Preliminary Testing: The Devil of Statistics?

Jack Pearce

Jack is graduated from the University of the West of England, Bristol, in July 2018 with a first class BSc(Hons) Mathematics, winning the Institute of Mathematics and its Applications (IMA) prize for outstanding achievement.

Ben Derrick

Ben is a PhD student and lecturer at the University of the West of England, Bristol.

Abstract

In quantitative research, the selection of the most appropriate statistical test for the comparison of two-independent samples can be problematic. There is a lack of consensus in the Statistics community regarding the appropriate approach; particularly towards assessing assumptions of normality and equal variances. The lack of clarity in the appropriate strategy affects the reproducibility of results. Statistical packages performing different tests under the same name, only adds to this issue.

The process of preliminary testing assumptions of a test using the sample data, before performing a test conditional upon the preliminary test, is performed by some researchers; this practice is often criticised in the literature. Preliminary testing is typically performed at the arbitrary 5% significance level. In this paper this process is reviewed, and additional results are given using simulation, examining a procedure with normality and equal variance preliminary tests.

Key Words

Statistics, robustness, t-test, preliminary testing, conditional tests, independent samples

Introduction

In statistical hypothesis testing, the literature rarely reaches an agreement on the most appropriate analysis strategy for any given scenario. To illustrate the problems faced, this paper will focus on comparing the central location of two independent samples. For example, some researchers may use an independent samples t-test with pooled variances (Independent t-test), some may use a form of the independent samples t-test not constrained to equal variances (Welch's test), some may use the Mann-Whitney or the Yuen-Welch test due to concerns over normality.

Each of these two-sample tests have accompanying assumptions. The Independent t-test assumes both normality and equal variances. Welch's test assumes normality, but not equal variances. The Mann-Whitney test assumes equal variances, but not normality. Yuen-Welch's test has no assumptions regarding normality or equal variances.

Assessment of the assumptions to determine the appropriate two sample test can occur at the design stage, or after the data has been collected in the form of preliminary tests of the assumptions. A researcher could have a plan to perform one of the above tests based on pre-existing knowledge of the assumptions, or they might plan to perform preliminary tests on the assumptions to determine the correct test, or they may have no plan at all.

There is no consensus as to the correct method of preliminary analyses, which results in researchers choosing tests in ad hoc ways, even selecting methods of analysis after the data has been compiled that provides the desired conclusion. This has contributed to the reproducibility crisis in the sciences.

The Independent t-test is taught as the 'standard' two-sample test. Undergraduate students are taught how to run the test, but not necessarily the definitive set of conditions when it might be appropriate, or the knowledge to evaluate the appropriateness of the test. Along with many practical users, undergraduate students will follow a set of arbitrary instructions based on an arbitrary decision tree provided by their lecturer, or other resource. Many decision trees can be found on the internet outlining a two-sample test procedure (Martz, 2017), but rarely in academic papers. One example of a two-sample test decision tree in an academic paper is Marusteri and Bacarea (2010), which involves both normality and equal variance preliminary tests.

Before an informed decision can be made as to whether the Independent t-test is the most appropriate two-sample test, questions regarding the assumptions of the Independent t-test must be answered, namely checking if the data are normally distributed and the group

variances are equal. Preliminary tests can be used to answer these questions. However, there are many different tests that could be performed to check the assumptions. To check the normality assumption, the Shapiro-Wilk test or the Kolmogorov-Smirnov test could be performed, among others. The tests for equality of variances assumptions could be Levene's test using deviations from the group means or Levene's test using deviations from group medians, among others.

Another issue with regards to reproducibility is the fact that different software run different tests under the same name. For assessing equality of variances, SPSS runs Levene's test using deviations from means, whereas Minitab runs Levene's test using deviations from medians. This affects reproducibility, because both SPSS and Minitab are widely used statistical packages in quantitative research. Researchers may run what they believe is the same Levene's test for equal variances, but receive conflicting conclusions, affecting their chosen conditional two-sample test and thus the final conclusions.

For example, data has been collected from an exam consisting of 20 multiple choice questions, taken by two different tutorial groups, i.e. there are two independent samples. The scores awarded by the participants of the exam can be found in Table 1.

Group 1	9	12	12	12	12	12	13	13	13	14	14	14
Group 2	9	10	11	14	15	15	15	16	16	17	18	19

Table 1: Number of correctly answered multiple choice questions out of 20.

