1-1-2011

Approximate Vs. Monte Carlo Critical Values For The Winsorized T-Test

Michael Lance
Wayne State University,

APPROXIMATE VS. MONTE CARLO CRITICAL VALUES FOR THE WINSORIZED T-TEST

by

MICHAEL W. LANCE

DISSERTATION

Submitted to the Graduate School

of Wayne State University,

Detroit, Michigan

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

2011

MAJOR: EDUCATION EVALUATION
AND RESEARCH

Approved by:

____
Advisor Date

____

____

____

بِسۡمِ اللّٰهِ الرَّحۡمٰنِ الرَّحِيۡمِ

DEDICATION

To Ihsaan and Ziyaan

ACKNOWLEDGMENTS

I have been fortunate enough to have Dr. Shlomo Sawilowsky as my dissertation advisor. For that I am grateful. He is an ideal mentor and professor and is always there for me to provide direction, feedback, and insightful guidance.

I am also fortunate to have the following committee members:

Dr. Gail Fahoome has taught me most of what I know of statistical models. She also introduced me to Fortran programming and provided much needed guidance in the process. Dr. Benjamin Kelcey taught me Experimental Design and after the course, I knew I had to have him on my committee due to his clear, thorough explanations and high expectations. I'd like to thank Dr. Claude Schochet for agreeing to be part of my committee on short notice and being helpful in the process.

At several key points as a graduate student, Dr. Peter Bahr has provided me with candid and helpful feedback and encouraged me to always do my personal best, for which I am most thankful.

Elliot P. Chandler (and comp.lang.fortran) deserves special praise for the tireless help with Fortran coding only for the sake and enjoyment of writing code that works and nothing more. I can safely say that this project would have taken at least an extra year to complete were it not for his help (and the comp.lang.fortran group).

I thank Mrs. Nawal Hamadeh for introducing me to this field and guiding me in its direction (and away from being an Assistant Principal, which could have been disastrous). I also thank Ms. Leila Hamade for introducing me to the wonderful world of Excel, which allowed me to complete the analysis of this project quite efficiently.

I thank my family for always encouraging me in my academic career and especially my parents for teaching me what disciplined, hard work is.

Finally, I’d like to thank my wife who has both inspired me to continue my education and provided me with the time to do so and my children who remind why I do what I do.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

Outliers, Trimming, and Winsorizing

The mean is a well-known and “cherished” (Dixon & Yuen, 1974, p.158) estimator of location due to its ease of calculation. However, it is not robust due to its finite breakdown point of $\frac{1}{N}$ (Wilcox, 1996), meaning that only one arbitrarily large or small value, or outlier, can significantly reduce its accuracy. For over a century, there have been many attempts to create algorithms and rules to identify and reject outliers, often for the purpose of eliminating (trimming) or adjusting (Winsorizing) them in order to increase the accuracy of the mean. An example is to use least squares and reject any value that (in magnitude) exceeds five times the probable error. This method assumes a Gaussian distribution (Anscombe & Guttman, 1960) which can be as common as a unicorn (see Micceri, 1989). Kruskal (1960) advised to trim the outliers using least squares and then analyze the remaining data (but recommended to record the outlying values).

Every method of identifying outliers has strengths and weaknesses due to the relative accuracy they provide given a specific distribution of data. In the case of using least squares, for example, the probable error may be too small to be of use (for detection) in the case of multiple outliers that are more extreme in value.

An underlying assumption of the *t*-test for independent samples is that the data are normally distributed. Because this test compares means, it also helps if the mean is an accurate estimate of location. If the mean is inaccurate due to the presence of outliers (and as a result, the distribution departs from normality) the comparison of means can become less robust to type I and II errors (Zimmerman, 1994).

Outliers increase variability about the mean (whereas inliers decrease such variability). However, the removal of outliers is equated with a decrease in the degrees of freedom, and may decrease comparative statistical power. This loss of power can potentially be exacerbated if asymptotic critical *t* values are used since the critical values will be higher, leading to conservative p-values.

The impact of outliers on mean comparisons is common in research. Orr et al. (1991) identified many studies where all data points were retained with no attempt to detect outliers. When outliers were addressed in these studies, they were treated inconsistently. A problem in comparing outliers of various data sets is the fact that not all outliers occur for the same reason. Therefore, the detection and treatment of outliers should vary depending on their causes, should they be known. Yet when causes are common, treatment should be consistent.

Two common procedures for dealing with outliers are trimming and Winsorizing. Trimming involves sorting an array of data, dropping the outliers, and calculating the mean, or other statistic, based on what remains. Winsorizing involves taking those same values that would otherwise be trimmed and replacing them with the values that would remain at the end(s) of the sorted, trimmed array. This serves to pull the mean toward the middle of the distribution (Dixon, 1960) while at the same time preserving the sample size. “In essence, the Winsorized mean pays more attention to the central portion of a distribution by transforming the tails” (Wilcox, 2005, p. 28). For example, the following array would be trimmed and Winsorized as follows:

Original: 1, 5, 6, 7, 10

20% Trimmed: 5, 6, 7

20% Winsorized: 5, 5, 6, 7, 7

The above sample is shown to be both trimmed and Winsorized 20% symmetrically, because 1 out of 5 values are trimmed or Winsorized at each end. This will be referred to as a 20% Winsorization, although the exact percentage will vary. Also, the Winsorized amount is 1, though it occurs at both ends of the data.

In addition to the idea of a mean being robust, another widely-spread misconception is that naturally-occurring data tend to be normally-distributed and that parametric tests (that assume normality) are robust to Type I and II errors under non-normal conditions (Micceri, 1989). For type II errors, the misconception exists even when compared to nonparametric competitors. The contrary has been demonstrated by Sawilowsky (1990). If Type I (or II) error rate of a test is inexact, one may be mistakenly rejecting (or failing to reject) the null hypothesis more or less than the specified significance level, making the test non-robust.

Statement of the Problem

This study aims to compare approximate and Monte Carlo-derived critical values with respect to types I and II error robustness properties. Because one or more outliers may greatly reduce the accuracy of the mean, lead to inexact Type I and/or II error rates, and may either be inconsistently or not addressed in research, the scientific community is in need of robust procedures that can easily be used when one or more outliers are present. The Winsorized *t*-test for independent samples is one such procedure. As the contrived example in table 1 shows, Winsorizing samples before running a *t*-test can

Table 1

Contrived example of the impact of Winsorizing samples before comparing means via the independent samples t-test (a=.025 per tail)

	Sample Data:		Winsorized Data (1/end):		Winsorized Data (3/end):	
	-2000	-100	-100	1	0	3
	-100	1	-100	1	0	3
	0	2	0	2	0	3
	0	3	0	3	0	3
	0	4	0	4	0	4
	0	55	0	55	0	55
	0	60	0	60	0	60
	8	61	8	61	8	61
	8	62	8	62	8	62
	9	63	9	63	9	63
	9	64	9	64	9	64
	12	65	12	65	12	65
	12	100	12	100	12	65
	12	100	12	100	12	65
	13	200	12	100	12	65
Winsorized Amount:	0	0	1 (7.5%)	1 (7.5%)	3 (20%)	3 (20%)
Mean:	-134.47	49.33	-7.87	49.40	5.47	42.73
Variance:	267119.41	4285.95	1424.12	1421.83	29.84	843.92
t-obtained:	-1.37		-4.16		-4.88	
Critical Value:	2.05					
$(h_1+h_2)-2)$ df C.V.:			2.06		2.12	
Winsorized C.V.:			2.39		3.64	
Lower Tail Decision:	Fail to Reject H_0		Reject H_0		Reject H_0	
Upper Tail Decision:	Fail to Reject H_0		Fail to Reject H_0		Fail to Reject H_0	

serve to increase power (since the lower-tail decision became to reject the null after Winsorizing). It is important to note that *t*-obtained changed from -1.37 to -4.16 and -4.88 for one and three Winsorized values (respectively). Though the decisions were the same for traditional (based on the adjusted degrees of freedom) and Winsorized critical values, these values differ. One may be left wondering if this difference in critical values matters (in terms of types I and II error), if this example is realistic, and if such results may occur with "real life" data. This study aims to explore this by drawing samples from eight commonly-occurring distributions (in Education and Psychology) as estimated by Micceri (1989). To be sure, the null hypothesis when using a Winsorized *t* test is that the *Winsorized* population means are the same.

Historically, knowledge of the robustness properties of the Winsorized *t* has been based on traditional, estimated (($h_1 + h_2$) – 2 degrees of freedom) critical values (Dixon & Tukey, 1968). This study aims to determine whether the Winsorized *t* formula should be used with traditional, estimated critical values or if (and/or when) it would be better (more robust) to use the traditional *t* formula with Farrell-Singleton's (2010) table of Winsorized *t* critical values. The use of Micceri's real population distribution estimates from Educational and Psychometric data will help to generate results that account for real-world data encountered by researchers in these areas and possibly beyond (where measures are often discrete or bounded). This study will also generate samples from Mathematical distributions for comparison with results from other such studies.

Assumptions and Limitations:

The real population distributions used in these simulations (from Micceri, 1989) are estimated from specific types of measures (achievement and psychometric) and therefore any results obtained regarding robustness should pertain only to statistical analyses conducted with those types of data sets and for Mathematical distributions, though their occurrence may be rare in real data sets.

Definitions

Alpha (a): See type I error.

Assumption: A requirement for a statistical test in order that type I errors will be as specified (i.e. p = .05)

Critical Value: A value used to determine if the formulaic result of a statistical procedure is significant.

Conservative: When a test does not reject the null hypothesis as much as it should for a given type I error rate.

Contamination: When values in a distribution occur due to the presence of a separate distribution.

Degrees of Freedom (df): The number of values in a sample that are free to vary after all others are constrained by a formula.

g-times: The number of values being Winsorized from each end of a sample (also called *k-times)*.

Heavy-Tailed: A characteristic of a distribution where the probability for extreme values exceeds that of the distribution assumed by the test (i.e. normal distribution).

Least Squares: A procedure that minimizes the squared differences between each value in a sample and the mean for that sample.

Liberal(1): When a test rejects the null hypothesis more than it should for a given type I error rate.

Liberal (2): According to Bradley (1978), when type I error falls within plus or minus half of the nominal alpha level.

Lower Tail: The lower extreme of a set of values in a distribution.

Monte Carlo Simulation: The use of a computer program to simulate some aspect of reality to make determinations of the nature of or change in reality through repeated sampling via Monte Carlo methods (Sawilowsky & Fahoome, 2003).

Normality: A state of a data distribution where it fits the normal or Gaussian curve.

Nonparametric Test: A statistical test of significance that makes no assumption about the shape of the sampling distribution.

Outlier: A datum whose value is outside of what is considered to be the normal range of a given distribution.

Parametric Test: A statistical test of significance that assumes specific population parameters about the shape of the sampling distribution.

Power: The ability to detect a significant difference or effect between sets of data (also known as the inverse of type II error).

Robustness: The degree to which a statistical test maintains types I and II error rates in light of assumption violations.

Robust Test (Procedure): A statistical test that maintains type I error rates in light of assumption violations.

Stringent: According to Bradley (1978), when type I error falls within plus or minus one-tenth of the nominal alpha level.

Sum of Squares: The sum of the squared differences between each value and the mean of all values within a sample.

Skewed Distribution: A distribution that has extremely high or low scores which pull the distribution to one side or the other.

Significance Level: The probability of making a type I error when conducting a hypothesis test (Triola, 1997).

Type I Error: The rejection of the null hypothesis when it is true.

Type II Error: The failure to reject the null hypothesis when it is false (known as the inverse of power).

Upper Tail: The upper extreme of a set of values in a distribution.

CHAPTER II

LITERATURE REVIEW

Addressing Outliers

According to Kruskal (1960), the decision to include or exclude outliers from an analysis should depend on the goal of the study. If the goal is to, for example, calculate the average value, he recommended omitting the outlier (noting the frequency, magnitude, and value). If the goal is to compare methods for measuring the average, then he recommended including the outlier. Kruskal also recommended running tests with and without outliers and if the null is not rejected or accepted in both cases, the researcher “should view any conclusions from the experiment with very great caution” (p. 3). No mention of experiment-wise error inflation (as a result of running two tests) was mentioned in the article, which of course, could be a potentially serious limitation to Kruskal’s approach. Also, the critical values for each experiment (with and without outliers) or the formula for obtaining the *t* statistic may need to differ from the traditional values or formula, which is the topic of this dissertation.

In many cases, the representativeness of any potential outlier is not known because population parameters are often not known. Therefore, it is best to attempt to classify outlier(s) as follows:

Outliers within a sample usually occur due to one of the following reasons:

1. Typographical error
2. Measurement error
3. A heavy-tailed population distribution
4. Contamination

Hawkins (1980) suggested that “when deciding whether to use Winsorization or outright rejection (trimming) of outlying observations, one should be guided by the underlying model of the data” (p. 5).

Typographical errors can occur during data entry. For example, an “83” may unintentionally be typed instead of a “38”, causing it to be an outlier. When this occurs, the outlier is neither representative of the sample nor the population and (according to Hawkins) should be trimmed because it “contain(s) no information about the basic distribution” (Hawkins, p. 5). With data entry being a common practice in research, it follows that such outliers are common.

Measurement error is prevalent in any tool used to measure attitudes, knowledge, or hypothetical actions or decisions. Though items, indices, and scales may be rated for reliability and validity, resulting data may (and often will) contain outliers for a variety of reasons. For example, if a student always has scores of 90% or above and on a similar test receives a 50% on a test covering the same content, this is an obvious case of an outlier likely attributable to measurement error. As with the example of the typographical error above, the outlier should be trimmed, according to Hawkins, for the same reason.

Some populations are naturally heavy-tailed.

> If the observations are generated by mechanism (i), the heavy-tailed distribution, and one wishes to estimate the parameters of this distribution, then the outliers represent valid observations. Thus one should be reluctant to discard them entirely, and hence prefer to use Winsorization, which is robust, but does make partial use of the outliers (Hawkins, p. 5)

Contamination occurs when the outliers are present as a result of sampling from a different distribution in addition to the distribution of interest. Hawkins describes this as mechanism (ii).

> If, on the other hand, one believes that mechanism (ii) is operative, and one is interested in estimating the parameters of the basic distribution, then Winsorization should not be used. In this case, the outliers may be presumed to come from the contaminating distribution, and hence to contain no information about the basic distribution. Thus to the extent that one is sure that they are not from the basic distribution, one should ignore them. This may be done in a classical way by deleting them as soon as one rejects the null hypothesis that they come from the basic distribution, or in a Bayesian way by assigning them smaller weights as they deviate more from the basic distribution. (Hawkins, p. 5)

An example of a contaminated distribution can be a set of reading scores from an eighth grade standardized test. If the researcher is interested in measuring the efficacy of a reading intervention on native English speakers and the distribution is mixed normal due to 20% of the students being non-native English speakers, the detection and treatment of outliers becomes an obvious necessity if all data points from the contaminating distribution can be identified. Dixon (1950) proposed a model for identifying values resulting from contamination that aimed to allow such separation and analysis. If this cannot be done, the contaminated model must be analyzed with a robust procedure that maintains power.

In summary, the most appropriate time to Winsorize outliers (according to Hawkins) is when that sample is drawn from a heavy-tailed distribution. Typographical error, measurement error, and contamination warrant trimming. This is because Winsorizing effectively draws the values closer to the center of the distribution and as such, only values that are representative of the distribution by using existing data (outliers) of interest should be Winsorized. If an outlier is not representative and is kept (or if a representative outlier is dropped), the sample becomes biased (Dixon, 1950). According to Micceri's (1989) study of 440 Education and Psychology data sets, many distributions in Education and Psychology are in fact heavy-tailed (moderately or extremely). There are also instances of contaminated distributions (mixed-normal, for example). Thus, at least for some types of measures, Winsorizing seems to be the best choice according to Hawkins, 1980. This will be tested for the Winsorized *t* for independent samples in similar situations.

Dixon & Massey (1969) showed the Winsorized mean to be more efficient than the trimmed mean for Gaussian and close to Gaussian distributions, but less efficient for distributions with very long tails, which may imply contamination and, if so, is consistent with Hawkins's (1980) suggestion to trim instead of Winsorize in such instances. Rivest (1994) showed that Winsorizing is particularly helpful for skewed distributions.

Some researchers prefer to use a rule of thumb method for Winsorizing or trimming, but this can lead to inexact estimates of location as the method (or criterion) used to detect outliers can impact subsequent inferences (Carey et al., 1997). Though it involves a process that is slightly more complex than using a simple rule of thumb, Maximum likelihood (M-) estimators can be used to identify the exact number of values at each end of a sample to Winsorize or trim (Wilcox, 1998) and have been shown by Sawilowsky (2002) to produce narrower bracketed (confidence) intervals with real Educational and Psychological data distributions (see Micceri, 1989) for sample sizes less than 50 (and in most cases, for samples greater than 50).

