Paper 228-2013

Variance Heterogeneity and Non-Normality: How SAS PROC TTEST® Can Keep Us Honest

Anh P. Kellermann, Aarti P. Bellara, Patricia Rodríguez de Gil,
Diep Nguyen, Eun Sook Kim, Yi-Hsin Chen, Jeffrey D. Kromrey
University of South Florida, Tampa, FL

ABSTRACT

The independent samples *t*-test is one of the most used tests for detecting true mean differences. The SAS system provides the PROC TTEST procedure which is an easy way to conduct a test for the difference between two population means by assuming homogeneity of variance or avoiding it. However, the *t*-test and its alternatives (the Satterthwaite's approximate test and Conditional *t*-test) assume population normality. Although past research has provided evidence of the *t*-test's robustness to departures of normality, questions about the performance of conditional testing when the assumption of normality is not met remain. This paper describes previous research on preliminary tests under the normality assumption, extends this research to the evaluation of conditional testing to departures of normality, and provides guidance to researchers on the proper use of this test with non-normal, heteroscedastic population distributions.

Keywords: STATISTICAL ASSUMPTIONS, ROBUSTNESS, NON-NORMALITY, VARIANCE HETEROGENEITY.

INTRODUCTION

While statistical procedures have become more complex (e.g., multilevel modeling), the independent samples *t*-test remains as one of the best known and widely used statistical procedures to compare two-group means (Hayes & Cai, 2007; Heiman, 2011; Sawilowsky & Blair, 1992). Under the assumption of normality, the *t*-test is "the most powerful unbiased test" (Bridge & Sawilowsky, 1999; p. 229) for detecting true mean differences. The syntax for PROC TTEST is quite simple; it requires only a CLASS statement to identify the independent variable and a VAR statement to identify the dependent variable. Yet, the default values of PROC TTEST provide a test of variance homogeneity (the Folded *F*-test). Nguyen et al. (2012) conducted a simulation study to investigate the effectiveness of preliminary tests and provided an example of the use of the Folded *F*-test to determine which *t*-test statistic (Independent *t*-test or Satterthwaite approximate *t*-test) is appropriate when the assumption of variance homogeneity is violated with the data being normally distributed.

PROC TTEST EXAMPLE

```
* +-------------------------------------------------------------------------------+
   SAS program to perform an independent-samples t-test. This simple SAS code tests
   the null hypothesis that there is no difference between two groups with respect to
   their mean scores on the survey measuring level of anxiety in a statistic course.
  +-------------------------------------------------------------------------------+;

PROC TTEST   DATA=Survey;
 class       Gender;
 var         Anxiety;
run;
```

Where:

```
class   independent-variable;
var     dependent-variable;
```

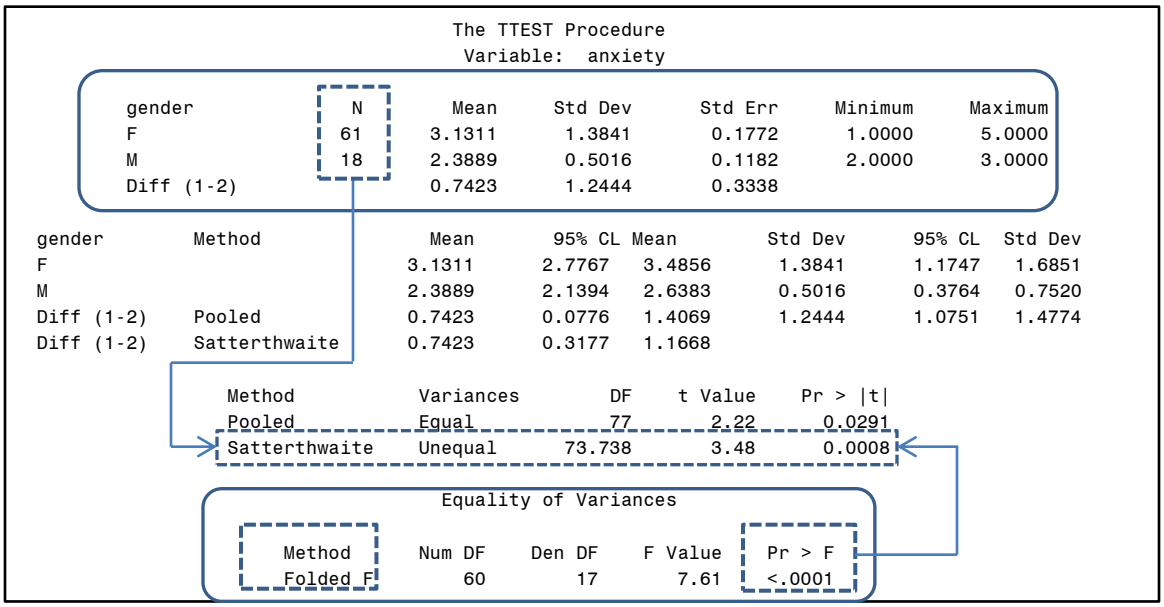

Figure 1. Results of PROC TTEST: Statistically Significant Differences in Variances Observed

Data in Figure 1 suggest heterogeneity of variance, F (60,17) = 7.61, p <.0001. In addition, with unequal sample sizes (n_1 = 61, n_2 = 18), the results from the test of means using Satterthwaite's approximate t-test may be the most appropriate.

