Paper PO-10

PROC TTEST® (Old Friend), What Are You Trying to Tell Us?

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

ABSTRACT

The SAS® procedure for Student's *t*-test (PROC TTEST) has been a part of the SAS system of statistical procedures since its mainframe computer days. The procedure provides hypothesis testing and confidence interval estimation for the difference between two population means. By default, the procedure provides two estimates of standard errors, two hypothesis tests, and two interval estimates: one that assumes homogeneity of variance and the other that avoids this assumption. In addition, PROC TTEST provides a test of variance homogeneity (the Folded *F* test) that ostensibly provides guidance in the choice between the two estimation methods. This paper describes past research on the accuracy of this conditional testing procedure, provides new simulation research results, and suggests guidelines for the use of the Folded *F* test in selecting between the two *t*-test approaches.

Keywords: STATISTICAL ASSUMPTIONS, ROBUSTNESS, VARIANCE HETEROGENEITY

THE INDEPENDENT MEANS T-TEST AND ALTERNATIVES

Elementary statistics courses typically introduce significance testing and inferential techniques using the independent means *t*-test, which provides a smooth transition into concepts such as statistical assumptions, robustness, Type I error control, and power. The independent means *t*-test relies on a strong assumption of equal variances (homoscedasticity), as the test statistic is a ratio of the difference in sample means to an estimate of the standard error of the difference, using a pooled variance estimate. Alternative approaches (e.g., Satterthwaite's approximate test) relax this assumption, approximating the *t* distribution and the corresponding degrees of freedom. Although the *t*-test may be one of the most basic and widely used statistical procedures to compare two group means (Hayes & Cai, 2007; Heiman, 2011), statisticians to date are still evaluating the various conditions and factors for which this test is robust under the violation of the equality of variances assumption. Many statistical textbooks (e.g., Cody & Smith, 1997; Schlotzhauer & Littell, 1997) continue recommending what Hayes and Cai (2007) call the "conditional decision rule" (p. 217), that researchers screen their samples for variance homogeneity by conducting preliminary tests (e.g., the Folded *F*-test). That is, the *t*-test assumes that the distributions of the two groups being compared are normal with equal variances. The preliminary test of the null hypothesis that $\sigma_1^2 = \sigma_2^2$ versus the alternative $\sigma_1^2 \neq \sigma_2^2$ is conducted using the test statistic: $F = \frac{S_1^2}{S_2^2}$.

Common practice has been that if the Folded *F*-test is not statistically significant (e.g., $p > .05$), then the test of $\mu_1 = \mu_2$ versus $\mu_1 \neq \mu_2$ is calculated using the independent means *t*-test:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{S_{X_1X_2} \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

On the other hand, if the preliminary test is statistically significant ($p < .05$) and in addition there are unequal sample sizes, the independent means *t*-test should be avoided and the Satterthwaite's approximate *t*-test should be used instead (Moser, Stevens, & Watts, 1989):

$$t' = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}}} \quad \text{with} \quad df = \frac{\left(\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}\right)^2}{\frac{\left(\frac{\hat{\sigma}_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{\hat{\sigma}_2^2}{n_2}\right)^2}{n_2 - 1}}$$

However, many authors (e.g., Moser et al., 1989; Zimmerman, 2004) argue about the serious disadvantages of performing preliminary tests of equality of variances and strongly recommend their discontinuance because the test of variances does not develop sufficient power to accurately avoid the independent means *t*-test.

PROC TTEST EXAMPLE

The syntax for PROC TTEST is quite simple, requiring only a CLASS statement (to identify the independent or grouping variable for the analysis) and a VAR statement (to identify the dependent or outcome variable). The CLASS statement must contain only two values or levels. If multiple variables are identified on the VAR statement, a separate analysis is conducted for each variable.

```
* +-------------------------------------------------------------------------------+
   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` predictor-variable;
`var` criterion-variable;

The TTEST Procedure
Variable: anxiety

gender	N	Mean	Std Dev	Std Err	Minimum	Maximum
F	61	3.1311	1.3841	0.1772	1.0000	5.0000
M	18	2.3889	0.5016	0.1182	2.0000	3.0000
Diff (1-2)		0.7423	1.2444	0.3338		

gender	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
F		3.1311	2.7767	3.4856	1.3841	1.1747	1.6851
M		2.3889	2.1394	2.6383	0.5016	0.3764	0.7520
Diff (1-2)	Pooled	0.7423	0.0776	1.4069	1.2444	1.0751	1.4774
Diff (1-2)	Satterthwaite	0.7423	0.3177	1.1668			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	77	2.22	0.0291
Satterthwaite	Unequal	73.738	3.48	0.0008

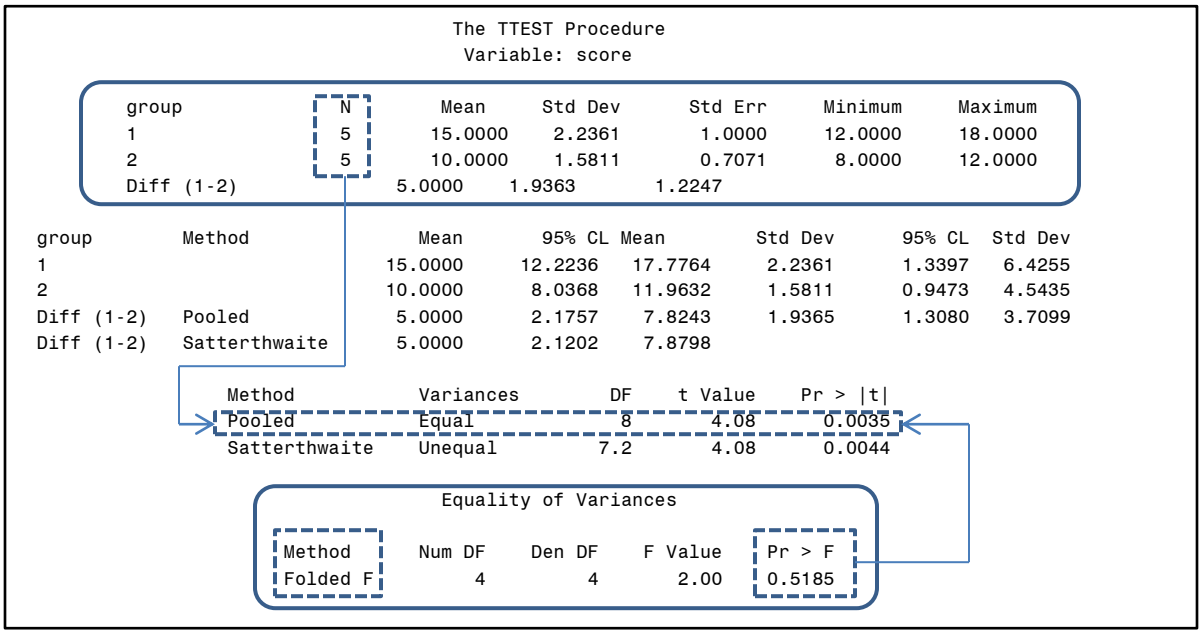
Equality of Variances

Method	Num DF	Den DF	F Value	Pr > F
Folded F	60	17	7.61	<.0001

Output1. Results of PROC TTEST: Significant Differences in Variances Observed

To determine which *t* statistic is appropriate, outputs 1 and 2 show that PROC TTEST by default performs the Folded *F* statistic to evaluate the equality of variances. If the *p*-value indicates that the difference in variances is statistically significant (e.g., less than .05) as in Output 1, the data suggest heterogeneity of variance; in addition, note that the

sample sizes are unequal (n_1 = 61, n_2 = 18). Thus, results using the Satterthwaite's test, based on unequal variances, may be most appropriate: t (73.74) = 3.48; p < .001.



