Journal of Modern Applied Statistical Methods

Volume 9 | Issue 1 Article 4

5-1-2010

The Influence of Data Generation on Simulation Study Results: Tests of Mean Differences

Tim Moses
Educational Testing Service, Princeton, NJ, tmoses@ets.org

Alan Klockars
University of Washington, klockars@u.washington.edu

Journal of Modern Applied Statistical Methods
May 2010, Vol. 9, No. 1, 15-27

1538 – 9472/10/$95.00

REGULAR ARTICLES

The Influence of Data Generation on Simulation Study Results: Tests of Mean Differences

Tim Moses
Educational Testing Service,
Princeton, NJ

Alan Klockars
University of Washington

Type I error and power of the standard independent samples t-test were compared with the trimmed and Winsorized t-test with respect to continuous distributions and various discrete distributions known to occur in applied data. The continuous and discrete distributions were generated with similar levels of skew and kurtosis but the discrete distributions had a variety of structural features not reflected in the continuous distributions. The results showed that the Type I error rates of the t-tests were not seriously affected, but the power rate of the trimmed and Winsorized t-test varied greatly across the considered distributions.

Key words: Nonnormality, independent samples t-test, trimming, Winsorizing.

Introduction

Monte Carlo simulation studies are commonly used to assess the performance of statistical strategies under defined and controlled conditions. Often the question of interest involves the performance of one or more strategies under violations of the assumptions associated with the mathematical model on which a procedure is based. While simulation studies are informative, their conditions and results may be generated in ways that are not relevant for applied research settings. Of particular concern is the accuracy of simulation studies' recommendations about the impact of assumption violations in continuous and unbounded distributions for applied distributions that are primarily discrete and bounded.

A number of traditional statistical procedures assume a normal distribution for the underlying population from which scores were drawn (e.g., t-test, ANOVA). In simulation studies that evaluate the robustness of statistical significance tests of mean differences, nonnormality is usually created in smooth, continuous and theoretically unbounded distributions. Several methods exist for transforming normally distributed random numbers into nonnormal distributions, including Hoaglin's (1985) g and h method, Fleishman's (1978) power method, and the use of Chi-square distributions with varying degrees of freedom.

The nonnormality generated with these methods can primarily be defined in terms of skew and kurtosis. In contrast to simulated data, applied distributions of psychometric tests and achievement tests are usually discrete with bounded score ranges and are noted to have features such as lumps, bimodalities, or popular, unpopular or impossible scores (Holland & Thayer, 2000; Micceri, 1989). While these discrete distributions can be described in terms of their skew and kurtosis, a complete description would require more attention to their structural features. Continuous and discrete

Tim Moses is a Senior Psychometrician at Educational Testing Service where he works on several testing programs. He completed his Ph.D. in Educational Psychology at the University of Washington. Email: tmoses@ets.org. Alan Klockars is Professor of Educational Psychology at the University of Washington. His research concerns multiple comparisons and, more recently, methods of conducting ATI research. Email: klockars@u.washington.edu.

distributions with similar skew and kurtosis can reflect very different shapes.

Simulation studies that have evaluated significance tests of mean differences for nonnormal continuous distributions have produced different recommendations than simulation studies that consider nonnormal discrete distributions. Studies based on nonnormal continuous distributions have recommended that standard tests of mean differences be abandoned in favor of robust tests of trimmed mean differences (Keselman, Othman, Wilcox & Fradette, 2004; Lix & Keselman, 1998). In contrast, Sawilowsky and Blair (1992) used a variety of discrete distributions as population distributions and found that the standard t-test's Type I error rate was relatively unaffected by their populations.

The interest of this study is to investigate how the data generation method and population distributions used in a simulation study influence the results and recommendations of statistical strategies. Data were generated from the continuous distributions commonly considered in simulation studies and from various discrete and bounded distributions noted to occur in applied data (Holland & Thayer, 2000; Micceri, 1989; Sawilowsky, & Blair, 1992). The continuous and discrete distributions were generated with similar levels of skew and kurtosis but the discrete distributions had structural features not reflected in the continuous distributions.

Type I error and power were assessed in the standard independent samples t-test and one of its most recommended alternatives for nonnormal data, Yuen's (1974) trimmed and Winsorized t-test (Keselman, et al., 2004). In addition, this article considers the relevance of simulation studies' recommendations of statistical strategies for applied data.

Methodology

The objective of this study was to compare the Type I error and power rates for the standard t-test and the trimmed and Winsorized t-test when used to compare means in discrete distributions noted to occur in applied data and in continuous distributions of equal skew and kurtosis typically considered in simulation studies. The Type I error and power rates were computed from 10,000 replications where in each replication two random samples of size 30 were drawn from one of nine population distributions and the groups' means were compared using the standard t-test and the trimmed and Winsorized t-test. The nine population distributions included one continuous distribution and three discrete distributions of symmetric shape and one continuous distribution and four discrete distributions of asymmetric shape.

Population Distributions

The population distributions reflected two basic shapes, asymmetric and symmetric. The two shapes were modeled with bounded and discrete distributions and one accompanying continuous distribution. The asymmetric shape is skewed (approximately −1.75) and leptokurtic (kurtosis approximately 3.75). The asymmetric continuous and unbounded population distribution is shown in Figure 1. One of the asymmetric discrete distributions is smooth (Figure 2), and the others have structures such as teeth (Figure 3), a lump at score zero (Figure 4) and favorite scores (Figure 5). The means, standard deviations, skews and kurtosis of these five distributions are summarized in Table 1.