The decision rule applied by both SPSS and Minitab is; if the null hypothesis of equal variances is failed to be rejected, the Independent t-test is performed, and conversely Welch's test is performed when variances are found to be unequal. If one researcher uses SPSS and the other uses Minitab, the following would occur, as per Table 2.

Test for equal variances

| Levene's test using means (SPSS)	Levene's test using medians (Minitab)
$p = .030$	$p = .071$
Reject null hypothesis of equal variances.	Fail to reject null hypothesis of equal variances.

Two-sample test depending on preliminary test

Welch's test (SPSS)	Independent t-test (Minitab)
$p = .051$	$p = .046$
Fail to reject the null hypothesis that the two samples means do not differ.	Reject the null hypothesis that the two samples means do not differ.

Table 2: Two-sample test procedure with test for equal variances preliminary test on multiple-choice scores, where normality is assumed.

As seen in Table 2, testing at the 5% significance level, performing the procedure on SPSS with Levene's preliminary test (using means), the researcher would reject the assumption of equal variances ($p = .030$); the conditional test is therefore Welch's test which finds no significant difference in the mean scores ($p = .051$). However, performing the procedure on Minitab with Levene's preliminary test (using medians), the researcher would fail to reject the assumption of equal variances ($p = .071$); then run the Independent t-test and find a significant difference in the mean exam scores ($p = .046$).

Therefore, two researchers with the same data arrive at different conclusions, simply due to the software used. Hence, even if there was a consensus as to the correct preliminary test procedure to run, researchers can have a hard time producing the same results. It is apparent that a lack of a plan and user apathy as to which statistical tests are being performed is dangerous. Moreover, a researcher could reverse engineer the software used and statistical test performed in order to achieve their desired conclusion.

A two-sample test procedure is often presented in the form of a decision tree. Figure 1 shows a two-step test procedure when comparing two independent samples. The test procedure includes both equal variance and normality preliminary tests.
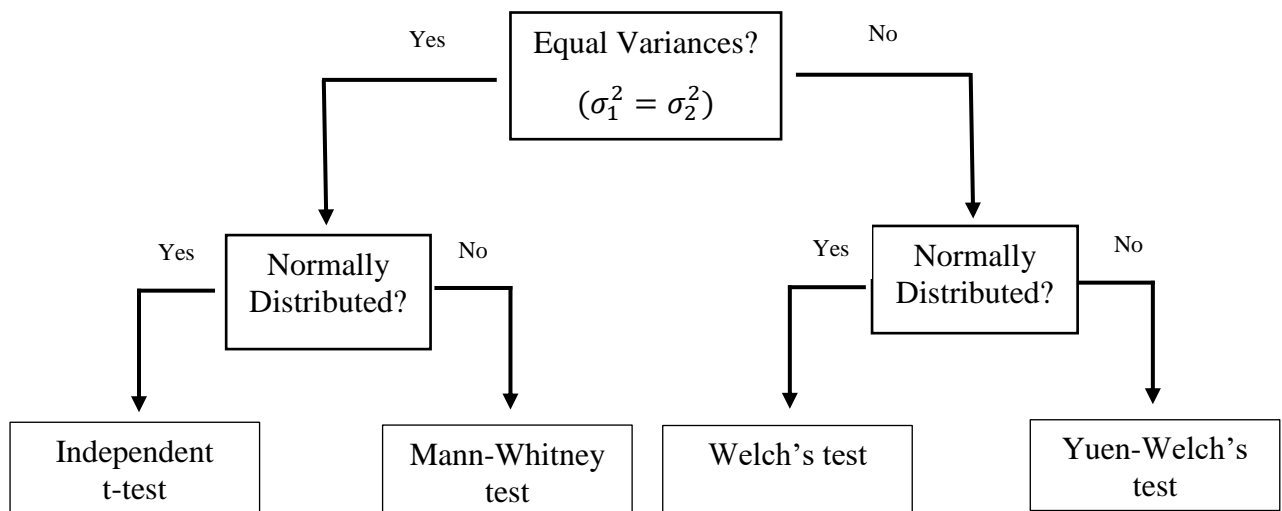
Figure 1: Two-Step test procedure, with both equal variance and normality preliminary tests.

Notice in Figure 1 that the Independent t-test is the default test, because without evidence to reject the assumption of normality or equal variances, the Independent t-test is performed.