The TTEST Procedure
Variable: score

group	N	Mean	Std Dev	Std Err	Minimum	Maximum
1	5	15.0000	2.2361	1.0000	12.0000	18.0000
2	5	10.0000	1.5811	0.7071	8.0000	12.0000
Diff (1-2)		5.0000	1.9363	1.2247		

group	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
1		15.0000	12.2236	17.7764	2.2361	1.3397	6.4255
2		10.0000	8.0368	11.9632	1.5811	0.9473	4.5435
Diff (1-2)	Pooled	5.0000	2.1757	7.8243	1.9365	1.3080	3.7099
Diff (1-2)	Satterthwaite	5.0000	2.1202	7.8798			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	8	4.08	0.0035
Satterthwaite	Unequal	7.2	4.08	0.0044

Equality of Variances

Method	Num DF	Den DF	F Value	Pr > F
Folded F	4	4	2.00	0.5185

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

Using data from another experiment, if the p-value of the Folded F statistic indicates that the difference in variances is not statistically significant (e.g., greater than .05) as in Output 2, the data do not suggest unequal population variances. Note that in this example, sample sizes are equal (n_1 = 5, n_2 = 5); thus, the t test for equal variances is reported: t (8) = 4.08; p < 0.004.

PREVIOUS RESEARCH ON CONDITIONAL TESTING

While some statistics textbooks do not even mention the assumption of homogeneity of variance (e.g., Gravetter & Wallnau, 2011) as one required for the t-test, which might mislead researchers into thinking that the t-test is robust to the violation of this assumption, homoscedasticity is basic and necessary for hypothesis testing because the violations of this assumption "alter Type I error rates, especially when sample sizes are unequal" (Zimmerman, 2004; p. 173). However, some research on preliminary tests suggests that the choice between the t-test and the Satterthwaite's test, conditioning on a preliminary test of the assumption of homogeneity of variance is not effective.

Moser et al. (1989) examined the effect of the significance level of the preliminary test of variance on the size and power of the t-test and Satterthwaite's tests of means and noted that when $\alpha = 0$ or $\alpha = 1$ was established for the significance level of the test of variances, it allowed applying directly the t-test or Satterthwaite's, respectively. In addition, they suggested that for equal sample sizes ($n_1 = n_2$), the t-test and the Satterthwaite's had the same power and provided very stable sizes close to the nominal alpha prescribed for the test of means. For unequal sample sizes ($n_1 \neq n_2$), the Satterthwaite's test still provided reasonable and stable sizes close to the nominal significance level. Based on their study, Moser et al. (1989) recommended applying directly the Satterthwaite's test for testing the equality of means from two independent and normally distributed populations where the ratio of the variance is unknown. Both Zimmerman (2004) and Rasch, Kubinger, and Moder (2011) found similar optimal results for the Welch-Satterthwaite separate-variance t-test if applied unconditionally whenever sample sizes were unequal and noted that the power of this test deteriorated if it was conditioned by a preliminary test. Grissom (2000) argued that it is realistic to expect heteroscedasticity in data as well as outliers, and examined the effect of these factors on variance. He also addressed issues of robustness (i.e., control of Type I error rate) in the presence of heteroscedasticity and departures from normality, for which he suggested trimming as a way to stabilize variances.

THE SIMULATION STUDY

A Monte Carlo simulation was conducted to investigate the performance of the t-test, Satterthwaite's approximate t-test, and the conditional t-test. 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 (from 1:1

to 1:20), (d) difference in means between populations (from no difference to a large effect size, Δ = 0.80), (e) alpha set for testing treatment effect or group mean difference (from α = .01 to α = .25), and (f) alpha set for testing homogeneity assumption for the conditional *t*-test (from α = .01 to α = .50).

Type I Error Control

An overall view of the Type I error control of the tests is provided in Figure 1. These boxplots describe the distributions of Type I error rate estimates under a nominal alpha level of .05 across all conditions in which the population means were identical. The first two 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.

Note in Figure 1 the great dispersion of Type I error rates for the independent means *t*-test. In some conditions, the test provides appropriate control of Type I error probability while in others the Type I error rate is very different from the nominal error rate. In contrast, Satterthwaite's approximate *t*-test provides adequate Type I error control in nearly all of the conditions simulated. The series of plots for the conditional *t*-test illustrate that the conditional test provides 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. This improvement occurs because the Folded *F*-test increases in statistical power as the alpha level is increased (see Figure 2). That is, the ability of this test to detect variance heterogeneity (and to subsequently steer us away from the independent means *t*-test and steer us to Satterthwaite's approximate *t*-test) increases with the alpha level for this test, which supports the argument of insufficient power when using a more conservative nominal alpha.

