Journal of Modern Applied Statistical Methods

Volume 16 | Issue 1 Article 15

5-1-2017

Outlier Impact and Accommodation on Power

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

Yanju Li
Western Carolina University, Cullowhee, NC, yanjuli@email.wcu.edu

Gordon P. Brooks
Ohio University, brooksg@ohio.edu

Outlier Impact and Accommodation on Power

Cover Page Footnote
This research was supported by School of English for Specific Purposes, Beijing Foreign Studies University, grant ZJ1513

Journal of Modern Applied Statistical Methods
May 2017, Vol. 16, No. 1, 261-278.
doi: 10.22237/jmasm/1493597640

ISSN 1538 − 9472

Outlier Impact and Accommodation on Power

Hongjing Liao
Beijing Foreign Studies University
Beijing, China

Yanju Li
Western Carolina University
Cullowhee, NC

Gordon P. Brooks
Ohio University
Athens, OH

The outliers' influence on power rates in ANOVA and Welch tests at various conditions was examined and compared with the effectiveness of nonparametric methods and Winsorizing in minimizing the impact of outliers. Results showed that, considering both power and Type I error, a nonparametric test is the safest choice to control the inflation of Type I error with a decent sample size and yield relatively high power.

Keywords: Outlier, Monte Carlo simulation, nonparametric, Winsorizing, Type I error, power

Introduction

Outliers are defined as "observations (or subset of observations) which appears to be inconsistent with the remainder of that set of data" (Barnett & Lewis, 1994, p. 4). They are often present in datasets of educational research, and could have disproportionate influence on statistical conclusions. Therefore, outlier detection and outlier treatment have become important issues in the practice of statistical analysis (Bakker, & Wicherts, 2014; Rousseeuw & van Zomeren, 1990). Detection of outliers has been the focus of outlier research for decades, and there is abundant literature on outlier detecting approaches (Berkane & Bentler, 1988; Barnett & Lewis, 1994; Cook, 1986; Gnanadesikan, 1997; Jarrell, 1991).

In practice, the most widely used method is to detect an outlier using the absolute *Z* value in standard normal distributions; a threshold value of *Z* beyond 3 is often used. Other methods include using the median absolute deviation statistic (MAD), the interquartile range (IQR), and different kinds of residuals (Bakker, & Wicherts, 2014; Barnett & Lewis, 1994; Berkane & Bentler, 1988; Cook, 1986; Gnanadesikan, 1997; Jarrell, 1991). There are also bivariate and multivariate techniques for outlier detection, such as principal components, hat matrix, minimum volume ellipsoid, minimum covariance determinant, minimum

Dr. Hongjing Liao is a lecturer. Email her at: hongjing.liao@139.com.

generalized variance, and Mahalanobis distance (Hawkins, 1974; Hoaglin & Welsch, 1978; Stevens, 1984; Wilcox, 2012). Methods of outlier detection may vary depending on research design, methods, and contexts. Yet after detecting an outlier, the researcher faces another challenge of dealing with the outliers. It is suggested that before any treatment on outliers, the unusual observations should be examined and evaluated under the specific context and try to find the reason for their occurrence. Outlier occurrence is usually from the following four sources: (a) errors, such as erroneous data entries, analysis errors, or equipment problems; (b) failure to specify missing values; (c) including a case that does not belong to the target population; (d) an actual value of the target population but the population has more extreme scores than a normal distribution (Freedman, Pisani, & Purves, 2007; Tabachnick & Fidell, 2001; Hampel, 2001; Warner, 2012).

Warner (2012) suggested three approaches of dealing with an outlier: "to retain, omit, or modify" (p. 287). When reasons for outlier occurrence are deterministic, that is, due to apparent errors in execution of data that are controllable, the approach to deal with the outliers is to correct or delete erroneous values. However, when reasons for outlier occurrence are less apparent, it is often recommended to decide on outlier handling before seeing the results of the main analyses and to report transparently about how outliers were handled (Bakker, & Wicherts, 2014; Liao, Li, & Brooks, 2016). Under these circumstances, thoughtless removal of the outliers is often not recommended, as outlying data can be legitimate data points (Orr, Sackett, & DuBois, 1991). When outliers are unusual but substantively meaningful aspects of the intended study, deleting the outliers causes loss of useful information and often increases the probability of finding a false positive (Chow, Hamaker, & Allaire, 2009; Hampel, 2001). If outliers have to be removed, it is suggested to compare the resulting analyses with and without outliers, and then report an assessment of the influence of outliers through deletion (Allison, Gorman, & Primavera, 1993; Bakker, & Wicherts, 2014).

Many other studies suggest outlier accommodation is a more reliable method to address outliers than simple removal (Analytical Methods Committee, 1989). Accommodation of outliers includes using a robust approach to reduce the impact of the outlying observations and treating outliers to lower their impact in statistical tests. Nonparametric statistical ranking is a commonly-used robust test that is shown to be less influenced by outliers; other robust tests also include the Mann-Whitney-Wilcoxon test and the Yuen-Welch test (Zimmerman & Zumbo, 1990).

