Journal of Modern Applied Statistical Methods

Volume 17 | Issue 2 Article 6

3-6-2019

Striving for Simple but Effective Advice for Comparing the Central Tendency of Two Populations

Graeme Ruxton
University of St Andrews, gr41@st-andrews.ac.uk

Markus Neuhäuser
Koblenz University of Applied Sciences, RheinAhrCampus

Recommended Citation
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

Journal of Modern Applied Statistical Methods
November 2018, Vol. 17, No. 2, eP2567
doi: 10.22237/jmasm/1551908612

בס"ד

ISSN 1538 − 9472

INVITED DEBATE

Striving for Simple but Effective Advice for Comparing the Central Tendency of Two Populations

Graeme Ruxton
University of St. Andrews
St. Andrews, UK

Markus Neuhäuser
Koblenz University of Applied Sciences,
RheinAhrCampus
Remagen, Germany

Nguyen et al. (2016) offered advice to researchers in the commonly-encountered situation where they are interested in testing for a difference in central tendency between two populations. Their data and the available literature support very simple advice that strikes the best balance between ease of implementation, power and reliability. Specifically, apply Satterthwaite's test, with preliminary ranking of the data if a strong deviation from normality is expected, or is suggested by visual inspection of the data. This simple guideline will serve well except when dealing with small samples of discrete data, when more sophisticated treatment may be required.

Keywords: Type I error control, statistical power, Satterwaite's test, Welch test, *t*-test, conditional testing

Nguyen et al. (2016) offer advice on selecting an appropriate method to compare the central tendencies of two populations. We believe their data and the available literature both support much simpler advice.

Nguyen et al. (2016) compare three methods: (i) the classical Student's *t*-test; (ii) Satterthwaite's test (more commonly called Welch's *t*-test, the unequal variances *t*-test, or the Aspin-Welch-Satterthwaite test); and (iii) conditional use of either (i) or (ii) depending on the outcome of an *F*-test for equality of variance. Their advice to researchers is as follows:

doi: 10.22237/jmasm/1551908612 | Accepted: August 2, 2017; Published: March 6, 2019.
Correspondence: Graeme Ruxton, gr41@st-andrews.ac.uk
Graeme Ruxton is Professor of Ecology in the School of Biology at the University of St Andrews, UK.

> *With equal sample size the independent means* t*-test is the appropriate testing procedure to examine the difference of two independent group means because it provides adequate Type I error control and more statistical power. With unequal sample size the Folded* F*-test can provide reasonable guidance in the choice between the independent* t*-test and Satterthwaite's 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 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 of 2). For less skewed populations, a total sample size of at least 100 is recommended. With a total sample size smaller than these recommended in the corresponding conditions, the Type I error control resulting from any of these testing procedures may be questionable.* (pp. 157-158)

We believe that the consensus of the available literature and of their simulations supports a much simpler set of recommendations, at least for continuous data or when there are a few ties only:

> *Satterthwaite's test can always be applied, with preliminary ranking of the data, if a strong deviation from normality is expected or is suggested by visual inspection of the data.*

The justification for our stance is as follows: When populations are normal and variances are equal then Satterthwaite's test gives near-identical performance to Student's *t*-test in terms of both type I error rate and power (e.g. Moser, Stevens, & Watts, 1989). However, if variances differ, then it is well known that Satterthwaite's test maintains the type I error rate at the nominal level but the *t*-test often shows substantial deviations (Coombs, Algina, & Oltman, 1996; Zimmerman & Zumbo, 1989). It is also well established that the power of the *t*-test is generally larger than that of the Satterthwaite test, but the difference is never substantial (Moser et al., 1989; Moser & Stevens, 1992; Coombs et al., 1996). From these results no conditional strategy of switching between these two tests will offer

substantially better performance than always adopting Satterthwaite's test, and such conditional strategies could easily perform worse. These conclusions are entirely congruent with the results presented by Nguyen et al. 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 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). This is in line with the conclusions of previous studies (Gans, 1991; Moser & Stevens, 1992). There are further reasons for not recommending procedures based on preliminary testing for equality of variance (see discussions in Markowski & Markowski, 1992; Quinn & Keough, 2002; Rasch, Kubinger, & Moder, 2011). 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).