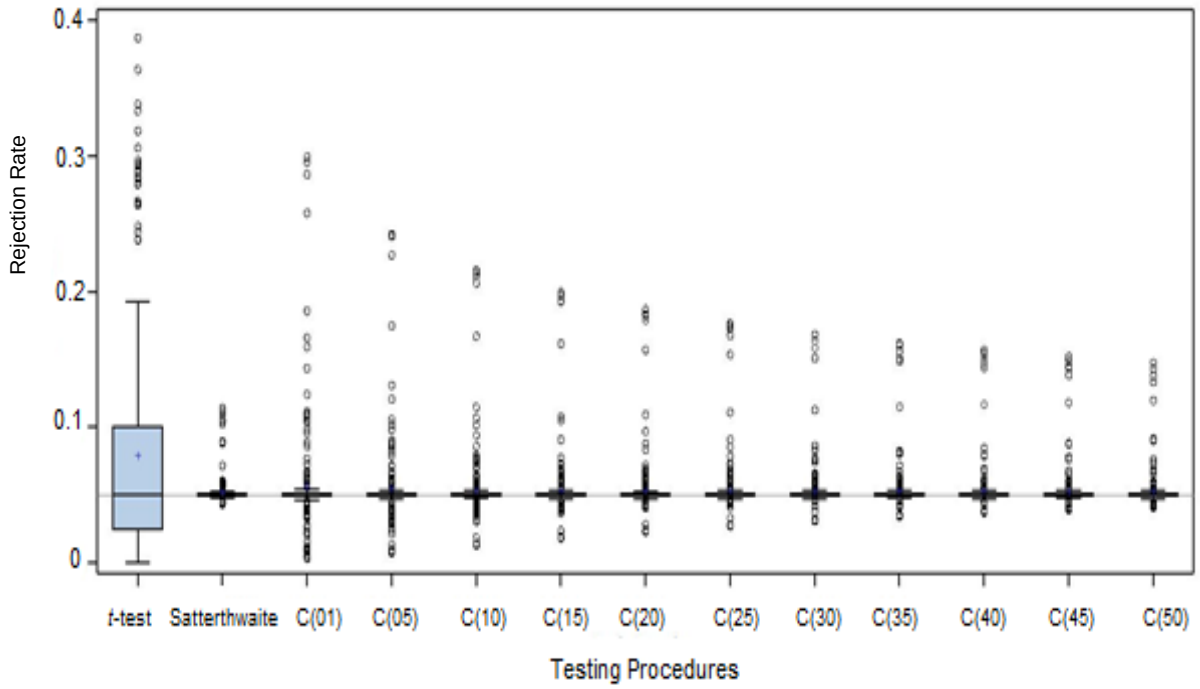


Figure 1. Distributions of Estimated Type I Error Rates Across All Simulation Conditions.

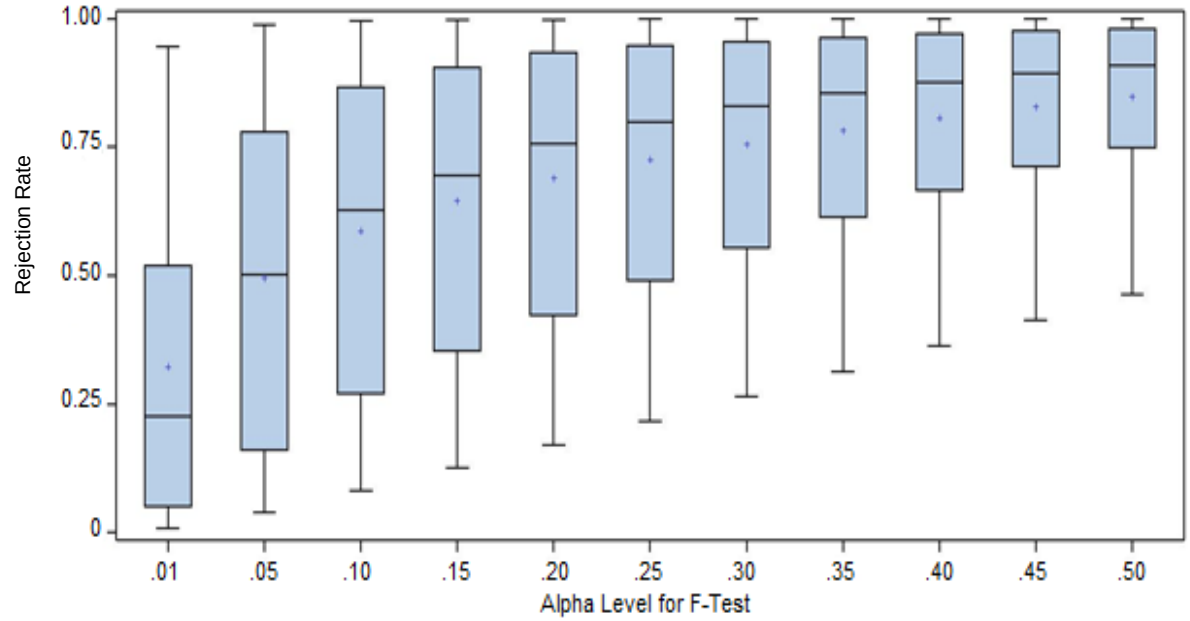


Figure 2. Distributions of Statistical Power for the Folded *F*-Test by Nominal Alpha Level.

The large dispersion of Type I error rates for the independent means *t*-test is a result of the variance heterogeneity that was included in the simulation conditions. Figure 3 presents the distributions of Type I error rates for the independent means *t*-test with the results disaggregated by population variance ratio. Note that the larger the population variance ratio, the larger the Type I error. Figures 4 and 5 present the analogous distributions of Type I error rates for Satterthwaite's approximate *t*-test and the conditional *t*-test, respectively, with an alpha level of .20 as a decision rule for the Folded *F*-test. Both the Satterthwaite's approximate *t*-test and the conditional *t*-test provided good control of Type I error rate even though the population variances in the two groups are heterogeneous.

