the number of hypotheses in the family. With rare exception (e.g., Harris, 2001), textbooks tend not to mention that the Bonferroni procedure controls the PFER, and instead recommend it only as a method for controlling the FWER. It is true that the Bonferroni procedure controls the FWER (as does any method that controls the PFER), but using a PFER controlling method to control the FWER prompts two questions: (1) If the objective is to control the PFER, then why not say so, and (2) if the objective is to control the FWER, then why not use a procedure that is more optimized for that purpose? After all, several methods for controlling the FWER are more powerful (meaning they can produce significance in more comparisons) than the Bonferroni procedure. Among the most popular of these methods are stepwise procedures, such as the Holm and Hochberg procedures, which are described in the following section.

Controlling the FWER using stepwise procedures

Holm's (1979) procedure first arranges the *m* hypotheses from lowest to highest *p*-value. Then the hypotheses are tested sequentially in that order, each at a nominal alpha level of $\alpha / (m - b + 1)$, where *b* is a number between 1 and *m* indicating the position of the given hypothesis in the sequence. Thus, the first hypothesis is tested at level α / m, the next at $\alpha / (m - 1)$, the next at $\alpha / (m - 2)$, and so on until the last hypothesis is tested at level α. Testing is conditional, meaning that if any *p*-value in the sequence is nonsignificant, then all larger *p*-values are also declared nonsignificant and testing stops. Holm's method controls the FWER, is more powerful than the Bonferroni procedure, and requires only slightly more computation. Like the Bonferroni procedure, Holm's method also allows computation of confidence intervals (Strassburger & Bretz, 2008; Guilbaud, 2008).

Hochberg's (1988) procedure is essentially the reverse of Holm's: The hypotheses are arranged from highest to lowest *p*-value, then tested sequentially in that order, each at a nominal alpha level of α / b, where *b* is a number between 1 and *m* indicating the position of the given hypothesis in the sequence. Thus, the first hypothesis is tested at level α, the second at $\alpha / 2$, the third at $\alpha / 3$, and so on until the last hypothesis is tested at level α / m. If any *p*-value in the sequence is significant, then all smaller *p*-values are also declared significant and testing stops. Hochberg's procedure controls the FWER (except in certain situations; see Dmitrienko et al., 2010) and is more powerful than Holm's, but generally does not allow computation of confidence intervals (Dmitrienko et al., 2010; Guilbaud, 2012).

Some other stepwise procedures for controlling the FWER are more powerful than Hochberg's (e.g., Hommel, 1988; Rom, 1990), but they are more computationally complex and, like Hochberg's method, generally do not allow computation of confidence intervals (Dmitrienko et al., 2010; Guilbaud, 2012). There are also methods that control the FWER in specific contexts. For example, Dunnett's (1955) procedure and its variations (see Dmitrienko et al., 2010) can be used when comparing multiple treatment groups to a placebo group. There are also Šidák based methods (see Bird & Hadzi-Pavlovic, 2013), which are not necessarily applicable to one sided tests.

Given the variety of multiple comparisons procedures available, the simplicity and versatility of the Bonferroni procedure—which works for any *p*-values regardless of how they were obtained—make the Bonferroni procedure useful to teach as a default method of Type I error control (Harris, 2001). However, it is important to note that the Bonferroni procedure controls not only the FWER but also the PFER. Failing to understand this may lead to confusion such as that discussed in the following section.

Confusion about the utility of the Bonferroni procedure

The Bonferroni procedure is often described as "overly conservative" (as noted by Gordon, Glazko, & Yakovlev, 2007), or as being "improved" through modifications such as Holm's and Hochberg's (see Dickhaus, 2014; Posch & Futschik, 2008; Simes, 1986). This framing is legitimate if the goal is to control the FWER. However, if the goal is to control the PFER, then the Bonferroni procedure is not overly conservative (and hence is not improved by modifications that make it more liberal). Thus, the Bonferroni procedure is perhaps better depicted not as a "blunt tool (Miles & Banyard, 2007, p. 263)" for controlling the FWER—but rather as a precise and efficient tool for controlling the PFER.

Psychological researchers that have touted the superior power of stepwise methods over the Bonferroni procedure (e.g., Blakesley et al., 2009; Eichstaedt, Kovatch, and Maroof, 2013; Seaman, Levin, & Serlin, 1991) have rarely mentioned that such methods—though useful—do not control the PFER and therefore are not adequate substitutes for the Bonferroni procedure when control of the PFER is desired. For example, Eichstaedt and colleagues (2013, p. 693) explicitly stated, "The Holm's sequential procedure corrects for Type I error as effectively as the traditional Bonferroni method"—which is only true if the PFER is not considered (see Barnette & McLean, 2005). In fact, the sometimes dramatically inflated PFERs associated with stepwise procedures are so widely unknown among researchers that Klockars and Hancock (1994) were moved to call inflated PFERs "the hidden costs" of stepwise procedures.

In summary, lack of acknowledgment for the PFER may be causing unnecessary controversy and confusion: Some present the Bonferroni procedure as an appropriate method for controlling the FWER; others present the Bonferroni procedure as underpowered and obsolete; and neither of these opposing views takes into account the procedure's usefulness for controlling the PFER. However, if the Bonferroni procedure were presented as a method for controlling the PFER, then there would be no dissonance between: (1) recommending the Bonferroni procedure for controlling the PFER, and (2) recommending more powerful methods for controlling the FWER.

The PFER may be more relevant now than in the past

There was a time when choosing between the FWER and the PFER appeared to be relatively inconsequential. Miller (1966, p. 10) called the choice "essentially a matter of taste," and acknowledged that he preferred the FWER "for feelings he [could not] entirely analyze." Similarly, Tukey (1953, p. 5) wrote that either error rate could be used in practice and that the FWER merely had "theoretical advantages". Ryan (1959, p. 40) called the choice between FWER and PFER "merely a matter of computational convenience." Indeed, the Bonferroni procedure's maximum FWER is known to be only trivially different from its maximum PFER. However, selecting an error rate is no longer simply an inconsequential matter of personal preference, given the development of procedures—such as the Holm, Hochberg, and Hommel methods—that can control the FWER while allowing considerable inflation of the PFER. The following simulations demonstrate this inflation in multivariate designs (for demonstrations of analogous PFER inflation in other contexts, see Barnette & McLean, 2005; Klockars & Hancock, 1994; Shaffer, Kowalchuk, & Keselman, 2013).

Methodology

Monte Carlo simulations were conducted in R (R Core Team, 2013) of two-group designs with 50 subjects per group. Three numbers of multivariate normal outcome variables were used: $m = 2$, $m = 5$, and $m = 10$. Equal population correlations (ρ) between outcome variables were set at 200 values between 0 and 1. All effect sizes (i.e., population mean differences) were set at zero so that any statistically significant sample mean difference between groups would be a Type I error. There were 100,000 simulations for each combination of m and ρ. These simulations generated pseudorandom sample mean differences and sample covariance matrices.

Two sided univariate tests of the sample mean differences were conducted at $\alpha = .05$ using each of the following four procedures: Bonferroni, Holm, Hochberg, and Hommel. For each of these procedures at each combination of m and ρ, the FWER was computed by dividing the number of simulations in which significance occurred by 100,000, and the PFER was computed by dividing the number of significant tests by 100,000.

Results

At each value of m, each of the four procedures had a maximum FWER less than .050, but the PFER could differ notably from the FWER when outcome variables were correlated. For example, Figure 1B shows that for five outcome variables, even a moderate correlation of .6 inflated the Hommel procedure's PFER to approximately 0.067. In other words, although the chance of making a Type I error in a given family remained less than one in 20, the rate of Type I errors per family was approximately one in 15. The stepwise procedures can allow even greater PFER inflation at higher values of m and ρ, but the Bonferroni procedure's maximum PFER is always equal to α and is insensitive to correlation.

Note that in Figures 1B and 1C, the maximum PFERs of the Hochberg and Hommel procedures are well beyond the upper limits of the graphs. At any value of m, the maximum PFER for both procedures approaches $\alpha \times m$ as ρ goes to 1. However, extending the range of the vertical axes to accommodate the extremely inflated PFERs at impractically high correlations would have sacrificed detail in the busier portions of the graphs while adding little useful information.

Discussion

Previous studies (Barnette & McLean, 2005; Klockars & Hancock, 1994; Shaffer, Kowalchuk, & Keselman, 2013) showed that the PFER can be substantially inflated in multigroup designs even when the FWER is controlled. This article has built on those findings in three principal ways: (1) by demonstrating through simulation that those findings extend to multivariate designs, (2) by graphically illustrating how the population correlation between outcome variables can enhance the disparity between the PFER and the FWER, and (3) by using the applied statistics literature to show that inadequate acknowledgement of the PFER may be causing unnecessary controversy and confusion, particularly with regard to the utility of the Bonferroni procedure.

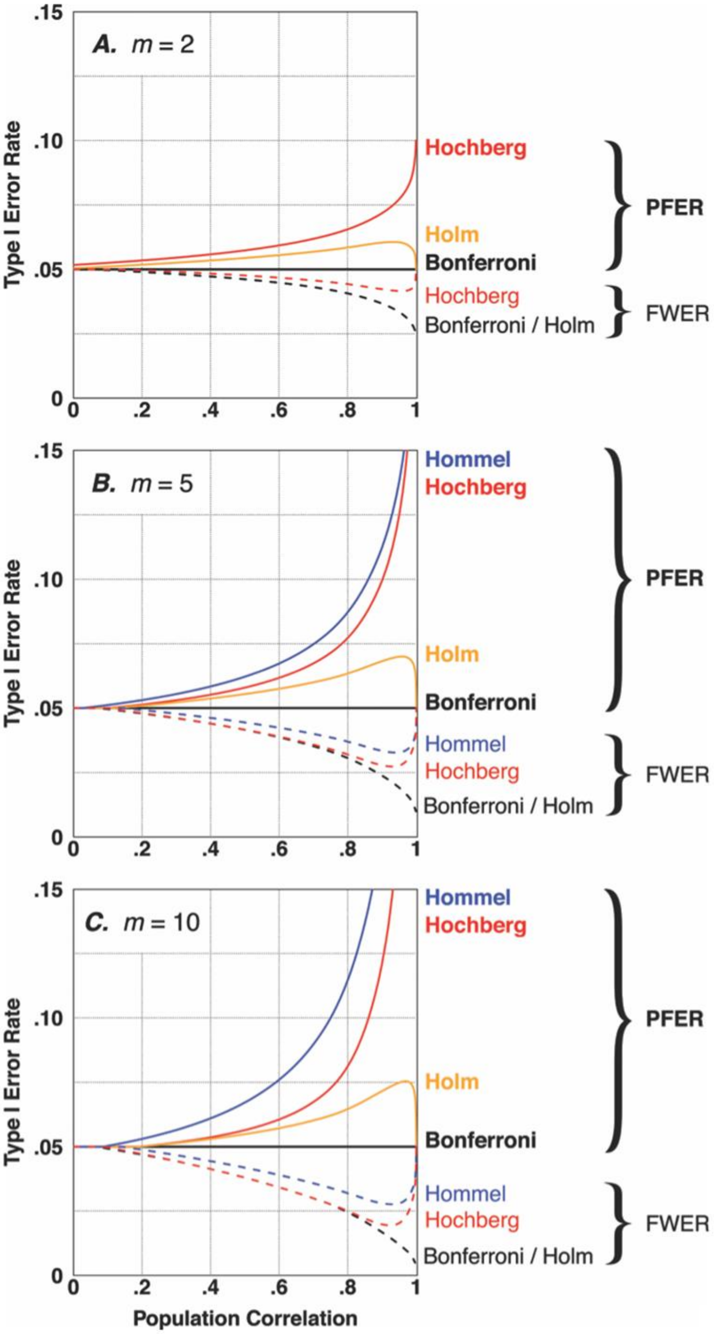


Figure 1. Per-family and familywise Type I error rates for the Bonferroni, Holm, Hochberg, and Hommel procedures in a two-group design with *m* outcome variables (*α* = .05, all null hypotheses true). Note that Hommel is equivalent to Hochberg for m = 2.

Implications for research practice

This article proposes that, depending on the research situation, either the PFER or the FWER may be more relevant than the other. Controlling the PFER (i.e., using the Bonferroni procedure) is appropriate when every mistake hurts—as is frequently the case in social and behavioral science research. For example, if a psychological therapy is found to significantly improve multiple symptoms, then it would be worse for many of those purported improvements to be Type I errors than for only one to be a Type I error. If statistical power is of concern, then improving the measures and manipulations or increasing the sample size would be a better solution than using a more liberal error rate that increases the toleration of false findings.

Controlling the FWER may be sufficient when, given one Type I error, additional Type I errors are not costly, or perhaps when dependency among the tests is known to be sufficiently low that FWER and PFER are only negligibly different. In such situations, a method more powerful than the Bonferroni procedure may be used, such as the Holm procedure (if confidence intervals are required), the Hochberg or Hommel procedure (if no confidence intervals are required), or a context specific method appropriate for the given situation (see Dmitrienko et al., 2010 for an extensive list). An important caveat is that the Hochberg and Hommel procedures do not necessarily control the FWER for one sided tests that can be negatively correlated (see Samuel-Cahn, 1996), whereas the Bonferroni and Holm methods do not have this limitation.

Implications for applied statistics pedagogy

If the PFER is to be addressed more in practice, then it must also be addressed more in pedagogy. Therefore, this article recommends that professors and textbook authors include discussion of the PFER along with discussion of the FWER. Additionally, when a multiple comparisons procedure is described, the specific error rates that it controls (and does not control) should be accurately identified. It is no longer sufficient to simply refer to “the Type I error rate.”

Limitations

This study did not examine every Type I error rate that has been defined. For example, the comparisonwise Type I error rate (Tukey, 1953) is the probability of Type I error for a single hypothesis irrespective of the number of hypotheses in the family. Thus, controlling Type I error at the comparisonwise level effectively means disregarding Type I error inflation altogether and simply conducting each hypothesis test at the unadjusted alpha level. Another error rate that has been proposed is the false discovery rate (Benjamini & Hochberg, 1995), which is, loosely speaking, the expected proportion of significant hypothesis tests in the family that are Type I errors (except when all null hypotheses are true, in which case the false discovery rate is equivalent to the FWER). Both the comparisonwise Type I error rate and the false discovery rate are more liberal than the FWER and thus beyond the scope of this article, but there are contexts in which these error rates may be appropriate.

It should also be acknowledged that the simulations examined neither a variety of alpha levels, nor an exhaustive variety of multiple comparisons procedures, nor an exhaustive variety of parameter combinations. However, to do so would have made exceedingly long and complex an article that required only a finite number of examples to support its conclusion that the PFER can be relevant. Future articles may examine in detail issues such as which Type I error rates are more relevant in particular contexts.

References

Barnette, J. J., & McLean, J. E. (2005). Type I error of four pairwise mean comparison procedures conducted as protected and unprotected tests. *Journal of Modern Applied Statistical Methods, 4*(2), 446-459. Available at http://digitalcommons.wayne.edu/jmasm/vol4/iss2/10/

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B (Methodological), 57*(1), 289-300. Retrieved from http://www.jstor.org/stable/2346101

Bird, K. D., & Hadzi-Pavlovic, D. (2013). Controlling the maximum familywise Type I error rate in analyses of multivariate experiments. *Psychological Methods*. Advance online publication. doi:10.1037/a0033806

Blakesley, R. E., Mazumdar, S., Dew, M. A., Houck, P. R., Tang, G., Reynolds III, C. F., & Butters, M. A. (2009). Comparisons of methods for multiple hypothesis testing in neuropsychological research. *Neuropsychology, 23*(2), 255-264. doi:10.1037/a0012850

Dickhaus, T. (2014). *Simultaneous statistical inference*. Berlin, Germany: Springer.

Dmitrienko, A., Bretz, F., Westfall, P. H., Troendle, J., Wiens, B. L., Tamhane, A. C., & Hsu, J. C. (2010). Multiple testing methodology. In A.

Dmitrienko, A. C. Tamhane, & F. Bretz (Eds.), *Multiple testing problems in pharmaceutical statistics* (pp. 35-98). Boca Raton, FL: Chapman & Hall.

Dunnett, C. W. (1955). A multiple comparison procedure for comparing several treatments with a control. *Journal of the American Statistical Association, 50*(272), 1096-1121. doi:10.1080/01621459.1955.10501294

Eichstaedt, K. E., Kovatch, K., & Maroof, D. A. (2013). A less conservative method to adjust for familywise error rate in neuropsychological research: The Holm's sequential Bonferroni procedure. *NeuroRehabilitation, 32*(3), 693-696. doi:10.3233/NRE-130893

Goodwin, C. J. (2010). *Research in psychology: Methods and design* (6th ed.). Hoboken, NJ: John Wiley & Sons.

Gordon, A., Glazko, G., Qiu, X. & Yakovlev, A. (2007). Control of the mean number of false discoveries, Bonferroni and stability of multiple testing. *The Annals of Applied Statistics*, *1*(1), 179-190. doi:10.1214/07-AOAS102

Guilbaud, O. (2008). Simultaneous confidence regions corresponding to Holm's stepdown procedure and other closed-testing procedures. *Biometrical Journal, 50*(5), 678-692. doi:10.1002/bimj.200710449

Guilbaud, O. (2012). Simultaneous confidence regions for closed tests, including Holm-, Hochberg-, and Hommel-related procedures. *Biometrical Journal, 54*(3), 317-342. doi:10.1002/bimj.201100123

Harris, R. J. (2001). *A primer of multivariate statistics* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Hinton, P. R. (2004). *Statistics explained* (2nd ed.). New York, NY: Routledge.

Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika, 75*(4), 800-802. doi:10.1093/biomet/75.4.800

Hochberg, Y., & Tamhane, A. C. (1987). *Multiple comparison procedures*. New York, NY: John Wiley & Sons.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandanavian Journal of Statistics, 6*, 65-70.

Hommel, G. (1988). A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika, 75*(2), 383-386. doi:10.1093/biomet/75.2.383

Howell, D. C. (2014). *Fundamental statistics for the behavioral sciences* (8th ed.). Belmont, CA: Wadsworth.

Hsu, J. C. (1996). *Multiple comparisons*. Boca Raton, FL: Chapman & Hall.

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. doi:10.1177/0013164494054002004

Mertler, C. A., & Vannatta, R. A. (2010). *Advanced and multivariate statistical methods* (4th ed.). Glendale, CA: Pyrczak.

Meyers, L. S., Gamst, G., & Guarino, A. J. (2006). *Applied multivariate research*. Thousand Oaks, CA: Sage.

Miles, J., & Banyard, P. (2007). *Understanding and using statistics in psychology*. London, England: Sage.

Miller, R. G., Jr. (1966). *Simultaneous statistical inference*. New York, NY: McGraw-Hill.

Posch, M., & Futschik, A. (2012). A uniform improvement of Bonferroni-type tests by sequential tests. *Journal of the American Statistical Association, 103*(481), 299-308. doi:10.1198/016214508000000012

R Core Team. (2013). R (Version 3.0.2) [Computer software]. Vienna, Austria: R Foundation for Statistical Computing.

Rom, D. M. (1990). A sequentially rejective test procedure based on a modified Bonferroni inequality. *Biometrika*, *77*(3), 663-665. doi:10.1093/biomet/77.3.663

Ryan, T. A. (1959). Multiple comparisons in psychological research. *Psychological Bulletin*, *56*(1), 26-47. doi:10.1037/h0042478

Ryan, T. A. (1962). The experiment as the unit for computing rates of error. *Psychological Bulletin*, *59*(4), 301-305. doi:10.1037/h0040562

Samuel-Cahn, E. (1996). Is the Simes improved Bonferroni procedure conservative? *Biometrika*, *83*(4), 928-933. doi:10.1093/biomet/83.4.928

Seaman, M. A., Levin, J. R., & Serlin, R. C. (1991). New developments in pairwise multiple comparisons: Some powerful and practicable procedures. *Psychological Bulletin*, *110*(3), 577-586. doi:10.1037/0033-2909.110.3.577

Shaffer, J. P., Kowalchuk, R. K., & Keselman, H. J. (2013). Error, power, and cluster separation rates of pairwise multiple testing procedures. *Psychological Methods, 18*(3), 352-367. doi:10.1037/a0032478

Simes, R. J. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika*, *73*(3), 751-754. doi:10.1093/biomet/73.3.751

Sirkin, R. M. (2006). *Statistics for the social sciences* (2nd ed.). Thousand Oaks, CA: Sage.

Stevens, J. P. (2009). *Applied multivariate statistics for the social sciences* (5th ed.). New York, NY: Taylor & Francis.

Strassburger, K., & Bretz, F. (2008). Compatible simultaneous lower confidence bounds for the Holm procedure and other Bonferroni based closed tests. *Statistics in Medicine, 27*(24), 4914-4927. doi:10.1002/sim.3338

Tabachnick, B. G., & Fidell, L. S. (2012). *Using multivariate statistics* (6th ed.). Boston, MA: Pearson.

Tukey, J. W. (1953). The problem of multiple comparisons. In H. Braun (Ed.), *The collected works of John W. Tukey volume VIII, multiple comparisons: 1948-1983* (pp. 1-300). New York, NY: Chapman & Hall.

Wetcher-Hendricks, D. (2011). *Analyzing quantitative data*. Hoboken, NJ: John Wiley & Sons.

Journal of Modern Applied Statistical Methods
May 2015, Vol. 14, No. 1, 24-37.

ISSN 1538 − 9472

Invited Debate
Per Family or Familywise Type I Error Control: “Eether, Eyether, Neether, Nyther, Let's Call the Whole Thing Off!”[1]

H. J. Keselman
University of Manitoba
Winnipeg, Manitoba

Frane (2015) pointed out the difference between per-family and familywise Type I error control and how different multiple comparison procedures control one method but not necessarily the other. He then went on to demonstrate in the context of a two group multivariate design containing different numbers of dependent variables and correlations between variables how the per-family rate inflates beyond the level of significance. In this article I reintroduce other newer better methods of Type I error control. These newer methods provide more power to detect effects than the per-family and familywise techniques of control yet maintain the overall rate of Type I error at a chosen level of significance. In particular, I discuss the False Discovery Rate due to Benjamini and Hochberg (1995) and k-Familywise Type I error control enumerated by Lehmann and Romano (2005), Romano and Shaikh (2006), and Sarkar (2008). I conclude the article by referring readers to articles by Keselman, et al. (2011, 2012) which presented R computer code for determining critical significance levels for these newer methods of Type I error control.

Keywords: Type I error, multiple comparisons, simultaneous inference

Introduction

Frane (2015) presented an article which clarified the difference between the per-family (PFER) and familywise (FWER) Type I error rates (See also Klockars & Hancock, 1994). It is important that applied researchers understand the difference between the rates and how different multiple comparison procedures may control

[1] From the film “Shall We Dance?” Words by Ira Gershwin; music by George Gershwin. Introduced by Fred Astaire and Ginger Rogers.

H. J. Keselman is a Professor of Psychology and Associate Editor Emeritus of this journal. Email him at kesel@ad.umanitoba.ca.

one rate of error but not the other. For example, as he notes, the typical Dunn (1961)-Bonferroni method controls the overall rate of Type I error per-family, whereas other Bonferroni methods of Type I error control (e.g., Holm, 1979) control the familywise rate of error. Through simulation methods he then shows that in a multivariate design containing two groups, multiple dependent measures, and various correlations between the dependent variables, the FWER may be controlled, yet the PFER can be very large. The author also notes in the article that other issues could have been discussed such as newer methods of controlling Type I errors and other multiple comparison procedures themselves; some issues were noted but not discussed in detail.

My intention in this article is to take the reader further into the topics of Type I error control and multiple comparison procedures that Frane (2015) did not have the space to discuss. I believe these additional topics are very important to discuss since the issue of Type I error control has advanced immeasurably since the early discussions related to PFER and FWER control.

Per-experiment and experimentwise Type I error control

At the outset I want to expand on the definitions of per-family and familywise presented by Frane (2015). But first, I want to re-introduce the per-experiment (PEER) and the experimentwise (EWER) Type I error rates, rates applied researchers are more likely to be familiar with. Ryan (1959, 1960, 1962) in his seminal articles regarding overall Type I error control versus comparisonwise (CWE) (i.e., per test or per comparison) control, used the terminology per-experiment and experimentwise to indicate that these rates applied to controlling the maximum overall rate of Type I error for multiple tests of significance assessed within an experiment. Later in the history of methods for controlling the overall rate of Type I error, per-family and familywise became equated with per-experiment and experimentwise (See Hochberg & Tamhane, 1987).

The distinction is important because it allows one to adopt per-family and familywise control in more interesting and dynamic ways. For example, in a one-way design where a researcher computes pairwise and complex comparisons between group means, one can set a per-family or familywise error rate over each family of tests (i.e., the pairwise tests and complex comparisons tests), and thus maintain the per-experiment or experimentwise rates at some overall maximum value. So a .05 level of significance can be tied to each family of tests and consequently the maximum overall joint per-experiment or experimentwise probability of Type I error can be fixed at .10. To further illustrate the nuances of familywise and experimentwise control consider an A × B design. In such a design a researcher can set familywise rates of error over all tests performed on the A effect, B effect, and A × B effects. Collectively, the overall or experimentwise Type I error rate would be a function of the three familywise rates. For example, suppose the researcher chose to perform all possible pairwise comparisons on the A main effect, a number of complex comparisons on the B main effect, and a number of interaction contrasts on the A × B effects setting a .05 value on each set. Collectively therefore, the overall experimentwise Type I error rate would be controlled at the .15 level. Clearly by thinking about the familywise or per-family rate as rates for related families of tests, the researcher can see the flexibility that s/he is afforded. I will have more to say on how researchers should define a family shortly.

Newer definitions of Type I error control

Background

Multiplicity of testing. The multiplicity problem in statistical inference refers to selecting the statistically significant findings from a large set of findings (tests) to either support or refute one's research hypotheses. Discussions on how to deal with multiplicity of testing have permeated many literatures for decades. There are those who believe that the occurrence of any false positive must be guarded at all costs (see Games, 1971; Ryan, 1960, 1962; Westfall & Young, 1993). That is, as promulgated by Thomas Ryan, pursuing a false lead can result in the waste of much time and expense, and is an error of inference that accordingly should be stringently controlled. Those in this camp deal with the multiplicity issue by setting α for the entire set of tests computed. This type of control has been referred to in the literature as experimentwise (EWER) or familywise (FWER) control. Those in the opposing camp maintain that stringent Type I error control results in a loss of statistical power and consequently important treatment effects go undetected (see Rothman, 1990; Saville, 1990). Members of this camp typically believe the error rate should be set per comparison [the probability of rejecting a given comparison] (the CWE rate) and usually recommend a five percent level of significance, allowing the overall error rate (i.e., EWER or FWER) to inflate with the number of tests computed. In effect, those who adopt comparisonwise control ignore the multiplicity issue.

Family size. Specifying family size is a very important component of multiple testing. As Westfall et al. (1999, p. 10) note, differences in conclusions reached from statistical analyses that control for multiplicity of testing (FWER) and those that do not (CWE) are directly related to family size. Specifically, the larger the family size, the less likely individual tests will be found to be statistically significant with FWER control. Accordingly, to achieve as much sensitivity as possible to detect true differences and yet maintain control over multiplicity effects, Westfall et al. recommend that researchers "choose smaller, more focused families rather than broad ones, and (to avoid cheating) that such determination must be made *a priori*..." (p. 10).

Not only does the FWER rate depend on the number of null hypotheses that are true but as well on the distributional characteristics of the data and the correlations among the test statistics. Because of this, an assortment of multiple comparison procedures have been developed, each intended to provide FWER control.

As I indicated at the outset, since the per-family/per-experiment and familywise/experimentwise error rates were introduced, researchers have defined new ways of controlling Type I errors which by-in-large are intended to provide control over multiple tests of significance that one does not achieve with comparisonwise control and more power to detect effects than is provided by the familywise and experimentwise rates.

The false discovery rate (FDR)

It was noted by Frane (2015) that this is a new definition of Type I error control that affords the user more power to detect true effects though at the cost of allowing a greater number of Type I errors. However, Frane believes that if researchers want more power they should exert better experimental control and/or use more subjects in their studies. Presuming that applied researchers are always attuned to controlling extraneous variance and accordingly adopt the best experimental control that is feasible for their studies, the remaining avenue to increase power to detect effects is to increase the number of participants examined in their studies. Not always however, possible. In my department the subject pool is limited and experimenters do not have access to as many subjects that comprise the pool. Thus, achieving more statistical power through more liberal definitions of Type I error control and more sensitive multiple comparison procedures should be a viable option for researchers to consider.

As indicated, several different error rates have been proposed in the multiple comparison literature. The majority of discussion in the literature has focused on the FWER, although other error rates, such as the FDR also have been proposed (e.g., Benjamini & Hochberg, 1995). The FDR is defined by these authors as the expected proportion of the number of erroneous rejections to the total number of rejections.

Use of the false discovery rate criterion has become widespread when making inferences in research involving the human genome, where family sizes in the thousands are common. See the review by Dudoit, Shaffer and Boldrick (2003), and references contained therein. Another area of research where FDR controlling procedures have had a significant impact is functional magnetic resonance imaging. In these experiments researchers are conducting numerous (often more than 100,000) significance tests that relate to tests of activation on specific voxels (i.e., areas) within the brain (e.g., Callan, Jones, Munhall, Callan, Kroos, & Vatikiotis-Bateson, 2003).

The Benjamini and Hochberg (1995) procedure has been shown to control the FWER for several situations of dependent tests, that is, for a wide variety of multivariate distributions that make their procedure applicable to most testing situations scientists might encounter (see Sarkar, 1998; Sarkar & Chang, 1997). In addition, simulation studies comparing the power of the Benjamini and Hochberg procedure to several FWER controlling procedures have shown that as the number of treatment groups increases (beyond 4 treatment groups), the power advantage of their procedure over the FWER controlling procedures becomes increasingly large (Keselman et al., 1999). The power of FWER controlling procedures is highly dependent on the family size (i.e., number of comparisons), decreasing rapidly with larger families (Holland & Cheung, 2002; Miller, 1981). Therefore, control of the FDR results in more power than FWER controlling procedures in experiments with many treatment groups, but yet provides more control over Type I errors than CWE controlling procedures.

Suppose for n means, $\mu_1, \mu_2, \ldots, \mu_J$, and our interest is in testing the family of $m = [J(J-1)]/2$ pairwise hypotheses, $H_0 : \mu_i - \mu_j = 0$, of which m_0 are true. Let S equal the number of correctly rejected hypotheses from the set of R rejections; the number of falsely rejected pairs will be V. In terms of the random variable V, the CWE is $E(V / m)$, while the FWER is given by $P(V \geq 1)$. Thus, testing each and every comparison at α guarantees that $E(V / m) \leq \alpha$, while according to the Bonferroni inequality, testing each and every comparison at level α / m guarantees that $P(V \geq 1) \leq \alpha$.

According to Benjamini and Hochberg (1995) the proportion of errors committed by falsely rejecting null hypotheses can be expressed through the random variable $Q = V / R$, that is, the proportion of rejected hypotheses that are erroneously rejected. (It is important to note that Q is defined to be zero when $R = 0$; that is, the error rate is zero when there are no rejections.) The FDR was defined by Benjamini and Hochberg as the mean of Q, that is

$$E(Q) = E\left(\frac{V}{R}\right), \text{ or } E(Q) = E\left(\frac{\text{Number of false rejections}}{\text{Number of rejections}}\right).$$

That is, the FDR is the expected proportion of false discoveries or false positives.

As Benjamini and Hochberg (1995) indicate, this error rate has a number of important properties:

a) If $\mu_1 = \mu_2 = \cdots = \mu_J$, then all m (pairwise) comparisons truly equal zero, and therefore the FDR is equivalent to the FWER; that is, in the case of the complete null being true, FDR control implies FWER control. Specifically, in the case of the complete null hypothesis being true, $S = 0$ and therefore $V = R$. So, if $V = 0$, then $Q = 0$, and if $V > 0$ then $Q = 1$ and accordingly $P(V \geq 1) = E(Q)$.

b) In testing the family of (pairwise) hypotheses, of which m_0 are true, when $m_0 < m$, the FDR is smaller than or equal to the FWER. The FDR is smaller than or equal to the FWER because in this case $\text{FWER} = P(R \geq 1) \geq E(V / R) = E(Q)$. This indicates that if the FWER is controlled for a procedure, then the FDR is as well. Moreover, if one adopts a procedure that provides FDR control, rather than strong (i.e., over all possible mean configurations) FWER control, then based on the preceding relationship, a gain in power can be expected.

c) V / R tends to be smaller when there are fewer pairs of equal means and when the non-equal pairs are more divergent, resulting in a greater differences in the FDR and the FWER values and thus a greater likelihood of increased power by adopting FDR control.

With the BH FDR procedure, the p-values corresponding to the m (pairwise) statistics for testing the hypotheses $H_1, H_2, \ldots, H_m$ are ordered from smallest to largest, that is, $p_1 \leq p_2 \leq \cdots \leq p_m$. Let k be the largest value of i for which $p_i \leq (i / m)\alpha$ and then reject all H_i, $i = 1, 2, \ldots, k$. On the basis of this procedure, one begins by assessing the largest p-value, p_m, and then proceeds to smaller p-values as long as $p_i > (i / m)\alpha$. Testing stops when $p_i \leq (k / m)\alpha$.

The *k*-FWER criterion and procedures for its control[2]

The classical approach for controlling Type I errors for a family of many (say m) hypothesis tests is FWER control. Once the family is defined, control of the FWER requires that

$$\text{FWER} \leq \alpha$$

for all configurations of true and false hypotheses. It is well known that for non-independent tests the probability (Pr) of making one or more Type I errors is

$$\text{FWER} = \Pr(\text{One or more Type I errors for } m \text{ tests}) < 1 - (1 - \alpha)^m$$

Examples of procedures that control the overall rate of Type I error when many tests of hypotheses are examined are the single-stage Bonferroni procedures (e.g., Dunn, 1961) and stepwise Bonferroni procedures (Hochberg, 1988; Holm, 1979). However, when there are many hypotheses to be examined they can be deficient in power to detect non-null hypotheses. Indeed, when the size of the family of hypotheses to be tested becomes large, FWER becomes very restrictive and not very powerful at detecting false null hypotheses. For example, for m tests of significance, the single-stage Bonferroni level of significance would be α / m and when m is large detecting non-null effects will be difficult. As Lehmann & Romano (2005) note "control of the FWER at conventional levels becomes so stringent that individual departures from the hypothesis have little chance of being detected" (p. 1139).

Accordingly, Type I error control is not the only issue researchers must consider when testing a hypothesis or set of hypotheses. As in the case of testing a single hypothesis, researchers must also consider the ability of a procedure to detect departures from the hypothesis when they do occur (Lehmann & Romano, 2005, p. 1139). To address this issue, Lehmann & Romano, as well as others (See the references cited in Lehmann & Romano) developed the *k*-FWER method of

[2] Keselman et al. (2012) previously introduced these procedures to the psychological audience. Their article also includes the mathematical underpinnings of the procedures.

Type I error control. As they note, with a larger family of hypotheses, one might be willing to allow the possibility of falsely rejecting k true null hypotheses. With the possibility of falsely rejecting more than one, two, three, etc. null hypothesis(es), one obtains more power to detect false null hypotheses. Lehmann and Romano (2005) define k-FWER as the probability of rejecting at least k true null hypotheses.

$$k\text{-FWER} = \Pr\{\text{reject at least } k \text{ hypotheses } H_i \text{ with } i \in I(P)\}$$

Here $I(P)$ denotes the set of true null hypotheses when P is the true probability distribution. Control of the k-FWER requires that $k\text{-FWER} \leq \alpha$ for all P. When $k = 1$, then k-FWER reduces to 1-FWER or FWER which controls the probability of rejecting at least one true null hypothesis.

To help the reader to fully appreciate k-FWER, I note the following. Consider what it means to control 2-FWER instead of 1-FWER (or simply FWER) at $\alpha = .05$? This would be equivalent to specifying that the probability of 2 or more false rejections is controlled at .05, whereas FWER controls the probability of any (i.e., 1 or more) false rejections at .05. In essence, then, 2-FWER implicitly tolerates 1 false rejection and makes no explicit attempt to control the probability of its occurrence, unlike FWER which tolerates no false rejections at all. More generally, then, k-FWER tolerates $k - 1$ false rejections, but controls the probability of k or more false rejections at an $\alpha = .05$.

Before presenting these newer methods I provide some additional clarification of the k-FWER. First, remember that FWER control treats rejections of multiple true null hypotheses as being no more serious than the rejection of only one (i.e., at least one) true null hypothesis. The newer procedures have the same conceptual underpinning; however, for them falsely rejecting multiple true null hypotheses is no more serious than the rejection of only two, three, etc. true null hypotheses (i.e., at least 2, 3, etc.). Accordingly, a clean outcome from an analysis controlling the FWER is an outcome with no Type I errors. A clean outcome from a k-FWER analysis is an outcome with no more than $k - 1$ Type I errors. Note that in both cases, the number of Type I errors produced when at least k are produced (1 in the case of FWER) is of no concern as far as the error rate criterion is concerned.

Keselman, Miller and Holland (2011) describe four procedures that utilize the k-FWER method of multiple testing control. Technical descriptions can be found in Keselman et al. (2011). As well these authors provide R code for running the newer procedures (See also Keselman et al., 2012).[3]

The Holm and generalized Holm (Lehmann and Romano) procedures

Lehmann and Romano (2005) provided a generalization of the Holm (1979) procedure. Just as the Holm procedure controls FWER under all dependency conditions, the generalized procedure controls *k*-FWER under the same dependency conditions (i.e., there are no dependency conditions).

The ordered *p*-values for the *m* individual tests denoted $p_{(1)} \leq \cdots \leq p_{(k)} \leq \cdots \leq p_{(m)}$ correspond to hypotheses, $H_{(1)}, \ldots, H_{(k)}, \ldots, H_{(m)}$. The generalized Holm procedure is defined stepwise as follows:

Step 0. Let $i = 1$, k and α are chosen by the experimenter.

Step 1. If $i \leq k$, go to step 2. If $k < i \leq m$, go to step 3. Otherwise, stop and reject all of the hypotheses.

Step 2. If $p_{(i)} > \dfrac{k\alpha}{m}$, go to step 4. Otherwise, set $i = i + 1$ and go to step 1.

Step 3. If $p_{(i)} > \dfrac{k\alpha}{m+k-i}$, go to step 4. Otherwise, set $i = i + 1$ and go to step 1.

Step 4. Reject $H_{(j)}$ for $j < i$ and accept $H_{(j)}$ for $j \geq i$.

The Hochberg and generalized Hochberg (Sarkar 1) procedures

The generalization of the Hochberg (1988) procedure is a step up version of the generalized Holm procedure presented by Lehmann and Romano. Sarkar (2008) states that it controls *k*-FWER when the test statistics are independent or when they satisfy the multivariate totally positive order of two (MTP_2) condition.[4]

A step up procedure based on the same set of critical values as a step down procedure will always reject at least as many hypotheses and therefore will be

[3] The R code provides users with adjusted *p*-values. In its typical application, researchers compare a test statistic to a FWER critical value. Another approach for assessing statistical significance is with adjusted *p*-values, $\tilde{p}_i$, $i = 1, \ldots, m$ (Westfall et al., 1999; Westfall & Young, 1993). As Westfall and Young note "$\tilde{p}_i$ is the smallest significance level for which one still rejects a given hypothesis (H_i) in a family, given a particular (familywise) controlling procedure." (p. 11) The advantage of adjusted *p*-values for multiple comparison procedures, as with *p*-values for tests in comparisonwise contexts, is that they are more informative than merely declaring retain or reject H_i; they are a measure of the weight of evidence for or against the null hypothesis when controlling FWER. For example, if $\tilde{p}_i = 0.09$, the researcher/reader can conclude that the test is statistically significant at the FWER = 0.10 level, but not at the FWER = 0.05 level. Adjusted *p*-values are provided by the SAS system for many popular multiple comparison procedures (See Westfall et al., 1999). SPSS also provides adjusted *p*-values for most multiple comparison procedures.

[4] Keselman et al. (2012) define MTP_2 in their article.

more powerful at detecting false null hypotheses. I therefore recommend using the generalized Hochberg procedure over the generalized Holm procedure as long as the Hochberg procedure is appropriate to use.

The generalized Hochberg procedure is defined stepwise as follows:

Step 0. Let $i = m$, k and α are chosen by the experimenter.

Step 1. If $i > k$, go to step 2. If $1 \leq i \leq k$, go to step 3. Otherwise, stop and accept all of the hypotheses.

Step 2. If $p_{(i)} \leq \dfrac{k\alpha}{m+k-i}$, go to step 4. Otherwise, set $i = i - 1$ and go to step 1.

Step 3. If $p_{(i)} \leq \dfrac{k\alpha}{m}$, go to step 4. Otherwise, set $i = i - 1$ and go to step 1.

Step 4. Reject $H_{(j)}$ for $j \leq i$ and accept $H_{(j)}$ for $j > i$.

Romano and Shaikh procedure Romano and Shaikh (2006) developed a generalized version of the Hochberg procedure that has no dependency restrictions associated with it. This fact makes it attractive in situations with complex dependency conditions, i.e., such as when the family of tests are that the elements of a correlation matrix are zero. Step up tests such as the Hochberg are more powerful at detecting false null hypotheses than the step down test using the same critical values. However, since this generalized Hochberg test is valid to use under all dependency conditions, it does not use the same critical values as the generalized Holm procedure. The critical values are approximately halved. This negatively affects power to detect false null hypotheses since the *p*-values must be less than the critical values to be declared statistically significant. See Keselman et al.'s (2011) Appendix A for more information.

Sarkar 2 procedure The Sarkar (2008) procedure is another generalized version of the Hochberg procedure. It controls *k*-FWER when the joint distribution of the *p*-values is multivariate totally positive of order two (MTP_2) in addition to having identical k^{th}-order joint distributions under the null hypotheses. MTP_2 is a somewhat restrictive condition that is violated if any of the test statistics are negatively correlated, but met if the tests are pairwise independent (Sarkar, 2000). An example of a MTP_2 procedure would be many to one contrasts in a balanced design as is found in a Dunnett's one-sided comparisons with a control.

When the p-values are independent, this procedure has been found to be a more powerful generalized Hochberg procedure than a step up version of the generalized Holm procedure when $2 \leq k \leq 1 / \alpha$ (Sarkar, 2008). When $k = 1$, the Sarkar procedure is equivalent to the Hochberg procedure. Although, the Sarkar procedure is valid to use as long as the p-values have a MTP_2 distribution, we only recommend its use when the p-values are independent [See Keselman et al.'s (2011) Table 1 for a description of k-FWER method and type of dependency assumed to exist between the test statistics and associated p-values]. (Note: The R code provided in their Appendix B is only valid for the Sarkar procedure when the p-values are independent.)

Discussion

As the reader can see, the way in which Type I errors can be controlled for families of tests goes way beyond the per-family and familywise rates discussed by Frane (2015). The intention of my article was to review methods previously presented in the statistical and psychological literatures, with the intention of letting the reader see that researchers have many techniques that can be adopted to control the overall rate of Type I error. I recommend that applied researchers give serious consideration to the newer techniques (FDR and k-FWER) because they provide more power to detect non-null effects and yet limit the overall rate of Type I error at some specified value. So referring back to the title of this article I would say with regard to per-family or familywise control—eether, eyether, or perhaps neether, nyther.[5] The reader should note that the R code provided in Keselman et al. (2011, 2012) provides adjusted p-values for all of the newer methods discussed in this article. Users must select a method of control before cherry-picking the method that has the greatest number of statistically significant findings as reported through the R code.

[5] The methods described in this paper do not provide confidence intervals as compared to simultaneous MCPs [procedures that use one critical value to assess statistical significance such as Tukey's (1953) method]; they, nonetheless, should be considered an important tool in any data analyst's arsenal of viable methods for investigating treatment effects through many tests of significance.

References

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, 57*(1), 289-300. doi:10.2307/2346101

Callan, D. E., Jones, J. A., Munhall, K., Callan, A. M., Kroos, C., & Vatikiotis-Bateson, E. (2003). Neural processes underlying perceptual enhancement by visual speech gestures. *Neuroreport*, *14*(17), 2213-2218.

Dunn, O. J. (1961). Multiple comparisons among means. *Journal of the American Statistical Association*, *56*(293), 52-64. doi:10.1080/01621459.1961.10482090

Dudoit, S., Shaffer, J. P. & Boldrick, J. C. (2003). Multiple hypothesis testing in microarray experiments. *Statistical Science*, *18*(1), 71-103. doi:10.1214/ss/1056397487

Frane, A. V. (2015). Are per-family Type I error rates relevant in social and behavioral science? *Journal of Modern Applied Statistical Methods, 14*(1).

Games, P. A. (1971). Multiple comparisons of means. *American Educational Research Journal*, *8,* 531-565. doi:10.3102/00028312008003531

Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika, 75*(4), 800-802. doi:10.1093/biomet/75.4.800

Hochberg, Y., & Tamhane, A. C. (1987). *Multiple comparison procedures*. New York, NY: John Wiley & Sons.

Holland, B. & Cheung, S. H. (2002). Familywise robustness criteria for multiple comparison procedures. *Journal of the Royal Statistical Society*, *B, 64*(1), 63-77. doi:10.1111/1467-9868.00325

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics, 6*, 65-70.

Keselman, H. J., Cribbie, R., & Holland, B. (1999). The pairwise multiple comparison multiplicity problem: An alternative approach to familywise/comparisonwise Type I error control. *Psychological Methods*, *4*(1), 58-69. doi:10.1037/1082-989X.4.1.58

Keselman, H. J., Miller, C. E., & Holland, B. (2011). Many tests of significance: New methods for controlling Type I errors. *Psychological Methods*, *16*(4), 420-431. doi:10.1037/a0025810

Keselman, H. J., Miller, C. E., & Holland, B. (2012). Correction to many tests of significance: New methods for controlling Type I errors. *Psychological Methods*, 17(4), 679. doi:10.1037/a0030995

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. doi:10.1177/0013164494054002004

Lehmann, E. L., & Romano, J. P. (2005). Generalizations of the familywise error rate. *Annals of Statistics*, *33*(3), 1138–1154. doi:10.1214/009053605000000084

Miller, R. G. (1981). *Simultaneous statistical inference.* (2nd ed.) New York: McGraw-Hill.

Romano, J. P., & Shaikh, A. M. (2006). Stepup procedures for control of generalizations of the familywise error rate. *Annals of Statistics*, *34,* 1850–1873. doi:10.1214/009053606000000461

Rothman, K. (1990). No adjustments are needed for multiple comparisons. *Epidemiology*, *1*(1), 43-46.

Ryan, T. A. (1959). Multiple comparisons in psychological research. *Psychological Bulletin*, *56*(1), 26-47. doi:10.1037/h0042478

Ryan, T. A. (1960). Significance tests for multiple comparison proportions, variances and other statistics. *Psychological Bulletin*, *57*(4), 318-328. doi:10.1037/h0044320

Ryan, T. A. (1962). The experiment as the unit for computing rates of error. *Psychological Bulletin*, *59*(4), 301-305. doi:10.1037/h0040562

Sarkar, S. K. (1998). Some probability inequalities for ordered MTP2 random variables: A proof of the Simes conjecture. *Annals of Statistics*, *26*(2), 494–504. doi:10.1214/aos/1028144846

Sarkar, S. K., (2000). A note on the monotonicity of the critical values of a step–up test. *Journal of Statistical Planning Information, 87*(2), 241-249. doi:10.1016/S0378-3758(99)00200-1

Sarkar, S. K. (2008). Generalizing Simes' test and Hochberg's stepup procedure. *Annals of Statistics*, *36*(1), 337–363. doi:10.1214/009053607000000550

Sarkar, S. K., & Chang, C. K. (1997). The Simes method for multiple hypothesis testing with positively dependent test statistics. *Journal of the American Statistical Association*, *92*(440), 1601–1608.

Saville, D. J. (1990). Multiple comparison procedures: The practical solution. *The American Statistician*, *44*(2), 174-180. doi:10.1080/00031305.1990.10475712

Tukey, J. W. (1953). The problem of multiple comparisons. In H. Braun (Ed.), *The collected works of John W. Tukey volume VIII, multiple comparisons: 1948-1983* (pp. 1-300). New York, NY: Chapman & Hall.

Westfall, P. H., Tobias, R. D., Rom, D., Wolfinger, R. D., & Hochberg, Y. (1999). *Multiple comparisons and multiple tests*. Cary, NC: SAS Institute, Inc.

Westfall, P. H., & Young, S. S. (1993). *Resampling-based multiple testing: Examples and methods for p-value adjustment.* New York: Wiley.

Journal of Modern Applied Statistical Methods
May 2015, Vol. 14, No. 1, 38-42.

ISSN 1538 − 9472

Invited Debate
Per Family Error Rates: A Response

James F. Troendle
Nat'l. Heart, Lung, & Blood Inst.
Bethesda, MD

Keshia-Lee Martin
American University
Washington, DC

Vance W. Berger
National Cancer Institute
Rockville, MD

As the authors note, the familywise error rate (FWER) is used rather often, whereas the per-family error rate (PFER) is not. Is this as it should be? It would seem that no universal answer is possible, as context determines which is more appropriate in any given application. In the general scenario of testing the benefit of an intervention, one might ideally want an error rate that aligns with the decision for benefit. In most cases the FWER does this pretty well, while allowing one to identify those endpoints for which benefit exists. The PFER does not seem to have any advantage over the FWER in this general testing scenario. Perhaps in some other scenarios the PFER might have some reasonable role.

Keywords: Familywise error rate, per-family error rate

Introduction

As Berger (2004) notes, the alpha level should be selected strategically, based on the ramifications of committing a Type I error relative to a Type II error. The entire testing framework becomes more complicated when dealing with multiple hypothesis tests, and in this case various circumstances must be taken into account. Apart from choosing the proper alpha level for the specific situation, one must also define (prospectively) what constitutes a win (so to speak). Is it enough to find statistical significance on any one endpoint? Or do we instead combine the results in some way to obtain an overall finding?

The familywise Type-I error rate (FWER) is the probability of at least one Type I error in a family of hypotheses occurring, and is used rather often. The

Dr. Troendle is a Mathematical Statistician at the Office of Biostatistics Research of NHLBI. Email him at jt3t@nih.gov. Keshia-Lee Martin is a 2014 alumni of the Department of Statistics of American University. Dr. Berger is a Mathematical Statistician at the NCI, an Adjunct Professor at University of Maryland Baltimore County, and an Assistant Editor of this journal. Email him at vb78c@nih.gov.

per-family Type I error rate (PFER) is the sum of probabilities of Type I errors in the family for all hypotheses, and is almost never used in practice (Frane, 2015).

When performing multiple hypothesis tests, various circumstances must be taken into account. Apart from choosing the proper alpha level for the specific situation (preferably strategically, rather than based on the one size fits all precedent of 0.05), there is a risk that a Type I (false positive) or Type II (false negative) error may occur. The familywise Type-I error rate (FWER), the probability of at least one Type I error in a family of hypotheses occurring, is used rather often. Meanwhile, the per-family Type I error rate (PFER), the sum of probabilities of Type I errors in the family for all hypotheses, is almost completely ignored (Frane, 2015). Does the PFER deserve as much attention as the FWER receives? We do not attempt any general answer to this question, but, instead, focus on one specific application. For the commonly encountered scenario of testing the benefit of an intervention with several possible endpoints, we think there is a good reason why PFER is not used.

As the author (Frane, 2015) states, committing numerous Type I errors simultaneously is worse than committing only one, with FWER unable to differentiate between creating one Type I error and multiple Type I errors in a family of hypotheses. We suggest that the choice between controlling the FWER or the PFER should be based on the specific situation. The FWER works well for the commonly encountered scenario of testing an intervention with several possible endpoints of interest. The PFER does not appear to have any advantage over the FWER in this scenario, but perhaps in some other scenarios it might. The purpose of this response is not to determine which error rate is superior to the other, but how to establish which error rate should be controlled based on a testing situation. We first consider the scenario of testing an intervention for benefit due to any of several endpoints and then discuss the choice of alpha level.

Tests of an intervention with multiple endpoints of interest

Consider a study designed to test whether an intervention or exposure is beneficial or detrimental to patient health, compared to some comparison condition. Suppose that benefit can be measured by using any of several endpoints. This is quite a general scenario, which applies equally to clinical trials as well as to behavioral intervention studies or in fact to many observational studies. In this case, it is easy to see that control of the FWER is sufficient to guarantee that if any endpoint is identified as significant, and if biases can be suitably removed by the study design, then either any such endpoint is truly affected by the intervention or an unlikely event has occurred. This is also true if the PFER is controlled. However, control of the PFER is more restrictive (less powerful) than control of the FWER. Thus, there is no reason to prefer the PFER to the FWER in this general scenario.

An interesting observation about this scenario is that control of the FWER is not necessary to guarantee the type of concordance desired. One might consider testing an intersection hypothesis whose rejection corresponds with evidence of an intervention benefit. To make this clearer, suppose that there are two endpoints, and let H_1 (H_2) be the null hypothesis that the first (second) endpoint is unaffected by the intervention. If one would recommend the intervention if either endpoint is beneficial, then one really wants to claim benefit if either H_1 or H_2 are false. This argues for testing the intersection null hypothesis $H_0 = H_1 \cap H_2$. Rejection of this null hypothesis corresponds to benefit. This approach circumvents multiple comparison altogether as only a single hypothesis is tested.

The downside to this approach is that rejection of H_0 leaves one unable to conclude improvement on any specific endpoint. As Durkalski and Berger (2009) note, success on a composite endpoint leaves one "unable to determine which outcome is driving the claim". The other caveat to this approach is one must decide how to test H_0, which in general could be difficult. An adaptive testing approach could prove useful (Berger and Ivanova, 2002), but the usual solution for testing H_0 involves rejecting if $\min(p_1, p_2) \leq \alpha/2$, where p_1 (p_2) is the p-value for testing H_1 (H_2). With this solution, one is once again controlling the FWER, although in general such an approach could lead to more powerful testing procedures. This observation is a major reason why FWER is the predominantly used error rate for publications of confirmatory findings for studies that test an intervention. Bloch et al. (2001) describe one way of testing a single null hypothesis, although rejecting their null also allows one to conclude non-inferiority on all endpoints.

Choosing an alpha level

Returning now to the strategic selection of the alpha level, we note that cancer therapy often involves both high risk and high reward. The promise of meaningful improvement is counterbalanced by the almost certain toxicity of the treatment which, in some cases, may have the potential to do more harm than good. That said, false positives and false negatives can both result in grave consequences, including illnesses left untreated, illnesses over-treated, and ultimately higher mortality rates for patients. So the calculation has to consider the relative harm likely caused by each type of error.

As one extreme example (following Berger, 2004), one may conduct a trial to determine if broccoli will prevent arthritis. If broccoli is found, rightfully or wrongfully, to prevent arthritis, then the result would simply be increased consumption of broccoli. Since broccoli is known to have other health benefits, and few (if any) drawbacks, this will still lead to substantial health benefits, regardless if it helps to treat the symptoms of arthritis. So here, a Type I error would not result in very much harm at all. Alpha can be set to a much larger level than the usual 0.05. Another example is Glucosamine and Chondroitin. Like broccoli, these substances have no known side effects and are known to be generally good for cartilage health. Despite no strong evidence of a benefit for sufferers of osteoarthritis pain, many people take Glucosamine and Chondroitin because of the low risk involved coupled with some possible benefit. Conversely, if an aggressive and highly toxic cancer treatment is found to be beneficial, then its increased use will incur additional costs and also result in toxicity, so the benefit should offset this risk, and we should be fairly certain that it does (Berger, 2004). A Type I error in this case would result in severe consequences, so alpha should be small, 0.05 or perhaps even 0.01. These are simple examples, but the concept is that alpha should be carefully considered, and not just set at the usual level of 0.05 as a matter of course (Berger & Hsieh, 2005).

References

Berger, V. W. (2004). On the generation and ownership of alpha in medical studies. *Controlled Clinical Trials*, *25*(6), 613-619. doi:10.1016/j.cct.2004.07.006

Berger, V. W., Hsieh, G. (2005). Rethinking statistics: basing efficacy alpha levels on safety data in randomized trials. *Israeli Journal of Emergency Medicine, 5*(3), 55-60. http://isrjem.org/IJEM_Aug_AlphaLevels_Proof.pdf

Berger, V. W., Ivanova, A. (2002). Adaptive tests for ordered categorical data. *Journal of Modern Applied Statistical Methods, 1*(2), 269-280. http://digitalcommons.wayne.edu/jmasm/vol1/iss2/36/

Bloch, D. A., Lai, T. L., Tubert-Bitter, P. (2001). One-sided tests in clinical trials with multiple endpoints. *Biometrics, 57*(4), 1039-1047. doi:10.1111/j.0006-341X.2001.01039.x

Durkalski, V., Berger, V. W. (2009). Re-formulating equivalence trials as superiority trials: the case of binary outcomes. *Biometrical Journal, 51*(1), 185-192. doi:10.1002/bimj.200810499

Frane, A. V. (2015). Are Per-Family Type I Error Rate Relevant in Social and Behavioral Science? *Journal of Modern Applied Statistical Methods*, *14*(1).

Journal of Modern Applied Statistical Methods
May 2015, Vol. 14, No. 1, 43-52.

ISSN 1538 − 9472

Comparison of Bayesian Credible Intervals to Frequentist Confidence Intervals

Kathy Gray
California State University
Chico, CA

Brittany Hampton
California State University
Chico, CA

Tony Silveti-Falls
California State University
Chico, CA

Allison McConnell
California State University
Chico, CA

Casey Bausell
Oregon State University
Corvallis, OR

Frequentist confidence intervals were compared with Bayesian credible intervals under a variety of scenarios to determine when Bayesian credible intervals outperform frequentist confidence intervals. Results indicated that Bayesian interval estimation frequently produces results with precision greater than or equal to the frequentist method.

Keywords: Confidence intervals, credible intervals, Bayesian statistics, mean square error

Introduction

Although mathematicians introduced the field of Bayesian statistics in the 1700s, Bayesian methods gained most of its popularity in practice fairly recently (McCarthy & Parris., 2004; Smyth, 2004; Stoyan & Penttinan, 2000). Researchers have used frequentist methods for statistical analysis until technological advances and the introduction of certain algorithms, such as Markov chain Monte Carlo, gave way to increased computational power that enabled complex calculations to be done using Bayesian procedures (Little, 2006). This resulted in an increase in the interest of Bayesian statistics and sparked much controversy and debate regarding which method should be used by researchers (Little, 2006).

The frequentist approach relies on properties based on repeated sampling and takes only sample data into account to estimate the population parameter. Bayesian statistics, however, adds the component of a prior distribution based on prior knowledge and/or expert opinion of the subject. Using the prior information and the observed data, Bayesian methods calculate a refined estimate of the

Dr. Gray is an Associate Professor of Department of Mathematics and Statistics. Email her at klgray@csuchico.edu.

population parameter. Some claim that this subjective prior is key to most accurately estimating the population parameter while others claim that the lack of objectivity of Bayesian statistics interferes with the results (Choy et al., 2009).

The goal of this study was to compare Bayesian credible intervals to frequentist confidence intervals under a variety of scenarios to determine when Bayesian credible intervals outperform frequentist confidence intervals. The Central Limit Theorem (CLT) states that when a large enough random sample is taken the distribution of the sample means will be approximately normal. This theorem has been widely researched and it is generally accepted that as long as the sample size is around 25 we can rely on the CLT when performing inference on the population mean when the population is not normal (Stonehouse & Forrester, 1998).

Although not as well studied as the CLT, there exists a Bayesian Central Limit Theorem (BCLT) which states that under certain conditions the posterior probability distribution is approximately normal for large enough sample sizes (Walker, 1969). For Bayesian credible intervals, if the data are assumed to follow a normal distribution and if the prior distribution is also assumed to be normal then the calculations are straightforward because the posterior distribution for the population mean will also follow a normal distribution (Kruschke, 2010). If the data are not normal and transformations of the data do not achieve normality, then a more appropriate distribution could be used to model the data, however, this leads to a more complicated analysis. Furthermore, there is no guarantee that an appropriate distribution can be found that models the data. The goal of our study is to examine the robustness of Bayesian credible intervals when the assumption of normally distributed data is violated and to determine under what scenarios Bayesian credible intervals outperform frequentist confidence intervals.

Methodology

In order to investigate the BCLT we generated populations from three different distributions: 1) Standard normal distribution; 2) Beta distribution with parameters $\alpha = 2$ and $\beta = 5$ (moderately skewed distribution); 3) Exponential distribution with parameter $\lambda = 0.5$ (strongly skewed distribution). We repeatedly and randomly sampled from each population for various sample sizes (n = 10, 15, 20, 25, 30, 40, 50, and 75). For each scenario we calculated Bayesian credible intervals and frequentist confidence intervals. The frequentist confidence interval was calculated using the following formula:

$$\bar{x} \pm t_{n-1} \frac{s}{\sqrt{n}} \tag{1}$$

where t_{n-1} is the critical value for a 95% confidence interval with $n - 1$ degrees of freedom. Bayesian confidence intervals were calculated as follows:

$$\mu_1 \pm t_{n-1}\sigma_1 \tag{2}$$

where

$$\mu_1 = \frac{1/\sigma_0^2}{n/\sigma^2 + 1/\sigma_0^2}\mu_0 + \frac{n/\sigma^2}{n/\sigma^2 + 1/\sigma_0^2}\bar{x} \tag{3}$$

and

$$\sigma_1^2 = \frac{\sigma^2\sigma_0^2}{\sigma^2 + n\sigma_0} \tag{4}$$

where σ^2 is the population variance, μ_0 is the prior mean and σ_0^2 is the prior variance.

The population variance was always estimated with the sample variance, s^2. For each population distribution and sample size we calculated Bayesian credible intervals using a prior mean that wasn't biased, a prior that had a low bias, and a prior that had a large bias. We use the term bias to represent how far off the prior mean is from the population mean. A bias of 0.25 times the standard deviation was considered as a small bias in the prior mean and a bias of 0.50 times the standard deviation as a large bias in the prior mean. For the normal distribution, the bias was added to the prior before running the simulations. For the skewed distributions, we looked at both positive (prior mean > population mean) and negative biases (prior mean < population mean). The prior variance can be thought of as how confident one is in the prior mean. For instance, if there is a lot of confidence in the prior mean then the prior variance would be small since the researcher has honed in on the population mean. If there is little confidence in their prior mean then the prior variance would be large to reflect this. A confidence in the precision was considered to be equal to a value that would be equivalent to a sample size of about 12. In other words, about as much confidence was placed in the prior as would be if there was a sample of 12 from the population. This value was somewhat arbitrary; however, it represents the typical confidence in a prior mean. Thus, the prior variance was calculated as

$$\sigma_0^2 = \frac{\sigma^2}{12} \tag{5}$$

For each scenario (combination of sample size, bias, and population shape) we computed capture rate as the percent of the intervals that contained the true population mean. Additionally, the mean squared error (MSE) was calculated for each scenario. The MSE combines both the bias of an estimator as well as the variance. The MSE was calculated as

$$\text{MSE}\left(\hat{\mu}\right) = \text{bias}^2 + \text{var}\left(\hat{\mu}\right) \tag{6}$$

The bias is the difference between the estimated value and the true mean of the population. For the frequentist method it can be shown that the bias of the sample mean is 0, therefore, the MSE is var($\bar{y}$) for frequentist methods. All statistical analyses were performed using R (R Development Core Team, 2007).

Results

The capture rates of the frequentist and Bayesian intervals are shown in Figure 1 for various scenarios and sample sizes. As expected, the frequentist intervals have a 95% capture rate when the population distribution is normal. The frequentist method does quite well for the moderately skewed population where a sample size of 30 is needed to obtain a 95% capture rate. For the strongly skewed population, the frequentist intervals do not capture the parameter at the stated 95% level, however, when the sample size is 75 the capture rate remains at about 94%. For all scenarios, the no bias and positive, low bias scenarios performed best for small sample sizes with capture rates above 95% for both the normal population and the moderately skewed population. These capture rates decreased when sample size increased since the credible intervals were weighted more heavily by the data rather than the prior information and thus conform to frequentist properties.

For the strongly skewed data, the scenario that performed the best was the positive, low bias prior. This scenario captured the mean about 95% of the time for all sample sizes. A negative bias increased the capture rate when compared to a positive bias. As expected, the high bias scenarios performed the worst with respect to capture rates. For all sample sizes, the high bias scenarios gave worse results than the frequentist intervals, indicating that sample sizes need to be larger than 75 to dilute the bias in the prior. The results indicate that as long as the bias in the prior distribution is not too large then one can have results better than frequentist's methods even for strongly skewed distributions.

The MSE is given in Figure 2 for each sample size. The MSE accounts for both bias and variance of an estimator and, therefore, the smaller the MSE the better the performance of the statistic in estimating the parameter. Since the sample mean is an unbiased estimate of the population mean, the frequentist confidence interval has a bias equal to zero and the MSE is only based on the variance of the estimator. The bias for the Bayesian credible intervals varied from no bias to a bias of half of a standard deviation. For all three population distributions, there were no significant differences between MSE when the sample size reached 75. Similar results were obtained for the normal and moderately skewed population.

The largest difference in MSE between the different scenarios occurred for small sample sizes. The MSE was significantly larger for the frequentist confidence intervals than the Bayesian credible intervals until a sample size of 40 for the high biased Bayesian scenario. When comparing the frequentist confidence intervals to the low bias and no bias credible intervals, they are significantly lower until the sample size reaches 75. All Bayesian scenarios performed better than the frequentist intervals until a sample size of 30 was reached. The no bias and low bias continued to perform better than the frequentist interval until a sample size of 75 was reached.

The degree of bias needed before the capture rate drops below 0.95 is investigated in Figure 3. Iterations were performed using samples sizes of 15, 30, 50, and 75. Surprisingly, there was not much difference between the three different population shapes (normal, moderately skewed, strongly skewed) even for smaller sample sizes. In addition, the sample size did not have much effect on capture rate. When the sample size is 15 both the normal distribution and the moderately skewed distribution are above the 95% capture rate until the bias was equal to 0.4, at which point the capture rate dropped very quickly. The strongly skewed population performed only slightly below the 95% capture rate when $n = 15$ until a bias equal to 0.4 at which point it dropped off significantly. The differences between the three distributions were very small for all sample sizes. For sample sizes larger than 15 the capture rate dropped below the 95% level at a

0.3 level of bias. Thus, it appears that the effects of the bias are slightly worse for larger sample sizes.

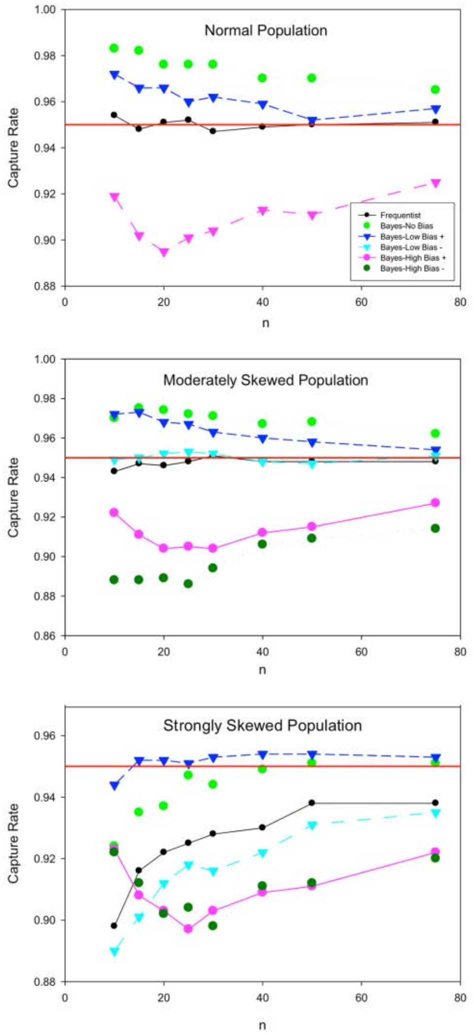


Figure 1. Capture rates for confidence intervals and Bayesian credible intervals.

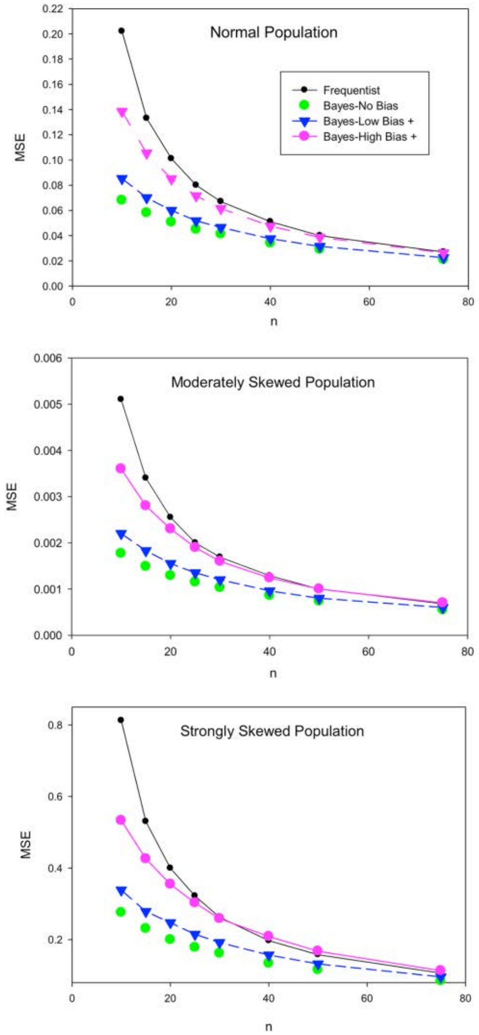


Figure 2. Mean squared error (MSE) for each scenario.

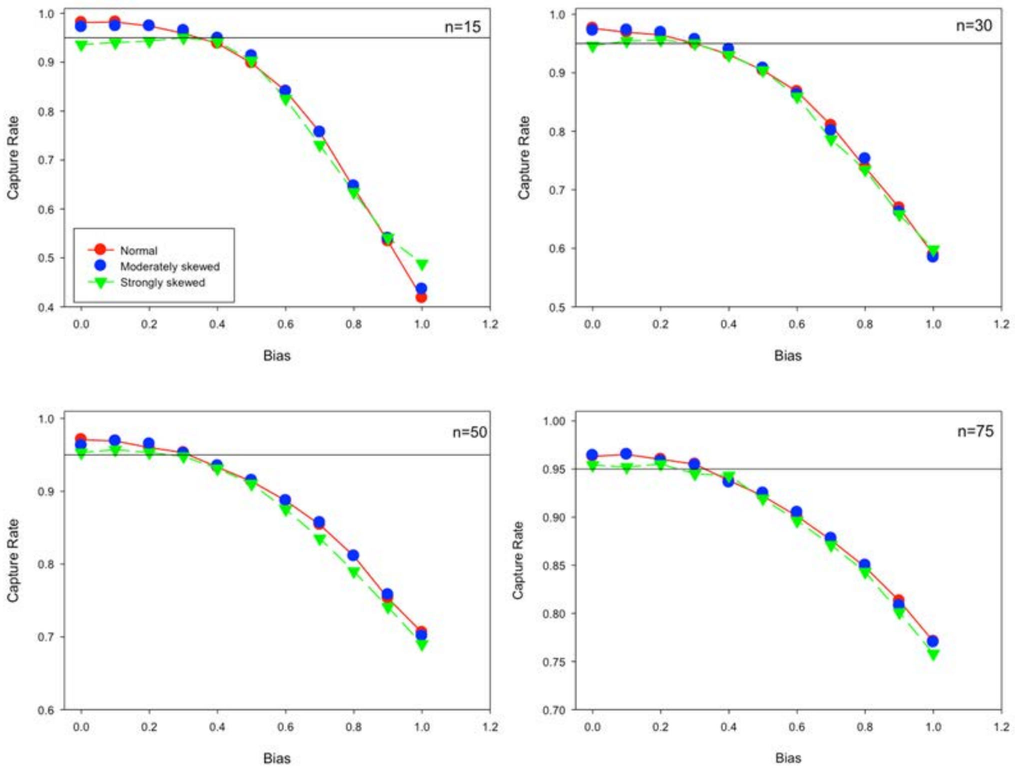


Figure 3. Capture rates for different sample sizes and different degrees of bias. The bias is calculated by the number of standard deviations above the prior's mean.

The MSE for increasing levels of bias in the prior mean in shown in Figure 4. The three distributions are shown on separate graphs and within each graph are three separate sample sizes. The solid line represents the MSE for frequentist methods for each sample size. For the normal distribution, when the Bayesian methods reach a bias of 0.6 the MSE is about equal to the frequentist methods with the same sample size. After a bias of 0.6 the Bayesian methods perform worse than frequentist methods with respect to the MSE. The differences were minor when comparing distributions. For strongly skewed distributions a smaller biased is required to perform better than frequentist methods. Surprisingly, for all distributions there was not much difference between the bias cutoff for different sample sizes. For the strongly skewed distribution, after a bias of 0.4 the frequentist methods performs better for $n = 50$ compared to a bias of 0.6 for a sample size of $n = 15$.

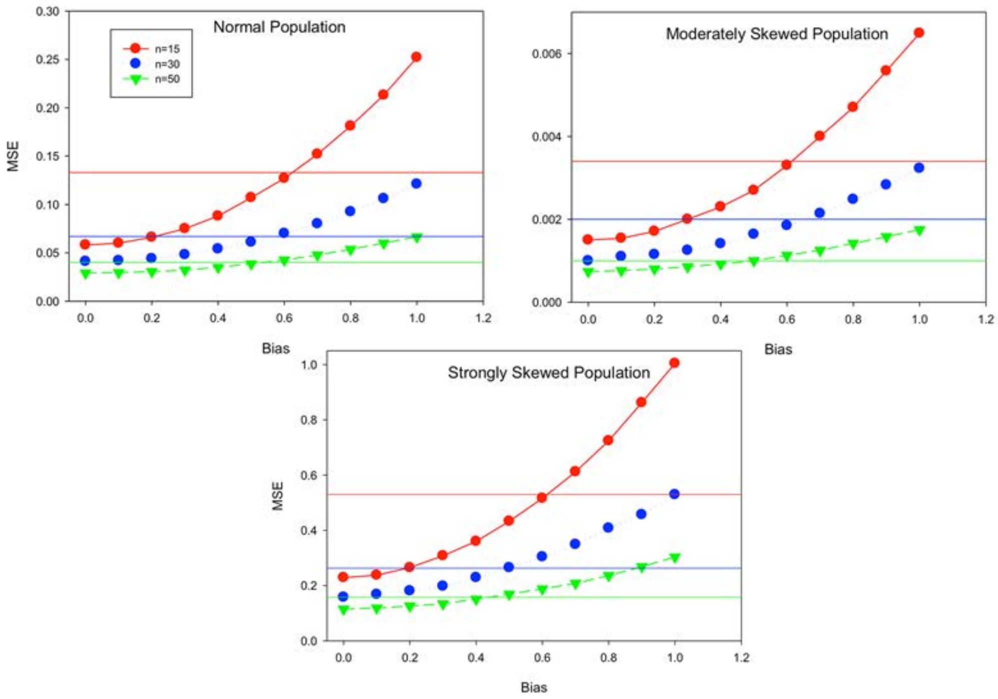


Figure 4. MSE calculated for different degrees of bias. The horizontal line reflects the MSE for frequentist methods of the same sample size.

Conclusion

These results indicate that when a prior mean is less than 0.4 to 0.6 standard deviations from the population mean then Bayesian credible intervals outperform frequentist confidence intervals with respect to MSE and capture rate for most scenarios that we looked at. For larger biases, frequentist confidence intervals will perform better with respect to MSE. Additionally, the distribution of the data did not have a large effect on the results even though the methods used assumed that the data came from a normal distribution. For strongly skewed data, neither frequentist nor Bayesian intervals performed at the optimal 95% capture rate with the exception being the Bayesian scenario with small, positive bias. Thus, for strongly skewed data it is suggested to seek a transformation for the data no matter which technique is used. In conclusion, this research demonstrates that Bayesian credible interval can have desirable properties for small sample sizes when the bias can be kept within about 0.5 standard deviations of the mean.

Because researchers will never know the bias of the prior mean they should only use Bayesian techniques when they have good information about the subject being researched.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. 1156612.

References

Choy, S. L., O'Leary, R., & Mengersen, K. (2009). Elicitation by design in ecology: using expert opinion to inform priors for Bayesian statistical models. *Ecology*, *90*(1), 265-277. doi:10.1890/07-1886.1

Kruschke, J. (2010). *Doing Bayesian data analysis: A tutorial introduction with R*. Waltham, MA: Academic Press.

Little, R. J. (2006). Calibrated Bayes: a Bayes/frequentist roadmap. *The American Statistician*, *60*(3), 213-223. doi:10.1198/000313006X117837

McCarthy, M. A. & Parris, K. M. (2004). Clarifying the effect of toe clipping on frogs with Bayesian statistics. *Journal of Applied Ecology*, *41*, 780–786. doi:10.1111/j.0021-8901.2004.00919.x

R Development Core Team (2007). *R: a language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing. Available at http://www.R-project.org.

Smyth, G. K. (2004). Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, *3*(1), 1-25. doi:10.2202/1544-6115.1027

Stonehouse, J. M., & Forrester, G. J. (1998). Robustness of the t and U tests under combined assumption violations. *Journal of Applied Statistics*, *25*(1), 63-74. doi:10.1080/02664769823304

Stoyan, D. & Penttinen, A. (2000). Recent applications of point process methods in forestry statistics. *Statistical Science 15*(1), 61-78. doi:10.1214/ss/1009212674. http://projecteuclid.org/euclid.ss/1009212674.

Walker, A. M. (1969). On the asymptotic behaviour of posterior distributions. *Journal of the Royal Statistical Society. Series B (Methodological)*, *31*(1). 80-88.

Journal of Modern Applied Statistical Methods
May 2015, Vol. 14, No. 1, 53-69.

ISSN 1538 − 9472

Modified Lilliefors Test

A. Adhikari
University of Northern Colorado
Greeley, Colorado

J. Schaffer
University of Northern Colorado
Greeley, Colorado

A new exponentiality test was developed by modifying the Lilliefors test of exponentiality. The proposed test considered the sum of all the absolute differences between the exponential cumulative distribution function (CDF) and the sample empirical distribution function (EDF). The proposed test is simple to understand and easy to compute.

Keywords: Cumulative distribution function, empirical distribution function, exponentiality test, critical value, significance level, and power

Introduction

Exponential distributions are quite often used in duration models and survival analysis, including several applications in macroeconomics, finance and labor economics (optimal insurance policy, duration of unemployment spell, retirement behavior, etc.). Quite often the data-generating process for estimating these types of models is assumed to behave as an exponential distribution. This calls for developing tests for distributional assumptions in order to avoid misspecification of the model (Acosta & Rojas, 2009). “The validity of estimates and tests of hypotheses for analyses derived from linear models rests on the merits of several key assumptions. The analysis of variance can lead to erroneous inferences if certain assumptions regarding the data are not satisfied” (Kuehl, 2000, p. 123).

As statistical consultants we should always consider the validity of the assumptions, be doubtful, and conduct analyses to examine the adequacy of the model. “Gross violations of the assumptions may yield an unstable model in the sense that different samples could lead to a totally different model with opposite conclusions” (Montgomery, Peck, & Vining, 2006, p. 122).

In this study we developed a new Goodness-of-Fit Test (GOFT) of exponentiality and compare it with four other existing GOFTs in terms of

Both authors are in the Applied Statistics and Research Methods. Email Dr. Adhikari at adhi2939@gmail.com. Email Dr. Schaffer at Jay.Schaffer@unco.edu.

computation and performance. This study also derived the critical values of the proposed test. The proposed test considered the sum of all the absolute differences between the empirical distribution function (EDF) and the exponential cumulative distribution function (CDF).

Methodology

To generate critical values, this study used data simulation techniques to mimic the desired parameter settings. Three different scale parameters ($\theta = 1$, 5, and 10) were used to generate random samples from an exponential distribution. Sample sizes 4, 5, 6, 7, 8, 9, 10, 15, 20, 25, 30, 35, 40, 45 and 50 were used. The study considered three different significance levels (α) (0.01, 0.05 and 0.10). For each sample size and significance level, 50,000 trials were run from an exponential distribution which generated 50,000 test statistics. The 50,000 test statistics were then arranged in the order from smallest to largest. The proposed test is a right tail test. So, this study used the 99th, 95th, and 90th percentile of the test statistics as the critical values for the given sample size for the 0.01, 0.05, and 0.10 significance levels respectively.

To verify the accuracy of the intended significance levels and to compare the power of the proposed test with other four exponentiality tests, data were produced from varieties of 12 distributions (Weibull (1,0.50), Weibull (1,0.75), Gamma (4,0.25), Gamma (0.55,0.275), Gamma (0.55,0.412), Gamma (4,0.50), Gamma (4,0.75), Gamma (4,1), Chi-Square (1), Chi-Square (2), t (5) and log-normal (0,1)) to see how the proposed test statistic works. Fifty thousand replications were drawn from each distribution for sample sizes 5, 10, 15, 20, 25, 30, 40, 50, 60, 70, 80, 90, 100, 200, 300, 400, 500, 1000, and 2000. For each sample size, the proposed test statistic and critical values were compared to make decisions about the null hypothesis. There were 50,000 trials for each sample size. The study tracked the number of rejections (rejection yes or no) in 50,000 trials to evaluate capacity of the proposed test to detect the departure from exponentiality.

The study used R 3.0.2 for most of the simulations to generate test statistics, critical values and power comparisons. Microsoft Excel 2010 was also used to make tables and charts. Monte Carlo simulation techniques were used to generate random numbers which were used to approximate the distribution of critical values for each test.

The proposed modified Lilliefors exponentiality test statistic (PML) takes the form,

$$PML = \sum_{i=1}^{n} \left| F^{*}\left(x_i\right) - S\left(x_i\right) \right|, \tag{1}$$

where $F^{*}\left(x_i\right)$ is the CDF of exponential distribution using the maximum likelihood estimator for the scale parameter θ and $S(x_i)$ is the sample cumulative distribution function. The estimator $\hat{\theta}$ is the uniformly minimum variance unbiased estimator (UMVUE) of the scale parameter θ.

The CDF, $F^{*}\left(x_i\right)$, is given by 2

$$F^{*}\left(x_i\right) = 1 - exp\left(-\frac{xi}{\overline{x}}\right), \tag{2}$$

where $\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$. The EDF is given by equation 3

$$S\left(x_i\right) = i/n \tag{3}$$

Lilliefors test (LF-test) statistic (Lilliefors, 1969) is given by:

$$D = \underset{x}{Sup} \left| F^{*}\left(x_i\right) - S\left(x_i\right) \right|, \tag{4}$$

where, $F^{*}(x_i) = 1 - \exp\left(-\frac{xi}{\overline{x}}\right)$, $\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$, and $S(x_i)$ is the empirical distribution function (EDF). Finkelstein & Schafers test (S-test) statistics (Finkelstein & Schafer, 1971) is given by:

$$S = \sum_{i=1}^{n} \max\left\{ \left| F_0\left(X_{(i)}, \hat{\theta}\right) - \frac{i}{n} \right|, \left| F_0\left(X_{(i)}, \hat{\theta}\right) - \frac{i-1}{n} \right| \right\}, \tag{5}$$

where, $\hat{\theta} = \overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$. Van-Soest test (VS-test) statistics (Soest, 1969) is given by:

$$W^2 = \frac{1}{12n} + \sum_{i=1}^{n}\left[t_i - \left(\frac{i-0.5}{n}\right)\right]^2, \tag{6}$$

where, $t_i = 1 - \exp\left(-\frac{xi}{\bar{x}}\right)$, and $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$. Srinivasan test ($\tilde{D}_n$ - test) statistics (Srinivasan, 1970) is given by:

$$\tilde{D}_n = \max 1 \le i \le n \left| S_n\left(x_i\right) - \tilde{F}\left(x\,;\lambda\right) \right|, \tag{7}$$

where, λ is a scale parameter, $\tilde{F}\left(x\,;\lambda\right)$ = 1 - $\left\{1 - \frac{x_i}{\left(n\bar{x}\right)}\right\}^{n-1}$, $S_n\left(x_i\right)$ is the EDF. According to Pugh (1963), the test statistic, $\tilde{D}_n$ -test, is based on the Rao-Blackwell and Lehman-Scheffe theorems which give the best unbiased estimate. Schafer, Finkelstein and Collins (1972) corrected the critical points of this test statistic originally proposed by Srinivasan (1970).

Results

Development of critical values

The critical values from the simulated data generated for the three different values of the scale parameters (θ = 1, 5, and 10) are exactly the same for the set of parameters. It appeared that the critical values for the proposed test are the functions of the sample size (n) and the significance levels (α) but invariant with the choice of the scale parameter (θ). Table 1 shows the critical values for the proposed test. Due to space limitations, only five digits are shown on Table 1.

Table 1. Critical Values for the Proposed Exponentiality Test (θ = 1)

n	α = 0.01	α = 0.05	α = 0.10
4	1.0567	0.8331	0.7409
5	1.1760	0.9315	0.8202
6	1.2703	1.0109	0.8931
7	1.3642	1.0856	0.9562
8	1.4647	1.1580	1.0189
9	1.5403	1.2209	1.0757
10	1.6274	1.2875	1.1310
15	1.9444	1.5561	1.3653
20	2.2271	1.7731	1.5636
25	2.4762	1.9682	1.7342
30	2.7097	2.1624	1.9066
35	2.9111	2.3291	2.0584
40	3.1062	2.4837	2.1904
45	3.3216	2.6331	2.3204
50	3.4557	2.7526	2.4309

Accuracy of significance levels

The simulated significance levels are presented on Table 2. Due to the limitations of the space, the simulated significance levels are rounded to three digits. The results showed that all five tests of exponentiality worked very well in terms of controlling the intended significance levels. The study found that the proposed test performs very closely to other four tests of exponentiality in terms of the accuracy of the intended significance levels (for each sample size and overall averages across the 19 different sample sizes). To allow for a better view of the five exponentiality tests across all sample sizes and significance levels, the columns for Lilliefors test are labelled by “LF”, Van-Soest test by “VS”, proposed modified Lilliefors test by “PML”, Srinivasan test by “D” and Finkelstein & Schafers test by “S” for the rest of the tables and figures presented in this study.

Table 2. Average Simulated Significance Levels

α	LF	D	CVM	S	PML
0.010	0.010	0.010	0.010	0.010	0.010
0.050	0.051	0.051	0.051	0.051	0.051
0.100	0.100	0.100	0.101	0.101	0.101

Power analysis

First, consider the relationship between the alternative distribution, Weibull (1, 0.50) and the simulated power. Figure 1 summarizes the power analysis for the Weibull (1, 0.50) alternative distribution. The PML-test outperformed the power for all other four exponentiality tests across all significance levels and sample sizes. The power of all four exponentiality tests exceeded the LF-test. The VS-test, the D-test, and the S-test showed similar performance in power. It appears that for sample sizes 40 or more, the powers for all five exponentiality tests close to 1.