The symmetric distributions included three discrete and bounded distributions and one continuous and unbounded distribution (Table 2, Figures 6-9). All four symmetric distributions have skews of 0. The symmetric continuous distribution is shown in Figure 6. One of the symmetric discrete distributions is smooth (Figure 7); the others have peaks (Figure 8) and bimodality (Figure 9).

Data Generation Methods

The first data generation method produced data (i.e., *Y* scores for two groups) that reflected the discreteness and shapes of the discrete distributions where only the integer scores in defined score ranges were possible and where each possible score had a corresponding population probability (Figures 2-5 & 7-9). Samples of 30 scores were randomly drawn from these population distributions with the scores' population probabilities defining the probabilities of those scores appearing in the sample datasets.

Table 1: Summary Statistics for Four Negatively Skewed Discrete Distributions and One Continuous Distribution

Distribution	Mean	Std. Dev.	Skew	Kurtosis
Continuous	15.00	4.00	-1.75	3.75
Smooth & Discrete	15.73	2.90	-1.85	3.88
Teeth	14.46	3.45	-1.81	3.94
Lump at Zero	12.08	3.79	-1.97	3.85
Favorite Scores	17.36	4.13	-1.92	3.73

Table 2: Summary Statistics for Three Symmetric Discrete Distributions and One Continuous Distribution

Distribution	Mean	Std. Dev.	Skew	Kurtosis
Continuous	15.00	4.00	0.00	0.00
Smooth & Discrete	15.00	4.00	0.00	-0.15
7 Peaks	10.50	4.88	0.00	0.06
Bimodal	15.00	6.42	0.00	-1.18

