COVERAGE AND INTERVAL LENGTH OF WELCH'S AND YUEN'S PROCEDURES FOR SHIFT IN LOCATION AND CHANGE IN SCALE FOR (UN)EQUAL SAMPLE SIZES

by

SAYDEE JONATHAN MENDS-COLE

DISSERTATION

Submitted to the Graduate School

of Wayne State University,

Detroit, Michigan

in partial fulfillment of the requirements

for the degree of

DOCTOR OF EDUCATION

2006

MAJOR: EDUCATIONAL EVALUATION & RESEARCH

Approved by:

Advisor	Date

UMI Number: 3210997

ProQuest Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346

ACKNOWLEDGEMENTS

I appreciated the assistance provided by the members of my committee. The members included my advisor, Shlomo Sawilowsky, Gail Fahoome and David Moxley.

TABLE OF CONTENTS

Section **Page**

Section **Page**

Section **Page**

Section **Page**

Section **Page**

LIST OF TABLES

LIST OF FIGURES

Figure **Page**

CHAPTER 1

Introduction

Background

How well does the sample effect $\overline{X}_1 - \overline{X}_2$ estimate the population effect $\mu_1 - \mu_2$? When inferring from a sample, the range of values plausible for the population effect, a confidence interval is the appropriate statistical technique. A confidence interval provides estimates of the range of values that permit a statement about the population parameter (Knapp & Sawilowsky, 2001; Hinkle, Wiersma, & Jurs, 1998).

Knowing the range of plausible values for the population parameter. There are many influences motivating the use of confidence intervals. *First*, if no hypothesis exists, however one seeks statistics on the range of values that are plausible for the population parameter, Knapp (1999) recommended using a confidence interval.

Second, Hoenig and Heisey (2001), Stiedl, Hayes, and Schauber (1997), and Wilkinson and Taskforce on Statistical Inference (1999) recommended using confidence intervals instead of retrospective power analysis. Retrospective power analysis employs statistical power (a) following the statistical analysis, and (b) with a sample estimate of effect

size. The use of statistical power in retrospective power analysis is untenable (Zumbo & Hubley, 1998; Knapp, 1999, Hoenig & Heisey, 2001).

(a) *Employing statistical power following the statistical analysis*. Statistical power analysis serves as a tool for planning a study to achieve a level of power. The use of statistical power analysis is defined (Cohen, 1988, pp. 14-16): (a) Knowing the alpha level, sample size, and a conjecture about the population effect size, statistical power can be determined. (b) Selecting a level of statistical power, a conjecture of effect size and alpha level can be used to find the sample size needed, before doing the statistical test.

(b) *Employing statistical power with a sample estimate of effect size*. Falk et al., Taylor, and Muller (as cited in Knapp, 1999) advocated using retrospective power analysis for determining the statistical power given the sample size and the effect size observed for the data. Applying retrospective power analysis to the sample effect size is not valid as defined and applied in statistical power analysis. Use of the sample effect size to estimate statistical power contradicts the definition of statistical power. Statistical power is the probability of rejecting the null hypothesis $H_o : \mu_1 - \mu_2 = 0$, if it is false (Wilcox, 1996; Zumbo & Hubley, 1998) and detecting an effect if it exists, that is $\mu_1 - \mu_2 \neq 0$. As defined, statistical

power depends on the hypothesized value of the population effect, and statistical power does not depend on sample effect. Hoenig and Heisey (2001) and Stiedl, Hayes, and Schauber (1997) made further arguments based on results from the formula for statistical power using sample effects.

As such, Hoenig and Heisey (2001), Stiedl, Hayes, and Schauber (1997), and Wilkinson and Taskforce on Statistical Inference (1999) recommended using confidence intervals instead of retrospective power analysis. In this respect, confidence intervals provide an indication of statistical precision. The interval is more precise if the length is narrower.

Replacing hypothesis tests with confidence intervals. *Third*, Cohen (1994), and Schmidt (1996) endorsed the use of confidence intervals instead of statistical hypothesis tests:

(a) Cohen (1994) stated that confidence intervals provide evidence for rejecting H_0. That is, values within the confidence limits are tenable values of the population parameter. Furthermore, the values are not rejected, based on the data (Hinkle, Wiersma, & Jurs, 1998). One rejects values that are not within the limits.

(b) Cohen (1994) stated that interval lengths provide an indication of statistical power. Because sample size inversely relates to the standard

error of the effect, increasing sample size yields narrower confidence intervals (Wilcox, 1996, 2001; Hinkle, Wiersma, & Jurs, 1998). A narrower interval length implies a smaller range of values. A smaller range implies that it is less likely that the interval will enclose zero. This implies that it is more likely that one will reject the null hypothesis, if it is false, that is, $\mu_1 - \mu_2 \neq 0$.

The preceding statements may suggest that I advocate the use of confidence intervals over statistical hypothesis tests. I do not advocate the use of confidence intervals over statistical hypothesis tests. The purpose of this study was to assess the statistical properties of confidence intervals, i.e., interval length and probabilities of coverage. The next section describes the relation between confidence intervals and hypothesis tests.

Type I Error Rate & Probability of Coverage

Confidence intervals and hypothesis tests are related. (1) The probability of coverage is $1-\hat{\alpha}$. Violations of assumptions such as heteroscedasticity and skewness that influence the Type I error rate also influence the probability of coverage. (2) Outliers and size of the sample, that influence the size of the standard error, influence statistical power

(Wilcox, 2001). Likewise, outliers and sample size can have an influence upon the interval length (Hinkle, Wiersma, & Jurs, 1998; Wilcox, 1996).

Skewness (Blair, 1981; Micceri, 1989) and heteroscedasticity (Kirk, 1995; Sawilowsky, 2002b) are likely to occur. This will mean that probability of coverage and interval length will be less accurate than results observed when comparing distributions that exhibit normality and homoscedasticity, with the use of the independent samples t test.

Problem Statement

The statement of the problem addressed key relationships related to the probability of coverage for confidence intervals. First, the relation between confidence intervals and significance tests was addressed. Second, the relation between skewness of the distribution and Type I error rate was addressed. Third, the relation between heteroscedasticity and Type I error rate was addressed. Fourth, alternative procedures for handling skewness and heteroscedasticity were addressed.

Confidence Intervals & Parametric Tests

Because confidence intervals use sample size calculations for the degrees of freedom used in accessing the critical value of the t-statistic, $t_{1-\alpha/2}$, and the pooled estimate of the population variance (Hinkle,

Wiersma, Jurs, 1998, p. 256-257), assumption violations that influence the rate of Type I errors also influence the probability of coverage (May, 2003). The rate of Type I errors is effected by the violation of parametric test assumptions (Glass, Peckham, & Sanders, 1972; Algina, Oshima, Lin, 1994; Sawilowsky & Blair, 1992).

Hinkle and Wilcox explained parametric test assumptions with respect to the nature of the distribution of the population and the group sampling procedure. The assumptions stated: (a) Random samples were taken from normally distributed populations (Wilcox, 1996, Hinkle, Wiersma, Jurs, 1998). (b) The variances for each population were equal (Wilcox, 1996, Hinkle et al., 1998). (c) The observations were random samples from defined populations (Hinkle et al., 1998). (d) The samples were independent (Wilcox, 1996, Hinkle et al., 1998).

This study is primarily concerned with violations of the assumptions of normality and homoscedasticity. Violation of the assumption of normality occurs frequently (Blair, 1981; Micceri, 1989).

Skewed Distributions

Samples from skewed and nonnormal populations are quite frequent in quantitative analysis (Blair, 1981). Micceri (1989) surveyed 440 published data sets. The p-value of the Kolmogorov-Smirnov test showed

the distributions of each data set to be significantly different from a normal distribution ($p < .01$).

For the independent samples t test, nonnormality influences the Type I error rate (Sawilowsky & Blair, 1992, Algina, Oshima, & Lin, 1994). Monte Carlo Type I error results (Sawilowsky & Blair, 1992) suggested that probability of coverage will be greater than $1-\alpha$ for skewed distributions, particularly if absolute skewness is within the range from 1.25 to 1.75. Setting the alpha level at 0.05, if Type I error rate is less than 0.05, probability of coverage is greater than 0.95. Yet, Sawilowsky and Blair observed that the independent samples t-test was robust: (a) if the test was two tailed rather than one tailed, (b) if sample sizes were about equal, and (c) if sample sizes were 25 or more.

Further, outliers present in skewed distributions inflate the variance and the interval length. Positively or negatively skewed distributions are characterized by a relatively small number of observations in the right or left tails of the distributions, respectively. That is, opposite the orientation of skewness is a greater mass of observations. If the distribution is skewed right, the mass of observations is to the left and vice versa. A researcher may deem tail values to be outliers based on the median absolute deviation (Wilcox, 2001, p. 36):

$$| X_i - M_X | > 2(MAD) / 0.6745 \quad (1)$$

X is deemed an outlier if the absolute difference between X and the median (M_X) is greater than the product of 2 and the median absolute deviation (MAD) divided by 0.6745. Where the estimate of the variance depends on the sum of squared deviations from the mean $\sum_{i=1}^{n}(X_i - \overline{X})^2$, outliers inflate the variance (and standard deviation).

Outliers that influence the size of the standard deviation will also influence the length of the confidence interval. Where the standard error of the mean is the estimate of the variance divided by its respective sample size, interval length will be determined consequently:

$$S_{\bar{x}i} \rightarrow SE_{\bar{x}1-\bar{x}2} \rightarrow 2(t_{1-\alpha/2})(SE_{\bar{x}1-\bar{x}2}).$$

If outliers exist, a researcher may want to identify and discard them. However, I examined a procedure that omits the most and least extreme values before statistical analysis – Yuen's procedure. Another violation that impinges upon the probability of coverage and interval length of the independent samples t test is heteroscedasticity.

Heteroscedasticity

Heteroscedasticity is likely to occur - why. Heteroscedasticity is likely to occur for three reasons. The first reason is that if groups differ

concerning their means, it is likely that they differ concerning their variances. That is, educational and psychological statisticians – such as, Sawilowsky and Blair, 1992, p. 358; Sawilowsky, 2002b, p. 467, Wilcox, 1996, p. 149; and Wilcox, 2001, p. 87 – experienced with the analysis of like data in their respective disciplines made this observation.

The second reason is that statistical tests of $H_0 : \sigma_1 = \sigma_2$ do not have enough statistical power (Wilcox, 1996; p. 135; Grissom, 200, p. 157). In particular, C. Markowski and E. Markowski (1990) found that when $\sigma_2 / \sigma_1 \geq 1.4$ or $\sigma_2 / \sigma_1 \leq 0.7$, the power of the ratio of variances F-test to detect this difference was less than 0.50. They obtained results for unequal group sample sizes ranging from 5 to 40. Their results suggested that one is less likely to detect heteroscedasticity via the statistic.

Third, another situation that may result in heteroscedasticity is standardization of treatment application. From one measurement to the next, if the manner in which the treatment is applied differs, the effects of treatment on the outcome measures will differ (Cook & Campbell, 1979; Kirk, 1995). Because of the different effects of treatment on outcome measures, the scores will vary widely from one another than if treatment implementation was standardized.

Scores spread-out from one another differ widely from the mean (Wilcox, 2002, pp. 26-27). Compared with scores obtained from standardized treatment implementation, the scores (from nonstandardized implementation) will exhibit a wider range. Endpoints of this wider range will increase the sum of squared deviations from the mean, and the variance.

Variance is a function of the sum of squared deviations from the mean $\sum_{i=1}^{n}(X_i - \overline{X})^2$. Scores that deviate farther from the mean inflate the variance (Wilcox, 2002, pp. 26-27). Therefore, for the treatment group, these set of conditions inflated the variance. If the variance for the treatment group is inflated, the treatment group variance will be larger than the control group variance, resulting in heteroscedasticity.

Effects of Heteroscedasticity

Effect on probability of coverage. Heteroscedasticity attenuates probability of coverage and augments interval length. When using the independent-samples t test, heteroscedasticity will influence the Type I error rate for the test statistic and the confidence interval length. The Type I error rate can be influenced by 20% or more. That is, alpha level equal to 0.05, but Type I error rate is greater or equal to 0.06 (Algina,

Oshima, & Lin, 1994; Penfield, 1994). It can be more than twice the alpha level when sample sizes are unequal. That is, alpha level is 0.05, but Type I error rate greater or equal to 0.10 (Algina, Oshima, & Lin, 1994; Penfield, 1994). Where the probability of coverage is $1-\hat{\alpha}$ (May, 2003), heteroscedasticity implies probability of coverage less than $1-\alpha$. This implies that when one attempts to estimate the population parameter for an effect $\mu_1 - \mu_2$, stating that the probability of coverage was $1-\alpha$ is incorrect when, in fact, the probability of coverage was less than $1-\alpha$.

Effect on interval length. Heteroscedasticity results in wider interval lengths. The standard error of the effect ($SE_{\bar{x}1-\bar{x}2}$) used to calculate the confidence interval is the square root of the sum of the squared standard errors of the mean ($S_{\bar{x}i}$) for each group (Hinkle et al., 1998). The standard error of the mean for each group relates directly to the variance for each group (Hinkle et al., 1998). To the extent that the variance for the treatment group increases, the standard error increases. To the extent that the standard error increases, the standard error of the effect increases. Interval length will increase (Wilcox, 2001), where interval length relates directly to the standard error of the effect: $2(t_{1-\alpha/2})(SE_{\bar{x}1-\bar{x}2})$. (Where application of the treatment results in a more homogeneous

group (Sawilowsky, 2002b), heteroscedasticity may result in narrower interval lengths.)

Wider interval lengths mean less precise confidence intervals (Hinkle et al., 1998). It may be the difference between being able to state that the population parameter is within the range of 34 to 36 as opposed to 33 to 39, for example.

Recommendations for Handling Heteroscedasticity & Skewness

If samples exhibit nonnormality or heteroscedasticity, Kuehl (1994) and Kirk (1995) recommended transformations for changing the data so that it is normal and homoscedastic in form (Kuehl, 1994; Kirk, 1995). One problem with this procedure is that the measures of central tendency compared in the transformed samples are different from those compared in the original sample (Wilcox, 1996). Another problem with these procedures is that they do not handle the low statistical power caused by nonnormality (Wilcox, 1996, 2001). Wilcox (1996) recommended alternative procedures for handling problems of heteroscedasticity and nonnormality.

Alternative Procedures

Wilcox (1996) and Hinkle, Wiersma, and Jurs (1998) recommended

Welch's procedure for a more accurate Type I error rate (than the Type I error rate for the independent samples t test) when random samples exhibit heteroscedasticity. Wilcox (1996; 2003) recommended Yuen's procedure instead of Welch's procedure. Wilcox recommended Yuen's procedure because, under nonnormality, it exhibited probability coverage that was more accurate.

For the t test for independent-samples design (Kirk, 1995, p. 29), no alternative statistical tests are robust to heteroscedasticity (S. Sawilowsky, personal communication, June, 2005). Yet Yuen's and Welch's procedures exhibited Type I error rates that better approximated the alpha level (Algina, Oshima, & Lin, 1994; Luh & Guo, 2000), given heteroscedasticity. Both procedures displayed Type I error rates near alpha level, 0.5(alpha level)≤Type I error rate≤ 1.5(alpha level), for normally distributed samples (Yuen, 1974, Luh & Guo, 2000). In contrast, Yuen's procedure outperformed Welch's procedure when samples were nonnormal (Yuen, 1974, Luh & Guo, 2000, Wilcox, 1996).

For example, both Welch's and Yuen's procedures exhibited probabilities of coverage within the range from 0.925 to 0.975 when skewness and kurtosis of the distributions are normal (Luh & Guo, 2000), (n_1, n_2)=(12, 24), (σ_1: σ_2)=(1:4). Yuen's procedure did well when skewness is as

extreme as 6.2 and kurtosis is 111. Results by Luh and Guo (2000) suggested that Welch's procedure under performs Yuen's procedure in terms of Type I error and probability of coverage.

The problem with these recommendations that is central to this study is that (1) they were based on random samples generated using mathematical functions (Sawilowsky & Blair, 1992; Sawilowsky & Fahoome, 2003). Accordingly, the results may not generalize to that observed in applied settings. (2) The studies that provided the recommendations emphasized Type I error rate and statistical power of the procedures. The Type I error and statistical power indirectly relate to the confidence interval (Knapp & Sawilowsky, 2001). However, interval length and probability of coverage directly pertain to the confidence interval.

Further, recommendations for using Yuen's and Welch's procedures were based on Monte Carlo studies that generated random samples using mathematical functions (Wilcox, 1994, Luh & Guo, 2000, Algina, Oshima, & Lin, 1994; Keselman, Wilcox, Kowalchuk, & Olejnik, 2002), termed functional samples. However, functional samples are not representative of the kind observed in quantitative analysis (Sawilowsky & Blair, 1992; Micceri, 1989). To the extent that the sample observations

represent data in applied situations, the results generalize to such settings (Sawilowsky & Blair, 1992; Sawilowsky & Fahoome, 2003).

In addition, these studies emphasized the Type I error rate and statistical power of the procedures (Wilcox, 1994, Luh & Guo, 2000, Algina, Oshima, & Lin, 1994; Keselman, Wilcox, Kowalchuk, & Olejnik, 2002). However, if a researcher is not interested in testing a hypothesis, but wants find the range of values wherein the parameter may lie (Sawilowsky, 2003; Knapp & Sawilowsky, 2001), that researcher should compute a confidence interval (Knapp, 1999). In computing a confidence interval, sizes of probability of coverage and interval length are directly relevant (Wilcox, 1996; Hinkle et al., 1998). That is, evidence that when population variances are heterogeneous the probability of coverage was 0.85 (Wilcox, 2001) would dissuade one from using Student's procedure.

Similar statements can be made about the interval length. Yet, publications in education and psychology did not furnish relevant data on the interval lengths for Yuen's and Welch's procedures. The next section summarizes the problem.

Summary of the Problem Statement

Samples from nonnormal populations were quite frequent in quantitative analysis (Blair, 1981, Micceri, 1989). Nonnormality influences

the rate of Type I errors (Sawilowsky & Blair, 1992, Algina, Oshima, & Lin, 1994), implying that nonnormality influences the probability of coverage (May, 2003).

Heteroscedasticity is likely to occur: (1) if groups differ concerning their means (Sawilowsky & Blair, 1992, p. 358; Sawilowsky, 2002B, p. 467, Wilcox, 1996, p. 149; Wilcox, 2001, p. 87), (2) statistical tests of $H_0 : \sigma_1 = \sigma_2$ possess low statistical power (Wilcox, 1996; p. 135; Grissom, 200, p. 157), implying that variance heterogeneity can occur without being detected, and (3) outliers occur in sample data (Wilcox, 1996; Wilcox, 2001). Heteroscedasticity will influence the rate Type I error for the test statistic and the confidence interval length.

Welch's and Yuen's procedures are recommended when samples exhibit heteroscedasticity and nonnormality (Wilcox, 1996; Hinkle et al., 1998). The problem is that (1) the recommendations are based on functional samples that may not generalize to that observed in research practice. (2) The recommendations emphasize Type I error rate and the statistical power of the procedures. When using a confidence interval, interval length and probability of coverage define pertinent characteristics of the interval. Accordingly, the purpose of this study was

to assess the probability of coverage and interval length for Yuen's and Welch's procedures under different conditions.

Value & Purpose of Study

This study adds to knowledge about the probability of coverage and interval length for Yuen's and Welch's procedures. This knowledge was gained using empirical distributions (e.g., Sawilowsky & Blair, 1992).

Specifically, the purpose of this study was to assess the probability of coverage and the interval length for:

(a) Welch's procedure, and

(b) Yuen's procedure.

This was achieved: (a) using data sets that were not normally distributed (i.e., Sawilowsky & Blair, 1992), (b) under conditions of heteroscedasticity, and (c) for unequal group sample sizes. The next section defines the terms used throughout the study.

Definition of Terms

The following terms were used throughout the study. Table 1 lists the terms. Throughout the definitions, the terms reference the analysis of the difference between means.

