Statistical inference with paired observations and independent observations in two samples

Benjamin Fletcher Derrick

A thesis submitted in partial fulfilment of the requirements of the University of the West of England, Bristol, for the degree of Doctor of Philosophy

Applied Statistics Group, Engineering Design and Mathematics, Faculty of Environment and Technology, University of the West of England, Bristol

February, 2020

Abstract

A frequently asked question in quantitative research is how to compare two samples that include some combination of paired observations and unpaired observations. This scenario is referred to as ‘partially overlapping samples’. Most frequently the desired comparison is that of central location. Depending on the context, the research question could be a comparison of means, distributions, proportions or variances. Approaches that discard either the paired observations or the independent observations are customary. Existing approaches evoke much criticism. Approaches that make use of all of the available data are becoming more prominent. Traditional and modern approaches for the analyses for each of these research questions are reviewed. Novel solutions for each of the research questions are developed and explored using simulation. Results show that proposed tests which report a direct measurable difference between two groups provide the best solutions. These solutions advance traditional methods in this area that have remained largely unchanged for over 80 years. An R package is detailed to assist users to perform these new tests in the presence of partially overlapping samples.

Acknowledgments

I pay tribute to my colleagues in the Mathematics Cluster and the Applied Statistics Group at the University of the West of England, Bristol. Some intriguing and friendly characters have assisted and entertained during the process.

I appreciate all of the hard working and enthusiastic students at the University of the West of England, Bristol. Particularly my outstanding final year project students, Jack Pearce and Annalise Ruck, along with additional publication co-authors; Anselma Dobson-McKittrick, Bethan Russ and Antonia Broad.

Acknowledgement is given to the journal editors and peer reviewers for their valuable contributions. This includes Denis Cousineau, Claudiu Herteliu, Shlomo Sawilowsky, and numerous reviewers who remain anonymous.

I give considerable thanks to the guidance of my supervisory team, Paul White and Deirdre Toher, for their encouragement and help on all matters PhD, and beyond. They have been a pleasure to work with and it is to their credit my achievement is maximised.

I am grateful to my friends and family for showing an interest, whether they wanted to or not, and assisting in the proof reading.

I am eternally grateful to Mum and Dad for being so supportive and proud. To Dad, for devoting your life to your family, and always believing in me.

Contents

List of Tables

List of Figures

Commonly used parameters

Table 1: Commonly used notation, simulation parameters.

Notation	**Description**
n_a	Number of observations exclusive to Sample 1
n_b	Number of observations exclusive to Sample 2
n_c	Number of pairs
n_1	$n_a + n_c$
n_2	$n_b + n_c$
$\bar{X}_1$	Mean of all observations in Sample 1
$\bar{X}_2$	Mean of all observations in Sample 2
$\bar{X}_a$	Mean of n_a observations
$\bar{X}_b$	Mean of n_b observations
$\bar{X}_{1c}$	Mean of n_c paired observations in Sample 1
$\bar{X}_{2c}$	Mean of n_c paired observations in Sample 2
S_1^2	Variance of all observations in Sample 1
S_2^2	Variance of all observations in Sample 2
S_a^2	Variance of n_a observations
S_b^2	Variance of n_b observations
S_{1c}^2	Variance of n_c paired observations in Sample 1
S_{2c}^2	Variance of n_c paired observations in Sample 2
r	Pearson's sample correlation coefficient
μ_i	Population mean where $i = (1, 2)$
σ_i	Population variance where $i = (1, 2)$
ρ	Pearson's population correlation coefficient

Notation consistent with recommended standards for
statistical symbols and notation (Halperin, Hartley, and Hoel, 1965)

Commonly used test statistics

Table 2: Commonly used notation, test statistics.

Notation	**Description**
T_1	Paired samples t-test
v_1	Degrees of freedom for T_1
T_2	Independent samples t-test, equal variances assumed
v_2	Degrees of freedom for T_2
T_3	Independent samples t-test, equal variances not assumed (Welch's test)
v_3	Degrees of freedom for T_3
W_1	Wilcoxon rank sum test (Wilcoxon test)
W_2	Pratt's test
MW	Mann-Whitney-Wilcoxon test (Mann-Whitney test)
$Z_{\text{corrected}}$	Looney and Jones (2003) method
T_{new1}	Partially overlapping samples t-test, equal variances assumed
v_{new1}	Degrees of freedom for T_{new1}
T_{new2}	Partially overlapping samples t-test, equal variances not assumed
v_{new2}	Degrees of freedom for T_{new2}
T_{RNK1}	Partially overlapping samples t-test on ranks, equal variances assumed
T_{RNK2}	Partially overlapping samples t-test on ranks, equal variances not assumed
T_{INT1}	Partially overlapping samples t-test, equal variances assumed, following inverse Normal transformation
T_{INT2}	Partially overlapping samples t-test, equal variances not assumed, following inverse Normal transformation
z_8	Partially overlapping samples z-test, for dichotomous data

Chapter 1

Introduction

Partially overlapping samples are defined as samples that contain both paired observations and independent observations. Research into the framework surrounding two partially overlapping samples is motivated in this chapter. Extant solutions to the partially overlapping samples problem are reviewed. Aims and objectives are articulated regarding the provision of solutions to the two partially overlapping samples framework.

1.1 Motivation

A question that is often asked in research is how to compare the means of two samples that include some combination of paired observations and unpaired observations. These scenarios are referred to as 'partially overlapping samples' (Martinez-Camblor, Corral, and De La Hera, 2013, p.78). Various other terminology used in the literature to refer to this scenario include; 'correlated variates with missing observations' (Bhoj, 1978), 'combined samples of correlated and uncorrelated data' (Looney and Jones, 2003), 'partially paired data' (Samawi and Vogel, 2011), 'partially observed data' (Ramosaj, Amro, and Pauly, 2018), a 'mixed design' (Mantilla and Terpstra, 2018), and 'partly depending data' (Stigler et al., 2018).

Consider Figure 1.1 which depicts scenarios where there are two samples, each with a different number of paired observations and independent observations.

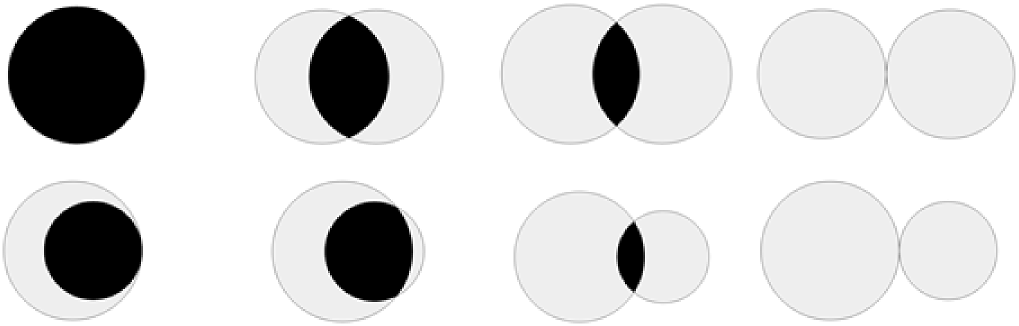


Figure 1.1: Examples of partially overlapping samples. In each scenario each of two samples are represented by a circle. The paired observations are represented by the overlap and shaded black. From left to right the graphic shows a decreasing number of paired observations. The relative sample sizes are represented by the size of the circle.

It is not well established how to proceed for the scenarios represented in Figure 1.1 where there is partial overlap. If the number of pairs is large, a 'standard' approach is to perform the paired samples t-test on only the paired observations (Looney and Jones, 2003). Conversely, if the number of independent samples is large, a 'standard' approach is to perform the independent samples t-test on only the independent observations (Looney and Jones, 2003). The exclusion of data can lead to information loss (Ramosaj and Pauly, 2017). These approaches have adverse consequences on the power of the test. Approaches that discard data are likely to maintain adequate power if the number of discarded observations is relatively 'small' and the sample sizes are relatively 'large'. Approaches that discard observations to perform a basic traditional test are referred to as naive approaches (Mantilla and Terpstra, 2018; Guo and Yuan, 2017).

An alternative approach that is commonly applied, is to perform the independent samples t-test on all of the available data (Samawi and Vogel, 2014a). However, this is less powerful than a paired samples approach and ignores the fact that there are pairs. A further alternative approach is to impute data, but this can lead to incorrect and inconsistent conclusions (Ramosaj, Amro, and Pauly, 2018; Zhu, Xu, and Ahn, 2019). Basic imputation approaches are biased solutions (Schafer, 1997). Mean imputation reduces the variation in the data set. Regression imputation inflates the correlation between variables. Some of the more sophisticated techniques, expected maximisation and multiple imputation, minimise the bias of the parameter estimates (Musil et al., 2002; Dong and Peng, 2013). However, these methods do not fully control the Type I error rates of the resulting paired samples t-test (Ramosaj, Amro, and Pauly, 2018). In addition, imputation techniques assume that data is missing and does not take into account that samples may be partially overlapping by design.

To demonstrate the breadth of the problem, the following list illustrates some situations where partially overlapping samples can occur:

1. A matched pairs design where some participants have no similar attributes. In a matched pairs design, pairs are determined based on similar attributes, but it may not be possible to find an appropriate match for all observations (Cochran, 1953).

2. An independent samples design, which inadvertently contains paired observations. In an example by Looney and Jones (2003), participants were randomly allocated to either placebo or active treatment and were each to provide one measurement on the response variable, however some participants allocated to the active treatment group received the placebo by mistake. For the participants where the error was made the response variable was recorded following the placebo, these participants were then given the active treatment and the response variable again recorded. The appropriateness of this may not be without question, but the researchers decided to treat the participants providing response for both treatments as paired observations.

3. Two groups with some common element between both groups. Paired observations and independent observations may be acquired by design. For example, when comparing a treatment for myopia with a treatment for hyperopia, where some individuals require treatment for both. Another example would be the comparison between prices on Amazon and eBay, where not all products are sold by both.

4. A design which includes both paired observations and unpaired observations, due to limited resource of paired observations. When a resource is scarce, researchers may only be able to obtain a limited number of paired observations, but would want to avoid wastage and also make use of the independent observations. For example, in a clinical trial assessing the performance of kidneys following transplantation, one group incorporates a new technique that reconditions the kidney prior to the transplant, and one group is the control group of standard cold storage (Hosgood et al., 2017). When the kidneys arrive at the transplanting centre in pairs, one is randomly allocated to each of the two groups. When a single kidney arrives at the transplanting centre, this is randomly allocated to one of the two groups in a 1:1 ratio. Another example is given in cancer genomic experiments where either the normal tissue or tumour tissue for an individual is not large enough for extraction (Qi, Yan, and Tian, 2018). Further examples are given by Stigler et al. (2018) where the femurs and mandibles between young and old mice are compared, and where samples are taken from a human cadaver.

5. The observations of a paired samples design and a separate independent samples design are combined. In empirical research, paired samples and independent samples may be obtained in separate tranches of a study. This could arise in a situation where practices are different, for example a clinic taking measurements at baseline and followup, a clinic only taking measurements at baseline and a clinic only taking measurements at follow up.

6. Observations taken at two points in time, where the population membership changes over time but retains some common members. When observations are taken on the same study unit at two points in time, it is anticipated that the dependent variable is recorded on both occasions, thus forming paired observations. However, where there is a natural turnover of membership of a group there may be study units that are only available to provide a response on one of the two occasions, thus

forming independent observations. For example, in a study by Banks et al. (2014) to assess dementia care, a questionnaire was conducted on two occasions. Due to the turn-out of participants changing at each session, 83 responses were obtained at Time 1, and 89 responses were obtained at Time 2. The approach taken by Banks et al. (2014) was to perform the independent samples t-test on all of the data. This approach ignored the fact that some observations are paired.

7. A paired samples design, where some observations are untraceable. Scenarios may occur when it may not be able to identify the natural pairing, for example if the pair includes the biological mother and biological father of a child where the latter is 'unknown'. In a medical context, the status for a response variable in some participants may be difficult to detect (Tian, Zhang, and Jiang, 2018).

8. A paired samples design, which inadvertently contains independent observations. Perhaps the most frequent occurrence of partially overlapping samples is a paired samples design with missing observations (Martinez-Camblor, Corral, and De La Hera, 2013; Ramosaj, Amro, and Pauly, 2018). In a medical context where data is missing due to participant drop-out, this can often lead to independent observations in one sample only.

The consequence of poor study design can be the presence of partially overlapping samples, as exhibited in Scenario (2). The need for good research design cannot be over-stressed.