Figure 1: Asymmetric Continuous Distribution

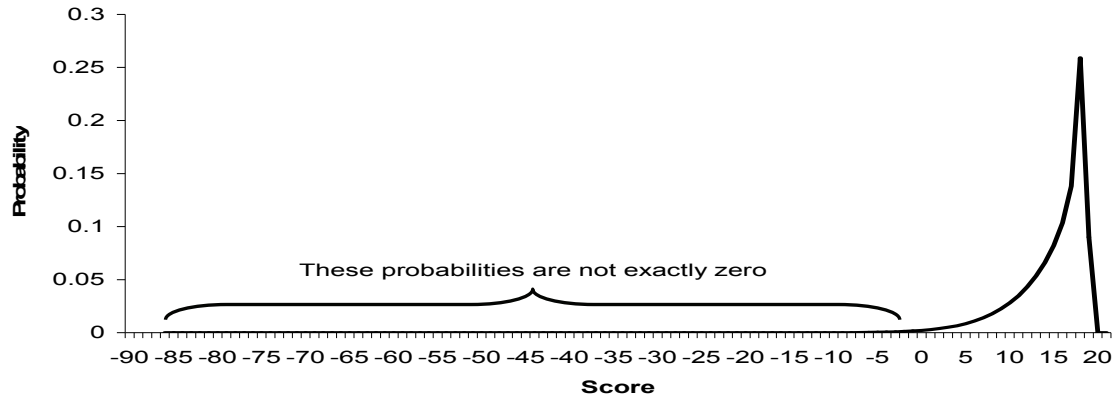


Figure 2: Asymmetric Smooth & Discrete Distribution

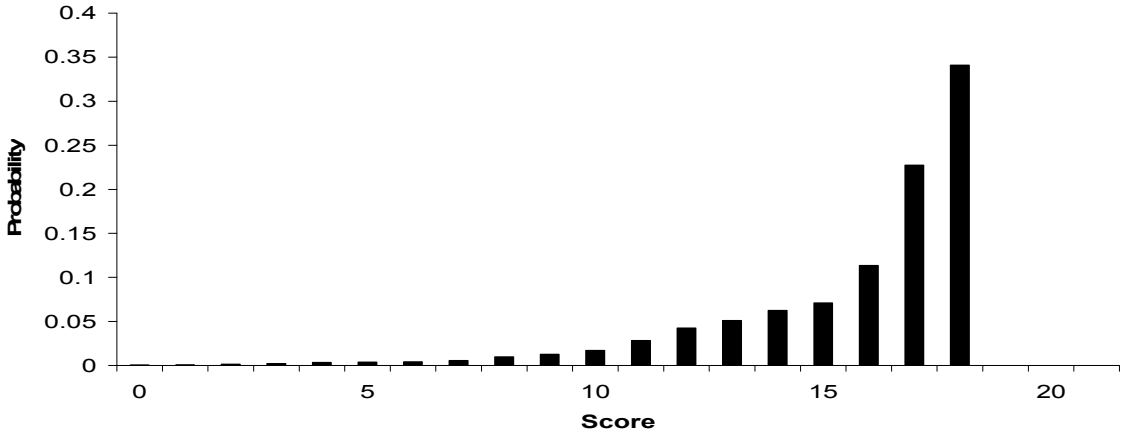


Figure 3: Asymmetric Teeth Distribution

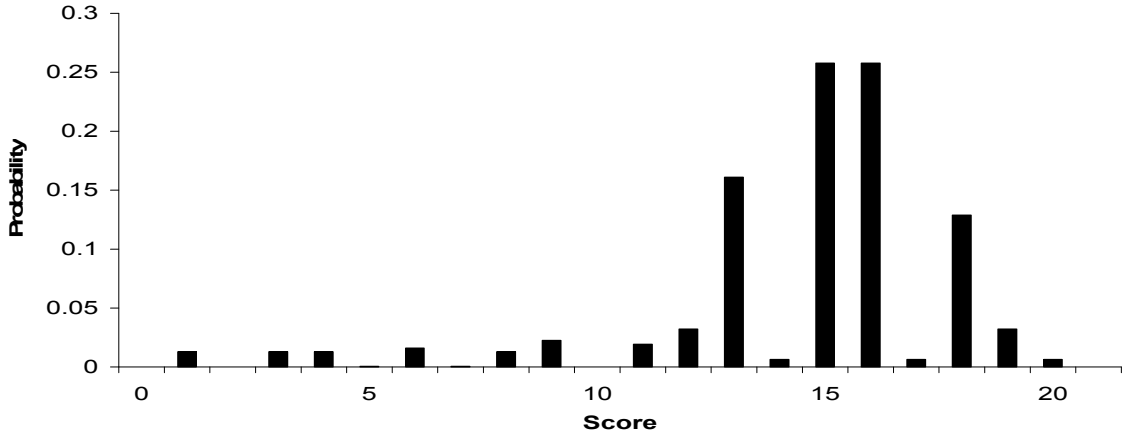


Figure 4: Asymmetric Lump at Zero Distribution

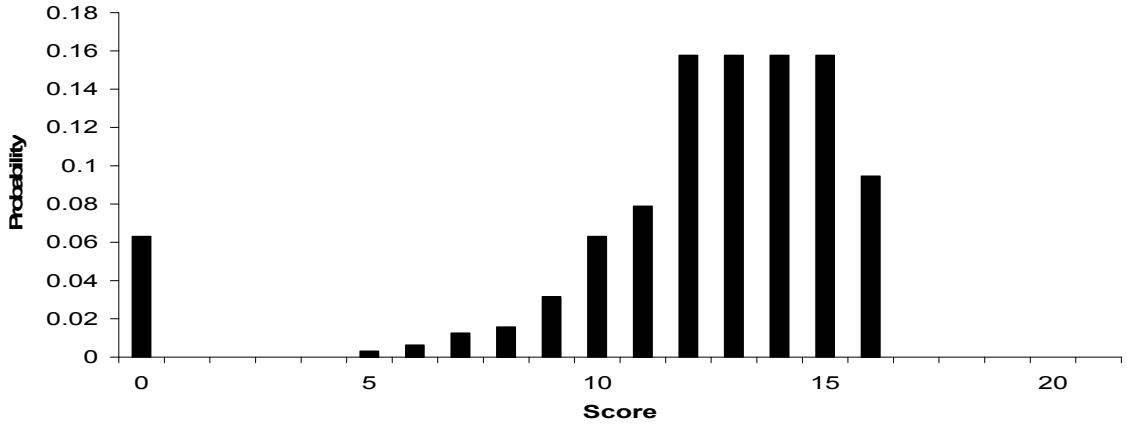