Using data from another experiment, the Folded F statistic shown in Figure 2 did not suggest unequal population variances, F (4,4) = 2.00, p = .5185. In addition, with equal sample sizes (n_1 = 5, n_2 = 5) reporting the independent means t-test seems more appropriate.

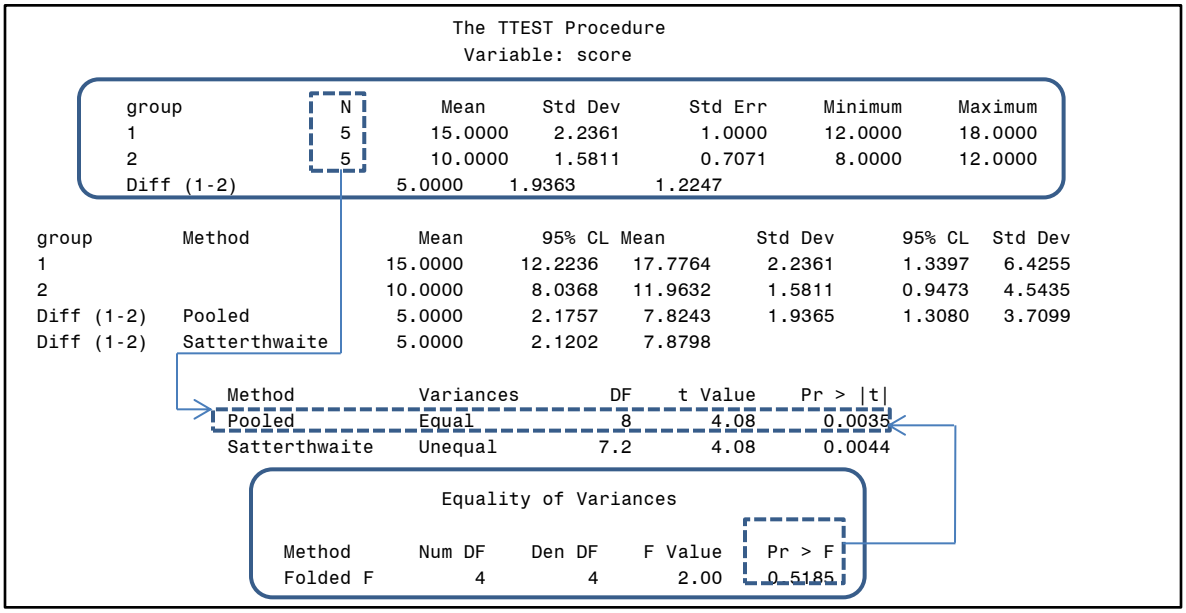

Figure 2. Results of PROC TTEST: Nonsignificant Differences in Variances Observed

Under a nominal alpha level of .05 across all studied conditions, Nguyen et al. (2012), found that the independent means t-test showed a large dispersion of Type I error rates as a result of variance heterogeneity included in the study. On the other hand, the Satterthwaite approximate t-test showed adequate Type I error control in most conditions studied. As for the conditional t-test, it provided an improved Type I error control as the alpha level for the Folded F-test was increased. Satterthwaite's approximate t-test performed very well regardless of the ratios of

population variances and sample sizes in the two groups used in the study (the exception being conditions with very small sample sizes such as 10). Overall, Satterthwaite's approximate *t*-test performed best in control of Type I error rates under all conditions, whereas the performance of the independent means *t*-test and the conditional *t*-test depended on the experimental conditions. As expected, under equal sample sizes, the independent *t*-test was relatively robust to violations of the homogeneity of variance assumption and both Satterthwaite's approximate *t*-test and the conditional *t*-test gave evidence of a much improved Type I error rate under variance heterogeneity when samples were unequal.

However, due to the unlikeliness of encountering real data that are normally distributed (see Micceri, 1989) researchers have questioned the robustness of the conditional *t*-test with respect to Type I error and statistical power when the assumption of normality is not met. Assessment of the tenability of the population normality assumption is easily made with PROC UNIVARIATE. The present study extends the Nguyen et al. (2012) research on conditional testing (homogeneity of variance assumption) to investigate the performance of the independent means *t*-test and alternatives under departures of normality as well as heterogeneous variances.

THE SIMULATION STUDY

A Monte Carlo simulation was conducted to investigate the performance of the independent means *t*-test and alternatives from departures of normality. The simulation conditions manipulated in this study were: (a) total sample size (from 10 to 400), (b) sample size ratio between groups (1:1, 2:3, and 1:4), (c) variance ratio between populations (1, 2, 4, 8, 12, 16, and 20), (d) effect size for mean difference between populations ($\Delta = 0, .2, .5, .8$), (e) alpha set for testing treatment effect (from $\alpha = .01$ to $\alpha = .25$), (f) alpha set for testing homogeneity assumption for the conditional *t*-test (from $\alpha = .01$ to $\alpha = .50$), and (g) population distributions with varying skewness and kurtosis values (i.e., $\gamma_1 = 1.00$ and $\gamma_2 = 3.00$, $\gamma_1 = 1.50$ and $\gamma_2 = 5.00$, $\gamma_1 = 2.00$ and $\gamma_2 = 6.00$, $\gamma_1 = 0.00$ and $\gamma_2 = 25.00$, as well as $\gamma_1 = 0.00$ and $\gamma_2 = 0.00$ for the normal distribution).