Robustness to Type I Error:

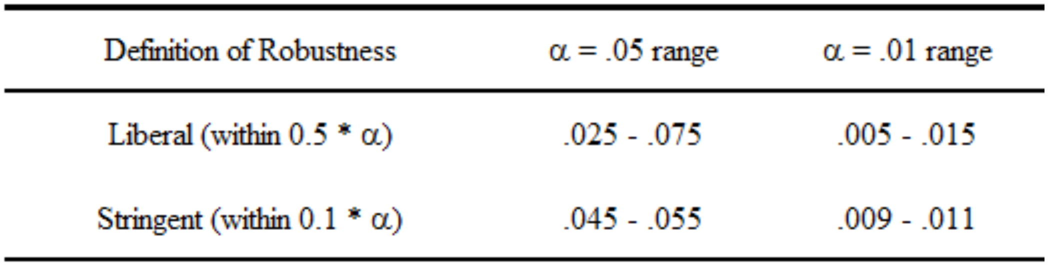
Bradley (1978) asserted that robustness, as discussed in many statistics textbooks, has lacked a solid definition. Even after decades of research aimed at honing the definition for specific tests, many textbooks still only give general guidelines. He did, however, provide some definitions of his own. The need for a study like this to apply

Bradley's definitions of robustness for the Winsorized *t* exists as those for the regular *t* have existed (and continues so for tests conducted with real data distributions not examined by Micceri, 1989).

Used in several studies of type I error (see Maxwell, 1980 and Ramsey et al., 2010), Bradley's (1978) criteria identifies stringent and liberal type I robustness. According to Bradley, p-values between 0.5α and 1.5α $(|p-\alpha| \leq \frac{\alpha}{2})$ (i.e. for α=.05, between .025 and .075) are considered to be meet a liberal criteria for robustness, while p-values between 0.1α and 1.1α $(|p-\alpha| \leq \frac{\alpha}{10})$ are considered to meet a stringent criteria. Table 2 shows liberal and stringent robustness definitions for both α=.05 and α =.01:

Table 2

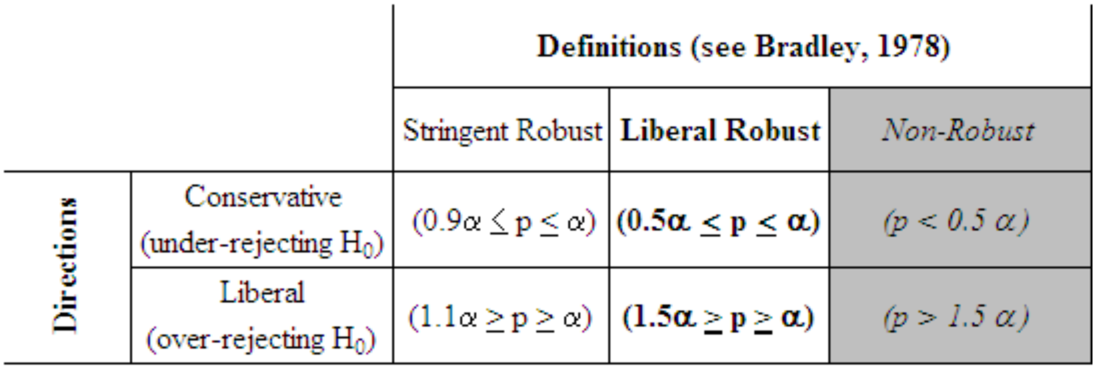
Definitions of Robustness to type I error resulting from Monte Carlo Simulations



Definition of Robustness	α = .05 range	α = .01 range
Liberal (within 0.5 * α)	.025 - .075	.005 - .015
Stringent (within 0.1 * α)	.045 - .055	.009 - .011

Note: See Bradley, 1978 (p. 146)

The directions of non-robustness are conservative (not rejecting the null hypothesis as much as the alpha level allows) and liberal (rejecting the null hypothesis more than allowed by the alpha level). Figure 1 illustrates the interactions between range (definition) and direction of (non-)robustness and how these will be identified (no formatting, bolded, and italicized with highlighting) in this chapter:



		Definitions (see Bradley, 1978)		
		Stringent Robust	**Liberal Robust**	*Non-Robust*
Directions	Conservative (under-rejecting H_0)	$(0.9\alpha \leq p \leq \alpha)$	$(\mathbf{0.5\alpha \leq p \leq \alpha})$	$(p < 0.5\ \alpha)$
	Liberal (over-rejecting H_0)	$(1.1\alpha \geq p \geq \alpha)$	$(\mathbf{1.5\alpha \geq p \geq \alpha})$	$(p > 1.5\ \alpha)$

Figure 1: Directions and definitions of type I error ranges.

It has long been known by those who study robustness that real data seldom approximate the normal curve. Even Gossett (who originally formulated the *t*-test) acknowledged this and encouraged robustness studies with this knowledge in mind (Pearson & Please, 1975). Pearson & Please, among others (i.e. Boneau, 1960, Bradley, 1977), conducted robustness studies on real data, but the benchmark for such studies is Micceri (1989). Micceri's advantage was not only the number of real distributions collected; he also had modern computer sampling and data sets from journals, test publishers, school districts, the Florida Department of Education, and the University of South Florida's institutional research department. This allowed for more accurate estimation of distributions common to achievement and psychometric measures. It should also be noted here that data from Education and Psychology has especially extreme departures from normality.

One of the most commonly applied statistical procedures, the *t*-test is used across most fields of research. Yet, since most data distributions do not meet the assumption of normality, this is cause for concern. Sawilowsky & Blair (1992) noted that the *t*-test is robust to type I error under the following conditions:

1. Sample sizes must be equal or nearly so.

2. Sample sizes must be fairly large (25 to 30 according to Boneau, 1960)

3. Tests should be two-tailed instead of one-tailed.

With these conditions met, Sawilowsky & Blair (in referencing Efron, 1969; Gayen, 1949, 1950; Geary, 1936, 1947; Pearson & Please, 1975) observed results to be of a conservative nature (Sawilowsky and Blair, 1992).

The next several pages will review the study of Sawilowsky and Blair (1992), in which robustness of the *t*-test was examined when data were randomly drawn (with replacement) from Micceri's real data sets.

Using eight real distributions identified and estimated by Micceri (1989), Sawilowsky & Blair (1992) conducted Monte Carlo studies on the independent samples *t*-test for instances of departures from normal distributions. Samples were drawn with replacement and the *t* obtained computed per iteration (of 10,000) of each sample size combination. Next, the obtained value was compared to the critical value in order to determine if a rejection of the null hypothesis was warranted. The authors noted that "These real distributions highlight situations in which the *t*-test was, by any definition, non-robust to Type I error. The degree of non-robustness in these instances was at times more severe than has been previously reported" (p. 359).

It is important to note that the non-robust results occurred when the above three conditions were not met and that skew was the biggest factor in such cases. Overall and specific descriptive information for each distribution is provided in table 3.

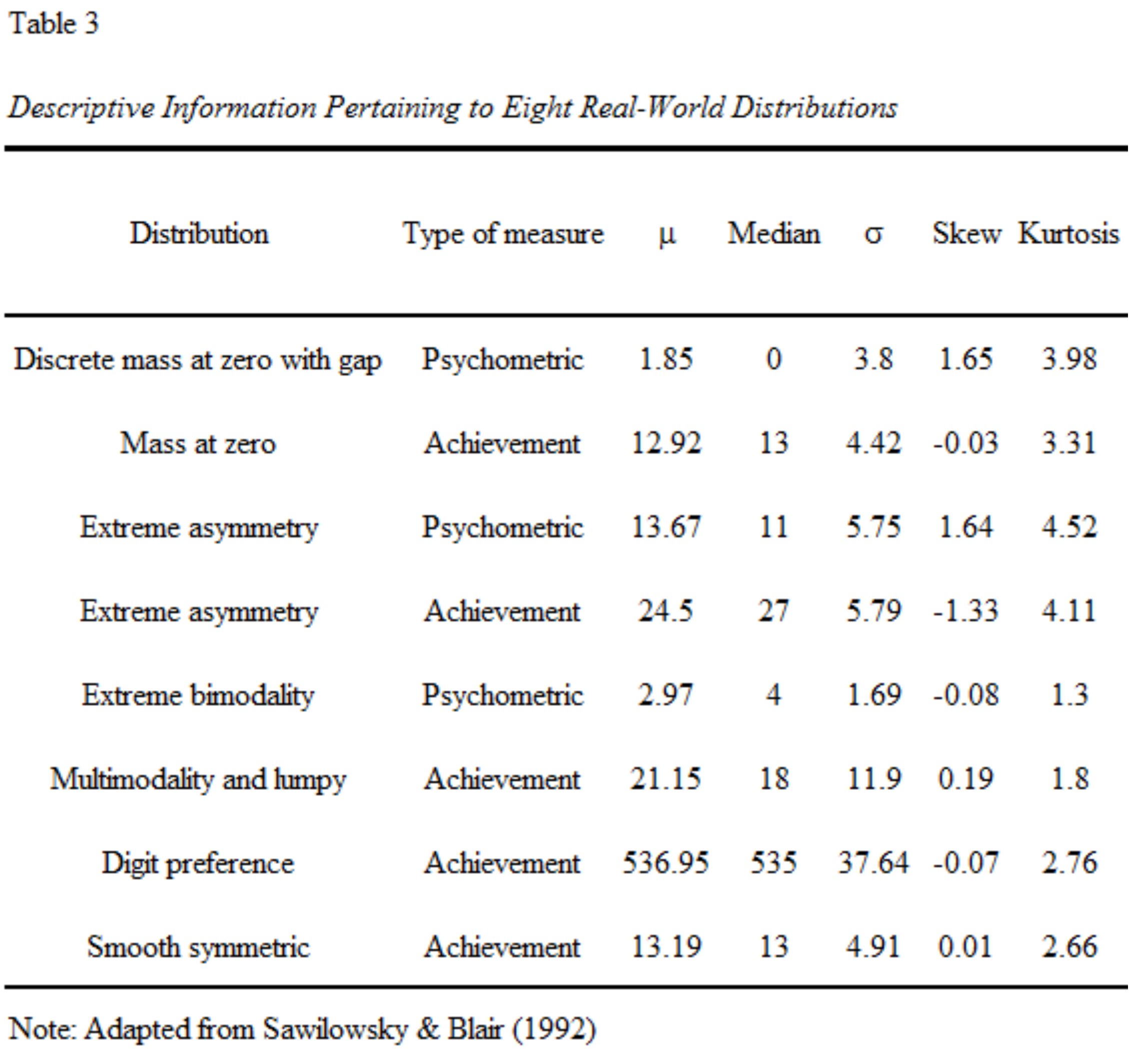


Table 3

Descriptive Information Pertaining to Eight Real-World Distributions

Distribution	Type of measure	µ	Median	σ	Skew	Kurtosis
Discrete mass at zero with gap	Psychometric	1.85	0	3.8	1.65	3.98
Mass at zero	Achievement	12.92	13	4.42	-0.03	3.31
Extreme asymmetry	Psychometric	13.67	11	5.75	1.64	4.52
Extreme asymmetry	Achievement	24.5	27	5.79	-1.33	4.11
Extreme bimodality	Psychometric	2.97	4	1.69	-0.08	1.3
Multimodality and lumpy	Achievement	21.15	18	11.9	0.19	1.8
Digit preference	Achievement	536.95	535	37.64	-0.07	2.76
Smooth symmetric	Achievement	13.19	13	4.91	0.01	2.66

Note: Adapted from Sawilowsky & Blair (1992)

As can easily be seen with the Discrete Mass at Zero with Gap distribution (figure 2, table 4), one-tailed and two-tailed (total) results are generally conservative (with notable exceptions from upper-tail results for uneven samples) with many outside of Bradley's (1978) liberal criteria of robustness. The Mass at Zero distribution (figure 3, table 5) yielded mostly stringently-robust results at .05 alpha level, while the .01 level yielded many liberally-robust results in both conservative (mostly) and liberal directions.

For the Extreme Asymmetric (Psychometric) distribution (figure 4, table 6), two-tailed results tended to be more exact or within the liberal criteria of robustness.

However, for unbalanced samples, upper-tail results tended to be liberal and lower-tail results conservative. This makes sense when looking at the skew of this distribution. For $\alpha = .10$, this trend is accentuated by non-robust (by liberal definition) results in both respective directions.

Results for the Extreme Asymmetric (Achievement) distribution (figure 5, table 7) tended to be the opposite of those from its Psychometric counterpart. A brief glance at this distribution explains why; they are both skewed yet in opposite directions.

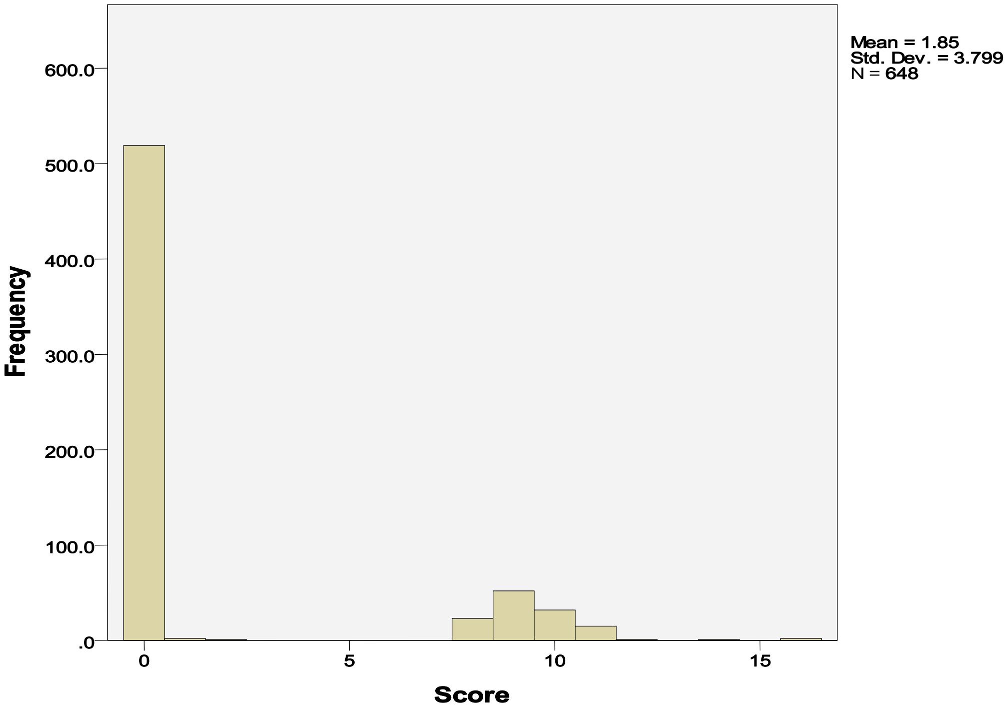


Figure 2: Discrete Mass at Zero with Gap. Adapted from Sawilowsky & Blair, 1992 (p. 356).

For the Extreme Bimodal (Psychometric) distribution (figure 6, table 8), two-tailed results were liberal or exact but rarely ever conservative.

Table 4

Type I Error Rates for Independent-Samples t Test for Various Sample Sizes and Alpha Levels When Sampling Is From a Discrete Mass at Zero With Gap (Psychometric) Distribution, 10,000 Repetitions

	α = .050			α = .010		
Sample Size	U025	L025	Total	U005	L005	Total
5, 15	*0.003*	*0.000*	*0.003*	*0.001*	*0.000*	*0.001*
10, 10	*0.007*	*0.006*	*0.013*	*0.000*	*0.001*	*0.001*
10,30	**0.036**	*0.000*	**0.036**	*0.011*	*0.000*	0.011
20,20	**0.02**	**0.02**	**0.04**	*0.002*	*0.002*	*0.004*
15,45	**0.03**	*0.004*	**0.034**	*0.01*	*0.000*	0.01
30,30	**0.021**	0.024	0.045	**0.004**	**0.003**	**0.007**
20,60	**0.03**	*0.009*	**0.039**	*0.009*	*0.000*	0.009
40,40	0.027	0.025	0.052	**0.006**	0.005	0.011
30,90	**0.031**	**0.018**	0.049	**0.007**	*0.002*	0.009
60,60	0.025	0.024	0.049	0.005	**0.006**	0.011

Note. Adapted from Sawilowsky & Blair (1992). Sample size = n1, n2. U025, U005 = proportion of rejections in the upper-tail. L025, L005 = proportion of rejections in the lower tail.

Outside of Bradley's (1978) stringent definition of type I robustness.

Outside of Bradley's (1978) liberal definition of type I robustness.