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). Alternative tests show better qualities than either the *t*-test and Satterthwaite's test (e.g. Coombs et al., 1996; Keselman, Othman, Wilcox, & Fradette, 2004; Neuhäuser & Ruxton, 2009), but none of these are commonly used. We do not recommend formal preliminary testing for normality (see Ruxton, Wilkinson, & Neuhäuser, 2015). However, Zimmerman and Zumbo (1993) demonstrated reasonably good performance of Satterthwaite's test when normality was violated, providing the test was carried out after ranking the data. They found that this procedure also outperformed the non-parametric Wilcoxon's rank sum test when variances were unequal across a simulation study involving eight different non-normal distribution types. A more recent study (Cribbie, Wilcox, Bewell, & Keselman, 2007) found that applying the Satterthwaite's test to ranked data offered

better power in many situations than even recently-developed methods such as the Brunner-Munzel test (Brunner & Munzel, 2000). Zimmerman (1998) also suggests an alternative procedure for non-normal data again involving pre-processing the data before applying Satterthwaite's test, involving downweighting values from the extremes of the sample. Although he demonstrates the effectiveness of this procedure, its performance is not compared with the ranking procedure; pending such an investigation we recommend the ranking approach because of its simplicity. We do not, however, recommend ranking unless there is concern (based on prior knowledge of the measured variable or visual inspection of the data) of substantial deviation from normality, since working with unranked data allows more straightforward interpretation of test results. Another possibility could be a *t*-test evaluated by randomization. However, a significant result in this test, called the Fisher-Pitman permutation test, does not necessarily provide evidence for a difference in means when variances differ (Boik, 1987; Neuhäuser & Manly, 2004). When variances are homogeneous, this test can be outperformed by the Wilcoxon's rank sum test, equivalent to ranking the data before Student's *t*-test (Weber & Sawilowsky, 2009). Thus, a combination of the two tests is useful if variances do not differ (Neuhäuser, 2015). Specifically, Neuhäuser (2015) demonstrated that a test based on the maximum of *t*-statistics calculated from Student's *t*-test and from Wilcoxon's rank-sum test is a more powerful strategy that always selecting either of the single tests across a range of distributions and avoids complex selection protocols. Further, since its power is close to the more powerful of the two tests, little advantage over the maximization test could be achieved by a protocol that allowed effective selection of one or other of these tests.

It should also be noted that, to this point, we have essentially considered testing in the Behrens-Fisher situation where we are interested in exploring whether a difference in central tendency might occur without making the assumption that the scale (i.e., the spread) of values will necessarily be the same. There is another situation that some (e.g., Sawilowsky, 2002) consider to be more realistic in many applied settings: where we are still interested in exploring whether there is a difference in central tendency, but crucially we also expect that if there is such a difference then it will also be accompanied by a change in scale. That is, we expect that the mechanism that might induce a change in average value will also affect the spread of values in a predictable direction. It can be argued that such a situation is especially appropriate when homogeneous experimental units are randomly assigned to different treatments or groups (e.g., Neuhäuser, 2002). If on the basis of understanding of the system this situation, called informative variance heterogeneity by Hothorn and Hauschke (1998), applies, then there are more

effective alternatives to Satterthwaite's test (see Blair & Sawilowsky, 1993b). However, notice that these alternatives assume that in the event of no effect both the means and variances of the two populations would be the same. Some methods assume that the population with the higher mean also has a higher variance. This approach can also be extended to comparing more than two populations (Blair & Sawilowsky, 1993a). When a so-called location-scale test rejects the null hypothesis that the means as well as the variances of the two populations are the same, both a location test and a scale test could be additionally performed in a closed testing procedure with level α (i.e. without adjustment: Neuhäuser & Hothorn, 2000). That is, in situations where the null hypothesis is rejected, researchers can often gain insight on the relative importance of difference in means and variances in driving the rejection of the null hypothesis. At the second stage of this procedure the Welch-Satterthwaite *t*-test might be carried out to illuminate the difference in central tendency.