Other popular approaches to treating outliers include trimming and Winsorizing (Wilcox, 1998; Dixon & Yuen, 1974). Trimming involves removing the extreme values and often results in a loss in sample size and power. Winsorizing is another popular method to reduce the weights of outliers by replacing them with a specific percentile of data-dependent values (Orr, Sackett, & DuBois, 1991). One-end Winsorizing means that, when outliers are all positive or negative, they are replaced from only one end; two-end Winsorizing means replacing outliers from each end. These different approaches of outlier accommodation may well vary in usefulness of producing consistent study results, and may affect both Type I error and power.

Some researchers studied the robustness of nonparametric tests in the presence of outliers (Zimmerman, 1994, 1995; Li et al., 2009), and Zimmerman (1995) found that nonparametric methods based on ranks have an advantage for outlier-prone densities over ANOVA. However, few studies have focused on multiple comparisons of different outlier accommodation methods. In 2014, the authors conducted a Monte Carlo simulation study and examined the influence of outliers on Type I error rates in ANOVA and Welch tests, and compared nonparametric test and Winsorizing at different locations in controlling outlier impact (Liao et al., 2016). In the current study, the authors followed up their previous simulation study to add new approaches to outlier accommodation methods on Type I error, and further explored outlier impact and accommodation methods on power.

Purpose of the Study

The purpose of this study is to look for answers to two practical questions by means of Monte Carlo methods: (1) what is the impact of outliers on statistical power with different effect sizes, sample sizes, and number of outliers? (2) Among the commonly-used outlier accommodation methods, such as nonparametric rank-based test and Winsorizing (one-end and two-end), which method is more effective in reducing outlier impact, and under what circumstances?

In this study, outliers' influence on statistical power in ANOVA and Welch tests were examined with different effect sizes, sample sizes, and number of outliers. Furthermore, two basic approaches in handling outliers, nonparametric tests and Winsorizing, and their effectiveness in controlling outlier impact were investigated. More specifically, the study compared the statistical power in the following two conditions: when the outliers were retained and non-parametric

methods were then applied to the data, and when outliers were treated using Winsorizing. As there has been no consensus regarding the percentile of Winsorizing and little information provided on how to decide the locations in existing literature, this study explored both one-end and two-end Winsorizing, and compared their difference in statistical power and Type I error.

Compared with outlier detection, there are few studies that concentrate on outlier treatment methods and even fewer on comparisons of outlier accommodation techniques. This study ventures to explore some new areas based on existing studies. From the research design perspective, when the reason for outlier occurrence 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. Moreover, this study focused on multiple comparisons of outlier accommodation techniques and presents simulation results for comparisons of outlier accommodation methods in order to provide recommendations for practice.

Methodology

In this study, a Monte Carlo program developed in the R programming language was conducted to simulate data, extract samples and calculate the statistics indices under a variety of conditions. First, three groups of univariate standard normal distribution data under different conditions were simulated by using the built-in R function rnorm. Samples of varied sample size and varied number of outliers were drawn from the same univariate normally-distributed data. For each condition, equal sample sizes were manipulated for three groups and a varied number of outliers were injected in only one group. ANOVA and Welch tests were performed using the same group of simulated data with both outliers included but with no treatment, and with outliers accommodated by the two types of Winsorization methods. Nonparametric tests were also performed using the same sample data with outliers included. For each condition, 10,000 replications were conducted. Type I error rate and statistical power for different outlier accommodation techniques and two different effect size conditions were computed and compared, and advantages and disadvantages of the outlier treatment techniques under different conditions noted.

Data Generation and Outlier Injection

The sample sizes (n = 20, 40, 60, 80, and 100) were manipulated in such a way that the three groups for statistical test always had equal sample sizes with the outlier(s) being inserted into only one group (group one). 200,000 normally distributed N(0, 1) cases were generated using the function rnorm. 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 sample were randomly selected from these two data sets. Previous research has investigated outlier impact on Type I error rate (Liao et al., 2016); this study repeated the Monte Carlo methods under the true null condition. Distinct from that effort, however, the performance of the Type I error rate by adding both one-end and two-end Winsorizing methods was considered in the current study. The mean for each group was 0. Additionally, two false null conditions were examined to display the performance of power rate under different treatment methods such as ANOVA, Welch, Nonparametric test and two types of Winsorizing. For the first false null condition, the means for group one, group two, and group three were set as 0.0, 0.3, and 0.6, respectively. For the second false null condition, 0.0, -0.3 and -0.6 were assigned respectively to the means of group one, two and three. The two conditions have equal magnitude of effect size.

Outliers were sampled from data beyond 3 standard deviations in both directions of the generated data, and the absolute value of outliers were injected into each sample; that is, all the inserted outliers are positive 1, 2, 3, 4 and 5 outliers and, for each sample size mentioned above, were investigated for both the Type I error analysis and power study under various treatment methods.

Monte Carlo Analysis

Under the true null hypothesis for each sample from the simulated population (e.g., $u_1 = 0$, $u_2 = 0$, $u_3 = 0$, $n = 40$, $n_{\text{outliers}} = 1, 2, 3, 4, 5$, sd = 1), two types of Winsorizing methods were used to examine the extent to which the inflation of the Type I error could be controlled: Winsorizing one end of data (the side with outliers) and Winsorizing both ends of data. The specified percentiles of Winsorizing for each condition are listed in Table 1, and the percentiles were performed through setting the value of parameter lambda (λ) in the R program. For example, when $N = 40$ and $n_{\text{outliers}} = 2$, $\lambda = 0.05$ (5th percentile) was employed. Under the conditions of one-end Winsorizing, only the outlier(s) were Winsorized; under the conditions of two-end Winsorizing, both the outlier(s) and the corresponding number of data on the opposite side were Winsorized.

