Journal of Modern Applied Statistical Methods

Volume 10 | Issue 1 Article 4

5-1-2011

Number of Replications Required in Monte Carlo Simulation Studies: A Synthesis of Four Studies

Daniel J. Mundform
New Mexico State University, daniel.mundfrom@eku.edu

Jay Schaffer
University of Northern Colorado, jay.schaffer@unco.edu

Myoung-Jin Kim
Illinois State University, mkim2@ilstu.edu

Dale Shaw
University of Northern Colorado, dale.shaw@unco.edu

Ampai Thongteeraparp
Kasestart University, Bangkean, Bangkok, Thailand, fsciamu@ku.ac.th

See next page for additional authors

DOI: 10.22237/jmasm/1304222580

Number of Replications Required in Monte Carlo Simulation Studies: A Synthesis of Four Studies

Authors
Daniel J. Mundform, Jay Schaffer, Myoung-Jin Kim, Dale Shaw, Ampai Thongteeraparp, and Pornsin Supawan

Journal of Modern Applied Statistical Methods
May 2011, Vol. 10, No. 1, 19-28

1538 – 9472/11/$95.00

Regular Articles

Number of Replications Required in Monte Carlo Simulation Studies: A Synthesis of Four Studies

Daniel J. Mundfrom
New Mexico State University,
Las Cruces, NM USA

Jay Schaffer
University of Northern Colorado,
Greeley, CO USA

Myoung-Jin Kim
Mennonite College of Nursing
at Illinois State University,
Normal, IL USA

Dale Shaw
University of Northern Colorado,
Greeley, CO USA

Ampai Thongteeraparp
Kasetsart University,
Bangkean, Bangkok, Thailand

Chana Preecha
Thepsatri Rajabhat University,
Muang, Lopburi, Thailand

Pornsin Supawan
Rajabhat Rajanagarindra University,
Chacoengsao, Chaiyaphum, Thailand

Monte Carlo simulations are used extensively to study the performance of statistical tests and control charts. Researchers have used various numbers of replications, but rarely provide justification for their choice. Currently, no empirically-based recommendations regarding the required number of replications exist. Twenty-two studies were re-analyzed to determine empirically-based recommendations.

Key words: Simulation studies, number of replications, ANOVA, multiple comparisons, regression, control charts.

Introduction

Monte Carlo simulation has become an important and popular research tool used by quantitative researchers in a variety of disciplines (Fan, Felsővályi, Sivo & Keenan, 2002). The Monte Carlo method provides approximate solutions to a variety of mathematical problems by performing statistical sampling experiments via computer. Monte Carlo simulation offers researchers an alternative to the theoretical approach; this is important because many situations exist in which implementing a theoretical approach is difficult – and finding an exact solution is even more difficult. In addition, computing power has become increasingly less expensive and computers are more widely available than ever before.

An important question to address when conducting a Monte Carlo simulation study is how many replications are needed to obtain accurate results. With advanced computers, researchers are able to run in excess of 10,000 replications in their studies (see, for example, Kaplan, 1983; Klockars & Hancock, 1992; Gamage & Weerahandi, 1998; Alyounes, 1999).

Daniel J. Mundfrom is an Associate Professor of Applied Statistics, Department of Economics and Professor Emeritus of Applied Statistics and Research Methods at the University of Northern Colorado. E-mail: daniel.mundfrom@eku.edu. Jay Schaffer is an Associate Professor of Applied Statistics and Research Methods. E-mail: jay.schaffer@unco.edu. Myoung-Jin Kim is an Assistant Professor. E-mail: mkim2@ilstu.edu. Dale Shaw is Professor Emeritus of Applied Statistics and Research Methods. E-mail: dale.shaw@unco.edu. Ampai Thongteeraparp is an Assistant Professor of Statistics in the Department of Statistics, Faculty of Science. E-mail: fsciamu@ku.ac.th. Chana Preecha is an Assistant Professor of Statistics on the Faculty of Science and Technology. E-mail: chanaua@yahoo.com. Pornsin Supawan is an Assistant Professor of Statistics on the Faculty of Science. E-mail: pornsin.s@gmail.com.

