Journal of Modern Applied Statistical Methods

Volume 17 | Issue 2 Article 9

3-11-2019

Can One Test Fit All? Responses to the Article "Striving for Simple but Effective Advice for Comparing the Central Tendency of Two Populations" (Ruxton & Neuhäuser, 2018)

Diep Nguyen
University of South Florida, diepnguyen@usf.edu

Eun Sook Kim
University of South Florida, ekim3@usf.edu

Yi-Hsin Chen
University of South Florida, ychen5@usf.edu

Recommended Citation
Nguyen, D., Kim, E. S., & Chen, Y.-H. (2018). Can One Test Fit All? Responses to the Article "Striving for Simple but Effective Advice for Comparing the Central Tendency of Two Populations" (Ruxton & Neuhäuser, 2018). *Journal of Modern Applied Statistical Methods, 17*(2), eP2822. doi: 10.22237/jmasm/1552331481

Journal of Modern Applied Statistical Methods
November 2018, Vol. 17, No. 2, eP2822
doi: 10.22237/jmasm/1552331481

בס"ד

ISSN 1538 − 9472

INVITED DEBATE

Can One Test Fit All? Responses to the Article "Striving for Simple but Effective Advice for Comparing the Central Tendency of Two Populations" (Ruxton & Neuhäuser, 2018)

Diep Nguyen
University of South Florida
Tampa, FL

Eun Sook Kim
University of South Florida
Tampa, FL

Yi-Hsin Chen
University of South Florida
Tampa, FL

Responses to suggestions made by Ruxton & Neuhäuser (2018) regarding Nguyen et al. (2016) are given.

Keywords: Type I error control, statistical power, parametric tests, independent means *t*-test, Satterthwaite's approximate *t*-test, conditional *t*-test

Ruxton and Neuhäuser (2018) provided some comments and recommendations for comparing the two group means. Although they offered some helpful information on this topic, some are not consistent or supported by Nguyen et al. (2016). It would be helpful to respond to these issues to provide a clearer understanding the motivating article.

Comment #1:

> *Specifically, their Figures 1-2 (pp. 148-149) demonstrate Satterthwaite's test having better control of type I error than the* t*-test and better or broadly equivalent control to any the 11 variants of their conditional procedure considered. Their Figure 7 (p. 154) compares the power of Satterthwaite's test with that of the conditional procedure, and the dominant feature of the graph is the very strong similarity of*

doi: 10.22237/jmasm/1552331481 | Accepted: May 15, 2018; Published: March 11, 2019.
Correspondence: Diep Nguyen, diepnguyen@usf.edu

> *performance in almost all test scenarios. There are no substantial parts of the extensive set of scenarios explored where the conditional procedure demonstrated considerably better performance in either control of type I error rate or power (and definitely not in both).* (Ruxton & Neuhäuser, 2018, p. 4)

Response: This conclusion is not in line with results from our study, which provided recommendations based on specific simulation conditions conducted in this study, and we found that no single test among three examined ones (i.e. the independent means *t*-test, Satterthwaite's test, and the conditional *t*-test) was superior in all or most scenarios. Specifically, although Satterthwaite's test, on average, demonstrated better performance in terms of Type I error control, it is not always better than the independent *t*-test or conditional *t*-test in terms of statistical power. With equal group sizes, statistical power of the independent means *t*-test was higher (in 83% of all investigated conditions) or equal (in 17% of all investigated conditions) than power of Satterthwaite's test. In other words, the power of Satterthwaite's test was never better than that of the independent means *t*-test with balanced design. The Type I error rates for the independent means *t*-test was also adequately controlled when the two groups had equal sizes (91% of the conditions met Bradley's liberal criterion for Type I error control) or equal variances (all conditions satisfied Bradley's liberal criterion) as shown in Table 2 (p. 153) and Figures 3 and 4 (p. 150) of our article (Nguyen et al., 2016). In order to see the detailed performance of the conditional *t*-test and Satterthwaite's test by certain scenarios (i.e., simulation factors) in terms of statistical power, readers should refer to Table 3 (p. 155) where it demonstrated the power estimate comparison of these two tests by simulation factors explored in our study. As shown in this table, the conditional *t*-test was more powerful in more design factor conditions than Satterthwaite's test. In addition, as indicated from Tables 2 and 3, for equal variance conditions the conditional *t*-test evidenced larger proportions of meeting Bradley's liberal criterion as well as higher power than Satterthwaite's test. Clearly, under these conditions of balanced samples and equal variances, the independent mean *t*-test or the conditional *t*-test not only adequately controlled for Type I error but also outperformed Satterthwaite's test in terms of statistical power.

Comment #2:

> *Some authors consider preliminary testing of both equality of variance and normality before selecting a test of the means of two independent*

> *samples (e.g. Perry, 2003), but we do not feel that this offers any attraction over the approach suggested here. Given this line of reasoning, it is no surprise that the function t.test in R calculates the Welch-Satterthwaite test rather than the classical* t *test by default as the "Welch procedure is generally considered the safer one" (Dalgaard, 2002, p. 89).* (Ruxton & Neuhäuser, 2018, p. 4)

Response: Two popular statistical software programs (SAS and SPSS) use the independent means *t*-test as default to compare the two group means. In these two programs, the independent means *t*-test is used when equal variances are assumed, and Satterthwaite's test result is recommended when equal variances are not assumed. The independent means *t*-test is also provided for group mean comparisons in other commonly-used statistical software programs such as Minitab and SYSTAT. In the R package t.test, the *t*-test output is produced when selecting the option var.equal = True, i.e. with equal variances (Taeger & Kuhnt, 2014).