TYPE I ERROR CONTROL

An overall view of the Type I error control of the tests is provided in Figures 3, 4, and 5. Figure 3 describes the distributions of Type I error rate estimates under a nominal alpha level of .01 across all conditions. The first two box plots are for the independent means *t*-test and Satterthwaite's approximate *t*-test, respectively. The remaining plots delineate the Type I error rate estimates for the conditional *t*-test across the different conditioning rules that were investigated. That is, the plot for C (01) provides the distribution of Type I error rates for the conditional *t*-test when an alpha level of .01 was used with the Folded *F*-test as the rule to choose between the independent means *t*-test and Satterthwaite's approximate *t*-test.

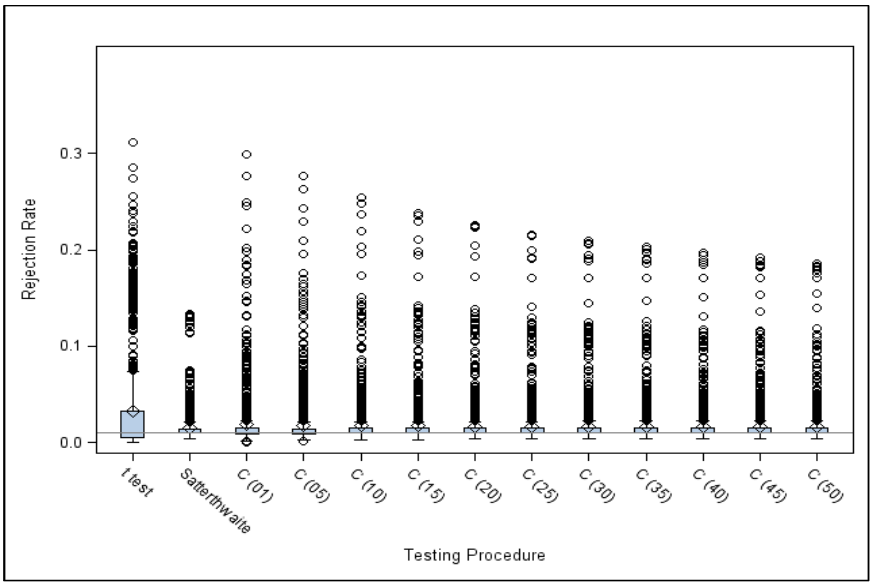

Figure 3. Distributions of Estimated Type I Error Rates (Nominal Alpha = .01)

Note in Figures 4 and 5 the great dispersion of Type I error rates for the independent means estimates under a nominal alpha level of .05 and .10. In contrast, Satterthwaite's approximate *t*-test provided better Type I error control. As for the conditional *t*-test, the box plots showed that this test provided a notable improvement in Type I error control relative to the independent means *t*-test, and the improvement increases as the alpha level for the Folded *F*-test is increased.

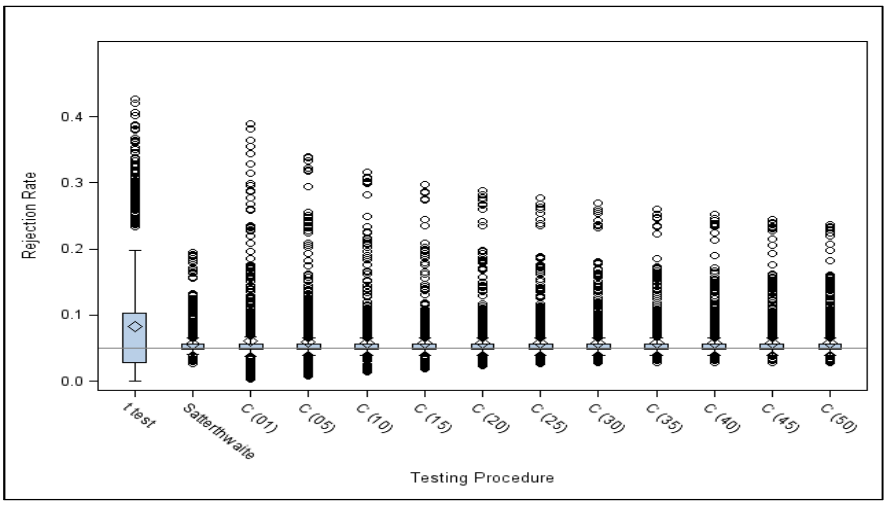

Figure 4. Distributions of Estimated Type I Error Rates (Nominal Alpha = .05)