Table 1

Definition of Statistical Terms

Term	Definition
Alpha level (α)	The prespecified level of significance used in selecting the critical value.
Behrens-Fisher problem	Violation of the homoscedasticity assumption of the independent-samples t-test that occurs when the ratio of variances is not equal to one (Sawilowsky, 2002b, p. 461) and population means are not equal $\mu_1 \neq \mu_2$.
Central limit theorem	As *n* gets larger, the distribution of the mean approaches a normal distribution with finite mean and variance (Wilcox, 1996, p. 85).
Confidence interval	An upper and lower limit that, before obtaining a random sample, had a $1-\alpha$ probability of containing the population effect. The upper and lower limits that result from a specific random sample either will or will not contain the population parameter (Sawilowsky, 2003, pp. 128-129).
Confidence level	The $1-\alpha$ probability that the confidence interval will enclose the population parameter before using sample estimates to obtain the limits for an interval.
Critical value	An extreme value of the test statistic that is unlikely to occur if the null hypothesis is true (Sheskin, 2000, p. 27).
Effect size	The absolute value of the difference between the means for the treatment and control conditions divided by the standard deviation (Cohen, 1988).

Table 1

Definition of Statistical Terms (continued)

Term	Definition
Heteroscedasticity	$\sigma_{max}/\sigma_{min} \geq 2$.
Interval length or width	$2(t_{1-\alpha/2})(SE_{\bar{x}1-\bar{x}2})$. Where $t_{1-\alpha/2}$ is the critical value; and $SE_{\bar{x}1-\bar{x}2}$ is the standard error of the effect.
Kurtosis	A measure of the degree to which a distribution is peaked (Spiegel, 1994): $Kurtosis = \frac{M_4}{M_2^2} - 3$. Where $M_r = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^r}{n}$; X_i are sample observations; n is the number of observations; & $\overline{X}$ is the mean of the observations.
Monte Carlo experimentation	A process that involves the use of a computer program that serves as a computer model. The computer model represents the application of a statistical technique (e.g., a significance test or confidence interval) to a sample of observations. The computer program uses sampling with replacement (from a frequency distribution) to determine the long-run average of a parameter.

Table 1

Definition of Statistical Terms (continued)

Term	Definition
Normal distribution	A bell-shaped curve having a skewness value of 0 and a kurtosis value of 0 (Spiegel, 1994).
Parametric tests	Statistical tests of hypotheses about the population parameter that require assumptions about the distribution(s) of the population (Hinkle et al., 1998).
Probability of coverage	The ratio of the number of times that the population parameter was within the confidence limits compared with the number of confidence intervals calculated.
Retrospective power analysis	(1) The use of sample estimates of effect ($\overline{X}_1 - \overline{X}_2$), variability (S_p), and sample size (n) to estimate statistical power. (2) A researcher evaluates the statistical power as evidence of the null hypothesis that was not rejected (Hoenig & Heisey, 2001).
Robustness	A condition satisfied when the Type I error rate is within the range from 0.9α to 1.1α or, alternatively, it is within the range from 0.5α to 1.5α (Bradley, 1978).
Skewness	A measure of the degree to which a distribution is asymmetric (Spiegel, 1994): $Skewness = \frac{M_3}{M_2^{3/2}}$.
Statistical power	The ratio of the number of significant statistical tests to the number of statistical tests done. The ratio is obtained having knowledge that the population means are not equal $\mu_1 - \mu_2 = k, k \neq 0$.
Trimming	Omitting a number of the largest scores and an equivalent number of the smallest scores from the sample.
Type I error rate ($\hat{\alpha}$)	The ratio of the number of statistically significant t-tests to the number t-tests done. The ratio is obtained when it is known that the population means are equal $\mu_1 = \mu_2$.

Table 1

Definition of Statistical Terms (continued)

Term	Definition
Welch's procedure	A statistical technique for computing the confidence interval for the population effect $\mu_1 - \mu_2$ for the difference between population means. Use of this procedure does not assume homoscedasticity.
Winsorization	(1) Replacing a number of the largest scores with the maximum score for the trimmed version of the same sample; and, (2) replacing an equivalent number of the smallest scores with the minimum score for the trimmed version of the same sample.
Yuen's procedure	A statistical technique for computing the confidence interval for the population effect $\mu_{t1} - \mu_{t2}$.

Throughout this study, a procedure was labeled 'robust'. I would like to distinguish the use of the term 'robust' from the branch of statistics known as 'robust statistics'. When I make such statements, robustness concerns violation of the assumptions of homoscedasticity and normality. Here, robustness refers to the Type I error rate or probability of coverage. In particular, for $\alpha = 0.05$, if the Type I error rate is within the range of 0.025-0.075, the test is considered robust. This implies that probability of coverage within the range of 0.925-0.975 is robust. The next section outlines the assumptions of this study.

Assumptions

The assumptions of this study relate to inferences made from confidence intervals, the application of confidence intervals to a specific situation, and the sampling distribution. First, this study assumed that the major statistical inference of a researcher is that of making an inference about the size of the population effect and the precision with which the effect is measured. Second, in using the distributions examined by Sawilowsky and Blair (1992), this study assumed that these are of the kind observed in quantitative analysis.

Criteria for the Inclusion of Studies

The next chapter reviews the literature concerning the results for Student's, Yuen's, and Welch's procedures. I gave four main considerations to the inclusion of studies in this review of these procedures. One, the studies involved Monte Carlo Experimentation. Two, the results dealt with either Student's, Yuen's or Welch's procedure. Three, the results entailed the effects of unequal group sample sizes. Four, the results entailed the effects of nonnormality or heteroscedasticity. Five, the results presented information about the Type I error rate.

CHAPTER 2

Literature Review

Confidence Intervals

Type I error: Significance Test & the Confidence Interval

The equation for the confidence interval is (Hinkle et al., 1998):

$$\overline{X}_1 - \overline{X}_2 \mp t_{1-\alpha/2} SE_{\bar{x}1-\bar{x}2} . \quad (2)$$

Where $\overline{X}_i$ refers to the mean for group i; and

$t_{1-\alpha/2}$ refers to the critical value of the test distribution.

$$SE_{\bar{x}1-\bar{x}2} = \sqrt{s^2_{\bar{x}1} + s^2_{\bar{x}2}} . \quad (3)$$

$$s^2_{\bar{x}i} = \frac{s_i^2}{n_i} . \quad (4)$$

Where n_i is the sample size for group i; and

s_i^2 is the variance for group i.

Skewness and heteroscedasticity effect the probability of coverage for a confidence interval that depends on the critical value of a parametric test. Normality and homoscedasticity are assumptions of parametric tests (Hinkle et al., 1998). Violations of the assumptions of parametric tests are often reported concerning the effects on the Type I error rate, not the effects on probability of coverage (e.g., Glass,

Peckham, & Sanders, 1972; Sawilowsky & Blair, 1992; Wilcox, 1998; Algina, Oshima, & Lin, 1994). Because the probability of coverage is $1-\hat{\alpha}$ (May, 2003), violations that influence the Type I error rate also influence the probability of coverage.

For example, Glass and consociates have shown that when n's are unequal and variances are unequal, the Type I error rate of the independent samples t test differs from the alpha level:

1. When *n's* are unequal and variances are heterogeneous, the actual significance level may greatly exceed the nominal significance level *when samples with smaller n's come from populations with larger variances.*
2. When *n's* are unequal and variances are heterogeneous, the actual significance level may be greatly exceeded by the nominal significance level *when samples with smaller n's come from populations with smaller variances.* (Glass, Peckham, & Sanders, 1972, p. 245).

According to the results from Glass et al. (1972), (1) the probability of coverage should be less than $1-\alpha$ when smaller samples are drawn from populations with larger variation. (2) The probability of coverage should

be greater than $1-\alpha$ when smaller samples are drawn populations with smaller variances.

As an aside, Glass et al. (1972) reviewed the literature to determine the effect of violating parametric test assumptions on fixed effects analysis of variance and analysis of covariance. The authors purposely avoided findings related to nonparametric statistics (Glass et al., 1972, pp. 237, 255). Blair (1981) diverged from the views of Glass et al. Blair suggested that Glass et al. overemphasized the Type I error rates in their decision to focus on analysis of variance and covariance. Blair showed that the Wilcoxon Rank Sum test exhibited greater statistical power than the independent samples t test, under nonnormality.

Furthermore, Glass et al. (1972) contended that the Type I error rates for analysis of variance do well, assuming nonnormality. Bradley (1984) explained how Glass et al. overlooked studies done by him that show otherwise. Bradley explained that under heteroscedasticity and unequal sizes, the t-test was not robust.

Heteroscedasticity also influences interval length. The length of a confidence interval is directly related to the size of the standard deviation (Hinkle et al., 1998). The next section discusses violations of normality observed in quantitative analysis.

Distributional Types Observed in Practice

Skewed distributions are frequent in quantitative analysis. The experience of researchers such as Blair (1981) and Micceri (1989) suggested that skewed distributions occur frequently. Blair (1981) surmised that skewed data are "of particular interest because they are examples of shapes that tend, in the experience of this researcher, to recur with some frequency in educational data" (p. 504).

By way of 440 datasets, Micceri (1989) confirmed the frequency of nonnormality. He found that "The Kolmogorov-Smirnov test of normality…found [all] of the distributions to be statistically significantly nonnormal at the 0.01 alpha level" (Micceri, 1989, p. 161). Findings from Micceri (1989) showed that more than 70% of the 440 distributions displayed some form of skewness, i.e., $0.30 \leq \text{skewness} \leq 2.00$. The implications of skewness are important for the validity of the t-test (Blair, 1981; Wilcox, 1998). In the next section, the results concerning the probability of coverage for the independent samples t (under violations of normality and homoscedasticity) are outlined.

Monte Carlo Results for the Independent Samples t Test

Results for the Independent Samples t Test

Beginning, take into consideration what results of the t-test implied

about probability of coverage. The results were observed for $\alpha = 0.05$. Studies showed that when distributions were skewed, sample sizes were unequal, sample sizes were greater than 30, and variances were homogeneous, probability of coverage was within the range from 0.925 to 0.975 (Boneau, 1960, Sawilowsky & Blair, 1992, Algina, Oshima, & Lin, 1994).

When distributions were skewed, sample sizes were uneven, and variances differed, probability of coverage was outside the range of 0.925 to 0.975. Furthermore, the probability of coverage was outside the range from 0.925 to 0.975 for small group sample sizes ranging from 5 to 15, samples for which the population skewness was 1.65 (Sawilowsky & Blair, 1992).

Boneau's study. Boneau (1960) compared the difference between the Type I error rate and alpha level. Boneau calculated the t-test 1,000 times to find the rates of Type I errors. For sizes of (5, 15) and standard deviations of (2:1), the probability of coverage was 0.84. This finding was observed from a normal distribution. This suggests that the effect of small and unequal sample sizes and heteroscedasticity inflated Type I error and reduced the probability of coverage.

For sizes of (5, 5) and standard deviations of (1:2), the probability of coverage was 0.917. Boneau (1960) called this distribution an exponential distribution. Mooney (1997) provided the specifications for skewness and

kurtosis. The skewness is 2.00. The kurtosis is 6.00. This suggests that the effect of small sample sizes and heteroscedasticity inflate Type I error rate and reduce the probability of coverage where skewness was 2.00.

Sawilowsky & Blair Study. Sawilowsky and Blair (1992) examined the Type I error rate of the independent samples t test. The skewness for one distribution was 1.65. The kurtosis was 0.98. When the confidence level was set at 0.95, the probability of coverage ranged from 0.997, for sample sizes of 5 and 15, to 0.948, for sample sizes of 40 and 40. For small sample sizes, the probability of coverage was not robust. For larger sample sizes, the probability of coverage was robust. Sawilowsky and Blair (1992) did 10,000 iterations of the independent-samples t test when sampling from the same population.

Sawilowsky and Blair (1992) also found that when sampling from a distribution with a skew of –1.33 and a kurtosis of 1.11, the probability of coverage ranged from 0.956, for sample sizes of 5 and 15, to 0.947, for sample sizes of 40 and 40. The probabilities of coverage were within the range from 0.925 to 0.975. The findings suggested that when distributions exhibited skewness values of -1.33 or 1.65, and, sample sizes of 5 and 15, the Type I error rate was less than 0.05 and the probability of coverage was greater than 0.95. For moderate and equal sample sizes (n_i>25),

negligible differences were found between the Type I error rate and alpha level.

Algina et al.'s study. Algina, Oshima, and Lin (1994) examined the Type I error rate of the independent samples t test. The skewness of the distribution was 6.10. A sample of size 33 was associated with a variance of 3. A sample of size 67 was associated with a variance of 1. The probability of coverage was 0.77. When the ratio of variances was 2:1, instead of 3:1, the probability of coverage was 0.84. When variances were equal, the probability of coverage was 0.95. The findings suggested that for skewed distributions with unequal sample sizes and heteroscedasticity, Type I error rate was inflated and probability of coverage was reduced.

Penfield's study. Penfield (1994) gauged the Type I error rate of the independent samples t test. Though Penfield gauged Type I error rate for different levels of kurtosis: $-1.00 \leq \text{kurtosis} \leq 3.75$, Penfield observed that kurtosis negligibly effected the Type I error rate. Consequently, Penfield reported the median Type I error rate across kurtosis levels for a specified level of skewness: $0.50 \leq skewness \leq 1.50$. Penfield did 10,000 calculations of the t test.

The Type I error rates from Penfield's (1994) study suggested that

when n_1=10, n_2=20, probability of coverage ranged from 0.950 (skewness=0.5) to 0.956 (skewness = 1.5). When n_1 and n_2 was 20, probability of coverage ranged from 0.946 (skewness=0) to 0.950 (skewness = 1.50). When n_1=10, n_2=20, $\sigma_1=2\sigma_2$, probability of coverage ranged from 0.888 (skewness = 0) to 0.895 (skewness = 1.0). When $\sigma_2=2\sigma_1$, probability of coverage ranged from 0.979 (skewness = 1.5) to 0.983 (skewness = 0.5).

One arrives at several conclusions from the Penfield study. Assuming homoscedasticity, $1-\hat{\alpha}$ was within the range 0.925-0.975. When $S^2_{\min}$ was divided by $n_{\min}$ (for the purpose of computing the standard error of the effect), $1-\hat{\alpha}$ was greater than 0.975. When $S^2_{\max}$ was divided by $n_{\min}$, $1-\hat{\alpha}$ was less than 0.925.

Summary. In summary, for $1-\hat{\alpha}$ greater than the confidence level, skewness of the distribution resulted in Type I error rates that were less than expected. For probabilities of coverage that were less than the alpha level, heteroscedasticity resulted in Type I error rates that were greater than the alpha level. The next section discussed the results for the independent samples t test.

Discussion of the Results for the independent samples t test

Skewness increases probability of coverage (Sawilowsky & Blair,

1992). Larger probabilities of coverage also imply larger interval lengths (Hinkle et al., 1998; Wilcox, 1996). For greater statistical precision, narrow interval length is preferred. Without evidence of interval length (e.g., Sawilowsky & Blair, 1992; Algina, Oshima, & Lin, 1994; & Penfield, 1994), no statement can be made that higher probability of coverage did not result at a cost of wider interval length.

Presenting confidence intervals that depend on heterogeneous variances can mean that the confidence level is greater than the probability of coverage. When estimating the confidence interval, probability of coverage may be as low as 0.770 (Algina, Oshima, & Lin, 1994) or 0.895 (Penfield, 1994). Yet, the researcher selects a 0.95 confidence level.

Heteroscedasticity can increase probability of coverage. The smaller variance divided by the smaller sample size results in Type I error rates less than expected (Penfield, 1994; Glass et al., 1972). Statistical procedures have been recommended for accurate probabilities of coverage. The procedures were developed for heteroscedastic and skewed data.

Recommendations for Alternative Procedures

Hinkle et al. (1998), and Wilcox (1996) recommended Welch's

procedure for dealing with heteroscedasticity. Wilcox (1996) recommended Yuen's procedure because Yuen's procedure exhibited accurate probabilities of coverage for skewed distributions. The computational details of Yuen's and Welch's procedures were outlined in several applied statistics textbooks, (e.g., Hinkle et al., 1998; Wilcox, 1996; Wilcox, 2001, Kirk, 1995).

Wilcox (1996) recommended several other procedures for comparing the median and one-step M-estimators. Yet, these procedures are computationally intensive. The procedures involve probabilities from a beta distribution or the use of a bootstrap procedure (Wilcox, 1996, pp. 71-72, 148). For these reasons, attention was given to Yuen's and Welch's procedures. Furthermore, Wilcox (1996) suggested that Yuen's procedure does as well for different criteria – such as, statistical power and probability of coverage – for skewed distributions. The next section reviewed the literature to show how well these procedures did under conditions of nonnormality and heteroscedasticity.

It is noted that authors of these studies also studied different procedures. For example, Luh and Guo (2000) examined Johnson's trimmed t test. Keselman, Wilcox, Kowalchuk, and Olejnik (2002) studied Zhou's statistic. Penfield (1994) examined the Wilcoxon Rank Sum and

Terry-Hoeffding tests. The focus of this study was on Yuen's and Welch's procedures. As a result, other procedures were not included.

Monte Carlo Results for Yuen's & Welch's Procedures

Trimming & Winsorization

Before reviewing the results concerning Yuen's and Welch's procedures, a discussion of the adjustments for outliers for Yuen's procedure was provided. Computations for Yuen's procedure involved both trimming and Winsorization. Trimming was used for the computation of the 20% trimmed mean. Winsorization was relevant in solving for the standard error of the trimmed mean. Described in this section are the following: (a) how much trimming was recommended, (b) the procedure used to trim observations for a specific sample, and (c) the procedure used to Winsorize observations.

The findings of Wilcox (2001) implied that the probability of coverage of the trimmed mean was closer to the confidence level up to 20% trimming, under conditions of nonnormality. Though this study did not address statistical power, Wilcox (1994) provided information to show that Yuen's procedure with 20% trimming was as (if not more) statistically powerful than Welch's procedure and a procedure for comparing medians. For this study, 20% trimming was selected.

The trimmed mean and the Winsorized standard deviation were used for calculating Yuen's procedure. Suppose that n-scores concern students rating the effectiveness of a drug-use prevention program: X_1, $X_2 \ldots X_n$. To do a 20% trim of the highest and lowest values in a data set, a second vector is formed that corresponds to the size of the trimmed data set. First, the data are sorted in ascending order: $X_{(1)}$, $X_{(2)} \ldots X_{(n)}$. The subscripts in parentheses represent the order of the scores. Let tau (τ) represent the number of values trimmed on either side of the data set. The scores may be listed: $X_{(1)}$, $X_{(2)} \ldots X_{(\tau+1)} \ldots X_{(n-\tau)} \ldots X_{(n)}$. The $\tau+1$ to the $n-\tau$ values represent the 20% trimmed data set. For this study, τ was the integer portion of 0.20n.

A similar procedure was followed for Winsorizing the dataset with the exception that the original size of the data was maintained. The values corresponding to indices 1, 2… τ, that is, $X_{(1)}$, $X_{(2)} \ldots X_{(\tau)}$ were set to the value indexed by $\tau+1$. Likewise, the $n-\tau+1$, $n-\tau+2 \ldots n$ values -- that is, $X_{(n-\tau+1)}$, $X_{(n-\tau+2)} \ldots X_{(n)}$ -- were set to the value indexed by $n-\tau$. In the next section, the probabilities of coverage for Yuen's and Welch's procedures are provided.

Results for Yuen's & Welch's Procedures

Yuen (1974), Algina et al. (1994), Penfield (1994), Wilcox (1994), Luh and Guo (2000), Guo and Luh (2000), and Keselman et al. (2000) studied these procedures. Table 2 delineates the sample size, heteroscedasticity, and skewness and kurtosis conditions as they related to the probabilities of coverage.