Hoekstra, Kiers and Johnson (2012) studied whether 30 Ph.D. students checked fictitious data for violations of the assumptions of the statistical tests they used. Hoekstra *et al.* found that the assumptions were rarely checked; in fact, the assumptions of normality and equal variances were formally checked only in 12% and 23% of cases respectively. When the Ph.D. students were asked the reason behind them not checking the assumptions, for the assumption of normality, approximately 90% of them said it was because they were unfamiliar with the assumption; similarly, approximately 60% of the Ph.D. students gave the same reason for not checking the assumption of equal variances.

Wells and Hintze (2007) suggested that the assumptions should be considered in the planning of the study, as opposed to being treated almost as an afterthought. Considering assumptions at the planning stage by: testing using prior data from the same/similar source; using theoretical knowledge or reasoning; addressing the assumptions before the data are collected, which can avoid the issues surrounding preliminary testing. Wells and Hintze finished by suggesting that studies should be designed, and statistical analyses selected that are robust to assumption violations, i.e. equal sized groups or large sample sizes, whenever possible. Equal sized groups are desirable due to most two sample tests that assume equal variances being robust against violations when there are equal sized groups, for example the Independent t-test (Nguyen *et al.*, 2012; Derrick, Toher and White, 2016).

Zumbo and Coulombe (1997) warned of at least two scenarios where equal variances cannot be assumed: when the groups of experimental units are assembled based on important differences such as age groups, gender, or education level; or the experimental units differ by an important, maybe unmeasured variable. Thus, ideally it is the design of the experiment that should determine whether this assumption is true, not the samples collected.

At the 5% significance level, a valid test procedure should reject the null hypothesis approximately 5% of the time; this would represent Type I error robustness. Rochon, Gondan and Kieser (2012) investigated the Type I error robustness of the Independent t-test and the Mann-Whitney test. Interestingly, the unconditional test (i.e. no preliminary test) controlled Type I error rates for both two-sample tests, under normality, and exponentially distributed data. There may be little need for preliminary tests, if the conditional tests are robust to minor deviations from the assumptions.

Garcia-Perez (2012) and Rasch, Kubinger and Moder (2011) highlighted the ramifications of checking assumptions using the same data that is to be analysed; if the researchers do not test the assumptions, they could suffer uncontrolled Type I error rates; or they can test the assumptions but will surrender control of the Type I error rates too. 'It thus seems that a researcher must make a choice between two evils' (Garcia-Perez, 2012, 21). Any preliminary assessment of assumptions can affect the Type I error rate of the final conditional test of interest; Ruxton (2006) and Zimmerman (2004) advise against preliminary testing.

Many textbooks recommend checking the assumptions of normality and equal variances graphically, e.g. Moore, Notz and Fligner (2018). However, Garcia-Perez (2012) emphasized that the problem of distorted Type I errors still persists because the decision on what technique to use is conditioned on the results of this graphical preliminary analysis, just like a formal hypothesis test. A graphical approach also introduces a further element of researcher subjectivity.

When preliminary testing for normality was performed, Rochon *et al.* (2012) show that the conditional Mann-Whitney test had elevated Type I error rates for the normally distributed data. Similarly, when preliminary testing, the conditional Independent t-test had large Type I error rates for the exponential distribution and uniform distribution; likely due to the lack of times it is performed, where tests for normality are performed on non-normal data. Rochon *et al.* concluded that for small samples, the Shapiro-Wilk test for normality lacks power to detect deviations from normality. However, this may be a good thing for a preliminary test due to the Independent t-test's robustness against violations of normality and its high power; in fact, the Kolmogorov-Smirnov test has less power than the Shapiro-Wilk test (Razali and

Wah, 2011) and is often preferred. Rochon *et al.* also suggested if the application of the Independent t-test is advised against due to potential concerns over normality, then the unconditional use of the Mann Whitney test is the most appropriate choice.

Other ad-hoc methods for selecting a test to compare the central location of two samples include looking at sample size or skewness. Fagerland (2012) recommend the Mann-Whitney test for small sample sizes. Rasch, Teuscher and Guiard (2007) suggest to always perform Welch's test when there are unequal sample sizes. Penfield (1994) recommend the Mann-Whitney test when the samples are highly skewed (i.e. non-normal). However, Fagerland and Sandvik (2009) found there was no clear best test across different combinations of variance and skewness. Another ad-hoc method for selecting the most appropriate test is to assess for outliers, and perform the Mann-Whitney test or the Yuen-Welch test if an outlier is identified (Derrick *et al.*, 2017)

A further complication is the choice of the 5% significance level mostly used in all preliminary tests. The 5% significance level is an arbitrary level suggested by statisticians, so it is not necessarily the optimal significance level for every application. Standard thinking regarding statistical inference at the 5% significance level is to be challenged (Wasserstein and Lazar, 2016).