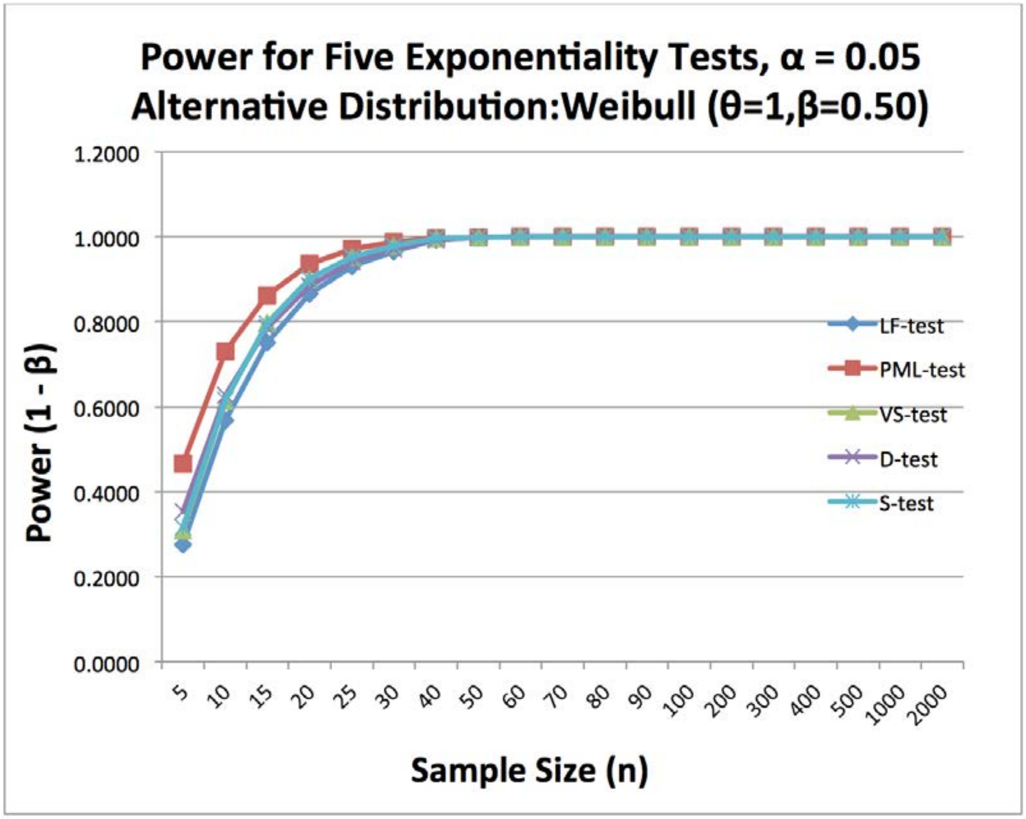


Figure 1. Power for Alternative Distribution: Weibull (1, 0.50)

Second, consider the relationship between the alternative distribution, Weibull (1, 0.75) and the simulated power. Figure 2 summarizes the power analysis for the Weibull (1, 0.75) alternative distribution. This distribution has the same scale parameter ($\theta = 1$) with the previous Weibull (1, 0.50) distribution but the shape parameter (β) is changed from 0.50 to 0.75. This caused the power to reduce substantially across all sample sizes and all significance levels under consideration.

The PML-test outperformed the power for all other four exponentiality tests across all sample sizes and significance levels. In all parameter settings under investigation, the powers for the LF-test were the lowest as compared to other four exponentiality tests. The powers of the S-test and VS-test were almost identical across all sample sizes and significance levels. For a fixed significance level, the powers for the D-test were greater than the S-test and VS-test for small sample sizes but this relationship was reversed for medium to large sample sizes. For all significance levels with sample sizes at least 200, the powers for all five exponentiality tests were almost equal and they approach 1.

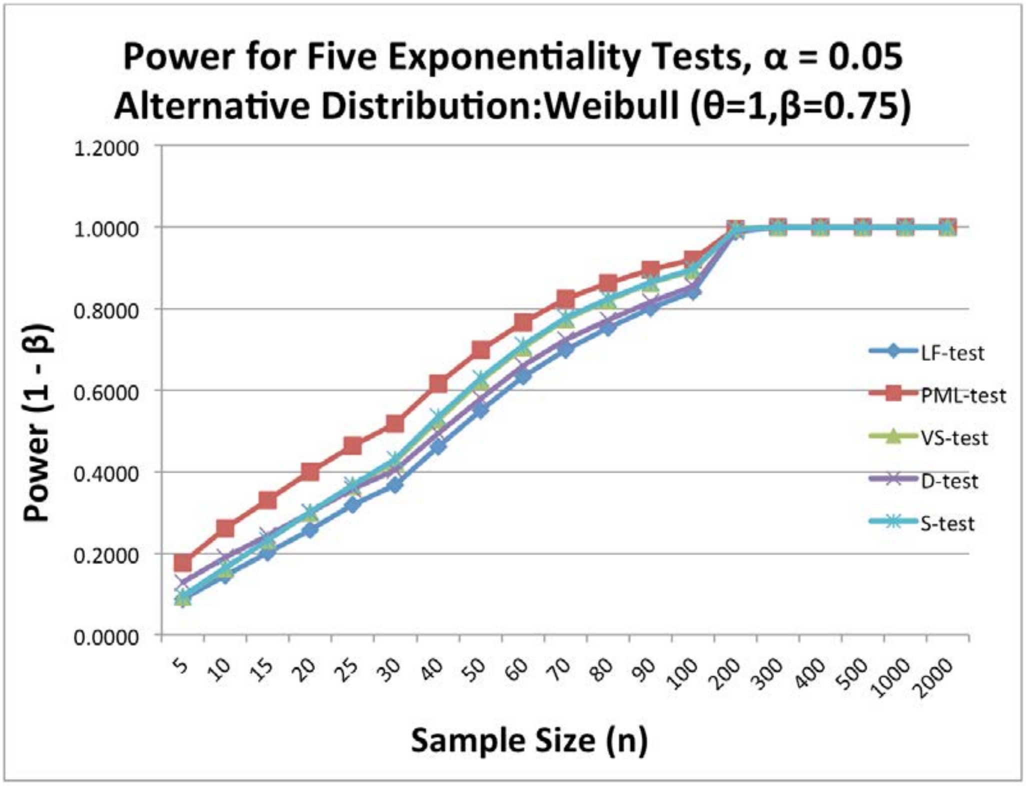


Figure 2. Power for Alternative Distribution: Weibull (1, 0.75)

Third, consider the relationship between the alternative distribution, Gamma (4, 0.25) and the simulated power. Figure 3 summarize the power analysis for the Gamma (4, 0.25) alternative distribution. According to Bain and Engelhardt (1992), the shape parameter, *k*, in the Gamma distribution determines the basic shape of the graph of the probability distribution function (PDF). The value of the shape parameter in null distribution is 1 and the shape parameter in this alternative distribution is 0.25 which are much different. The PML-test outperformed the powers of all other four exponentiality tests across all sample sizes and all significance levels under consideration. For a fixed significance level, the powers of the D-test, VS-test, and S-test exceeded the powers of the LF-test for small sample sizes. For medium to large sample sizes, the LF-test, D-test, S-test, and the VS-test exhibited the identical power across all significance levels. In all parameter settings, the powers of the D-test, the VS-test and the S-test were similar. For sample sizes at least 40, the powers of all five exponentiality tests were found almost equal which were close to 1 across all significance levels.

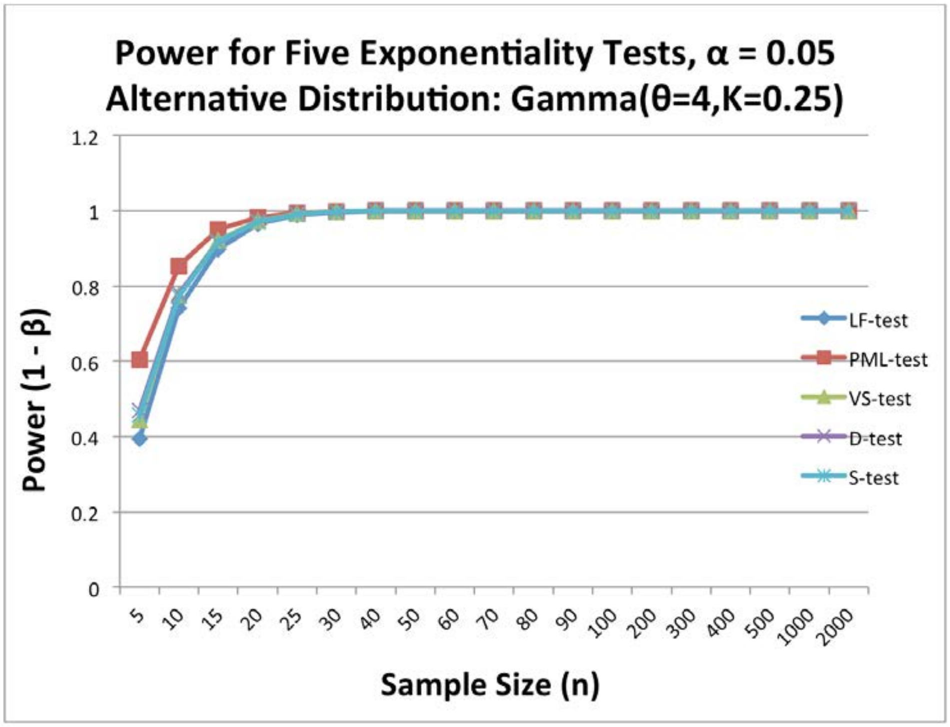


Figure 3. Power for Alternative Distribution: Gamma (4, 0.25)

Fourth, consider the relationship between the alternative distribution, Gamma (0.55, 0.275) and the simulated power. Figure 4 summarizes the power analysis for the Gamma (0.55, 0.275) alternative distribution. The PML-test outperformed other four exponentiality tests across all sample sizes and significance levels. The LF-test exhibited the lowest power across all sample sizes and significance levels. For sample sizes at least 50, the powers for all five tests were found almost equal which were close to 1 across all significance levels. In all parameter settings, the powers for the VS-test, the D-test, and the S-test were identical but all these three tests outperformed the LF-test across all sample sizes and significance levels.

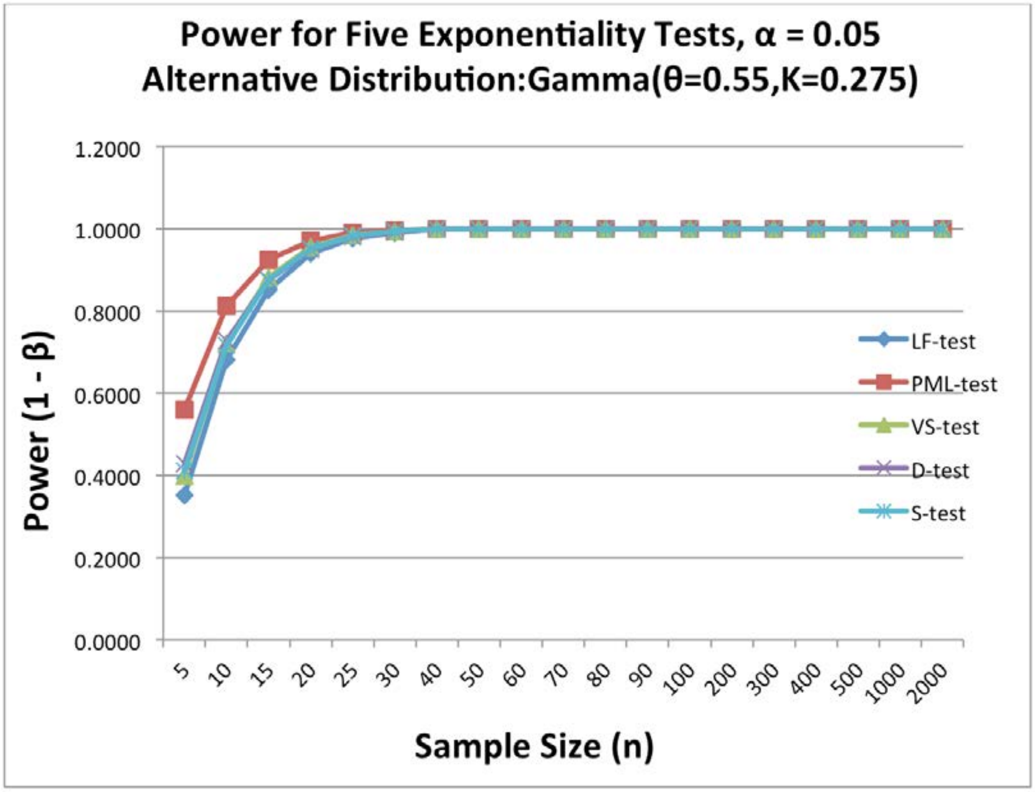


Figure 4. Power for Alternative Distribution: Gamma (0.55, 0.275)

Although the overall power trends in the previous alternative distribution (Gamma (4, 0.25)) and this distribution were similar among five exponentiality tests, the powers for this distribution was lower than the previous alternative distribution across all sample sizes and significance levels. In the previous alternative distribution, the value of the shape parameter (K) is 0.25 which is 0.275 in this alternative distribution.

Fifth, consider the relationship between the alternative distribution, Gamma (0.55, 0.412) and the simulated power. Figure 5 summarizes the power analysis for the Gamma (0.55, 0.412) alternative distribution. The PML-test outperformed other four exponentiality tests across all sample sizes and significance levels. The LF-test exhibited the lowest power across all sample sizes and significance levels. For sample sizes at least 80, the powers for all five tests were found almost equal which were close to 1 across all significance levels. In all parameter settings, the powers for the VS-test, the D-test, and the S-test were identical but all three tests outperformed the LF-test across all sample sizes and significance levels. Comparing the powers for this alternative distribution with the previous alternative distribution (Gamma (0.55, 0.275)), the powers were reduced in this alternative distribution across all sample sizes and significance levels. This is due to only the change in shape parameter (k) from 0.275 to 0.412. The scale parameters (θ) were the same on these two alternative distributions. It is relevant to argue that for Gamma alternative distribution, the powers for these five exponentiality tests depend only on the shape parameter (k). It is also important to note that the shape parameter (k) in the null distribution was 1. So, this study showed that as the shape parameter in the alternative distribution is close to the shape parameter of the null distribution, the simulated powers would be decreased.

Before considering the power for next two alternative distributions, it is imperative to discuss that the Chi-Square distribution is a special case of Gamma distribution. According to Bain and Engelhardt (1992), if a variable Y is a special Gamma distribution with scale parameter ($\theta = 2$) and shape parameter ($k = \text{v}/2$), the variable Y is said to follow a Chi-Square distribution with v degrees of freedom. So, if $Y \sim$ Gamma ($\theta = 2$, $k = v/2$), a special notation for this distribution can be written as:

$$Y \sim \chi^2(\nu) \tag{8}$$

Using equation 8, the Gamma (4, 0.5) and the Chi-Square (1) distributions are equivalent. This study previously showed that the power for the Gamma distribution depends only on the shape parameter (k). So, the powers of the Gamma (4, 0.5) and Chi-Square (1) alternative distributions must be equivalent.

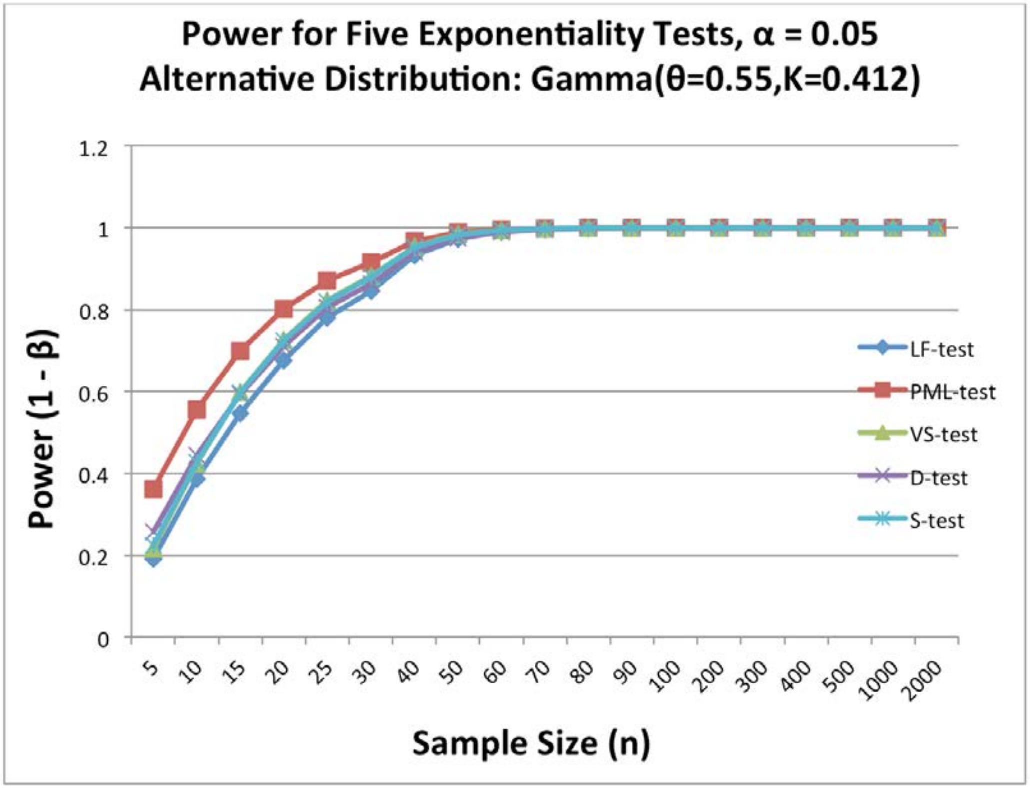


Figure 5. Power for Alternative Distribution: Gamma (0.55, 0.412)

Sixth, consider the relationship between the alternative distributions, Gamma (4, 0.5), Chi-Square (1) and the simulated power. Figure 6 summarizes the power analysis for the Gamma (4, 0.5) and Chi-Square (1) alternative distributions. For a fixed sample size and a significance level, powers for these two alternative distributions were exactly the same. As in the previous alternative distributions, the PML-test outperformed all other four exponentiality tests across all sample sizes and significance levels. The LF-test was in the last place on the power curve. The powers for the VS-test and S-test were identical for a fixed sample size and a significance level. The D-test demonstrated the superior power than the VS-test and the S-test for small sample sizes across all significance levels but this relationship was reversed for medium to large sample sizes. For sample sizes at least 200, the powers for all five tests were equivalents which were close to 1. As compare with the previous alternative distribution (Gamma (0.55, 0.412)), powers for these two alternative distributions decrease across all sample sizes and significance levels. It is relevant to note that the shape parameter (k) was changed from 0.412 to 0.50 which caused the decrease in power. It appears that as the value of the shape parameter (k) approaches that of the null distribution ($k = 1$), the simulated powers decreases.

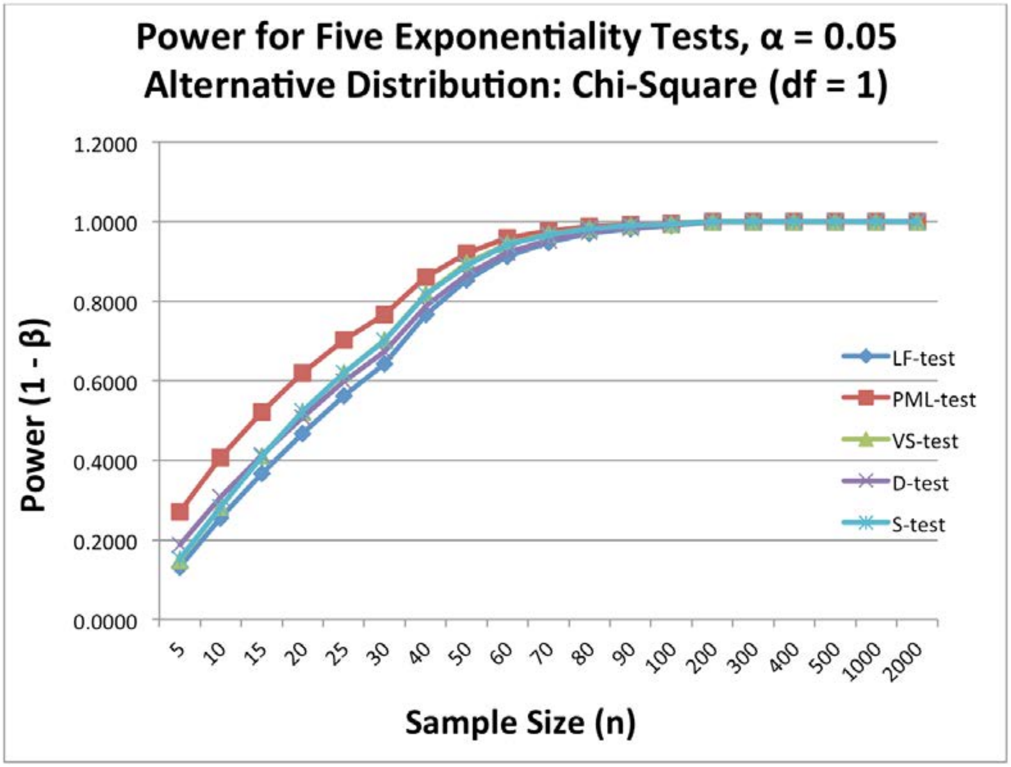


Figure 6. Power for Alternative Distribution: Chi-Square (1)

Seventh, consider the relationship between the alternative distribution Gamma (4, 0.75) and the simulated power. Figure 7 summarizes the power analysis for the Gamma (4, 0.75) alternative distribution. The PML-test outperformed all other four exponentiality tests across all sample sizes and significance levels. The LF-test was in the last place on the power curve. The powers for the VS-test and S-test were identical for a fixed sample size and significance level. The D-test demonstrated the superior power than the VS-test and the S-test for small sample sizes across all significance levels but this relationship was reversed for medium to large sample sizes. For sample size at least 1,000, the powers of all five tests were equivalents which were close to 1. As compare with the previous alternative distribution (Gamma (4, 0.5)), powers of this alternative distributions were significantly decrease across all sample sizes and significance levels. It is relevant to note that the shape parameter (*k*) was changed from 0.5 to 0.75 which caused the decrease in power. Among five Gamma alternative distributions discussed in this chapter, this alternative distribution exhibited the lowest power across all sample sizes and significance levels.

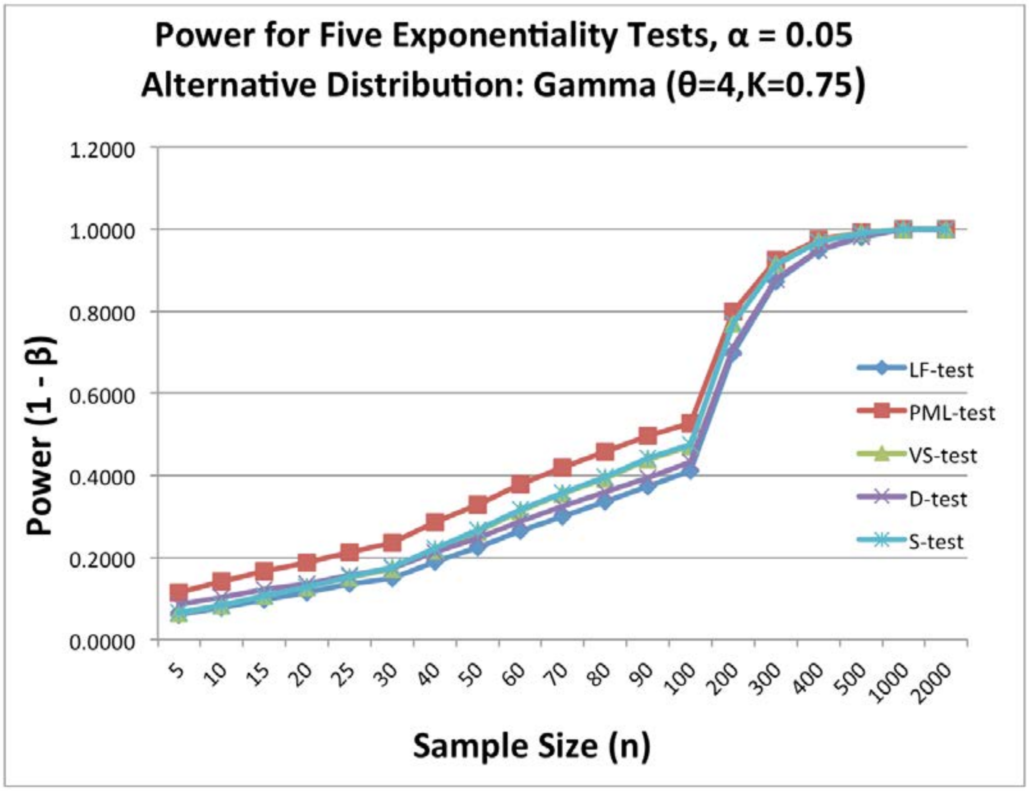


Figure 7. Power for Alternative Distribution: Gamma (4, 0.75)

Before considering the power for next two alternative distributions, it is indispensable to revisit that the Chi-Square distribution is a special case of Gamma distribution (equation 8). This study previously showed that the power for the Gamma distribution depends only on the shape parameter (*k*). Null distributions were generated using the exponential ($\theta = 5$) for power simulation.

Using 8, Gamma (4, 1) and Chi-Square (2) alternative distributions must produce similar powers for the set of parameters (n and α). In other words Gamma (4, 1) and Chi-Square (2) alternative distributions can be used for the simulation of significance levels.

Eighth, consider the relationship between the alternative distributions, Gamma (4, 1), Chi-Square (2) and the simulated power. Figure 8 summarizes the power analysis for the Gamma (4, 1) and Chi-Square (2) alternative distributions. The powers of all five exponentiality tests across all sample sizes and significance levels were too low which were pretty close to their significance levels. It is due to the fact that the power of these five exponentiality tests depends only on the shape parameter (k). It appears that the scale parameter (θ) does not have any role on the simulated powers.

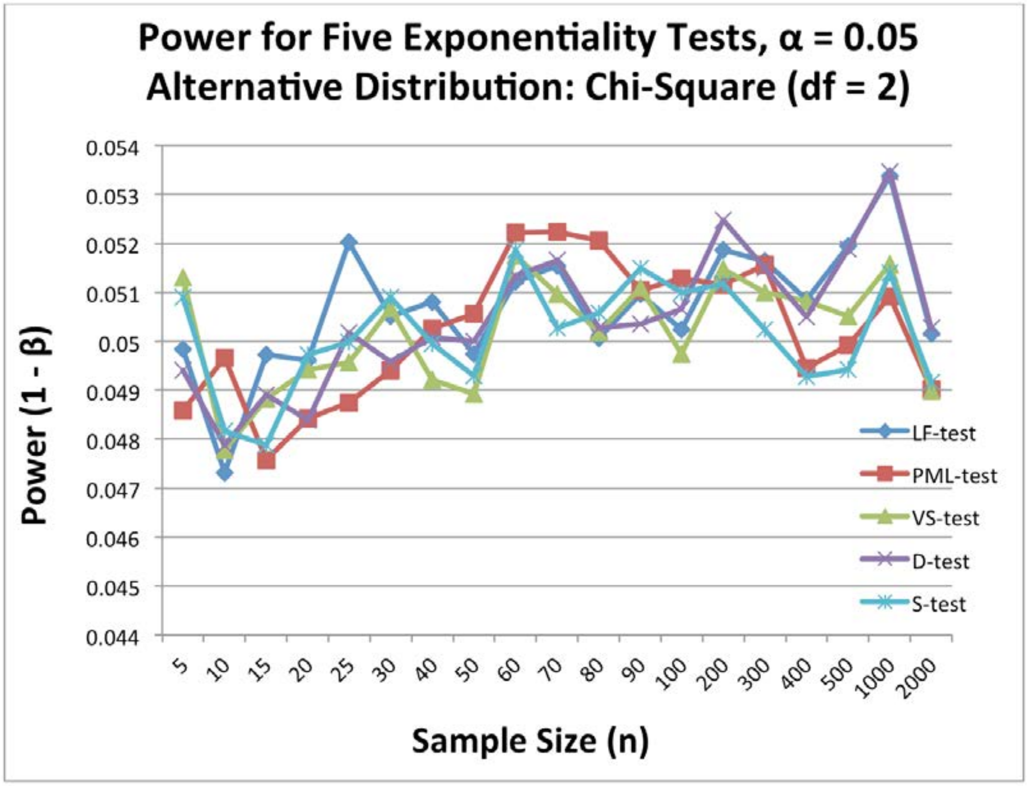


Figure 8. Power for Alternative Distribution: Chi-Square (2)

Ninth, consider the relationship between the alternative distribution *t* (5) and the simulated power. Figure 9 summarizes the power analysis for the *t* (5) alternative distribution. This is the only one symmetric distribution used in the power analyses. All five exponentiality tests quickly detected non-exponentiality. For sample sizes at least 15, the powers for all five tests were almost identical which were close to 1. The range of the powers was found to be very narrow across all sample sizes for a fixed significance level.

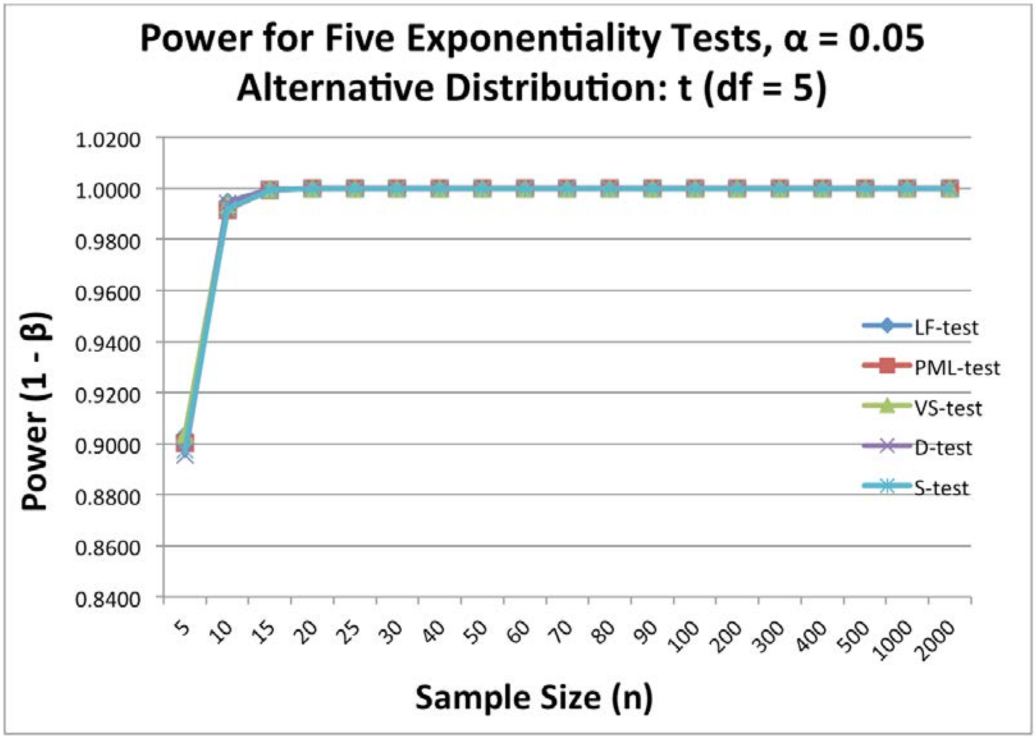


Figure 9. Power for Alternative Distribution: *t* (5)

Finally, consider the relationship between the alternative distribution log-normal (0, 1) and the simulated power. Figure 10 summarizes the power analysis for the log-normal (0, 1) alternative distribution. For small sample sizes, all five exponentiality tests demonstrated similar power across all significance levels. For medium to large sample sizes, the PML-test and S-test were in the top, the VS-test was in the middle and the D-test and LF-test were in the bottom of the power curve. It appears that the PML-test exhibited equal or better power among five exponentiality tests in the set of parameters considered in this study. For sample sizes at least 1000, the powers for all five tests were almost identical which were close to 1.

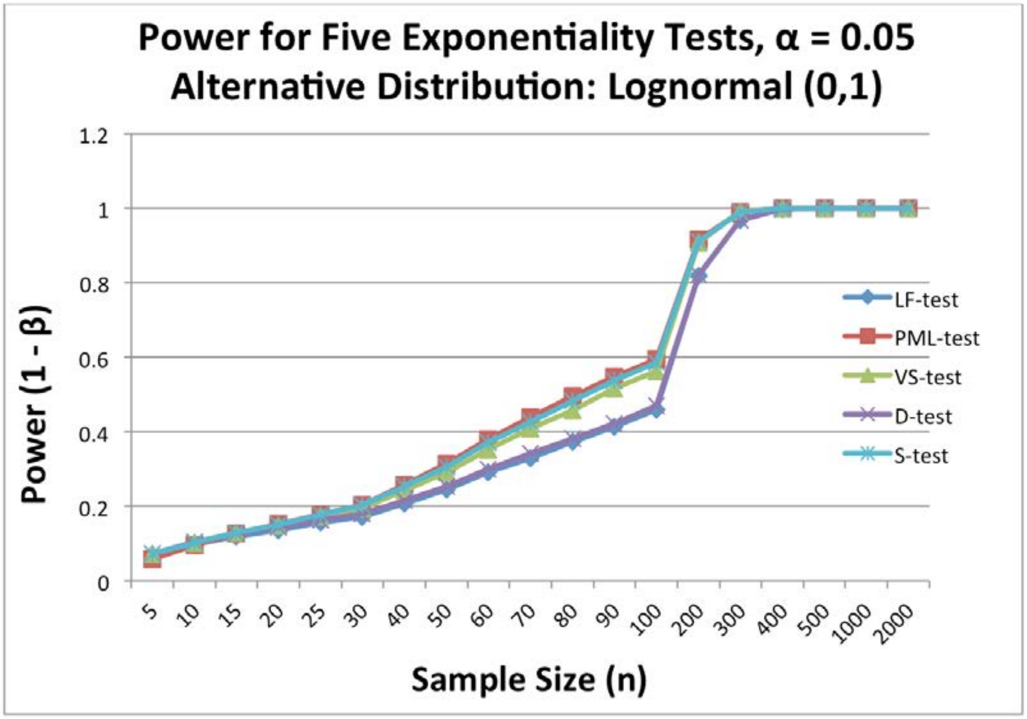


Figure 10. Power for Alternative Distribution: log-normal (0, 1)

Conclusion

This study claimed that the PML-test demonstrated consistently superior power over the S-test, LF-test, VS-test, and D-test for most of the alternative distributions presented in this study. The D-test, VS-test, and S-test exhibited similar power for a fixed sample size and a significance level. The LF-test consistently showed the lowest power among five exponentiality tests. So, practically speaking the proposed test can hope to replace the other four exponentiality tests discussed throughout this study while maintaining a very simple form for computation and easy to understand for those people who have limited knowledge of statistics.

References

Acosta, P., & Rojas, G. M. (2009). A simple IM test for exponential distributions. *Applied Economics Letters, 16*(2), 109-112. doi:10.1080/13504850601018221

Bain, L. & Engelhardt, M. (1992). *Introduction to Probability and Mathematical Statistics* (2nd ed.). MA: PWS-KENT Publishing Company.

Finkelstein, J. M. & Schafer, R. E. (1971). Improved goodness-of-fit tests. *Biometrika, 58*(3), 641-645. doi:10.1093/biomet/58.3.641

Kuehl, R. (2000). *Design of Experiments: Statistical Principles of Research Design and Analysis* (2nd ed.). CA: Duxbury.

Lilliefors, H. W. (1969). On the Kolmogorov-Smirnov test for the exponential distribution with mean unknown. *Journal of the American Statistical Association, 64*(325), 387-389. doi:10.1080/01621459.1969.10500983

Montgomery, D. C., Peck, E. A. & Vining, G. G. (2006). *Introduction to Linear Regression Analysis* (4th ed.). NJ: John Wiley and Sons, Inc.

Pugh, E. L. (1963). The best estimate of reliability in the exponential case. *Operations Research, 11*(1), 57-61.doi:10.1287/opre.11.1.57

Schafer, R. E., Finkelstein, J. M. & Collins, J. (1972). On a goodness-of-fit test for the exponential distribution with mean unknown. *Biometrika, 59*(1), 222-224. doi:10.1093/biomet/59.1.222

Soest, J. v. (1969). Some goodness of fit tests for the exponential distribution. *Statistica Neerlandica, 23*(1), 41-51. doi:10.1111/j.1467-9574.1969.tb00072.x

Srinivasan, R. (1970). An approach to testing the goodness of fit of incompletely specified distributions. *Biometrika, 57*(3), 605-611. doi:10.1093/biomet/57.3.605

Journal of Modern Applied Statistical Methods
May 2015, Vol. 14, No. 1, 70-82.

ISSN 1538 − 9472

Test for the Equality of Partial Correlation Coefficients for Two Populations

Madhusudan Bhandary
Columbus State University
Columbus, GA

Arjun K. Gupta
Bowling Green State University
Bowling Green, OH

A likelihood ratio test for the equality of two partial correlation coefficients based on two independent multinormal samples has been derived. The large sample Z-test for the same problem has also been discussed. The power analysis of the two tests is obtained. It has been found that the approximate likelihood ratio (ALR) test showed consistently better results than Z -test in terms of power. The size of the ALR test is slightly more than the alpha level. The ALR test is recommended strongly for use in practice.

Keywords: Likelihood ratio test, partial correlation coefficients, asymptotic distribution

Introduction

The partial correlation coefficient is frequently used to measure the correlation of two variables after eliminating the effect of other variable(s) in a set of correlated variables. For example, it may be of interest to know the correlation between intelligence and weight of people after eliminating the effect of age. In this case, the partial correlation coefficient will give the appropriate measure of the required correlation.

Statistical inference concerning the partial correlation coefficient for a single sample problem has been studied by Fisher (1924). Some discussions are also given in Anderson (2003). Surprisingly, the extension of inference problem concerning partial correlation coefficient to two-sample as well as multi-sample problems has received very little attention. Test for the equality of several multiple and partial correlation coefficients based on several independent Wishart densities has been derived by Gupta and Kabe (2001).

Dr. Bhandary is a Professor of Mathematics. Email him at: bhandary_madhusudan@columbusstate.edu, Dr. Gupta is a Distinguished Professor in the Department of Mathematics and Statistics. Email him at: gupta@bgnet.bgsu.edu.

In this paper, it has been considered the problem of testing the equality of two partial correlation coefficients based on two independent multinormal samples. It could be of interest to see whether the partial correlation of intelligence and weight of children after eliminating the effect of age in United States differs from the same in Asia and therefore, it is needed to develop test for the equality of partial correlation coefficients.

In the next section, the likelihood ratio test for the equality of two partial correlation coefficients is derived and the large sample test is also discussed. The power analysis of the tests is obtained through simulation and is discussed in the final section.

It has been found that the approximate likelihood ratio (ALR) test shows consistently better results than Z-test in terms of power. The size of the ALR test is slightly more than the alpha level. The ALR test is recommended strongly for use in practice.

Test of H_0: $\rho_{12.3}^{(1)} = \rho_{12.3}^{(2)}$ Versus H_1: $\rho_{12.3}^{(1)} \neq \rho_{12.3}^{(2)}$

Likelihood ratio test

Let $\underset{\sim}{x} = \left(x_1 \dots x_p\right)'$ be a $p \times 1$ vector of observations and $\underset{\sim}{\mu} = \left(\mu_1 \dots \mu_p\right)'$ be a $p \times 1$ vector of unknown means. It is assumed that $\underset{\sim}{x} \sim N_p\left(\underset{\sim}{\mu}, \Sigma\right)$, where Σ is a $p \times p$ unknown positive definite matrix and N_p denotes p-variate normal distribution. It will be considered the case of $p = 3$ in this article, i.e.

$$\underset{\sim}{x} = \begin{pmatrix} x_1 & x_2 & x_3 \end{pmatrix}' \sim N_3\left(\begin{pmatrix} \mu_1 & \mu_2 & \mu_3 \end{pmatrix}', \begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} \end{pmatrix} \right) \tag{1}$$

It follows from (1) that

$$\begin{pmatrix} x_{1.3} \\ x_{2.3} \end{pmatrix} \sim N_2\left(\begin{pmatrix} \mu_{1.3} \\ \mu_{2.3} \end{pmatrix}, \begin{pmatrix} \sigma_{11.3} & \rho_{12.3}\sqrt{\sigma_{11.3}}\sqrt{\sigma_{22.3}} \\ \rho_{12.3}\sqrt{\sigma_{11.3}}\sqrt{\sigma_{22.3}} & \sigma_{22.3} \end{pmatrix} \right) \tag{2}$$

where,

$$\begin{pmatrix} x_{1.3} \\ x_{2.3} \end{pmatrix} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \text{ given } X_3 = x_3$$

$$\mu_{1.3} = \mu_1 + \frac{\sigma_{13}}{\sigma_{33}}(x_3 - \mu_3)$$

$$\mu_{2.3} = \mu_2 + \frac{\sigma_{23}}{\sigma_{33}}(x_3 - \mu_3) \text{ and}$$

$$\rho_{12.3} = \frac{\sigma_{12.3}}{\sqrt{\sigma_{11.3}}\sqrt{\sigma_{22.3}}} = \text{partial correlation coefficient}$$

between X_1 and X_2 given $X_3 = x_3$.

Now, the joint probability density function of $x_{1.3}$ and $x_{2.3}$ is given by

$$f(x_{1.3}, x_{2.3}; \theta) = \frac{1}{2\pi\sqrt{\sigma_{11.3}\sigma_{22.3}\left(1-\rho_{12.3}^2\right)}} e^{-\frac{1}{2\left(1-\rho_{12.3}^2\right)}\left[\frac{(x_{1.3}-\mu_{1.3})^2}{\sigma_{11.3}} - 2\rho_{12.3}\frac{(x_{1.3}-\mu_{1.3})}{\sqrt{\sigma_{11.3}}}\frac{(x_{2.3}-\mu_{2.3})}{\sqrt{\sigma_{22.3}}} + \frac{(x_{2.3}-\mu_{2.3})^2}{\sigma_{22.3}}\right]} \qquad (3)$$

where, θ denotes the parameter vector of the distribution. Let $\underset{\sim}{X}^{(1)}, \underset{\sim}{X}^{(2)}, \ldots, \underset{\sim}{X}^{(n)}$ be 3×1 vector of observations i.i.d. $\sim N_3\left(\underset{\sim}{\mu}, \Sigma\right)$, where $\underset{\sim}{\mu}$ and Σ are given by (1).

From (2), it can be said that

$$\begin{pmatrix} x_{1.3}^{(i)} \\ x_{2.3}^{(i)} \end{pmatrix} \sim N_2\left(\begin{pmatrix} \mu_{1.3} \\ \mu_{2.3} \end{pmatrix}, \begin{pmatrix} \sigma_{11.3} & \rho_{12.3}\sqrt{\sigma_{11.3}}\sqrt{\sigma_{22.3}} \\ \rho_{12.3}\sqrt{\sigma_{11.3}}\sqrt{\sigma_{22.3}} & \sigma_{22.3} \end{pmatrix}\right) \qquad (4)$$

and $\begin{pmatrix} x_{1.3}^{(i)} \\ x_{2.3}^{(i)} \end{pmatrix}$'s are independent, $i = 1,2,\ldots,n$. Using (3), the likelihood function of $\begin{pmatrix} x_{1.3}^{(i)} \\ x_{2.3}^{(i)} \end{pmatrix}$, $i = 1,2,\ldots,n$ is obtained as follows:

$$L\left(x_{1.3}^{(i)}, x_{2.3}^{(i)}, i=1,\ldots,n;\theta\right)=$$

$$\frac{1}{\left(2\pi\sqrt{\sigma_{11.3}\sigma_{22.3}\left(1-\rho_{12.3}^{2}\right)}\right)^{n}} e^{-\frac{1}{2\left(1-\rho_{12.3}^{2}\right)}\left[\sum_{i=1}^{n}\frac{\left(x_{1.3}^{(i)}-\mu_{1.3}\right)^{2}}{\sigma_{11.3}}-2\rho_{12.3}\sum_{i=1}^{n}\frac{\left(x_{1.3}^{(i)}-\mu_{1.3}\right)}{\sqrt{\sigma_{11.3}}}\frac{\left(x_{2.3}^{(i)}-\mu_{2.3}\right)}{\sqrt{\sigma_{22.3}}}+\sum_{i=1}^{n}\frac{\left(x_{2.3}^{(i)}-\mu_{2.3}\right)^{2}}{\sigma_{22.3}}\right]} \quad (5)$$

Now, it is considered two sample problem with n_1 and n_2 observations respectively from each population. Let $\underset{\sim}{x}^{(i)}=\left(x_1^{(i)} \quad x_2^{(i)} \quad x_3^{(i)}\right)'$ be the ith observation from population 1; i = 1,2,…,n_1 and $\underset{\sim}{x}^{(i)} \sim N_3\left(\underset{\sim}{\mu}^{(1)}, \Sigma^{(1)}\right)$, where, $\underset{\sim}{\mu}^{(1)}$ and $\Sigma^{(1)}$ are the mean vector and dispersion matrix respectively and hence

$\begin{pmatrix} x_{1.3}^{(i)} \\ x_{2.3}^{(i)} \end{pmatrix} \sim N_2\left(\begin{pmatrix} \mu_{1.3}^{(i)} \\ \mu_{2.3}^{(i)} \end{pmatrix}, \begin{pmatrix} \sigma_{11.3}^{(1)} & \rho_{12.3}^{(1)}\sqrt{\sigma_{11.3}^{(1)}}\sqrt{\sigma_{22.3}^{(1)}} \\ \rho_{12.3}^{(1)}\sqrt{\sigma_{11.3}^{(1)}}\sqrt{\sigma_{22.3}^{(1)}} & \sigma_{22.3}^{(1)} \end{pmatrix}\right), i=1,2,\ldots,n_1$ where, $\begin{pmatrix} x_{1.3}^{(i)} \\ x_{2.3}^{(i)} \end{pmatrix} = \begin{pmatrix} x_1^{(i)} \\ x_2^{(i)} \end{pmatrix}$ given $X_3^{(i)} = x_3^{(i)}$

Let $\underset{\sim}{z}^{(j)}=\left(z_1^{(j)} \quad z_2^{(j)} \quad z_3^{(j)}\right)'$ be the jth observation from population 2; j = 1, 2, …, n_2 and $\underset{\sim}{z}^{(j)} \sim N_3\left(\underset{\sim}{\mu}^{(2)}, \Sigma^{(2)}\right)$, where, $\underset{\sim}{\mu}^{(2)}$ and $\Sigma^{(2)}$ are the mean vector and dispersion matrix respectively and hence $\begin{pmatrix} z_{1.3}^{(j)} \\ z_{2.3}^{(j)} \end{pmatrix} \sim N_2\left(\begin{pmatrix} \mu_{1.3}^{(2)} \\ \mu_{2.3}^{(2)} \end{pmatrix}, \begin{pmatrix} \sigma_{11.3}^{(2)} & \rho_{12.3}^{(2)}\sqrt{\sigma_{11.3}^{(2)}}\sqrt{\sigma_{22.3}^{(2)}} \\ \rho_{12.3}^{(2)}\sqrt{\sigma_{11.3}^{(2)}}\sqrt{\sigma_{22.3}^{(2)}} & \sigma_{22.3}^{(2)} \end{pmatrix}\right)$, where $\begin{pmatrix} z_{1.3}^{(j)} \\ z_{2.3}^{(j)} \end{pmatrix} = \begin{pmatrix} z_1^{(j)} \\ z_2^{(j)} \end{pmatrix}$ given $Z_3^{(j)} = z_3^{(j)}$. Under the above setup, the likelihood ratio test for testing H_0 Vs. H_1 is derived as follows: Under H_1, the log-likelihood function based on $\begin{pmatrix} x_{1.3}^{(i)} \\ x_{2.3}^{(i)} \end{pmatrix}, i=1,2,\ldots,n_1$ and $\begin{pmatrix} z_{1.3}^{(j)} \\ z_{2.3}^{(j)} \end{pmatrix}, j=1,2,\ldots,n_2$ is

$$\begin{aligned}\log L_1 = &-\left(n_1+n_2\right)\log 2\pi - \frac{n_1}{2}\left\{\log\sigma_{11.3}^{(1)} + \log\sigma_{22.3}^{(1)} + \log\left(1-\rho_{12.3}^{(1)2}\right)\right\}\\ &-\frac{n_2}{2}\left\{\log\sigma_{11.3}^{(2)} + \log\sigma_{22.3}^{(2)} + \log\left(1-\rho_{12.3}^{(2)2}\right)\right\}\\ &-\frac{1}{2\left(1-\rho_{12.3}^{(1)2}\right)}\left[\begin{array}{l}\sum_{i=1}^{n_1}\frac{\left(x_{1.3}^{(i)}-\mu_{1.3}^{(1)}\right)^2}{\sigma_{11.3}^{(1)}} - 2\rho_{12.3}^{(1)}\sum_{i=1}^{n_1}\frac{\left(x_{1.3}^{(i)}-\mu_{1.3}^{(1)}\right)}{\sqrt{\sigma_{11.3}^{(1)}}}\frac{\left(x_{2.3}^{(i)}-\mu_{2.3}^{(1)}\right)}{\sqrt{\sigma_{22.3}^{(1)}}}+\\ \sum_{i=1}^{n_1}\frac{\left(x_{2.3}^{(i)}-\mu_{2.3}^{(1)}\right)^2}{\sigma_{22.3}^{(1)}}\end{array}\right]\\ &-\frac{1}{2\left(1-\rho_{12.3}^{(2)2}\right)}\left[\begin{array}{l}\sum_{j=1}^{n_2}\frac{\left(z_{1.3}^{(j)}-\mu_{1.3}^{(2)}\right)^2}{\sigma_{11.3}^{(2)}} - 2\rho_{12.3}^{(2)}\sum_{j=1}^{n_2}\frac{\left(z_{1.3}^{(j)}-\mu_{1.3}^{(2)}\right)}{\sqrt{\sigma_{11.3}^{(2)}}}\frac{\left(z_{2.3}^{(j)}-\mu_{2.3}^{(2)}\right)}{\sqrt{\sigma_{22.3}^{(2)}}}+\\ \sum_{j=1}^{n_1}\frac{\left(z_{2.3}^{(j)}-\mu_{2.3}^{(2)}\right)^2}{\sigma_{22.3}^{(2)}}\end{array}\right]\end{aligned} \qquad (6)$$

Maximizing log L_1 in (6) w.r.t. $\mu_{1.3}^{(r)}$, $\mu_{2.3}^{(s)}$, $\sigma_{11.3}^{(r)}$, $\sigma_{22.3}^{(s)}$, $\rho_{12.3}^{(r)}$; r, s = 1, 2, it can be obtained that

$$\underset{H_1}{Sup}L_1 = \frac{1}{\left(2\pi\right)^{n_1+n_2}\left(\frac{a_{11.3}}{n_1}\right)^{\frac{n_1}{2}}\left(\frac{a_{22.3}}{n_1}\right)^{\frac{n_2}{2}}\left(1-\hat{\rho}_{12.3}^{(1)2}\right)^{\frac{n_1}{2}}\left(\frac{b_{11.3}}{n_2}\right)^{\frac{n_2}{2}}\left(\frac{b_{22.3}}{n_2}\right)^{\frac{n_2}{2}}\left(1-\hat{\rho}_{12.3}^{(2)2}\right)^{\frac{n_2}{2}}}.e^{-\left(n_1+n_2\right)} \qquad (7)$$

where,

$$a_{\alpha\beta.3} = \sum_{i=1}^{n_1}\left(x_{\alpha.3}^{(i)}-\bar{x}_{\alpha.3}\right)\left(x_{\beta.3}^{(i)}-\bar{x}_{\beta.3}\right)$$

$$\bar{x}_{\alpha.3} = \frac{1}{n_1}\sum_{i=1}^{n_1}x_{\alpha.3}^{(i)};\alpha,\beta=1,2$$

$$b_{mn.3} = \sum_{j=1}^{n_2}\left(z_{m.3}^{(j)}-\bar{z}_{m.3}\right)\left(z_{n.3}^{(j)}-\bar{z}_{n.3}\right)$$

$$\bar{z}_{m.3} = \frac{1}{n_2}\sum_{j=1}^{n_2}z_{m.3}^{(j)};m,n=1,2$$

and $\hat{\rho}_{12.3}^{(1)} = \dfrac{a_{12.3}}{\sqrt{a_{11.3}a_{22.3}}}, \hat{\rho}_{12.3}^{(2)} = \dfrac{b_{12.3}}{\sqrt{b_{11.3}b_{22.3}}}$.

Similarly, under H_0, the log-likelihood function is given by

$$\begin{aligned}\log L_0 = &-(n_1+n_2)\log 2\pi - \frac{n_1}{2}\left\{\log\sigma_{11.3}^{(1)} + \log\sigma_{22.3}^{(1)} + \log(1-\rho^2\right\} \\ &-\frac{n_2}{2}\left\{\log\sigma_{11.3}^{(2)} + \log\sigma_{22.3}^{(2)} + \log(1-\rho^2)\right\} \\ &-\frac{1}{2(1-\rho^2)}\left[\begin{array}{l}\displaystyle\sum_{i=1}^{n_1}\frac{\left(x_{1.3}^{(i)}-\mu_{1.3}^{(1)}\right)^2}{\sigma_{11.3}^{(1)}} - 2\rho\sum_{i=1}^{n_1}\frac{\left(x_{1.3}^{(i)}-\mu_{1.3}^{(1)}\right)}{\sqrt{\sigma_{11.3}^{(1)}}}\frac{\left(x_{2.3}^{(i)}-\mu_{2.3}^{(1)}\right)}{\sqrt{\sigma_{22.3}^{(1)}}} + \\ \displaystyle\sum_{i=1}^{n_1}\frac{\left(x_{2.3}^{(i)}-\mu_{2.3}^{(1)}\right)^2}{\sigma_{22.3}^{(1)}}\end{array}\right] \\ &-\frac{1}{2(1-\rho^2)}\left[\begin{array}{l}\displaystyle\sum_{j=1}^{n_2}\frac{\left(z_{1.3}^{(j)}-\mu_{1.3}^{(2)}\right)^2}{\sigma_{11.3}^{(2)}} - 2\rho\sum_{j=1}^{n_2}\frac{\left(z_{1.3}^{(j)}-\mu_{1.3}^{(2)}\right)}{\sqrt{\sigma_{11.3}^{(2)}}}\frac{\left(z_{2.3}^{(j)}-\mu_{2.3}^{(2)}\right)}{\sqrt{\sigma_{22.3}^{(2)}}} + \\ \displaystyle\sum_{j=1}^{n_2}\frac{\left(z_{2.3}^{(j)}-\mu_{2.3}^{(2)}\right)^2}{\sigma_{22.3}^{(2)}}\end{array}\right]\end{aligned} \tag{8}$$

where, ρ = common value of $\rho_{12.3}^{(1)}$ and $\rho_{12.3}^{(2)}$ under H_0. Maximizing log L_0 in (8) w.r.t. $\mu_{1.3}^{(1)}, \mu_{2.3}^{(1)}, \mu_{1.3}^{(2)}, \mu_{2.3}^{(2)}$ it is obtained that

$\hat{\mu}_{1.3}^{(1)} = \bar{x}_{1.3}, \hat{\mu}_{2.3}^{(1)} = \bar{x}_{2.3}, \hat{\mu}_{1.3}^{(2)} = \bar{z}_{1.3}, \hat{\mu}_{2.3}^{(2)} = \bar{z}_{2.3}$

Now, $\dfrac{\partial\log L_0}{\partial\sigma_{11.3}^{(1)}} = 0 \Rightarrow -\dfrac{n_1}{\sigma_{11.3}^{(1)}} + \dfrac{1}{(1-\rho^2)}\left\{\dfrac{a_{11.3}}{\sigma_{11.3}^{(1)2}} - \dfrac{\rho a_{12.3}}{\sigma_{11.3}^{(1)3/2}\sigma_{22.3}^{(1)1/2}}\right\} = 0$

$$\text{i.e. } \frac{a_{11.3}}{\sigma_{11.3}^{(1)}} - \frac{\rho a_{12.3}}{\sqrt{\sigma_{11.3}^{(1)}}\sqrt{\sigma_{22.3}^{(1)}}} = n_1(1-\rho^2) \tag{9}$$

Similarly,

$$\frac{\partial \log L_0}{\partial \sigma_{22.3}^{(1)}} = 0 \Rightarrow \frac{a_{22.3}}{\sigma_{22.3}^{(1)}} - \frac{\rho a_{12.3}}{\sqrt{\sigma_{11.3}^{(1)}}\sqrt{\sigma_{22.3}^{(1)}}} = n_1\left(1-\rho^2\right) \tag{10}$$

It is obtained by adding the equations (9) and (10) that

$$\frac{a_{11.3}}{\hat{\sigma}_{11.3}^{(1)}} + \frac{a_{22.3}}{\hat{\sigma}_{22.3}^{(1)}} - \frac{2\rho a_{12.3}}{\sqrt{\hat{\sigma}_{11.3}^{(1)}}\sqrt{\hat{\sigma}_{22.3}^{(1)}}} = 2n_1\left(1-\rho^2\right). \tag{11}$$

From (9) and (10), it follows that

$$\frac{a_{11.3}}{\hat{\sigma}_{11.3}^{(1)}} = \frac{a_{22.3}}{\hat{\sigma}_{22.3}^{(1)}} = k \text{ (say)}. \tag{12}$$

From (11) and (12), it can be obtained that $2k - 2\rho\hat{\rho}_{12.3}^{(1)}k = 2n_1\left(1-\rho^2\right)$

$$\text{i.e. } k = \frac{n_1\left(1-\rho^2\right)}{1-\rho\hat{\rho}_{12.3}^{(1)}} \tag{13}$$

Similarly,

$$k^* = \frac{n_1\left(1-\rho^2\right)}{1-\rho\hat{\rho}_{12.3}^{(2)}} \text{ where, } k^* = \frac{b_{11.3}}{\hat{\sigma}_{11.3}^{(2)}} = \frac{b_{22.3}}{\hat{\sigma}_{22.3}^{(2)}} \tag{14}$$

From (12), and (14) it follows that

$\hat{\sigma}_{11.3}^{(1)} = \dfrac{a_{11.3}}{k}, \hat{\sigma}_{22.3}^{(1)} = \dfrac{a_{22.3}}{k}, \hat{\sigma}_{11.3}^{(2)} = \dfrac{b_{11.3}}{k^*}$, and $\hat{\sigma}_{22.3}^{(2)} = \dfrac{b_{22.3}}{k^*}$. Given,

$$\rho,\ \underset{H_0}{Sup} \log L_0\left(\rho\right) = -\left(n_1+n_2\right)\log 2\pi - \frac{n_1}{2}\left[\log\left(\frac{a_{11.3}}{k}\right) + \log\left(\frac{a_{22.3}}{k}\right)\right]$$
$$-\frac{n_2}{2}\left[\log\left(\frac{b_{11.3}}{k^*}\right) + \log\left(\frac{b_{22.3}}{k^*}\right)\right] - \left(\frac{n_1+n_2}{2}\right)\log\left(i-\rho^2\right) - \left(n_1+n_2\right)$$
$$= -\left(n_1+n_2\right)\log 2\pi - \frac{n_1}{2}\left[\log\left(\frac{a_{11.3}}{n_1}\right) + \log\left(\frac{a_{22.3}}{n_1}\right)\right] - n_1\log\left(1-\rho\hat{\rho}_{12.3}^{(1)}\right)$$

$$-\frac{n_2}{2}\left[\log\left(\frac{b_{11.3}}{n_2}\right)+\log\left(\frac{b_{22.3}}{n_2}\right)\right]-n_2\log\left(1-\rho\hat{\rho}_{12.3}^{(2)}\right)+\left(\frac{n_1+n_2}{2}\right)\log\left(1-\rho^2\right)-\left(n_1+n_2\right) \quad (15)$$

(15) is obtained by using (13) and (14).

Now, taking partial derivative of the expression in (15) w.r.t. ρ and setting it to zero, it follows that

$$-\frac{n_1\left(-\hat{\rho}_{12.3}^{(1)}\right)}{1-\rho\hat{\rho}_{12.3}^{(1)}}-\frac{n_2\left(-\hat{\rho}_{12.3}^{(2)}\right)}{1-\rho\hat{\rho}_{12.3}^{(2)}}+\frac{\left(n_1+n_2\right)\left(-2\rho\right)}{2\left(1-\rho^2\right)}=0$$

$$\Rightarrow \frac{n_1\hat{\rho}_{12.3}^{(1)}k}{n_1\left(1-\rho^2\right)}+\frac{n_2\hat{\rho}_{12.3}^{(2)}k^*}{n_2\left(1-\rho^2\right)}-\frac{\left(n_1+n_2\right)\rho}{1-\rho^2}=0 \ \left(\text{using } (13) \text{ and } (14)\right)$$

$$\text{i.e., } \hat{\rho}=\frac{k\hat{\rho}_{12.3}^{(1)}+k^*\hat{\rho}_{12.3}^{(2)}}{n_1+n_2}. \quad (16)$$

Note: Asymptotically, $k \simeq n_1$ and $k^* \simeq n_2$ under H_0.
Hence,

$$\underset{H_0}{Sup}\, L_0=\frac{\left(1-\hat{\rho}^2\right)^{\frac{n_1+n_2}{2}}}{\left(2\pi\right)^{n_1+n_2}\left(\frac{a_{11.3}}{n_1}\right)^{\frac{n_1}{2}}\left(\frac{a_{22.3}}{n_1}\right)^{\frac{n_1}{2}}\left(1-\hat{\rho}\hat{\rho}_{12.3}^{(1)}\right)^{n_1}\left(\frac{b_{11.3}}{n_2}\right)^{\frac{n_2}{2}}\left(\frac{b_{22.3}}{n_2}\right)^{\frac{n_2}{2}}\left(1-\hat{\rho}\hat{\rho}_{12.3}^{(2)}\right)^{n_2}}.e^{-\left(n_1+n_2\right)} \quad (17)$$

Using (7) and (17), likelihood ratio test statistic for testing H_0 Vs. H_1 is obtained as follows:

$$\Lambda=\frac{\left(1-\hat{\rho}^2\right)^{\frac{n_1+n_2}{2}}\left(1-\hat{\rho}_{12.3}^{(1)2}\right)^{\frac{n_1}{2}}\left(1-\hat{\rho}_{12.3}^{(2)2}\right)^{\frac{n_2}{2}}}{\left(1-\hat{\rho}\hat{\rho}_{12.3}^{(1)}\right)^{n_1}\left(1-\hat{\rho}\hat{\rho}_{12.3}^{(2)}\right)^{n_2}} \quad (18)$$

where $\hat{\rho}$ is given by (16), $\hat{\rho}_{12.3}^{(1)}$ and $\hat{\rho}_{12.3}^{(2)}$ are given by (7).

Asymptotic distribution of Λ:

Lemma 1. Suppose that A, A_n, $n = 1, 2, \ldots$ are all $p \times p$ symmetric matrices such that $A_n - A = O(\alpha_n)$ and $\alpha_n \to 0$ as $n \to \infty$. Denote by $g(A_n)$ and $g(A)$ as real valued continuous function of A_n and A respectively. Then we have $g(A_n) - g(A) = O(\alpha_n)$ as $n \to \infty$.

Proof: The proof of Lemma 1 can be done the same way as in Zhao, Krishnaiah and Bai (1986).

Lemma 2. Let $\underset{\sim}{X}_1, \underset{\sim}{X}_2, \ldots, \underset{\sim}{X}_n$ be i.i.d. $\sim N_3\left(\underset{\sim}{\mu}, \Sigma\right)$ where, $\underset{\sim}{\mu} = \left(\mu_1\mu_2\mu_3\right)'$ and $\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} \end{pmatrix}$. Then $S \xrightarrow{a.s} \Sigma$ where $S = \frac{1}{n}\sum_{i=1}^{2}\left(\underset{\sim}{X}_i - \bar{X}\right)\left(\underset{\sim}{X}_i - \bar{X}\right)'$

Proof: Proof of Lemma 2 follows from the Strong Law of Large Numbers.

Lemma 3. Let $\rho_{12.3}$ be the partial correlation coefficient between X_1 and X_2, given $X_3 = x_3$. It can be written that $\rho_{12.3} = \frac{\sigma_{12.3}}{\sqrt{\sigma_{11.3}}\sqrt{\sigma_{22.3}}} = \frac{\operatorname{cov}\left(x_{1.3}, x_{2.3}\right)}{\sqrt{\operatorname{var}\left(x_{1.3}\right)\operatorname{var}\left(x_{2.3}\right)}}$ i.e., $\rho_{12.3}$ is a continuous function of Σ. Let $\hat{\rho}_{12.3}$ = estimate of $\rho_{12.3} = \frac{s_{12.3}}{\sqrt{s_{11.3}}\sqrt{s_{22.3}}}$, where $\underset{2x2}{S_{1.2}} = \begin{pmatrix} s_{11} & s_{12} \\ s_{21} & s_{22} \end{pmatrix} - \frac{1}{s_{33}}\begin{pmatrix} s_{13} \\ s_{23} \end{pmatrix}\begin{pmatrix} s_{12} & s_{23} \end{pmatrix} = \begin{pmatrix} s_{11.3} & s_{12.3} \\ s_{21.3} & s_{22.3} \end{pmatrix}$ i.e., $\hat{\rho}_{12.3}$ is a continuous function of S. Then $\hat{\rho}_{12.3} \xrightarrow{a.s.} \rho_{12.3}$.

Proof: Since $\hat{\rho}_{12.3}$ is a continuous function of S and $\rho_{12.3}$ is a continuous function of Σ, the proof of Lemma 3 follows easily from Lemma 1 and Lemma 2.

Theorem 1: Let Λ be the likelihood ratio test statistic given by (18) for testing H_0 vs. H_1. Then $-2\log\Lambda \sim \chi_1^2$ under H_0 as $n_1, n_2 \to \infty$, where the symbol χ_1^2 denotes chi-square distribution with 1 degree of freedom.

Proof: Using Lemma 3, it follows under H_0 that,

$$\begin{aligned} \hat\rho^{(1)}_{12.3} &\xrightarrow{a.s.} \rho \\ \hat\rho^{(2)}_{12.3} &\xrightarrow{a.s.} \rho \\ \hat\rho &\xrightarrow{a.s.} \rho \end{aligned} \tag{19}$$

where, $\hat\rho^{(1)}_{12.3}$, $\hat\rho^{(2)}_{12.3}$, $\hat\rho$ and ρ are given by (7), (8) and (16). Now, the expression of Λ in (18) is asymptotically equivalent to

$$\Lambda = \frac{\left(1-\hat\rho^2\right)^{\frac{n_1+n_2}{2}}\left(1-\rho^2\right)^{\frac{n_1}{2}}\left(1-\rho^2\right)^{\frac{n_2}{2}}}{\left(1-\rho^2\right)^{n_1}\left(1-\rho^2\right)^{n_2}} \text{ (using (19)) } = \frac{\left(1-\hat\rho^2\right)^{\frac{n_1+n_2}{2}}}{\left(1-\rho^2\right)^{\frac{n_1+n_2}{2}}}.$$

Hence, $-2\log\Lambda = -n\left[\log\left(1-\hat\rho^2\right)-\log\left(1-\rho^2\right)\right]$, where $n = n_1 + n_2$

$$= -n\left[\log(1-\hat\rho)+\log(1+\hat\rho)-\log(1-\rho)-\log(1+\rho)\right]$$

$$= -n\left[\left\{\log(1-\hat\rho)-\log(1-\rho)\right\}+\left\{\log(1+\hat\rho)-\log(1+\rho)\right\}\right]$$

$$= -n\left[\begin{array}{l} \left\{(\rho-\hat\rho)\left(-\dfrac{1}{1-\rho}\right)+\dfrac{(\rho-\hat\rho)^2}{2!}\left(-\dfrac{1}{(1-\rho)^2}\right)-O(\alpha_n)\right\} \\ +\left\{(\hat\rho-\rho)\left(\dfrac{1}{1+\rho}\right)+\dfrac{(\hat\rho-\rho)^2}{2!}\left(-\dfrac{1}{(1+\rho)^2}\right)+O(\alpha_n)\right\} \end{array}\right]$$

(by Taylor Series expansion and $\alpha_n \to 0$ as $n \to \infty$)

$$= -n\left[\frac{2(\hat\rho-\rho)}{1-\rho^2}-(\hat\rho-\rho)^2\frac{\left(1+\rho^2\right)}{\left(1-\rho^2\right)^2}\right] \text{ as } n \to \infty$$

$$= -nO(\alpha_n)+\frac{n(\hat\rho-\rho)^2}{\left(1-\rho^2\right)^2}+\frac{n(\hat\rho-\rho)^2\rho^2}{\left(1-\rho^2\right)^2}$$

$$= -nO(\alpha_n)+\frac{\dfrac{(\hat\rho-\rho)^2}{\left(1-\rho^2\right)^2}}{n}+nO(\alpha_n) \tag{20}$$

where $\alpha_n \to 0$ as $n \to \infty$ (using (19)). Since $\dfrac{\sqrt{n}\left(\hat{\rho}-\rho\right)}{1-\rho^2} \xrightarrow{L} N\left(0,1\right)$ as $n \to \infty$ (Anderson (2003), p.133), it is obvious that

$$\frac{\left(\hat{\rho}-\rho\right)^2}{\dfrac{\left(1-\rho^2\right)^2}{n}} \xrightarrow{L} \chi_1^2 \text{ as, } n \to \infty \text{ where } \xrightarrow{L} \text{ denotes convergence in distribution.} \quad (21)$$

Theorem 1 follows from (20) and (21).

Large sample Z-test:

Under this case of p = 3, it can be shown that (Anderson (2003)) for large sample sizes, $z_1 \sim N\left(\varsigma_1, \dfrac{1}{n_1-4}\right)$ and $z_2 \sim N\left(\varsigma_2, \dfrac{1}{n_2-4}\right)$

$$\text{where, } z_1 = \frac{1}{2}\log\frac{1+\hat{\rho}_{12.3}^{(1)}}{1-\hat{\rho}_{12.3}^{(1)}}, \varsigma_1 = \frac{1}{2}\log\frac{1+\rho_{12.3}^{(1)}}{1-\rho_{12.3}^{(1)}}$$

$$\text{and } z_2 = \frac{1}{2}\log\frac{1+\hat{\rho}_{12.3}^{(2)}}{1-\hat{\rho}_{12.3}^{(2)}}, \varsigma_2 = \frac{1}{2}\log\frac{1+\rho_{12.3}^{(2)}}{1-\rho_{12.3}^{(2)}}.$$

The following large sample Z-test for testing H_0 Vs. H_1 is proposed:

$$Z = \frac{\left|z_1 - z_2\right|}{\sqrt{\dfrac{1}{n_1-4}+\dfrac{1}{n_1-4}}} > 1.96 \text{ at 5\% level of significance.} \quad (22)$$

The two tests given by (18) and (22) are compared by power analysis in the next section.

Simulation Results

Multivariate normal random vectors using R program are generated in order to evaluate the power and size of the two tests given by (18) and (22). The R program produced estimates of $\rho_{12.3}^{(1)}, \rho_{12.3}^{(2)}$ and ρ (given by (7) and (16)) along with the Approximate Likelihood Ratio (ALR) statistic given by (18) and the

Z-statistic given by (22) 5,000 times for each particular combination of population parameters ($\rho_{12.3}^{(1)}$ and $\rho_{12.3}^{(2)}$). The frequency of rejection of each test statistic at $\alpha = 0.05$ was noted and the proportion of rejections (power) are reported in Table 1 for various combinations of ρ_1 and ρ_2 ($\rho_{12.3}^{(1)}$ and $\rho_{12.3}^{(2)}$).

On the basis of our study, it is found that the ALR-test showed consistently better results than Z - test in terms of power. The size of the ALR test is slightly more than alpha level. The ALR test is recommended strongly for use in practice

Table 1. Empirical significance level and power of the Approximate Likelihood Ratio (APR) test and the Z-test (ZT) for p = 3 and α = 0.05

					*N*₁=*N*₂=25					
ρ_1	ρ_2	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0.1	ALR	0.050	0.067	0.107	0.187	0.307	0.453	0.639	0.758	0.998
	ZT	0.050	0.054	0.068	0.101	0.158	0.259	0.431	0.684	0.945
0.3	ALR	0.100	0.062	0.054	0.069	0.125	0.242	0.414	0.640	0.981
	ZT	0.069	0.055	0.049	0.057	0.086	0.148	0.280	0.530	0.884
0.5	ALR	0.300	0.195	0.127	0.074	0.053	0.081	0.178	0.420	0.727
	ZT	0.159	0.119	0.085	0.061	0.051	0.066	0.130	0.319	0.743
0.7	ALR	0.621	0.528	0.426	0.297	0.184	0.089	0.052	0.118	0.472
	ZT	0.425	0.354	0.279	0.202	0.131	0.076	0.050	0.101	0.428
0.9	ALR	0.998	0.995	0.981	0.956	0.902	0.782	0.586	0.282	0.063
	ZT	0.945	0.920	0.884	0.828	0.742	0.613	0.429	0.202	0.050
					*N*₁=25, *N*₂=50					
0.1	ALR	0.059	0.077	0.143	0.263	0.469	0.670	0.876	0.977	0.999
	ZT	0.050	0.056	0.076	0.119	0.202	0.338	0.552	0.814	0.986
0.3	ALR	0.122	0.065	0.052	0.065	0.145	0.284	0.489	0.721	0.802
	ZT	0.076	0.058	0.049	0.060	0.099	0.186	0.364	0.661	0.959
0.5	ALR	0.466	0.305	0.186	0.095	0.059	0.010	0.257	0.610	0.965
	ZT	0.200	0.145	0.098	0.064	0.050	0.072	0.163	0.414	0.863
0.7	ALR	0.881	0.782	0.633	0.453	0.279	0.128	0.066	0.166	0.692
	ZT	0.548	0.461	0.363	0.261	0.164	0.085	0.050	0.120	0.551
0.9	ALR	1.000	0.999	0.997	0.992	0.960	0.885	0.707	0.356	0.069
	ZT	0.987	0.977	0.958	0.925	0.864	0.750	0.550	0.261	0.050

Acknowledgement

Authors would like to thank Mr. Mohammad Aziz for computational help.

References

Anderson, T. W. (2003). *An introduction to multivariate statistical analysis* (3rd ed.). Hoboken, N.J: John Wiley & Sons.

Fisher, R. A. (1924). The distribution of the partial correlation coefficient. *Metron, 3*, 329-332.

Gupta, A. K. and Kabe, D. G. (2001). On testing the equality of K multiple and partial correlation coefficients. *ACTA Mathematica Scientia, 21*(2), 221-223.

Zhao, L. C., Krishnaiah, P. R. & Bai, Z. D. (1986). On detection of number of signals in presence of white noise. *Journal of Multivariate Analysis, 20*(1), 1-25. doi:10.1016/0047-259X(86)90017-5

Journal of Modern Applied Statistical Methods
May 2015, Vol. 14, No. 1, 83-106.

ISSN 1538 − 9472

Convergent and Discriminant Validity with Formative Measurement: A Mediator Perspective

Xuequn Wang
Murdoch University
Perth, Australia

Brian F. French
Washington State University
Pullman, WA

Paul F. Clay
Fort Lewis College
Durango, CO

The ability to validate formative measurement has increased in importance as it is used to develop and test theoretical models. A method is proposed to gather convergent and discriminant validity evidence of formative measurement. Survey data is used to test the proposed method.

Keywords: Causal indicators, formative measurement, construct validity, convergent validity, discriminant validity, mediator

Introduction

There has been a vigorous debate and discussion about the issues surrounding the application of formative measurement (Bollen, 2007; Howell et al., 2007a, 2007b; Petter et al., 2007) and how to validate this specific kind of measurement model (Hardin et al. 2011). Because procedures used to validate reflective measurement are not appropriate for formative measurement, there is a need to develop measurement theory to validate formative measurement (Hardin et al., 2011).

Formative measurement has been applied in multiple disciplines, including Marketing (e.g., Chandon et al., 2000), Entrepreneurship (e.g., Brettel et al., 2011), and Information Systems (IS) (e.g., Pavlou & Gefen, 2005). For example, Pavlou and Gefen (2005) measured perceived effectiveness of institutional structures with formative measurement, which included four dimensions: feedback technologies, escrow services, credit card guarantees and trust in intermediary.

Dr. Wang is a Lecturer in School of Engineering and Information Technology. Email at xuequnwang1600@gmail.com. Dr. French is a Professor and Director of the Learning & Performance Research Center. Email at frenchb@wsu.edu. Dr. Clay is an Assistant Professor in the School of Business. Email at pfclay@fortlewis.edu.

Although some researchers question the appropriateness of such models (e.g., Edwards, 2011), others have shown that formative measurement can be appropriate in certain contexts. For example, for multidimensional constructs, causal indicators can be developed to "comprise all essential aspects of the focal construct's definition" (MacKenzie et al., 2011, p. 304).

Using only global reflective indicators may, however, "diminish the correspondence between the empirical meaning of the construct and its nominal meaning, because there is no way to know whether the respondent is considering all of the subdimensions (facets) of the focal construct that are part of the nominal definition when responding to the global question" (MacKenzie et al., 2011, p. 327). Therefore, though there remain several issues related to the adoption of formative measurement, given that formative measurement can be appropriate in many contexts (Cadogan & Lee, 2013; Diamantopoulos et al., 2008; Jarvis et al., 2003; MacKenzie et al., 2011), developing corresponding methods is necessary so that researchers can validate formative measurement.

There are multiple aspects of construct validity that require evaluation using various methods to develop and maintain a strong validity argument. Having such evidence does not and cannot rely on a single method. According to Messick (1995), there are six aspects of construct validity: content, substantive, structural, generalizability, external, and consequential aspects of construct validity. In this paper, external aspect of validity evidence is focused upon, which deals with "convergent and discriminant evidence" (Messick, 1995, p. 745). More recently, Cizek et al. (2008) examined various aspects of validity from previously published indicators. They discussed validity including the traditional division of construct validity evidence (convergent and discriminant evidence), criterion-related evidence, content evidence, evidence based on response process, evidence based on consequences, face validity evidence and evidence based on internal structure, supporting the need for various forms of evidence. In this study associations with other variables (convergent and discriminant evidence) rather than all possible sources of validity evidence is focused on. Note that this is only one step toward developing a comprehensive validity argument to support inferences from formative measurement.

Previous studies have paid little attention to convergent and discriminant validity of formative measurement (Bollen, 2011). This may be attributed to the fact that formative measurement is quite different from reflective measurement. Although there are relatively mature and sophisticated methods to gather convergent and discriminant validity evidence for reflective measurement based on classical test theory (CTT) (Kane, 2006), there lacks an agreed method or set of procedures to gather convergent and discriminant validity evidence for formative measurement (Barki et al., 2007; Diamantopoulos & Winklhofer, 2001; Jarvis et al., 2003; Petter et al., 2007). Thus, a researcher and practitioner can often faces difficulty in dealing with convergent and discriminant validity when one moves from reflective measurement to formative measurement (Diamantopoulos et al., 2008).

In this study, constructs are used to refer to "a conceptual term used to describe a phenomenon of theoretical interest" (Edwards & Bagozzi, 2000, p. 156-157), and latent variable is used to refer to the representation of a certain construct in a model. Indicators are used to refer to "observed variables that measure a latent variable" (Bollen, 2011, p. 360). The kind of indicators depends on "whether the indicator is influenced by the latent variable or vice versa" (Bollen, 2011, p.360). Reflective indicators are used to refer to those influenced by the latent variable, and causal indicators are used to refer to those influencing the latent variable.

The focus in this study is on formative measurement with causal indicators. As Bollen (2011) illustrated, formative measurement may include causal indicators or formative indicators. The key difference between these two types of indicators is that "causal indicators should have conceptual unity in that all the variables should correspond to the definition of the concept whereas formative indicators are largely variables that define a convenient composite variable where conceptual unity is not a requirement" (Bollen, 2011, p. 360). Variables consisting of formative indicators may not have any meaningful conceptualization. Therefore, formative measurement with causal indicators is focused upon in this study (Bollen, 2011).

Although formative measurement have been recognized in the literature (Diamantopoulos et al., 2008); there are no agreed upon methods to provide convergent and discriminant validity evidence for formative measurement. Because construct validity is "a necessary condition for theory development and testing" (Jarvis et al., 2003, p. 199), it is important to gain validity evidence before one tests theory. This paper adds to the current validity literature by proposing and testing a method to gain validity evidence (convergent and discriminant evidence) for formative measurement. Note that the proposed method does not aim to challenge or replace CTT when testing reflective measurement. After testing our method with real data for formative measurement, construct validity for reflective measurement is also examined following our new method. The results from our method and those from Confirmatory Factor Analysis (CFA) are consistent.

Reflective vs. Formative Measurement

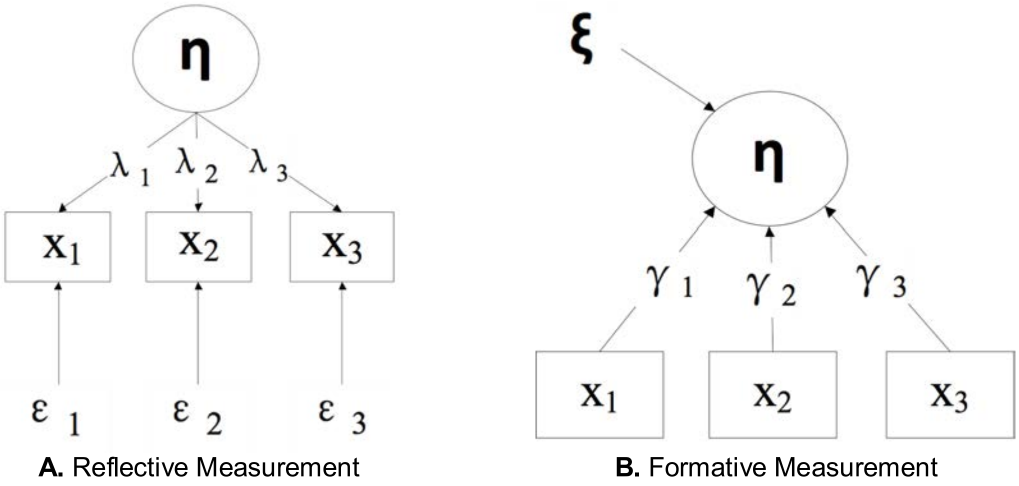


Figure 1. Two kinds of measurement models.

Many measurement models that social science deals with are reflective (Panel A from Figure 1; Diamantopoulos et al., 2008, Petter et al., 2007). For reflective measurement, the direction of causality is from the latent variable to the indicators. Because all indicators are the effects of the same latent variable, they are expected to be highly correlated (internal consistency reliability) (Bollen, 1984). The deletion of an indicator will probably not alter the meaning of the latent variable given that there are sufficient and similar functioning indicators to represent the latent variable. Ideally the indicators are interchangeable. Measurement errors are taken into account at the indicator level (c.f. Edwards and Bagozzi (2000), Jarvis et al. (2003), MacKenzie et al. (2005), for a more detailed description). Thus, the equation for a measurement model with reflective indicators is given as (Bollen & Lennox, 1991):

$$x_i = \lambda_i \eta + \varepsilon_i \qquad (1)$$

where η is the latent variable, x_i is the i^{th} reflective indicator for the latent variable η, λ_i represents the effect of η on that indicator (coefficient) and ε_i is the measurement error for x_i.

In contrast, for formative measurement the latent variable is influenced by these causal indicators (Bollen, 1984; Chin, 1998). Thus, deleting an indicator will alter the meaning of the latent variable (Bagozzi, 2007; Bollen, 2007; Diamantopoulos et al., 2008; Howell et al., 2007b; Jarvis et al., 2003). Additionally, there is no reason to expect that these causal indicators are necessarily highly correlated with each other, which makes internal consistency reliability inappropriate. Unlike reflective indicators, causal indicators are assumed to be error free (c.f. Edwards and Bagozzi (2000), Jarvis et al. (2003), and MacKenzie et al. (2005)) and that there may be a disturbance term representing "non-modeled causes" (Diamantopoulos, 2006, p. 7). Thus, the equation for a measurement model with causal indicators is (Bollen & Lennox, 1991):

$$\eta = \gamma_1 x_1 + \ldots + \gamma_i x_i + \zeta \tag{2}$$

where η represents the latent variable, x_i is the i^{th} causal indicator for latent variable η, γ_i represents the path weights for indicators x_i and ζ is the disturbance term which includes other variance not accounted for by the indicators (MacKenzie et al., 2005). For example, job satisfaction can be measured with indicators such as "I am very satisfied with my pay", "I am very satisfied with the nature of my work", and "I am very satisfied with my opportunities for promotion", and so on, and these three indicators influences one's job satisfaction level (MacKenzie et al., 2011). Because the covariance between causal indicators could be any value, the way to examine construct validity (convergent validity and discriminant validity) for reflective measurement based on CTT (e.g., CFA) cannot be used. Therefore, a new method is required to validate formative measurement.

For reflective measurement, convergent evidence is provided when "different indicators of theoretically similar or overlapping constructs are strongly interrelated" (Brown, 2006, p. 2), and discriminant evidence is provided when "indicators of theoretically distinct constructs are not highly intercorrelated" (Brown, 2006, p. 3). In other words, convergent validity essentially refers to whether indicators from a latent variable do belong to that latent variable, and discriminant validity essentially refers to whether indicators from a latent variable do not belong to other latent variables.

However, for formative measurement, high correlations are *not* required between its indicators (Jarvis et al., 2003). Furthermore, correlations among causal indicators within a measurement model need not be higher compared to correlations between them and indicators from other measurement models (Bollen, 2011; Bollen & Lennox, 1991). Therefore, the traditional approach toward establishing convergent and discriminate validity from CTT is not appropriate. In this study, an adaptation of the definition of convergent and discriminant validity is proposed to accommodate the context of formative measurement. Convergent validity is used to specify that causal indicators from a measurement model should explain a significant proportion of variance from the latent variable that they measure; discriminant validity is used to specify that these same indicators should explain a much lower proportion of variance from other latent variables. That is, indicators that are associated with the target latent variable will explain much more variance of that latent variable and those indicators should not explain a large amount of variance of other latent variables relative to the target latent variable.

These definitions adapt Brown (2006)'s definition by reversing the direction of relationship between the latent variable and the indicators. Discriminant evidence is particularly important because it indicates that these indicators do not belong to other latent variables.

The Context of Validation

Identification is always an issue for structural equation models with latent variables, and there are two general identification rules: First, each latent variable must be assigned a scale; Second, the number of free parameters estimated in a model must be no more than the number of unique pieces of information in the covariance matrix of manifest variables (Bollen & Davis, 2009). Thus, for a reflective measurement model, the minimum number of indicators should be at least three. However, there is one more identification requirement raised by formative measurement. MacCallum and Browne (1993) showed that an additional requirement for the identification of the disturbance from formative measurement was that the latent variable measured by causal indicators must emit two paths to its reflective indicators or other latent variables. Therefore, a model is proposed in which the latent variable measured by causal indicators predicts two or more outcome variables measured by reflective indicators as the context in which to gather convergent and discriminant validity evidence (Bollen & Davis, 2009). Our model is consistent with the circumstances identified by Bagozzi (2011) under which formative measurement are appropriate to be used.

The example model proposed is shown in Figure 2, where latent variable η_1 is measured by causal indicators and its convergent and discriminant validity evidence is to be examined. Note that the actual research model may be different from this test model: The model is used to gather convergent and discriminant validity evidence only; and its structural paths may differ widely from those of the research model. What the model is trying to do is to examine the indicators from latent variable η_1 in terms of convergent and discriminant validity.