Figure 5: Asymmetric Favorite Scores Distribution

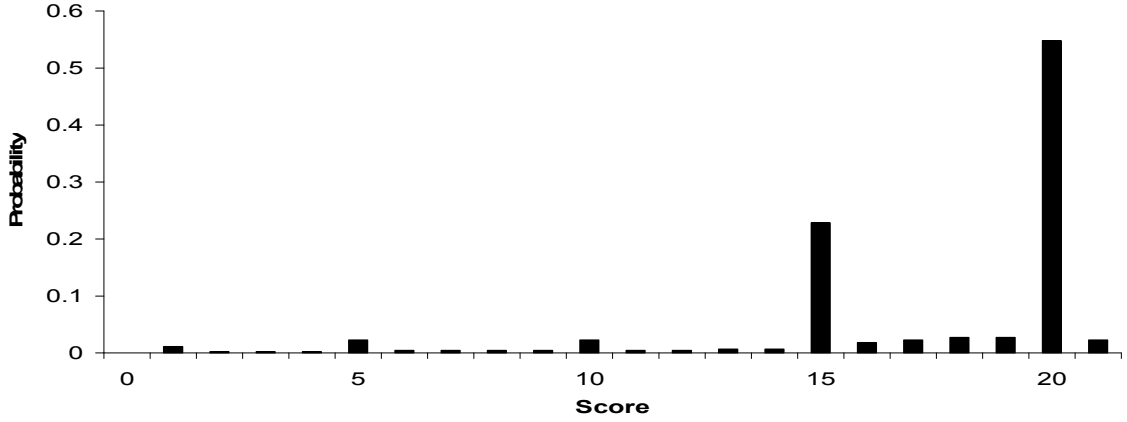


Figure 6: Symmetric Continuous Distribution

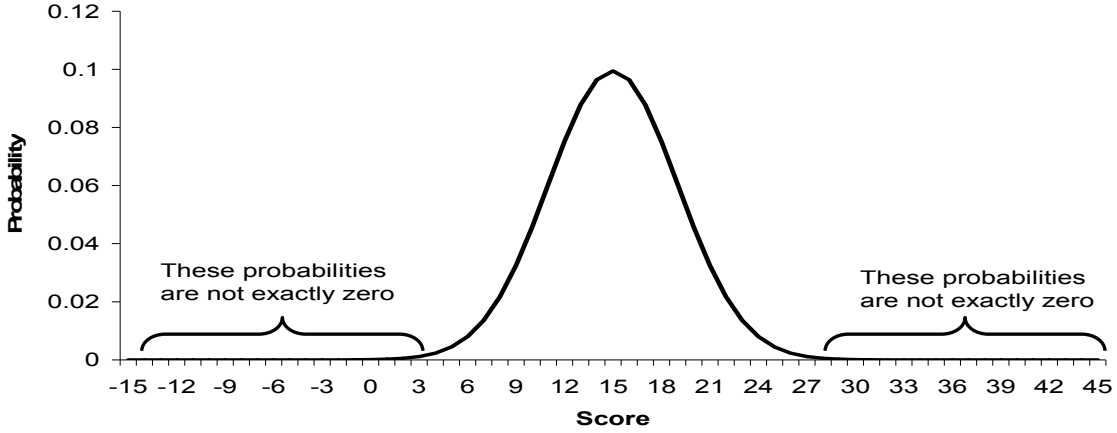


Figure 7: Symmetric Smooth & Discrete Distribution

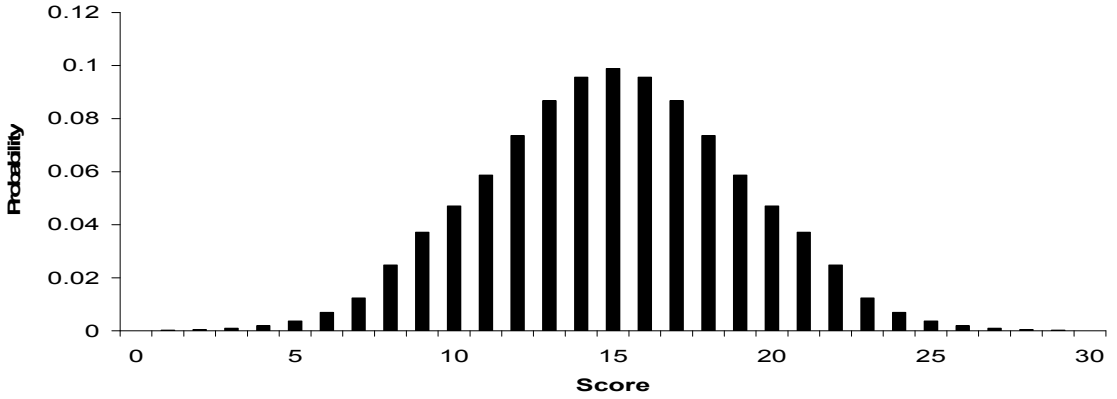


Figure 8: Symmetric 7 Peaks Distribution

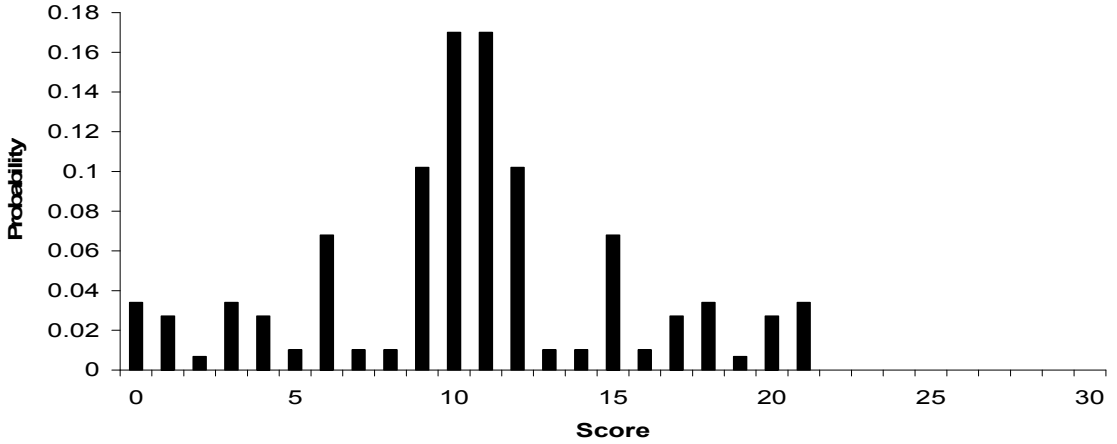
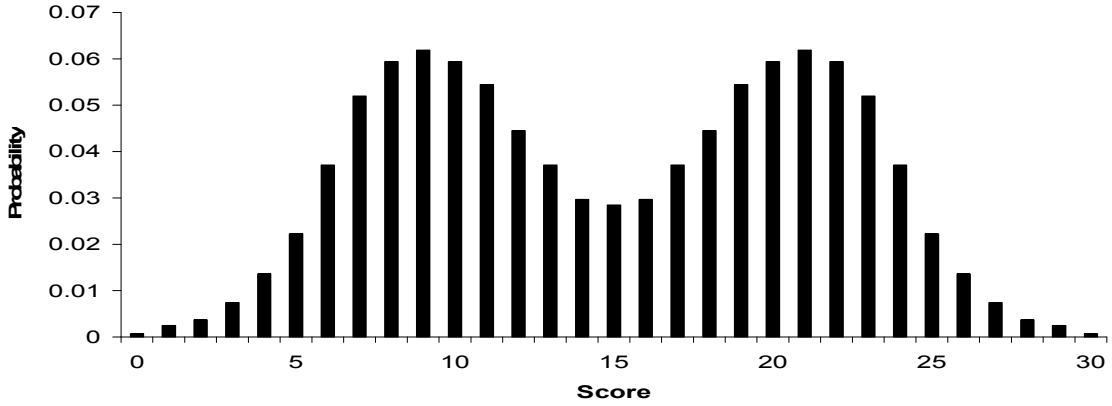


