British Journal of Applied Science & Technology
4(1): 113-126, 2014

SCIENCEDOMAIN ***international***
www.sciencedomain.org

New Critical Values for the Winsorized *t*-Test

Michael Lance[1*], Piper Farrell-Singleton[2] and Shlomo S. Sawilowsky[1]

[1]Education Evaluation and Research, Wayne State University, Detroit, MI, 48282, USA.
[2]Office of Education Improvement and Innovation, Michigan Department of Education, Lansing, MI, 48909, USA.

Authors' contributions

This work was carried out in collaboration between all authors. Authors ML and SSS designed the study, conducted the literature review, wrote the Fortran code, carried out the Monte Carlo experiments, and wrote all versions of the manuscript. Authors PF-S and SSS designed and carried out the prior Monte Carlo study necessary to produce the new critical values for the Winsorized t-Test. All authors read and approved the final manuscript.

Research Article

Received 25th June 2013
Accepted 31st July 2013
Published 1st October 2013

ABSTRACT

Aims: To determine if (and in which situations) Monte Carlo or asymptotically derived critical values are more robust for the Winsorized t-test.
Study Design: A Monte Carlo simulation via FORTRAN 90 was used to test type I and II error properties across 14 unique distributions for various combinations of sample sizes and effect sizes for alpha = .01 and .05. Both Monte Carlo and asymptotically derived sets of critical values were used. Each combination of parameters was used to run 1 million iterations.
Place and Duration of Study: Windows PC for a duration of 6.5 days (to obtain results generated per each set of iterations).
Methodology: FORTRAN 90 code was used to do the following: For 1 (value) and 10% of n_1 + n_2, samples were drawn per distribution and Winsorized. Next, t-tests were conducted per the parameters specified above in the study design.
Results: Results generally supported the use of the new table of Monte Carlo derived critical values over the classical asymptotically-derived critical values.
Conclusion: The Monte Carlo-derived Winsorized critical values are generally preferable to asymptotically derived critical values.

**Corresponding author: E-mail: michael.lance@gmail.com;*

Keywords: Outliers; trimming; winsorizing; t-test; critical values; Monte Carlo simulation.

1. INTRODUCTION

The arithmetic mean is a well-known and “cherished” [1, p. 158] estimator of location due to its ease of calculation. However, it is not robust due to its finite breakdown point of $\frac{1}{N}$ [2] meaning that only one value, an outlier, can make the result arbitrarily large or small. Hawkins [3] noted that outliers occur due to typographical error, measurement error, and heavy-tailed population distribution. Grace and Sawilowsky [4, p. 306] maintained that when a substantial portion of a data set can be modeled by a different mean and variance it is dubious to view the dataset contaminated with outliers, and instead should be handled as a mixed distribution.

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) to increase the accuracy of the arithmetic mean. Trimming involves sorting an array of data and dropping the outliers. Winsorizing involves taking those same values that would otherwise have been trimmed and replacing them with the values that remain at the end(s) of the sorted, trimmed array. Whereas typographical error, measurement error, and minor contamination warrant trimming, Hawkins [3] suggested it is more appropriate to Winsorize outliers when the sample is drawn from a heavy-tailed distribution. Hence, the null hypothesis of the Winsorized *t*-test is that the Winsorized population means are the same. Dixon and Massey [5] 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. Rivest [6] showed that Winsorizing is particularly helpful for skewed distributions.

Some researchers prefer to use arbitrary rules of thumb in determining when to Winsorize or trim [7]. Maximum likelihood estimators, however, can be used to identify the exact number of values at each end of a sample to Winsorize and has been shown by Sawilowsky [8] to produce narrower bracketed (confidence) intervals with the real sets observed by Micceri [9] for sample sizes < 50 than either light or heavy trimming (as defined by Wilcox [10]).

In the context of a two sample test, Dixon & Tukey [11] proposed the Winsorized *t*

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

with the Winsorized variance

$$S^2_{wk} = (k+1)\left(x_{(k+1)} - \bar{x}_{wk}\right)^2 + \sum_{i=k+2}^{n-k-1}\left(x_{(i)} - \bar{x}_{wk}\right)^2 + (k+1)\left(x_{(n-k)} - \bar{x}_{wk}\right)^2 \qquad (2),$$

where $\bar{x}_w$ is the Winsorized mean, $y_1, \ldots, y_n$ are y ordered observations from a sample, and k is the number of Winsorized values, with $(h_1 + h_2) - 2$ degrees of freedom. (Alternative formulas were provided by Fung & Rahman [12], and Gans [13]). Obviously, if outliers are not representative and are kept, the test becomes biased [14]. Fung and Rahman [12] showed the trimmed and Winsorized *t*-tests to have immaterially small power differences. Yuen and Dixon [15] reached the same conclusion with trimming.

Unfortunately, the removal of outliers through Winsorizing results in a decrease in variability, and trimming results in both a decrease in degrees of freedom and variability. It is hypothesized that a power loss will surface in subsequent use of the classical asymptotically obtained critical *t* values, because the resultant values will be larger than they should be, producing conservative Type I errors and inflated Type II error rates. It is also postulated that using more precise critical values would better preserve correct Types I and II error rates, and hence, increase comparative statistical power.

1.1 Statement of the Problem

The robustness properties of the Winsorized *t* have been based on estimated $(h_1 + h_2) - 2$ degrees of freedom and traditional critical values [10]. The aim here is to compare them with Monte Carlo critical values recently derived by Farrell-Singleton [16]. The use of real social and behavioral science data sets [9] as the referent distributions will provide results researchers in these areas are likely to encounter. Various mathematical distributions will also be sampled, as is more commonly done in Monte Carlo studies.

2. METHODOLOGY

For each Monte Carlo study, the number of Winsorized outliers is equal across samples per iteration, even when samples are unbalanced. For example, when $n_1 + n_2 = 20$ and the Winsorized amount (denoted *k*) is 2 per end (10% of $n_1 + n_2$), 6 original values (denoted *h*) will remain per sample in addition to 4 Winsorized (*k* = 2) values. When this is applied to $n_1 = 5$ and $n_2 = 15$, n_1 will be composed of 5 of the same values per the middle value from the sorted sample, because two values are to be Winsorized at each end, leaving the middle value as the only original value to use in the process. This is the only instance in the study where Winsorizing will produce a sample composed of equal values throughout. In this case, n_2 is composed of 11 original values (*h* = 11) with 4 Winsorized (*k* = 2) values.