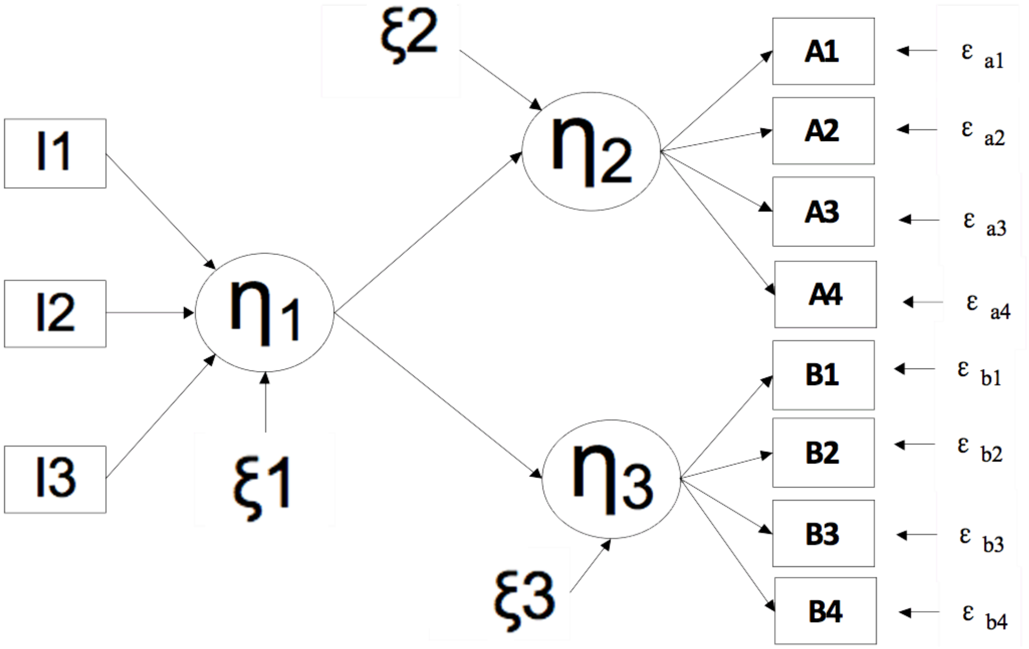


Figure 2. An example model of the proposed method.

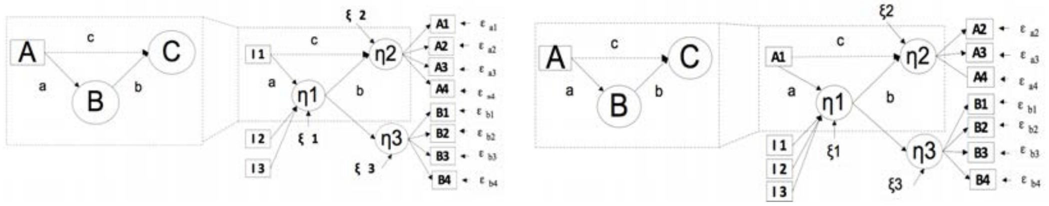
A Mediator Perspective

Psychologists have recognized the concept of a mediator for quite a long time (e.g., Woodworth, 1928). Furthermore, Baron and Kenney (1986) clarified the nature of a mediator: a given variable functioned as a mediator if it accounted for the relationship between an independent variable and a dependent variable. To be a mediator, a variable needs to meet three conditions: (a) Variance of independent variable *A* significantly accounts for variance of mediator *B*. In other words, the path coefficient of Path A is significant. (b) Variance of mediator *B* significantly accounts for variance of the dependent variable *C*. In other words, the path coefficient of Path B is significant. (c) When Paths A and B are controlled, the previous significant relation (Path C) between the independent variable *A* and dependent variable *B* significantly decreases (or even becomes zero).

By applying the mediator perspective, the relevant latent variable η_1 can be seen as a mediator which accounts the influence of causal indicators I1-I3 on the other latent variables (e.g., η_2; Panel A from Figure 3) (Bollen, 2007; Bollen & Davis, 2009; Howell et al., 2007b). Then, latent variable η_1's construct validity (i.e., convergent and discriminant evidence) can be examined. Note that our method is justified based on previous literature. Bollen (2007), for example, argued that the latent variables measured by causal indicators mediated "the effect of causal indicators on these other variables" (p. 222). MacKenzie et al. (2011) also argued that "the adequacy of the hypothesized multidimensional structure can be assessed by testing whether the sub-dimensions of the multidimensional focal construct have significant direct effects on a consequence construct, over and above the direct effect that the focal construct has on the consequence" (p. 323). Specifically, the causal indicators "*must* share the latent variable η as a common consequence and, moreover, η must fully mediate the effects of" their indicators "on other observed or latent variables that are modeled as outcomes of η" (Diamantopoulos, 2011, p. 340). Also as Franke et al. (2008, p. 1230) argued, the latent variables measured by causal indicators "mediate the effects of their indicators on other variables, constraining their indicators to have the same proportional influence on the outcome variables….If the formative indicators could have direct as well as mediated effects on the outcome variables, then the proportionality constraint would not necessarily hold". (Here formative indicators refer to causal indicators in Bollen (2011)'s terminology.)

In the proposed method, the validity of formative measurement is supported even if causal indicators have direct influence on the outcomes variables, as long as "the magnitude of the effect of the focal construct on the consequence construct is substantially larger than the combined magnitudes of the direct effects" of its indicators on the outcome variables (MacKenzie et al., 2011, p. 323). In other words, the latent variable can fully or partially mediate the influence of causal indicators I1-I3 on latent variable η_2. It is similar to the context in which the research model only contains reflective measurement and construct validity is supported even if cross-loadings exist as long as these cross-loadings are much less then loadings between reflective indicators and the focal latent variables.

Therefore, to gather η_1's convergent evidence, if indicator I1 indeed belongs to η_1, the influence of I1 on η_2 should be mediated by η_1 (Panel A from Figure 3). In other words, I1 should explain a significant amount of variance of η_1. That is consistent with the definition of formative measurement: Indicator I1 influences η_1, and then η_1 influences η_2. Following Baron and Kenny's instruction, we can examine convergent validity in three steps. See Table 1 for each step. Especially, significant indicator weight is the first step. If indicator weights (Path A) are not significant, there is no need to go further, given that the strength of indicator weight is the statistical metric used to judge indicator retention (Bollen & Lennox, 1991; Chin, 1998; Diamantopoulos et al., 2008; Diamantopoulos & Winklhofer, 2001).



A. Convergent Validity

B. Discriminant Validity

Figure 3. A mediator perspective.

Table 1. A mediator perspective to gather validity evidence for formative measurement.

Step	Description
Step 1	Examine if path coefficient for Path A is significant • If path coefficient for Path A is not significant, then I1 does not significantly cause η_1. There is no need to go further. • If path coefficient for Path A is significant, then
Step 2	Examine the coefficient for Path C (without controlling B) • If path coefficient for Path C is not significant, then I1 and η_2 do not share a significant amount of variance. There is no need to go further. • If path coefficient for Path C is significant, then
Step 3	Examine the coefficient for Path C by controlling A and B • If path coefficient for Path C becomes less or insignificant, then η_1 mediates the influence of I1 on η_2. Therefore I1 probably belongs to η_1. • If path coefficient for Path C remains the same or changes little, then η_1 does not mediate the influence of I1 on η_2. Therefore I1 may not belong to Y_1.

To gather η_1's discriminant evidence, the same process is gone through by examining if η_1 mediates indicators from other measurement models. For example, indicators A1-A4 from latent variable η_2 can be examined and confirmed that η_1 cannot mediate these indicators' influences on η_2 (Panel B from Figure 3). Indicators from η_2 should explain a much less amount of variance of η_1 than I1 - I3. The same process in Table 1 is followed. When path coefficient for Path C is tested controlling for Path A and Path B, if path coefficient for Path C does not change significantly, then the influences of indicator A1- A4 are not mediated by η_1. Therefore, indicators A1- A4 do not belong to η_1. In contrast, if the path coefficient for Path C reduces significantly or even becomes insignificant, A1- A4 may belong to η_1. Here content analysis is needed to further examine these indicators, and indicators A1- A4 are problematic in the sense that the results are not consistent with developed theory.

Methodology

Participants

Participants (*N* = 337) from an entry level business class at a large state university in the Northwest of the U.S. completed the scales described below. The demographic information collected includes age and gender. The mean age of the participants was 20.35, with the range between 18 and 36 years. The percentage of male students was 62.00%.

Measures

Perceived Effectiveness of Institutional Structures (PE) (Pavlou & Gefen, 2005), a correctly modeled formative measurement (Petter et al., 2007), was selected as our example of formative measurement. Two other constructs (Trust and Trust Propensity (TP), where Trust is Trust in the Community of Sellers, and TP is Trust Propensity). For a detailed description of PE, Trust and TP and their indicators, please refer to Pavlou and Gefen (2005).) were chosen to form the model to test in Figure 2. The instruments from original studies were adapted to fit the new study environment. The indicators of PE and Trust were reworded to focus on online shopping behaviors.

Procedures

Participants were given class credit to participate in the study (less than 1% of their final grade) with other options if they selected not to participate. Data collection occurred in laboratories for the business class. After participants arrived in the laboratories, the administrator read aloud the purpose and procedures for the study. Then participants accessed a website to complete the questionnaire. The questionnaire contained a randomized sequence of indicators from PE, Trust, TP and other constructs from Pavlou and Gefen (2005) as well as demographic information questions. Once the questionnaire was completed (about 10 mins), participants were thanked and exited the laboratory.

Data Analysis

M*plus* (Muthén & Muthén, 1998-2012) was used to analyze the data. Our analysis had two components. First, our proposed method was tested with the model including PE, Trust and TP. Second, the proposed method was applied to gain convergent and discriminant evidence for Trust, to show that the proposed method is consistent with CTT when examining measurement models with reflective indicators.

For the first component of the analysis, CFA was first performed to gather the convergent and discriminant evidence of the two latent variables measured by reflective indicators: Trust and TP (Brown, 2006). The global fit was assessed and the following fit indices were used: chi-square statistic (χ^2), Comparative Fit Index (*CFI*), and the Standardized Root Mean Squared Residual (*SRMR*). The χ^2 test is significant when p value is less than 0.05. In such contexts, the model may not represent data reasonably well. *CFI* equal to or greater than .90 indicates reasonable global fit (Rigdon, 1996). The *SRMR* less than .05 indicates acceptable fit (Byrne, 1998). Because the result of chi-square test is likely inflated by sample size, the result of χ^2 test is routinely significant with large sample size, even if the differences between S and Σ are negligible (Brown, 2006). Therefore, other fit indices were used in combination with the chi-square test. Standardized loadings were then used to gather the convergent evidence and cross loadings were used to gather the discriminant evidence. For the size of item loadings, suggestions given by Straub et al. (2004) were followed, who suggest that loadings should be "above .707 so that over half of the variance is captured by the latent construct" (p. 410).

Next the model including PE, TP and Trust was examined to gather convergent and discriminant validity evidence for PE, which is measured by causal indicators. The global fit of the model was first examined. Here acceptable overall goodness of model fit is important to show that the baseline model can fit the data well (Brown, 2006). The convergent and discriminant validity evidence for PE was then gathered following the method proposed above (refer to Table 1).

For convergent evidence, proposed indicators for PE should converge on PE. From a mediator perspective, PE should mediate the influence of its indicators on the other two latent variables (Figure 4). For discriminant evidence, indicators from other measurement models should not belong to PE. From a mediator perspective, PE should not mediate the influence of indicators from other latent variables on these two latent variables.

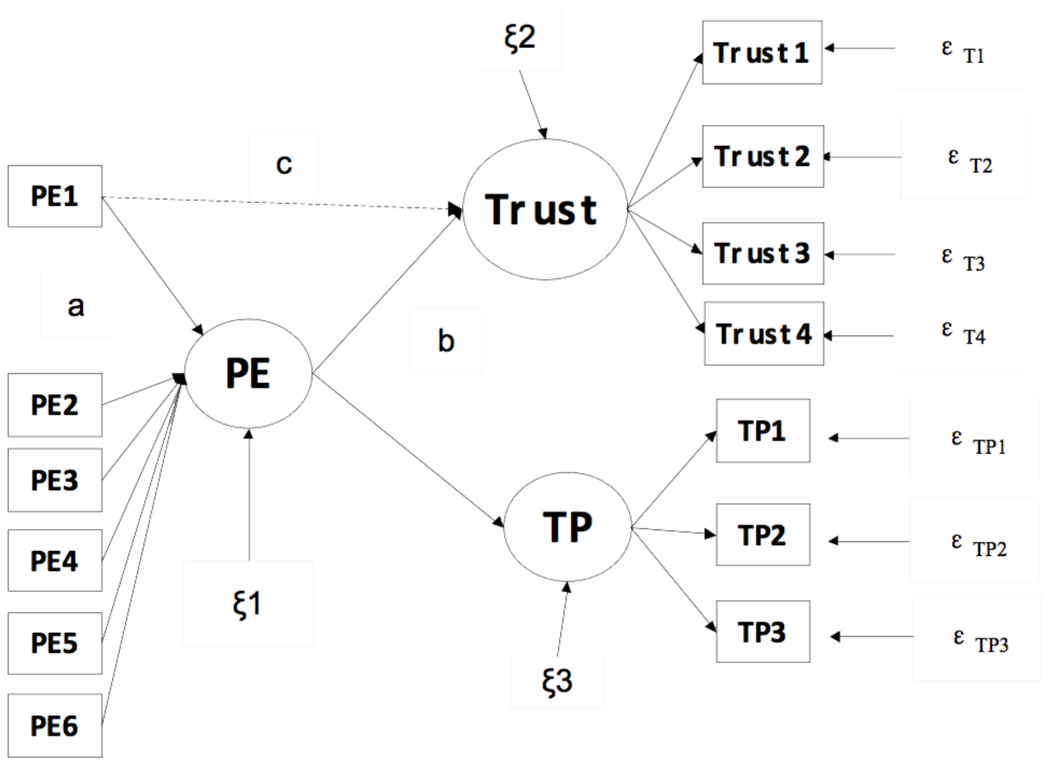


Figure 4. Model to gather convergent and discriminant evidence for PE.

In the second component of the analysis, the convergent and discriminant validity evidence of Trust were gathered with the method proposed in this study.

These analyses demonstrated that our proposed method was consistent with CTT when gathering convergent and discriminant evidence from reflective measurement as well. First convergent validity of Trust was examined to check if Trust1-Trust4 belonged to Trust (Figure 5). Next discriminant validity was examined to check if TP1-TP3 belonged to Trust.

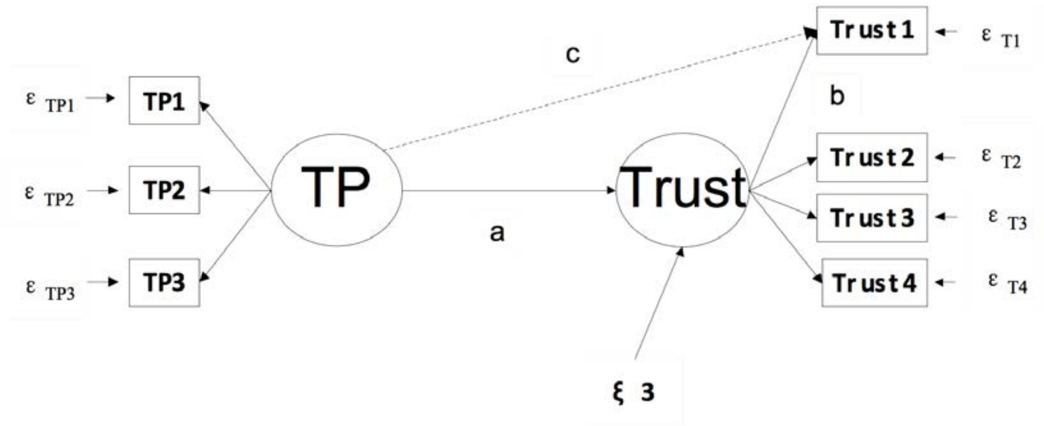


Figure 5. A mediator method to gather convergent and discriminant evidence for trust.

Results

CFA

The global fit of the model was acceptable ($\chi^2(13) = 85.779$, $NC = 6.60$, $p < 0.0001$, $CFI = 0.943$, $SRMR$ is 0.040). Although the result of χ^2 test was significant, it was largely due to the large sample size (337). Other fit indices met stated criteria.

For convergent evidence, indicators' standardized loadings were examined. The standardized loadings for all indicators are shown in Table 2: all loadings were significant and most loadings were above 0.707 (except for Trust2 and TP2), which indicates that the latent variables explain more than 50% of variance for most indicators. This indicated reasonable convergent evidence. For discriminant evidence, the cross loadings between indicators and other latent variables were examined, requiring that indicators load much higher on the latent variables they measure than on other latent variables (Gefen & Straub, 2005). From the results of Modification Indices (M.I.), no M.I.s for cross loading are significant, indicating good discriminant evidence. (In M*plus*, M.I. is the amount chi-square which would drop if the parameter is estimated as part of the model. 3.84 is the chi-square value which is significant at the .05 level for one degree of freedom. When the M.I. is significant, we also want to examine the size of completely standardized expected parameter change. Usually, values more than 0.300 are considered large and should be included in the model. Value less than 0.200 indicates a trivial change of parameter, and we may not include it into the model, even if M.I. is significant.) To summarize, Trust and TP have good convergent and discriminant evidence.

Table 2. Loadings.

	Trust		TP
Trust1	0.786	*TP1*	0.750
Trust2	0.687	*TP2*	0.595
Trust3	0.907	*TP3*	0.803
Trust4	0.928		

Construct Validity (Convergent and Discriminant Evidence): Formative Measurement

The fit for baseline model was first examined. The model met fit criteria ($\chi^2(48) = 145.439$, $p < 0.0001$, $NC = 3.03$, $CFI = .92$, SRMR is 0.039). Therefore, the global fit of baseline model was reasonable.

The method outlined in Table 1 was followed. For convergent validity, PE1-PE6 were considered as independent variable, PE as the mediator, and Trust (or TP) as the dependent variable. In the first model (Trust as the dependent variable, refer to Table 3), the path coefficient for Path A was first examined. According to the second column, the path coefficients from PE1 and PE6 to PE were significant, indicating that PE1 and PE6 significantly influenced PE in this context. Next, the path coefficient for Path C was examined, without controlling Path A. According to the forth column, path coefficients from PE1 and PE6 to Trust were significant, indicating that the PE1 and PE6 explained a significant amount of variation of Trust. Finally, the path coefficient for Path C was examined, controlling Path A and B. According to the third column in Table 3, the path coefficient for Path B (from PE to Trust) was significant. According to the last column, when controlling Path A and Path B, all path coefficients were insignificant, indicating that there were no direct effects from PE1 and PE6 to Trust. Therefore, PE fully mediated the influence of PE1 and PE6 on Trust. In the second model (TP as the dependent variable, refer to Table 4), the same procedures were followed, and the results also indicated full mediation. Specially, path coefficients for Path C were not significant according to the forth column, indicating that PE1 and PE6 could not explain a significant amount of variance of TP even before controlling Path A and Path B. Therefore, PE1 and PE6 belonged to PE, indicating good convergent evidence.

Table 3. Path coefficient between PE, PE's indicators and Trust.

	Path A	Path B	Path C (before controlling Path A)	Path C (after controlling Path A)
PE1	0.239*	0.764*	0.148*	0.082
PE2	0.173	0.764*	-	0.098
PE3	0.142	0.764*	-	-0.131
PE4	0.046	0.764*	-	-0.136
PE5	-0.020	0.764*	-	0.007
PE6	0.355*	0.764*	0.163*	0.000

*Note: $p < 0.05$

Table 4. Path coefficient between PE, PE's indicators and TP.

	Path A	Path B	Path C (before controlling Path A)	Path C (after controlling Path A)
PE1	0.239*	0.629*	0.011	-0.069
PE2	0.173	0. 629*	-	-0.094
PE3	0.142	0. 629*	-	0.097
PE4	0.046	0. 629*	-	0.099
PE5	-0.020	0. 629*	-	-0.004
PE6	0.355*	0. 629*	0.091	0.001

*Note: $p < 0.05$

For discriminant validity, Trust1-Trust4 were considered as independent variable, PE as the mediator, and Trust as the dependent variable (refer to Table 5). First, the path coefficient for Path A was examined. According to the second column, path coefficients from Trust1-Trust4 to PE were significant, indicating that Trust1-Trust4 significantly influenced PE. Next, the path coefficient for Path

C was examined, without controlling Path A. According to forth column, Trust1-Trust4 significantly influenced Trust.

Finally, the path coefficient for Path C was examined, controlling Path A and Path B. According to third column, the path coefficient for Path B (from PE to Trust) was significant. According to the last column, path coefficient for Path C (from Trust1-Trust4 to Trust) was still significant and decreased little after controlling for Path B, indicating that PE did not mediate the influence of Trust1-Trust4 on Trust. Therefore, indicators Trust1-Trust4 did not belong to PE, and discriminant evidence was supported.

Table 5. Path coefficient between PE, Trust and Trust's indicators.

	Path A	Path B	Path C (before controlling Path A)	Path C (after controlling Path A)
Trust1	0.755*	0.967*	0.715*	0.636*
Trust2	0.633*	0. 931*	0.575*	0.445*
Trust3	0.867*	0. 964*	0.837*	0.620*
Trust4	0.883*	0. 985*	0.868*	0.678*

*Note: $p < 0.05$

Another evidence of discriminant validity was that after adding Trust1 (to Trust4) to PE, the path coefficient from PE to Trust was more than 0.900, indicating bad discriminant validity (Now PE and Trust cannot discriminate from each other). Therefore, to keep PE as a meaningful and separate latent variable, Trust1 (to Trust4) should be removed from PE. However, this argument should be based on the previous step in that PE could mediate several indicators' influence on Trust and TP. If PE could not function as mediator in previous steps, then indicators could be problematic.

Construct Validity (Convergent and Discriminant Evidence): Reflective Measurement

In this section the proposed method was applied to gather convergent and discriminant evidence of reflective measurement (Trust), to confirm that Trust1-Trust4 belonged to Trust and TP1-TP3 did not belong to Trust. To gather convergent evidence, TP was considered as the independent variable, Trust as the mediator and Trust1-Turst4 as the dependent variable (refer to Table 6).

Table 6. Path coefficient between Trust, Trust's indicators and TP.

	Path A	Path B	Path C (before controlling Path A)	Path C (after controlling Path A)
Trust1	0.473*	0.786*	0.375*	0.004
Trust2	0.473*	0.687*	0.321*	-0.006
Trust3	0.473*	0.907*	0.435*	0.012
Trust4	0.473*	0.928*	0.435*	-0.011

*Note: $p < 0.05$

The path coefficient for Path A was first examined. According to the second columns in Table 6, the path coefficients were significant and not more than 0.800, which indicated that TP explained a significant amount of variance of Trust, and TP and Trust were discriminant from each other. Next the path coefficient for Path C was examined, without controlling Path A. According to the forth column, path coefficients for Path C were significant, indicating that Trust1-Trust4 loaded on TP significantly. Finally, the path coefficient for Path C was examined, controlling Path A and Path B. According to the third column, path coefficients for Path B were significant and more than 0.707 (except for Trust2). According to the last column, all path coefficients for Path C were insignificant, which indicated that Trust fully mediated TP's effect on Trust1-Trust4. Therefore, good convergent evidence was supported.

To gather discriminant evidence, TP was considered as the independent variable, Trust as the mediator and TP1-TP3 as the dependent variable (refer to Table 7). The path coefficient for Path A was first examined. According to the second column, the path coefficient was significant and less than 0.800, indicating that TP explained a significant amount of variance from Trust, and they were discriminant from each other. Next, the path coefficients for Path C were examined, without controlling Path A. According to the forth column, path coefficients for Path C were all significant, indicating that TP1-TP3 loaded on TP significantly. Finally, the path coefficients for Path C was examined, controlling Path A and Path B. According to the third column, the path coefficients for Path B (from Trust to TP1-TP3) were significant. However, no path coefficients (loading) were more than 0.707. According to the last column, all path coefficients for Path C were significant and decreased little, indicating Trust could not mediate TP's effect on TP1-TP3. Therefore, TP1-TP3 did not belong to Trust. Thus, good discriminant evidence was supported.

Table 7. Path coefficient between Trust, TP and TP's indicators.

	Path A	Path B	Path C (before controlling Path A)	Path C (after controlling Path A)
TP1	0.437*	0.432*	0.750*	0.642*
TP2	0.500*	0.269*	0.595*	0.625*
TP3	0.525*	0.366*	0.803*	0.920*

*Note: $p < 0.05$

To summarize, our results showed that Trust1-Trust4 are indicators of Trust but TP1-TP3 were not. These conclusions are consistent with the results of CFA in the framework of CTT. Therefore, the method proposed is consistent with CTT when we gather convergent and discriminant evidence for reflective measurement.

Discussion

Formative measurement has been recognized in previous literature (Bollen, 1984; Bollen, 2011; Petter et al., 2007; Wang, Jessup, & Clay, 2015). However, there has not been an agreed method to gain convergent and discriminant validity evidence for formative measurement. The purpose of this study was to propose a method to gain convergent and discriminant evidence for formative measurement. A mediator perspective was adopted to propose a series of steps to test the validity of formative measurement. The data collected supports our method and showed that the method could keep those indicators which should belong to a formative measurement model and teasing out those which should not be part of the measurement. Our method can guide further social and behavioral research on how to gather convergent and discriminant validity evidence for formative measurement, and contribute a potential solution to one of the issues surrounding the application of formative measurement raised by recent literature (Edwards, 2011).

It is admitted that conclusions drawn from our method are dependent upon the data from a single example with one data set. In the results above that we showed that PE2, PE3, PE4 and PE5 did not significantly influence PE. Therefore, those four indicators may not belong to PE. However, the decision whether PE2, PE3, PE4 and PE5 are to be retained based on statistical results (convergent and discriminant validity) and other validity evidences (e.g., content validity) would be necessary. Any scale refinement should be based on both empirical and theoretical information and not rely solely on empirical data. For formative measurement, indicator weights are dependent on specified structural models (Bollen &Davis, 2009), and the relative contribution of indicator weights is model dependent (Bollen et al., 2001; Hauser & Warren, 1997). Therefore, the choice should be based on “theoretical relevance” (Cenfetelli & Bassellier, 2009). If PE2, PE3, PE4 and PE5 represent unique and important domain of PE, they should be kept despite the fact that they do not significantly influence PE in this context with an eye in refining how they are assessed.

Because the procedures of measurement development and validation are quite complex, researchers may find that the focal latent variable cannot mediate the relationship between certain causal indicators and outcome variables. Consider the context with reflective measurement only. Even if researchers have followed strict procedures to develop indicators, it is still possible for several reflective indicators to have insufficient discriminant validity (e.g., cross-loadings are high) (MacKenzie et al., 2011). Based on previous discussions, cross-loadings for reflective indicators are similar to direct effects which cannot be mediated by the latent variable from a formative measurement model (Figure 4 and 5). When the latent variable measured with causal indicators cannot mediate the relationship between certain causal indicators and outcome variables, these corresponding indicators are problematic (Diamantopoulos, 2011; MacKenzie et al., 2011). Our method can detect these indicators and warn researchers that their measurement models are not be supported.

Limitation and Directions for Future Research

A few limitations should be recalled when applying the proposed method. First, the application of statistical testing is based on relevant literature (e.g., Bollen, 1989; Bollen & Lennox, 1991). As MacKenzie et al. (2011) argue, “indicator validity is captured by the significance and strength of the path from the indicator to composite latent construct” (p. 315). Bollen (2011) also argued that “a coefficient of a causal indicator with the wrong sign or that is not statistically significant would appear to be invalid and a candidate for exclusion” (p. 365). A significance test was relied on in the first stage of examining convergent and discriminant validity (Table 1). After the first stage, it is the difference of path coefficients between the second and the third stage that is important in supporting validity claims (Table 1). It is fully acknowledged that the exclusive focus on statistical significance ignores the problem that in large samples, effects that are trivial in magnitude can be statistically significant. However, in smaller samples where power is too low to be effective, even appreciably large effects may not be statistically significant in smaller samples. Therefore, when researchers apply our method and are in the first stage of our method, they may also want to check the statistical power to ensure that there is adequate power to detect medium to large effects.

Second, because the residual from formative measurement can only be identified when there are at least two paths emitting from the formative measurement model, at least two other latent variables measured by reflective indicators are needed. This limitation is due to the underlying attribute of formative measurement. One potential way to solve that issue is to add a reflective indicator to that measurement model so that only one other latent variable is needed. In this context, the formative measurement model still emits two paths: one to its reflective indicator and one to another outcome latent variable. Note that our method is fully consistent with recent debate of the disturbance term for formative measurement (Cadogan & Lee, 2013). Specifically, Cadogan and Lee (2013) suggested that using formative latent variables (formative measurement with the disturbance term) should be suspended until researchers developed corresponding measurement theories; meanwhile, other alternatives could be used, such as formative composite variables (formative measurement without the disturbance term). Therefore, after gathering convergent and discriminant validity evidence for formative measurement, researchers should apply formative composite variables in their model testing. As discussed above, our model is just to validate formative measurement, not to test theories developed containing formative measurement.

Third, for our method, the number of indicators used in reflective measurement should be at least four. As discussed above, for reflective measurement, the minimum number of indicators should be at least three. However, if there are only three indicators in a reflective measurement model (like TP in the previous data), the number of indicators from that measurement model will become two when we move one indicator to the formative measurement model and test if the latent variable measured with causal indicators can mediate the effect from that indicator. With only two indicators a latent variable will be unidentifiable.

Fourth, the analysis employed indicators from previously published studies. There was no control over model fit, strength of relationship between variables, and so on. Even though this may reflect reality, future studies can employ Monte

Carlo techniques to further validate the proposed under a variety of conditions (e.g. degree of model misspecification, strength of loadings).

References

Bagozzi, R. P. (2007). On the meaning of formative measurement and how it differs from reflective measurement: Comment on Howell, Breivik, and Wilcox. *Psychological Methods, 12*(2), 229-237. doi:10.1037/1082-989X.12.2.229

Bagozzi, R. P. (2011). Measurement and meaning in information systems and organizational research: Methodological and philosophical foundations. *MIS Quarterly, 35*(2), 261-292.

Barki, H., Titah, R., & Boffo, C. (2007). Information system use-related activity: An expanded behavioral conceptualization of individual-level information system use. *Information Systems Research, 18*(2), 173-192. doi:10.1287/isre.1070.0122

Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology, 51*, 1173-1182. doi:10.1037/0022-3514.51.6.1173

Bollen, K. (1984). Multiple indicators: internal consistency or no necessary relationship? *Quality and Quantity, 18*(4), 377-385. doi:10.1007/BF00227593

Bollen, K. (1989). *Structural equations with latent variables*. New York: John Wiley & Sons.

Bollen, K. (2007). Interpretational confounding is due to misspecificaiton, not to type of indicator: Comment on Howell, Breivik, and Wilcox (2007). *Psychological Methods, 12*(2), 219-228. doi:10.1037/1082-989X.12.2.219

Bollen, K. (2011). Evaluating effect, composite, and causal indicators in structural equation models. *MIS Quarterly, 35*(2), 359-372.

Bollen, K., & Davis, W. R. (2009). Causal indicator models: identification, estimation, and testing. *Structural Equation Modeling, 16*(3), 498-522. doi:10.1080/10705510903008253

Bollen, K., Glanville, J., & Stecklov, G. (2001). Socioeconomic status and class in studies of fertility and health in developing countries. *Annual Review of Sociology, 27*, 153-185. doi:10.1146/annurev.soc.27.1.153

Bollen, K., & Lennox, R. (1991). Conventional wisdom on measurement: A structural equation perspective. *Psychological Methods, 110*(2), 305-314. doi: 10.1037/0033-2909.110.2.305

Brettel, M., Engelen, A., Müller, T., & Schilke, O. (2011). Distribution channel choice of new entrepreneurial ventures. *Entrepreneurship: Theory and Practice, 35*(4), 683–708. doi:10.1111/j.1540-6520.2010.00387.x

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York: The Guilford Press.

Byrne, B. M. (1998). *Structural equation modeling with LISREL, PRELIS, and SIMPLIS: Basic concepts, applications, and programming*. Mahwah, NJ: Lawrence Erlbaum.

Cadogan, J. W., & Lee, N. (2013). Improper use of endogenous formative variables. *Journal of Business Research, 66*(2), 233-241. doi:10.1016/j.jbusres.2012.08.006

Cenfetelli, R. T., & Bassellier, G. (2009). Interpretation of formative measurement in information systems research. *MIS Quarterly, 33*(4), 689-707.

Chandon, P., Wansink, B., & Laurent, G. (2000). A benefit congruency framework of sales promotion effectiveness. *Journal of Marketing, 64*(4), 65-84. doi:10.1509/jmkg.64.4.65.18071

Chin, W. (1998). The partial least squares approach to structural equation modeling. In G. A. Marcoulides (Ed.), *Modern methods for business research* (p. 295-336). Mahwah, NJ: Lawrence Erlbaum.

Cizek, G. J., Rosenberg, S. L., & Koons, H. H. (2008). Sources of validity evidence for educational and psychological tests. *Educational and Psychological Measurement, 68*(3), 397-412. doi:10.1177/0013164407310130

Curtis, R. F., & Jackson, E. F. (1962). Multiple indicators in survey research. *American Journal of Sociology, 68*(2), 195-204.

Diamantopoulos, A. (2006). The error term in formative measurement models: interpretation and modeling implications. *Journal of Modelling in Management, 1*(1), 7-17. doi:10.1108/17465660610667775

Diamantopoulos, A. (2011). Incorporating formative measures into covariance-based structural equation models. *MIS Quarterly, 35*(2), 335-358.

Diamantopoulos, A., Riefler, P., & Roth, K. P. (2008). Advancing formative measurement models. *Journal of Business Research, 61*(12), 1201-1302. doi:10.1016/j.jbusres.2008.01.009

Diamantopoulos, A., & Winklhofer, H. M. (2001). Index construction with formative indicator. *Journal of Marketing Research, 38*(2), 269-277. doi:10.1509/jmkr.38.2.269.18845

Edwards, J. R. (2011). The fallacy of formative measurement. *Organizational Research Methods, 14*(2), 370-388. doi:10.1177/1094428110378369

Edwards, J. R., & Bagozzi, R. P. (2000). On the nature and direction of relationships between constructs and measures. *Psychological Methods, 5*(2), 155-174. doi:10.1037/1082-989X.5.2.155

Franke, G. R., Preacher, K. J., & Rigdon, E. E. (2008). Proportional structural effects of formative indicators. *Journal of Business Research, 61*(12), 1229-1237. doi:10.1016/j.jbusres.2008.01.011

Gefen, D., & Straub, D. (2005). A practical guide to factorial validity using PLS-Graph: Tutorial and annotated example. *Communications of the Association for Information Systems, 16*, 91-109.

Hardin, A. M., Chang, J. C., Fuller, M. A., & Torkzadeh, G. (2011). Formative measurement and academic research: In search of measurement theory. *Educational and Psychological Measurement 71*(2), 281-305. doi:10.1177/0013164410370208

Hauser, R. M., & Warren, J. R. (1997). Socioeconomic indexes for occupations: A review, update, and critique. *Sociological Methodology, 27(*1*)*, 177-298. doi:10.1111/1467-9531.271028

Howell, R. D., Breivik, E., & Wilcox, J. B. (2007a). Is formative measurement really measurement? Reply to Bollen (2007) and Bagozzi (2007) *Psychological Methods 12*(2), 238-245. doi:10.1037/1082-989X.12.2.238

Howell, R. D., Breivik, E., & Wilcox, J. B. (2007b). Reconsidering formative measurement. *Psychological Methods 12*(2), 205-218. doi:10.1037/1082-989X.12.2.205

Jarvis, C. B., MacKenzie, S. B., & Podsakoff, P. M. (2003). A critical review of construct indicators and measurement model misspecification in marketing and consumer research. *Journal of Consumer Research, 30*(2), 199-218. doi:10.1086/376806

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement: American Council on Education / Praeger Series on Higher Education* (4th ed., pp. 17-64). Westport, CT: Praeger Publishers.

MacCallum, R. C. & Browne, M. W. (1993). The use of causal indicators in covariance structure models: some practical issues. *Psychological Bulletin 114*(3), 533-541. doi:10.1037/0033-2909.114.3.533

MacKenzie, S. B., Podsakoff, P. M., & Jarvis, C. B. (2005). The problem of measurement model misspecification in behavioral and organizational research and some recommended solutions. *Journal of Applied Psychology, 90*(4), 710-730. doi:10.1037/0021-9010.90.4.710

MacKenzie, S. B., Podsakoff, P.M., & Podsakoff, N. P. (2011). Construct measurement and validation procedures in MIS and behavioral research: integrating new and existing techniques. *MIS Quarterly, 35*(2), 293-334.

Messick, S. (1995). Validity of psychological assessment, *American Psychologist*, *50*(9), 741-749. doi:10.1037/0003-066X.50.9.741

Muthén, L.K. & Muthén, B.O. (1998-2012). *Mplus User's Guide*. (7th Ed.). Los Angeles, CA: Muthén and Muthén.

Pavlou, P. A., & Gefen, D. (2005). Psychological contract violation in online marketplaces: antecedents, consequences, and moderating role. *Information Systems Research, 16*(4), 372-299. doi:10.1287/isre.1050.0065

Petter, S., Straub, D., & Rai, A. (2007). Specifying formative constructs in information systems research. *MIS Quarterly, 31*(4)*,* 623-656.

Rigdon, E. E. (1996). CFI versus RMSEA: A comparison of two fit indexes for structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal, 3*(4), 369-379. doi:10.1080/10705519609540052

Straub, D., Boudreau, M., & Gefen, D. (2004). Validation guidelines for IS positivist research. *Communications of the Association for Information Systems 13*, 380-427.

Wang, X., Jessup, L. M., & Clay, P. F. (2015). Measurement model in entrepreneurship and small business research: a ten year review. *International Entrepreneurship and Management Journal, 11*(1), 183-212. doi:10.1007/s11365-013-0285-0

Woodworth, R. S. (1928). Dynamic psychology. In C. Murchison (Ed.), *Psychologies of 1925* (pp. 111-126). Worcester, MA: Clark University Press.

Journal of Modern Applied Statistical Methods
May 2015, Vol. 14, No. 1, 107-121.

ISSN 1538 − 9472

Method of Estimation in the Presence of Non-response and Measurement Errors Simultaneously

Rajesh Singh Singh
Banaras Hindu University
Varanasi, India

Prayas Sharma
Banaras Hindu University
Varanasi, India

The problem of estimating the finite population mean of in simple random sampling in the presence of non-response and response error was considered. The estimators use auxiliary information to improve efficiency, assuming non–response and measurement error are present in both the study and auxiliary variables. A class of estimators was proposed and its properties studied in the simultaneous presence of non-response and response errors. It was shown that the proposed class of estimators is more efficient than the usual unbiased estimator, ratio and product estimators under non-response and response error together. A numerical study was carried out to compare its performance.

Keywords: Population mean, Study variable, Auxiliary variable, Mean squared error, Measurement errors, Non-response.

Introduction

Over the past several decades, statisticians were interested in the problem of estimating the parameters of interest in the presence of response error (measurement errors). In survey sampling the properties of the estimators based on data usually presuppose that the observations are the correct measurements on characteristics being studied. However, this assumption is not satisfied in many applications and data is contaminated with measurement errors, such as reporting errors and computing errors. These measurement errors make the result invalid, which are meant for no measurement error case. If measurement errors are very small and we can neglect it, then the statistical inferences based on observed data continue to remain valid. On the contrary, when they are not appreciably small and negligible, the inferences may not be simply invalid and inaccurate but may

Rajesh Singh is Assistant Professor in Department of Statistics. Email him at rsinghstat@gmail.com. Prayas Sharma is Faculty of Science in the Department of Statistics. Email at prayassharma02@gmail.com.

often lead to unexpected, undesirable and unfortunate consequences (see Srivastava & Shalabh 2001). Some important sources of measurement errors in survey data are discussed in Cochran (1968), Shalabh (1997), Sud and Srivastva (2000). Singh and Karpe (2008, 2010), Kumar, Singh, and Smarandache (2011), Kumar, Singh, Sawan, and Chauhan (2011) and Sharma and Singh (2013) studied the properties of some estimators of population parameters under measurement error.

Consider a finite population $U = (U_1, U_2, ..., U_N)$ of N units. Let Y and X be the study variate and auxiliary variate, respectively. Suppose that we have a set of n paired observations obtained through simple random sampling procedure on two characteristics X and Y. Further it is assumed that x_i and y_i for the i^{th} sampling units are observed with measurement error instead of their true values (X_i, Y_i). For a simple random sampling scheme, let $(x_{i,} y_i)$ be observed values instead of the true values (X_i, Y_i) for i^{th} $(i = 1.2 ,\ldots, n)$ unit, as

$$u_i = y_i - Y_i \tag{1}$$

$$v_i = x_i - X_i \tag{2}$$

where u_i and v_i are associated measurement errors which are stochastic in nature with mean zero and variances σ_u^2 and σ_v^2, respectively. Further, let the u_i's and v_i's are uncorrelated although X_i's and Y_i's are correlated.

Let the population means of X and Y characteristics be μ_x and μ_y, population variances of (x, y) be (σ_x^2, σ_y^2) and let ρ be the population correlation coefficient between x and y respectively (see Manisha and Singh 2002).

In sample surveys, the problem of non-response is common and is more widespread in mail surveys than in personal interviews. The usual approach to overcome non-response problem is to contact the non-respondent and obtain the information as much as possible. Hansen and Hurwitz (1946) were the first to deal with the problem of non-response. They proposed a sampling scheme that involves taking a subsample of non-responds after the first mail attempt and then obtain the information by personal interview.

For a finite population $U = \{U_1, U_2, \ldots, U_N\}$ of size N and a random sample of size n is drawn without replacement. Let the characteristics under study, say, y takes value y_i on the unit U_i $(I = 1, 2, \ldots, N)$. In survey on human population it is often the case that n_1 unit respond on the first attempt while n_1 $(= n - n_1)$ units do not provide any response. In the case of non-response of at initial stage Hansen and Hurwitz (1946) suggested a double sampling plan for estimating the population mean comprising the following steps:

i. A simple random sample of size n is drawn and the questionnaire is mailed to the sample units;

ii. A sub-sample of size $r = (n_2 / k)$, $(k > 1)$ from the n_2 non responding units in the initial step attempt is contacted through personal interviews.

Note that Hansen and Hurwitz (1946) considered the mail surveys at the first attempt and the personal interviews at the second attempt. In the Hansen and Hurwitz method the population is supposed to be consisting of response stratum of size N_1 and the non-response stratum of size $N_2 = (N - N_1)$. Let $\bar{Y} = \sum_{i=1}^{N} y_i/N$ and $S_y^2 = \sum_{i=1}^{N}\left(y_i - \bar{Y}\right)^2 / (N-1)$ denote the mean and the population variance of the study variable y. Let $\bar{Y}_1 = \sum_{i=1}^{N_1} y_i/N_1$ and $S_{y(1)}^2 = \sum_{i=1}^{N_1}\left(y_i - \bar{Y}\right)^2 / (N_1-1)$ denote the mean and variance of response group. Similarly, let $\bar{Y}_2 = \sum_{i=1}^{N_2} y_i/N_2$ and $S_{y(2)}^2 = \sum_{i=1}^{N_2}\left(y_i - \bar{Y}\right)^2 / (N_2-1)$ denote the mean and variance of the non-response group. The population mean can be written as $\bar{Y} = W_1\bar{Y}_1 + W_2\bar{Y}_2$, where $W_1 = (N_1 / N)$ and $W_2 = (N_2 / N)$. The sample mean $\bar{y}_1 = \sum_{i=1}^{n_1} y_i/n_1$ is an unbiased for $\bar{Y}_1$, but has a bias equal to $W_2\left(\bar{Y}_1 - \bar{Y}_2\right)$ in estimating the population mean $\bar{Y}$.

The sample mean $\bar{y}_{2r} = \sum_{i=1}^{r} y_i/r$ is unbiased for the mean y_2 for the n_2 units. Hansen and Hurwitz (1946) suggested an unbiased estimator for the population mean $\bar{Y}$ is given by $\bar{y}^* = w_1\bar{y}_1 + w_2\bar{y}_{2r}$.

Where $w_1 = (n_1 / n)$ and $w_2 = (n_2 / n)$ are responding and non-responding proportions in the sample. The variance of $\bar{y}^*$ is given by $V\left(\bar{y}^*\right) = \left(\frac{1-f}{n}\right)S_y^2 + \frac{W_2\left(k-1\right)}{n}S_{y(2)}^2$; where $f = (n / N)$.

In the sampling literature, it is known that efficiency of the estimator of population mean of a study variable *y* can be increased by the use of auxiliary information related to *x* which is highly correlated with study variable *y*. Cochran (1977) suggested the ratio and regression estimator of the population mean $\bar{Y}$ of study variable *y* in which information on the auxiliary variable is obtained from all sample units, and the population mean of auxiliary variable *x* is known, while some units do not provide any information on study variable *y*. Rao (1986), Khare and Srivastava (1995,1997), Okafor and Lee (2000) and Singh and Kumar (2008, 2009, 2010) have suggested some estimator for population mean of the study variable *y* using auxiliary information in presence of non-response.

Let x_i, $(i = 1, 2, \ldots, N)$ denote a auxiliary characteristics correlated with the study variable y_i, $(i = 1, 2, \ldots, N)$ the population mean of auxiliary variable is $\bar{X} = \sum_{i=1}^{N} x_i / N$. Let $\bar{X}_1$ and $\bar{X}_2$ denote the population means of the response and non-response groups. Let $\bar{x}_1 = \sum_{i=1}^{n_2} x_i / n_2, \bar{x}_2 = \sum_{i=1}^{n_2} x_i / n_2, \bar{x}_{2r} = \sum_{i=1}^{r} x_i / r$ denote the means of the n_1 responding units, n_2 non-responding units, and $r = (n_2 / k)$ sub-sampled units respectively. In this paper we have merged two major concepts for improvement of estimation techniques that is consideration of measurement error and non-response in the estimation procedure and proposed a class of estimators.

Notations

Let $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$, be the unbiased estimator of population means $\bar{X}$ and $\bar{Y}$, respectively but $s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$ and $s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2$ are not unbiased estimator of (σ_x^2, σ_y^2), respectively. The expected values of s_x^2 and s_y^2 in the presence of measurement error are, given by,

$$E\left(s_x^2\right) = \sigma_x^2 + \sigma_v^2$$

$$E\left(s_y^2\right) = \sigma_y^2 + \sigma_u^2$$

and for non-response group

$$E\left(s_{x_2}^2\right) = \sigma_{x_2}^2 + \sigma_{v_2}^2$$
$$E\left(s_{y_2}^2\right) = \sigma_{y_2}^2 + \sigma_{u_2}^2.$$

When the error variance σ_v^2 is known, the unbiased estimator of σ_x^2, is $\hat{\sigma}_x^2 = s_x^2 - \sigma_v^2 > 0$, and when σ_u^2 is known, then the unbiased estimator of σ_y^2 is $\hat{\sigma}_y^2 = s_y^2 - \sigma_u^2 > 0$.

Similarly, for the non-response group the unbiased estimator of $\sigma_{x_2}^2$, is $\hat{\sigma}_{x_2}^2 = s_{x_2}^2 - \sigma_{v_2}^2 > 0$, and when $\sigma_{u_2}^2$ is known, then the unbiased estimator of $\sigma_{y_2}^2$ is $\hat{\sigma}_{y_2}^2 = s_{y_2}^2 - \sigma_{u_2}^2 > 0$.

$$E\left(s_{x_2}^2\right) = \sigma_{x_2}^2 + \sigma_{v_2}^2$$
$$E\left(s_{y_2}^2\right) = \sigma_{y_2}^2 + \sigma_{u_2}^2.$$

Define

$$\bar{y} = \mu_y\left(1 + e_0\right)$$
$$\bar{x} = \mu_x\left(1 + e_1\right)$$

such that

$$E\left(e_0\right) = E\left(e_1\right) = 0,$$

and up to the first degree of approximation (when finite population correction factor is ignored)

$$E\left(e_0^2\right)=\frac{C_y^2}{n}\left(1+\frac{S_u^2}{S_y^2}\right)+\frac{W_2\left(k-1\right)}{n}C_{y_2}^2\left(1+\frac{S_{u_2}^2}{S_{y_2}^2}\right)$$

$$E\left(e_1^2\right)=\frac{C_x^2}{n}\left(1+\frac{S_v^2}{S_x^2}\right)+\frac{W_2\left(k-1\right)}{n}C_{x_2}^2\left(1+\frac{S_{v_2}^2}{S_{x_2}^2}\right)$$

$$E\left(e_0e_1\right)=\frac{\rho_{yx}C_yC_x}{n}+\frac{W_2\left(k-1\right)}{n}\rho_{yx_2}C_{y_2}C_{x_2}$$

$$C_y=S_y/\bar{Y},C_x=S_x/\bar{X},C_{y_2}=S_{y_2}/\bar{Y},C_{x_2}=S_{x_2}/\bar{X},\rho_{xy}=S_{xy}/S_xS_y$$

Adapted estimator

A traditional estimator for estimating population mean in the simultaneous presence of response and non-response error is given by,

$$t_1=\bar{y} \tag{3}$$

Expression (3) can be written as

$$t_1-\bar{Y}=\bar{Y}^2\left(1+e_0\right) \tag{4}$$

Taking expectation both sides of (4), we get bias of estimator t_1 given as

$$Bias\left(t_1\right)=0 \tag{5}$$

Squaring both sides of (4)

$$\left(t_1-\bar{Y}\right)^2=\bar{Y}^2e_0^2 \tag{6}$$

and taking expectation and using notation, mean square error of t_1 is obtained up to first order of approximation, as

$$MSE\left(t_1\right)=\frac{S_y^2}{n}\left(1+\frac{\sigma_u^2}{S_y^2}\right)+AS_{y2}^2\left(1+\frac{\sigma_{u2}^2}{S_{y2}^2}\right) \tag{7}$$

or

$$MSE\left(t_1\right)=M \tag{8}$$

where, $A=\dfrac{\left(k-1\right)W_2}{n}$ and $M=\dfrac{S_y^2}{n}\left(1+\dfrac{\sigma_u^2}{S_y^2}\right)+AS_{y2}^2\left(1+\dfrac{\sigma_{u2}^2}{S_{y2}^2}\right)$.

In the case, when the measurement error is zero or negligible, MSE of estimator t_1 is given by,

$$MSE^*\left(t_1\right)=\frac{S_y^2}{n}+AS_{y2}^2 \tag{9}$$

where, $M_{t_1}=\dfrac{\sigma_u^2}{n}+A\sigma_{u2}^2$ is the contribution of measurement errors in t_1.

When there is non-response and response error both are present, a ratio type estimator for estimating population mean is, given by

$$t_r=\frac{\overline{y}^*}{\overline{x}^*}\overline{X} \tag{10}$$

Expressing the estimator t_r in terms of e's

$$t_r=\overline{Y}\left(1+e_0\right)\left(1+e_1\right)^{-1} \tag{11}$$

Expanding equation (11) and simplifying,

$$\left(t_r-\overline{Y}\right)=\overline{Y}\left[e_0-e_1-e_0e_1+e_1^2\right] \tag{12}$$

and taking expectation both sides of (12), the bias of estimator t_r is

$$Bias\left(t_r\right)=\left[\frac{S_x^2}{n}\left(1+\frac{\sigma_v^2}{S_x^2}\right)+AS_{x2}^2\left(1+\frac{\sigma_{v2}^2}{S_{x2}^2}\right)-2\left(\frac{1}{n}\rho_{xy}S_xS_y+A\rho_{xy2}S_{x2}S_{y2}\right)\right] \tag{13}$$

Squaring both sides of (12),

$$\left(t_r - \sigma_y^2\right)^2 = \bar{Y}^2\left[e_0^2 + e_1^2 - 2e_0e_1\right] \tag{14}$$

Taking expectations of (14) and using notations, we get the MSE of estimator t_r as

$$\begin{aligned} MSE\left(t_r\right) = \frac{1}{n}&\left[S_y^2\left(1+\frac{\sigma_u^2}{S_y^2}\right)+S_x^2\left(1+\frac{\sigma_v^2}{S_x^2}\right)-2\rho_{xy}S_xS_y\right] \\ &+A\left[S_{y2}^2\left(1+\frac{\sigma_{u2}^2}{S_{y2}^2}\right)+S_{x2}^2\left(1+\frac{\sigma_{v2}^2}{S_{x2}^2}\right)-2\rho_{xy2}S_{x2}S_{y2}\right] \end{aligned} \tag{15}$$

$$\begin{aligned} &=\left[\begin{aligned} &\frac{S_y^2}{n}\left(1+\frac{\sigma_u^2}{S_y^2}\right)+AS_{y2}^2\left(1+\frac{\sigma_{u2}^2}{S_{y2}^2}\right)+\frac{S_x^2}{n}\left(1+\frac{\sigma_v^2}{S_x^2}\right) \\ &+AS_{x2}^2\left(1+\frac{\sigma_{v2}^2}{S_{x2}^2}\right)-2\left(\frac{1}{n}\rho_{xy}S_xS_y+A\rho_{xy2}S_{x2}S_{y2}\right) \end{aligned}\right] \\ &= M+N-2O \end{aligned} \tag{16}$$

where,

$$\begin{aligned} M &= \left\{\frac{1}{n}S_y^2\left(1+\frac{\sigma_u^2}{S_y^2}\right)+AS_{y2}^2\left(1+\frac{\sigma_{u2}^2}{S_{y2}^2}\right)\right\} \\ N &= \left\{\frac{1}{n}S_x^2\left(1+\frac{\sigma_v^2}{S_x^2}\right)+AS_{x2}^2\left(1+\frac{\sigma_{v2}^2}{S_{x2}^2}\right)\right\} \\ O &= \left\{\frac{1}{n}\rho_{xy}S_xS_y+A\rho_{xy2}S_{x2}S_{y2}\right\}. \end{aligned}$$

A regression estimator under measurement error and non-response is defined as

$$t_{lr} = \bar{y}^* + b\left(\bar{X}-\bar{x}^*\right) \tag{17}$$

Expressing the estimator t_r in terms of e's, $t_{lr} = \bar{Y}\left(1+e_0\right) - b\bar{X}e_1$,

and expanding equation (17) and simplifying,

$$\left(t_{lr}-\bar{Y}\right)=\left(\bar{Y}e_0-b\bar{X}e_1\right) \tag{18}$$

Squaring both sides of (18) and after simplification,

$$\left(t_{lr}-\bar{Y}\right)^2=\left[\bar{Y}^2e_0^2+b^2\bar{X}^2e_1^2-2b\bar{X}\bar{Y}e_0e_1\right] \tag{19}$$

Taking expectations both sides of (19) the MSE of estimator t_{lr} is obtained as

$$MSE\left(t_{lr}\right)=M+b^2R^2N-2bRO \tag{20}$$

The optimum value of b is obtained by minimizing (20) and is given by

$$b^*=\frac{1}{R}\left[\frac{O}{N}\right] \tag{21}$$

Substituting the optimal value of b in equation (20), the minimum MSE of the estimator t_{lr} is obtained as

$$MSE\left(t_{lr}\right)_{\min}=M\left[1-\frac{O^2}{MN}\right] \tag{22}$$

In the case, when the measurement error is zero or negligible, MSE of estimator t_1 is given by

$$MSE\left(t_{lr}\right)=\frac{1}{n}S_y^2\left[1-\rho_{xy}^2\right]+\frac{(k-1)W_2}{n}\left[S_{y2}^2+b^2S_{x2}^2-2b\rho_{xy2}S_{x2}S_{y2}\right] \tag{23}$$

Proposed class of estimator

A proposed class of estimators given by

$$t_p=m_1\bar{y}^*+m_2\frac{\bar{y}^*}{\bar{x}^*}\bar{X} \tag{24}$$

Note for $(m_1, m_2) = (1, 0)$ $t_1 = \bar{y}^*$ (usual unbiased estimator), and for $(m_1, m_2) = (0, 1)$ $t_2 = \frac{\bar{y}^*}{\bar{x}^*}\bar{X}$ (usual ratio estimator). Thus, the proposed class of estimators is generalized version of usual unbiased estimator and ratio estimator. Expressing the estimator t_p in terms of e's

$$t_p = m_1\bar{Y}\left(1+e_0\right)+m_2\bar{Y}\left(1+e_0\right)\left(1+e_1\right)^{-1} \tag{25}$$

Expanding equation (25) and simplifying,

$$\left(t_p-\bar{Y}\right)=\bar{Y}\left[e_0+m_2\left(-e_1+e_1^2-e_0e_1\right)\right] \tag{26}$$

Squaring both sides of (26) and after simplification,

$$\left(t_p-\bar{Y}\right)^2=\bar{Y}^2\left[e_0^2+m_2^2e_1^2-2m_2e_0e_1\right] \tag{27}$$

Taking expectations of (27) and using notations, the MSE of estimator t_r is obtained as

$$MSE\left(t_p\right)=M+m_2R^2N-2m_2RO \tag{28}$$

The optimum value of m_2 is obtained by minimizing (28), given by

$$m_2^*=\frac{1}{R}\left[\frac{O}{N}\right] \tag{29}$$

and $m_1^* = 1 - m_2^*$.

Substituting the optimal value of m_2 in equation (28) the minimum MSE of the estimator t_p is obtained as

$$MSE\left(t_p\right)_{\min}=M\left[1-\frac{O^2}{MN}\right] \tag{30}$$

The minimum MSE of proposed class of estimator t_p given in (30) is same as the MSE of regression estimator under simultaneous presence of non-response and measurement error, given in equation (22).

Efficiency comparisons

First, the efficiency of the proposed estimator t_p is compared with usual unbiased estimator,

$$\begin{aligned} &MSE(t_1) - MSE(t_P)_{\min} > 0 \\ &\text{If } \left[M - M\left(1 - \frac{O^2}{MN}\right)\right] > 0, \left[\frac{O^2}{MN}\right] > 0 \end{aligned} \tag{31}$$

The condition listed in (31) shows that proposed family of estimators is always better than the usual estimator under the non-response and measurement error.

Next, the ratio estimator is compared with proposed family of estimators t_p,

$$\begin{aligned} &MSE(t_2)_{\min} - MSE(t_P)_{\min} > 0, \left[(M + N - 2O) - M\left(1 - \frac{O^2}{MN}\right)\right] > 0 \\ &[N - O]^2 > 0 \end{aligned} \tag{32}$$

Observe that the condition (32) holds and shows proposed family of estimators is better than the ratio estimator under the non-response and measurement error.

Empirical study

Data statistics

The data used for empirical study was taken from Gujarati and Sangeetha (2007, pg, 539) where,

Y_i = True consumption expenditure,

X_i = True income,

y_i = Measured consumption expenditure,

x_i = Measured income.

From the data given we get the following parameter values:

Table 1. Value of the parameters

n	μ_y	μ_x	S_y	S_x	ρ	σ_u^2	σ_v^2
70	981.29	1755.53	613.66	1406.13	0.778	36.00	36.00
μ_{y2}	μ_{x2}	S_{y2}	S_{x2}	ρ_2	R	W_2	
597.29	1100.24	244.11	631.51	0.445	0.5589	0.25	

Table 2. Showing the MSE of the estimators with and without measurement errors

Estimators	MSE Without Error	Contribution of meas. error in MSE	Contribution of non-response	MSE including me. Errors & non-response
$t_1 = \bar{y}^*$	10759.39	1.03	2553.840	13313.58
t_r	6967.135	1.35	4607.335	11574.92
t_{lr}	4246.903	0.86	2527.751	6775.036
t_p	4246.903	0.86	2527.751	6775.036

Table 2 exhibits that measurement error and non-response plays an important role in increasing the MSE of an estimator. We also conclude that contribution of measurement error and non-response in usual estimator is less than in comparison to the ratio estimator; these observations have interesting implication where the ratio estimator performs better than sample mean under the absence of any measurement error in X characteristics. There may be a case when ratio estimator is poor than sample mean under the consideration of any measurement error. It is observed from Table 2 that the performance of our proposed estimator t_p is better than usual estimator t_1 and ratio estimator t_r under non-response and measurement error. Further it is observed that contribution of non-response error is larger than the response error in increasing the MSE of the estimators.

Conclusion

A class of estimator of the population mean of study variable *y* was proposed using auxiliary information. The estimators use auxiliary information to improve efficiencies, assuming non-response and measurement error are present in both the study and auxiliary variables. In addition, some known estimator of population mean such as usual unbiased estimator and ratio estimator for population mean are found to be members of the proposed class of estimators. The MSEs of the proposed class of estimators were obtained up to the first order of approximation in the simultaneous presence of non-response and response error. The proposed class of estimators are advantageous in the sense that the properties of the estimators which are members of the proposed class of estimators can be easily obtained from the properties of the proposed class of estimators.

References

Allen, J., Singh, H. P. & Smarandache, F. (2003). A family of estimators of population mean using multiauxiliary information in presence of measurement errors. *International Journal of Social Economics, 30*(7), 837–849. doi:10.1108/03068290310478775

Cochran, W. G. (1968). Errors of measurement in statistics. *Technometrics, 10*(4), 637-666. doi:10.1080/00401706.1968.10490621

Cochran, W. G. (1977). *Sampling Techniques* (3rd ed.). New York: John Wiley and Sons.

Das, A. K. & Tripathi, T. P. (1978). Use of auxiliary information in estimating the finite population variance. *Sankhya, 40*(C), 139-148.

Diana, G. and Giordan, M. (2012). Finite Population Variance Estimation in Presence of Measurement Errors. *Communication in Statistics-Theory and Methods, 41*(23), 4302-4314. doi:10.1080/03610926.2011.573165

Gujarati, D. N. & Sangeetha (2007). *Basic econometrics*. Tata McGraw-Hill.

Hansen, M.H. & Hurwitz, W.N. (1946). The problem of non-response in sample surveys. *Journal of the American Statistical Association, 41*(236), 517–529. doi:10.1080/01621459.1946.10501894

Khare, B.B. & Srivastava, S. (1995). Study of conventional and alternative two-phase Sampling ratio, product and regression estimators in presence of non-response. *Proceedings of the Indian National Science Academy, 65*, 195–203.

Khare, B. B., & Srivastava, S. (1997). Transformed ratio type estimators for the population mean in the presence of nonresponse. *Communications in Statistics - Theory and Methods, 26*(7), 1779-1791. doi:10.1080/03610929708832012

Koyuncu, N. & Kadilar, C. (2010). On the family of estimators of population mean in stratified random sampling. *Pakistan Journal of Statistics, 26*(2), 427-443.

Kumar, M., Singh, R., Singh, A. K. & Smarandache, F. (2011). Some ratio type estimators under measurement errors. *World Applied Sciences Journal, 14*(2), 272-276.

Kumar, M., Singh, R., Sawan, N. & Chauhan, P. (2011). Exponential ratio method of estimators in the presence of measurement errors. *International Journal of Agricultural and Statistical Sciences, 7*(2), 457-461.

Manisha & Singh, R. K. (2002). Role of regression estimator involving measurement errors. *Brazilian Journal of Probability and Statistics, 16*, 39-46.

Okafor, F. C. & Lee, H. (2000). Double sampling for ratio and regression estimation with sub-sampling the non-respondents, *Survey Methodology, 26*(2), 183–188.

Rao, P. S. R. S. (1986). Ratio estimation with sub sampling the non-respondents. *Survey Methodology, 12*(2), 217–230.

Shalabh (1997). Ratio method of estimation in the presence of measurement errors. *Journal of Indian Society of Agricultural Statistics, 50*(2), 150–155.

Sharma, P. & Singh, R. (2013). A generalized class of estimator for population variance in presence of measurement errors. *Journal of Modern Applied Statistical Methods, 12*(2), 231-241. Retrieved from http://digitalcommons.wayne.edu/jmasm/vol12/iss2/13

Singh, H.P. & Karpe, N. (2008). Ratio-product estimator for population mean in presence of measurement errors. *Journal of Applied Statistical Science, 16*(4), 437-452.

Singh, H. P. and Karpe, N. (2009). A class of estimators using auxiliary information for estimating finite population variance in presence of measurement errors. *Communication in Statistics - Theory and Methods, 38*(5),734-741. doi:10.1080/03610920802290713

Singh, H. P. and Karpe, N. (2010). Effect of measurement errors on the separate and combined ratio and product estimators in stratified random sampling. *Journal of Modern Applied Statistical Methods, 9*(2), 338-402. Retrieved from http://digitalcommons.wayne.edu/jmasm/vol9/iss2/8

Singh, H. P. & Kumar, S. (2008). Estimation of mean in presence of non-response using two phase sampling scheme. *Statistical Papers, 50*, 559–582. doi:10.1007/s00362-008-0140-5

Singh, H. P. & Kumar, S. (2009). A general class of estimators of the population mean in survey sampling using auxiliary information with sub sampling the non-respondents, *The Korean Journal of Applied Statistics, 22*(2), 387–402. doi:10.5351/KJAS.2009.22.2.387

Srivastava, A., K., & Shalabh (2001). Effect of measurement errors on the regression method of estimation in survey sampling. *Journal of Statistical Research, 35*(2), 35-44.

Sud, U. C. & Srivastava, A. K. (2000): Estimation of population mean in repeated surveys in the presence of measurement errors. *Journal of the Indian Society of Agricultural Statistics, 53*(2), 125-133.

Journal of Modern Applied Statistical Methods
May 2015, Vol. 14, No. 1, 122-151.

ISSN 1538 − 9472

Estimating the Accuracy of Automated Classification Systems Using Only Expert Ratings that are Less Accurate than the System

Paul E. Lehner
The MITRE Corporation
McLean, VA, USA

A method is presented to estimate the accuracy of an automated classification system based only on expert ratings on test cases, where the system may be substantially more accurate than the raters. In this method an estimate of overall rater accuracy is derived from the level of inter-rater agreement, Bayesian updating based on estimated rater accuracy is applied to estimate a ground truth probability for each classification on each test case, and then overall system accuracy is estimated by comparing the relative frequency that the system agrees with the most probable classification at different probability levels. A simulation analysis provides evidence that the method yields reasonable estimates of system accuracy under diverse and predictable conditions.

Keywords: Inter-rater reliability, Kappa, artificial intelligence

Introduction

Information technology is advancing to develop systems that address problems of increasing sophistication and complexity. A quick scan of programs sponsored by research funding agencies (e.g., www.nih.gov, www.nsf.gov, www.darpa.mil, www.iarpa.gov) showed new systems being developed to address complex problems as diverse as automated medical and clinical diagnoses, technology readiness evaluation, detection of emerging technologies, classification of the behavioral contents of unstructured video segments, recognition and classification of metaphors used in natural language text and many others.

The complexities of the problems that these advanced systems address make it difficult to evaluate the accuracy of such systems. It is usually necessary to

Dr. Lehner is a Consulting Scientist with The MITRE Corporation. Email him at plehner@mitre.org.

resort to using expert raters to assign ground truth for test cases. However, the complexity of these problems also challenge to the expert raters. Raters often disagree as to which is the correct category. Furthermore as future systems address problems of ever increasing sophistication and complexity, it seems likely that the experts will be even more challenged and exhibit even lower levels of agreement. Ground truth data sets based on expert assignments are fallible and are likely to become more so in the future.

Using expert raters to assign ground truth to test cases is a well-established practice. For classification problems, which are the focus of this paper, a statistic such as Kappa is used to measure inter-rater agreement; and then the rating process is refined until a satisfactory level of agreement is reached. Once the agreement threshold is reached, assignments of individual raters or collaborating teams of raters are treated as truth and system accuracy is measured by the level of agreement with the assigned ground truth (See Gwet, 2010 for review).

For several reasons, this common scientific practice does not adequately meet the needs of advanced system evaluation. First, the level of agreement amongst raters will rarely meet a satisfactory level. The problems that these systems address are simply too complex. About the only way to increase the level of agreement is to select relatively simple and therefore non-representative test cases.

Second, estimating system accuracy by measuring the level of agreement with expert raters makes the *de facto* assumption that the experts are more accurate than the system. This assumption runs contrary to a substantial body of empirical research where it is often found that simple algorithms outperform human experts in complex judgments (Dawes, 1979; Grove, Zald, Lebow, Snitz, & Nelson 2001; Tetlock, 2005). It should not be presumed that the experts are more accurate than the system.

Third, there is considerable evidence to suggest that for a wide variety of judgment tasks collaborative team judgments are not substantially more accurate than the judgments of randomly selected individual team member (e.g., Surowiecki, 2005; Armstrong, 2006). In judgment tasks, where there is no obvious correct answer, it should not be presumed that collaboration will reliably lead the raters to converge to the correct answer.

Finally, when evaluating a classification system the statistic of greatest interest is the accuracy of the system - the proportion of system assignments that are correct. Unfortunately there is an unclear relationship between inter-rater reliability statistics such as Kappa, the probability of correct ground truth assignments and the accuracy of any systems tested against error-prone ground truth assignments.

A different approach is presented here to using expert ratings to estimate the accuracy of classification systems. Rather than treat expert ratings as a surrogate for ground truth, expert ratings are treated as error prone estimates of ground truth where independent ratings are fused to estimate ground truth probabilities, and the ground truth probabilities are then used to estimate system accuracy.

One practical instantiation of this estimation approach is described below. In addition simulation test results are provided that support several claims. First, under diverse conditions, this approach reliably yields estimates of system accuracy that are approximately correct. If a system is 90% accurate then this approach will yield an estimate of system accuracy that is close to 90%. Second, the accuracy of the estimate of system accuracy is largely independent of whether the expert raters are more or less accurate than the system. If a system is in fact 90% accurate, and the raters are individually 60% accurate, then the estimate of system accuracy will still be approximately 90%. Third, reliable estimates of system accuracy can often be obtained with a reasonably small number of test cases (e.g. fifty test cases with three expert raters). In complex domains it is important to keep sample sizes as small as possible, since it may be time consuming and costly to obtain expert ratings. Fourth, and importantly, the conditions under which the above three claims may break down are predictable. Therefore test data sets can be intentionally constructed to ensure that the conditions are met that are needed for accurate estimation of system accuracy.

Estimating the accuracy of system classifications

The method for estimating accuracy described below was derived from the following assumptions.

AA1. For each case there is a unique correct classification.
AA2. For each case raters independently assign classifications.
AA3. Expected agreement between raters increases as expected rater accuracy increases.

Assumption AA3 refers to *expected* agreement and accuracy. Here "accuracy" refers to the total proportion of correct classifications made by all the raters, irrespective of which raters are making correct and incorrect classifications. And "agreement" refers to the total proportion of pairwise agreement among all of the raters and cases. For any particular set of cases, accuracy may be low yet agreement high (the raters made the same mistakes), but AA3 asserts that *in general* there is an expected positive relationship between accuracy and agreement.

Theorem 1:

AA1-AA3 are ensured if and only if the raters behave as though their selection for each case is determined by a single confusion matrix where the conditional probability of correct assignment is constant and the conditional probability of all incorrect assignments is equal.

That is to say all raters on all problems are selecting from a single confusion matrix with a structure such as shown in Table 1.

The proof of this theorem is found in the Appendix. The general structure of the proof shows that if the raters are assigning classifications using any process other than selecting from a common confusion matrix with the structure illustrated in Table 1, then it is always possible to construct a classification process with lower expected accuracy and higher expected agreement, or higher accuracy and lower agreement; thereby violating the assumed monotonic relationship between expected accuracy and expected agreement.

Table 1. Implied Structure of Rater Confusion Matrices for Four Category Problem (A to D are true categories and "A" to "D" are selected categories.)

	"A"	"B"	"C"	"D"
A	P_c	$(1-P_c)/3$	$(1-P_c)/3$	$(1-P_c)/3$
B	$(1-P_c)/3$	P_c	$(1-P_c)/3$	$(1-P_c)/3$
C	$(1-P_c)/3$	$(1-P_c)/3$	P_c	$(1-P_c)/3$
D	$(1-P_c)/3$	$(1-P_c)/3$	$(1-P_c)/3$	P_c

AA1 through AA3 also seem to be assumed implicitly in many contexts where the Kappa statistic is applied. Indeed it is AA3 that would seem to warrant the common practice of using expert ratings as surrogates for ground truth when high levels of inter-rater agreement are found. Consequently it is reasonable to claim that the estimation method described below is derived from assumptions implicit in the Kappa statistic and how Kappa is often used. Because of this relationship to the Kappa statistic, in the remainder of this paper AA1-AA3 will be referred to as *K-assumptions*. Furthermore, the properties of equal rater accuracy, equal error probabilities and equal problem difficulty that are implied by the *K-assumptions* will be referred to as *K-properties*.

Table 2. Sample data of expert ratings and system assignments for 10 test cases

Case #	Rater 1	Rater 2	Rater 3	Rater 4	System
1	“C”	“D”	“C”	“C”	“A”
2	“B”	“D”	“C”	“C”	“C”
3	“C”	“C”	“D”	“C”	“C”
4	“B”	“B”	“D”	“D”	“B”
5	“A”	“B”	“B”	“B”	“B”
6	“C”	“B”	“D”	“A”	“A”
7	“A”	“A”	“A”	“A”	“A”
8	“A”	“D”	“B”	“C”	“C”
9	“D”	“B”	“A”	“A”	“D”
10	“A”	“D”	“A”	“B”	“B”

The estimation method is straightforward to explain in the context of an example. Consider the test data in Table 2. There are 10 test cases, 4 categories, 4 raters and the system’s proposed answers. When referring to ground truth the four categories are labeled *A, B, C, D*; when referring to rater and system assignments they are labeled “A”, “B”, “C”, “D”.

As described below the estimation method is composed of four basic steps.

Estimate rater accuracy

Given that each rater has an identical confusion matrix, with the structure shown in Table 1, the probability that two raters will agree on any one case is

$$P_a = P_c^2 + \frac{(1-P_c)^2}{N-1} \tag{1}$$

Here P_a is the probability of agreement, P_c is the probability that a rater will make the correct assignment, and N is the number of categories. Solving for P_c yields

$$P_c = \left(\frac{1}{N}\right) + \sqrt{\left(\frac{(N-1)*P_a - \frac{N-1}{N}}{N}\right)} \tag{2}$$

Eq. 2 is used to estimate rater accuracy. In the 10 cases in Table 1 there was 33% agreement (20 pairs out of 60). Setting P_a to .33 and solving for P_c yields P_c = 0.5; which is the estimate of rater accuracy.

Estimate base rates

The probability that a rater will assert a category, say "A", is as follows:

$$P(\text{"A"}) = P(\text{"A"}|A)*P(A) + \left(1 - \frac{P(\text{"A"}|A)}{N-1}\right)*(1-P(A)) \tag{3}$$

Here P("A") is the marginal probability that the rater asserts "A", P("A"|A) is the conditional probability that the rater will assert "A" if A is true, and P(A) is the marginal probability of A. Solving for P(A) yields

$$P(A) = \frac{(N-1)*P(\text{"A"}) - 1 + P(\text{"A"}|A)}{N*P(\text{"A"}|A) - 1} \tag{4}$$

Setting P("A") to be the observed relative frequency of "A", and P("A"|A) to be the estimate of P_c from above, yields

$$P(A) = \frac{(N-1)*P(\text{"A"}) - 1 + P_c}{N*P_c - 1} \tag{5}$$

Eq. 5 is used to estimate the base rate for each category by setting P_c to be the estimate from above and P("X") to be the observed relative frequency across all raters and ratings that category X was assigned. In Table 1 there are 11 instances of each of the categories; so the estimated base rate is 0.325 for category A. Applying Eq. 5 to the other categories yields base rates of 0.25, 0.25 and 0.175 for B, C and D respectively.

Estimate ground truth probabilities

Use Bayes rule, assuming conditional independence for each rater, to estimate ground truth probabilities. For example, in case 1 above the raters selected “CCDC”. So for each possible ground truth value calculate *P*(…|”CDCC”) and normalize.

$$
\begin{aligned}
P\left(A\middle|\text{”CDCC”}\right) &\sim P(A) * P\left(\text{“C”}\middle|A\right) * P\left(\text{“D”}\middle|A\right) * P\left(\text{“C”}\middle|A\right) * P\left(\text{“C”}\middle|A\right) \\
&= .325 * .167 * .167 * .167 * .167 = .00025 \rightarrow .041 \\
P\left(B\middle|\text{”CDCC”}\right) &\sim P(B) * P\left(\text{“C”}\middle|B\right) * P\left(\text{“D”}\middle|B\right) * P\left(\text{“C”}\middle|B\right) * P\left(\text{“C”}\middle|B\right) \\
&= .25 * .167 * .167 * .167 * .167 = .00019 \rightarrow .032 \\
P\left(C\middle|\text{”CDCC”}\right) &\sim P(C) * P\left(\text{“C”}\middle|C\right) * P\left(\text{“D”}\middle|C\right) * P\left(\text{“C”}\middle|C\right) * P\left(\text{“C”}\middle|C\right) \\
&= .25 * .5 * .167 * .5 * .5 = .00521 \rightarrow .860 \\
P\left(D\middle|\text{”CDCC”}\right) &\sim P(D) * P\left(\text{“C”}\middle|D\right) * P\left(\text{“D”}\middle|D\right) * P\left(\text{“C”}\middle|D\right) * P\left(\text{“C”}\middle|D\right) \\
&= .175 * .167 * .5 * .167 * .167 = .00041 \rightarrow .067
\end{aligned}
$$

Repeating this step for the other 9 cases yields the estimated probability distributions shown in Table 3.

Table 3. Estimated ground truth probabilities for sample data

	Ground Truth Probability				System
Case #	**A**	**B**	**C**	**D**	**Answer**
1	0.041	0.032	0.860	0.067	“A”
2	0.084	0.195	0.584	0.136	“C”
3	0.041	0.032	0.860	0.067	“C”
4	0.074	0.511	0.057	0.358	“B”
5	0.120	0.828	0.031	0.021	“B”
6	0.325	0.250	0.250	0.175	“A”
7	0.975	0.009	0.009	0.006	“A”
8	0.325	0.250	0.250	0.175	“C”
9	0.657	0.169	0.056	0.118	“D”
10	0.657	0.169	0.056	0.118	“B”

Estimate system accuracy

Assume any probability distribution over the categories for each test case. For any test case, let P_g be the probability of the classification with the highest probability, *P*P_s be the probability that the system will assign the correct answer, P_a be the probability that the system will assign the same classification as the highest ground truth probability. It follows that

$$P_a = P_g * P_s + \left(1 - P_g\right) * \frac{1 - P_s}{N - 1} \tag{6}$$

Note that this relationship holds whether or not the classification with the highest probability is correct. Solving for P_s yields

$$P_s = \frac{\left(N - 1\right) * P_a - 1 + P_g}{N * P_g - 1} \tag{7}$$

Eq. 7 is used to estimate system accuracy as follows. First separate the test cases into bins with approximately the same highest estimated ground truth probability. In this paper the ranges (.9, 1.0], (.8, .9], (.7, .8], etc. are used. For example, in Table 3 there is one case in the (.9, 1.0] range, 3 cases in the (.8, .9] range, 2 cases in the (.6, .7] range, etc. Second for each bin calculate the average ground truth probability within the bin; record the proportion of system assignments that agree with the most probable answer; then estimate system accuracy for each bin using equation Eq. 7. Third estimate overall system accuracy by taking the average of the estimated accuracy in each bin weighted by the number of cases in each bin. This is shown in Table 4.

Table 4. Estimate of System Accuracy for Sample Data

Probability	Average Ground	Number	Proportion of	Estimated
Bin	**Truth Probability**	**in Bin**	**Agreement**	**Accuracy**
.9 – 1.0	0.975	1	1.000	1.000
.8 - .9	0.849	3	0.667	0.776
.6 - .7	0.657	2	0.000	0.000
.5 - .6	0.548	2	0.333	0.452
.2 - .3	0.325	2	0.500	1.000
			Weighted Average =	0.731

The reader may be curious as to why the estimate of system accuracy is not simply the average of the estimated ground truth probabilities for the system answers. The reason is that taking the average will consistently underestimate system accuracy; because the system's answer is itself additional evidence for each category. So, for example, if the system answer is "C" and the estimated ground truth probability for C is 0.6; then a better estimate for C would be somewhat higher than .6. But until system accuracy is estimated it cannot be determined how much more than .6 is appropriate. In the above example, the average estimated ground truth probability of the system answers is .466, but the estimate of system accuracy in Table 4 is 0.731.

Note that the value of Kappa (using 1/number-categories to determine random agreement) for the data in Table 2 is

$$\begin{aligned}\text{Kappa} =&\\ &= (\text{Observed Agreement - Random Agreement})/(1.0 \text{ - Random Agreement})\\ &= (.333 - .25)/(1 - .25) = 0.11\end{aligned}$$

Standard thresholds normally require a level of $\text{Kappa} = 0.7$ before the expert ratings are considered usefully reliable (Gwet 2010). $\text{Kappa} = 0.11$ is considered "slight agreement" and is far too low for the ratings to be considered useful for establishing ground truth.

Overall then, in the sample data provided in Table 2; inter-rater agreement is "slight" ($\text{Kappa} = 0.11$), estimated rater accuracy is 0.50, and estimated system accuracy is 0.731.

Performance and robustness

The above example illustrates how to estimate system accuracy for classification problems even when inter-rater agreement and estimated rater accuracy are very low. This section examines the accuracy of estimates of system accuracy, and the robustness of those estimates, through a series of simulations.

All of the simulations described below use the following procedure to assign the confusion matrix for each rater and the system, based on values set to four parameters: an initial probability of correct assignment (IPC), a problem difficulty adjustment (PDA), degree of asymmetric dispersion (AD), and a proportional error range (PER).

Each confusion matrix is constructed as follows:

1. Initially assign the conditional probability of a correct classification to be IPC for all categories.

2. Add PDA to the conditional probabilities of correct assignment.
3. For each category distribute the remaining probability (1 - IPC - PDA) to the incorrect classifications in a manner that is proportional to the distance from the correct classification, where the probability of a classification that is M steps removed from the correct classifications is AD times more likely than a classification that is M+1 steps removed.
4. For each conditional probability of incorrect assignment (IC) set the range to be [IC - PER*IC, IC + PER*IC], then randomly select a new probability by uniform sampling over this range.
5. Normalize the modified confusion matrix after the random changes in step 4 so that expected accuracy is equal to IPC + PDA.

For example, if there are five categories and (IPC, PDA, AD, PER) = (.6, 0, 1.0, 0), then the resulting confusion matrix is shown in Table 5.

Table 5. Confusion matrix where (IPC, PDA, AD, PER) = (0.6, 0, 1.0, 0)

Correct			Classification		
Category	**"A"**	**"B"**	**"C"**	**"D"**	**"E"**
A	0.6	0.1	0.1	0.1	0.1
B	0.1	0.6	0.1	0.1	0.1
C	0.1	0.1	0.6	0.1	0.1
D	0.1	0.1	0.1	0.6	0.1
E	0.1	0.1	0.1	0.1	0.6

On the other hand, if (IPC, PDA, AD, PER) = (.6, -.2, 2.0, 1.0), then the confusion matrix after the first three steps would be as shown in Table 6.

Table 6. Confusion matrix where (IPC, PDA, AD, PER) = (0.6, -0.2, 2.0, 0)

Correct			Classification		
Category	**"A"**	**"B"**	**"C"**	**"D"**	**"E"**
A	0.400	0.320	0.160	0.080	0.040
B	0.218	0.400	0.218	0.109	0.055
C	0.100	0.200	0.400	0.200	0.100
D	0.055	0.109	0.218	0.400	0.218
E	0.040	0.080	0.160	0.320	0.400

Then after adding random variation around the incorrect probability assignments in step 4, and renormalizing in step 5, the resulting confusion matrix would look something like the randomly generated confusion matrix shown in Table 7.

Table 7. Example of randomly generated confusion matrix where (IPC, PDA, AD, PER) = (0.6, -0.2, 2.0, 1.0)

Correct Category	Classification "A"	"B"	"C"	"D"	"E"
A	0.349	0.438	0.106	0.082	0.025
B	0.015	0.439	0.291	0.183	0.073
C	0.034	0.225	0.377	0.301	0.064
D	0.107	0.088	0.085	0.512	0.207
E	0.010	0.008	0.098	0.469	0.415

For a selected sample size, *N*, a "simulation run" executes the following:

1. Randomly select the base rate probability for each classification
2. Generate the confusion matrices for each rater and the system
3. Use the base rate probability and confusions matrices to randomly generate *N* cases.
4. Estimate system accuracy (using method described above)
5. Compare estimated system accuracy to "true" system accuracy, where there are two measures of true system accuracy
    a. Expected accuracy (i.e. *P*(*A*)**P*("A"|*A*) + *P*(*B*)**P*("B"|*B*) + …)
    b. Proportion correct in sample

When *K-Assumptions* are satisfied

This section examines circumstances where the assumptions implicit in Kappa are satisfied. That is to say where the raters are selecting from a single confusion matrix of the structure shown in Table 1 and where the system confusion matrix also has the same well-behaved structure.

Illustrated in Figure 1 is the asymptotic behavior of the estimation method. The simulation results depicted in Figure 1 had five categories, three experts each with 60% accuracy, 5000 test cases for each run, and where there are 10 runs each with system accuracy set to .1, .3, .5, .7 and .9 respectively.

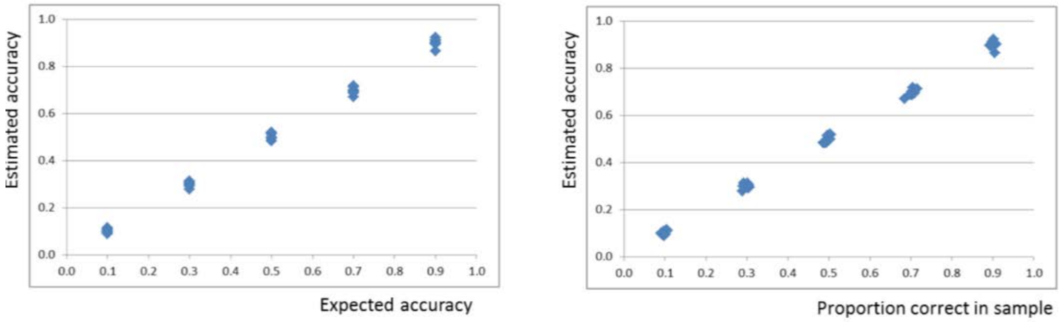


Figure 1. Estimated vs. true system accuracy from simulations with accuracy of three experts each at 0.6, sample size at 5000, with equal error probabilities and equal problem difficulty. (Kappa = 0.251)

The results depicted in Figure 1 indicate that estimates of system accuracy cluster tightly around true system accuracy. When true system accuracy is 0.1, which is less accurate than random guessing (0.2), estimates of system accuracy cluster tightly around 0.1. When true system accuracy is 0.9, which is far better than the raters' accuracy (0.6), estimates of system accuracy cluster tightly around 0.9. Across all fifty simulation runs the average value of Kappa was just 0.251.

The results below depict what happens when sample size and rater accuracy are varied. Figures 2-4 depict the results of fifty simulation runs with a sample size of 200 per run and rater expert accuracy is set to .4, .6 and .8 respectively.