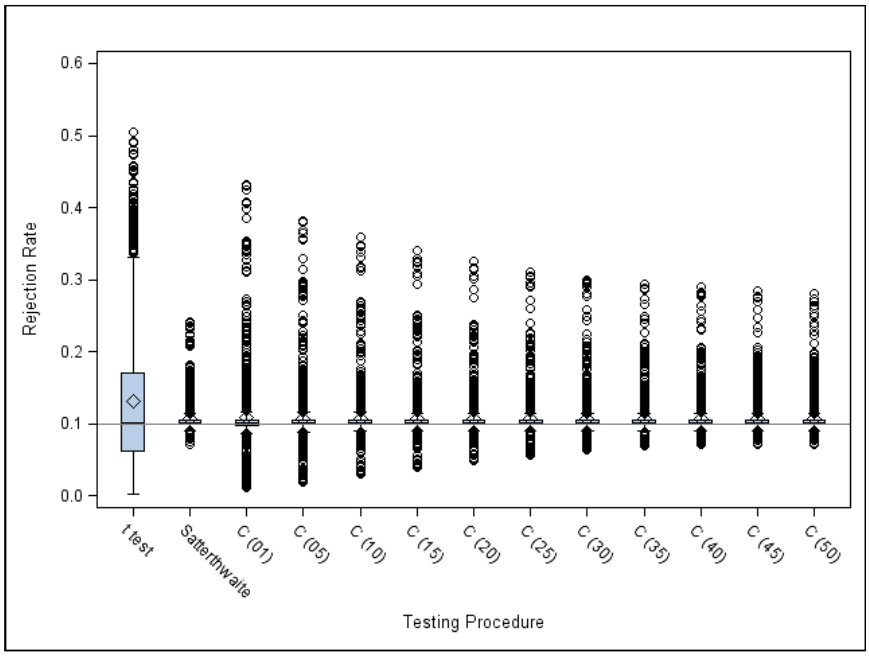

Figure 5. Distributions of Estimated Type I Error Rates (Nominal Alpha = .10)

An eta-squared analysis on Type I error rates of the three test types showed that for the independent means *t*-test, sample size ratio and the interaction between sample size ratio and variance ratio were major factors with sample size ratio being the strongest factor. As for Satterthwaite's approximate *t*-test and the conditional *t*-test, total sample size, sample size ratio, and the interaction between sample size ratio and total sample size were major factors with combined sample size ratio together with total sample size being the strongest factor. In considering the impact of distribution shape alone on Type I error rates of the three tests, Type I error rate of Satterthwaite's approximate *t*-test was most impacted. Type I error rate of the conditional *t*-test was also impacted but to a much lesser degree compared to that of Satterthwaite's approximate *t*-test, whereas Type I error rate of the independent means t-test was only marginally impacted by the distribution shape.

Figures 6, 7, and 8 present the mean Type I error rates by distribution shape and total sample size for the independent means *t*-test, Satterthwaite's approximate *t*-test, and the conditional *t*-test, respectively, under the nominal alpha level of .05. In Figure 6, Type I error rates of the independent means *t*-test are far above the nominal alpha level regardless of the distribution shapes and total sample sizes. In contrast, Satterthwaite's approximate *t*-test provided much better Type I error control except for extremely small sample size (i.e., total sample size = 10) or the

extremely skewed distribution (i.e., skewness = 2).(see Figure 7). Figure 8 shows that the conditional *t*-test provided similar Type I error control with the same exceptions as those of Satterthwaite's approximate *t*-test.

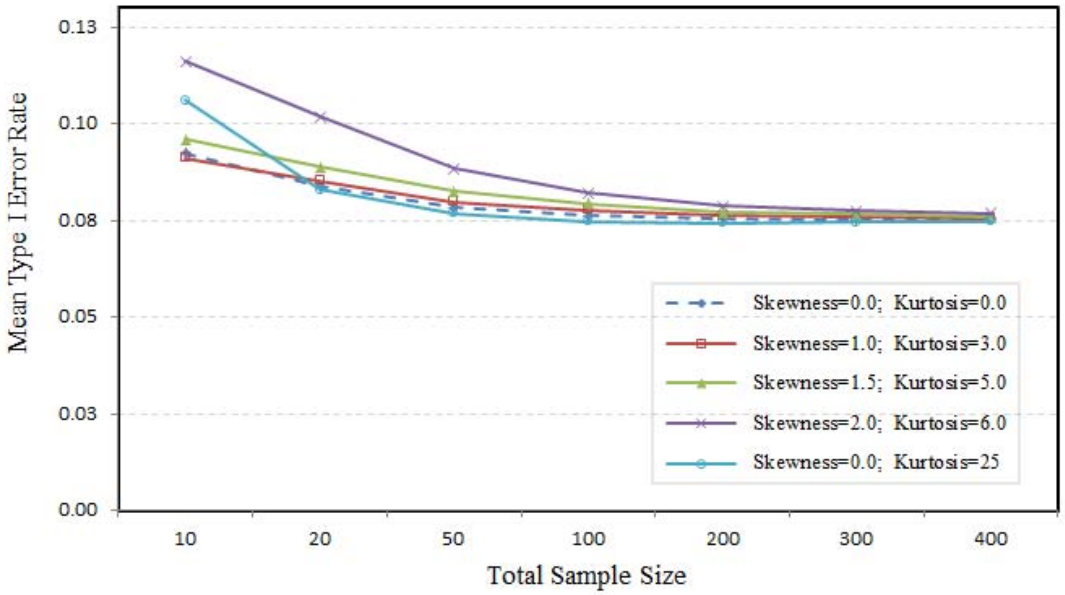

Figure 6. Mean Type I Error Rate for the Independent Means *T*-Test by Total Sample Size and Distribution Shape.