The characteristics of the data sets from Micceri [9] are depicted in Table 1 below.

Table 1. 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: Table adapted from Sawilowsky & Blair [17].

Data were also sampled from mathematical distributions, including Gaussian (normal), Cauchy, *t* ($\nu = 3$), Chi-Squared ($\nu = 2$), Exponential ($\mu = \sigma = 1$), and Uniform ($[0,1]$) for comparison with Boneau [18] and other published Monte Carlo studies. Effect sizes of 0.2σ, 0.5σ, 0.8σ, and 1.2σ [19] were modeled for the Type II error and comparative power portions of the study.

A program was written in a Fortran 2008 superset to conduct all simulations. The Rangen 2.0 subroutines [20], a Fortran 90/95 update from the original Fortran 77 version [21] provided the mathematical distributions. The Fortran 90 Realpops subroutines [22] was used to provide real data sets from Micceri [9]. For detailed information on how the Monte Carlo-derived critical values were generated, as compiled in Table 2 below, see Farrell-Singleton [16].

Table 2. Critical values for the two sample t and Winsorized t

x(n)	y(n)	Outliers	Df	0.01	0.01 ((*h*1+*h*2)-2 df)	.01 (Monte Carlo)	0.05	0.05 ((*h*1+*h*2)-2) df)	.05 (Monte Carlo)
5	5	1	8	3.36	4.60	9.38	2.31	2.78	5.42
5	15	1	18	2.88	2.98	3.81	2.10	2.14	2.72
5	15	2	18	2.88	3.17	5.71	2.10	2.23	3.93
10	10	1	18	2.88	2.98	3.81	2.10	2.14	2.72
10	10	2	18	2.88	3.17	5.71	2.10	2.23	3.93
15	15	1	28	2.76	2.80	3.24	2.05	2.06	2.39
15	15	3	28	2.76	2.92	5.07	2.05	2.12	3.64
10	30	1	38	2.71	2.73	3.04	2.02	2.03	2.25
10	30	4	38	2.71	2.82	4.82	2.02	2.07	3.51
20	20	1	38	2.71	2.73	3.04	2.02	2.03	2.25
20	20	4	38	2.71	2.82	4.82	2.02	2.07	3.51
25	25	1	48	2.68	2.69	2.93	2.01	2.02	2.19
25	25	5	48	2.68	2.76	4.68	2.01	2.05	3.44
15	45	1	58	2.66	2.67	2.85	2.00	2.00	2.14
15	45	6	58	2.66	2.73	4.59	2.00	2.03	3.40
30	30	1	58	2.66	2.67	2.85	2.00	2.00	2.14
30	30	6	58	2.66	2.73	4.59	2.00	2.03	3.40
45	45	1	88	2.63	2.64	2.76	1.99	1.99	2.08
30	90	1	118	2.62	2.62	2.71	1.98	1.98	2.05
30	90	12	118	2.62	2.65	4.39	1.98	1.99	3.30
60	60	1	118	2.62	2.62	2.71	1.98	1.98	2.05
60	60	12	118	2.62	2.65	4.39	1.98	1.99	3.30
90	90	1	178	2.60	2.60	2.67	1.97	1.97	2.02
90	90	18	178	2.60	2.62	4.31	1.97	1.98	3.26
120	120	1	238	2.60	2.60	2.64	1.97	1.97	2.00
120	120	24	238	2.60	2.61	4.29	1.97	1.98	3.25

Note. See Farrell-Singleton [16].

All descriptions of robustness refer to direction (conservative or liberal) and magnitude (liberal or stringent) according to Bradley's [23] definitional ranges. P-values between 0.9 α and 1.1 α ($|p-\alpha| \le \frac{\alpha}{10}$) are considered to meet a stringent criteria for robustness, whereas p-values between 0.5 α and 1.5 α ($|p-\alpha| \le \frac{\alpha}{2}$) (i.e. for α =.05, between .025 and .075) are considered to be meet a liberal criteria. To describe non-robust results that fall outside of Bradley's liberal range, the phrase "outside of the liberal range" is used.

The three sets of simulations being compared are:

1. Student's *t*-test with no outliers present in each sample.
2. Winsorized *t*-test with Dixon and Tukey's [11] ((h_1 + h_2) – 2 degrees of freedom) *t* critical values (outliers Winsorized).
3. Winsorized *t*-test with Monte Carlo critical values (outliers Winsorized).

3. RESULTS AND DISCUSSION

3.1 Type I Error: Without Winsorizing

These results verified previous findings of factors that impact robustness properties, such small sample size, the interaction of skew and unbalanced samples, and the Discrete Mass at Zero with Gap distribution results being less robust for smaller samples see [17].

The results obtained from sampling the mathematical distributions were also consistent with previous studies. Normal distribution results were generally stringently robust for the Monte Carlo critical values (as shown in Fig. 1) and all liberal, outside of Bradley's liberal range of robustness for approximate critical values.

3.2 Type I Error: With Winsorizing

Generally, Winsorizing samples and using alternative critical values led to less conservative (yet robust) p-values. Using the Monte Carlo-derived critical values produced results that were generally more robust than for the adjusted critical values as suggested by Dixon and Tukey [11].

The 10% 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). Such cases are outlined in Table 3.

In Fig. 2, results for the Discrete Mass at Zero with Gap distribution show how skew can interact with small sample size and unbalanced samples to impact robustness. Fig. 3 shows that increased Winsorizing can amplify the impact of these factors. Because 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 so than in the larger one. This created samples with unequal variances and negatively impacted robustness of the results. The role of skew in this respect was more pronounced with non-mathematical distributions.

Results were still generally more robust when Winsorizing and using the Winsorized (more so with Monte Carlo-derived) critical values. Fig. 4 shows just how vulnerable p-values are for the normal distribution with just one outlier per end.