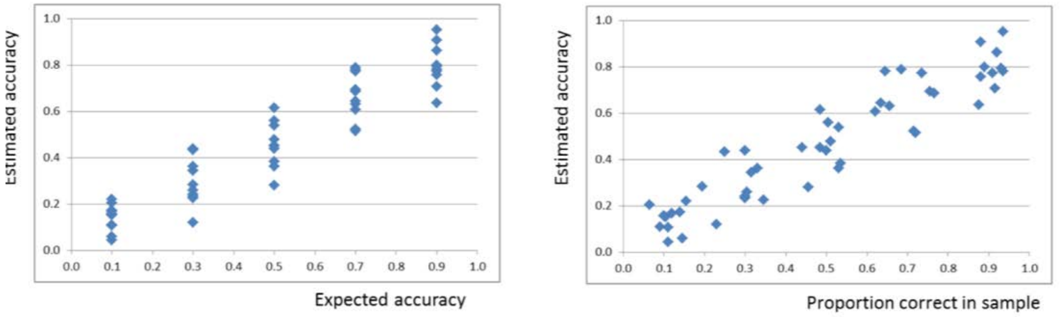


Figure 2. Estimated vs. true system accuracy from simulations with accuracy of three raters each at 0.4, sample size at 200, with equal error probabilities and equal problem difficulty. (Kappa = .065)

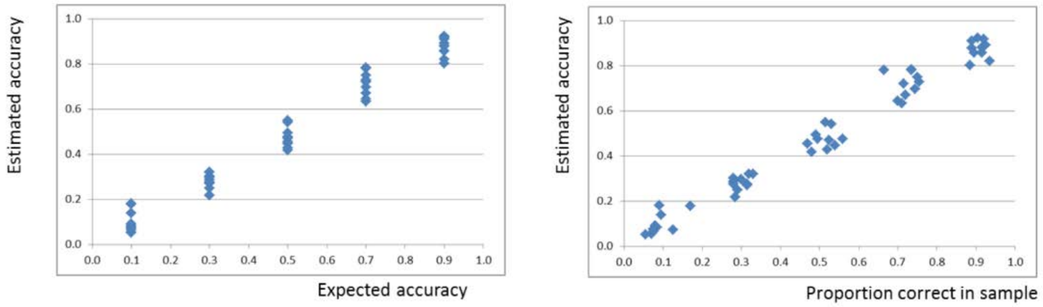


Figure 3. Estimated vs. true system accuracy from simulations with accuracy of three raters each at 0.6, sample size at 200, with equal error probabilities and equal problem difficulty. (Kappa = .255)

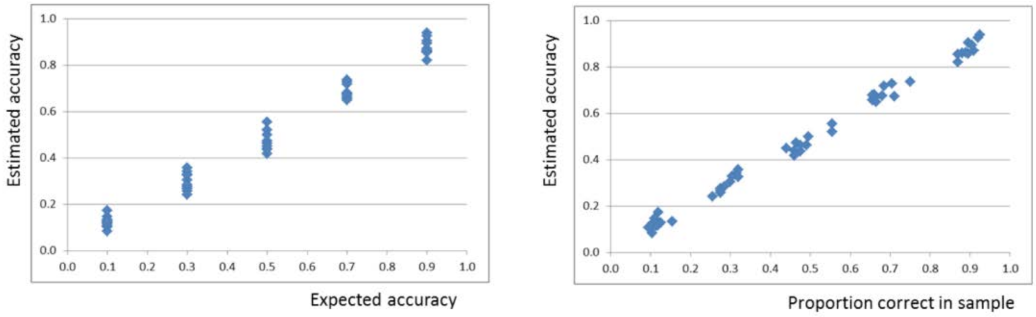


Figure 4. Estimated vs. true system accuracy from simulations with accuracy of three raters each at 0.8, sample size at 200, with equal error probabilities and equal problem difficulty. (Kappa = .562)

The results shown in Figures 2-4 indicate that the correspondence between estimated and true system accuracy improves rapidly as rater accuracy improves. Even when the raters are just 60% accurate, estimates of system accuracy are consistently within ± 0.1 of true system accuracy.

Figures 5-7 depict results when sample size is further reduced to just 50 cases per run. When rater accuracy is 0.4 there is little correspondence between estimated and true system accuracy. However when rater accuracy is 0.6 and 0.8 this correspondence improves quickly.

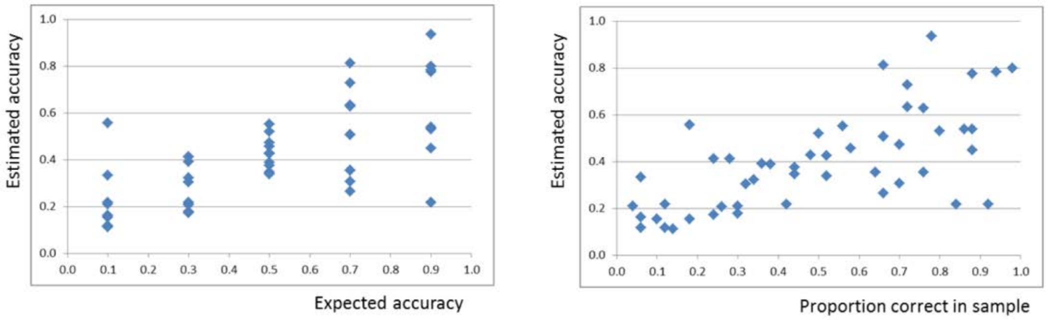


Figure 5. Estimated vs. true system accuracy from simulations with accuracy of three experts each at 0.4, sample size at 50, with equal error probabilities and equal problem difficulty. (Kappa = .060)

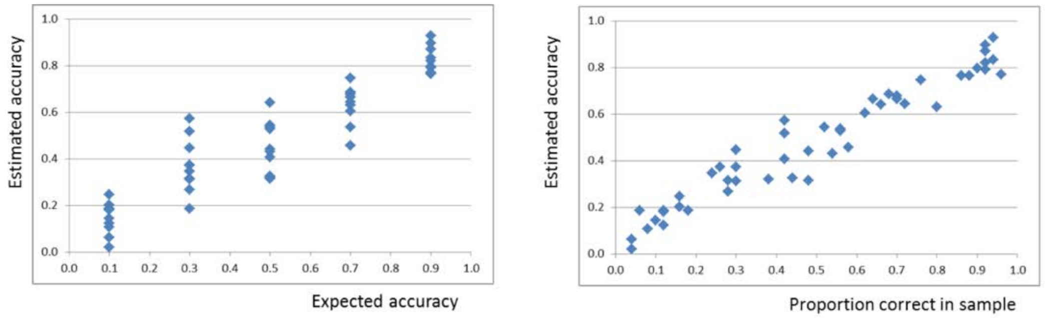


Figure 6. Estimated vs. true system accuracy from simulations with accuracy of three experts each at 0.6, sample size at 50, with equal error probabilities and equal problem difficulty. (Kappa = .244)

Note that in Figures 6 and 7 the two measures of true system accuracy yield slightly different results. Estimated accuracy corresponds more closely to proportion correct in sample than to expected accuracy. This occurs because the proportion correct in a sample varies according to a binomial distribution defined by system accuracy. So even if there is perfect correspondence between estimated accuracy and proportion correct (as is the case when rater accuracy is set to 1.0), the standard deviation of the estimate around expected accuracy (E_a) would still be equal to $(E_a \cdot (1-E_a)/N)^{\frac{1}{2}}$.

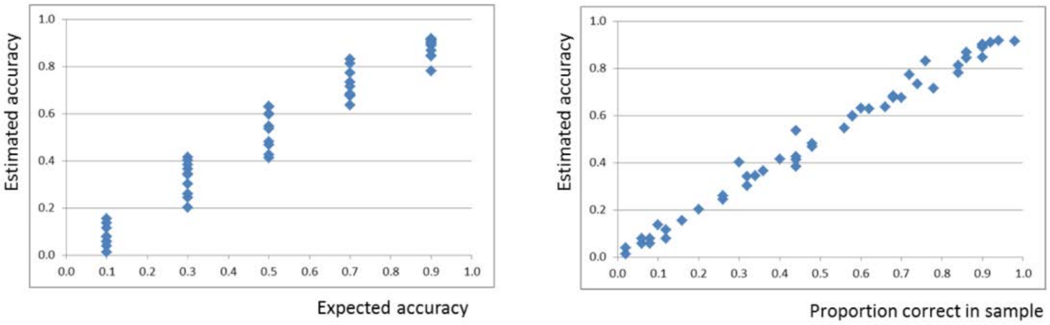


Figure 7. Estimated vs. true system accuracy from simulations with accuracy of three experts each at 0.8, sample size at 50, with equal error probabilities and equal problem difficulty. (Kappa = .546)

In summary, when the *K-assumptions* are satisfied, the estimation method exhibits an orderly relationship between estimated and true system accuracy. Estimates of system accuracy are unbiased, and the correspondence between true and estimated system accuracy improve rapidly as rater accuracy and sample size increase.

When *K-Assumptions* are substantially violated

In practice it is difficult to imagine a circumstance where the *K-assumptions* and the implied *K-properties* are satisfied. All raters are not equally accurate; some are typically more experienced and expert than others. All types of errors are not equally probable; this property is certainly false when the categories are naturally ordered or when the raters have some idea of which categories have the highest base rates. And all problems are not equally difficult; unless the test cases are carefully pre-selected and therefore unrepresentative of real world diversity.

In this section the behavior of the estimation method is examined in cases where the *K-properties* are violated. In all of the simulation runs summarized below the *K-properties* of equal rater accuracy, equal problem difficulty, and equal error probabilities are substantially violated. Specifically:

Rater accuracy (IPC) was varied by .1. For example, instead of three raters with .6 accuracy, initial accuracy would be set to .5, .6 and .7 respectively.

Problem difficulty (PDA) was varied by .2. For about a third of the test cases rater and system accuracy were reduced by .2 (or set to a minimum of 0.0) and for about another third accuracy was increased by .2 (or set to the maximum of 1.0).

Asymmetric dispersion (AD) was set to 2.0. An incorrect answer that is 'next to' the correct answer is twice as likely as one two steps removed and 4 times as likely as one 3 steps removed, etc.

Error probabilities were randomly varied by up to 100% (PER=1.0). For example, if the error probability is initially set to .2 then that error probability would be randomly selected from the range [0, .4]. This random variation is done independently for each error probability.

To appreciate the magnitude of impact of these parameter settings consider again Tables 5 and 7 above. Table 5 is exactly the confusion matrix that results when initial rater accuracy is set to .6 and the *K-properties* are satisfied. Table 7 is representative of about 1/3 of the cases when initial rater accuracy is set to .6 but with the above parameter settings. It seems fair to characterize Table 7 as a substantial variation from Table 5.

All of the simulation runs in this section use the above parameter settings to systematically and then randomly vary the rater and system confusion matrices. The results shown in Figure 8 illustrate the asymptotic behavior of the estimation method when the *K-properties* are substantially violated. Note that when system accuracy is preset to .1 and .9, expected accuracy is .133 and .867 respectively. This occurs because problem difficulty is varied plus and minus 0.2, but accuracy can be no lower than 0.0 or higher than 1.0. So for example when system accuracy is preset to 0.1, one third of the problems have system accuracy reset to 0.3, one third stay at 0.1 and the remaining third are reset to 0.0; then averaged expected system accuracy is then .133.

There is a linear relationship between estimated and true accuracy. There is also some bias in the estimates; estimated accuracy is too high when true system accuracy is low and estimated accuracy is to low when true system accuracy is high. Note though that when the system was more accurate than the raters the estimates of system accuracy were still consistently higher than the raters' accuracy. The estimate of system accuracy may be conservative, but it is not bounded by the raters' accuracy.

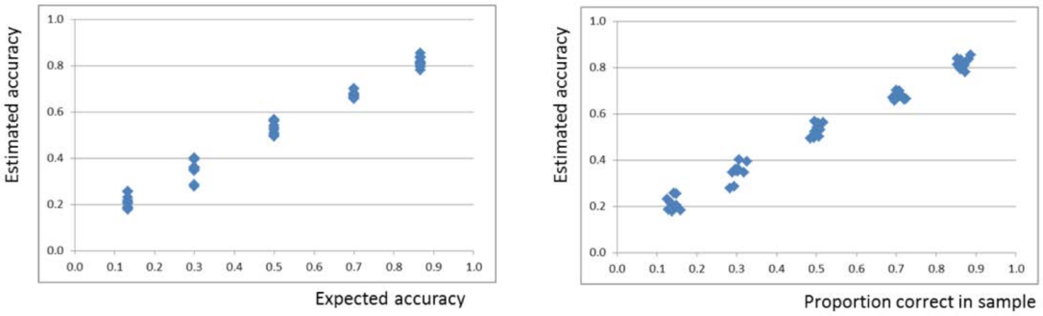


Figure 8. Estimated vs. true system accuracy from simulations with accuracy of three raters at .5, .6 and .7; sample size at 5000 and confusion matrices systematically then randomly varied. (Kappa = 0.305)

There is a straightforward explanation for this estimation bias. The violations of the *K-properties* inflated inter-rater agreement. Because inter-rater agreement is used to estimate rater accuracy, as per Eq. 2, this leads to a slightly inflated estimate of rater accuracy. Inflated estimates of rater accuracy in turn lead to overestimates of the ground truth probabilities for the categories with the highest estimated ground truth probabilities. Finally given the equation for deriving system accuracy from the ground truth probabilities (Eq. 7) this leads to the estimation bias. In comparing Figures 1 and 8, note that Kappa was .251 and .305 respectively; and the average estimated accuracy for the runs in Figure 1 was exactly 0.60 and the average estimated rater accuracy for the runs in Figure 8 was 0.64.

In general violations of the *K-properties* will inflate expected inter-rater agreement with one exception. Differences between rater accuracy decreases rather than increases expected inter-rater agreement, but the net effect is small when compared to the larger opposite effect of the other violations. For example, if overall rater accuracy is set to .6 and then varied by.2 (i.e. rater accuracy set to .4, .6, .8 respectively) and true system accuracy is 0.9 then estimated accuracy will be approximately 0.924 – a 0.024 overestimate. But if instead problem difficulty is varied by the same amount (.4, .6, .8 respectively) then system accuracy will be approximately 0.857 – a 0.043 underestimate. Varying dispersion by 100% around the error probabilities results in an approximate 0.036 underestimate, and setting asymmetric dispersion to 2.0 results in a 0.068 underestimate.

In Figures 9-11 the sample size is 200 cases per run and expected rater accuracy is set to .4, .6 and .8 respectively. In Figures 12-14 sample size is reduced to 50 cases per run. Except for the bias toward underestimating high system accuracy (and overestimating low system accuracy) these results are similar to the results with the matrices that satisfy the *K-properties*. Increasing rater accuracy and sample size both decrease the variance of the estimate. The estimation bias is pronounced when rater accuracy is very low (0.4), noticeable when rater accuracy is moderate (0.6), and appears negligible when rater accuracy is high (0.8).

In practice, most efforts to evaluate system accuracy address systems that are hypothesized to perform well. For such evaluations the estimates derived from this method become increasingly conservative as the ratings of the experts are increasingly suspect.

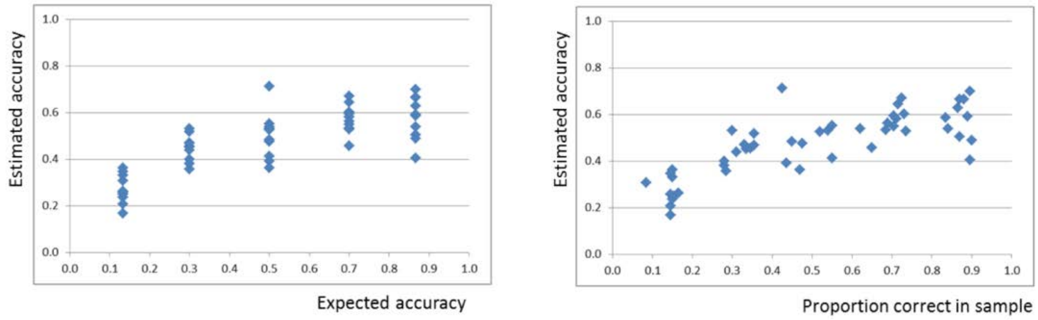


Figure 9. Estimated vs. true system accuracy from simulations with accuracy of three raters at .3, .4 and .5; sample size at 200 and confusion matrices systematically then randomly varied. (Kappa = .142)

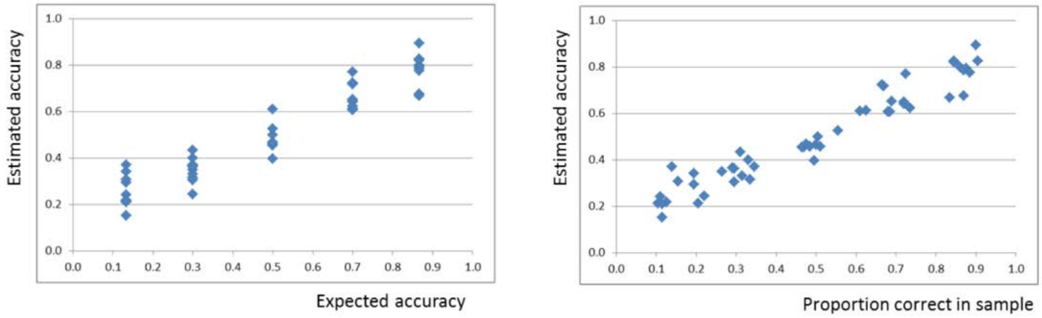


Figure 10. Estimated vs. true system accuracy from simulations with accuracy of three raters at .5, .6 and .7; sample size at 200 and confusion matrices systematically then randomly varied. (Kappa = .306)

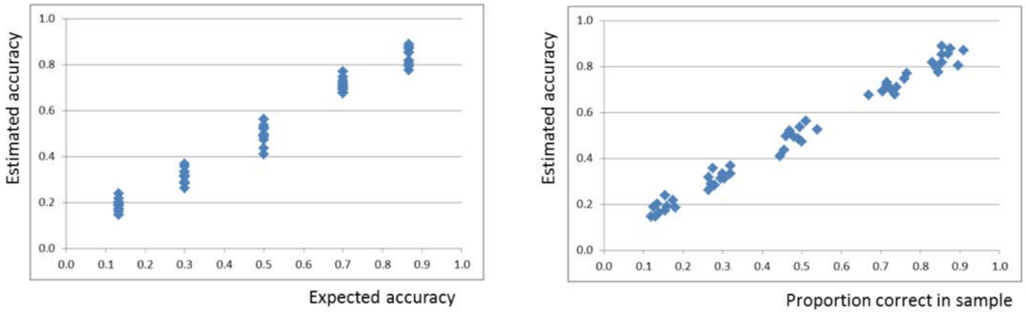


Figure 11. Estimated vs. true system accuracy from simulations with accuracy of three raters at .7, .8 and .9; sample size at 200 and confusion matrices systematically then randomly varied. (Kappa = .578)

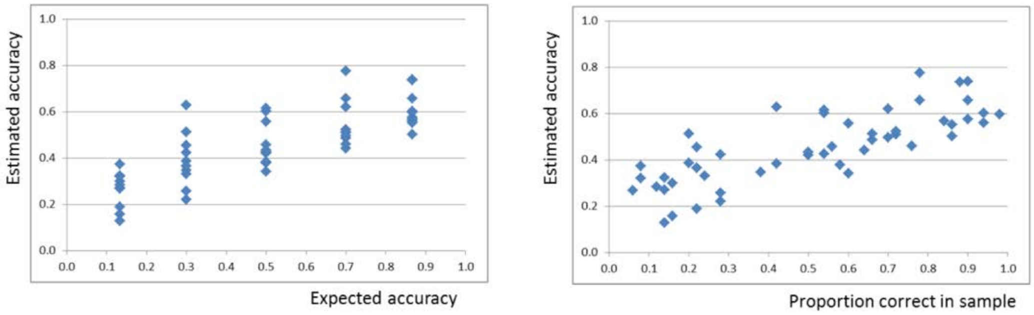


Figure 12. Estimated vs. true system accuracy from simulations with accuracy of three raters at .3, .4 and .5; sample size at 50 and confusion matrices systematically then randomly varied. (Kappa = .144)

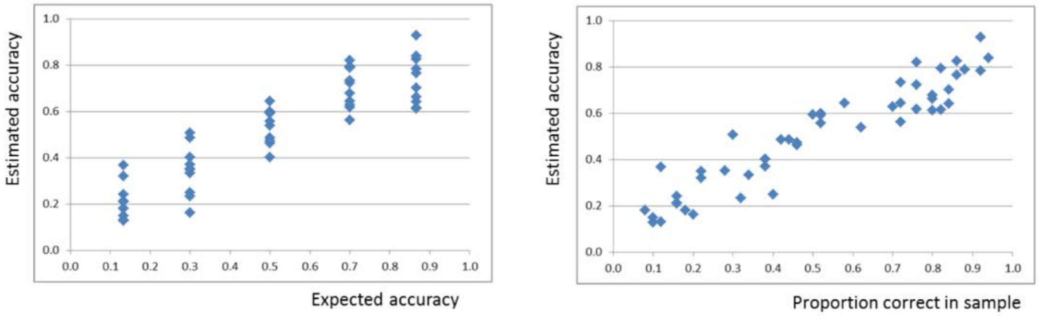


Figure 13. Estimated vs. true system accuracy from simulations with accuracy of three raters at .5, .6 and .7; sample size at 50 and confusion matrices systematically then randomly varied. (Kappa = .311)

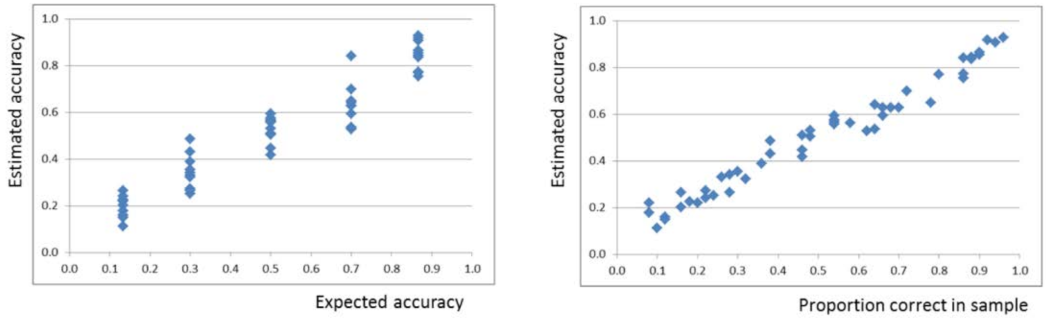


Figure 14. Estimated vs. true system accuracy from simulations with accuracy of three raters at .7, .8 and .9; sample size at 50 and confusion matrices systematically then randomly varied. (Kappa = .586)

Discussion

The objective in this study was to demonstrate that it is feasible to reliably estimate the accuracy of system classifications when ground truth can only be estimated with fallible expert ratings. The simulation results described herein provide evidence for the claims stated in the introduction, namely that reliable estimates of system accuracy can be obtained from fallible expert ratings under a diverse conditions, that the reliability of these estimates is approximately the same whether the system is more or less accurate than the expert raters, and that the conditions under which these accuracy estimates become unreliable are predictable (e.g., inter-rater agreement is low and sample size is small).

In the estimation method the level of inter-rater agreement is used to estimate the overall accuracy of the expert ratings, Bayesian updating based on the estimated expert accuracy is used to estimate a "ground truth" probability for each classification, and finally system accuracy is estimated by comparing the relative frequency that the system assignment agrees with the most probable classification at different probability levels.

Although the estimation method was derived from assumptions that are implicit in the Kappa statistic (and how it is often used), a simulation analysis shows that the accuracy of the estimates of system accuracy are robust against substantial variations from the rater behavior implied by those assumptions. The accuracy of the estimates of system accuracy is driven primarily by overall rater accuracy (which can be estimated from inter-rater agreement) and sample size.

Recommended use and uses to avoid

The simulation results presented herein suggest an overall data collection and estimation approach where measured inter-rater agreement is used to determine the number of test cases needed to obtain high confidence in system accuracy estimates. For example for five category problems with three raters if initial data collection indicates that Kappa is around .3 then data collection should continue for at least 200 cases. This would be a sufficient number of cases to obtain 90% "confidence" that estimated accuracy is within .1 of true accuracy. On the other hand, if Kappa is around .55 then a sample size of 100 cases is sufficient to ensure the same "confidence interval." As the number of raters and categories varies, so does the parametric relationship between sample size and confidence in estimates of system accuracy; so additional simulation runs such as those shown here would be needed to determine sample size requirements.

In this approach all test cases are useable, even ones where raters substantially disagree. This makes it feasible to randomly select test cases from the population of problems from which the system is likely to be applied which in turn should facilitate the ability generalize test results to practice.

As noted above, violations of the *K-properties* (equal rater accuracy, problem difficulty and error probabilities) will bias the estimate of system accuracy. The magnitude of this bias interacts with overall rater accuracy. If system accuracy is high and rater accuracy low then the estimation procedure described herein will likely substantially under estimate system accuracy. In the above simulations, for example, on five category problems when true system accuracy was .9 and rater accuracy was .4 the estimate of system accuracy was around .6. Consequently when Kappa is very low (e.g. less than .2) it would be helpful to examine the inter-rater agreement data for patterns that suggest violations of the *K-properties*. For example, the K-property of equal error probabilities implies that all pairwise disagreements are equally likely (e.g. "AB" as likely as "AE") and a statistical test can be performed to help determine if this pattern is violated. If it is, then the estimate of system accuracy can be adjusted upwards. There is much work to be done to determine how and when such adjustments should be made, but doing so seems feasible.

The estimation method described herein is specifically intended for cases where each rater is an independent measure of ground truth classifications. The procedure assumes the causal structure shown in Figure 15-10a.

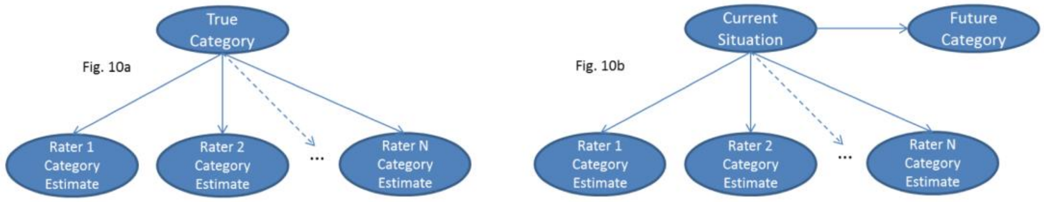

True Category
Fig. 10a
Rater 1 Category Estimate
Rater 2 Category Estimate
...
Rater N Category Estimate
Current Situation
Future Category
Fig. 10b
Rater 1 Category Estimate
Rater 2 Category Estimate
...
Rater N Category Estimate

Figure 15. Assumed causal relationship between ground truth and expert ratings vs. causal structure of forecasting tasks

There are many applications that involve aggregation of independent estimates from multiple individuals but do not have the causal structure shown in Figure 15-10a. For many such applications use of the estimation method described here would be inappropriate. For example, it is becoming common practice in forecasting to systematically combine the ratings of multiple independent forecasters (e.g. Surowieki, 2005). Although the estimation method presented here could be mechanically applied to such forecasting tasks, such an application may yield spurious results. Forecasting tasks do not have the causal structure shown in Figure 15-10a, but have a causal structure closer to the one shown in Figure 15-10b where expert ratings are not in any sense direct measures of the future outcomes. On the other hand the estimation method can and has been used to retrospectively assess whether a forecasted outcome actually occurred. For example Lehner et al. (2012) examined the accuracy of the imprecise forecasts typically found in published forecasts by using multiple raters to retrospectively assess whether the forecasted outcome occurred and then using an estimation method similar to the one presented here to estimate the accuracy of a collection of forecasts. Similarly Levitt and Lehner (2011) applied a variation of this method to resolve disagreeing historical judgments as to the timeframe when key developments occurred in the maturation of new technologies.

The distinction between Figures 15-10a and 15-10b is essentially the distinction between medical diagnosis and medical prognosis. It would be appropriate to apply the method to estimate the accuracy of a new diagnostic system by comparing system diagnoses to those of medical professionals, but it would be inappropriate to use it to estimate the accuracy of a new system's prognoses by comparing them to the prognoses of medical professionals.

In general it is important that the causal structure relating the rater and system selections to ground truth match the structure assumed by the estimation method. The process of collecting ratings from the experts should be engineered to ensure this causal structure; such as by ensuring that the expert ratings are independent and to the extent possible having available the same data for each rater for each test case.

The estimation method presented here was developed to address test and evaluation of an automated classification system after development. However it does seem feasible to also employ this approach during system development. Specifically the estimation method could be used to develop training data sets with a probability distribution of correct classifications for each training case.

Related and future research

The research presented in this paper had the very specific goal of demonstrating that it is feasible to reasonably estimate system accuracy using fallible expert ratings even when the system is substantially more accurate than the experts. Nothing in this paper would support a claim that the estimation method presented here is in any sense optimal. There are many opportunities for improvement. Three suggestions are offered below.

First, the estimation method was designed for use with classification problems for which there is no natural ordering to the categories. The simulation results suggest that the method is robust even when there is a natural ordering, but the accuracy of estimates of system accuracy would likely be improved if the method is modified to specifically account for the fact that certain types of errors are more likely than others. For example, if the natural ordering is *A, B, C, D, E*, then a rating of “A” should be more evidence for category B than for category E. The method presented here treats *B* and *E* equally.

Second, as noted above, it should be feasible to develop statistical procedures to estimate whether and to what degree *K-properties* are violated. From these estimates it should be also feasible to adjust the system accuracy estimates to correct for bias. This area is unexplored.

Third, the estimation method presented here is entirely algebraic. Everything is derived directly from some percent-of-agreement statistics. No effort was made to estimate base rates and confusion matrices that represent a “best fit” to the inter-rater agreement data. But there are best fit methods that could be used for this purpose. For example, the non-linear optimization methods in Latent Class Analysis (McCutcheon, 1987) could be used to find maximum likelihood estimates for the base rate and confusion matrix probabilities. Both Uebersax (1988) and Carpenter (2008) applied this approach to binary classification problems; and Carpenter also used Bayes inference to aggregate ratings and estimate classification probabilities. Similarly one could use non-linear optimization to find base rates and confusion matrix probabilities that minimize the difference between expected and observed relative frequency of each inter-rater pair (relative frequency of "AA", "AB", "AC" …). It remains an open and interesting question as to whether use of such optimization methods would yield better results.

Acknowledgements

This research was funded by The MITRE Corporation under project number 05MSR003.

References

Armstrong, J. S. (2006). How to make better forecasts and decisions: Avoid face-to-face meetings. *The International Journal of Applied Forecasting*, *5*, 3-15.

Bishop, M. A., & Trout, J. D. (2002). 50 years of successful prediction modeling should be enough. *Philosophy of Science*, *69*(S3), S197-S208.

Carpenter, B. (2008). *Multi-level Bayesian models of categorical data annotation*. Manuscript found at http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.174.1374&rep=rep1&type=pdf.

Dawes, R. M. (1979). The robust beauty of improper linear models. *American Psychologist*, *34*, 571- 582.

Grove, W. H., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2001). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment, 12*, 19-30.

Gwet, K. L. (2010). *Handbook of inter-rater reliability* (2nd Ed.). Advanced Analytics, LLC.

Lehner, P., Michelson, A., Adelman, L., & Goodman, A. (2012). Using Inferred Probabilities to Measure the Accuracy of Imprecise Forecasts. *Judgment and Decision Making*, *7*(6), 728-740.

Levitt, T. & Lehner, P. (2011) Baseline Judgment Estimation and Challenge Question Answer Assignment for the FUSE Program. Technical Report, The MITRE Corporation, December 28, 2011.

McCutcheon, A. L. (1987). *Latent class analysis*. Thousand Oaks, California: Sage Publications.

Surowiecki, J. (2005). *The wisdom of crowds*. Anchor Books.

Tetlock, P. E. (2005). *Expert political judgment: How good is it? How can we know?* Princeton, N.J.: Princeton University Press.

Uebersax, J. S. (1988). Validity inferences from inter observer agreement. *Psychological Bulletin*, *104*(3), 405-416.

Appendix

Proof of Theorem 1

Restating the assumptions:

AA1. For each case there is a unique correct classification

AA2. For each case raters independently assign classifications

AA3. Expected agreement between raters increases as expected rater accuracy increases.

Begin with a few definitions.

Definition of *correct classification* in AA1: For each case there is a vector <c_1, c_2 … c_n> where for some index i, $c_i = 1$ and the remaining values are 0.

Definition of *independent assignment* in AA3: For each case, the probability that a rater will select a class is conditionally independent of the other raters' selections.

Independent assignments allow the description of each rater's selection behavior as a probability vector. That is to say, for each case each rater has a selection probability for each category. These will be called *selection vectors*.

Definition of *rater accuracy* in AA3: For M raters and N cases, rater accuracy is defined as the total proportion of correct selections.

For example, if there are 10 cases and three raters who make correct assignments in 7, 5 and 9 of the cases respectively, then rater accuracy = 0.7.

The three lemmas below all use the same proof strategy. Begin with any two selection vectors that are not identical. Construct a selection vector that is the average of the two. The average vector will necessarily have the same expected accuracy but a different level of expected agreement than the original two vectors. If the average vector has higher/lower expected agreement, then create a new vector by slightly reducing/increasing the probability of correct assignment in the average vector. When the change is sufficiently small the new vector will have higher/lower expected accuracy and lower/higher expected agreement than the original two vectors. Most of the algebraic complexity in these proofs is the result of showing one way to calculate a change that is always “sufficiently small”.

Lemma 1:

To ensure AA1-AA3 within each case all raters must behave as though they are selecting a category using the same selection vector.

Proof: Let $<p_{11}, p_{12} \ldots p_{1n}>$ and $<p_{21}, p_{22} \ldots p_{2n}>$ be the selection vectors of 2 raters for a specific case; where some probabilities do not agree (e.g. $p_{11} \neq p_{21}$). For purposes of the proofs below, assume that category 1 is the correct category. (The arguments below apply no matter which category is correct.)

Below it is shown how to construct from two different selection vectors a common selection vector for both raters where expected accuracy is lower but expected agreement higher. Consequently unless the two raters have the same selection vector, there will always be another pair of vectors with lower expected accuracy and higher expected agreement – violating AA3.

Set $p_i = (p_{1i} + p_{2i})/2$, $e_i = (p_{1i} - p_{2i})/2$, $d = \left(e_1^2/\left(2*(p_2 - p_1)\right)\right)$,if $p_1 < p_2$, $d = -\left(e_1^2/\left(2*(p_2 - p_1)\right)\right)$, and $d = 0$ if $p_1 = p_2$

$$
\begin{aligned}
&\text{For selection vectors } < p_{11}, p_{12} \ldots p_{1n} > \text{ and } < p_{21}, p_{22} \ldots p_{2n} > \\
&\quad \text{Expected accuracy} = p_1 \\
&\text{Expected agreement} = p_{11} * p_{21} + p_{12} * p_{22} + \ldots + p_{1n} * p_{2n} \\
&\qquad = (p_1 + e_1)*(p_1 - e_1) + (p_2 + e_2) \\
&\qquad\ *(p_2 - e_2) + \ldots + (p_n + e_n)*(p_n - e_n) \\
&\qquad = p_1^2 + p_2^2 + \ldots + p_n^2 - e_1^2 - e_2^2 - \cdots - e_n^2
\end{aligned}
\tag{A1}
$$

$$
\begin{aligned}
&\text{For selection vectors } < p_1, p_2 \ldots p_n > \text{ and } < p_1, p_2 \ldots p_n > \\
&\quad \text{Expected accuracy } = p_1 \\
&\text{Expected agreement } = p_1^2 + p_2^2 + \ldots + p_n^2
\end{aligned}
\tag{A2}
$$

$$\begin{aligned} &\text{For selection vectors } < p_1, p_2 \ldots p_n > \text{ and} < p_1, p_2 \ldots p_n > \\ &\text{Expected accuracy } = p_1 \\ &\text{Expected agreement } = p_1^2 + p_2^2 + \ldots + p_n^2 \end{aligned} \qquad \text{(A3)}$$

Expected accuracy in (A1) is higher than in (A3), but expected agreement is lower; where the common selection vector in (A3) was constructed from a difference between the vectors in (A1). Consequently, whenever there is a difference between the selection vectors of two raters a selection probability vector for the two raters can be constructed with lower expected accuracy but high expected agreement.

Within each case if the selection vectors of the raters differ AA3 is not guaranteed. ***

Lemma 2:

To ensure AA1-AA3 within each case the error probability is the same for all incorrect categories.

Proof: From Lemma 1 it is known that AA1-AA3 imply that for each case all raters have the same selection vector. Let that vector be <p_1, p_2 ... p_n>. Assume category 1 is the correct assignment and that the remaining probabilities are not all equal.

Below it is shown how to construct selection vector, with equal probability for all incorrect assignments, where expected accuracy is higher but expected agreement lower. Consequently the error probabilities are unequal, there will always be a vector with higher expected accuracy and lower expected agreement – violating AA3.

Set $p_e = (p_2 + \ldots + p_n)/(n-1)$, $e_i = (p_i - p_e)$ for all $i > 1$, set $e_{\min} = \min(|e_2| \ldots |e_n|)$ and $d = e_{\min}^2 / 2$.

Note that (e_2 + … e_n) = 0 and that there are at least 2 e_i that are not zero.

$$\begin{aligned}
\text{For the vector } &< p_1, p_2 \ldots p_n >, \\
\text{Expected accuracy } &= p_1 \\
\text{Expected agreement } &= p_1^2 + p_2^2 + \ldots + p_n^2 \\
&= p_1^2 + \left(p_e + e_2\right)^2 + \ldots + \left(p_e + e_n\right)^2 \qquad \text{(A4)} \\
&= p_1^2 + p_e^2 + \ldots + p_e^2 + 2p_e \begin{pmatrix} e_2 + e_3 \\ + \ldots + e_n \end{pmatrix} + e_2^2 + e_3^2 + \ldots e_n^2 \\
&= p_1^2 + \left(n-1\right) p_e^2 + e_2^2 + e_3^2 + \ldots e_n^2
\end{aligned}$$

$$\begin{aligned}
\text{For the vector } &< p_1, p_e \ldots p_e > \\
\text{Expected accuracy } &= p_1 \qquad \text{(A5)} \\
\text{Expected agreement } &= p_1^2 + \left(n-1\right) p_e^2
\end{aligned}$$

$$\begin{aligned}
\text{For the vector } &< p_1 + d,\ p_e - d,\ p_e \ldots\ p_e > \\
\text{Expected accuracy } &= p_1 + d \ = p_1 + \ e_{\min}^2 / 2 \\
\text{Expected agreement } &= \left(p_1 + d\right)^2 + \left(p_e - d\right)^2 + p_e^2 + \ldots + p_e^2 \\
&= p_1^2 + \ \left(n-1\right) p_e^2 + \ 2p_1 d \ - \ 2p_e d \ + \ 2d^2 \qquad \text{(A6)} \\
&= p_1^2 + \ \left(n-1\right) p_e^2 + \ 2d\left(p_1 - p_e\right) \ + \ 2d^2 \\
&= p_1^2 + \ \left(n-1\right) p_e^2 + \ e_{\min}^2 * \left(p_1 - p_e\right)\Big) \ + \ e_{\min}^4 / 2
\end{aligned}$$

Since $e_{\min}{}^2*(p_1 - p_e)) + e_{\min}{}^4/2 < e_{\min}{}^2 + e_{\min}{}^2 <= e_2{}^2 + e_3{}^2 + \ldots\ e_n{}^2$, expected agreement in (A4) is higher than expected agreement in (A6) even though expected accuracy is lower.

Consequently, whenever the probability of incorrect assignment is unequal, there will always be a selection vector with higher expected accuracy and lower expected agreement, violating AA3.

Within each case and selection vector if the error probabilities are unequal AA3 is not guaranteed.

***

Lemma 3:

To ensure AA1-AA3 the selection vector must be the same across all cases.

Proof: Lemmas 1 and 2 show that AA1-AA3 imply that for each case the raters have identical selection vectors of the form <p_e … p_c … p_e> where p_c is the probability of assigning the correct category and $p_e = (1\text{-}p_c)/(n\text{-}1)$ where n is the number of categories.

Below it is shown that across different cases the selection vectors must have the same values for p_c (and therefore p_e) else a violation of AA3 can be constructed.

Let p_{c1} and p_{c2} be the probability of correct assignment on two different cases, and p_{e1} and p_{e2} the corresponding error probabilities. For each case, order the cases such that the correct assignment is first. So for all raters the probability vector is <$p_{c1}, p_{e1}, \ldots\ p_{e1}$> for case 1 and <$p_{c2}, p_{e2}, \ldots, p_{e2}$> for case 2, but the categories may be in a different order. The proof below makes no reference to matching categories across cases so this ordering does not affect the proof.

Set $p_c = (p_{c1} + p_{c2})/2$, $p_e = (p_{e1} + p_{e2})/2$, $e_c = (p_{c1} - p_c)$, $e_e = (p_{e1} - p_e)$, $e_{\min} = \min(|e_c|, |e_e|)$, $d = e_{\min}^2/2$

For two cases with accuracy $p_{c1} \neq p_{c2}$

$$\begin{aligned}
\text{Expected accuracy} &= p_c \\
\text{Expected agreement} &= \left(p_{c1}^2 + (n-1)\,p_{e1}^2 + p_{c2}^2 + (n-1)\,p_{e2}^2\right)/2 \\
&= \left(\begin{array}{l} (p_c + e_c)^2 + (n-1)(p_e + e_e)^2 \\ +(p_c - e_c)^2 + (n-1)(p_e - e_e)^2 \end{array}\right)\Big/2 \qquad \text{(A7)} \\
&= \left(2p_c^2 + 2(n-1)\,p_e^2 + 2e_c^2 + 2(n-1)\,e_e^2\right)/2 \\
&= p_c^2 + (n-1)\,p_e^2 + e_c^2 + (n-1)\,e_e^2
\end{aligned}$$

For two cases with accuracy $p_{c1} = p_{c2}$

$$\begin{aligned}
\text{Expected accuracy} &= p_c \qquad \text{(A8)} \\
\text{Expected agreement} &= p_c^2 + (n-1)\,p_e^2
\end{aligned}$$

$$\begin{aligned}
\text{For two cases with accuracy vectors } &< p_c + d,\ p_e - d,\ p_e \ldots p_e\big) > \\
\text{Expected accuracy } &= p_c + d \\
\text{Expected agreement } &= \left(p_c + d\right)^2 + \left(p_e - d\right)^2 + \left(n-2\right) p_e^2 \\
&= p_c^2 + \left(n-1\right) p_e^2 + 2 p_c d + d^2 - 2 p_e d + \ d^2 \\
&= p_c^2 + \left(n-1\right) p_e^2 + 2d\left(p_c - p_e\right) + 2d^2 \qquad \text{(A9)} \\
&= p_c^2 + \left(n-1\right) p_e^2 + 2\left(e_{\min}^2/2\right) \\
&\left(p_c - p_e\right) + 2\left(e_{\min}^2/2\right)^2 \\
&= p_c^2 + \left(n-1\right) p_e^2 + \left(p_c - p_e\right) e_{\min}^2 + e_{\min}^4/2
\end{aligned}$$

Since $e_{\min}^2 * \left(p_c - p_e\right)\big) + e_{\min}^4/2 < e_{\min}^2 + e_{\min}^2 <= e_c^2 + e_e^2$, expected agreement in (A7) is higher than expected agreement in (A9) even though expected accuracy is lower.

Consequently, whenever the probability of correct assignment across cases is unequal, there will always be a probability vector that is the same across cases with higher expected accuracy and lower expected agreement, violating AA3. Across cases, if the selection vectors differ then AA3 is not guaranteed. ***

Theorem 1:

AA1-AA3 are ensured if and only if the raters behave as though their selection for each case is determined by a single confusion matrix where the conditional probability of correct assignment is constant and the conditional probability of all incorrect assignments is equal.

Proof: The "only if" necessity portion follows directly from Lemmas 1-3. Sufficiency follows the fact that with a constant conditional probability of correct assignment (P_c) and incorrect assignments (P_e), expected accuracy is P_c and expected agreement is $P_c^2 + \left(n-1\right) P_e^2 = P_c^2 + \left(1 - P_c\right)^2 / \left(n-1\right)$. Clearly expected agreement increases monotonically with P_c. ***

Journal of Modern Applied Statistical Methods
May 2015, Vol. 14, No. 1, 152-167.

ISSN 1538 − 9472

Comparison of Model Fit Indices Used in Structural Equation Modeling Under Multivariate Normality

Sengul Cangur
Duzce University
Duzce, Turkey

Ilker Ercan
Uludag University
Bursa, Turkey

The purpose of this study is to investigate the impact of estimation techniques and sample sizes on model fit indices in structural equation models constructed according to the number of exogenous latent variables under multivariate normality. The performances of fit indices are compared by considering effects of related factors. The Ratio Chi-square Test Statistic to Degree of Freedom, Root Mean Square Error of Approximation, and Comparative Fit Index are the least affected indices by estimation technique and sample size under multivariate normality, especially with large sample size.

Keywords: Structural equation modeling, multivariate normality

Introduction

Modeling methods are employed for studying the phenomena than require the utilization of complex variable set. Structural Equation Modeling (SEM) is preferred when studying the causal relations and the latent constructs among the variables is in question. The reason is it can be used to analyze complex theoretical models and its practicability.

The objective of SEM is to explain the system of correlative dependent relations between one or more manifest variables and latent constructs simultaneously. It serves to determine how the theoretical model that denotes relevant systems is supported by sample data, i.e., estimation of relations between the main constructs. Because there is no single criterion for the theoretical model fit evaluation obtained as a result of SEM, a wide array of fit indices was developed (Schermelleh-Engel and Moosbrugger, 2003; Ding et al., 1995; Sugawara and MacCallum, 1993). Studies conducted through SEM were

Dr. Cangur is an Assistant Professor in the Department of Biostatistics and Medical Informatics. Email at sengulcangur@duzce.edu.tr. Dr. Ercan is a Professor in the Department of Biostatistics. Email at ercan@uludag.edu.tr.

undertaken by using empirical and non-empirical data so as to develop and confirm theory (Bentler and Dudgeon, 1996; Wang et al., 1996; Bentler, 1994).

Simulation studies were conducted to test the robustness of SEM, because the assumptions required usually cannot be verified in practice. Because these studies were conducted in order to verify hypothesis, a known theoretical model was taken as a reference and the behaviors of the most commonly used techniques in specific conditions were observed. The parameter estimations obtained through the estimation techniques based on various distributional conditions and sample size, standard errors and the bias of model fit indices were researched in the studies conducted.

Studies were conducted for recommending and improving the parameter estimation techniques used in SEM and selecting the conditions in which these are to be used (Boomsma and Hoogland, 2001; Wang et al., 1996; Chou and Bentler, 1995; Bentler, 1994). Other studies were conducted by employing various empirical designs so as to examine the effects of factors such as estimation techniques, sample sizes, distributional conditions, number of latent variables, number of manifest variables, the misspecification degree of the model, factor loads, factor correlations, improper solutions, convergence errors on model fit indices make contribution to the SEM literature (e.g., Herzog & Boomsma, 2009; Fan & Sivo, 2007; Sivo et al., 2006; Lei & Lomax, 2005; Marsh et al., 2004; Boomsma and Hoogland, 2001; Fan et al., 1999; Hu & Bentler, 1998, 1999; Wang et al., 1996; Chou and Bentler, 1995; Ding et al., 1995; Marsh & Balla, 1994; Sugawara and MacCallum, 1993; Gerbing & Anderson, 1992).

Hence, a wide array of simulation studies were conducted on model fit indices through various estimation techniques. Unlike these studies, in the current study the inclusion of a higher number of estimation techniques was used. Furthermore, the differentiation of the model structure was agreed to be studied as exogenous factor rather than an effect so as to reach a mutual interpretation. The effects of estimation technique and sample size factors on model fit indices were examined in circumstances in which the multivariate normality assumption was ensured and in the models which were established by taking exogenous (independent) latent variables into consideration in the research. The model fit indices were compared to recommend appropriate model fit indices in line with the effects of these factors.

Methodology

Maximum likelihood estimation technique

Maximum likelihood estimation (MLE) technique is one of the normal theory estimation techniques that is able to provide model parameter estimations simultaneously (Kline, 2011; Chou and Bentler, 1995). Assume a $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ random sample is derived from multivariate normal distribution $\mathrm{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$. In order to achieve $\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}(\boldsymbol{\theta}_0)$, assumed there is population (true) matrix function with $\boldsymbol{\Sigma}_0$, $q \times 1$ size and $\boldsymbol{\theta}_0$ unknown parameter. In this case, MLE function can be defined as in equation (1).

$$\mathrm{F}_{\mathrm{MLE}}(\boldsymbol{\theta}) = log\left|\boldsymbol{\Sigma}(\boldsymbol{\theta})\right| + tr\left(\mathbf{S}\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}\right) - log\left|\mathbf{S}\right| - p \tag{1}$$

$\mathbf{S}$ denotes sample covariance matrix while $\boldsymbol{\Sigma}(\boldsymbol{\theta}_0)$ indicates the covariance matrix of the hypothesized model, tr denotes the trace of matrix and p represents the number of manifest variables (Lee, 2007).

Generalized least squares technique

The GLS technique makes multivariate normality assumption flexible compared to MLE technique, yet also features the assumptions of MLE technique. GLS function can be given as follows.

$$\mathrm{F}_{\mathrm{GLS}}(\boldsymbol{\theta}) = 2^{-1}\ tr\left\{(\mathbf{S} - \boldsymbol{\Sigma})\mathbf{V}\right\}^2 \tag{2}$$

The population and sample covariance matrices are indicated with $\boldsymbol{\Sigma}$ and $\mathbf{S}$ respectively. The $\mathbf{V}$ matrix can be a constant positive definite matrix or a stochastic matrix which converges to $\boldsymbol{\Sigma}_0^{-1}$. The GLS function reduces to the least squares function when $\mathbf{V}$ equals to identity matrix ($\mathbf{I}$) (Lee, 2007).

Asymptotically distribution-free technique

The Asymptotically Distribution-Free (ADF) technique does not require multivariate normality assumption and is based on the calculation of $\mathbf{W}$ weighted matrix and GLS estimation. Accordingly, assume $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ are the independent identically distributed observations of a sample with mean vector μ, covariance matrix $\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}(\boldsymbol{\theta}_0)$ and finite eighth-order moments that is not obliged to be selected from a multivariate normal distribution. $\tilde{\boldsymbol{\theta}}_A$ ADF estimator of $\boldsymbol{\theta}_0$ will be defined as in equation (3) as the vector which minimizes GLS function:

$$\mathrm{F}_{\mathrm{ADF}}(\boldsymbol{\theta}) = 2^{-1}\left[vecs\left\{\mathbf{S}-\boldsymbol{\Sigma}(\boldsymbol{\theta})\right\}\right]' \mathbf{W}^{-1}\left[vecs\left\{\mathbf{S}-\boldsymbol{\Sigma}(\boldsymbol{\theta})\right\}\right] \tag{3}$$

Here *vecs* denotes the column vector which is obtained through derivation of lower triangle matrix components row by row. **W** is the stochastic weighted matrix with positive definite and is assumed to converge to $\boldsymbol{\Sigma}^*$ (Lee, 2007). Many researchers emphasized the requirement to work with large sample sizes so as to ensure that ADF estimations have the desired asymptotical properties (i.e., Bentler & Dudgeon, 1996).

Satorra-Bentler scaled chi square test statistic

The normal theory chi-square statistic can be adjusted for its convergence to the referenced chi-square distribution even if it is not fit for the expected chi-square distribution in circumstances where the normality assumption is violated. Satorra−Bentler scaled χ^2 test statistic can be indicated as follows:

$$\chi^2_{\mathrm{SB}} = \frac{\chi^2_{\mathrm{MLE}}}{\varpi} \tag{4}$$

χ^2_{MLE} denotes the chi-square value of MLE technique. The ϖ constant, also known as the scaling factor, is a function of the model-implied weighted matrix, the multivariate kurtosis index and the degree of freedom for the model (Finney and Distefano, 2006; Chou and Bentler, 1995). Provided that multivariate kurtosis is not in question χ^2_{MLE} value is equal to χ^2_{SB} value, and two chi-square values are obtained as different from each other only on the event of the degree of multivariate kurtosis increases (Finney and Distefano, 2006).

Commonly-used model fit indices in SEM

χ^2 and χ^2 / v Ratio The χ^2 test statistic is an absolute fit index which assumes multivariate normality and is sensitive to sample size (Gerbing and Anderson, 1992). This test statistic

$$\chi^2 = -2\left\{-\tfrac{1}{2}(n-1)\left[tr\left(\mathbf{S}\boldsymbol{\Sigma}^{-1}\right) + log\left|\boldsymbol{\Sigma}\right| - log\left|\mathbf{S}\right| - p\right]\right\} = (n-1)\mathrm{F} \qquad (5)$$

is distributed the central χ^2 with degree of freedom $\{½\,p\,(p + 1)\} - t$ in large samples. Here p, denotes the number of observed variables and t symbolizes the number of estimated independent parameters. **S** denotes unrestricted sample covariance matrix whereas $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ denotes restricted covariance matrix. It is said that the larger the likelihood related to χ^2, the closer the fit between the hypothesized model and the perfect model (Herzog and Boomsma, 2009; Hu and Bentler, 1995). This statistic is dependent on sample size. With increasing sample size and a fixed number of degree of freedom, the χ^2 value increases. This signs to the problem that plausible models might be rejected (Schermelleh-Engel and Moosbrugger, 2003).

χ^2 / v, χ^2 is an index obtained by dividing the test statistic value by the degree of freedom (v). It is known as parsimony and stand-alone fit index. The development of Tucker-Lewis Index is also based on this ratio. The value of this ratio gives information on the fit between data and model. It is said that with smaller index value of χ^2 / v ratio, the consistency will be better. Schermelleh-Engel and Moosbrugger (2003) stated that this ratio indicates good fit when it produces 2 or a smaller value while it indicates an acceptable value when it produces a value of 3. Ding et al. (1995) stated that this ratio should be close to 1 or have a smaller value.

Standardized Root Mean Square Residual (SRMR) Index

The Standardized Root Mean Square Residual (SRMR) is an index of the average of standardized residuals between the observed and the hypothesized covariance matrices (Chen, 2007). This absolute fit index can be indicated as follows:

$$SRMR = \sqrt{\frac{\sum_{i=1}^{p}\sum_{j=1}^{i}\left[\left(s_{ij} - \hat{\sigma}_{ij}\right) / \left(s_{ii}s_{jj}\right)\right]^2}{p\left(p+1\right)/2}} \qquad (6)$$

where s_{ij} indicates a component of **S** sample covariance matrix and $\hat{\sigma}_{ij}$ shows a component of $\boldsymbol{\Sigma}\left(\hat{\boldsymbol{\theta}}\right)$ hypothesized model whereas p is the number of observed variables. SRMR does not give any information about the direction of discrepancies between $\mathbf{S}$ and $\boldsymbol{\Sigma}\left(\hat{\boldsymbol{\theta}}\right)$ (Kline, 2011; Schermelleh-Engel and Moosbrugger, 2003).

Although SRMR indicates the acceptable fit when it produces a value smaller than 0.10, it can be interpreted as the indicator of good fit when it produces a value lower than 0.05 (Kline, 2011; Hu and Bentler, 1999; Schermelleh-Engel and Moosbrugger, 2003; Lacobucci, 2010). One of the reasons of preferring SRMR index in studies is its relative independence from sample size (Chen, 2007).

Root Mean Square Error of Approximation (RMSEA) Index The RMSEA is an index of the difference between the observed covariance matrix per degree of freedom and the hypothesized covariance matrix which denotes the model (Chen, 2007). This absolute fit index is estimated as follows:

$$RMSEA = \sqrt{max\left\{\left(\frac{\mathrm{F}\left(\mathbf{S},\boldsymbol{\Sigma}\left(\hat{\boldsymbol{\theta}}\right)\right)}{v} - \frac{1}{n-1}\right),\ 0\right\}} \qquad (7)$$

Here $\mathrm{F}\left(\mathbf{S},\boldsymbol{\Sigma}\left(\hat{\boldsymbol{\theta}}\right)\right)$ indicates the fit function is minimized whereas *max* points to the maximum value of the values given in brackets. While l is the number of known parameters and t is the number of independent parameters, $v = l - t$ indicates the value of the degrees of freedom and n indicates the sample size (Schermelleh-Engel and Moosbrugger, 2003).

Observe in equation (7) that RMSEA produces a better quality of estimation when the sample size is large compared to smaller sample sizes. When the sample size is large, the term $[1/(n - 1)]$ gets closer to zero asymptotically (Rigdon, 1996).

The RMSEA also takes the model complexity into account as it reflects the degree of freedom as well. RMSEA value smaller than 0.05, it can be said to indicate a convergence fit to the analyzed data of the model while it indicates a fit close to good when it produces a value between 0.05 and 0.08. A RMSEA value falling between the range of 0.08–0.10 is stated to indicate a fit which is neither good nor bad. Hu and Bentler (1999) remarked that RMSEA index smaller than 0.06 would be a criterion that will suffice. A few researchers stated that RMSEA is among the fit indexes which are affected the least by sample size (Marsh et al., 2004; Schermelleh-Engel and Moosbrugger, 2003).

Tucker-Lewis Index (TLI) The Tucker-Lewis Index (TLI) is an incremental fit index. Non-Normed Fit Index (NNFI) which is also known as TLI was developed against the disadvantage of Normed Fit Index regarding being affected by sample size. TLI is calculated as given below (Schermelleh-Engel and Moosbrugger, 2003; Ding et al., 1995; Gerbing & Anderson, 1992).

$$TLI=\frac{\left(\chi_i^2/\nu_i\right)-\left(\chi_t^2/\nu_t\right)}{\left(\chi_i^2/\nu_i\right)-1}=\frac{\left(\mathrm{F}_i/\nu_i\right)-\left(\mathrm{F}_t/\nu_t\right)}{\left(\mathrm{F}_i/\nu_i\right)-\left(1/\left(n-1\right)\right)} \qquad (8)$$

Here χ_i^2 belongs to the independence model whereas χ_t^2 belongs to the target model. v_i and v_t are the number of degrees of freedom for the independence and target models respectively, in relation to the chi-square test statistics. F is the value of appropriate minimum fit function, and *n* indicates sample size.

The bigger TLI value indicated better fit for the model. Although values larger than 0.95 are interpreted as acceptable fit, 0.97 is accepted as the cut-off value in a great deal of researches. Furthermore TLI is not required to be between 0 and 1 as it is non-normed. The key advantage of this fit index is the fact that it is not affected significantly from sample size (Schermelleh-Engel and Moosbrugger, 2003; Ding et al., 1995; Gerbing & Anderson, 1992).

Comparative Fit Index (CFI) The Comparative Fit Index (CFI) is an incremental fit indices. CFI is a corrected version of relative non-centrality index. The extent to which the tested model is superior to the alternative model established with manifest covariance matrix is evaluated (Chen, 2007) and the equation can be given as in (9).

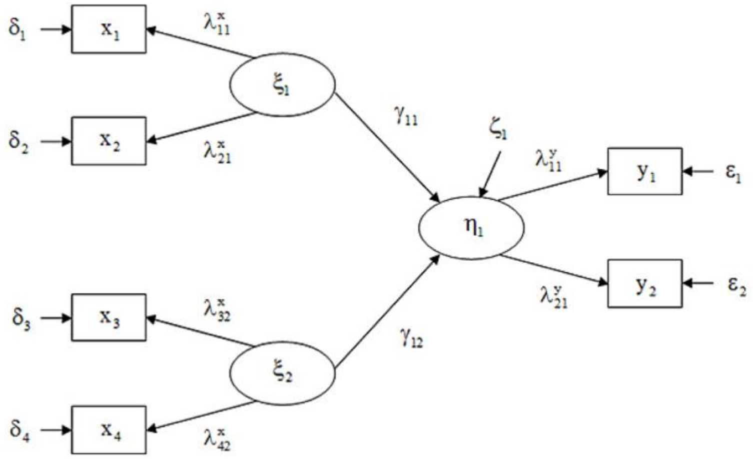
$$CFI=1-\frac{max\left[\left(\chi_t^2-\nu_t\right),\ 0\right]}{max\left[\left(\chi_t^2-\nu_t\right),\left(\chi_i^2-\nu_i\right),\ 0\right]} \qquad (9)$$

Here *max* indicates the maximum value of the values given in brackets. χ_i^2 and χ_t^2 are test statistics of the independence model and the target model respectively. v_i and v_t are the degrees of freedom of the independence model and the target model in relation to chi-square test statistics respectively (Schermelleh-Engel and Moosbrugger, 2003; Ding et al., 1995; Gerbing & Anderson, 1992).

The CFI produces values between 0−1 and high values are the indicators of good fit. When CFI value is 0.97, it means that the fit in question is better compared to the independence model. An acceptable fit is in question provided that CFI value is larger than 0.95 (Schermelleh-Engel and Moosbrugger, 2003). This index is relatively independent from sample size and yields better performance when studies with small sample size (Chen, 2007; Hu and Bentler, 1998).

Hypothesized models

Two structural equation models (SEMs) with different structures of mean and covariance, and constructed in accordance with exogenous latent variable number were established. Model 1 is the model with two exogenous and one endogenous latent variables with each of the exogenous variable having two indicators (Figure 1). Model 2 is the other model established through the addition of one exogenous variable with two indicators to the structure given in Model 1 (Figure 2).

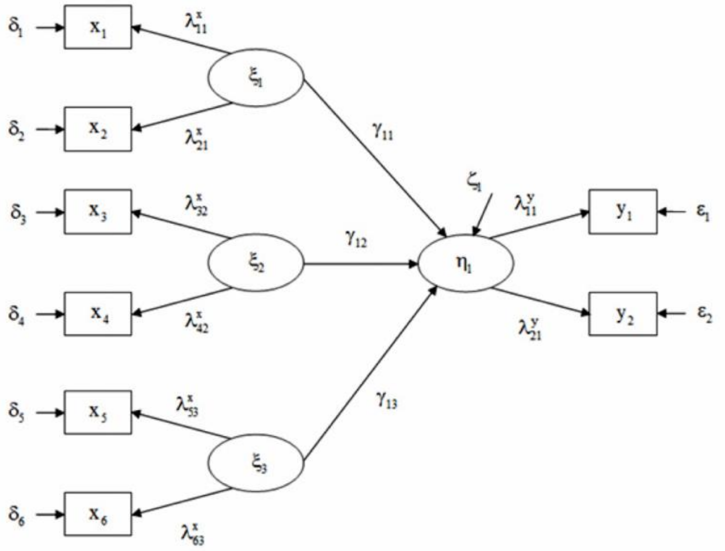


$$\eta_1 = \gamma_{11}\,\xi_1 + \gamma_{12}\,\xi_2 + \zeta_1$$

$$y_1 = \lambda^y_{11}\,\eta_1 + \varepsilon_1 \qquad x_1 = \lambda^x_{11}\,\xi_1 + \delta_1 \qquad x_3 = \lambda^x_{32}\,\xi_2 + \delta_3$$

$$y_2 = \lambda^y_{21}\,\eta_1 + \varepsilon_2 \qquad x_2 = \lambda^x_{21}\,\xi_1 + \delta_2 \qquad x_4 = \lambda^x_{42}\,\xi_2 + \delta_4$$

Figure 1. Structural equation model with three latent variables, with observed variables each (Model 1)



$$\eta_1 = \gamma_{11}\,\xi_1 + \gamma_{12}\,\xi_2 + \gamma_{13}\,\xi_3 + \zeta_1$$

$$y_1 = \lambda^y_{11}\,\eta_1 + \varepsilon_1 \quad x_1 = \lambda^x_{11}\,\xi_1 + \delta_1 \quad x_3 = \lambda^x_{32}\,\xi_2 + \delta_3 \quad x_5 = \lambda^x_{53}\,\xi_3 + \delta_5$$

$$y_2 = \lambda^y_{21}\,\eta_1 + \varepsilon_2 \quad x_2 = \lambda^x_{21}\,\xi_1 + \delta_2 \quad x_4 = \lambda^x_{42}\,\xi_2 + \delta_4 \quad x_6 = \lambda^x_{63}\,\xi_3 + \delta_6$$

Figure 2. Structural equation model with four latent variables, with observed variables each (Model 2)

Sample generation

The mean vectors and covariance matrices which were used for generating data are given in Table 1 for identification model. Multivariate normal distribution data were generated by taking Model 1 and Model 2 into consideration for the sample sizes determined as 100, 500 and 1000 units. MLE, GLS, ADF and SB_χ^2 techniques were applied to the derived data. SEMs which are significant in accordance to the test statistics were included in the study ($p > 0.05$). χ^2 / v ratio, SRMR, RMSEA, TLI, and CFI model fit indices which were obtained from the significant SEMs were recorded. A total of 1200 significant SEMs were examined in the research. The simulation and all of the remaining statistical analyses were performed in R software through the utilization of *MSBVAR*, *mvShapiroTest*, *QRMlib* and *lavaan* packages.

Table 1. Covariance matrices and Mean vectors of Model 1 and Model 2

Model 1	y_1	y_2	x_1	x_2	x_3	x_4
y_1	1.50					
y_2	1.18	1.50				
x_1	0.95	0.90	1.50			
x_2	0.95	0.90	1.20	1.50		
x_3	0.95	0.90	0.50	0.50	1.50	
x_4	0.95	0.90	0.50	0.50	1.30	1.50
μ_1 =	(100	100	100	100	100	100)

Model 2	y_1	y_2	x_1	x_2	x_3	x_4	x_5	x_6
y_1	1.50							
y_2	1.18	1.50						
x_1	0.95	0.90	1.50					
x_2	0.95	0.90	1.20	1.50				
x_3	0.95	0.90	0.50	0.50	1.50			
x_4	0.95	0.90	0.50	0.50	1.30	1.50		
x_5	0.95	0.90	0.50	0.50	0.50	0.50	1.50	
x_6	0.95	0.90	0.50	0.50	0.50	0.50	1.25	1.50
μ_2 =	(100	100	100	100	100	100	100	100)

μ_1: Mean vector of Model 1; μ_2: Mean vector of Model 2

Study design

The study was designed as 4×3 so as to examine the effects of 4 different estimation techniques (MLE, GLS, ADF and SB_ χ^2) and 3 different sample sizes (100, 500 and 1000) under multivariate normal distribution condition by taking both structural models into consideration.

A rank transform was applied to each index, and then Factorial Analysis of Variance (Factorial ANOVA) was conducted so as to find out the effects of estimation technique and sample size factors on χ^2 / v ratio, SRMR, RMSEA, TLI and CFI model fit indices based on the models established. Tukey's Honestly Significant Difference (Tukey's HSD) was used for the pairwise comparisons of the factors in which statistically significant differences were found.

Results

Out of the simulation results obtained by applying SEM estimation techniques to Model 1 and Model 2 under multivariate normality condition, 3.17%, 8.60% and 7.6% comprise of the convergence error of model, improper solutions, and the simulations excluded from the study (non-significant SEMs) respectively. As well as the significance of the models included in the study, it was found that fit indices also have good fit and acceptable fit.

The comparative summarized table of model fit indices based on estimation techniques (p-values) is given in Table 2. While no significant differentiation was identified in respect to χ^2 / v ratio obtained from Model 1 based on the estimation techniques and RMSEA indices, differentiations were identified in SRMR, TLI and CFI. Although the CFI was the least affected one from the estimation techniques among the model fit indices which were identified to have differentiations, SRMR was the most affected one. No significant differentiation between the normal theory techniques MLE and GLS or between SB_ χ^2 and each normal theory was found in respect to CFI. However, CFI obtained with ADF technique was identified to be different from those achieved by the other techniques. In terms of TLI, no significant differentiation was determined between MLE and SB_ χ^2 techniques and, as for SRMR index, between MLE and GLS techniques (Table 2).

When the entirety of the model fit indices were examined based on the estimation techniques in the structure given in Model 2, it was found that χ^2 / v ratio index was different compared to GLS and ADF techniques, yet these produced similar values in all of the remaining techniques. As for the RMSEA and CFI indices, these were identified to show no difference compared to MLE, GLS and SB_ χ^2 techniques, yet all of the values obtained with ADF were different from those obtained with the other techniques. In respect to TLI, only MLE and SB_ χ^2 did not show any significant difference in between (Table 2).

Table 2. The comparative summarized table of model fit indices based on estimation techniques (p-values for Tukey’s HSD)

	Model 1					***Model 2***				
	Fit Indices					Fit Indices				
Technique	χ^2 / v¤£	***SRMR***¤	***RMSEA***¤£	***TLI***¤	***CFI***¤	χ^2 / v¤	***SRMR***¤	***RMSEA***¤	***TLI***¤	***CFI***¤
MLE-GLS		0.191		<0.001	0.372	0.42	<0.001	0.471	<0.001	0.72
MLE-ADF		<0.001		<0.001	<0.001	0.068	<0.001	0.022	<0.001	<0.001
MLE-SB_ χ^2		<0.001		1.000	0.999	1.000	<0.001	0.999	1.000	0.999
GLS-ADF		<0.001		0.002	0.038	<0.001	<0.001	<0.001	<0.001	<0.001
GLS-SB_ χ^2		<0.001		<0.001	0.457	0.401	<0.001	0.551	<0.001	0.629
ADF-SB_ χ^2		<0.001		<0.001	<0.001	0.074	<0.001	0.015	<0.001	<0.001

MLE: Maximum Likelihood Estimation; GLS: Generalized Least Squares; ADF: Asymptotically Distribution Free; SB_ χ^2: Satorra-Bentler Scaled Chi-Square; χ^2 / v :(Chi-Square test statistic/degree of freedom) ratio; SRMR: Standardized Root Mean Square Residual; RMSEA: Root Mean Square Error of Approximation; TLI: Tucker – Lewis Index; CFI: Comparative Fit Index; ¤ : Ranked Value; Degree of Freedom of Model 1 (v_1)= 6; Degree of Freedom of Model 2 (v_2)= 14; £: p>0.05 value for Factorial ANOVA

Table 3. The comparative summarized table of model fit indices based on sample sizes (*p*-values for Tukey's HSD)

	Model 1					***Model 2***				
	Fit Indices					Fit Indices				
Sample Size	χ^2/v¤	**SRMR**¤	**RMSEA**¤	**TLI**¤	**CFI**¤	χ^2/v¤	**SRMR**¤	**RMSEA**¤	**TLI**¤	**CFI**¤
100-500	0.006	<0.001	0.005	<0.001	<0.001	0.005	<0.001	0.217	<0.001	0.004
100-1000	0.001	<0.001	0.049	<0.001	<0.001	0.024	<0.001	0.003	<0.001	<0.001
500-1000	0.786	<0.001	0.705	<0.001	0.862	0.863	<0.001	0.236	<0.001	0.126

(χ^2/v): (Chi-Square test statistic/degree of freedom) ratio; SRMR: Standardized Root Mean Square Residual; RMSEA: Root Mean Square Error of Approximation; TLI: Tucker – Lewis Index; CFI: Comparative Fit Index; ¤ : Ranked Value; Degree of Freedom of Model 1 (ν_1)= 6; Degree of Freedom of Model 2 (ν_2)= 14

The summarized comparative table of model fit indices based on sample size (*p*-values) is given in Table 3. The index values of SRMR and TLI obtained from Model 1 under multivariate normality condition was found to be significantly different according to sample sizes. However, while χ^2/v ratio, RMSEA and CFI obtained with a sample size of 100 units were observed to be significantly different from those obtained with the sample sizes of 500 and 1000 units, no significant differentiation was observed in none of the three indices obtained in sample sizes of 500 and 1000 units. With the increasing sample size, and in particular, when the sample size was above 500 units, it can be said that no significant change is seen in χ^2/v, RMSEA and CFI values. All model fit indices showed significant differences based on sample size. However, while no significant differentiation was identified when they were examined in respect to χ^2/v ratio, RMSEA and CFI values based on large sample size ($n > 500$), significant differentiation was determined according to small and large sample sizes (100 and 1000). Additionally, it was found that there is no difference between the values obtained with small sample sizes (100 and 500) in RMSEA.

Discussion

The empirical evaluation of the proposed models is an important aspect of theory development process. It was determined that the χ^2/v ratio index based on the structures given in Model 1 and Model 2 was not affected from MLE and SB_ χ^2 techniques, and RMSEA and CFI were not affected from MLE, GLS and SB_ χ^2. TLI was determined to be insensitive to MLE and SB_ χ^2 techniques, yet SRMR index was affected from all estimation techniques. When the compliance of our findings with the literature is evaluated on the basis of models, it is seen that they are generally in compliance with the results of the studies conducted by Sugawara and MacCallum (1993), Hu and Bentler (1998, 1999), Fan et al. (1999), and Lei and Lomax (2005) yet entirely incompliant with the results produced by Ding et al. (1995).

When both model structures are taken into consideration in multivariate normal distribution condition and in the event of studying with large sample size; χ^2 / v rate, RMSEA and CFI were determined to be independent from sample size while SRMR and TLI were dependent. When the compliance of our findings with the literature is examined on the basis of models, it was generally in parallel to the study results produced by Lacobucci (2010), Herzog et al. (2009), Jackson, (2001, 2007), Beauducel and Wittmann (2005), Curran et al. (2003), Kenny and McCoach (2003), Curran et al. (2002), Hu and Bentler (1999), Fan et al. (1999), Ding et al. (1995), Marsh and Balla (1994). Yet our findings except *RMSEA* were quite different from the study results of Fan and Sivo (2007). Furthermore, Rigdon (1996) emphasized the requirement to prefer RMSEA with large sample sizes and researches conducted to develop theory in his study in which RMSEA and CFI were compared.

The difference of model structure was accepted as an exogenous factor rather than a primary effect. Therefore, it can be stated that particular model fit indices obtained with only ADF technique are negatively affected from the increase of the number of latent variables when the result is evaluated in respect to the factors examined in this study.

In conclusion, it would be appropriate to prefer χ^2 / v ratio, RMSEA and CFI in the event of studying with large samples and utilization of MLE, GLS and SB_χ^2 techniques under multivariate normal distribution condition. Furthermore, we do not recommend using SRMR in model fit research as it is the most affected index from estimation technique and sample size.

References

Beauducel, A., & Wittmann, W. W. (2005). Simulation study on fit indices in confirmatory factor analysis based on data with slightly distorted simple structure. *Structural Equation Modeling: A Multidisciplinary Journal*, *12*(1), 41-75. doi:10.1207/s15328007sem1201_3

Bentler, P. M. (1994). A testing method for covariance structure analysis. In T. W. Anderson, K. T. Fang & I. Olkin (Eds.), Multivariate Analysis and Its

Applications. *Institute of Mathematical Statistics Lecture Notes-Monograph series*, *24*, 123-136. doi:10.1214/lnms/1215463790

Bentler, P. M., & Dudgeon, P. (1996). Covariance structure analysis: statistical practice, theory and directions. *Annual Review of Psychology*, *47*(1), 563-592. doi:10.1146/annurev.psych.47.1.563

Boomsma, A., & Hoogland, J. J. (2001). The robustness of LISREL modeling revisited. In R. Cudeck, S. Du Toit & D. Sörbom (Eds.), *Structural equation models: Present and future* (pp. 139-168). Chicago: Scientific Software International Inc.

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*(3), 464-504. doi:10.1080/10705510701301834

Chou, C. P., & Bentler, P. M. (1995). Estimates and tests in structural equation modeling. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 37-55). Thousand Oaks: Sage Publications.

Curran, P. J., Bollen, K. A., Paxton, P., Kirby, J., & Chen, F. (2002). The noncentral chi-square distribution in misspecified structural equation models: Finite sample results from a Monte Carlo simulation. *Multivariate Behavioral Research*, *37*(1), 1-36. doi:10.1207/S15327906MBR3701_01

Curran, P. J., Bollen, K. A., Chen, F., Paxton, P., & Kirby, J. (2003). Finite sampling properties of the point estimates and confidence intervals of the RMSEA. *Sociological Methods & Research*, *32*(2), 208-252. doi:10.1177/0049124103256130

Ding, L., Velicer, W. F., & Harlow, L. L. (1995). Effects of estimation methods, number of indicators per factor, and improper solutions on structural equation modeling fit indices. *Structural Equation Modeling: A Multidisciplinary Journal*, *2*(2), 119-143. doi:10.1080/10705519509540000

Fan, X., Thompson, B., & Wang, L. (1999). Effects of sample size, estimation methods, and model specification on structural equation modeling fit indexes. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 56-83. doi:10.1080/10705519509540000

Fan, X., & Sivo, S. A. (2007). Sensitivity of fit indices to model misspecification and model types. *Multivariate Behavioral Research*, *42*(3), 509-529. doi:10.1080/00273170701382864

Finney, S. J., & Distefano, C. (2006). Nonnormal and categorical data in structural equation modeling. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modelling: second course* (pp. 269-314). Greenwich: Information Age Publishing.

Gerbing, D. W., & Anderson, J. C. (1992). Monte Carlo evaluations of goodness of fit indices for structural equation models. *Sociological Methods and Research*, *21*(2), 132-160. doi:10.1177/0049124192021002002

Herzog, W., & Boomsma, A. (2009). Small-sample robust estimators of non-centrality-based and incremental model fit. *Structural Equation Modeling: A Multidisciplinary Journal*, *16*(1), 1-27. doi:10.1080/10705510802561279

Hu, L. T., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to under parameterized model misspecification. *Psychological Methods*, *3*(4), 424-453. doi:10.1037/1082-989X.3.4.424

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1-55. doi:10.1080/10705519909540118

Jackson, D. L. (2001). Sample size and number of parameter estimates in maximum likelihood confirmatory factor analysis: A Monte Carlo investigation. *Structural Equation Modeling: A Multidisciplinary Journal*, *8*(2), 205-223. doi:10.1207/S15328007SEM0802_3

Jackson, D. L. (2007). The effect of the number of observations per parameter in misspecified confirmatory factor analytic models. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*(1), 48-76. doi:10.1080/10705510709336736

Kenny, D. A., & McCoach, D. B. (2003). Effect of the number of variables on measures of fit in structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *10*(3), 333-351. doi:10.1207/S15328007SEM1003_1

Kline, R. B. (2011). *Principles and practice of structural equation modeling*. 3rd edition. New York: The Guilford Press.

Lacobucci, D. (2010). Structural equations modeling: Fit indices, sample size, and advanced topics. *Journal of Consumer Psychology*, *20*(1), 90-98. doi:10.1016/j.jcps.2009.09.003

Lei, M., & Lomax, R. G. (2005). The effect of varying degrees of nonnormality in structural equation modelling. *Structural Equation Modeling: A Multidisciplinary Journal*, *12*(1), 1-27. doi:10.1207/s15328007sem1201_1

Lee, S. Y. (2007). *Structural equation modeling: A bayesian approach*. New York, NJ: John Wiley & Sons.

Marsh, H. W., & Balla, J. R. (1994). Goodness of fit indices in confirmatory factor analysis: The effects of sample size and model parsimony. *Quality & Quantity*, *28*(2), 185-217. doi:10.1007/BF01102761

Marsh, H. W., Hau, K. T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Structural Equation Modeling: A Multidisciplinary Journal*, *11*(3), 320-341. doi:10.1207/s15328007sem1103_2

Rigdon, E. E. (1996). CFI versus RMSEA: A comparison of two fit indexes for structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *3*(4), 369-379. doi:10.1080/10705519609540052

Schermelleh-Engel, K., & Moosbrugger, H. (2003). Evaluating the fit of structural equation models: tests of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research Online*, *8*(2), 23-74.

Sivo, S. A., Fan, X., Witta, E. L., & John, T. (2006). The search for "optimal" cutoff properties: Fit index criteria in structural equation modeling. *The Journal of Experimental Education*, *74*(3), 267-288. doi:10.3200/JEXE.74.3.267-288

Sugawara, H. M., & MacCallum, R. C. (1993). Effect of estimation method on incremental fit indexes for covariance structure models. *Applied Psychological Measurement*, *17*(4), 365-377. doi:10.1177/014662169301700405

Wang, L., Fan, X., & Willson, V. L. (1996). Effects of nonnormal data on parameter estimates and fit indices for a model with latent and manifest variables: An empirical study. *Structural Equation Modeling: A Multidisciplinary Journal*, *3*(3), 228-247. doi:10.1080/10705519609540042

Journal of Modern Applied Statistical Methods
May 2015, Vol. 14, No. 1, 168-179.

ISSN 1538 − 9472

Applying Penalized Binary Logistic Regression with Correlation Based Elastic Net for Variables Selection

Zakariya Yahya Algamal
Universiti Teknologi Malaysia
Skudai, Johor, Malaysia

Muhammad Hisyam Lee
Universiti Teknologi Malaysia
Skudai,Johor, Malaysia

Reduction of the high dimensional classification using penalized logistic regression is one of the challenges in applying binary logistic regression. The applied penalized method, correlation based elastic penalty (CBEP), was used to overcome the limitation of LASSO and elastic net in variable selection when there are perfect correlation among explanatory variables. The performance of the CBEP was demonstrated through its application in analyzing two well-known high dimensional binary classification data sets. The CBEP provided superior classification performance and variable selection compared with other existing penalized methods. It is a reliable penalized method in binary logistic regression.

Keywords: high dimensional, penalization, binary classification, correlation based penalty, LASSO, elastic net, ridge

Introduction

With advances in technology, data are becoming larger, resulting in high dimensional problems. One of these problems facing researchers in application is the number of variables *p*, exceeding the number of sample size *n*. In classical statistical theory, it is assumed that the number of observations is much larger than the number of explanatory variables, so that large-sample asymptotic theory can be used to derive procedures and analyze their statistical accuracy and interpretability. For high-dimensional data, this assumption is violated.

To overcome this challenge, various penalized methods have been proposed beginning with ridge penalty (Hoerl & Kennard, 1970). It estimates the regression

Zakariya Algamal is a Ph.D student in the Department of Mathematical Sciences at Universiti Teknologi Malaysia. Email him at zak.sm_stat@yahoo.com. Muhammad Lee is Professor of Statistics in the Department of Mathematical Sciences at Universiti Teknologi Malaysia. Email him at: mhl@utm.my.

coefficients through ℓ_2-norm penalty. It is well known that ridge regression shrinks the coefficients of correlated predictor variables toward each other, allowing them to borrow strength from each other (Friedman, Hastie, & Tibshirani, 2010). The least absolute shrinkage and selection operator (LASSO) was proposed by Tibshirani (1996) to estimate the regression coefficients through ℓ_1 -norm penalty. While demonstrating promising performance for many problems, the LASSO estimator does have some shortcomings (Zou & Hastie, 2005). Firstly, the LASSO tends to have problems when explanatory variables are highly correlated. Secondly, it cannot select more explanatory variables than the sample size.

Zou and Hastie (2005) proposed the elastic net penalty which is based on a combined penalty of LASSO and ridge regression penalties in order to overcome the drawbacks of using the LASSO and ridge regression on their own. Tutz and Ulbricht (2009) proposed correlation based penalty to encourage a grouping effect by using correlation between explanatory variables as weights through making a group of highly correlated explanatory variables to either be selected together or to left out altogether. Although this penalty does well when there is high correlation among explanatory variables, it doesn’t do as well when the correlation is perfect (Tan, 2012). This study applies a new penalized penalty proposed by Tan (2012), namely Correlation Based Elastic Penalty (CBEP), in penalized logistic regression, and compares it with elastic net, LASSO, and ridge penalties. We apply these four methods and test the classification performance on two well-known data sets.

This paper is organized as follows. Methodology covers the penalized logistic regression methods. Data description is explained in the following section. The second to last section is devoted to results and discussions. Finally we end this paper with a conclusion. All implementations are done using *elasticnet package* in R.

Methodology

Penalized Logistic Regression Methods

Logistic regression is considered one of the most important methods in several fields such as medicine, social science, and financial banking. It is widely used in binary classification problems, where the response variable has two values coded as 0 and 1. One of the problems that researchers face in applying logistic regression is the high dimensionality of the data, where the number of variables *p*, exceeds the number of sample size n, in fields such as genomics, biomedical imaging, and DNA micro-arrays. Selecting an optimal subset of explanatory variables in order to improve the classification accuracy and to make the model's interpretation easier is the main objective of the variable selection in high dimensional data (Pourahmadi, 2013). A procedure called penalization, which is always used in variable selection in high dimensional data, attaches a penalty term $P_\lambda(\beta)$ to the log-likelihood function to get a better estimate of the prediction error by avoiding overfitting. Recently, there is growing interest in applying the penalization method in logistic regression models. In order to extract the most important explanatory variables in classification problems, a series of penalized logistic regression methods have been proposed. For example, Shevade and Keerthi (2003) proposed the sparse logistic regression based on the LASSO penalty. Similar to sparse logistic regression with the LASSO penalty, Cawley and Talbot (2006) investigated sparse logistic regression with Bayesian penalty. Liang et al. (2013) did another investigation in the sparse logistic regression model using a $\ell_{1/2}$ penalty. There are varieties of different forms of the penalty term, depending on the application requirements.

In a high dimensional classification using logistic regression, our goal is to classify the response variable y, which is coded as 0 and 1, from high dimensional explanatory variables $x \in \Re^p$. In general, in logistic regression, the response variable y is a Bernoulli random variable, and the conditional probability that y is equal to 1 given x, which is denoted as $\pi(x)$, is

$$p\left(y_i = 1 \middle| x_{ij}\right) = \pi\left(x_j\right) = \frac{e^{x'_j\beta}}{1+e^{x'_j\beta}}, j = 1, 2, \ldots, p \tag{1}$$

$$f\left(y_i\right) = \pi_i^{y_i}\left(1-\pi_i\right)^{1-y_i}, i = 1, 2, \ldots, n \tag{2}$$

The likelihood will be

$$L\left(\beta, y_i\right) = \prod_{i=1}^{n} f\left(y_i\right) = \prod_{i=1}^{n} \pi_i^{y_i}\left(1-\pi_i\right)^{1-y_i}. \tag{3}$$

Then, the log-likelihood becomes

$$\ell(\beta, y_i) = \sum_{i=1}^{n} \left\{ y_i \log \pi(x_i) + (1 - y_i) \log(1 - \pi(x_i)) \right\} \tag{4}$$

The penalized logistic regression (PLR) is defined as

$$PLR = \sum_{i=1}^{n} \left\{ y_i \log(\pi(x_i) + (1 - y_i) \log(1 - \pi(x_i)) \right\} + \lambda P(\beta) \tag{5}$$

where λ is defined as a tuning parameter ($\lambda \geq 0$). It controls the strength of shrinkage in the explanatory variables: when λ takes larger value, more weight will be given to the penalty term. Because the value of λ depends on the data, it can be computed using cross-validation method (James, Witten, Hastie, & Tibshirani, 2013). Before solving the PLR, it is worth centering to the y and standardizating to x_j, so that $\sum_{i=1}^{n} y_i = 0$, $\frac{1}{n}\sum_{i=1}^{n} X_{ij} = 0$, and $\sum_{i=1}^{n} X_{ij}^2 = 1$ for $j = 1,2,\ldots, p$, in order to make the intercept (β_0) equal zero. Many different forms of the penalty term have been introduced in the literature, including ridge penalty, LASSO, elastic net, and correlation based penalty.