Table 1. The percentile of Winsorizing (lambda λ)

	Number of outlier(s)				
Sample size	**1**	**2**	**3**	**4**	**5**
20	0.0500	0.1000	0.1500	0.2000	0.2500
40	0.0250	0.0500	0.0750	0.1000	0.1250
60	0.0200	0.0400	0.0500	0.0700	0.0900
80	0.0125	0.0250	0.0375	0.0500	0.0625
100	0.0100	0.0200	0.0300	0.0400	0.0500

Under the false null hypothesis, for each sample from the simulated population (e.g., $u_1 = 0$, $u_2 = 0.3$, $u_3 = 0.6$, $n = 20$, $n_{\text{outliers}} = 1, 2, 3, 4, 5$, sd = 1), ANOVA and Welch tests were used to explore the statistical power, that is, true rejection rates for the false null hypothesis. Statistical *p*-values were documented for data with no outliers, data with outliers, and data treated by two commonly-used outlier accommodation methods: nonparametric and Winsorizing.

Apart from the simulation procedures and data analyses, this study also adopted different verification approaches to validate data generation and collection. A small sample size (e.g., $N = 10$), small outlier number (e.g., 1 outlier), and small replications (e.g., 10 iterations) were carried out for generating dataset. Total rejection rates computed by hand were compared with the solution acquired from a cyber-program in order to manually verify data generation. A few normally- distributed sample data sets, simulated by the R program, were tested via the Statistical Package for the Social Sciences (SPSS) program. The data were confirmed to be indeed distributed normally. Various trials such as 1, 10, 100, and 1000 were employed for the stress-testing of R codes. All the results obtained from the specific R testing codes exhibited good performances under varied conditions.

Results

Simulation results are compiled in Table 2 and Table 3. The results include statistical power of parametric significance tests and different outlier accommodation techniques under two effect sizes (0, 0.3, 0.6; 0, -0.3, -0.6), five sample sizes (20, 40, 60, 80, and 100), and with six outlier conditions (outlier = 0, 1, 2, 3, 4, 5).

Table 2. Power of parametric significance tests and different outlier accommodation techniques under varied sample size, outlier conditions and effect size 0.0, 0.3, 0.6

Sample Size	Outlier	Parametric		Non-parametric	Wins.: one-end		Wins: two-end	
		ANOVA	Welch		ANOVA	Welch	ANOVA	Welch
N = 20	0	0.3720	0.3629	0.3473	0.3720	0.3629	0.3720	0.3629
	1	0.1555	0.1536	0.2250	0.2677	0.2619	0.2618	0.2573
	2	0.0663	0.0927	0.1312	0.1929	0.1939	0.1871	0.1879
	3	0.0451	0.0859	0.0831	0.1523	0.1589	0.1588	0.1578
	4	0.0573	0.0992	0.0653	0.1341	0.1421	0.1604	0.1614
	5	0.0989	0.1296	0.0720	0.1276	0.1378	0.2057	0.2042
N = 40	0	0.6723	0.6620	0.6419	0.6723	0.6620	0.6723	0.6620
	1	0.4966	0.4800	0.5425	0.5754	0.5623	0.5674	0.5561
	2	0.3313	0.3265	0.4393	0.4741	0.4648	0.4572	0.4500
	3	0.2148	0.2349	0.3493	0.3898	0.3859	0.3684	0.3655
	4	0.1365	0.1826	0.2712	0.3209	0.3249	0.2991	0.2980
	5	0.1003	0.1643	0.2127	0.2686	0.2785	0.2505	0.2561
N = 60	0	0.8507	0.8480	0.8245	0.8507	0.8480	0.8507	0.8480
	1	0.7520	0.7429	0.7699	0.7918	0.7852	0.7875	0.7815
	2	0.6340	0.6208	0.7070	0.7203	0.7143	0.7060	0.7025
	3	0.5068	0.5006	0.6380	0.6475	0.6397	0.6261	0.6201
	4	0.3817	0.3976	0.5605	0.5739	0.5673	0.5411	0.5378
	5	0.2818	0.3203	0.4877	0.5739	0.5673	0.4652	0.4677
N = 80	0	0.9386	0.9376	0.9235	0.9386	0.9376	0.9386	0.9376
	1	0.8940	0.8888	0.8958	0.9108	0.9081	0.9086	0.9054
	2	0.8301	0.8199	0.8652	0.8718	0.8689	0.8628	0.8611
	3	0.7449	0.7349	0.8211	0.8246	0.8197	0.8103	0.8056
	4	0.6509	0.6473	0.7748	0.7745	0.7681	0.7509	0.7469
	5	0.5483	0.5595	0.7232	0.7198	0.7133	0.6845	0.6830
N = 100	0	0.9748	0.9741	0.9658	0.9748	0.9741	0.9748	0.9741
	1	0.9575	0.9547	0.9536	0.9634	0.9619	0.9618	0.9608
	2	0.9253	0.9197	0.9373	0.9450	0.9431	0.9414	0.9392
	3	0.8811	0.8740	0.9180	0.9211	0.9162	0.9128	0.9088
	4	0.8234	0.8168	0.8955	0.8897	0.8863	0.8757	0.8729
	5	0.7561	0.7525	0.8630	0.8543	0.8511	0.8331	0.8305