According to Brooks (2002), simulations may produce inaccurate estimates if an insufficient number of replications are used. Hutchinson and Bandalos (1997) also criticized:

> With too few replications, idiosyncratic results based on a particular sample are more likely to arise. Unfortunately for simulation researchers there are no definitive guidelines for selecting the appropriate number of replications. The specific number will depend on the type of phenomenon being studied, the extent to which the steps of the simulation can be automated, as well as available computer resources. (p. 238)

The choice of the number of replications used in simulation studies appear to be made solely by the judgment of the researchers; this is surmised due to the many simulation studies that have been conducted without any justification provided for the number of replications used (see, for example, Fellner, 1990; Neubauer, 1997; Khoo & Quah, 2002; Khoo & Quah, 2003; Khoo, 2003; Khoo, 2004). Currently, however, no empirically-based recommendations for general guidelines regarding the required number of replications a researcher should use in order to achieve accurate results exist. The obtained results from a Monte Carlo study might be invalid if too few replications were used, whereas time and resources may have been wasted if more replications were used than were necessary. In addition, with the same amount of time and resources but fewer replications, more conditions could be investigated.

The purpose of this synthesis was to: (1) provide information regarding the minimum number of replications required to reproduce a reported statistic, within a specified degree of accuracy, in 22 published Monte Carlo studies from a variety of areas, and (2) provide general recommendations regarding the minimum number of replications needed for future simulation studies.

Methodology

An extensive review of the literature was conducted in various fields of study, identifying research that used Monte Carlo simulations to estimate characteristics of interest (e.g., Type I error rates, power and average run length). Through four dissertations, 22 studies were selected such that each provided sufficient information regarding methodology to replicate.

Each study was re-analyzed using the same number of replications as in the original study to produce results that were considered the standard to be met by the re-analyses using a different number of replications. Using a decreasing (or increasing) number of replications, the simulations were repeated until the minimum number of replications was found that produced stable results.

For example, if the original study used 10,000 replications, the process started with 10,000 replications to reproduce the original results and identify the standard to be met, and then the study was re-done with the number of replications cut in half to 5,000. If the results were reproduced, the replications were cut to 2,500; conversely, if the results were not reproduced the replications were increased to 7,500. This iterative process, either reducing the number of replications by cutting in half the number of replications used in the previous step, or increasing the number of replications used by splitting the difference between the last two numbers of replications used (e.g., 5,000 and 10,000), continued until stable results were obtained. After the simulations were completed, recommendations were put forth for the minimum number of replications necessary to estimate a particular parameter within a defined degree of accuracy.

In order to define a specified degree of accuracy, an error band was created by adding/subtracting some percentage to/from each statistic of interest. Bradley (1978) presented two intervals to examine the robustness of hypothesis testing by examining Type I error rate, α. These two intervals were described as a fairly stringent error band, $\alpha \pm 0.1\,\alpha$, and a fairly liberal error band, $\alpha \pm 0.4\,\alpha$. If $\alpha = 0.05$, these error bands become $\pm$ 0.005 and $\pm$ 0.02 respectively. Bradley's criteria were used in these dissertations.

Dissertation I: ANOVA Simulation Studies (Preecha, 2004)

This study replicated 5 simulation studies related to ANOVA. The studies included:

(1) Brown and Forsythe (1974) examined the small sample behavior of various statistics testing the equality of several means. They used 10,000 replications and examined both Type I error rate and power. No justification was provided for the number of replications used or for how the accuracy of results was determined. The four statistics compared were the:

   (a) ANOVA F-statistic;
   (b) Modified F-statistic;
   (c) Welch and James statistic (Welch, 1947); and the
   (d) Welch and James statistic (Welch, 1951).

(2) Alyounes (1999) compared the Type I error rate and power for the Kruskal-Wallis test and the Welch test to the F-test, followed by four post hoc procedures. They used 21,000 replications but provided no justification for that number. Bradley's stringent criterion and Robey and Barcikowski's intermediate criterion were used to examine the robustness of the tests compared. The parametric and nonparametric omnibus tests and the post hoc comparisons used were the:

   (a) ANOVA F-test;
   (b) Welch test;
   (c) Kruskal Wallis test;
   (d) Tukey-Kramer test;
   (e) Games-Howell test;
   (f) Joint ranking (Improved Dunn) test; and
   (g) Separate ranking test.

