DIGITALCOMMONS @WAYNESTATE

Journal of Modern Applied Statistical Methods

Volume 15 | Issue 1 Article 23

5-2016

Outlier Impact and Accommodation Methods: Multiple Comparisons of Type I Error Rates

Hongjing Liao
Beijing Foreign Studies University, hl346309@ohio.edu

Yanju Li
Western Carolina University, yl323205@ohio.edu

Gordon Brooks
Ohio University

Journal of Modern Applied Statistical Methods
May 2016, Vol. 15, No. 1, 452-471.

ISSN 1538 − 9472

Outlier Impact and Accommodation Methods: Multiple Comparisons of Type I Error Rates

Hongjing Liao
Beijing Foreign Studies University
Beijing, China

Yanju Li
Western Carolina University
Cullowhee, NC

Gordon Brooks
Ohio University
Athens, OH

A Monte Carlo simulation study was conducted to examine outliers' influence on Type I error rates in ANOVA and Welch tests, and the effectiveness of two outlier accommodation methods: nonparametric rank based method and Winsorizing. Recommendations are given regarding outlier handling with different sample sizes and number of outliers.

Keywords: outliers, type I error, Monte Carlo simulation, outlier accommodation, nonparametric, Winsorizing

Introduction

Extreme data points, or outliers, requires attention and investigation (Barnett & Lewis, 1994). Outliers are often inevitably seen in data sets of educational research projects, even when data come from reputable sources and the data collection is carefully executed. The existence of outliers has been recognized and noted for centuries, and the outlier problem is generally seen as "reducing and distorting the information about the data source or generating mechanism" (Barnett & Lewis, 1994, p. 4). To put it in a statistical context, there are concerns about the disproportionate influence of outliers on statistical analyses, based on sample means and variance. Studies have provided evidence that shows the effect of outliers resulted in inflation of Type I error rates and reduced power in parametric t and F tests (Barnett & Lewis, 1994; Hampel, Ronchetti, Rousseeuw, & Stahel, 1986; Wilcox, 1998; Zimmerman, 1994b).

Because distortions of statistical significance tests could lead to faulty conclusions if indications of outliers are not carefully examined, it is natural to seek a means of identifying and explaining outliers. A number of studies are devoted to

Hongjing Liao is a lecturer. Email her at: hongjing.liao@139.com. Yanju Li is a Research Development Specialist. Email her at: yanjuli@wcu.edu. Dr. Brooks is an Associate Professor of Educational Research and Evaluation.

investigating the sources of outliers and detecting their presence in various data sets and distributions (Beckman & Cook, 1983). However, very few give emphasis on how to handle outliers, and there are even fewer studies that compare different outlier accommodation techniques. It is useful to investigate the circumstances under which outliers can be treated, as well as the effectiveness of outlier treatment methods. Hence, the purpose of this study is to focus on the impact of outliers on significance tests, and presents simulation results for comparisons of outlier accommodation methods in order to provide recommendations for practice.

Outliers: Definition, Detection, and Accommodation

An outlier refers to an observation that "appears to be inconsistent with the remainder of that set of data" (Barnett & Lewis, 1994, p. 7). Although problems in statistical analyses caused by outliers are a concern in the development of statistical methods (Barnett & Lewis, 1994), perceptions about outliers evolve with the development of educational research methodologies. The restrictive view of outliers being erroneous and contaminating has changed. In a present perspective, outliers are an "empirical reality" (Rousseeuw & Van Zomeren, 1990, p. 650), and instead of being misleading and wrong, they could provide useful information about the sample and, in some situations, indicate that a different model or distribution may fit the data better (Barnett & Lewis, 1994).

In parametric analyses, outliers are often identified according to how particular data points deviate from the center (the mean) of the distribution of the data set. Thus, for a normally distributed data set, the common rule is that an outlier is any value that is beyond $\pm$ 3 standard deviations from the mean. In addition, for different research designs and methods of analysis, there are different approaches developed to detect outliers (Barnett & Lewis, 1994; Berkane & Bentler, 1988; Cook, 1986; Gnanadesikan, 1997; Jarrell, 1991). Some approaches are adapted from univariate methods, such as frequency tables, histograms, and box plots (Allison, Gorman, & Primaverya, 1993; Jarrell, 1991); some use residuals of various kinds (Cook, 1986; David, 1978); others suggest bivariate and multivariate techniques such as Cook's distance (Allison et al., 1993), principal components (Hawkins, 1974), hat matrix (Hoaglin & Welsch, 1978), and Mahalanobis distance (Stevens, 1984). However, with such a variety of approaches available, it is still the researcher's decision to define outliers depending on research contexts, and researchers should always seek meaning and interpretation of outliers before rejecting or choosing any techniques to deal with the deviant observations. The

reason for doing so is because, as the nature and origins of outliers differ, the approaches to handle outliers vary accordingly.

Outliers may arise for deterministic reasons or for less tangible reasons. Deterministic reasons refer to apparent errors in execution of data that are controllable and correctable. Examples of deterministic outliers include recording and calculating errors, erroneous data entries, and failure to specify missing values (Barnett & Lewis, 1994; Tabachinick & Fidell, 2001; Warner, 2008). For outliers that arise as a result of deterministic reasons, the remedy is simple and straightforward: to replace outliers with correct values. However, more often than not, the reasons for the existence of outliers are less clear-cut. Scholars suggested three major sources of outliers: inherent variability, measurement error, and execution error (Anscombe, 1960; Barnett, 1978; Grubbs, 1969; Hampel et al., 1986; Tabachinick & Fidell, 2001). First, inherent variability refers to the variations demonstrated by outliers as a natural feature of the population under study. In other words, the outlying observations are representative of the target population, because the population has more extreme scores than a normal distribution. Of course, outliers are also possible as part of a normal distribution. Second, the occurrence of outliers could also be due to measurement error, such as rounding errors, recording errors, or variability imposed due to an inadequate measuring instrument. Finally, an execution error could be another source of outlying observations, such as a biased sample that includes individuals who are not truly representative of the population. Although theoretically, measurement and execution errors could be examined and corrected, in many circumstances it is very difficult, or even impossible, in practical research projects to distinguish from which sources outliers truly rise.