Ridge Regression

One of the most popular penalties is ridge regression, which was introduced by Hoerl and Kennard (1970) as an alternative solution to ordinary least square when there is multicollinearity between explanatory variables. The ridge regression solves the logistic log-likelihood in Eq. (4) using ℓ_2-norm penalized logistic log-likelihood (i.e., $\lambda P(\beta) = \lambda \sum_{j=1}^{p} \beta_j^{2}$)

$$PLR = \sum_{i=1}^{n} \left\{ y_i \log(\pi(x_i) + (1 - y_i) \log(1 - \pi(x_i)) \right\} + \lambda \sum_{j=1}^{p} \beta_j^{2} \tag{6}$$

In ridge regression, the tuning parameter λ controls the amount of shrinkage, but never sets explanatory variable coefficients to be exactly equal zero. So, in high dimensional data when $p > n$, the ridge regression will not provide the sparsity model. Although ridge regression doesn't have the sparsity property, it is preferred in high dimensional data because we expect high correlation between explanatory variables. The maximum likelihood solution of Eq. (6) is

$$\hat{\beta}_{Ridge} = \arg\min_{\beta}\left\{\ell\left(\beta, y_i\right) + \lambda\sum_{j=1}^{p}\beta_j^{\,2}\right\} \tag{7}$$

Least Absolute Shrinkage and Selection Operator

Tibshirani (1996) proposed the least absolute shrinkage and selection operator (LASSO), as a penalty for variables selection by setting some variable coefficients' to zero. It does both continuous shrinkage and automatic variable selection simultaneously. As with the ridge regression the LASSO estimates are obtained by maximizing the log-likelihood. Instead of using ℓ_2-norm, the LASSO uses the ℓ_1-norm on the logistic regression coefficients (i.e., $\lambda P\left(\beta\right) = \lambda\sum_{j=1}^{p}\left|\beta_j\right|$). The penalized logistic regression using LASSO is

$$PLR = \sum_{i=1}^{n}\left\{y_i \log\pi\left(x_i\right) + \left(1-y_i\right)\log\left(1-\pi\left(x_i\right)\right\} + \lambda\sum_{j=1}^{p}\left|\beta_j\right| \tag{8}$$

Depending on the property of the LASSO penalty, some coefficients will be exactly equal zero. Hence, LASSO does the variable selection. Consequently, LASSO has sparsity property. Although LASSO is widely used in many applications, it has some drawbacks. One of these drawbacks is that it is not robust to high correlation among explanatory variables and will randomly choose one of these variables and ignore the rest. Another drawback of LASSO is that in high dimensional data when *p*>*n*, it chooses at most *n* explanatory variables, whereas there may be more explanatory variable coefficients than *n* with non-zero values in the final model (Zhou, 2013). Solving Eq. (8) will depend on optimization methods. So,

$$\hat{\beta}_{LASSO} = \arg\min_{\beta}\left\{\ell\left(\beta, y_i\right) + \lambda\sum_{j=1}^{p}\left|\beta_j\right|\right\} \tag{9}$$

Elastic Net

Elastic net is a penalized method for variable selection, which is introduced by Zou and Hastie (2005) to deal with the drawbacks of LASSO. Elastic net tries to merge the ℓ_2-norm and the ℓ_1-norm penalties, by using ridge regression penalty to deal with high correlation problem while taking advantage of LASSO penalty in variable selection property. The elastic net logistic regression is defined by

$$PLR = \sum_{i=1}^{n}\left\{y_i \log \pi(x_i) + (1-y_i)\log\left(1-\pi(x_i)\right\} + \lambda_1 \sum_{j=1}^{p}\left|\beta_j\right| + \lambda_2 \sum_{j=1}^{p}\beta_j^{2}. \quad (10)$$

As we observe from Eq. (10), elastic net is dependent on non-negative two tuning parameters λ_1, λ_2 and leads to penalized logistic regression solution

$$\hat{\beta}_{Elastic} = \arg\min_{\beta}\left\{\ell(\beta, y_i) + \lambda_1 \sum_{j=1}^{p}\left|\beta_j\right| + \lambda_2 \sum_{j=1}^{p}\beta_j^{2}\right\}. \quad (11)$$

According to lemma 1 in Zou and Hastie (2005), to find the estimates of $\beta_{Elastic}$ in Eq. (11), the given data set $(\mathbf{y}, \mathbf{X})$ is extended to an augmented data $(\mathbf{y}^*, \mathbf{X}^*)$ and is defined by

$$X^{*}_{(n+p,p)} = (1+\lambda_2)^{-\frac{1}{2}}\begin{pmatrix} X \\ \sqrt{\lambda_2}\,\mathrm{I} \end{pmatrix}, y^{*}_{(n+p,1)} = \begin{pmatrix} y \\ 0 \end{pmatrix} \quad (12)$$

As a result of this augmentation the elastic net can be written as a LASSO penalty and solved. Hence, the elastic net can select all p explanatory variables in the high dimensional when $p > n$ and not only n explanatory variables as in the LASSO, because $\mathbf{X}^*$ has rank p.

Correlation Based Penalty

Similar to elastic net, this penalty encourages a grouping effect by using correlation between explanatory variables as weights. This penalty is proposed by Tutz and Ulbricht (2009), their contribution is to make a group of highly correlated explanatory variables to be either selected together or to left out altogether. Tan (2012) reported that although the elastic net penalty does well when there is high correlation among explanatory variables, it doesn't do well when there is perfect correlation. An extension of the correlation-based penalty to deal with this drawback was made in elastic net penalty. The penalty is defined as

$$\lambda P(\beta)=\lambda_1\sum_{j=1}^{p}\left|\beta_j\right|+\lambda_2\left\{\sum_{j=1}^{p-1}\left(\beta_j-r_{j,j+1}\beta_{j+1}\right)^2+\beta_p^2\right\} \tag{13}$$

where $r_{j,j+1}$ is the correlation between x_j and x_{j+1}. The penalized logistic regression using this penalty and the estimate of β_{CBEP} be, respectively

$$\begin{aligned} PLR=&\sum_{i=1}^{n}\left\{y_i\log\pi\left(x_i\right)+\left(1-y_i\right)\log\left(1-\pi\left(x_i\right)\right)\right\}\\ &+\lambda_1\sum_{j=1}^{p}\left|\beta_j\right|+\lambda_2\left\{\sum_{j=1}^{p-1}\left(\beta_j-r_{j,j+1}\beta_{j+1}\right)^2+\beta_p^2\right\}\end{aligned} \tag{14}$$

$$\hat{\beta}_{CBEP}=\arg\min_{\beta}\left\{\ell\left(\beta,y_i\right)+\lambda_1\sum_{j=1}^{p}\left|\beta_j\right|+\lambda_2\left[\sum_{j=1}^{p-1}\left(\beta_j-r_{j,j+1}\beta_{j+1}\right)^2+\beta_p^2\right]\Bigg]\right\} \tag{15}$$

CBEP is reduced to LASSO like elastic net after applying augmentation to the original data set for different values of λ_2.

Data Set Description

To evaluate the four used methods, two binary classification microarray data sets are used: colon tumor data set and diffuse large B-cell lymphoma (DLBCL) data set. The colon tumor microarray data set describes the expression of 2000 genes in 40 tumor and 22 normal tissue samples, the aim being to construct a classifier capable of distinguishing between cancer and normal tissues. This set is described in Alon et al. (1999), and publicly available at http://genomics-pubs.princeton.edu/oncology/affydata/index.html. For the DLBCL data set, the gene expression values of 77 samples were measured by high-density oligonucleotide microarrays of the two most prevalent adult lymphoid malignancies which 58 samples of diffuse large B-cell lymphomas (DLBCL) and 19 samples of follicular lymphoma (FL). Each sample contains 7,129 gene expression values. More information on this data can be found in Shipp et al. (2002) and it is freely available at http://www.genome.wi.mit.edu/cgi-bin/cancer/datasets.cgi. To apply the binary classification using the four methods that we are considered, the type of the response variable for each data set is coded as a 0 and 1, where in colon data the normal equals 0 and tumor equals 1, while in

DLBCL data, FL is set to code 0 and DLBCL is set to code 1. The classification function is defined as $I\left(\hat{y} > 0.5\right)$.

Results

To examine the performance of the correlation based elastic penalty we compare it with three well-known penalization methods; elastic net, LASSO, and ridge. We use a randomly drawn test data set. Each data set at hand was split into 10%, 20%, and 30% to form the test data set, respectively. This procedure is repeated 100 times. The required tuning parameters by the ridge, LASSO, elastic net, and CBEP methods were performed by 10-fold cross-validation on the training data set. Specifically, for ridge and LASSO, the tuning parameter was $\lambda_{Ridge} = 5.460, 3.197, 5.590)$ and $\lambda_{Lasso} = (0.055, 0.091, 0.068)$ for each training data set respectively. For the tuning parameters of elastic net and CBEP, the solution is different, because these two methods require prior value of λ_2 to transform the original training data set to the new augmented training data set. A sequence of values for λ_2 is given, where $0 \le \lambda_2 \le 100$. For each value of λ_2 a 10-fold cross-validation was performed to select the remaining tuning parameters. Here the best value for λ_2 is 0.01 for colon data set and 0.025 for DLBCL data set. Therefore, the tuning parameters for elastic net are $(0.30, 0.15, 0.40)$ and $(0.50, 0.40, 0.30)$ for colon and DLBCL data sets corresponding to each percentage of test data set, and for CBEP are $(0.40, 0.30, 0.38)$ and $(0.60, 0.50, 0.35)$ for colon and DLBCL data sets corresponding to each percentage of test data set.

The deviance test error is computed as the criterion of evaluation. Figure 1 displays the corresponding boxplots of the deviance test error for the four used methods for both data sets, (a) colon tumor and (b) DLBCL. It is clear that CBEP has less variability among the three penalization methods. Also, it can be seen that LASSO and ridge are more variable than CBEP and elastic net. Table 1 summarizes the averaged deviance test error (Mean) and the standard deviation (Std. Dev.) of the estimation of the response variable. Furthermore, coefficient of variation (CV), classification accuracy, and the numbers of selected variables are listed. When the sample size of the test set increases, the mean of the deviance test error decreases for the CBEP and the other three methods in both data sets. For example, in colon data the means for CBEP are 0.108, 0.104, and 0.102 with the sample size of the test set 10%, 20%, and 30% respectively.

Concerning the deviance test error, we observed that for colon and DLBCL data the CBEP method has mean with standard deviation smaller than the results of the elastic net, LASSO, and ridge for all test set sizes. For example, in DLBCL data, when the test data size is 10%, the mean of the CBEP is 0.118 with standard deviation equal to 0.032, which is smaller than 0.124 (0.045), 0.340 (0.265), and 0.292 (0.268) for the elastic net, ridge, and LASSO methods respectively. With both data sets and test set sizes, the results of CV show that the CBEP method yields less variation than the other three methods. Furthermore, we see that the CBEP method outperforms the elastic net, LASSO, and ridge for both colon and DLBCL data sets in term of accuracy classification. It can even classify with accuracy of 100% for colon data set at percentage 10% and 20% of test set, and also for DLBCL data set at 20% and 30% percentages of test set.

In terms of the number of selected variables (model complexity), the penalized logistic regression using CBEP includes explanatory variables less than using elastic net, although in some cases CBEP includes variables same as elastic net. Moreover, LASSO selects more variables than CBEP and elastic, and of course penalized logistic regression using ridge includes the whole explanatory variables. Because of several correlation coefficients among explanatory variables above 0.5, we have seen that the CBEP and elastic net methods prevail against the LASSO.

It is obvious that the CBEP method performs better in term of averaged deviance test error by obtaining smaller values of deviance error, classification accuracy, and the number of selected variables followed by elastic net, LASSO, and ridge for various percentages of test data sets for both colon and DLBCL data sets.

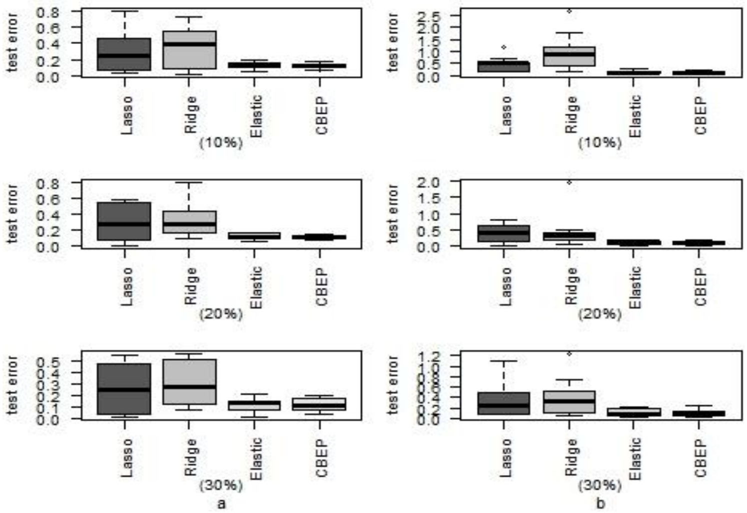


Figure 1: Percentages comparison of the deviance test error

Table 1: Deviance test error, classification accuracy, and no. of variables selected over 100 random split

		Colon				DLBCL			
		LASSO	Ridge	Elastic	CBEP	LASSO	Ridge	Elastic	CBEP
Deviance test error									
	Mean	0.483	0.958	0.134	**0.108**	0.292	0.340	0.124	**0.118**
10%	Std. Dev.	0.295	0.785	0.079	0.069	0.268	0.265	0.045	0.032
	CV	1.154	2.687	0.277	0.226	0.806	0.724	0.198	0.176
	Mean	0.422	0.447	0.119	**0.104**	0.288	0.331	0.122	**0.116**
20%	Std. Dev.	0.297	0.552	0.067	0.060	0.227	0.218	0.042	0.023
	CV	0.829	1.968	0.200	0.187	0.589	0.810	0.172	0.155
	Mean	0.354	0.395	0.107	**0.102**	0.265	0.296	0.117	**0.112**
30%	Std. Dev.	0.337	0.375	0.066	0.069	0.220	0.186	0.053	0.054
	CV	1.088	1.237	0.208	0.248	0.538	0.558	0.203	0.195
Classification Accuracy (%)									
	10%	50.00	33.34	100.00	**100.00**	75.00	62.50	75.00	**87.50**
	20%	83.34	66.67	91.69	**100.00**	86.67	80.00	100.00	**100.00**
	30%	89.47	73.68	89.47	**94.73**	86.95	82.60	95.65	**100.00**
No. of selected variables									
	10%	28	All	21	**21**	42	All	40	**40**
	20%	26	All	23	**24**	44	All	39	**38**
	30%	24	All	16	**14**	40	All	40	**38**

Finally, Figure 2 displays the path solution of the CBEP and elastic net for the colon tumor data set of 70% training data set in one run. The doted horizontal line represents the best value of elastic net ($s = 0.40$) and CBEP penalty ($s = 0.38$) that selected by cross-validation. The figure also shows, the elastic net path is very similar to CBEP path.

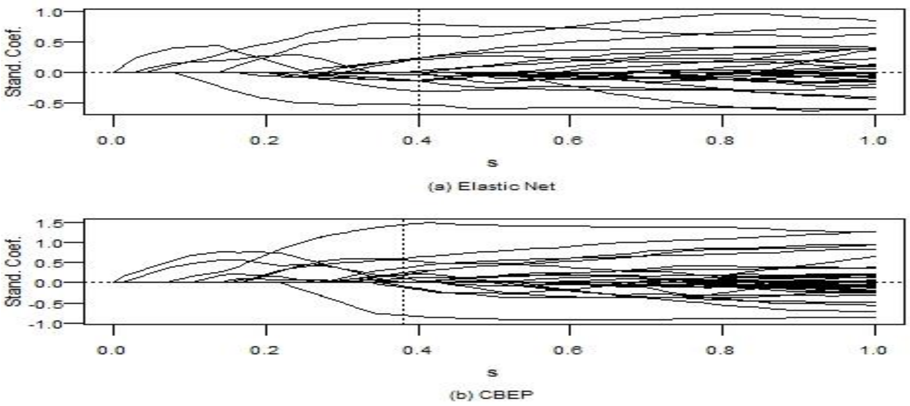


Figure 2: Solution paths for 30% test of colon tumor

Conclusion

A study of a new penalization method based on CBEP was done by application to binary logistic regression. Three penalization methods in addition to CBEP, including elastic net, LASSO, and ridge, were compared by applying two high dimensional real data sets. The results show that the CBEP outperforms the other three methods in term of deviance test error, classification accuracy, and model complexity. Also, the different percentages of the test data size do not affect the performance of CBEP. It was concluded the CBEP is more reliable in applying penalized binary logistic regression.

References

Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D., & Levine, A. J. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences, 96*(12), 6745-6750. doi:10.1073/pnas.96.12.6745

Cawley, G. C., & Talbot, N. L. C. (2006). Gene selection in cancer classification using sparse logistic regression with Bayesian regularization. *Bioinformatics, 22*(19), 2348-2355. doi:10.1093/bioinformatics/btl386

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software, 33*(1), 1-22.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics, 12*(1), 55-67. doi:10.1080/00401706.1970.10488634

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. New York: Springer.

Liang, Y., Liu, C., Luan, X. Z., Leung, K. S., Chan, T. M., Xu, Z. B., & Zhang, H. (2013). Sparse logistic regression with a L-1/2 penalty for gene selection in cancer classification. *Bmc Bioinformatics, 14*, 1-12. doi:10.1186/1471-2105-14-198

Pourahmadi, M. (2013). *High-dimensional covariance estimation: with high-dimensional data*. Hoboken, New Jersey: John Wiley & Sons.

Shevade, S. K., & Keerthi, S. S. (2003). A simple and efficient algorithm for gene selection using sparse logistic regression. *Bioinformatics, 19*(17), 2246-2253. doi:10.1093/bioinformatics/btg308

Shipp, M. A., Ross, K. N., Tamayo, P., Weng, A. P., Kutok, J. L., Aguiar, R. C., ...Golub, T. R. (2002). Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nature Medicine, 8*(1), 68-74. doi:10.1038/nm0102-68

Tan, Q. E. A. (2012). *Correlation adjusted penalization in regression analysis.* PhD. dissertation, The University of Manitoba, Canada. Retrieved from http://mspace.lib.umanitoba.ca/handle/1993/9147

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological), 58*(1), 267-288.

Tutz, G., & Ulbricht, J. (2009). Penalized regression with correlation-based penalty. *Statistics and Computing, 19*(3), 239-253. doi:10.1007/s11222-008-9088-5

Zhou, D. X. (2013). On grouping effect of elastic net. *Statistics & Probability Letters, 83*(9), 2108-2112. doi:10.1016/j.spl.2013.05.014

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67*, 301-320. doi:10.1111/j.1467-9868.2005.00503.x

Journal of Modern Applied Statistical Methods
May 2015, Vol. 14, No. 1, 180-195.

ISSN 1538 − 9472

Modeling Probability of Causal and Random Impacts

Stan Lipovetsky
GfK Custom Research North America
Minneapolis, MN

Igor Mandel
Telmar Group Inc.
New York, NY

The method of the estimation of the probability of an event occurring under the influence of the causal and random effects is considered. Epistemological differences from the traditional approaches to causality are discussed, and a new model of the statistical estimation of the parameters of each effect is proposed. The simple and effective algorithms of the model parameters estimation are presented, and numerical simulations are performed. A practical marketing example is analyzed. The results support the validity of the estimation procedure and open the perspective for the application of the method for various decision making problems, where different causes can yield the same outcome.

Keywords: causal and random effects, categorical data, causal modeling

Introduction

Modern decision making actively uses statistical methods, but there is one paradoxical aspect in it. To apply the results of statistical modeling and forecasting in practice, a decision maker, or a manager should be sure that the decision is based on a causal relationship: for instance, a positive correlation between advertising and sales could mean that it makes sense to increase spending on advertising for getting higher revenue. However, most of the statistical methods do not produce "causal models", they only agree that "correlation is not causation". For instance, Leo Breiman (2001) emphasized the indifference of the statistical learning to causal problematics (see also Hastie, Tibsharani, & Friedman 2009). So, a positive relationship between advertising and sales may simply indicate that with bigger sales, a company has a higher profit and thus is able to spend more on advertising. More questions related to statistical and causal approaches in sociosystemics and mediaphysica are considered in (Kuznetsov & Mandel, 2007, Mandel, 2011).

Email Mr Lipovetskty at: stan.lipovetsky@gfk.com. Email Mr. Mandel at: igor.mandel@gmail.com.

The past two decades have witnessed a burst of works on various causality problems and methods. Three main approaches intensively used in causality studies are: simultaneous structural equations founded by S. Wright (1921, 1960; more references within Kline, 2010); potential outcomes proposed by J. Splawa-Neyman (1990), and advanced by D. Rubin (1974, 2006), and the concept of do-operators and associated with them acyclic graphs developed by J. Pearl (2000, 2013). There are many other authors and proposals combining and modifying these ideas, although according to J. Pearl, almost all of these approaches in fact talk about the same things, using different terms and stressing different aspects of the problem. One thing is common for most of these works is that they consider a situation when many variables are interlinked, and the main aim of the causal analysis consists in disentangling of these influences and evaluating the pure impact of each cause on the effect. For instance, in the influential J. Pearl's book (2000), all descriptions begin only when graphs have complex structures, with several arrows targeting each node, but it is not clear what to do, if there is only one outcome and many potential causes.

While most applications of causal inference focus on a complex situation with multiple outcomes, the current paper revisits a seemingly simple case of a single binary outcome variable with multiple sources of causal and additional random effects. Randomness is understood here not as a "remaining part" of the unexplained variance, which is typical in statistics, but as the source of the unknown (not associated with any variables) causes, resulting in the same effect. This concept and a general model was proposed in (Mandel, 2013), where one can also find a discussion about the correct definition of causes and effects, the differences between individual and statistical causes, and other methodological issues, partly touched on here. This current paper considers the problems of the parameters estimation in such a model.

The Concept of the Causal Intrinsic Probabilities

Consider a model of the direct impact of multiple causes onto the binary outcome Y with $Y = 1$ and $Y = 0$ meaning that the effect of the interest has occurred or has not, respectively. Consider a case of K attributes $A_1, A_2,..., A_K$ (where $A_k = 1$ and $A_k = 0$ denote the presence and the absence of a k-th attribute, with $k = 1, 2,…, K$). The attributes are represented by the categorical variables which may be binary, ordinal, or nominal variables. A vector of the realized values of such attributes can be denoted as $a = (a_1, a_2, ..,a_K)$, and this may represent levels of the same and/or different categorical variables, e.g., $A_1 = 1$ means male, $A_2 = 1$ means female, $A_3 = 1$ means kids, $A_4 = 1$ means teenagers, $A_5 = 1$ means adults, etc. Let us assume that the attribute A_k creates the causal effect $Y = 1$ with probability p_k. In the simplest case $k = 2$, the probability that $Y = 1$ would follow the rule of the union of the independent events: $S = p_1 + p_2 - p_1 * p_2$. In essence, it just reflects the fact that the coincidence of two causes does not produce anything more than one effect. Respectively, the probability of not having the causal effect would be presented as $1 - S = (1 - p_1)(1 - p_2)$.

For any K, the causal effect of outcome $Y = 1$ is defined as an intrinsic (latent) probability $S_{\text{causal}}(a)$, where a is a vector of the realized set of attributes, so that the probability of not-occurring of the event is:

$$1 - S_{causal}(a) = \prod_{\{k:\, a_k = 1\}} (1 - p_k) \tag{1}$$

where $p_1, p_2, .., p_K$ are parameters which represent the causal strength associated with the presence of each attribute A_k. Note that the absence of an attribute may imply the presence of the opposite attribute (e.g., the absence of the "male" attribute A_1 contributes to the presence of the "female" attribute, A_2). In other situations it could vary: for instance, a road accident may happen due to fog (A_1), reckless driving (A_2), ice conditions (A_3), and other non-mutually exclusive causes.

There is also an irreducible latent probabilistic "random cause" that represents other factors that are not explicitly accounted for by the set of attributes. It is assumed that this random effect is: a) independent of other attributes; b) its outcome (denoted as r in the sequel) is constant across all configuration of attributes that may be present for a particular individual. These assumptions yield the expected probability at the population level as the union of the causal and random sources, $S(a) = S_{causal}(a) + S_{random} - S_{causal}(a) \cdot S_{random}$, or in the explicit form:

$$\begin{aligned} S(a) &= S_{causal}(a) + r - r \cdot S_{causal}(a) \\ &= 1 - (1 - r)(1 - S_{causal}(a)) \\ &= 1 - (1 - r) \prod_{\{k:\, a_k = 1\}} (1 - p_k) \end{aligned} \tag{2}$$

The aim of the proposed causal model is the estimation of $K + 1$ parameters, $p_1, .., p_K$, and r, on the basis of the sample of the realized outcomes $Y_i(a) = \{1, 0\}$ and the associated attribute vectors.

Concerning the motivation for the model, we can see the following arguments. Our setup acknowledges the *asymmetric* nature of causality, and the model (1)-(2) for intrinsic (causal) probability assumes that a *single* cause is sufficient for an event to happen ("fire"), whereas for an event not to occur, all potential causes should be ineffectual. It can be seen in a diagram with parallel pathways, where at least one of them would fire the event. It contrasts with a common binary logistic regression, where all the attributes contribute additively to the probability of the event occurring, or not occurring. Also, the model assumes that a random cause is *irreducible* and is presented within the sample probabilities $S(a)$. Finally, in the considered model, the main role is played by the *presence of attributes*, rather than by the changing levels of the factors in classical methods based on the concept of regression, potential outcomes, and other models.

Thus, each cause works as an independent entity and is associated not with the whole variable (like a binary “gender”), but with the separated levels (grades) of the variable (like two variables of “males” and “females”). It is different from the traditional statistical way of making models: one should look at these “grades related yields” rather than at the coefficients of general association (or regression), linking the whole “gender” to the outcome. Each level of the potentially causal variable produces an outcome with its own intrinsic probability. And if there are some causes, which cannot be associated with any measured variables, but still produce the outcome, then we relate them to the random cause. A typical example of such random causes is as follows: customers can buy a product regardless of advertising or promotions (a “baseline” which is hard to estimate). The purpose of the causal analysis is to evaluate the intrinsic probabilities, or the parameters of the outcome $Y = 1$ generated by different causes, including the random ones, with the observed data.

Causal Analysis and Parameters Estimation

The causes and the effect are associated with the usual statistical variables. Consider a data set containing variables X – the attribute causes of the outcome variable Y. With categorical causal variables, each grade of a causal X variable has its probability of generating the occurrence of the event, or the value $Y = 1$ in the outcome. A categorical variable with n grades can be presented as a set of n binary variables $x_1, x_2, \ldots, x_n$, or the so-called Gifi-system (Gifi, 1990; Lipovetsky, 2012), where each j-th of these binary variables has 1s in the positions of j-th grade, and 0s in other positions. It allows the estimation of the causal effect only for values 1 for each variable, and the random cause can also have the impact inducting the appearance of the event $Y = 1$. So, the outcome $Y = 1$ occurs as a union of the independent events coming from two different sources – those associated with the measured variables and random noise (2).

As an explicit example, consider data with three x, so in total there are eight cells of all combinations of their values, and in each cell we find the proportions S_i of the outcome variable $S(a)$, so the proportion of $Y = 1$ in the base size of each cell. The cells and corresponding proportions S_i are presented in Table 1. Of course, in a particular real data set, some cells can be empty. The variables in Table 1 are orthogonal (see in Appendix A), so they are statistically independent.

Table 1. Example of data set with three binary variables.

i	x_1	x_2	x_3	S_i
1	0	0	0	0.09141
2	1	0	0	0.73409
3	0	1	0	0.25630
4	1	1	0	0.80300
5	0	0	1	0.57608
6	1	0	1	0.86570
7	0	1	1	0.63409
8	1	1	1	0.89563

In a general case of many variables, each presented via the Gifi-system of binaries with their total number of K variables, model (2) can be presented in a generalized form:

$$S_i = 1 - (1 - r)\prod_{k=1}^{K}(1 - p_k)^{x_{ik}} \tag{3}$$

where $k = 1, 2, \ldots, K$ is a number of variable x_{ik} identifying the k-th parameter of the probability in the i-th cell ($i = 1, 2, \ldots, N$). The values of x_{ik} are 1 or 0 when the variable is presented or not, respectively, as in Table 1. So, for the i-th value S_i, with values $x_{ik} = 1$, the term $1\text{-}p_k$ enters the product in (3), and for values $x_{ik} = 0$ the term $1\text{-}\ p_k$ is absent in the i-th row of the data. The cells as the new units are denoted by the current index $i = 1, 2, \ldots, N$. The relations (3) show that if any probability p_k or r is close to 1, the total probability of event occurrence S_i is close to 1 as well. This system corresponds to the feature of the independence of the variables' levels when the value of union S_i is defined by the criterion of "at least" one variable impacting on the event appearance. It is important to note that due to the definition (3), any additional cause with the term (1-p_k) can only increase S_i, as can be expected.

Consider how to estimate parameters of the model (3) by data like those given it Table 1. Regrouping and taking logarithm of equation (3), and using notations

$$y_i = \ln\left(1-S_i\right), \quad b_0 = \ln\left(1-r\right), \quad b_k = \ln\left(1-p_k\right) \tag{4}$$

we represent (3) in the linearized form:

$$y_i = b_0 + b_1 x_{i1} + \ldots + b_K x_{iK} \tag{5}$$

So, the problem of estimation of the parameters b_k is reduced to the ordinary least squares (OLS) linear regression, with the known solution

$$b = \left(X'X\right)^{-1} X'y \tag{6}$$

where y (4) is a vector of N-th order, X is the design matrix of x_{ik} values (completed by the additional column of all 1s, which corresponds to the intercept b_0 in the model), b is the vector of all $K+1$ parameters in (5). If there are not enough observations, the matrix of the second moments $X'X$ in (6) could be close to singular, and its inversion is impossible, or it yields too inflated coefficients. In such a case, we can use a regularization imposed onto the parameters which produces the so-called ridge-regression:

$$b = \left(X'X + qI\right)^{-1} X'y \tag{7}$$

When the profile parameter of the ridge regression q is close to zero, the solution (7) reduces to OLS (6). More complicated ridge-regressions with a high quality of fit see in (Lipovetsky, 2010).

By the estimated coefficients b (6)-(7), each original parameter of probability can be obtained from the relations (4) by transformation:

$$r = 1 - \exp\left(b_0\right), \quad p_k = 1 - \exp\left(b_k\right) \tag{8}$$

The relations (8) show that the parameters b should be negative which can be achieved by their special parameterization (for instance, each b is substituted by another unknown parameter c in the relation $b = -c^2$, and a nonlinear estimation is performed for the free parameters c). But usually the solutions (6)-(7) are feasible for the meaningful values (8).

To illustrate this approach, return to Table 1, take $y_i = \ln(1 - S_i)$ as the dependent variable (4), and construct the model (6). Its coefficients are presented in the first column of Table 2. These coefficients are transformed by (8) to the probabilities r of the random impact and p_i of the causes, which are given in the second numerical column in Table 2. In the next column, Table 2 also presents the original values of cause probabilities used in this simulated data. Comparison of the estimated and the original values shows a pretty good quality of the estimation with the relative errors of several percent or less shown in the last column of Table 2. The coefficient of multiple determination in this model (6) equals 0.998, and its value adjusted by degrees of freedom equals 0.995, so the quality of the model is indeed very high.

Table 2. Regression coefficients and probability estimates.

Coefficients of regression		Estimates of cause probability		Original values used in simulation	Relative error, % to original values
b_0	-0.102722	r	0.09762	0.10	2.38
b_1	-1.240261	p_1	0.71069	0.70	1.53
b_2	-0.224870	p_2	0.20138	0.20	0.69
b_3	-0.697456	p_3	0.50215	0.50	0.43

Note that a design matrix like in Table 1 is orthogonal, so the x-variables have zero correlations. In such situation, coefficients of multiple linear regression equal the coefficients in the pair regression of y on each x separately, which makes calculations even simpler, as shown in Appendix A. If a cell of certain variables' combination is empty, the number of rows in the design table can be reduced. But even in such a case, it is possible to hold the whole design matrix substituting zero by a small proportion value, say, $S = 0.005$.

In application, the interest may be in estimating an additive share of influence of each cause in the effect. In order to achieve this, use the formula:

$$S_{ik} = S_i \frac{\ln\left(1 - p_k\right)^{x_{ik}}}{\ln\left(1 - S_i\right)} \tag{9}$$

where the total of the causes (including the random one corresponding to the index $k = 0$) in each cell equals the predicted proportion:

$$S_i = \sum_{k=0}^{K} S_{ik} \tag{10}$$

The derivation and other properties of the relations (9)-(10) are presented in Appendix B.

Methodology

Numerical simulations

To test validity of the proposed estimation procedure, a series of experiments on the generated data were performed. The varied parameters are described in Table 3. Not all combinations of these parameters (there are about 1700 scenarios) were estimated, some of them are simply impossible. For each combination of factors, several random runs (from one to forty) were performed. In each case, the assignment of value 1 to Y was done, if any of X variables was equal 1. For correlations in Table 3, both signs were used; correlation -1 means that two variables represent in fact one binary variable.

Table 3. Different parameters of simulation and estimation

	Parameters	Value 1	Value 2	Value 3	Value 4
1	Number of observations in a data set	100	500	10,000	
2	Number of causal variables	1	2	3	8
3	Correlations between certain X variables	Low (0-0.3)	Middle (0.3-0.7)	High (> 0.7)	-1
4	Random causal coefficients	0.1	0.5	0.8	Any
5	Causal coefficients for X variables	Equal	Different		

After the modeling, the estimated in (6)-(8) parameters of the causal yields were compared with the original values used in data generation. The estimated and the original parameters for models with one, two, or three causal variables on ten datasets, together with the relative errors, are presented in Table 4.

Table 4. Probability estimates for datasets with 1, 2, or 3 variables, by 10,000 observations.

		Estimated parameters				Original parameters				Relative error, %			
Data set	Model	r	$p1$	$p2$	$p3$	r	$p1$	$p2$	$p3$	r	$p1$	$p2$	$p3$
1	OLS	0.571	0.773			0.5	0.77			14.3	0.5		
2	OLS	0.107	0.269			0.1	0.31			7.1	14.0		
3	OLS	0.857	0.514			0.8	0.71			7.1	27.4		
4a	OLS	0.893	-0.426			0.8	0			11.6			
4b	Ridge	0.847	0.001			0.8	0			5.9			
5	OLS	0.104	0.026	0.716		0.1	0.03	0.72		4.4	8.2	0.4	
6	OLS	0.803	0.527	0.041		0.8	0.48	0.15		0.4	10.1	72.1	
7	OLS	0.095	0.837	0.911		0.1	0.87	0.87		4.8	4.2	4.3	
8	OLS	0.527	0.883	0.887		0.5	0.87	0.87		5.5	1.1	1.6	
9	OLS	0.489	0.677	0.393	0.677	0.5	0.53	0.53	0.97	2.2	28.8	25.2	30.5
10	OLS	0.099	0.498	0.559	0.613	0.1	0.53	0.53	0.65	0.7	5.2	6.2	5.5

In most cases, the OLS regression (6) works well, producing probabilities close to the original values used for the data simulation. In one dataset #4, the OLS yields the negative probability value (row 4a), so we run the ridge regression (7), which yields all positive probabilities (row 4b). It is interesting to note that the original p_1 in this case equals zero. The relative errors of the estimated probabilities to their original values show a reasonable precision mostly of several percent, but sometimes more (it often corresponds to close to zero or one original values).

What is especially important here, when the causes are dominantly random (like in rows 3, 4, and 6, when $r = 0.8$), the estimation procedure still yields very good results, separating causal related events with low intensity from this very high level (80%) of "noise". In fact, even the biggest deviation (72%) in row 6 for the estimate 0.04 vs. 0.15 doesn't seem bad with this high random influence. The other important feature: the procedure works even when coefficients are equal to each other, like in rows 7 and 8, with different level of randomness. It is remarkable because in traditional statistics, if two values (i.e., males and females) produce the same marginal frequency, the gender is considered having no causal interpretation. Actually, we can say that each cause works with the same intensity, and they both differ from the random cause.

In another experiment with eight variables, the original coefficients might take any values (not controlled). This situation matches the typical data sets in many applied research. The results of 40 simulations are shown in Table 5, where the average correlation of original *Y* with *X*s is 0.05, and the maximum correlation equals 0.15.

Table 5. Quality of the parameters estimation.

	*p*1	*p*2	*p*3	*p*4	*p*5	*p*6	*p*7	*p*8	*r*
Correlations between original and estimated values among 40 runs	0.69	0.87	0.78	0.81	0.83	0.86	0.74	0.80	0.64
Median error, % to original value	35	20	23	32	27	33	21	21	40

The first row in Table 5 shows that correlations between original and estimated values are quite big, so the procedure definitely captures the main features of the data. It is especially important because the original datasets (10,000 observations) have practically no correlations among *Y* and *X* variables, so in this situation the traditional statistical methods fail. The second row in Table 5 shows that median error is about 20-30% of the original values, similar to those in row 9 in Table 4. Of course, it is not an ideal but a good enough result in a situation where original data are uncorrelated. Other experiments showed that the estimations only slightly depend on the level of the mutual correlations between *X* variables, so the problem of multicollinearity is not so troubling in this approach as in common regression modeling.

Example of estimation of advertising efficiency

A typical phase in media planning is the analysis of mutual frequency distribution of the media vehicles (TV shows, magazines, websites, etc.) and of the particular brand consumption. The high brand frequency for some vehicle is considered as a good indicator, and this vehicle is included in the list of the candidates for making advertising via it. Table 6 in its left part presents cross-tabulation of five products and four media vehicles (all data are real and represent popular magazines and different important products; the number of the respondents is measured in tens of thousands). For example, in a cell Product 1 - Vehicle 2, or *P1*-*V2* (Table 6, left half), 13.8% means that this fraction of the readers of *V2* magazine have bought *P1*, so *V2* is the most promising vehicle (not accounting for circulation, frequency of advertising, and other factors).

Five causal models were constructed using each product as a target – the resulting parameters are presented in Table 6, the right part, with estimates of the random causes in the last column. Comparing the two parts of Table 6 shows a rather dramatic difference. The most promising cell *P1-V2* suggests that just about 3.1% of buyers (instead of 13.8%) might have bought the product due to this magazine's advertising, while the other customers could buy regardless of it. A similar diminishing we see in any cell, for instance, the Vehicle 1 is even not important at all, so all buyers have no relation to this magazine, they would buy the product anyway, without this advertising. This type of analysis shows completely different picture of the media performance, and the decisions about advertising distribution could be changed accordingly.

Table 6. Modeling of the advertising efficiency

Product *P*/ Vehicles *V*	Fraction of vehicle audience consuming particular product, %					Estimated causal coefficients, %				
	V1	*V2*	*V3*	*V4*	Total	*V1*	*V2*	*V3*	*V4*	Random causes
P1	7.2	13.8	9.2	11.8	8.5	0	3.1	1.9	6.5	8.3
P2	2.9	5.1	3.9	5	3.2	0	4.8	2.6	0	3.1
P3	0.3	1.2	1.8	0.5	0.5	0	1.1	1.5	0	0.5
P4	6.8	12.6	10.3	10.6	7.5	0	4.3	0	0	7.1
P5	2.2	3.4	5.9	3.8	2.9	0	0	0.7	1.9	3

For each product, the total number of positive outcomes was decomposed by different magazines, according to (9), (10) and (23) from Appendix B. As expected, the found shares reflect the importance of the magazines, as shown in Table 6. For example, for the product *P2* the vehicles *V2*, *V3* and random effect contribute as much as 17%, 7% and 76%, respectively; the random causes dominate (up to 95%) for all the considered products.

Conclusion

A new approach to causal modeling was considered based on the direct accounting for the internal relationship between the causal impacts and the outcome effect. The proposed model is a significant departure from the regular regression, or statistical learning models, as well as from the traditional models of causal analysis. In the suggested model, each causal variable effects the outcome individually, not cumulatively with others, which contrasts with the traditional statistics, where the outcome cumulates the combined effect of all the variables of influence, and adding variables improves the goodness of fit. Also, unlike in the traditional methods, the random cause is not considered as something to be "minimized", but rather as a reflection of all causes which were not captured by the introduced variables. The proposed approach to the analysis and estimation of causal relations demonstrates several important features:

- it offers a way to estimate the causal relationships, when many possible causes generate one effect – a situation very typical for numerous applications;
- it allows to estimate the intensity of the causal relationships in the data, even if there is no correlation between *Y* and *X* variables, when causal variables are highly correlated among themselves, when coefficients of variables are equal to each other, when random component in the data is very high; all these features make it very different from the traditional statistical and causal approaches;
- it works just with frequency tables (providing they exist for all or many combinations of the predictors), so there is no need for the original observational data sets, that may be very useful in many practical situations;
- parameter estimation is simple and could be performed with any available software.
- it works with data of high dimensionality, since the orthogonal design matrix allows to reduce estimation to paired regressions.

Future generalization of the main problems of the parameter evaluation for causal and random impacts can be seen in using numerical *Y* and *X* variables, and in the framework of complex causal relationships (as in structural equations, or acyclic graphs with do-operators).

References

Breiman, L. (2001). Statistical modeling: The two cultures. *Statistical Science*, *16*(3) 199-231 (with comments and rejoinder). doi:10.1214/ss/1009213726

Gifi, A. (1990). *Nonlinear Multivariate Analysis*. Chichester, England: Wiley.

Hastie, T., Tibshirani, R., & Friedman, J. (2006). *The elements of statistical learning: Data mining, inference, and prediction*. New York: Springer.

Kline, R. B. (2010). *Principles and practice of structural equation modeling*. New York, NY: Guilford Press.

Kuznetsov D. & Mandel, I. (2007), Statistical physics of media processes: Mediaphysics. *Physica A*, *377*(1), 253-268. doi:10.1016/j.physa.2006.10.098

Lipovetsky, S. (2010). Enhanced ridge regressions. *Mathematical and Computer Modeling*, *51*(5/6), 338-348. doi:10.1016/j.mcm.2009.12.028

Lipovetsky, S. (2012), Regression split by levels of the dependent variable. *Journal of Modern Applied Statistical Methods*, *11*(2), 319-324. Available at: http://digitalcommons.wayne.edu/jmasm/vol11/iss2/4

Mandel, I. (2011), Sociosystemics, statistics, decisions. *Model Assisted Statistics and Applications*, *6*(3), 163–217. doi:10.3233/MAS-2011-0203

Mandel, I. (2013), Fusion and causal analysis in big marketing data sets. *Proceedings of JSM - Section on Statistics in Marketing*. Montreal, Canada: American Statistical Association.

Pearl, J. (2000), *Causality: Models, reasoning, and inference*. Cambridge: Cambridge University Press.

Pearl, J. (2013). Linear models: A useful "microscope" for causal analysis. *Journal of Causal Inference. 1*(1), 155-170. doi:10.1515/jci-2013-0003

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*(5), 688-701. doi:10.1037/h0037350

Rubin, D. B. (2006). *Matched sampling for causal effects*. Cambridge, MA: Cambridge University Press.

Splawa-Neyman, J. (1990). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. (D. M. Dabrowska & T. P. Speed, Trans. and Ed.). *Statistical Science, 5*(4), 465 472. (Reprinted from 1923, *Annals of Agricultural Sciences, 10*, 1-51.)

Wright, S. (1921). Correlation and causation. *Journal of Agricultural Research*, *20*(7), 557–585.

Wright, S. (1960). Path coefficients and path regressions: Alternative or complimentary concepts? *Biometrics*, *16*(2), 189–202. doi:10.2307/2527551

Appendix A: Closed-form OLS solution for the orthogonal design

The OLS solution (6) for a multiple linear regression can be presented in analytical closed-form, if we have a total set of all possible cells, that is: for K variables there are $N = 2^K$ cells of all possible combinations of 0 and 1 values by each variable. For instance, if $K = 3$, there are eight cells as those presented in Table 1. For the orthogonal design matrix, each coefficient of regression can be estimated by the paired relation. For that we need covariance of y with each x_j:

$$\operatorname{cov}\left(y, x_j\right) = N^{-1}\sum_{i=1}^{N}\left(y_i - \bar{y}\right)\left(x_{ij} - \bar{x}_j\right) = N^{-1}\left(\sum_{i=1}^{N} y_i x_{ij} - N\bar{y}\bar{x}_j\right) \qquad \text{(A1)}$$

And the variance of each x is:

$$\operatorname{var}\left(x_j\right) = N^{-1}\sum_{i=1}^{N}\left(x_{ij} - \bar{x}_j\right)^2 = N^{-1}\left(\sum_{i=1}^{N} x_{ij}^2 - N\bar{x}_j^{\,2}\right) \qquad \text{(A2)}$$

Covariance of two binary predictors is:

$$\begin{aligned} \operatorname{cov}\left(x_j, x_k\right) &= N^{-1}\sum_{i=1}^{N}\left(x_{ij} - \bar{x}_j\right)\left(x_{ik} - \bar{x}_k\right) \\ &= N^{-1}\left(\sum_{i=1}^{N} x_{ij} x_{ik} - N\bar{x}_j\bar{x}_k\right) \qquad \text{(A3)} \\ &= \frac{1}{2^K}\left(2^{K-2} - \frac{2^K}{4}\right) = 0 \end{aligned}$$

Then each coefficient of regression (6) equals:

$$\begin{aligned} b_j &= \frac{\sum_{i=1}^{N} y_i x_{ij} - N\bar{y}\bar{x}_j}{\sum_{i=1}^{N} x_{ij}^2 - N\bar{x}_j^{\,2}} = \frac{\sum_{i=1}^{N} y_i x_{ij} - 2^{K-1}\bar{y}}{2^{K-1} - 2^{K-2}} \qquad \text{(A4)} \\ &= \frac{1}{2^{K-2}}\sum_{i=1}^{N} y_i x_{ij} - 2\bar{y} = \frac{1}{2^{K-2}}\left(y' x_j\right) - 2\bar{y} \end{aligned}$$

where $y'x_j$ is the scalar product of the vector and vector x_j. With all mean values of xs equal 0.5, the intercept in the model (5) equals:

$$b_0 = \bar{y} - b_1\bar{x}_1 - \ldots - b_K\bar{x}_K = \bar{y} - 0.5\left(b_1 + \ldots + b_K\right) \tag{A5}$$

Using (14) in (15) yields:

$$\begin{aligned} b_0 &= \bar{y} - 0.5\sum_{j=1}^{K}\left(\frac{1}{2^{K-2}}\left(y'x_j\right) - 2\bar{y}\right) \\ &= \bar{y} - \frac{1}{2^{K-1}}\sum_{j=1}^{K} y'x_j + K\bar{y} \\ &= \left(K+1\right)\bar{y} - \frac{1}{2^{K-1}}\sum_{j=1}^{K} y'x_j \end{aligned} \tag{A6}$$

The parameters of probability (8) are also related. Indeed, rewriting r using (A5) yields:

$$\begin{aligned} r &= 1 - \exp\left(b_0\right) \\ &= 1 - \exp\left(\bar{y} - 0.5\left(b_1 + \ldots + b_K\right)\right) \\ &= 1 - e^{\bar{y}}\left(e^{b_1}\right)^{-0.5}\ldots\left(e^{b_K}\right)^{-0.5} \\ &= 1 - \frac{e^{\bar{y}}}{\sqrt{(1-p_1)\ldots(1-p_K)}} \end{aligned} \tag{A7}$$

So, the relations (A5), or (A7) between the coefficients should be taken into account in simulations of the model parameters.

Appendix B. Decomposition to the additive shares of influence of each cause in the effect.

Consider (3) in a generalized form, where we denote $r = p_0$ and $x_{i0} = 1$ identically:

$$1 - S_i = \left(1-r\right)\prod_{k=1}^{K}\left(1-p_k\right)^{x_{ik}} = \prod_{k=0}^{K}\left(1-p_k\right)^{x_{ik}} \tag{B1}$$

with the aim to present S_i as a total of the items, each related to one of the causes:

$$S_i = \sum_{k=0}^{K} S_{ik} \tag{B2}$$

For the additive decomposition of S_i we take shares proportional to the ratio of logarithms:

$$S_{ik} = S_i \frac{\ln\left(1-p_k\right)^{x_{ik}}}{\ln\left(1-S_i\right)} \tag{B3}$$

The total of (B3), due to (B1), coincides with S_i itself:

$$\begin{aligned} \sum_{k=0}^{K} S_{ik} &= \sum_{k=0}^{K} S_i \frac{\ln\left(1-p_k\right)^{x_{ik}}}{\ln\left(1-S_i\right)} \\ &= S_i \frac{\sum_{k=0}^{K} \ln\left(1-p_k\right)^{x_{ik}}}{\ln\left(1-S_i\right)} = S_i \frac{\ln\left(\prod_{k=0}^{K}\left(1-p_k\right)^{x_{ik}}\right)}{\ln\left(1-S_i\right)} = S_i \end{aligned} \tag{B4}$$

If S_i was defined as the quotient of the counts $S_i = n_i / N_i$, where n_i is the counts of $Y = 1$ in the base size N_i of each cell. Then by using it in (B3), we obtain the estimation for the counts n_{ik} related to each k-th cause:

$$n_{ik} = N_i S_{ik} = N_i S_i \frac{\ln\left(1-p_k\right)^{x_{ik}}}{\ln\left(1-S_i\right)} = n_i \frac{\ln\left(1-p_k\right)^{x_{ik}}}{\ln\left(1-S_i\right)} \tag{B5}$$

Total of (B5) by k, similarly to (B4) yields:

$$\begin{aligned} \sum_{k=0}^{K} n_{ik} &= \sum_{k=0}^{K} n_i \frac{\ln\left(1-p_k\right)^{x_{ik}}}{\ln\left(1-S_i\right)} \\ &= n_i \frac{\sum_{k=0}^{K} \ln\left(1-p_k\right)^{x_{ik}}}{\ln\left(1-S_i\right)} = n_i \frac{\ln\left(\prod_{k=0}^{K}\left(1-p_k\right)^{x_{ik}}\right)}{\ln\left(1-S_i\right)} = n_i \end{aligned} \tag{B6}$$

Journal of Modern Applied Statistical Methods
May 2015, Vol. 14, No. 1, 196-207.

ISSN 1538 − 9472

A Comparison of Semi-Parametric and Nonparametric Methods for Estimating Mean Time to Event for Randomly Left Censored Data

Farzana Chowdhury
Northern University Bangladesh
Dhaka, Bangladesh

Jahida Gulshan
University of Dhaka
Dhaka, Bangladesh

Syed Shahadat Hossain
University of Dhaka
Dhaka, Bangladesh

The aim of this study was to make a comparison among existing estimation methods (Kaplan-Meier, Nelson-Aalen and Regression on Ordered Statistics (ROS)) for randomly left censored time to event data under selected distributions and for different level of censoring and sample sizes in order to determine the strength of these methods based on simulated data. Comparisons among the methods are made on the basis of unbiasedness and Monte Carlo Standard Error of the summary statistics (mean time to event) obtained by those methods under different conditions.

Keywords: Time to event data, Left censoring, detection limit, bias, Monte Carlo Standard Error

Introduction

Time to event data arises in a number of applied fields, such as medicine, biology, public health, epidemiology, engineering, economics, demography, actuarial science and many other scientific areas in which time to the occurrence of some event is of interest for some population of individuals. The most typical characteristic of time to event data is incompleteness where it arises either by censoring or by truncation. Censoring, a very common feature of time to event data broadly indicates the situation that some events are known to have occurred only within certain intervals but the exact time of occurrence is unknown (Klein & Moeschberger, 2003). Among different censoring situations, left censoring provides information indicating only that the event of interest has occurred prior

Farzana Chowdhury is a Masters graduate in Applied Statistics and currently working as a lecturer in Department of Business Administration. Email her at: fchowdhury@isrt.ac.bd. Jahida Gulshan is an Associate Professor. Email her at: gulshan@isrt.ac.bd. Dr. Hossein is a Professor. Email him at: shahadat@isrt.ac.bd .

to entry into the study (Klein & Moeschberger, 2003). In other words, left censored data are commonly encountered as values below a detection limit and hence are often termed as non-detects. A detection limit is a threshold below which measured values are not considered significantly different from a blank value, at a specified level of probability (Helsel, 2005a).

Although the analysis of left-censored data has important applications in various fields of study, very few studies focused on left censoring. Owen and DeRouen (1980) used Monte Carlo simulation techniques for estimating the average exposure of industrial workers to an air contaminant. Another study on water-quality data containing multiple detection limits used a robust approach to estimate the summary statistics and model the distributions of trace-level environmental data (Lee & Helsel, 2005). Popovic, Nie, Chettle, and McNeill (2007) used inverse variance weighting (IVW) of measurements to estimate the mean and standard error of the randomly left censored data on bone lead concentrations in order to provide valid inference about bone lead concentrations. A comparison based simulation study was done by Annan, Liu, and Zhang (2009) to compare a non-parametric, a semi parametric and a parametric approach to obtain estimates of summary statistics in different censoring situations and varying sample sizes

The Kaplan-Meier (Kaplan & Meier, 1958), Nelson-Aalen (Nelson, 1972 and Aalen, 1978), Maximum Likelihood (Cohen, 1959) and the Regression on Order Statistics (ROS) (Helsel & Cohn, 1988) are the methods available in literature for computing summary statistics on data with non-detects. The objective of this study is to compare three nonparametric and one semi-parametric estimation methods for finding summary statistics.

In this study, two different algorithms of Kaplan-Meier (1958) methods, one (denoted as KM-I in the rest of this paper) proposed by Helsel (2005a) and the other one (KM-II) by Popovic et al. (2007), was compared with another non parametric method based on modified Nelson Aalen method proposed by Popovic et.al (2007) and a semi parametric method based on Regression on Order Statistics (denoted as ROS) suggested by Helsel and Cohn (1988). A Monte Carlo simulation study was conducted to determine the efficiency of these methods for analyzing left-censored data under different distributions in terms of Bias and Monte Carlo Standard Error of the mean time to event in which the methods were employed for different sample sizes and different censoring levels.

Non-parametric Estimation of Mean

Let $S(x)$ be the survivorship function that gives the proportion of subjects expected to live at least x units of time. The survival probability is a product of incremental probabilities indicating the probabilities of surviving to the next lowest detection limit, given the number of observations at and below that detection limit. The mean of survival time x is calculated by

$$\mu\left(x^*\right)=\int_{b_1}^{b_2} S\left(u\right)du \qquad (1)$$

where $\mu(x^*)$ signifies that the mean of variable x is a function of the chosen interval $x_i : \{b_1 \le x_i \le b_2\}$. Parameter b_1 is the chosen lower boundary for the set of measurements.

Kaplan-Meier (KM) method

The Kaplan-Meier (KM) method proposed by Kaplan and Meier (1958) is a nonparametric method frequently considered as a standard method for estimating summary statistics of censored time to event data. The method has primarily been used for right-censored data. However, for calculation of summary statistics of left-censored data, the basic algorithm of Kaplan Meier method (used for right-censored data) has been modified. The modifications suggested are:

i. to transform left censored data to right censored one (Helsel, 2005b)
ii. to directly use left censored data with modified formulae (Popovic et al. 2007).

Formulation of KM method 1 According to the transformation method suggested by Helsel (2005b), the following steps are carried out to obtain the KM estimator of the survival probability:

i. All left-censored values are first arranged in descending order and subtracted from an arbitrarily chosen value larger than maximum value of the data set. Consequently, the left-censored data will automatically be transformed into right-censored data arranged in ascending order. All observations are then ranked from lowest to highest. For each subject $i = 1, \ldots, n$ (considering both censored and observed values), the transformed value will be

$$\omega = A_i - x_i \tag{2}$$

where A_i is an arbitrary constant, greater than the maximum observed value of the data set and x_i is the i^{th} observed value.

ii. The number of both detected and censored data that are at and below each observed value (observations at risk) are then computed as

$$b_j = n - r_j + 1 \tag{3}$$

where n is the total number of observations regarding both observed and censored and r_j is the rank of observed values only.

iii. If d_j denotes the number of observations at the j^{th} value (for tied values it is greater than 1), the incremental probabilities are given by

$$\frac{b_j - d_j}{b_j}, j = 1, \ldots, k \;, \tag{4}$$

and the product of the k incremental probabilities, going from high to low values for the k detected observations will give the KM estimator

$$\hat{S}(x) = \prod_{j=1}^{k} p_j \tag{5}$$

iv. The mean survival time is then estimated as

$$\begin{aligned}\hat{\mu}(x_j) = \hat{S}(x_0)x_1 + \hat{S}(x_1)(x_2 - x_1)\\ + \ldots + \hat{S}(x_{k-1})(x_k - x_{k-1}).\end{aligned} \tag{6}$$

Generally we consider $\hat{S}(x_0)=1$ and $\hat{S}(x_n)=0$.

v. The estimated mean survival time for original data will thus be

$$\hat{\mu}(x_j) = A_j - \hat{\mu}(x_j) \tag{7}$$

Formulation of KM method 2 The algorithm of this process was developed by Popovic et al. (2007) for estimating the survival function based, primarily, on the work of Kaplan and Meier (1958), Hosmer and Lemeshow (1999) and Ware and Demets (1976). According to this method, the following steps are to be carried out for obtaining this estimator:

i. For each subject $i = 1, \ldots, n, x_i$ is ordered in ascending order regarding both censored and observed data, and a censoring level δ_i is assigned such that $\delta_i = 1$, if the subject is observed and $\delta_i = 0$ if it is censored. Therefore, in case of a tie, censored entries should precede the observed events. Only the observed values along with their rank order r_i and censoring level δ_i from previous step will be considered. Thus the subjects with $\delta_i = 1$ are selected. For each entry, the incremental probabilities are calculated as

$$p_i = \frac{r_i - \delta_i}{r_i} \tag{8}$$

ii. Conventionally, $\hat{S}(x)$ is computed starting with the highest ranked entry X_n which is given as

$$\hat{S}(x) = \prod_{i=n}^{1} p_i \tag{9}$$

and the estimator of the mean for the given range $\{x_i : \{b_1 \leq x_i \leq b_2\}$is given by

$$\hat{\mu}(x^*) = \hat{\mu}(b_2) - \sum_{i=n}^{1} \hat{S}(x)(x_i - x_{i-1}), \text{ where } x_0 = b_1 \tag{10}$$

Since the survivorship function for left censored data equals unity for observations greater than the maximum observed event, $\hat{\mu}(b_2)$ is equal to the maximum observation in the set. As a result, the probability of having detected all observations greater than the maximum value of the data set is one. The probability decreases as x becomes progressively closer to b_1, with discontinuities at each measured event.

Nelson-Aalen method According to Popovic et al. (2007), computation method of Nelson-Aalen estimator (Nelson, 1972 and Aalen, 1978) for left-censored data set is similar to the KM method that uses left censored data directly. The basic difference between these two methods lies in the process of computing the survival probability, which instead of equation (7), is computed as

$$p_i = \frac{\delta_i}{r_i} \tag{11}$$

Semi-parametric Method (Regression on Order Statistics (ROS))

The algorithm of Regression on Order Statistics (ROS) method, developed by Helsel and Cohn (1988) can be summarized into following steps:

i. Let E_j be the probability of exceeding the j^{th} detection limit, by A_j the total number of uncensored observations in the range $[j, j+1)$ and by B_j the total number of observations, censored and uncensored, less than or equal to the j^{th} detection limit. Note that for highest detection limit, $E_{j+1} = 0$ and $A_j + B_j = n$. The exceedance probability E_j for each detection limit can be utilized for the computation of plotting positions for both censored and uncensored data using the relation

$$E_j = E_{j+1} + \frac{A_j}{A_j + B_j}\left(1 - E_{j+1}\right) \tag{12}$$

and the number of non-detects below the j^{th} detection limit is defined as

$$C_j = B_j - B_{j-1} - A_{j-1} \tag{13}$$

ii. A Weibull-type plotting position p can be calculated for a given uncensored observation by

$$p(i) = \left(1 - E_j\right) + \frac{\left(E_j - E_{j+1}\right)}{A_j + 1} r_i \tag{14}$$

where, E_j is the exceedance probability of the censoring limit below the observation, E_{j+1} is the exceedance probability of the censoring limit above the observation and r_i is the rank of the observation falling within the j^{th} and $(j + 1)^{\text{th}}$ detection limit.

iii. The Weibull-type plotting positions for censored observations are generally given by

$$p(i) = \frac{\left(1 - E_j\right)}{C_j + 1} r_i \tag{15}$$

iv. The normal quantiles of the plotting positions are known as the order statistics of the ROS method. A linear regression of the uncensored observations against the normal quantiles of the uncensored plotting positions is formed and the regression equation for predicting the unobserved data can be obtained as

$$\text{Predicted log-value} = \beta + \alpha \times \text{normal scores of the plotting positions} \tag{16}$$

v. The censored concentrations are modeled using the parameters of the linear regression and normal quantiles of the censored data. These modeled censored observations are used along with the uncensored observations, to model the distribution of the sample population. Individually, they are not considered the values that would have existed in the absence of censoring. The observed uncensored values are then combined with modeled censored values to corporately estimate summary statistics of the entire population. By combining both types of values this method avoids transformation bias.

Methodology

Simulation study

In this study, randomly left censored time to event data was simulated from exponential, Weibull and lognormal distribution where 1000 simulations were conducted for different combinations of sample sizes and censoring levels. The levels of censoring were considered to be 15%, 25% and 50% and the sizes of samples considered are small (25), moderately large (80) and large (200).

Results and Findings

A comparison of the methods by this simulation is made on the basis of the performances of the four methods, KM-I, KM-II, N-A and ROS in terms of absolute bias and MCSE of the estimates. Note that the performances of the four methods according to the two criteria have a nested factorial structure of its own, the factors that are taken under consideration of the simulation are:

1. Three different populations, namely exponential ($\lambda = 0.5$), Weibull ($\lambda = 1$, $k = 2$) and lognormal distribution ($\mu = 0.19$ and $\sigma = 1$)
2. Three different sample sizes 25, 80 and 200,
3. Three different levels (15%, 25% and 50%) of censored observations, and
4. Any possible interaction between the above factors.

The major findings of the simulation studies are summarized in Table 1. From these findings, it can be observed that when the populations mean is estimated using a sample drawn from an exponential (0.5) distribution, the ROS method performs the best in terms of absolute bias for all sample sizes and censoring levels considered in the study. For sample size 80, with 15%, 25% and 50% censored observations, the ROS method produced an absolute bias of 0.017, 0.037 and 0.112 respectively, which are lowest among the four methods, whereas the corresponding highest (among the four methods) absolute biases, 0.028, 0.083 and 0.412 respectively are observed for the KM-I method. Similar observations can be made for sample sizes 25 and 200 from exponential population, where ROS method produced the least absolute bias for estimate of mean for each of the censoring levels 15%, 25% and 50% and KM-I method produced the corresponding highest absolute bias.

Table 1. Comparison of Bias and Monte Carlo Standard Error (MCSE) of mean time to event for KM-I, KM-II, N-A and ROS methods under three different distributions (exponential with $\lambda = 0.5$, Weibull with $\lambda = 1$, $k = 2$ and lognormal with $\mu = 0.19$ and $\sigma = 1$)

			Absolute Bias				MCSE			
Distribution	Sample size	Cens. level	KM-I	KM-II	N-A	ROS	KM-I	KM-II	N-A	ROS
Exponential	25	0.15	0.047	0.163	0.206	0.025	0.397	0.422	0.414	0.401
		0.25	0.112	0.232	0.282	0.042	0.390	0.415	0.409	0.401
		0.50	0.486	0.194	0.301	0.108	0.367	0.401	0.387	0.418
	80	0.15	0.028	0.174	0.188	0.017	0.216	0.229	0.228	0.217
		0.25	0.083	0.234	0.252	0.037	0.211	0.225	0.225	0.217
		0.50	0.412	0.157	0.215	0.112	0.190	0.208	0.203	0.221
	200	0.15	0.025	0.169	0.175	0.017	0.141	0.148	0.148	0.142
		0.25	0.077	0.233	0.242	0.038	0.138	0.146	0.146	0.142
		0.50	0.395	0.127	0.164	0.120	0.124	0.134	0.131	0.146
Weibull	25	0.15	0.046	0.155	0.198	0.025	0.388	0.408	0.401	0.392
		0.25	0.104	0.239	0.290	0.035	0.395	0.416	0.411	0.409
		0.50	0.476	0.209	0.313	0.094	0.377	0.414	0.339	0.440
	80	0.15	0.033	0.157	0.172	0.023	0.219	0.228	0.227	0.221
		0.25	0.092	0.230	0.249	0.046	0.221	0.236	0.255	0.227
		0.50	0.416	0.155	0.215	0.119	0.192	0.210	0.207	0.227
	200	0.15	0.027	0.168	0.173	0.019	0.137	0.145	0.144	0.138
		0.25	0.079	0.231	0.240	0.041	0.133	0.140	0.140	0.137
		0.50	0.392	0.133	0.169	0.118	0.122	0.134	0.131	0.143
Lognormal	25	0.15	0.029	0.147	0.183	0.001	0.427	0.418	0.411	0.428
		0.25	0.070	0.218	0.260	0.004	0.425	0.402	0.396	0.426
		0.50	0.302	0.273	0.353	0.020	0.422	0.371	0.359	0.427
	80	0.15	0.032	0.133	0.145	0.009	0.247	0.246	0.245	0.247
		0.25	0.065	0.200	0.216	0.008	0.245	0.236	0.235	0.245
		0.50	0.265	0.228	0.271	0.001	0.237	0.208	0.204	0.242
	200	0.15	0.022	0.136	0.141	0.002	0.155	0.151	0.151	0.155
		0.25	0.055	0.203	0.211	0.001	0.154	0.147	0.146	0.154
		0.50	0.248	0.213	0.239	0.003	0.148	0.135	0.133	0.151

In case of Weibull (1, 2) population, the absolute bias produced by the ROS method is, again, the least among those of the four methods for each of the sample sizes and each of the censoring levels considered in the simulation. In comparison between methods, we can observe that for sample size 25 with 25% censored observations, absolute bias for the KM-I, KM-II, N-A and ROS methods are 0.104, 0.239, 0.289 and 0.035 respectively. For sample size 80, the computed absolute bias for the ROS method for 15%, 25% and 50% censored observations are 0.023, 0.046 and 0.119 respectively.

Considering the lognormal (0.19, 1) population, the absolute bias produced by the ROS method is still the least among those of the four methods for each of the sample sizes and each of the all censoring levels considered in the simulation. In comparison between methods, we observe for sample size 80 with 25% censored observations, absolute bias for the KM-I, KM-II, N-A and ROS methods are 0.065, 0.200, 0.216 and 0.008 respectively. For sample size 25, the computed absolute bias for the ROS method for 15%, 25% and 50% censored observations are 0.001, 0.004 and 0.020 respectively.

For all the methods and for all the sample sizes from lognormal (0.19, 1) population, the simulation results conform to the almost obvious affirmation that the absolute bias decreases as the censoring levels increases. When the samples are drawn from an exponential (0.5) or Weibull (1, 2) population, the same observation, that is, the absolute bias decreases as the censoring level increases, can be made for the KM-I and ROS methods and for all the sample sizes. The KM-II and N-A methods in cases of both exponential (0.5) or Weibull (1, 2) population, however, surprisingly showed inconsistency where the absolute bias decreases for 50% censoring levels.

The effect of increasing sample size on the absolute bias of the estimate of mean for the three methods other than the ROS method seems to be apparent for all the parent populations. For example, with exponential (0.5) population, the ROS method produces an absolute bias of 0.025, 0.017 and 0.017 for the sample sizes 25, 80 and 200 respectively at a censoring level of 15%. This eventually is indicating evidence of ROS method being insensitive to the increase of sample size from 80 to 200. The method has also been observed to be robust to the change of sample sizes with 25% and 50% of censoring levels and with Weibull (1, 2) and lognormal (0.19, 1) populations.

Although, the four methods differ substantially in terms of the bias of the estimated mean, it is noticeable that for lognormal (0.19, 1) population, the Monte Carlo Standard Error (MCSE) of the estimated mean is almost the same for the methods for same sample size and level of censoring. However, for exponential (0.5) and Weibull (1, 2) populations, slight differences in MCSEs is observed, and these differences reveal that the KM-I and ROS methods have a marginal advantage over the KM-II and N-A method. For example, for Weibull (1, 2) population, the MCSE for the four methods, KM-I, KM-II, N-A and ROS, for sample size 80 with 15% censoring level are 0.054, 0.057, 0.057 and 0.054 respectively. The difference of MCSE for different methods is seemingly higher for smaller sample sizes and higher level of censoring. The generally anticipated feature that the MCSE would be smaller for larger sample has been observed throughout.

Conclusion

The discussion in the earlier section can be summarized to reach to the following conclusions:

1. The ROS method produces the least absolute bias among those of the four methods for all sample sizes, all level of censoring for exponential (0.5), Weibull (1, 2) and lognormal (0.19, 1) populations.
2. The ROS method is more robust to the level of censoring. For increasing level of censoring, absolute bias of the estimate of mean increase for all sample sizes and all methods except for the ROS method.
3. For larger sample sizes, the MCSE of the estimate of mean of ROS method is the least among those of the four methods, although the differences of MSE are trivially small.
4. The ROS method is more robust to the change of sample size. For increasing sample size, absolute bias of both the estimates of mean increase for all levels of censoring and all methods except for the ROS method.

References

Aalen, O. (1978). Nonparametric inference for a family of counting processes. *Annals of Statistics, 6*(4), 701-726. doi:10.1214/aos/1176344247

Annan, S. Y., Liu, P. & Zhang, Y. (2009). *Comparison of the Kaplan-Meier, Maximum Likelihood, and ROS Estimators for left-censored data using simulation studies*. http://homepage.divms.uiowa.edu/~kcowles/s166_2009/Annan.pdf.

Cohen, A. C. (1959). Simplified estimators for the normal distribution when samples are singly censored or truncated. *Technometrics*, *1*(3), 217-237. doi:10.1080/00401706.1959.10489859

Helsel, D. R. & Cohn, T. A. (1988). Estimation of descriptive statistics for multiply censored water quality data. *Water Resources Research*, *24*(12), 1997-2004. doi:10.1029/WR024i012p01997

Helsel, D. R. (2005a). More than obvious: Better methods for interpreting nondetect data. *Environmental science & technology*, *39*(20), 419A-423A. doi:10.1021/es053368a

Helsel, D. R. (2005b). *Nondetects and data analysis: Statistics for censored environmental data*. Hoboken, NJ: Wiley-Interscience.

Hosmer, D. W. & Lemeshow, S. (1999). *Applied survival analysis: Regression modelling of time to event data*. New York, NY: Wiley.

Kaplan, E. L. & Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, *53*(282), 457-481. doi:10.1080/01621459.1958.10501452

Klein, J. P. & Moeschberger, M. L. (2003). *Survival analysis: techniques for censored and truncated data*. New York: Springer-Verlag.

Lee, L. & Helsel, D. (2005). Statistical analysis of water-quality data containing multiple detection limits: S-language software for regression on order statistics. *Computers & Geosciences*, *31*(10), 1241-1248. doi:10.1016/j.cageo.2005.03.012

Nelson, W. (1972). Theory and applications of hazard plotting for censored failure data. *Technometrics, 14*(4), 945-966. doi:10.1080/00401706.1972.10488991

Owen, W. J. & DeRouen, T. A. (1980). Estimation of the mean for lognormal data containing zeroes and left-censored values, with applications to the measurement of worker exposure to air contaminants. *Biometrics*, *36*(4), 707-719. doi:10.2307/2556125

Popovic, M., Nie, H., Chettle, D. R. & McNeill, F. E. (2007). Random left censoring: A second look at bone lead concentration measurements. *Physics in Medicine and Biology*, *52*(17), 5369-5378. doi:10.1088/0031-9155/52/17/018

Ware, J. H. & Demets, D. L. (1976). Reanalysis of some baboon descent data. *Biometrics,* 32(2), 459-463. doi:10.2307/2529516

Journal of Modern Applied Statistical Methods
May 2015, Vol. 14, No. 1, 208-214.

ISSN 1538 − 9472

Maximum Likelihood Estimation of the Kumaraswamy Exponential Distribution with Applications

K. A. Adepoju
University of Ibadan
Ibadan, Nigeria

O. I. Chukwu
University of Ibadan
Ibadan, Nigeria

The Kumaraswamy exponential distribution, a generalization of the exponential, is developed as a model for problems in environmental studies, survival analysis and reliability. The estimation of parameters is approached by maximum likelihood and the observed information matrix is derived. The proposed models are applied to three real data sets.

Keywords: Information matrix, Maximum likelihood, Moment generating function.

Introduction

A random variable X has the exponential distribution if its cumulative distribution function for $x > 0$ is given by

$$F(x) = 1 - \ell^{-\lambda x} \tag{1}$$

where $\lambda > 0$ is a scale parameter, the probability density function is

$$f(x) = \lambda \ell^{-\lambda x} \tag{2}$$

Using the Kumaraswamy link function by Cordeiro and de Castro (2011) given as

$$g(x) = a,b\,f(x)\left[F(x)\right]^{b-1}\left[1 - F(x)^{b}\right]^{a-1} \tag{3}$$

K.A. Adepoju is in the Department of Statistics. Email at ka.adepoju@ui.edu.ng. O.I. Chukwu is in the Department of Statistics. Email at unnachuks2002@yahoo.co.uk.

By inserting (1) and (2) in (3) we have

$$g\left(x\right)=a,b\,\lambda\ell^{-\lambda\ell}\left(1-\ell^{-\lambda x}\right)^{b-1}\left[1-\left(1-\ell^{-\lambda x}\right)^{b}\right]^{a-1} \tag{4}$$

$$a,\, b,\, \lambda>0$$

Another term of Kumaraswamy distribution can be obtained using the binomial series expansion. The Kumaraswamy exponential distribution in equation (4) can be expanded as follows:

$$\left(1-m\right)^{K}=\sum_{j=1}^{K}\left(-1\right)^{j}\binom{K}{j}M^{j} \tag{5}$$

as

$$g\left(x\right)=a\,b\lambda\ell^{-\lambda x}\left(1-\ell^{-\lambda x}\right)^{b-1}\sum_{j=1}^{K}\left(-1\right)^{j}\binom{K}{j}\left(1-\ell^{-\lambda x}\right)^{bj}$$

$$=a\,b\lambda\ell^{-\lambda x}\sum_{j=1}^{K}\left(-1\right)^{j}\binom{K}{j}\left(1-\ell^{-\lambda x}\right)^{bj+b-1}$$

Statistical inference

Given a random variable X following equation (4), the likelihood function is obtained as

$$L=a^{n}b^{n}\lambda^{n}\,\ell\prod_{i=1}^{n}\ell^{-\lambda x}\left(1-\ell^{-\lambda x}\right)^{b-1}\left[1-\left(1-\ell^{-\lambda x}\right)^{b}\right]^{a-1}$$

Taking log-likelihood of the above

$$\log L=n\log a+n\log b+n\log\lambda-\lambda\sum_{i=1}^{n}x+\left(b-1\right)\sum_{i=1}^{n}\log\left(1-\ell^{-\lambda x}\right)$$
$$+\left(a-1\right)\sum_{i=1}^{n}\log\left[1-\left(1-\ell^{-\lambda x}\right)^{b}\right]$$

$$\frac{\partial \log L}{\partial a} = \frac{n}{a} + \sum_{i=1}^{n} \log\left[1 - \left(1 - \ell^{-\lambda x}\right)^{b}\right]$$

$$\frac{\partial \log L}{\partial b} = \frac{n}{b} + \sum_{i=1}^{n} \log\left(1 - \ell^{-\lambda x}\right) - (a-1)\sum_{i=1}^{n} \frac{\left(1-\ell^{-\lambda x}\right)^{b} \log\left(1-\ell^{-\lambda x}\right)}{1-\left(1-\ell^{-\lambda x}\right)^{b}}$$

$$\frac{\partial \log L}{\partial \lambda} = \frac{n}{\lambda} - \sum_{i=1}^{n} x + (b-1)\sum_{i=1}^{n} \frac{x\,\ell^{-\lambda x}}{1-\ell^{-\lambda x}} - (a-1)b\sum_{i=1}^{n} \frac{x\ell^{-\lambda x}\left(1-\ell^{-\lambda x}\right)^{b-1}}{1-\left(1-\ell^{-\lambda x}\right)^{b}}$$

Fisher information

$$\frac{\partial^2 \log L}{\partial a^2} = -\frac{n}{a^2}$$

$$\frac{\partial^2 \log L}{\partial b^2} = -\frac{n}{b^2} - (a-1)\sum_{i=1}^{n} \frac{\left(1-\ell^{-\lambda x}\right)^{b}\left[\log\left(1-\ell^{-\lambda x}\right)\right]^{2}}{\left[1-\left(1-\ell^{-\lambda x}\right)^{b}\right]^{2}}$$

$$\frac{\partial^2 \log L}{\partial \lambda^2} = -\frac{n}{\lambda^2} - (b-1)\sum_{i=1}^{n} \frac{x}{\left(1-\ell^{-\lambda x}\right)^{2}} -$$

$$b(a-1)\sum_{i=1}^{n} \frac{x^2\ell^{-\lambda x}\left(1-\ell^{-\lambda x}\right)^{b-1}}{1-\left(1-\ell^{-\lambda x}\right)^{b}}\left[\frac{(b-1)\,\ell^{-\lambda x}\left(1-\ell^{-\lambda x}\right)^{b-2}}{\left(1-\ell^{-\lambda x}\right)^{b-1}} - \frac{b\ell^{-\lambda x}\left(1-\ell^{-\lambda x}\right)^{b-1}}{\left(1-\ell^{-\lambda x}\right)^{b}}\right]$$

$$\frac{\partial^2 \log L}{\partial a \partial b} = -\left(1-\ell^{-\lambda x}\right)^b \log\left(1-\ell^{-\lambda x}\right)$$

$$\frac{\partial^2 \log L}{\partial b \partial \lambda} = \sum_{i=1}^{n} \frac{x_i \ell^{-\lambda x_i}}{1-\ell^{-\lambda x_i}}$$

$$\frac{\partial^2 \log L}{\partial a \partial \lambda} = -b\sum_{i=1}^{n} \frac{x_i \ell^{-\lambda x_i}\left(1-\ell^{-\lambda x_i}\right)^{b-1}}{1-\left(1-\ell^{-\lambda x_i}\right)^b}$$

Application

For the sake of numerical illustrations, the two data sets used by Raja and Mir (2011) are considered. The first data set is on the failure time of the conditioning system of an airplane and the second is the runs scored by a Cricketer in 27 innings at national level.

Table 1. Descriptive Statistics on Failure Time on Conditional System

Min	Q_1	Q_2	Mean	Q_3	Max	Var
1.0	12.5	22.0	59.6	83.0	261.0	5167.421

Skewness	Kurtosis
1.693605	4.966655

Table 2. Descriptive Statistics in runs scored by a Cricketer

Min	Q_1	Q_2	Mean	Q_3	Max	Var
2.00	8.00	25.00	36.41	50.00	127.00	1149.02

Skewness	Kurtosis
1.124548	3.492725

Table 3. Failure Time on Conditional System

		Statistics	
Model	**Estimates**	**Log-likelihood**	**AIC**
Weibull	$\hat{\alpha}$=0.8536, $\hat{\lambda}$=0.0183	-151.970	305.94
Lognormal	$\hat{\mu}$=3.358, $\hat{\lambda}$=1.3190	151.621	305.242
Exponentiated Weibull	$\hat{\alpha}$=3.824, $\hat{\theta}$=0.1732, $\hat{\delta}$=82.235	-149.567	308.134
Exponentiated Gumbel	$\hat{\alpha}$=1.9881, $\hat{\lambda}$=49.0638	-148537	299.074
Exponentiated Lognormal	$\hat{\alpha}$=0.1542, $\hat{\mu}$=3.1353, $\hat{\delta}$=0.3648	-148.659	303.318
Lehman Type II Exponential	$\hat{\alpha}$=0.3439, $\hat{\lambda}$=0.0057	-152.6097	309.2593
Exponential	$\hat{\lambda}$=0.0168	-152.6297	307.2593
Kumaraswamy Exponential Distribution	$\hat{\alpha}$=10.142, $\hat{b}$=0.9129, $\hat{\lambda}$=0.0005	-107/9653	221.9306

Table 4. Runs Scored by a Cricketer

		Statistics	
Model	**Estimates**	**Log-likelihood**	**AIC**
Gamma	$\hat{\alpha}$=0.7235, λ=0.0127	-125.654	253.308
Weibull	$\hat{\alpha}$=1.040, λ=36.985	-124.021	250.042
Lognormal	$\hat{\mu}$=3.0534, λ=1.174	-125.059	252.118
Exponentiated exponential	$\hat{\alpha}$=0.8126, λ=0.0153	-125.945	253.93
Exponentiated Lognormal	$\hat{\alpha}$=0.578, $\hat{\mu}$=3.1836, $\hat{\delta}$=0.7834	-125.965	257.93
Exponentiated Gumbel	$\hat{\alpha}$=1.873, λ=45.264	-124.843	251.686
Exponential	$\hat{\lambda}$=0.0275	-124.0589	250.1177
Kumaraswamy Exponential	$\hat{\alpha}$=0.13006, $\hat{b}$=0.9557, $\hat{c}$=0.00014	-108.7224	223.4449

Conclusion

The probability density function of Kumaraswamy-exponential distribution was discussed and applied for two data sets. In first data set regarding failure times of the conditioning system of an aeroplane. Kumaraswamy exponential provided the best fit followed by exponentiated Gumbel. In second data set regarding runs scored by a cricketer Kumaraswamy exponential, Weibull and exponential distributions provided better fit.

References

Aarts, R. M. (2000). Lauricella functions. *From MathWorld -- A Wolfram Web Resource, created by Eric W. Weisstein.* *http://mathworld.wolfram.com/LauricellaFunctions.html.*

Akinsete, A., Famoye, F. & Lee, C. (2008). The beta-Pareto distribution. *Statistics, 42*(6), 547-563. doi:10.1080/02331880801983876

Chaudhry, M. A. & Zubair, S. M. (2002). *On a class of incomplete gamma functions with applications.* Chapman and Hall/CRC: Boca Raton, Florida.

Cooray, K. & Ananda, M. M. A. (2008). A generalization of the half-normal distribution with applications to lifetime data. *Communications in Statistics - Theory and Methods, 37*(9), 1323-1337. doi:10.1080/03610920701826088

Cordeiro, G. M & de Castro, M. (2009). A new family of generalized distributions. *Journal of Statistical Computation & Simulation*, 00(00), 1-17.

Cordeiro, G. M. & Nadarajah, S. (2011). Closed-form expressions for moments of a class of beta generalized distributions. *Brazilian Journal of Probability and Statistics, 25*(1), 14-33. doi:10.1214/09-BJPS109

Eugene, N., Lee, C. & Famoye, F. (2002). Beta-normal distribution and its applications. *Communications in Statistics-Theory and Methods, 31*(4), 497-512. doi:10.1081/STA-120003130

Exton, H. (1978). *Handbook of hypergeometric integrals: Theory, applications, tables, computer programs*. Halsted Press: New York.

Gupta, R. C., Gupta, P. L. & Gupta, R. D. (1998). Modeling failure time data by Lehman alternatives. *Communications in Statistics-Theory and Methods, 27*(4), 887-904. doi:10.1080/03610929808832134

Gupta, R. D. & Kundu, D. (1999). Theory & methods: Generalized exponential distributions. *Australian and New Zealand Journal of Statistics, 41*(2), 173-188. doi:10.1111/1467-842X.00072

Gupta, R. D. & Kundu, D. (2001). Exponentiated exponential family: An alternative to gamma and Weibull distributions. *Biometrical Journal, 43*(1), 117-130. doi:10.1002/1521-4036(200102)43:1<117::AID-BIMJ117>3.0.CO;2-R

Raja, T. A. & Mir, A. H. (2011). On extension of some exponentiated distributions with application. *International Journal Of Contemporary Mathematical Sciences, 6*(8), 393-400.

Journal of Modern Applied Statistical Methods
May 2015, Vol. 14, No. 1, 215-237.

ISSN 1538 − 9472

Estimation for the Parameters of the Exponentiated Exponential Distribution Using a Median Ranked Set Sampling

Monjed H. Samuh
Palestine Polytechnic University
Palestinian Territories

Areen Qtait
Palestine Polytechnic University
Palestinian Territories

The method of maximum likelihood estimation based on Median Ranked Set Sampling (MRSS) was used to estimate the shape and scale parameters of the Exponentiated Exponential Distribution (EED). They were compared with the conventional estimators. The relative efficiency was used for comparison. The amount of information (in Fisher's sense) available from the MRSS about the parameters of the EED were be evaluated. Confidence intervals for the parameters were constructed using MRSS.

Keywords: Exponentiated exponential distribution; Fisher information; Maximum likelihood estimation; Median ranked set sampling; Ranked set sampling

Introduction

One of the most common approaches of data collection is that of a simple random sample (SRS). Other more structured sampling designs, such as stratified sampling or probability sampling, are also available to help make sure that the obtained data collection provides a good representation of the population of interest. Any such additional structure of this type revolves around how the sample data themselves should be collected in order to provide an informative image of the larger population. With any of these approaches, once the sample items have been chosen, the desired measurements are collected from each of the selected items.

Many efforts are made to develop statistical techniques for data collection that generally leads to more representative samples (samples whose characteristics accurately reflect those of the underlying population). To this end, ranked set sampling and some of its variations were developed.

Dr. Samuh is Assistant Professor of Statistics in the College of Applied Sciences. Email him at monjedsamuh@ppu.edu.

In this section, the ranked set sampling (RSS) and median ranked set sampling (MRSS) are presented. The exponentiated exponential distribution (EED) and its properties are also discussed.

Ranked set sampling

RSS was proposed by McIntyre (1952) to estimate a pasture yield in Australia. This method was not used for a long time, but in the last 30 years a lot of research work was done using this method, which has become very important in different aspects.

SRS and RSS are both independent, but they differ in several ways, like:

1. RSS is more efficient than SRS with the same number of measured elements.
2. Development of RSS procedure is more difficult than that of SRS.
3. In SRS, just m elements are needed but in RSS m elements are chosen out of m^2 to achieve the desired sample.

Also stratified random sampling and RSS are different in some things like:

1. In stratified sampling we limited with no more six strata but in RSS we are not restricted ourselves with the number of sets.
2. In both of them SRS is used but in RSS ordering the elements in each set is needed before selecting the sample.

RSS as a method used basically for infinite population where the set of sampling units drawn from a population can ranked in a cheap way which is not costly and/or time consuming. The steps of choosing RSS are as follows:

1. Randomly select m sets each of size m elements from the population under study.
2. The elements for each set in Step (1) are ranked visually or by any negligible cost method that does not require actual measurements.
3. Select and quantify the i^{th} minimum from the i^{th} set, $i = 1, 2, \ldots, m$ to get a new set of size m, which is called the ranked set sample.
4. Repeat Steps (1) − (3) h times (cycles) until obtaining a sample of size $n = mh$.