Table 3. Power of parametric significance tests and different outlier accommodation techniques under varied sample size, outlier conditions and effect size 0.0, -0.3, -0.6

Sample Size	Outlier	Parametric		Non-parametric	Wins.: one-end		Wins: two-end	
		ANOVA	Welch		ANOVA	Welch	ANOVA	Welch
N = 20	0	0.3724	0.3640	0.3488	0.3724	0.3640	0.3724	0.3640
	1	0.4864	0.4405	0.4185	0.4564	0.4391	0.4994	0.4846
	2	0.6272	0.5462	0.4984	0.5549	0.5247	0.6414	0.6251
	3	0.7641	0.6590	0.5877	0.6471	0.6096	0.7680	0.7572
	4	0.8791	0.7709	0.6777	0.7253	0.6902	0.8670	0.8608
	5	0.9508	0.8712	0.7665	0.7925	0.7610	0.9314	0.9347
N = 40	0	0.6691	0.6621	0.6352	0.6691	0.6621	0.6691	0.6621
	1	0.7557	0.7354	0.6878	0.7343	0.7230	0.7515	0.7423
	2	0.8307	0.8015	0.7382	0.7924	0.7760	0.8241	0.8166
	3	0.8934	0.8613	0.7895	0.8402	0.8253	0.8842	0.8754
	4	0.9413	0.9081	0.8311	0.8835	0.8685	0.9273	0.9209
	5	0.9702	0.9456	0.8714	0.9148	0.9017	0.9587	0.9536
N = 60	0	0.8542	0.8472	0.8265	0.8542	0.8472	0.8542	0.8472
	1	0.9008	0.8909	0.8576	0.8902	0.8833	0.8979	0.8930
	2	0.9350	0.9245	0.8836	0.9189	0.9135	0.9296	0.9253
	3	0.9613	0.9495	0.9075	0.9411	0.9317	0.9555	0.9502
	4	0.9786	0.9677	0.9259	0.9588	0.9516	0.9727	0.9693
	5	0.9882	0.9804	0.9451	0.9706	0.9646	0.9837	0.9821
N = 80	0	0.9424	0.9398	0.9243	0.9424	0.9398	0.9424	0.9398
	1	0.9619	0.9585	0.9399	0.9581	0.9553	0.9606	0.9585
	2	0.9762	0.9723	0.9534	0.9695	0.9678	0.9738	0.9723
	3	0.9863	0.9824	0.9640	0.9793	0.9767	0.9843	0.9817
	4	0.9922	0.9896	0.9725	0.9862	0.9838	0.9905	0.9897
	5	0.9951	0.9940	0.9790	0.9910	0.9890	0.9941	0.9937
N = 100	0	0.9793	0.9790	0.9719	0.9793	0.9790	0.9793	0.9790
	1	0.9881	0.9872	0.9773	0.9861	0.9854	0.9872	0.9864
	2	0.9925	0.9907	0.9821	0.9903	0.9898	0.9915	0.9910
	3	0.9954	0.9943	0.9867	0.9933	0.9925	0.9943	0.9937
	4	0.9977	0.9968	0.9901	0.9954	0.9948	0.9969	0.9965
	5	0.9983	0.9977	0.9925	0.9968	0.9966	0.9979	0.9977

Outlier Impact on Power

Results for the first false null condition (mean = 0.0, 0.3, 0.6) are summarized in Figure 1 and Figure 2. As is shown in the figures, under the first false null condition, the presence of outliers caused significant decrease in the power of statistical testing. When sample size is as small as 20, with the presence of one

outlier, the power dropped by about 60% in ANOVA from 0.372 to 0.156. As sample size increases, the power decrease slowed down. When sample size is as large as 100, the power dropped only by less than 2% at the presence of an outlier.

Shown in Figure 2 is the statistical power when nonparametric and Winsorizing two-end methods were used under the first false null condition. From what is shown in the figure, outlier accommodation methods, though slightly different in effectiveness, can help diminish the impact of outlier on power. However, these outlier-robust measures can only diminish the impact but can hardly eliminate the impact.

Simulation results for the second false null condition (mean = 0.0, -0.3, -0.6) are summarized in Figure 3 and Figure 4. In contrast to the first null condition, under the second false null condition power was increased with the presence of outliers. The results further confirmed the impact of outliers on power rates, and indicated that, as the number of outliers increase, their impact on power increases as well.