If partially overlapping samples do not occur by design, for example in Scenario (8), it is necessary to consider why the paired samples are incomplete (Kang, 2013). When data are Missing Completely At Random (MCAR), the reason for missing data is not related to the value of the observation itself, or other variables recorded. The assumption of MCAR is often unlikely to be valid, nevertheless it is often assumed (Leon et al., 2006). One approach to verify this assumptions is to compare the Sample 1 paired observations against the Sample 1 observations where the corresponding Sample 2 observations are not present (Leon et al., 2006). Examples of data that are

MCAR include; a question in a survey that is accidentally missed, or data in a laboratory experiment that is accidentally lost or damaged. If incomplete observations are MCAR, it is reasonable to discard the corresponding paired observations without causing bias (Donders et al., 2006).

If data are Missing At Random (MAR), they are missing based on characteristics not directly associated with the missing observation itself. However, the missing data is related to another variable in the dataset. The discarding of information that are MAR is likely to cause bias, therefore the standard approach of pairwise deletion is not recommended (Schafer, 1997; Donders et al., 2006). If data are Missing Not At Random (MNAR), the probability that an observation is missing, directly depends on the value of the observation being recorded. When data are MNAR, there is no statistical procedure that can eliminate potential bias (Musil et al., 2002). This is particularly of concern for analyses with missing data because it is difficult to distinguish between data that is MAR and data that is MNAR. Nevertheless, if the amount of missing data is small, the bias is likely to be inconsequential. The literature suggests that up to 5% of observations missing is acceptable (Graham, 2009; Schafer, 1997). There are some that take a more liberal stance suggesting that up to 20% of data missing may be acceptable (Schlomer, Bauman, and Card, 2010).

Standard statistical software often perform the paired samples *t*-test discarding unpaired observations (Zhu, Xu, and Ahn, 2019). This is often done without any warning to the user. Examples of this include SPSS, SAS and Unistat. Caution should be exercised when employing these software because users may be tempted to analyse only the complete pairs when readily presented with the opportunity, and not realise the consequences of not using all of the data. The ‘scipy.stats’ module within Python will not perform a related samples *t*-test with unequal length arrays. Minitab and the default ‘t.test’ in R present similar error messages when a paired samples *t*-test is selected for unequal sample sizes. Python, Minitab and R at the very least make users aware there are considerations to take into account with the analysis they are trying to perform.

As an aside, Bedeian and Feild (2002) propose three ways that the paired samples t-test could be used, if anonymity means that the pairings are not known; i) arranging observations from one group in ascending order, and from the other group in descending order, ii) arranging observations in both groups in such a way that the correlation between them is zero, iii) using previously known estimates for the correlation. These methods do not solve the problem of using all of the available data if the number of observations in both groups differ. There are also issues how to order the data if there are tied scores. Although the authors state that $\rho = 0$ is achieved by randomly sorting, algorithms may be required to find the arrangement for this, and it may not always be possible to achieve the required correlation by these methods of re-ordering. A random pairing will have a zero correlation structure on average, but the genuine paired data would most likely have a non-zero correlation structure. Furthermore, Bedeian and Feild (2002) cite Zimmerman (1997) so they should be aware of the perils of negatively correlated data and the fact that the paired samples t-test is less powerful when $\rho = 0$. Bedeian and Feild (2002) are concerned with a special case where the pairings are not known, but this highlights the issues with ad-hoc approaches that may also be performed by some researchers in the scenarios above. These methods do not represent justification for forcing the use of the paired samples t-test, but due to the power of the paired samples t-test, a solution that can encompass the paired t-test when the design warrants, would be desirable.

The pitfalls of these existing approaches emphasise the need to form statistically valid tests in the partially overlapping samples case, and to build a consensus regarding best practice.

1.2 Previous research into the comparison of central location for partially overlapping samples

The partially overlapping samples framework is misunderstood (Martinez-Camblor, Corral, and De La Hera, 2013). As part of the partially overlapping samples framework, the research question could be a comparison of means, but could also be a comparison of distributions or variances. This is an overview of the literature regarding the comparison of central location, which has received the bulk of previous attention within the partially overlapping samples framework.

The approach within recent literature is to explore 'robustness' of proposed solutions via simulation. This term usually refers to Type I error robustness. The 'Type I error rate' of a statistical test is the proportion of times a true null hypothesis is rejected. Thus 'Type I error robustness' is defined as a statistical test that rejects a true null hypothesis at the same rate as the nominal significance level. For example, for 10,000 iterations of sample generation where the null hypothesis is true, a Type I error robust solution would reject the null hypothesis approximately 500 times at the 5% significance level, indicating a Type I error rate of approximately 5%. An illustration of this is given in Section 2.4.

Amro and Pauly (2017) define three categories of solution to the partially overlapping samples problem that use all available data and do not rely on resampling methods or Bayesian inference. The categories are; tests based on maximum likelihood estimators (Section 1.2.1), weighted combination tests (Section 1.2.2), and tests based on a simple mean difference (Section 1.2.3). Bayesian techniques could be explored, but this research focuses on frequentist methods because they are commonly applied in many disciplines.

1.2.1 Maximum likelihood estimators

Early literature on partially overlapping samples focused on maximum likelihood estimates for Normal distributions, when data are missing by accident rather than by design. Lin (1973) use maximum likelihood estimates for the specific case where data is missing from one of the two groups. Lin and Stivers (1974) apply this to a more general case, but find no single solution is applicable. For normally distributed data, Ekbohm (1976) compared Lin and Stivers (1974) solutions with similar proposals based on maximum likelihood estimators. He used simulation methods to compare these methods to the standard approaches of using the independent samples t-test using all of the data, and the paired samples t-test discarding incomplete pairs. The results reveal that the most powerful test of those that maintain nominal 5% Type I error rates when $\rho = 0$ is the independent samples t-test. For $\rho \geq 0.5$ the paired samples t-test has greater power than the other tests considered.

Guo and Yuan (2017) reviewed parametric solutions under the condition of normality, and recommend the Lin and Stivers (1974) maximum likelihood approach when the normality assumption is met. However, Amro and Pauly (2017) demonstrate that this maximum likelihood estimator approach has an inflated Type I error rate under normality and non-normality.

Maximum likelihood proposals are complex mathematical procedures, which would be a barrier to some analysts in a practical setting. A related but more practical solution available in most standard software is to fit a mixed model using all of the available data. In a mixed model, effects are assessed using Restricted Maximum Likelihood estimators 'REML'. Within the mixed model the group is declared as a repeated measures fixed effect and the observation units are declared as a random effect. Mehrotra (2004) indicates that for positive group correlation, this REML approach is a Type I error robust and more powerful approach than competing approaches.

1.2.2 Weighted combination tests

Weighted combination tests are where test statistics for independent samples and paired samples are combined, often weighted using complex methods. These tests do not answer the fundamental question of the difference between the two groups on the numerator. Neither do these proposals have a denominator that represents the standard error of the parameter difference. It would be difficult to obtain confidence intervals for the difference between means using these weighted approaches. Further issues arise with the creation of a non-parametric test based on these approaches.

A method by Bhoj (1978) demonstrates reasonable Type I error robustness, although they did not consider situations that violate the normality assumption (Derrick, White, and Toher, in press). Uddin and Hasan (2017) optimised the weighting constants used by Bhoj (1978) so that the combined variance of the two elements is minimised. Yu et al. (2012) reveal that a similar technique proposed by Kim et al. (2005) does not always satisfy liberal robustness criteria.

Samawi and Vogel (2014b) and Martinez-Camblor, Corral, and De La Hera (2013) proposals are in principle adding two t-tests together, paired samples and independent samples, and treating the combination as a t-statistic. It is not statistically correct for small sample sizes that two t-distributions can be added together to form a t-distribution. The weights Samawi and Vogel (2014b) use of $\sqrt{\gamma}$ and $\sqrt{1-\gamma}$ are not ideal, when taking a square root of the weighting function, γ, the weights will not accurately reflect the sample size ratio of the observations in the two samples.

A familiar weighted combination based approach from meta-analysis is to obtain the p-value for a paired samples test (discarding unpaired observations) and the p-value for an independent samples test (discarding paired observations). These are then combined using a weighted z-test (Stouffer et al., 1949). Under Stouffer's method, Z_{combined}, the p-values are transformed to Normal scores. Let ϕ be the standard Normal cumulative distribution, then the sum of two independent normally distributed variables is a normally distributed variable as follows:

$$Z_{\text{combined}} = \frac{\sum \gamma_i Z_i}{\sqrt{\gamma_i^2}} \text{ where } Z_i = \phi^{-1}(1 - p_i).$$

In general it is usual that the weights γ_i are determined by the sample size (Chen, 2011). Alternatively these weights could be calculated so as to maximise power, but there is no one way of trying to decide optimal weights and this would be calculation intensive. Z_{combined} would be a method that practitioners would be more likely to buy in to if weights are based on sample sizes. Key advantages for this technique are that; it can be performed without the requirement of bootstrapping, it can be more easily extended to the situation where there are more than two groups to be compared, and it can be more easily extended to the non-parametric situation.

There are other methods for combining p-values of independent tests, and there is no uniformly most powerful test (Whitlock, 2005). A noteworthy alternative is the generalised Fisher test proposed by Lancaster (1961). When used to combine p-values from independent tests, the latter method is similar but more powerful (Chen, 2011). The key disadvantage of these techniques is that confidence intervals for the mean difference are not easy to obtain. Also of note is that multiple separate tests are required before applying this test.

1.2.3 Tests based on a simple mean difference

These test statistics have a form where the numerator is the difference between two means with a denominator representing the standard error of the difference, thus easing interpretation of the results.

Looney and Jones (2003) proposed a test statistic construction formulated as a linear interpolation between the paired samples z-test and the independent samples z-test. This uses the standard Normal distribution to calculate p-values. This is known as the corrected z-test, $Z_{\text{corrected}}$, and is given as:

$$Z_{\text{corrected}} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_a+n_c} + \frac{S_2^2}{n_b+n_c} - \frac{2n_c S_x}{(n_a+n_c)(n_b+n_c)}}}$$

As per Table 1 on page xi, $\bar{X}_1$ is the mean of Sample 1, $\bar{X}_2$ is the mean of Sample 2, S_1^2 is the variance of Sample 1, S_2^2 is the variance of Sample 1, n_a is the number of observations in Sample 1 only, n_b is the number of observations in Sample 2 only, n_c is the number of pairs. S_x is a measure of the covariance between the paired observations only i.e. $S_x = \rho \times S_{1c} \times S_{2c}$.

In extreme scenarios where there are no paired observations or alternatively no independent observations, this test defaults to the independent samples z-test or alternatively the paired samples z-test. For example, for a value of $n_c = 0$ this would result in the test statistic defaulting to the independent samples z-test as below:

$$Z_{\text{corrected}} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_a+0} + \frac{S_2^2}{n_b+0} - \frac{0}{(n_a+0)(n_b+0)}}} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

The authors demonstrate that when only 10% of the sample is paired, the naive independent samples t-test performs equally as well as $Z_{\text{corrected}}$. When the paired sample is 50% or 90%, $Z_{\text{corrected}}$ to a better extent maintains the Type I error rate relative to the independent samples t-test on all of the data. However, the simulation design considered equal variances only.

Looney and Jones (2003) do not give guidance to how 'large' the samples should be for their z-test, but the paired sample size must be ≥ 3 so that covariance can be calculated. The covariance is calculated based only on the paired observations. Uddin and Hasan (2017) offer a minor adjustment to the calculation of the covariance, however the issue for small sample sizes remain. The test constructed by Looney and Jones (2003) gives credence to the theory that a t-statistic constructed in a similar manner could be used in a greater number of conditions for smaller sample sizes. The method by Looney and Jones (2003) continues to be promoted as one of the best solutions to the partially overlapping samples problem (Looney and McCracken, 2018).

1.3 Aims and Objectives

The principle aim is to develop a complete framework of recommendations for the problem of two groups with partially overlapping samples. To achieve this goal, preliminary investigation into the robustness of extant solutions is performed [see Chapter 2], which informs the derivation of proposed solutions [see Chapter 3]. The experimental design for assessing the proposed solutions is then defined [see Chapter 4].

Within this framework, the aims are to facilitate statisticians and practitioners in the analyses of partially overlapping samples for:

I a comparison of means between two groups, assuming normality [see Chapter 5].

II a comparison of means and/or distributions between two groups, for continuous data, not assuming normality [see Chapter 6].

III a comparison of means and/or distributions between two groups, in the presence of a single or multiple outliers [see Chapter 7].