Figure 1 illustrates the procedure of RSS in terms of matrices. Let $Y_i = \{X_{(ii)}; i = 1, \ldots, m\}$; that is, the obtained RSS, $\{X_{(11)}, X_{(22)}, \ldots, X_{(mm)}\}$, is denoted by $\{Y_1, Y_2, \ldots, Y_m\}$. If the process is repeated h cycles, then the RSS can be represented as a matrix of size $n = hm$ as it is shown in Step 4 of Figure 1.

$$\text{Step 1: } \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1m} \\ X_{21} & X_{22} & \cdots & X_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ X_{m1} & X_{m2} & \cdots & X_{mm} \end{bmatrix} \qquad \text{Step 2: } \begin{bmatrix} X_{(11)} & X_{(12)} & \cdots & X_{(1m)} \\ X_{(21)} & X_{(22)} & \cdots & X_{(2m)} \\ \vdots & \vdots & \ddots & \vdots \\ X_{(m1)} & X_{(m2)} & \cdots & X_{(mm)} \end{bmatrix}$$

$$\text{Step 3: } \left\{X_{(11)}, X_{(22)}, \ldots, X_{(mm)}\right\} \qquad \text{Step 4: } \begin{bmatrix} Y_{11} & Y_{12} & \cdots & Y_{1m} \\ Y_{21} & Y_{22} & \cdots & Y_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ X_{h1} & X_{h2} & \cdots & X_{hm} \end{bmatrix}$$

Figure 1. Ranked set sampling procedure

RSS as a method is applicable where ranking and sampling units is much cheaper than the measurement of the variable of interest. In particular RSS can be used in the following situation:

1. Ranking units in a set can be done easily by judgment in the variable of interest through visual inspection or with the help of certain auxiliary means.
2. If there is a concomitant variable can be obtained easily (concomitant is a variable which is not of major concern but are correlated with the variable of interest).

Median ranked set sampling

MRSS was suggested by Muttlak (1997) as a method to estimate the population mean instead of RSS to reduce the errors, and increase the efficiency over RSS and SRS. It is described by the following steps:

1. Randomly select m^2 sample units from the target population.
2. Allocate the m^2 units into m sets each of size m, and rank the units within each set.
3. From each set in Step (2), if the sample size m is odd select from each set the $\left(\frac{m+1}{2}\right)^{th}$ smallest rank unit i.e the median of each set. While if the sample size m is even select from the first $\frac{m}{2}$ sets the $\left(\frac{m}{2}\right)^{th}$ smallest rank unit and from the second $\frac{m}{2}$ sets take the $\left(\frac{m+2}{2}\right)^{th}$ smallest rank unit. This step yields m sample elements which is the median RSS.
4. Repeat Steps (1) − (3) h times (cycles) until obtaining a sample of size $n = mh$.

Figure 2 illustrates the procedure of MRSS when $m = 4$ in terms of matrices. Let us denote the MRSS $\{X_{(12)}, X_{(22)}, X_{(33)}, X_{(43)}\}$ by $\{Y_1, Y_2, Y_3, Y_4\}$. If the process is repeated h cycles, then the RSS can be represented as a matrix of size $n = 4h$ as it is shown in Step 4 of Figure 2.

Step 1: $\begin{bmatrix} X_{11} & X_{12} & X_{13} & X_{14} \\ X_{21} & X_{22} & X_{23} & X_{24} \\ X_{31} & X_{32} & X_{33} & X_{34} \\ X_{41} & X_{42} & X_{43} & X_{44} \end{bmatrix}$ Step 2: $\begin{bmatrix} X_{(11)} & X_{(12)} & X_{(13)} & X_{(14)} \\ X_{(21)} & X_{(22)} & X_{(23)} & X_{(24)} \\ X_{(31)} & X_{(32)} & X_{(33)} & X_{(34)} \\ X_{(41)} & X_{(42)} & X_{(43)} & X_{(44)} \end{bmatrix}$

Step 3: $\{X_{(12)}, X_{(22)}, X_{(33)}, X_{(44)}\}$ Step 4: $\begin{bmatrix} Y_{11} & Y_{12} & Y_{13} & Y_{14} \\ Y_{21} & Y_{22} & Y_{23} & Y_{24} \\ \vdots & \vdots & \vdots & \vdots \\ Y_{h1} & Y_{h2} & Y_{h3} & Y_{h4} \end{bmatrix}$

Figure 2. Median ranked set sampling procedure

The exponentiated exponential distribution

The exponentiated exponential distribution (EED) introduced by Gupta and Kundu (1999) as a generalization of the exponential distribution. It is of great interest and is popularly used in analyzing lifetime or survival data. Consider the random variable X that is exponentiated exponential-distributed with scale parameter $\lambda > 0$ and shape parameter $\alpha > 0$. The probability density function of X is given by

$$f\left(x;\alpha,\lambda\right)=\alpha\lambda e^{-\lambda x}\left(1-e^{-\lambda x}\right)^{\alpha-1};x>0.$$

The corresponding cumulative distribution function is given by

$$F\left(x;\alpha,\lambda\right)=\left(1-e^{-\lambda x}\right)^{\alpha}.$$

It is clear that the EED is simply the α^{th} power of the exponential cumulative distribution. So, the case where $\alpha = 1$ is called the exponential distribution. The mean, variance, skewness, kurtosis and the pdf's curves of the EED for different values of the scale and shape parameters are shown in Table 1.

The properties of the EED have been studied by many authors, see for example Gupta and Kundu (2001), Nadarajah (2011), Ghitany et al. (2013), and Ristić and Nadarajah (2014).

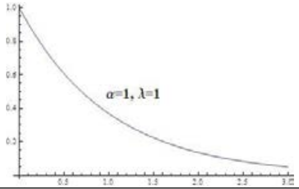
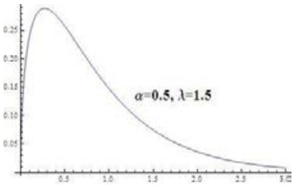
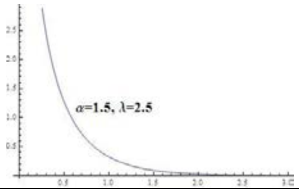
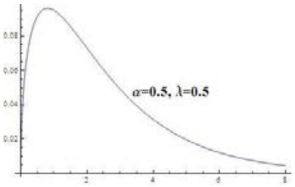
Literature Review

Stokes (1976) used RSS for estimating the parameters in a location-scale family of distributions. The RSS estimators of the location and scale parameters are shown to be more efficient than the SRS estimators. She also used RSS to estimate the correlation coefficient of a bivariate normal distribution.

Lam et al. (1994) used RSS for estimating two-parameter exponential distribution.

$$f\left(y\right)=\frac{1}{\sigma}\exp\left\{\frac{-\left(y-\theta\right)}{\sigma}\right\} \qquad (1)$$

Table 1. The mean, variance, skewness, kurtosis and the pdf's curves of the EED for different values of *α* and *λ*.

(*α*,*λ*)	**Properties of the EED**		
(1,1)	*Mean:* 1 *Variance:* 1	*Skewness:* 2 *Kurtosis:* 9	*PDFs Curve:* 
(0.5,1.5)	*Mean:* 0.2845 *Variance:* 0.2790	*Skewness:* 3.8514 *Kurtosis:* 19.6675	*PDFs Curve:* 
(1.5,2.5)	*Mean:* 0.7364 *Variance:* 1.0640	*Skewness:* -0.5053 *Kurtosis:* 0.5123	*PDFs Curve:* 
(0.5,0.5)	*Mean:* 0.8536 *Variance:* 2.5109	*Skewness:* 3.8514 *Kurtosis:* 19.6675	*PDFs Curve:* 

An unbiased estimators of θ and σ based on RSS with their variances are derived. They made a comparison between these estimators with their counterpart in SRS.

Stokes (1995) considered the location-scale distribution, $F\left(\dfrac{x-\mu}{\sigma}\right)$, and estimated μ and σ using the methods of maximum likelihood estimation and best linear unbiased estimation within the framework of RSS. Sinha et al. (1996) used RSS to estimate the parameters of the normal and exponential distributions. Their work assumed partial knowledge of the underlying distribution without any knowledge of the parameters. For each parameter, they proposed best linear unbiased estimators for full and partial RSS.

Samawi and Al-Sagheer (2001) studied the use of Extreme RSS (ERSS) and MRSS for distribution function estimation. For a random variable *X*, it is shown that the distribution function estimator when using ERSS and MRSS are more efficient than when using SRS and RSS for some values of a given *x*.

Abu-Dayyeh and Sawi (2009) considered the maximum likelihood estimator and the likelihood ratio test for making inference about the scale parameter of the exponential distribution in case of moving extreme ranked set sampling (MERSS). The estimators and test cannot be written in closed form. Therefore, a modification of the maximum likelihood estimator using the technique suggested by Maharota and Nanda (1974) was considered. It was used to modify the likelihood ratio test to get a test in closed form for testing a simple hypothesis against one-sided alternatives.

Al-Omari and Al-Hadhrami (2011) used ERSS to estimate the parameters and population mean of the modified Weibull distribution. The maximum likelihood estimators are investigated and compared to the corresponding one based on SRS. It was found that the estimators based on ERSS are more efficient than estimators using SRS. The ERSS estimator of the population mean was also found to be more efficient than the SRS based on the same number of measured units.

Haq et al. (2013) proposed a partial ranked set sampling (PRSS) method for estimation of population mean, median and variance. On the basis of perfect and imperfect rankings, Monte Carlo simulations from symmetric and asymmetric distributions are used to evaluate the effectiveness of the proposed estimators. It was found that the estimators under PRSS are more efficient than the estimators based on simple random sampling.

Abu-Dayyeh et al. (2013) used RSS for studying the estimation of the shape and location parameters of the Pareto distribution. The estimators were compared with their counterpart in SRS in terms of their biases and mean square errors. It was shown that the estimators based on RSS can be real competitors against those based on SRS.

Sarikavanij et al. (2014) considered simultaneous comparison of the location and scale estimators of a two-parameter exponential distribution based on SRS and RSS by using generalized variance (GV). They suggested various RSS strategies to estimate the scale parameter. Their performances in terms of GV were compared with SRS strategy. It was shown that the minimum values of set size, *m*, based on RSS, which would result in smaller GV than that based on SRS.

Maximum likelihood estimation and fisher information based on SRS

Consider a random sample coming from the EED *f*(*x*; *α*; *λ*) where the values of *α* and *λ* are unknown. The likelihood function is given by

$$L_{SRS}(\alpha,\lambda)=\alpha^{n}\lambda^{n}\prod_{i=1}^{n}(1-e^{-\lambda x_i})^{\alpha-1}e^{-\lambda\sum_{i=1}^{n}x_i};\quad \alpha>0,\lambda>0.$$

Thus, the log likelihood function is

$$\log L_{SRS}(\alpha,\lambda)=n\log\alpha+n\log\lambda+(\alpha-1)\sum_{i=1}^{n}\log(1-e^{-\lambda x_i})-\lambda\sum_{i=1}^{n}x_i. \qquad (2)$$

The normal equations become

$$\frac{\partial\log L_{SRS}(\alpha,\lambda)}{\partial\lambda}=\frac{n}{\lambda}+(\alpha-1)\sum_{i=1}^{n}\frac{x_i e^{-\lambda x_i}}{1-e^{-\lambda x_i}}-\sum_{i=1}^{n}x_i=0. \qquad (3)$$

$$\frac{\partial\log L_{SRS}(\alpha,\lambda)}{\partial\alpha}=\frac{n}{\alpha}+\sum_{i=1}^{n}\log(1-e^{-\lambda x_i})=0, \qquad (4)$$

From Equation 4, the maximum likelihood estimator (MLE) of *α* as a function of *λ*, say $\hat{\alpha}(\lambda)$, is

$$\hat{\alpha}(\lambda)=\frac{-n}{\sum_{i=1}^{n}\log\left(1-e^{-\lambda x_i}\right)}.$$

Substituting $\hat{\alpha}(\lambda)$ in Equation 2, we obtain the profile log-likelihood of *λ* as

$$\begin{aligned}\log L_{SRS}(\hat{\alpha}(\lambda),\lambda)&=n\log n-n\log-\sum_{i=1}^{n}\log(1-e^{-\lambda x_i})\\&\quad+n\log\lambda-n-\sum_{i=1}^{n}\log(1-e^{-\lambda x_i})-\lambda\sum_{i=1}^{n}x_i.\end{aligned} \qquad (5)$$

The MLE of *λ*, can be obtained by maximizing (5) w.r.t *λ* as

$$\frac{\partial \log L_{SRS}(\hat{\alpha}(\lambda),\lambda)}{\partial \lambda}=\frac{n}{\lambda}-\sum_{i=1}^{n}\frac{x_i e^{-\lambda x_i}}{1-e^{-\lambda x_i}}-\sum_{i=1}^{n}x_i-\frac{n}{\sum_{i=1}^{n}\log(1-e^{-\lambda x_i})}\sum_{i=1}^{n}\frac{x_i e^{-\lambda x_i}}{1-e^{-\lambda x_i}}. \quad (6)$$

However, the solutions are not in closed forms, in order to obtain estimates for α and λ, the normal equations can be solved numerically.

Fisher information (*FI*) number is used to measure the amount of information that an observable sample carries about the parameter(s). The *FI* number for the parameter θ is defined as

$$FI(\theta)=\frac{\partial^2 \log L(\theta)}{\partial \theta^2}.$$

Based on the random sample $X_1, X_2, \ldots, X_n$ the *FI* numbers of α and λ are, respectively, given by

$$FI_{SRS}(\alpha)=\frac{\partial^2 \log L_{SRS}(\alpha,\lambda)}{\partial \alpha^2}=\frac{n}{\alpha^2},$$
$$FI_{SRS}(\lambda)=\frac{\partial^2 \log L_{SRS}(\alpha,\lambda)}{\partial \lambda^2}=\frac{n}{\lambda^2}+(\alpha-1)\sum_{i=1}^{n}\frac{x_i^2 e^{-\lambda x_i}}{(1-e^{-\lambda x_i})^2}.$$

Maximum likelihood estimation and fisher information based on MRSS

Consider the maximum likelihood estimation of the parameters α and λ of EED under MRSS paying attention to the odd and even set sizes.

Odd set sizes

Suppose $\{Y_{ji}; j = 1, 2, \ldots, m\}$ is a MRSS from an EED, where h is the number of cycles and m is the set size. Since the set size m is assumed to be odd, the Y_{ji} are independent and identically distributed as the distribution of the $\left(\frac{m+1}{2}\right)^{th}$ order statistics of the random sample $X_1, X_2, \ldots, X_m$; that is

$$f_Y\left(y\right)=f_{X_{\left(\frac{m+1}{2}\right)}}\left(y\right)=\frac{m!}{\left(\left(\frac{m-1}{2}\right)!\right)^2}\alpha\lambda e^{-\lambda y}\left(1-e^{-\lambda y}\right)^{\alpha\left(\frac{m+1}{2}\right)-1}\left(1-\left(1-e^{-\lambda y}\right)^{\alpha}\right)^{\frac{m-1}{2}}.$$

The likelihood function of MRSS for odd set size m is given by

$$\begin{aligned}L_{MRSSO}\left(\alpha,\lambda\right)&=\prod_{j=1}^{h}\prod_{i=1}^{m}f_Y\left(y_{ji}\right)=c_1\alpha^{mh}\lambda^{mh}e^{-\lambda\sum_{j=1}^{h}\sum_{i=1}^{m}y_{ji}}\\&\times\prod_{j=1}^{h}\prod_{i=1}^{m}\left(1-e^{-\lambda y_{ji}}\right)^{\alpha\left(\frac{m+1}{2}\right)^{-1}}\prod_{j=1}^{h}\prod_{i=1}^{m}\left(1-\left(1-e^{-\lambda y_{ji}}\right)^{\alpha}\right)^{\frac{m-1}{2}}\end{aligned}\tag{7}$$

where c_1 is a constant. Thus, the log likelihood function is

$$\begin{aligned}\log L_{MRSSO}\left(\alpha,\lambda\right)&=d_1+mh\log\alpha+mh\log\lambda-\lambda\sum_{j=1}^{h}\sum_{i=1}^{m}y_{ji}\\&+\sum_{j=1}^{h}\sum_{i=1}^{m}y_{ji}\left(\alpha\left(\frac{m+1}{2}\right)-1\right)\sum_{j=1}^{h}\sum_{i=1}^{m}\left(1-e^{\lambda y_{ji}}\right)\\&+\left(\frac{m-1}{2}\right)\sum_{j=1}^{h}\sum_{i=1}^{m}\log\left(1-\left(1-e^{-\lambda y_{ji}}\right)^{\alpha}\right)\end{aligned}\tag{8}$$

where d_1 is a constant.

The normal equations become

$$\begin{aligned}\frac{\partial\log L_{MRSSO}\left(\alpha,\lambda\right)}{\partial\alpha}&=\frac{mh}{\alpha}-\sum_{j=1}^{h}\sum_{i=1}^{m}y_{ji}\\&+\left(\alpha\left(\frac{m+1}{2}\right)-1\right)\sum_{j=1}^{h}\sum_{i=1}^{m}\frac{y_{ji}e^{-\lambda y_{ji}}}{1-e^{\lambda y_{ji}}}\\&-\alpha\left(\frac{m-1}{2}\right)\sum_{j=1}^{h}\sum_{i=1}^{m}\frac{y_{ji}e^{\lambda y_{ji}}\left(1-e^{\lambda y_{ji}}\right)^{\alpha-1}}{1-\left(1-e^{\lambda y_{ji}}\right)\alpha}=0\end{aligned}\tag{9}$$

$$\frac{\partial \log L_{MRSSO}(\alpha,\lambda)}{\partial \alpha}=\frac{mh}{\alpha}+\left(\frac{m+1}{2}\right)\sum_{j=1}^{h}\sum_{i=1}\log\left(1-e^{-\lambda y_{ji}}\right)$$
$$-\left(\frac{m-1}{2}\right)\sum_{j=1}^{h}\sum_{i=1}\frac{\left(1-e^{-\lambda y_{ji}}\right)^{\alpha}\log\left(1-e^{-\lambda y_{ji}}\right)}{1-\left(1-e^{-\lambda y_{ji}}\right)^{\alpha}}=0, \qquad (10)$$

The MLEs of the parameters α and λ are the solutions of the Equations (9) and (10). However, the solutions are not in closed forms, in order to obtain estimates for α and λ, the normal equations can be solved numerically. Based on the MRSS $\{Y_{ji}; j = 1, 2, \ldots, h; i = 1, 2, \ldots, m\}$, for odd set size m, the *FI* numbers of α and λ are, respectively, given by

$$FI_{MRSSO}(\alpha)=\frac{\partial^2 \log L_{MRSSO}(\alpha,\lambda)}{\partial \alpha^2},$$
$$FI_{MRSSO}(\lambda)=\frac{\partial^2 \log L_{MRSSO}(\alpha,\lambda)}{\partial \lambda^2}.$$

The observed *FI* numbers are evaluated at the maximum likelihood estimates.

Even set sizes

Because the set size m is assumed to be even, for each $j=1,2,\ldots,h; Y_{ji}\sim f_{Y_i}(y)$,

$$f_{Y_i}(y)=\begin{cases} f_{X_{\left(\frac{m}{2}\right)}}(y) & \text{for } i=1,\ldots,\frac{m}{2} \\ f_{X_{\left(\frac{m+2}{2}\right)}}(y) & \text{for } i=\frac{m+2}{2},\ldots,m \end{cases}$$

where $X_{\left(\frac{m}{2}\right)}$ and $X_{\left(\frac{m+2}{2}\right)}$ are the $\left(\frac{m}{2}\right)^{th}$ and the $\left(\frac{m+2}{2}\right)^{th}$ order statistics of the random sample $X_1, X_2, \ldots, X_m$; therefore, for $i=1,\ldots,\frac{m}{2}$; Y_{ji} are independent and identically distributed as

$$f_{Y_i}(y) = \frac{m!}{\frac{m}{2}\left(\left(\frac{m}{2}-1\right)!\right)^2}\alpha\lambda e^{-\lambda y}\left(1-e^{-\lambda y}\right)^{\alpha\frac{m}{2}-1}\left(1-(1-e^{-\lambda y})^{\alpha}\right)^{\frac{m}{2}}, \qquad (11)$$

$$f_{Y_i}(y) = \frac{m!}{\frac{m}{2}\left(\left(\frac{m}{2}-1\right)!\right)^2}\alpha\lambda e^{-\lambda y}\left(1-e^{-\lambda y}\right)^{\alpha\left(\frac{m}{2}+1\right)-1}\left(1-(1-e^{-\lambda y})^{\alpha}\right)^{\frac{m}{2}-1}. \qquad (12)$$

and for $i = \frac{m+2}{2}, \ldots, m;$ Y_{ji} are independent and identically distributed as

Note $\{Y_{ji}; j = 1, 2, \ldots, h; i = 1, 2, \ldots, m\}$ are independent. Thus, the likelihood function of MRSS for even set size m is given by

$$\begin{aligned} L_{MRSSE}(\alpha,\lambda) &= \left(\prod_{j=1}^{h}\prod_{i=1}^{\frac{m}{2}} f_{Y_i}(y_{ji})\right)\times\left(\prod_{j=1}^{h}\prod_{i=\frac{m+2}{2}}^{m} f_{Y_i}(y_{ji})\right) \\ &= c_2\alpha^{mh}\lambda^{mh}e^{-\lambda\sum_{j=1}^{h}\sum_{i=1}^{\frac{m}{2}}y_{ji}}\times e^{-\lambda\sum_{j=1}^{h}\sum_{i=\frac{m+2}{2}}^{m}y_{ji}}\prod_{j=1}^{h}\prod_{i=1}^{\frac{m}{2}}\left(1-e^{-\lambda y_{ji}}\right)^{\alpha\frac{m}{2}-1} \\ &\times\prod_{j=1}^{h}\prod_{i=\frac{m+2}{2}}^{m}\left(1-e^{-\lambda y_{ji}}\right)^{\alpha\left(\frac{m}{2}+1\right)-1}\prod_{j=1}^{h}\prod_{i=1}^{\frac{m}{2}}\left(1-\left(1-e^{-\lambda y_{ji}}\right)^{\alpha}\right)^{\frac{m}{2}} \\ &\prod_{j=1}^{h}\prod_{i=\frac{m+2}{2}}^{m}\left(1-\left(1-e^{-\lambda y_{ji}}\right)^{\alpha}\right)^{\frac{m}{2}-1}, \end{aligned} \qquad (13)$$

where c_2 is a constant. Thus, the log likelihood function is

$$
\begin{aligned}
\log L_{MRSSE}(\alpha,\lambda) = d_2 + mh\log\alpha + mh\log\lambda & \\
-\lambda\sum_{j=1}^{h}\sum_{i=1}^{\frac{m}{2}} y_{ji} - \lambda\sum_{j=1}^{h}\sum_{i=\frac{m+2}{2}}^{m} y_{ji} & \\
+\left(\alpha\frac{m}{2}-1\right)\sum_{j=1}^{h}\sum_{i=1}^{\frac{m}{2}}\log\left(1-e^{-\lambda y_{ji}}\right) & \\
+\left(\alpha\left(\frac{m}{2}+1\right)-1\right)\sum_{j=1}^{h}\sum_{i=\frac{m+2}{2}}^{m}\log\left(1-e^{-\lambda y_{ji}}\right) & \\
+\frac{m}{2}\sum_{j=1}^{h}\sum_{i=1}^{\frac{m}{2}}\log\left(1-\left(1-e^{-\lambda y_{ji}}\right)^{\alpha}\right) & \\
+\left(\frac{m}{2}-1\right)\sum_{j=1}^{h}\sum_{i=\frac{m+2}{2}}^{m}\log\left(1-\left(1-e^{-\lambda y_{ji}}\right)^{\alpha}\right), &
\end{aligned}
\tag{14}
$$

where d_2 is a constant.

The normal equations become

$$
\begin{aligned}
\log L_{MRSSE}(\alpha,\lambda) = d_2 + mh\log\alpha + mh\log\lambda & \\
-\lambda\sum_{j=1}^{h}\sum_{i=1}^{\frac{m}{2}} y_{ji} - \lambda\sum_{j=1}^{h}\sum_{i=\frac{m+2}{2}}^{m} y_{ji} & \\
+\left(\alpha\frac{m}{2}-1\right)\sum_{j=1}^{h}\sum_{i=1}^{\frac{m}{2}}\log\left(1-e^{-\lambda y_{ji}}\right) & \\
+\left(\alpha\left(\frac{m}{2}+1\right)-1\right)\sum_{j=1}^{h}\sum_{i=\frac{m+2}{2}}^{m}\log\left(1-e^{-\lambda y_{ji}}\right) & \\
+\frac{m}{2}\sum_{j=1}^{h}\sum_{i=1}^{\frac{m}{2}}\log\left(1-\left(1-e^{-\lambda y_{ji}}\right)^{\alpha}\right) & \\
+\left(\frac{m}{2}-1\right)\sum_{j=1}^{h}\sum_{i=\frac{m+2}{2}}^{m}\log\left(1-\left(1-e^{-\lambda y_{ji}}\right)^{\alpha}\right), &
\end{aligned}
\tag{15}
$$

$$\begin{aligned}\frac{\partial \log L_{MRSSE}(\alpha,\lambda)}{\partial\alpha} &= \frac{mh}{\alpha}+\frac{m}{2}\sum_{j=1}^{h}\sum_{i=1}^{\frac{m}{2}}\log\left(1-e^{-\lambda y_{ji}}\right)\\ &+\left(\frac{m}{2}+1\right)\sum_{j=1}^{h}\sum_{i=\frac{m+2}{2}}^{m}\log\left(1-e^{-\lambda y_{ji}}\right)\\ &-\frac{m}{2}\sum_{j=1}^{h}\sum_{i=1}^{\frac{m}{2}}\frac{\left(1-e^{-\lambda y_{ji}}\right)^{\alpha}\log\left(1-e^{-\lambda y_{ji}}\right)}{1-\left(1-e^{-\lambda y_{ji}}\right)^{\alpha}} \qquad (16)\\ &-\left(\frac{m}{2}-1\right)\sum_{j=1}^{h}\sum_{i=\frac{m+2}{2}}^{m}\frac{\left(1-e^{-\lambda y_{ji}}\right)^{\alpha}\log\left(1-e^{-\lambda y_{ji}}\right)}{1-\left(1-e^{-\lambda y_{ji}}\right)^{\alpha}}\\ &=0\end{aligned}$$

The MLEs of the parameters α and λ are the solutions of the Equations (15) and (16). However, the solutions are not in closed forms, in order to obtain estimates for α and λ, the normal equations can be solved numerically. Based on the MRSS $\{Y_{ji}; j = 1, 2, \ldots, h; i = 1, 2, \ldots, m\}$, for even set size m, the *FI* numbers of α and λ are, respectively, given by

$$FI_{MRSSE}(\alpha)=\frac{\partial^2 \log L_{MRSSE}(\alpha,\lambda)}{\partial\alpha^2},$$

$$FI_{MRSSE}(\lambda)=\frac{\partial^2 \log L_{MRSSE}(\alpha,\lambda)}{\partial\lambda^2}.$$

The observed *FI* numbers are evaluated at the maximum likelihood estimates.

The comparison between the resulting estimators under MRSS and SRS can be done using the asymptotic efficiency (see Stokes, 1995). The asymptotic efficiency of MRSS w.r.t SRS for estimating θ is defined by

$$Aeff(\hat{\theta}_{MRSS};\hat{\theta}_{SRS})=\lim_{n\to\infty} eff(\hat{\theta}_{MRSS};\hat{\theta}_{SRS})=\frac{FI_{MRSS}(\theta)}{FI_{SRS}(\theta)}.$$

Interval Estimates

Let $X_1, \ldots, X_n$ be a random sample from $f(x;\theta)$, where θ is an unknown quantity. A confidence interval for the parameter θ, with confidence level or confidence coefficient $1-\gamma$, is an interval with random endpoints $[S_L(X_1, \ldots, X_n), S_u(X_1, \ldots, X_n)]$. It is given by

$$P\left(S_L\left(X_1,\ldots,X_n\right) \le \theta \le S_U\left(X_1,\ldots,X_n\right)\right) = 1-\gamma.$$

The interval $[S_L(X_1, \ldots, X_n), S_u(X_1, \ldots, X_n)]$ is called a $100(1-\gamma)\%$ confidence interval for θ.

For large sample size, the maximum likelihood estimator, under appropriate regularity conditions (see Davison, 2008, p.118), has many useful properties, including reparametrization-invariance, consistency, efficiency, and the sampling distribution of a maximum likelihood estimator $\hat{\theta}_{MLE}$ is asymptotically unbiased and also asymptotically normal with its variance obtained from the inverse Fisher information number of sample size 1 at the unknown parameter θ; that is, $\hat{\theta}_{MLE} \to N\left(\theta, FI^{-1}(\theta)\right)$ as $n \to \infty$. Therefore, the approximate $100(1-\gamma)\%$ confidence limits for the $\hat{\theta}_{MLE}$ of θ can be constructed as

$$P\left(-z_{\frac{\gamma}{2}} \le \frac{\hat{\theta}-\theta}{\sqrt{FI^{-1}(\theta)}} \le z_{\frac{\gamma}{2}}\right) = 1-\gamma,$$

where z_γ is the γ^{th} upper percentile of the standard normal distribution. Therefore, the approximate $100(1-\gamma)\%$ confidence limits for the scale and location parameters of the EED are given, respectively, by

$$P\left(\hat{\alpha} - z_{\frac{\gamma}{2}}\sqrt{FI^{-1}(\alpha)} \le \alpha \le \hat{\alpha} + z_{\frac{\gamma}{2}}\sqrt{FI^{-1}(\alpha)}\right) = 1-\gamma, \tag{17}$$

$$P\left(\hat{\lambda} - z_{\frac{\gamma}{2}}\sqrt{FI^{-1}(\lambda)} \le \lambda \le \hat{\lambda} + z_{\frac{\gamma}{2}}\sqrt{FI^{-1}(\lambda)}\right) = 1-\gamma. \tag{18}$$

Then, the approximate confidence limits for α and λ will be constructed using Equation (17) and (18), respectively.

Simulation Study

To investigate the properties of the maximum likelihood estimators of the scale and locations parameters of the EED a simulation study is conducted. Monte Carlo simulation is applied for different sample sizes, $m = \{2,3,4,5\}$ and $h = \{10,50,100\}$, and for different parameter values, $(\alpha, \lambda) = \{(1,1),(0.5,1.5),(1.5,2.5)\}$. The estimates of α and λ, the bias estimates, the MSEs, and the efficiency values are computed over 2000 replications for different cases. The results are reported in Tables 2-4. Moreover, the observed Fisher information matrices and the asymptotic efficiency in estimating α and λ under SRS and MRSS are calculated and the results reported in Table 5. The observed Fisher information numbers of α and λ based on SRS are denoted by $FI_{SRS}(\hat{\alpha})$ and $FI_{SRS}(\hat{\lambda})$, respectively, and the observed information numbers of α and λ based on MRSS are denoted by $FI_{MRSS}(\hat{\alpha})$ and $FI_{MRSS}(\hat{\lambda})$, respectively. The asymptotic efficiency, *Aeff*, for estimating α is found as the ratio

$$Aeff(\hat{\alpha}_{MRSS};\hat{\alpha}_{SRS}) = \frac{FI_{MRSS}(\alpha)}{FI_{SRS}(\alpha)},$$

and for estimating λ is found as the ratio

$$Aeff(\hat{\lambda}_{MRSS};\hat{\lambda}_{SRS}) = \frac{FI_{MRSS}(\lambda)}{FI_{SRS}(\lambda)}.$$

Confidence intervals based on SRS and MRSS for $(\alpha, \lambda) = (1.5,2.5)$ for different sample sizes are constructed at $1-\gamma = 0.95$ level of confidence using Equation (17) and (18), respectively, and the results are shown in Table 5.

Table 2. The Bias, MSE, and Efficiency values of estimating the parameters (α = 1, λ = 1) under SRS and MRSS.

			$\alpha = 1$				$\lambda = 1$			
m	**h**	**Sampling**	$\hat{\alpha}$	**Bias($\hat{\alpha}$)**	**MSE($\hat{\alpha}$)**	**eff**	$\hat{\lambda}$	**Bias($\hat{\lambda}$)**	**MSE($\hat{\lambda}$)**	**eff**
2	*10*	*SRS*	1.1731	0.1731	0.2063		1.1346	0.1346	0.1490	
		MRSS	0.7835	-0.2165	0.1270	1.6247	0.7520	-0.2480	0.1446	1.0300
	50	*SRS*	1.0298	0.0298	0.0213		1.0231	-0.3187	0.1147	
		MRSS	0.7075	-0.2925	0.0954	0.2228	0.6813	-0.3187	0.1147	0.1791
	200	*SRS*	1.0066	0.0066	0.0044		1.0090	0.0090	0.0047	
		MRSS	0.6915	-0.3085	0.0973	0.0451	0.6626	-0.3374	0.1168	0.0401
3	*10*	*SRS*	1.1015	0.1015	0.1038		1.0782	0.0782	0.0650	
		MRSS	1.0949	0.0949	0.0880	1.1795	1.0730	0.0730	0.0650	1.1968
	50	*SRS*	1.0229	0.0229	0.0125		1.0160	0.0160	0.0117	
		MRSS	1.0165	0.0165	0.0116	1.0770	1.0114	0.0114	0.0103	1.1317
	200	*SRS*	1.0056	0.0056	0.0030		1.0037	0.0037	0.0028	
		MRSS	1.0052	0.0052	0.0027	1.1180	1.0036	0.0036	0.0026	1.1019
4	*10*	*SRS*	1.0612	0.0612	0.0599		1.0550	0.0550	0.0561	
		MRSS	0.8719	-0.1281	0.0492	1.2174	0.8553	-0.1447	0.0561	0.9920
	50	*SRS*	1.0154	0.0154	0.0091		1.0141	0.0141	0.0091	
		MRSS	0.8298	-0.1702	0.0347	0.2612	0.8189	-0.1811	0.0396	0.2291
	200	*SRS*	1.0027	0.0027	0.0022		1.0034	0.0034	0.0022	
		MRSS	0.8230	-0.1770	0.0326	0.0669	0.8122	-0.1878	0.0369	0.0600
5	*10*	*SRS*	1.0514	0.0514	0.0474		1.0460	0.0460	0.0415	
		MRSS	1.0571	0.0571	0.0396	1.1966	1.0486	0.0486	0.0351	1.1821
	50	*SRS*	1.0099	0.0099	0.0076		1.0076	0.0076	0.0074	
		MRSS	1.0099	0.0099	0.0064	1.1812	1.0075	0.0075	0.0060	1.2246
	200	*SRS*	1.0044	0.0044	0.0019		1.0028	0.0028	0.0018	
		MRSS	1.0025	0.0025	0.0014	1.3314	1.0014	0.0014	0.0014	1.2649

Table 3. The Bias, MSE, and Efficiency values of estimating the parameters (α = 0.5, λ = 1.5) under SRS and MRSS.

			***α* = 0.5**				***λ* = 1.5**			
m	***h***	**Sampling**	$\hat{\alpha}$	***Bias(*** $\hat{\alpha}$ ***)***	***MSE(*** $\hat{\alpha}$ ***)***	***eff***	$\hat{\lambda}$	***Bias(*** $\hat{\lambda}$ ***)***	***MSE(*** $\hat{\lambda}$ ***)***	***eff***
2	*10*	*SRS*	0.5687	0.0687	0.0349		1.7938	0.2938	0.6328	
		MRSS	0.4062	-0.0938	0.0240	1.4557	1.0823	-0.4177	0.4572	1.3842
	50	*SRS*	0.5124	0.0124	0.0041		1.5499	0.0499	0.0730	
		MRSS	0.3748	-0.1252	0.0178	0.2317	0.9367	-0.5633	0.3566	0.2048
	200	*SRS*	0.5024	0.0024	0.0009		1.5181	0.0181	0.0160	
		MRSS	0.3680	-0.1320	0.0179	0.0493	0.9006	-0.5994	0.3676	0.0445
3	*10*	*SRS*	0.5410	0.0410	0.0188		1.6688	0.1688	0.3043	
		MRSS	0.5366	0.0366	0.0146	1.2872	1.6565	0.1565	0.2594	1.1732
	50	*SRS*	0.5100	0.0100	0.0025		1.5341	0.0346	0.0418	
		MRSS	0.5067	0.0067	0.0021	1.1871	1.5246	0.0246	0.0382	1.0930
	200	*SRS*	0.5024	0.0024	0.0006		1.5080	0.0080	0.0099	
		MRSS	0.5021	0.0021	0.0005	1.2382	1.5077	0.0077	0.0094	1.0532
4	*10*	*SRS*	0.5238	0.0238	0.0110		1.6165	0.1165	0.2066	
		MRSS	0.4463	-0.0537	0.0090	1.2301	1.2424	-0.2576	0.1895	1.0905
	50	*SRS*	0.5062	0.0062	0.0018		1.5284	0.0284	0.0318	
		MRSS	0.4290	-0.0710	0.0062	0.2929	1.1656	-0.3344	0.1345	0.2365
	200	*SRS*	0.5010	0.0010	0.0004		1.5070	0.0070	0.0077	
		MRSS	0.4263	-0.0737	0.0057	0.0775	1.1520	-0.3480	0.1263	0.0608
5	*10*	*SRS*	0.5205	0.0205	0.0091		1.5984	0.0984	0.1544	
		MRSS	0.5219	0.0219	0.0065	1.3956	1.6029	0.1029	0.1371	1.1265
	50	*SRS*	0.5041	0.0041	0.0015		1.5166	0.0166	0.0260	
		MRSS	0.5039	0.0039	0.0011	1.3306	1.5160	0.0160	0.0225	1.1578
	200	*SRS*	0.5019	0.0019	0.0004		1.5055	0.0055	0.0062	
		MRSS	0.5010	0.0010	0.0003	1.5124	1.5033	0.0033	0.0053	1.1736

Table 4. The Bias, MSE, and Efficiency values of estimating the parameters (α = 1.5, λ = 2.5) under SRS and MRSS.

			$\alpha = 1.5$				$\lambda = 2.5$			
m	*h*	Sampling	$\hat{\alpha}$	*Bias*($\hat{\alpha}$)	*MSE*($\hat{\alpha}$)	*eff*	$\hat{\lambda}$	*Bias*($\hat{\lambda}$)	*MSE*($\hat{\lambda}$)	*eff*
2	*10*	*SRS*	1.8041	0.3041	0.6164		2.7886	0.2886	0.7157	
		MRSS	1.1505	-0.3495	0.3413	1.8061	1.9210	-0.5790	0.7692	0.9304
	50	*SRS*	1.5507	0.0507	0.0575		2.5493	0.0493	0.1040	
		MRSS	1.0212	-0.4788	0.2538	0.2264	1.7673	-0.7327	0.6082	0.1710
	200	*SRS*	1.5119	0.0119	0.0116		2.5198	0.0198	0.0238	
		MRSS	0.9948	-0.5052	0.2605	0.0445	1.7256	-0.7744	0.6157	0.0387
3	*10*	*SRS*	1.6762	0.1762	0.2947		2.6681	0.1681	0.3835	
		MRSS	1.6680	0.1680	0.2600	1.1336	2.6581	0.1581	0.3202	1.1976
	50	*SRS*	1.5380	0.0380	0.0329		2.5341	0.0341	0.0589	
		MRSS	1.5285	0.0285	0.0320	1.0270	2.5245	0.0245	0.0520	1.1341
	200	*SRS*	1.5092	0.0092	0.0078		2.5080	0.0080	0.0145	
		MRSS	1.5088	0.0088	0.0074	1.0628	2.5078	0.0078	0.0130	1.1103
4	*10*	*SRS*	1.6086	0.1086	0.1688		2.6194	0.1194	0.2785	
		MRSS	1.2888	-0.2112	0.1343	1.2568	2.1679	-0.3321	0.2911	0.9568
	50	*SRS*	1.5267	0.0267	0.0241		2.5311	0.0311	0.0461	
		MRSS	1.2173	-0.2827	0.0951	0.2531	2.0889	-0.4111	0.2045	0.2255
	200	*SRS*	1.5049	0.0049	0.0057		2.5074	0.0074	0.0113	
		MRSS	1.2054	-0.2946	0.0901	0.0636	2.0742	-0.4258	0.1897	0.0594
5	*10*	*SRS*	1.5897	0.0897	0.1286		2.5990	0.0990	0.2071	
		MRSS	1.6012	0.1012	0.1160	1.1095	2.6059	0.1059	0.1740	1.1900
	50	*SRS*	1.5169	0.0169	0.0201		2.5163	0.0163	0.0375	
		MRSS	1.5172	0.0172	0.0181	1.1081	2.5160	0.0160	0.0307	1.2210
	200	*SRS*	1.5073	0.0073	0.0050		2.5063	0.0063	0.0093	
		MRSS	1.5041	0.0041	0.0040	1.2501	2.5030	0.0030	0.0072	1.2845

Table 5. The observed Fisher information matrix, the variance-covariance matrix, a 95% confidence interval of the parameters (α = 1.5, λ = 2.5), and the asymptotic efficiency under SRS and MRSS.

				Fisher Information			Variance-Covariance		95% CI		
m	*h*	Sampling		$\hat{\alpha}$	$\hat{\lambda}$	*Aeff*	$\hat{\alpha}$	$\hat{\lambda}$	Lower	Upper	Width
2	10	*SRS*	$\hat{\alpha}$	8.67	-4.23	3.60[a]	0.2067	0.1870	0.9130	2.6952	1.7822
			$\hat{\lambda}$	-4.23	4.68	2.17[b]	0.1870	0.3830	1.5756	4.0016	2.4260
		MRSS	$\hat{\alpha}$	31.26	-12.88		0.0670	0.0851	0.6432	1.6578	1.0147
			$\hat{\lambda}$	-12.88	10.15		0.0851	0.2065	1.0303	2.8117	1.7813
	200	*SRS*	$\hat{\alpha}$	177.59	-89.25	3.37	0.0111	0.0110	1.3054	1.7184	0.4130
			$\hat{\lambda}$	-89.25	90.67	2.16	0.0110	0.0218	2.2304	2.8094	0.5788
		MRSS	$\hat{\alpha}$	597.86	-257.35		0.0039	0.0051	0.8724	1.1172	0.2448
			$\hat{\lambda}$	-257.35	195.78		0.0051	0.0118	1.5127	1.9385	0.4258
3	10	*SRS*	$\hat{\alpha}$	13.31	-6.53	1.91	0.1389	0.1299	0.9457	2.4067	1.4610
			$\hat{\lambda}$	-6.53	6.98	2.00	0.1299	0.2649	1.6593	3.6769	2.0176
		MRSS	$\hat{\alpha}$	25.45	-15.57		0.1235	0.1376	0.9793	2.3568	1.3776
			$\hat{\lambda}$	-15.57	13.98		0.1376	0.2249	1.7286	3.5876	1.8590
	200	*SRS*	$\hat{\alpha}$	266.10	-134.32	1.92	0.0075	0.0073	1.3395	1.6789	0.3395
			$\hat{\lambda}$	-134.32	136.78	2.04	0.0073	0.0145	2.2720	2.7440	0.4720
		MRSS	$\hat{\alpha}$	512.14	-320.54		0.0070	0.0080	1.3448	1.6728	0.3280
			$\hat{\lambda}$	-320.54	278.70		0.0080	0.0128	2.2861	2.7296	0.4435
4	10	*SRS*	$\hat{\alpha}$	17.86	-8.77	3.73	0.1052	0.1002	0.9729	2.2443	1.2714
			$\hat{\lambda}$	-8.77	9.22	2.98	0.1002	0.2039	1.7344	3.5044	1.7701
		MRSS	$\hat{\alpha}$	66.59	-36.32		0.0540	0.0714	0.8333	1.7443	0.9109
			$\hat{\lambda}$	-36.32	27.45		0.0714	0.1309	1.4588	2.8770	1.4183
	200	*SRS*	$\hat{\alpha}$	355.82	-179.45	3.74	0.0056	0.0055	1.3582	1.6516	0.2933
			$\hat{\lambda}$	-179.45	182.44	2.99	0.0055	0.0109	2.3028	2.7120	0.4093
		MRSS	$\hat{\alpha}$	1329.79	-734.72		0.0029	0.0040	1.0999	1.3109	0.2111
			$\hat{\lambda}$	-734.72	545.12		0.0040	0.0072	1.9079	2.2405	0.3326

(continued)

5	10	*SRS*	$\hat{\alpha}$	22.49	-11.06	2.79	0.0843	0.0811	1.0206	2.1588	1.1382
			$\hat{\lambda}$	-11.06	11.49	3.06	0.0811	0.1651	1.8026	3.3954	1.5928
		MRSS	$\hat{\alpha}$	62.70	-41.62		0.0748	0.0886	1.0651	2.1373	1.0721
			$\hat{\lambda}$	-41.62	35.12		0.0886	0.1335	1.8898	3.3220	1.4323
	200	*SRS*	$\hat{\alpha}$	443.02	-223.87	2.88	0.0045	0.0044	1.3758	1.6388	0.2630
			$\hat{\lambda}$	-223.87	227.76	3.11	0.0044	0.0087	2.3235	2.6891	0.3656
		MRSS	$\hat{\alpha}$	1277.40	-854.20		0.0041	0.0049	1.3786	1.6296	0.2510
			$\hat{\lambda}$	-854.20	707.46		0.0049	0.0073	2.3355	2.6705	0.3349

***Note:** a) *Aeff* ($\hat{\alpha}_{MRSS}$, $\hat{\alpha}_{SRS}$), b) *Aeff* ($\hat{\lambda}_{MRSS}$, $\hat{\lambda}_{SRS}$)

Conclusion

The method of maximum likelihood estimation for estimating the shape and scale parameters of the EED is studied in the MRSS framework. The new obtained estimators are com-pared with the conventional estimators obtained by SRS. The relative efficiency are calculated for comparing the estimators. The amount of information available from the MRSS about the parameters of the EED is evaluated. Confidence intervals for the parameters are constructed using SRS and MRSS. More specifically, we have the following conclusions.

1. From Tables 2-4 it can be concluded that:

   a. For each sampling method, the MSEs of the estimators decrease as the set size increases and as the number of cycle increases.
   b. It is clear, from the biases, that MRSS overestimate and when the set size is odd and underestimate and when the set size is even.
   c. The biases of the estimators based on MRSS when the set size is odd decrease as the number of cycle increases. When the set size is even the biases of the estimators based on MRSS increase as the number of cycle increases.
   d. The efficiency is always greater than 1 when the set size is odd; that is, MRSS is more efficient than SRS in estimating the parameters of the EED.

2. From Table 5 it can be concluded that:

a. Fisher information numbers obtained from MRSS are greater than that from SRS.

b. The asymptotic variances of the estimators decrease as the set size increases and as the number of cycle increases.

c. The interval width of the estimators decreases as the set size increases and as the number of cycle increases.

d. The interval width obtained by MRSS is narrower than the one obtained by SRS.

References

Abu-Dayyeh, W., Assrhani, A., & Ibrahim, K. (2013). Estimation of the shape and scale parameters of Pareto distribution using ranked set sampling. *Statistical Papers, 54*(1), 207-225. doi:10.1007/s00362-011-0420-3

Abu-Dayyeh, W. & Sawi, E. A. (2009). Modified inference about the mean of the exponential distribution using moving extreme ranked set sampling. *Statistical Papers, 50*(2), 249-259. doi:10.1007/s00362-007-0072-5

Al-Omari, A. & Al-Hadhrami, S. A. (2011). On maximum likelihood estimators of the parameters of a modified Weibull distribution using extreme ranked set sampling. *Journal of Modern Applied Statistical Methods, 10*(2), 607-617. http://digitalcommons.wayne.edu/jmasm/vol10/iss2/18/

Davison, A. C. (2008). *Statistical Models*. New York: Cambridge University Press.

Ghitany, M. E., Al-Jarallah, R. A., & Balakrishnan, N. (2013). On the existence and uniqueness of the MLEs of the parameters of a general class of exponentiated distributions. *Statistics: A Journal of Theoretical and Applied Statistics, 47*(3), 605-612. doi:10.1080/02331888.2011.614950

Gupta, R. D. & Kundu, D. (1999). Generalized exponential distributions. *Australian and New Zealand Journal of Statistics, 41*(2), 173-188. doi:10.1111/1467-842X.00072

Gupta, R. D. & Kundu, D. (2001). Exponentiated exponential family: An alternative to gamma and Weibull distributions. *Biometrical Journal, 43*(1), 117-130. doi:10.1002/1521-4036(200102)43:1<117::AID-BIMJ117>3.0.CO;2-R

Haq, A., Brown, J., Moltchanova, E., & Al-Omari, A. I. (2013). Partial ranked set sampling design. *Environmetrics, 24*(3), 201-207. doi:10.1002/env.2203

Sinha, B. K., Sinha, B. K., & Purkayastha, S. (1996). On some aspects of ranked set sampling for estimation of normal and exponential parameters. *Statistics & Risk Modeling, 14*(3), 223-240. doi:10.1524/strm.1996.14.3.223

Lam, K., Sinha, B. K., & Wu, Z. (1994). Estimation of parameters in two-parameter exponential distribution using ranked set sampling. *Annals of the Institute of Statistics and Mathematics, 46*(4), 723-736. doi:10.1007/BF00773478

Maharota, K. & Nanda, P. (1974). Unbiased estimator of parameter by order statistics in the case of censored samples. *Biometrika, 61*(3)*,* 601-606. doi:10.1093/biomet/61.3.601

McIntyre, G. (1952). A method for unbiased selective sampling, using ranked sets. *Australian Journal of Agricultural Research, 3*(4)*,* 385-390. doi:10.1071/AR9520385

Muttlak, H. A. (1997). Median ranked set sampling. *Journal of Applied Statistical Science, 6*(4)*,* 245-255.

Nadarajah, S. (2011). The exponentiated exponential distribution: a survey. *Advances in Statistical Analysis, 95*(3)*,* 219-251. doi:10.1007/s10182-011-0154-5

Ristić, M. M. & Nadarajah, S. (2014). A new lifetime distribution. *Journal of Statistical Computation and Simulation, 84*(1)*,* 135-150. doi:10.1080/00949655.2012.697163

Samawi, H. M. & Al-Sagheer, O. (2001). On the estimation of the distribution function using extreme and median ranked set sampling. *Biometrical Journal, 43*(3)*,* 357-373. doi:10.1002/1521-4036(200106)43:3<357::AID-BIMJ357>3.0.CO;2-Q

Sarikavanij, S., Kasala, S., & Sinha, B. K. (2014). Estimation of location and scale parameters in two-parameter exponential distribution based on ranked set sample. *Communications in Statistics - Simulation and Computation, 43*(1)*,* 132-141. doi:10.1080/03610918.2012.698776

Stokes, L. (1995). Parametric ranked set sampling. *Annals of the Institute of Statistical Mathematics, 47*(3)*,* 465-482. doi:10.1007/BF00773396

Stokes, S. L. (1976). An investigation of the consequences of ranked set sampling (Unpublished doctoral thesis). University of North Carolina, Chapel Hill, NC.

Journal of Modern Applied Statistical Methods
May 2015, Vol. 14, No. 1, 238-252.

ISSN 1538 − 9472

Special Education Distributions and Analysis

Valerie Felder
Department of Education
State of Michigan

Shlomo S. Sawilowsky
Wayne State University
Detroit, MI

Micceri (1989) examined the distributional characteristics of 440 large sample general education achievement and psychometric measures. All the distributions were found to be statistically significantly different from the normal distribution. In this study, 395 special education datasets were examined. Although there were some normally distributed datasets, most were not, and some were markedly different in shape from those found by Micceri (1989). Implications for statistical testing and making special education policy decisions were given.

Keywords: Nonnormal data sets, statistical testing, special education

Special education distributions

Micceri (1989) conducted an investigation of the distributional characteristics of 440 large sample educational achievement and psychometric measures. The data sets were obtained from general education and the behavioral and social sciences, including ability tests, achievement tests, criterion or mastery level tests, psychometric measures, and pre- and post-intervention scores. All were found to be non-normal based on the Kolmogorov-Smirnov test with nominal $\alpha = 0.01$. Factors that contributed to a non-Gaussian error distribution in the population include (a) subpopulations within a target population, (b) ceiling/floor effects, and (c) variability in the items within a measure. This has implications in terms of statistical testing, because classical parametric tests require normality in order to maintain acceptable robustness and comparative power properties (Sawilowsky & Blair, 1992). If ignored, costly errors may occur in making policy decisions.

The prevalence of non-normally distributed data permeates many fields. Previous studies that demonstrated this include Bradley (1977, 1982), Hill and

Dr. Valerie Felder is Research Consultant with the Michigan Department of Education. Email her at: felderv@michigan.gov. Dr. Sawilowsky is a Professor in the Department of Evaluation and Research, College of Education. Email him at shlomo@wayne.edu.

Dixon (1982), Ito (1980), Pearson and Please (1975) and Tan (1982). However, they, as well as Micceri (1989), did not have special education and disability assessments as a focus.

Assessment of students in special education is frequently different than for students in general education, because often the focus is on process or progress as opposed to specific learning outcomes. This may include adaptive behavior, development, and screening. Adaptive behavior skills are those skills that are useful in daily functioning. Developmental skills pertain to fine- and gross-motor, communication and language, social, cognitive, and self-help skills. Screening helps find children who might be below the norm in different areas (Rosenberg, Westling, & McLeskey, 2010).

Purpose of the study

Given the paucity of representation of special education data sets in the studies mentioned above, the purpose of this study is to canvass that literature to determine the distributional shape commonly encountered. This will help inform the appropriate statistical method (i.e., parametric or nonparametric) to be used in measuring the progress of students in special education.

Methodology

The distribution patterns of special education data sets were obtained from published, peer-reviewed journal articles from the years of 2007-2011. In addition, research studies that focused on special education assessment were considered for inclusion. A Google Scholar search with the key terms “special education” and “data” returned 396,397 related publications.

To construct a confidence level of 95% and margin of error of ±5%, a sample size of 384 data sets was needed from that population. It was estimated a return response rate of 25% was needed to accommodate lack of responses, and therefore 1,540 survey requests were made from selected authors of those published studies. Assessment data sets were also solicited from various state departments of education. Requests were made via email and telephone. The request included instructions to de-identify student information. Initial contact via email and phone was made from October - December, 2012. Follow-up phone calls and email messages were made in January, 2013.

Criteria for inclusion

Potential studies were reviewed to determine if the instrument used to collect data was supported by adequate reliability and validity information. However, there was no preset type or minimum reliability index or validity methodology required for inclusion.

Reliability is "the consistency that a test measures whatever it measures" (Sawilowsky, 2007, p. 516). As noted by Sawilowsky (2000), reliability is a psychometric property of a test. If the test produces similar results under consistent conditions then it is considered reliable. There were different types of reliability information obtained:

- Internal consistency, which is the extent items on an instrument relate to each other.
- Test-retest, which is the consistency over time (i.e., stability) of an instrument.
- Inter-rater reliability, which is the degree of agreement among raters.

Validity is "the degree that a test measures what it purports to measure (Sawilowsky, 2007, p. 166). There are different types of validity, including content-related validity, construct validity, and predictive validity (Cicchetti, 1994):

- Content-related validity, which is how well the content of the test relates to what is being assessed.
- Construct validity. "A construct is a fiction that is used to explain reality" (Cuzzocrea & Sawilowsky, 2009, p. 215), such as aptitude, intelligence, or self-determination. Hence, construct validity is the degree that a test measures that fiction used to explain reality.
- Predictive validity, which is the extent a test predicts some criterion measure.

Results

There were 744 authors contacted via email. Note that many authors had obtained multiple data sets in their study, exceeding the 1,540 data set requirement. Follow-up phone calls and emails were conducted where necessary after 3 months. There were *n* = 333 data sets collected from journal article authors, as compiled in

Table 1. In addition, academic achievement special education assessment test scores were requested from state education departments. Twenty four state departments of education, randomly selected, were contacted from which an additional $n = 62$ data sets were obtained from Alaska, Florida, Michigan, Minnesota, Missouri, and South Carolina, as compiled in Table 2. Thus, there were a total $N = 395$ data sets. Based on an estimated accessible population, the obtained sample size yielded a confidence level of 95% with a ±4.25% margin of error.

Table 1. Summary of Canvassed Authors (744) and Data Sets (4,362)

	Total	**Total % of Articles**
Acceptable Reliability	1760	40.30%
Acceptable Validity	1600	36.70%
Acceptable Articles*	1002	23.00%
Acceptable Data Sets	333	7.60%

*__Note__: An acceptable article required acceptable reliability and validity evidence.

Table 2. Data Sets from State Departments of Education

Florida	16	Minnesota	19
South Carolina	8	Alaska	15
Missouri	3	Michigan	1
		Total	**62**

Cronbach alpha coefficients for the instruments used to obtain these data sets ranged from .70 to .93. Test-retest reliability coefficients ranged from .65 to .97, and alternate-forms reliability ranged from .91 to .92. Concurrent validity indices ranged from .70 to .89, and predictive validity indices ranged from .65 to .86. (The author of one study used Item response theory (IRT) in a measurement model (i.e., Rasch one-parameter logistic (1PL) partial credit model for polytomous scoring).

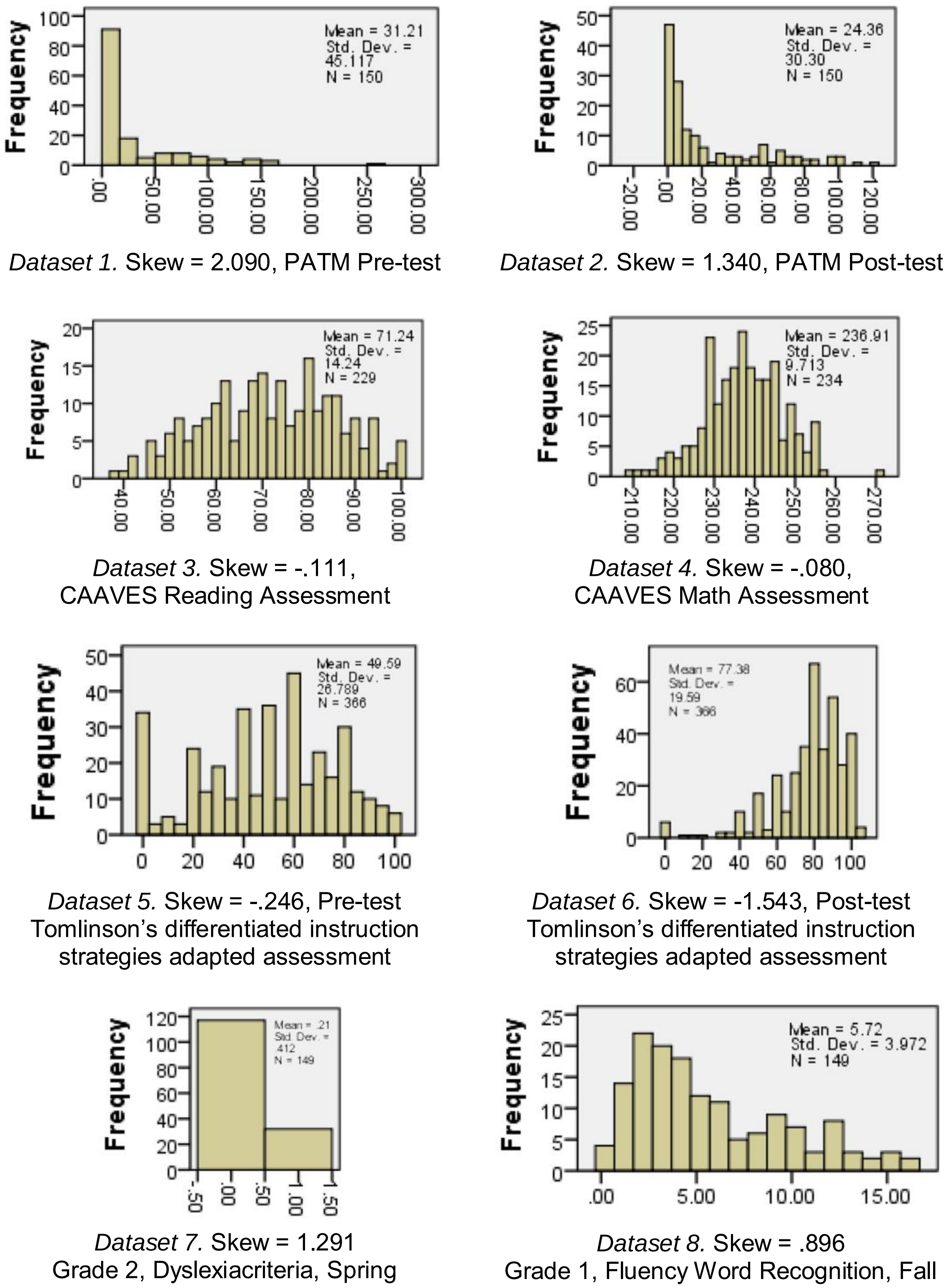
Distribution shapes

The histograms was analyzed and categorized. Histograms that resembled Micceri's (1986) distributions were named accordingly. Histograms that did not resemble Micceri's distributions were given a name based on the shape of each distribution. Figure 1 contains typical shapes obtained from the data sets. The types of distributions and the percentage of each distribution that were collected are indicated in Table 3. There were 258 (65.31%) special education data sets that were different and 137 (34.67%) similar to Micceri's (1989) shapes.

The data sets were also analyzed for normality and compared with Micceri's data sets. Based on the Kolmogorov-Smirnov and Shapiro-Wilks tests, 318 (81%) data sets were non-normally distributed and 77 (19%) data sets were normally distributed. Recall that Micceri (1986, 1989) found 100% of the distributions to be significantly non-normally distributed at the $\alpha = .01$ level. There were 19 out of 440 distributions, or 4.3%, that were considered reasonable approximations to the Gaussian distribution only in the sense that they were smooth symmetric with light tails. As compared with Micceri's (1986, 1989) results, this study shows special education assessment data sets were somewhat more likely to be normally distributed, but the number of different data sets shapes was higher than those found by Micceri (1986, 1989).

Table 3. Type, Number, and Percentage and Distribution Shapes

Type of Distribution	Number	Percentage
Extreme Bimodality	106	26.84%
Equimodal	96	24.30%
Unimodal and Smooth	79	20.00%
Bimodal and Smooth	31	7.85%
Slight Asymmetry	25	6.33%
Multimodal and Lumpy	19	4.81%
Unimodal and Slightly Smooth	10	2.53%
Extreme Asymmetry	6	1.52%
Slightly Asymmetric and Digit Preference	6	1.52%
Digit Preference	4	1.01%
Unimodal and Slightly Lumpy	4	1.01%
Equimodal and Symmetric	3	0.76%
Extreme Mass at Zero	2	0.51%
Mass at Zero	1	0.25%
Smooth Symmetric	1	0.25%
Equimodal and Slight Asymmetry	1	0.25%
Slightly Smooth and Symmetric	1	0.25%



Dataset 1. Skew = 2.090, PATM Pre-test

Dataset 2. Skew = 1.340, PATM Post-test

Dataset 3. Skew = -.111,
CAAVES Reading Assessment

Dataset 4. Skew = -.080,
CAAVES Math Assessment

Dataset 5. Skew = -.246, Pre-test
Tomlinson's differentiated instruction
strategies adapted assessment

Dataset 6. Skew = -1.543, Post-test
Tomlinson's differentiated instruction
strategies adapted assessment

Dataset 7. Skew = 1.291
Grade 2, Dyslexiacriteria, Spring

Dataset 8. Skew = .896
Grade 1, Fluency Word Recognition, Fall

Figure 1. Special Education Data Sets

Discussion

There were more classifications of special education data sets as extreme bimodality (*n* = 106, uni-modal, and smooth and equimodal than found in other disciplines. There were 106 extreme bimodality distributions and 57%, or 60 data sets, were non-normal. There were 46 distributions that were normal. There were 79 unimodal and smooth distributions and 29%, or 23 data sets, were non-normal. The remaining category, which had a large amount of distributions, is the equimodal category. There were 96 distributions and 70%, or 67, were non-normal. Thirty percent of the equimodal distributions were normally distributed based on the Kolmogorov-Smirnov and/or Shapiro-Wilks normality tests.

These data sets that were non-normally shaped pertained to curriculum-based assessments in writing, alternative assessments, applied problem solving, calculations, mathematics operations, reading, letter-word identification, segmenting words, and letter naming. Assessments of achievement, and fine- and gross-motor skills tended to be shaped normally.

In terms of policy, it is important to consider statistical robustness and comparative power when analyzing special education assessments. The results of this survey confirm the importance of considering nonparametric alternatives to parametric methods. As has been conducted throughout the Monte Carlo literature of the past century for data in many disciplines (e.g., general education, psychology, medicine, nursing), a study is warrant to determine the extent to which robustness and power of parametric tests may be compromised when analyzing special education data.

The new special education data shapes in this study may overlap with Micceri's (1989) data shapes. Due to the small sample size of the special education data sets, some of the shapes were different than Micceri's data shapes, but a larger sample sizes may show the data converges to one of Micceri's shapes.

For example, consider the data sets from the Florida Alternate Assessment. They were separated by grade level and a distribution was created for each data set, because the achievement of students in special education is measured based on a set of academic standards for each grade level. However, if the sample size is increased by combining a single grade with all grade levels, the resulting shape, identified by Micceri (1989) as a discrete mass at zero with gape, will result, as noted in Figure 2.

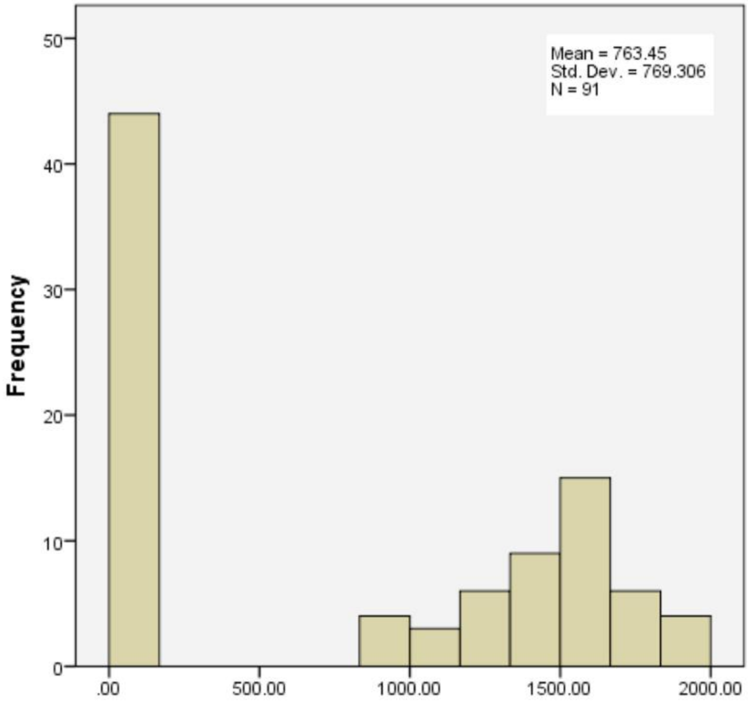


Figure 2. Concatenated Florida Alternate Assessment Special Education Data Set for All Grade Levels

In summary, Micceri's (1989) seminal article on 440 real data sets from general education achievement and psychometric constructs, shockingly, found them all to be non-normally distributed. This led to a major overhaul in techniques for analyzing quantitative data, as is known in the statistical literature, in those fields. Unfortunately, progress in revising and updating statistical strategies into other fields has been slow. Workers have the tendency to hold fast to techniques learned many years prior in graduate school, and furthermore, with the uptick in qualitative research, the lessons learned from Micceri (1989) obtain little voice until such surveys are replicated in their fields. On the basis of 395 special education data sets obtained in this study, differences from Micceri's (1989) rubric were noted, particularly the emergence of new non-normal distribution shapes. We believe this survey will help motivate quantitative workers in the special education field update their data analytic choices.

References

Aldridge, J. (2008). Narrowing the gaps for special-needs students. *Childhood Education, 84*(3), 182.

Barkley, R.A. (1997). *ADHD and the nature of self-control.* New York: Guilford Press.

Biancarosa, G. & Snow, C.E. (2004). *Reading next: A vision for action and research in middle and high school literacy. A report to the Carnegie Corporation of New York*. Washington, DC: Alliance for Excellent Education.

Bluman, A. (2007). *Elementary statistics. A step by step approach*. New York, NY: McGraw-Hill Higher Education.

Bradley, J. W., (1977). A common situation conducive to bizarre distribution shapes. *The American Statistician*, *31*(4), 147-150. doi:10.2307/2683535

Bradley, J. W. (1982). The-insidious L-shaped distribution. *Bulletin of the Psychonomic Society*, *20*, 85-88. doi:10.3758/BF03330089

Browder, D., Wakeman, S., & Flowers, C. (2006). Assessment of progress in the general curriculum for students with disabilities. *Theory Into Practice, 45*(3), 249-259.

Caffrey, E., Fuchs, D., & Fuchs, L. S. (2008). The predictive validity of dynamic assessment: A review. *The Journal of Special Education, 41*(4), 254-270. doi:10.1177/0022466907310366

Calhoon, M.B., Sandow, A., & Hunter, C. (2010). Reorganizing the instructional reading components: could there be a better way to design remedial reading programs to maximize middle school students with reading disabilities' response to treatment? *Annals of Dyslexia, 60,* 57-85. doi:10.1007/s11881-009-0033-x

Cicchetti, D. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment, 6*(4), 284-290. doi:10.1037/1040-3590.6.4.284

Clarke, B., Baker, S. & Smolkowski, K. (2008). An analysis of early numeracy curriculum-based measurement: Examining the role of growth in student outcomes. *Remedial and Special Education, 29*(1), 46-57. doi:10.1177/0741932507309694

Cuzzocrea, J., & Sawilowsky, S. S. (2009). Robustness to non-independence and power of the I test for trend in construct validity. *Journal of Modern Applied*

Statistical Methods, *8*(1), 215-225. Available at: http://digitalcommons.wayne.edu/jmasm/vol8/iss1/19

Eckes, S., Swando, J. (2009). Special education subgroups under NCLB: Issues to consider. *Teachers College Record, 111*(11), 2479-2504.

Elbaum, B. (2007). Effects of an oral testing accommodation on the mathematics performance of secondary students with and without learning disabilities. *The Journal of Special Education, 40*(4), 218-219. doi:10.1177/00224669070400040301

Fore, C., Boon, R. T., Burke, M. D. & Martin, C. (2009). Validating curriculum-based measurement for students with emotional and behavioral disorders in middle school. *Assessment for Effective Intervention, 34*(2), 67-73. doi:10.1177/1534508407313234

Fuchs, D., Mock, D., Morgan, P. L., & Young, C. L. (2003). Responsiveness-to-intervention: Definitions, evidence, and implications for the learning disabilities construct. *Learning Disabilities Research & Practice, 18*(3), 157-171. doi:10.1111/1540-5826.00072

Graham, S., & Harris, K. R. (1989). Components analysis of cognitive strategy instruction: Effects on learning disabled students' compositions and self-efficacy. *Journal of Educational Psychology, 81,* 353-361. doi:10.1037/0022-0663.81.3.353

Hardman, M. L., Drew, C. J., & Egan, M. W. (2002). *Human exceptionality: Society, school, and family.* Boston: Allyn & Bacon.

Heckaman, K., Conroy, M., Fox, J., & Chait, A. (2000). Functional assessment-based intervention research on students with or at risk for emotional and behavioral disorders in school settings. *Behavioral Disorders, 25*(3), 196-210.

Helwig, R. & Tindal, G. (2003). An experimental analysis of accommodation decisions on large-scale mathematics tests. *Council for Exceptional Children, 69*(2), 211-225. doi:10.1177/001440290306900206

Hill, M., & Dixon, W. J. (1982). Robustness in real life: A study of clinical laboratory data. *Biometrics*, *38*(2), 377-396. doi:10.2307/2530452

Hosp, J. L., Howell, K. W. & Hosp, M. K. (2003). Characteristics of behavior rating scales: Implications for practice in assessment and behavioral support. *Journal of Positive Behavior Interventions, 5*(4), 201. doi:10.1177/10983007030050040301

Hughes, C.A., Schumaker, J.B., & Deshler, D.D. (2005). *The essay test-taking strategy.* Lawrence, KS: Edge Enterprises, Inc.

Ito, P. K. (1980). Robustness of ANOVA and MANOVA test procedures. In P. R. Krishnaiah (Ed.), *Handbook of Statistics* (Vol. 6, p. 199-236). Amsterdam: North-Holland.

Jacobson, L. & Reid, R., (2010). Improving the persuasive essay writing of high school students with ADHD. *Exceptional Children, 76*(2), 157.

Katz, L. A., Stone, C. A., Carlisle, J. F., Corey, D. L. & Zeng, J. (2008). Initial progress of children identified with disabilities in Michigan's Reading First schools. *Exceptional Children, 74*(2), 235-256. doi:10.1177/001440290807400206

Kohl, F. L., McLaughlin, M. J., & Nagle, K. (2006). Alternate achievement standards and assessments: A description investigation of 16 states. *Exceptional Children, 73*(1), 107-123. doi:10.1177/001440290607300106

Kover, S. T. & Atwood, A. K. (2013). Establishing equivalence: Methodological progress in group-matching design and analysis. *American Journal on Intellectual and Developmental Disabilities, 118*(1), 3-15. doi:10.1352/1944-7558-118.1.3

Lane, K. L., Carter, E. W., Pierson, M. R. & Glaeser, B. C. (2006). Academic, social, and behavioral characteristics of high school students with emotional disturbances or learning disabilities. *Journal of Emotional and Behavioral Disorders, 14*(2), 108-117. doi:10.1177/10634266060140020101

Mayes, S. D., Calhoun, S. L., & Crowell, E. W. (2000). Learning disabilities and ADHD: Overlapping spectrum disorders. *Journal of Learning Disabilities, 33*(5), 417-424. doi:10.1177/002221940003300502

McConaughy, S., & Ritter, D. (2002). Best practices in multidimensional assessment of emotional or behavioral disorders. *Best practices in school psychology IV* (pp. 1303-1336). Bethesda, MD: National Association of School Psychologists.

McGinnis, E., Kiraly, J., & Smith, C. R. (1984). The types of data used in identifying public school students as behaviorally disordered. *Behavioral Disorders, 9*(4), 239-246.

Mertens, D. M. & McLaughlin, J. A. (2004). *Research and evaluation methods in special education*. (pp. 170-178). Thousand Oaks, CA: Sage Publication Ltd.

Mervis, C. B. & Klein-Tasman, B. P. (2004). Methodological issues in group-matching designs: Alpha levels for control variable comparisons and measurement characteristics of control and target variables. *Journal of Autism and*

Developmental Disorders, 34(1), 7-17. doi:10.1023/B:JADD.0000018069.69562.b8

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

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin, 105*(1), 156-166. doi:10.1037/0033-2909.105.1.156

Mosteller, F., & Tukey, J. (1977). *Data Analysis and Regression*, (pp. 55). Reading, MA: Addison-Wesley Publishing Company.

No Child Left Behind (NCLB) Act of 2001, Pub. L. No. 107-110, § 115, Stat. 1425 (2002).

Olson, L. (2000). Worries of a standards "backlash" grow. *Education Week, 30,* 1-13.

Pearson, E. S., & Please, N. W. (1975). Relation between the shape of population distribution and the robustness of four simple test statistics. *Biometrika*, *62*(2), 223-241. doi:10.1093/biomet/62.2.223

Rosenberg, M., Westling, D. & McLeskey, J. (2010). *Special education for today's teachers: An introduction*. Upper Saddle River, NJ: Pearson Education.

Runyon, R., Coleman, K & Pittenger, D. (2000). *Fundamentals of behavioral statistics*. New York, NY: McGraw-Hill Higher Education.

Salahu-Din, D., Persky, H., & Miller, J. (2008). *The nation's report card: Writing 2007* (NCES 2008-468). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education, Washington, DC.

Sawilowsky, S. S. (2000). Psychometric versus datametrics: Comment on Vacha-Haase's 'Reliability Generalization' method and some EPM editorial policies. *Educational and Psychological Measurement*, *60*(2), 157-173. doi:10.1177/00131640021970439

Sawilowsky, S. S. (2007). KR-20 and KR-21. In (N. J. Salkind, Ed.), *Encyclopedia of Measurement and Statistics*. Thousand Oaks, CA: Sage, p. 516-519.

Sawilowsky, S. S. & Blair, R. C. (1992). A more realistic look at the robustness and type II error properties of the t test departures from population normality. *Psychological Bulletin. 111*(2), 352-360. doi:10.1037/0033-2909.111.2.352

Sawilowsky, S. S., Blair, R.C., & Micceri, T. (1990). REALPOPS.LIB: a PC Fortran library of eight real distributions in psychology and education. *Psychometrika, 55*(4), 729.

Sawilowsky, S. S. & Fahoome, G.F. (2003). *Statistics through Monte Carlo simulation with Fortran.* Michigan: JMASM, Inc.

Silberglitt, B. & Hintze, J. M., (2007). How Much Growth Can We Expect? A Conditional Analysis of R-CBM Rates by Level of Performance. *Exceptional Children, 74*(1), 71. doi:10.1177/001440290707400104

Tan, W. Y. (1982). Sampling distributions and robustness of t, F and variance-ratio in two samples and ANOVA models with respect to departure from normality. *Communications in Statistics*. *A11*, 2485-2511.

Therrien, W. J., Hughes, C., Kapelski, C. & Mokhtari, K. (2009). Effectiveness of a test-taking strategy on achievement in essay tests for students with learning disabilities. *Journal of Learning Disabilities 42*(1), 14-23. doi:10.1177/0022219408326218

Thorndike, R., Hagen, E. (1986). *The Stanford-Binet intelligence scale, fourth edition: Guide for administering and scoring*. Chicago, IL: Riverside Publishing Co.

Tindal, G. & Fuchs, L.S. (1999). *A summary of research on test change: An empirical basis for defining accommodation.* Lexington: University of Kentucky, Mid-South Regional Resource Center.

Tomlinson, C. A. (1995). Deciding to differentiate instruction in the middle school: One school's journey. *Gifted Child Quarterly, 39*(2), 77-114. doi:10.1177/001698629503900204

Tukey, J.W. (1977). *Exploratory data analysis*. (pp. 63) Reading, MA: Addison-Wesley Publishing Company.

Wechsler, D. (1991). *The Wechsler intelligence scale for children*. San Antonio, TX: Psychological Corporation.

Woodcock, R., Mather, N., McGrew, K. (2001). *Woodcock-Johnson III Tests of Cognitive Abilities Examiner's Manual*. Itasca: Riverside.

Ysseldyke, J., Thurlow, M., Langenfield, K., Nelson, J.R., Teelucksing, E., & Seyfarth, A. (1998). *Educational results for students with disabilities: What do the data tell us?* (Tech. Rep. No. 23). Minneapolis: University of Minnesota, National Center of Educational Outcomes.

Appendix: Journals used in the survey

Journals marked with an "*" were used in the survey. The data is available from the first author of this study.

*American Annals of Deaf
*American Educational Research Journal
*American Journal on Intellectual and Developmental Disabilities
*Annals of Dyslexia
*Applied Measurement in Education
Australasian Journal of Special Education
Behavioral Disorders
British Journal of Special Education
Career Development for Exceptional Individuals
Child Development Perspectives
Developmental Psychology
Early Childhood Research Quarterly
Education and Training in Mental Retardation and Developmental Disabilities
*Education and Treatment of Children
Educational Assessment
*Educational and Psychological Measurement
*Elementary School Journal
*Exceptional Children
*Exceptionality: A Research Journal
International Journal of Disability
*Journal of Adolescent and Adult Literacy
*Journal of Applied Behavior Analysis
Journal of Applied Developmental Psychology
Journal of the Association for Persons with Severe Handicaps
Journal of Attention Disorders
*Journal of Autism and Developmental Disorders
Journal of Deaf Studies and Deaf Education
*Journal of Disability Policy Studies
*Journal of Early Intervention
Journal of Educational Psychology
Journal of Educational and Behavioral Statistics
Journal of Educational Measurement
*Journal of Emotional and Behavioral Disorders

Journal of Intellectual Disability Research

*Journal of the International Association of Special Education

*Journal of Learning Disabilities

Journal of Policy and Practice in Intellectual Disabilities

*Journal of Positive Behavior Interventions

*Journal of Psychoeducational Assessment

Journal of Research and Development in Education

*Journal of School Psychology

*Journal of Special Education

Journal of Speech and Hearing Research

*Journal of Visual Impairment and Blindness

*Learning and Individual Differences

*Learning Disability Quarterly

*Learning Disabilities Research and Practice

Mental Retardation

Peabody Journal of Education

*Preventing School Failure

*Psychology in the Schools

*Reading and Writing

Reading Psychology

Reading Research Quarterly

*Remedial and Special Education

Research in Developmental Disabilities

*Review of Educational Research

*School Psychology Quarterly

*School Psychology Review

Teachers College Record

Teaching Exceptional Children

*Volta Review

Journal of Modern Applied Statistical Methods
May 2015, Vol. 14, No. 1, 253-275.

ISSN 1538 − 9472

JMASM Statistical Software Applications and Review
SPSS Programs for Addressing Two Forms of Power for Multiple Regression Coefficients

Christopher Aberson
Humboldt State University
Arcata, California

This paper presents power analysis tools for multiple regression. The first takes input of correlations between variables and sample size and outputs power for multiple predictors. The second addresses power to detect significant effects for all of the predictors in the model. Both employ user-friendly SPSS Custom Dialogs.

Keywords: Power, sample size, simulation, SPSS, multiple regression, Power(all)

Introduction

Power analysis came to prominence with Jacob Cohen's seminal work on the topic (e.g., Cohen, 1988). Since that time, an extensive literature and several software packages and other resources focused on power (e.g., PASS, nQuery, Sample Power, G*Power, PiFace) emerged. Despite these advances, surveys across fields such as abnormal psychology (e.g., Sedlemeier & Gigerenzer, 1989), consulting, clinical, and social psychology (Rossi, 1990), and neuroscience (Button et al., 2013) suggest that low power remains common in published literature.

One explanation for the persistence of underpowered studies, suggested by Cohen is that "researchers find too complicated … reference material for power analysis (1992, p. 156)." The development of software approaches for power analysis allows researchers to move beyond some of the difficulties in understanding power analysis for many designs. With regard to power analyses for multiple regression designs, many approaches exist for estimating adequate

Christopher Aberson is Professor of Psychology. Email him at: cla18@humboldt.edu.

power for multiple R^2 (often termed R^2 model) based on considerations such as the number of predictors and sample size (see Algina & Olejnik, 2003; Dunlap, Xin, & Myers, 2004; Krishnamoorthy & Xia, 2008; Mendoza & Stafford, 2001; Murphy & Myors, 2004; Shieh & Kung, 2007).

Although many tools exist for power analyses focused on R^2 model, power analyses focused on multiple regression coefficients remains challenging. Existing resources for detecting power for coefficients are of limited utility, as most require input of complicated statistical values. For example, G*Power (Faul, Erdfelder, Buchner, & Lang, 2009) provides protocols to address power for an individual predictor. This approach is accurate but requires that users input either partial R^2 or its components. The partial R^2 is a function of the proportion of variance uniquely explained by the predictor (squared semi-partial correlation) and the variance explained in the dependent measure by the other predictors in the model. This value is not particularly intuitive, nor is it commonly provided by most commercial packages. Similarly, the PiFace regression applet (Lenth, 2006-9) also provides a complex approach that requires entry of the variance inflation factor (VIF) and several other values. The VIF is an index of overlap between predictors. Although common to most statistical packages, the VIF statistic, reflecting one divided by the residual variance from an analysis regressing the predictor of interest on the other predictors, is also not intuitive to most researchers. Additionally, both approaches require separate estimates for each predictor of interest. That is, to get accurate power estimates, users must repeat a complex set of calculations for each predictor. It is my impression that most researchers find it difficult to estimate values such as partial R^2 and VIF accurately for power analysis. These tools are well designed and accurate; however, the complexity of the required inputs limits their usability.

The estimates required by these protocols are "endpoint" values. Endpoint values are statistical values that require extensive computation for accurate estimation. Endpoint values such as the partial R^2 and VIF are a function of the correlation between the predictors and the dependent variable and the strength of correlations between the predictor of interest and other predictors in the model (i.e., a correlation matrix). Although partial R^2 and VIF are difficult to estimate, the zero-order correlations that produce these values are not. A researcher basing power analyses on previous work on the variables of interest is far more likely to find presentation of zero-order correlations between variables than VIF or partial R^2 statistics. For this reason, the protocols introduced in this paper focus on input of correlations as the primary statistical values for power analysis.

Another explanation for low power in designs with multiple predictors is a lack of attention to power for detecting a set of outcomes. Researchers using multiple regression models with three predictors commonly want to detect significant coefficients for all of the predictors. However, applications of power analyses for designs with multiple predictors typically yield an estimate of power for each predictor (e.g., Aberson, 2010), but not power to detect all of them in the same study. Problematically, power to detect multiple effects differs considerably from power for individual effects. In most research situations, power to detect multiple effects is considerably lower than the power for individual effects. The lack of attention to this form of power is a likely source underpowered research in the behavioral sciences (Maxwell, 2004).

The paper introduces tools to calculate simultaneous power estimates for two or more multiple regression coefficients (MRPower), power for detecting significant effects on all coefficients in a model (MRPower Simulate), and presents analyses using a series of SPSS Custom Dialogs based on the syntax found in Appendices A and B and available from http://users.humboldt.edu/chris.aberson/Index.html. All tools require entry of zero-order correlations with several additional optional values.

Equations for power calculations

Power for multiple regression coefficients is a function of the regression coefficient and its standard error with these values being a function of the correlations among variables in the model. The calculation of the standardized regression coefficient (Eq.1) involves both the correlations between the predictors (represented with numbers) and the criterion or dependent variable (represented with *y*). In this equation, r_{y1} is the correlation between the first predictor and the *dv*, r_{y2} is the correlation between the second predictor and the *dv*, and r_{12} is the correlation between the first predictor and the second predictor.

$$b^*_{y1.2} = \frac{r_{y1} - r_{y2}r_{12}}{1 - r_{12}^2} \qquad (1)$$

A simplified explanation of Equation 1 is that the coefficient is larger when correlations between the predictor and *DV* are large but becomes smaller when predictors correlate in the same direction as in the second predictor-*dv* relationship. In terms of the influence on power analysis, larger coefficients produce more power.

The standard error of the standardized regression coefficient (Eq. 2) is a function of the total variance explained by the two predictors in the analyses (often termed R^2 model, represented as $R^2_{y.12}$) and the squared correlation of the two predictors (r^2_{12}). The standard error is smaller when the variables explain more variance, when the correlation between predictors is smaller, and when sample size (n) is larger.

$$se_{b^*} = \sqrt{\frac{1-R^2_{y.12}}{\left(1-r^2_{12}\right)*\left(n-3\right)}} \tag{2}$$

Calculation of the standard error requires R^2 for a model with all the predictors (Eq. 3). This value increases as correlations between predictors and the *DV* increase and gets smaller as correlations between predictors rise, provided that correlations all run in the same direction.

$$R^2_{y.12} = \frac{r^2_{y1} + r^2_{y2} - 2r_{y1}r_{y2}r_2}{1-r^2_{12}} \tag{3}$$

The ratio of coefficient to standard error produces the non-centrality parameter (δ). Larger δ values represent more power. This value allows for calculation of power. Power calculations require application of non-central distribution probability density functions that are beyond the scope of simple calculations. However, SPSS and other packages provide the calculation (see next section for application).

$$\delta = \frac{b^*_{y1.2}}{se_{b^*}} \tag{4}$$

These formulae demonstrate several important concepts relevant to power analysis with multiple predictors. First, larger regression coefficients (i.e., larger effect size) promote more power. Larger coefficients result from stronger correlations between predictors and the *DV*. Correlation between predictors drives coefficient size downward and thus reduces power. Broadly this means that collinearity (or with three or more predictors, multicollinearity) reduces statistical power.