However, the default or selection of statistical tests in software programs might not always be the optimal choice. That is why there is a need and responsibility for researchers to inform the developers of those statistical software programs about the behaviors of selected statistical methods through studies from different aspects. Our simulation study was such an attempt to examine the performance of three tests to compare two independent group means and report in which situations a particular test behaves well or not.

It was suggested in other studies to examine the assumptions of homogeneity of variance and/or normality before selecting statistical techniques to compare two independent group means and mentioned serious consequences of ignoring these assumptions (e.g. Olsen, 2003; Choi, 2005; Nimon, 2012; Hoekstra, Kiers, & Johnson, 2012).

Comment #3:

> *If distributions deviate strongly from normality (and especially if these distributions are skewed), then both the* t*-test and Satterthwaite's test become unreliable in terms of control of type I error rate. No conditional strategy selecting between them will thus provide good control, especially not one conditional on an* F*-test (which itself not only rests on the assumption that both populations are normally distributed but is also known to be extremely sensitive to non-normality, e.g. Box, 1953).* (Ruxton & Neuhäuser, 2018, p. 4)

Response: Results from our study do not completely support this statement. First, our simulation results showed that the average power of the Folded *F*-test remained consistent irrespective of distribution shapes.

Second, based on different population shape conditions examined in our study (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 $\gamma_2 = 5.00$, as well as $\gamma_1 = 0.00$ and $\gamma_2 = 0.00$ for the normal distribution, where γ_1 and γ_2 represent skewness and kurtosis, respectively), Type I error control for both Satterthwaite's test and the conditional *t*-test, but NOT for the independent means *t*-test, was impacted by skewness (e.g., skewness = 2). The impact of skewness on Satterthwaite's test on Type I error control, especially when the larger sample size is associated with smaller heterogeneity of variance, was also in line with results from the study of Zimmerman (2006). In fact, as stated in our study, the independent means *t*-test adequately controls Type I error rates (i.e., 100% of total conditions met Bradley's liberal criterion when the variances of two groups were equal and 91% satisfied Bradley's criterion when the group sizes were identical) across all population shapes. The independent means *t*-test became questionable in terms of Type I error control when either the variances or the group sizes were not equal. The robustness of the independent means *t*-test with heterogeneity of variance when group sizes are equal with large sample sizes has been known for a long time and is mentioned in several studies such as Boneau (1960), Glass, Peckham, and Sanders (1972), Ruxton (2006), Nimon (2012), and Delacre, Lakens, and Leys (2017).

In conclusion, based on results in our study we do not recommend the sole use of Satterthwaite's test to compare two independent group means in all scenarios. Although Satterthwaite's test, on average, outperformed the independent means *t*-test and the conditional *t*-test in controlling for Type I error, it was not the only optimal or the best choice for all investigated conditions. As indicated in the results of our simulation study, the independent means *t*-test achieved acceptable Type I error control in all or most conditions when the group variances or group sizes were identical, regardless of the population shapes. Moreover, in those conditions (i.e., balanced samples), the conditional *t*-test always had greater power than Satterthwaite's test.

References

Boneau, C. A. (1960). The effects of violations of assumptions underlying the t test. *Psychological Bulletin, 57*(1), 49-64. doi: 10.1037/h0041412

Choi, P. T. (2005). Statistics for the reader: What to ask before believing the results. *Canadian Journal of Anesthesia, 52*(Suppl 1), R46-R46. doi: 10.1007/bf03023086

Delacre, M., Lakens, D., & Leys, C. (2017). Why psychologists should by default use Welch's *t*-test instead of Student's *t*-test. *International Review of Social Psychology, 30*(1), 92-101. doi:10.5334/irsp.82

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. doi: 10.2307/1169991

Hoekstra, R., Kiers, H. A., & Johnson, A. (2012). Are assumptions of well-known statistical techniques checked, and why (not)? *Frontiers in Psychology, 3*, 137. doi: 10.3389/fpsyg.2012.00137

Nguyen, D. T., Kim, E. S., Rodriguez de Gil, P., Kellermann, A., Chen, Y.-H., Kromrey, J. D., & Bellara, A. (2016). Parametric tests for two population means under normal and non-normal distributions. *Journal of Modern Applied Statistical Methods, 15*(1), 141-159. doi: 10.22237/jmasm/1462075680

Nimon, K. F. (2012). Statistical assumptions of substantive analyses across the general linear model: A mini-review. *Frontiers in Psychology, 3*, 322. doi: 10.3389/fpsyg.2012.00322

Olsen, C. H. (2003). Review of the use of statistics in infection and immunity. *Infection and Immunity, 71*(12), 6689-6692. doi: 10.1128/iai.71.12.6689-6692.2003

Ruxton, G. D. (2006). The unequal variance *t*-test is an underused alternative to Student's *t*-test and the Mann-Whitney *U* test. *Behavioral Ecology, 17*(4), 688-690. doi: 10.1093/beheco/ark016

Ruxton, G., & Neuhäuser, M. (2018). Striving for simple but effective advice for comparing the central tendency of two populations. *Journal of Modern Applied Statistical Methods, 17*(2), eP2567. doi: 10.22237/jmasm/1551908612

Taeger, D., & Kuhnt, S. (2014). Statistical hypothesis testing with SAS and R. West Sussex, UK: John Wiley & Sons. doi: 10.1002/9781118762585

Zimmerman, D. W. (2006). Two separate effects of variance heterogeneity on the validity and power of significance tests of location. *Statistical Methodology, 3*(4), 351-374. doi: 10.1016/j.stamet.2005.10.002