We should also sound a note of caution with regard to our recommendation to sometimes apply ranking prior to applying Satterthwaite's test. There are dangers associated with ranking prior to application of an essentially parametric method (see Sawilowsky, 2000). Although the rank transformation looks like a convenient bridge between parametric and nonparametric methods, it is in general not valid in the Behrens-Fisher problem (Brunner & Munzel, 2013). Hence, Satterthwaite's test on ranked data has a heuristic justification only (Delaney & Vargha, 2002): its appropriateness and robustness are based on empirical studies only. Indeed, the rank Welch test can become liberal according to the simulation results presented by Delaney and Vargha (2002). It cannot be recommended for discrete distributions when sample sizes are small or moderate, hence our caution that our simple guidance offered at the start of this piece only applies to continuous data. For discrete data, when there are many ties or sample sizes are small or moderate, nonparametric methods such as those investigated by Delaney and Vargha should be preferred.

References

Blair, R. C., & Sawilowsky, S. (1993a). A note on the operating characteristics of the modified *F* test. *Biometrics, 49*(3), 935-939. doi: 10.2307/2532215

Blair, R. C., & Sawilowsky, S. (1993b). Comparison of two tests useful in situations where treatment is expected to increase variability relative to controls. *Statistics in Medicine, 12*(23), 2233-2243. doi: 10.1002/sim.4780122308

Boik, R. J. (1987). The Fisher-Pitman permutation test: A non-robust alternative to the normal theory *F* test when variances are heterogeneous. *British Journal of Mathematical and Statistical Psychology, 40*(1), 26-42. doi: 10.1111/j.2044-8317.1987.tb00865.x

Box, G. E. P. (1953). Non-normality and tests on variances. *Biometrika, 40*(3-4), 318-335. doi: 10.1093/biomet/40.3-4.318

Brunner, E., & Munzel, U. (2000). The nonparametric Behrens-Fisher problem: Asymptotic theory and a small sample approximation. *Biometrical Journal, 42*(1), 17-25. doi: 10.1002/(SICI)1521-4036(200001)42:1<17::AID-BIMJ17>3.0.CO;2-U

Brunner, E., & Munzel, U. (2013). *Nichtparametrische datenanalyse* (2nd ed.). Berlin, Germany: Springer. doi: 10.1007/978-3-642-37184-4

Coombs, W. T., Algina, J., & Oltman, D. O. (1996). Univariate and multivariate omnibus hypothesis tests selected to control Type I error rates when population variances are not necessarily equal. *Review of Educational Research, 66*(2), 137-179. doi: 10.3102/00346543066002137

Cribbie, R. A., Wilcox, R. R., Bewell, C., & Keselman, H. J. (2007). Tests for treatment group equality when data are nonnormal and heteroscedastic. *Journal of Modern Applied Statistical Methods, 6*(1), 117-132. doi: 10.22237/jmasm/1177992660

Dalgaard, P. (2002). *Introductory statistics with R*. New York, NY: Springer.

Delaney, H. D., & Vargha, A. (2002). Comparing several robust tests of stochastic equality with ordinally scaled variables and small to moderate sized samples. *Psychological Methods, 7*(4), 485-503. doi: 10.1037/1082-989x.7.4.485

Gans, D. J. (1991). Letters to the Editor: Preliminary test on variances. *The American Statistician, 45*(3), 258.

Hothorn, L. A., & Hauschke, D. (1998). Principles in statistical testing in randomized toxicological studies. In S. C. Chow & J. P. Liu (Eds.), *Design and analysis of animal studies in pharmaceutical development* (pp. 79-133). New York, NY: Marcel Dekker.