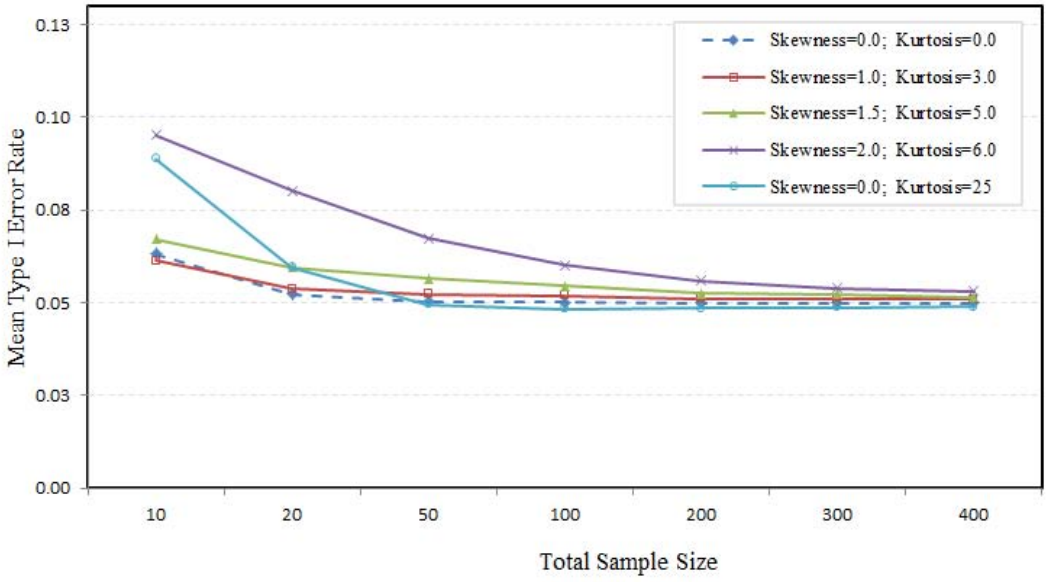

Figure 7. Mean Type I Error Rate for Satterthwaite's Approximate *T*-Test by Total Sample Size and Distribution Shape.

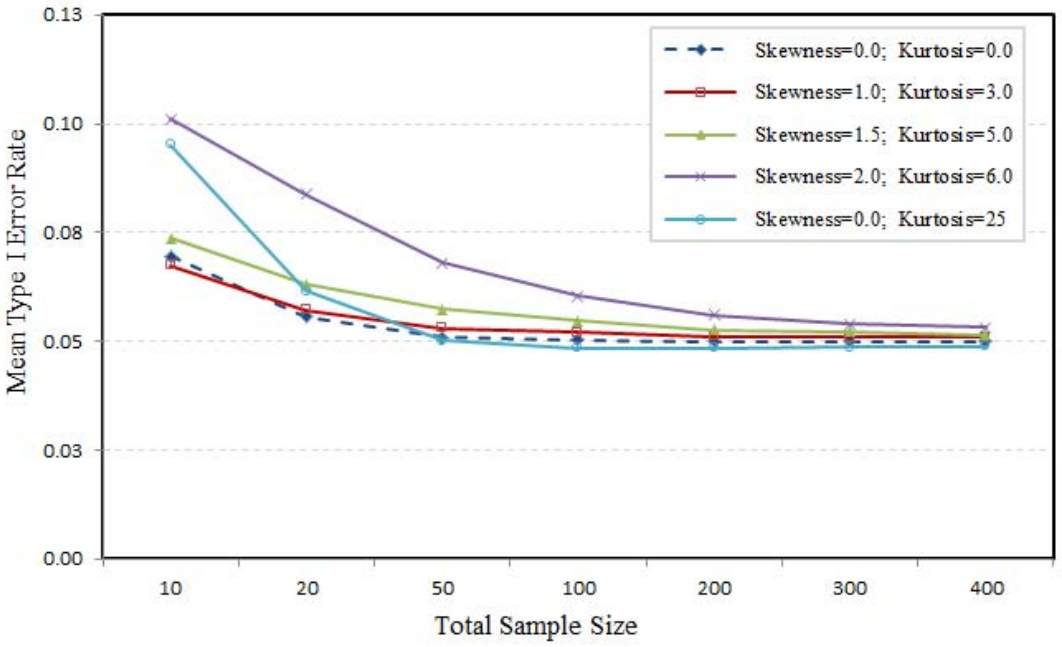

Figure 8. Mean Type I Error Rate for the Conditional *T*-Test by Total Sample Size and Distribution Shape.