Results for Welch's procedure. The findings for these studies showed that Welch's procedure displayed probabilities of coverage ranging from 0.925 to 0.975 when sample sizes were unequal, skewness ranges from 0.00 to 2.00, and kurtosis ranges from -1.20 to 0.00. To arrive at the results, Yuen (1974) did 5,000 iterations using Welch's procedure. Algina, Oshima, and Lin (1994) and Penfield (1994), Luh and Guo (2000), Guo and Luh (2000) did 10,000 iterations.

However, probabilities of coverage less than 0.925 were associated with heteroscedasticity conditions where the sample sizes were unequal. In addition, the skewness for most of the distributions was greater than 2.00. This suggested that skewness, unequal group sizes, and heteroscedasticity inflated the Type I error rate and reduced the probability of coverage for Welch's procedure (Luh & Guo, 2000, & Guo & Luh, 2000).

Results for Yuen's procedure. Yuen's procedure appeared to do better than Welch's procedure when sample sizes were unequal, variances were unequal, and skewness was greater than 2.00 (Luh & Guo, 2000). Values of probability of coverage ranged from 0.92 to 0.95. The rate of Type I errors -- the probabilities of coverage -- were less influenced by unequal sizes, heteroscedasticity, and skewness greater than 2.00.

Yuen's procedure did well when skewness and kurtosis of the distributions differed (Wilcox, 1994b). The first group entailed samples from a standard normal distribution. The skewness conditions for the second group were 2.00 and 3.90, respectively. The respective kurtoses conditions were 6.0 and 42.2. Sample size pairs ranged from (12, 12) to (80, 20). Wilcox (1994b) did 100,000 iterations using Yuen's procedure. The probabilities of coverage ranged from 0.94 to 0.95. Unequal sample sizes and skewness had less of an influence on the probabilities of coverage.

Keselman, Wilcox, Kowalchuk, and Olejnik (2002) examined the Type I error rates of Yuen's and Welch's procedures. Keselman et al. examined one distribution with a skewness of 1.52, a kurtosis of 1.35. A second distribution was examined with a skewness of 1.63, a kurtosis of 1.00. A third distribution was examined with a skewness of 6.2, a kurtosis of 111.

Furthermore, Keselman et al. used total sample sizes ranging from 30 to 50. Keselman et al. labeled these cases small. Keselman et al. labeled total sample sizes ranging from 200 to 600 large cases. Keselman et al. decussated the levels of size and the levels of skewness and kurtosis pairs with heteroscedasticity levels ranging from (σ_1: σ_2)=(2:1) to (4:1). Keselman et al. summated the findings by size and variance conditions, given 162 conditions.

The results imply that Yuen's procedure did better than Welch's procedure. For the small size condition, Yuen's procedure displayed probabilities of coverage ranging from 0.931 to 0.950. Welch's procedure showed probabilities of coverage ranging from 0.882 to 0.946. For the large size conditions, Yuen's procedure showed probabilities of coverage ranging form 0.948 to 0.950. Welch's procedure exhibited probabilities of coverage ranging from 0.933 to 0.950.

Summary. In summary, for sample sizes as small as (n_1, n_2)=(10, 20), (σ_1:$\sigma_{2)}$=(1:2), Welch's procedure was robust. Probabilities of coverage ranged from 0.95 to 0.96, for values of skewness ranging from 0.00 to 1.50 (Penfield, 1994). Yet, when size was (n_1, n_2)=(33, 67), (σ_1:σ_2)=(3:1), probability of coverage was 0.88, for a skew value of 6.10 (Algina et al., 1994). Luh and Guo (2000) and Guo and Luh (2000) observed similar results. Under similar conditions, Yuen's procedure did better.

Table 2

Probability of Coverage of Yuen's & Welch's Procedures Reported in the Literature

Test	Citation	n1	n2	σ1	σ2	Skew.	Kurt.	PC
Welch's	Yuen(1974)	10	10	1.00	0.71	0.00	-1.20	0.95
		20	10	1.00	0.71	0.00	-1.20	0.95
		20	20	1.41	0.71	0.00	-1.20	0.95
		10	20	4.00	1.00	0.00	0.00	0.95
		10	10	2.00	1.00	0.00	0.00	0.95
	Algina (1994)	33	67	3.00	1.00	6.10	np	0.88
	et al.	33	67	2.00	1.00	6.10	np	0.90
		33	67	1.00	1.00	6.10	np	0.94
	Penfield (1994)	10	20	1.00	1.00	0.00	np	0.96
		10	20	1.00	1.00	1.50	np	0.96
		20	20	1.00	2.00	0.00	np	0.95
		20	20	1.00	2.00	1.50	np	0.95
		10	20	1.00	2.00	1.50	np	0.95

Table 2

Probability of Coverage of Yuen's & Welch's Procedures Reported in the Literature (continued)

Test	Citation	n1	n2	σ1	σ2	Skew.	Kurt.	PC
Welch's		10	20	1.00	2.00	1.00	np	0.95
		10	20	2.00	1.00	0.00	np	0.95
	Penfield (1994)	10	20	2.00	1.00	1.50	np	0.96
	Luh&Guo(2000)	12	24	1.00	4.00	6.20	111.00	0.91
		12	24	4.00	1.00	6.20	111.00	0.85
	Guo&Luh(2000)	18	12	1.00	6.00	1.75	5.90	0.92
		18	12	1.00	6.00	6.20	111.00	0.85
Yuen's	Luh&Guo(2000)	12	24	1.00	4.00	6.20	111.00	0.95
		12	24	4.00	1.00	6.20	111.00	0.92
	Wilcox(1994)	12	12	1.00	1.00	2.00	6.00	0.95
		40	12	1.00	1.00	2.00	6.00	0.95
		80	20	1.00	1.00	2.00	6.00	0.94
		12	12	1.00	1.00	3.90	42.20	0.95
		40	12	1.00	1.00	3.90	42.20	0.95
		80	20	1.00	1.00	3.90	42.20	0.95

Note. np=not provided/PC=Probability of Coverage/Skewness and Kurtosis specifications for Wilcox (1994b) represent the second group data. The first group is sampled from a standard normal distribution.

When size was $(n_1, n_2)=(12, 24)$, $(\sigma_1:\sigma_2)=(4:1)$, probability of coverage was 0.92 (Luh & Guo, 2000). This suggests that under conditions of skewness greater than 2.00, Yuen's procedure did better than Welch's procedure, for the accuracy of probability of coverage. These results were presented with reservation.

Discussion of the Monte Carlo Results for Yuen's & Welch's Procedures

One reservation regarding the results for Welch's and Yuen's procedures relates to a motivation for studying the robustness of the independent samples t test (Sawilowsky & Blair, 1992; Micceri, 1989). The authors did not obtain their results from the type of data researchers observed in practice. As a result, these findings may not generalize to the data that researchers observe in practice (Sawilowsky & Blair, 1992).

A second reservation concerns the alpha levels. Some of these studies (Keselman, et al., 2002; Luh & Guo, 2000; Guo & Luh, 2000; and Penfield, 1994) did not examine the procedures at $\alpha = 0.01$. Bradley (1978) observed that larger sample sizes were required for the t-test to exhibit robustness under skewness and heteroscedasticity at $\alpha = 0.01$ than at $\alpha = 0.05$.

Given the cost of committing a Type I error (Hinkle et al., 1998, p. 148), a researcher may select $\alpha = 0.01$ instead of $\alpha = 0.05$. Given the

conditions of heteroscedasticity, skewness, and size observed in the preceding studies, the robustness of these findings may no longer hold when $\alpha = 0.01$. Bradley (1978) observed that the t-test did not exhibit robustness assuming skewness and heteroscedasticity (σ_1:σ_2)=(4:1) until each condition included 1,024 subjects. Bradley did not specify the degrees of skewness; instead, Bradley termed these distributions L-shaped.

If a researcher selected a more stringent alpha level, such as $\alpha = 0.01$, these procedures can yield results that are not robust, where the procedures exhibited robustness at 0.05 alpha level. Rates of probability of coverage may differ from the confidence level. Higher rates of Type I error outside the bounds deemed robust and lower rates of probability of coverage may result.

A third reservation is that these results leave unanswered: How well does each procedure allow for the consideration of statistical precision (Hinkle et al., 1998)? Yet, statistical precision is important for the following reasons. (a) Confidence intervals provide an indication of statistical significance when no hypothesis has been specified (Knapp, 1999). In this respect, the question of statistical precision is important for narrow interval length.

(b) Alternatively, a utilization of the confidence interval that agrees

with the presentation of a number of applied statistics texts (Hinkle et al., 1998; Wilcox, 1996; Wilcox, 2002) is the following. Having achieved statistical significance, the researcher wants to know the statistical precision of the treatment effect.

For example, an observed effect of 3 was observed to be significant (e.g., $p<0.05$). The question to be answered takes the following form, 'as a researcher, am I 95% confident that the effect lies within the range 2.75-3.25, 2.00-5.00, or 0.05-5.95?'. The lengths of the values express the differences in the narrowness of the interval.

A Study of the Statistical Precision of the Trimmed t

One study examined the interval length of the trimmed mean. Sawilowsky (2002a) obtained samples from seven empirical distributions. Sawilowsky computed the interval length for the 20% trimmed mean 10,000 times.

Sawilowsky (2002a) compared the interval for the trimmed mean to that of a second estimate. Sawilowsky obtained these estimates for a single sample. The second estimate involved trimming. Trimming was done by whether or not a sample value was unusually large or small compared with all values in the sample (Wilcox, 1996, p. 146). This second estimate will be called the adaptive trimmed mean. The outcome

measure for the study was the median ratio of the interval length for the adaptive trimmed mean to that of the 20% trimmed mean.

In most situations, the confidence interval for the adaptive trimmed mean exhibited narrower interval lengths. Under conditions of asymmetry, $|skewness| > 1.25$, this ratio never exceeded 0.75. This suggests that the adaptive trimmed mean exhibited narrower interval lengths than the trimmed mean. Under conditions of lesser asymmetry, $|skewness| < 0.50$, this ratio was not lower than 0.890. This suggests that the interval length for the trimmed mean nearly approximated that of the adaptive trimmed mean. The finding suggests that Yuen's procedure may display wider interval lengths for skewed distributions.

Summary of the Literature Review

Summarizing, skewness resulted in probabilities of coverage that were greater than expected. Heteroscedasticity resulted in probabilities of coverage that were less than expected.

Hinkle, Wiersma, and Jurs (1998) recommended Welch's procedure for dealing with heteroscedasticity. Wilcox recommended Yuen's procedure be used with skewed distributions (Wilcox, 1996). Results of simulation studies showed that Yuen's procedure does better than Welch's procedure under conditions of skewness greater than 2.00.

However, (1) results from these studies were not derived from the types of data that researchers observed in practice. (2) Robustness of results when alpha is 0.05 does not necessarily hold when alpha is 0.01. None of the studies dealt with the problem of statistical precision. This study adds to knowledge about the probability of coverage and interval length for Yuen's and Welch's procedures using empirical distributions.

CHAPTER 3

Method

Overview of the Method

This section provided an overview of the study methodology. The first section described the applicability of the methods used in studies of Student's, Yuen's, and Welch's procedures. The second section discussed the Monte Carlo design. The specifications for sample size, heteroscedasticity, skewness and kurtosis, nominal alpha, and the number of iterations were provided. The third section discussed trimming, Winsorization, and the confidence interval for the trimmed mean. In the fourth section, the information about trimming, Winsorization, and the confidence interval provided an outline of Welch's and Yuen's procedures. The fifth section outlined the computer algorithm that solves for interval length and probability of coverage.

Applicability of Method

This study replicated, modified, and improved upon the methods of previous studies. This study replicated the sampling with replacement methodology used previously (e.g., Sawilowsky & Blair, 1992; Wilcox, 1994; Penfield, 1994; Algina, Oshima, & Lin, 1994). Whereas other studies used Monte Carlo methods to study the average rate of Type I errors or the

average rate of true rejections, it was applied here to study the average rate at which confidence limits enclosed the population parameter and the average interval length. An improvement from previous studies was that the procedures entailed repeated sampling with empirical distributions. The empirical distributions represented those observed in quantitative analysis. Another improvement from previous studies was that the study entailed a greater number of iterations. This helped ensure that the standard errors for probability of coverage and average interval length were small compared with that observed in previous studies.

There were other procedures for estimating the interval length and Type I error rate. The procedures included the bootstrap and the jackknife. The procedures applied repeated sampling to a single random sample (Sawilowsky & Fahoome, 2003) to estimate the statistical properties of a statistical technique. However, the generalizations of the results were limited to the quality of the initial sample. Consequently, these methods were not selected. The next section provides information about the Monte Carlo design.

Monte Carlo Design

The conditions included: sample size, heteroscedasticity, skewness and kurtosis, and alpha levels. These conditions were chosen to represent

the Monte Carlo Experiments included in the literature review (i.e., Algina, et al., 1994, Luh & Guo, 2000, & Guo & Luh, 2000, Sawilowsky, 2002a, Sawilowsky & Blair, 1992, Wilcox, 1994b).

Sample Size Specifications

Three sample size ratios were selected. To represent a range of potential situations, ratios of 1:1, 3:1, and 1:3 were selected. The respective sample sizes were (n_1, n_2) = (13, 13), (13, 39), (39, 13), and (39, 39). Algina et al. (1994) ((n_1, n_2)=(10,30)), Sawilowsky and Blair (1992) ((n_1, n_2)=(10,10), (15,45)), and Wilcox (1994b) ((n_1, n_2)=(12,12), (40,12)), for example, used sample size ratios of 1:1, 3:1, and 1:3.

The selected sample sizes provided an indication of the probability of coverage and interval length properties of Yuen's and Welch's procedures for small (n=13) and moderate (n=39) sample sizes. That is, knowledge of the small sample properties for Yuen's and Welch's procedures was gleaned about the probability of coverage and interval length. The use of sizes (n_1, n_2) = (39, 13), and (13, 39) allowed different pairings of unequal sample size and heteroscedasticity (for the purpose of computing the standard error of the effect): $S^2_{\min} / n_{\min}$ and $S^2_{\max} / n_{\min}$. The next section discusses the levels of heteroscedasticity.

Heteroscedasticity Specifications

The ratios of standard deviations included (σ_1:σ_2) = (1:4), (1:2), and (1:1). Sawilowsky and Blair (1992) set a 2:1 ratio (Sawilowsky & Blair used the 2:1 ratio to study the statistical power of the t-test.), and Algina, et al. (1994). After reviewing articles in the *American Educational Research Journal*, Wilcox (1987) suggested that ratios of standard deviations greater or equal to four did occur; furthermore, results from Keselman et al. (1998) observed an average ratio of 2.0 and a maximum of 23.8 – for journals in education and psychology. Variance ratios of 1:1, 1:2 and 1:4 allowed a comparison of the probability of coverage and interval lengths for each procedure under homoscedasticity and heteroscedasticity. The added effect of skewness and heteroscedasticity was learned.

Skewness & Kurtosis Specifications

Micceri (1989) took to task the tradition of Monte Carlo studies that used mathematical functions to generate random samples, termed functional samples. Sawilowsky and Blair (1992) argued that this type of sample was not representative of that observed in research practice. (Wilcox (1995, p. 109) did not believe that the empirical distributions of the Sawilowsky and Blair (1992) study were a random a sample of all existing empirical studies. By stating that these empirical distributions were not a

random sample, Wilcox implies that they are not representative of existing studies. Wilcox recommended using functional samples, to help ensure that the distributions used in Monte Carlo studies represented that observed in quantitative analysis, based on some underlying rationale or theory.) Next, the eight distributions presented by Micceri were identified.

Micceri (1986) identified eight distributions prevalent in research practice. Table 8 contains the means, standard deviations, skewness, and kurtoses for seven of the eight distributions. Estimates of interval length and probability of coverage were obtained by sampling from the seven distributions. The next section provides specifications for the levels of alpha.

Alpha Level Specifications

Probabilities of coverage and interval length were examined for the 0.01, and 0.05 alpha levels. Monte Carlo experiments entailed an alpha level of 0.05 (i.e., Algina, et al., 1994, Luh & Guo, 2000, & Guo & Luh, 2000, Sawilowsky, 2002a, Sawilowsky & Blair, 1992, Wilcox, 1994b).

Sawilowsky and Blair (1992) and Wilcox (1994b) set an alpha level of 0.01. The 0.01 alpha level was included here for comparison purposes. Comparisons were made between the probabilities of coverage and between interval lengths for alpha set at 0.01 and 0.05.

Table 3

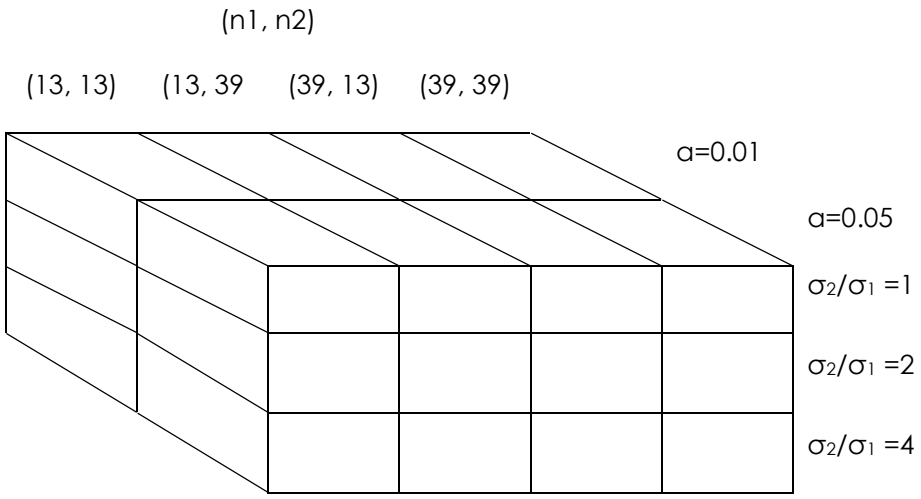
Descriptive Information Pertaining to Eight Real World Distributions

Distribution	M	SD	Skew.	Kurt.
Mass at Zero	12.92	4.42	-0.03	0.31
Extreme Asymmetry-Psychometric	13.67	5.75	1.64	1.52
Extreme Asymmetry-Achievement	24.5	5.79	-1.33	1.11
Extreme Bimodality	2.97	1.69	-0.08	-1.70
Multimodal & Lumpy	21.15	11.9	0.19	-1.20
Digit Preference	536.95	37.64	-0.07	-0.24
Smooth Symmetric	13.19	4.91	0.01	-0.34

Note. Adapted from "A More Realistic Look at the Robustness and Type II Error Properties of the t Test to Departures From Population Normality", by S. S. Sawilowsky and R. C. Blair, 1992, Psychological Bulletin, 2, p. 353. Copyright 1992 by the American Psychological Association

Comparisons were made between the probabilities of coverage for the conditions of this study to that of other studies. Moreover, qualifications concerning the robustness of a procedure were different for alpha set at 0.01 and 0.05 (Bradley, 1978). The levels of sample size, heteroscedasticity, and skewness and kurtosis were decussated. Next, the number of iterations was specified (see Figure 1).

(skewness, kurtosis)=(-1.33, 1.10)



• • •

(skewness, kurtosis)=(1.64, 1.51)

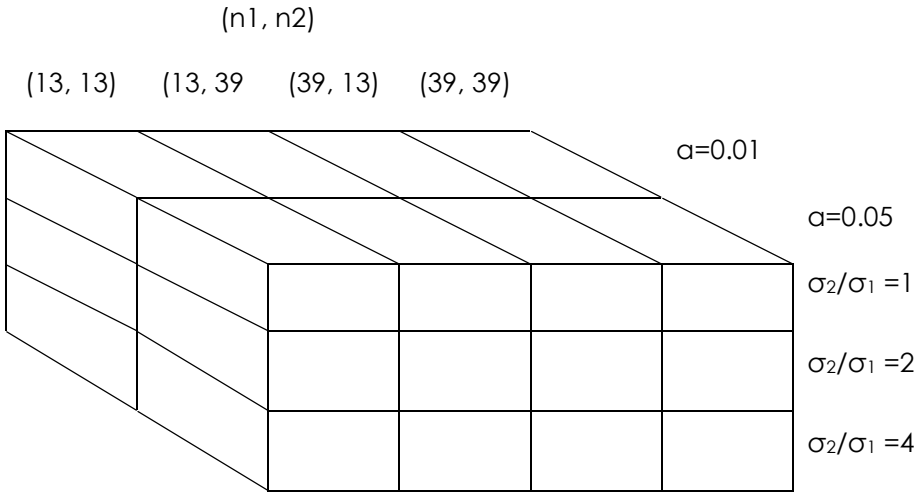


Figure 1. Monte Carlo Design & Conditions

Monte Carlo Iterations

Once designed, the Monte Carlo experiment must be iterated. The number of iterations specified was 1,000,000. Computational steps for Welch's and Yuen's procedures are discussed in the next section.