For less tangible outliers, the reasons for their occurrences are often not clear; there are two basic approaches to handle such outliers: to reject the outliers or to retain and accommodate the outliers to reduce their effect (Jarrell, 1991; Warner, 2008). Rejection of outliers includes simple removal of outliers after taking into account the appropriateness of all data (Field, 2011). Or, as Allison et al. (1993) suggested, rejecting outliers can also include running the analyses with and without outliers, comparing the results and reporting an assessment of the influence of outliers through deletion.

However, it is often not encouraged to reject outliers, especially when there is no tangible explanation about the occurrence of outliers. Outliers can be legitimate data points and removal may cause loss of useful information (Orr, Sackett, & DuBois, 1991). Sometimes outliers may reflect unusual but substantively meaningful aspects of the intended study (Chow, Hamaker, & Allaire,

2009; Hampel, 2001). An alternative approach is to use accommodation methods to reduce the impact of the outlying observations, including utilizing robust tests and outlier treatment methods. However, even with outlier accommodation approaches effectively applied, it is uncertain that the influence of outliers can be removed completely, but the aim is to minimize such influence.

There are several approaches that can be used to diminish or lower the impact of outliers, such as log transformation (Warner, 2008), nonparametric statistical ranking (Zimmerman & Zumbo, 1990), and Winsorizing (Dixon & Tukey, 1969; Dixon & Yuen, 1974). Among statistical tests, nonparametric methods based on ranks are argued to effectively control Type I error rates in the presence of outliers (Zimmerman & Zumbo, 1990). For example, Zimmerman (1994b; 1995) reported that compared to parametric methods, the Mann-Whitney-Wilcoxon test can effectively control Type I error, and nonparametric methods based on ranks exhibited slightly better Type I error rates than ANOVA methods for several outlier-prone and non-normal distributions. Zimmerman and Zumbo (1990) showed that the Type I error rates of Mann-Whitney-Wilcoxon and pooled-variance Student t test were relatively equivalent under simple bounded transformations used to handle outlier-prone distributions.

Furthermore, Winsorizing is another popular method to reduce the weights of outliers by replacing them with a specific percentile of data-dependent values (Dixon & Yuen, 1974; Orr et al., 1991). In practice, the location of Winsorization often depends on prior knowledge, and is suggested to be adjusted according to the shapes of the distribution (Dixon & Yuen, 1974; Tukey, 1962). This study, therefore, explores the effectiveness of the Winsorizing approach with different percentiles.

Purpose of the Study

This study is primarily motivated by two very practical questions: what is the impact of outliers on Type I error rates with different sample sizes and number of outliers, and which outlier accommodation methods can effectively control Type I error rates under varying sample size and outlier number conditions? Therefore, the purpose and contributions of this study are three-fold:

First, this study started by examining outliers' influence on Type I error rates in ANOVA and Welch tests with different sample sizes and number of outliers, and further explored distinct features of such influence in various combinations of conditions. Outlier impact in previous studies is often treated as a type of violation of normality, and the number of outliers in data sets was not studied separately

(Zimmerman, 1994a; 1994b; 1995). This study highlighted the association of outlier number and its influence on Type I error rates. Moreover, given outliers' influence on sample variance, both ANOVA and Welch tests are included and compared to take into consideration the problem of unequal variance in mean comparison analyses.

Secondly, although outlier accommodation methods are available and effective for different conditions and distributions (Dixon & Tukey, 1969; Dixon & Yuen, 1974; Orr et al., 1991; Zimmerman, 1995), no comparison between the methods are provided in terms of the effectiveness of Type I error control under the same conditions. This study investigated two basic approaches in handling outliers and how effective they were in controlling Type I error rates. Specifically, the study compared the Type I error rates when outliers are removed and retained using nonparametric methods and Winsorizing. Comparing the sensitivity of nonparametric and Winsorizing methods on outlier impact not only fills the current gap about the two methods, but can also provide basic information for guidelines of the use of outlier treatment methods. Finally, our study was also conducted to explore the Winsorizing methods with different Winsorization percentiles because no consensus has been reached regarding Winsorization locations, and little information was provided on how to decide the locations in existing literature.

In short, this study ventured to explore some new areas on outlier impact and outlier treatment based on existing studies. From the research design perspective, when the occurrence of outliers cannot be traced, which frequently happens in statistical analyses of educational research, it is reasonable to retain the outliers but give less weight to their influence. Therefore, understanding the impact brought by the presence of outliers and choosing an appropriate method for outlier accommodation are critical for credible analysis and conclusion.

Methods

A Monte Carlo program was developed in the R language for data simulation and computation of statistical results for different outlier and accommodation conditions. As a useful approach to evaluate the quality of statistical procedures (in this case the Type I error rate), a Monte Carlo program allows sample data to be drawn with many iterations in simulation. Rejection rates of significant tests could be counted with many iterations, through which Type I error rate under the true null hypotheses would be obtained (Mooney, 1997). In addition, R as an open-source computer statistical package and programming language has built-in functions to perform the ANOVA F test and the nonparametric Kruskal-Wallis rank sum test.

The general procedures for this simulation study are as follows: first, as this study focuses on multiple comparisons of Type I errors, samples of varied sample sizes and varied number of outliers were drawn from the same univariate normally distributed data simulated using the built-in R function rnorm. For each condition, equal sample sizes were manipulated for three groups and a varied number of outliers are included in only one group (group three). Second, ANOVA and Welch tests were performed using the same group of simulated data for analysis with outliers excluded, outliers included but with no treatment, and outliers accommodated by the nonparametric test and the Winsorizing method. For each condition, 10,000 replications were conducted and Type I error rates for different conditions were computed. Finally, simulation and statistical results were analyzed to examine outlier impact on Type I error rates, as well as advantages and disadvantages of the outlier treatment techniques under different conditions. Details about data generation, outlier injection, replication, and analysis procedures are provided in the following sections.