In this paper, a simulation study investigates the Type I error robustness of the two sample test procedure outlined in Figure 1. The procedure is investigated for two commonly used normality tests and two commonly used tests for equal variances, each performed at varying significance levels.

Simulation Methodology

Serlin (2000) explained that in testing robustness, running simulations is the standard and appropriate approach. The simulation approach, where numerous iterations are run, generates the long-run probability of a Type I error; because for each individual test there is either a Type I error or not; performing this process numerous times allows us to calculate the Type I error rate.

In a two independent samples design, each of the Independent t-test (IT), Welch's test (W), the Mann-Whitney test (MW), and the Yuen-Welch test (YW) are performed.

The normality preliminary tests considered are the Shapiro Wilk test (SW) and the Kolmogorov-Smirnov test (KS). The tests for equality of variances considered are Levene's test using means (LMean) and Levene's test using medians (LMed). Each preliminary test is performed on each conditional test. The preliminary tests are performed at the 1% and 5% significance levels. The conditional test is selected based on the results of each of the preliminary tests and performed at the 5% significance level.

To account for both normally distributed data and skewed distributions, the distributions considered are Normal, Exponential and Lognormal. The Normal distribution is considered for both groups sampled from distributions with mean of zero. Firstly where both groups have variances equal to one. Secondly groups sampled from the Normal distributions with unequal variances {1, 2} are considered. Exponential deviates are generated with a mean and variance of one. Lognormal deviates are generated with a mean of 0 and variance of 1. Thus in effect four separate sets of simulations are performed.

For each set of simulations, sample sizes for each of the two groups are generated in a factorial design {5,10,20,30}, i.e. 16 sample size combinations. 10,000 iterations are performed for each combination. Emphasis is on small sample sizes, reflecting practical application.

To calculate the Type I error rates of the conditional test procedure in Figure 1, for each combination of sample size the weighted averages of the Type I error rates for the two-sample tests performed are taken; this provides one overall value to represent the test procedure's performance. The weighting for the Type I error rates is how often the test is performed; the two-sample test performed most often is likely the most appropriate test (i.e. its assumptions match the characteristics of the two samples distributions) and should have the largest influence on the Type I error rate. Simply taking averages is not fair because one test may only be performed a small percentage of the time; similarly, reporting each

conditional test Type I error rate separately is not fair because it does not consider how often the test is performed.

The Type I error rates in this study are ideally 5% because two-sample tests are designed so that their Type I error rate should match that of the significance level being tested at. These will be scrutinised in conjunction with Bradley's liberal criterion (Bradley, 1978), which says that a robust or stable Type I error rate is between 2.5% and 7.5% when testing at the 5% significance level. To determine what two-sample test or test procedure has the most robust Type I errors across the four distributions, it is proposed that the average absolute deviation from 5% across the four distributions is examined.

Results

Before the conditional test procedure which uses preliminary tests is assessed, first the unconditional performance of the four tests across the four distributions is considered. The unconditional performance refers to the different two-sample tests Type I error rates when performed regardless of whether the assumptions are met or not, no preliminary tests are performed. In Table 3 the 'Overall Type I Error Rate' refers to the average of the Type I error rates for the combinations of sample size, and variances (when using the Normal distribution) for the two samples.

Distribution	Overall Type I Error Rate			
	IT	MW	W	YW
Normal (Equal Variances)	5.01%	4.54%	5.16%	5.99%
Normal (Unequal Variances)	6.24%	5.05%	5.09%	6.00%
Exponential	4.54%	4.58%	6.04%	5.82%
Lognormal	4.17%	4.43%	5.61%	5.18%
Average absolute deviations from 5%	0.0063	0.0038	0.0047	0.0075

Table 3: Two-Sample tests unconditional average Type I error rates.