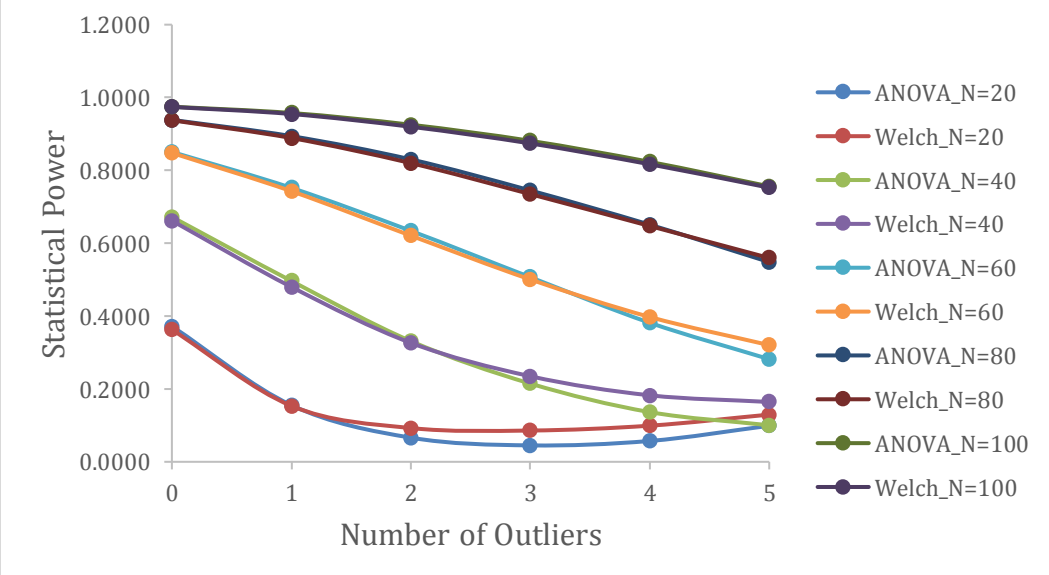

Figure 1. Statistical power for ANOVA and Welch with varied sample size and number of outliers when standardized group mean equals to 0.0, 0.3 and 0.6

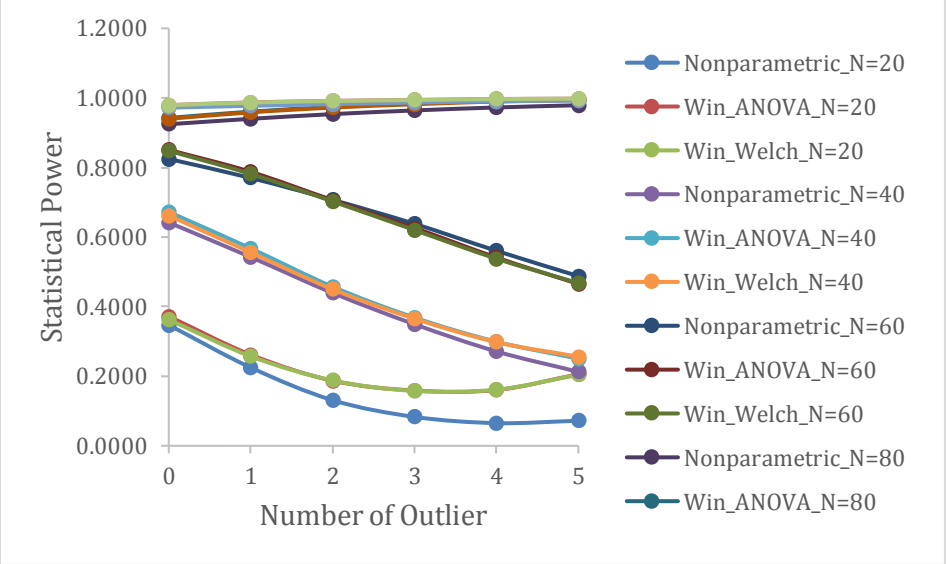

Figure 2. Statistical power for Nonparametric and Wisorizing two-end method with varied sample size and number of outliers when standardized group mean equals to 0.0, 0.3 and 0.6

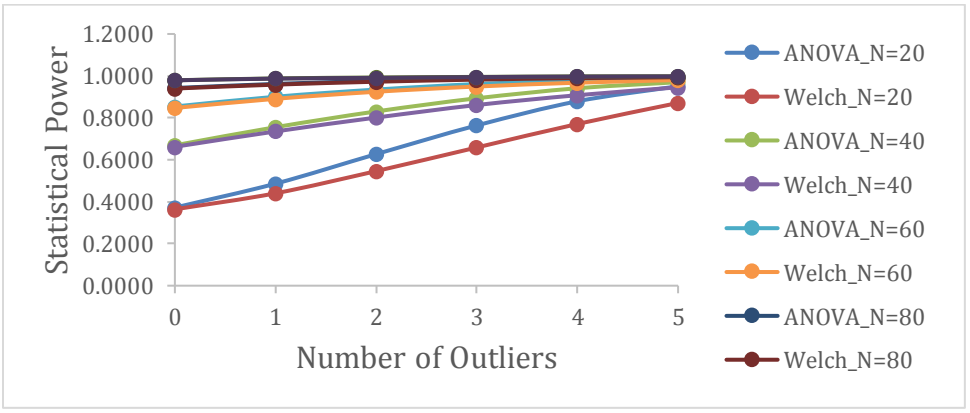

Figure 3. Statistical power for ANOVA and Welch with varied sample size and number of outliers when standardized group mean equals to 0.0, -0.3 and -0.6