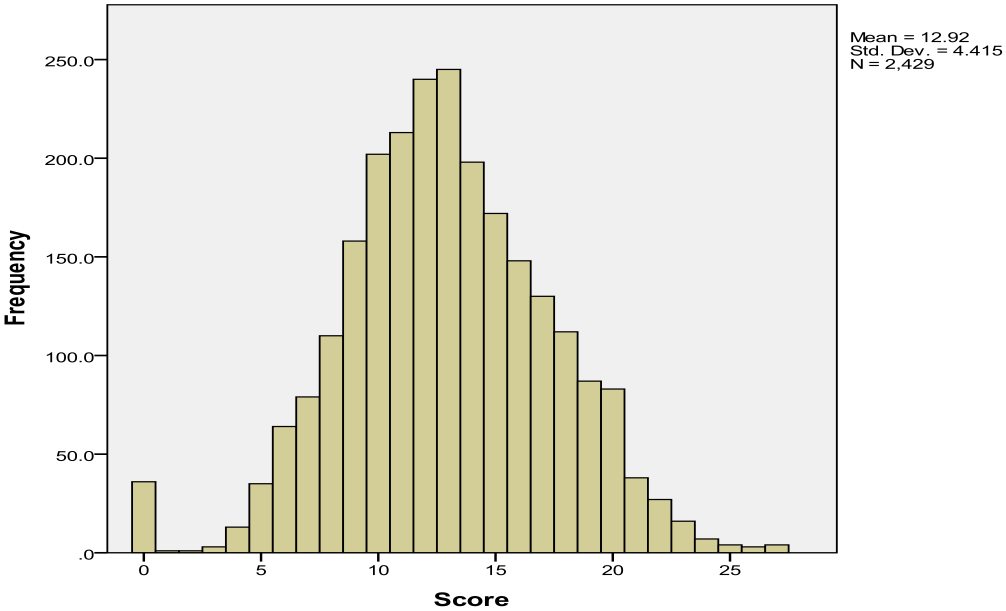


Figure 3: Mass at Zero. Adapted from Sawilowsky & Blair, 1992 (p. 356).

Multimodal and Lumpy (achievement, figure 7, table 9) results were generally stringently-robust for α = .05, though there are many examples of liberally-robust results in the liberal direction for α = .01.

For the Digit Preference (achievement) distribution (figure 8, table 10), results were generally robust for α = .05 with a mixture of liberally-robust (both directions) and stringently-robust results for α = .10, especially for samples of 20 or less and for upper-tail results.

The Smooth Symmetric (achievement) distribution (figure 9, table 11) yielded mostly stringently-robust results for α = .05. Results for α = .10 were all either stringently or liberally robust, with liberally-robust results occurring in both directions for upper and lower-tail results.

When running a *t*-test, the most important part of a distribution is in the tails, since the decision regarding the null hypothesis depends on them. This explains why

Table 5

Type I Error Rates for Independent-Samples t Test for Various Sample Sizes and Alpha Levels When Sampling Is From a Mass at Zero (Achievement) Distribution, 10,000 Repetitions.

	α = .050			α = .010		
Sample Size	U025	L025	Total	U005	L005	Total
5, 15	0.026	0.024	0.050	0.005	**0.003**	**0.008**
10, 10	0.024	0.025	0.049	0.005	0.005	0.010
10,30	**0.022**	0.027	0.049	0.005	**0.006**	0.011
20,20	0.023	0.025	0.048	**0.004**	**0.004**	**0.008**
15,45	**0.022**	0.024	0.046	**0.006**	**0.004**	0.010
30,30	0.023	0.027	0.050	**0.004**	0.005	0.009
20,60	0.024	0.026	0.050	**0.004**	0.005	0.009
40,40	0.025	0.027	0.052	0.005	**0.006**	0.011
30,90	0.026	0.026	0.052	0.005	**0.004**	0.009
60,60	0.024	0.026	0.050	0.005	0.005	0.010

Note. Adapted from Sawilowsky & Blair (1992). Sample size = n1, n2. U025, U005 = proportion of rejections in the upper-tail. L025, L005 = proportion of rejections in the lower tail.

Outside of Bradley's (1978) stringent definition of type I robustness.

Outside of Bradley's (1978) liberal definition of type I robustness.

skewness (in combination with uneven samples) impacts the robustness of the *t* more than kurtosis (Sawilowsky & Blair, 1992), which can be seen when results in each tail (for unbalanced samples) tend to be in opposite directions for skewed distributions, with the tail containing the most values giving more conservative results for that tail.

Interestingly, Micceri (1989) observed 97% of empirical distributions studied in Psychology and Education having longer tails than the normal distribution. Also, the conditions (outlined by Sawilowsky & Blair) necessary for a *t*-test to be robust are not the norm. Taken together, this illustrates the need for a robust test such as the two-sample Winsorized *t* for studies conducted with real data in Education, Psychology, and probably most other disciplines, in order to obtain results that are robust to type I error.

Although they are based on much smaller samples, the distributions (figure 10) from Pearson & Please (1975, p. 225) illustrate how data from other disciplines can be relatively more normal. This, of course, depends on the type (and sensitivity) of the measure.

Boneau (1960) showed the *t* to be remarkably robust for equal samples and variances. However, the distributions used in the study were from normal, Exponential, and rectangular (Uniform) distributions (see table 12). While the *t* proved to be relatively non-robust for the Exponential distribution in most cases and to a lesser degree for the rectangular distribution for samples of 15, the applied researcher must consider how common such distributions are with real data.

Robustness to type II error

If the researcher is unsure about the population distribution parameters of either sample, then if outliers are found on either sample, both tails from both samples should

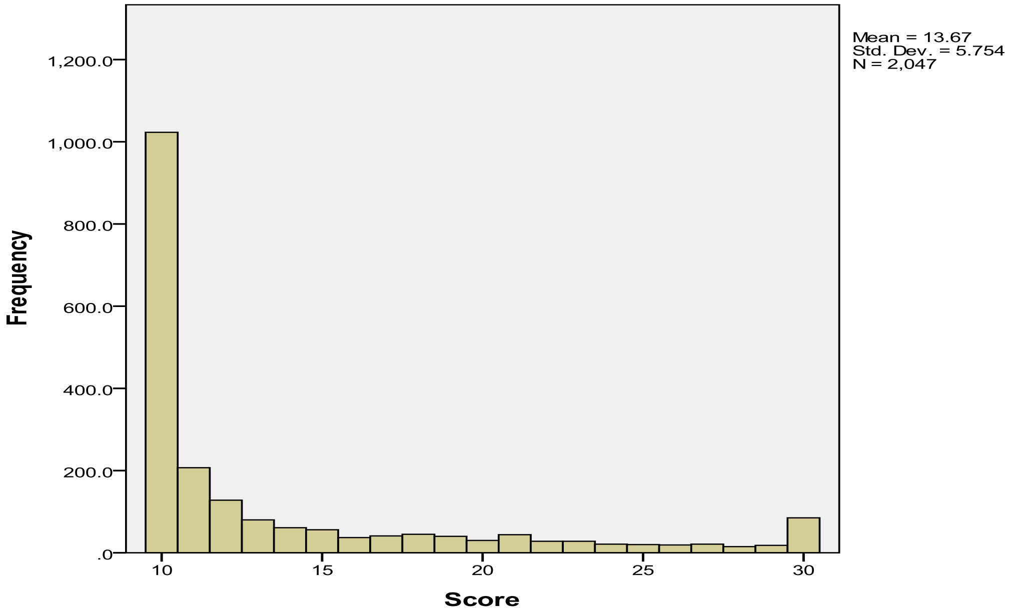


Figure 4: Extreme Asymmetry (Psychometric). Adapted from Sawilowsky & Blair, 1992 (p. 356).

be Winsorized to compare the central and common data characteristics. Under normality, Winsorizing can lead to a minimal power loss, but with long-tailed distributions, it can lead to a great gain in power (Fung & Rahman, 1980). Yuen and Dixon (1973) reached the same conclusion with trimming. Fung and Rahman showed the trimmed and Winsorized *t*-tests to have immaterially small power differences. Yuen & Dixon (1973) found that for samples equal-to or greater-than 10, the loss of power for both trimmed and Winsorized *t*-tests is negligible under normality and for samples equal-to or less-than 5, the regular *t* is recommended except for instances of substantial departures from normality. Sawilowsky and Blair (1992) found the *t*-test to be robust to type II error under non-normal conditions, but suggested that robust nonparametric competitors may be a better choice. The Winsorized *t* may be an example of such a robust competitor.

Table 6

Type I Error Rates for Independent-Samples t Test for Various Sample Sizes and Alpha Levels When Sampling Is From an Extreme Asymmetric (Psychometric) Distribution, 10,000 Repetitions.

	α = .050			α = .010		
Sample Size	U025	L025	Total	U005	L005	Total
5, 15	*0.039*	*0.002*	**0.041**	*0.012*	*0.000*	**0.012**
10, 10	**0.022**	0.023	0.045	*0.002*	**0.003**	**0.005**
10,30	**0.037**	*0.011*	0.048	*0.011*	*0.000*	0.011
20,20	0.023	0.024	0.047	**0.004**	0.005	0.009
15,45	**0.031**	**0.015**	0.046	*0.009*	*0.001*	0.010
30,30	0.025	0.025	0.050	**0.006**	0.005	0.011
20,60	**0.033**	**0.015**	0.048	*0.009*	*0.001*	0.010
40,40	0.023	0.025	0.048	0.005	0.005	0.010
30,90	0.027	**0.019**	0.046	**0.006**	*0.002*	**0.008**
60,60	0.026	0.024	0.050	**0.006**	**0.004**	0.010

Note. Adapted from Sawilowsky & Blair (1992). Sample size = n1, n2. U025, U005 = proportion of rejections in the upper-tail. L025, L005 = proportion of rejections in the lower tail.

Outside of Bradley's (1978) stringent definition of type I robustness.

Outside of Bradley's (1978) liberal definition of type I robustness.

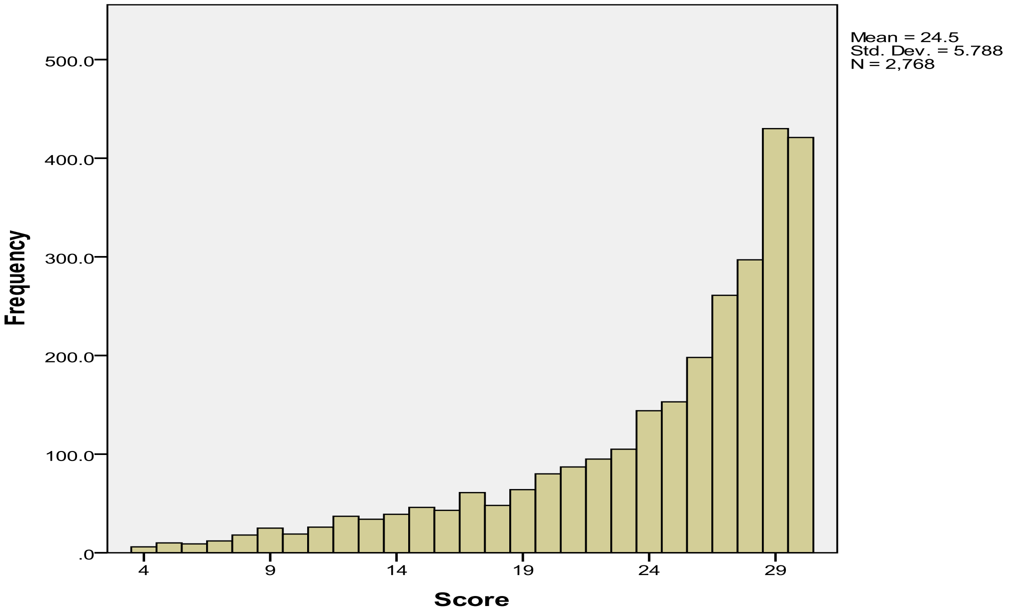


Figure 5: Extreme Asymmetry (Achievement). Adapted from Sawilowsky & Blair, 1992 (p. 357).

Winsorized t

For long-tailed underlying distributions, Dixon & Tukey (1968) recommended using the Winsorized *t*. "When the population is normally distributed, Winsorized *t* also behaves, to a satisfactory approximation, to Student's *t* with h – 1 degrees of freedom. Asymptotically, the ratio tends to a Gaussian variate and standard normal tables can be used" (Dixon & Tukey, 1968). For two samples, this is extended to ($h_1 + h_2$) – 2 degrees of freedom. As such, the formula for the independent samples Winsorized *t* used is identical to that of the regular *t* for independent samples with a few slight changes in notation (Farrell-Singleton, 2010):

$$t_w = \frac{\overline{x}_{w1} - \overline{x}_{w2}}{\sqrt{\frac{(n_1 - 1)S^2{}_{xwk1} + (n_2 - 1)S^2{}_{xwk2}}{n_1 + n_2 - 2}\left[\frac{n_1 + n_2}{n_1 n_2}\right]}}$$

Table 7

Type I Error Rates for Independent-Samples t Test for Various Sample Sizes and Alpha Levels When Sampling Is From an Extreme Asymmetric (Achievement) Distribution, 10,000 Repetitions

	$\alpha = .050$			$\alpha = .010$		
Sample Size	U025	L025	Total	U005	L005	Total
5, 15	*0.011*	**0.033**	**0.044**	*0.000*	*0.010*	0.010
10, 10	0.025	0.025	0.050	0.005	**0.003**	**0.008**
10,30	**0.015**	**0.032**	0.047	*0.001*	**0.007**	**0.008**
20,20	0.024	0.023	0.047	0.005	0.005	0.010
15,45	**0.021**	**0.030**	0.051	*0.001*	*0.008*	0.009
30,30	0.024	0.025	0.049	0.005	0.005	0.010
20,60	**0.020**	**0.029**	0.049	**0.003**	*0.008*	0.011
40,40	0.026	0.027	0.053	**0.006**	0.005	0.011
30,90	**0.019**	0.026	0.045	**0.004**	**0.006**	0.010
60,60	0.025	0.026	0.051	**0.004**	0.005	0.009

Note. Adapted from Sawilowsky & Blair (1992). Sample size = n1, n2. U025, U005 = proportion of rejections in the upper-tail. L025, L005 = proportion of rejections in the lower tail.

Outside of Bradley's (1978) stringent definition of type I robustness.

Outside of Bradley's (1978) liberal definition of type I robustness.

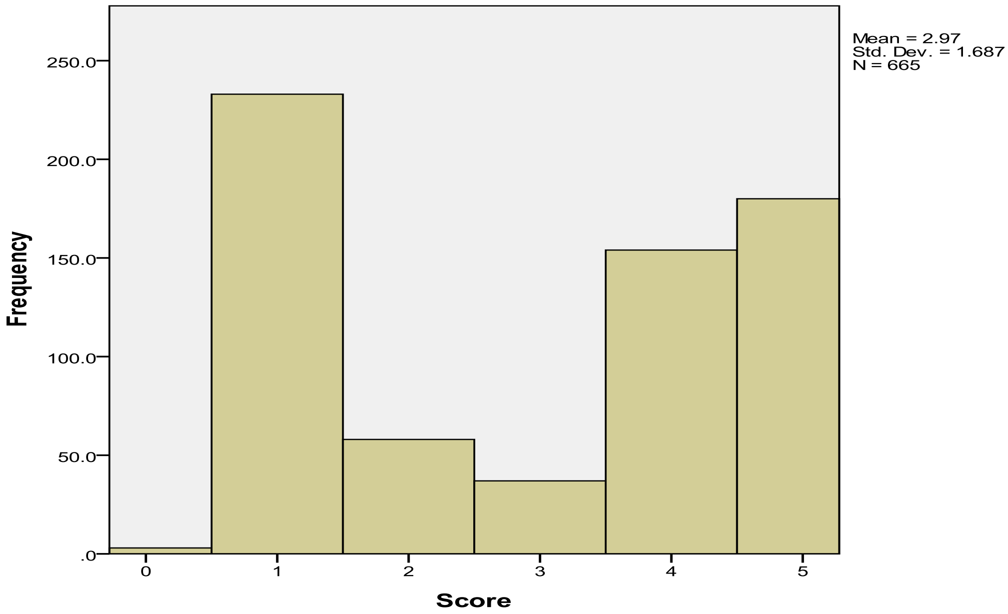


Figure 6: Extreme Bimodal. Adapted from Sawilowsky & Blair, 1992 (p. 357).

with the Winsorized variance:

$$s_{wk}^{2} = (k+1)\left(x_{(k+1)} - \bar{X}_{wk}\right)^{2} + \sum_{i=k+2}^{n-k-1}\left(x_{(i)} - \bar{X}_{wk}\right)^{2} + (k+1)\left(x_{(n-k)} - \bar{X}_{wk}\right)^{2}$$

where $\bar{X}_w$ is the Winsorized mean, $y_1,\mathrm{K}\ ,y_n$ are y ordered observations from a sample, and k is the number of Winsorized values. There have been alternative formulas recommended by Fung and Rahman (1980) and Gans (1988), but none of these have been subject to verification (in generating critical values) via Monte Carlo methods with 1,000,000 iterations as those by Farrell-Singleton (2010), as was done in this study. This is not to say that these alternative formulas are not as good or superior and they may be worthy for use in replicating this study in the future.