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. doi: 10.1111/j.0963-7214.2004.01501008.x

Markowski, C. A., & Markowski, E. P. (1990). Conditions for the effectiveness of a preliminary test of variance. *The American Statistician, 44*(4), 322-326. doi: 10.1080/00031305.1990.10475752

Moser, B. K., & Stevens, G. R. (1992). Homogeneity of variance in the two-sample means test. *The American Statistician, 46*(1), 19-21. doi: 10.1080/00031305.1992.10475839

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*(11), 3963-3975. doi: 10.1080/03610928908830135

Neuhäuser, M. (2002). Two-sample tests when variances are unequal. *Animal Behaviour, 63*(4), 823-825. doi: 10.1006/anbe.2002.1993

Neuhäuser, M. (2015). Combining the *t* test and Wilcoxon's rank sum test. *Journal of Applied Statistics, 42*(12), 2769-2775. doi: 10.1080/02664763.2015.1070809

Neuhäuser, M., & Hothorn, L. A. (2000). Parametric location-scale and scale trend tests based on Levene's transformation. *Computational Statistics & Data Analysis, 33*(2), 189-200. doi: 10.1016/s0167-9473(99)00051-1

Neuhäuser, M., & Manly, B. F. J. (2004). The Fisher-Pitman permutation test when testing for differences in mean and variance. *Psychological Reports, 94*(1), 189-194. doi: doi: 10.2466/pr0.94.1.189-194

Neuhäuser, M., & Ruxton, G. D. (2009). Distribution-free two-sample comparisons in the case of heterogeneous variances. *Behavioral Ecology and Sociobiology, 63*(4), 617-623. doi: 10.1007/s00265-008-0683-4

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

Perry, K. T. (2003). A critical examination of the use of preliminary tests in two-sample tests of location. *Journal of Modern Applied Statistical Methods, 2*(2), 314-328. doi: 10.22237/jmasm/1067645100

Quinn, G. P., & Keough, M. J. (2002). *Experimental design and data analysis for biologists*. Cambridge, UK: Cambridge University Press. doi: 10.1017/cbo9780511806384

Rasch, D., Kubinger, K. D., & Moder, K. (2011). The two-sample *t* test: Pre-testing its assumptions does not pay off. *Statistical Papers, 52*(1), 219-231. doi: 10.1007/s00362-009-0224-x

Ruxton, G. D., Wilkinson, D. M., & Neuhäuser, M. (2015). Advice on testing the null hypothesis that a sample is drawn from a normal distribution. *Animal Behaviour, 107*, 249-252. doi: 10.1016/j.anbehav.2015.07.006

Sawilowsky, S. S. (2000). Review of the rank transform in designed experiments. *Perceptual and Motor Skills, 90*(2), 489-497. doi: 10.2466/pms.90.2.489-497

Sawilowsky, S. S. (2002). Fermat, Schubert, Einstein, and Behrens-Fisher: The probable difference between two means when $\sigma_1^2 \neq \sigma_2^2$. *Journal of Modern Applied Statistical Methods, 1*(2), 461-472. doi: 10.22237/jmasm/1036109940

Weber, M., & Sawilowsky, S. S. (2009). Comparative power of the independent t, permutation t, and Wilcoxon tests. *Journal of Modern Applied Statistical Methods, 8*(1), 10-15. doi: 10.22237/jmasm/1241136120

Zimmerman, D. W. (1998). Invalidation of parametric and nonparametric statistical tests by concurrent violation of two assumptions. *The Journal of Experimental Education, 67*(1), 55-68. doi: 10.1080/00220979809598344

Zimmerman, D. W., & Zumbo, B. D. (1993). Rank transformations and the power of the Student *t* test and Welch *t′* test for non-normal populations with unequal variances. *Canadian Journal of Experimental Psychology, 47*(3), 523-539. doi: 10.1037/h0078850