Data Generation

The sample sizes (n = 20, 40, 60, 80, and 100) were manipulated in the way that the three groups for ANOVA test always had equal sample sizes with the outlier(s) being inserted into only one group. 200,000 normally distributed N(0, 1) cases were generated using the function rnorm (sample size, mean, standard deviation). The generated population data were split into two data sets: data without outliers ($u - 3\sigma \leq x \leq u + 3\sigma$) and data with outliers ($x < u - 3\sigma$ and $x > u + 3\sigma$). Data for each group of a sample were randomly selected from these two data sets. The built-in R function sample was used to randomly sample the required number of observations from different data sets. The random selection procedures were performed in the following way: first, for the first two groups that contain no outliers, n points of data were randomly sampled from each data set ($u_1 - 3\sigma \leq x \leq u_1 + 3\sigma$) and ($u_2 - 3\sigma \leq x \leq u_2 + 3\sigma$), respectively. For the third group that has inserted outliers, $n_{outliers}$ outliers were sampled from the data set ($x < u_3 - 3\sigma$ and $x > u_3 + 3\sigma$), and the absolute value of each was taken to ensure positive outliers. Then the rest of data for group three, $n - n_{outliers}$ ($n_{outliers}$ = 0, 1, 2, 3, 4, and 5) number of data were sampled from the data set ($u_3 - 3\sigma \leq x \leq u_3 + 3\sigma$). To study Type I error rates, the null hypothesis is set to be true. Therefore, each group was randomly drawn from the same normal distribution N(0, 1).

Outlier Injection

As indicated in the sampling procedures, outliers were sampled from data beyond 3 standard deviations on both directions of the generated data, and were injected into each sample. Generating data only between $u - 3\sigma$ and $u + 3\sigma$ results in a slightly decreased standard deviation than the population value but provides a certain gap between normal data and outliers. By "injecting" outliers into the normal data, we can ensure the required number of outliers for the research purpose. This differs from the approach adopted by many to use a "contaminated" standard normal distribution, where some data are generated N(0, 1) while other data are generated at perhaps N(0, 3) or N(2, 1). The difference between the "injection" and the "contamination" methods lie in that the "injection method" guarantees that outliers are from the same normally distributed population and that they are included in every sample. The design is important to study Type I errors because the null hypothesis is held true when drawing the whole sample, including normal data and outliers from the same population.

Replication

10,000 replications were conducted for each condition to minimize the Monte Carlo sampling impact. Robey and Barcikowski (1992) tabulated the number of iterations required for examining departures from varied nominal Type I error rates. Mooney (1997) proposed that the more the better in choosing the number of iterations for Monte Carlo simulations. Thus, in order to sufficiently ensure the stability and generalizability of the results and, meanwhile, to avoid inefficiency in excessive iterations, 10,000 iterations were used for the current study.

Monte Carlo Analysis

For each sample from the simulated population (e.g., $u_1 = u_2 = u_3 = 0$, $n = 20$, $n_{outliers} = 1, 2, 3, 4, 5$, sd = 1), ANOVA and Welch tests were used to test the Null hypothesis. Statistical *p* values were documented for data with no outliers, data with outliers, data with outliers deleted, and data treated by two commonly used outlier accommodation methods: nonparametric and Winsorizing (Winsorized at 95^{th}, 90^{th}, 85^{th}, 80^{th}, and 75^{th} percentile). In R codes, the function anova was used except for the nonparametric Kruskal-Wallis test, which used the built-in R function kruskal.test.

Type I error rates were calculated at the nominal significance level $\alpha = 0.05$. The calculated p values were compared to the liberal criterion $\alpha \pm 1/2\alpha$ with an interval of [0.025, 0.075], the stringent criterion $\alpha \pm 1/10\alpha$ with an interval of [0.045, 0.055] (Bradley, 1978), and the intermediate criterion $\alpha \pm 1/4\alpha$ with an interval of [0.0375, 0.0625] (Robey & Barcikowski, 1992).

Table 1. Type I error rates of parametric significance tests and outlier removed under varied sample sizes and outlier conditions

		Parametric		Outlier Removed	
Sample Size	**Outlier Number**	**Anova**	**Welch**	**Anova**	**Welch**
N = 20	0 outliers	0.0492	0.0467	0.0492	0.0467
	1 outlier	0.0455	0.0459	0.0503	0.0479
	2 outliers	0.0846	0.0702	0.0504	0.0489
	3 outliers	0.1599	0.1182	0.0486	0.0475
	4 outliers	0.3002	0.1940	0.0486	0.0475
	5 outliers	0.4881	0.3168	0.0486	0.0471
N = 40	0 outliers	0.0528	0.0528	0.0528	0.0528
	1 outlier	0.0533	0.0513	0.0513	0.0514
	2 outliers	0.0767	0.0707	0.0530	0.0525
	3 outliers	0.1233	0.1038	0.0517	0.0520
	4 outliers	0.1920	0.1536	0.0523	0.0525
	5 outliers	0.2918	0.2222	0.0516	0.0512
N = 60	0 outliers	0.0497	0.0522	0.0497	0.0522
	1 outlier	0.0509	0.0518	0.0511	0.0512
	2 outliers	0.0659	0.0638	0.0516	0.0516
	3 outliers	0.0981	0.0890	0.0516	0.0518
	4 outliers	0.1480	0.1288	0.0512	0.0514
	5 outliers	0.2156	0.1803	0.0507	0.0509
N = 80	0 outliers	0.0546	0.0535	0.0546	0.0535
	1 outlier	0.0520	0.0514	0.0538	0.0530
	2 outliers	0.0647	0.0625	0.0545	0.0531
	3 outliers	0.0925	0.0844	0.0546	0.0532
	4 outliers	0.1319	0.1161	0.0538	0.0528
	5 outliers	0.1820	0.1571	0.0531	0.0535
N = 100	0 outliers	0.0489	0.0483	0.0489	0.0483
	1 outlier	0.0507	0.0489	0.0499	0.0492
	2 outliers	0.0590	0.0568	0.0494	0.0486
	3 outliers	0.0809	0.0761	0.0505	0.0498
	4 outliers	0.1104	0.1018	0.0501	0.0489
	5 outliers	0.1523	0.1338	0.0503	0.0495

Table 2. Type I error rates of different outlier accommodation techniques under varied sample sizes and outlier conditions