Table 8

Type I Error Rates for Independent-Samples t Test for Various Sample Sizes and Alpha Levels When Sampling Is From an Extreme Bimodal (Psychometric) Distribution, 10,000 Repetitions

	$\alpha = .050$			$\alpha = .010$		
Sample Size	U025	L025	Total	U005	L005	Total
5, 15	0.025	0.023	0.048	**0.004**	0.005	0.009
10, 10	**0.028**	0.027	0.055	*0.008*	*0.008*	*0.016*
10,30	0.025	0.025	0.050	0.005	0.005	0.010
20,20	0.025	0.027	0.052	**0.006**	**0.006**	**0.012**
15,45	0.026	0.024	0.050	**0.004**	0.005	0.009
30,30	0.023	0.025	0.048	0.005	0.005	0.010
20,60	0.024	0.025	0.049	**0.006**	**0.006**	**0.012**
40,40	0.027	0.024	0.051	0.005	0.005	0.010
30,90	0.025	0.024	0.049	0.005	0.005	0.010
60,60	0.025	0.026	0.051	0.005	0.005	0.010

Note. Adapted from Sawilowsky & Blair (1992). Sample size = n1, n2. U025, U005 = proportion of rejections in the upper-tail. L025, L005 = proportion of rejections in the lower tail.

Outside of Bradley's (1978) stringent definition of type I robustness.

Outside of Bradley's (1978) liberal definition of type I robustness.

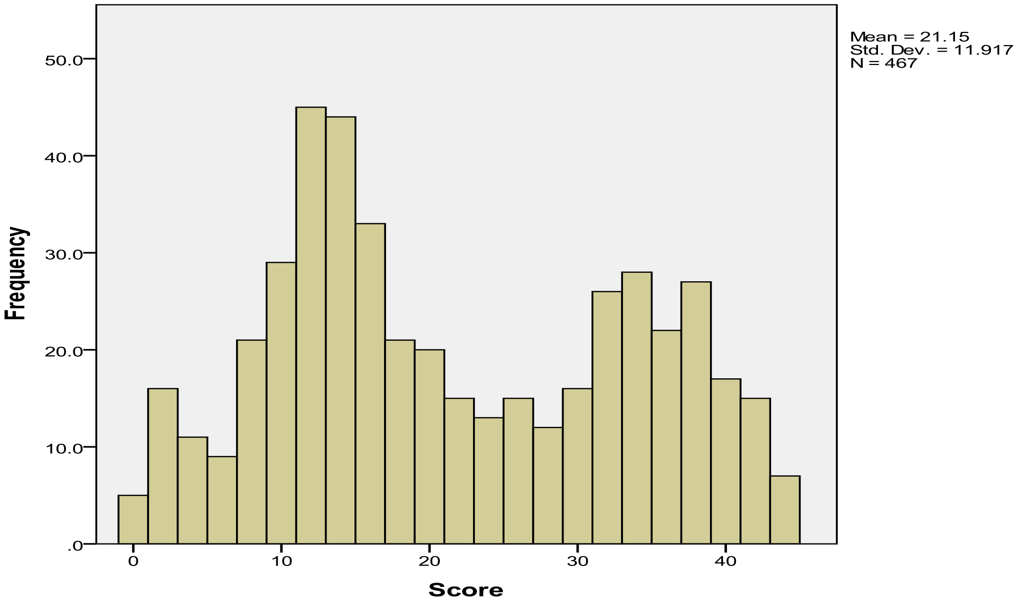


Figure 7: Multimodal and lumpy. Adapted from Sawilowsky & Blair, 1992 (p. 357).

Monte Carlo Methods:

The Monte Carlo method has its modern roots in particle physics, where it was first used by Scientists at the Los Alamos Laboratory to detect the location (or distance traveled) of neutrons (Metropolis, 1987) and was instrumental in research leading up to the development of the atomic bomb. Metropolis & Ulam (1949) described it as a technique that is made possible with the help of modern (at the time, punch card-based) computers. This modern analytical method eclipsed the previously tedious method of sampling (Metropolis) and served to exponentially increase the efficiency of the process. “Monte Carlo refers to repeated sampling from a probability distribution to determine the long run average of some parameter or characteristic” (Sawilowsky & Fahoome, 2003, p. 46). Sampling is done with replacement such that each value has an equal chance of selection for every sample. Upon sampling, data are analyzed and results are recorded

Table 9

Type I Error Rates for Independent-Samples t Test for Various Sample Sizes and Alpha Levels When Sampling Is From a Multimodal and Lumpy (Achievement) Distribution, 10,000 Repetitions

	α = .050			α = .010		
Sample Size	U025	L025	Total	U005	L005	Total
5, 15	0.025	0.023	0.048	0.005	**0.004**	0.009
10, 10	0.026	0.026	0.052	**0.006**	0.005	0.011
10,30	0.025	0.023	0.048	0.005	0.005	0.010
20,20	0.025	0.027	0.052	**0.006**	**0.006**	**0.012**
15,45	**0.029**	0.025	0.054	**0.007**	0.005	**0.012**
30,30	0.025	0.026	0.051	0.005	0.005	0.010
20,60	0.025	0.025	0.050	**0.006**	0.005	0.011
40,40	0.025	0.024	0.049	**0.006**	**0.006**	**0.012**
30,90	0.025	0.026	0.051	**0.006**	**0.006**	**0.012**
60,60	**0.029**	0.026	0.055	0.005	**0.006**	0.010

Note. Adapted from Sawilowsky & Blair (1992). Sample size = n1, n2. U025, U005 = proportion of rejections in the upper-tail. L025, L005 = proportion of rejections in the lower tail.

Outside of Bradley's (1978) stringent definition of type I robustness.

Outside of Bradley's (1978) liberal definition of type I robustness.

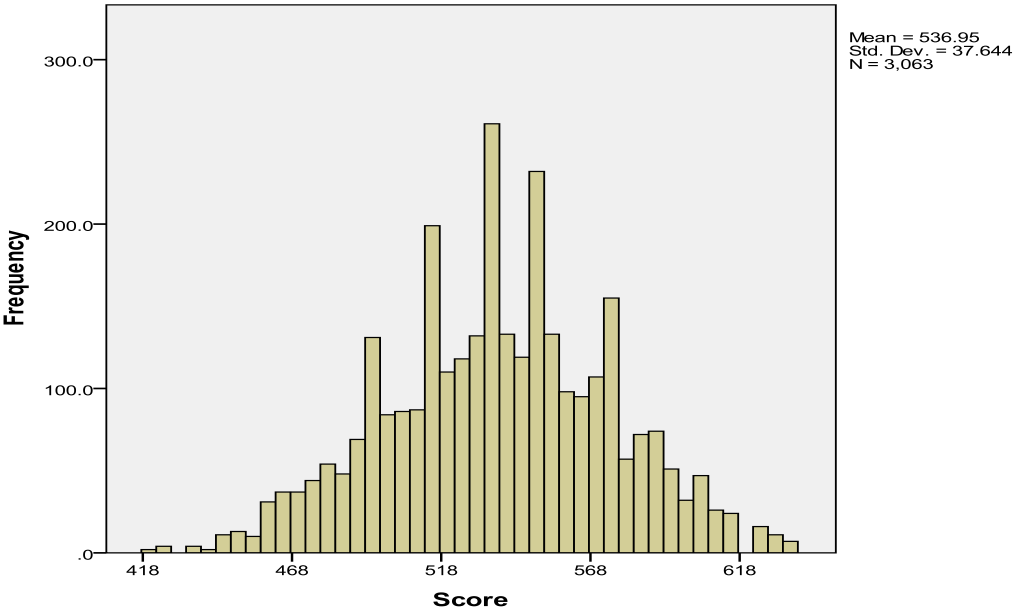


Figure 8: Digit Preference. Adapted from Sawilowsky & Blair, 1992 (p. 358).

before proceeding to the next sample/analysis. The overall results (from all samples) are tallied in some form.

“Simulation is the representation of reality with a model that can be manipulated” (Sawilowsky & Fahoome, p. 46). The accuracy of a simulation improves with the accuracy of its constituent parameters. Simulations serve a variety of purposes across many fields of study as they enable one to approximate reality in order to predict potential outcomes. Such predictions can inform anything from where to build a factory to the probable location of a subatomic particle.

A Monte Carlo simulation can be defined as “the use of a computer program to simulate some aspect of reality (to make) determinations of the nature of reality or change in reality through the repeated sampling via Monte Carlo methods” (Sawilowsky & Fahoome, 2003, p. 46). In statistics, Monte Carlo simulations are often used to

Table 10

Type I Error Rates for Independent-Samples t Test for Various Sample Sizes and Alpha Levels When Sampling Is From a Digit Preference (Achievement) Distribution, 10,000 Repetitions

	α = .050			α = .010		
Sample Size	U025	L025	Total	U005	L005	Total
5, 15	0.023	**0.028**	0.051	**0.004**	**0.007**	0.011
10, 10	0.025	0.025	0.050	**0.004**	**0.004**	**0.008**
10,30	0.026	0.024	0.050	**0.006**	**0.006**	**0.012**
20,20	0.025	0.025	0.050	**0.006**	**0.006**	**0.012**
15,45	**0.021**	0.024	0.045	**0.004**	0.005	0.009
30,30	0.027	0.026	0.053	**0.006**	0.005	0.011
20,60	0.024	0.026	0.050	0.005	0.005	0.010
40,40	0.025	0.025	0.050	**0.006**	**0.006**	0.011
30,90	0.025	0.025	0.050	**0.004**	0.005	0.009
60,60	0.024	**0.029**	0.053	**0.003**	0.005	**0.008**

Note. Adapted from Sawilowsky & Blair (1992). Sample size = n1, n2. U025, U005 = proportion of rejections in the upper-tail. L025, L005 = proportion of rejections in the lower tail.

Outside of Bradley's (1978) stringent definition of type I robustness.

Outside of Bradley's (1978) liberal definition of type I robustness.

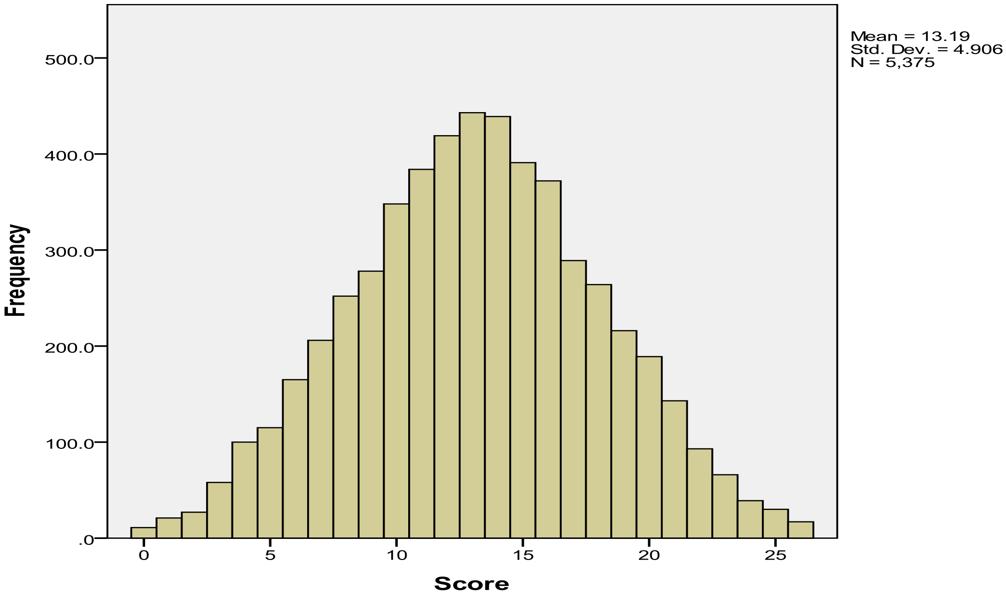


Figure 9: Smooth Symmetric. Adapted from Sawilowsky & Blair, 1992 (p. 358).

evaluate the robustness and/or power of a statistical test under certain conditions (i.e. assumption violations). The purpose of such studies are to gain knowledge into what tests work best under certain conditions so that researchers from all disciplines benefit in their ability to make discoveries and/or avoid false positives.

Resultant Topic

Instead of adjusting the degrees of freedom to compare an approximate critical value to *t* obtained, one could instead just modify the critical value. Based on this line of thinking, Farrell-Singleton (2010) developed the table of Winsorized *t* values. This study aims to evaluate these methods to determine which is more robust to type I and II errors across sample sizes, Winsorized amounts, various distributions, alpha levels (.01 and .05), and effect sizes through Monte Carlo simulations conducted via Fortran. In addition, Micceri's real data sets will be used to benefit the generalization of results to Educational and Psychological data sets.

Table 11

Type I Error Rates for Independent-Samples t Test for Various Sample Sizes and Alpha Levels When Sampling Is From a Smooth Symmetric (Achievement) Distribution, 10,000 Repetitions

	α = .050			α = .010		
Sample Size	U025	L025	Total	U005	L005	Total
5, 15	0.025	0.024	0.049	**0.004**	0.005	0.009
10, 10	0.026	0.027	0.053	**0.006**	0.005	0.011
10,30	0.026	0.023	0.049	**0.006**	0.005	0.011
20,20	0.025	0.025	0.050	0.005	0.005	0.010
15,45	0.025	0.025	0.050	0.005	0.005	0.010
30,30	0.024	0.024	0.048	0.005	**0.004**	0.009
20,60	0.026	**0.029**	0.055	**0.004**	**0.006**	0.010
40,40	0.023	**0.022**	0.045	0.005	0.005	0.010
30,90	0.023	0.023	0.046	**0.004**	0.005	0.009
60,60	0.026	0.023	0.049	0.005	**0.004**	0.009

Note. Adapted from Sawilowsky & Blair (1992). Sample size = n1, n2. U025, U005 = proportion of rejections in the upper-tail. L025, L005 = proportion of rejections in the lower tail.

Outside of Bradley's (1978) stringent definition of type I robustness.

Outside of Bradley's (1978) liberal definition of type I robustness.

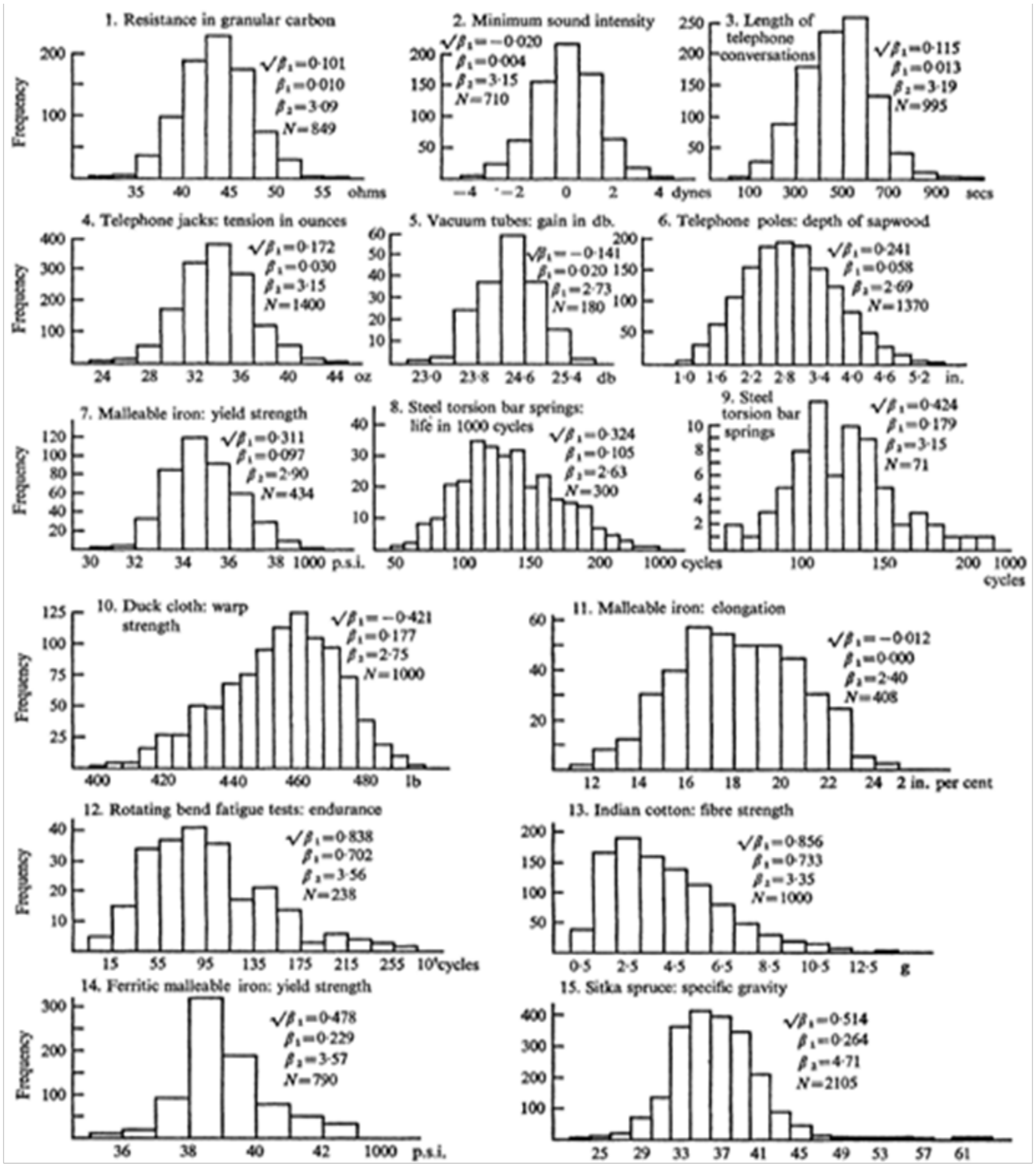


Figure 10. Histogram distributions of some industrial data. Reprinted from “Relation between the shape of population distribution and the robustness of four simple test statistics,” by E. S. Pearson and N. W. Please, 1975, *Biometrika*, 62, p. 22.