(3) Gamage and Weerahandi (1998) examined the size performance of four tests in a one-way ANOVA. They compared the Type I error rate and power of the Generalized F-test to the classical F-test, the F-test using weighted least squares to adjust for heteroscedasticity, the Brown-Forsythe test, and the Welch test using 20,000 replications. No justification was provided for the number of replications used or for how the accuracy of results was determined. The statistics compared were the:

   (a) Generalized F-test;
   (b) ANOVA F-test;
   (c) F-test using weighted-least squares;
   (d) Brown-Forsythe test; and the
   (e) Welch test.

(4) Kim (1997) examined three robust tests for ANOVA using weighted likelihood estimation, comparing Type I error rate and power with 5,000 replications. No justification was provided for the number of replications used or for how the accuracy of results was determined. The statistics compared were the:

   (a) Basu-Sarkar-Basu test;
   (b) Modified Welch Test with weighted likelihood estimators; and the
   (c) Modified Brown-Forsythe test using weighted likelihood estimators.

(5) Kaplan (1984) examined the comparative effects of violations of homogeneity of variance on two tests when the underlying populations were normal, but sample sizes were unequal. She compared Type I error rate and power using 20,000 replications. She provided no justification for the number of replications used, but used the estimated standard error when examining a single proportion and the estimated standard error of the difference between two proportions when comparing two independent proportions. The tests compared were the:

   (a) χ^2-approximation of the Kruskal-Wallis statistic; and the
   (b) Incomplete Beta approximation of the Kruskal-Wallis statistic.

Dissertation I: Results and Discussion

Each of the five studies investigated Type I error rates and power. Using ± 0.005 for Type I error (Bradley's fairly stringent criterion) and ± 0.02 for power (Bradley's fairly liberal criterion), the minimum number of replications

were found that produced stable results. Table 1 displays the number of replications used in the original study along with the recommended minimum number of replications needed to produce similar results. In each situation, it appears that fewer replications could have been used to predict power and in all but one situation, fewer replications could have been used to estimate Type I error. In that one situation a larger number of replications was required to get a stable estimate of the Type I error rate.

Table 1: Number of Replications Used By the Original Study Along With the Recommended Minimum Number of Replications Required To Produce Stable Results

	Original	Replications Recommended	
Study	Replications Used	Type I Error	Power
1	10,000	5,000 – 10,000	5,000
2	21,000	10,500	5,250
3	20,000	7,500	5,000
4	5,000	7,500	2,500
5	20,000	5,000	5,000

Dissertation II: Multiple Comparison Simulation Studies (Ussawarujikulchai, 2004)

The second dissertation replicated 5 simulation studies related to multiple comparison tests after a significant ANOVA was found. The studies included:

(1) Seaman, Levin and Serlin (1991) examined the Type I error rate of several multiple comparison procedures using 5,000 replications to compare 5 treatment groups with sample sizes of n = 10. Three groups had means set equal to 0 and the other groups had means set to 0.8560. The procedures compared were:

(a) Standard Bonferroni;
(b) Tukey test;
(c) Holm test;
(d) Fisher LSD test;
(e) Hayter-Fisher Modified LSD test;
(f) REGWQ test;
(g) Newman-Kuels test;
(h) Duncan test;
(i) Shaffer test;
(j) Protected Shaffer test; and
(k) Ramsey's Model-Testing approach.

(2) Klockars and Hancock (1992) examined the power of five multiple comparison procedures against the standard Bonferroni procedure when applied to complete sets of orthogonal contrasts. They used 20,000 replications with both k = 4 and k = 5 treatment groups partitioned into k–1 orthogonal contrasts. The procedures they compared were:

(a) Holm test;
(b) Hochberg test;
(c) Hommel test;
(d) Protected Shaffer test;
(e) Modified Stagewise Protected test; and
(f) Standard Bonferroni procedure.

(3) Hsiung and Olejnik (1994) examined the Type I error rate of several multiple comparison procedures for all pairwise contrasts when population variances differed in both balanced and unbalanced one-factor designs. They used 10,000 replications for each of k = 4 and k = 6 treatment groups. The multiple comparison procedures they compared were:

(a) Games-Howell test;
(b) Dunnett T3 test;
(c) Dunnett C test;
(d) Holland-Copenhaver test;
(e) Shaffer test; and
(f) Protected Shaffer test.