Welch's & Yuen's Procedures

Confidence Interval for the Yuen's Procedure

The study examined the interval length, $2(t_{1-\alpha/2})(SE_{\bar{x}1-\bar{x}2})$, and the probability of coverage $1-\hat{\alpha}$. Interval length and probability of coverage were examined for the confidence interval of trimmed means for two groups (Wilcox, 1996):

$$(\overline{X}_{t1}-\overline{X}_{t2}) \mp t_{1-\alpha/2} SE_{\bar{x}t1-\bar{x}t2}. \tag{5}$$

Where $\overline{X}_{ti}$ is the trimmed mean for group i.

$$\overline{X}_t = \frac{\sum_{i=\tau+1}^{n-\tau} X_{(i)}}{h_i}. \tag{6}$$

$t_{1-\alpha/2}$ is the critical value of the *t*-statistic at the alpha level (a/2).

$$SE_{\bar{x}t1-\bar{x}t2} = \sqrt{\frac{S'^2_{w1}}{h_1} + \frac{S'^2_{w2}}{h_2}}. \tag{7}$$

$$S'^2_{wi} = \frac{SSD_{wi}}{(h_i - 1)}. \tag{8}$$

$$SSD_{wi} = (\tau+1)[(X_{(\tau+1)} - \overline{X}_w)^2 + (X_{(n-\tau)} - \overline{X}_w)^2] + \sum_{i=\tau+2}^{n-\tau-1} (X_{(i)} - \overline{X}_w)^2 . \quad (9)$$

$$\overline{X}_{wi} = \frac{(\tau+1)(X_{(\tau+1)} + X_{(n-\tau)}) + \sum_{i=\tau+2}^{n-\tau-1} X_{(i)}}{n_i} . \quad (10)$$

$$h_i = n_i - 2(\tau_i) . \quad (11)$$

Where h_i is the sample size after trimming for the ith group;

n_i is the sample size for group i; and

tau (τ_i) is the integer portion of $0.20(n_i)$.

The alpha levels for the critical values, $t_{1-\alpha/2}$, were set at $\alpha/2 =$ 0.025 and 0.005, respectively. The confidence levels were $1-\alpha$ = 0.95 and 0.99, respectively. Next, Welch's procedure is explained.

Welch's Procedure

Welch's procedure does not use trimming. Equation (2) calculates the confidence interval for Welch's procedure. The difference, when computing the confidence interval for Welch's procedure, is that the degrees of freedom were computed using the following equation (Hinkle et al., 1998):

$$df_{wlch} = \frac{(s_{x1}^2 + s_{x2}^2)^2}{\frac{(s_{x1}^2)^2}{n_1 - 1} + \frac{(s_{x2}^2)^2}{n_2 - 1}} . \quad (12)$$

Yuen's Procedure

Yuen's procedure was applied with the 20% trimmed mean. Equation (5) provided calculations for the confidence interval for Yuen's procedure. The degrees of freedom were estimated using the following equation (Wilcox, 1996):

$$df_{yuen} = \frac{(\frac{S'^2_{w1}}{h_1} + \frac{S'^2_{w2}}{h_2})^2}{\frac{S'^4_{w1}}{h_1^2(h_1 - 1)} + \frac{S'^4_{w2}}{h_2^2(h_2 - 1)}}. \quad (13)$$

The next section outlines the algorithm for interval length and probability of coverage.

Description of Algorithm

This section outlined the computer algorithm having as its solution the interval length and probability of coverage. The steps include:

1. random sampling,
2. standardizing scores,
3. modeling effects and heterogeneity
4. trimming and Winsorizing the dataset,
5. computing the interval,
6. number of iterations,

7. summing values of interval length and probability of coverage, and
8. averaging values of interval length and probability of coverage.

Random Samples

Generating random samples in Monte Carlo experiments. Before presenting the method of random sampling, the issue of random sample generation was addressed. (1) Mathematical functions (Mooney, 1997) were used to generate random samples (e.g., Algina, Oshima, & Lin, 1994, or Wilcox, 1994b). (2) Alternatively, mathematical transformations of functional sample values were calculated (e.g., Algina, Oshima, & Lin, 1994, or Wilcox, 1994b, Luh & Guo, 2000, & Guo & Luh, 2000). Values from these mathematical functions conformed to known values of skewness and kurtosis.

For example, one transformation used by Luh and Guo (2000) and Guo and Luh (2000) resulted in a very high kurtosis value. The kurtosis value was 111. Guo and Luh (2000) and Luh and Guo, (2000) used g- and h- transformations to generate their samples. Using a mathematical function to generate values from a standard normal distribution, set values of g and h were entered to a mathematical transformation that yielded

values that have a certain distributional shape (Wilcox, 1994a). As such, set values of g and h resulted in set values of skewness and kurtosis.

These estimates of skewness and kurtosis were transformed into the third and fourth moment estimates using Table 1 of Wilcox (1994a, p. 297). From Table 1 of Wilcox (1994a), skew and kurtosis values of 1.75 and 8.90 were associated with g and h values of 0.50 and 0.00, respectively. Skewness and kurtosis values of 6.20 and 111.00 were associated with g and h values of 1.00 and 0.00, respectively. Next, the steps of the algorithm were outlined.

Random sampling method. *First*, the IMSL routine specified as RNUND generated pseudorandom numbers from a discrete uniform distribution. These numbers served as indices. The size of the indices was between one and the size of the sampling distribution. Sample values were set to pseudorandom values from the population, using the indices. This procedure was repeated separately for each of the independent samples.

Standardizing Scores

Second, equation (14) provided the operations for standardizing the scores for each sample.

$$Z = \frac{X_i - \mu}{\sigma} \quad (14)$$

The mean and standard deviation for the standard score were zero and one, respectively.

Standardization for Yuen's procedure. The values for Yuen's procedure were standardized with the population trimmed mean μ_t (Keselman et al., 2000; Luh & Guo, 2000) and Winsorized standard deviation (Lix & Keselman, 1999). Lix and Keselman (1999) suggested that the Winsorized variance is the appropriate estimate of the variance for the trimmed mean.

Modeling Effects & Heterogeneity

Third, equation (15) provided the operations for modeling a treatment mean of one for the second group. It provided the operations for modeling a standard deviation of two and four, respectively, for the second group (Fan, Felsovalyi, Sivo, & Keenan, 2002, p. 17).

$$X' = \mu' + \sigma' Z \quad (15)$$

Where μ' is the mean for the transformed observations;

σ' is the standard deviation for the transformed observations; and

Z is a standard score.

Trimming & Winsorizing

Fourth, the IMSL (1998) routine SVRGN sorted the data by increasing value. Next, the sorted values were trimmed. The maximum and minimum scores of the trimmed values were used to obtain the Winsorized values. The trimmed mean and the Winsorized standard deviation were for calculating Yuen's procedure.

Means & Standard Deviations

Welch's procedure. *Fifth*, the IMSL routine TWOMV provided the group means and variances. The group means were for calculating the effect $\overline{X}_1 - \overline{X}_2$. Using the group variances, equation (3) provided operations for $SE_{\bar{x}1-\bar{x}2}$.

Yuen's procedure. Equation (6) provided operations for the trimmed means. The trimmed means were used to calculate the effect $\overline{X}_{t1} - \overline{X}_{t2}$. TWOMV was used to obtain SSD_{wi}. Equation (7) provided operations for $SE_{\bar{x}t1-\bar{x}t2}$, using SSD_{wi}.

Computing the Confidence Interval

Welch's procedure. Sixth, TWOMV provided the degrees of freedom for Welch's procedure. The IMSL function TIN provided the

critical value. Equation (2) provided the confidence limits. The equation $2(t_{1-\alpha/2})(SE_{\bar{x}1-\bar{x}2})$ resulted in the interval length.

Yuen's procedure. Equation (13) resulted in the degrees of freedom for Yuen's procedure. TIN gave forth the critical value. Equation (5) provided operations for the confidence limits. The equation $2(t_{1-\alpha/2})(SE_{\bar{x}t1-\bar{x}t2})$ resulted in interval length.

Student's procedure. For comparison purposes, results for Student's procedure were calculated. The degrees of freedoms were given by the formula: $n_1 + n_2 - 2$.

Summing Values of Parameter & length

Seventh, the population parameter was compared with each of the confidence limits:

$$(\overline{X}_{t1} - \overline{X}_{t2}) - t_{1-\alpha/2} SE_{\bar{x}t1-\bar{x}t2} \leq \mu_{t1} - \mu_{t2} \leq (\overline{X}_{t1} - \overline{X}_{t2}) + t_{1-\alpha/2} SE_{\bar{x}t1-\bar{x}t2}.$$

This was done to determine whether the parameter was within the limits. If the parameter was within the limits, the sum for the probability of coverage was incremented. Otherwise, the sum was not incremented.

Replicating the Experiments & Obtaining Results

Eighth, each of these steps was performed for the levels of each variable decussated. The variables included sample size,

heteroscedasticity, nominal alpha, and each distribution. The steps were repeated 1,000,000 times. After 1,000,000 iterations, the probabilities of coverage and interval lengths were calculated. The steps were executed as part of the Monte Carlo experiment (Grier, 1986). These steps were summarized in Figure 2.

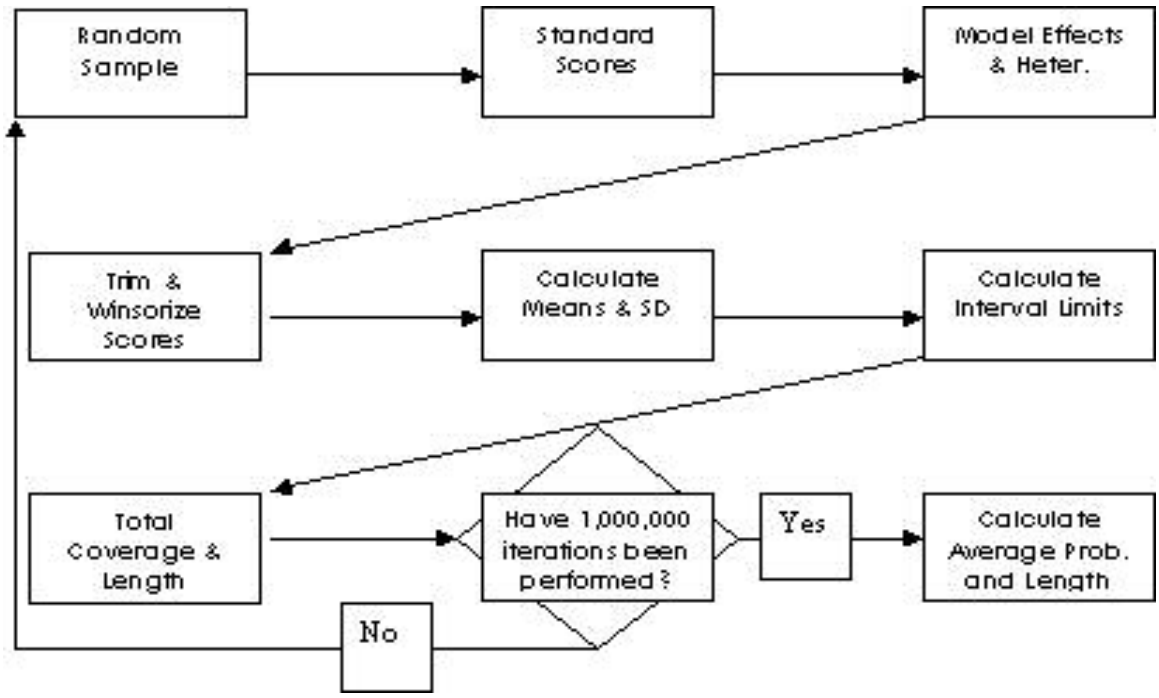


Figure 2. Summary of Algorithm

Verifying the Code

Each procedure was examined with standard normal distribution. The examination served as a test the veracity of the code. An IMSL routine was used to generate pseudo random numbers from a standard

normal distribution (RNNOR). The veracity of the code was verified using decussations of size, heteroscedasticity, and alpha conditions.

Length Ratios

For evaluating the interval length, the ratio of the average length for Student's procedure divided by the average length for the comparison procedure was calculated:

$$LR = \frac{average\_length_{Student}}{average\_length_{Comparison}}, \tag{16}$$

the comparison procedure being either Welch's or Yuen's procedure.

Sawilowsky (2002a) provided the impetus for using the ratio of lengths. The difference, between Sawilowsky's location relative efficiency and the length ratios is that Sawilowsky computed the ratio of lengths after each length was obtained from different procedures. Here, the ratio of lengths was obtained after obtaining the average interval length. The ratios were obtained for each condition.

Length ratios greater than one suggested that Welch's or Yuen's procedure yielded interval lengths that were narrower than the lengths observed using Student's procedure. Vice versa, length ratios less than one suggested that Student's procedure yielded interval lengths that were narrower.

Limitations & Delimitations

Information on the delimitations and limitations of the methodology used in this study was provided. The delimitations pertained to the methods and distributions. The limitations pertained to the skewness of the distributions and the random sampling procedure.

Delimitations

The delimitations entailed the following. First, Yuen's and Welch's procedures were examined. Second, seven empirical distributions were studied – as was done by Sawilowsky (2002a). These empirical distributions have skewness ranging from -1.33 to 1.64 and kurtosis ranging from -1.70 to 1.52. One distribution presented by Sawilowsky and Blair (1992) was omitted. Sawilowsky and Blair (1992) labeled this distribution: 'Discrete Mass at Zero with Gap'. Figure 1 of Sawilowsky and Blair (1992) depicted that more than 80% of these values were zeros. For many random samples of size 13, scores for the 20% trimmed mean were (a) constant or undifferentiated, and (b) resulted in zero variance. Third, the study assumed that the confidence interval applied to the independent samples t-test.

Limitations

The primary limitations of the study included the following: The study did not consider skewness values outside the range of -1.33 to 1.65. The study did not consider kurtosis values outside the range of -1.70 to 1.52. These results may not generalize to skewness and kurtosis values outside these ranges.

Important properties of a pseudorandom number generator are (1) the length of the sequence of numbers produced before the sequence was repeated and (2) that successive values (of the sequence) be independent (Sawilowsky & Fahoome, 2003). To the extent that the IMSL subroutine RNUND had a longer sequence (than the generators used in previous studies) and successive values were independent, the quality of this Monte Carlo study was improved.

Computer Usage

Essential Lahey Fortran 90 (Lahey Computer Systems, 1995-2000) was used to perform the analysis. The analysis was done using the IMSL F90 library. Table 4 provides information about the computer environment.

Table 4

Programs and Computing Environment

Type	Specification
Program Name:	Essential Lahey Fortran 90, Version 4.00 ®
	IMSL(R) F90 MP Library 3.0 ®
Computer Environment:	Pentium IV
	74 GB Hard Drive
	960 MB RAM

CHAPTER 4

Results

Introduction

Skewness and heteroscedasticity (a) lower the probability of coverage (May, 2003) and (b) increase the interval length. Wilcox (1996) recommended Welch's procedure when populations are heteroscedastic and Wilcox recommended Yuen's procedure when populations are skewed and exhibit heteroscedasticity. The purpose of this study was to assess the probability of coverage and the interval length for Welch's and Yuen's procedures under skewness, heteroscedasticity, and unequal sizes.

The algorithm that produced these results included the following steps. First, a random sample was obtained and the scores were standardized. Second, means and standard deviations were modeled. Third, specific to Yuen's procedure, the samples were trimmed and Winsorized. Fourth, confidence interval limits were calculated using means and standard deviations. Fifth, the interval length and probability coverage were summed. Sixth, steps one to five were repeated 1,000,000 times and the results were averaged.

The results for probability of coverage and interval length were presented. For comparison purposes, the results of Student's procedure

were presented also. The results for interval length were presented using length-ratios. The results for probabilities of coverage were presented first.

Results for Probability of Coverage

Veracity of Code

The procedures were applied to samples from a standard normal distribution to validate the code. If Student's procedure did not maintain the appropriate Type I error rate, the Type I error rate suggested incorrect code. These results supported the veracity of the code.

The probability of coverage for Student's procedure met expectations (Wilcox, 1996, p. 131). The findings are presented in Table 5. Students' procedure exhibited probabilities of coverage of 0.95, assuming normality and homoscedasticity.

Under heteroscedasticity and unequal size, probabilities of coverage for Student's procedure were outside the range from 0.925 to 0.975. Probability coverage for Student's procedure exceeded 0.975 for $(n_1 : n_2) = (13 : 39)$. Under the same conditions, the probabilities of coverage for Welch's and Yuen's procedures were 0.95, however. Probabilities of coverage were less than 0.925, $(n_1 : n_2) = (39 : 13)$, for Student's procedure. Yet, probabilities of coverage for Welch's and Yuen's procedures were between 0.945 to 0.955. The results for Student's

procedure confirmed results of previous studies (Wilcox, 1996, p. 131): that is, when variances were unequal, sizes were unequal, and variances and sizes were disproportional, the Type I error rate for the t-test will exceed the nominal level.

Table 5

Probability of coverage for Student's, Welch's, & Yuen's procedures under normality

			Alpha Level					
			0.05			0.01		
n_1	n_2	σ_2 / σ_1	Student	Welch	Yuen	Student	Welch	Yuen
13	13	1	0.950	0.951	0.950	0.990	0.990	0.990
13	13	2	0.946	0.950	0.948	0.988	0.990	0.988
13	13	4	0.946	0.950	0.948	0.988	0.990	0.988
13	39	1	0.950	0.949	0.947	0.990	0.989	0.988
13	39	2	0.991	0.950	0.950	0.999	0.990	0.990
13	39	4	0.991	0.950	0.950	0.999	0.990	0.990
39	13	1	0.950	0.949	0.947	0.990	0.989	0.988
39	13	2	0.847	0.949	0.945	0.940	0.989	0.987
39	13	4	0.847	0.949	0.945	0.940	0.989	0.987
39	39	1	0.950	0.950	0.950	0.990	0.990	0.990
39	39	2	0.949	0.950	0.949	0.989	0.990	0.989
39	39	4	0.949	0.950	0.949	0.989	0.990	0.989

The results for probabilities of coverage are provided in the next section.

Probability of Coverage

Homoscedasticity. The following results pertained to the 0.05 alpha level. Under homoscedasticity, probabilities of coverage for Student's, Welch's, and Yuen's procedures were within the range 0.925-0.975 for most of the distributions. Probabilities of coverage were presented in Table 6. Probabilities of coverage for Yuen's procedure exceeded 0.975 for the extreme asymmetric (psychometric) distribution.

Heteroscedasticity and unequal sizes. Given heteroscedasticity and unequal sizes, the probabilities of coverage for Welch's and Yuen's procedures deviated less from the nominal level than Student's procedure. Except conditions that involved extreme skewness and heteroscedasticity, probabilities of coverage were within the range 0.925-0.975.