Figure 9: Symmetric Bimodal Distribution



The second data generation method was a continuous data generation method. The continuous data generation method used in this study is known as Fleishman's (1978) power method. Sample datasets of 30 standard normal deviates (*Z*) were generated and these normal deviates were transformed into samples from the desired population distributions,

$$Y = \mu + \sigma(a + bZ + cZ^2 + dZ^3). \quad (1)$$

Sets of μ, σ, *a*, *b*, *c*, and *d* values were used to produce *Y* values that had means, standard deviations, skews and kurtoses that reflected the symmetric and asymmetric discrete distributions.

For the Asymmetric Continuous distribution (Figure 1), μ and σ were 15 and 4, respectively, and constants of *a*, *b*, *c*, and *d* values of 0.3995, 0.9297, −0.3995 and −0.0365 were used to achieve the asymmetry and non-normality (skew = −1.75; kurtosis =3.75). For the Symmetric Continuous distribution (Figure 6), μ and σ were 15 and 4, and *a*, *b*, *c*, and *d* values of 0, 1, 0 and 0 were used to achieve the symmetry and normality (skew = 0; kurtosis = 0).

Statistical Strategies for Testing Mean Differences

Two statistical tests were considered for evaluating the mean differences in *Y* for groups *j* = 1 and 2. The standard independent samples t-test assuming homogeneous variances is defined as,

$$t_{Standard} = \frac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{s^2}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \quad (2)$$

where $\bar{Y}_1$ and $\bar{Y}_2$ are the groups' sample means,

$$\frac{1}{n_j}\sum_i Y_{i,j}, \quad (3)$$

s_1^2 and s_2^2 are the groups' sample variances,

$$\frac{1}{n_j - 1}\sum_i (Y_{i,j} - \bar{Y}_j)^2. \quad (4)$$

used to compute the pooled variance, s^2,

$$\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}, \quad (5)$$

The statistical significance of $t_{Standard}$ is determined by computing its percentile on a t distribution with $n_1 + n_2 - 2$ degrees of freedom.

Yuen's (1974) trimmed and Winsorized t-test was also considered. First the *Y* scores are ordered within each treatment group,

$Y_{1,j} \le Y_{2,j} \le \ldots \le Y_{n_j,j}$, $g_j = \gamma n_j$ is then defined where γ indicates the proportion of individuals trimmed in each tail of the distribution (γ = 0.1 & 0.2 in this study) and the effective sample size for group *j* is $h_j = n_j - 2g_j$. The trimmed mean for group *j* is computed as,

$$\bar{Y}_{t.,j} = \frac{1}{h_j} \sum_{i=g_j+1}^{n_j - g_j} Y_{i,j} \,. \qquad (6)$$

The data for group *j* are Winsorized as,

$$\begin{aligned} X_{ij} &= Y_{g_j+1,j} \quad \text{if } Y_{ij} \le Y_{g_j+1,j} \\ &= Y_{ij} \qquad \text{if } Y_{g_j+1,j} < Y_{ij} < Y_{n_j-g_j,j} \,, \\ &= Y_{n_j-g_j,j} \quad \text{if } Y_{ij} \ge Y_{n_j-g_j,j} \end{aligned} \qquad (7)$$

and the Winsorized data are used to compute group *j*'s Winsorized mean,

$$\bar{X}_{w.j} = \frac{1}{n_j} \sum_{i=1}^{n_j} X_{i,j} \,, \qquad (8)$$

and variance,

$$s^2_{w.j} = \frac{1}{n_j - 1} \sum_{i=1}^{n_j} \left(X_{i,j} - \bar{X}_{w.j} \right)^2 . \qquad (9)$$

Finally, the t-test for comparing groups' trimmed means is computed as,

$$t_{trimWinsorized} = \frac{\bar{Y}_{t.,1} - \bar{Y}_{t.,2}}{\sqrt{d_1 + d_2}} \,, \qquad (10)$$

where

$$d_j = \frac{\left(n_j - 1\right) s^2_{w.j}}{h_j \left(h_j - 1\right)} .$$

The statistical significance of the $t_{trimWinsorized}$ statistic is determined by computing its percentile on a t distribution with

$$\frac{\left(d_1 + d_2\right)^2}{d_1 / \left(h_1 - 1\right) + d_2 / \left(h_2 - 1\right)}$$

degrees of freedom.

Both the standard and the trimmed and Winsorized t-tests were implemented as two-tailed significance tests with nominal Type I error rates of 0.05. The trimmed and Winsorized t-test was based on symmetric trimming and Winsorizing of 10% and 20% of the most extreme lowest and highest observations of the two groups' *Y* distributions.

Type I Error and Power Evaluations

The standard and trimmed and Winsorized t-tests were used to evaluate the statistical significance of the differences in means of two groups whose scores were generated as samples from one of the nine population distributions. The t-tests were evaluated with respect to their Type I error (where the population difference in groups' means was zero) and power (where the population difference in groups' means was not zero).