			Winsorizing									
			95th percentile		90th percentile		85th percentile		80th percentile		75th percentile	
Sample size	**Outlier number**	**Nonparametric**	**Anova**	**Welch**	**Anova**	**Welch**	**Anova**	**Welch**	**Anova**	**Welch**	**Anova**	**Welch**
N = 20	0 outliers	0.0480	0.0492	0.0467	0.0492	0.0467	0.0492	0.0467	0.0492	0.0467	0.0492	0.0467
	1 outlier	0.0459	0.0514	0.0504	0.0539	0.0523	0.0536	0.0533	0.0540	0.0542	0.0551	0.0534
	2 outliers	0.0583	0.0832	0.0700	0.0711	0.0663	0.0675	0.0645	0.0654	0.0635	0.0654	0.0642
	3 outliers	0.0873	0.1601	0.1182	0.1535	0.1167	0.1090	0.0974	0.0928	0.0870	0.0850	0.0801
	4 outliers	0.1348	0.3081	0.1973	0.2912	0.1917	0.2804	0.1885	0.1752	0.1470	0.1288	0.1197
	5 outliers	0.2098	0.5090	0.3259	0.4839	0.3174	0.4670	0.3115	0.4525	0.3052	0.2766	0.2245
N = 40	0 outliers	0.0507	0.0528	0.0528	0.0528	0.0528	0.0528	0.0528	0.0528	0.0528	0.0528	0.0528
	1 outlier	0.0508	0.0555	0.0533	0.0562	0.0544	0.0565	0.0545	0.0561	0.0544	0.0552	0.0539
	2 outliers	0.0593	0.0674	0.0646	0.0621	0.0610	0.0615	0.0609	0.0602	0.0600	0.0605	0.0602
	3 outliers	0.0748	0.1176	0.1019	0.0818	0.0780	0.0742	0.0710	0.0708	0.0676	0.0674	0.0654
	4 outliers	0.0988	0.1847	0.1509	0.1297	0.1172	0.0990	0.0956	0.0864	0.0847	0.0794	0.0766
	5 outliers	0.1298	0.2872	0.2217	0.2664	0.2124	0.1429	0.1322	0.1118	0.1073	0.0964	0.0939
N = 60	0 outliers	0.0508	0.0497	0.0522	0.0497	0.0522	0.0497	0.0522	0.0497	0.0522	0.0497	0.0522
	1 outlier	0.0511	0.0518	0.0527	0.0520	0.0531	0.0519	0.0528	0.0520	0.0528	0.0527	0.0531
	2 outliers	0.0559	0.0593	0.0596	0.0562	0.0572	0.0550	0.0558	0.0550	0.0558	0.0553	0.0551
	3 outliers	0.0644	0.0780	0.0745	0.0682	0.0679	0.0626	0.0624	0.0602	0.0590	0.0591	0.0586
	4 outliers	0.0805	0.1379	0.1215	0.0872	0.0839	0.0773	0.0744	0.0684	0.0689	0.0646	0.0650
	5 outliers	0.1004	0.2049	0.1735	0.1221	0.1146	0.0940	0.0920	0.0812	0.0806	0.0729	0.0732
N = 80	0 outliers	0.0514	0.0546	0.0535	0.0546	0.0535	0.0546	0.0535	0.0546	0.0535	0.0546	0.0535
	1 outlier	0.0511	0.0554	0.0537	0.0549	0.0542	0.0550	0.0540	0.0550	0.0536	0.0551	0.0531
	2 outliers	0.0548	0.0591	0.0599	0.0586	0.0574	0.0576	0.0566	0.0566	0.0563	0.0566	0.0557
	3 outliers	0.0620	0.0730	0.0695	0.0654	0.0643	0.0627	0.0619	0.0599	0.0600	0.0594	0.0591
	4 outliers	0.0742	0.1002	0.0932	0.0765	0.0756	0.0721	0.0700	0.0677	0.0661	0.0631	0.0621
	5 outliers	0.0913	0.1674	0.1483	0.0992	0.0942	0.0820	0.0802	0.0758	0.0749	0.0718	0.0693
N = 100	0 outliers	0.0509	0.0489	0.0483	0.0489	0.0483	0.0489	0.0483	0.0489	0.0483	0.0489	0.0483
	1 outlier	0.0513	0.0507	0.0501	0.0509	0.0504	0.0511	0.0501	0.0512	0.0504	0.0516	0.0499
	2 outliers	0.0531	0.0542	0.0538	0.0531	0.0534	0.0532	0.0525	0.0530	0.0517	0.0522	0.0511
	3 outliers	0.0606	0.0632	0.0619	0.0593	0.0591	0.0575	0.0569	0.0560	0.0551	0.0556	0.0540
	4 outliers	0.0697	0.0823	0.0799	0.0689	0.0679	0.0633	0.0635	0.0598	0.0606	0.0591	0.0580
	5 outliers	0.0795	0.1125	0.1052	0.0818	0.0800	0.0719	0.0715	0.0658	0.0658	0.0628	0.0622

Results

The Monte Carlo simulation results are summarized in Table 1 and Table 2. The results include Type I error rates of parametric significance tests and different outlier accommodation techniques under five sample sizes (20, 40, 60, 80, and 100) with six outlier conditions (outlier = 0, 1, 2, 3, 4, 5). Each entry in the tables is the probability of falsely rejecting the null hypothesis under the situation of a true Null (the Type I error rate). The rows represent sample sizes and the number of outliers, and the columns represent the significance tests with either untreated or treated outliers.

The Influence of Outliers on Statistical Results

The first two major columns in Table 1, "parametric" and "outlier removed" columns, provide the Type I error rates in ANOVA and Welch tests for untreated outliers and after outliers being removed from the data set. The results show a clear influence of outliers on the statistical results of significance tests, which is illustrated by inflated Type I error rates. When outliers are removed from the data set, the Type I error rates of both ANOVA and Welch tests drop back to a significance level of around 0.05. Results in the parametric column show the general trend that with an increasing number of outliers being "injected" into the sample, the probability of Type I error increases from the significance level of 0.05 to a maximum of 0.4881 (ANOVA, $n = 20$, $n_{outliers} = 5$).

There are several notable features in the parametric test results regarding the influence of outliers on statistical results. First, the impact of outliers reflected in inflated Type I error rates varies significantly depending on the number of outliers present in the data set. Table 1 shows the number of outlier conditions from 0 to 5. Figure 1 shows the Type I error rates of ANOVA and Welch tests when no outlier and only one outlier is present. As it is shown in Figure 1, when there is only one outlier, the Type I error rates maintain around the significance level 0.05 regardless of the sample size and significance tests. For both ANOVA and Welch tests, a single outlier exerts little modification on the false rejection rates. Compared with a single outlier in the data set, there is an apparent inflation of Type I error rates when there are two outliers.