Table 12

Obtained Percentages of Cases Falling Outside the Appropriate Tabled t Values for the 5% and 1% Levels of Significance

		Obtained Percentage at	
Sample Size	Distribution	5% level	1% level
5, 5	Normal	5.3	0.9
15, 15	Normal	4	0.8
5, 15	Normal	4	0.6
5, 5	Exponential	3.1	0.3
15, 15	Exponential	4	0.4
5, 5	Rectangular	5.1	1
15, 15	Rectangular	5	1.5

Note: See Boneau (1960)

CHAPTER III

METHODOLOGY

Purpose:

A Monte Carlo study will be used to evaluate the robustness properties of the Winsorized *t*-test using a newly developed table of critical values (Farrell-Singleton, 2010) to verify the robustness of the two independent samples *t*-test for $\alpha = .05$ and .01 for a selection of mathematical distributions and real data sets, for various sample sizes, amounts of Winsorized values, effect sizes, and distribution shapes. The results will be compared with the same set of parameters of the *t* with outliers as well as with the Winsorized *t*-obtained where the formula (instead of the critical value) with adjusted degrees of freedom is used to accommodate the outliers. When used for two samples, the Winsorized *t* table assumes symmetrically-Winsorized samples due to the presence of outliers (in equal amounts) at both tails of each sample. Therefore, every simulation across all conditions will account for this assumption.

The four sets of simulations being compared are:

1. Ordinary *t*-test with no outliers present in each sample.
2. Ordinary *t*-test with outliers present in each sample (equally per end).
3. Winsorized *t*-test with regular (($h_1 + h_2$) – 2 degrees of freedom) *t* critical values (outliers Winsorized).
4. Winsorized *t*-test with Winsorized *t* critical values (outliers will be Winsorized).

For each study, the number of outliers and Winsorized amounts will be equal across samples per iteration. This will also apply to unbalanced samples. For example, if

2 values are to be Winsorized per end for n =10, when n = 5, only one original value will be left. This is practiced due to the nature of the table of Winsorized critical values since it gives the degrees of freedom and Winsorized amount.

The magnitude of generated outliers will be equal per sample and based on the lowest value not transformed to an outlier. Individual combinations of parameters as well as overall sets of simulations will be compared so that a detailed picture of the results may emerge in addition to overall trends. Alpha levels of .01 and .05 were chosen to reflect those most common in applied research. Distributions will include normal, those from Micceri (1989) (to generalize results to real world data), Cauchy, t (3 df), Chi-Squared (2 df), and Exponential and Uniform for comparison with Boneau (1960). Effect sizes of 0.2, 0.5, 0.8, and 1.2 will be added for the type II error portion. The normal distribution will serve as a control since the table of *t* and Winsorized *t* critical values are based on normality.

Simulation of Outliers:

The terms “one-out” and “one-wild” have been used by several researchers to indicate an obvious outlier, though exact values vary in usage. Examples of the how the "one-wild" value has been generated include: any value randomly drawn from a population of N (0, 100) (Lax, 1985), of N (10, 0) (Carey et al., 1997), or a number drawn from a normal distribution that is multiplied by ten (Wilcox, 1998). The first two above examples involve using an "outlier-generating model" in which outliers are drawn from a separate theoretical population (Davies & Gather, 1993). This is actually simulating contamination, which is a unique situation whereby a separate mechanism generates outliers due to the presence of a separate distribution (Hawkins, 1980).

Hawkins recommends trimming in such instances and Winsorizing when outliers occur as a result of a heavy-tailed distribution (though Fung and Rahman's (1980) results suggest the Winsorized *t* to be nearly identical to the trimmed *t* with respect to robustness under contamination). To simulate such a distribution in this study, and methodologically more in line with Wilcox (1998), outliers will instead be generated as follows:

1. Each sample will first be drawn and sorted.
2. Low values (to be transformed to outliers) will be replaced by the next highest value minus ten times itself (m).
3. High values (to be transformed to outliers) will be replaced by the next lowest value plus ten times itself (y).
4. The amount of outliers generated will be equal to the amount of values being Winsorized (in equal amounts per end) for that comparison per α level, sample size, and distribution.

Though there were instances where Sawilowsky and Blair found the *t*-test to be non-robust to type I error, it was for the most part robust by Bradley's (1978) definitions. The distribution with the most obvious outliers (Discrete Mass at Zero with Gap) produced the most non-robust results in Sawilowsky and Blair (1992), so it follows that adding outliers in general will produce less-robust results. However, if outliers are added, robust alternatives may be needed, as robustness in the presence of added (simulated) outliers will be examined. Such alternatives include the trimmed (Yuen & Dixon, 1973) and Winsorized (Fung & Rahman, 1980) *t*-test for independent samples.

To be sure, simulating outliers will introduce data points that are not part of the parent populations (in magnitude and/or frequency). Also, standard deviations used as multipliers for effect size are based on the original population distributions, which often do not have outliers. These practices are common in articles cited in the above literature review (for examples, see Lax, 1985 and Carey et al., 1997). Computationally, this reinforces the status of the outliers as not from the parent distributions. To test how a Winsorized *t*-test helps in this situation seems to go against the above literature review since it was noted that Winsorizing is more appropriate for when outliers represent the parent population distribution. However, Fung and Rahman (1980) found the two-sample Winsorized *t* to be as robust as the two-sample trimmed *t* under contaminated and long-tailed distributions. Since Winsorized distributions will be compared to real parent distributions with no outliers, if the parent distributions have outliers themselves then this will be accounted for (i.e. parent distribution outliers vs. parent plus simulated outliers will be examined).

Though some of the real distributions to be used in this study already have outliers, simulating additional outliers will serve to exaggerate them to accentuate their presence. In the case of discrete populations (such as with Extreme Bimodality), the frequency of highest and lowest values will be increased. In the case of more continuous distributions, such as the mixed-normal, outliers will likely increase in frequency, and definitely in magnitude.

Procedures:

Table 13 shows the parameters to be used for each type of distribution. Note that the regular critical values are the same for any amount of outliers per degrees of freedom. The critical values (regular, Winsorized, or adjusted) will be used with the regular *t*-test (traditional *t*-obtained formula). For each iteration (of 1,000,000 total) per simulation (each line or set of conditions in table 13 represents one simulation), samples of specified sizes will first be drawn (with replacement) from a specified population distribution. Next, samples will be sorted and either Winsorized, given outliers, or neither of these, depending on the set. After this, a *t*-test (or Winsorized *t*-test) will be conducted on the samples and the result will either be rejected or accepted, based on the critical value being used. If the result is rejected, it will be added to a running tally. When this is done 1,000,000 times, the portion of rejected results out of 1,000,000 will be reported for upper and lower-tails. If the critical value comes from the .05 column, then a result of .025 will be expected for each tail. If the value differs from expected, then the test, under these parameters, Bradley's (1978) definition of robustness will be applied for type I error.

Fortran 90 programming with Compaq Visual Fortran 6.6 will be used to conduct all simulations. The Rangen 2.0 subroutines (Fahoome, 2002), a Fortran 90/95 update from the original Fortran 77 version (Blair, 1987) will provide random numbers and normal and mathematical distributions. The adapted/modified Realpops subroutines (Sawilowsky & Fahoome, 2003) will be used to provide real data sets from Micceri (1989). For detailed information on how the Monte Carlo-derived critical values were generated, see Farrell-Singleton (2010).

Table 13

Critical Values for the two sample t and Winsorized t

x(n)	y(n)	Outliers	df	0.01	0.01 ((h1+h2)-2 df)	.01 (Winsorized)	0.05	0.05 ((h1+h2)-2) df)	.05 (Winsorized)
5	5	1	8	3.35539	4.60410	9.38220	2.30600	2.77645	5.42160
5	15	1	18	2.87844	2.97684	3.80900	2.10092	2.14479	2.71970
5	15	2	18	2.87844	3.16927	5.71300	2.10092	2.22814	3.92500
10	10	1	18	2.87844	2.97684	3.80900	2.10092	2.14479	2.71970
10	10	2	18	2.87844	3.16927	5.71300	2.10092	2.22814	3.92500
15	15	1	28	2.76326	2.79694	3.24400	2.04841	2.06390	2.39120
15	15	3	28	2.76326	2.92078	5.07470	2.04841	2.11991	3.63510
10	30	1	38	2.71156	2.72839	3.03550	2.02439	2.03225	2.25460
10	30	4	38	2.71156	2.81876	4.82200	2.02439	2.07387	3.50560
20	20	1	38	2.71156	2.72839	3.03550	2.02439	2.03225	2.25460
20	20	4	38	2.71156	2.81876	4.82200	2.02439	2.07387	3.50560
25	25	1	48	2.68220	2.69228	2.93060	2.01064	2.01537	2.18640
25	25	5	48	2.68220	2.76326	4.68490	2.01064	2.04841	3.44150
15	45	1	58	2.66329	2.66999	2.85050	2.00172	2.00488	2.14210
15	45	6	58	2.66329	2.72839	4.58780	2.00172	2.03225	3.39970
30	30	1	58	2.66329	2.66999	2.85050	2.00172	2.00488	2.14210
30	30	6	58	2.66329	2.72839	4.58780	2.00172	2.03225	3.39970
45	45	1	88	2.63286	2.63563	2.75580	1.98729	1.98861	2.07820
45	45	9	88	2.63286	2.67373	4.46050	1.98729	2.00665	3.32910
30	90	1	118	2.61814	2.61965	2.70790	1.98027	1.98099	2.04740
30	90	12	118	2.61814	2.64791	4.38790	1.98027	1.99444	3.29510
60	60	1	118	2.61814	2.61965	2.70790	1.98027	1.98099	2.04740
60	60	12	118	2.61814	2.64791	4.38790	1.98027	1.99444	3.29510
90	90	1	178	2.60373	2.60438	2.66620	1.97338	1.97369	2.01690
90	90	18	178	2.60373	2.62301	4.31360	1.97338	1.98260	3.25730
120	120	1	238	2.59664	2.59700	2.63670	1.96998	1.97015	2.00340
120	120	24	238	2.59664	2.61090	4.28760	1.96998	1.97681	3.24800

CHAPTER IV

RESULTS

All descriptions of robustness will refer to direction (conservative or liberal) and magnitude (liberal or stringent) according to Bradley's (1978) definitional ranges. To describe non-robust results that fell outside of Bradley's liberal range, the phrase "outside of the liberal range" will be used.

Type I Error: Distributions with No Outliers:

With the exception of unbalanced samples of 5 and 15 (where upper tail results were liberal instead of conservative (.0355 and.0124 instead of .003 and .001)), the results from this study echo those from Sawilowsky & Blair (1992) for Micceri's (1989) real data sets. The mathematical distributions in this study were not used in Sawilowsky & Blair and results are summarized below.

Error rates from the normal distribution matched their alpha (and 0.5 alpha for one-tailed) levels. This served as a "control" (in addition to comparing with results from Sawilowsky & Blair, 1992) to ensure that resulting data from the Fortran 90 program are accurate.

The Uniform distribution produced all stringently-robust results save for at the .01 alpha level where they were mostly stringent save for smaller samples which tended to be liberal within liberal range. For the Exponential distribution (two-tailed at the .05 alpha level), smaller samples tended to be conservative (up to 10, 10) in liberal range. Upper tail results tended to be liberal in the liberal range for unbalanced samples yet conservative in the liberal or stringent range for balanced samples. Lower tail results tended to be conservative (either in the liberal or stringent range) except for samples of 5 and 15, which were conservative non-liberal. At the .01 alpha level, results were mostly stringent save for smaller samples which tended to be liberal within liberal range.

At the .05 alpha level, smaller samples tended to be conservative (up to 10, 10) within the liberal range for the Exponential distribution. Upper tail results tended to be liberal outside of the liberal range for unbalanced samples yet conservative in the liberal or stringent range for balanced samples. Lower tail results were conservative outside of the liberal range for unbalanced samples and within the range for balanced samples. For both upper and lower tail results, balanced samples of 118 degrees of freedom and higher were robust by the stringent definition. At the .01 alpha level, up to 118 degrees of freedom for balanced samples, two-tailed results tended to be conservative within the stringent or liberal range. Upper tail results tended to be liberal outside of the liberal range for unbalanced samples yet conservative in the liberal or stringent range for balanced samples. Lower tail results were to be conservative (non-robust) outside of the liberal range for unbalanced samples and within the range for balanced samples. For both upper and lower tail results, balanced samples of 118 degrees of freedom and higher were robust by the stringent definition.

For the *t* distribution with three degrees of freedom, at the .05 alpha level the two-tailed results tended to be conservative within the liberal range (except for samples 5 and 15, which were within the stringent range) up to 38 degrees of freedom, where results were stringently robust. Both lower and upper tail results reflected the same trends as those for two-tailed results. At the .01 alpha level, all results tended to be conservative within the liberal range with a few exceptions for unbalanced samples (which were actually within the stringent range of robustness).

For the Chi-Squared distribution with two degrees of freedom (.05 alpha level), two-tailed results were conservative (within the liberal range) up to 28 degrees of freedom, after which results were within the stringent range of robustness. For one-tailed results, balance sample results are within the conservative within the liberal range through samples of 10 and 10. After that point, they tended to be stringently robust. Unbalanced samples tended to be liberal (within the liberal range) for upper tails yet conservative (within the liberal range) for lower tails. At the .01 alpha level, one and two-tailed balanced sample results were conservative (within the liberal range) up to 118 degrees of freedom, after which results were within the stringent range of robustness. Unbalanced samples (two-tailed) were all robust.

The Cauchy distribution (at the .05 alpha level) produced balanced sample results that were outside of the liberal range of conservative, yet unbalanced samples were conservative within the liberal range. At the .01 alpha level, the only difference was that unbalanced samples 15 and 45 and 30 and 90 were also conservative outside of the liberal range.

Type I Error: Simulated Outliers:

Simulating outliers in discrete distributions posed unique challenges due to their bounded nature. With a discrete mass at zero, for example, Winsorized data points (for that tail) can look exactly like outliers. The main differences between distributions with outliers vs. Winsorized values are in the longer tail as well as the variance, since outlier-simulated discrete distributions have values from the center recoded to the extremes.

For Micceri’s (1989) eight real data sets, simulating one outlier per end tended to make results more conservative, especially for smaller samples, since outliers inflate the variance and reduce the probability of a significant *t*-obtained value. With symmetrical Mathematical distributions (normal, *t* with three degrees of freedom, and Cauchy), all results with 20% outliers were conservative, non-robust. For one outlier per end, most small sample results were the same. Winsorizing 10% per end generally exaggerated the effect of adding one per end. Winsorizing in these cases generally brought p-values within stringent or liberal ranges of robustness except for large samples of the Cauchy distribution, which tended to be liberal, non-robust when Winsorized. For the Exponential distribution and Chi-Squared distributions, adding one outlier per end made results generally more liberal for this distribution, which was unique yet not unexpected since these distributions are in themselves uniquely skewed.