All t-tests' Type I error and power rates were rates at which the t-tests indicated that the groups' mean differences were statistically significant across 10,000 replications (i.e., 10,000 statistical significance tests of groups' mean differences). The t-tests' Type I error rates were computed in conditions where the sample datasets for the two groups were drawn from one population distribution and were not altered prior to their analyses with the t-tests. The robustness of the t-tests' Type I error rates were considered with respect to two criteria, the Type I error range defined as ±2 standard errors of the nominal 0.05 rate for a simulation study based on 10,000 replications (i.e., $= 0.05 \pm 2\sqrt{\frac{(0.05)(0.95)}{10,000}} = 0.0456$ to 0.0544), and a wider robustness criterion proposed by Bradley (= 0.025 to 0.075, 1978). The t-tests' power rates were computed in the simulated conditions where the sample datasets for the two groups were drawn from one population distribution and then 1/2 of the population

distribution's standard deviation was added to one of the groups' scores.

Results

Type I Error

Table 3 presents the t-tests' Type I error rates across this study's nine population distributions. Comparisons of the standard and trimmed and Winsorized t-tests for the two continuous distributions pertain to the t-test evaluations of interest in most simulation studies. Comparisons of the t-tests across the discrete distributions are unconsidered in most simulation studies.

The Type I error rates of the three t-tests across all population distributions were within the 0.025 to 0.075 range defined by Bradley's (1978) criterion, but several fell outside of the ±2 standard error range (0.0456 to 0.0544). The nonrobust Type I error rates were conservative (less than 0.05) rather than the liberal (greater than 0.05) Type I error rates that would prompt the greatest concern of the t-tests' robustness. The trimmed and Winsorized t-test had more nonrobust, conservative Type I error rates than the standard t-test across the continuous and discrete distributions.

The extent of trimming had distribution-dependent influences on Type I error, where 20% trimming versus 10% trimming reduced Type I error for some distributions (i.e., the Asymmetric Continuous, Asymmetric Smooth & Discrete, and the Symmetric 7 Peaks distributions) and increased Type I error for other distributions (i.e., the Asymmetric Favorite Scores, Asymmetric Lump at Zero, Asymmetric Teeth, Symmetric Continuous, Symmetric Smooth & Discrete and the Symmetric Bimodal distributions).

Power

Table 4 presents the t-tests' power rates across this study's nine population distributions. The t-tests' power rates were most clearly affected by whether the distributions were symmetric or asymmetric. For the asymmetric distributions, the trimmed and Winsorized t-test was more powerful than the standard t-test. The greater power of the trimmed and Winsorized t-test held across the asymmetric continuous and asymmetric discrete distributions, and was especially apparent in the Asymmetric Teeth and Asymmetric Lump at Zero distributions. For the Asymmetric Teeth and Asymmetric Lump at Zero distributions, 20% trimming resulted in increased power relative to 10% trimming. For most of the symmetric distributions, the trimmed and Winsorized t-test was less powerful than the standard t-test. For all but the Symmetric 7 Peaks distribution, 20% trimming reduced power relative to 10% trimming.

Conclusion

In simulation research considerable attention has been devoted to the effects of nonnormality on the accuracy of statistical significance tests for groups' mean differences (Glass, Peckham & Saunders, 1972; Keselman, et al., 2004; Lix, & Keselman, 1998; Lix, Keselman & Keselman, 1996). In this research nonnormality is predominantly characterized in terms of the level of skew and kurtosis of continuous and theoretically unbounded distributions.

Recent results and proposals from simulation research have suggested that standard significance tests should be abandoned in favor of alternative significance tests that are designed to be robust to nonnormality (Lix, Keselman & Keselman, 1996; Wilcox, 1995). However, a somewhat unique simulation study found that the standard t-test can be quite robust with respect to the types of nonnormality noted to occur in real world distributions of psychometric and achievement tests, where score ranges are discrete and bounded and where nonnormality cannot be completely characterized with respect to skew and kurtosis (Sawilowsky & Blair, 1992). This study was designed to reconsider the Type I error and power of standard and trimmed and Winsorized t-tests of mean differences with respect to the types of distributions considered in the majority of simulation studies and the types of distributions noted to occur in applied psychometric and achievement test data.

In terms of Type I error, the results show that the standard and trimmed and Winsorized t-tests did not exhibit extreme lack of robustness for any of the considered distributions. Type I error rates obtained for the continuous distributions considered in simulation studies were reasonably representative of the Type I error rates obtained

Table 3: Type I Error Results

Symmetry	Distribution	Standard t-test	Trimmed & Winsorized t-test (10% trimming)	Trimmed & Winsorized t-test (20% trimming)
Asymmetric	Continuous	0.0424*	0.0431*	0.0393*
	Favorite Scores	0.0454*	0.0360*	0.0502
	Lump at Zero	0.0476	0.0333*	0.0460
	Smooth & Discrete	0.0471	0.0435*	0.0431*
	Teeth	0.0473	0.0364*	0.0455*
Symmetric	Continuous	0.0447*	0.0450*	0.0452*
	7 Peaks	0.0493	0.0451*	0.0379*
	Smooth & Discrete	0.0494	0.0469	0.0498
	Bimodal	0.0478	0.0477	0.0495