Table 3 shows that simply disregarding all assumptions and performing the Independent t-test unconditionally may not be the most robust approach. The Mann-Whitney test has the most robust Type I error rates across the four distributions since the average of the absolute deviations from 5% across the four distributions is the smallest. When looking at the test with the most Type I error control for each of the four distributions, the Independent t-test is most robust under the Normal distribution and equal variances; the Mann-Whitney test is most robust under the Normal distribution with unequal variances and the Exponential distribution; Yuen-Welch's test is the most robust under the Lognormal distribution. It is worth noting all these Type I error rates are within Bradley's liberal criterion, so they all control the Type I error.

Figure 2 offers a more in-depth look at the Type I error rates of the two-sample tests across the four distributions, compared to the average Type I error rates displayed in Table 3. 'Normal EV' and 'Normal UEV', refer to the Normal distribution under equal and unequal variances respectively. The dotted horizontal lines represent Bradley's liberal criterion boundaries.
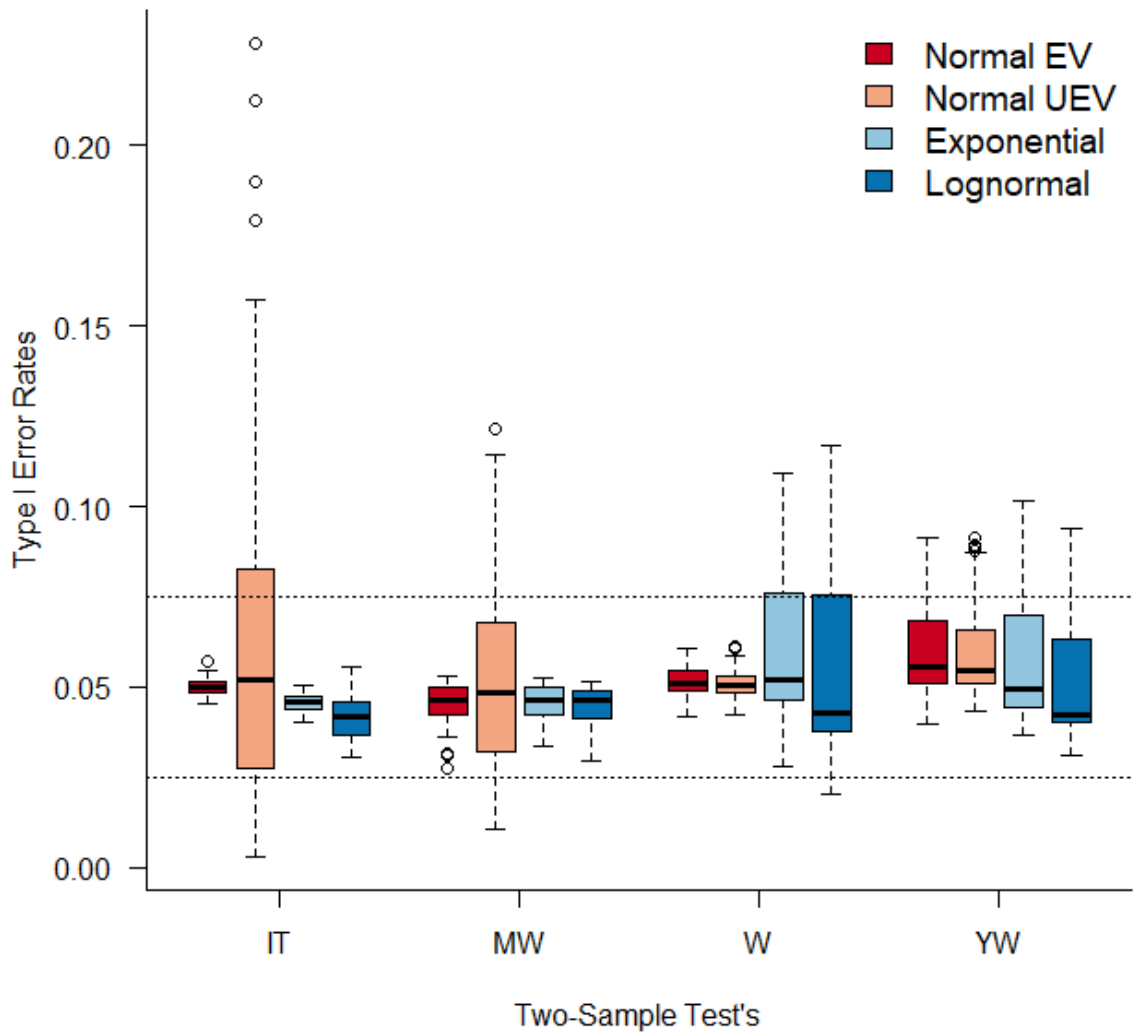
Figure 2: Two-Sample tests unconditional Type I error rates.