IV a comparison of means and/or distributions between two groups, for ordinal data [see Chapter 8].

V a comparison of proportions between two groups [see Chapter 9].

VI a comparison of variances between two groups [see Chapter 10].

1.4 Declaration

This thesis, and the publications cited in the coming chapters where I am the lead author, are my own work.

Chapter 2

Literature review and preliminary investigation

The literature on the assumptions of the paired samples t-test and the independent samples t-test are discussed. This leads to the consideration of alternative approaches when the assumptions of these tests are violated. Tests reviewed in this chapter inform solutions given in Chapter 3. Given the concern for violations to assumptions, and the alternative tests available, common practice is to perform preliminary tests of the assumptions before deciding on an appropriate test. This chapter concludes with a review of the preliminary testing process.

2.1 History of the t-test

The t-test is a long established statistical inference technique. In its inception, it was used to test if the mean of a sample is equal to a hypothesised population mean. The t-statistic and t-distribution were developed by the British statistician William Gosset, published alias 'Student' (1908). The context was studying the yields of varieties of barley at the Guinness brewery in Ireland for whom he worked. This original paper has received relatively few citations compared to its frequent use[1]. If more papers making use of

[1] Citations = 1,202. Google Scholar [Feb 2020].

the t-test were to cite its origins it would be among the most cited of all time. The most cited statistics papers are not necessarily the papers that statisticians would deem to be the most important (Van Noorden, Maher, and Nuzzo, 2014).

Before his death in 1937 aged 61, William Gosset went on to produce several works under his pseudonym. His criticism of a famous milk experiment documents the importance of random allocation in a study design to protect internal validity ('Student', 1931).

It was Fisher (1925) and in subsequent editions of his book, who was able to show that the t-test could be used in many circumstances, including where there is more than one sample to be compared. The t-test in its various forms is believed to be the most commonly performed statistical test (Creech, 2018).

There is some colloquial belief that the 't' in 't-distribution' and 't-statistic' is an acronym for the word 'test', however this has the bizarre connotation that the 't-test' would be correctly known as the 'test-test'. The first known reference to Gosset's test as the 't-test' was Fisher (1925). He does not explain his reason for the choice of name for the t-statistic, a likely argument given by some historians is that 't' was simply an available letter.

The t-test has been developed into many forms, for which there are extensive comparisons in the literature. The paired samples t-test and the independent samples t-test are parametric tests of central location, the parameter being the mean.

2.2 Assumptions of the t-test

For an experimental design consisting of randomly selected naturally occurring pairs of observations, a paired samples t-test is typically used to take the pairing into account. Examples include; a split plot design, randomised block design or a repeated measures design. A paired samples t-test may also be referred to as a 'dependent samples' t-test or a 'correlated samples' t-test. A paired samples design can also occur where observations are not naturally paired, but are matched based on similar characteristics, thus the terminology ‘matched pairs’ t-test is also appropriate. Fradette et al. (2003) show that the paired samples t-test, T_1, can be written as:

$$T_1 = \frac{\bar{X_1} - \bar{X_2}}{\sqrt{\frac{S_1^2}{n_c} + \frac{S_2^2}{n_c} - 2\rho\frac{S_1 S_2}{n_c}}} \tag{2.1}$$

The paired samples t-test follows Student’s t-distribution with degrees of freedom equal to $v_1 = n_c - 1$. A simplified, identical form of the paired samples t-test can be written as: $T_1 = \dfrac{\bar{d}}{S_d/\sqrt{n_c}}$

where $\bar{d}$ is the mean of the differences i.e. $\bar{X_1} - \bar{X_2}$, and S_d is the one sample standard deviation of the differences. Given two normally distributed populations with means μ_1 and μ_2, and variances σ_1^2 and σ_2^2, the sampling distribution of the differences, are normally distributed with mean $\mu_1 - \mu_2$, and variance $\sigma_1^2/n_1 + \sigma_2^2/n_2$. In other words, independently and identically distributed $N(\mu,\sigma^2)$ differences can be analysed using the paired samples t-test (British Standards Institution, 1975).

The concept of normality requires observations to form a continuous bell shape curve around the mean to $\pm\infty$. In reality such a distribution does not exist, so when the assumption applies it requires an approximation to normality. The assumption stated in textbooks that observations have to be normally distributed is not strictly correct for two sample tests (Totton and White, 2011). If sample observations are approximately normally distributed, then the difference in the sample means are approximately normally distributed. Thus the question is how closely approximated to the Normal distribution the differences need to be in order for the t-test to be valid, and the extent of the impact to violations. Under the law of the Central Limit Theorem, the distribution of sample means can be approximated by the Normal distribution regardless of the shape of the population distribution. Thus as sample size increases, the differences in means approximate to the Normal distribution. Sawilowsky and Blair (1992) suggest that sample sizes of 30 (or more) are adequate if the data is not highly skewed or contaminated by outliers. The t-distribution asymptotically approximates to the Normal distribution. The t-test is exact under normality, and asymptotically exact when the assumption of normality is relaxed (Ramosaj, Amro, and Pauly, 2018).

Inherent within the normality assumption is the condition that observations are from a continuous metric distribution. Data should not be subjected to excessive rounding which would lead to the data being on a discrete scale, typically the rounding interval should not exceed $\sigma/4$ (Eisenhart, 1947).

Another assumption of the t-test is that observations are independently and identically distributed. This means that observations must be mutually exclusive to other observations.

Laerd (2018) list an additional assumption as the presence of 'no significant outliers'. However, they do not suggest a formal test for outliers, this assumption is subject to further investigation in Chapter 7.

The design of an experiment using the independent samples t-test should have observations sampled from the population or distribution in two mutually exclusive groups. There should be no reason to believe observations from one sample are correlated to observations from the other sample (Zimmerman, 1997). The independent samples t-test, T_2, is devised as the differences in the sample means divided by the standard error of the differences so that:

$$T_2 = \frac{\bar{X_1} - \bar{X_2}}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{2.2}$$

where S_p is the pooled standard deviation $\sqrt{\dfrac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{(n_1-1)+(n_2-1)}}$

Note that Bessel's correction factor is used, i.e. $(N-1)/N$, as per the relationship between the population variance and the sample variance (Kenney and Keeping, 1951, p.161).

The independent samples t-test follows Student's t-distribution with degrees of freedom equal to $v_2 = n_1 + n_2 - 2$.

An additional assumption when performing the independent samples t-test is that the variances are equal between the two groups. The population variances are assumed to be equal and thus the pooled standard error is appropriate. If subjects are randomly assigned to groups this assumption should hold, but it may not hold in naturally occurring samples.

When the assumptions of the test are true, the independent samples t-test is the 'most powerful unbiased test' for detecting differences in means (Sawilowsky and Blair, 1992). For this to be true a further implicit assumption of the independent samples t-test is that there is no correlation between the two groups.

Repeated measures designs can have compromised internal validity, e.g. learning or memory effects. Likewise a matched design could have compromised internal validity through poor matching. However, if a dependent design can avoid extraneous systematic bias, then paired designs can be advantageous when contrasted with between subjects or independent designs. These advantages arise by each pair acting as its own control helping to have a fair comparison. This allows differences or changes between the two samples to be directly examined, i.e. focusing directly on the phenomenon of interest. This has the result of removing systematic between pairs differences, leading to increased power or a reduction in the sample size to retain power compared with the alternative independent design. A paired design is usually more efficient because less subjects are required to collect the same amount of observations. In a paired design the degrees of freedom are less than in an independent design with the same number of observations, potentially resulting in a wider confidence interval. However, with effective pairing the reduction in the standard error more than compensates for this and the result is narrower confidence intervals (Johnson and Bhattacharyya, 1996).

The power of the paired samples t-test relative to the independent samples t-test can be observed by considering relative effect sizes. Cohens's d effect size index is calculated as $\frac{\mu_1 - \mu_2}{\sigma}$. Cohen (1992) shows that to identify a difference in means for a small effect size of 0.2 with 80% power, the number of observations required in each sample is 383 ($v_2 = 784$), whereas under the same conditions in a paired design 199 pairs are required ($v_1 = 198$). Similarly for medium and large effect sizes the number of observations required for a paired samples design is lower.

The impact of performing the independent samples t-test ignoring any pairing, T_2^{all}, was considered by Zimmerman (1997). His findings were that for T_2^{all}, the Type I error rate decreases as ρ increases through -0.5 to 0.5, being around the nominal 5% significance level when $\rho = 0$, whereas for the paired samples t-test the Type I error rate remains close to the nominal significance level for the entire range of correlations considered. The power of the paired samples t-test increases as ρ increases through -0.5 to 0.5. Zimmerman (1997) found that when $\rho = 0$ and for increasingly negative ρ, T_2^{all} is more powerful than the paired t-test for all sample sizes considered. However, for negative values of ρ this is not a fair comparison due to the different Type I error rates.

Even if there was a preconceived assumption that observations form pairs, if the correlation is very small, Fradette et al. (2003) states that the independent samples t-test should be used instead of the paired samples t-test. However, Vonesh (1983) demonstrate that the paired samples t-test is more powerful than the independent samples test when $\rho = 0.25$. In addition Zimmerman (1997) demonstrates that small between group correlations can distort the Type I error rate when using the independent samples t-test.

An alternative to the t-test, the point biserial correlation coefficient could be calculated, this is Pearson's product moment correlation coefficient with one dichotomous variable and one continuous variable. The point biserial correlation coefficient is a useful measure of effect size, and the p-value form an independent samples t-test is equivalent to that from an assessment of the point biserial correlation (Kornbrot, 2005).

As it forms part of the the expanded formula for the paired samples t-test in Equation 2.1, the assumptions of Pearson's correlation coefficient apply. The sample value r is the maximum likelihood estimate of the population correlation coefficient for bivariate normal data (Binder, 1984). However the assumption that data are from a bivariate Normal distribution is not crucial, 'the correlation coefficient is informative about the degree of linear association between the two random quantities regardless of whether their joint distribution is Normal' (Puth, Neuhäuser, and Ruxton, 2014, p.185).

It is known that Pearson's correlation coefficient is only suitable for linear correlation. This would suggest that designs that do not result in linear correlations may not effectively be analysed using the paired samples t-test.

For negative correlation, the Type I error rate of the independent samples t-test is distorted, and the power is lower for the paired samples t-test. Negative correlation should be avoided where possible in the design of an experiment.

Pearson's correlation coefficient is Type I error robust to non-normality when testing for correlation (Duncan and Layard, 1973). This includes correlation on interval and ordinal measurement scales (Havlicek and Peterson, 1977). Exception to this Type I error robustness is when the sample size is very small or the data are extremely non-normal (Zimmerman and Zumbo, 1993). Bishara and Hittner (2012) also found that Pearson's correlation coefficient is Type I error robust, except in extreme scenarios when the number of pairs is small ($n_c = 5$) and both distributions are long tailed. Nevertheless for non-normal data, textbooks frequently give Spearman's rank order correlation as an alternative (Bishara and Hittner, 2012). Fowler (1987) and Zimmerman and Zumbo (1993) found that for non-normal distributions, Spearman's rank correlation coefficient is more powerful than Pearson's correlation coefficient. However, Bishara and Hittner (2012) found that with small sample sizes Spearman's rank correlation coefficient is not Type I error robust.

2.3 Violations of the assumptions

When assumptions of parametric tests are violated, Graybill (1976) identified four potential responses: i) Ignore the violation and proceed with the planned statistical test. ii) Modify the test to account for the violation. iii) Design a new model to satisfy the assumptions. iv) Use a non-parametric or distribution free procedure. These four potential responses are considered in turn.

i) In Section 2.3 the robustness of the independent samples t-test and the paired samples t-test to violations of their assumptions is discussed alongside the consequences of disregarding the assumptions.

ii) A well-known modification to a test statistic, applied when equal variances are not assumed, is considered in Section 2.4.

iii) Examples of designing a new model to satisfy assumptions include, taking trimmed means or transforming the data. These are considered in Section 2.5.

iv) Non-parametric methods are summarised in Section 2.6.

2.3.1 Normality assumption

Zumbo and Jennings (2002) identify two types of non-normality: (i) samples from non-normal distributions, and (ii) samples from inherently Normal distributions, but with outliers. Robustness of the *t*-tests for these two types of non-normality are considered in turn.

Samples from non-normal distributions

The title of the article by Micceri (1989), 'The Unicorn, The Normal Curve, and Other Improbable Creatures', sums up their assertion that normality is a myth. Totton and White (2011) exhibit the serpentine nature of the normality assumption by referring to normality as 'ubiquitous' but also 'mythical'.