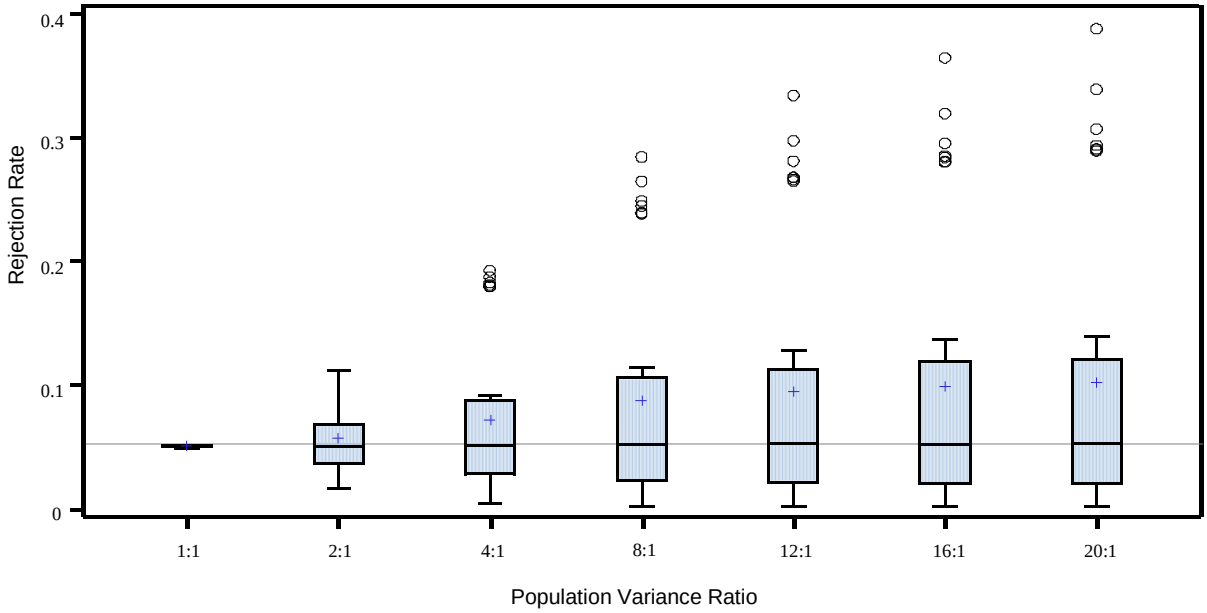


Figure 3. Distributions of Estimated Type I Error Rates by Variance Ratio at α = .05 for the Independent Means *T*-Test

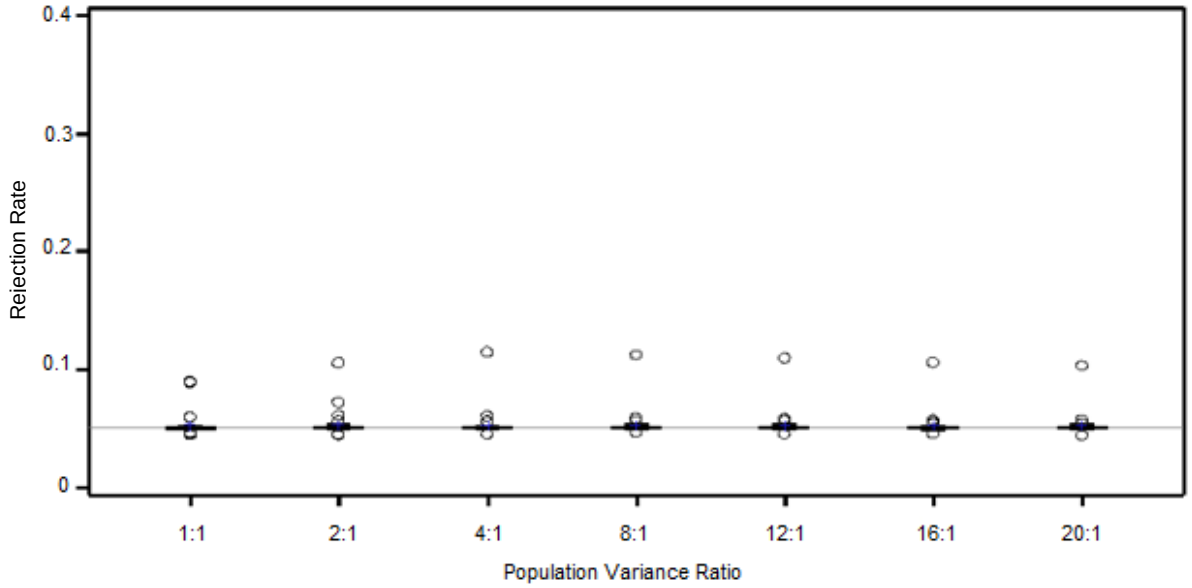


Figure 4. Distributions of Estimated Type I Error Rates by Variance Ratio at α = .05 for the Satterthwaite's Approximate *T*-Test

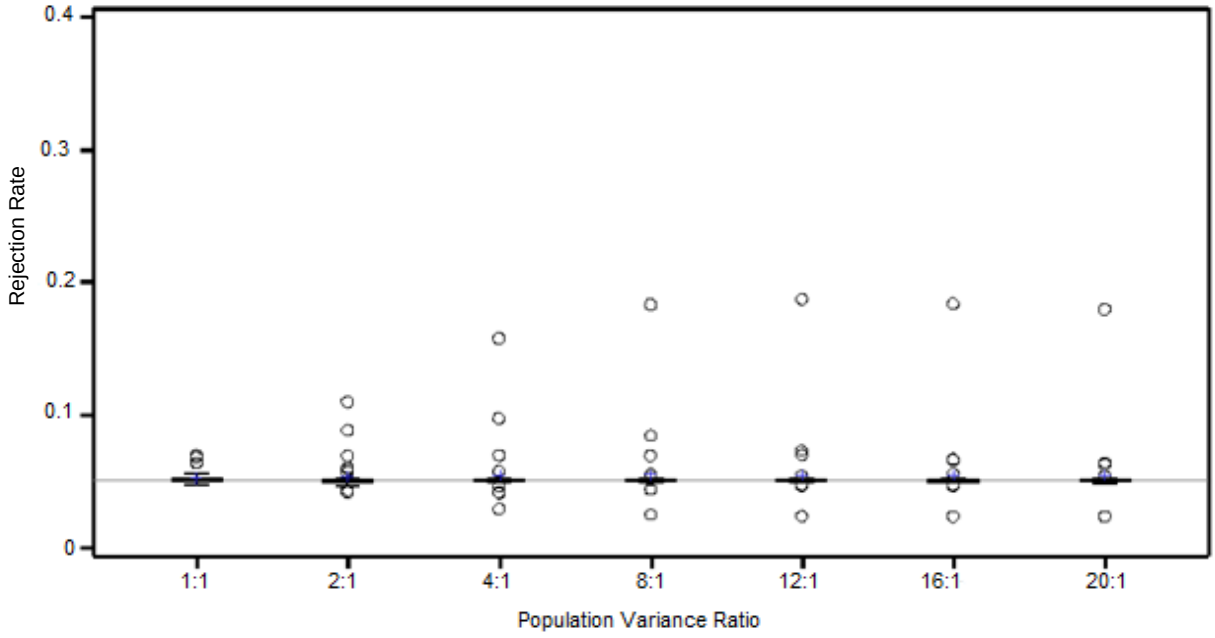


Figure 5. Distributions of Estimated Type I Error Rates by Variance Ratio at α = .05 for the Conditional *T*-Test Using α = .20 for the Folded *F*-Test

Of course, the independent means *t*-test is known to be relatively robust to violations of the assumption of variance homogeneity if the sample sizes in the two groups are equal. This phenomenon is illustrated in Figure 6. Note that the Type I error rate for the independent means *t*-test is maintained near the nominal .05 level if sample sizes are equal. With disparate sample sizes in the two groups, the independent means *t*-test either becomes conservative (Type I error control lower than the nominal alpha level) or liberal (Type I error control higher than the nominal level) depending upon the relationship between sample size and population variance. In contrast, both Satterthwaite's approximate *t*-test (Figure 7) and the conditional *t*-test (Figure 8) evidence much improved Type I error control under variance heterogeneity when samples sizes are unequal.