CONDITIONS MEETING BRADLEY'S LIBERAL CRITERION

Bradley's (1978) liberal criterion of robustness was used to examine the Type I error rates across conditions of the study. Using this criterion, a test is considered to maintain adequate Type I error control if the Type I error rate is within the range of $\alpha \pm 0.5\alpha$. The proportions of cases meeting the Bradley's liberal criterion (Bradley rate hereafter) are reported in Table 1. The major factors affecting Bradley rates depend on the test type. For the independent means *t*-test, variance ratio and sample size ratio play major roles in Bradley rates. With homogeneous variance between groups, the independent means *t*-test meets Bradley's criterion perfectly. In case of balanced groups, Bradley rate for the independent means *t*-test remains high (91%). However, as variance ratio or sample size ratio becomes disproportionate, the independent means *t*-tests show substantially low Bradley rates. As presented in Figure 9, high Bradley rates are warranted for the independent means *t*-test only under homogeneity of variance and/or with equal sample size between groups. Other than these conditions, the independent means *t*-test frequently does not meet Bradley's criterion. For example, when sample sizes are unbalanced between groups, Bradley rates are virtually zero even with moderate heterogeneity of variance.

On the other hand, the Bradley rates of the conditional *t*-test and Satterthwaite's approximate *t*-test are greatly associated with the shape of data distribution. When the data are skewed (i.e., skewness = 2), Satterthwaite's approximate *t*-test showed low Bradley rates (78%). In other conditions Bradley rates of the Satterthwaite test are consistently high. For example, when data are not skewed including extremely high kurtosis (i.e., skewness = 0 and kurtosis = 25), Bradley rates are constantly high regardless of variance ratio. However, when data are skewed, Satterthwaite does not work well even under homogeneity of variance (see Figure 10). For conditional *t*-tests, when the alpha level for the Folded *F*-tests is set larger, the tests are closer to Satterthwaite test results. The conditional *t*-test with the Folded *F*-test alpha set at .25 showed very comparable results to Satterthwaite tests. Under certain circumstances (e.g., extremely small sample size, N=10) conditional *t*-tests at .25 showed slightly better Bradley rates (68% vs. 65%, respectively). Both testing procedures had Bradley rates over 90% when total sample size was 50 or more, and 100% when total sample size was 200 or more.

Condition	t-test	Conditional	Satterthwaite	Condition	t-test	Conditional	Satterthwaite
N				Variance ratio			
10	0.451	0.680	0.651	1:1	1.000	0.943	0.920
20	0.486	0.760	0.817	1:2	0.617	0.931	0.949
50	0.451	0.931	0.949	1:4	0.400	0.909	0.926
100	0.434	0.971	0.971	1:8	0.286	0.897	0.914
200	0.417	1.000	1.000	1:12	0.269	0.886	0.891
300	0.406	1.000	1.000	1:16	0.246	0.886	0.891
400	0.406	1.000	1.000	1:20	0.234	0.891	0.897
N ratio				Shape			
1:4	0.180	0.976	0.967	0,0	0.433	0.963	0.967
2:3	0.665	0.976	0.976	1,3	0.437	0.955	0.967
1	0.910	0.967	0.967	1.5,5	0.437	0.931	0.939
3:2	0.282	0.910	0.914	2,6	0.461	0.771	0.780
4:1	0.143	0.702	0.739	0,25	0.412	0.910	0.910

Table 1. The proportions of cases meeting the Bradley's liberal criterion by Tests and Conditions at α = .05

Note. Conditional = the conditional *t*-test at α = .25 for the Folded *F*-test. For shape, two values indicate skewness and kurtosis, respectively.

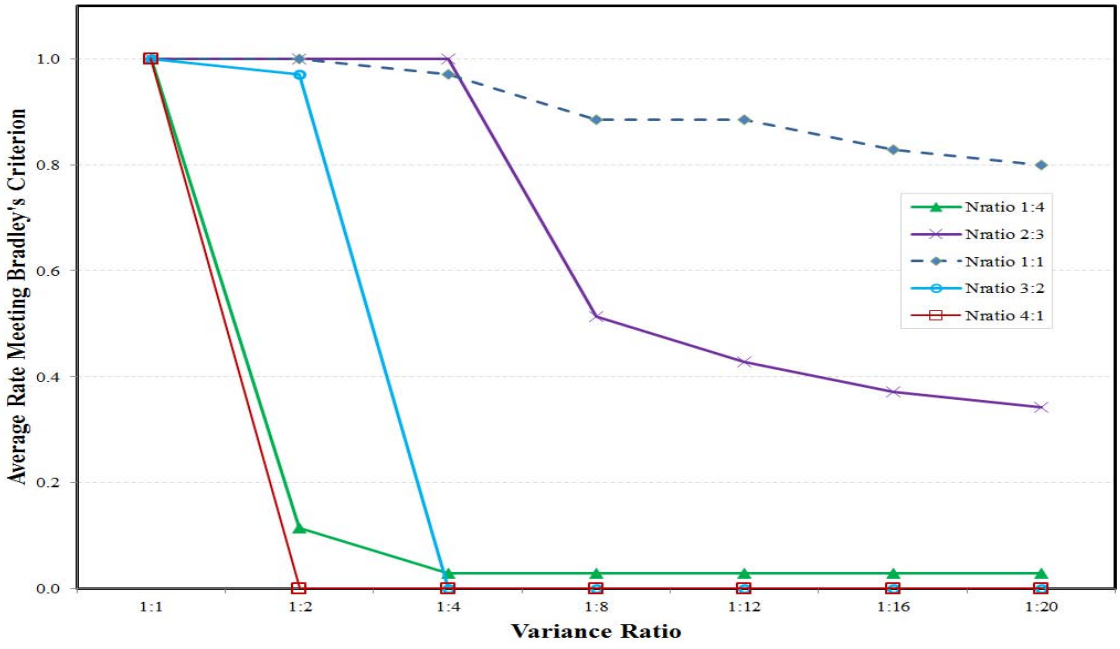

Figure 9. The Proportion of Cases Meeting Bradley's Liberal Criterion for the Independent Means *T*-Test.