Scenarios where distributions may not be normally distributed are identified by Nunnally (1994, p.160) as '(a) the existence of undefined subpopulations within a target population having different abilities or attitudes, (b) ceiling or floor effects, (c) variability in the difficulty of items within a measure, and (d) treatment effects that change not only the location parameter and variability but also the shape of a distribution'.

To provide evidence that normality is a theoretical construct invented by statisticians, Micceri (1989) obtained distributions from 440 real world scenarios. It was found that only 28.4% were relatively symmetric, and about 15.2% had tails that were approximately Normal. Micceri (1989, p.161) concluded that 'No distributions among those investigated passed all tests of normality, and very few seem to be even reasonably close approximations'.

Whilst Micceri (1989) states that the implications of non-normality are unclear, he does concede that the independent samples *t*-test could be Type I error robust to departures from normality, even for extreme exponential asymmetry seen in psychometric measures. Micceri (1989) also concedes that the independent samples t-test maintains reasonable power under these conditions.

Sawilowsky and Blair (1992) performed simulation studies using various sample sizes on eight non-normal distributions characterised by Micceri (1989), and conclude that t-tests are robust with respect to Type I error rates and power when sample sizes are equal or nearly equal, or samples sizes are 'fairly large'. Boneau (1960) suggest sample sizes of 25-30 suffice. A common non-normal distribution identified by Micceri (1989) is where there is a discrete mass at zero with a gap between the distribution with the rest of the observations, known as an L-shaped distribution. Sawilowsky and Hillman (1992) investigated this distribution and concluded that except for the smallest extreme unequal sample size they considered (5, 15), Type I error rates are similar as observed for normally distributed data. Power curves in all cases were a good match of those expected under normality. Sawilowsky and Hillman (1992, p.242) conclude that the independent samples t-test is robust to non-normality. 'Although the power to find a treatment is diminished with small samples, at least researchers can be assured that the power of the t-statistic will be similar to that in Cohen's tables for dependent variables with this radically non-normal population shape'. This finding is consolidated by Sullivan and D'Agostino (1992) who found that the independent samples t-test is Type I error robust for small samples, even for a distribution where as many as 50% of observations were zero.

Chaffin and Rhiel (1993) found no impact of kurtosis on the Type I error rate of the one sample t-test. Wilcox (1990) also found that kurtosis has little impact on the independent samples t-test. However for extreme levels of skewness, Wilcox (1990) demonstrated that a two tailed test is not as robust for smaller sample sizes.

When sample sizes are equal, the independent samples t-test is robust for two samples from the Lognormal distribution or the Exponential distribution (Zimmerman, 2004). Wilcox (1990) also show that the independent samples t-test is robust if sample sizes are approximately equal, skew approximately equal and sample sizes not too small.

Delaney and Vargha (2000) found that the independent samples t-test is not robust under moderate to high skew and/or kurtosis when sample size <20, or when the groups have different levels of skewness. In addition, Sawilowsky and Blair (1992) and Delaney and Vargha (2000) found that one tailed tests are more sensitive to violations of non-normality. Bradley (1982) found that the independent samples t-test is not robust for the L-shaped distribution, even when applied to very large sample sizes. The L-shaped distribution is highly positively skewed, this suggests that the robustness of validity argument may be reasonable to some degree of the violation of non-normality. Fagerland and Sandvik (2009a) recommend the independent samples t-test for unequal sample sizes, only when the skewness is approximately equal in both samples.

Fradette et al. (2003) consider two groups of observations generated from the Exponential distributed and then two groups of observations generated from the Chi-squared distribution. Both the paired samples t-test and the independent samples t-test are robust when $\rho = 0$. For both distributions, the independent samples t-test does not maintain the nominal Type I error rate when ρ deviates from 0. For the paired samples t-test, there are some failures to maintain Type I error rate within Bradley's liberal criteria, particularly for the Chi-squared distribution, as $\rho \to -1$. Discussion of the power properties reveals negligible difference between the independent samples t-test and paired samples t-test when $\rho = 0$. For these non-normal distributions, the paired samples t-test performs slightly better than the independent samples t-test when $\rho = 0.1$, and the superiority of the paired samples t-test gradually increases as $\rho \to 1$.

The literature reveals that conclusions with respect to comparisons of the independent samples t-test against the paired samples t-test tests are the same under non-normality as under normality. The adverse effects on both tests are increased with severe non-normality.

Samples from inherently normal distributions but with outliers

Jennings, Zumbo, and Joula (2002) apply varying degrees of outlier contamination to simulated normally distributed data, and conclude that generally the paired samples t-test retains robustness when outlier contamination is symmetric. They note that the Type I error rate is quite stable for most degrees of symmetric contamination considered. However, an inflation in the Type I error rate is observed when extreme asymmetric contamination is applied. If Type I error robustness is maintained, then power values are maintained for medium and large effect sizes. Asymmetric contamination results in a power loss for small effect sizes. However, the authors noticed a paradox for low levels of symmetric contamination which increases the power. This power advantage is exacerbated when sample size is small.

If there is some reason to believe outliers occur due to inaccurate data, or that some observations come from separate populations, a proposed solution to the violation is for the rejection of the offending observations, prior to running statistical tests (Preece, 1982). Techniques are available for the detection and removal of outliers. The removal of outliers can increase power. However, it may not always be acceptable to remove outliers as this may bias the result.

Simulations showing the relevance of this assumption are the subject of Chapter 7.

2.3.2 Within sample independence assumption

Although subjects may be allocated randomly to a group, subjects within that group may influence each other and thus the observations lose their independence. This independence of observations within a group assumption is considered in less detail in the literature. Typically this assumption is taken as being true without the performance of preliminary tests to check this. However, the implications of violations to this assumption are grave (Zimmerman, 1997).

Type I error rates increase with positive within group correlations, and Type I error rates decrease with negative within group correlation (Wiedermann and von Eye, 2013; Keller, 2014; Lissitz and Chardos, 1975). This is true for the independent samples t-test and also for non-parametric tests (Wiedermann and von Eye, 2013).

The independent samples t-test is so non-robust to violations of the independence within groups assumption that Keller (2014) suggests that regular critical values from the t-distribution should not be used when this assumption is violated. Keller (2014) performed 95 million t-tests to produce alternative tables of critical values. The discrepancy in Type I error rates from the nominal significance level depends not only on the extent and direction of the correlation but also sample size. It is likely that in the case of non-independence it is the wrong choice to be testing based on the number of subjects within a group, instead the group itself should be one observation unit.

The implications of violations to the independence assumption are so fatal that it is taken as given that should there be a violation of this assumption in the partially overlapping samples case, statistical techniques cannot be easily applied.

2.3.3 Equal variances assumption

Even if random allocation ensures baseline equality of variances, the treatment may increase the variance for the experimental condition compared to the control condition (Delacre, Lakens, and Leys, 2017).

It is well known that a violation to the equal variances assumption impacts the validity of the independent samples t-test. Complications arise for the independent samples t-test when variances are unequal, particularly when the sample sizes are unequal. When the smaller sample size has the greater variance, the probability of rejecting the null hypothesis when it is true is higher than the nominal Type I error rate, the opposite is true when the larger sample size has the greater variance. This is particularly problematic when the smaller sample size has the greater variance (Zimmerman and Zumbo, 1993; Coombs, Algina, and Oltman, 1996).

This gives rise to the dilemma how to compare means in the presence of unequal variances (heteroscedasticity). This question, applied to two independent samples from Normal populations is known as the Behrens-Fisher problem. Behrens (1928) provided a solution for this problem, which was validated using Fishers fiducial inference (Fisher, 1935; Fisher, 1941). However these solutions, namely the fiducial inference, were disputed by Welch (1938), Welch (1947), and Welch (1951).

2.4 Modifying the test to overcome the equal variances assumption

Derrick, Toher, and White (2016) explored the properties of Welch's test to confirm and identify the properties that make Welch's test robust under normality. The themes within that paper are explored in this section.

The form of the t-test not constrained to equal variances is:

$$T_3 = \frac{\bar{X_1} - \bar{X_2}}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \tag{2.3}$$

As an alternative solution to the Behrens-Fisher problem, Welch (1938) derived an asymptotic test that is highly accurate but not exact. The complexity of this asymptotic test limits its practical use (Grimes and Federer, 1982). Welch (1938) also developed an approximation using degrees of freedom v_3, which are an independent random variable equivalent to:

$$v_3 = \frac{(n_1 - 1)(n_2 - 1)}{(n_2 - 1)c^2 + (n_1 - 1)(1 - c)^2} \text{ where } c = \frac{\frac{S_1^2}{n_1}}{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

In a practical environment, Welch's approximation can be used with little loss of accuracy (Wang, 1971; Scheffé, 1970). Additionally, Welch's test maintains better Type I error robustness and has better power properties than the Behrens-Fisher solution (Lee and Gurland, 1975). Fay and Proschan (2010, p.14) confirm that Welch's test 'is approximately valid for the Behrens-Fisher perspective'. Furthermore 'in the case of comparing two sample means, the consensus in the literature seems to be the approval of Welch's approximate solution' (Grimes and Federer, 1982, p.10). Thus the most commonly used solution to the Behrens-Fisher problem is the separate variances t-test with Welch's approximate degrees of freedom, and is referred to henceforth as Welch's test.

Derrick, Toher, and White (2016) show that the minimum value of v_3 is $\min\{n_1,\ n_2\} - 1$.

The maximum value of v_3 is derived as follows:

Let $y_2 = (n_2 - 1)c^2 + (n_1 - 1)(1 - c)^2$, then $\max v_3 \rightarrow \min y_2$.
The turning point where $\frac{dy_2}{dc} = [2(n_2 - 1)c] - [2(n_1 - 1)(1 - c)] = 0$ is:

$$(n_2 - 1)c = (n_1 - 1)(1 - c)$$

$$(n_2 - 1)\left(\frac{\frac{S_1^2}{n_1}}{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}\right) = (n_1 - 1)\left(1 - \frac{\frac{S_1^2}{n_1}}{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}\right)$$

$$\frac{(n_2 - 1)}{(n_1 - 1)}\left(\frac{\frac{S_1^2}{n_1}}{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}\right)\left(\frac{\frac{S_2^2}{n_1}}{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}\right) = 1$$

$$\frac{(n_2 - 1)}{(n_1 - 1)}\left(\frac{S_1^2}{n_1}\right) = \frac{S_2^2}{n_2}$$

$$\frac{S_1^2}{S_2^2} = \frac{n_1(n_1 - 1)}{n_2(n_2 - 1)}$$

The outcome of this property is that the degrees of freedom used in Welch's test are always less than or equal to the degrees of freedom used in the independent samples t-test.

Preliminary simulations are performed to satisfy the use of Welch's approximate degrees of freedom as a basis of a solution to the Behrens (1928) problem. To demonstrate the impact of the degrees of freedom, the independent samples test statistic T_2 but with v_3 degrees of freedom is considered. Likewise, the test statistic T_3 but with v_2 degrees of freedom is considered. These are compared against the standard approaches for the independent samples t-test and Welch's test. Two samples are generated from the Standard Normal distribution $N(0, 1)$, and tested for equal means at the $\alpha = 0.05$ significance level, two sided. This is repeated for 10,000 iterations. This is applied to eight sample size and variance combinations. Table 2.1 summarises the proportions of iterations where H_0 is rejected, i.e. the Type I error rates. Liberal robustness criteria by Bradley (1978) states that the Type I error rate when the nominal $\alpha = 0.05$ should be in the interval [0.025, 0.075]. Values within this interval are highlighted in bold. More details on the simulation

Table 2.1: Type I error robustness of independent samples t-tests, various t-statistic and degrees of freedom (df) combinations.

n_a, n_b	σ_1, σ_2	T_2, v_2	T_2, v_3	T_3, v_2	T_3, v_3
5,5	1,1	**0.050**	**0.045**	**0.050**	**0.045**
5,5	1,2	**0.056**	**0.047**	**0.056**	**0.047**
5,100	1,1	**0.053**	0.012	0.110	**0.056**
5,100	1,2	0.001	0.000	0.093	**0.060**
100,5	1,1	**0.050**	0.011	0.108	**0.055**
100,5	1,2	0.295	0.153	0.118	**0.052**
100,100	1,1	**0.049**	**0.049**	**0.049**	**0.049**
100,100	1,2	**0.050**	**0.049**	**0.050**	**0.049**

Tests performed at α=5% significance level
T_2, v_2 : independent samples t-test
T_2, v_3 : equal variances t-statistic, Welch's df
T_3, v_2 : unequal variances t-statistic, independent samples t-test df
T_3, v_3 : Welch's test

mechanics used throughout can be found in Chapter 4.