For example, when there are two outliers, the Type I error rates of ANOVA tests for sample size 20, 40, 60 and 80 are 0.0846, 0.0767, 0.0659 and 0.0647, respectively. All of them exceed the upper bound of Robey and Barcikowski's

(1992) intermediate criterion 0.0625. When there are three outliers or more, there is an even more dramatic increase in the Type I error rate across all sample sizes, all of which become greater than the upper bound of the liberal criterion 0.075 (Bradley, 1978), with the lowest Type I error rates as 0.0761 (Welch, $n = 100$) and 0.0809 (ANOVA, $n = 100$). This tendency of inflated Type I errors with an increased number of outliers can be clearly shown in the graphical representations of Figure 2, the Type I error rates of ANOVA and Welch tests with 0 to 5 number of outliers across five different sample sizes.

In addition to the number of outliers casting an impact on the Type I error rate, the second feature involves sample sizes. As it can also be shown in Figure 2, the magnitude of Type I error rate inflation decreases with the growth of sample size. In other words, the impact of outliers on the false rejection rates is substantially greater with smaller sample sizes, and as sample size increases, the impact of outliers decreases although it is still inflated.

When other conditions hold the same, the Welch test showed a better control of Type I error rates when compared with the ANOVA test in presence of outliers. Although the Type I error rates are inflated beyond Bradley's (1978) criteria for both ANOVA and Welch tests when there are more than three outliers, at each sample size level with the same number of outliers, the Welch test has a less inflated Type I error rates than the ANOVA test.

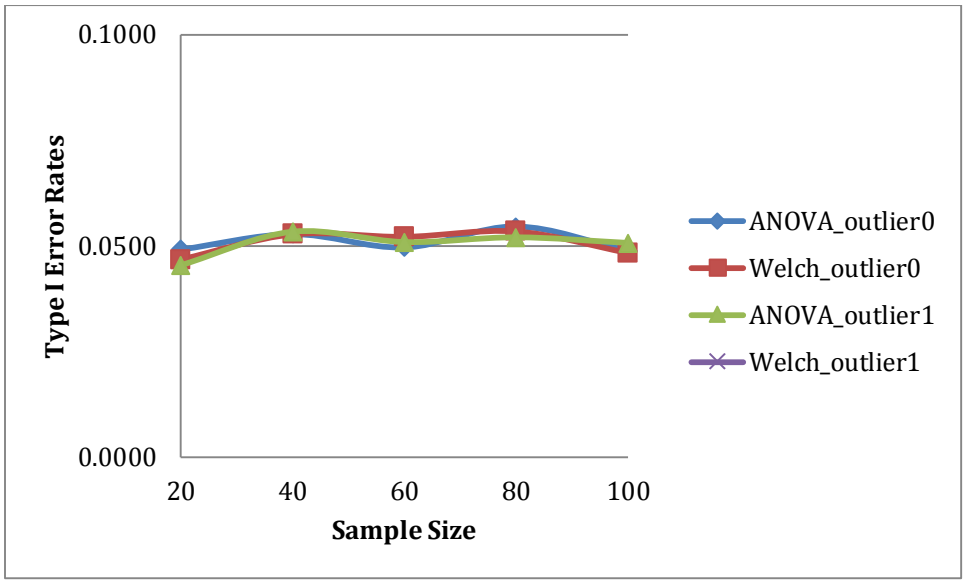

Figure 1. Type I error rates for ANOVA and Welch when zero and one outlier exist

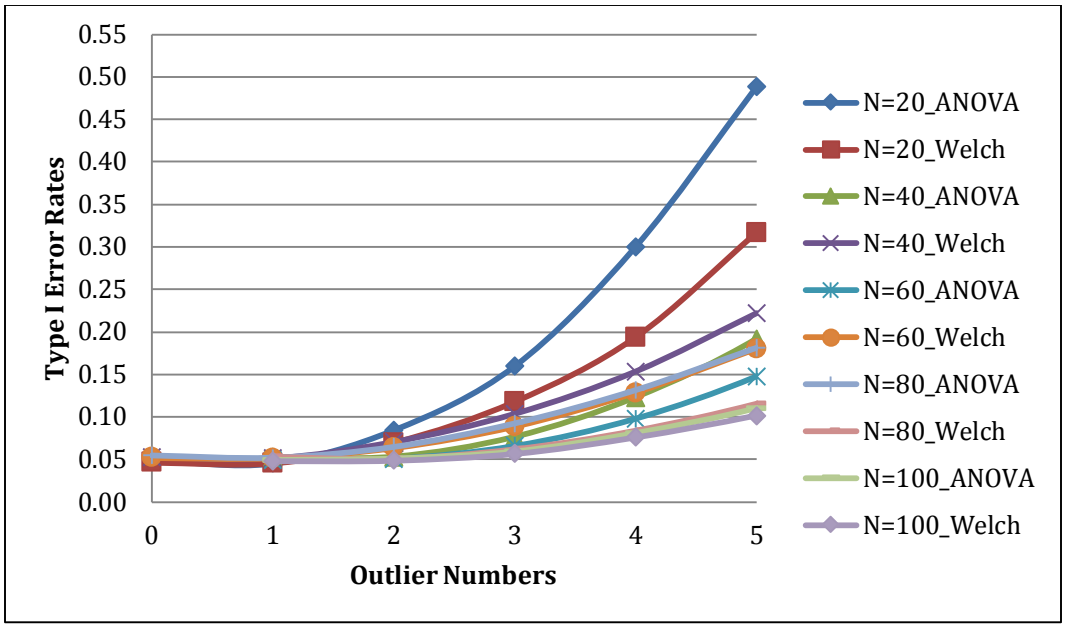

Figure 2. Type I error rates for ANOVA and Welch with varied sample sizes and number of outliers.

Outlier Treatments Methods can Reduce Outlier Influence on Statistical Results

Table 2 provides Type I error rates of significance tests with outlier accommodation methods being applied: the robust approach of using nonparametric Kruskal-Wallis test and the outlier treatment method of Winsorizing at the 95th, 90th, 85th, 80th, and 75th percentiles.

Overall, both the nonparametric Kruskal-Wallis test and the Winsorizing method are effective in reducing outlier influence on statistical results. Figure 3 shows graphical comparisons of Type I error rates of ANOVA and Welch with untreated and treated outliers under various conditions. For sample size equal to 20, 40, 60, 80 and 100, the outlier treatment method of Winsorizing is illustrated at 75th, 80th, 85th, 85th and 90th percentiles as examples, at which the Type I error rates are acceptable. It can be seen from Figure 3, with untreated outliers, the Type I error rates inflate rapidly as the number of outliers increase. Comparatively, the Type I error rate inflation is reduced to the acceptable intervals of criteria when outlier accommodation methods are used. This tendency in the results is not only accurate for the examples in Figure 3; it is also consistent across all sample size and methods conditions.