*The Type I error rate is outside of the +/- 2 standard error range (0.0456 to 0.0544)

Table 4: Power Results

Symmetry	Distribution	Standard t-test	Trimmed & Winsorized t-test (10% trimming)	Trimmed & Winsorized t-test (20% trimming)
Asymmetric	Continuous	0.4910	0.5241	0.5135
	Favorite Scores	0.5001	0.6144	0.5012
	Lump at Zero	0.4980	0.6698	0.7437
	Smooth & Discrete	0.5014	0.5352	0.5254
	Teeth	0.5030	0.6511	0.7543
Symmetric	Continuous	0.4810	0.4527	0.4213
	7 Peaks	0.4756	0.4590	0.5849
	Smooth & Discrete	0.4813	0.4391	0.4104
	Bimodal	0.4746	0.3805	0.2813

from different types of discrete distributions. The Type I error rates of the t-tests were more likely to be slightly conservative rather than liberal. The trimmed and Winsorized t-test had a Type I error that was usually more conservative than that of the standard t-test.

This study's power results were more extreme than the Type I error results, and varied by the type of t-test, by whether the population distribution was symmetric or asymmetric, and by the specific features of the population distribution. To assess the power results in more detail, this study's power simulations were re-run and analyzed with respect to issues such as the expected mean differences in the samples, the standard error of the mean differences in the samples, and the accuracy of the estimated standard error of the mean differences. To simplify the analyses, all of the simulated data were transformed so that all population standard deviations were four, all population mean differences were two and the standard errors of these population untrimmed mean differences were about 1.03 (given the group sample sizes of 30). The score transformations had negligible effects on the power rates reported in Table 4 and no effect on the discreteness and structures of the distributions.

The results of the re-run power analyses are presented in Table 5, where the 27 power rates corresponding to the nine population distributions and three t-tests are sorted from highest to lowest. Along with the power rates, the standard errors of the mean differences are shown (i.e., the standard deviation of the differences in the means evaluated by the t-tests across the 10,000 replications of the simulations). These 27 standard errors correlated −0.97 with the 27 power rates and provide a useful basis for understanding how power was affected by the population distributions and t-tests considered in this study. The major power results can be described as follows,

- Power was highest for the distributions and t-tests where the standard error of mean differences was lowest. Power was lowest for the distributions and t-tests where the standard error of mean differences was highest.
- The trimmed and Winsorized t-test had high power and a low standard error when used with all of the asymmetric distributions. The trimmed and Winsorized t-test had low power and a high standard error when used with all of the symmetric distributions except for the Symmetric 7 Peaks distribution.
- The extent of trimming had mixed results, in that for some distributions increased trimming resulted in increased power and decreased standard errors while for other distributions increased trimming resulted in decreased power and increased standard errors.
- The issue of continuous and discrete distributions had an influence on the power of the trimmed and Winsorized t-test such that power rates were less extreme for the continuous distributions of comparable levels of skew. That is, the power for the Asymmetric Continuous distribution was lower than the power for the asymmetric discrete distributions while the power for the Symmetric Continuous distribution was greater than the power for the symmetric discrete distributions.
- The standard t-test's power and standard errors were less influenced than the trimmed and Winsorized t-test across the distributions, being less powerful than the trimmed and Winsorized t-test for the asymmetric distributions and more powerful than the trimmed and Winsorized t-test for the symmetric distributions. In contrast to the trimmed and Winsorized t-test, the standard t-test was slightly less powerful for the symmetric distributions than for the asymmetric distributions.

Implications for Practice

This study's findings regarding how a data generation method affects the relative power of different t-tests have implications for practice. The trimmed and Winsorized t-test is more complexly affected by the type of distribution than the standard t-test. Some of the power issues with the trimmed and Winsorized t-test could be anticipated with careful examination of the data at hand. Specifically, for

Table 5: Power Rates Sorted by the Standard Error of the Difference in Means

Distribution	Statistical Method	Std. Error	Power
Asymmetric Teeth	Trimmed & Winsorized (20%)	0.7273	0.7543
Asymmetric Lump at Zero	Trimmed & Winsorized (20%)	0.7462	0.7437
Asymmetric Teeth	Trimmed & Winsorized (10%)	0.8579	0.6511
Asymmetric Lump at Zero	Trimmed & Winsorized (10%)	0.8817	0.6698
Asymmetric Favorite Scores	Trimmed & Winsorized (10%)	0.8971	0.6144
Symmetric 7 Peaks	Trimmed & Winsorized (20%)	0.9030	0.5849
Asymmetric Favorite Scores	Trimmed & Winsorized (20%)	0.9633	0.5011
Asymmetric Smooth & Discrete	Trimmed & Winsorized (10%)	0.9693	0.5352
Asymmetric Smooth & Discrete	Trimmed & Winsorized (20%)	0.9794	0.5254
Asymmetric Continuous	Trimmed & Winsorized (10%)	0.9800	0.5244
Asymmetric Continuous	Trimmed & Winsorized (20%)	0.9832	0.5137
Symmetric Smooth & Discrete	Standard t-test	1.0258	0.4813
Symmetric 7 Peaks	Standard t-test	1.0260	0.4756
Symmetric Bimodal	Standard t-test	1.0287	0.4746
Symmetric Continuous	Standard t-test	1.0298	0.4811
Asymmetric Continuous	Standard t-test	1.0308	0.4910
Asymmetric Favorite Scores	Standard t-test	1.0309	0.5001
Asymmetric Teeth	Standard t-test	1.0317	0.5030
Asymmetric Lump at Zero	Standard t-test	1.0332	0.4981
Asymmetric Smooth & Discrete	Standard t-test	1.0332	0.5014
Symmetric 7 Peaks	Trimmed & Winsorized (10%)	1.0499	0.4590
Symmetric Continuous	Trimmed & Winsorized (10%)	1.0578	0.4528
Symmetric Smooth & Discrete	Trimmed & Winsorized (10%)	1.0705	0.4390
Symmetric Continuous	Trimmed & Winsorized (20%)	1.0968	0.4216
Symmetric Smooth & Discrete	Trimmed & Winsorized (20%)	1.1168	0.4104
Symmetric Bimodal	Trimmed & Winsorized (10%)	0.9981	0.3805
Symmetric Bimodal	Trimmed & Winsorized (20%)	1.0021	0.2813