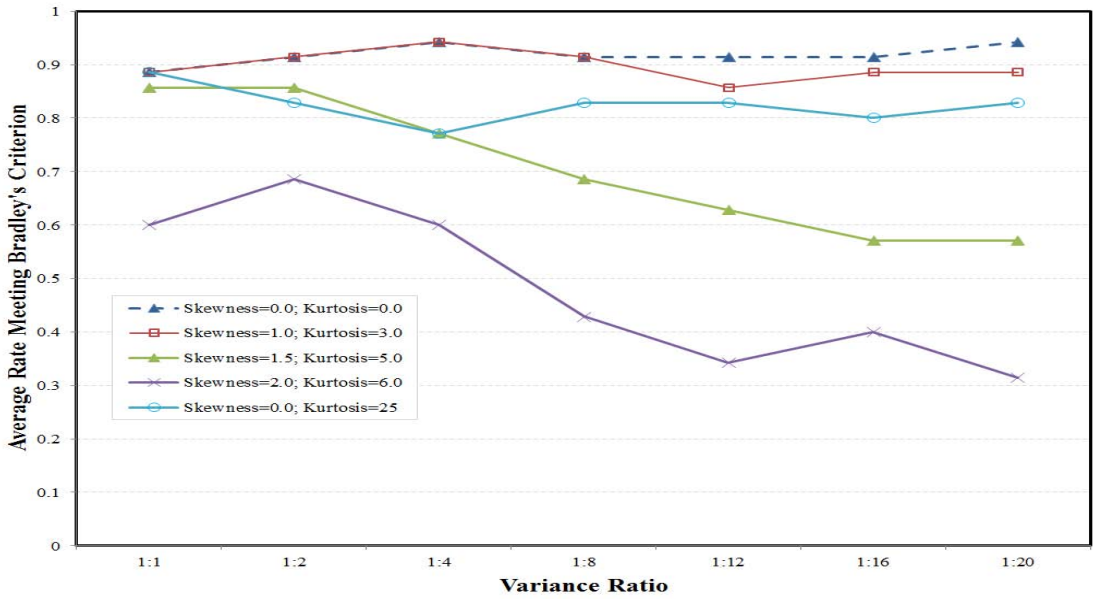

Figure 10. The Proportion of Cases Meeting Bradley's Liberal Criterion for Satterthwaite's Approximate *T*-Test at Alpha=.01.

STATISTICAL POWER

Although Satterthwaite's approximate *t*-test provides superior Type I error control, it is not always the best test to select because of the potential for power differences. When the assumptions are met, the independent means *t*-test is the most powerful test for mean differences. For this simulation study, power comparisons were made only for conditions in which both testing procedures evidenced adequate Type I error control by Bradley's (1978) benchmark. In 1,551 simulation conditions, both the independent means *t*-test and Satterthwaite's approximate *t*-test provided adequate Type I error control. Among these conditions, the independent means *t*-test was more powerful in 57% of the conditions and was less powerful in only 32% of the conditions.

In 3,312 simulation conditions, both the conditional *t*-test (using alpha = .25 for the Folded *F*-test of variances) and Satterthwaite's approximate *t*-test maintained adequate Type I error control. Figure 11 presents a scatter plot of the power estimates for these conditions. As evident in this figure, all of the differences in power were small. However, the conditional *t*-test was more powerful in 29% of the conditions while Satterthwaite's approximate *t*-test was more powerful in only 23% of the conditions (identical power estimates were obtained in the other conditions). Although the power differences were small, the use of the conditional testing procedure clearly may provide a power advantage over the use of Satterthwaite's approximate *t*-test.