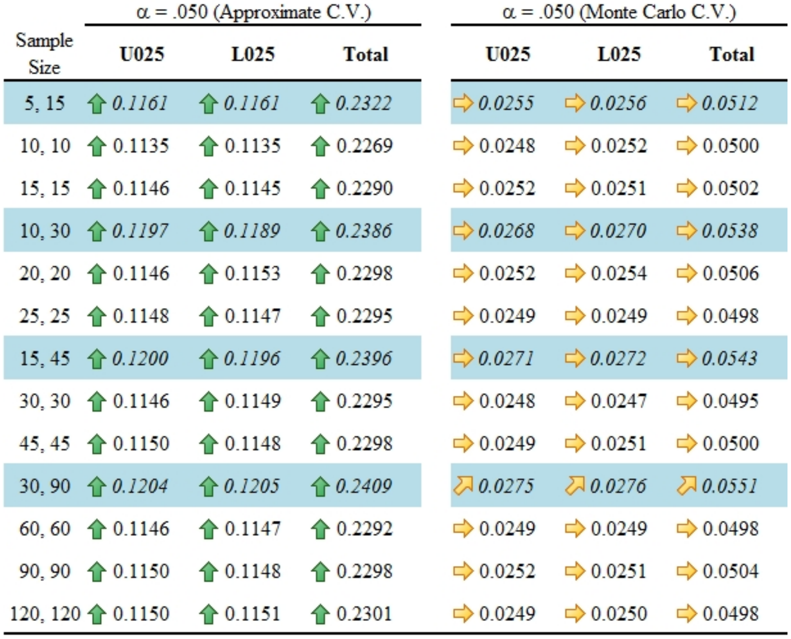
3.3 Type II Error: Without Winsorizing

In general, as Fig. 5 shows, an increase in effect size led to larger portions of data points that fell into the upper tails and less for the lower tails since larger effect sizes mean a greater the shift in mean (or distribution). When examining the results, it is noticeable that the Uniform distribution had the highest rejection (of the null) rates and the Cauchy distribution had 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 tended to be relatively close in rejection rates, which increased as a function of effect size, alpha level, and sample size.

3.4 Type II Error: With Winsorizing

Winsorizing, in tandem with using the new Monte Carlo-derived critical values, reduced type II error while providing more exact type I error rates. As previous research has shown, Winsorized results generally involved less type II error due to decreases in variance.

As with the results for type I error, increased Winsorizing amplified the impact of skew combined with unbalanced samples on type II error (Figs. 6 and 7). 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 shows that for distributions with high concentrations of values on the upper tail, unequal samples were a benefit to rejection rates and for those with concentrations on the lower tail, the opposite was true. Again, this trend is more pronounced among the non-mathematical distributions.



Sample Size	α = .050 (Approximate C.V.) U025	L025	Total	α = .050 (Monte Carlo C.V.) 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

Fig. 1. Type I error rates for independent-samples t test for various sample sizes and alpha levels when sampling is from a normal distribution, 1,000,000 repetitions, 10% *(of n_1 + n_2)* Winsorized critical values

Note: SU025 = proportion of rejections in upper-tail. L025 = proportion of rejections in the lower-tail. Excerpted from Lance [24, p. 60].

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

⇧ Liberal, Outside the Liberal Range ⬀ Liberal, Inside the Liberal Range

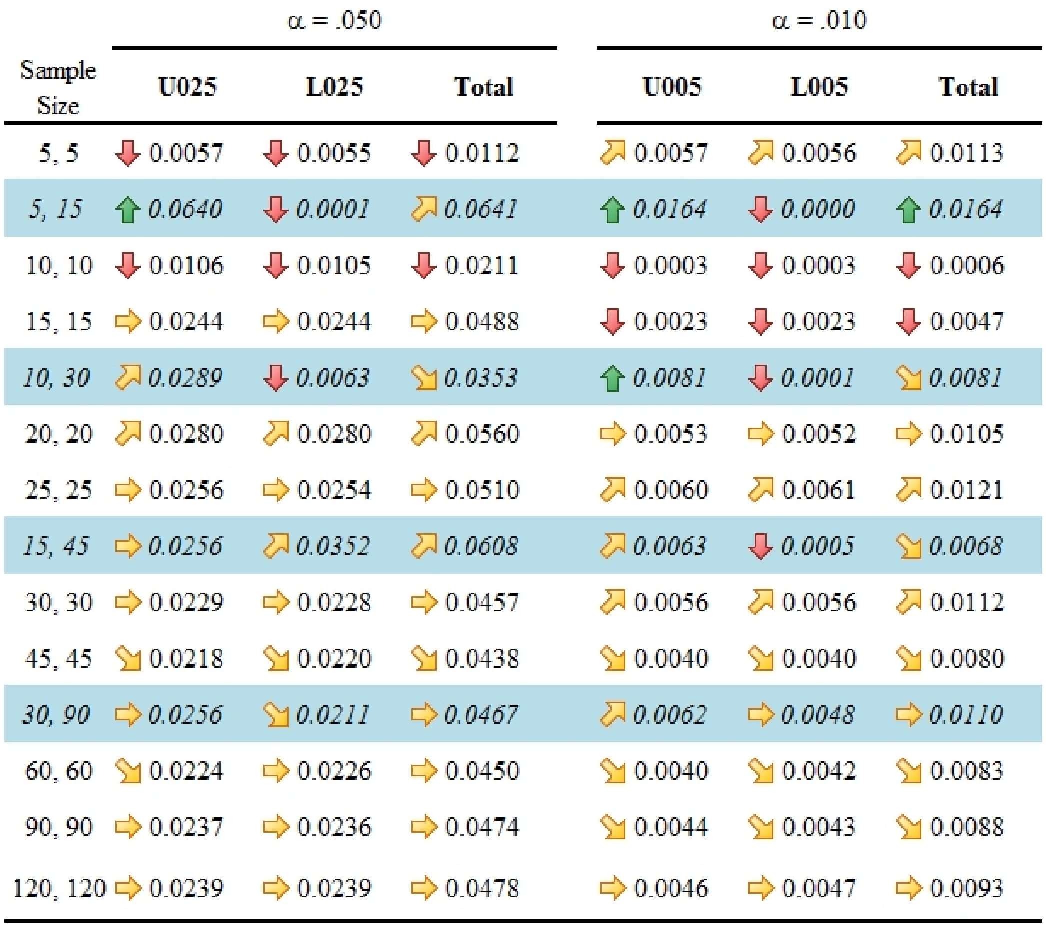
⇨ Within the Stringent Range

⇩ Conservative, Outside the Liberal Range ⬂ Conservative, Inside the Liberal Range