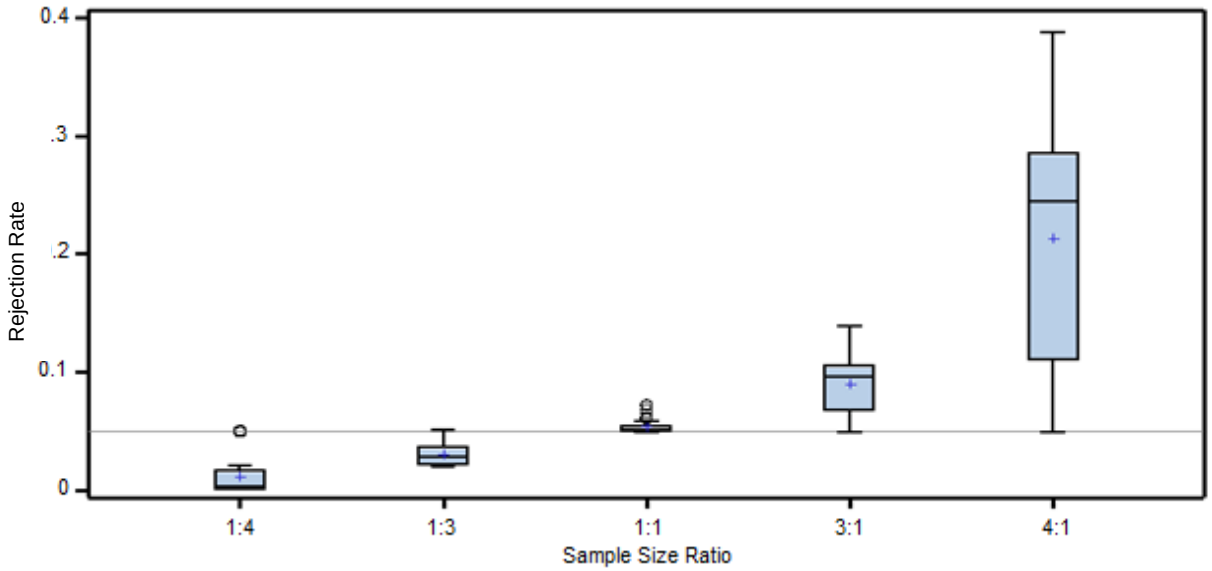


Figure 6. Distributions of Estimated Type I Error Rates by Sample Size Ratio for the Independent Means *T*-Test

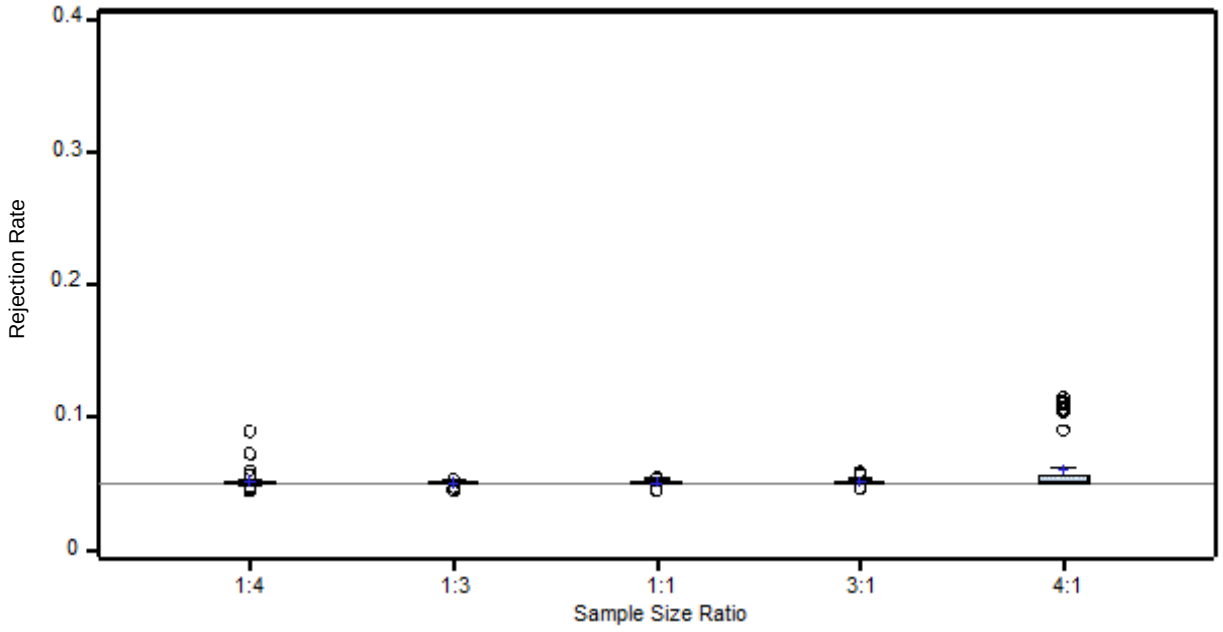


Figure 7. Distributions of Estimated Type I Error Rates by Sample Size Ratio for Satterthwaite's Approximate *T*-Test

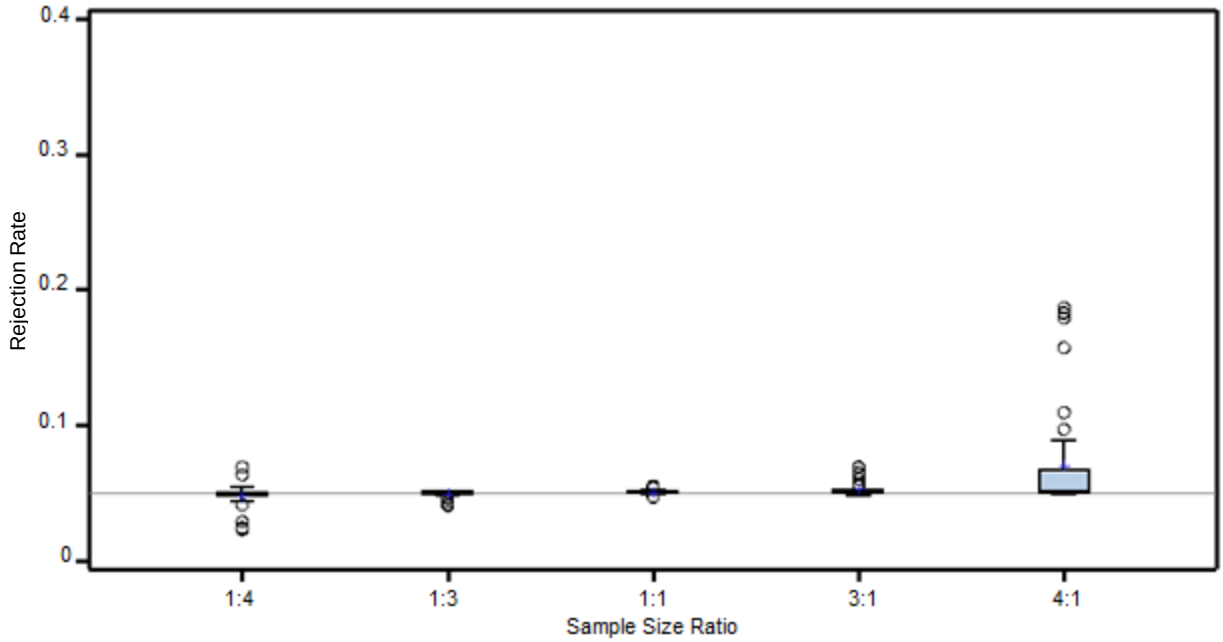


Figure 8. Distributions of Estimated Type I Error Rates by Sample Size Ratio for the Conditional *T*-Test

Using a larger sample size does not improve the performance of the independent means *t*-test (Figure 9), but larger samples provide substantial improvements to the Type I error control of both Satterthwaite's approximate *t*-test (Figure 10) and the conditional *t*-test (Figure 11).