Welch's and Yuen's procedures deviated by 3% from the probability of coverage under extreme skewness. Probabilities of coverage for Welch's procedure deviated by 3% where: (a) skewness=1.64, $(n_1 : n_2) = (39 : 13)$, $(\sigma_1 : \sigma_2) = (1 : 4)$; (b) skewness=-1.33, $(n_1 : n_2) = (13 : 13)$, $(\sigma_1 : \sigma_2) = (1 : 4)$; and (c) skewness=-1.33, $(n_1 : n_2) = (39 : 13)$, $(\sigma_1 : \sigma_2) = (1 : 2)$ or $(1 : 4)$.

Table 6

Probabilities of coverage for each procedure by sizes, standard deviations, and alpha levels

			$\alpha = 0.05$			$\alpha = 0.01$		
n_1	n_2	σ_2/σ_1	Student	Welch	Yuen	Student	Welch	Yuen
Distribution: Extreme Asymmetry - Achievement; (Skewness, Kurtosis)= (-1.33, 1.11)								
13	13	1	0.952	0.956a	0.964a	0.992a	0.994a	0.995a
13	39	1	0.952	0.937c	0.948	0.991	0.980d	0.988c
39	13	1	0.952	0.937c	0.948	0.991	0.980d	0.988c
39	39	1	0.950	0.950	0.955	0.990	0.991	0.993a
13	13	2	0.933c	0.935c	0.947	0.978d	0.979d	0.988c
13	39	2	0.986b	0.952	0.958a	0.997b	0.992a	0.994a
39	13	2	0.838d	0.922d	0.932c	0.934d	0.965d	0.974d
39	39	2	0.945	0.946	0.948	0.986c	0.986c	0.988c
13	13	4	0.914d	0.921d	0.931c	0.961d	0.965d	0.974d
13	39	4	0.992b	0.945	0.948	0.998b	0.986c	0.988c
39	13	4	0.753d	0.918d	0.929c	0.863d	0.962d	0.972d
39	39	4	0.938c	0.941c	0.943c	0.981d	0.982d	0.983d
Distribution: Extreme Bimodality; (Skewness, Kurtosis)= (-0.08, -1.70)								
13	13	1	0.962a	0.962a	0.961a	0.994a	0.994a	0.993a
13	39	1	0.959a	0.958a	0.952	0.994a	0.993a	0.988c
39	13	1	0.959a	0.958a	0.952	0.994a	0.993a	0.988c
39	39	1	0.950	0.950	0.949	0.990	0.990	0.989
13	13	2	0.958a	0.961a	0.953	0.993a	0.994a	0.989
13	39	2	0.993b	0.955	0.953	0.999b	0.991	0.990

Table 6

Probabilities of coverage for each procedure by sizes, standard deviations, and alpha levels (continued)

			$\alpha = 0.05$			$\alpha = 0.01$		
N_1	n_2	σ_2/σ_1	Student	Welch	Yuen	Student	Welch	Yuen
Distribution: Extreme Bimodality; (Skewness, Kurtosis)= (-0.08, -1.70)								
39	13	2	0.857d	0.960a	0.948	0.949d	0.994a	0.984d
39	39	2	0.948	0.950	0.947	0.989	0.989	0.987c
13	13	4	0.953	0.961a	0.949	0.991	0.994a	0.984d
13	39	4	0.997b	0.951	0.948	1.000b	0.990	0.987c
39	13	4	0.781d	0.961a	0.949	0.893d	0.994a	0.982d
39	39	4	0.946	0.949	0.946	0.988c	0.989	0.985c
Distribution: Digit Preference; (Skewness, Kurtosis)= (-0.07, -0.24)								
13	13	1	0.950	0.951	0.951	0.990	0.991	0.990
13	39	1	0.950	0.949	0.947	0.990	0.989	0.988c
39	13	1	0.950	0.949	0.947	0.990	0.989	0.988c
39	39	1	0.950	0.950	0.949	0.990	0.990	0.990
13	13	2	0.946	0.950	0.948	0.988c	0.990	0.989
13	39	2	0.991b	0.950	0.949	0.999b	0.990	0.989
39	13	2	0.846d	0.949	0.945	0.940d	0.989	0.988c
39	39	2	0.948	0.950	0.948	0.989	0.990	0.989
13	13	4	0.941c	0.949	0.945	0.985c	0.989	0.988c
13	39	4	0.998b	0.950	0.949	1.000b	0.990	0.989
39	13	4	0.765d	0.949	0.946	0.881d	0.989	0.988c
39	39	4	0.947	0.950	0.948	0.989	0.990	0.989

Table 6

Probabilities of coverage for each procedure by sizes, standard deviations, and alpha levels (continued)

			$\alpha = 0.05$			$\alpha = 0.01$		
n_1	n_2	σ_2/σ_1	Student	Welch	Yuen	Student	Welch	Yuen
Distribution: Mass at Zero; (Skewness, Kurtosis)= (-0.03, 0.31)								
13	13	1	0.950	0.951	0.951	0.990	0.991	0.990
13	39	1	0.950	0.950	0.947	0.990	0.990	0.988c
39	13	1	0.950	0.950	0.947	0.990	0.990	0.988c
39	39	1	0.950	0.950	0.950	0.990	0.990	0.990
13	13	2	0.947	0.951	0.948	0.989	0.990	0.989
13	39	2	0.991b	0.950	0.950	0.999b	0.990	0.990
39	13	2	0.847d	0.950	0.945	0.941d	0.990	0.987c
39	39	2	0.949	0.950	0.948	0.990	0.990	0.989
13	13	4	0.942c	0.950	0.945	0.986c	0.990	0.987c
13	39	4	0.998b	0.950	0.948	1.000b	0.990	0.989
39	13	4	0.765d	0.950	0.945	0.882d	0.990	0.988c
39	39	4	0.947	0.950	0.948	0.989	0.990	0.989
Distribution: Smooth Symmetric; (Skewness, Kurtosis)= (0.01, -0.34)								
13	13	1	0.950	0.950	0.950	0.990	0.990	0.990
13	39	1	0.950	0.948	0.946	0.990	0.989	0.988c
39	13	1	0.950	0.949	0.947	0.990	0.989	0.988c
39	39	1	0.950	0.950	0.950	0.990	0.990	0.990
13	13	2	0.945	0.949	0.947	0.988c	0.989	0.988c
13	39	2	0.991b	0.950	0.950	0.999b	0.990	0.990

Table 6

Probabilities of coverage for each procedure by sizes, standard deviations, and alpha levels (continued)

			$\alpha = 0.05$			$\alpha = 0.01$		
n_1	n_2	σ_2/σ_1	Student	Welch	Yuen	Student	Welch	Yuen
Distribution: Smooth Symmetric; (Skewness, Kurtosis)= (0.01, -0.34)								
39	13	2	0.846d	0.949	0.945	0.939d	0.989	0.987c
39	39	2	0.949	0.950	0.949	0.989	0.990	0.989
13	13	4	0.941c	0.949	0.946	0.985c	0.989	0.988c
13	39	4	0.998b	0.950	0.949	1.000b	0.990	0.989
39	13	4	0.765d	0.949	0.946	0.881d	0.989	0.988c
39	39	4	0.947	0.950	0.948	0.988c	0.990	0.989
Distribution: Multimodal Lumpy; (Skewness, Kurtosis)= (0.19, -1.20)								
13	13	1	0.949	0.949	0.949	0.989	0.989	0.989
13	39	1	0.950	0.947	0.939c	0.990	0.987c	0.981d
39	13	1	0.950	0.947	0.939c	0.990	0.987c	0.981d
39	39	1	0.950	0.950	0.950	0.990	0.990	0.989
13	13	2	0.944c	0.947	0.937c	0.986c	0.987c	0.980d
13	39	2	0.991b	0.950	0.948	0.999b	0.989	0.989
39	13	2	0.845d	0.947	0.930c	0.937d	0.986c	0.972d
39	39	2	0.948	0.949	0.947	0.989	0.989	0.987c
13	13	4	0.938c	0.946	0.929c	0.982d	0.986c	0.971d
13	39	4	0.997b	0.950	0.948	1.000b	0.989	0.987c
39	13	4	0.767d	0.946	0.929c	0.880d	0.986c	0.971d
39	39	4	0.947	0.949	0.946	0.988c	0.989	0.986c

Table 6

Probabilities of coverage for each procedure by sizes, standard deviations, and alpha levels (continued)

			$\alpha = 0.05$			$\alpha = 0.01$		
n_1	n_2	σ_2/σ_1	Student	Welch	Yuen	Student	Welch	Yuen
Distribution: Extreme Asymmetry - Psychometric; (Skewness, Kurtosis)= (1.64, 1.52)								
13	13	1	0.969a	0.973a	0.989b	0.996b	0.998b	0.999b
13	39	1	0.961a	0.949	0.979b	0.991	0.985c	0.998b
39	13	1	0.960a	0.948	0.979b	0.991	0.985c	0.998b
39	39	1	0.952	0.953	0.969a	0.992a	0.992a	0.997b
13	13	2	0.943c	0.946	0.982b	0.983d	0.984d	0.999b
13	39	2	0.983b	0.960a	0.976b	0.996b	0.995a	0.998b
39	13	2	0.861d	0.928c	0.958a	0.950d	0.965d	0.994a
39	39	2	0.944c	0.945	0.947	0.985c	0.985c	0.989
13	13	4	0.921d	0.925c	0.952	0.961d	0.963d	0.995a
13	39	4	0.989b	0.944c	0.944c	0.997b	0.984d	0.987c
39	13	4	0.779d	0.923d	0.940c	0.880d	0.960d	0.987c
39	39	4	0.936c	0.938c	0.927c	0.977d	0.979d	0.968d

a. $1-\hat{\alpha} > 0.955, \alpha = 0.05$ or $1-\hat{\alpha} > 0.991, \alpha = 0.01$
b. $1-\hat{\alpha} > 0.975, \alpha = 0.05$ or $1-\hat{\alpha} > 0.995, \alpha = 0.01$
c. $1-\hat{\alpha} < 0.945, \alpha = 0.05$ or $1-\hat{\alpha} < 0.989, \alpha = 0.01$
d. $1-\hat{\alpha} < 0.925, \alpha = 0.05$ or $1-\hat{\alpha} < 0.985, \alpha = 0.01$

Under these conditions, probabilities of coverage were less than 0.925. Probabilities of coverage for Yuen's procedure deviated by 3% where

skewness was 1.64. Probabilities of coverage deviated where: (a) $(n_1 : n_2) = (13 : 13)$, $(\sigma_1 : \sigma_2) = (1 : 1)$ or $(1 : 2)$; (b) $(n_1 : n_2) = (13 : 39)$, $(\sigma_1 : \sigma_2) = (1 : 1)$ or $(1 : 2)$; and (c) $(n_1 : n_2) = (39 : 13)$, $(\sigma_1 : \sigma_2) = (1 : 1)$. Probabilities of coverage exceeded 0.975.

Welch's and Yuen's procedures. Observations of the probabilities of coverage for Welch's and Yuen's procedures showed that: probabilities of coverage for Welch's procedure never exceeded 0.975; probabilities of coverage for Yuen's procedure were never less than 0.925.

The 0.01 alpha level. Similar results were observed at the 0.01 alpha level with exceptions. The following results pertained to Welch's and Yuen's procedures. Given extreme skewness (i.e., -1.33 or 1.64) and heteroscedasticity, probabilities of coverage that were within the range of 0.925-0.975 ($\alpha = 0.05$) were less than 0.985 ($\alpha = 0.01$). The following results were for Yuen's procedure. Given extreme bimodality, probabilities of coverage that were within the range of 0.925-0.975 ($\alpha = 0.05$) were less than 0.985 ($\alpha = 0.01$). Where size was inversely paired with variance, under a multimodal lumpy distribution, probabilities of coverage were within the range 0.925-0.975 at the 0.05 alpha level. At the 0.01 alpha level, probabilities of coverage were less an 0.985. The next section presents the results for interval length.

Interval Length

Sizes inversely proportional to variances. The ratio of lengths showed that the interval lengths for Welch's and Yuen's procedures were wider than the lengths for Student's procedure when sample sizes were inversely proportional to variance. The ratios of lengths were represented in Table 7. Length ratios for Welch's procedure ranged from: (a) 0.568, skewness=-0.08, $(\sigma_1 : \sigma_2) = (1:4)$, to (b) 0.694, skewness=-1.33, $(\sigma_1 : \sigma_2) = (1:2)$. Length ratios for Yuen's procedure ranged from: (a) 0.416, skewness=1.64, $(\sigma_1 : \sigma_2) = (1:4)$, to (b) 0.594, skewness=-0.08, $(\sigma_1 : \sigma_2) = (1:2)$.

Equal sizes & (homo-) heteroscedasticity; unequal sizes & homoscedasticity. Where either sample sizes were equal, given heteroscedasticity or homoscedasticity, or samples were unequal, given homoscedasticity: the ratios of lengths for Student's and Welch's procedures were around 1.000. Specifically, the ratio of lengths ranged from: (a) 0.959, skewness=0.19, $(n_1 : n_2) = (13:13)$, $(\sigma_1 : \sigma_2) = (1:4)$; to (b) 1.000, skewness=0.08, $(n_1 : n_2) = (39:39)$, $(\sigma_1 : \sigma_2) = (1:1)$. Under the same conditions, the ratio of lengths ranged from: (a) 0.692, skewness=1.64, $(n_1 : n_2) = (13:13)$, $(\sigma_1 : \sigma_2) = (1:4)$; to (b) 0.886, skewness=-0.08, $(n_1 : n_2) = (13:13)$, $(\sigma_1 : \sigma_2) = (1:1)$, for Yuen's procedure.

Table 7

Ratio of average lengths for Student's procedure Compared with that for Welch's and Yuen's procedures by sizes, standard deviations, and alpha levels

			$\alpha = 0.05$		$\alpha = 0.01$	
n_1	n_2	σ_2/σ_1	Welch	Yuen	Welch	Yuen
Distribution: Extreme Asymmetry - Achievement; (Skewness, Kurtosis): (-1.33, 1.11)						
13	13	1	0.993a	0.733a	0.989a	0.712a
13	39	1	0.975c	0.727	0.954d	0.696c
39	13	1	0.975c	0.727	0.954d	0.696c
39	39	1	0.999	0.763	0.999	0.757a
13	13	2	0.979c	0.724	0.966d	0.695c
13	39	2	1.359	1.023a	1.353a	1.005a
39	13	2	0.694d	0.514c	0.667d	0.480d
39	39	2	0.994	0.760	0.991c	0.751c
13	13	4	0.960d	0.709c	0.935d	0.670d
13	39	4	1.610	1.224	1.608c	1.211c
39	13	4	0.573d	0.424c	0.547d	0.390d
39	39	4	0.987c	0.755c	0.980d	0.742d

Table 7

Ratio of average lengths for Student's procedure Compared with that for Welch's and Yuen's procedures by sizes, standard deviations, and alpha levels (continued)

			$\alpha = 0.05$		$\alpha = 0.01$	
n_1	n_2	σ_2/σ_1	Welch	Yuen	Welch	Yuen
Distribution: Extreme Bimodality; (Skewness, Kurtosis): (-0.08, -1.70)						
13	13	1	0.999a	0.886a	0.999a	0.870a
13	39	1	0.964a	0.840	0.943a	0.809c
39	13	1	0.964a	0.840	0.944a	0.809c
39	39	1	1.000	0.827	1.000	0.822
13	13	2	0.981a	0.864	0.969a	0.833
13	39	2	1.360	1.162	1.356	1.148
39	13	2	0.682a	0.594	0.655a	0.556d
39	39	2	0.994	0.820	0.991	0.811c
13	13	4	0.959a	0.837	0.934a	0.790d
13	39	4	1.613	1.343	1.612	1.330c
39	13	4	0.568a	0.493	0.542a	0.455d
39	39	4	0.987	0.811	0.980	0.797c

Table 7

Ratio of average lengths for Student's procedure Compared with that for Welch's and Yuen's procedures by sizes, standard deviations, and alpha levels (continued)

			$\alpha = 0.05$		$\alpha = 0.01$	
N_1	n_2	σ_2/σ_1	Welch	Yuen	Welch	Yuen
Distribution: Digit Preference; (Skewness, Kurtosis): (-0.07, -0.24)						
13	13	1	0.996	0.768	0.994	0.750
13	39	1	0.971	0.749	0.950	0.718c
39	13	1	0.971	0.749	0.950	0.718c
39	39	1	1.000	0.775	0.999	0.770
13	13	2	0.980	0.753	0.968	0.725
13	39	2	1.361	1.054	1.356	1.038
39	13	2	0.688	0.527	0.662	0.492c
39	39	2	0.994	0.769	0.991	0.761
13	13	4	0.959	0.733	0.935	0.692c
13	39	4	1.612	1.247	1.610	1.234
39	13	4	0.570	0.435	0.544	0.400c
39	39	4	0.987	0.763	0.980	0.749

Table 7

Ratio of average lengths for Student's procedure Compared with that for Welch's and Yuen's procedures by sizes, standard deviations, and alpha levels (continued)

			$\alpha = 0.05$		$\alpha = 0.01$	
n_1	n_2	σ_2 / σ_1	Welch	Yuen	Welch	Yuen
Distribution: Mass at Zero; (Skewness, Kurtosis): (-0.03, 0.31)						
13	13	1	0.995	0.812	0.993	0.793
13	39	1	0.972	0.793	0.951	0.761c
39	13	1	0.972	0.794	0.951	0.761c
39	39	1	1.000	0.824	0.999	0.819
13	13	2	0.980	0.796	0.967	0.767
13	39	2	1.360	1.119	1.355	1.102
39	13	2	0.690	0.558	0.663	0.521c
39	39	2	0.994	0.819	0.991	0.809
13	13	4	0.959	0.775	0.935	0.731c
13	39	4	1.611	1.325	1.610	1.312
39	13	4	0.571	0.460	0.545	0.423c
39	39	4	0.987	0.811	0.980	0.797

Table 7

Ratio of average lengths for Student's procedure Compared with that for Welch's and Yuen's procedures by sizes, standard deviations, and alpha levels (continued)

			$\alpha = 0.05$		$\alpha = 0.01$	
n_1	n_2	σ_2/σ_1	Welch	Yuen	Welch	Yuen
Distribution: Smooth Symmetric; (Skewness, Kurtosis): (0.01, -0.34)						
13	13	1	0.996	0.735	0.994	0.718
13	39	1	0.971	0.716	0.950	0.687c
39	13	1	0.971	0.716	0.950	0.687c
39	39	1	1.000	0.742	0.999	0.738
13	13	2	0.980	0.721	0.968	0.694c
13	39	2	1.361	1.009	1.356	0.994
39	13	2	0.688	0.504	0.661	0.470c
39	39	2	0.994	0.737	0.991	0.729
13	13	4	0.959	0.702	0.935	0.662c
13	39	4	1.612	1.195	1.611	1.183
39	13	4	0.570	0.416	0.544	0.383c
39	39	4	0.987	0.731	0.980	0.718

Table 7

Ratio of average lengths for Student's procedure Compared with that for Welch's and Yuen's procedures by sizes, standard deviations, and alpha levels (continued)

			$\alpha = 0.05$		$\alpha = 0.01$	
n_1	n_2	σ_2/σ_1	Welch	Yuen	Welch	Yuen
Distribution: Multimodal Lumpy; (Skewness, Kurtosis): (0.19, -1.20)						
13	13	1	0.998	0.808	0.997	0.790
13	39	1	0.968	0.777c	0.947c	0.747d
39	13	1	0.968	0.777c	0.947c	0.747d
39	39	1	1.000	0.784	1.000	0.779
13	13	2	0.980	0.790c	0.969c	0.762d
13	39	2	1.362	1.085	1.357	1.071
39	13	2	0.685	0.548c	0.658c	0.512d
39	39	2	0.994	0.778	0.991	0.769c
13	13	4	0.959	0.768c	0.935c	0.725d
13	39	4	1.613	1.268	1.612	1.256c
39	13	4	0.569	0.454c	0.543c	0.418d
39	39	4	0.987	0.770	0.980	0.756c

Table 7

Ratio of average lengths for Student's procedure Compared with that for Welch's and Yuen's procedures by sizes, standard deviations, and alpha levels (continued)

			$\alpha = 0.05$		$\alpha = 0.01$	
n_1	n_2	σ_2 / σ_1	Welch	Yuen	Welch	Yuen
Distribution: Extreme Asymmetry - Psychometric; (Skewness, Kurtosis): (1.64, 1.52)						
13	13	1	0.992a	0.698b	0.988b	0.673b
13	39	1	0.967	0.702b	0.946c	0.669b
39	13	1	0.967	0.702b	0.945c	0.668b
39	39	1	0.999	0.772a	0.999a	0.765b
13	13	2	0.978	0.696b	0.965d	0.666b
13	39	2	1.349a	0.998b	1.343a	0.974b
39	13	2	0.691c	0.501a	0.664d	0.467a
39	39	2	0.994	0.774	0.990c	0.765
13	13	4	0.960c	0.692	0.935d	0.654a
13	39	4	1.605c	1.230c	1.604d	1.213c
39	13	4	0.573d	0.416c	0.547d	0.383c
39	39	4	0.988c	0.777c	0.980d	0.763d

Note.

a. $1-\hat{\alpha} > 0.955, \alpha = 0.05$ or $1-\hat{\alpha} > 0.991, \alpha = 0.01$

b. $1-\hat{\alpha} > 0.975, \alpha = 0.05$ or $1-\hat{\alpha} > 0.995, \alpha = 0.01$

c. $1-\hat{\alpha} < 0.945, \alpha = 0.05$ or $1-\hat{\alpha} < 0.989, \alpha = 0.01$

d. $1-\hat{\alpha} < 0.925, \alpha = 0.05$ or $1-\hat{\alpha} < 0.985, \alpha = 0.01$

Sizes proportional to variances. If sample sizes were proportional to heteroscedastic variances, the lengths for Welch's procedure were narrower than the lengths for Student's procedure. More precisely, the ratio of lengths for Welch's procedure ranged from: (a) 1.349, skewness=1.64, $(\sigma_1 : \sigma_2) = (1:2)$; to (b) 1.613, skewness=0.19, $(\sigma_1 : \sigma_2) = (1:4)$. Where unequal sizes were proportional to heteroscedastic variances, the ratio of lengths for Yuen's procedure approximated and was exceeded by the lengths for Student's procedure. The ratio of lengths ranged from: (a) 0.998, skewness=1.64, $(\sigma_1 : \sigma_2) = (1:2)$; to (b) 1.343, skewness=-0.08, $(\sigma_1 : \sigma_2) = (1:4)$.