None of the two-sample tests considered, control the Type I error rates for all combinations of sample size, across the four distributions. The largest violations occur when there are large disparities in sample size. Therefore, performing any of the four two-sample tests unconditionally will provide Type I error rates outside of Bradley's liberal criterion for specific combinations of sample size and variances. Thus, a preliminary testing procedure may be required.

The test procedure with both normality and equal variance preliminary testing as per Figure 1 is considered. This two-stage preliminary test procedure provides 16 combinations of preliminary tests considered in a factorial design, i.e. two normality tests (SW and KS), two equal variances tests (LMean and LMed), two significance levels for the normality tests (1% and 5%), and two significance levels for the equal variances tests (1% and 5%).

The average absolute deviation from 5% across the four distributions is given in Table 4. The best preliminary test combinations for each distribution are displayed.

| | Overall Type I Error Rate | | | |
| | LMean 5% | LMean 1% | LMed 1% | LMed 1% |
Distribution	KS 5%	KS 1%	SW 1%	SW 5%
Normal (Equal Variances)	5.11%	5.00%	5.07%	5.14%
Normal (Unequal Variances)	5.73%	5.81%	6.15%	6.24%
Exponential	5.10%	4.20%	5.11%	4.97%
Lognormal	4.34%	3.74%	4.68%	4.61%
Average absolute deviations from 5%	0.0040	0.0072	0.0041	0.0045

Table 4: Two-Sample tests procedures average Type I error rates.

For the two non-normal distributions, the Shapiro-Wilk normality test is preferred. When drawing from the non-normal distributions, normality needs to be rejected as often as possible to provide better Type I error rate control, therefore the Shapiro-Wilk test which does this more often, is the better normality test in this case. The average weighted Type I error rates are comfortably within Bradley's liberal criterion, with the Type I errors from the Normal distribution with unequal variances being the worst.

Table 4 shows that the two-step test procedure with Kolmogorov-Smirnov and Levene's (Mean) preliminary tests, both at the 5% significance level, achieves the most Type I error rate control. However, there is negligible difference between each of the preliminary test combinations to be of real practical consequence.

Conclusion

This paper examines some of the standard statistical tests for comparing two samples. Results show that the Independent t-test's Type I errors were less robust than the Mann-Whitney's and Welch's, but still within Bradleys liberal robustness criterion; therefore, it is not necessarily a bad choice for the default two-sample test, just not necessarily the best. Wells and Hintze (2007) and Rasch *et al.* (2011) also question why the Independent t-test is considered the default two-sample test and suggested using Welch's test as the default. These results further advocate a theory that the approach be revised so that Welch's test is the default.

In this paper procedures with preliminary hypotheses tests are examined to replicate the conditions many users face when comparing two independent samples. The weighted average Type I error rates for each combination of preliminary tests was considered. Taking averages with Type I error rates does have its limitations, since robust Type I error rates are defined in a range; the limitations of this is that it is possible to have equally non-robust Type I error rates either side of 5%, that when averaged provide a robust Type I error rate, which is not the case. However, it is more likely the test procedure has either consistently liberal or conservative Type I errors, due to the changes in sample size and variances considered being relatively small, making switches from liberal to conservative Type I errors less likely. The implication of this is that when averaged, the weighted Type I error rate will identify either a liberal or conservative Type I error rate, if the set of Type I error rates are truly liberal or conservative, instead of showing robust Type I errors when the set of Type I errors is not.

When comparing the two-sample tests performed unconditionally to the conditional testing procedure, the weighted Type I errors across the four distributions for the recommended conditional test procedures were comparable and more robust in most cases. This implies that despite the test procedures introducing compounded errors caused by the preliminary tests, the weighted Type I error rates were better for it, because the most appropriate test was performed more often.

For the scenarios considered, the benefits of implementing a test procedure to find the most appropriate two-sample test may outweigh that of performing a two-sample test unconditionally, in terms of controlled Type I error rates across the four distributions. However, it is advised if possible to follow Wells and Hintze (2007) advice of: determining whether the sample size is large enough to invoke the Central Limit Theorem; considering the assumptions in the planning of the study; testing assumptions if necessary from a similar previous data source.