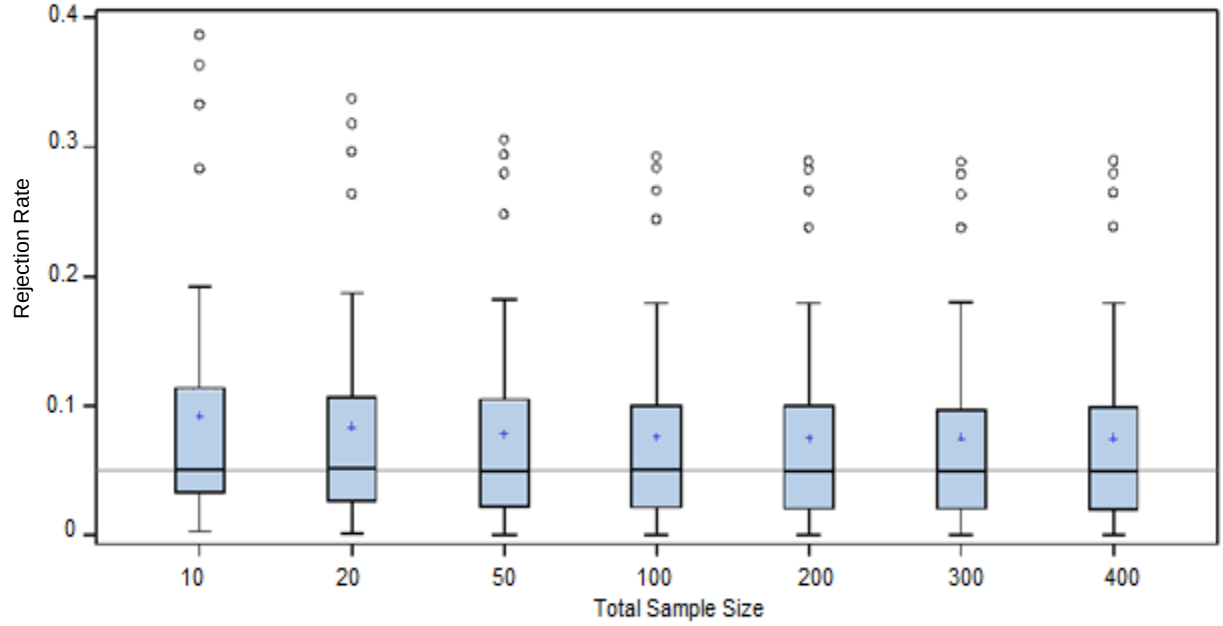


Figure 9. Distributions of Estimated Type I Error Rates for Independent Means *T*-Test by Total Sample Size

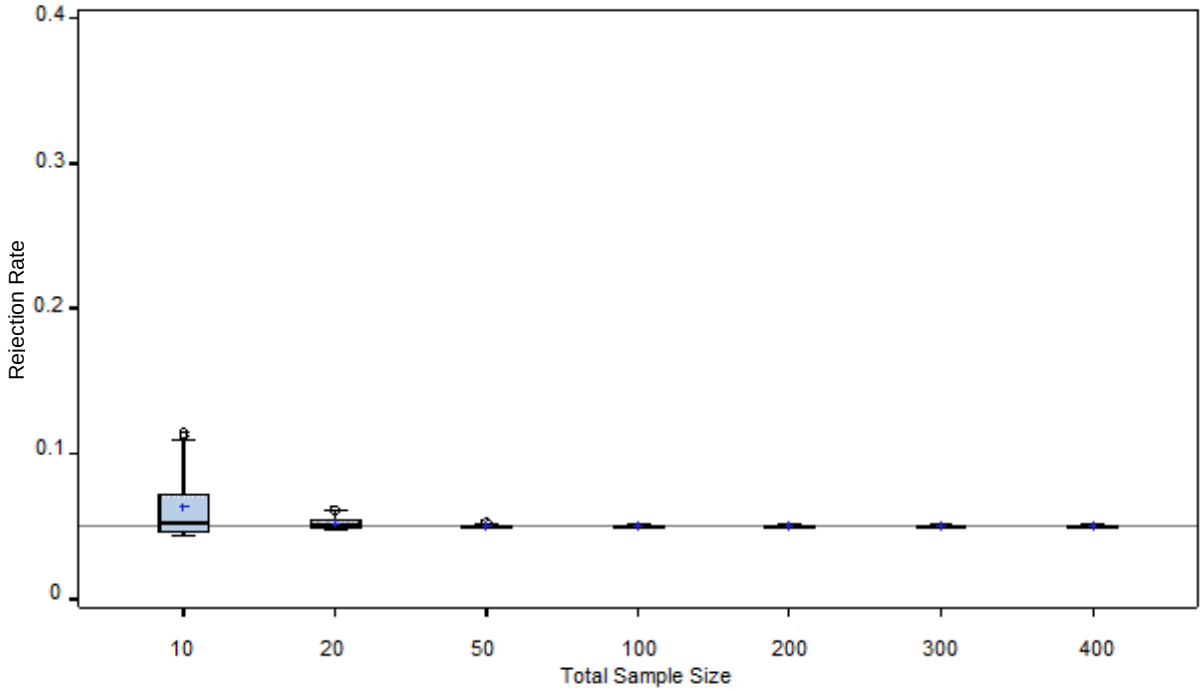


Figure 10. Distributions of Estimated Type I Error Rates for Satterthwaite Test by Total Sample Size

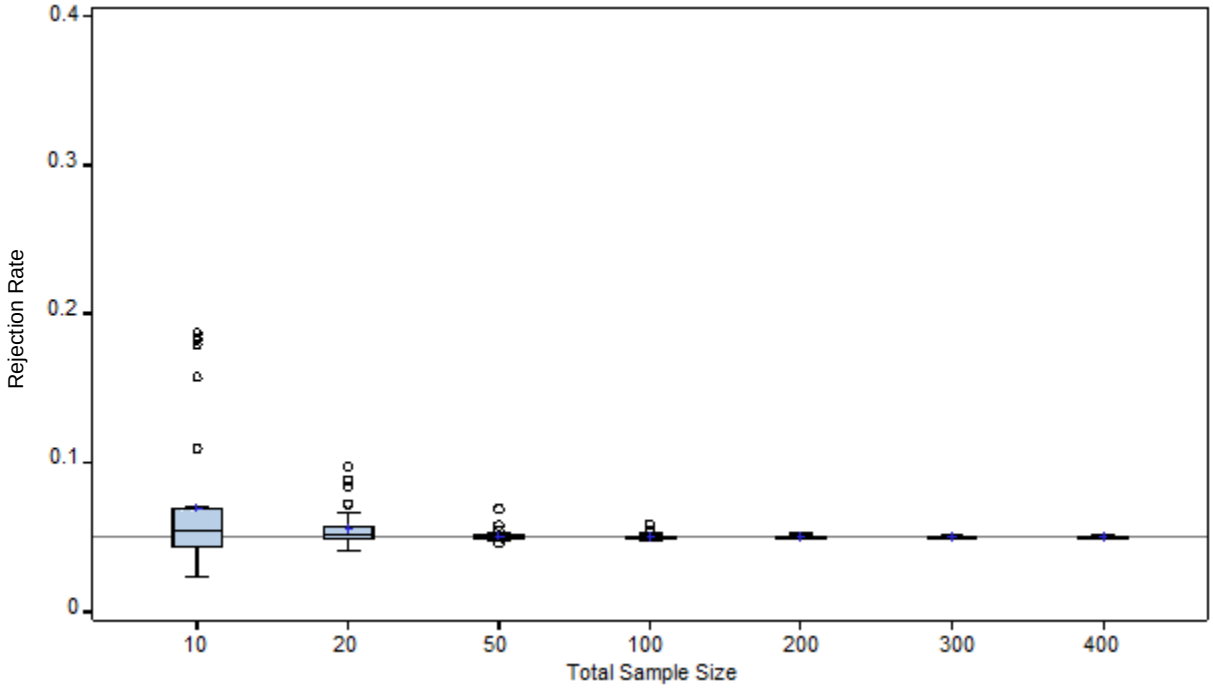


Figure 11. Distributions of Estimated Type I Error Rates for Conditional *T*-Test by Total Sample Size

Bradley (1978) suggested guidelines for determining adequate Type I error control. Specifically, his "liberal criterion" for robustness suggests that an actual Type I error rate that is within the interval of $\alpha_{nominal} \pm 0.5\ \alpha_{nominal}$ represents acceptable Type I error control. Using this criterion, the Type I error control of these tests are summarized in Table 1. With equal sample sizes, all three approaches provided adequate control of Type I error in all simulation conditions. With a sample size ratio of 2:3 or 3:2, both the conditional test and Satterthwaite's approximate *t*-test provided adequate control under all conditions, while the independent means *t*-test provided adequate control in only 59% of the conditions with a 2:3 sample size ratio and in only 29% of the conditions with a 3:2 sample size ratio. With the most extreme sample size ratios examined, the Satterthwaite approximate *t*-test provided slightly better results than the conditional test (98% vs. 92% with a 1:4 ratio, and 86% vs. 82% with a 4:1 ratio), but both procedures were strikingly better than the independent means *t*-test which provided adequate control in only 14% of the conditions.

Sample Size Ratio	Independent means *t*-test	Conditional *t*-test [C (20)]	Satterthwaite's approximate *t*-test
1:4	0.14	0.92	0.98
2:3	0.59	1.00	1.00
1:1	1.00	1.00	1.00
3:2	0.29	1.00	1.00
4:1	0.14	0.82	0.86

Table 1. Proportion of Conditions with Adequate Type I Error Control by Bradley's Criterion

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 the Satterthwaite's approximate *t*-test and the conditional *t*-test procedures evidenced adequate Type I error control by Bradley's (1978) benchmark. Figure 12 presents a scatter plot of the power estimates for the conditional *t*-test and Satterthwaite's approximate *t*-test (at a nominal alpha of .05). As evident in this figure, all of the differences in power were small. However, the conditional *t*-test, using alpha = .20 for the Folded *F*-test of variances was more powerful in 31% of the conditions while Satterthwaite's approximate *t*-test was more powerful in only 15% 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.