Table 2.1 shows that Welch's test, i.e. test statistic T_3 with degrees of freedom v_3, is Type I error robust across all eight scenarios considered. For unequal sample sizes and unequal variances, T_2 used in conjunction with v_2 or v_3, and T_3 used in conjunction with v_2, do not satisfy Bradley's liberal Type I error robustness criteria. Welch's degrees of freedom therefore represent an important property for controlling Type I error rates. However, the composition of the test statistic, which takes into account the two separate sample variances, is also important.

When sample sizes are equal or variances are equal, for any given data set the test statistics for the independent samples *t*-test and Welch's test are equivalent. Therefore, the difference in *p*-values is a direct result of the degrees of freedom used to calculate the critical value.

When variances are not equal, Welch's estimated standard error impacts the critical value, but this effect is smaller than the impact on the test statistic. When the smaller sample size is associated with the larger variance, the effect on the value of the test statistic is exacerbated.

When sample sizes are equal and variances are equal, both the independent samples t-test and Welch's test perform similarly (Zimmerman and Zumbo, 1993; Moser, Stevens, and Watts, 1989). Welch's test maintains nominal Type I error rates for comparing groups of discrete numerical data, i.e. where the frequency of a number of events is recorded (Fagerland, Sandvik, and Mowinckel, 2011).

Welch's test is Type I error robust for normally distributed data, in scenarios when the independent samples t-test is not. Additionally, in situations where the independent samples t-test is Type I error robust, Welch's test is also. For the comparisons of two means from assumed Normal populations, a general rule to preserve Type I error robustness is to use Welch's test if in doubt about the equality of variances.

When reporting the results of Welch's test, historical convention is to round the degrees of freedom down to the nearest integer (Ruxton, 2006). This convention arises from the historical process of checking against statistical tables. This convention ensures that the degrees of freedom estimate is conservative. This practice remains current when reporting results from statistical software (Weir, 2018). In addition, $v_3 \leq v_2$ as shown earlier and stated by Howell (2012). This property also ensures that the degrees of freedom estimate v_3 is a conservative estimate of the true degrees of freedom.

2.5 Designing a new model

Potential solutions regarding remodelling the data include taking trimmed means or transformations.

2.5.1 Trimmed means

A trimmed mean is the mean of an ordered data set, after a percentage of observations have been removed from each tail.

The median is a consistent, unbiased estimator of the population mean, but is less efficient than the sample mean or trimmed mean, but is more sufficient than the trimmed mean. Trimmed means can be described as a compromise between the mean and the median (Bunner and Sawilowsky, 2002).

The removal of data in the calculation of a trimmed mean, warrants the requirement for winsorizing variances in a test statistic making use of trimmed means.

Due to its resistance and efficiency properties, trimming a data set and then winsorizing the variances is often performed to remove extreme values or skew from each tail of the distribution, and reducing the variability. Yuen (1974) first proposed a test statistic such that Welch's test is performed on the trimmed means and winsorized variances. This is known as the as the Yuen-Welch test, sometimes referred to simply as Yuen's test. It is regarded as a potential solution when the assumptions of normality and/or equal variances are violated.

An important consideration when performing the Yuen-Welch test is what the form of the null hypothesis being assessed is. Possible null hypotheses include; the means are equal, the trimmed means are equal, the distributions are equal, or the medians are equal. Depending which null hypothesis is considered, the robustness of the Yuen-Welch test differs (Fagerland and Sandvik, 2009a). Although it is unlikely that a practitioner will be particularly interested in testing whether the trimmed means of two samples are equal, Keselman et al. (2004) state that this is a reasonable null hypothesis because it provides an estimate for typical or the majority of observations. Keselman et al. (2002) and Lix and Keselman (1998) found that when the null hypothesis is that trimmed means are equal, the Yuen-Welch test is Type I error robust for the Normal distribution and the Lognormal distribution, and has greater power than Welch's test for the latter.

The recommended amount of trimming in the literature varies from 10%-25% (Keselman et al., 2004). 10% or 15% trimming from each tail should suffice to tightly control Type I error (Keselman et al., 2004). In the books and articles reviewed, the typical default trimming applied is 10% or 20% per tail. Alternative trimming procedures for the Yuen-Welch test are available. However, when the trimming is less than 20% per tail, there is no conclusive evidence of a practical advantage of replacing Yuen-Welch based approaches with some other method (Wilcox, 2012). It should be noted that a confidence intervals for a trimmed mean are based on winsorized variances and so are computationally complex and further detract from interpretation.

Fagerland and Sandvik (2009a) suggest that that under a null hypothesis of equal means, Welch's test is superior to the Yuen-Welch's test. Under a null hypothesis of equal distributions, Welch's test again performs well, In the case of unequal sample sizes when both sample distributions are highly skewed with a small difference in variance between the groups, the Yuen-Welch test is preferred over Welch's test. If the null hypothesis is equal medians, the Yuen-Welch test performs well.

Wiedermann and Alexandrowicz (2007) found when variances are unequal or the samples are from Lognormal distributions, test statistics using trimmed means are not Type I error robust for assessing equality of means.

2.5.2 Transformations

Skovlund and Fenstad (2001) recommend that transformations are considered to obtain normally distributed data in each sample, so that Welch's test can be performed. Scenarios where transformations may be appropriate are heavily skewed distributions or for two sample distributions that do not have the same shape.

Transforming data, for example using the Box-Cox transformation, often overcomes violations of the normality assumption, so that traditional parametric tests can be applied. However, the Box-Cox transformation is not robust with respect to unequal variances (Zarembka, 1990). Although popular, the Box-Cox transformation rarely results in both normality and equal variances at the same time (Sakia, 1992).

Cohen and Arthur (1991) found that the independent samples t-test performed on log transformed or squared transformed data exhibits satisfactory Type I error robustness.

Even if practitioners are comfortable with the hypothesis of comparing means of transformed data, a suitable transformation may not always be found. However one transformation, that should always give the appearance of normally distributed data is Normal scores transformations (McSweeney and Penfield, 1969). An example test is the Bell and Doksum (1965) test. This involves ranking the data against an ascending random sample of random Normal deviates. However, because a random sample is used, a different result occurs each time the analysis is run.

Inverse Normal Transformations (INTs) based on Fisher and Yates (1938) approximate Normal scores are the most powerful (Beasley, Erickson, and Allison, 2009). The INT by Blom (1958) is the most commonly used. However it makes little difference which method is used because most are linear transformations of one-another (Tukey, 1962).

Beasley, Erickson, and Allison (2009) apply INTs to multiple test statistics and conclude that the use of transformed and rank data maintains the Type I error rate, if the data prior to transformation would likewise. However, Beasley, Erickson, and Allison (2009) show that if the assumptions of Welch's test are violated, the INT procedure does not maintain the Type I error rate, and it is often close to 100%. Looking more closely at the results it is apparent that they are using groups with differing variances. When applying the different variances to data from the Chi-squared distribution for example, the means are also be different, so what is actually being reflected is very high power. This reiterates the importance of being clear what the null hypothesis being tested is.

2.6 Non-parametric tests

When the normality assumption is violated, statisticians often turn to non-parametric or distribution free methods, which make no assumptions about the underlying distribution.

When there are three or more groups containing both paired observations and independent observations, one non-parametric approach is the Skillings-Mack test. This test is equivalent to the Freidman test when data are balanced (Chatfield and Mander, 2009). For an unbalanced design the Skillings-Mack test requires that any block with only one observation is removed. The Skillings-Mack test therefore cannot be used in the two group situation. This gives motivation for the development of appropriate tests for the two sample scenario.

The best known non-parametric test for two independent samples is the Mann-Whitney-Wilcoxon U test, referred to henceforth as the Mann-Whitney test. In textbooks by Mendenhall, Beaver, and Beaver (2012) and Howell (2012), the null hypothesis of the Mann-Whitney test is reported as 'the distributions are equal'. Fagerland and Sandvik (2009b) assert that the null hypothesis is more correctly reported as probability $(X > Y) = 0.5$. For a comparison of two distributions, it is possible that the latter null hypothesis is true, but for the samples to be from distributions of different shape. When the distributions are equal other than in central location, the Mann-Whitney test is a comparison of central location (Skovlund and Fenstad, 2001). However, the Mann-Whitney test is not recommended as a test for a location shift when variances are not equal (Zimmerman, 1987; Penfield, 1994; Moser, Stevens, and Watts, 1989). Ultimately, the Mann-Whitney test can detect differences in the shape of the two sample distributions, or their medians, or their means (Hart, 2001).

The best known non-parametric test for paired samples is the Wilcoxon rank sum test, referred to henceforth as the Wilcoxon test. A variation to this for discrete data known as the Pratt test is introduced in Chapter 8.

Non-parametric tests typically involve ranking the data from the smallest to largest and performing a test on the ranks. These methods are not without criticism because they do not give an indication of the extent of the difference between any two consecutive ranks (Derrick and Toher, 2016). When the central location is the same for each group, non-parametric tests can have have poor power for detecting seemingly obvious distributional differences. For example, Derrick and Toher (2016) compared the points given to countries in the 2016 Eurovision song contest by the jury and by the televote. In this example the observations are paired by country. The mean and median number of points awarded by both the jury and the televote is fixed by design. A Wilcoxon test comparing the distribution of the points awarded by the televote against the points awarded by the juries, gives no evidence that the distributions differ (Z = -0.546, p = 0.585). These conclusions are counter-intuitive to the widely held belief that the jury and televote opinions show a wide disparity, countries with high diaspora perform better in the televote. Similarly, Derrick and Toher (2016) contrast a new scoring system that was used in this edition of the contest against the previous method. The results of a test on the distribution of the ranks for the old method compared to the distribution of the ranks for the new system, suggests that the distributions for the two methods do not differ (Z = -1.500, p = 0.134). However, in reality the positions of the countries in each scoring system are quite different, including crucially the winner.

Fagerland and Sandvik (2009a) compared the independent samples t-test, Welch's test, Mann-Whitney test, Brunner-Munzel test, and the Yuen-Welch test. Fagerland and Sandvik (2009a) conclude that there is not one test that fits all conditions. The authors chart 16 tables of scenarios with varying sample sizes, skewness and variance in two samples. For each scenario they recommend the most appropriate test, with Welch's test occurring most frequently. When comparing two equally skewed distributions with equal variances, the Brunner-Munzel test and Mann-Whitney test are both identified as appropriate tests. Seldom are either the Mann-Whitney test or the Brunner-Munzel test the best test when the two sample variances are unequal, because they are constantly outperformed by Welch's test.

The Mann-Whitney test should only be used when variances are assumed to be equal (Penfield, 1994; Moser, Stevens, and Watts, 1989). An alternative to the Mann-Whitney test is the Fligner-Policello test. It is an adaption of the Mann-Whitney test for tied values and unequal variances. Mickelson (2013) confirms that this test also fails to control Type I error rates under unequal variances, even for large sample sizes.

The minimum sample size that is required for the independent samples t-test is two in each group, whereas for the Mann-Whitney test four per group are required to allow for any possibility of rejecting the null hypothesis (Fay and Proschan, 2010). The independent samples t-test is more efficient than the Mann-Whitney test (McSweeney and Penfield, 1969). In addition, Cohen and Arthur (1991) found that the independent samples t-test on transformed data has greater power than the Mann-Whitney test.

Non-parametric tests are not necessarily the optimum choice even under non-normality (Murphy, 1976). Rasch and Guiard (2004, p.175) state that 'generally the results are such that in most practical cases the parametric approach for inferences about means is so robust that it can be recommended in nearly all applications'. Concurrent violation of both the normality assumption and equal variance assumption can distort the Type I error rate of non-parametric tests to a greater extent than parametric tests (Zimmerman, 1998).

Whereas non-parametric tests make inferences regarding the distributions, distribution free tests allow for the population parameter to be directly compared, without specifying the underlying distribution. Tests following an INT fulfill this property. Penfield (1994) found that the Van der Wearden (1952) INT test, performs similarly to the Mann-Whitney test with respect to maintaining Type I error rates. The power properties are also similar, however the Mann-Whitney test is more powerful for moderate levels of skew. When comparing means, Welch's test may be superior to an INT test (Zimmerman, 2011).