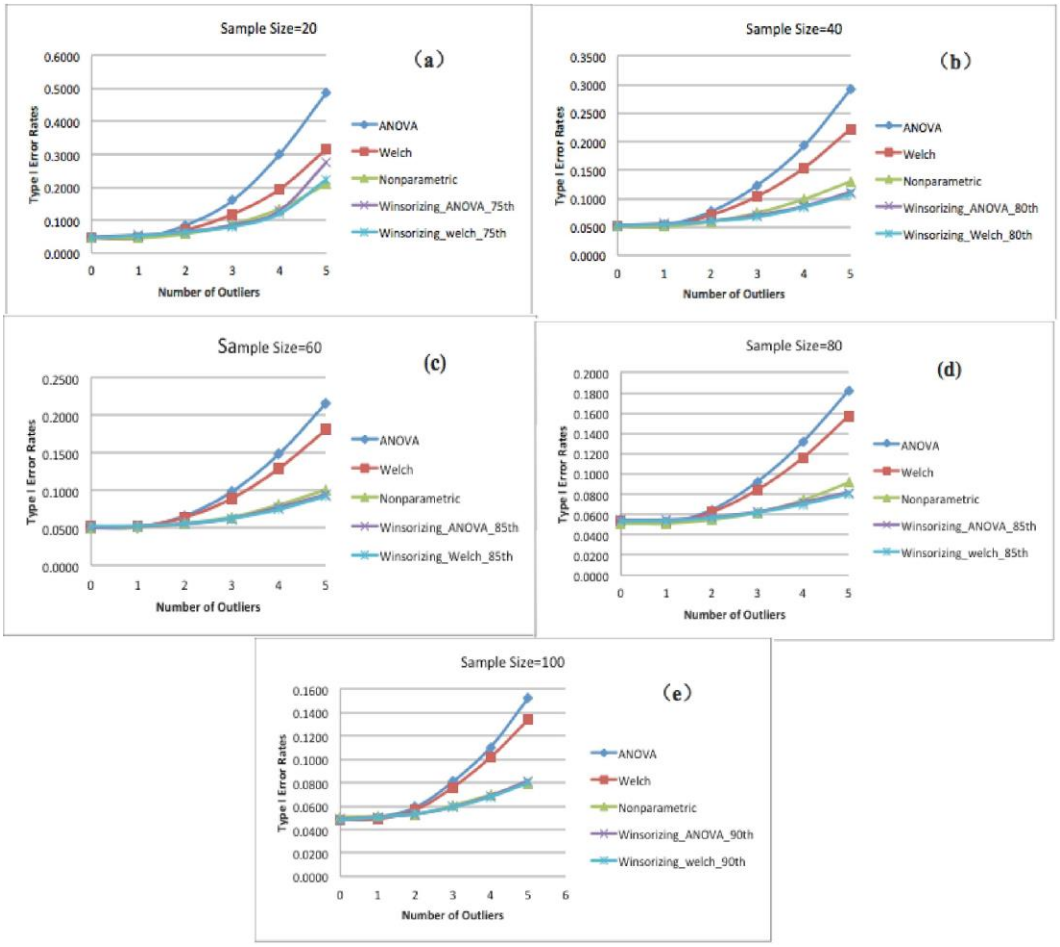

Figure 3. Type I error rates for ANOVA, Welch, nonparametric, and (a) 75 percentile of Winsorizing with sample size of 20; (b) 80 percentile of Winsorizing with sample size of 40; (c) 85 percentile of Winsorizing with sample size of 60; (d) 85 percentile of Winsorizing with sample size of 80; (e) 90 percentile of Winsorizing with sample size of 100.

Outlier Accommodation Methods: Sensitivity

To a certain extent, the outlier treatment methods "corrected" the influence of outliers on the statistical results, although the degree of correction varies for different methods. The two outlier accommodation techniques perform differently in minimizing the impact of outliers under different conditions.

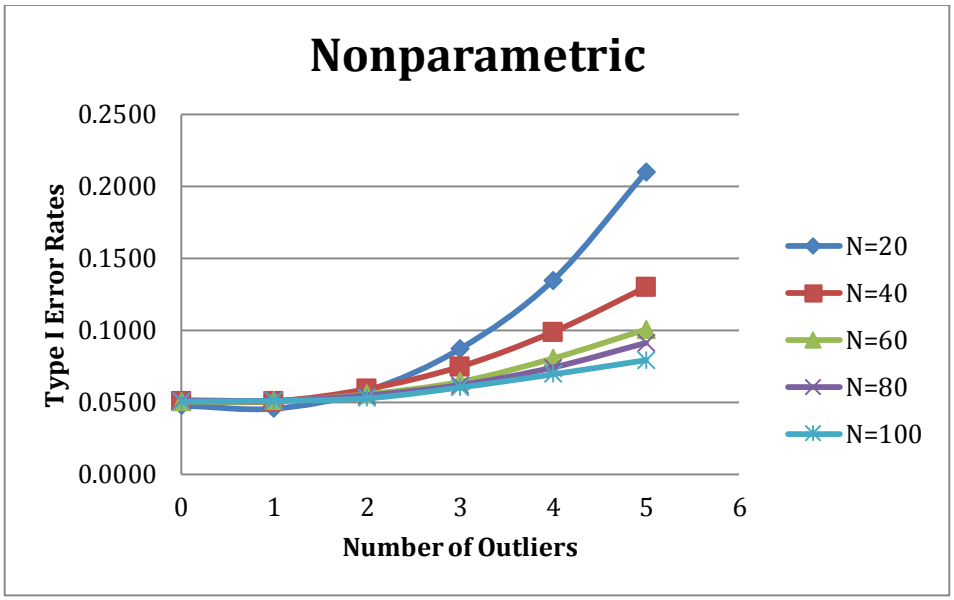

Figure 4. Type I error rates for nonparametric with varied sample size and number of outliers.

The effectiveness of the Kruskal-Wallis test in controlling Type I error rates depends jointly on the number of outliers and sample size. The intertwining effect of sample size and outlier numbers can be observed in Figure 4, which shows the Type I error rates across five different sample sizes of the nonparametric Kruskal-Wallis test in the presence of a different number of outliers.