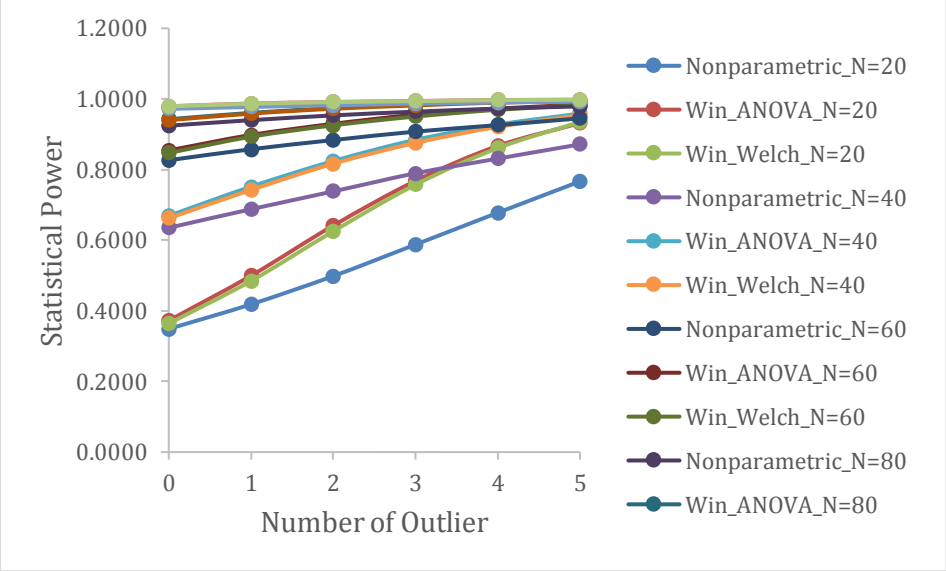

Figure 4. Statistical power for Nonparametric and Wisorizing two-end method with varied sample size and number of outliers when standardized group mean equals to 0.0, -0.3 and -0.6

Regarding the impact of outliers and effect size, in this study we inserted only positive outliers, and the results show that outlier impact is different for positive (mean = 0.0, 0.3, 0.6) and negative effect sizes (mean = 0.0, -0.3, -0.6).

Shown in Figure 5 is one of the examples of outlier impact on power with two effect sizes, and other results of other sample sizes showed similar trends. For positive effect size (0.0, 0.3, 0.6), the presence of outliers decreases power; for negative effect size (0.0, -0.3, -0.6), outliers increase power. Similar simulations with the effect size (-0.3, 0.0, 0.3) were conducted and yielded similar results as the effect size (0.0, 0.3, 0.6). Note that, in Figure 5: P1_ANOVA corresponds to Parametric ANOVA under the effect size 0.0, 0.3 and 0.6; P1_Welch corresponds to Parametric Welch under the effect size 0.0, 0.3 and 0.6; W11_ANOVA corresponds to Winsorizing one-end ANOVA under the effect size 0.0, 0.3 and 0.6; W11_Welch corresponds to Winsorizing one-end Welch under the effect size 0.0, 0.3 and 0.6; W12_ANOVA corresponds to Winsorizing two-end ANOVA under the effect size 0.0, 0.3 and 0.6; W12_Welch corresponds to Winsorizing two-end Welch under the effect size 0.0, 0.3 and 0.6; and P2_ANOVA corresponds to Parametric ANOVA under the effect size 0.0, -0.3 and -0.6.

Across all effect sizes, sample sizes, and numbers of outliers, ANOVA yields more power than Welch tests (see Table 2 and Table 3).

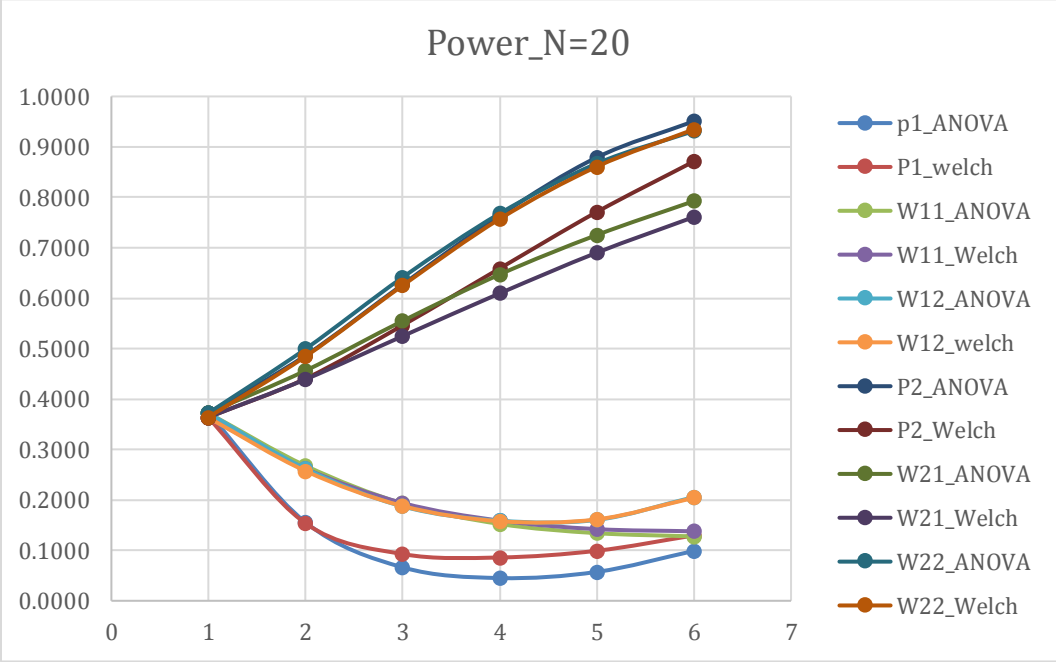

Figure 5. Statistical power for ANOVA and Welch with sample size = 20, number of outliers = 0, 1, 2, 3, 4, 5, and two effect sizes (standardized group mean equals to 0.0, 0.3, 0.6 and 0.0, -0.3, -0.6)