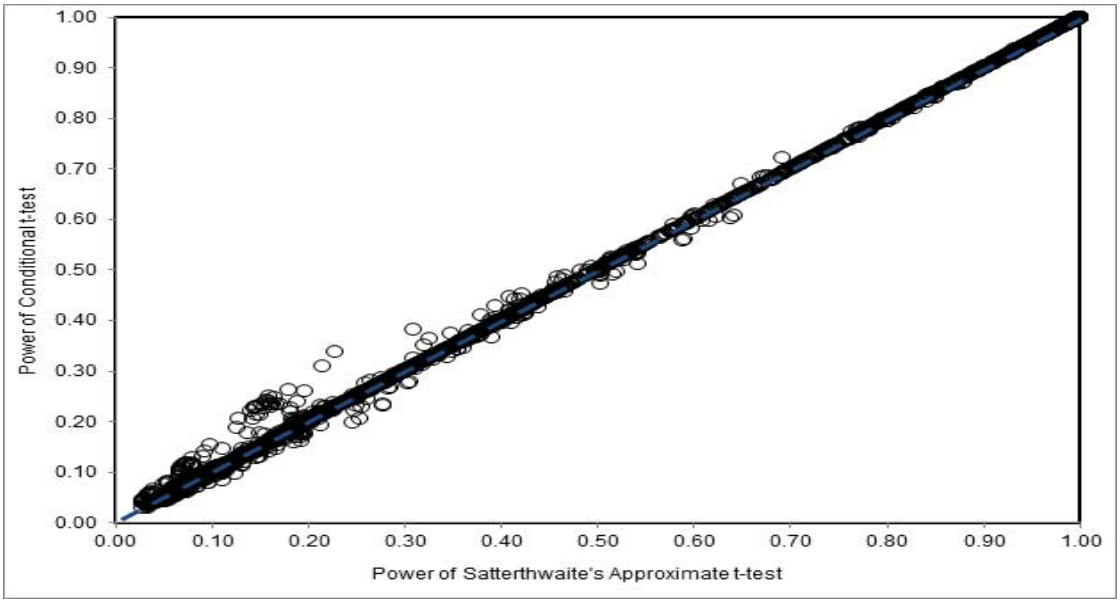

Figure 11. Scatterplot of Power Estimates for the Conditional *T*-Test and Satterthwaite's Approximate *T*-Test.

CONCLUSIONS AND IMPLICATIONS

The simulation study was intended to explore the performance of the independent means *t*-test and alternatives under departures of normality as well as heterogeneous variances. In addition to distribution shape and variance ratio, the manipulated variables in this study included total sample size, sample size ratio, effect size for mean difference, alpha level for testing treatment effect, and alpha level for testing homogeneity assumption. As expected, the independent means *t*-test performed very well on the Type I error control when the homogeneity assumption was met regardless of the tenability of the normality assumption. This reminds us a well-known fact that the independent means *t*-test is robust to violation of the normality assumption when two group variances are equal. Under this condition, the independent means *t*-test is the best method to examine the difference of two independent means. This testing procedure also provides more statistical power. However, under the condition of violating the homogeneity assumption, the independent mean *t*-test could not adequately control for Type I error. Thus, two alternatives, Satterthwaite's approximate *t*-test and the conditional *t*-test, were considered in this study.

This study found that Satterthwaite's approximate *t*-test maintained adequate Type I error control and the conditional *t*-test also yielded comparable results using a large alpha level of .25 for the Folded *F*-test of variances. Both alternatives made a huge improvement of Type I error control, compared to the independent means *t*-test, when group variances were unequal. Extreme skewness (e.g., skewness = 2) contaminated the Type I error control for both alternative testing procedures. Kurtosis seemed not to have this kind of impact. Increasing total sample sizes was found in this study to be able to improve the control of Type I error rates for both testing procedures. When total sample size was 200 or more, Bradley's rates were 100% for both alternative testing procedures. Overall speaking, although Satterthwaite's approximate *t*-test provides slightly better Type I error control, the use of the conditional *t*-test may have a slight power advantage.

So, how can SAS PROC TEST keep us honest? First, with equal variances (i.e., homogeneous variances) the independent means *t*-test is the best testing procedure to examine the difference of two independent group means because it provides adequate Type I error control regardless of the tenability of the normality assumption and more statistical power. With unequal variances (heterogeneous variances) the Folded *F*-test can provide reasonable guidance in the choice between the independent *t*-test and Satterthwaite's approximate *t*-test. A large alpha level of .25 is recommended to evaluate the results of the Folded *F*-test. If the *F* value is not statistically significant at this large alpha level, then the independent means *t*-test should be used. In contrast, if the *F* value is statistically significant at this large alpha level, then Satterthwaite's approximate *t*-test should be chosen. Finally, the confidence in this conditional testing procedure increases as the sample sizes become larger. To adequately control for Type I error rate in the conditional testing procedure, a total sample size of at least 200 is recommended with extremely skewed populations (e.g., skewness = 2). For less skewed populations, a total sample size of at least 100 is recommended. With a total sample size of less than these recommended numbers in the corresponding conditions, the Type I error control resulting from any of these testing procedures may be questionable.

REFERENCES

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

Bridge, P. D., & Sawilowsky, S. S. (1999). Increasing physicians' awareness of the impact of statistics on research outcomes: Comparative power of the t-test and Wilcox Rank-Sum test in small samples applied research. *Journal of Clinical Epidemiology, 52*(3), pp. 229-235.

Hayes, A. F. & Cai, L. (2007). Further evaluating the conditional decision rule for comparing independent means. *British Journal of Mathematical and Statistical Psychology*, *60*, 217-244.

Heiman, G. W. (2011). *Basic Statistics for the Behavioral Sciences* (6th Ed.). Belmont, CA: Wadsworth Cengage Learning.

Nguyen, D., Rodriguez de Gil, P., Kim, E. S., Bellara, A. P., Kellermann, A., Chen, Y-H., Kromrey, J. D. (2012). PROC TTEST® (Old Friend), what are you trying to tell us?. Paper presented at the Southeast SAS Users Group (SESUG) Conference, Cary, NC.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin, 105*(1), 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*, 352–360.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Anh P. Kellermann
PO Box 92181
Lakeland, FL 33804-2181
E-mail: napham@mail.usf.edu