Type I Error: Using Adjusted Critical Values:

Using the critical values based on adjusted degrees of freedom generally produced liberal, non-robust results save for some instances with large sample sizes. There were few instances where using the adjusted critical values produced more robust results than using the Winsorized critical values. This was due to the fact that, though both approaches served to allow more rejection of the null, if the Winsorized critical value approach still produced conservative results (due to the unique properties of the distribution), the adjusted critical value approach produced more liberal results that happened to be closer to the nominal alpha level.

Type I Error: Using Winsorized Critical Values:

Generally, Winsorizing samples led to less conservative (and more robust) p-values. Using the Winsorized critical values produced results that were consistently (with few exceptions) more robust than using the adjusted critical values. Results from the Uniform distribution were mixed while the Exponential and Chi-Squared distributions became more conservative. However, for all distributions, Winsorized results were generally more robust than their regular critical value (with outliers) counterparts.

There were some rare exceptions where samples with simulated outliers produced more robust results due to interactions of small sample size, unequal samples, and skew. Examples came mainly from the following distributions: Discrete Mass at Zero with Gap (for both .01 and .05 alpha levels), Extreme Asymmetry (psychometric/decay, .01 alpha), and Extreme Bimodality (.01 alpha).

In tables 14 and 15, the Discrete Mass at Zero with Gap distribution shows how skew combined with increased Winsorizing serve to hinder type I robustness by using the Winsorized critical values. In the case of Discrete Mass at Zero with Gap, samples of 30 or greater seem more robust if simply keeps the outliers and using the old critical value. However, results were generally better when Winsorizing and using the Winsorized critical values. Table 16 shows just how vulnerable p-values are for the normal distribution with just one outlier per end.

Type II Error: Distributions With No Outliers:

Table 17 shows that in general, an increase in effect size led to larger portions of data points that fall into the upper tails and less for the lower tails since the greater the effect, the greater the shift in mean (or distribution) that should occur. When looking at the results, it is quite noticeable that the Uniform distribution has the highest rejection (of

Table 14

Type I Error Rates for Independent-Samples t Test for Various Sample Sizes and Alpha Levels When Sampling Is From a Discrete Mass at Zero With Gap (Psychometric) Dist., 1,000,000 Repetitions, 1 outlier/end Winsorized, Winsorized C.V.

	α = .050			α = .010		
Sample Size	**U025**	**L025**	**Total**	**U005**	**L005**	**Total**
5, 5	⇩ 0.0057	⇩ 0.0055	⇩ 0.0112	↘ 0.0148	↘ 0.0147	↘ 0.0295
5, 15	⇧ *0.0640*	⇩ *0.0001*	↗ *0.0641*	⇧ *0.0390*	⇩ *0.0003*	↘ *0.0393*
10, 10	⇩ 0.0106	⇩ 0.0105	⇩ 0.0211	↘ 0.0156	↘ 0.0155	↘ 0.0312
15, 15	⇨ 0.0244	⇨ 0.0244	⇨ 0.0488	⇨ 0.0259	⇨ 0.0259	⇨ 0.0518
10, 30	↗ *0.0289*	⇩ *0.0063*	↘ *0.0353*	↗ *0.0281*	⇩ *0.0124*	↘ *0.0405*
20, 20	↗ 0.0280	↗ 0.0280	↗ 0.0560	↗ 0.0281	↗ 0.0281	↗ 0.0562
25, 25	⇨ 0.0256	⇨ 0.0254	⇨ 0.0510	↗ 0.0275	↗ 0.0280	↗ 0.0555
15, 45	⇨ *0.0256*	↗ *0.0352*	↗ *0.0608*	⇨ *0.0267*	⇨ *0.0250*	⇨ *0.0517*
30, 30	⇨ 0.0229	⇨ 0.0228	⇨ 0.0457	⇨ 0.0271	⇨ 0.0270	⇨ 0.0541
45, 45	↘ 0.0218	↘ 0.0220	↘ 0.0438	⇨ 0.0248	⇨ 0.0250	⇨ 0.0498
30, 90	⇨ *0.0256*	↘ *0.0211*	⇨ *0.0467*	⇨ *0.0260*	⇨ *0.0261*	⇨ *0.0521*
60, 60	↘ 0.0224	⇨ 0.0226	⇨ 0.0450	⇨ 0.0238	⇨ 0.0238	⇨ 0.0476
90, 90	⇨ 0.0237	⇨ 0.0236	⇨ 0.0474	⇨ 0.0232	⇨ 0.0233	⇨ 0.0465
120, 120	⇨ 0.0239	⇨ 0.0239	⇨ 0.0478	⇨ 0.0233	⇨ 0.0235	⇨ 0.0468

Note. SU025, U005 = proportion of rejections in the upper- tail. L025, L005 = proportion of rejections in the lower-tail.

Based on Bradley's (1978) definitions of type I robustness:

⇧ Liberal, Outside the Liberal Range	↗ Liberal, Inside the Liberal Range
⇨ Within the Stringent Range	↘ Conservative, Inside the Liberal Range
⇩ Conservative, Outside the Liberal Range	

Table 15

Type I Error Rates for Independent-Samples t Test for Various Sample Sizes and Alpha Levels When Sampling Is From a Discrete Mass at Zero With Gap (Psychometric) Dist., 1,000,000 Repetitions, 20% Winsorized Outliers, Winsorized C.V.

	α = .050			α = .010		
Sample Size	**U025**	**L025**	**Total**	**U005**	**L005**	**Total**
5, 5	n/a	n/a	n/a	n/a	n/a	n/a
5, 15	↑ ***0.0516***	↓ ***0.0000***	→ ***0.0516***	↑ ***0.0340***	↓ ***0.0000***	↑ ***0.0340***
10, 10	↓ **0.0006**	↓ **0.0006**	↓ **0.0012**	↓ **0.0001**	↓ **0.0000**	↓ **0.0001**
15, 15	↓ 0.0025	↓ 0.0026	↓ 0.0051	↓ 0.0000	↓ 0.0000	↓ 0.0001
10, 30	↘ ***0.0145***	↓ ***0.0000***	↓ ***0.0145***	↑ ***0.0132***	↓ ***0.0000***	↗ ***0.0132***
20, 20	↘ **0.0176**	↘ **0.0176**	↘ **0.0351**	↓ **0.0001**	↓ **0.0001**	↓ **0.0001**
25, 25	→ 0.0261	→ 0.0263	→ 0.0524	↓ 0.0007	↓ 0.0007	↓ 0.0013
15, 45	↓ ***0.0045***	↓ ***0.0000***	↓ ***0.0045***	↘ ***0.0041***	↓ ***0.0000***	↓ ***0.0041***
30, 30	↑ **0.0702**	↑ **0.0700**	↑ **0.1401**	↓ **0.0014**	↓ **0.0014**	↓ **0.0029**
45, 45	↑ 0.2463	↑ 0.2464	↑ 0.4927	↑ 0.0139	↑ 0.0144	↑ 0.0283
30, 90	↓ ***0.0005***	↑ ***0.0384***	↘ ***0.0390***	↓ ***0.0003***	↓ ***0.0000***	↓ ***0.0003***
60, 60	↑ **0.2540**	↑ **0.2539**	↑ **0.5078**	↑ **0.1043**	↑ **0.1036**	↑ **0.2079**
90, 90	↑ 0.2610	↑ 0.2616	↑ 0.5227	↑ 0.2546	↑ 0.2548	↑ 0.5094
120, 120	↑ 0.2648	↑ 0.2646	↑ 0.5293	↑ 0.2629	↑ 0.2621	↑ 0.5250

Note. SU025, U005 = proportion of rejections in the upper- tail. L025, L005 = proportion of rejections in the lower-tail.

Based on Bradley's (1978) definitions of type I robustness:

↑ Liberal, Outside the Liberal Range

↗ Liberal, Inside the Liberal Range

→ Within the Stringent Range

↘ Conservative, Inside the Liberal Range

↓ Conservative, Outside the Liberal Range

Table 16

Type I Error Rates for Independent-Samples t Test for Various Sample Sizes and Alpha Levels When Sampling Is From a Normal Dist., 1,000,000 repetitions, 1 outlier vs. 1 Winsorized value (both per end).

	α = .050 (1 outlier/end)			α = .050 (1 Winsorized/end)		
Sample Size	**U025**	**L025**	**Total**	**U025**	**L025**	**Total**
5, 5	⇩ 0.0000	⇩ 0.0000	⇩ 0.0000	⇨ 0.0253	⇨ 0.0252	⇨ 0.0506
5, 15	⇩ ***0.0000***	⇩ ***0.0000***	⇩ ***0.0000***	⇨ ***0.0266***	⇨ ***0.0266***	⇨ ***0.0532***
10, 10	⇩ **0.0000**	⇩ **0.0000**	⇩ **0.0000**	⇨ **0.0249**	⇨ **0.0250**	⇨ **0.0499**
15, 15	⇩ 0.0000	⇩ 0.0000	⇩ 0.0000	⇨ 0.0249	⇨ 0.0250	⇨ 0.0499
10, 30	⇩ ***0.0000***	⇩ ***0.0000***	⇩ ***0.0000***	⇨ ***0.0257***	⇨ ***0.0259***	⇨ ***0.0516***
20, 20	⇩ **0.0000**	⇩ **0.0000**	⇩ **0.0000**	⇨ **0.0252**	⇨ **0.0251**	⇨ **0.0504**
25, 25	⇩ 0.0000	⇩ 0.0000	⇩ 0.0000	⇨ 0.0251	⇨ 0.0250	⇨ 0.0501
15, 45	⇩ ***0.0000***	⇩ ***0.0000***	⇩ ***0.0000***	⇨ ***0.0256***	⇨ ***0.0257***	⇨ ***0.0513***
30, 30	⇩ **0.0000**	⇩ **0.0000**	⇩ **0.0000**	⇨ **0.0253**	⇨ **0.0255**	⇨ **0.0508**
45, 45	⇩ 0.0000	⇩ 0.0000	⇩ 0.0000	⇨ 0.0249	⇨ 0.0249	⇨ 0.0498
30, 90	⇩ ***0.0000***	⇩ ***0.0000***	⇩ ***0.0000***	⇨ ***0.0251***	⇨ ***0.0250***	⇨ ***0.0501***
60, 60	⇩ **0.0000**	⇩ **0.0000**	⇩ **0.0000**	⇨ **0.0249**	⇨ **0.0250**	⇨ **0.0499**
90, 90	⇩ 0.0000	⇩ 0.0000	⇩ 0.0000	⇨ 0.0252	⇨ 0.0249	⇨ 0.0501
120, 120	⇩ 0.0000	⇩ 0.0000	⇩ 0.0000	⇨ 0.0247	⇨ 0.0249	⇨ 0.0496

Note. SU025 = proportion of rejections in the upper- tail. L025 = proportion of rejections in the lower-tail.

Based on Bradley's (1978) definitions of type I robustness:

⇧ Liberal, Outside the Liberal Range ⇗ Liberal, Inside the Liberal Range

⇨ Within the Stringent Range ⇘ Conservative, Inside the Liberal Range

⇩ Conservative, Outside the Liberal Range

Table 17

Rejection Rates for Independent-Samples t Test (No Outliers) for Various Sample Sizes, Effect Sizes, and Distributions (1.000.000 Repetitions).

Sample Size	**Discrete Mass at Zero w/ Gap**	**Mass at Zero**	**Extreme Asymmetry (P/D)**	**Extreme Asymmetry (A/G)**	**Extreme Bimodality**	**Multimodal/Lumpy**	**Digit Preference**	**Smooth Symmetric**	**Normal**	**Uniform**	**Exponential**	**t w/ 3df**	**Chi-squared w/2df**	**Cauchy**
n_1, n_2	**E.S. = 0.5**		**α = .05**		**No Outliers or Winsorized Values**									
5, 5	0.0094	0.1090	0.1657	0.1261	0.0993	0.1028	0.1075	0.1070	0.1078	0.4780	0.1359	0.1378	0.1363	0.0733
5, 15	***0.2208***	***0.1515***	***0.1740***	***0.1657***	***0.1417***	***0.1434***	***0.1498***	***0.1497***	***0.1506***	***0.7431***	***0.1839***	***0.1898***	***0.1831***	***0.1422***
10, 10	**0.1142**	**0.1887**	**0.2171**	**0.2020**	**0.1721**	**0.1749**	**0.1833**	**0.1829**	**0.1860**	**0.8639**	**0.2253**	**0.2389**	**0.2255**	**0.2000**
15, 15	0.2605	0.2657	0.2826	0.2757	0.2486	0.2514	0.2608	0.2603	0.2626	0.9728	0.3033	0.3234	0.3033	0.2609
10, 30	***0.2945***	***0.2682***	***0.2683***	***0.2866***	***0.2586***	***0.2578***	***0.2651***	***0.2650***	***0.2663***	***0.9742***	***0.2880***	***0.3164***	***0.2876***	***0.1902***
20, 20	**0.3453**	**0.3405**	**0.3518**	**0.3481**	**0.3270**	**0.3287**	**0.3367**	**0.3370**	**0.3382**	**0.9953**	**0.3755**	**0.3999**	**0.3740**	**0.3081**
25, 25	0.4154	0.4122	0.4200	0.4180	0.4002	0.4021	0.4091	0.4081	0.4104	0.9992	0.4417	0.4690	0.4415	0.3546
15, 45	***0.3841***	***0.3809***	***0.3739***	***0.3965***	***0.3723***	***0.3708***	***0.3774***	***0.3772***	***0.3779***	***0.9982***	***0.3897***	***0.4250***	***0.3892***	***0.2339***
30, 30	**0.4830**	**0.4804**	**0.4851**	**0.4840**	**0.4712**	**0.4721**	**0.4768**	**0.4770**	**0.4775**	**0.9999**	**0.5045**	**0.5301**	**0.5045**	**0.4022**
45, 45	0.6528	0.6518	0.6524	0.6530	0.6485	0.6473	0.6505	0.6497	0.6506	1.0000	0.6630	0.6793	0.6624	0.5610
30, 90	***0.6467***	***0.6543***	***0.6458***	***0.6620***	***0.6522***	***0.6490***	***0.6518***	***0.6523***	***0.6526***	***1.0000***	***0.6512***	***0.6737***	***0.6519***	***0.3897***
60, 60	**0.7769**	**0.7772**	**0.7749**	**0.7762**	**0.7764**	**0.7744**	**0.7750**	**0.7754**	**0.7752**	**1.0000**	**0.7781**	**0.7821**	**0.7781**	**0.6399**
90, 90	0.9156	0.9158	0.9146	0.9152	0.9179	0.9162	0.9161	0.9164	0.9154	1.0000	0.9114	0.9014	0.9121	0.7181
120, 120	0.9705	0.9710	0.9702	0.9704	0.9726	0.9717	0.9713	0.9713	0.9708	1.0000	0.9680	0.9543	0.9679	0.7599

Note. Darker shading and longer bars indicate higher rejection rates.

the null) rate and the Cauchy distribution has the lowest. For balanced, small sample sizes, the Discrete Mass at Zero with Gap distribution also had noticeably lower rejection rates. The rest of the distributions tend to be relatively close in rejection rates. Rejection rates increased as a function of effect size, alpha level, and sample size.

Type II Error: Simulated Outliers:

As noted by Zimmerman and Zumbo (1993), outliers can reduce the power of the *t* test. Table 18 shows that the results of this study generally support this observation except for the Uniform distribution, which was largely immune (in terms of type II error) to the effect of simulated outliers. Scenarios with parameters associated with the .05 alpha level produced higher rejection rates, as expected.

In addition to the Uniform distribution, the Exponential and Chi-Squared with two degrees of freedom distributions showed higher rejection rates than all other distributions when outliers were added, even more than when there were no simulated outliers.

When twenty percent of the values of each end are recoded to outliers, a greater variety of rejection rates emerged. The *t*, Cauchy, and normal distributions were the most negatively impacted by outliers (for one per end more so than 10% per end), even with greater effect sizes. This may be attributable to outliers being added at both ends which contained negative and positive values, thus inflating the variance even more so when taking the sum of squares (since negative values are positive when squared) of absolute values.

Also negatively affected were the Mass at Zero, Extreme Asymmetry (P/D), Digit Preference, and Smooth Symmetric distributions. The common results between these distributions were that even with greater effect sizes, they were still negatively impacted by outliers. Sample size was also a mediator, but less-so for the above distributions.

Positive skew, in combination with a high frequency of lower extreme values, seems to have positively impacted the ability of effect size and sample size to mediate the effect of outliers. Distributions with such unique properties tended reject more when effect size and sample size increased. These distributions include Discrete Mass at Zero With Gap, Extreme Asymmetry (P/D), Extreme Bimodality, Multimodal/Lumpy, Exponential, and Chi-Squared. The Uniform distribution was also less impacted by outliers, but this distribution is also unique in other ways described above and has no skew.