Welch's and Yuen's procedures. Interval lengths for Yuen's procedure were wider than the lengths for Welch's procedure. The length ratios for Yuen's procedure were smaller than the ratios of Welch's procedure. In addition, lengths for Welch's procedure ranged from: (a) 0.897, skewness=-1.33, $(n_1 : n_2) = (39:39)$, $(\sigma_1 : \sigma_2) = (1:1)$; to (b) 4.908, skewness=-0.08, $(n_1 : n_2) = (13:13)$, $(\sigma_1 : \sigma_2) = (1:4)$. Lengths for Yuen's procedure ranged from: (a) 1.088, skewness=-0.08, $(n_1 : n_2) = (39:39)$, $(\sigma_1 : \sigma_2) = (1:1)$; to (b) 6.709, skewness=1.64, $(n_1 : n_2) = (13:13)$, $(\sigma_1 : \sigma_2) = (1:4)$.

Summary

Probabilities of Coverage

The results showed that under homoscedasticity, probabilities of coverage for Student's, Welch's, and Yuen's procedures were within the range 0.925-0.975 for most of the distributions. Given heteroscedasticity and unequal sizes, the probabilities of coverage for Welch's and Yuen's procedures deviated less from the nominal level than Student's procedure. Under extreme skewness (e.g., -1.33 or 1.64), probabilities of coverage were less than 0.925 for Welch's procedure; probabilities of coverage exceeded 0.975 for Yuen's procedure.

Observations of the probabilities of coverage for Welch's and Yuen's procedures showed that: probabilities of coverage for Welch's procedure never exceeded 0.975; probabilities of coverage for Yuen's procedure were never less than 0.925.

Given extreme skewness (i.e., -1.33 or 1.64) and heteroscedasticity, probabilities of coverage that were within the range of 0.925-0.975 ($\alpha = 0.05$) were less than 0.985 ($\alpha = 0.01$), for Welch's and Yuen's procedures. For Yuen's procedure, at the 0.01 alpha level, where size was inversely paired with variance, probabilities of coverage were less than

0.985 for an extreme bimodal distribution and a multimodal lumpy distribution.

Interval Length

The results for interval length showed that (a) the ratio of lengths showed that the interval lengths for Welch's and Yuen's procedures were wider than the lengths for Student's procedure when sample sizes were inversely proportional to variance. (b) Where either sample sizes were equal, given heteroscedasticity or homoscedasticity, or samples were unequal, given homoscedasticity: the interval lengths for Welch's procedure approximated the lengths for Student's procedure. (c) Under the same conditions, interval lengths for Yuen's procedure approximated that of Student's procedure to a lesser extent than Welch's procedure did. (d) If sample sizes were proportional to heteroscedastic variances, the lengths for Welch's procedure were less than the lengths for Student's procedure. (d) Under the same conditions, the ratio of lengths for Yuen's procedure approximated and was exceeded by the lengths for Student's procedure. (e) Interval lengths for Yuen's procedure were wider than the lengths for Welch's procedure.

CHAPTER 5

Discussion

Introduction

The purpose of this study was to assess the statistical precision for Welch's and Yuen's procedures. Several results have implications for applied practice. First, Welch's procedure displayed probabilities of coverage less than the confidence level when size was inversely proportional to heteroscedasticity and skewness was –1.33 or 1.64. Second, lower Type I and Type II error rates have been the basis for recommending Yuen's procedure instead of Welch's procedure (Wilcox, 1994; Luh & Guo, 2000); these results showed that unless skewness and heteroscedasticity were great in magnitude (i.e., |skewness| > 1.25 & $\sigma_2 / \sigma_1 \geq 4$), Welch's procedure displayed better statistical precision than Yuen's procedure. Third, if kurtosis of the population defined by the trimmed mean was less than –1.25, probabilities of coverage were attenuated at the 0.01 alpha level.

A discussion of the findings was presented. Findings were discussed relative to previous studies, to statistical theory, to unanticipated findings, and for practice implications.

Summary of Findings

The findings addressed how:

1. skewness appeared to inflate probability of coverage for Yuen's procedures;
2. for Welch's procedure, both skewness (i.e., absolute skewness greater than 1.25) and heteroscedasticity attenuated probabilities of coverage;
3. negative kurtosis appeared to decrease the probability of coverage for Yuen's procedures;
4. for both Yuen's and Welch's procedures, heteroscedasticity did not appear to be a major determinant of probability of coverage;
5. for Welch's and Yuen's procedures, the effect of weighting the variance by the inverse size resulted in a standard error that was greater or less than the standard error for Student's procedure;
6. interval lengths for Yuen's procedure were wider than the lengths for Welch's procedure; and
7. larger interval lengths were observed for the heteroscedastic condition than for the homoscedastic condition.

First, skewness appeared to inflate probability of coverage for Yuen's procedures. Probability of coverage was greater than the

confidence level when skewness was greater than 1.25, sample sizes were equal and less than 25 or sample sizes were unequal. These results were observed under homoscedasticity. In addition, probability of coverage was greater than the confidence level when skewness was greater than 1.25 and heteroscedasticity was proportional to size or sample sizes were equal, less than 25, and heteroscedastic. The probability of coverage exceeded the upper bound of the Bunner-Bradley criterion, i.e., $(1-\hat{\alpha}) > (1-0.5\alpha)$. These results were not observed where $\sigma_2 / \sigma_1 = 4$.

Second, for Welch's procedure, both skewness (i.e., absolute skewness greater than 1.25) and heteroscedasticity ($\sigma_2 / \sigma_1 = 4$) attenuated probabilities of coverage. That is, probabilities of coverage were less than 0.925 or 0.985. This result occurred where sample sizes were inversely proportional to variances; alternatively, where group sample sizes were less than 25.

Third, negative kurtosis appeared to decrease the probability of coverage for Yuen's procedures. For the population defined by the trimmed mean, kurtoses values less than –1.25 were observed with probabilities of coverage less than 0.985. Given extreme bimodality, probabilities of coverage that were within the range of 0.925-0.975 ($\alpha = 0.05$) were less than 0.985 ($\alpha = 0.01$). These results occurred under

both homoscedastic and heteroscedastic conditions. Where size was inversely paired with variance, under a multimodal lumpy distribution, probabilities of coverage were within the range 0.925-0.975 at the 0.05 alpha level. At the 0.01 alpha level, probabilities of coverage were less an 0.985.

Fourth, for both Yuen's and Welch's procedures, heteroscedasticity did not appear to be a major determinant of probability of coverage. When Yuen's and Welch's procedures exhibited probabilities of coverage outside the bounds of $(1-1.5\alpha)$ - $(1-0.5\alpha)$, absolute skewness was greater than 1.25 or kurtosis was less than –1.25.

Skewness and kurtosis appeared to have a greater effect on probability of coverage than did heteroscedasticity. When skewness was less than 1.25 (in absolute value) and kurtosis was greater than –1.25, probability of coverage exhibited robustness (i.e., according to the Bunner-Bradley criterion) across the conditions of heteroscedasticity, size, and alpha. Under these conditions, probability of coverage was outside the bounds of $(1-1.5\alpha)$ - $(1-0.5\alpha)$ for Student's procedure but not for either Welch's or Yuen's procedure. For $S^2_{\min}$ divided by $n_{\min}$, probability of coverage for Student's procedure exceeded $(1-0.5\alpha)$. For $S^2_{\max}$

divided by $n_{\min}$, probability of coverage for Student's procedure was less than $(1-1.5\alpha)$.

Fifth, for Welch's and Yuen's procedures, the effect of weighting the variance by the inverse size resulted in a standard error that was less than the standard error for Student's procedure, where $S^2_{\min}$ was divided by $n_{\min}$. Therefore, the lengths for Welch's and Yuen's procedures were less than the lengths for Student's procedure. If $S^2_{\max}$ was divided by $n_{\min}$, the reverse was true. For example, given a random sample from the extreme asymmetric psychometric distribution and (σ_1: σ_2) = (1:4): (a) (n_1, n_2) = (13, 39), $SE_{student}$ = 6.70, SE_{welch} = 4.09, SE_{yuen} = 4.11; or (b) (n_1, n_2) = (39, 13), $SE_{student}$ = 4.22, SE_{welch} = 6.92, SE_{yuen} = 4.51.

Sixth, Interval lengths for Yuen's procedure were wider than the lengths for Welch's procedure. The length ratios for Yuen's procedure were smaller than the ratios of Welch's procedure. This occurred because the standard error of the trimmed mean was larger than the standard error of the mean for Welch's procedure. These estimates were obtained using a smaller (trimmed) sample size (Wilcox, 2003, pp. 126-127).

Seventh, larger interval lengths were observed for the heteroscedastic than for the homoscedastic condition. This result can be explained by the increase in variance that resulted in an increase in the

standard error. The increase in the standard error resulted in an increase in the interval length.

The results from this study depicted seven findings. First, skewness appeared to inflate probability of coverage for Yuen's procedures. Second, for Welch's procedure, both skewness (i.e., absolute skewness greater than 1.25) and heteroscedasticity ($\sigma_2 / \sigma_1 = 4$) attenuated probabilities of coverage. Third, negative kurtosis appeared to decrease the probability of coverage for Yuen's procedures. Fourth, for both Yuen's and Welch's procedures, heteroscedasticity did not appear to be a major determinant of probability of coverage. Skewness and kurtosis appeared to have a greater effect on probability of coverage than did heteroscedasticity.

Interval length findings depicted the following: Fifth, for Welch's and Yuen's procedures, the effect of weighting the variance by the inverse size resulted in a standard error that was less than the standard error for Student's procedure, where $S^2_{\min}$ was divided by $n_{\min}$. If $S^2_{\max}$ was divided by $n_{\min}$, the reverse was true. Sixth, interval lengths for Yuen's procedure were wider than the lengths for Welch's procedure. Seventh, larger interval lengths were observed for the heteroscedastic condition

than for the homoscedastic condition. The next section discusses the findings.

Discussion

The topics to be discussed included: (a) relationship of results to prior research, (b) alignment with existing theories, (c) explanation of unanticipated findings, and (d) implications for research practice.

Relationship of Results to Prior Research

The discussion of the relationship of results to prior research entails: (a) the effect of skewness upon probability of coverage, (b) the effect of heteroscedasticity, and skewness upon probability of coverage, (c) the effect on probability of coverage at different alpha levels, (d) the relationship of these results on interval length to previous recommendations for Yuen's and Welch's procedures, and (e) the relationship of these results to results on statistical power.

Skewness & Type I error. Findings by Sawilowsky and Blair (1992, p. 359) showed that skewness attenuated the Type I error rates for the t-test. In a similar vein, skewness augmented probability of coverage for Yuen's procedure. The results of the present study showed that if skewness was greater than 1.25, (e.g., skewness of the extreme asymmetric - psychometric distribution was 1.417 after trimming), probabilities of

coverage were augmented (i.e., $(1-\hat{\alpha}) > (1-0.5\alpha)$). Such findings would be problematic if they effected interval length and statistical power. Yet, the results from this study showed that heteroscedasticity augmented length more than skewness. In addition, results by Wilcox (1994) and Luh and Guo (2000) showed that Yuen's procedure was more powerful than Welch's procedure under a contaminated normal distribution and skewed distributions, respectively.

Heteroscedasticity, skewness & Type I error. Type I error rates by Luh and Guo (2000), Guo and Luh (2000), and Algina et al. (1994) showed that when size was inversely proportional to heteroscedasticity and skewness was greater or equal to 2.00, Welch's procedure displayed probabilities of coverage less than the confidence level (i.e., $(1-\hat{\alpha}) < (1-0.5\alpha)$). Probabilities of coverage observed in this study showed that when size was inversely proportional to heteroscedasticity and absolute skewness was greater than 1.25, Welch's procedure displayed probabilities of coverage less than the confidence level. When sizes were proportional to standard deviations, probabilities of coverage were between 0.925-0.975. Repeating what was observed in previous studies (Algina et al., 1994; Luh & Guo, 2000), the results for this study suggested that if samples were unequal, heteroscedastic, and exhibited absolute skewness greater than

1.25, one should not apply Welch's procedure because of the Type I error inflation.

Alpha levels. The augmentation or attenuation of probability of coverage for both procedures occurred more frequently at 0.01 than at 0.05 alpha levels. This finding was consistent with results from Bradley (1978, p. 147). Bradley showed that larger sample sizes were required for the t-test to exhibit robustness at the 0.01 level than at the 0.05 level. The results of this study suggested that larger sample sizes were required for Welch's and Yuen's procedures to maintain the appropriate Type I error rate, given absolute skewness greater than 1.25 and heteroscedasticity (i.e., $\sigma_2 / \sigma_1 = 4$). One recommendation would be that of doubling or tripling the largest sizes from this study. Without re-performing the study with larger sample sizes, doubling or tripling the largest sample sizes would occur at a cost in terms of liberal or conservative estimates of Type I error or extra number of subjects. Even at (n_1, n_2) = (39, 39), probability of coverage was less than the confidence level for Welch's and Yuen's procedures (i.e., $(1-\hat{\alpha}) < (1-0.5\alpha)$).

Interval length. In previous studies, interval length was not a measure of the effectiveness of Welch's and Yuen's procedures (Wilcox, 1994; Penfield, 1994; Guo & Luh, 2000; Luh & Guo, 2000; Algina et al.,

1994). Type I and Type II error rates were measures of the effectiveness of these procedures. Yet, interval length was important to the question of statistical precision (Hinkle, et al., 1998) – that is, probability of coverage and interval length. The lower Type I and Type II error rates have been the basis for recommending Yuen's procedure instead of Welch's procedure (Wilcox, 1994; Luh & Guo, 2000). The results of this study showed that the interval lengths for Welch's procedure were narrower than the interval lengths for Yuen's procedure. One cannot discount the fact that under heteroscedasticity and skewness probabilities of coverage were lower for Welch's procedure. Yet, under negative kurtosis and lesser asymmetric deviation from normality, probabilities of coverage for Welch's procedure appeared more accurate than probabilities of coverage for Yuen's procedure at the 0.01 and 0.05 alpha levels. Here, the results showed that unless skewness and heteroscedasticity were great in magnitude (i.e., $|\text{skewness}| > 1.25$ & $\sigma_2 / \sigma_1 \geq 4$), one should apply Welch's procedure instead of Yuen's procedure.

Statistical power. Wilcox (2001, pp. 71-72) observed higher statistical power given a smaller standard error of the mean. In addition, Cohen (1994) and Knapp and Sawilowsky (2001) explained that higher statistical power should be observed between procedures, given narrower interval

lengths. Under nonnormality, previous studies (Wilcox, 1994; Luh & Guo, 2000) depicted Yuen's procedure to be more powerful than Welch's procedure. Here, the results should bring into question the mechanisms (i.e., larger effect or smaller standard error) by which Yuen's procedure was more powerful. Yuen's procedure exhibited larger interval length and standard error. Given the magnitude of interval length and standard error, it cannot be concluded that Yuen's procedure was more powerful due to narrower interval length or smaller standard error. One explanation, evidence for which was not provided here, the effect for the trimmed means was wider than the effects for the mean, in preceding studies (Wilcox, 1994; Luh & Guo, 2000).

The discussion for this section was summarized.

1. First, similar to findings by Sawilowsky and Blair (1992, p. 359) showing that skewness attenuated the Type I error rates for the t-test; the results of the present study showed that if skewness was greater than 1.25, e.g., (skewness of the extreme asymmetric - psychometric distribution was 1.417 after trimming), probabilities of coverage were augmented (i.e., $(1-\hat{\alpha}) > (1-0.5\alpha)$).
2. Second, similar to findings by Luh and Guo (2000), Guo and Luh (2000), and Algina et al. (1994) showing that when size was inversely

proportional to heteroscedasticity and skewness was greater or equal to 2.00, Welch's procedure displayed probabilities of coverage less than the confidence level when size was inversely proportional to heteroscedasticity and skewness was –1.33 or 1.64.

3. Third, the augmentation or attenuation of probability of coverage for both procedures occurred more at 0.01 than at 0.05 alpha levels; this finding was consistent with results from Bradley (1978, p. 147) showing that larger sample sizes were required for the t-test to exhibit robustness at the 0.01 level than at the 0.05 level.
4. Fourth, the lower Type I and Type II error rates have been the basis for recommending Yuen's procedure instead of Welch's procedure (Wilcox, 1994; Luh & Guo, 2000); these results showed that unless skewness and heteroscedasticity were great in magnitude (i.e., $|\text{skewness}| > 1.25$ & $\sigma_2 / \sigma_1 \geq 4$), Welch's procedure should be recommended instead of Yuen's procedure.
5. Fifth, under nonnormality, previous studies (Wilcox, 1994; Luh & Guo, 2000) depicted Yuen's procedure to be more powerful than Welch's procedure; these results should bring into question the mechanisms (i.e., larger effect or smaller standard error) by which Yuen's procedure was more powerful.

Alignment with Existing Theories

This section presented a discussion of the alignment of results with existing theory concerning the t-test. Specifically, this section discussed the results for Welch's and Yuen's procedures under skewness and heteroscedasticity compared with the results under normality and homoscedasticity (i.e., the normal curve theory of how the procedure should perform). This was done to explain the probability of coverage for Yuen's and Welch's procedure under skewness and heteroscedasticity.

Results from previous Monte Carlo studies explained deviations from nominal alpha by comparing the approximate distribution of the t-statistic for skewed populations with the distribution of the t-statistic for normal populations (Boos & Hughes-Oliver, 2000; Wilcox, 2003, pp. 118-124). The 97.5 percentile of the distribution of the t-under normality was the statistic applied; the 97.5 percentile of the t under nonnormality represented the statistic needed to maintain the appropriate Type I error rate.