Table 3. Minimum sample sizes per distribution where approximate critical values are recommended (for *k*=1, n_1 = n_2)

Distribution	n_1 (α = .05)
Discrete mass at zero with gap	45
Mass at zero	n/a
Extreme asymmetry (P/D)	90
Extreme asymmetry (A/G)	90
Extreme bimodality	20
Multimodality and lumpy	45
Digit preference	90
Smooth symmetric	90
Normal	90
Uniform	45
Exponential	n/a
t with 3 degrees of freedom	n/a
Chi-squared with 2 degrees of freedom	n/a
Cauchy	n/a

Note: At α=.01, there were no sufficient trends to warrant such recommendations.

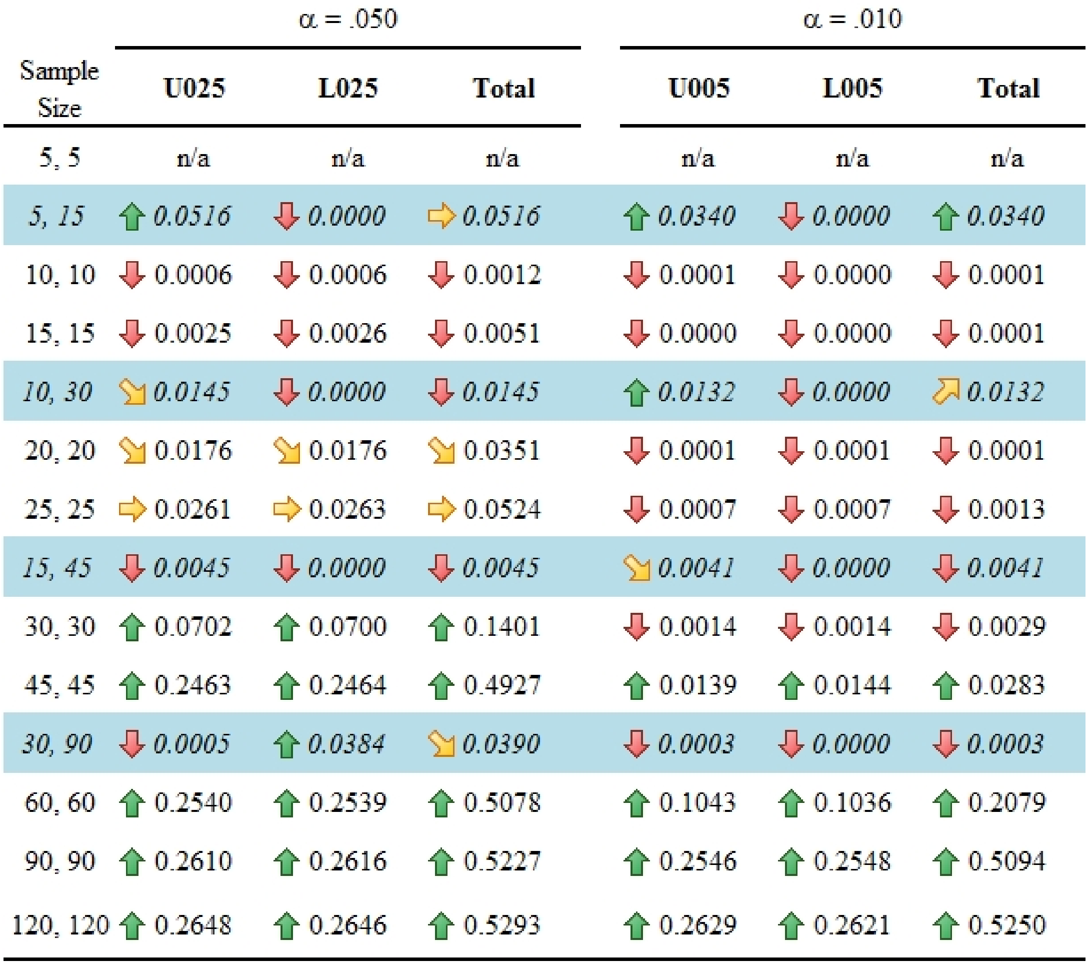


Sample Size	α = .050			α = .010		
	U025	L025	Total	U005	L005	Total
5, 5	↓ 0.0057	↓ 0.0055	↓ 0.0112	↗ 0.0057	↗ 0.0056	↗ 0.0113
5, 15	↑ *0.0640*	↓ *0.0001*	↗ *0.0641*	↑ *0.0164*	↓ *0.0000*	↑ *0.0164*
10, 10	↓ 0.0106	↓ 0.0105	↓ 0.0211	↓ 0.0003	↓ 0.0003	↓ 0.0006
15, 15	⇨ 0.0244	⇨ 0.0244	⇨ 0.0488	↓ 0.0023	↓ 0.0023	↓ 0.0047
10, 30	↗ *0.0289*	↓ *0.0063*	↘ *0.0353*	↑ *0.0081*	↓ *0.0001*	↘ *0.0081*
20, 20	↗ 0.0280	↗ 0.0280	↗ 0.0560	⇨ 0.0053	⇨ 0.0052	⇨ 0.0105
25, 25	⇨ 0.0256	⇨ 0.0254	⇨ 0.0510	↗ 0.0060	↗ 0.0061	↗ 0.0121
15, 45	⇨ *0.0256*	↗ *0.0352*	↗ *0.0608*	↗ *0.0063*	↓ *0.0005*	↘ *0.0068*
30, 30	⇨ 0.0229	⇨ 0.0228	⇨ 0.0457	↗ 0.0056	↗ 0.0056	↗ 0.0112
45, 45	↘ 0.0218	↘ 0.0220	↘ 0.0438	↘ 0.0040	↘ 0.0040	↘ 0.0080
30, 90	⇨ *0.0256*	↘ *0.0211*	⇨ *0.0467*	↗ *0.0062*	⇨ *0.0048*	⇨ *0.0110*
60, 60	↘ 0.0224	⇨ 0.0226	⇨ 0.0450	↘ 0.0040	↘ 0.0042	↘ 0.0083
90, 90	⇨ 0.0237	⇨ 0.0236	⇨ 0.0474	↘ 0.0044	↘ 0.0043	↘ 0.0088
120, 120	⇨ 0.0239	⇨ 0.0239	⇨ 0.0478	⇨ 0.0046	⇨ 0.0047	⇨ 0.0093