First, with respect to the number of outliers, two or fewer outliers show little modification in the probability of Type I error rates for the Kruskal-Wallis test, and this result is in accord with conclusions of previous studies about the robustness of nonparametric tests under violations of normality (e.g., Zimmerman, 1994b; 1995). However, when there are three or more outliers present, there is still discernable inflation of Type I errors, and the Kruskal-Wallis test is not able to effectively control Type I error rates to be within the interval of Bradley's (1978) standards.

Second, similar to the impact of untreated outliers on the probability of Type I errors, sample size plays a role regarding the magnitude of change: the larger the sample size, the relatively less inflation in Type I error rates.

Table 2 shows the Type I error rates of ANOVA and Welch tests after outliers being Winsorized at five different percentiles under varying sample size and outlier number conditions. In other words, the injected outliers in each data set are replaced by the scores at the assigned percentile (95^{th}, 90^{th}, 85^{th}, 80^{th}, and 75^{th}). Similarly to

the nonparametric test, Winsorizing also shows an effective control of Type I error rates, and the effectiveness of Winsorizing varies at different Winsorization locations. How much to Winsorize in order for a reasonable control of Type I error is jointly affected by the sample size and the number of outliers.

A smaller Winsorization percentile, such as the 75th or 80th percentile, is necessary to control Type I error rates when sample sizes are small. As the sample size increases, the impact of outliers on probability of Type I errors decreases, and a relatively larger percentile (90th or 95th) of Winsorization is sufficient to accommodate the effects of outliers to achieve an acceptable Type I error rate. Regarding the number of outliers, with two or fewer outliers, Winsorizing at the 95th percentile shows a good control of Type I error rates across all sample sizes except when n = 20. At each sample size, with growing number of outliers in the data set, a smaller percentile is necessary for a good control of Type I error rates. However, it is important to note that when sample sizes are small, such as n = 20 and 40, even Winsorization at the 75th percentile does not show very effective control of Type I error rates when there are four or more outliers. As sample sizes grow bigger, a 75th percentile Winsorizing can reduce the inflated Type I error to meet the intermediate or liberal standards (Bradley, 1978; Robey & Barcikowski, 1992).

Conclusion

Based on the results and figures presented in the result section, certain statements of existing studies regarding the impact of outliers were replicated. In addition, it was confirmed that outliers can change the probability of Type I errors by exerting a disproportionate influence on means and variances in parametric F tests such as ANOVA and Welch tests. The current study provides new evidence in two ways: first, positive outliers inserted into one of the three groups can inflate the Type I error rates of F tests when the null hypothesis is true. Secondly, for previous studies, the impact of outliers was investigated using the contamination method in an outlier-prone data set (Zimmerman, 1994a; 1994b; 1995), in which the precise number of outliers or the extremity of outliers are not specified at each condition. The current study, by adopting the injection method, specified the number and relative extremity of the outliers, and made sure that the inserted outliers did belong to the population. The current study comes to similar conclusions with studies using contamination methods, and further confirms the impact of outliers under a different circumstance.

Furthermore, regarding outlier impact on a nonparametric test, different from conclusions drawn from mixed-normal distributions where outliers were studied in Zimmerman's (1994b; 1995) studies, the current study investigated different outlier number and sample size conditions, and presented similar results under certain conditions and different conclusions in other conditions. The impact of outliers on nonparametric tests in terms of Type I error rates depends on sample size and the number of outliers. When sample size is relatively large (n = 80 and 100), a nonparametric test has a good control of Type I error even when there are five outliers, which is in accord with the results of previous studies (Zimmerman, 1994b; 1995). However, when the sample size is small, there is non-ignorable inflation of Type I error caused by outlier influence, especially with two and more outliers present. Therefore, the nonparametric test is robust against outlier influence, but more attention should be paid when the sample size is small.

It is the number of outliers that seems to matter on the issue of outlier impact on the statistical results, regardless of the sample size. In other words, no matter how large the sample size is, the false rejection rates almost adhere to the nominal significance level (0.05) when the number of outliers is less than two, indicating that no accommodation techniques are necessary. As the number of outliers increases, the inflation of Type I errors begins to appear. Different outlier accommodation techniques have similar effect when the number of outliers was less than two, but the effect began to differ greatly as the number of outliers increased.

As for the comparison of sensitivity to outlier influence between nonparametric test and Winsorizing, it largely depends on the number of outliers in the data set and the location of Winsorization. When there are only two outliers, both the nonparametric test and Winsorizing methods show an effective control of Type I error rates. Yet, when sample sizes are small, the nonparametric test shows a better control of Type I errors than Winsorizing at the 95^{th} and 90^{th} percentile, but the two accommodations methods yield similar results at the 85^{th} and 80^{th} percentile of Winsorization. Therefore, with relatively small sample sizes, nonparametric could have an advantage over Winsorizing in controlling Type I error rates, especially when a large Winsorization percentile is preferred. When there are more than three outliers, a nonparametric test is still relatively robust with large sample sizes, but it does not show a good control of Type I errors when sample sizes are as small as 20 and 40. Comparatively, the Winsorizing with different percentiles can still maintain a good control of the probability of Type I error except that, as the number of outliers increases, the Winsorization location requires a smaller percentile. Thus, when encountering a relatively small sample size with three or

more outliers present, the Winsorizing method offers a more accurate control of Type I error rates than the nonparametric test.

Recommendations can be made regarding the choices of parametric tests when outliers exist in the data and outlier accommodation methods. First, since a single outlier can have little impact on the Type I error rates under a true Null condition, researchers can keep the outlier in the data regardless of the sample size and the Type of F tests applied. However, if there is more than one outlier, the Welch test shows a better performance in controlling Type I error rates and is therefore recommended over ANOVA. Second, with regard to the choice of outlier accommodation methods, the nonparametric test is recommended for small sample sizes when two or less outliers are identified, or for large sample sizes when the number of outliers exceeds three. In addition, the method of Winsorizing is able to accommodate different sample size and outlier number conditions with different Winsorization percentiles. The smaller the sample size and the more outliers, the smaller percentile of Winsorization is required to have a better control of Type I error rates.