Power for two predictors

This section presents calculations of power for a two predictor example and then introduces the MRPower SPSS program to perform power calculations.

Calculation example

This example predicts voting intentions relevant to a hypothetical proposition to continue or discontinue affirmative action (on a scale where 0 = Absolutely will vote to eliminate to 10 = Absolutely will vote to continue) from beliefs that AA is fair and rejection of the merit principal. For the predictors, higher scores mean more fairness and stronger perceptions that merit should not be the only consideration in hiring. Based on earlier work, the example uses for $r_{y1} = .5$ (the correlation between fairness and intention), $r_{y2} = .4$ (the correlation between merit and intention), and $r_{12} = .3$ (the correlation between fairness and merit). The section that follows demonstrates calculation of power for a sample of $n = 50$.

$$R^2_{y.12} = \frac{r^2_{y1} + r^2_{y2} - 2r_{y1}r_{y2}r_{12}}{1 - r^2_{12}} = \frac{.5^2 + .4^2 - 2(.5)(.4)(.3)}{1 - .3^2} = .3187$$

$$b^*_{y1.2} = \frac{r_{y1} - r_{y2}r_{12}}{1 - r^2_{12}} = \frac{.5 - (.4 * .3)}{1 - .3^2} = .4176$$

$$b^*_{y2.1} = \frac{r_{y2} - r_{y1}r_{12}}{1 - r^2_{12}} = \frac{.4 - (.5 * .3)}{1 - .3^2} = .2747$$

$$se_{b^*} = \sqrt{\frac{1 - R^2_{y.12}}{(1 - r^2_{12}) * (n - 3)}} = \sqrt{\frac{1 - .3187}{(1 - .3^2)(50 - 3)}} = .1262$$

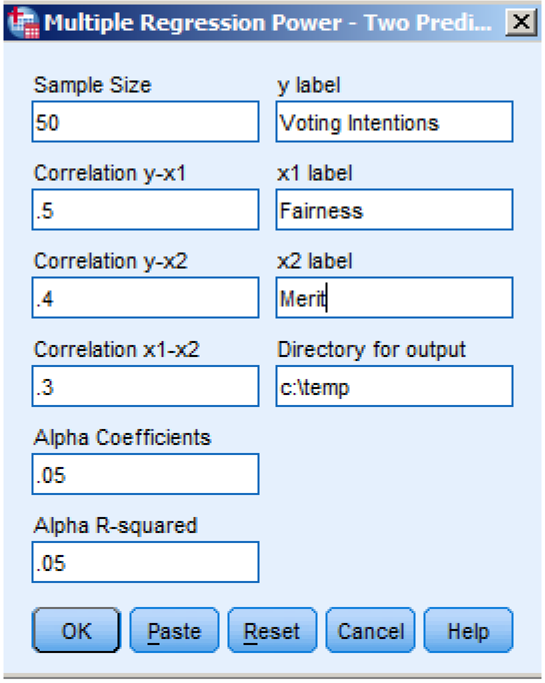
$$\delta_1 = \frac{b^*_{y1.2}}{se_{b^*}} = \frac{.4176}{.1262} = 3.309$$

$$\delta_2 = \frac{b^*_{y2.1}}{se_{b^*}} = \frac{.2747}{.1262} = 2.177$$

With alpha = .05, Power $x_1 = .90$ (fairness) and Power $x_2 = .57$ (merit). To obtain these values, provide SPSS with the following syntax for the first predictor: Compute Power = 1 - NCDF.T (2.012, 47, 3.309). The value 2.012 represents the critical value of t for rejection of the null, using two-tailed $\alpha = .05$. The value 47 represents degrees of freedom and 3.309 is δ.

Two predictor power using MRPower

The MRPower Two dialog provides a user-friendly interface that takes input of correlation values and sample size and returns power for each coefficient and R^2 model. The interface also allows users to enter labels for each variable, desired Type I error level for tests of the model and for coefficients, and the directory for files generated by the analyses. These values are optional. Figure 1 demonstrates entry of values into MRPower Two. Figure 2 presents the output from the dialog, yielding values consistent with calculations as well as an estimate for R^2 model power. The output provides power for all coefficients simultaneously. To obtain a desired level of power, increase sample size until reaching the target value. Power ≥ .80 for both coefficients requires a sample of 83, whereas Power ≥ .90 for both coefficients requires 110 participants.



Multiple Regression Power - Two Predi...

Sample Size: 50

y label: Voting Intentions

Correlation y-x1: .5

x1 label: Fairness

Correlation y-x2: .4

x2 label: Merit

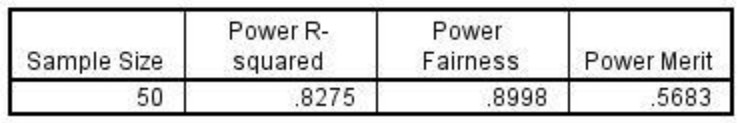
Correlation x1-x2: .3

Directory for output: c:\temp

Alpha Coefficients: .05

Alpha R-squared: .05

Figure 1. MRPower two interface demonstrating calculation of power for two individual predictors.



Sample Size	Power R-squared	Power Fairness	Power Merit
50	.8275	.8998	.5683

Figure 2. MRPower two output for the analysis specified in Figure 1

Models with three predictors

Calculations for two predictor models are relatively straightforward. Models with three or more predictors require approaches that are substantially more complex. For three or more predictions, calculations involve matrix inversion and other approaches that likely go beyond the backgrounds of most researchers (see Cohen, Cohen, West, & Aiken, 2003 for calculator approaches). The syntax and custom dialogs presented in this paper provide researchers with tools to obtain power estimates for multiple regression designs with three variables through a simple extension of the approach employed in the two predictor section. Although not demonstrated in this paper, dialogs for four through ten predictors (named MRPower Four, MRPower Five, etc.) are in development.

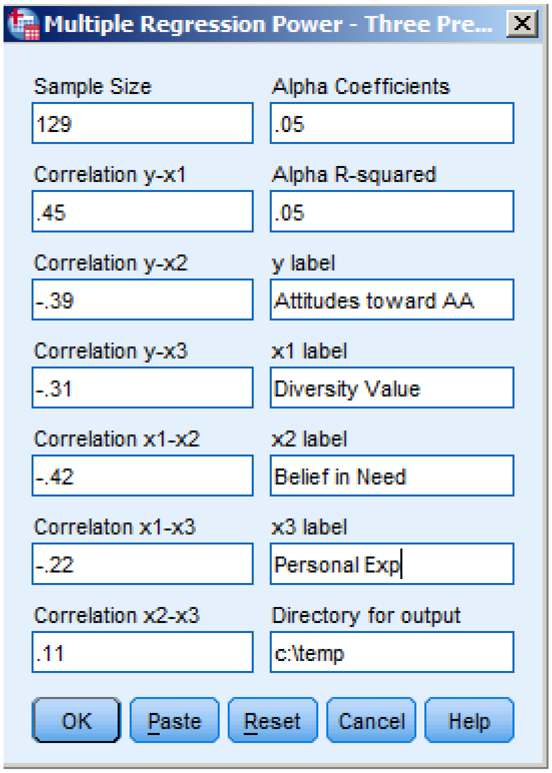
Three predictors with MRPower

The example that follows demonstrates use of MRPower to determine adequate sample size. This example takes results from Aberson (2007) and uses those values to determine power for a new study involving three predictors of general attitudes toward affirmative action. The predictors are diversity valuation, belief in the need for affirmative action, and personal experiences of discrimination with their expected population correlations shown in Table 1.

Figure 3 demonstrates the MRPower Three interface. In this example, to obtain power of .80 or greater for each predictor requires a sample size of 129. Specifically, as shown in Figure 4, the analysis reports power of .94 for diversity, .82 for belief in need, and .80 for experience of discrimination.

Table 1. Correlations between variables in three predictor example.

	General Policy	Diversity	Belief in Need
General			
Diversity	.45 (r_{y1})		
Belief in Need	-.39 (r_{y2})	-.42 (r_{12})	
Exp of Disc	-.31 (r_{y3})	-.22 (r_{13})	.11 (r_{23})



Multiple Regression Power - Three Pre...

Sample Size: 129

Alpha Coefficients: .05

Correlation y-x1: .45

Alpha R-squared: .05

Correlation y-x2: -.39

y label: Attitudes toward AA

Correlation y-x3: -.31

x1 label: Diversity Value

Correlation x1-x2: -.42

x2 label: Belief in Need

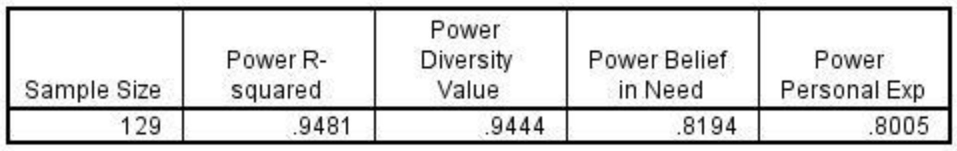
Correlaton x1-x3: -.22

x3 label: Personal Exp

Correlation x2-x3: .11

Directory for output: c:\temp

Figure 3. MRPower Three interface demonstrating calculation of individual power for three predictors.



Sample Size	Power R-squared	Power Diversity Value	Power Belief in Need	Power Personal Exp
129	.9481	.9444	.8194	.8005

Figure 4. MRPower Three output for the analysis specified in Figure 3.

Power for detecting significant effects for all predictors in the model

Often researchers using multiple regression want to detect significant effects for all of the predictors in a model. However, existing power analysis approaches only address power for individual predictors. This section details how power to detect effects for all of the predictors in a model differs from power to detect individual effects and present tools for addressing this form of power. The primary issue relevant to detecting significant effects for multiple predictor variables is the role of Beta error inflation (or Familywise Beta error; see Maxwell, 2004 for a technical discussion). This issue is similar to inflation of α or Type I error. When conducting multiple significance tests, Type I error rates for the family of tests (a.k.a., familywise alpha) increase. Equation 5 provides an estimate of familywise α error for multiple comparisons and is the conceptual basis for development of tests such as the Bonferroni adjustment. According to the formula, with three tests using a pairwise alpha (α_{pw}) of .05, familywise alpha (α_{fw}) is .14.

$$\alpha_{fw} = 1 - \left(1 - \alpha_{pw}\right)^{c} \tag{5}$$

The same process is at work with regard to the familywise probability of making a β or Type II error (Equation 6), a value referred as β_{fw} throughout the paper. For example, take a study designed for β of .20 (called β_{ind} for Beta individual) for each of its three predictors (a.k.a., Power = .80 for each predictor). The likelihood of making a single β error among those three tests is substantially higher than the error rate of .20 for the individual tests. Just as with α error, multiple tests inflate the chances to make a single β error among a set of significance tests. The β_{fw} value easily converts to power to detect all of the effects in the design by taking 1 - β_{fw}. Throughout the paper, this value is referred to as Power(All).

$$\beta_{fw} = 1 - \left(1 - \beta_{ind}\right)^{c} \tag{6}$$

Table 2. Familywise Type II error (Beta) rates for predictors using β_{pw} = .20 (Power = .80)

Number of Predictors	β_{fw}	Power(All)
2	.360	.640
3	.488	.512
4	.590	.410
5	.672	.328
6	.738	.262
7	.790	.210
8	.832	.168
9	.866	.134
10	.893	.107

* Note. All predictors uncorrelated. This table is not accurate for correlated predictors.

Table 2 shows β_{fw} and Power(All) for two through 10 predictors. One clear result here is that in models with four predictors or more, if the researcher designs for Power = .80 for each individual predictor, the study will more likely than not fail to find significance on at least one of the predictors. This table is useful for a conceptual understanding of β_{fw}, however these results (and Eq. 6) are only accurate for calculations where all tests have the same power and predictors are uncorrelated.

Power(All) for designs with correlated predictors

Calculation of β_{pw} and Power(All) is straightforward for situations where predictors are uncorrelated. However, in most multiple regression applications predictors do correlate. How this influences Power(All) is a function of the strength and direction of correlations between predictors. Broadly, when predictors correlate positively with each other, Power(All) decreases. If predictors negatively correlate, Power(All) increases.

Calculations of Power(All) given correlated predictors are best handled by simulation. Simulations draw a large number of independent samples (e.g., 10,000) from a population with parameters used in the power analysis (defined by a correlation matrix). From those samples, count how many allow rejection of null hypotheses relevant to all of the predictors in the study. The proportion of samples producing results allowing for rejection of all hypotheses reflects Power(All).

Table 3 demonstrates the impact of predictor correlations on Power(All) for a two predictor model. Power for each predictor is constant across each situation at .80 (the correlation between the predictors and *DV* changes to create this level of power) and the sample size is 50. The Reject All column reflects Power(All) estimates derived by simulation of 10,000 samples drawn from a population with the given correlations. Since this approach is empirical, there is some deviation from theoretical probabilities. For example, Power(All) for two predictors with Power = .80 and no correlation between predictors is theoretically .64. The simulation provides a value of .6348. Although not exact with 10,000 replications, the simulated values provide a clear demonstration of the patterns of expected results. The range of values for Power(All) is roughly .59 to .72 with more power generated as correlations between predictors move from strongly positive to strongly negative.

These values suggest that negative correlations between predictors are advantageous. However, is important to recognize that it is unlikely to find predictors that correlate strongly in the negative direction when both predictors have a consistent (i.e., all positive or all negative) relationships with the *DV*.

Table 3. Power(All) for two predictors with power = .80 and varying levels of correlation.

Correlation between predictors	Required x-y correlations	Reject None	Reject One	Reject All
-.80	.1274	.1294	.1492	.7214
-.60	.1891	.1074	.2029	.6897
-.40	.2445	.0816	.2458	.6726
-.20	.2999	.0564	.2912	.6524
.00	.3594	.0463	.3189	.6348
.20	.4266	.0279	.3518	.6203
.40	.5070	.0190	.3708	.6102
.60	.6102	.0102	.3864	.6034
.80	.7561	.0033	.4107	.5860

* Note. Required *x-y* correlation is the correlation between each predictor and the *dv* to produce Power = .80 with n = 50.

Table 4 demonstrates Power(All) for models with three predictors. In each situation, Power = .80 for each predictor and the sample size is 100. One striking finding here is that Power(All) can be as low as .44 for a model with strongly correlated predictors, despite the relatively high level of power for individual predictors. As with the two predictor model, Power(All) rises as correlations among predictors move from positive to negative. However, Power(All) tends to be smaller with more predictors. For two predictors, Power(All) ranges from .59 to .72 whereas with three predictors, Power(All) goes from .44 to .64.

Table 4. Power(All) for three predictors with power = .80 and varying levels of correlation.

Correlation between predictors	Required *x-y* correlations	Reject None	Reject One	Reject Two	Reject All
-.80	n/a				
-.60	n/a				
-.40	.0804	.0793	.1030	.1800	.6377
-.20	.1692	.0268	.1129	.3046	.5557
.00	.2583	.0091	.1005	.3678	.5226
.20	.3569	.0033	.0892	.4251	.4824
.4	.4703	.0008	.0678	.4681	.4633
.6	.6057	.0001	.0506	.5000	.4493
.8	.7747	.0000	.0435	.5211	.4354

* Note. Required *x-y* correlation is the correlation between each predictor and the *dv* to produce Power = .80 with *n* = 100.

Also of note is that some values in Table 4, represented as n/a, are not possible. For example, there is no predictor-*DV* correlation where it is possible to have correlations of -.60 or -.80 between the predictors (given *n* = 100). Additionally, models with substantial positive correlations among multiple predictors likely violate regression assumptions regarding multicollinearity.

MRPower Simulate dialogs

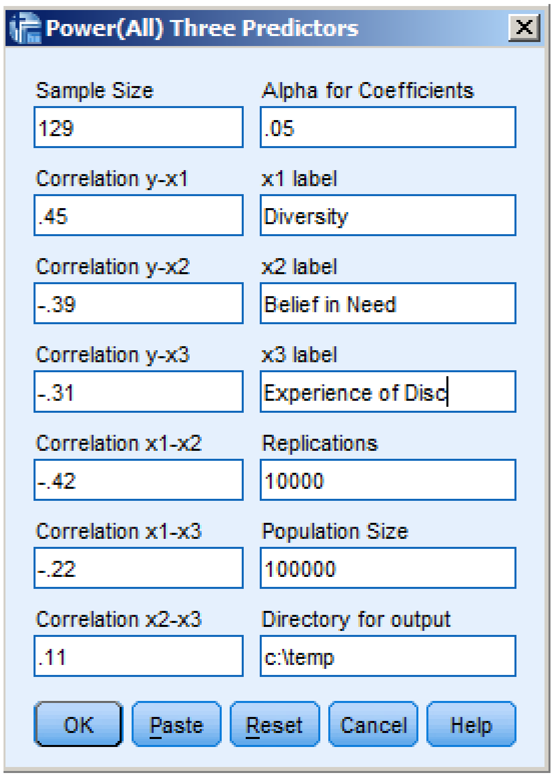
The previous section demonstrated how correlations between predictors impact Power(All). However, the values presented in those tables are limited as they reflect situations wherein correlations between predictors and power for individual predictors are constant. Practically predictors might show different levels of power and varying levels of correlation. The MRPower Simulate dialogs allow for such input and address Power(All) for designs with two to ten predictors.

In the example from the previous section, power exceeded .80 for three predictors with a sample of 129. However, power for detecting significant effects for all three predictors in the same sample [termed Power(All)] is likely substantially smaller. The MRPower Simulate dialog creates a population based on user-supplied correlations. Next, the program takes a sample of size *n* from the population (*n* is specified by the user) and generates an analysis predicting the *DV* from the set of IVs for that sample. The results of the analysis are output to a datafile (stored in the directory c:\temp as a default). The program repeats this process 10,000 times. Finally, the program compiles rejection rates and provides output representing power for individual coefficients (total times rejecting null divided by total number of replications) and power for rejecting zero to all coefficients.

The number of replications and population size are modifiable. Although population is theoretically infinite, a finite population of 100,000 is, for most purposes, large enough to produce an accurate result. In testing the dialog, there was little difference between the default settings and simulations using larger populations (e.g., 10 million) and more replications (e.g., 100,000). However, more replications substantially increased processing time. If sample sizes begin to approach even a small percentage of population size, it would likely be beneficial to increase the population size. For quick analyses (e.g., trying to determine whether the sample size for Power(All) = .80 is closer to 300 than 400), replications might be reduced initially then increased in subsequent runs for a precise result.

MRPower Simulate example.

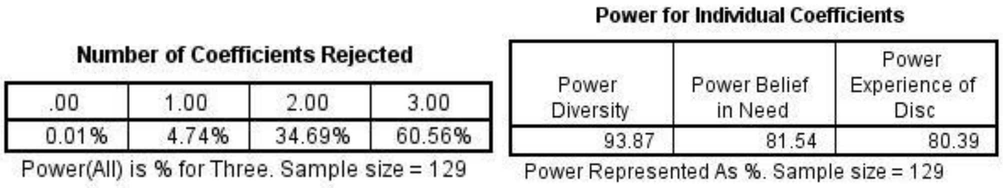
Figure 5 demonstrates the MRPower Simulate dialog using a sample of 129 and the correlations from Table 1. As shown in Figure 6, this analysis generates Power(All) = .6056 to detect all three effects in the same model. The output also indicates the number of samples rejecting null hypotheses for zero, one, or two coefficients. On a positive note, the likelihood of finding no significant effects is .0001.


Power(All) Three Predictors

Sample Size: 129
Alpha for Coefficients: .05
Correlation y-x1: .45
x1 label: Diversity
Correlation y-x2: -.39
x2 label: Belief in Need
Correlation y-x3: -.31
x3 label: Experience of Disc
Correlation x1-x2: -.42
Replications: 10000
Correlation x1-x3: -.22
Population Size: 100000
Correlation x2-x3: .11
Directory for output: c:\temp

Figure 5. MRPower Simulate three interface for calculation of Power(All).

Figure 6 also presents power for each individual predictor. This value is the number of times rejecting the null for the predictor over total number of replications. These values provide a useful check against the results of the MRPower Three dialog. In this case, power for Diversity (.9387 vs. .9444), Power for Belief in Need (.8154 vs. .8194), and Power for Personal Experience (.8039 vs. .8005) are all consistent with the MRPower estimates. If these values are not consistent, it suggests incorrect specification of the parameters of the model (i.e., something not entered correctly in the dialog).



Number of Coefficients Rejected

.00	1.00	2.00	3.00
0.01%	4.74%	34.69%	60.56%

Power(All) is % for Three. Sample size = 129

Power for Individual Coefficients

Power Diversity	Power Belief in Need	Power Experience of Disc
93.87	81.54	80.39

Power Represented As %. Sample size = 129

Figure 6. MRPower Simulate three output for Power(All) and individual predictors given specification from Figure 5.

A final question is how large a sample is necessary for Power(All) of a specific value (e.g., .80). Using the simulation tool, Power(All) hits .80 with $n = 171$. For $n = 171$, power for the individual predictors are .98, .91, and .90 respectively. This represents an increase of roughly one-third of the original sample size estimate.

References

Aberson, C. L. (2007). Diversity, merit, fairness, and discrimination beliefs as predictors of support for affirmative action policy actions. *Journal of Applied Social Psychology, 37*(10), 2451-2474. doi:10.1111/j.1559-1816.2007.00266.x

Aberson, C. L. (2010). *Applied power analysis for the behavioral sciences*. New York: Psychology Press.

Algina, J., & Olejnik, S. (2003). Sample size tables for correlation analysis with applications in partial correlation and multiple regression analysis. *Multivariate Behavioral Research, 38*(3), 309-323. doi:10.1207/S15327906MBR3803_02

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience, 14*, 365-376. doi:10.1038/nrn3475

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

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155-159. doi:10.1037/0033-2909.112.1.155

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). Applied multiple regression/Correlation analysis for the behavioral sciences (3rd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Dunlap, W. P., Xin, X., & Myers, L. (2004). Computing aspects of power for multiple regression. *Behavior Research Methods, Instruments & Computers, 36*(4), 695-701. doi:10.3758/BF03206551

Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41*(4), 1149-1160. doi:10.3758/BRM.41.4.1149

Krishnamoorthy, K, & Xia, Y. (2008). Sample size calculation for estimating or testing a nonzero squared multiple correlation coefficient. *Multivariate Behavioral Research, 43*(3), 382-410. doi:10.1080/00273170802285727

Lenth, R. V. (2006-9). *Java applets for power and sample size* [Computer software]. Retrieved from http://www.stat.uiowa.edu/~rlenth/Power.

Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: Causes, consequences, and remedies. *Psychological Methods, 9*(2), 147-163. doi:10.1037/1082-989X.9.2.147

Mendoza, J. L., & Stafford, K. L. (2001). Confidence interval, power calculation, and sample size estimation for the squared multiple correlation coefficient under the fixed and random regression models: A computer program and useful standard tables. *Educational and Psychological Measurement, 61*(4), 650-667. doi:10.1177/00131640121971419

Murphy, K. R., & Myors, B. (2004). *Statistical power analysis: A simple and general model for traditional and modern hypothesis tests (2nd ed.)*. Hillsdale, NJ: Laurence Erlbaum Associates.

Rossi, J. S. (1990). Statistical power of psychological research: What have we gained in 20 years? *Journal of Consulting and Clinical Psychology, 58*(5), 646-656. doi:10.1037/0022-006X.58.5.646

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin, 105*(2), 309-316. doi:10.1037/0033-2909.105.2.309

Shieh, G., & Kung, C. F. (2007). Methodological and computational considerations for multiple correlation analysis. *Behavior Research Methods, 39*(4), 731-734. doi:10.3758/BF03192963

Appendix A

MRPower Three Syntax

```
*Values noted with %% are user supplied values from the dialog. For example if n
= 60 is entered *in the dialog, the %%n%% is replaced by 60 for analyses.

*OMS command suppresses output
OMS SELECT ALL
 /DESTINATION VIEWER=NO.

*Creates correlation matrix for analysis
*Means are set at 1, 2, 3, and 4 to facilitate SPSS processing.
*0s sometimes cause SPSS to terminate
MATRIX DATA VARIABLES = ROWTYPE_  y x1 x2 x3.
BEGIN DATA
Mean 1 2 3 4
STDEV 1 1 1 1
N %%n%% %%n%% %%n%%  %%n%%
Corr 1
Corr %%ry1%% 1
Corr %%ry2%% %%r12%%  1
Corr %%ry3%% %%r13%%  %%r23%% 1
END DATA.
DATASET CLOSE %%dir%%\resultsC.sav.

*Captures coefficient and R² values for power calculations
OMS SELECT TABLES
 /destination format = sav numbered = "Table_Number" outfile =
"%%dir%%\resultsC.sav"
 /if commands = ['regression'] subtypes = ['Coefficients']
 /tag = "reg".
OMS SELECT TABLES
 /destination format = sav numbered = "Table_Number" outfile =
"%%dir%%\resultsC.sav"
 /if commands = ['regression'] subtypes = ['ANOVA']
 /tag = "regF".
```

```
*Runs regression to obtain non-centrality parameter values (equivalent to F and
t)
REGRESSION
  /MATRIX=IN(*)
  /STATISTICS COEFF OUTS R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT y
  /METHOD=ENTER x1 x2 x3 .
OMSEND.
OMS SELECT ALL
 /DESTINATION VIEWER=NO.
GET FILE "%%dir%%\resultsC.sav".

*Extracts test statistic
FILTER OFF.
USE ALL.
SELECT IF ( ~ NMISS(Sig)).
EXECUTE.
IF (nmiss(t)) lambdaF=F.
IF  (nmiss(F)) lambdaC=t*t.
EXECUTE.

*Computer power from non-centrality parameter, df, and alpha
COMPUTE pred = 3.
COMPUTE dfe=%%n%%-pred-1.
COMPUTE sample = %%n%%.
COMPUTE F_critM = IDF.F(1-%%alphaR%%,pred, dfe) .
COMPUTE F_critC = IDF.F(1-%%alphaC%%,1, dfe) .
COMPUTE PowerF = 1-NCDF.F(F_critM,pred, dfe, lambdaF) .
COMPUTE PowerC = 1-NCDF.F(F_critC,1, dfe, lambdaC) .
If (nmiss(lambdaC)) Power = PowerF.
If (nmiss(lambdaF)) Power = PowerC.
COMPUTE ID=$CASENUM.
EXECUTE.
If (ID = 3) PowerX1=PowerC.
If (ID = 4) PowerX2=PowerC.
If (ID = 5) PowerX3=PowerC.
```

```
EXECUTE .
OMSEND.

*Creates output for power analysis
CTABLES
  /VLABELS VARIABLES=sample PowerF PowerX1 PowerX2 PowerX3 DISPLAY=NONE
  /TABLE BY sample [MAXIMUM 'Sample Size' F40.0] + PowerF [S][MAXIMUM 'Power R-
squared' F40.4] + PowerX1 [S][MAXIMUM 'Power %%x1lab%%' F40.4]
    + PowerX2 [S][MAXIMUM 'Power %%x2lab%%' F40.4] + PowerX3 [S][MAXIMUM
'Power %%x3lab%%' F40.4].

*Deletes files created to run analysis
OMS SELECT ALL
  /DESTINATION VIEWER=NO.
NEW FILE.
ERASE FILE ='%%dir%%\resultsC.sav'.
OMSEND.
```

Appendix B

MRPower Simulate Three Syntax

```
*Values noted with %% are user supplied values from the dialog.
*This command suppresses output
OMS SELECT ALL
/DESTINATION VIEWER=NO.
*The data generation approach used here modifies syntax presented in an IBM SPSS
support
*file at http://www-01.ibm.com/support/docview.wss?uid=swg21480900 . Based on
personal *correspondence and references to edstat-l archives, I believe this
approach was developed *by David Nichols.
matrix data variables=v1 to v4
/contents=corr.
begin data.
1
%%ry1%% 1
%%ry2%%  %%r12%%  1
%%ry3%%  %%r13%%  %%r23%%  1
end data.
save outfile='%%dir%%\corrmat.sav'
/keep=v1 to v4.

*Generate raw data. Loop # generates desired population size.
*Vector x() and #j reflect number of variables (1 dv, 3 predictors in this
example)
new file.
input program.
loop #i=1 to %%popsize%%.
vector x(4).
loop #j=1 to 4.
compute x(#j)=rv.normal(0,1).
end loop.
end case.
end loop.
end file.
end input program.
```

```
execute.

*FACTOR procedure generates principal components, which will be uncorrelated and
have *mean 0 and standard deviation 1 for each variable.
factor var=x1 to x4
/criteria=factors(4)
/save=reg(all z).

matrix.
get z /var=z1 to z4.
get r /file='%%dir%%\corrmat.sav'.
compute out=z*chol(r).
save out /outfile='%%dir%%\giant_datafile.sav'.
end matrix.

*End data generation portion
*Gets the generated data and test correlations.
get file='%%dir%%\giant_datafile.sav'.

*Rename variables
RENAME variables col1 = y.
RENAME variables (col2 to col4=x1 to x3).
COMPUTE ID=$CASENUM .
SAVE OUTFILE='%%dir%%\giant_datafile.sav'
  /COMPRESSED.

*This piece draws random samples of size n. Creates number of samples equal to
reps.
*Puts everything in one file then splits it by sample number
INPUT PROGRAM .
LOOP SAMP=1 to %%reps%%.
LOOP V = 1 to %%n%%.
COMPUTE ID=TRUNC(UNIFORM(%%popsize%%)) + 1.
END CASE.
LEAVE SAMP.
END LOOP.
END LOOP.
END FILE.
END INPUT PROGRAM .
SORT CASES BY ID .
MATCH FILES / FILE * / TABLE  '%%dir%%\giant_datafile.SAV' / BY ID .
SORT CASES BY SAMP.
SPLIT FILE BY SAMP.
DATASET CLOSE %%dir%%\boot1.sav.
*Runs regression on each sample. Outfile command saves results in datafile
called boot1.sav

REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT y
  /METHOD=ENTER x1 x2 x3
  /OUTFILE=COVB('%%dir%%\boot1.sav').

USE ALL.
GET
  FILE='%%dir%%\boot1.sav'.
DATASET NAME boot1 WINDOW=FRONT.

**Takes the information saved in the outfile and does some analyses based on the
sig of each test
**After that, just count up how many results were significant out of 10,000 -
that's the power
USE ALL.
COMPUTE filter_$=(ROWTYPE_="SIG").
VARIABLE LABEL filter_$ 'ROWTYPE_="SIG" (FILTER)'.
VALUE LABELS filter_$  0 'Not Selected' 1 'Selected'.
FORMAT filter_$ (f1.0).
FILTER BY filter_$.
EXECUTE .
COMPUTE Sig_Coeff1 = 0 .
EXECUTE .
IF (x1<%%alpha%%) Sig_Coeff1 = 1 .
EXECUTE .
```

```
COMPUTE Sig_Coeff2 = 0 .
EXECUTE .
IF (x2<%%alpha%%) Sig_Coeff2 = 1 .
EXECUTE .
COMPUTE Sig_Coeff3 = 0 .
EXECUTE .
IF (x3<%%alpha%%) Sig_Coeff3 = 1 .
EXECUTE .
COMPUTE Total_reject=Sig_Coeff1 + Sig_Coeff2 + Sig_Coeff3.
EXECUTE.
COMPUTE b1pct=(Sig_Coeff1 / %%reps%%)*100.
COMPUTE b2pct=(Sig_Coeff2 / %%reps%%)*100.
COMPUTE b3pct=(Sig_Coeff3 / %%reps%%)*100.
VARIABLE LEVEL b1pct b2pct b3pct(SCALE).
EXECUTE.

OMSEND.

*Custom Tables to produce individual power and Power(All)
CTABLES
  /VLABELS VARIABLES=b1pct b2pct b3pct DISPLAY=NONE
  /TABLE BY b1pct [SUM 'Power %%x1lab%%' F40.2] + b2pct [SUM 'Power %%x2lab%%'
F40.2] + b3pct [SUM 'Power %%x3lab%%' F40.2]
  /TITLES
    TITLE='Power for Individual Coefficients'
    CAPTION='Power Represented As %. Sample size = %%n%%'.
CTABLES
  /VLABELS VARIABLES=Total_reject DISPLAY=NONE
  /TABLE BY  Total_reject [C][ROWPCT.COUNT PCT40.2]
  /SLABELS VISIBLE=NO
  /CATEGORIES VARIABLES=Total_reject ORDER=A KEY=VALUE
    EMPTY=EXCLUDE
  /TITLES
    TITLE='Number of Coefficients Rejected'
    CAPTION='Power(All) is % for Three. Sample size = %%n%%'.

*Delete all files created.
OMS SELECT ALL
```

```
/DESTINATION VIEWER=NO.
New File.
DATASET CLOSE boot1.
Erase File='%%dir%%\corrmat.sav'.
Erase File='%%dir%%\giant_datafile.sav'.
Erase File='%%dir%%\boot1.sav'.
Omsend.
```

Journal of Modern Applied Statistical Methods
May 2015, Vol. 14, No. 1, 276-307.

ISSN 1538 − 9472

JMASM Algorithms and Code Algorithms for Assessing Intervention Effects in Single-Case Studies

Chao-Ying Joanne Peng
Indiana University at Bloomington
Bloomington, IN

Li-Ting Chen
Indiana University at Bloomington
Bloomington, IN

Free web-based resources or popular software to assess six data features recommended by the *What Works Clearinghouse: Procedures and Standards Handbook* (IES, 2013 February) to determine intervention effects in a single-case study (Lambert, Cartledge, Heward, & Lo, 2006) are demonstrated. Lambert et al. (2006) employed a reversal (or ABAB) design and visual inspection to investigate the effectiveness of the report-card treatment in reducing disruptive behaviors in students. In our demonstration, we assessed each of the six data features separately; then integrated six assessments into one comprehensive analysis of the intervention effect. A simple approach to the determination of intervention effects illustrates how researchers and practitioners can be empowered to interpret data comprehensively and formulate evidence-based conclusions logically from well-designed and well-executed single-case studies.

Keywords: algorithm, intervention effect, single-case studies, level, trend, variability, immediacy, overlap, effect size, Spearman rank correlation, Page test, confidence interval

Introduction

Horner et al. (2005) defined a single-case design (SCD) as a “rigorous, scientific methodology used to define basic principles of behavior and establish evidence-based practice.” (p. 165). SCDs are particularly important to clinical studies in which detailed information about aspects of a few participants’ behavior is gathered over an extended period of time in order to determine effects of an intervention. Yet determining intervention effects in SCD studies presents unique challenges due to the small sample size, the correlated nature of outcome measures, and the difficulty of applying statistical methods to SCD data. Visual inspection has been traditionally used by researchers and practitioners to assess an

Dr. Peng is an Adjunct Professor of Statistics. Email her at peng@indiana.edu. Dr. Chen is a Research Associate. Email her at litchen@indiana.edu.

intervention effect. Indeed, according to the Institute of Education Sciences' publication, *What Works Clearinghouse: Procedures and Standards Handbook* (IES, 2013 February, hereafter abbreviated as the *WWC Handbook*), "Single-case researchers traditionally have relied on visual analysis of the data to determine (a) whether evidence of a relation between an independent variable and an outcome variable exists, and (b) the strength or magnitude of that relation." (p. E.5).

The subjectivity associated with visual analysis and its lack of a theoretical framework for testing a scientific hypothesis have hampered the generalizability of SCD findings. The *WWC Handbook* actually recommends the examination of six data features both within and between phases in order to determine the effectiveness of an intervention effect. The six data features include: level/level change, trend, variability, immediacy of the effect, overlap, and consistency of data in similar phases. These six features should be assessed collectively to determine if (1) the observed pattern of data in the intervention phase is indeed due to the intervention effects and (2) the observed pattern of data in the intervention phase is different from the predicted pattern of data, predicated from data collected in the baseline phase. The *WWC Handbook* further recommends that a measure of the strength of the relation between an independent variable and an outcome be computed and reported to accompany the assessment of that relation.

Given the importance of the WWC's initiative "to be a central and trusted source of scientific evidence for what works in education." (IES, 2013 February, p. 1) and the intended purpose of the *WWC Handbook* to provide "a detailed description of the standards and procedures of the WWC" (IES, 2013 February, p. 2), it is imperative that researchers and practitioners be empowered to evaluate evidence of intervention effects in any SCD study according the WWC standards and recommendations. In this paper, we demonstrate how to assess each of these six features in a real world data set (Lambert et al., 2006). In our demonstration, we assessed each of the six data features separately first. We subsequently integrated six assessments into one comprehensive analysis of the intervention effect. These assessments were conducted using free or commercially available software. The computing algorithms for these assessments appear in Appendices A to C. We conclude this paper by discussing relative advantages of our simple and straightforward approach, compared to visual analysis or complex statistical modeling and methods.

The Lambert data set

The Lambert data set was first reported and analyzed in *Journal of Positive Behavior Interventions* by Lambert et al. (2006). In Lambert et al. (2006)'s study, nine students from two classrooms were observed in baseline (the single-student responding or SSR) phase and the treatment (the response card or RC) phase for their disruptive behaviors during the teacher' instruction. A disruptive behavior, such as engaging in a conversation, provoking others, laughing or touching others, was recorded in 10 intervals of a study session (p. 89 of Lambert et al., 2006). The study employed a reversal (or an ABAB) design with two baseline phases (SSR1 and SSR2), each followed by an intervention phase (RC1 or RC2). The number of intervals in which a disruptive behavior was recorded was the outcome or the dependent measure. Figure 1 presents the findings reproduced from pp. 93-94 of the Lambert et al. (2006) article with permission. Using visual analyses, Lambert et al. (2006) concluded that the use of report cards was successful in decreasing disruptive behaviors for these nine students.

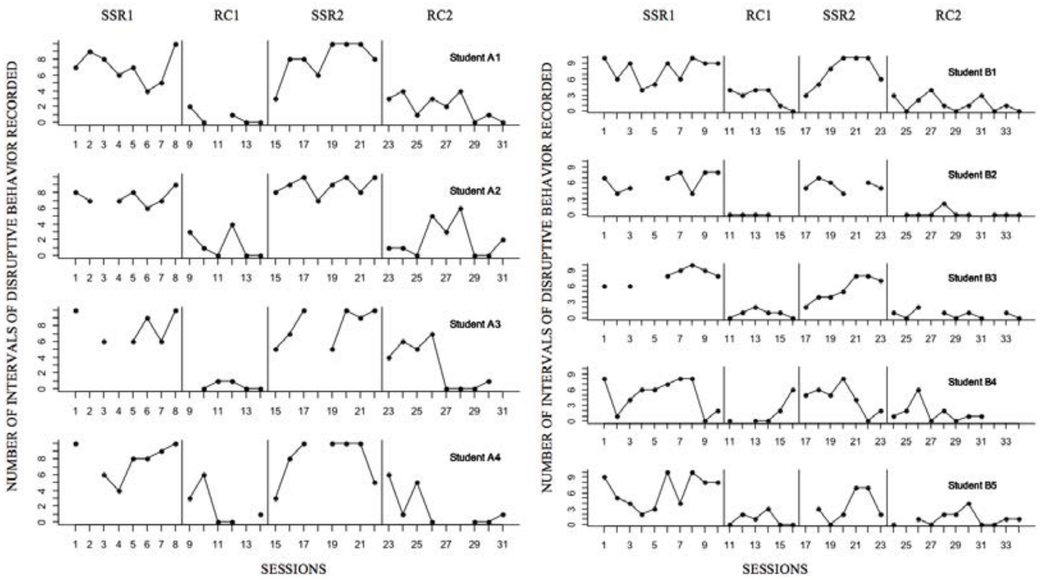


Figure 1. Number of intervals of disruptive behaviors during single-student responding (SSR) and response card (RC) condition. Adapted from "Effects of Response Cards on Disruptive Behavior and Academic Responding During Math Lessons by Fourth-Grade Urban Students," by Lambert et al., 2006, *Journal of Positive Behavior Interventions*, *8*, pp. 93-94, Copyright 2006 by Sage Publications. Used with permission.

Notice that there are breaks in Figure 1 due to student absences (p. 93 of Lambert et al., 2006). These breaks were ignored in the reanalysis of this data set by the special issue of *Journal of School Psychology* (Shadish, 2014). In this paper, we treat these breaks as missing data in order to retain the initial structure of this data set. Because there were different numbers of sessions implemented in the two baseline phases (SSR1 and SSR2) and the intervention phases (RC1 and RC2) in Classrooms A and B, we decided to analyze the two classroom data sets separately. Data collected from four students (A1 to A4) in Classroom A are hereafter referred to as the Lambert-A data set. B1 to B5 students' data from Classroom B are referred to as the Lambert-B data set. Both Lambert-A and -B data sets were systematically analyzed using SAS (Appendix A), a free web-based calculator (Appendix B; Vannest, Parker, & Gonen, 2011), and SPSS (Appendix C).

Assessment of level/level change

The *WWC Handbook* defines "level" as the mean score for data within a phase (2013, p. E.6). A level change between phases therefore indicates a change in the outcome measure due to the intervention. To assess the level and level change, we applied six paired-samples *t*-tests to means obtained from adjacent phases in Lambert-A and -B data sets (Table 1). The SAS computing codes for assessing levels and level changes are shown in Part A of Appendix A. The *t*-statistics and their corresponding *p*-values were further verified by two free web-sites located at http://www.statdistributions.com/chisquare/ and http://www.graphpad.com/quickcalcs/ttest1.cfm, respectively.

According to Table 1 results, the three paired-samples *t*-tests for Lambert-A data ranged from 18.57 to −16.99 with $df = 3$ (or 4−1). For Lambert-B data, the three paired-samples *t*-tests ranged from 8.52 to −6.70 with $df = 4$ (or 5−1). All six paired-samples *t*-tests were statistically significant at $\alpha = .05$ (one-tailed), suggesting that there was a level change between phases for both data sets. And the level changes supported the effectiveness of the intervention, namely, the report card treatment.

Table 1. Means, *SD*s, *t*-tests of differences between phases in Lambert-A and –B data sets

	SSR1-RC1		RC1-SSR2		SSR2-RC2	
	Set A	**Set B**	**Set A**	**Set B**	**Set A**	**Set B**
Mean[a]	6.45	5.46	−7.26	−4.01	6.19	4.21
SD[b]	0.69	1.43	0.85	1.34	0.70	1.62
m[c]	4	5	4	5	4	5
t-test[d]	18.57 (*df*=3)	8.52 (*df*=4)	−16.99 (*df*=3)	−6.70 (*df*=4)	17.81 (*df*=3)	5.82 (*df*=4)
p-value	0.00015	0.0005	0.0002	0.0013	0.0002	0.00215

Note. [a] Means are computed as an average of individuals' difference score over sessions between the two adjacent phases. Missing scores are left as missing.
[b] *SD*s are computed as the square root of the variance of individuals' difference scores. Missing scores are left as missing.
[c] *m* = number of participants.
[d] *t*-test of adjacent phases, *df* = *m*−1.

Assessment of trend

"*Trend* refers to the slope of the best-fitting straight line for the data within a phase," according to The *WWC Handbook* (2013, p. E.6). Because a best-fitting straight line is a narrow definition for trends, we elected to assess monotonic trends in the Lambert data set using the Page test. A monotonic trend can be either increasing or decreasing. It is more general than a linear trend because a monotonic trend incorporates different slopes throughout a data pattern to reflect an upward (or increasing), or a downward (or decreasing), trend in data. Marascuilo and McSweeney (1977) and Page (1963) recommended the Page test for testing monotonic changes over time in SCD. The type of measurement required by the Page test is ranks of data or ranked data. Marascuilo and Busk (1988) and Busk and Marascuilo (1992) effectively applied the Page test to assess trends in the simple AB design, the multiple-baseline AB designs and replicated ABAB designs across participants. Recently, Peng and Chen (2014) proposed a measure of effect sizes (ES) and its confidence interval (CI) to accompany the Page test. Both the ES and its CI are derived from the Page test statistic to further determine an increasing, or a decreasing, trend in data.

To assess trends in the Lambert data set, we conducted six Page tests, computed six corresponding ES measures and their CIs. These results appear in Tables 2-7. SAS computing codes for assessing trends in Lambert-A data are shown in Part B of Appendix A.

Six Page tests of trends

The Page test was applied to three adjacent phases (SSR1-RC1, RC1-SSR2, SSR2-RC2) in both Lambert A and B data sets. A total of six Page tests were performed. According to Lambert et al. (2006), the RC intervention should minimize a student's disruptive behavior. Therefore, for two of the three adjacent phrases (i.e., SSR1-RC1 and SSR2-RC2), we proposed to test the null hypothesis of no trend against the alternative of a monotonic decreasing trend. For the RC1-SSR2 adjacent phrases, the null hypothesis is the same as before; yet the alternative hypothesis states that there is a monotonic increasing trend. Thus, all alternative hypotheses were directional. For demonstration purposes, we describe the Page test of the SSR1-RC1 phases from the Lambert-A data first (Table 2). The results of the other two adjacent phases from Set A are presented in Tables 3 and 4. Parallel analyses of the Lambert-B data appear in Tables 5-7.

For data obtained from the SSR1-RC1 phases in Lambert-A data, the following null and alternative hypotheses are specified, in (1) and (2), respectively:

$$H_0 : \bar{R}_1 = \bar{R}_2 = \ldots = \bar{R}_{14} \tag{1}$$

$$H_1 : \bar{R}_1 \geq \bar{R}_2 \geq \ldots \geq \bar{R}_{14}, \text{ with at least one strict inequality.} \tag{2}$$

Note that the null and alternative hypotheses specify mean ranks of students' scores only. Furthermore, the rejection of H_0 requires no more than one inequality in the ranked data, a decline in this case. In order to apply the Page test to test H_0 in (1), the raw data in the upper panel of Table 2 were converted to ranks for each student, shown in the middle panel of Table 2. Ranks are assigned from high to low within each student. If scores were tied, we broke the tie by averaging the two corresponding ranks, such as assigning the rank of 10.5 to the two 7s for Student A1 in both Sessions 1 and 5 during the SSR1 phase. Missing data were treated conservatively in the sense of supporting the null hypothesis, instead of the alternative hypothesis. Thus, if the H_0 of no trend can be rejected at $\alpha = .05$ with this conservative approach, it can be rejected at the same or a lower α level, if the missing data were replaced by a score in support of the alternative hypothesis. Thus, for Student A1 in Session 11 in the RC1 phase (upper panel of Table 2), we treated the missing score, shown as a period (.), with a score of 2, appearing in parenthesis. The score of 2 was the highest score of Student A1 in the RC1 phase. Replacing the missing score by 2 supported the null hypothesis of no trend, more than other scores taken from Student A1 for this phase. This replacement led to a rank of 5.5, shown in parenthesis, in the middle panel of Table 2. Likewise, for Student A2 in Session 3 in the SSR1 phase, we treated the missing score with 6, in parenthesis. The score of 6 was the lowest score of Student A2 in the SSR1 phase. Other missing data were treated similarly in either the SSR1 or the RC1 phase.

Table 2. Number of intervals of disruptive behaviors and their ranks in 8 sessions (1 to 8) of the SSR1 phase and 6 sessions (9 to 14) of the RC1 phase of Class A (Lambert et al., 2006)

	SSR1								RC1					
Session	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**
A1	7	9	8	6	7	4	5	10	2	0	.(2)	1	0	0
A2	8	7	.(6)	7	8	6	7	9	3	1	0	4	0	0
A3	10	.(6)	6	.(6)	6	9	6	10	.(1)	0	1	1	0	0
A4	10	.(4)	6	4	8	8	9	10	3	6	0	0	.(6)	1
Mean	8.75	6.5	6.5	5.75	7.25	6.75	6.75	9.75	2.25	1.75	0.75	1.5	1.5	0.25
SD	1.5	2.08	1	1.26	0.96	2.22	1.71	0.5	0.96	2.87	0.96	1.73	3	0.5
	SSR1 Ranks								**RC1 Ranks**					
Session	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**
A1	10.5	13	12	9	10.5	7	8	14	5.5	2	(5.5)	4	2	2
A2	12.5	10	(7.5)	10	12.5	7.5	10	14	5	4	2	6	2	2
A3	13.5	(9)	9	(9)	9	12	9	13.5	(5)	2	5	5	2	2
A4	13.5	(5.5)	8	5.5	10.5	10.5	12	13.5	4	8	1.5	1.5	(8)	3
Total Rank $\left(\sum_{i=1}^{m=4} \bar{R}_j\right)$	50	37.5	36.5	33.5	42.5	37	39	55	19.5	16	14	16.5	14	9
Expected Rank (*Y_j*)	14	13	12	11	10	9	8	7	6	5	4	3	2	1

H_0	$H_0 : \bar{R}_1 = \bar{R}_2 = \ldots = \bar{R}_{14}$	***m, n***	4, 14	**Standardized *L* (or *z*)**	z = 5.06[c]
H_1	$H_1 : \bar{R}_1 \geq \bar{R}_2 \geq \ldots \geq \bar{R}_n$, w / ≥ 1strict inequality	***χ^2(df=1)***	25.60 [b]	***z*-upper**	7.02[d]
Page *L*	*L* = 3788.5[a]	***p*-value**	< .0001	***z*-lower**	3.10[d]

Note: Missing data are denoted as a period (.). Its rank is based on the score shown in parenthesis next to the period (.). The score in the parenthesis is assigned a rank, also shown in parenthesis, based on the lowest score of SSR1 phase or the highest score of the RC1 phase, in support of the H_0. Tied scores are assigned the average rank of the corresponding ranks.

[a]
$$L = 3788.5 = \sum_{j=1}^{n=14}\left[Y_j \times \left(\sum_{i=1}^{m=4} \bar{R}_j\right)\right] = \left[\begin{array}{l} 14\times(50)+13\times(37.5)+12\times(36.5)+11\times(33.5) \\ +10\times(42.5)+9\times(37)+8\times(39)+7\times(55)+6\times(19.5) \\ +5\times(16)+4\times(14)+3\times(16.5)+2\times(14)+1\times(9) \end{array}\right].$$

[b]
$$25.60 = \chi_L^2 = \frac{\left[12L - 3mn(n+1)^2\right]^2}{mn^2(n^2-1)(n+1)} = \frac{\left[12\times 3788.5 - 3\times 4\times 14\times(14+1)^2\right]^2}{4\times 14^2\times(14^2-1)\times(14+1)} = \frac{[45462-37800]^2}{4\times 196\times 195\times 15} = \frac{58706244}{2293200} = 25.60014129.$$

[c] $5.06 = z = \sqrt{\chi_L^2} = \sqrt{25.60014129} = 5.059658.$

[d] 95% CI for Standardized L$= z \pm 1.96 =$ 5.06± 1.96$= [\,3.10, 7.02\,]$

Table 3. Number of intervals of disruptive behaviors and their ranks in 8 sessions (15 to 22) of the SSR2 phase and 9 sessions (23 to 31) of the RC2 phase of Class A (Lambert et al., 2006)

	SSR2							RC2									
Session	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**	**31**
A1	8	8	8	6	10	10	10	8	3	4	1	3	2	4	0	1	0
A2	8	9	10	7	9	10	8	10	1	1	0	5	3	6	0	0	2
A3	5	7	10	.(5)	5	10	9	10	4	6	5	7	0	0	0	0	. (7)
A4	3	8	10	.(3)	10	10	10	5	6	1	5	0	. (6)	. (6)	0	0	1
Mean	6.00	8.00	9.50	5.25	8.50	10.00	9.25	8.25	3.50	3.00	2.75	3.75	2.75	4.00	0.00	0.25	2.50
SD	2.45	0.82	1.00	1.71	2.38	0	0.96	2.36	2.08	2.45	2.63	2.99	2.50	2.83	0.00	0.50	3.11
	SSR2 Ranks							**RC2 Ranks**									
Session	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**	**31**
A1	12.5	12.5	12.5	10	16	16	16	12.5	6.5	8.5	3.5	6.5	5	8.5	1.5	3.5	1.5
A2	11.5	13.5	16	10	13.5	16	11.5	16	4.5	4.5	2	8	7	9	2	2	6
A3	7.5	12	16	(7.5)	7.5	16	14	16	5	10	7.5	12	2.5	2.5	2.5	2.5	(12)
A4	6.5	13	15.5	(6.5)	15.5	15.5	15.5	8.5	11	4.5	8.5	2	(11)	(11)	2	2	4.5
Total Rank	38	51	60	34	52.5	63.5	57	53	27	27.5	21.5	28.5	25.5	31	8	10	24
Expected Rank	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1

***H*₀**	$H_0: \bar{R}_1 = \bar{R}_2 = \ldots = \bar{R}_n$	***m, n***	4, 17	**Standardized *L* (or *z*)**	z2 = 5.08[c]
***H*₁**	$H_1: \bar{R}_1 \geq \bar{R}_2 \geq \ldots \geq \bar{R}_n$, w / ≥ 1 strict inequality	**χ^2 (*df*=1)**	25.77 [b]	***z*-upper**	7.04[d]
Page *L*	*L2* = 6543.5[a]	***p*-value**	< .0001	***z*-lower**	3.12[d]

***Note*:** Missing data are denoted as a period (.). Its rank is based on the score shown in parenthesis next to the period (.). The score in the parenthesis is assigned a rank, also shown in parenthesis, based on the lowest score of SSR2 phase or the highest score of the RC2 phase, in support of the H_0. Tied scores are assigned the average rank of the corresponding ranks.

[a]
$$L2 = 6543.5 = \sum_{j=1}^{n=17}\left[Y_j \times \left(\sum_{i=1}^{m=4} \bar{R}_j\right)\right] = \begin{bmatrix} 17\times(38)+16\times(51)+15\times(60)+14\times(34)+13\times(52.5) \\ +12\times(63.5)+11\times(57)+10\times(53)+9\times(27)+8\times(27.5) \\ +7\times(21.5)+6\times(28.5)+6\times(28.5)+5\times(25.5) \\ +4\times(31)+3\times(8)+2\times(10)+1\times(24) \end{bmatrix}$$

[b]
$$25.77 = \chi^2_{L2} = \frac{\left[12L - 3mn(n+1)^2\right]^2}{mn^2(n^2-1)(n+1)} = \frac{\left[12\times 6543.5 - 3\times 4\times 17\times(17+1)^2\right]^2}{4\times 17^2\times(17^2-1)\times(17+1)} = \frac{[78522-66096]^2}{4\times 289\times 288\times 18} = \frac{154405476}{5992704} = 25.76557694.$$

[c]
$$z2 = 5.08 = \sqrt{\chi^2_{L2}} = \sqrt{25.76557694} = 5.075980392.$$

[d]
$$95\%\ \text{CI for StandardizedL2} = z2 \pm 1.96 = 5.08 \pm 1.96 = [3.12, 7.04].$$

Table 4. Number of intervals of disruptive behaviors and their ranks of 6 sessions (9 to 14) of the RC1 phase and 8 sessions (15-22) of the SSR2 phase of Class A (Lambert et al., 2006)

	RC1						SSR2							
Session	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**
A1	2	0	.(2)	1	0	0	8	8	8	6	10	10	10	8
A2	3	1	0	4	0	0	8	9	10	7	9	10	8	10
A3	.(1)	0	1	1	0	0	5	7	10	. (5)	5	10	9	10
A4	3	6	0	0	.(6)	1	3	8	10	. (3)	10	10	10	5
Mean	2.25	1.75	0.75	1.50	1.50	0.25	6.00	8.00	9.50	5.25	8.50	10.00	9.25	8.25
SD	0.96	2.87	0.96	1.73	3.00	0.50	2.45	0.82	1.00	1.71	2.38	0.00	0.96	2.36
	RC1 Ranks						**SSR2 Ranks**							
Session	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**
A1	5.5	2	(5.5)	4	2	2	9.5	9.5	9.5	7	13	13	13	9.5
A2	5	4	2	6	2	2	8.5	10.5	13	7	10.5	13	8.5	13
A3	(5)	2	5	5	2	2	8	10	13	(8)	8	13	11	13
A4	5	8.5	1.5	1.5	(8.5)	3	5	10	12.5	(5)	12.5	12.5	12.5	7
Total Rank	20.5	16.5	14	16.5	14.5	9	31	40	48	27	44	51.5	45	42.5
Expected Rank	1	2	3	4	5	6	7	8	9	10	11	12	13	14

H_0	$H_0 : \bar{R}_1 = \bar{R}_2 = \ldots = \bar{R}_n$	***m, n***	4, 14	**Standardized *L* (or *z*)**	z3 = 5.22[c]
H_1	$H_1 : \bar{R}_1 \geq \bar{R}_2 \geq \ldots \geq \bar{R}_n$, w / ≥ 1 strict inequality	**χ^2 (*df*=1)**	27.27 [b]	***z*-upper**	7.18[d]
Page *L*	*L3* = 3809[a]	***p*-value**	< .0001	***z*-lower**	3.26[d]

***Note*:** Missing data are denoted as a period (.). Its rank is based on the score shown in parenthesis next to the period (.). The score in the parenthesis is assigned a rank, also shown in parenthesis, based on the lowest score of SSR2 phase or the highest score of the RC1 phase, in support of the H_0. Tied scores are assigned the average rank of the corresponding ranks.

[a]
$$L3 = 3809 = \sum_{j=1}^{n=14}\left[Y_j \times \left(\sum_{i=1}^{m=4} \bar{R}_j\right)\right] = \begin{bmatrix} 1\times(20.5)+2\times(16.5)+3\times(14)+4\times(16.5)+5\times(14.5) \\ +6\times(9)+7\times(31)+8\times(40)+9\times(48)+10\times(27) \\ +11\times(44)+12\times(51.5)+13\times(45)+14\times(42.5) \end{bmatrix}.$$

[b]
$$27.27 = \chi^2_{L3} = \frac{\left[12L - 3mn(n+1)^2\right]^2}{mn^2(n^2-1)(n+1)} = \frac{\left[12\times 3809 - 3\times 4\times 14(14+1)^2\right]^2}{4\times 14^2\times(14^2-1)\times(14+1)} = \frac{[45708-37800]^2}{4\times 196\times 195\times 15} = \frac{62536464}{2293200} = 27.27039246.$$

[c]
$$z3 = 5.22 = \sqrt{\chi^2_{L3}} = \sqrt{27.27039246} = 5.222106133.$$

[d]
$$95\% \text{ CI for StandardizedL3} = z3 \pm 1.96 = 5.22 \pm 1.96 = [3.26, 7.18].$$

Table 5. Number of intervals of disruptive behaviors and their ranks in 10 sessions (1 to 10) of the SSR1 phase and 6 sessions (11 to 16) of the RC1 phase of Class B (Lambert et al., 2006)

	SSR1										RC1					
Session	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**
B1	10	6	9	4	5	9	6	10	9	9	4	3	4	4	1	0
B2	7	4	5	.(4)	.(4)	7	8	4	8	8	0	0	0	0	.(0)	.(0)
B3	6	.(6)	6	.(6)	.(6)	8	9	10	9	8	0	1	2	1	1	0
B4	8	1	4	6	6	7	8	8	0	2	0	.(6)	0	0	2	6
B5	9	5	4	2	3	10	4	10	8	8	0	2	1	3	0	0
Mean	8.00	4.40	5.60	4.40	4.80	8.20	7.00	8.40	6.80	7.00	.80	2.40	1.40	1.60	.80	1.20
SD	1.58	2.07	2.07	1.67	1.30	1.30	2.00	2.61	3.83	2.83	1.79	2.30	1.67	1.82	0.84	2.68
	SSR1 Ranks										**RC1 Ranks**					
Session	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**
B1	15.5	9.5	12.5	5.5	8	12.5	9.5	15.5	12.5	12.5	5.5	3	5.5	5.5	2	1
B2	12.5	8.5	11	(8.5)	(8.5)	12.5	15.0	8.5	15	15	3.5	3	3.5	3.5	(3.5)	(3.5)
B3	9	(9)	9	(9)	(9)	12.5	14.5	16	14.5	12.5	1.5	4	6	4	4	1.5
B4	15	5	8	10.5	10.5	13	15	15	2.5	6.5	2.5	(10.5)	2.5	2.5	6.5	10.5
B5	14	11	9.5	5.5	7.5	15.5	9.5	15.5	12.5	12.5	2	5.5	4	7.5	2	2
Total Rank	66	43	50	39	43.5	66	63.5	70.5	57	59	15	26.5	21.5	23	18	18.5
Expected Rank	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1

H₀	$H_0: \bar{R}_1 = \bar{R}_2 = \ldots = \bar{R}_n$	***m, n***	5, 16	**Standardized *L* (or *z*)**	z4 = 4.82[c]
H₁	$H_1: \bar{R}_1 \geq \bar{R}_2 \geq \ldots \geq \bar{R}_n$, w / ≥ 1 strict inequality	χ^2 **(*df*=1)**	23.25 [b]	***z*-upper**	6.78[d]
Page *L*	*L4* = 6726.5[a]	***p*-value**	< .0001	***z*-lower**	2.86[d]

Note: Missing data are denoted as a period (.). Its rank is based on the score shown in parenthesis next to the period (.). The score in the parenthesis is assigned a rank, also shown in parenthesis, based on the lowest score of SSR1 phase or the highest score of the RC1 phase, in support of the H_0. Tied scores are assigned the average rank of the corresponding ranks.

[a]
$$L4 = 6726.5 = \sum_{j=1}^{n=16}\left[Y_j \times \left(\sum_{i=1}^{m=4} \bar{R}_j\right)\right] = \begin{bmatrix} 16\times(66)+15\times(43)+14\times(50)+13\times(39)+12\times(43.5)+11\times(66) \\ +10\times(63.5)+9\times(70.5)+8\times(57)+7\times(59)+6\times(15) \\ +5\times(26.5)+4\times(21.5)+3\times(23)+2\times(18)+1\times(18.5) \end{bmatrix}$$

[b]
$$23.25 = \chi^2_{L4} = \frac{\left[12L - 3mn(n+1)^2\right]^2}{mn^2(n^2-1)(n+1)}$$

$$= \frac{\left[12\times 6726.5 - 3\times 5\times 16\times(16+1)^2\right]^2}{5\times 16^2\times(16^2-1)\times(16+1)} = \frac{[80718-69360]^2}{5\times 256\times 255\times 17} = \frac{129004164}{5548800} = 23.24902033.$$

[c] $z4 = 4.82 = \sqrt{\chi^2_{L4}} = \sqrt{23.24902033} = 4.821723792.$

[d] $95\%\ \text{CI for StandardizedL4} = z4 \pm 1.96 = 4.82 \pm 1.96 = [2.86, 6.78].$

Table 6. Number of intervals of disruptive behaviors and their ranks in 7 sessions (17 to 23) of the SSR2 phase and 11 sessions (24 to 34) of the RC2 phase of Class B (Lambert et al., 2006)

	SSR2							**RC2**										
Session	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**	**31**	**32**	**33**	**34**
B1	3	5	8	10	10	10	6	3	0	2	4	1	0	1	3	0	1	0
B2	5	7	6	4	.(4)	6	5	.(2)	0	0	0	2	0	0	.(2)	0	0	0
B3	2	4	4	5	8	8	7	1	0	3	.(3)	1	0	1	0	.(3)	1	0
B4	5	6	5	8	4	0	2	1	2	6	0	2	0	1	1	.(6)	.(6)	.(6)
B5	.(0)	3	0	2	7	7	2	0	.(4)	1	0	2	2	4	0	0	1	1
Mean	3.00	5.00	4.60	5.80	6.60	6.20	4.40	1.40	1.20	2.40	1.40	1.60	0.40	1.40	1.20	1.80	1.80	1.40
SD	2.12	1.58	2.97	3.19	2.61	3.77	2.30	1.14	1.79	2.30	1.95	0.55	0.89	1.52	1.30	2.68	2.39	2.61
	SSR2 Ranks							**RC2 Ranks**										
Session	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**	**31**	**32**	**33**	**34**
B1	10	13	15	17	17	17	14	10	2.5	8	12	6.0	2.5	6	10	2.5	6	2.5
B2	14.5	18	16.5	12.5	(12.5)	16.5	14.5	(10)	4.5	4.5	4.5	10	4.5	4.5	(10)	4.5	4.5	4.5
B3	9	13.5	13.5	15	17.5	17.5	16	6.5	2.5	11	(11)	6.5	2.5	6.5	2.5	(11)	6.5	2.5
B4	11.5	15	11.5	18	10	2	8	5.0	8	15	2	8	2	5	5	(15)	(15)	(15)
B5	(3.5)	14.0	3.5	11.5	17.5	17.5	11.5	3.5	(15.5)	8	3.5	11.5	11.5	15.5	3.5	3.5	8	8
Total Rank	48.5	73.5	60	74	74.5	70.5	64	35	33	46.5	33	42	23	37.5	31	36.5	40	32.5
Expected Rank	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1

H_0	$H_0 : \bar{R}_1 = \bar{R}_2 = \ldots = \bar{R}_n$	***m, n***	5, 18	**Standardized *L* (or *z*)**	z5 = 4.42[c]
H_1	$H_1 : \bar{R}_1 \geq \bar{R}_2 \geq \ldots \geq \bar{R}_n$, w / ≥ 1 strict inequality	χ^2 **(*df*=1)**	19.51 [b]	***z*-upper**	6.38[d]
Page *L*	*L5* = 9283[a]	***p*-value**	< .0001	***z*-lower**	2.46[d]

Note: Missing data are denoted as a period (.). Its rank is based on the score shown in parenthesis next to the period (.). The score in the parenthesis is assigned a rank, also shown in parenthesis, based on the lowest score of SSR2 phase or the highest score of the RC2 phase, in support of the H_0. Tied scores are assigned the average rank of the corresponding ranks.

[a]
$$L5 = 9283 = \sum_{j=1}^{n=18}\left[Y_j \times \left(\sum_{i=1}^{m=5}\bar{R}_j\right)\right] = \begin{bmatrix} 18\times(48.5)+17\times(73.5)+16\times(60)+15\times(74)+14\times(74.5)+13\times(70.5) \\ +12\times(64)+11\times(35)+10\times(33)+9\times(46.5)+8\times(33)+7\times(42) \\ +6\times(23)+5\times(37.5)+4\times(31)+3\times(36.5)+2\times(40)+1\times(32.5) \end{bmatrix}.$$

[b]
$$19.51 = \chi^2_{L5} = \frac{\left[12L - 3mn(n+1)^2\right]^2}{mn^2(n^2-1)(n+1)}$$

$$= \frac{\left[12\times 9283 - 3\times 5\times 18\times(18+1)^2\right]^2}{5\times 18^2\times(18^2-1)\times(18+1)} = \frac{[111396-97470]^2}{5\times 324\times 323\times 19} = \frac{193933476}{9941940} = 19.50660294.$$

[c] $z5 = 4.42 = \sqrt{\chi^2_{L5}} = \sqrt{19.50660294} = 4.416628005.$

[d] $95\%\text{ CI for StandardizedL5} = z5 \pm 1.96 = 4.42 \pm 1.96 = [2.46, 6.38].$

Table 7. Number of intervals of disruptive behaviors and their ranks of 6 sessions (11 to 16) of the RC1 phase and 7 sessions (17-23) of the SSR2 phase of Class B (Lambert et al., 2006)

	RC1						SSR2						
Session	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**
B1	4	3	4	4	1	0	3	5	8	10	10	10	6
B2	0	0	0	0	. (0)	.(0)	5	7	6	4	.(4)	6	5
B3	0	1	2	1	1	0	2	4	4	5	8	8	7
B4	0	.(6)	0	0	2	6	5	6	5	8	4	0	2
B5	0	2	1	3	0	0	.(0)	3	0	2	7	7	2
Mean	.80	2.40	1.40	1.60	.80	1.20	3.00	5.00	4.60	5.80	6.60	6.20	4.40
SD	1.79	2.30	1.67	1.82	.84	2.68	2.12	1.58	2.97	3.19	2.61	3.77	2.30
	RC1 Ranks						**SSR2 Ranks**						
Session	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**
B1	6	3.5	6	6	2	1	3.5	8	10	12	12.	12	9
B2	3.5	3.5	3.5	3.5	(3.5)	(3.5)	9.5	13	11.5	7.5	(7.5)	11.5	9.5
B3	1.5	4	6.5	4	4	1.5	6.5	8.5	8.5	10	12.5	12.5	11
B4	2.5	(11)	2.5	2.5	5.5	11	8.5	11	8.5	13	7	2.5	5.5
B5	3	8	6	10.5	3	3	(3)	10.5	3	8	12.5	12.5	8
Total Rank	16.5	30	24.5	26.5	18	20	31	51	41.5	50.5	51.5	51	43
Expected Rank	1	2	3	4	5	6	7	8	9	10	11	12	13

***H*₀**	$H_0: \bar{R}_1 = \bar{R}_2 = \ldots = \bar{R}_n$	***m, n***	5, 13	**Standardized *L* (or *z*)**	z6 = 4.44[c]
***H*₁**	$H_1: \bar{R}_1 \geq \bar{R}_2 \geq \ldots \geq \bar{R}_n$, w / ≥ 1 strict inequality	**χ^2 (*df*=1)**	19.74 [b]	***z*-upper**	6.40[d]
Page *L*	*L6* =3707[a]	***p*-value**	< .0001	***z*-lower**	2.48[d]

Note: Missing data are denoted as a period (.). Its rank is based on the score shown in parenthesis next to the period (.). The score in the parenthesis is assigned a rank, also shown in parenthesis, based on the lowest score of SSR2 phase or the highest score of the RC1 phase, in support of the H_0. Tied scores are assigned the average rank of the corresponding ranks.

[a] $$L6 = 3707 = \sum_{j=1}^{n=13}\left[Y_j\left(\sum_{i=1}^{m=5}\bar{R}_j\right)\right] = \begin{bmatrix} 1\times(16.5)+2\times(30)+3\times(24.5)+4\times(26.5) \\ +5\times(18)+6\times(20)+7\times(31)+8\times(51)+9\times(41.5) \\ +10\times(50.5)+11\times(51.5)+12\times(51)+13\times(43) \end{bmatrix}.$$

[b] $$19.74 = \chi^2_{L6} = \frac{\left[12L-3mn(n+1)^2\right]^2}{mn^2(n^2-1)(n+1)}$$

$$= \frac{\left[12\times3707-3\times5\times13\times(13+1)^2\right]^2}{5\times13^2\times(13^2-1)\times(13+1)} = \frac{[44484-38220]^2}{5\times169\times168\times14} = \frac{39237696}{1987440} = 19.74283299.$$

[c] $z6 = 4.44 = \sqrt{\chi^2_{L6}} = \sqrt{19.74283299} = 4.443290784.$

[d] $95\%\ \text{CI for StandardizedL6} = z6 \pm 1.96 = 4.44 \pm 1.96 = [2.48, 6.40].$

Next, we computed the total rank for each of the 14 sessions. The total ranks $\left(\sum_{i=1}^{m=4} R_{ij}\right)$ were subsequently weighted by their expected ranks (Y_j), suggested by

H_1. The product of the total rank weighted by its expected rank was subsequently summed over all 14 sessions into the Page statistic, L, according to (3) below:

$$\begin{aligned} L &= \sum_{j=1}^{n=14}\left[Y_j \times \left(\sum_{i=1}^{m=4} R_{ij}\right)\right] \\ &= \left[\begin{array}{l} 14\times(50)+13\times(37.5)+12\times(36.5)+11\times(33.5) \\ +10\times(42.5)+9\times(37)+8\times(39)+7\times(55)+6\times(19.5) \\ +5\times(16)+4\times(14)+3\times(16.5)+2\times(14)+1\times(9) \end{array}\right] \\ &= 3788.5 \end{aligned} \qquad (3)$$

where, n = the number of sessions, m = the number of students/participants, Y_j = the expected rank of the j^{th} session, and R_{ij} = the observed rank of the i^{th} student's score in the j^{th} session.

The exact significance level of the L statistic can be obtained from Page (1963), if n ranges from 3 to 10 and m ranges from 2 to 50. Given the present values of $n = 14$ and $m = 4$, the significance level can be approximated by a chi-square distribution with $df = 1$, according to (4) below (Page, 1963, p. 224):

$$\begin{aligned} \chi_L^2 &= \frac{\left[12L - 3mn(n+1)^2\right]^2}{mn^2(n^2-1)(n+1)} \\ &= \frac{\left[12\times3788.5 - 3\times4\times14\times(14+1)^2\right]^2}{4\times14^2\times(14^2-1)\times(14+1)} \\ &= \frac{[45462-37800]^2}{4\times196\times195\times15} = \frac{58706244}{2293200} = 25.60 \end{aligned} \qquad (4)$$

The above chi-square statistic is statistically significant at $p < .0001$ leading to a rejection of H_0 of no trend at $\alpha = .05$ (one-tailed), specified in (1) above. We therefore concluded that there was a monotonic decreasing trend across these 14 sessions, as specified in H_1 of (2).

The large-sample approximation to the sampling distribution of Page's L statistic yields acceptable Type I error rates for a directional Page test, as long as $n > 11$ for $\alpha = .05$, or $n > 18$ for $\alpha = .01$, according to Fahoome (2002). An acceptable Type I error rate was defined in Fahoome (2002) as within 10% of the nominal α rate, in reference to Bradley (1978)'s work. Page (1963) also suggested that the large-sample chi-square approximation be used under one of three conditions: (1) for $m > 20$ with any n, (2) for $m > 12$ and $n \geq 4$, or (3) for any m when $n \geq 9$. Because $m = 4$ and $n = 14$, the Page test result and its statistical significance level were judged to be acceptable, according to Bradley (1978), Fahoome (2002), and Page (1963).

Summary of six Page tests of trends. The Page test was applied similarly to two other adjacent phases from Lambert-A data and to the three adjacent phases from Lambert-B data. Results of these Page tests are summarized in Tables 3 to 7, including their corresponding H_0s and H_1s. All six Page tests shown in Tables 2 to 7 were statistically significant at $p < .0001$, rejecting all H_0s at $\alpha = .05$ (one-tailed) and confirming a trend as specified in the corresponding H_1s. For data in the SSR1-RC1 and the SSR2-RC2 adjacent phases, the Page test results of L, $L2$, $L4$, and $L5$ suggested a monotonic decreasing trend from the baseline phase (i.e., SSR) to the intervention phase (i.e., RC) in both Lambert-A and -B data sets. For data in the RC1-SSR2 adjacent phases, the Page test results of $L3$ and $L6$ suggested a monotonic increasing trend from the intervention phase (i.e., RC1) to the baseline phase (i.e., SSR2) again for both A and B data sets.

Six ES measures derived from Page's *L*

The L statistic defined in (3) is conceptually and algebraically equivalent to the average Spearman rank correlation coefficient (ρ) between Students' ranked scores (i.e., the frequency of disruptive behaviors) and the expected ranks according to a monotonic decreasing or increasing trend (Page, 1963; van de Wiel & Di Bucchianico, 2001). It is an unstandardized ES measure of a monotonic trend in data. To convert L into a standardized ES, one divides Page's L (i.e., the average ρ) by its standard deviation (Lyerly, 1952; Page, 1963, p. 227) to yield a standardized normal z, as in (5):

$$\frac{\rho}{SD_\rho} = \rho \times \sqrt{m \times (n-1)} = z = \sqrt{\chi_L^2} = \sqrt{25.60} = 5.06 \qquad (5)$$

where χ_L^2 is defined in (4) above. This normalized z statistic is similar to Cohen's d, in the sense of being scale-free and ranging from negative to positive values without bounds. They differ, however, in their assumptions. Cohen's d and its population parameter δ assume normality and equal variances for underlying populations (Cohen, 1988), whereas the standardized L, or the normalized z in (5), does not, because the latter is based on ranks of the data.

CI for the standardized ES derived from Page's *L*

Since the standardized L, or z from (5), follows a standard normal distribution (e.g., Fahoome, 2002; Lyerly, 1952), a nondirectional 95% CI for the standardized L can be constructed using (6) below:

$$\text{95\% CI for Standardized } L = z \pm 1.96 = 5.06 \pm 1.96 = \left[3.10, 7.02\right] \qquad (6)$$

Because the upper and the lower limits of the 95% CI are both positive, the 95% CI supports the earlier rejection of the H_0 of no trend at $\alpha = .05$, in favor of a monotonic decreasing trend across the 14 sessions from the SSR1-RC1 phases of the Lambert-A data.

Summary of six ESs and six CIs. The standardized ESs (or zs) and their corresponding CIs further confirmed the rejection of the H_0 of no trend and in favor of the H_1 of a monotonic trend. Taken together, the six Page test results, their corresponding ESs and CIs provided multiple evidence for monotonic decreasing trends in students' disruptive behaviors due to the intervention.

Assessment of variability

According to the *WWC Handbook* (2013), "*Variability* refers to the range or standard deviation of data about the best-fitting straight line." (p. E.6). Even though we did not fit a straight regression line to the Lambert data, the variability of scores was assessed within and between phases using SAS—see Part A of Appendix A; results are presented in Table 1. In five out of six instances, the intervention phases (RC1 and RC2) yielded less variability than their corresponding baseline phases, namely, SSR1 and SSR2 respectively. The only exception occurred in Lambert-A data set between SSR2 and RC2. We did not test the differences in variability because these statistical tests (e.g., Levene's F' test) are not robust under nonnormal conditions, which might be the case for the Lambert data.

Assessment of immediacy of the effect

According to the *WWC Handbook* (2013), "*Immediacy of the effect* refers to the change in level between the last three data points in one phase and the first three data points of the next. The more rapid (or immediate) the effect, the more convincing the inference that change in the outcome measure was due to manipulation of the independent variable." (p. E.6). Applying this definition to Figure 1 using the visual analysis, we determined that data patterns in the intervention phases (i.e., RC1 and RC2) exhibited an immediate decreasing effect on disruptive behaviors, compared to data patterns in the baseline phases (i.e., SSR1 and SSR2). Even though the last three data points of Student B4's from the SSR2 phase, compared to the first three data points of the RC2 phase, suggested an exception, the overall profile of this student's data supported a decline in disruptive behavior during the intervention phase. Thus, we concluded that there was an immediacy effect due to the intervention in both A and B data sets.

Assessment of overlap

According to the *WWC Handbook* (2013), "*Overlap* refers to the proportion of data from one phase that overlaps with data from the previous phase. The smaller the proportion of overlapping data points (or conversely, the larger the separation), the more compelling the demonstration of an effect." (p. E.6). To assess this data feature, we computed the degree of nonoverlap for all data pairs (NAP) in adjacent phases for each student (Table 8). NAP is defined as the number of pairs of data showing no overlap between a baseline phase and an intervention phase, divided by the total number of pairs (Parker & Vannest, 2009). Each NAP corresponds to two adjacent phases, such as SSR1 and RC1. Values of NAP range from 0 to 1. A value of 0 indicates that all data points in phase A (e.g., SSR1) are greater than the points in phase B (e.g., RC1). In contrast, a value of 1 indicates that all data points in phase A (e.g., RC1) are smaller than the points in phase B (e.g., SSR2). According to Table 8, all NAP results were statistically significant at $\alpha = .05$ (two-tailed), except for two students (B4 and B5) in two adjacent phases (RC1-SSR2, and SSR2-RC2). We therefore concluded that there was a statistically significant lack of overlap in students' outcome measures between phases, supporting the effectiveness of the intervention in decreasing disruptive behaviors. The NAPs and their corresponding statistical significance were computed using a free web-based calculator from http://www.singlecaseresearch.org/. The web-based calculator was developed by

Vannest, Parker, and Gonen (2011) and its functionalities are shown in Appendix B. The NAP results were subsequently verified by SPSS, shown in Appendix C.

Table 8. Nonoverlap of All Pairs (NAP) between phases in Lambert-A and -B data sets

	SSR1-RC1		RC1-SSR2		SSR2-RC2	
	NAP [a]	*p*-value [b]	NAP	*p*-value	NAP	*p*-value
Student A1	0.0000	0.0034	1.0000	0.0034	0.0417	0.0015
Student A2	0.0000	0.0027	1.0000	0.0019	0.0000	0.0005
Student A3	0.0000	0.0062	1.0000	0.0045	0.0982	0.0092
Student A4	0.0429	0.0094	0.9286	0.0149	0.0714	0.0073
Student B1	0.0250	0.0020	0.9167	0.0124	0.0260	0.0009
Student B2	0.0000	0.0066	1.0000	0.0105	0.0000	0.0015
Student B3	0.0000	0.0027	0.9881	0.0034	0.0079	0.0010
Student B4	0.1800	0.0500	0.7571	0.1439	0.2232	0.0728
Student B5	0.0333	0.0024	0.7778	0.1093	0.2167	0.0652

Note: Missing scores are left as missing.
[a] NAPs were computed using a web-based calculator developed by Vannest, Parker, and Gonen (2011)—see Appendix B, and verified by SPSS's Receiver Operator Characteristics module—see Appendix C.
[b] *p*-values were obtained from the web-based calculator developed by Vannest, Parker, and Gonen (2011)—see Appendix B, and verified by SPSS's Receiver Operator Characteristics module and its option called Area Under the Curve (AUC)—see Appendix C

Assessment of consistency of data in similar phases

According to the *WWC Handbook* (2013, p. E.6), "*Consistency of data in similar phases* involves looking at data from all phases within the same condition… and examining the extent to which there is consistency in the data patterns from phases with the same conditions. The greater the consistency, the more likely the data represent a causal relation." To determine the consistency of data, we employed the visual analysis of the Lambert-A and –B data sets and determined that data patterns were similar in the same phase between these two sets. Furthermore, we applied four independent-samples t-tests to each phase between means of sets A and B, whether it was baseline or intervention (Table 9). According to Table 9, the *t*-test was not statistically significant for any phase at $\alpha = .05$ (two-tailed with $df = 7 = 4+5-2$). These statistically insignificant t-test results suggested that the mean scores obtained from sets A and B were not statistically significantly different from each other. Thus, we concluded that there was consistency of data patterns within similar phases for both data sets. SAS programming codes for assessing consistency in the Lambert-A data are shown in Part C of Appendix A.

Table 9. Means, *SD*s, *t*-tests of differences within phases in Lambert-A and -B data sets

	SSR1		RC1		SSR2		RC2	
	Set A	Set B	Set A	Set B	Set A	Set B	Set A	Set B
Mean[a]	7.53	6.68	1.08	1.22	8.34	5.23	2.15	1.02
SD[b]	$\sqrt{3.48}=1.87$	$\sqrt{5.76}=2.40$	$\sqrt{2.67}=1.63$	$\sqrt{2.41}=1.55$	$\sqrt{4.23}=2.06$	$\sqrt{5.91}=2.43$	$\sqrt{5.83}=2.42$	$\sqrt{1.76}=1.33$
m[c]	4	5	4	5	4	5	4	5
n[d]	8	10	6	6	8	7	9	11
|*t*|[e]	0.5824 (*SE*=1.468)		0.1317 (*SE*=1.063)		2.0345 (*SE*=1.529)		0.8978 (*SE*=1.259)	
***p*-value**	0.579		0.899		0.081		0.4	

[a] Means are computed as an average of individuals' mean score over sessions within each phase. Missing scores are left as missing.
[b] *SD*s are computed as the square root of the averaged variance of individuals' variances of scores within each phase. Missing scores are left as missing.
[c] *m* = number of participants or students.
[d] *n* = number of sessions.
[e] two-tailed *t*-test of Set A vs. Set B with *df* = 7.

Conclusions based on six assessments

The analyses summarized in Tables 1-9 and interpreted above collectively examined all data features recommended by the *WWC Handbook* (2013) for documenting an intervention effect. These assessments led to the same conclusion, as Lambert et al. (2006) did based on visual analysis alone. Next, we discuss the simplicity and rationality of the demonstrated approach, compared to visual analysis or complex statistical modeling and methods for determining intervention effects.

Discussion

In this paper, we demonstrated how to use free web-based resources or popular software to assess six data features recommended by the *WWC Handbook* (IES, 2013 February) to determine intervention effects in a single-case study (Lambert et al., 2006). The six data features are level and level change between phases, trend, variability, immediacy of the effect, overlap, and consistency of data in similar phases. Lambert et al. (2006) employed a reversal (or ABAB) design to collect data on the effectiveness of the report-card intervention in reducing students' disruptive behaviors in classrooms. The intervention was judged to be effective by Lambert et al. (2006) based on visual inspection alone. Our approach was to assess each of the six data features separately; then integrate six assessments into one comprehensive analysis of the intervention effect.

Among the six data features, the assessment of trends is probably most discussed but least agreed upon in the literature. To assess trends in the Lambert data, we employed the Page test and computed its ES and CI, proposed by Peng and Chen (2014). The Page test has been shown in the literature to be applicable to a variety of SCD contexts, such as, the simple AB designs, multiple-baseline AB designs, or replicated ABAB designs. They are equally applicable to one participant as well as to multiple participants, to one study as well as to multiple studies in a meta-analytic framework. The versatile Page test requires only ranked data. It can be computed and interpreted even when data have no variance (namely, there is uniformity in scores), display ceiling or floor effects, or are incomplete (Peng & Chen, 2014). Likewise, its proposed ES and CI are interpretable as they are direct derivatives from Page's *L* statistic. The proposed ES is a meaningful measure of intervention effects and its precision is expressed by the CI (Peng & Chen, 2014). Both ES and CI can be computed simply using SAS algorithms shown in Appendix A. The reporting of ES and its precision, expressed as CI, have been required or highly recommended by refereed journals and professional organizations, such as the American Psychological Association (APA) and American Educational Research Association (AERA) (AERA, 2006; APA Publications and Communications Board Working Group on Journal Article Reporting Standards, 2008; APA, 2010; Peng, Chen, Chiang, & Chiang, 2013).

The Lambert et al. (2006) data were recently reanalyzed in five articles published in a special issue of *Journal of School Psychology* (Shadish, 2014) to demonstrate alternative ways of analyzing and reporting SCD data, beyond the initial visual analysis. Each article published in that special issue employed complex statistical models (such as, the hierarchical linear modeling) and/or methods (such as, the Bayesian approach). These complex models and methods are often difficult to conceptualize or implement by practitioners not specially trained for these methodologies. In our demonstration, we assessed each of the six data features separately; then integrated six assessments into one comprehensive analysis. The separate assessments and the final integration were carried out using tools free from the Internet, or from the popular statistical software, such as SAS and SPSS. Thus, our approach to the determination of intervention effects is both simple and comprehensive. It illustrates how researchers, clinicians, teachers, parents, or policy makers can be empowered to interpret data efficiently and formulate evidence-based conclusions logically from well-designed and well-executed single-case studies.

Acknowledgements

This research was supported in part by two Maris M. Proffitt and Mary Higgins Proffitt Endowment Grants of Indiana University awarded to C.-Y. J. Peng while L.-T. Chen served as a research assistant on these two grants.

References

American Educational Research Association. (2006). Standards for reporting on empirical social science research in AERA publications. *Educational Researcher*, *35*, 33-40. doi:10.3102/0013189X035006033

American Psychological Association. (2010). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: Author.

APA Publications and Communications Board Working Group on Journal Articles Reporting Standards. (2008). Reporting standards for research in psychology: Why do we need them? What might they be? *American Psychologist*, *63*, 839-851. doi:10.1037/0003-066X.63.9.839

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology*, *31*, 144-152. doi:10.1111/j.2044-8317.1978.tb00581.x

Busk, P. L., & Marascuilo, L. A. (1988). Autocorrelation in single-subject research: A counterargument to the myth of no autocorrelation. *Behavioral Assessment*, *10*, 229-242.