Fig. 2. 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, 1,000,000 repetitions, 1 outlier/end Winsorized, Monte Carlo C.V

Note: SU025 = proportion of rejections in upper-tail. L025 = proportion of rejections in the lower-tail.
Excerpted from Lance [24, p. 47]
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, Outside the Liberal Range ↘ Conservative, Inside the Liberal Range



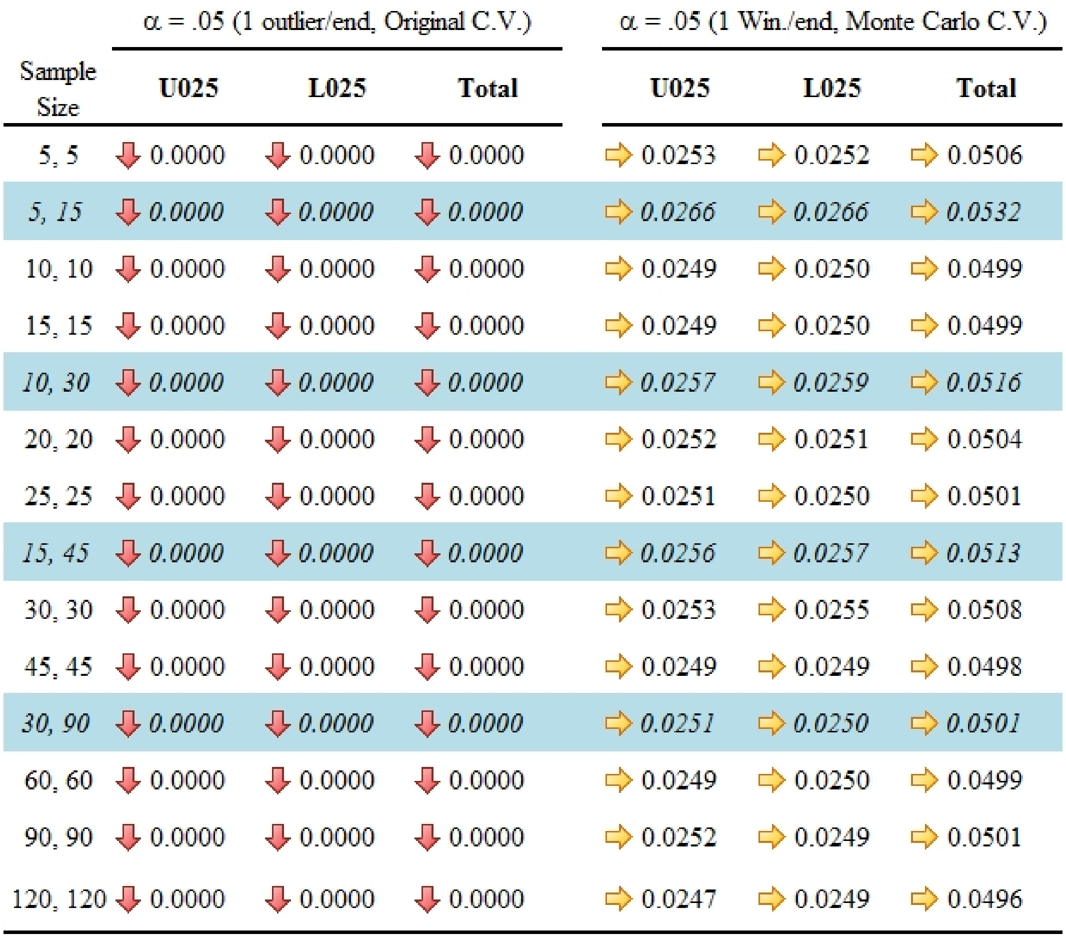
	α = .050			α = .010		
Sample Size	U025	L025	Total	U005	L005	Total
5, 5	n/a	n/a	n/a	n/a	n/a	n/a
5, 15	↑ *0.0516*	↓ *0.0000*	→ *0.0516*	↑ *0.0340*	↓ *0.0000*	↑ *0.0340*
10, 10	↓ 0.0006	↓ 0.0006	↓ 0.0012	↓ 0.0001	↓ 0.0000	↓ 0.0001
15, 15	↓ 0.0025	↓ 0.0026	↓ 0.0051	↓ 0.0000	↓ 0.0000	↓ 0.0001
10, 30	↘ *0.0145*	↓ *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

Fig. 3. 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, 1,000,000 repetitions, 10% Winsorized Outliers, Monte Carlo C.V

Note: SU025 = proportion of rejections in upper-tail. L025 = proportion of rejections in the lower-tail. Excerpted from Lance [24, p. 48].

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, Outside the Liberal Range	↘ Conservative, Inside the Liberal Range



Sample Size	α = .05 (1 outlier/end, Original C.V.)			α = .05 (1 Win./end, Monte Carlo C.V.)		
	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

Fig. 4. Type I error rates for independent-samples t test for various sample sizes and alpha levels when sampling is from a normal distribution, 1,000,000 repetitions, 1 outlier vs. 1 Winsorized value (both per end)