(4) Morikawa, Terao and Iwasaki (1996) examined the Type I error rate and power of several multiple comparison procedures for pairwise comparisons. They used 1,000 replications with each of k = 3 and k = 4

treatment groups and sample sizes of 10, 20 and 50 to examine both any-pair power and all-pairs power. The procedures they compared were:

(a) Tukey test;
(b) Standard Bonferroni test;
(c) Holm test;
(d) Shaffer test;
(e) Hommel test;
(f) Hochberg test; and the
(g) Rom test.

(5) Ramsey (2002) examined the power of five pairwise multiple comparison procedures using 10,000 replications with 4 treatment groups and a sample size of 16. Both any-pair power and all-pairs power were examined for three different mean configurations-maximum range, equally spaced, and minimum range. The procedures compared were:

(a) Tukey test;
(b) Hayter-Fisher Modified LSD test;
(c) Shaffer-Welsch test;
(d) Shaffer test; and the
(e) Holland-Copenhaver test.

Dissertation II: Results and Discussion

Each of these five studies investigated either Type I error rate, power, or both. Using ± 0.005 for Type I error (Bradley's fairly stringent criterion) and ± 0.02 for power (Bradley's fairly liberal criterion), the minimum number of replications were found that produced stable results. Table 2 displays the number of replications used by the original study along with the recommended minimum number of replications needed to produce stable results. It appears that fewer replications could have been used to predict power in studies 2 and 5, while too few replications were used in study 4. To predict Type I error, it appears that study 3 could have used fewer replications, whereas study 4 again could have used more replications.

Table 2: Number of Replications Used By the Original Study Along With the Recommended Minimum Number of Replications Required To Produce Stable Results

	Original	Replications Recommended	
Study	Replications Used	Type I Error	Power
1	5,000	5,000	---
2	20,000	---	3,750
3	10,000	5,000	---
4	1,000	8,000	4,000
5	10,000	---	3,750

Dissertation III: Regression Simulation Studies (Supawan, 2004)

The third dissertation replicated 6 simulation studies related to multiple linear regression. The studies included:

(1) Griffiths and Surekha (1986) examined the Type I error rate and power of three tests for heteroscedasticity. They used 5,000 replications, but provided no justification for that choice. The tests they compared were:

(a) Szroeter Test;
(b) Breusch-Pagan Test; and
(c) Goldfeld-Quandt Test.

(2) Pfaffenberger and Dielman (1991) examined the Type I error rate and power of the Filliben test for normality of regression residuals using 6 different statistics. They used 5,000 replications, justifying this choice by their desire to control the maximum standard deviation of the rejection percentage to be < 1.0%. The six statistics they examined were:

(a) Means and the z-transformed residuals;
(b) Medians and the z-transformed residuals;
(c) Means and standardized residuals;
(d) Medians and standardized residuals;

(e) Means and studentized deleted residuals; and
(f) Medians and studentized deleted residuals.

(3) Godfrey (1978) examined the power of the $\chi^2(1)$ heteroscedasticity test for two multiplicative models, Uniform (1,31) and Lognormal (3, 1) using 1,000 replications, but providing no justification for this choice.

(4) Flack and Chang (1987) examined the effects of sample size and the number of noise variables on the frequency of selecting noise variables by using R^2 selection. They used 50 replications, justifying the choice by their belief that it was sufficient to give reliable results.

(5) Hurvich and Tsai (1990) examined the effect of Akaike's Information Criterion (AIC) for model selection on the coverage rates of confidence regions of linear regression. They used 500 replications with no justification provided for their choice.

(6) Olejnik, Mills and Keselman (2000) examined the accuracy of using stepwise regression compared with Wherry's $R^2_{adjusted}$ and Mallow's C_p to select the model in all possible regressions by considering the effect of sample size, the number of noise variables and the correlation between authentic variables. They used 1,000 replications, but provided no justification for their choice.