Busk, P. L., & Marascuilo, L. A. (1992). Statistical analysis in single-case research: Issues, procedures, and recommendations, with applications to multiple behaviors. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case research design and analysis: New directions for psychology and education* (pp. 159-186). Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

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

Fahoome, G. (2002). Twenty nonparametric statistics and their large sample approximations. *Journal of Modern Applied Statistical Methods*, *1*(2), 248-268. Retrieved from http://digitalcommons.wayne.edu/jmasm/vol1/iss2/35

Horner, R. H., Carr, E. G., Halle, J., McGee, G., Odom, S., & Wolery, M. (2005). The use of single-subject research to identify evidence-based practice in special education. *Exceptional Children*, *71*(2), 165-179. Retrieved from http://journals.cec.sped.org/ec/

Institute of Education Sciences (2013 February). *What Works Clearinghouse: Procedures and standards handbook* (Version 3.0). Retrieved from http://ies.ed.gov/ncee/wwc/DocumentSum.aspx?sid=19

Lambert, M. C., Cartledge, G., Heward, W. L., & Lo, Y.-Y. (2006). Effects of response cards on disruptive behavior and academic responding during math lessons by fourth-grade urban students. *Journal of Positive Behavior Interventions*, *8*, 88-99. doi:10.1177/10983007060080020701

Lyerly, S. B. (1952). The average Spearman rank correlation coefficient. *Psychometrika*, *17*(4), 421-428.

Marascuilo, L. A., & Busk, P. L. (1988). Combining statistics for multiple-baseline AB and replicated ABAB designs across subjects. *Behavioral Assessment*, *10*, 1-28.

Marascuilo, L. A., & McSweeney, M. (1977). *Nonparametric and distribution-free methods for the social sciences*. Monterey, CA: Brooks/Cole Publishing Company.

Page, E. B. (1963). Ordered hypotheses for multiple treatments: A significance test for linear ranks. *Journal of the American Statistical Association*, *58*, 243-254. doi:10.2307/2282965

Parker, R. I., & Vannest, K. (2009). An improved effect size for single-case research: Nonoverlap of all pairs. *Behavior Therapy*, *40*, 357-367. doi:10.1016/j.beth.2008.10.006

Peng, C.-Y. J. & Chen, L.-T. (2014). *Assessing intervention effects in single-case studies using the Page test*. Manuscript retrieved from https://oncourse.iu.edu/access/content/user/peng/CI for page test.condensed.Peng_11_Chen_6_+tablefigure.7-15-2014.pdf (July 19th, 2014).

Peng, C.-Y. J., Chen, L.-T., Chiang, H.-M., & Chiang, Y.-C. (2013). The impact of APA and AERA guidelines on effect size reporting. *Educational Psychology Review, 25,* 157-209. doi:10.1007/s10648-013-9218-2

SAS Institute Inc. (2014). *SAS/STAT 13.1 user's guide*. Cary, NC: SAS Institute Inc.

Shadish, W. R. (Ed.) (2014). Analysis and meta-analysis of single-case designs [Special issue]. *Journal of School Psychology, 52*(2).

van de Wiel, M. A., & Di Bucchianico, A. (2001). Fast computation of the exact null distribution of Spearman's ρ and Page's L statistic for samples with and without ties. *Journal of Statistical Planning and Inference*, *92*, 133-145. doi:10.1016/S0378-3758(00)00166-X

Vannest, K. J., Parker, R. I., & Gonen, O. (2011). *Single Case Research: web based calculators for SCR analysis*. (Version 1.0) [Web-based application]. College Station, TX: Texas A&M University. Retrieved Thursday 10th July 2014. Available from http://www.singlecaseresearch.org

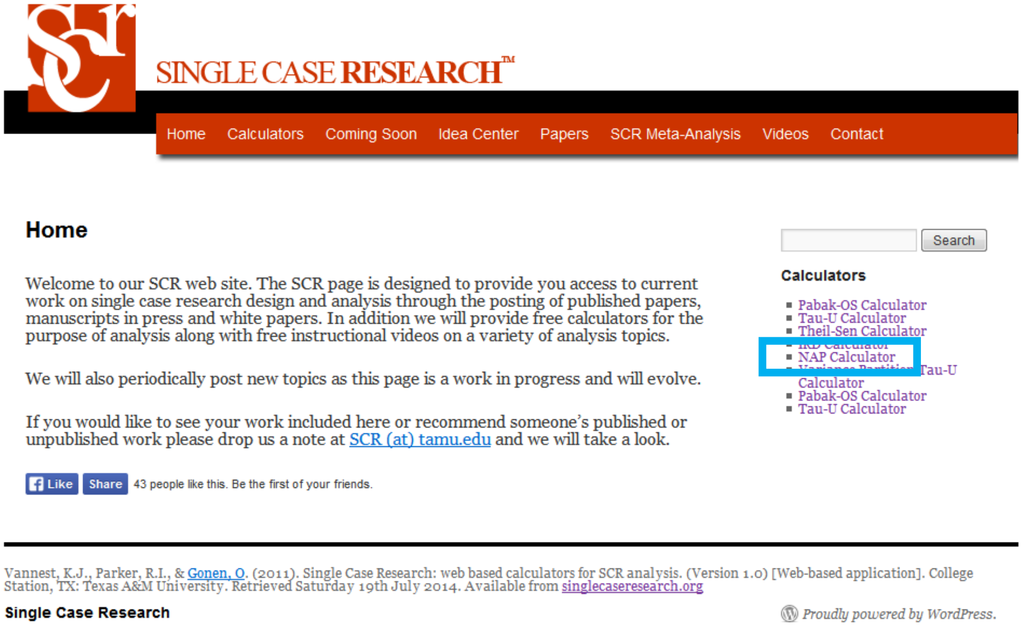
Appendix A: SAS Program for Assessing Level/Level Change, Trends, Variability, and Consistency in Lambert-A Data

```
*-----------------------------------------------------------------------------------------------
*Data came from Lambert et al., (2006) with two Classrooms, A and B.
*Class A data are analyzed in this program. Class B data can be analyzed similarly.
*Each class has two baselines, SSR1 and SSR2, each followed by an intervention: RC1 and RC2.
*Class A has 4 participants, A1 to A4 and 31 sessions: 1-8 in SSR1, 9-14 in RC1, 15-22 in SSR2,
* sessions 23-31 in RC2.
*Class B has 5 participants, B1 to B5 and 34 sessions: 1-10 in SSR1, 11-16 in RC1, 17-23 in SSR2,
* sessions 24-34 in RC2.
* 3 Page tests, their Chi-square tests, and p-values are computed in this program, for class A.
*
*----------------------------------------------------------------------------------------------;

OPTIONS LS=80 PAGENO=1;
TITLE ‘Lambert A Data analyzed using Page test’;

DATA A;                        /*Classroom A data of 4 students*/
     INPUT id $ score1-score31;

*Class A has 4 participants, A1 to A4 and 31 sessions: 1-8 in SSR1, 9-14 in RC1, 15-22 in SSR2, 23-31
in RC2;

     minssr1=min (OF score1-score8);
     maxrc1=max (OF score9-score14);
     minssr2 = min (OF score15-score22);
     maxrc2=max (OF score23-score31);

*Compute the mean of each student for each phase--------------------------------------;

     meanssr1=mean(OF score1-score8);
     meanrc1=mean(OF score9-score14);
     meanssr2=mean(OF score15-score22);
     meanrc2=mean(OF score23-score31);

*Compute differences of adjacent phases---------------------------------------------;

     diff_ssr1_rc1=meanssr1-meanrc1;
     diff_rc1_ssr2=meanrc1-meanssr2;
     diff_ssr2_rc2=meanssr2-meanrc2;

*Compute the variance of each student for each phase----------------------------------;

     varssr1=VAR (OF score1-score8);
     varrc1=VAR (OF score9-score14);
     varssr2=VAR (OF score15-score22);
     varrc2=VAR (OF score23-score31);

* Create new variables for single imputation missing data----------------------------;

     ARRAY score{*} score1-score31;
     ARRAY new{*} new1-new31;

     DO i = 1 to 31 by 1;
        new{i} = score{i};
     END;

DATALINES;
A1 7  9  8  6  7  4  5 10  2  0  .  1  0  0  8  8  8  6 10  10 10   8  3  4  1  3  2  4  0  1  0
A2 8  7  .  7  8  6  7  9  3  1  0  4  0  0  8  9 10  7  9  10  8  10  1  1  0  5  3  6  0  0  2
A3 10 .  6  .  6  9  6 10  .  0  1  1  0  0  5  7 10  .  5  10  9  10  4  6  5  7  0  0  0  0  .
A4 10 .  6  4  8  8  9 10  3  6  0  0  .  1  3  8 10  . 10  10 10   5  6  1  5  0  .  .  0  0  1
;

* Compute descriptive stat. in the data set-------------------------------------------;
* Part A -------------------------------------------------------------;

PROC MEANS DATA=A; RUN;

* Replace missing scores in each phase by the min or max of that phase for each participant from
Lambert-A data set -------------;

DATA A1; SET A;
         ARRAY score{*} score1-score31;
         ARRAY new{*} new1-new31;
         Do i = 1 to 31 by 1;
            session = i;
                IF 1<= session <= 8 THEN phase = 'SSR1';
                ELSE IF 9 <=session <=14 THEN phase = 'RC1';
                ELSE IF 15<= session<=22 THEN phase = 'SSR2';
                ELSE IF 23<=session <=31 THEN phase = 'RC2';

                IF phase = 'SSR1' AND new{i}=. THEN new{i}=minssr1;
                ELSE IF phase = 'RC1' AND  new{i}=. THEN new{i}=maxrc1;
                    ELSE IF phase = 'SSR2' AND new{i}=. THEN new{i}=minssr2;
                ELSE IF phase = 'RC2' AND  new{i}=. THEN new{i}=maxrc2;

         END;
         KEEP id new1-new31;


*Create three data sets for two adjacent phases for Lambert A data set -------------------------;


DATA A_SSR1_RC1; set A1; KEEP id new1-new14;
DATA A_SSR2_RC2; set A1; KEEP id new15-new31;
DATA A_RC1_SSR2; set A1; KEEP id new9-new22;

*Rank data in SSR1-RC1 phases from SAS data set A_SSR1_RC1 of Lambert A data set ---------------;

PROC TRANSPOSE DATA=A_SSR1_RC1 OUT=Table1;
/* transpose the data matrix in order to rank scores*/
     ID id;
RUN;

PROC RANK DATA=Table1 OUT=Table1;
     VAR A1-A4;

PROC TRANSPOSE DATA=Table1 OUT=Table1 PREFIX=rank;   /* transpose the ranked data back */

* Compute total ranks for 14 sessions in SSR1-RC1 phases from SAS data set A_SSR1_RC1 of Lambert A
data set ----------;

PROC MEANS DATA=Table1;                           /* compute the total of rank1 to rank14 */
     VAR rank1-rank14;
     OUTPUT OUT=Table1 SUM=sum1-sum14;

PROC PRINT DATA = Table1; RUN;

* Part B-------------------------------------------------------------------;

* Compute Page L, chi-square, z and CI of z for SSR1-RC1 phases from Lambert A data set -------;

*Page test for SSR1-RC1 phases in Lambert A data set-------------------------------------------;

DATA L_1; SET Table1;

L1 =
14*sum1+13*sum2+12*sum3+11*sum4+10*sum5+9*sum6+8*sum7+7*sum8+6*sum9+5*sum10+4*sum11+3*sum12+2*sum13+1*
sum14;   /*Page L stat */
m = 4;
n = 14;
n1= n+1;
p = (n1)**2;        /* n+1 squared */
```

```
q = n**2;           /* n squared */
q1 = n**2 - 1;      /* n squared -1 */

Chi1= ((12*L1 - 3*m*n*p)**2)/((m*q)*q1*n1);
chi_p1 = probchi(Chi1,1);
z1 = sqrt(chi1);    /* 95% of z CI for L1  */

z1_lower = z1-1.96;  /* Lower bound of z    */
z1_upper = z1+1.96;  /* Upper bound of z    */

PROC PRINT DATA=L_1; RUN;

*Page test for SSR2-RC2 phases in Lambert A data set------------------------------------------;

*Rank data in SSR2-RC2 phases from SAS data set A_SSR2_RC2 of Lambert A data set -------------;

PROC TRANSPOSE DATA=A_SSR2_RC2 OUT=Table2;
/* transpose the data matrix in order to rank scores*/
     ID id;
RUN;

PROC RANK DATA=Table2 OUT=Table2;
     VAR A1-A4;RUN;

PROC TRANSPOSE DATA=Table2 OUT=Table2 PREFIX=rank;   /* transpose the ranked data back */

* Compute total ranks for 17 sessions in SSR2-RC2 phases from SAS data set A_SSR2_RC2 of Lambert A
data set ----------;

PROC MEANS DATA=Table2;                           /* compute the total of rank23 to rank31 */
     VAR rank1-rank17;
     OUTPUT OUT=Table2 SUM=sum15-sum31;

PROC PRINT DATA = Table2; RUN;

* Compute Page L, chi-square, z and CI of z for SSR2-RC2 phases from Lambert A data set --------;

DATA L_2; SET Table2;

L2 =
17*sum15+16*sum16+15*sum17+14*sum18+13*sum19+12*sum20+11*sum21+10*sum22+9*sum23+8*sum24+7*sum25+6*sum2
6+5*sum27+4*sum28+3*sum29+2*sum30+1*sum31;   /*Page L stat */
m = 4;
n = 17;
n1= n+1;
p = (n1)**2;        /* n+1 squared */
q = n**2;           /* n squared */
q1 = n**2 - 1;      /* n squared -1 */

Chi2= ((12*L2 - 3*m*n*p)**2)/((m*q)*q1*n1);
chi_p2 = probchi(Chi2,1);
z2 = sqrt(chi2);    /* 95% of z CI for L2  */

z2_lower = z2-1.96;  /* Lower bound of z2    */
z2_upper = z2+1.96;  /* Upper bound of z2    */

PROC PRINT DATA=L_2; RUN;

*Page test for RC1-SSR2 phases in Lambert A data set-------------------------------------------;

*Rank data in RC1-SSR2 phases from SAS data set A_RC1_SSR2 of Lambert A data set --------------;

PROC TRANSPOSE DATA=A_RC1_SSR2 OUT=Table3;
/* transpose the data matrix in order to rank scores*/
     ID id;
RUN;

PROC RANK DATA=Table3 OUT=Table3;
     VAR A1-A4;
```

```
PROC TRANSPOSE DATA=Table3 OUT=Table3 PREFIX=rank;   /* transpose the ranked data back */

* Compute total ranks for 14 sessions in RC1_SSR2 phases from SAS data set A_RC1_SSR2 of Lambert A
data set ----------;

PROC MEANS DATA=Table3;                           /* compute the total of rank9 to rank22 */
     VAR rank1-rank14;
     OUTPUT OUT=Table3 SUM=sum9-sum22; RUN;

PROC PRINT DATA = Table3; RUN;

* Compute Page L, chi-square, z and CI of z for SSR2-RC2 phases from Lambert A data set --------;

DATA L_3; SET Table3;

L3 =
1*sum9+2*sum10+3*sum11+4*sum12+5*sum13+6*sum14+7*sum15+8*sum16+9*sum17+10*sum18+11*sum19+12*sum20+13*s
um21+14*sum22;     /*Page L stat */
m = 4;
n = 14;
n1= n+1;
p = (n1)**2;        /* n+1 squared */
q = n**2;           /* n squared */
q1 = n**2 - 1;      /* n squared -1 */

Chi3= ((12*L3 - 3*m*n*p)**2)/((m*q)*q1*n1);
chi_p3 = probchi(Chi3,1);
z3 = sqrt(chi3);    /* 95% of z CI for L3  */

z3_lower = z3-1.96;  /* Lower bound of z3    */
z3_upper = z3+1.96;  /* Upper bound of z3    */


PROC PRINT DATA=L_3; RUN;
```

Appendix B: Assessing Overlap in Lambert-A Data Using a Web-Based Calculator (Vannest, Parker, & Gonen, 2011) at http://singlecaseresearch.org



Home

Welcome to our SCR web site. The SCR page is designed to provide you access to current work on single case research design and analysis through the posting of published papers, manuscripts in press and white papers. In addition we will provide free calculators for the purpose of analysis along with free instructional videos on a variety of analysis topics.

We will also periodically post new topics as this page is a work in progress and will evolve.

If you would like to see your work included here or recommend someone's published or unpublished work please drop us a note at SCR (at) tamu.edu and we will take a look.

43 people like this. Be the first of your friends.

Calculators

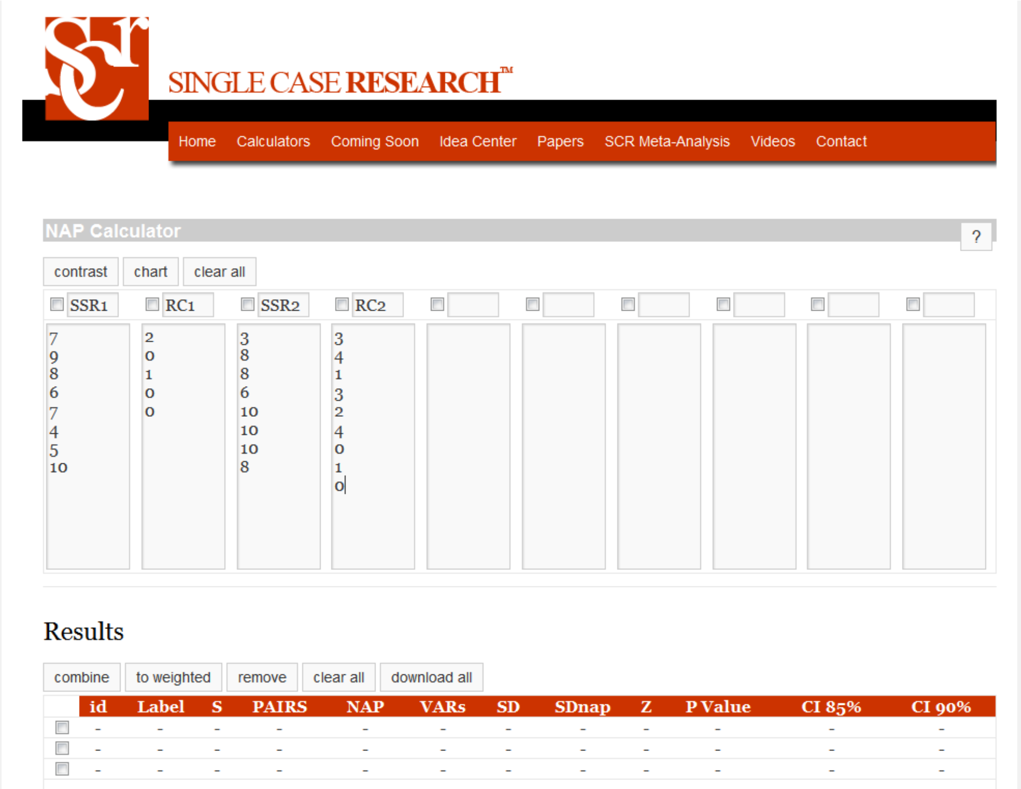
- Pabak-OS Calculator
- Tau-U Calculator
- Theil-Sen Calculator
- IRD Calculator
- NAP Calculator
- Variance Partition Tau-U Calculator
- Pabak-OS Calculator
- Tau-U Calculator

Vannest, K.J., Parker, R.I., & Gonen, O. (2011). Single Case Research: web based calculators for SCR analysis. (Version 1.0) [Web-based application]. College Station, TX: Texas A&M University. Retrieved Saturday 19th July 2014. Available from singlecaseresearch.org

Single Case Research

Proudly powered by WordPress.

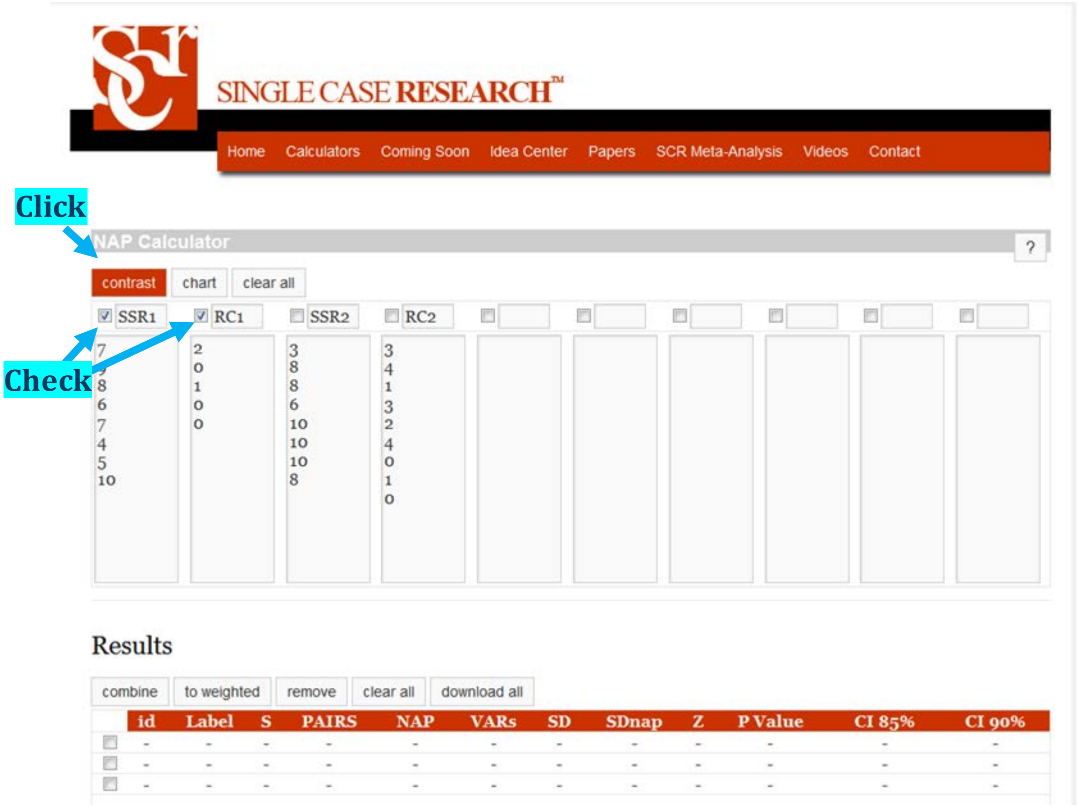
Figure B1. Web-based calculator for single-case studies developed by Vannest et al. (2011)



☐ SSR1	☐ RC1	☐ SSR2	☐ RC2	☐	☐	☐	☐	☐	☐
7 9 8 6 7 4 5 10	2 0 1 0 0	3 8 8 6 10 10 10 8	3 4 1 3 2 4 0 1 0						

	id	Label	S	PAIRS	NAP	VARs	SD	SDnap	Z	P Value	CI 85%	CI 90%
☐	-	-	-	-	-	-	-	-	-	-	-	-
☐	-	-	-	-	-	-	-	-	-	-	-	-
☐	-	-	-	-	-	-	-	-	-	-	-	-

Figure B2. Data entry for student A1 of Lambert-A data set in web-based calculator for single-case studies developed by Vannest et al. (2011)

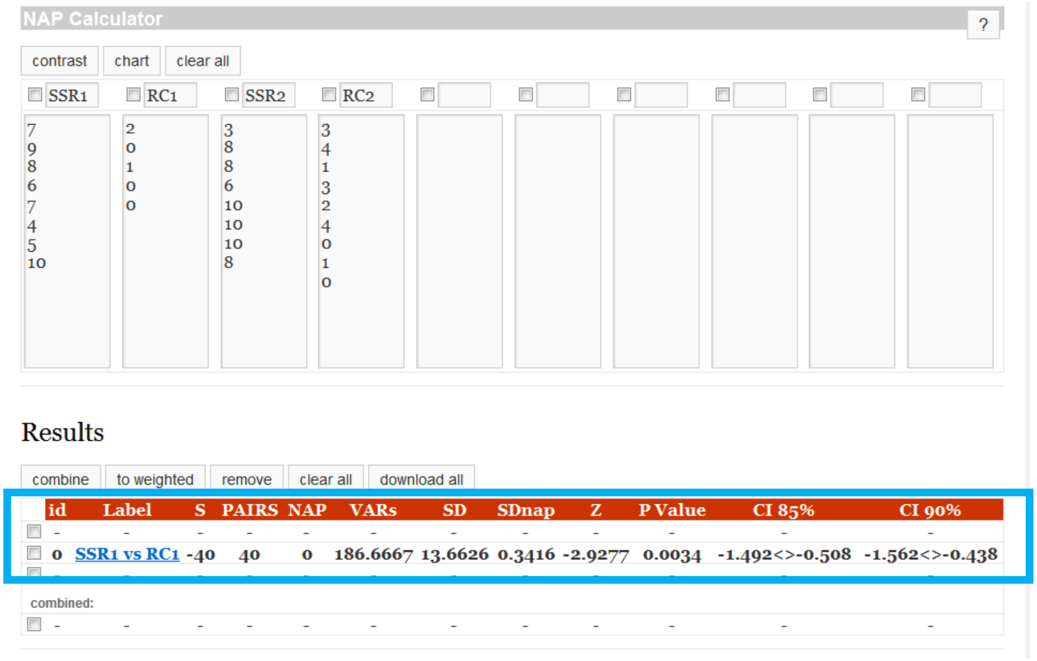


☑ SSR1	☑ RC1	□ SSR2	□ RC2	□	□	□	□	□	□
7 9 8 6 7 4 5 10	2 0 1 0 0	3 8 8 6 10 10 10 8	3 4 1 3 2 4 0 1 0						

Results

	id	Label	S	PAIRS	NAP	VARs	SD	SDnap	Z	P Value	CI 85%	CI 90%
□	-	-	-	-	-	-	-	-	-	-	-	-
□	-	-	-	-	-	-	-	-	-	-	-	-
□	-	-	-	-	-	-	-	-	-	-	-	-

Figure B3. Compute NAP of SSR1-RC1 for student A1 of Lambert-A data set in web-based calculator for single-case studies developed by Vannest et al. (2011)

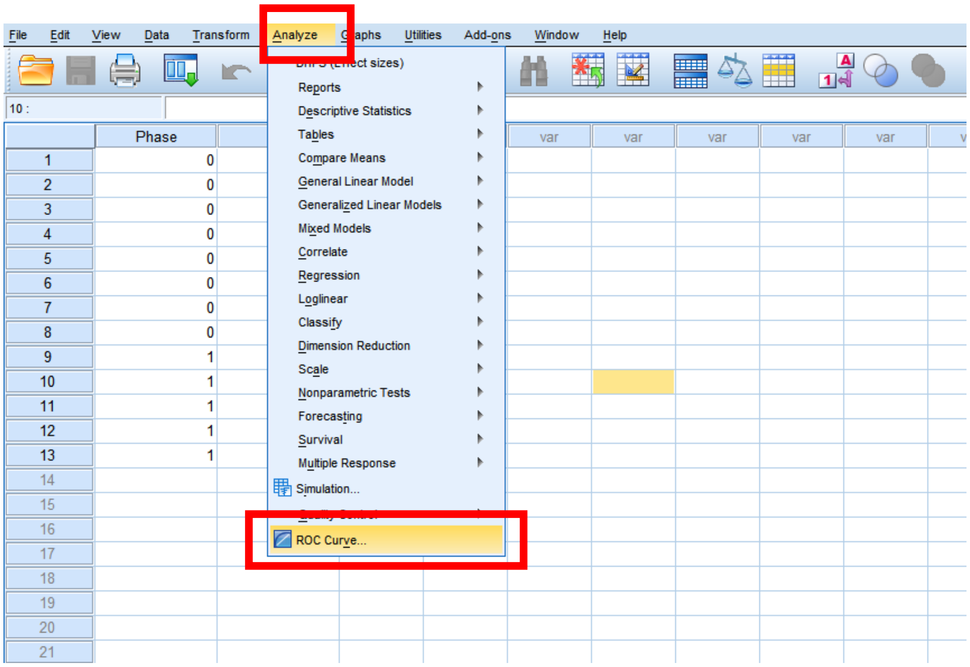


□ SSR1	□ RC1	□ SSR2	□ RC2	□	□	□	□	□	□
7	2	3	3						
9	0	8	4						
8	1	8	1						
6	0	6	3						
7	0	10	2						
4		10	4						
5		10	0						
10		8	1						
			0						

	id	Label	S	PAIRS	NAP	VARs	SD	SDnap	Z	P Value	CI 85%	CI 90%
□	-	-	-	-	-	-	-	-	-	-	-	-
□	0	SSR1 vs RC1	-40	40	0	186.6667	13.6626	0.3416	-2.9277	0.0034	-1.492<>-0.508	-1.562<>-0.438
□	-	-	-	-	-	-	-	-	-	-	-	-
combined:												
□	-	-	-	-	-	-	-	-	-	-	-	-

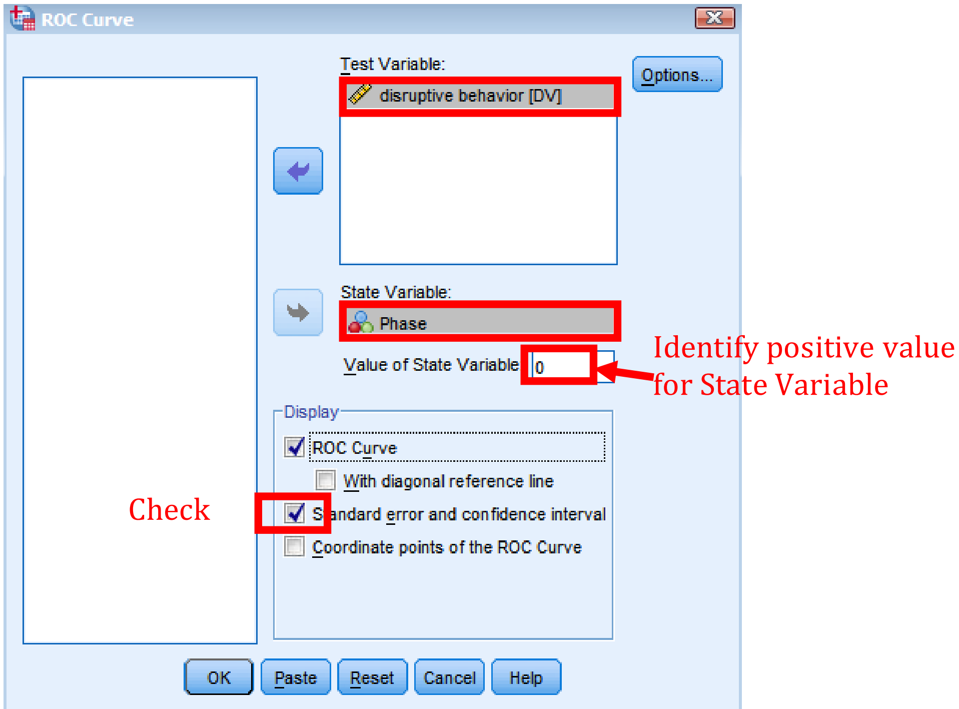
Figure B4. Obtain NAP of SSR1-RC1 for student A1 of Lambert-A data set from web-based calculator for single-case studies developed by Vannest et al. (2011)

Appendix C: Assessing Overlap in Lambert-A Data Using SPSS (Version 21)



	Phase
1	0
2	0
3	0
4	0
5	0
6	0
7	0
8	0
9	1
10	1
11	1
12	1
13	1
14	
15	
16	
17	
18	
19	
20	
21	

Figure C1. Compute NAP using SPSS Receiver Operator Characteristics module



Test Variable: disruptive behavior [DV]

State Variable: Phase

Value of State Variable: 0

Identify positive value for State Variable

Display

☑ ROC Curve

☐ With diagonal reference line

☑ Standard error and confidence interval

☐ Coordinate points of the ROC Curve

Check

Figure C2. Dialogue window after selecting ROC Curve

Area Under the Curve

Test Result Variable(s): disruptive behavior

			Asymptotic 95% Confidence Interval	
Area	Std. Error[a]	Asymptotic Sig.[b]	Lower Bound	Upper Bound
NAP 1.000	.000	***p*-value** .003	1.000	1.000

a. Under the nonparametric assumption

b. Null hypothesis: true area = 0.5

Figure C3. Obtain NAP of SSR1-RC1 for student A1 of Lambert-A data set from SPSS 21.0

Journal of Modern Applied Statistical Methods
May 2015, Vol. 14, No. 1, 308-334.

ISSN 1538 − 9472

JMASM Algorithms and Code
Pseudo-Random Number Generators for Vector Processors and Multicore Processors

Agner Fog
Technical University of Denmark
Ballerup, Denmark

Large scale Monte Carlo applications need a good pseudo-random number generator capable of utilizing both the vector processing capabilities and multiprocessing capabilities of modern computers in order to get the maximum performance. The requirements for such a generator are discussed. New ways of avoiding overlapping subsequences by combining two generators are proposed. Some fundamental philosophical problems in proving independence of random streams are discussed. Remedies for hitherto ignored quantization errors are offered. An open source C++ implementation is provided for a generator that meets these needs.

Keywords: Random number generation, SIMD, vector processors, multiprocessors, parallel generation, combination of generators, quantization errors, theoretical proofs, philosophy of science

Introduction

The exponential increase in the computing power of mainstream microprocessors over several decades, known as Moore's Law, has made large scale Monte Carlo applications feasible and common. The current trend in microprocessor technology goes towards parallel processing of data in mainly two ways: 1) microprocessors have vector registers that can do arithmetic operations on a whole vector with a single CPU instruction (Single Instruction Multiple Data, SIMD), and 2) microprocessor chips have multiple CPU cores that can execute multiple threads simultaneously. The design of pseudo-random number generators (PRNGs) has been improved considerably in recent decades, but few of the published designs are suitable for utilizing the parallel processing capabilities of

Dr. Fog is a Researcher at the Technical University of Denmark. Email him at agfo@dtu.dk.

today's microprocessors in large scale computations (Manssen, et al., 2012; Passerat-Palmbach, Mazel and Hill, 2011). The construction of pseudo-random number generator software capable of utilizing both vector processing and multithreading for the fast generation of large amounts of pseudo-random numbers of high quality, using the newest microprocessor technology are considered.

Choice of hardware

Several hardware platforms are available for parallel processing:

Mainstream CPUs for the PC market

These CPUs are quite powerful. They are universally available and cheap because of high production volumes. The size of vector registers in the common x86 family of microprocessors has grown exponentially in recent years, as illustrated in Table 1.

Table 1. Vector register size of x86 family microprocessors.

Year introduced	Instruction set for integer vector operations	Vector size, bits
1997	MMX	64
2001	SSE2	128
2013	AVX2	256
expected 2017	AVX-512	512

Vector sizes of 1024 bits and perhaps 2048 bits can be expected in mainstream CPUs in the coming years. However, the vector size will probably not keep growing exponentially because of diminishing returns and because the size of mask registers used for conditional execution is limited to 64 bits, corresponding to 64 elements of 32 bits each = 2048 bits, in current specifications from Intel (Intel, 2014a).

The high-end CPUs are currently available with 8 or more cores and a clock frequency of 3 – 4 GHz. Some models are capable of running two threads in each core, but this may not be useful for CPU-intensive code because both threads are competing for the same hardware resources (Fog, 2014a).

Graphic processors.

Graphics Processing Units (GPUs) are included in many PCs and designed mainly for the purpose of computer games. Contemporary GPUs are available in many different configurations with hundreds or thousands of parallel streams and clock frequencies ranging from 200 to 1600 MHz. GPUs have increasingly been applied to general computation tasks that involve large amounts of parallel data. Software libraries for random number generation in GPUs are available (Manssen, et al., 2012; Demchik, 2011; Barash and Shchur, 2014; Nandapalan, et al., 2012).

A serious limitation of GPUs is that each stream has access to only a small amount of RAM memory, and communication between streams is expensive. We have to consider that random number generation is typically only a small part of an application, using only a small part of the total CPU time. The other parts of a typical application, the ones that consume the random numbers, will typically be running in the same units that produced the random numbers and be subject to the same limitations on memory use and communication between streams. This is limiting the usefulness of GPUs for large scale Monte Carlo applications.

Many-core coprocessors

Intel's current Many Integrated Core (MIC) Xeon Phi coprocessor codenamed Knights Corner has up to 61 cores with 512-bit vector registers and a clock frequency of 1.2 GHz (Chrysos, 2012). The throughput per core is much lower than for a general purpose CPU, and the total throughput is rarely more than a few times the throughput of the best mainstream CPU configurations. In some cases, a mainstream CPU can even outperform the Knights Corner (Saule, Kamer and Çatalyürek, 2013; Chan, 2013; Karpiński 2014). The Knights Corner has its own instruction set, which makes it less attractive for portable software. The announced successor, codenamed Knights Landing, is expected to be faster and it will be using the same instruction set (AVX-512) as future mainstream CPUs (Anthony, 2013). This will make it possible to use the same software on MIC processors and mainstream CPUs.

Similar products from other vendors include Nvidia Tesla and AMD FireStream. These processors have more in common with GPUs.

Large vector processors

For most applications, clusters of general microprocessors have largely replaced the large and expensive supercomputers that were used decades ago for demanding scientific purposes.

Parallel generation of pseudo-random numbers in vector processors

A PRNG generally uses a generating function f of the form (L'Ecuyer, 1994)

$$x_i = \mathrm{f}\left(x_{i-1}, x_{i-2}, \ldots, x_{i-n}\right)$$

where each new value x_i is a function of the previous n values. The successive values x_i may be used directly as random numbers, or they may be transformed by an output function g of the form

$$y_i = \mathrm{g}\left(x_i, x_{i-1}, \ldots, x_{i-n}\right)$$

Not all of the values x_{i-1}, x_{i-2}, ..., x_{i-n} need to be included in f. We will say that f has a feedback path of length φ if f depends on $x_{i-\varphi}$. The function f can be implemented in a vector processor with registers of size v bits if $v \leq w\varphi$ for all feedback paths φ, where w is the number of bits needed to represent each x_i. For example, for a vector size v of 256 bits and a word size w of 32 bits, the shortest feedback path φ must be at least 8 for an efficient vectorized implementation of f. If $\varphi \geq 8$ and $n \geq 8$ then we can calculate 8 successive values of x_i with a vectorized function **f** of the form:

$$\left(x_{i+7}, x_{i+6}, \ldots, x_i\right) = \mathbf{f}\left(x_{i-1}, x_{i-2}, \ldots, x_{i-n}\right)$$

If $v > w\varphi$ then the vectorized function **f** needs to implement multiple steps of the generating function f. This is usually so complicated that it offsets the advantage of vectorized calculation.

The last n values of x_i are stored in a circular buffer, called the state buffer, which is updated by each call of the generating function f or **f**. The initial value of the state buffer is a function of an arbitrary number called the seed. This function is the so-called seeding procedure.

The size of the state buffer is at least wn and often extended to the nearest multiple of the vector size v. The implementation is most efficient if $w\varphi$ and wn are multiples of the vector size v.

Most of the commonly used PRNGs have a feedback path $\varphi = 1$, which makes them unsuited for vectorized calculation. Preferred generators are those with feedback paths corresponding to the largest vector size there is access to in available vector processors. A generator designed to match 128-bit vector registers has been published under the name SIMD-oriented Fast Mersenne Twister (SFMT) (Saito and Matsumoto, 2008, 2009).

Parallel generation of pseudo-random numbers in independent streams

The construction of generators suitable for vector processors has received relatively little attention in the literature, but the simultaneous generation of multiple pseudo-random streams has been discussed in several publications. Five different methods for producing independent streams have been proposed (L'Ecuyer, 1994; Salmon, 2011; L'Ecuyer, Oreshkin and Simard, 2014; Bauke and Mertens, 2007):

1. Use multiple instances of the same generator with different seeds. We want to avoid overlap between the generated subsequences. Assume that we are generating k subsequences of length ℓ from a generator with total cycle length ρ. If the seeding procedure is sufficiently random then we can calculate the probability that any of the subsequences are overlapping as (L'Ecuyer, Oreshkin and Simard, 2014)

   $$p \approx 1-\left(1-k\ell/\rho\right)^{k=1} \approx k^2\ell/\rho$$

   If the total cycle length ρ is sufficiently long then this probability can be very small. For example, for a Mersenne Twister MT19937 (Matsumoto and Nishimura, 1998) with cycle length $\rho = 2^{19937}-1$, $k = 1000$ and $\ell = 10^{10}$, we have $p = 2\cdot 10^{-5986}$. This means that we can safely ignore the risk of overlapping subsequences in such cases.
2. Use a generator with a jump-ahead feature. We use this jump-ahead feature to generate each stream as a subsequence of the same generator at an offset $q \geq \ell$ relative to the preceding stream (L'Ecuyer, 1994; L'Ecuyer and Côté, 1991). The jump-ahead feature is usually quite complicated and requires a significant amount of computing resources. Regularly spaced starting points may cause inferior randomness for some generators (Durst, 1989).

3. A variant of the jump-ahead method is to put all the randomness in the output function g, while the generating function f is a simple counting $x_i = x_{i-1} + 1 \bmod 2^w$ (Salmon, 2011). This makes it trivial to generate non-overlapping subsequences. The output function g is borrowed from cryptology. Instructions for AES encryption are implemented in hardware in many computers, using a vector size of 128 bits, but not higher (Intel, 2014b).
4. Leapfrogging. The first of k streams uses outputs x_i, x_{i+k}, x_{i+2k}, ... The next stream uses x_{i+1}, x_{i+1+k}, x_{i+1+2k}, ... and so on. This is useful when the k streams form a vector generated by a single vector processor. It is more complicated to use leapfrogging when the streams are generated in separate processors. Known multiprocessor implementations use prime modulus (Bauke and Mertens, 2007), which leads to quantization errors (see below).
5. Use different generators based on the same principle but with different sets of parameters in the generating function. If we have many streams then we need to either store many pre-calculated parameter sets, or include the necessary code to search for good parameter sets on the fly (Matsumoto and Nishimura, 2000). This so-called dynamic creation method requires a lot of computational resources, possibly even more than the resources needed to generate the random number streams, and it has been reported to make inferior parameter sets in some cases (Passerat-Palmbach, Mazel, Mahul and Hill, 2010).

There is disagreement among theorists about whether method 5 can be recommended. One would intuitively assume that random streams generated by different generators with different parameter sets are statistically independent, but some have argued that we have no theoretical proof that there is no unwanted correlation between such random streams (Passerat-Palmbach, Mazel and Hill, 2011; L'Ecuyer, 1994). However, those who make this objection seem to ignore that the same argument can be made about subsequences from the same generator. Perhaps they rely on the implicit (and arguably false) assumption that the most recommended generators are perfect, and conclude that non-overlapping subsequences from the same generator are statistically independent.

However, if subsequences are spaced by an offset of e.g. $q = 10^{15}$ and experimental tests for randomness have included no sequences longer than $\ell = 10^{10}$ then we have no experimental proof that all subsequences are independent, and no theoretical proof either (Bauke & Mertens, 2007). It is reasonable to assume that the probability of unwanted correlations between sequences from different generators (with different seeds) is not bigger than the probability of unwanted correlations between subsequences of the same generator. We will return to a more general discussion of theoretical proofs below.

6. A sixth method of making independent pseudorandom streams is now proposed. It involves the combination of two different PRNGs. We will have two different generators, G and H, and initialize them with seeds $s^1{}_{\mathrm{G}}$ and $s^1{}_{\mathrm{H}}$, respectively. G generates a pseudorandom sequence $x^1{}_{\mathrm{G}i}$ and H makes another sequence $x^1{}_{\mathrm{H}i}$, where each x is an integer of w bits, and $0 \leq i < \ell$. The two sequences are now combined into one stream by means of a bitwise XOR operation or addition modulo 2^w, e.g. $x^1{}_i = x^1{}_{\mathrm{G}i} + x^1{}_{\mathrm{H}i} \bmod 2^w$. The combined stream $x^1{}_i$ now depends on both seeds $s^1{}_{\mathrm{G}}$ and $s^1{}_{\mathrm{H}}$. We can make a second combined stream (indicated by superscript 2) $x^2{}_i$ by changing the seed for G, $s^1{}_{\mathrm{G}}$ to $s^2{}_{\mathrm{G}}$ and keeping the seed for H constant: $s^1{}_{\mathrm{G}} \neq s^2{}_{\mathrm{G}} \wedge s^1{}_{\mathrm{H}} = s^2{}_{\mathrm{H}}$. The second combined stream is $x^2{}_i = x^2{}_{\mathrm{G}i} + x^2{}_{\mathrm{H}i} = x^2{}_{\mathrm{G}i} + x^1{}_{\mathrm{H}i} \bmod 2^w$. Now consider the unlikely event that the seed $s^2{}_{\mathrm{G}}$ generates a sequence $x^2{}_{\mathrm{G}i}$ that is offset from $x^1{}_{\mathrm{G}i}$ by a distance $q < \ell$, perhaps because of a bad seeding procedure. In this case, the sequences $x^1{}_{\mathrm{G}i}$ and $x^2{}_{\mathrm{G}i}$ have a partial overlap of length $\ell - q$ because $x^2{}_{\mathrm{G}i} = x^1{}_{\mathrm{G}i+q}$. However, the contribution from H is $x^2{}_{\mathrm{H}i} = x^1{}_{\mathrm{H}i} \neq x^1{}_{\mathrm{H}i+q}$, except for random i-occurrences with expected frequency 2^{-w}. Therefore, the first and second combined sequences $x^1{}_i$ and $x^2{}_i$ will be statistically independent, even in the unlucky event that the G component of the sequences have a partial overlap.
7. A variant of method 6 is to change both seeds: $s^1{}_{\mathrm{G}} \neq s^2{}_{\mathrm{G}} \wedge s^1{}_{\mathrm{H}} \neq s^2{}_{\mathrm{H}}$. To see if this method is safe from overlaps, consider the coincidence of three unlucky events: 1) The sequence $x^2{}_{\mathrm{G}i}$ is offset from $x^1{}_{\mathrm{G}i}$ by a distance $|q_{\mathrm{G}}| < \ell$ so that the G-sequences have a partial overlap; 2) the sequence $x^2{}_{\mathrm{H}i}$ is offset from $x^1{}_{\mathrm{H}i}$ by a distance $|q_{\mathrm{H}}| < \ell$ so that the H-sequences have a partial overlap; and 3) the two overlaps are equal $q_{\mathrm{G}} = q_{\mathrm{H}}$. The two combined sequences $x^1{}_i$ and $x^2{}_i$ have a partial overlap only in this contrived scenario. This is a theoretical possibility, but it can only happen at the coincidence of three unlucky events, all of which are extremely unlikely. The probability of this coincidence happening between any of k

> combined sequences is approximately $k^2\ell / (\rho_G\rho_H)$ where ρ_G and ρ_H are the cycle lengths of G and H, respectively. With large cycle lengths, this probability is so low that there is room for human errors. Even in the event that both seeding procedures are seriously flawed, the coincidence of the three unlikely events seems no more than a theoretical possibility.

Method 7 has the advantage that the difference between two combined streams $d_i = x^2{}_i - x^1{}_i$ depends on both generators G and H, while d_i depends only on G if method 6 is used. This gives improved randomness in applications where differences between streams are involved. The possible improvement in randomness by combining two different generators is discussed in the next section.

Advantages of combined generators

The technique of combining two or more PRNGs is often used in order to improve randomness and cycle length. The cycle length of a combined generator is the least common multiple of the cycle lengths of the individual generators.

There are different opinions on the merits of combining two or more PRNGs. L'Ecuyer has argued that the combined output of two generators may conceivably be less random than the individual sequences (L'Ecuyer, 1990, 1994), while the acknowledged handbook *Numerical Recipes* emphasizes: "An acceptable random generator must combine at least two (ideally unrelated) methods" (Press, 2007, p. 342).

The combination of two random streams can only be less random than its components if the two streams are correlated in a certain way. The next section will discuss whether it is possible to prove that such an unfortunate correlation between two random streams does not exist.

It has been observed that the combination of two or more PRNGs produces a stream that is more random than either component. In fact, many good random generators have been made by combining inferior ones. Pragmatically speaking, we may say that if generator G has some defects and generator H has some other defects, then the combination of G and H has neither of these defects, as long as the defects of G and H are of different kinds. This is not a universal law of nature, of course, and it requires a more specific analysis to determine whether a particular kind of defect can be eliminated by combination of generators. There is plenty of theoretical evidence that various defects in random generators can be eliminated by combining with other generators that do not have the same kind of defects (Matsumoto and Nishimura, 2000; Deng, Lin, Wang and Yuan, 1997; L'Ecuyer and Granger-Piché 2003; Marsaglia, 1985). Experience shows that combining two generators is a very efficient way of improving randomness. For example, if generator G has a bias that makes certain values more frequent than others, and generator H has no such bias, then the combined output of G and H will have no bias. If Generator H has a correlation between subsequent numbers and generator G has no such correlation, then the combined output will be free from such correlations. The two generators should preferably be very different in their design in order to avoid that they both have the same kinds of defects (Press, 2007).

Combining two or more generators is also useful in applications where security is important. It is possible to reconstruct a complete sequence from a subsequence in many generators. This becomes very difficult or impossible when multiple generators are combined and only the combined output is accessible to the attacker.

How much can be proven?

It has been argued above that it is unreasonable to demand a theoretical proof that streams from different PRNGs are uncorrelated as long as we cannot even prove the same thing for different substreams of the same generator. This opens up a much more general discussion about what kind of proofs are actually possible in relation to PRNGs. There are three kinds of claims that we would like to prove for generators:

a) A particular generator G has no unwanted correlation with an application A, i.e. a correlation that would make A produce results that are significantly different from what perfectly random numbers would give.

b) There is no correlation between non-overlapping subsequences from the same generator G.

c) There is no correlation between the outputs of two different generators G and H.

Claims of type (a) are made implicitly or explicitly whenever a particular PRNG is recommended. Such claims may later be falsified when a particular weakness in a generator is discovered. For example, Linear congruential generators which have been widely used in commercial software were found after many years to have serious defects (Entacher, 1998). The popular and often recommended Mersenne Twister has the flaw that it can produce long sequences with more 0's than 1's if it comes into a state where the state buffer contains mostly 0's. This flaw was reported only after the Mersenne Twister had been the preferred generator for several years (Saito and Matsumoto, 2008). A tiny bias in the Multiply-with-carry generators was discovered a few years after this kind of generators had been recommended (Couture and L'Ecuyer, 1997). In fact, one defect reported by Bauke and Mertens (2004) applies to a large part of all known PRNGs.

The possibility cannot be ruled out that more such discoveries will be made in the future, no matter how good we believe that our generators are. Claims that a PRNG is good should therefore be regarded as falsifiable propositions in accordance with Popper's (1963) philosophy of science. The claim that a generator produces random output is never true in the strictest sense, because the output is deterministic. It may be proven experimentally that the output of a PRNG passes certain tests for randomness, but the possibility that it will fail some test if a larger sample size is used cannot be ruled out. If the sample size is increased to the entire cycle length then the total sample is no longer random because, typically, all output values occur the same number of times in a full cycle.

In science, theoretical proofs are often regarded as stronger than experimental proofs. However, for PRNGs there is a dilemma. If it is possible to prove theoretically that a PRNG has a certain desirable property, then the theoretical insight that allowed this analysis may also be used in the construction of an experimental test that defeats the same generator. For example, the construction of generators in the Mersenne Twister family usually relies on the Berlekamp-Massey algorithm for verification of the cycle length (Saito and Matsumoto, 2008). Therefore, it is no surprise that the Mersenne Twisters fail a test based on the Berlekamp-Massey algorithm, the so-called linear complexity test (L’Ecuyer and Simard, 2007). If a chaotic behavior with no recognizable mathematical structure is what characterizes a good PRNG, then perhaps the best generators are the ones that are most difficult to prove good (Fog, 2001). On the other hand, attempts to produce PRNGs without any theory have led to very bad results (Knuth, 1998).

Claims of type (a) are generally the easiest to falsify. Most of the generators described in the literature have weaknesses that have been discovered by either experimental of theoretical methods.

Claims of type (b) have occasionally been falsified. Durst (1989) demonstrated a correlation between regularly spaced subsequences of linear congruential generators.

Claims of type (c) are the most difficult to falsify. The more different two generators are, the more difficult it is to construct a mathematical framework that allows the simultaneous analysis of both, and the more unlikely it is that they have a common structural property that can produce a correlation (Press, 2007). A given generator is more likely to correlate with an application, which can have a lot of regularity, than with another generator that was designed with the goal of avoiding correlations.

The dilemma that mathematical tractability is good for theoretical analysis but bad for randomness seems to prevent us from making the best random generators, or at least from knowing which generators are best. Fortunately, we can get along with less than perfect generators as long as we can eliminate known defects by combining two different generators. This means that we can live with minor imperfections in (a) and (b) as long as we can rely on claims of type (c).

It is unreasonable to demand a theoretical proof of type (c) for three reasons. The first reason is that it is not clear what kind of theoretical proof is expected to prove the randomness of a pseudo-random sequence of numbers. The second reason is that the philosophy of science does not allow absolute proofs of this kind, only evidence and falsifiable hypotheses. And the third reason is that the mathematical tractability that would allow such a proof, would also defeat it.

All evidence, theoretical as well as experimental, supports the claim that we can improve randomness by combining the outputs of two or more very different generators. We will rely on this claim as long as it has not been falsified, because it is the best method we have so far for producing deterministic pseudo-random numbers. A more general philosophical discussion is needed about what kind of proofs are possible or desirable in relation to PRNGs.

Quantization effects

The minimum difference between two floating point numbers in the interval [½, 1] is $\delta = 2^{-24}$ for single precision, and 2^{-53} for double precision according to the IEEE-754 standard, which all modern computers support (IEEE Computer Society, 2008). The minimum difference for single precision is 2^{-25} in [¼, ½], 2^{-26} in [⅛, ¼], and so on. Many applications require random floating point numbers with uniform distribution in the interval [0,1). If we require equidistant points with the best possible resolution in single precision, then we will have 2^{24} possible values in the interval [0,1). For this, we need a generator capable of giving 2^{24} different values, all with the same frequency. If the generator outputs e.g. a 32-bit word then we can simply use 24 of these bits and discard the remaining 8 bits.

For most generators, the generating function f gives an integer output x_i in an interval [0, m). Typically f is some arithmetic function modulo m. If m is a power of 2 then we can easily extract the desired number of random bits. Unfortunately, many of the generators that are described in the literature have a modulus m which is not a power of 2. Often m is a prime because functions with prime modulus have advantageous mathematical properties. When converting a pseudorandom integer x_i modulo m to a floating point number in [0,1) it is common to just divide x_i by m. Unfortunately, this does not give equidistant points with equal frequency. If $m < 2^{24}$ then there will be some of the possible values that never occur. If $m > 2^{24}$ then some values between 0.5 and 1 will occur more frequently than other, and values less than 0.5 can be spaced less than $\delta = 2^{-24}$ apart. Such quantization effects can lead to systematic errors in applications that depend on the probability that a random number falls within a certain narrow interval.

For example, consider a generator with prime modulus $m = 2^{32}-5$ (e.g. L'Ecuyer, 1999). A floating point output from this generator will have the value 0.6 with frequency $255/m$, while the next value $0.6 + \delta$ occurs with frequency $256/m$. The value 0.2 occurs with frequency 63/m while the next value $0.2 + \delta/4$ occurs with frequency $64/m$.

Such inaccuracies may be unimportant in small applications, but in large applications that use billions of random numbers, the accumulated errors may actually be statistically significant. It is possible to eliminate the quantization errors by means of a rejection method, but this is quite costly in terms of efficiency (See below for an example of a rejection method). Alternatively, the quantization error may be tempered by an appropriate output function that uses multiple elements in the state buffer.

Why is the output interval half open?

The half-open intervals [0,1) and (0,1] can both be divided into 2^{24} equidistant points with the maximum resolution $\delta = 2^{-24}$ for single precision floating point numbers. This makes it easy to generate a uniformly distributed variable from 24 random bits. We will have quantization errors, as explained above, if we map a 24-bit random number to one of the symmetric intervals [0,1] and (0,1), which have $2^{24} + 1$ and $2^{24} - 1$ equidistant points, respectively.

A Monte Carlo application can generate an event with probability $p \in [0,1]$ by testing $x < p$, where $x \in [0,1)$ is a uniform random variable. If x is quantized as 2^{24} equidistant points in [0,1) with equal frequency and p is similarly quantized by $\delta = 2^{-24}$ then the event $x < p$ will occur with the exact frequency p. If $x \in (0,1]$ then $x \leq p$ will also occur with the exact frequency p. A uniformly distributed x in one of the symmetric intervals [0,1] or (0,1) will give rise to tiny rounding errors in the frequency of $x < p$.

A disadvantage of the half-open intervals is that the mean is not exactly ½, but $(1-\delta)/2$ and $(1+\delta)/2$, respectively. This is acceptable for most purposes since it will take a sample size of $8 \cdot 10^{14}$ to estimate the mean of x with enough precision to get a statistically significant error of 3 standard deviations.

Requirements for good generators

Consider some requirements that are important for the choice of PRNGs for large applications using vector processors, multicore processors and CPU clusters.

1. The generator should pass experimental tests for randomness.
2. The cycle length should be so high that the risk of overlapping subsequences is negligible, but not so high that the state buffer uses an excessive amount of data cache.
3. Good equidistribution, as determined by theoretical or experimental methods (L'Ecuyer, 1994).
4. Good diffusion. This is obtained if each bit in the state buffer depends on multiple bits in the previous state (Panneton, L'Ecuyer and Matsumoto, 2006). Diffusion is closely related to the concept of bifurcation in chaos theory (Fog, 2001; Černák, 1996). A good diffusion means highly chaotic behavior, which is a desirable property for a PRNG.
5. The shortest feedback path should be long enough to fit the largest available vector register. However, a long feedback path means poor diffusion. Therefore, the shortest feedback path should not be longer than necessary.
6. The modulus m should be a power of 2 to avoid quantization effects and rounding errors.
7. The generator should be reasonably fast.
8. It should be possible to generate independent streams from multiple instances of the generator.

Construction of a generator satisfying these requirements

There are many PRNGs described in the literature, but few that satisfy all the requirements listed above. Parallel generation has relied more on multiprocessors than on vector processors (L’Ecuyer, Oreshkin and Simard, 2014). The only generator explicitly designed for vector processors is the "SIMD-oriented Fast Mersenne Twister" (SFMT), which relies on 128-bit vectors (Saito and Matsumoto, 2008, 2009). Unfortunately, the feedback path of this generator does not allow implementations in larger vector registers, and there are no plans for an extended version (Saito, 2014). The general Mersenne Twisters have long feedback paths (Matsumoto and Nishimura, 1998; Nishimura, 2000) so that they can easily be implemented in vector processors. These generators have poor diffusion and slow recovery from a state of mostly 0's. The recently published variant "Mersenne Twister for Graphic Processors" (MTGP) (Saito and Matsumoto, 2013) has somewhat improved diffusion properties, and this appears to be the best choice. The chosen version has the Mersenne exponent 11213, which gives a state buffer size of 351 x 32 bits. The cycle length is $\rho = 2^{11213}-1$. This is more than enough to avoid overlapping subsequences, and higher values would be a waste of data cache. Smaller versions have not been published. The shortest feedback path is 84 x 32 bits, which makes implementation in large vector registers possible.

This generator has known weaknesses, which are common to the Mersenne Twister family: It is vulnerable to tests based on $\mathbb{F}_2$ algebra; it has relatively poor diffusion; and it has subsequences with more 0's than 1's. These weaknesses should be eliminated by combination with a second generator that does not have the same weaknesses.

Other generators with long feedback paths are difficult to find in the literature. The RANROT generator is a lagged Fibonacci generator with bit rotation (Fog, 2001). This generator is simple and fast, it can be constructed with any feedback path length, and most versions pass all tests for randomness. However, this is an example of a generator that is difficult to analyze theoretically. Assumptions about the cycle lengths of RANROT generators are based on extrapolations from experimental measurements on very small generators. The RANROT may be a good generator, but more research is needed before we can rely on this generator for demanding applications.

No other generator was found with a sufficiently long feedback path suitable for our purpose. Multiply-with-carry generators with lag have been described, but they have an extra feedback path of length 1 in the carry (Marsaglia, 2003). It may be possible to construct a multiply-with-carry generator where the carry feedback is also lagged.

Because no suitable candidate for the second generator has been found with a feedback path that allows vectorization, we have instead to rely on multiple parameter sets for the same kind of generator (method 5). Each vector position will have its own independent generator with different parameters for each. After rejecting generators with prime modulus, the best candidate we found was a multiply-with-carry (MWC) generator (Goresky and Klapper, 2003). This generator is relatively simple, it has excellent randomness and very high diffusion or bifurcation. Nine good multipliers for MWC are listed by Press (2007). Eight of these are used in order to implement eight generators of 64 bits each in a 512 bit vector. The output function is a 64-bit XOR-shift method as recommended by Press (2007). Unfortunately, there are not enough good multipliers for future implementations in larger vector registers. Each MWC generator delivers a 64-bit output which is divided into two 32-bit random numbers.

The eight MWC generators have different cycle lengths, ranging from $5 \cdot 10^{18}$ to $9 \cdot 10^{18}$. This is not enough to completely rule out overlapping subsequences in large applications when the MWC generator is used alone, but the MTGP generator has prime cycle length so that the cycle lengths are multiplied when the MWC and MTGP generators are combined.

The MWC generator has a very slight bias in the upper bits (Couture and L'Ecuyer, 1997). The bias is too small to have practical significance, and it is removed by the output function or by the combination with the MTGP generator anyway.

It can be concluded that the MTGP and MWC generators both have known defects, but they have no defects in common. There are no known defects in any of these two generators that cannot be removed by combination with the other generator. Therefore, it is expected that the combined output of these two generators is suitable for even the most demanding applications. Multiple independent streams can be generated from multiple instances of the combined generator by changing the seed of one or both generators, in accordance with method 6 or 7.

Practical implementation

It was decided to make an implementation that is suitable for the forthcoming AVX-512 instruction set, which will be common to the most relevant hardware platforms in a near future. Existing instruction sets with vector sizes smaller than

512 bits are supported by dividing the data into smaller vectors. C++ is the obvious choice of programming language for code that needs to be portable to several platforms and operating systems, highly optimized, and needs overloaded operators for vector operations. The code is integrated into the vector class library (VCL. Fog, 2014b) which provides efficient vector operators for the generator as well as for the application that uses it. Supported platforms include Windows, Linux and Mac OS with Microsoft, Intel, Gnu and Clang compilers.

The generator, named RANVEC1, is implemented as a C++ class so that an application can make a separate instance for each thread in a multiprocessor environment. Each instance can deliver random number vectors of up to 512 bits with integer or floating point elements.

The fastest way of generating a uniform floating point output with equidistant points from random bits is to set the exponent of a single precision floating point number in the IEEE-754 representation to (0+bias) and set the mantissa to 23 random bits. This gives a uniform random number in the interval [1,2). Subtracting 1 then gives a number in the desired interval [0,1) (Saito and Matsumoto, 2009). This method gives a resolution of 2^{-23}. The maximum resolution of $\delta = 2^{-24}$ can be obtained from 24 random bits by first using 23 bits to make a random number in the interval [1,2) as above, and then subtracting either 1 or $(1-\delta)$ depending on whether the last bit is 0 or 1. It is possible to make a double precision random number with the maximum resolution of 2^{-53} by the same method, but the current implementation gives only a resolution of 2^{-52} for double precision because it was decided that the last bit will have no significance for applications with a realistic sample size.

Many applications need a random integer u with uniform distribution in an interval $[a,b]$ of length $d = \text{b-a} + 1$. This can be obtained from a random 32-bit unsigned integer x by a 64-bit multiplication: $u = a + \left\lfloor xd / 2^{32} \right\rfloor$. However, this method is subject to a bias similar to the quantization error discussed above when the interval length d is not a power of 2. Floating point calculation methods give the same error because of the mapping of an interval of a power-of-2 length to another interval of incommensurable length d. Most standard random generator libraries have this error. The error may be negligible when d is small, but it can be quite serious for large d. The worst case is $d = 3 \cdot 2^{30}$. In this case, values of $(u - a)$ that are divisible by 3 occur twice as frequent as other values. This can obviously lead to serious errors in applications that happen to depend on u mod 3. This error can be eliminated by using a rejection method. Confine x to r possible values where r is a multiple of d. $r = \lfloor 2^{32} / d \rfloor \cdot d$. If xd mod $2^{32} \geq r$ then reject the value and generate a new x.

Rejection methods are also used for generating random variables with other distributions than uniform (Devroye, 1986). Algorithms that involve rejection methods may be implemented in vector processors as follows. First generate a random vector and execute the steps in the algorithm necessary to determine rejection. If any elements of the vector are rejected, then generate another random vector and repeat the calculations. Replace any rejected elements in the first vector by accepted elements from the second vector. Continue like this until we have a vector of only accepted elements. If calculations are expensive and not dependent on changing parameters then we may save any remaining accepted elements for the next round. If exact reproducibility across platforms is required then we must keep the vector size constant.

Tests of the constructed generator

The randomness of the generator outputs were tested using the powerful BigCrush battery of tests in the TestU01 software suite of experimental tests for randomness (L'Ecuyer and Simard, 2007). The MWC generators were tested in various configurations: each of the eight generators separately, the lower or upper 32-bit half of each generator output, as well as all eight generators in a round robin fashion. All tests were passed. The MWC generators failed several tests when the XOR-shift output function was removed.

The MTGP generator failed the linear complexity test as expected, but passed all other tests in the BigCrush battery of tests. The MTGP generator also failed a binary matrix rank test where the matrix size was increased to 12000×12000. The test results were the same when the output function (so called tempering) was removed. The combination of the MWC and MTGP generator passed all tests, with or without tempering.

The speed of the random generators were tested after compiling with different compilers and different vector register sizes. The test measured the time required to generate 2^{14} random 32-bit integers and computing their sum. The calculation time depends on the CPU clock frequency, which varies a lot due to the power-saving features of the CPU. In order to get consistent and reproducible time measurements, it was decided to use the core clock count as time unit. This time unit is defined by the frequency that the execution unit in the CPU is actually running at. Core clock counts were measured using the TESTP test program (Fog, 2014c). The calculation speed was measured for the MWC and MTGP generators as well as for the SFMT generator and the original Mersenne Twister (MT). The results are given in Table 2.

Table 2. Random number generation times for various generators using different compilers and register sizes. The unit is core clock cycles per 32 bits, single thread.

		Compiler			
Generator	**Register size bits**	**Gnu**	**Clang**	**Intel**	**Microsoft**
MWC	128	4.1	4.0	3.6	3.0
	256	1.8	2.2	2.6	3.1
MTGP	128	8.9	10.3	8.8	18.4
	256	4.0	4.5	4.5	43.1
MTGP w/o tempering	256	3.1	3.5	3.6	18.9
MWC + MTGP	128	10.4	12.4	10.4	20.3
	256	5.0	5.7	6.1	46.4
MWC + MTGP w/o tempering	256	3.9	4.6	5.1	20.7
SFMT	128	2.0	1.8	2.0	1.9
MT	32	9.3	14.2	8.5	12.8

Configuration: Intel Haswell microprocessor, 3.4 GHz. Windows 7, 64 bits. Gnu C++ compiler v. 4.8.3 Cygwin. Clang C++ compiler v. 3.4.2 Cygwin. Intel C++ compiler v. 15.0. Microsoft C++ compiler v. 17.0.

Notice that the combined generator takes 5 – 6 clock cycles per random number using a vector size of 256 bits when the Gnu, Clang or Intel compiler is used. This corresponds to approximately $6{\cdot}10^8$ random numbers per second per thread on a 3.4 GHz processor. This number can be multiplied by the number of cores in the CPU when each core is running one thread. It is possible to run two threads per core on some CPUs, but this may not be optimal if the two threads are competing for the same execution resources (Fog, 2014a).

Most Monte Carlo applications take much more time than this to process the random numbers, so that the random number generation will account for only a small fraction of the total execution time. A few clock cycles more or less is hardly important in this context. Therefore, we can afford the luxury of using a combined generator of very high quality. The convenient availability of random numbers as vectors can make it easier to vectorize the applications that use the random numbers, possibly leading to very significant speed gains for some applications.

The RANVEC1 code also supports a register size of 512 bits. This was verified using Intel Software Emulator version 7.1.0, but no meaningful speed measurement was possible because no microprocessor with the AVX-512 instruction set is available yet.

The SFMT generator is faster than the MTGP generator because the former is designed specifically for vector processing while the MTGP is designed for graphics processors. Unfortunately, the SFMT generator cannot be implemented with vector sizes higher than 128 bits.

Conclusion

There are two main principles for parallel processing: vector processing and multicore processing. Large Monte Carlo applications need to utilize both in order to get the maximum performance out of modern computers. A literature search revealed only one generator specifically designed for vector processing, and none that fits the growing vector size of modern processors. Fortunately, it is possible to utilize vector processors by adapting other generators with sufficiently long feedback paths or by implementing multiple similar generators in parallel. The combined generator described here (RANVEC1) utilizes both methods. A C++ implementation of this combined generator is available as part of the vector class library (VCL) at http://www.agner.org/optimize/#vectorclass.

As Monte Carlo applications get larger they also put higher demands on the quality of random number generators. The following qualities must be considered:

1. Quality of randomness.
2. Speed.
3. Avoid overlapping sequences.
4. Equidistant points with perfectly uniform distribution.
5. Portability among platforms.
6. Reproducibility.

The quality of randomness (1) can be improved by combining two generators with fundamentally different design. This enables us to overcome the flaws caused by the unsolvable dilemma between the need for mathematical tractability and the desire for chaotic behavior.

The speed (2) of the available generators is so high that the generation of random numbers accounts for only a small fraction of the total calculation time of a typical application. However, there is a pitfall when measuring the speed of a generator in isolation. The larger Mersenne Twister generators are consuming considerable amounts of data cache whereby they may slow down the applications that use them. The size of the state buffer should be a compromise between long cycle length and low data cache use.

The risk of overlapping sequences (3) gets higher as the number of simultaneous random streams is increasing. This risk can be made negligible by using a generator with an extremely long cycle length, or we can eliminate it completely by combining two different generators.

Quantization effects are often ignored in standard PRNG libraries, which makes them deviate from the perfectly uniform distribution (4). Undesired quantization effects are seen when the output of a generator with prime modulus is mapped onto an interval with power-of-2 modulus and when the output of any generator is used for generating a random integer in an interval of arbitrary (incommensurable) length. These undesired effects can be eliminated by avoiding generators with prime modulus or by using a rejection method.

Portability (5) is generally obtained by using a standardized programming language. The RANVEC1 generator is designed for the vector extensions to the x86 instruction set. This fits the most commonly used computer platforms today, as well as prospected future processors with 512-bit vectors. It cannot be used on platforms with other instruction sets without major reprogramming, and the target platform must have similar vector processing capabilities.

Reproducibility (6) is useful for replaying an interesting simulation event, for verifying results and for debugging. It is always possible to reproduce a random number stream by using the same generator again with the same seed. However, problems may arise when vector sizes change. For example, consider a simulation application that uses both integer and floating point random number vectors. First, it generates a vector of 8 integers, then a vector of 8 floats, then 8 integers, 8 floats, etc. If we now update the hardware to a processor that supports bigger vectors, we may generate first 16 integers and then 16 floats, etc. This means that the numbers are generated in a different order so that the simulation results will be different even though we have used the same seed. A remedy against this problem is to generate numbers in batches that correspond to the biggest possible vector size. The RANVEC1 software uses batches of 512 bits to fit the future AVX-512 instruction set, but the reproducibility will be lost in case of future extensions to 1024 bits or more. Reproducibility can also be lost in case of outputs that use a rejection method when the vector size is changed.

Scope for future research

We have found an acceptable solution to our needs for a good PRNG that utilizes both vector processing and multiprocessing, but we can predict the future need for a generator that fits larger vector sizes. We would also like a more efficient solution even though the speed is acceptable for current purposes.

The vector implementation of the MTGP is slower than the SFMT even though it can use a larger vector size. The difference in speed can be explained by the following factors.

- The size of the state buffer in the MTGP is not divisible by the vector size. Extra code is needed to handle the wrap-around situation where a vector spans part of the end of the buffer and part of the beginning. Memory access is misaligned for the same reason.
- The output function in the MTGP, called tempering, consumes a large fraction of the code and CPU time. The purpose of the tempering is to improve equidistribution, but this improvement is not visible in the test results. The SFMT generator obtains good equidistribution by an appropriate choice of parameters without a tempering function.
- The MTGP algorithm has longer dependency chains than the SFMT.
- The SFMT can use the state buffer also as output buffer in a block generation scheme. This is not possible with the MTGP because its tempering function needs to read two parts of the state buffer for each output value.

A better solution would have a state buffer size that is a multiple of the largest vector size we expect to be available in a reasonable future. It is possible to increase the state buffer size beyond the Mersenne exponent either by having some bits without feedback or by using the same method as the SFMT (Saito and Matsumoto, 2008, 2009). The state buffer size should not be excessive because of the data cache use. Parameters should be adjusted to give satisfactory equidistribution in order to eliminate the need for a tempering function.

The shortest feedback path should be at least as long as the largest possible vector size. There is a tradeoff here because a large feedback path is reducing the diffusion in the generator. The diffusion is already low in many variants of Mersenne Twisters because they use sparse matrixes in the algorithm. There are various ways to make more dense matrixes without excessive computation time. It is possible to implement a 4×32 bit $\mathbb{F}_2$ matrix multiplication with a single 512-bit vector permutation instruction, and this method is used in the RANVEC1 code. Another possibility, which has not been utilized so far, is to use carry-less multiplication. Modern x86 processors have such an instruction. The carry-less multiplication instruction multiplies two 64-bit vectors to give a 127-bit product (Intel, 2014b), and this corresponds to a dense matrix multiplication in $\mathbb{F}_2$. Unfortunately, there is no version of this instruction with larger vectors, but the result can easily be broadcast into a larger vector in order to increase diffusion.

The second generator in our combination, the MWC, cannot easily be expanded to larger vectors than 512 bits. There are nine known good multipliers for a 64-bit MWC (Press, 2007) and we have used eight of these for implementing eight parallel MWC generators. Future implementations with larger vector sizes need another generator with more good parameter sets—perhaps a variant of MWC with an addend, an extra term or a short lag.

These are very practical problems, which can definitely be solved. On a more philosophical level, we need a clarification of the role of proofs in PRNG research. Is it possible to prove that a generator has no defects? What kind of evidence can we accept? If all we have is falsifiable propositions, does it make sense to say that some propositions have more value than others if it is more difficult to find examples that falsify them? Does it make sense to require theoretical proofs, e.g. that two random number streams are statistically independent, when it is impossible to even prove the more fundamental assumptions about randomness of a single stream?

References

Anthony, S. (2013). *Intel unveils 72-core x86 Knights Landing CPU for exascale supercomputing* [Blog post]. Retrieved from http://www.extremetech.com/extreme/171678-intel-unveils-72-core-x86-knights-landing-cpu-for-exascale-supercomputing

Barash, L. Yu., & Shchur, L. N. (2014). PRAND: GPU accelerated parallel random number generation library: using most reliable algorithms and applying parallelism of modern GPUs and CPUs. *Computer Physics Communications*, *185*(4), 1343–53. doi:10.1016/j.cpc.2014.01.007

Bauke, H., & Mertens, S. (2004). Pseudo random coins show more heads than tails. *Journal of Statistical Physics*, *114*(3–4), 1149–69. doi:10.1023/B:JOSS.0000012521.67853.9a

Bauke, H., & Mertens, S. (2007). Random numbers for large-scale distributed Monte Carlo simulations. *Physical Review E*, *75*(6), 066701. doi:10.1103/PhysRevE.75.066701

Černák, J. (1996). Digital generators of chaos. *Physics Letters A*, *214*(3), 151–60. doi:10.1016/0375-9601(96)00179-X

Chan, E. Y. K. (2013). *Benchmarks for Intel MIC Architecture*. Retrieved from http://www.clustertech.com/wp-content/uploads/2014/01/MICBenchmark.pdf

Chrysos, G. (2012). *Intel Xeon Phi Coprocessor - the Architecture*. Retrieved from http://software.intel.com/en-us/articles/intel-xeon-phi-coprocessor-codename-knights-corner

Couture, R., & L'Ecuyer, P. (1997). Distribution properties of multiply-with-carry random number generators. *Mathematics of Computation*, *66*(218), 591–607.

Demchik, V. (2011). Pseudo-random number generators for Monte Carlo simulations on ATI graphics processing units. *Computer Physics Communications*, *182*(3), 692–705. doi:10.1016/j.cpc.2010.12.008

Deng, L. Y., Lin, D. K. J., Wang, J., & Yuan, Y. (1997). Statistical justification of combination generators. *Statistica Sinica*, *7*, 993–1003.

Devroye, L. (1986). *Non-uniform random variate generation*. New York: Springer.

Durst, M. J. (1989). Using linear congruential generators for parallel random number generation. In E. A. MacNair, K. J. Musselman, & P. Heidelberger (Eds.), *WSC '89 Proceedings of the 21st conference on Winter simulation* (pp. 462–66). New York: ACM. doi:10.1145/76738.76798

Entacher, K. (1998). Bad subsequences of well-known linear congruential pseudorandom number generators. *ACM Transactions on Modeling and Computer Simulation*, *8*(1), 61–70.

Fog, A. (2001). *Chaotic random number generators with random cycle lengths*. Publisher: Author. Retrieved from http://www.researchgate.net/publication/245642152

Fog, A. (2014a). *Optimizing software in C++. An optimization guide for Windows, Linux and Mac platforms*. Publisher: Author. Retrieved from http://www.agner.org/optimize/optimizing_cpp.pdf

Fog, A. (2014b). *C++ vector class library* [Software library]. Publisher: Author. Retrieved from http://www.agner.org/optimize/#vectorclass

Fog, A. (2014c). *Test programs for measuring clock cycles and performance monitoring* [Software library]. Publisher: Author. Retrieved from http://www.agner.org/optimize/#testp

Goresky, M., & Klapper, A. (2003). Efficient multiply-with-carry random number generators with maximal period. *ACM Transactions on Modeling and Computer Simulation*, *13*(4), 310–21. doi:10.1145/945511.945514

IEEE Computer Society. (2008). *IEEE Standard for Floating-Point Arithmetic* (IEEE Std. 754-2008) New York: IEEE.

Intel. (2014a). *Intel architecture instruction set extensions programming reference* (Doc. 319433-021). Retrieved from http://software.intel.com/en-us/intel-isa-extensions

Intel. (2014b). Intel 64 and IA-32 architectures software developer's manual (Doc. 325462-052US). http://www.intel.com/content/www/us/en/processors/architectures-software-developer-manuals.html

Karpiński, P. (2014). *Evaluation of Intel Xeon Phi (Knight's Corner) coprocessor's core performance using VCL*. Manuscript submitted for publication.

Knuth, D. E. (1998). *The art of computer programming, volume 2: seminumerical algorithms* (3rd Ed). Boston, MA: Addison-Wesley Professional.

L'Ecuyer, P. (1990). Random numbers for simulation. *Communications of the ACM, 33*(10), 85–97. doi:10.1145/84537.84555

L'Ecuyer. P. (1994). Uniform random number generation. *Annals of Operations Research, 53*(1), 77–120. doi:10.1007/BF02136827

L'Ecuyer, P. (1999). Tables of linear congruential generators of different sizes and good lattice structure. *Mathematics of Computation, 68*(225), 249–60. doi:10.1090/S0025-5718-99-00996-5

L'Ecuyer, P., & Côté, S. (1991). Implementing a random number package with splitting facilities. *ACM Transactions on Mathematical Software, 17*(1), 98–111. doi:10.1145/103147.103158

L'Ecuyer, P., & Granger-Piché, J. (2003). Combined generators with components from different families. *Mathematics and Computers in Simulation, 62*(3–6), 395–404. doi:10.1016/S0378-4754(02)00234-3

L'Ecuyer, P., Oreshkin, B., & Simard, R. (2014). *Random numbers for parallel computers: requirements and methods*. Manuscript submitted for publication.

L'Ecuyer, P., & Simard, R. (2007). TestU01: a C library for empirical testing of random number generators. *ACM Transactions on Mathematical Software, 33*(4), 22. doi:10.1145/1268776.1268777

Manssen, M., Weigel, M., & Hartmann, A. K. (2012). Random number generators for massively parallel simulations on GPU. *European Physical Journal: Special Topics, 210*(1), 53–71. doi:10.1140/epjst/e2012-01637-8

Marsaglia, G. (1985). A current view of random number generators. In L. Billard (Ed.), *Computer science and statistics: proceedings of the Sixteenth Symposium on the Interface, Atlanta, Georgia*, (pp. 3–10). Amsterdam: North-Holland.

Marsaglia, G. (2003). Random number generators. *Journal of Modern Applied Statistical Methods, 2*(1), 2–13. http://digitalcommons.wayne.edu/jmasm/vol2/iss1/2/

Matsumoto, M., & Nishimura, T. (1998). Mersenne Twister: A 623-Dimensionally Equidistributed Uniform Pseudo-Random Number Generator. *ACM Transactions on Modeling and Computer Simulation, 8*(1), 3–30.

Matsumoto, M., & Nishimura, T. (2000). Dynamic creation of pseudorandom number generators. In H. Niederreiter & J. Spanier (Eds.), *Monte Carlo and Quasi-Monte Carlo Methods, 1998: Proceedings of a Conference Held at the Claremont Graduate University, Claremont, California, USA* (pp. 55–69). New York: Springer-Verlag, Inc.

Nandapalan, N., Brent, R.P., Murray, L. M., & Rendell, A. P. (2012). High-performance pseudo-random number generation on graphics processing units. In R. Wyrzykowski, J. Dongarra, K. Karczewski, & J. Waśniewski (Eds.), *Lecture Notes in Computer Science 7203: Parallel Processing and Applied Mathematics (9th International Conference, PPAM 2011, Torun, Poland, September 11-14, 2011, Revised Selected Papers, Part I)* (pp. 609–18). New York: Springer-Verlag. doi:10.1007/978-3-642-31464-3_62

Nishimura, T. (2000). Tables of 64-Bit Mersenne twisters. *ACM Transactions on Modeling and Computer Simulation, 10*(4), 348–57. doi:10.1145/369534.369540

Panneton, F., L'Ecuyer, P., & Matsumoto, M. (2006). Improved long-period generators based on linear recurrences modulo 2. *ACM Transactions on Mathematical Software, 32*(1), 1–16. doi:10.1145/1132973.1132974

Passerat-Palmbach, J., Mazel, C., Mahul, A., & Hill, D. (2010). Reliable Initialization of GPU-Enabled Parallel Stochastic Simulations Using Mersenne Twister for Graphics Processors. In G. K. Janssens, K. Ramaekers, & A. Caris (Eds.), *European Simulation and Modelling 2010, Essen, Belgium* (pp. 187–95). Ostend, Belgium: Eurosis.

Passerat-Palmbach, J., Mazel, C., & Hill, D. R. C. (2011). Pseudo-random number generation on GP-GPU. In S. Strassburger (Ed.), *2011 IEEE Workshop on Principles of Advanced and Distributed Simulation (PADS)*, (pp. 1–8). New York: IEEE. doi:10.1109/PADS.2011.5936751

Popper, K. (1963). *Conjectures and refutations: the growth of scientific knowledge*. London, U.K.: Routledge.

Press, W. H. (2007). *Numerical recipes: the art of scientific computing* (3rd Ed.). Cambridge, U.K.: Cambridge University Press.

Saito, M. (2014). Personal communication.

Saito, M., & Matsumoto, M. (2008). SIMD-oriented fast Mersenne twister: a 128-Bit pseudorandom number generator. In A. Keller, S. Heinrich, & H. Niederreiter (Eds.), *Monte Carlo and Quasi-Monte Carlo Methods 2006*, (pp. 607–22). New York: Springer. doi:10.1007/978-3-540-74496-2_36

Saito, M., & Matsumoto, M. (2009). A PRNG specialized in double precision floating point numbers using an Affine transition. In P. L'Ecuyer and A. B. Owen (Eds.), *Monte Carlo and Quasi-Monte Carlo Methods 2008*, (pp. 589–602). New York: Springer. doi:10.1007/978-3-642-04107-5_38

Saito, M., & Matsumoto, M. (2013). Variants of Mersenne twister suitable for graphic processors. *ACM Transactions on Mathematical Software, 39*(2), 12. doi:10.1145/2427023.2427029

Salmon, J. K. (2011). Parallel random numbers: as easy as 1, 2, 3. In J. Costa & W. Kramer (Program Chairs), *Proceedings of 2011 International Conference for High Performance Computing, Networking, Storage and Analysis, Seattle, WA*. doi:10.1145/2063384.2063405

Saule, E., Kamer, K., & Çatalyürek, Ü. V. (2013). *Performance Evaluation of Sparse Matrix Multiplication Kernels on Intel Xeon Phi* (arXiv:1302.1078). http://arxiv.org/abs/1302.1078.

בס"ד

Journal of Modern Applied Statistical Methods
May 2015, Vol. 14, No. 1, 335-336

ISSN 1538 − 9472

Instructions for Authors

Authors wishing to submit to *JMASM* may do so using the submission form at the journal's website, http://digitalcommons.wayne.edu/jmasm. Three areas are appropriate for *JMASM*:

1. Development or study of new statistical tests or procedures, or the comparison of existing statistical tests or procedures, using computer-intensive Monte Carlo, bootstrap, jackknife, or resampling methods;
2. Development or study of nonparametric, robust, permutation, exact, and approximate randomization methods; and
3. Applications of computer programming, preferably in Fortran (all other programming environments are welcome), related to statistical algorithms, pseudo-random number generators, simulation techniques, and self-contained executable code to carry out new or interesting statistical methods.

Elegant derivations, as well as articles with no take-home message to practitioners, have low priority. Articles based on Monte Carlo (and other computer-intensive) methods designed to evaluate new or existing techniques or practices, particularly as they relate to novel applications of modern methods to everyday data analysis problems, have high priority.

Work appearing in *Regular Articles*, *Brief Reports*, and *Emerging Scholars* is externally peer reviewed, with input from the Editorial Board; work appearing in *Statistical Software Applications and Review* and *JMASM Algorithms and Code* is internally reviewed by the Editorial Board. *JMASM* charges neither article processing fees nor submission fees.

Please observe the following guidelines when preparing manuscripts:

1. *JMASM* uses a modified American Psychological Association style guideline.
2. Articles should be submitted without a title page or abstract. There should be no material identifying authorship except in the fields of the submission form. Include a statement in the cover letter indicating that proper human subjects protocols were followed where applicable, including informed consent.
3. Manuscripts should be prepared in Microsoft Word (.doc or .docx) 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.
4. The text maximum is 20 pages double spaced, not including tables, figures, graphs, and references. Use 11 point Times Roman font.
5. 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.
6. The submission form requires an Abstract with a 50 word maximum, and 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.

7. 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.

8. 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.

9. 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 the spacebar twice after a period.

Journal of **Modern Applied Statistical Methods**

ISSN: 1538−9472 http://digitalcommons.wayne.edu/jmasm

PUBLISHED biannually (May, November) in partnership by:

JMASM, Inc.
PO Box 48023
Oak Park, MI 48237
ea@jmasm.com

Wayne State University Library System
Purdy Library
Detroit, MI 48202
digitalcommons@wayne.edu

Copyrights, Attribution and Usage Policies

To Advertisers