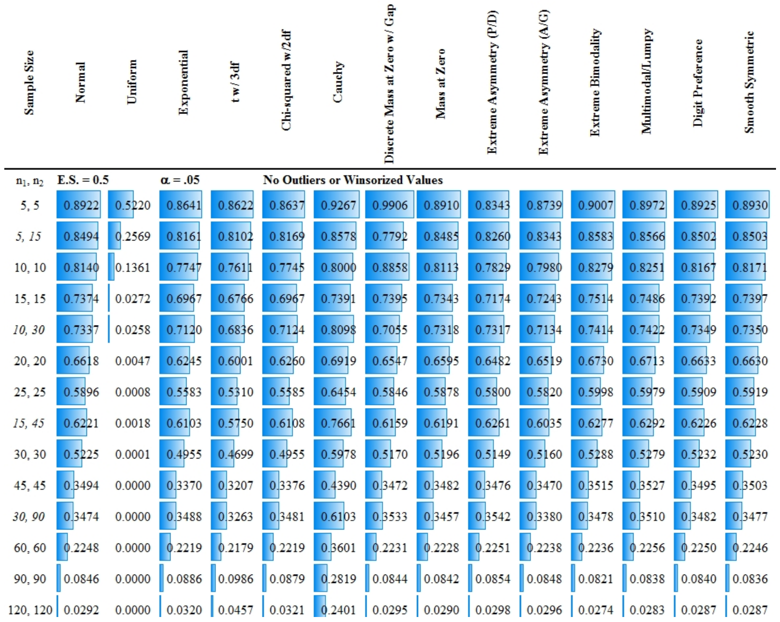
Note: SU025 = proportion of rejections in upper-tail. L025 = proportion of rejections in the lower-tail. Excerpted from Lance [24, p. 49].

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

⬆ Liberal, Outside the Liberal Range ↗ Liberal, Inside the Liberal Range

➡ Within the Stringent Range

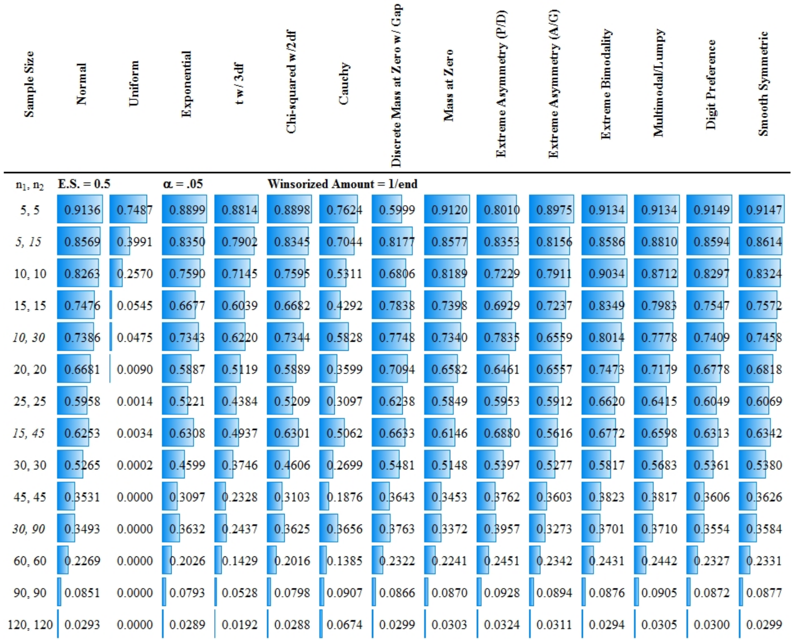
⬇ Conservative, Outside the Liberal Range ↘ Conservative, Inside the Liberal Range



Sample Size	Normal	Uniform	Exponential	t w/ 3df	Chi-squared w/2df	Cauchy	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
n_1, n_2	**E.S. = 0.5**		**α = .05**		**No Outliers or Winsorized Values**									
5, 5	0.8922	0.5220	0.8641	0.8622	0.8637	0.9267	0.9906	0.8910	0.8343	0.8739	0.9007	0.8972	0.8925	0.8930
5, 15	0.8494	0.2569	0.8161	0.8102	0.8169	0.8578	0.7792	0.8485	0.8260	0.8343	0.8583	0.8566	0.8502	0.8503
10, 10	0.8140	0.1361	0.7747	0.7611	0.7745	0.8000	0.8858	0.8113	0.7829	0.7980	0.8279	0.8251	0.8167	0.8171
15, 15	0.7374	0.0272	0.6967	0.6766	0.6967	0.7391	0.7395	0.7343	0.7174	0.7243	0.7514	0.7486	0.7392	0.7397
10, 30	0.7337	0.0258	0.7120	0.6836	0.7124	0.8098	0.7055	0.7318	0.7317	0.7134	0.7414	0.7422	0.7349	0.7350
20, 20	0.6618	0.0047	0.6245	0.6001	0.6260	0.6919	0.6547	0.6595	0.6482	0.6519	0.6730	0.6713	0.6633	0.6630
25, 25	0.5896	0.0008	0.5583	0.5310	0.5585	0.6454	0.5846	0.5878	0.5800	0.5820	0.5998	0.5979	0.5909	0.5919
15, 45	0.6221	0.0018	0.6103	0.5750	0.6108	0.7661	0.6159	0.6191	0.6261	0.6035	0.6277	0.6292	0.6226	0.6228
30, 30	0.5225	0.0001	0.4955	0.4699	0.4955	0.5978	0.5170	0.5196	0.5149	0.5160	0.5288	0.5279	0.5232	0.5230
45, 45	0.3494	0.0000	0.3370	0.3207	0.3376	0.4390	0.3472	0.3482	0.3476	0.3470	0.3515	0.3527	0.3495	0.3503
30, 90	0.3474	0.0000	0.3488	0.3263	0.3481	0.6103	0.3533	0.3457	0.3542	0.3380	0.3478	0.3510	0.3482	0.3477
60, 60	0.2248	0.0000	0.2219	0.2179	0.2219	0.3601	0.2231	0.2228	0.2251	0.2238	0.2236	0.2256	0.2250	0.2246
90, 90	0.0846	0.0000	0.0886	0.0986	0.0879	0.2819	0.0844	0.0842	0.0854	0.0848	0.0821	0.0838	0.0840	0.0836
120, 120	0.0292	0.0000	0.0320	0.0457	0.0321	0.2401	0.0295	0.0290	0.0298	0.0296	0.0274	0.0283	0.0287	0.0287

Fig. 5. Type II error rates for independent-samples t test (no outliers) for various sample sizes, effect sizes, and distributions (1,000,000 repetitions)