Comparison of Outlier Accommodation Methods on Power

Results on the effectiveness of the two outlier accommodation methods are now presented. As the results in Table 2 and Table 3 showed, outlier accommodation methods, including nonparametric tests and one-end and two-end Winsorizing, can help diminish the impact of outliers to a certain degree.

For the positive effect size, the decrease of power is a concern for parametric tests (ANOVA and Welch) when outliers are present. When sample sizes are small, the impact of outliers is stronger and outlier accommodation methods are relatively more effective in decreasing outlier impact; therefore they might be more useful in that case. For example, for positive effect size (0.0, 0.3, 0.6) and when $n = 40$ with 5 outliers, power decreased over 80%; outlier accommodation methods can increase the power by about 50%. Comparatively,

when sample sizes are large and when power decreases by about 50%, the outlier accommodation methods can increase the power by 10% at most.

Table 4. Type I error rates of nonparametric tests and Winsorizing method under varied sample sizes and outlier conditions

Sample Size	Outlier	Non-parametric	Winsorizing: one-end		Winsorizing: two-end	
			ANOVA	Welch	ANOVA	Welch
N = 20	0	0.0480	0.0492	0.0467	0.0492	0.0467
	1	0.0459	0.0522	0.0507	0.0615	0.0585
	2	0.0583	0.0707	0.0664	0.1043	0.1004
	3	0.0873	0.1034	0.0949	0.1940	0.1873
	4	0.1348	0.1563	0.1354	0.3339	0.3293
	5	0.2098	0.2282	0.1952	0.5053	0.5129
N = 40	0	0.0507	0.0528	0.0528	0.0528	0.0528
	1	0.0508	0.0556	0.0531	0.0597	0.0565
	2	0.0593	0.0674	0.0642	0.0856	0.0818
	3	0.0748	0.0898	0.0844	0.1288	0.1239
	4	0.0988	0.1247	0.1156	0.1950	0.1842
	5	0.1298	0.1709	0.1535	0.2878	0.2717
N = 60	0	0.0508	0.0497	0.0522	0.0497	0.0522
	1	0.0511	0.0517	0.0530	0.0546	0.0558
	2	0.0559	0.0617	0.0611	0.0713	0.0709
	3	0.0644	0.0776	0.0742	0.1015	0.0994
	4	0.0805	0.1052	0.0983	0.1461	0.1418
	5	0.0509	0.1399	0.1285	0.2087	0.1996
N = 80	0	0.0514	0.0546	0.0535	0.0546	0.0535
	1	0.0511	0.0543	0.0529	0.0566	0.0558
	2	0.0548	0.0621	0.0613	0.0694	0.0676
	3	0.0620	0.0770	0.0721	0.0931	0.0891
	4	0.0742	0.0990	0.0922	0.1315	0.1244
	5	0.0913	0.1303	0.1196	0.1787	0.1674
N = 100	0	0.0509	0.0489	0.0483	0.0489	0.0483
	1	0.0513	0.0506	0.0496	0.0527	0.0516
	2	0.0531	0.0551	0.0554	0.0609	0.0601
	3	0.0606	0.0685	0.0669	0.0832	0.0807
	4	0.0697	0.0879	0.0844	0.1121	0.1074
	5	0.0795	0.1116	0.1043	0.1503	0.1434

Regarding a comparison between nonparametric tests and Winsorizing, for the first false null condition with the effect size (0.0, 0.3, 0.6), Winsorizing

performed a little better in obtaining higher power than nonparametric test. In general, both nonparametric and Winsorizing show similar effects in increasing power. Similarly, a comparison of one-end and two-end Winsorizing methods shows that the two Winsorizing methods yield similar results, with one-end Winsorizing having slightly better performance in controlling outlier impact on power.

It is suggested by the simulation results and comparison of outlier accommodation methods above that, when examining the robustness and effectiveness of outlier accommodation methods, both power and Type I error should be taken into consideration. In our earlier study (Liao et al., 2016), we compared the effectiveness of nonparametric tests and one-end Winsorizing in controlling outlier impact on Type I error rates. In this study, based on earlier results, a comparison of Type I error rates with one-end and two-end Winsorizing was conducted. Table 4 is a summary of Type I error rates from previous studies with new results on the comparison of one-end and two-end Winsorizing methods. For effect size (0.0, 0.3, 0.6), although both nonparametric and Winsorizing show similar effects in increasing power, nonparametric methods yield the lowest Type I error rates across different sample sizes and numbers of outliers. For effect size (0.0, -0.3, -0.6), as the presence of outliers increases power, there is less concern regarding power but more regarding Type I error rate. Nonparametric tests were shown to be the most robust in controlling Type I error among all accommodation methods. Between one-end and two-end Winsorizing, one-end Winsorizing consistently performed better in controlling outlier impact on Type I error and power. In addition, one-end Winsorizing becomes more effective when the number of outliers gets bigger.