In the presence of non-normality, Hogg (1977) suggest that a parametric test should be carried out, as well as a non-parametric equivalent. If the results are similar, the authors would report the results of the parametric test. If results are not similar they recommend that the results of the non-parametric alternative are reported. This is a contentious viewpoint and could bias the choice of test based on the conclusion required. Instead, the appropriate test should be selected based on the underlying properties of the data and the robustness of the tests. Under non-normality, the comparisons of means may not be of interest, differences between medians may well be more relevant (Wilcox and Charlin, 1986).

2.7 Confidence intervals

Confidence intervals allow insight into the estimation of a difference and the precision of the estimate. A confidence interval can be useful when reported alongside statistical tests (Levine et al., 2008a). There is frequently too much focus on hypothesis testing, confidence intervals may be of more practical interest (Gardner and Altman, 1986).

An alternative to comparing the differences in means (or medians or trimmed means) between two groups, is to compare the confidence interval of the parameter for one sample with the confidence interval of the parameter for the other sample (Peró-Cebollero and Guàrdia-Olmos, 2013). This approach requires the null hypothesis to be defined in terms of the confidence interval overlap. The authors define 'non-strict' criteria as the parameter of either group is not within the confidence interval of the other group. 'Strict' criteria is defined as when the confidence intervals of the two groups do not overlap.

'Non-strict' criteria offers a novel extension to statistical research, traditionally the null hypothesis is the strict criteria. However, the authors results show that non-strict criteria gives particularly poor Type I error rates.

There are many different ways of calculating a confidence interval for a median. For large sample sizes of equal length, the strict criteria is Type I error robust only if using the 'Binomial' method (Peró-Cebollero and Guàrdia-Olmos, 2013). For unequal skew, the Type I error robustness is particularly violated, this may indicate that this procedure has good power properties for detecting if the distributions are not the same. However, given that this procedure is not Type I error robust for equal distributions, these procedures are not considered any further as potential solutions to the partially overlapping samples problem.

A further novel extension that could be considered is 'moderate' criteria, when the parameter of Group 1 is within the confidence interval of Group 2, but the parameter of Group 2 is not within the confidence interval of Group 1, or vice-versa. However, given the poor results obtained for the 'strict' and 'non-strict' criteria, it is unlikely that this alternative would add value.

2.8 Preliminary testing

The themes in this section were presented and discussed at the Research Students Conference in Probability and Statistics (Derrick, 2018b).

Despite the consequences of violations to the assumptions of statistical tests as outlined in Section 2.3, there is often neglect in published work to report on the assumptions of the tests being performed. The American Psychological Association for example does not instruct researchers to check for violations of assumptions or report the checks performed (APA, 2018).

Some researchers religiously perform a parametric test for central location, some researchers perform both a parametric test and a non-parametric test and then report the one that gives a significant result. These are both misguided approaches (Sawilowsky, 2005).

Some researchers perform preliminary hypothesis tests of assumptions to determine whether a parametric or non-parametric test is appropriate. When a preliminary test informs the user which comparison of central location test to perform, the resulting test is referred to as a conditional test. Some researchers choose to do a formal preliminary test of the assumptions, others will do a more informal approach looking at graphics. Advocates of a formal preliminary testing approach for the comparison of two independent samples include Gurland and McCullough (1962) and Gebski and Keech (2003). However, the approach also has its critics (Zimmerman, 2004; Delacre, Lakens, and Leys, 2017).

As an illustration, in the case of choosing a one sample test for central location, Weir, Gwynllyw, and Henderson (2015) advocate performing a formal hypothesis test for normality when samples sizes are small (preliminary test), to determine whether the one sample *t*-test or the one sample Wilcoxon test is performed (conditional test).

When performing a preliminary test, the null hypothesis is rejected as dictated by the significance level of the preliminary test. The conditional test also has a Type I error rate. This double testing increases the chances of Type I errors and thus can be detrimental (Moser and Stevens, 1992; Rasch, Kubinger, and Moder, 2011). A further limitation of preliminary testing identified by Hoekstra, Kiers, and Johnson (2012) is that assumptions are never strictly true, and those assumptions are made about the population, not the sample.

Although this current work is concerned with a frequentist approach, it is reported that Bayesian approaches are no better than frequentist decisions with respect to Type I errors, and is also subject to bias and lack of precision (Alcala-Quintana and Garcia-Perez, 2004).

Figure 2.1 shows a frequently performed test procedure for two independent samples (Weir, 2018).

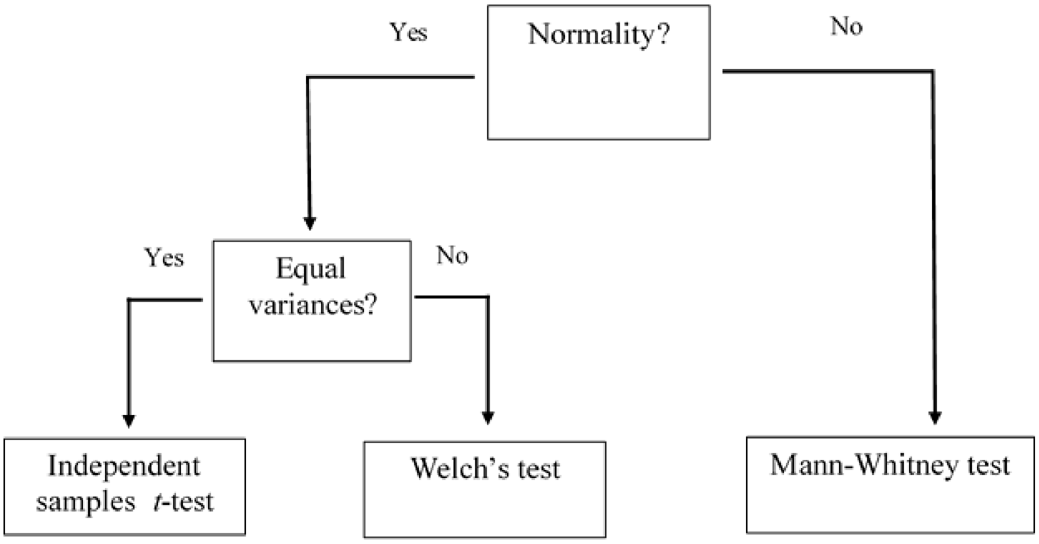


Figure 2.1: A typical two independent samples test procedure

There are numerous ways in which the assumptions in Figure 2.1 could be evaluated. Some two sample test procedures incorporate arbitrary cut-off values for skewness and kurtosis for informing the appropriate conditional test (Kim, 2013).

Fagerland (2012) disagree with the premise of assessing skewness. Fagerland (2012) also suggest that for large sample sizes the parametric test should always be applied. Other methods for determining which two sample test to perform are available throughout the Internet. Examples include Anderson (2014) or Mayfield (2013), which rarely have supporting references, and have the common theme of being relatively vague on how to assess the assumptions.

For the comparison of two independent samples, Ruxton and Neuhauser (2018, p.3) advise that for continuous data Welch'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'. Although this statement is intended to simplify the process, it gives rise to a preliminary check for normality.

Viewing graphical representation instead of formal preliminary testing does not eliminate the problem because the decision on the analysis is conditioned on the results of preliminary analysis (Garcia-Perez, 2012). Furthermore graphical assessment gives rise to subjective interpretation.

Hoekstra, Kiers, and Johnson (2012) investigated the approach by 30 psychology PhD students in the Netherlands when faced with a research question comparing two independent samples. Approximately 1 in 4 checked the assumption of normality (in these cases no formal preliminary test was performed), and approximately 1 in 3 checked the assumption of equal variances (in these cases a formal preliminary test was common). Of those that did not check the assumptions, approximately 2 in 3 were unfamiliar how to check the assumption, and less than 1 in 3 said they regarded the test as robust to violations of the assumption and therefore did not need to check.

Performing preliminary testing based on a pre-defined set of rules can lead to inertness and apathy with regards to the conditional test used. Conversely, some researchers may perform preliminary testing on an ad-hoc basis, and reverse engineer the preliminary tests performed to achieve their desired conclusions.

Publication bias, where only statistically significant findings are published, leads to a temptation by some researchers to adopt practices such as data dredging or data fishing. Publication bias also leads to a temptation by some researchers to report the findings of the statistical test that offers the most significant effect. Inconsistent advice regarding preliminary testing offers researchers opportunity to exploit this practice. This has contributed to the reproducibility crisis in the sciences (Baker, 2016). This adds weight to those that recommend against preliminary testing.

2.8.1 Preliminary tests for normality

Given the *t*-test assumption that two samples arise from the same normally distributed population and the debated robustness of statistical tests, standard practice is to first test the samples for normality (Mahdizadeh, 2018). It should be noted that it is more appropriate to test normality of the residuals rather than the data itself (Totton and White, 2011).

There are numerous tests for normality (Razali and Wah, 2011). Example tests for normality include the Shapiro-Wilk test and the Epps-Pulley test. These two tests are the recommended tests for normality, and in a practical sense there is little to choose between them (British Standards Institution, 1997).

The Shapiro-Wilk test is Type I error robust regardless of sample size (Mendes and Pala, 2003). However, for small sample sizes, the Shapiro-Wilk test lacks power to detect deviations from normality (Rochon, Gondan, and Kieser, 2012; Razali and Wah, 2011).

The most commonly applied normality test is the Kolmogorov-Smirnov test (Ghasemi and Zahediasl, 2012). This is likely because it is readily available in most statistical software, and can be used to test a data set against any distribution. When testing for normality, the Kolmogorov-Smirnov test is more conservative, and therefore less sensitive than the Shapiro-Wilk test (Shapiro, Wilk, and Chen, 1968). The Shapiro-Wilk test has good power and has therefore become the most widely advocated test for normality (Razali and Wah, 2011; Mendes and Pala, 2003; Ghasemi and Zahediasl, 2012). Tests for normality are widely researched, with authors striving for and continuing to develop more powerful tests for normality (Mahdizadeh, 2018). However, for preliminary testing of assumptions, the insensitive nature of the Komologorov-Smirnov test to minor deviations from normality could be advantageous in a practical environment, due to the robustness of parametric tests.

Lumley et al. (2002) suggest that for large samples in public health data there is no requirement for a normalilty assumption. With smaller sample sizes, Lumley et al. (2002) conclude that tests for non-normality are undesirable because they have low power and they detract from the real analyses.

Rochon, Gondan, and Kieser (2012) investigated the Type I error rates of conditional tests for two samples of equal size, performing the Shapiro-Wilk test for normality followed by the independent samples t-test or the Mann-

Whitney test as determined from the result of the normality test. Rochon, Gondan, and Kieser (2012) inform that performing a preliminary Shapiro-Wilk test maintains the nominal significance level of the conditional test for two samples of normally or uniformly distributed data. However, for exponentially distributed populations, this preliminary testing process increases the probability of making Type I error. Therefore preliminary tests for normality may be reasonable if the data are symmetric. Rochon, Gondan, and Kieser (2012) conclude that the t-test is robust in many situations, so preliminary testing does little harm but is a waste of time. Given their assertion that normality is a myth, Micceri (1989) also dismiss the preliminary testing process as futile.

2.8.2 Preliminary tests for equal variances

For unequal sample sizes, statisticians debate the conditions for which the independent samples t-test is robust when the assumption of equal variances is violated (Nguyen et al., 2012). As a result of this uncertainty, common practice is to test for equality of variances prior to performing a test of equal means.

The 'proc ttest' in SAS provides results from the independent samples t-test, Welch's test (referred to as Satterthwaite's test) and a conditional test based on the result of an F-test for equal variances. When sample sizes are unequal, if the F-test reports a significant difference in the variances at $\alpha = 0.05$, Welch's test is used, otherwise the independent samples t-test is used. This approach was investigated by Nguyen et al. (2012) for normally distributed data. They found that when sample sizes are equal, the independent samples t-test is the most Type I error robust. In addition, when sample sizes are unequal, Welch's test and the conditional test procedure perform best against Bradley's liberal robustness criteria. They also found that as the sample size imbalance between the two groups increases, Welch's tests maintains the nominal Type I error rate slightly better than the conditional test procedure, likely due to the double testing applied under conditional testing. Larger sample sizes do not improve the Type I error rate for the independent samples t-test, but do for the conditional and Welch's test. The conditional test procedure showed a very slight power advantage. By adding skewness and kurtosis, Kellermann et al. (2013) reasonably replicate the conclusions by Nguyen et al. (2012) for all nominal significance levels. Under non-normality their results suggests that the independent samples t-test is generally robust. However, in these scenarios the sample size required is greater for the conditional method and Welch's test to stay within Bradley's liberal criteria. The α value for the preliminary test was also investigated and found to be optimal at 0.20 by Nguyen et al. (2012) and 0.25 by Kellermann et al. (2013).