Type II Error: Simulated Outliers:

Unequal sample sizes had mixed effects on results. The Extreme Asymmetry (A/G) distribution showed a higher rejection rate for unequal samples, yet the Extreme Asymmetry (P/D) distribution showed the opposite effect. The trend seems to show that for distributions with high concentrations of values on the upper tail, unequal samples are a benefit to rejection rates and for those with concentrations on the lower tail, the opposite is true. This may be due to the fact that in the simulations, the effects were added to the larger of the unequal samples.

Extreme Bimodality and Multimodal/Lumpy distributions also shared another general trend; their rejection rates benefitted less from increased Winsorized amounts. Discrete Mass at Zero with Gap and Extreme Asymmetry (P/D) showed the greatest benefit from increased Winsorizing, yet they also showed (especially for larger samples) inflated type I error rates.

For 20% Winsorized samples, an interaction between skew and unequal sample sizes was apparent for skewed distributions (Discrete Mass at Zero with Gap, Extreme Asymmetry (P/D), Exponential, and Chi-Squared with 3 degrees of freedom). This was also partially true for Mass at Zero, but to a lesser degree. The interaction reduced rejection rates for the same degrees of freedom with balanced samples. The opposite effect as a result of the same interaction can be seen for Extreme Asymmetry (A/G) and Extreme Bimodality.

Table 18

Rejection Rates for Independent-Samples t Test (with Outliers) for Various Sample Sizes, Effect Sizes, Outlier Amounts, and Distributions (1,000,000 Repetitions)

Sample Size	Discrete Mass at Zero w/ Gap	Mass at Zero	Extreme Asymmetry (P/D)	Extreme Asymmetry (A/G)	Extreme Bimodality	Multimodal/Lumpy	Digit Preference	Smooth Symmetric	Normal	Uniform	Exponential	t w/ 3df	Chi-squared w/2df	Cauchy
n_1, n_2	**E.S. = 0.5**		**α = .05**		**Outlier Amount = 1/end**									
5, 5	0.0060	0.0418	0.0964	0.0724	0.0406	0.0659	0.0552	0.0563	0.0000	0.4502	0.3161	0.0000	0.3859	0.0000
5, 15	***0.1810***	***0.0688***	***0.1287***	***0.1001***	***0.0812***	***0.1190***	***0.0883***	***0.0929***	***0.0001***	***0.7203***	***0.2420***	***0.0003***	***0.1996***	***0.0348***
10, 10	**0.0815**	**0.0423**	**0.1200**	**0.0978**	**0.1252**	**0.1170**	**0.0658**	**0.0700**	**0.0000**	**0.8474**	**0.4378**	**0.0001**	**0.4422**	**0.0172**
15, 15	0.1719	0.0725	0.1879	0.1260	0.2134	0.2058	0.1048	0.1189	0.0000	0.9698	0.5042	0.0001	0.5039	0.0280
10, 30	***0.2498***	***0.1297***	***0.2720***	***0.1355***	***0.2004***	***0.2274***	***0.1525***	***0.1699***	***0.0003***	***0.9713***	***0.3267***	***0.0011***	***0.2883***	***0.0487***
20, 20	**0.2418**	**0.1493**	**0.3023**	**0.2155**	**0.2983**	**0.2979**	**0.1803**	**0.2061**	**0.0000**	**0.9949**	**0.5636**	**0.0000**	**0.5615**	**0.0363**
25, 25	0.3153	0.2496	0.3969	0.3170	0.3765	0.3817	0.2739	0.3031	0.0000	0.9992	0.6174	0.0000	0.6159	0.0436
15, 45	***0.3583***	***0.2557***	***0.3869***	***0.2201***	***0.3204***	***0.3485***	***0.2517***	***0.2886***	***0.0003***	***0.9981***	***0.4180***	***0.0009***	***0.3849***	***0.0585***
30, 30	**0.3906**	**0.3504**	**0.4756**	**0.4095**	**0.4502**	**0.4585**	**0.3666**	**0.3941**	**0.0000**	**0.9999**	**0.6672**	**0.0001**	**0.6643**	**0.0505**
45, 45	0.5975	0.5912	0.6524	0.6222	0.6371	0.6428	0.5943	0.6120	0.0000	1.0000	0.7836	0.0001	0.7816	0.0821
30, 90	***0.6595***	***0.6278***	***0.6536***	***0.5998***	***0.6212***	***0.6458***	***0.5972***	***0.6251***	***0.0007***	***1.0000***	***0.6626***	***0.0018***	***0.6421***	***0.0960***
60, 60	**0.7470**	**0.7503**	**0.7754**	**0.7631**	**0.7704**	**0.7729**	**0.7482**	**0.7584**	**0.0002**	**1.0000**	**0.8630**	**0.0004**	**0.8627**	**0.1090**
90, 90	0.9085	0.9119	0.9145	0.9125	0.9171	0.9167	0.9099	0.9130	0.0020	1.0000	0.9493	0.0029	0.9495	0.1540
120, 120	0.9697	0.9709	0.9700	0.9703	0.9723	0.9717	0.9701	0.9709	0.0151	1.0000	0.9825	0.0109	0.9824	0.1926

Note. Darker shading and longer bars indicate higher rejection rates.

Table 18 (continued)

Rejection Rates for Independent-Samples t Test (with Outliers) for Various Sample Sizes, Effect Sizes, Outlier Amounts, and Distributions (1,000,000 Repetitions)

Sample Size	**Discrete Mass at Zero w/ Gap**	**Mass at Zero**	**Extreme Asymmetry (P/D)**	**Extreme Asymmetry (A/G)**	**Extreme Bimodality**	**Multimodal/Lumpy**	**Digit Preference**	**Smooth Symmetric**	**Normal**	**Uniform**	**Exponential**	**t w/ 3df**	**Chi-squared w/2df**	**Cauchy**
n_1, n_2	**E.S. = 0.5**		**α = .05**		**Outlier Amount = 20%/end**									
5, 15	***0.1828***	***0.0702***	***0.1253***	***0.0797***	***0.0793***	***0.1188***	***0.0831***	***0.0890***	***0.0000***	***0.7091***	***0.3390***	***0.0000***	***0.2588***	***0.0332***
10, 10	**0.0828**	**0.0424**	**0.1167**	**0.0968**	**0.0789**	**0.0979**	**0.0661**	**0.0670**	**0.0087**	**0.8347**	**0.6085**	**0.0130**	**0.6477**	**0.0541**
15, 15	0.1775	0.0518	0.1366	0.1100	0.1400	0.1389	0.0872	0.0883	0.0246	0.9611	0.7723	0.0341	0.7956	0.1119
10, 30	***0.2138***	***0.1185***	***0.2383***	***0.1290***	***0.1190***	***0.2299***	***0.1343***	***0.1471***	***0.0056***	***0.9623***	***0.5746***	***0.0080***	***0.3967***	***0.0744***
20, 20	**0.2379**	**0.0634**	**0.1709**	**0.1337**	**0.2073**	**0.1865**	**0.1101**	**0.1129**	**0.0308**	**0.9918**	**0.8687**	**0.0459**	**0.8833**	**0.1537**
25, 25	0.2847	0.0771	0.2086	0.1604	0.2758	0.2378	0.1354	0.1394	0.0337	0.9985	0.9256	0.0529	0.9344	0.1871
15, 45	***0.2637***	***0.1628***	***0.3556***	***0.1657***	***0.1578***	***0.3374***	***0.1835***	***0.2034***	***0.0193***	***0.9963***	***0.7449***	***0.0244***	***0.5211***	***0.1095***
30, 30	**0.3297**	**0.0921**	**0.2495**	**0.1907**	**0.3442**	**0.2928**	**0.1611**	**0.1667**	**0.0360**	**0.9998**	**0.9585**	**0.0582**	**0.9634**	**0.2191**
45, 45	0.4555	0.1431	0.3899	0.2838	0.5290	0.4654	0.2451	0.2587	0.0408	1.0000	0.9933	0.0715	0.9942	0.3295
30, 90	***0.4436***	***0.2879***	***0.6650***	***0.2747***	***0.3032***	***0.6019***	***0.3282***	***0.3689***	***0.0877***	***1.0000***	***0.9601***	***0.0931***	***0.7874***	***0.2317***
60, 60	**0.5678**	**0.2051**	**0.5471**	**0.3814**	**0.6772**	**0.6213**	**0.3307**	**0.3554**	**0.0454**	**1.0000**	**0.9990**	**0.0840**	**0.9992**	**0.4138**
90, 90	0.7403	0.3530	0.7898	0.5734	0.8631	0.8363	0.4900	0.5351	0.0557	1.0000	1.0000	0.1072	1.0000	0.5326
120, 120	0.8515	0.5171	0.9116	0.7209	0.9483	0.9391	0.6233	0.6831	0.0657	1.0000	1.0000	0.1291	1.0000	0.6159

Note. Darker shading and longer bars indicate higher rejection rates.

Table 19

Rejection Rates for Winsorized Independent-Samples t Test (using Winsorized Critical Values) for Various Sample Sizes, Effect Sizes, and Distributions (1,000,000 Repetitions)

Sample Size	**Discrete Mass at Zero w/ Gap**	**Mass at Zero**	**Extreme Asymmetry (P/D)**	**Extreme Asymmetry (A/G)**	**Extreme Bimodality**	**Multimodal/Lumpy**	**Digit Preference**	**Smooth Symmetric**	**Normal**	**Uniform**	**Exponential**	**t w/ 3df**	**Chi-squared w/2df**	**Cauchy**
n_1, n_2	**E.S. = 0.5**		**α = .05**		**Winsorized Amount = 1/end**									
5, 5	0.4001	0.0880	0.1990	0.1025	0.0866	0.0866	0.0851	0.0853	0.0864	0.2513	0.1101	0.1186	0.1102	0.2376
5, 15	**0.1823**	**0.1423**	**0.1647**	**0.1844**	**0.1414**	**0.1190**	**0.1406**	**0.1386**	**0.1431**	**0.6009**	**0.1650**	**0.2098**	**0.1655**	**0.2956**
10, 10	**0.3194**	**0.1811**	**0.2771**	**0.2089**	**0.0966**	**0.1288**	**0.1703**	**0.1676**	**0.1737**	**0.7430**	**0.2410**	**0.2855**	**0.2405**	**0.4689**
15, 15	0.2162	0.2602	0.3071	0.2763	0.1651	0.2017	0.2453	0.2428	0.2524	0.9455	0.3323	0.3961	0.3318	0.5708
10, 30	**0.2252**	**0.2660**	**0.2165**	**0.3441**	**0.1986**	**0.2222**	**0.2591**	**0.2542**	**0.2614**	**0.9525**	**0.2657**	**0.3780**	**0.2656**	**0.4172**
20, 20	**0.2906**	**0.3418**	**0.3539**	**0.3443**	**0.2527**	**0.2821**	**0.3222**	**0.3182**	**0.3319**	**0.9910**	**0.4113**	**0.4881**	**0.4111**	**0.6401**
25, 25	0.3762	0.4151	0.4047	0.4088	0.3380	0.3585	0.3951	0.3931	0.4042	0.9986	0.4779	0.5616	0.4791	0.6903
15, 45	**0.3367**	**0.3854**	**0.3120**	**0.4384**	**0.3228**	**0.3402**	**0.3687**	**0.3658**	**0.3747**	**0.9966**	**0.3692**	**0.5063**	**0.3699**	**0.4938**
30, 30	**0.4519**	**0.4852**	**0.4603**	**0.4723**	**0.4183**	**0.4317**	**0.4639**	**0.4620**	**0.4735**	**0.9998**	**0.5401**	**0.6254**	**0.5394**	**0.7301**
45, 45	0.6357	0.6547	0.6238	0.6397	0.6177	0.6183	0.6394	0.6374	0.6469	1.0000	0.6903	0.7672	0.6897	0.8124
30, 90	**0.6237**	**0.6628**	**0.6043**	**0.6727**	**0.6299**	**0.6290**	**0.6446**	**0.6416**	**0.6507**	**1.0000**	**0.6368**	**0.7563**	**0.6375**	**0.6344**
60, 60	**0.7678**	**0.7759**	**0.7549**	**0.7658**	**0.7569**	**0.7558**	**0.7673**	**0.7669**	**0.7731**	**1.0000**	**0.7974**	**0.8571**	**0.7984**	**0.8615**
90, 90	0.9134	0.9130	0.9072	0.9106	0.9124	0.9095	0.9128	0.9123	0.9149	1.0000	0.9207	0.9472	0.9202	0.9093
120, 120	0.9701	0.9697	0.9676	0.9689	0.9706	0.9695	0.9700	0.9701	0.9707	1.0000	0.9711	0.9808	0.9712	0.9326

Note. Darker shading and longer bars indicate higher rejection rates.

Table 19 (continued)

Rejection Rates for Winsorized Independent-Samples t Test (using Winsorized Critical Values) for Various Sample Sizes, Effect Sizes, and Distributions (1,000,000 Repetitions)

Sample Size	Discrete Mass at Zero w/ Gap	Mass at Zero	Extreme Asymmetry (P/D)	Extreme Asymmetry (A/G)	Extreme Bimodality	Multimodal/Lumpy	Digit Preference	Smooth Symmetric	Normal	Uniform	Exponential	t w/ 3df	Chi-squared w/2df	Cauchy
n_1, n_2	**E.S. = 0.5**		**α = .05**		**Winsorized Amount = 20%/end**									
5, 15	**0.4354**	**0.1244**	**0.1883**	**0.1275**	**0.0848**	**0.1102**	**0.1162**	**0.1149**	**0.1185**	**0.4511**	**0.1566**	**0.1792**	**0.1573**	**0.3323**
10, 10	**0.4545**	**0.1644**	**0.3378**	**0.1949**	**0.0850**	**0.1147**	**0.1566**	**0.1521**	**0.1550**	**0.5895**	**0.2200**	**0.2724**	**0.2201**	**0.5461**
15, 15	0.6558	0.2390	0.4299	0.2842	0.0878	0.1415	0.2261	0.2207	0.2255	0.8245	0.3280	0.4108	0.3270	0.7296
10, 30	**0.2878**	**0.2010**	**0.1448**	**0.4020**	**0.2372**	**0.1643**	**0.2089**	**0.2051**	**0.2168**	**0.7888**	**0.2028**	**0.3603**	**0.2030**	**0.5839**
20, 20	**0.6502**	**0.3123**	**0.4896**	**0.3613**	**0.0999**	**0.1746**	**0.2936**	**0.2889**	**0.2950**	**0.9375**	**0.4207**	**0.5309**	**0.4228**	**0.8378**
25, 25	0.6430	0.3799	0.5384	0.4288	0.1221	0.2130	0.3565	0.3517	0.3617	0.9796	0.5042	0.6300	0.5040	0.9019
15, 45	**0.2102**	**0.2785**	**0.1266**	**0.6081**	**0.3572**	**0.1902**	**0.3034**	**0.2944**	**0.3100**	**0.9299**	**0.2503**	**0.5240**	**0.2504**	**0.7511**
30, 30	**0.6365**	**0.4443**	**0.5847**	**0.4916**	**0.1479**	**0.2568**	**0.4174**	**0.4137**	**0.4256**	**0.9940**	**0.5753**	**0.7103**	**0.5757**	**0.9416**
45, 45	0.6353	0.6150	0.7052	0.6475	0.2277	0.4002	0.5784	0.5757	0.5963	0.9999	0.7430	0.8705	0.7436	0.9875
30, 90	**0.0894**	**0.5005**	**0.1114**	**0.8912**	**0.4510**	**0.2723**	**0.5721**	**0.5350**	**0.5580**	**0.9986**	**0.3966**	**0.8373**	**0.3957**	**0.9537**
60, 60	**0.6618**	**0.7415**	**0.7956**	**0.7630**	**0.3225**	**0.5380**	**0.7012**	**0.7018**	**0.7253**	**1.0000**	**0.8509**	**0.9465**	**0.8504**	**0.9974**
90, 90	0.7209	0.8925	0.9070	0.9000	0.5435	0.7519	0.8609	0.8628	0.8859	1.0000	0.9544	0.9919	0.9541	0.9999
120, 120	0.7447	0.9584	0.9588	0.9605	0.7312	0.8778	0.9385	0.9398	0.9559	1.0000	0.9871	0.9989	0.9872	1.0000

Note. Darker shading and longer bars indicate higher rejection rates.

CHAPTER V

DISCUSSION

Major Findings

The purpose of this study was to compare approximate and Monte Carlo-derived critical values of the Winsorized *t* test for independent samples with respect to robustness to type I and II errors. As a whole, it was found that using the process of Winsorizing along with the Monte Carlo-derived Winsorized critical values produces more robust type I and II error rates than simply ignoring outliers or using the traditional, adjusted degrees of freedom critical values. Where distributions were not robust to type I error, type II error comparisons lose meaning (i.e. with Discrete Mass at Zero with Gap, as illustrated in table 14) since robustness to the former is a pre-requisite for interpreting the latter.