14

The preliminary testing procedure that most closely maintains the Type I error rate is preforming Kolmogorov-Smirnov normality test and Levene's (Mean) test for equal variances, both at the 5% significance level. The test procedure performs well, with robust Type I errors when data are from either the Normal distribution or the skewed distributions considered. However, the use of a flow diagram and this rule to select the 'appropriate' test can encourage inertia and restrict critical thinking from the user about the test being performed.

Given the capacity for different researchers to conduct potentially conflicting analyses, solutions which offer the most transparency and forward planning are recommended. This is leading to some disciplines requesting that analysis plans are pre-registered, examples include the Journal of Development Economics and the Center for Open Science. This would seem like an appropriate way forward.

References

Bradley, J.V. (1978), 'Robustness?', *British Journal of Mathematical and Statistical Psychology*. 31 (2), 144-152.

Derrick, B., Toher, D. and White, P. (2016), 'Why Welch's test is Type I error robust', *The Quantitative Methods in Psychology*. 12 (1), 30-38.

Derrick, B., Broad, A., Toher, D. and White, P. (2017), 'The impact of an extreme observation in a paired samples design', *Metodološki Zvezki - Advances in Methodology and Statistics*. 14 (2), 1-17.

Fagerland, M.W. (2012),'t-tests, non-parametric tests, and large studies—a paradox of statistical practice?', *BMC Medical Research Methodology* 12 (1), 78.

Fagerland, M.W. and Sandvik, L. (2009),'Performance of five two-sample location tests for skewed distributions with unequal variances', *Contemporary Clinical Trials*. 30 (5), 490-496.

Garcia-Perez, M.A. (2012),'Statistical conclusion validity: some common threats and simple remedies', *Frontiers in Psychology*. 3 325.

Hoekstra, R., Kiers, H.A. and Johnson, A. (2012), 'Are assumptions of well-known statistical techniques checked, and why (not)?', *Frontiers in Psychology*, 3, 137.

Martz, E. (2017), '*Three Common P-Value Mistakes You'll Never have to Make'*. Available from: http://blog.minitab.com/blog/understanding-statistics/three-common-p-value-mistakes-youll-never-have-to-make [Accessed 03 April 2018].

Marusteri, M. and Bacarea, V. (2010), 'Comparing groups for statistical differences: how to choose the right statistical test?', *Biochemia Medica*. 20 (1), 15-32.

Moore, D. S., Notz, W., and Fligner, M. A. (2018). *The basic practice of statistics*. WH Freeman.

Nguyen, D., Rodriguez de Gil, P., Kim, E., Bellara, A., Kellermann, A., Chen, Y. and Kromrey, J. (2012), 'PROC TTest (Old Friend), What are you trying to tell us', *Proceedings of the South East SAS Group Users*.

Penfield, D.A. (1994), 'Choosing a two-sample location test', *The Journal of Experimental Education*. 62 (4), 343-360.

Rasch, D., Kubinger, K.D. and Moder, K. (2011), 'The two-sample t test: pre-testing its assumptions does not pay off', *Statistical Papers*. 52 (1), 219-231.

Rasch, D., Teuscher, F. and Guiard, V. (2007), 'How robust are tests for two independent samples?', *Journal of Statistical Planning and Inference*. 137 (8), 2706-2720.

Razali, N.M. and Wah, Y.B. (2011), 'Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests', *Journal of Statistical Modeling and Analytics*. 2 (1), 21-33.

Rochon, J., Gondan, M. and Kieser, M. (2012), 'To test or not to test: Preliminary assessment of normality when comparing two independent samples', *BMC Medical Research Methodology*. 12 (1), 81.

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.

Serlin, R.C. (2000), 'Testing for robustness in Monte Carlo studies', *Psychological Methods*. 5 (2), 230.

Wasserstein, R.L. and Lazar, N.A., 2016. 'The ASA's statement on p-values: context, process, and purpose', The American Statistician, 70(2), 129-133.

Wells, C.S. and Hintze, J.M. (2007), 'Dealing with assumptions underlying statistical tests. *Psychology in the Schools',* 44 (5), 495-502.

Zimmerman, D.W. (2004), 'A note on preliminary tests of equality of variances', *British Journal of Mathematical and Statistical Psychology*. 57 (1), 173-181.

Zumbo, B.D. and Coulombe, D. (1997), 'Investigation of the robust rank-order test for non-normal populations with unequal variances: The case of reaction time'. *Canadian Journal of Experimental Psychology*. 51 (2), 139.