A widely used test for equality of variances is Levene's test, which in the two group case is equivalent to the independent samples t-test on absolute deviations from the mean. Brown and Forsythe (1974) proposed alternatives to Levene's test for when data are not normally distributed. They found that absolute deviations from trimmed means, maintains Type I errors far better for symmetric distributions with long tails (10% trimming was arbitrarily chosen). They also found that absolute deviations from the median maintains nominal Type I error far better than Levene's test, particularly for asymmetric distributions. The modified Levene's test using absolute deviations from the median, known as the Brown-Forsythe method is computationally more complex, but is more widely recognised for its robustness (Nordstokke and Zumbo, 2007; Zimmerman, 2004). Zimmerman (2004) found that when performing Levene's test as a preliminary test, the overall Type I error rate is less than the nominal significance level when the higher sample size was associated with the higher variance, but more than the nominal significance level when the reverse is true.

Generally it is not a good idea to test for homogeneity of variances, and this approach in its present form is no longer widely recommended (Zimmerman, 2004). The decision to either use the independent samples t-test or Welch's test should be made at the design stage of an experiment (Zumbo and Coulombe, 1997). Under normality, the independent samples t-test is clearly inadequate for increasingly unequal variances (Zimmerman and Zumbo, 2009; Kellermann et al., 2013), and Welch's test should be used in these situations instead. Zimmerman and Zumbo (2009) and Zimmerman (2004) both recommend performing Welch's test whenever sample sizes are unequal. Based on results similar to this, Ruxton (2006) suggest the routine use of Welch's test under normality. This approach results in a loss of power when the variances are equal, but a power gain when they are not. However, Fairfield-Smith (1936) state that there is no uniformly most powerful and unbiased test.

Another issue of performing a preliminary test is that different preliminary tests can give different conclusions, thus informing to use different conditional tests. For example, SPSS and Minitab both report values for 'Levene's test' but the results are not the same. SPSS uses Levene's test based on the absolute deviations from the mean, whereas Minitab uses the Brown-Forsythe modification which is based on the absolute deviations from the median. The choice of which test of variances to use therefore requires judgments about the distribution of the data. In fact there are dozens of proposed tests for equal variances (Conover, Johnson, and Johnson, 1981). In their preliminary testing procedure, Anderson (2014) cite three such tests without stating which to perform when. Given the vast array of potential preliminary tests, as an extreme approach, preliminary preliminary tests could be performed in order to select which preliminary tests to perform.

A judgment is required, so it could be argued that a practitioner could simply make a judgment on which form of the *t*-test to use from prior knowledge. For example, for a completely randomised design it is fair to assume equal variances given that both groups are being filled at random from the same population. For naturally occurring groups, for example if groups are split between male and female, a judgment is needed whether equal variances can be assumed, but it is likely that it is not reasonable in this instance (Zumbo and Coulombe, 1997).

2.8.3 Significance level for preliminary tests

A further consideration for preliminary testing is the optimum significance level to work at for the preliminary tests. The 5% significance level is usually used but this could be altered.

In the following simulation investigation, the significance level for the preliminary tests is considered for competing tests for assessing equality of variances, and competing tests for assessing normality, while performing the conditional test of interest each time at the 5% significance level. In a two independent samples design, each of the Mann-Whitney test, the independent samples t-test and Welch's test are performed to compare two generated samples. The preliminary tests for normality performed are the Shapiro-Wilk test (SW) and the Kolmogorov-Smirnov test (KS). The tests for equality of variances are Levene's test (L) and the Brown-Forsythe test (BF). Each preliminary test is performed on each conditional test. The preliminary tests are performed at the 0% to the 10% significance level in increments of 1%. The conditional test is calculated based on the results of each of the preliminary test combinations. The sample sizes varied within a factorial design are $\{5, 10, 20, 30\}$. For each sample size, preliminary test and significance level combination, the process is repeated for 10,000 iterations. The Type I error rate of the overall test procedure in Figure 2.1 is calculated as the weighted average Type I error rate across each of the conditional tests. The first set of simulations is performed where both samples are taken from a $N(0, 1)$. The process is repeated where one sample is taken form $N(0, 1)$ and the other is taken from $N(0, 4)$. The process is further repeated where both samples are taken form the Exponential distribution and then when both samples are taken from the Lognormal distribution. An overview of the results is given in Table 2.2.

Even for the most skewed distribution considered, Table 2.2 indicates that the procedure identified in Figure 2.1 is Type I error robust, because none of the Type I error rates greatly deviate from 5%.

When samples are drawn from the Normal distribution with equal variances, each of the decision rules applied are all approximately equally Type I error robust. This is because all of the tests are Type I error robust under normality and equal variances, therefore the choice of preliminary test is irrelevant. When the two samples come from Normal distributions with unequal variances, the Type I error rate is inflated by performing preliminary tests, however this inflation is within a liberal tolerable region as defined by

Table 2.2: Robustness of preliminary testing procedure in Figure 2.1.

Preliminary tests	Normal	Normal	Exponential	Lognormal
	$\sigma_1^2 = \sigma_1^2$	$\sigma_1^2 \neq \sigma_1^2$		
SW 1%, L 1%	**0.050**	**0.058**	**0.042**	**0.038**
SW 5%, L 5%	**0.051**	**0.057**	**0.052**	**0.046**
SW 10%, L 10%	**0.052**	**0.057**	**0.058**	**0.049**
SW 1%, L 10%	**0.053**	**0.058**	**0.056**	**0.050**
SW 10%, L 1%	**0.049**	**0.058**	**0.047**	**0.043**
KS 1%, L 1%	**0.051**	**0.060**	**0.050**	**0.046**
KS 5%, L 5%	**0.053**	**0.061**	**0.050**	**0.046**
KS 10%, L 10%	**0.054**	**0.061**	**0.049**	**0.046**
KS 1%, L 10%	**0.054**	**0.059**	**0.053**	**0.047**
KS 10%, L 1%	**0.051**	**0.061**	**0.049**	**0.046**
KS 1%, BF 1%	**0.051**	**0.062**	**0.051**	**0.048**
KS 5%, BF 5%	**0.053**	**0.063**	**0.050**	**0.047**
KS 10%, BF 10%	**0.055**	**0.063**	**0.050**	**0.046**
KS 1%, BF 10%	**0.054**	**0.062**	**0.053**	**0.047**
KS 10%, BF 1%	**0.051**	**0.063**	**0.050**	**0.047**
KS 1%, BF 1%	**0.050**	**0.060**	**0.044**	**0.045**
KS 5%, BF 5%	**0.051**	**0.060**	**0.047**	**0.043**
KS 10%, BF 10%	**0.052**	**0.059**	**0.053**	**0.045**
KS 1%, BF 10%	**0.054**	**0.060**	**0.048**	**0.041**
KS 10%, BF 1%	**0.049**	**0.060**	**0.049**	**0.047**

Type I error rates for the independent samples t-test (α=5%) following Shapiro-Wilk (SW) or Kolmogorov-Smirnov (KS) test for normality and Brown-Forsythe (BF) or Levene's (L) test for equal variances

Bradley (1978).

Across all of the distributions, the robustness of the overall test procedure is not greatly impacted by altering the significance level of the preliminary test within the simulated range. Furthermore, the robustness is not greatly impacted by the combination of preliminary tests.

The results in Table 2.2 suggest that the procedure as per Weir (2018), who suggests routine use of the Shapiro-Wilk test at the 5% significance level and Levene's test at the 5% significance level, is Type I error robust. Using the same simulation methodology but for a slightly different decision tree which incorporates the Yuen-Welch test when variances are unequal and the normality assumption is unreasonable, Pearce and Derrick (2019) recommend a two-stage preliminary testing procedure with a Kolmogorov-Smirnov normality test and Levene's test for equal variances, both at the 5% significance level. Akin with the strategy in Figure 2.1, Pearce and Derrick (2019) note that for this slightly different strategy there is not much to choose between different preliminary tests and significance levels. The recommendation of a specific approach within the procedure by Pearce and Derrick (2019) was for the purposes of encouraging a consistent approach only.

The range of different strategies for preliminary testing used are not necessarily poor strategies, so the problem becomes the potential for manipulation and the selection of a strategy for the wrong reasons. Some disciplines request that protocols and analysis plans are pre-registered, examples include the British Medical Journal, Trials, the Journal of Development Economics and the Center for Open Science. This is recommended as an appropriate course of action. When preparing the plan, the assumptions should be assessed based on prior knowledge or preliminary testing of test data (Wells and Hintze, 2007).

2.9 Summary

There are many approaches that could be followed in the comparison of central location for two samples, they each have merit in different scenarios depending on the underlying distributional properties of the data. Welch's test is most frequently the test of choice in Fagerland and Sandvik (2009a). A limitation of parametric and non-parametric tests is that in some scenarios none of the traditional tests are valid, particularly when the two sample distributions have unequal sample sizes and unequal variances (Fagerland and Sandvik, 2009a).

It is more desirable to report the result of a parametric test when normality exists (Huber, 2011). It is wrong to say that parametric tests are always more powerful than non-parametric tests, but is correct when the underpinning assumptions of the parametric test are true (Sawilowsky and Blair, 1992).

Solutions for the partially overlapping samples problem would be useful in parametric and non-parametric form. Then the choice of test can be based on sound underlying logic and assumptions regarding the context being analysed.

Allowing the sample to determine the analysis approach can lead to poor practices. Where methods for analysis are considered approximately equally robust, the analysis strategy should be determined in advance.

Chapter 3

Proposed solutions for the comparison of means with partially overlapping samples

Newly proposed solutions to the partially overlapping samples problem are defined. These new solutions are within the category of a test based on a simple mean difference as defined by Amro and Pauly (2017). The chapter concludes with example uses of these solutions for illustrative purposes.

3.1 Derivation of solutions

In accordance with Chapter 2, a method of comparing two partially overlapping samples that takes into account a paired design but does not lose the unpaired information is proposed.

Collectively, the tests in Section 3.1.1 and Section 3.1.2 are the newly proposed 'partially overlapping samples t-tests'. They are derived from the result for the difference between two random variables. E.g. for means $\bar{X}_1$ and $\bar{X}_2$ the standardised difference is given as:

$$\frac{\bar{X}_1 - \bar{X}_2}{\sqrt{Var(\bar{X}_1) + Var(\bar{X}_2) - 2Cov(\bar{X}_1, \bar{X}_2)}}$$

For both forms of the partially overlapping samples t-test, following calculation of the degrees of freedom, the percentage points of the t-distribution available in any set of statistical tables can be used to obtain the critical value.

For calculating effect size, Cohen's d can be written as $2t/\sqrt{v}$ (Rosenthal and Rosnow, 1991). Thus effect sizes can be readily calculated from the elements of the partially overlapping samples t-test in the same manner.

For two samples of size n_1 and n_2, in the following n_a represents the number of observations in Sample 1 only, n_b represents the number of observations in Sample 2 only, and n_c represents the number of pairs. Other notation is as per convention for the comparison of two samples, see Table 1 on page xi.

3.1.1 Partially overlapping samples t-test, equal variances

The partially overlapping samples t-test assuming equal variances, T_{new1}, is as follows:

$$T_{\text{new1}} = \frac{\bar{X}_1 - \bar{X}_2}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2} - 2\rho\frac{n_c}{n_1 n_2}}} \text{ where } S_p = \sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{(n_1-1)+(n_2-1)}} \quad (3.1)$$

The test statistic T_{new1} is referenced against the t-distribution with degrees of freedom, v_{new1}, derived by linear interpolation between v_1 and v_2:

$$v_{\text{new1}} = (n_c - 1) + \left(\frac{n_a + n_b + n_c - 1}{n_a + n_b + 2n_c}\right)(n_a + n_b) \quad (3.2)$$

Proof: let X = $n_a + n_b$, and let Y = v_{new_1}. If $n_a = 0$ and $n_b = 0$, then $v_{\text{new}_1} = n_c - 1$. The maximum number of observations, $n_a + n_b + 2n_c$, has $v_{\text{new}_1} = n_a + n_b + 2n_c - 2$.
The gradient $m = \dfrac{n_a + n_b + 2n_c - 2 - (n_c - 1)}{n_a + n_b + 2n_c}$.
Substituting into the equation for a straight line $Y = a + mX$ gives v_{new1} as per Equation 3.2.