datasets that have structures and asymmetry resulting in only a small number of the possible scores being observed (i.e., the Asymmetric Teeth and Asymmetric Lump at Zero distributions), trimming and Winsorizing of these observed scores will produce a dataset with even fewer unique scores, a standard error of the trimmed mean that is relatively small, and a power rate that may be large relative to the standard t-test.

For datasets where many of the possible scores are observed (i.e., the Symmetric Bimodal distribution), trimming and Winsorizing of the observed scores will produce a dataset with a large range of unique scores, a standard error of the trimmed mean that is relatively large, and a power rate that is small relative to the standard t-test. If the data at hand are so skewed and/or are based on a sample size that is extremely small, trimming and Winsorizing could remove all of the scores from the data and make a significance test of mean differences impossible.

Note that this study focused on creating distributions that reflect structures that have been observed in psychometric and achievement test data (Holland & Thayer, 2000; Micceri, 1989). While the discrete distributions considered in this study may be more realistic than the continuous distributions typically created in simulation studies, these discrete distributions clearly do not reflect all of the possible distributions encountered in applied data.

Important distributions that were not considered in this study are distributions of counted variables, such as individuals' income, individuals' total of social connections to other individuals, or websites' numbers of hits. Extreme observations are more likely in distributions of unbounded counted variables than in distributions of psychometric and achievement test scores. Simulations based on distributions where extreme observations are likely may show that the standard t-test has a nonrobust Type I error rate whereas the trimmed and Winsorized t-test is robust.

Implications for Simulation Research

This study's findings of how the data generation method affected the relative Type I error and power rates of different t-tests have implications for simulation research. One issue that could be reconsidered is how assumptions are violated in simulation studies. For example, in simulation studies' continuous and unbounded distributions, the levels of skew and kurtosis can be much greater than are possible to create in discrete and bounded distributions, such as skew values of 120 and kurtosis values ranging from 8.9 to beyond 18,000 (Keselman, et al., 2004; Wilcox, 1994). The current study suggests that simulation studies' results based on extreme levels of assumption violations do not always generalize to situations where levels of assumption violations are more limited. In particular, this study suggests that for the relatively limited levels of assumption violations that can occur in bounded distributions, the robustness of standard tests of mean differences is not likely to be as serious a concern as implied when robust strategies are promoted. This study also showed that more can be learned about robust statistical procedures proposed as replacements for standard statistical tests. Additional simulation studies that consider the distributions and assumption violations likely to be encountered in applied research are encouraged.

References

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

Fleishman, A. I. (1978). A method for simulating non-normal distributions. *Psychometrika*, *43*, 521-532.

Glass, G. V., Peckham, P. D., & Sanders, J. R. (1972). Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. *Review of Educational Research*, *42*(*3*), 237-288.

Hoaglin, D. C. (1985). Summarizing shape numerically: The g-and-h distributions. In D. Hoaglin, F. Mosteller, & J. Tukey (Eds.). *Exploring data tables, trends, and shapes.* New York: Wiley.

Holland, P. W. & Thayer, D. T. (2000). Univariate and bivariate loglinear models for discrete test score distributions. *Journal of Educational and Behavioral Statistics*, *25*, 133-183.

Keselman, H. J., Othman, A. R., Wilcox, R. R., & Fradette, K. (2004). The new and improved two-sample t test. *Psychological Science*, *15*(*1*), 47-51.

Lix, L. M., & Keselman, H. J. (1998). To trim or not to trim: Tests of location equality under heteroscedasticity and nonnormality. *Educational and Psychological Measurement*, *58*(*3*), 409-429.

Lix, L., Keselman, J., & Keselman, H. J. (1996). Consequences of assumption violations revisited: A quantitative review of alternatives to the one-way analysis of variance F test. *Review of Educational Research*, *66*(*4*), 579-619.

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

Sawilowsky, S. S., & Blair, R. C. (1992). A more realistic look at the robustness and Type II error properties of the t test to departures from population normality. *Psychological Bulletin*, *111*(*2*), 352-360.

Wilcox, R. R. (1994). A one-way random effects model for trimmed means. *Psychometrika*, *59*(*3*), 289-306.

Wilcox, R. R. (1995). ANOVA: A paradigm for low power and misleading measures of effect size? *Review of Educational Research*, *65*(*1*), 51-77.

Yuen, K. K. (1974). The two-sample trimmed t for unequal population variances. *Biometrika*, *61*, 165-170.