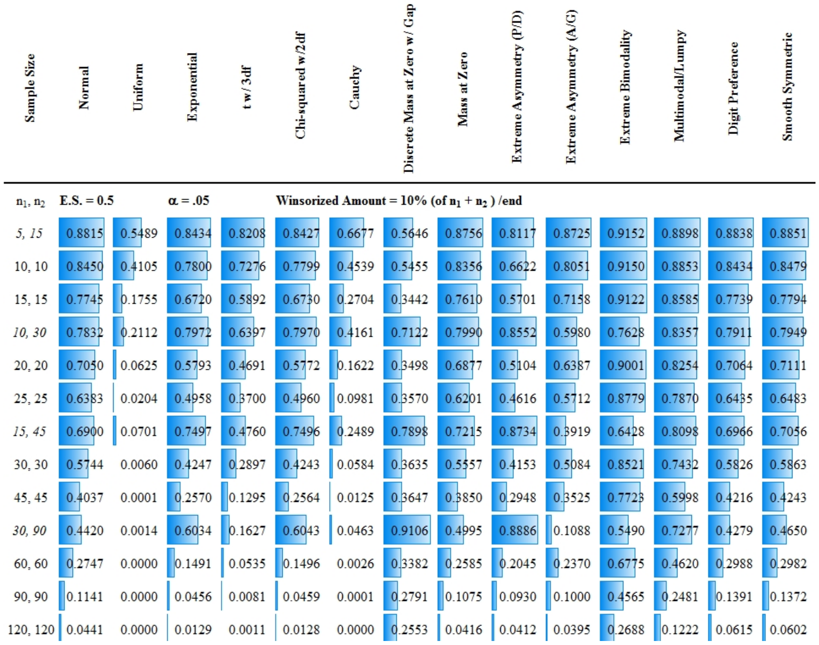
Note: Longer bars indicate higher rejection rates. Adapted from Lance [24, p. 50]



Sample Size	Normal	Uniform	Exponential	t w/ 3df	Chi-squared w/2df	Cauchy	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
n_1, n_2	**E.S. = 0.5**		**α = .05**		**Winsorized Amount = 1/end**									
5, 5	0.9136	0.7487	0.8899	0.8814	0.8898	0.7624	0.5999	0.9120	0.8010	0.8975	0.9134	0.9134	0.9149	0.9147
5, 15	0.8569	0.3991	0.8350	0.7902	0.8345	0.7044	0.8177	0.8577	0.8353	0.8156	0.8586	0.8810	0.8594	0.8614
10, 10	0.8263	0.2570	0.7590	0.7145	0.7595	0.5311	0.6806	0.8189	0.7229	0.7911	0.9034	0.8712	0.8297	0.8324
15, 15	0.7476	0.0545	0.6677	0.6039	0.6682	0.4292	0.7838	0.7398	0.6929	0.7237	0.8349	0.7983	0.7547	0.7572
10, 30	0.7386	0.0475	0.7343	0.6220	0.7344	0.5828	0.7748	0.7340	0.7835	0.6559	0.8014	0.7778	0.7409	0.7458
20, 20	0.6681	0.0090	0.5887	0.5119	0.5889	0.3599	0.7094	0.6582	0.6461	0.6557	0.7473	0.7179	0.6778	0.6818
25, 25	0.5958	0.0014	0.5221	0.4384	0.5209	0.3097	0.6238	0.5849	0.5953	0.5912	0.6620	0.6415	0.6049	0.6069
15, 45	0.6253	0.0034	0.6308	0.4937	0.6301	0.5062	0.6633	0.6146	0.6880	0.5616	0.6772	0.6598	0.6313	0.6342
30, 30	0.5265	0.0002	0.4599	0.3746	0.4606	0.2699	0.5481	0.5148	0.5397	0.5277	0.5817	0.5683	0.5361	0.5380
45, 45	0.3531	0.0000	0.3097	0.2328	0.3103	0.1876	0.3643	0.3453	0.3762	0.3603	0.3823	0.3817	0.3606	0.3626
30, 90	0.3493	0.0000	0.3632	0.2437	0.3625	0.3656	0.3763	0.3372	0.3957	0.3273	0.3701	0.3710	0.3554	0.3584
60, 60	0.2269	0.0000	0.2026	0.1429	0.2016	0.1385	0.2322	0.2241	0.2451	0.2342	0.2431	0.2442	0.2327	0.2331
90, 90	0.0851	0.0000	0.0793	0.0528	0.0798	0.0907	0.0866	0.0870	0.0928	0.0894	0.0876	0.0905	0.0872	0.0877
120, 120	0.0293	0.0000	0.0289	0.0192	0.0288	0.0674	0.0299	0.0303	0.0324	0.0311	0.0294	0.0305	0.0300	0.0299

Fig. 6. Type II error rates for Winsorized independent-samples t test (using monte carlo critical values) for various sample sizes, effect sizes, and distributions (1,000,000 repetitions).

Note: Longer bars indicate higher rejection rates. Adapted from Lance [24, p. 56].



Sample Size	Normal	Uniform	Exponential	t w/3df	Chi-squared w/2df	Cauchy	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
n_1, n_2	**E.S. = 0.5**		**α = .05**		**Winsorized Amount = 10% (of $n_1 + n_2$) /end**									
5, 15	0.8815	0.5489	0.8434	0.8208	0.8427	0.6677	0.5646	0.8756	0.8117	0.8725	0.9152	0.8898	0.8838	0.8851
10, 10	0.8450	0.4105	0.7800	0.7276	0.7799	0.4539	0.5455	0.8356	0.6622	0.8051	0.9150	0.8853	0.8434	0.8479
15, 15	0.7745	0.1755	0.6720	0.5892	0.6730	0.2704	0.3442	0.7610	0.5701	0.7158	0.9122	0.8585	0.7739	0.7794
10, 30	0.7832	0.2112	0.7972	0.6397	0.7970	0.4161	0.7122	0.7990	0.8552	0.5980	0.7628	0.8357	0.7911	0.7949
20, 20	0.7050	0.0625	0.5793	0.4691	0.5772	0.1622	0.3498	0.6877	0.5104	0.6387	0.9001	0.8254	0.7064	0.7111
25, 25	0.6383	0.0204	0.4958	0.3700	0.4960	0.0981	0.3570	0.6201	0.4616	0.5712	0.8779	0.7870	0.6435	0.6483
15, 45	0.6900	0.0701	0.7497	0.4760	0.7496	0.2489	0.7898	0.7215	0.8734	0.3919	0.6428	0.8098	0.6966	0.7056
30, 30	0.5744	0.0060	0.4247	0.2897	0.4243	0.0584	0.3635	0.5557	0.4153	0.5084	0.8521	0.7432	0.5826	0.5863
45, 45	0.4037	0.0001	0.2570	0.1295	0.2564	0.0125	0.3647	0.3850	0.2948	0.3525	0.7723	0.5998	0.4216	0.4243
30, 90	0.4420	0.0014	0.6034	0.1627	0.6043	0.0463	0.9106	0.4995	0.8886	0.1088	0.5490	0.7277	0.4279	0.4650
60, 60	0.2747	0.0000	0.1491	0.0535	0.1496	0.0026	0.3382	0.2585	0.2045	0.2370	0.6775	0.4620	0.2988	0.2982
90, 90	0.1141	0.0000	0.0456	0.0081	0.0459	0.0001	0.2791	0.1075	0.0930	0.1000	0.4565	0.2481	0.1391	0.1372
120, 120	0.0441	0.0000	0.0129	0.0011	0.0128	0.0000	0.2553	0.0416	0.0412	0.0395	0.2688	0.1222	0.0615	0.0602