Conclusion

It was concluded previously that the impact of outliers on nonparametric tests in terms of Type I error rates alone depends on sample size and the number of outliers (Liao et al., 2016). When sample size is relatively large (e.g., n = 80 and 100), a nonparametric test has a good control of Type I error. When the sample size is small, there is non-ignorable inflation in Type I error caused by outlier influence, especially with two and more outliers present. Furthermore, it is the number of outliers that seems to matter when it comes to the issue of outlier impact on the statistical results, regardless of the sample size. 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.

This simulation study built on the previous simulation stud. It further compared outlier accommodation with one-end and two-end Winsorizing and followed up with outlier impact on power to discuss outlier accommodation methods with consideration of both power and Type I error. This study has yielded new evidence regarding outlier impact on power, and the comprehensive effectiveness of the two commonly-used outlier accommodation methods in controlling outlier impact on Type I error and power.

First, the results show that the location of outliers could affect the direction of their impact. When only positive outliers were inserted, power decreases for positive effect size (mean = 0.0, 0.3, 0.6) and increases for negative effect size (mean = 0.0, -0.3, -0.6). Therefore, depending on the location of the outliers, the researcher needs to decide when outlier impact on power is a big concern.

Secondly, among parametric tests, ANOVA, and Welch tests yield similar results in the presence of outliers; Welch tests consistently have better control in Type I error rate. Winsorizing seems a little more effective compared with nonparametric tests in controlling outlier impact on power, but since the difference is less than 5% and nonparametic tests always have better control of Type I error inflation, the nonparametric tests seem the safest approach across most conditions.

Lastly, Winsorizing only one end seems better than both ends in controlling Type I error inflation and outlier impact on power. Therefore, it is recommended that when all outliers are on the same side, one-end Winsorizing is the most useful approach.

Both nonparametric and Winsorizing methods have similar effects in diminishing outlier impact on power, yet when deciding on an accommodation method, it is necessary to comprehensively consider both power and Type I error. Therefore, the nonparametric seems safest because the Type I error remains only a little inflated with more outliers but it generally has higher power.

Outliers will almost inevitable exist in educational datasets and, in practice, removing outliers is still a common approach (Bekker, 2014). It is therefore highly recommended to examine the reason for outlier occurrence and, if the reasons are obscure or cannot be traced, our recommendation is to retain the outlier and use appropriate outlier accommodation methods to minimize outlier impact in statistical testing.

Acknowledgments

This research was supported by School of English for Specific Purposes, Beijing Foreign Studies University, grant ZJ1513.

References

Allison, D. B, Gorman, B. S., & Primavera, 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), 153-185.

Analytical Methods Committee. (1989). Robust statistics-how not to reject outliers: Part 1. Basic concepts. *Analyst*, *114*, 1693-1697. doi: 10.1039/an9891401693

Bakker, M., & Wicherts, J. M. (2014). Supplemental material for outlier removal, sum scores, and the inflation of the type I error rate in independent samples t tests: The power of alternatives and recommendations. *Psychological Methods*. doi: 10.1037/met0000014.supp

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

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

Chow, S.-M., 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. Available from http://www.jstor.org/stable/2345711

Dixon, W. J., & Yuen, K. K. (1974). Trimming and Winsorization: A review. *Statistische Hefte, 15*(2), 157-170. doi: 10.1007/BF02922904

Freedman, D., Pisani, R., & Purves, R. (2007). *Statistics* (4th ed.). New York, NY: Norton.

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

Hampel, F. R. (2001). *Robust statistics: A brief introduction and overview*. Research Report No. 94. Zürich, Switzerland: Eidgenössische Technische Hochschule. Retrieved from ftp://ess.r-project.org/Research-Reports/94.pdf

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: 10.2307/2285654

Hoaglin, D. C., & Welsch, R. E. (1978). The hat matrix in regression and ANOVA. *The American Statistician, 32*(1), 17-22. doi: 10.2307/2683469

Jarrell, M. G. (1991, November). *Multivariate outliers: Review of the literature*. Paper presented at the Annual Meeting of the Mid-South educational research Association, Lexington, KY.

Li, Y., An, Q., Wanich, W., Lewis, M., Huang, Y., & Brooks, G. (2009, October). *Type I error rates and power for multiple comparisons when extreme values exist in data*. Paper presented at the Annual Meeting of the Mid-Western Educational Research Association, St. Louis, MO.

Liao, H., Li, Y., & Brooks, G. (2016). Outlier impact and accommodation methods: Multiple comparisons of type I error rates. *Journal of Modern Applied Statistical Methods, 15*(1), 452-471. Retrieved from http://digitalcommons.wayne.edu/jmasm/vol15/iss1/23

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

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

Tabachnick, B. G., & Fidell, L. S. (2001). *Using multivariate statistics*. New York, NY: Allyn & Bacon.

Warner R. M. (2012). *Applied statistics: From bivariate through multivariate techniques*. Thousand Oaks, CA: Sage.

Wilcox, 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

Wilcox, R. (2012). *Modern statistics for the social and behavioral sciences: A practical introduction*. Boca Raton, FL: CRC Press

Zimmerman, D. W. (1994). 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. *The 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