CONCLUSIONS AND IMPLICATIONS

This simulation study was intended to explore the performance of the independent means *t*-test, Satterthwaite's approximate *t*-test, and the conditional *t*-test under the manipulated conditions of total sample size, sample size ratio between groups, variance ratio between populations, the difference in means between populations, alpha level for testing the treatment effect, and alpha level for testing the homogeneity assumption for the conditional *t*-test. 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 conditions. To control for Type I error rate, Satterthwaite's approximate *t*-test performed very well regardless of the ratios of population variances and sample sizes in the two groups used in this study (the exception being conditions with very small sample sizes). More advantageously, increasing the total sample size (e.g., as few as 100) improved the control of Type I error rate for Satterthwaite's approximate *t*-test.

As expected, the independent means *t*-test showed adequate control of Type I error rate when population variances or sample sizes in the two groups were equal. These results re-emphasize two well-known factors: (1) the independent means *t*-test requires the homogeneity assumption to be met if Type I error control is to be maintained, and (2) the independent means *t*-test is robust to the violation of the homogeneity assumption when the sample sizes are equal. If the sample sizes are not equal, the Type I error rate of the independent means *t*-test either becomes conservative or liberal. This study also indicates that the Type I error rate of the conditional *t*-test is affected by the alpha level for the Folded *F*-test that is used to examine the homogeneity assumption of population variances. The more conservative alpha levels for the Folded *F*-test resulted in larger Type I error rates for the conditional test because of lower statistical power, such that the Folded *F*-test may not be able to detect the true difference between population variances. This leads us to re-consider the conventional procedures for examining the difference between two population means. Thus, the conditional *t*-test, the third approach in this study, that uses a relatively large alpha level for the Folded *F*-test may be an appropriate alternative. Furthermore, increasing the total sample size does not improve the control of Type I error rate for the independent means *t*-test, but larger samples provide better Type I error control for the conditional *t*-test.