Dissertation III: Results and Discussion

Studies 1-3 investigated either Type I error rate, power, or both. Using ± 0.005 for Type I error (Bradley's fairly stringent criterion) and ± 0.02 for power (Bradley's fairly liberal criterion), the minimum number of replications were found that produced stable results. Table 3 displays the number of replications used by the original study along with the recommended minimum number of replications needed. In all but two situations, it appears that fewer replications could have been used to predict Type I error and power, with only Study #1 needing substantially more replications than were used to get a stable prediction for power.

Table 3: Number of Replications Used By the Original Study Along With the Recommended Minimum Number of Replications Required To Produce Stable Results

	Original	Replications Recommended	
Study	Replications Used	Type I Error	Power
1	5,000	4,600	7,000
2	5,000	4,200	1,300
3	1,000	---	1,250

Studies 4-6 investigated the proportion of variables selected to be included in the multiple linear regression model. Using ± 0.005 for the proportion of variables selected (Bradley's fairly stringent criterion), the minimum number of replications were found that produced stable results. Table 4 displays the number of replications used by the original study along with the recommended minimum number of replications needed. In each instance, it appears that more replications than were used in the original studies were required to obtain stable results.

Table 4: Number of Replications Used By the Original Study Along With the Recommended Minimum Number of Replications Required To Produce Stable Results

	Original	Replications Recommended
Study	Replications Used	Proportion of Variables Selected
4	50	1,900
5	500	2,000
6	1,000	1,900

Dissertation IV: Quality Control Simulation Studies (Kim, 2005)

The fourth dissertation replicated 6 simulation studies examining the average run length, ARL, of various statistical process control charts. The studies included:

(1) Khoo (2004) examined the ARL property of the Shewhart chart using individual observations for 18 different shifts of size δ. They used 10,000 replications with no justification provided.

(2) Fellner (1990) examined the ARL property of the cumulative sum or CUSUM chart using individual observations for 6 different shifts of size δ. A two-sided CUSUM control chart using decision values H = 2, 3, 4, 5, 6 and reference value K = 0.5 was studied. A total of 30 different scenarios were simulated using 10,000 replications with no justification provided.

(3) Neubauer (1997) examined the ARL property of the exponentially weighted moving average (EWMA) chart using individual observations for 31 different shifts of size δ. The EWMA control chart studied used a weighting constant λ = 0.2 and width of the control limits L = 2.86; 10,000 replications were used with no justification provided.

(4) Khoo and Quah (2003) examined the ARL property of the Hotelling χ^2 chart using individual observation vectors for 18 different shifts of size δ. Only the bivariate case was considered for shifts of size δ. They used 10,000 replications, but provided no justification.

(5) Khoo and Quah (2002) examined the ARL property of two multivariate CUSUM or MCUSUM charts using individual observation vectors for 11 different shifts of size δ. The MC1 control chart studied used p = 2, 3, and 10 variables with reference value k = 0.5 and the MC2 control chart studied used p = 2, 3, and 10 variables with reference values k = 2.5, 3.5, and 10.5. A total of 33 different scenarios were simulated for each MCUSUM chart. They used 10,000 replications, but provided no justification.

(6) Khoo (2003) examined the ARL property of the multivariate EWMA or MEWMA chart using individual observation vectors for 6 different shifts of size δ. The MEWMA control chart studied used p = 2, 4, and 10 variables and weighting constants λ = 0.05, 0.10, and 0.20. A total of 54 different scenarios were simulated. They used 10,000 replications, but provided no justification.

Dissertation IV: Results and Discussion

Statistical control charts are based on the same principles as hypothesis testing. A process is said to be out-of-control if the test of hypotheses is rejected and in-control when it is not rejected, thus, control charts have Type I error rates and power. However, they are typically measured through a different metric, the average run length (ARL). When the process has not changed or shifted, type I error rates can be determined through an in-control ARL. However, when the process has shifted, power can be measured through an out-of-control ARL.

A modified error band, incorporating ARL (e.g. ARL ± 0.1ARL), was used by Chang & Gan (2004) to examine the robustness of the Shewhart control chart with respect to both ARL and SDRL (standard deviation of run length). Chakraborti & van de Wiel (2005) stated this 10% error band might be too wide to detect practical departures of the simulated results from the target value. They used a 2% error band, ARL ± 0.02ARL, to examine the robustness of a non-parametric control chart with respect to its ARL. The 2% error band was used in Dissertation IV.