Fig. 7. Type II error rates for Winsorized independent-samples t test (using monte carlo critical values) for various sample sizes, effect sizes, and distributions (1,000,000 repetitions)

Note: Longer bars indicate higher rejection rates. Adapted from Lance [24, p. 57].

4. CONCLUSION

When data are non-normally distributed, non-parametric tests are robust and powerful alternatives when the treatment alternative is a shift in means [25,26]. However, when the data essentially follow a parametric model with perturbations, the classical normal theory tests remain useful via trimming or Winsorizing. The purpose of this study, therefore, was to compare approximate critical values due to Dixon and Tukey [11] and Monte Carlo-derived critical values for the Winsorized *t* test for independent samples with respect to robustness to Type I and II errors. The Monte Carlo-derived Winsorized critical values produced more robust Type I error rates than using the Dixon and Tukey estimation, and led to dramatic improvement in Type I and II errors (Fig. 1).

Because 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. In general, the Monte Carlo-derived critical values are recommended for 10% (of n_1 + n_2) Winsorized samples.

COMPETING INTERESTS

Authors have declared that no competing interests exist.

REFERENCES

1. Dixon, WJ, Yuen, KK. Trimming and winsorization: a review. Statistische Hefte. 1974;2:157-170.
2. Wilcox RR. Statistics for the social sciences. San Diego: Academic Press; 1996.
3. Hawkins, DM. Identification of outliers. London: Chapman & Hall; 1980.
4. Grace TA, Sawilowsky SS. Data error prevention and cleansing: a comprehensive guide for instructors of statistics and their students. Model Assisted Statistics and Applications. 2009;4:303-312.
5. Dixon WJ, Massey FJ. Introduction to statistical analysis, McGraw-Hill, New York; 1969.
6. Rivest L. Statistical properties of winsorized means for skewed distributions. Biometrika. 1994;81(2):373-383.
7. Carey VJ, Walters, EE, Wager, CG, Rosner, BA. Resistant and test-based outlier rejection: effects on Gaussian one- and two-sample inference. Technometrics. 1997;39(3):320-330.
8. Sawilowsky, SS. A measure of relative efficiency for location of a single sample. Journal of Modern Applied Statistical Methods. 2002;1(1):52-60.
9. Micceri T. The unicorn, the normal curve, and other improbable creatures. Psychological Bulletin. 1989;105:156-166.
10. Wilcox RR. How many discoveries have been lost by ignoring modern statistical methods? American Psychologist. 1998;53(3):300-314.
11. Dixon WJ, Tukey JW. Approximate behavior of the distribution of winsorized t (trimming/winsorization 2). Technometrics. 1968;10(1):83-98.
12. Fung KY, Rahman SM. The two-wample winsorized t. Communications in Statistics: Simulation and Computation. 1980;89(4):337-347.
13. Gans D. Trimmed and winsorized means, tests for. In: Kots S, Johnson N, editors. Encyclopedia of statistical sciences. New York: Wiley; 1988;9.
14. Dixon WJ. Analysis of extreme values. The Annals of Mathematical Statistics. 1950;21(4):488-506.
15. Yuen KK, Dixon WJ. The approximate behavior and performance of the two-sample trimmed t. Biometrika. 1973;60:369-374.
16. Farrell-Singleton P. Critical values for the two independent samples winsorized t-test. Unpublished Doctoral Dissertation. Wayne State University; 2010.
17. Sawilowsky SS, Blair RC. A more realistic look at the robustness and Type II error properties of the t-test to departures from population normality. Psychological Bulletin. 1992;111:353-360.
18. Boneau, CA. The effects of violations of assumptions underlying the t test. Psychological Bulletin. 1960;57:49-64.
19. Sawilowsky SS. New effect size rules of thumb. Journal of Modern Applied Statistical Methods. 2009;8(2):597-599.
20. Fahoome GC. JMASM algorithms and code JMASM1: Rangen 2.0 (Fortran 90/95). Journal Of Modern Applied Statistical Methods. 2002;1:182-190.
21. Blair RC. Rangen: Version 1.0. Boca Raton, FL: IBM, 1987.
22. Sawilowsky SS, Fahoome GC. Statistics via monte carlo simulation with fortran. Rochester Hills, MI: JMASM; 2003
23. Bradley JV. Robustness? British Journal Mathematical and Statistical Psychology. 1978;31:114-152.
24. Lance MW. Approximate vs. monte carlo critical values for the winsorized t-test. Unpublished Doctoral Dissertation. Wayne State University; 2011.

25. Sawilowsky SS. Misconceptions leading to choosing the t test over the wilcoxon mann-whitney u test for shift in location parameter. Journal of Modern Applied Statistical Methods. 2005;4(2):598-600.
26. Fatal-Weber M, Sawilowsky SS. Comparative statistical power of the independent t, permutation t, and wilcoxon tests. Journal of Modern Applied Statistical Methods. 2009;8(1):21-26.

Peer-review history:
The peer review history for this paper can be accessed here:
http://www.sciencedomain.org/review-history.php?iid=268&id=5&aid=2100