For the purpose of this study, thousands of test statistics (i.e., Yuen's or Welch's procedures) were calculated using sampling with replacement. The statistics were calculated (a) using samples from a standard normal distribution; alternatively, (b) using samples from one of the empirical distributions studied here. No effect size was modeled. The

percentiles (i.e., the 97.5 percentile, $\alpha = 0.05$) for distributions of the test statistic were compared to assess how the magnitude of each would explain the probability of coverage for the normal curve statistic versus statistic obtained under non-normality that would have been necessary to maintain the appropriate Type I error rate.

The program that resulted in probability of coverage and average length was modified to provide the t-statistics. Specifications for the follow-up analysis included: (a) 16,000 iterations provided 16,000 t-statistics, (b) no effect was modeled, (c) standard deviation ratios examined included 1:1 and 1:4, (d) sample sizes included (n_1, n_2) = (13, 13), (13, 39), (39, 13), (39, 39), and (e) samples were obtained from an extreme asymmetric distribution, (skewness, kurtosis)=(1.64, 1.52).

The veracity of this procedure was tested using samples from a standard normal distribution, under homoscedasticity. That is, the 97.5 percentile of the observed t-statistic was compared with the inverse t-statistic from IMSL, TIN. Results were provided in the Table 8. The critical values and the observed values agree within 0.00 to 0.06 of one another. By increasing the Monte Carlo size from 16,000, greater precision would be achieved. Yet, the software that performed the analysis (NCSS/PASS, Hintz, 2000) accepts 16,500 cases at most. Increasing the number of

cases (i.e., test statistics calculated) would not be possible for values well greater than 16,000 (i.e., 18,000 or more).

Table 8

Critical t-statistic and percentiles of the t-distribution from a standard normal distribution

			Percentile	
n_1	n_2	$t_{1-\alpha/2}$	2.5	97.5
13	13	2.06	-2.00	2.02
13	39	2.01	-1.99	2.01
39	13		-2.01	2.00
39	39	1.99	-1.99	1.99

Note. $t_{\alpha/2} = -t_{1-\alpha/2}$

Based on both the Type I error rates for Welch's and Yuen's procedures under normality (Table 5) and the results from the approximation of the t-distribution (Table 8), it was decided that the results for Welch's procedure under skewness and heteroscedasticity would be compared with the results under normality and homoscedasticity. Under normality and homoscedasticity, probabilities of coverage for Welch's and Yuen's procedures were in the range 0.945-0.955. These results reflected the critical value for calculating the confidence interval. The

value obtained under heteroscedasticity and skewness was the value needed to maintain the appropriate Type I error rate.

Observed t-distributions were formed to assess (a) how skewness inflated probability of coverage for Yuen's procedure, and (b) how skewness and heteroscedasticity attenuated probability of coverage for Welch's procedure. First, how did skewness effect probability of coverage for Yuen's procedure?

The 97.5 percentile for Yuen's procedure under normality and homoscedasticity was 2.06, $(n_1, n_2) = (13, 13)$. Under skewness, the 97.5 percentile was 1.72; the 2.5 percentile was –1.76. The 97.5 percentile for Yuen's procedure under normality and homoscedasticity was 2.07, $(n_1, n_2) = (13, 39)$. Under skewness, the 97.5 percentile was 1.81; the 2.5 percentile was –1.83. Results similar in magnitude were observed for $(n_1, n_2) = (39, 13)$. Under homoscedasticity and normality, the critical value applied when using Yuen's procedure was greater than the critical value for Yuen's procedure given skewness. For this reason, the probabilities of coverage for Yuen's procedure were greater than the confidence level. Results were presented in Figure 3.

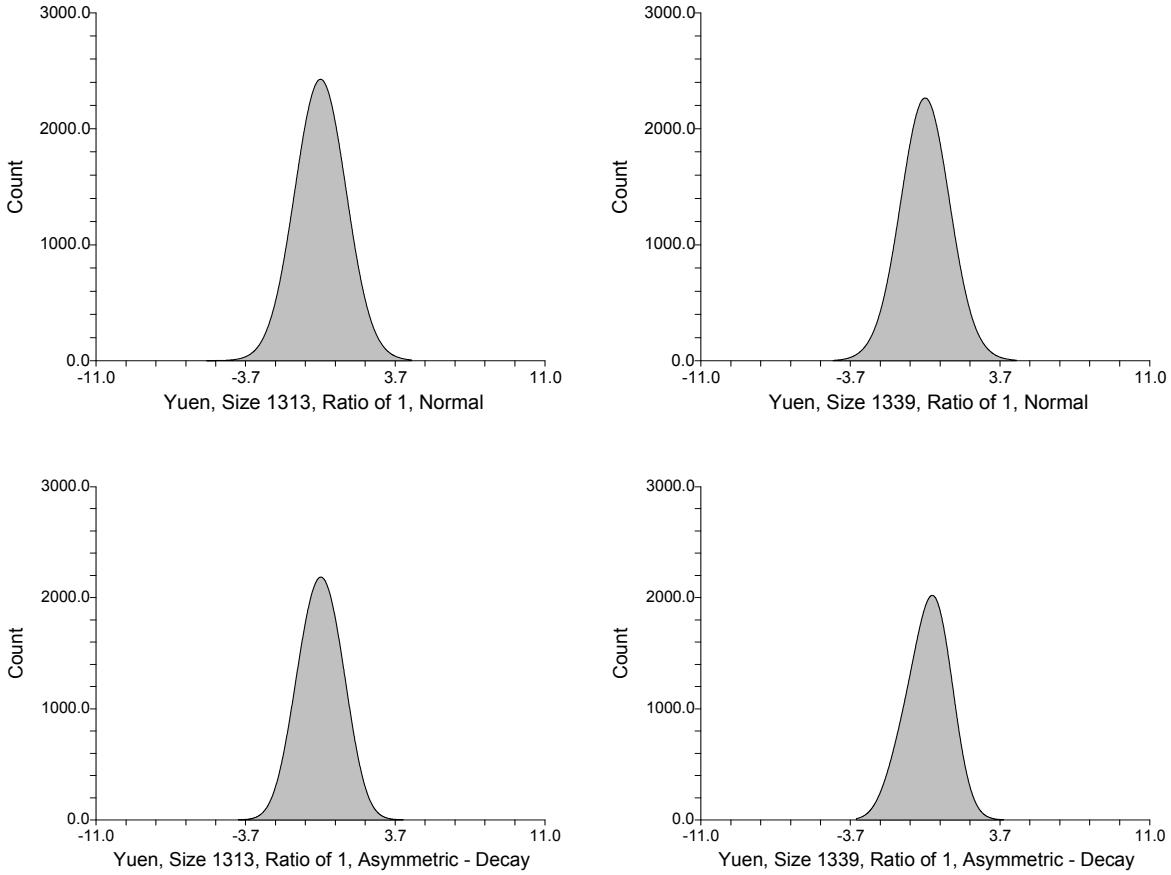


Figure 3. Distributions of Yuen's statistic under normality and nonnormality

How did skewness and heteroscedasticity attenuate probability of coverage? Assuming normality and homoscedasticity, the 97.5 percentile was 2.02, where $(n_1, n_2) = (13, 13)$. Under nonnormality and heteroscedasticity, the 97.5 percentile was 3.30; the 2.5 percentile was –1.68. Where $(n_1, n_2) = (39, 13)$, the critical value was 2.03, under normality and homoscedasticity. The 97.5 percentile was 3.65; the 2.5 percentile was –1.66, under skewness and heteroscedasticity. The critical values, under normality and homoscedasticity, were shifted left and narrower

than the percentiles obtained under heteroscedasticity and skewness. The interval limits for Welch's procedure likely missed left and were relatively narrow to accurately enclose the population parameter. Results were presented in Figure 4.

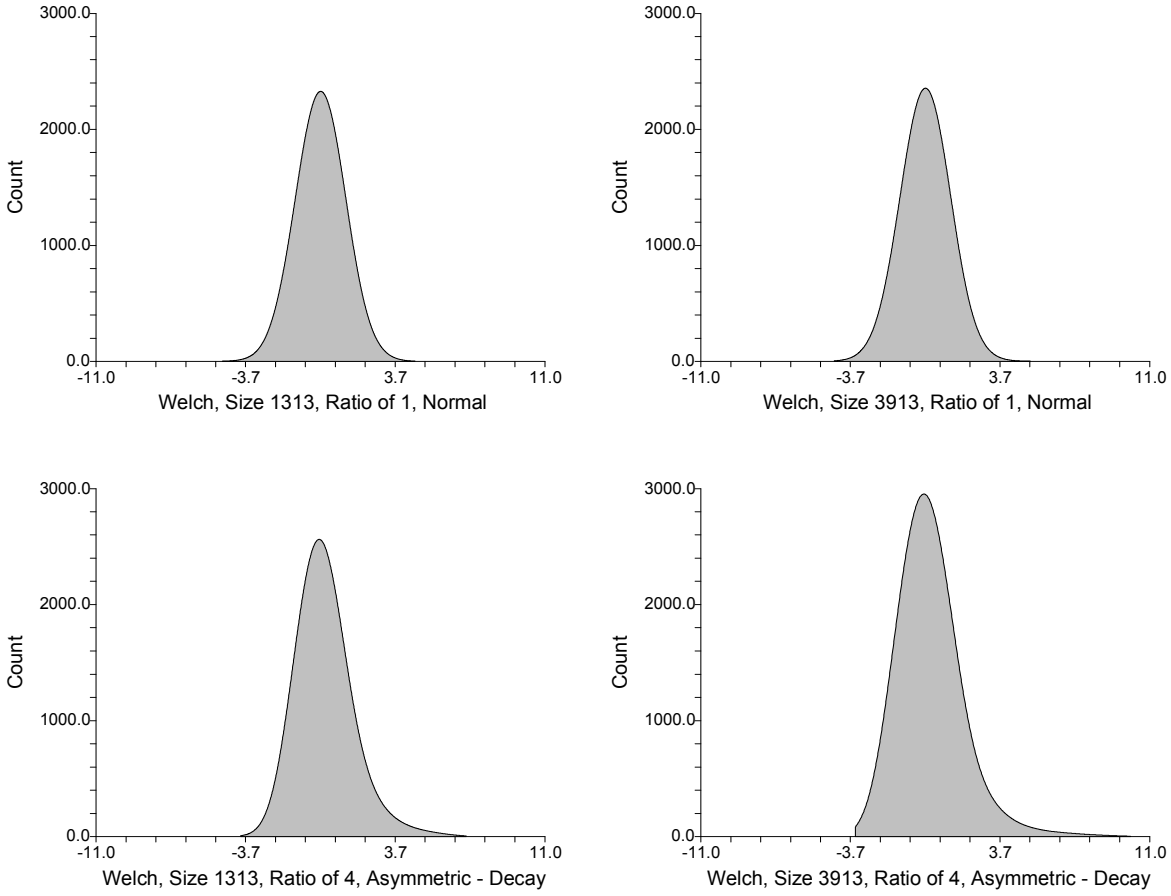


Figure 4. Distributions of Welch's statistic under normality and nonnormality

The discussion of the alignment of results with existing theory concerning the t-test was summarized. First, under homoscedasticity and skewness, the critical value applied when using Yuen's procedure was

greater than the critical value for Yuen's procedure given skewness. For this reason, the probabilities of coverage for Yuen's procedure were greater than the confidence level. Second, the critical values, under normality and homoscedasticity, were shifted left and narrower than the percentiles obtained under heteroscedasticity and skewness, for Welch's procedure. The interval limits likely missed left and were too narrow to accurately enclose the population parameter.

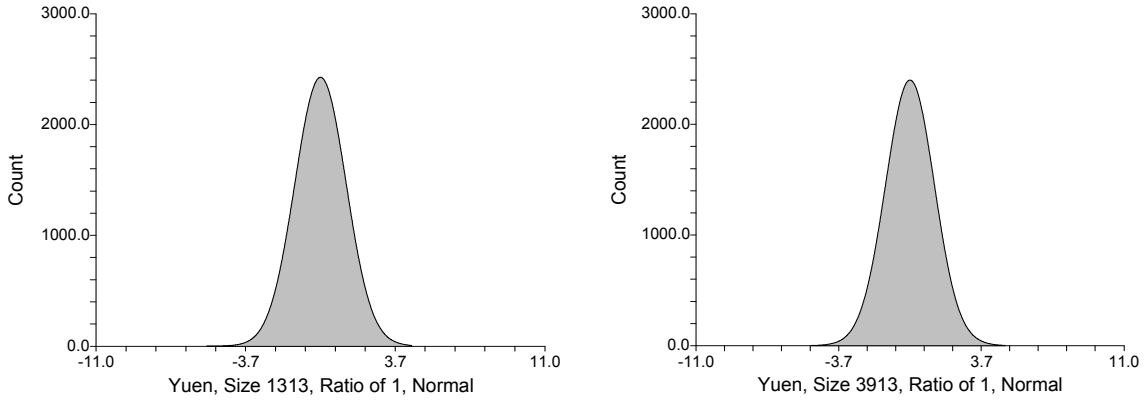
Explanation of Unanticipated Findings

The unanticipated findings to be discussed addressed how and why kurtoses effected Type I error rates for Yuen's procedure.

Kurtoses and Type I error. Not many studies have linked kurtoses and Type I error rates (e.g., one exception Glass et al., 1972). The major distributional description for Monte Carlo studies involving Welch's and Yuen's procedures was skewness (Penfield, 1994; Guo & Luh, 2000; Algina et al., 1994). Kurtosis would not be important for this study had the alpha level been restricted to 0.05. The results of this study showed that if kurtosis of the population defined by the trimmed mean was less than –1.25 (e.g. the kurtoses of the extreme bimodal & multimodal lumpy distributions were -1.45 & -1.27, respectively); probabilities of coverage were attenuated at the 0.01 alpha level (i.e., $(1-\hat{\alpha}) < (1-0.5\alpha)$). This implies that

if kurtosis of the population defined by the trimmed mean was less than –1.25, the Type I error rate exceeded Bradley's liberal criterion.

How did kurtosis attenuate probability of coverage for Yuen's procedure? Assuming normality and homoscedasticity, the 99.5 percentile for Yuen's procedure was 2.80, $(n_1, n_2) = (13, 13)$. Under bimodality and heteroscedasticity, the 99.5 percentile was 3.11; the 0.5 percentile was –3.69. Assuming normality and homoscedasticity, the 99.5 percentile for Yuen's procedure was 2.88, $(n_1, n_2) = (39, 13)$. Under bimodality and heteroscedasticity, the 99.5 percentile was 3.47; the 0.5 percentile was –3.81. The critical value for calculating Yuen's procedure was small (in absolute value) compared with the values needed to maintain the appropriate Type I error rate. The interval limits were narrower than the limits necessary to maintain the appropriate Type I error rate. Results were presented in the figure below.



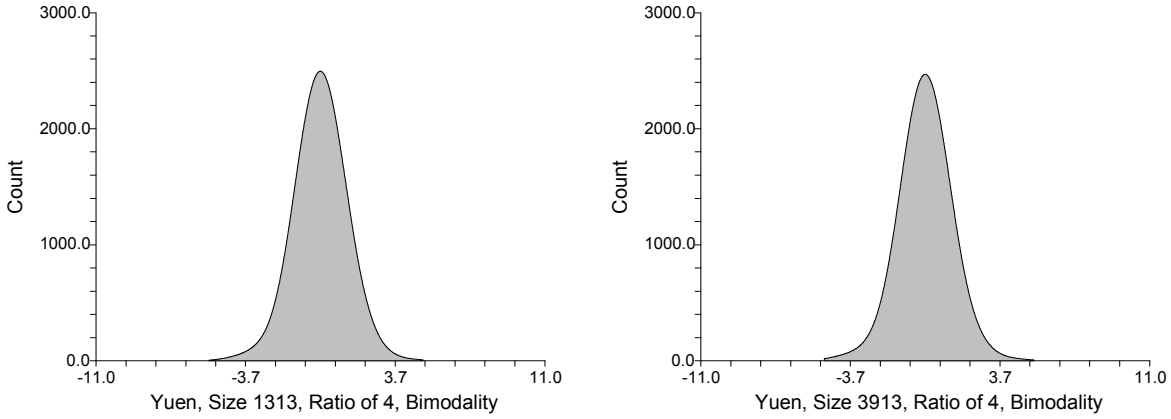


Figure 5. Distributions of Yuen's statistic under normality and nonnormality

The discussion of how and why kurtoses effected Type I error rates for Yuen's procedure was summarized. The results of this study showed that if kurtosis of the population defined by the trimmed mean was less than –1.25 (e.g. the kurtoses of the extreme bimodal & multimodal lumpy distributions were -1.45 & -1.27, respectively); probabilities of coverage were attenuated at the 0.01 alpha level. The critical value for calculating Yuen's procedure was small (in absolute value) compared with the values needed to maintain the appropriate Type I error rate. The interval limits were narrower than the limits necessary to maintain the appropriate Type I error rate. The next section presents a discussion of the implications of the results for applied practice

Implications for Research Practice

This section discusses what these results suggested about applying Welch's and Yuen's procedures, and statistical practice given results for Welch's/Yuen's procedures.

Applying Welch's and Yuen's procedures. Where the results of this study intersect statistical data analysis, a researcher intends to compare the means and assess the statistical precision of the mean difference. In this respect, implications for practice should be made so that when a procedure is applied with the intent of maintaining an appropriate confidence level, it does that with minimal deviation from the nominal level and narrow length. Recommendations for using either Yuen's or Welch's procedures would have to be specific to the bracketing of skewness, kurtosis, group sizes, and heteroscedasticity. The findings from this study suggested:

1. Given sizes disproportional to heteroscedastic variances and absolute skewness greater than 1.25, Welch's procedure should not be used. Yuen's procedure would exhibit coverage that is more accurate though interval length would be wider.

2. If kurtosis was less than –1.25 and alpha level was set at 0.01, Yuen's procedure should not be used. Welch's procedure would provide better statistical precision – accurate coverage and narrow length.
3. If these procedures were used at the 0.01 alpha level and populations exhibited absolute skewness greater than 1.25, larger samples would be required to maintain an acceptable probability of coverage (i.e., according to the Bunner-Bradley criterion). This recommendation was based on the central limit theorem (Wilcox, 1996, p. 85). Specifications regarding the magnitude of size cannot be specified given the results from this study. Yet, it was deemed necessary to be greater than 39 per group.
4. If absolute skewness was less than 1.25, and the ratio of standard deviations were less than 4, one should apply Welch's procedure. Here, it displayed accurate probabilities of coverage and narrower interval lengths compared with Yuen's procedure.

The next section discusses present statistical practices.

Statistical practice and Welch's/Yuen's procedures. Yuen's or Welch's procedures should be used when assumptions of normality and homoscedasticity are violated but it does not appear to be the case. The literature (i.e., articles published in educational journals) suggested that traditional methods (i.e., t-test, ANOVA) were used instead of

heteroscedastic methods (i.e., Welch's or Yuen's procedures). Reviewing between subjects univariate tests, Keselman et al. (1998) found that less than 20% of 61 studies considered assumptions of normality and homoscedasticity. In addition, Keselman observed that that researchers did not adopt procedures recommended in the statistical literature (Keselman et al., 1998). Statistical procedures have been studied and modified (e.g., Algina, et al., 1994; Penfield, 1994, & Luh & Guo, 2001, Guo & Luh, 2001). Applied statisticians assessed these procedures, given the data violated the assumptions of the procedure. For example, results from this study showed that when $S^2_{\min}$ was divided by $n_{\min}$, this situation attenuated the Type I error rate for Student's procedure. When $S^2_{\max}$ was divided by $n_{\min}$, this situation augmented the Type I error rate for Student's procedure.

Statistical procedures such as Student's, Welch's, and Yuen's procedures were developed on assumptions that when violated, the results from these procedures can be misleading. For example, Welch's procedure assumes that samples are randomly and independently selected from a normal population, although the assumption of homoscedasticity was untenable (Wilcox, 1996).