Many factors contribute to what approaches or methods should be taken in actual research, and the recommendations made in this study are solely based on the factor of controlling Type I error inflation and on the premise that both parametric and nonparametric approaches are available for use. It is recommended that the outliers be investigated as part of the research design before applying any accommodation techniques, and decisions on the choice of methods should consider the research design and methodology. Apart from Type I errors, other statistical factors such as power will also contribute to the effectiveness of outlier accommodation methods, which should be investigated in future studies of this topic.

References

Allison, D. B., Gorman, B. S., Primaverya, L. H. (1993). Some of the most common questions asked of statistical consultants: Our favorite responses and recommended readings. *Genetic, Social, and General Psychology Monographs, 119*(2), 155-184.

Anscombe, F. J. (1960). Rejection of outliers. *Technometrics, 2*(2), 123-147. doi: 10.1080/00401706.1960.10489888

Barnett, V. (1978). Convenient probability plotting positions for the normal distribution. *Journal of the Royal Statistical Society. Series C (Applied Statistics), 25*(1), 47-50. doi: 10.2307/2346518

Barnett, V., & Lewis, T. (1994). *Outliers in statistical data* (3rd ed.). Chichester, UK: Wiley.

Beckman, R., & Cook, R. D. (1983). Outlier.....s. *Technometrics, 25*(2), 119-149. doi: 10.2307/1268541

Berkane, M., & Bentler, P. M. (1988). Estimation of contamination parameters and identification of outliers in multivariate data. *Sociological Methods & Research, 17*(1), 55-64. doi: 10.1177/0049124188017001003

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology, 31*(2), 144-152. doi: 10.1111/j.2044-8317.1978.tb00581.x

Chow, S., Hamaker, E. L., & Allaire, J. C. (2009). Using innovative outliers to detect discrete shifts in dynamics in group-based state-space models. *Multivariate Behavioral Research, 44*(4), 465-496. doi: 10.1080/00273170903103324

Cook, R. D. (1986). Assessment of local influence. *Journal of the Royal Statistical Society. Series B (Methodological), 48*(2), 133-169. Retrieved from http://www.jstor.org/stable/2345711

David, H. A. (Ed.). (1978). *Contributions to survey sampling and applied statistics: Papers in honor of H. O. Hartley*. New York, NY: Academic Press.

Dixon, W. J., & Tukey, J. W. (1969). Approximate behavior of the distribution of Winsorized *t* (trimming/Winsorization 2). *Technometrics, 10*(1), 83-98. doi: 10.1080/00401706.1968.10490537

Dixon, W. J., & Yuen, K. K. (1974). Trimming and Winsorization: A review. *Statistische Hefte, 15*(2-3), 150-170. doi: 10.1007/BF02922904

Field, M. S. (2011). Application of robust statistical methods to background tracer data characterized by outliers and left-censored data. *Water Research, 45*(10), 3017-3118. doi: 10.1016/j.watres.2011.03.018

Gnanadesikan, R. (1997). *Methods for statistical data analysis of multivariate observations*. (2nd ed.). New York, NY: Wiley.

Grubbs, F. E. (1969). Procedures for detecting outlying observations in samples. *Technometrics, 11*(1), 1-21. doi: 10.1080/00401706.1969.10490657

Hampel, F. R. (2001). Robust statistics: A brief introduction and overview. *Robust Statistics and Fuzzy Techniques in Geodesy and GIS Symposium*. Retrieved from ftp://ess.r-project.org/Research-Reports/94.pdf

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., & Stahel, W. A. (1986). *Robust statistics: The approach based on influence functions*. New York, NY: Wiley.

Hawkins, D. M. (1974). The detection of errors in multivariate data using principal components. *Journal of the American Statistical Association, 69*(346), 340-344. doi: 10.1080/01621459.1974.10482950

Hoaglin, D. C., & Welsch, R. E. (1978). The hat matrix in regression and ANOVA. *The American Statistician, 32*(1), 17-22. doi: 10.1080/00031305.1978.10479237

Jarrell, M. G. (1991). Multivariate outliers: Review of the literature. *Annual Meeting of the Mid-South Educational Research Association*. Available from http://eric.ed.gov/?id=ED339754

Mooney, C. Z. (1997). *Monte Carlo Simulation*. Thousand Oaks, CA: Sage Publications.

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*(3), 473- 486. doi: 10.1111/j.1744-6570.1991.tb02401.x

Robey, R. R., & Barcikowski, R. S. (1992). Type I error and the number of iterations in Monte Carlo studies of robustness. *British Journal of Mathematical and Statistical Psychology, 45*(2), 283-288. doi: 10.1111/j.2044-8317.1992.tb00993.x

Rousseeuw, P. J., & Van Zomeren, B. C. (1990). Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association, 85*(411), 633-639. doi: 10.2307/2289995

Stevens, J. P. (1984). Outliers and influential data points in regression analysis. *Psychological Bulletin, 95*(2), 334-344. doi: 10.1037/0033-2909.95.2.334

Tabachinick, B. G., & Fidell, L. S. (2001). *Using multivariate statistics* (4th ed.). New York, NY: Allyn and Bacon.

Tukey, J. W. (1962). The future of data analysis. *The Annals Mathematical Statistics, 33*(1), 1-67. Retrieved from http://www.jstor.org/stable/2237638

Warner, R. M. (2008). *Applied statistics: From bivariate through multivariate techniques*. Thousand Oaks, CA: SAGE Publication.

Wilcox, R. R. (1998). How many discoveries have been lost by ignoring modern statistical methods? *American Psychologist, 53*(3), 300-314. doi: 10.1037/0003-066X.53.3.300

Zimmerman, D. W. (1994a). Increasing the power of the ANOVA F test for outlier-prone distributions by modified ranking methods. *Journal of General Psychology, 122*(1), 83-94. doi: 10.1080/00221309.1995.9921224

Zimmerman, D. W. (1994b). A note on the influence of outliers on parametric and nonparametric tests. *The Journal of General Psychology, 121*(4), 391-401. doi: 10.1080/00221309.1994.9921213

Zimmerman, D. W. (1995). Increasing the power of nonparametric tests by detecting and downweighting outliers. *Journal of Experimental Education, 64*(1), 71-85. doi: 10.1080/00220973.1995.9943796

Zimmerman, D. W., & Zumbo, B. D. (1990). The relative power of the Wilcoxon-Mann-Whitney test and student *t* test under simple bounded transformations. *The Journal of General Psychology, 117*(4), 425-436. doi: 10.1080/00221309.1990.9921148