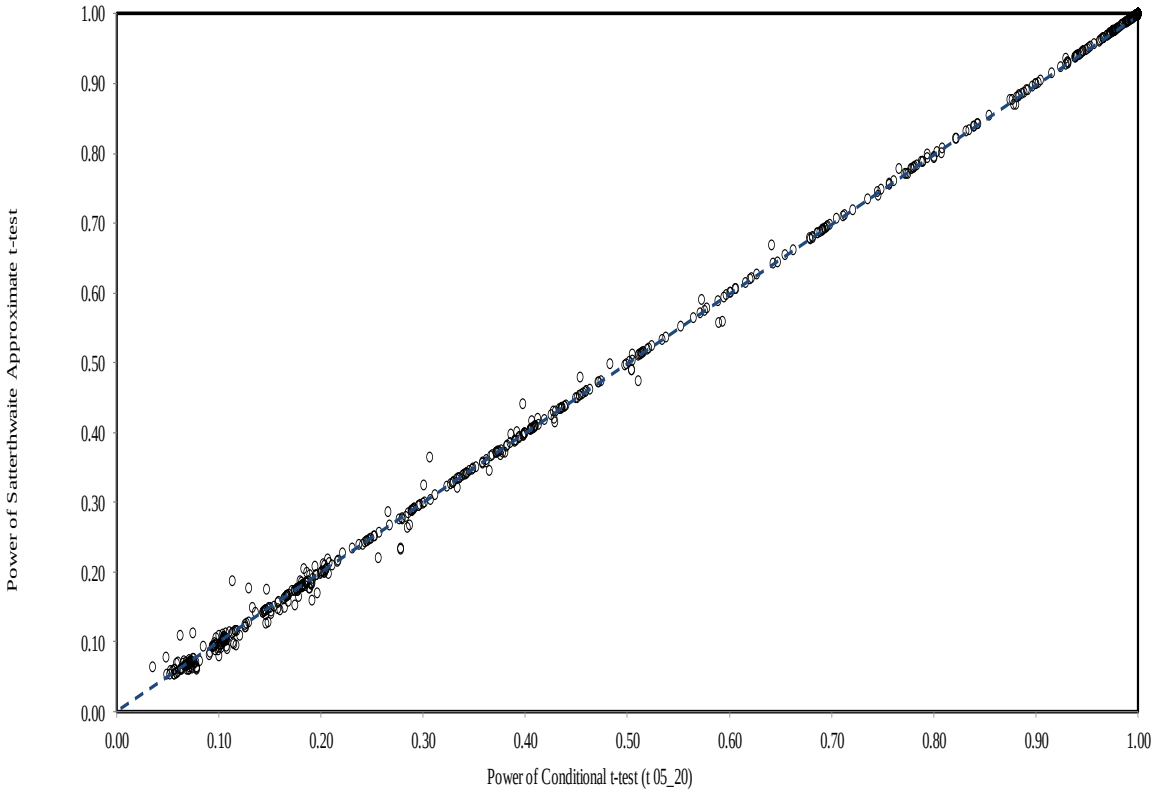


Figure 12. Scatterplot of Power Estimates for the Conditional *T*-Test and Satterthwaite's Approximate *T*-Test.

Although the conditional *t*-test did not perform as well as Satterthwaite's approximate *t*-test over all of the conditions examined, it evidenced a notable improvement in control of Type I error rate compared to the independent means *t*-test when the alpha level for the Folded *F*-test was increased (which led to a concomitant increase in the statistical power of the *F*-test). Under the conditions of different population variance ratios and samples sizes with an alpha level of .20 for the Folded *F*-test, the conditional *t*-test performed nearly as well as Satterthwaite's approximate *t*-test; that is, the conditional *t*-test provided acceptable Type I error rates despite the large ratio of population variances (e.g., 1:20) or of sample size (e.g., 1:4). Although Satterthwaite's approximate *t*-test provides superior Type I error control in the most extreme conditions, the conditional *t*-test may be the best choice because it can provide more power than Satterthwaite's approximate *t*-test.

So, what is our old friend PROC TTEST trying to tell us? First, with equal sample sizes the independent means *t*-test will probably provide adequate Type I error control regardless of the tenability of the homogeneity of variance assumption. With unequal sample sizes, the Folded *F*-test can provide reasonable guidance in the choice between the independent means *t*-test and Satterthwaite's approximate *t*-test. To evaluate the results of the Folded *F*-test, a relatively large alpha level is recommended (e.g., .20). If the *F* value is statistically significant at this large alpha level, then Satterthwaite's approximate *t*-test should be used. Conversely, if the *F* value is not statistically significant at this large alpha level, then the independent means *t*-test should be applied. Finally, our confidence in this conditional testing procedure should increase as our sample sizes become larger (with a total sample size of less than 20, 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.

Cody, R. P. & Smith, J. K. (1997). *Applied Statistics and the SAS Programming Language* (4th Ed.). Upper Saddle River, NJ: Prentice-Hall.

Grissom, R. J. (2000). Heterogeneity of variance in clinical data. *Journal of Consulting and Clinical Psychology*, *68*, 155-165.

Gravetter, F. J. (2011). *Essentials of Statistics for the Behavioral Sciences* (6th Ed.). Belmont, CA: Wadsworth, Cengage Learning.

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.

Moser, B. K., Stevens, G. R., & Watts, C. L. (1989). The two-sample t-test versus Satterthwaite's approximate F test. *Communications in Statistics: Theory and Methods*, *18*, 3963-3975.

Rasch, D., Kubinger, K. D., & Moder, K. (2011). The two-sample t-test: pre-testing its assumptions does not pay off. *Statistical Papers*, *52*, 219-231.

SAS Institute, Inc. (2002-2003). SAS version 9.2

Schlotzhauer, S. D., & Littell, R. C. (1997). *SAS System For Elementary Statistical Analysis* (2nd Ed.). Cary, NC: SAS Institute, Inc.

Zimmerman, D. W. (2004). A note on preliminary tests of equality of variances. *British Journal of Mathematical and Statistical Psychology*, *57*, 173-181.

CONTACT INFORMATION

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

Diep Nguyen
University of South Florida
4202 East Fowler Ave., EDU 105
Tampa, FL 33620

Work Phone: (813) 974-3220
Fax: (813) 974-4495
Email: diepnguyen@usf.edu