Table 5 displays the number of replications used by the original study along with the recommended ranges for the minimum number of replications needed to produce stable results for various size shifts within the process. Each process shift is recorded in standard deviations. It appears that fewer replications could have been used to predict ARL in each study, particularly when the shift in the process is large.

Table 5: Number of Replications Used By the Original Study Along With the Recommended Ranges for the Minimum Number of Replications Required To Produce Stable Results for Various Shifts

Study	Original Replications Used	Replications Recommended						
		Shift						
		0.0	0.1-1.0	1.1-2.0	2.1-3.0	3.1-4.0	4.1-5.0	5.1-6.0
1	10,000	6,329	2,500-3,985	2,031-4,375	1,093-2,265	546	78	---
2	10,000	2,187-7,891	1,093-3,203	312-1,953	78-390	78-703	78-312	---
3	10,000	5,391	1,797-6,407	312-1,641	78-703	---	---	---
4	10,000	5,391	3,125-5,703	3,593-5,703	1,171	937-1,015	312	234
5a	10,000	2,891-7,657	1,015-4,843	156-1,407	156-546	156-703	78-156	---
5b	10,000	4,845-5,157	2,187-5,547	937-2,109	546-2,109	156-625	156-390	---
6	10,000	2,036-9,921	625-5,235	625-1,797	312-703	78-546	78-390	---

Conclusion

Monte Carlo simulations have been used extensively in studying the performance of various statistical tests and control charts. Researchers have used a wide range (50-21,000 in the 22 studies replicated herein) of replications in their studies, but seldom provided justifications for the number of replications they used. Currently, there are no empirically based recommendations regarding the required number of replications to ensure accurate results.

Through 4 dissertations, 22 studies from various fields were re-analyzed to provide empirically based recommendations for future simulation studies. In many cases, fewer replications than were used in the original studies were needed to produce stable estimates of the results. In all but two of the situations in which more replications than what was used originally were needed, the original studies began with 1,000 or fewer replications. In general, for most of the studies replicated and most of the statistics calculated, the minimum recommended number of replications was always less than 10,000 and in many cases was less than 5,000. In several situations investigated in these dissertations, 5,000 replications were not sufficient, but seldom were more than 7,500 replications needed. It appears to be the case, generally, that 7,500 to 8,000 replications are sufficient to produce stable results, and in a number of situations, depending upon what characteristic is being estimated, 5,000 replications may be enough.

References

Alyounes, Y. (1999). *A realistic look at one-way ANOVA: A comparison of parametric and nonparametric omnibus tests and their post hoc comparison procedures.* Unpublished Doctoral Dissertation, Ohio University, Athens, OH.

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

Brooks, C. (2002). *Solutions to the review questions at the end of chapter 10 in introductory economics for finance*. Retrieved from http://www.cambridge.org/resources/0521790182/1565-69060.doc.

Brown, M. B., & Forsythe, A. B. (1974). The small sample behavior of some statistics which test the equality of several means. *Technometrics*, *16*(*1*), 129-132.

Chang, T. C., & Gan, F. F. (2004). Shewhart charts for monitoring the variance components. *Journal of Quality Technology*, *36*, 293-308.

Chakraborti, S., & van de Wiel, M. A. (2005). *A nonparametric control chart based on the Mann-Whitney Statistic*. Unpublished manuscript. Retrieved from http://projecteuclid.org/euclid.imsc/1207058271, Digital Object Identifier: doi:10.1214/193940307000000112.

Fan, X., Felsővályi, Á., Sivo, S. A., & Keenan, S. C. (2002). *SAS for Monte Carlo studies: A guide for quantitative researchers* (*1st Ed.*). North Carolina: SAS Institute Inc.

Flack, V. F., & Chang, P. C. (1987). Frequency of selecting noise variables in subset regression analysis: A simulation study. *The American Statistician*, *41*, 84-86.

Fellner, W. H. (1990). Algorithm AS 258: Average run lengths for cumulative sum schemes. *Applied Statistics*, *39*, 402-412.