As Fung and Rahman (1980) asserted, under normality, Winsorizing did account for a loss in power (or increased type II error rates) in some instances. Yet, when using the Monte Carlo-derived critical values, type I error robustness showed overall dramatic improvement (see table 20) and power generally increased. The Monte Carlo-derived critical values are larger, which account for lower rejection rates in general.

20% Winsorized results for approximate critical values were generally liberal, non-robust. Though the Monte Carlo-derived critical values were almost always more robust to type I error, there were cases, with one Winsorized value per end, where the approximate critical values were within Bradley's liberal or stringent definitions of robustness and, if used, would be more robust to type II error (for one Winsorized value per end only).

Since approximate critical values are smaller, they have higher rejection rates. When appropriately robust to type I error, they can be used to produce results with greater robustness to type II error. At the .05 alpha level (*again, for one Winsorized value per end*), for Micceri's real data sets, samples of 25 or more (samples of 45 at the .01 alpha level) produced liberally-robust results. For the Multimodal/Lumpy and Extreme Bimodality distributions, this can be said for samples as small as 15 (25 and 20, respectively for the .01 alpha level). In most cases for real distributions, samples of 90 or greater (for a = .05) produced stringently-robust results with approximate critical values. For the Multimodal/Lumpy and Extreme Bimodality distributions, this can be said for samples as small as 45 and 20, respectively.

For discrete distributions, simulating outliers in this study involved recoding inner to outer values. This inflated the variance more than simply recoding a highest or lowest value to increase its absolute value, especially for smaller samples. This effect of inflated variance due to the recoding process may have interacted with the unique skew of certain distributions to produce inconsistent results. At any rate, such results do not change the overall results addressing the purpose of this study.

Since Winsorizing serves to decrease variance and increase rejections, it follows that alternative critical values would be larger to offset the potential increase in rejected nulls. The degree to which the alternative critical values does this, however, makes a difference in how robust the results are to type I and II errors, as was found in this study.

Table 20

Type I Error Rates for Independent-Samples t Test for Various Sample Sizes and Alpha Levels When Sampling Is From a Normal Dist., 1,000,000 repetitions, 20% Winsorized Critical Values.

	α = .050 (Approximate C.V.)			α = .050 (Monte Carlo C.V.)		
Sample Size	**U025**	**L025**	**Total**	**U025**	**L025**	**Total**
5, 15	⇧ ***0.1161***	⇧ ***0.1161***	⇧ ***0.2322***	⇨ ***0.0255***	⇨ ***0.0256***	⇨ ***0.0512***
10, 10	⇧ **0.1135**	⇧ **0.1135**	⇧ **0.2269**	⇨ **0.0248**	⇨ **0.0252**	⇨ **0.0500**
15, 15	⇧ 0.1146	⇧ 0.1145	⇧ 0.2290	⇨ 0.0252	⇨ 0.0251	⇨ 0.0502
10, 30	⇧ ***0.1197***	⇧ ***0.1189***	⇧ ***0.2386***	⇨ ***0.0268***	⇨ ***0.0270***	⇨ ***0.0538***
20, 20	⇧ **0.1146**	⇧ **0.1153**	⇧ **0.2298**	⇨ **0.0252**	⇨ **0.0254**	⇨ **0.0506**
25, 25	⇧ 0.1148	⇧ 0.1147	⇧ 0.2295	⇨ 0.0249	⇨ 0.0249	⇨ 0.0498
15, 45	⇧ ***0.1200***	⇧ ***0.1196***	⇧ ***0.2396***	⇨ ***0.0271***	⇨ ***0.0272***	⇨ ***0.0543***
30, 30	⇧ **0.1146**	⇧ **0.1149**	⇧ **0.2295**	⇨ **0.0248**	⇨ **0.0247**	⇨ **0.0495**
45, 45	⇧ 0.1150	⇧ 0.1148	⇧ 0.2298	⇨ 0.0249	⇨ 0.0251	⇨ 0.0500
30, 90	⇧ ***0.1204***	⇧ ***0.1205***	⇧ ***0.2409***	⬈ ***0.0275***	⬈ ***0.0276***	⬈ ***0.0551***
60, 60	⇧ **0.1146**	⇧ **0.1147**	⇧ **0.2292**	⇨ **0.0249**	⇨ **0.0249**	⇨ **0.0498**
90, 90	⇧ 0.1150	⇧ 0.1148	⇧ 0.2298	⇨ 0.0252	⇨ 0.0251	⇨ 0.0504
120, 120	⇧ 0.1150	⇧ 0.1151	⇧ 0.2301	⇨ 0.0249	⇨ 0.0250	⇨ 0.0498

Note. SU025 = proportion of rejections in the upper- tail. L025 = proportion of rejections in the lower-tail.

Based on Bradley's (1978) definitions of type I robustness:

⇧ Liberal, Outside the Liberal Range

⬈ Liberal, Inside the Liberal Range

⇨ Within the Stringent Range

⬊ Conservative, Inside the Liberal Range

⇩ Conservative, Outside the Liberal Range

The tails of a distribution can impact robustness more than the shape itself, since decisions are made based on tails.

As mentioned above, there were some instances of interaction between unbalanced sampling, effect size, Winsorized amount, and distribution type where results were unique. For the discrete mass at zero with gap and extreme asymmetry (P/D) distributions, increased Winsorizing actually led to less robust results. For the Exponential and Chi-Squared distributions, more outliers led to more rejections. In general, skewed distributions that had higher concentrations of values on the upper tail had more rejections for unequal samples.

The definition of an outlier was different for discrete and continuous distributions in this study. In a real-life study, one may notice a mass of values at an extreme and wish to Winsorize to lessen their impact on the mean and variance. This is a slightly different approach than simply Winsorizing to recode particularly large or small values, yet the results of this study show that both situations can benefit from both Winsorizing and using the Monte Carlo-derived critical values.

The additive effect of skew and unequal sample size tended to increase upper tail rejections and decrease lower tail rejections when skew was positive (mass at lower end). An opposite trend was apparent for negatively-skewed distributions.

Since Winsorized, unbalanced samples were Winsorized by the exact same amount per end as with balanced samples, the process of Winsorizing reduced variance in the smaller sample more than in the larger one. This created samples with unequal variances and negatively impacted robustness of the results. Increased Winsorizing exacerbated this effect.

There were several factors associated with larger critical values (and hence, smaller rejection rates). These include lower degrees of freedom, decreased alpha level, increased Winsorization, and Monte Carlo critical values (as opposed to approximate or, to a larger degree, traditional). For each factor, more precision is expected from *t*-obtained.

Next Steps

This study addressed symmetrically-Winsorized samples. This body of knowledge would benefit from studies on non-symmetrically-Winsorized and/or trimmed samples. Future studies can examine the same robustness properties for comparing trimmed sample means (similar such studies have been done, but not with the same distributions and number or repetitions). Also, such studies can be extended to the analysis of variance and other more complex statistical procedures.

Other amounts of Winsorization can be examined, since 1 and 10% per end may be too far apart to account for all possible nuances. If critical values were derived for unbalanced samples (such that 20% Winsorizing would be based on the sample itself), unequal variances can be avoided.

A power study comparing the Winsorized *t* to nonparametric and other robust competitors would be beneficial and would build on these results as well as those from Blair et al. (1980). Subsequent studies could catalogue which test to use under specific parameters to ensure robustness to types I and II errors.

The simulation of outliers is a procedure that varies from study to study. Some researchers choose to contaminate samples by drawing outliers from different distribution (Lax, 1985, Carey et. al., 1997), while others simply multiply the highest and lowest values by three or ten (Wilcox, 1998). It would be beneficial to future research if such approaches were compared to see how different results can be when choosing one technique over another, though the purpose of a study that calls for contaminating a distribution may be different than one that does not.

There are other distributions from other fields (such as biology and medicine) that can be estimated and added to such studies. This can be especially beneficial since many researchers in laboratories are not familiar with robust methods or their great benefit to statistical analysis. For example, with recent biological databases emerging, distributions from such areas should become more estimable and robustness studies more feasible.

This study can also be repeated for a table of Winsorized critical values estimated from Yuen and Dixon's (1973) formula for the trimmed *t* for independent samples as recommended by Gans (1988). Though the results may differ, since the results from this study show robustness under normality, there should not be much of a difference.

REFERENCES

Anscombe, F. J. & Guttman, I. (1960). Rejection of outliers. *Technometrics*, 2, 123-147.

Blair, R. C. (1987). *Rangen*: Version 1.0. Boca Raton, FL: IBM.

Blair, R. C., Higgins, J. J., & Smitley, W. D. S. (1980). On the relative power of the U and *t* tests. *British Journal of Mathematical and Statistical Psychology*, 33, 114-120.

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

Bradley, J. V. (1977). A common situation conducive to bizarre distribution shape. *American Statistician*, 31, 147-150.

Bradley, J. V. (1978). Robustness? *British Journal Mathematical and Statistical Psychology*, 31, 114-152.

Carey, V. J., Walters, E. E., Wager, C. G., & Rosner, B. A. (1997). Resistant and test-based outlier rejection: effects on Gaussian one- and two-sample inference. *Technometrics* 39(3), 320-330.

Davies, Laurie & Gather, Ursula. (1993). The identification of multiple outliers. *Journal of the American Statistical Association*. 88(423), 782-792.

Dixon, W. J. (1950). Analysis of extreme values. *The Annals of Mathematical Statistics*, 21(4), 488-506.

Dixon, W. J. (1960). Simplified Estimation from Censored normal Samples. *The Annals of Mathematical Statistics,* 31, 385-391

Dixon, W. J., Tukey, J.W. (1968). Approximate Behavior of the Distribution of Winsorized t (Trimming/Winsorization 2), *Technometrics*, (10)1, 83-98.

Dixon, W. J. and Massey, F. J. (1969). *Introduction to Statistical Analysis*, McGraw-Hill, New York.

Dixon, W. J. and Yuen, K. K. (1974). Trimming and Winsorization: a review. *Statistische Hefte* 2, 157-170.

Efron, B. (1969). Student's t-test under symmetry conditions. *Journal of the American Statistical Association*, 64, 1278-1302.

Fahoome, Gail F. (2002). JMASM Algorithms and Code JMASM1: Rangen 2.0 (Fortran 90/95). *Journal Of Modern Applied Statistical Methods*, 1, 182-190.

Farrell-Singleton, Piper. (2010). Critical values for the two independent samples Winsorized *t*-test. Unpublished Doctoral Dissertation. Wayne State University.

Fung, K.Y. & Rahman, S.M. (1980). The Two-Sample Winsorized *t*. *Communications in Statistics: Simulation and Computation*, 89(4), 337-347.

Gans, D. (1988), *"Trimmed and Winsorized Means, Tests For", Encyclopedia of Statistical Sciences*. Vol. 9, (S. Kotz and N. Johnson, Eds.). New York: Wiley, 346-348.

Gayen, A. K. (1949). The distribution of 'Student' t in random samples of any size drawn from non-normal universes. *Biometrika*, 36, 353-369.

Gayen, A. K. (1950). The distribution of the variance ratio in random samples of any size drawn from non-normal universes. *Biometrika,* 37, 236-255.

Geary, R. C. (1936). The distribution of 'Student's' ratio from non-normal samples. *Journal of the Royal Statistical Society,* 3, 178-184.

Geary, R. C. (1947). Testing for normality. *Biometrika,* 34, 209-242.

Hawkins, D. M. (1980). *Identification of outliers.* London: Chapman & Hall.

Keselman, H. J., Othman, A. R., Wilcox, R.R., & Fradette, K. (2004). The new and improved two-sample t-test. *Psychological Science*, 15(1), 57–51.

Kruskal, W. H. (1960). Some remarks on wild observations. *Technometrics*, 2(1), 1-3.

Lax, David A. (1985). Robust estimators of scale: finite-sample performance in long-tailed symmetric distributions. *Journal of the American Statistical Association*, 80(391), 736-741.

Maxwell, Scott E. (1980). Pairwise Multiple comparisons in repeated measures designs. *Journal of Educational Statistics*, 5(3), 269-287.

Metropolis, N. (1987). The beginning of the Monte Carlo method. *Los Alamos Science*, 15, 125-130.

Metropolis, N. & Ulam, S. (1949). The Monte Carlo Method. *Journal of the American Statistical Association* 44 (247): 335–341.

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

Orr, J. M., Sackett, P. R., & DuBois, C. L. Z. (1991). Outlier detection and treatment in I/O Psychology: A survey of researcher beliefs and an empirical illustration. *Personnel Psychology, 44*, 473-486.

Pearson, E. S., & Please, N. W. (1975). Relation between the shape of population distribution and the robustness of four simple test statistics. *Biometrika*, 62, 223-241.

Ramsey, Philip H., Ramsey, Patricia P. & Barrera, Kyrstle. (2010). Choosing the best pairwise comparisons of means from non-normal populations, with equal variances, but equal sample sizes. *Journal of Statistical Computation and Simulation*, 80(6), 595-608.

Rivest, L. (1994), Statistical properties of Winsorized means for skewed distributions. *Biometrika*, 81(2), 373-383.

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

Sawilowsky, S. S. & Blair, R. C., (1992). A more realistic look at the robustness and Type II error properties of the *t*-test to departures from population normality. *Psychological Bulletin*, 111, 353-360.

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

Sawilowsky, S. S. & Fahoome, G. C. (2003). *Statistics via Monte Carlo Simulation with Fortran*. Rochester Hills, MI: JMASM.

Triola, M. (1997). *Elementary Statistics* (7th ed.). Reading, MA: Addison Wesely Longman Publishing Company.

Wilcox, R. R. (1996). *Statistics for the Social Sciences*. San Diego: Academic Press.

Wilcox, R. R. (1998). How many discoveries have been lost by ignoring modern statistical methods? *American Psychologist, 53*(3), 300-314.

Wilcox, R. R. (2005). New methods for comparing groups: strategies for increasing the probability of detecting true differences. *Current Directions in Psychological Science* 14(5), 272-275.

Yuen, K.K., & Dixon, W. J. (1973). The approximate behavior and performance of the two-sample trimmed t. *Biometrika*, 60, 369-374.

Zimmerman, D. (1994). A note on the influence of outliers on parametric and nonparametric tests. *Journal of General Psychology*, 121(4), 391.

Zimmerman, D. W., & Zumbo, B. D. (1993). The relative power of parametric and non-parametric statistical methods. In Keren, G. and Lewis, C. (Eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues.* Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

ABSTRACT

APPROXIMATE VS. MONTE CARLO CRITICAL VALUES FOR THE WINSORIZED T-TEST

by

MICHAEL W. LANCE

May 2011

Advisor: Dr. Shlomo Sawilowsky

Major: Education Evaluation & Research

Degree: Doctor of Philosophy

Historically, it has been accepted practice for critical values for the Winsorized *t* test for independent samples to be based on adjusted degrees of freedom depending on the number of total non-Winsorized (approximate) values. Recently, a new such table of Winsorized critical values has been developed via approximate randomization by Monte Carlo simulation.

Based on eight common data distributions estimated from Psychology and Education along with the normal and five Mathematical distributions, these two tables of values were compared with respect to robustness to types I and II errors through Monte Carlo simulations for one and 10% Winsorized values per end.

20% Winsorized results were generally non-robust for approximate critical values and mixed for Monte Carlo-derived critical values. With one Winsorized value per end, for small samples, type I error results generally support the use of the newly-developed table of Monte Carlo-derived critical values over the approximate critical values. For

larger samples (one Winsorized value per end), approximate critical values become increasingly robust (in most cases, stringently-so for samples of 90 or more) to type I error while maintaining an advantage over Monte Carlo-derived critical values with respect to type II error.

AUTOBIOGRAPHICAL STATEMENT

Michael Lance
15633 Northville Forest Dr. #Q182
Plymouth, MI 48170
(313) 523-0660
michael.lance@gmail.com

Education

Doctor of Philosophy in Ed. Evaluation & Research May, 2011

Wayne State University, Detroit, Michigan

Master of Arts in Sociology December, 2007

Wayne State University, Detroit, Michigan

Bachelor of Arts in Elementary Education August, 2002

Saginaw Valley State University, University Center, Michigan

Minors: Science and Mathematics

Presentations

Bay Mills Community College

Interpreting and using student achievement data at district & school levels. October, 2010

Wayne State University

Social Psychology: Identity Theory March, 2006

Work Experience

Director of Assessment & Evaluation August, 2007 – Present

Hamadeh Educational Services

Science/Math Teacher August, 2002 – June, 2007

Star International Academy

Crescent Academy International

Farquhar Middle School