When data meets the assumptions, a researcher can validly interpret the significance probabilities for Welch's procedure. When data does not meet the assumptions, significance probabilities (p values) deviate from their exact values. For example, results from this study showed that under normality and heteroscedasticity, the Type I error rate for Welch's procedure was 0.05 (Table 5). When skew of the distribution was –1.33, the Type I error rate was 0.08 (Table 6). The results were more likely to be deemed significant when in fact they were not. Whereas other studies verified the robustness of Student's procedure (Sawilowsky and Blair, 1992) procedure for Type I error rates, this study informed statistical practice by showing how under nonnormality and heteroscedasticity, Yuen's and Welch's procedures maintain probability of coverage and exhibit narrow interval length.

The situation specific application of Welch's, Yuen's, and other procedures should be recommended. For example, when considering statistical precision:

1. If group samples are equal, greater than 30, and the test is two-tailed, Student's procedure should be recommended (Sawilowsky & Blair, 1992; Sawilowsky & Fahoome, 2003).

2. If heteroscedasticity and unequal size were concerns, Yuen's or Welch's procedures should be recommended based on the results of this study.

The discussion of what these results suggested about the statistical practice was summarized. It does not appear that researchers adopted procedures (Keselman et al., 1998). Statistical procedures such as Student's, Welch's, and Yuen's procedures were developed on a set of assumptions. When data meet the assumptions, significance probabilities for each procedure can be validly interpreted. When these assumptions are not met, results from this study showed that significance probabilities (p values) deviate from their exact values. The situation specific application of Welch's, Yuen's, and other procedures should be recommended.

Recommendations for Research

In this study, it was explained how Type I and Type II error rates were evidence for recommending Yuen's procedure over Welch's procedure. Interval lengths for Welch's procedure were narrower than interval lengths for Yuen's procedure. From comments by Cohen (1995) and Wilcox (2001), one would conclude that because Yuen's procedure was more powerful it would have a smaller standard error and narrower interval

length. Therefore, when considering a procedure, individuals should account for interval length (or standard error) in addition to Type I (probability coverage) and Type II (statistical power) errors. Such evidence would serve well when discerning a procedure for computing a confidence interval or a hypothesis test.

APPENDIX

Table 9

Central Tendency, Variation, and Normality for the Eight Distributions

	Mean	SD	Skewness	Kurtosis
Distributions	Results for the Seven Distributions			
Smooth Symmetric	13.186	4.906	0.005	-0.337
Extreme Bimodality	2.971	1.687	-0.077	-1.697
Extreme Asymmetric - Achievement	24.497	5.788	-1.328	1.099
Multimodal Lumpy	21.148	11.917	0.193	-1.207
Mass at Zero	12.919	4.415	-0.034	0.306
Extreme Asymmetric - Psychometric	13.667	5.754	1.636	1.511
Digit Preference	536.947	37.644	-0.065	-0.244
	Results After Trimming			
Smooth Symmetric	13.173	2.310	-0.025	-1.051
Extreme Bimodality	2.960	1.428	-0.219	-1.454
Extreme Asymmetric - Achievement	26.021	2.498	-0.627	-0.663
Multimodal Lumpy	20.473	7.586	0.427	-1.269
Mass at Zero	12.834	1.985	0.153	-0.984

Table 9(continued)

Central Tendency, Variation, and Normality for the Eight Distributions

	Mean	SD	Skewness	Kurtosis
	Results After Trimming			
Extreme Asymmetric - Psychometric	11.445	2.000	1.417	1.017
Digit Preference	537.363	17.494	0.000	-0.977
	Results After Winsorization			
Smooth Symmetric	13.104	3.100	-0.051	-1.526
Extreme Bimodality	2.976	1.681	-0.069	-1.711
Extreme Asymmetric - Achievement	25.413	3.521	-0.490	-1.351
Multimodal Lumpy	21.281	9.399	0.267	-1.634
Mass at Zero	12.900	2.960	0.083	-1.421
Extreme Asymmetric - Psychometric	12.466	3.219	0.890	-0.940
Digit Preference	537.418	24.616	0.005	-1.476

Table 10

Interval lengths for Student's, Welch's, and Yuen's Methods

			Alpha Level					
			a=0.05			a=0.01		
Skew.	Sizes	SD_2/SD_1	Stud.	Welch	Yuen	Stud.	Welch	Yuen
-1.33	1313	1	1.592	1.603	2.171	2.157	2.181	3.028
-1.33	1313	2	2.502	2.556	3.458	3.391	3.509	4.878
-1.33	1313	4	4.576	4.769	6.450	6.202	6.631	9.259
-1.33	1339	1	1.276	1.308	1.754	1.701	1.782	2.443
-1.33	1339	2	2.307	1.698	2.255	3.076	2.273	3.060
-1.33	1339	4	4.482	2.785	3.661	5.976	3.716	4.935
-1.33	3913	1	1.276	1.308	1.755	1.701	1.783	2.444
-1.33	3913	2	1.666	2.401	3.240	2.221	3.329	4.629
-1.33	3913	4	2.694	4.699	6.361	3.592	6.565	9.204
-1.33	3939	1	0.897	0.897	1.176	1.190	1.191	1.571
-1.33	3939	2	1.415	1.424	1.863	1.877	1.895	2.500
-1.33	3939	4	2.604	2.637	3.451	3.454	3.524	4.658
-0.08	1313	1	1.618	1.619	1.825	2.192	2.196	2.521
-0.08	1313	2	2.556	2.606	2.958	3.464	3.574	4.156
-0.08	1313	4	4.710	4.908	5.628	6.383	6.831	8.082
-0.08	1339	1	1.283	1.331	1.528	1.711	1.814	2.114
-0.08	1339	2	2.322	1.708	1.999	3.097	2.283	2.696

Table 10

Interval lengths for Student's, Welch's, and Yuen's Methods (continued)

			Alpha Level					
			a=0.05			a=0.01		
Skew.	Sizes	SD_2/SD_1	Stud.	Welch	Yuen	Stud.	Welch	Yuen
-0.08	1339	4	4.515	2.799	3.361	6.019	3.734	4.526
-0.08	3913	1	1.283	1.331	1.528	1.711	1.813	2.114
-0.08	3913	2	1.684	2.469	2.833	2.245	3.427	4.042
-0.08	3913	4	2.753	4.851	5.582	3.671	6.776	8.074
-0.08	3939	1	0.899	0.900	1.088	1.194	1.194	1.452
-0.08	3939	2	1.422	1.430	1.734	1.886	1.904	2.327
-0.08	3939	4	2.621	2.655	3.230	3.477	3.548	4.362
-0.07	1313	1	1.604	1.610	2.089	2.173	2.187	2.899
-0.07	1313	2	2.527	2.579	3.356	3.424	3.538	4.723
-0.07	1313	4	4.642	4.839	6.333	6.291	6.730	9.092
-0.07	1339	1	1.281	1.319	1.711	1.707	1.797	2.377
-0.07	1339	2	2.318	1.703	2.200	3.090	2.279	2.977
-0.07	1339	4	4.504	2.794	3.611	6.005	3.729	4.865
-0.07	3913	1	1.281	1.319	1.711	1.707	1.797	2.378
-0.07	3913	2	1.675	2.434	3.180	2.234	3.376	4.544
-0.07	3913	4	2.723	4.774	6.265	3.631	6.672	9.066
-0.07	3939	1	0.899	0.900	1.161	1.193	1.194	1.550

Table 10

Interval lengths for Student's, Welch's, and Yuen's Methods (continued)

			Alpha Level					
			a=0.05			a=0.01		
Skew.	Sizes	SD_2/SD_1	Stud.	Welch	Yuen	Stud.	Welch	Yuen
-0.07	3939	2	1.420	1.429	1.846	1.884	1.902	2.477
-0.07	3939	4	2.616	2.650	3.430	3.471	3.541	4.632
-0.03	1313	1	1.598	1.605	1.968	2.165	2.181	2.731
-0.03	1313	2	2.516	2.568	3.160	3.410	3.525	4.447
-0.03	1313	4	4.617	4.812	5.960	6.257	6.692	8.558
-0.03	1339	1	1.277	1.314	1.610	1.703	1.790	2.238
-0.03	1339	2	2.311	1.699	2.066	3.081	2.274	2.797
-0.03	1339	4	4.491	2.787	3.388	5.987	3.719	4.564
-0.03	3913	1	1.277	1.313	1.609	1.703	1.790	2.237
-0.03	3913	2	1.670	2.421	2.993	2.227	3.358	4.277
-0.03	3913	4	2.712	4.748	5.900	3.615	6.635	8.539
-0.03	3939	1	0.898	0.898	1.089	1.191	1.192	1.454
-0.03	3939	2	1.417	1.425	1.731	1.880	1.898	2.323
-0.03	3939	4	2.609	2.642	3.216	3.461	3.531	4.342
0.01	1313	1	1.604	1.609	2.183	2.173	2.186	3.029
0.01	1313	2	2.526	2.578	3.506	3.423	3.537	4.934
0.01	1313	4	4.642	4.839	6.616	6.291	6.729	9.501

Table 10

Interval lengths for Student's, Welch's, and Yuen's Methods (continued)

			Alpha Level					
			a=0.05			a=0.01		
Skew.	Sizes	SD_2/SD_1	Stud.	Welch	Yuen	Stud.	Welch	Yuen
0.01	1339	1	1.280	1.319	1.787	1.707	1.797	2.484
0.01	1339	2	2.317	1.702	2.296	3.089	2.278	3.106
0.01	1339	4	4.502	2.793	3.767	6.002	3.727	5.073
0.01	3913	1	1.280	1.319	1.787	1.706	1.797	2.484
0.01	3913	2	1.675	2.434	3.324	2.233	3.376	4.750
0.01	3913	4	2.723	4.775	6.548	3.631	6.673	9.478
0.01	3939	1	0.899	0.899	1.211	1.192	1.193	1.616
0.01	3939	2	1.420	1.428	1.925	1.883	1.901	2.583
0.01	3939	4	2.615	2.648	3.579	3.469	3.540	4.833
0.19	1313	1	1.612	1.615	1.996	2.184	2.192	2.763
0.19	1313	2	2.545	2.595	3.219	3.448	3.560	4.527
0.19	1313	4	4.681	4.881	6.095	6.344	6.788	8.753
0.19	1339	1	1.284	1.326	1.652	1.712	1.807	2.291
0.19	1339	2	2.325	1.707	2.143	3.100	2.284	2.895
0.19	1339	4	4.519	2.801	3.562	6.024	3.737	4.798
0.19	3913	1	1.284	1.326	1.652	1.712	1.807	2.291
0.19	3913	2	1.682	2.456	3.067	2.243	3.406	4.379

Table 10

Interval lengths for Student's, Welch's, and Yuen's Methods (continued)

			Alpha Level					
			a=0.05			a=0.01		
Skew.	Sizes	SD_2/SD_1	Stud.	Welch	Yuen	Stud.	Welch	Yuen
0.19	3913	4	2.742	4.818	6.038	3.655	6.735	8.737
0.19	3939	1	0.901	0.901	1.150	1.196	1.196	1.534
0.19	3939	2	1.424	1.432	1.831	1.889	1.907	2.457
0.19	3939	4	2.625	2.658	3.409	3.482	3.552	4.603
1.64	1313	1	1.619	1.631	2.321	2.194	2.221	3.259
1.64	1313	2	2.542	2.598	3.655	3.445	3.569	5.175
1.64	1313	4	4.645	4.841	6.709	6.295	6.730	9.632
1.64	1339	1	1.282	1.325	1.825	1.709	1.807	2.554
1.64	1339	2	2.311	1.712	2.317	3.081	2.294	3.164
1.64	1339	4	4.484	2.793	3.646	5.978	3.728	4.927
1.64	3913	1	1.281	1.326	1.826	1.708	1.808	2.556
1.64	3913	2	1.683	2.435	3.357	2.243	3.377	4.804
1.64	3913	4	2.730	4.765	6.568	3.640	6.659	9.502
1.64	3939	1	0.898	0.899	1.163	1.191	1.193	1.557
1.64	3939	2	1.416	1.425	1.829	1.879	1.897	2.457
1.64	3939	4	2.605	2.638	3.354	3.456	3.526	4.527

REFERENCES

Algina, J., Oshima, T. C., & Lin, W-Y. (1994). Type I error rates for Welch's tests and James's second-order test under nonnormality and inequality of variance when there are two groups. *Journal of Educational and Behavioral Statistics*. 19, 275-291.

Barnett, V. (1978). The study of outliers: Purpose and model. *Applied Statistics*, 27, 242-250.

Blair, R. C. (1981). A reaction to "Consequences of failure to meet assumptions underlying the fixed effects analysis of variance and covariance." *Review of Educational Research*, 51, 499-507.

Boneau, C. A. (1960). The effects of violations of assumptions underlying the *t* test. *Psychological Bulletin*, 57, 49-64.

Boos, D. D. & Hughes-Oliver, J. M. (2000). How large does n have to be for z and t intervals? *The American Statistician*, 54, 121-128.

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

Bradley, J. V. (1984). Antinonrobustness: A case study in the sociology of science. *Bulletin of the Psychonomic Society,* 22, 463-466.

Bunner, J. M. (2003). Forming a bracketed interval around the trimmed mean: Alternatives to S_w (Doctoral Dissertation, Wayne State University, 2003). *Dissertation Abstracts International,* 64, 6147.

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

Cohen, J. (1994). The earth is round (p<.05). *American Psychologist*, 49, 997-1003.

Cook, T. D. & Campbell, D. T. (1979). *Quasi-experimentation: Design & analysis for field settings*. Boston, MA: Hougton Miffilin Company.

Glass, G. V., Peckham, P. D., & Sanders, J. R. (1972). Consequences of failure to meet assumptions underlying fixed effects analysis of variance and covariance. *Review of Educational Research*, 42, 237-288.

Grier, D. A. (1986). A system for monte carlo experimentation. In J. Wilson & J. Henriksen (Eds.), *Proceedings of the 1986 winter simulation conferences* (pp. 876-888). Piscataway, NJ: IEEE

Guenther, W. C. (1965). *Concepts of statistical inference*. New York: McGraw-Hill Book Company

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

Hintz, J. (2001). NCSS and PASS [Computer software]. Kaysville, UT: Number Cruncher Statistical Systems, Inc.

Hinkle, D. E., Wiersma, W., & Jurs, S. G. (1998). *Applied statistics for the behavioral sciences.* Boston, MA: Hougton Mifflin Company.

Hoenig, J. M. & Heisey, D. M. (2001). The abuse of power: The pervasive fallacy of power calculations for data analysis. *American Statistician*, 55, 19-23.

IMSL (1998). IMSL(R) F90 MP library 3.0. [Computer software]. Incline Village, NV: Lahey Computer Systems, Inc.

Kirk, R. E. (1995). *Experimental design: Procedures for the behavioral sciences.* Pacific Grove, CA: Brooks/Cole.

Keselman, H. J., Wilcox, R. R., Kowalchuk, R. K. & Olejnik, S. (2002). Comparing trimmed or least squares means of two independent skewed populations. *Biometrical Journal*, 4, 478-489.

Keselman, H. C., Huberty, C. J., Lix, L. M., Olejnik, S., Cribbie, R. A., Donahue, B., Kowalchuk, R. K., Lowman, L. L., Petoskey, M. D., Keselman, J. C., & Levin, J. R., (1998). Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA, and ANCOVA analyses. *Review of Educational Research*, 68, 350-386.

Knapp, T. R. & Sawilowsky, S. S. (2001) Constructive criticism of methodological and editorial practices. *Journal of Experimental Education*, 70, 65-79.

Kuehl, R. O. (1994). *Statistical principles of research design and analysis*. Belmont, CA: Duxbury Press.

Lahey Computer Systems. (1995-2000). Essential lahey fortran 90 [Computer software]. Incline Village, NV: Lahey Computer Systems, Inc.

Luh, W-M. & Guo, J-H. (2000). Johnson's transformation two-sample trimmed *t* and its bootstrap method for heterogeneity and nonnormality. *Journal of Applied Statistics*, 27, 965-973.

May, K. (2003). A note on the use of confidence intervals. *Understanding Statistics*, 2, 133-135.

Micceri, T. (1986). A futile search for that statistical chimera of normality. Paper presented at the 31st Annual Convention of the Florida Educational Research Association, Tampa.

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

Mooney, C. Z. (1997). *Monte carlo simulation*. Thousand Oaks, CA: Sage Publications, Inc.

Sawilowsky, S. S. (2002a). A measure of relative efficiency for location of a single sample. *Journal of Modern & Applied Statistical Methods*, 1, 52-60.

Sawilowsky, S. S. (2002b). Fermat, schubert, einstein and berhens-fisher: The probable difference between two means when $\sigma_1^2 \neq \sigma_2^2$ *Journal of Modern & Applied Statistical Methods*, 1, 461-472.

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

Sawilowsky, S. S. & Fahoome, G. (2003). Statistics through Monte Carlo experimentation with fortran. Oak Park, MI: JMASM, Inc.

Sheskin, D. J. (2000). *Handbook of parametric and nonparametric statistical procedures.* (2nd ed.) Boca Raton, FL: Chapman & Hall

Stiedl, R. J., Hayes, J. P. & Schauber, E. (1997). Statistical power analysis in wildlife research. *Journal of Wildlife Management*, 61, 270-279.

Wilcox, R. R. (1987). New designs in analysis of variance. *Annual Review of Psychology,* 38, 29-60.

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

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

Wilcox, R. R. (1994b). Some results on the Tukey-McLaughlin and Yuen methods for trimmed means when distributions are skewed. *Biometrical Journal*. 3, 259-273.

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

Wilcox, R. R. (1996). *Statistics for the social sciences.* San Diego, CA: Academic Press.

Wilcox, R. R. (1998). The goals and strategies of robust methods. *British Journal of Mathematical & Statistical Psychology*, 51, 55-62.

Wilcox, R. R. (2001). *Fundamentals of modern statistical methods.* New York, NY: Springer-Verlag.

Wilcox, R. R. (2002). *Applying contemporary statistical techniques.* San Diego, CA: Academic Press.

Wilkinson, L. & Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist.* 594-604.

Zumbo, B. D. & Hubley, A. M. (1998). A note on misconceptions concerning prospective and retrospective power. *The Statistician*, 47, 385-388.

ABSTRACT

COVERAGE AND INTERVAL LENGTH OF WELCH'S AND YUEN'S PROCEDURES FOR SHIFT IN LOCATION AND CHANGE IN SCALE FOR (UN)EQUAL SAMPLE SIZES

by

SAYDEE JONATHAN MENDS-COLE

May 2006

Advisor: Dr. Shlomo Sawilowsky

Major: Educational Evaluation and Research

Degree: Doctor of Education

The purpose of this dissertation study was to assess the statistical precision for Welch's and Yuen's procedures. Several results have practical implications in applied practice. First, Welch's procedure displayed probabilities of coverage less than the confidence level when size was inversely proportional to heteroscedasticity and skewness was –1.33 or 1.64. Second, lower Type I and Type II error rates have been the basis for recommending Yuen's procedure instead of Welch's procedure (Wilcox, 1994; Luh & Guo, 2000); these results showed that unless skewness and heteroscedasticity were great in magnitude (i.e., |skewness| > 1.25 & $\sigma_2 / \sigma_2 \geq 4$), Welch's procedure displayed better statistical precision than Yuen's procedure. Third, if kurtosis of the population defined by the

trimmed mean was less than –1.25, probabilities of coverage were attenuated at the 0.01 alpha level.

AUTOBIOGRAPHICAL STATEMENT

Saydee Jonathan Mends-cole

Education

Expected May 2006	E.D.D. Candidate, Educational Evaluation & Research, Wayne State University, Detroit, MI 4.00/4.00
1999	M.S. Ed. Educational Psychology, Southern Illinois University, Carbondale, IL 3.67/4.00 GPA
1997	B.A. Psychology, Minor: Mathematics, Southern Illinois University, Carbondale, IL