Gamage, J., & Weerahandi, S. (1998). Size performance of some tests in one-way ANOVA. *Communications in Statistics – Simulation and Computation*, *27*(*3*), 625-640.

Godfrey, L. G. (1978). Testing for multiplicative heteroscedasticity. *Journal of Econometrics*, *8*, 227-236.

Griffiths, W. E., & Surekha, K. (1986). A Monte Carlo evaluation of the power of some tests for heteroscedasticity. *Journal of Econometrics*, *31*, 219-231.

Hsiung, T. H., & Olejnik, S. (1994). Power of pairwise multiple comparison in the unequal variance case. *Communications in Statistics – Simulation and Computation*, *23*(*3*), 691-710.

Hurvich, C. M., & Tsai, C. L. (1990). The impact of model selection on inference in linear regression. *The American Statistician*, *44*, 214-217.

Hutchinson, S. R., & Bandalos, D. L. (1997). A guide to Monte Carlo simulation research for applied researchers. *Journal of Vocational Education Research*, *22*, 233-245.

Kaplan, A. M. (1984). *A comparison of the Kruskal-Wallis test with ANOVA under violations of homogeneity of variance*. Unpublished Doctoral Dissertation, Temple University, Philadelphia, PA.

Khoo, M. B. C. (2003). Increasing the sensitivity of MEWMA control chart. *Quality Engineering*, *16*, 75-85.

Khoo, M. B. C. (2004). Performance measures for the Shewhart $\bar{x}$ control chart. *Quality Engineering*, *16*, 585-590.

Khoo, M. B. C., & Quah, S. H. (2002). Computing the Percentage Points of the Run-Length Distributions of Multivariate CUSUM Control Charts. *Quality Engineering*, *15*, 299-310.

Khoo, M. B. C., & Quah, S. H. (2003). Incorporating run rules into Hotelling's χ^2 control charts. *Quality Engineering*, *15*, 671-675.

Kim, C. (1997). *Robust tests using weighted likelihood estimation*. Unpublished Doctoral Dissertation, Oklahoma State University, Stillwater, OK.

Kim, M. (2005). *Number of replications required in control chart Monte Carlo simulation studies*. Unpublished Doctoral Dissertation, University of Northern Colorado, Greeley, CO.

Klockers, A. J., & Hancock, G. R. (1992). Power of recent multiple comparison procedures as applied to a complete set of planned orthogonal contrasts. *Psychological Bulletin*, *111*(*3*), 505-510.

Morikawa, T., Terao, A., & Iwasaki, M. (1996). Power evaluation of various modified Bonferroni procedures by a Monte Carlo study. *Journal of Biopharmaceutical Statistics*, *6*(*3*), 343-359.

Olejnik, S., Mills, J., & Keselman, H. (2000). Using Wherry's adjusted R^2 and Mallows' C_p for model selection from all possible regressions. *The Journal of Experimental Education*, 68(4), 365-380.

Neubauer, A. S. (1997). The EWMA control chart; properties and comparison with other quality-control procedures by computer simulation. *Clinical Chemistry*, *43*, 594-601.

Pfaffenberger, R. C., & Dielman, T. E. (1991). Testing normality of regression disturbances: A Monte Carlo study of the Filliben test. *Computational Statistics & Data Analysis*, *11*, 265-273.

Preecha, C. (2004). *Numbers of replications required in ANOVA simulation studies*. Unpublished Doctoral Dissertation, University of Northern Colorado, Greeley, CO.

Ramsey, P. H. (2002). Comparison of closed testing procedures for pairwise testing of means. *Psychological Methods*, *7*(*4*), 504-523.

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*), 557-586.

Supawan, P. (2004). *An examination of the number of replications required in regression simulation studies*. Unpublished Doctoral Dissertation, University of Northern Colorado, Greeley, CO.

Ussawarujikulchai, A. (2004). *Number of replications required in multiple comparison procedures Monte Carlo simulation studies*. Unpublished Doctoral Dissertation, University of Northern Colorado, Greeley, CO.

Welch, B. L. (1947). The generalization of Students' problem when several different population variances are involved. *Biometrika*, *34*(*1/2*), 28-35.

Welch, B. L. (1951). On the comparison of several mean values: an alternative approach. *Biometrika*, *38*(*3/4*), 330-336.