In the extremes of $n_c = 0$ or $n_a = 0$ and $n_b = 0$, T_{new1} defaults to the independent samples t-test or the paired samples t-test. To demonstrate this, if the samples were independent then $n_a = n_1$, $n_b = n_2$ and $n_c = 0$. Thus T_{new1} defaults to the independent samples t-test as follows:

$$T_{\text{new1}} = \frac{\bar{X}_1 - \bar{X}_2}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2} - 2\rho\frac{0}{n_1 n_2}}} = \frac{\bar{X}_1 - \bar{X}_2}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = T_2$$

$$v_{\text{new1}} = (0 - 1) + \left(\frac{n_1 + n_2 + 0 - 1}{n_1 + n_2 + 0}\right)(n_1 + n_2) = n_1 + n_2 - 2 = v_2$$

Alternatively, if there are completely matched pairs then $n_a = 0$, $n_b = 0$ and $n_c = n_1 = n_2 = n$. Thus T_{new1} defaults to the paired samples t-test as follows:

$$S_p = \sqrt{\frac{(n-1)S_1^2 + (n-1)S_2^2}{(n-1) + (n-1)}} = \sqrt{\frac{S_1^2 + S_2^2}{2}}$$

$$T_{\text{new1}} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2} - 2\rho\frac{S_1 S_2}{n}}} = T_1$$

$$v_{\text{new1}} = (n - 1) + \left(\frac{0 + 0 + n - 1}{0 + 0 + 2n}\right)(0 + 0) = n - 1 = v_1$$

It is observed that the contribution towards the degrees of freedom from the independent observations can be attributed to the total number of independent observations, $n_a + n_b$. Figure 3.1 indicates the relationship between the number of observations, paired and independent, and the degrees of freedom.

3.1.2 Partially overlapping samples t-test, not constrained to equal variances

The partially overlapping samples t-test not constrained to equal variances, T_{new2}, is given as follows:

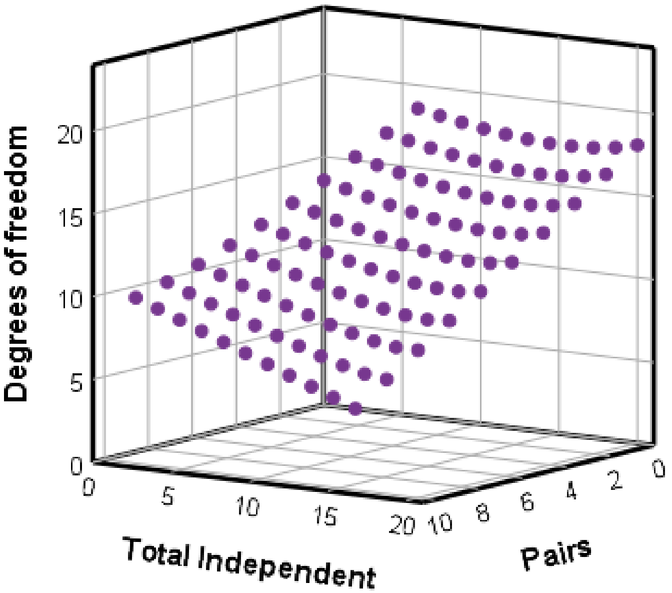


Figure 3.1: Degrees of freedom for T_{new1}.

$$T_{\text{new2}} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2} - 2\rho\frac{S_1 S_2 n_c}{n_1 n_2}}} \tag{3.3}$$

The test statistic T_{new2} is referenced against the t-distribution with degrees of freedom derived as a linear interpolation between v_1 and v_3 so that:

$$v_{\text{new2}} = (n_c - 1) + \left(\frac{\gamma - n_c + 1}{n_a + n_b + 2n_c}\right)(n_a + n_b) \tag{3.4}$$

$$\text{where } \gamma = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\left(\frac{S_1^2}{n_1}\right)^2 / n_1 - 1 + \left(\frac{S_2^2}{n_2}\right)^2 / n_2 - 1}$$

Akin with Welch's test, the degrees of freedom are a random variable, varying from sample to sample, dependent on the sample variance.

In the extremes of no paired observations or no independent observations, T_{new2} defaults to Welch's test or the paired samples t-test respectively.

3.1.3 Partially overlapping samples t-test applied to rank data

For the two sample situation, the relative means, variances, skewness and kurtosis maintain similar characteristics for a distribution transformed to ranks, as are observed in the original distribution (Zimmerman, 2011).

For proposed non-parametric solutions, all observations are pooled into one vector and assigned rank values in ascending order. This is equivalent to an RT-1 (Conover and Iman, 1981) ranking procedure. The rank values are substituted into the calculation of T_{new1} and T_{new2} in place of the observed values. Tied ranks are each given the median of the tied ranks. This gives the test statistics T_{RNK1} and T_{RNK2} respectively. The degrees of freedom are v_{new1} and v_{new2} respectively, with γ calculated using the pooled rank values. The calculation of r uses an RT-2 (Conover and Iman, 1981) ranking procedure, so that r represents Spearman's rank correlation coefficient between the paired observations.

3.1.4 Partially overlapping samples t-test applied to data following Inverse Normal Transformation

Using the Fisher and Yates INT procedure, the observations are pooled, sorted into ascending order and ranked so that $X_i = \Phi^{-1}\left(\dfrac{y_i - c}{N - 2c + 1}\right)$ where X_i is the ordinary rank of observation i, y_i is the total pooled sample size, Φ^{-1} is the standard Normal quantile function and c is a constant.

Calculating the Van der Wearden (1952) scores, i.e. $c = 0$, and using these scores within the calculation of T_{new1} and T_{new2}, gives distribution free test statistics T_{INT1} and T_{INT2} respectively. The degrees of freedom v_{new1} and v_{new2} respectively, are calculated using the pooled transformed values. The calculation of r is Pearson's correlation coefficient between the transformed paired observations.

3.2 Examples of application

Some practical examples based on existing problems, and some of the earliest known applications of the newly proposed solutions are given.

Example 1: Derrick et al. (2017a)

This is a plausible hypothetical example given in Derrick, Toher, and White (2017), acting as a tutorial on how to proceed when faced with two partially overlapping samples.

The sleep fragmentation index measures the quality of sleep for an individual over one night. A lower sleep fragmentation score represents less disrupted sleep. The research question is whether the genre of a movie watched before bedtime impacts the quality of sleep. Study participants are randomly allocated to either a between subjects design (stage 1) or a repeated measures design (stage 2). The two stages are then combined for analyses. In the first stage of the study, the sleep fragmentation score is taken over one night, for two groups of individuals. A sample of $n_a = 8$ individuals watch a 'horror' movie before bedtime. A separate sample of $n_b = 8$ individuals watch a 'feel good' movie before bedtime. In a second stage of the study, the sleep fragmentation index is recorded over two separate nights, for a sample of $n_c = 8$ individuals watching a 'feel good' movie and a 'horror' movie on two alternate nights before bedtime (with order counterbalanced). When the two stages of the study are combined, the total number of individuals who watched a 'horror' movie is $n_1 = n_a + n_c = 16$. The total number of individuals who watched a 'feel good' movie is $n_2 = n_b + n_c = 16$. The hypothesis being tested is whether the mean sleep fragmentation scores are the same between individuals watching a 'horror' movie and individuals watching a 'feel good' movie. Thus the null hypothesis is $H_0 : \mu_1 = \mu_2$. The alternative hypothesis, assuming a two-sided test, is $H_1 : \mu_1 \neq \mu_2$. This is an example of Scenario (5) in Chapter 1. The sleep fragmentation scores are given in Table 3.1.

In this scenario, from a missing data perspective it would be reasonable to assume MCAR. There are no missing data per se; it is the design of the study

Table 3.1: Sleep fragmentation scores.

Independent samples (stage 1)				Paired samples (stage 2)		
ID	Horror	ID	Feel good	ID	Horror	Feel good
I1	20	I9	10	P1	14	15
I2	21	I10	16	P2	15	10
I3	16	I11	18	P3	18	15
I4	18	I12	16	P4	20	17
I5	14	I13	15	P5	11	13
I6	12	I14	14	P6	19	19
I7	14	I15	13	P7	14	12
I8	17	I16	10	P8	15	13

that results in partially overlapping samples. Therefore standard approaches of discarding either the paired or independent samples are unbiased. However, performing either the paired samples t-test or the independent samples t-test requires discarding exactly half of the observations, and the power of the test is reduced. This therefore is a good example of where a test statistic that makes use of all available data, taking into account both paired and independent observations could be useful.

For either form of the partially overlapping samples t-test, if $\mu_1 > \mu_2$ (i.e. the population mean score for 'horror' movie is greater than the population mean score for 'feel good' movie), then it is anticipated that this will be reflected in the sample values above, and the expectation is to observe a large positive value of the test statistic. Conversely if $\mu_1 < \mu_2$, the expectation would be for a large but negative value of the test statistic to be observed. In absolute terms it is anticipated that large values of the test statistic are observed if H_0 is false. The null hypothesis is rejected if the observed value of the test statistic is greater than the critical value from a t-distribution with the degrees of freedom as defined by v_{new1} or v_{new2}.

To calculate elements for the partially overlapping samples t-test let; $\bar{x}_1$ = mean of all observations in Sample 1 (i.e. the mean for the n_1 observations for individuals watching a 'horror' movie), $\bar{x}_2$ = mean of all observations in Sample 2 (i.e. the mean for the n_2 observations for individuals watching a 'feel good' movie), s_1 = standard deviation of all observations in Sample 1,

s_2 = standard deviation of all observations in Sample 2, and r = Pearson's correlation coefficient for the paired observations only (i.e. in n_c).

The elements of the calculation of the test statistics are: $n_1 = 16$, $n_2 = 16$, $n_a = 8$, $n_b = 8$, $n_c = 8$, $\bar{x}_1 = 16.125$, $\bar{x}_2 = 14.125$, $s_1 = 2.986$, $s_2 = 2.778$, $r = 0.687$, $s_p = 2.884$, $\gamma = 29.845$, $t_{\text{new1}} = 2.421$, $t_{\text{new2}} = 2.419$, $v_{\text{new1}} = 18.500$, $v_{\text{new2}} = 18.422$ and from the t-distribution the critical value is 2.097.

The calculated value of t_{new1} is greater than the critical value, therefore the null hypothesis is rejected ($p = 0.026$). Likewise for t_{new2} ($p = 0.026$).

When using the partially overlapping samples t-test at the 5% significance level, there is a statistically significant difference in the mean sleep fragmentation index between individuals watching a 'horror' movie prior to bedtime, and individuals watching a 'feel good' movie prior to bedtime. The results suggest that individuals watching a 'feel good' movie before bedtime have less disrupted sleep compared to individuals watching a 'horror' movie before bedtime.

For this example the paired samples t-test ($t_1 = 1.821$, $v_1 = 7$, $p = 0.111$), the independent samples t-test ($t_2 = 1.667$, $v_2 = 14$, $p = 0.118$) and Welch's test ($t_3 = 1.667$, $v_1 = 13.912$, $p = 0.118$) all fail to reject the null hypothesis at the 5% significance level. Thus the choice of test to apply is important because the statistical decision is not the same. This example emphasises the lower power for these traditional approaches.

In general, the more observations used in the calculation of a test statistic, the greater the power of the test will be. However, rare situations may arise where the independent observations and the paired observations have mean differences in opposing directions. In these situations, the partially overlapping samples t-test may cancel out these differences, but to ignore either the paired observations or independent observations could create bias.

In this example, the two samples are partially overlapping by design. It is also possible to encounter a partially overlapping samples design, with incomplete observations, as per scenario (8) in Chapter 1. In these situations, the partially overlapping samples t-test can similarly be performed on all available observations, when the missing observations are MCAR. To demonstrate this, consider the situation where there are occasional errors with the machine recording sleep fragmentation. As a result of errors, let the 'horror' observations for individuals 'I1' and 'P1' be missing. There is now one missing independent 'horror' observation and one missing paired observation. The resulting reduction in sample size is further to the detriment of the paired samples t-test, the independent samples t-test and Welch's test. Using the partially overlapping samples t-test, the 'feel good' observation for individual 'P1' is not discarded. Revised elements of the partially overlapping samples t-test are; $n_1 = 14$, $n_2 = 16$, $n_a = 7$, $n_b = 9$, $n_c = 7$, $\bar{x}_1 = 16.000$, $\bar{x}_2 = 14.125$, $s_1 = 2.961$, $s_2 = 2.778$, $r = 0.736$, $s_p = 2.864$, $\gamma = 26.903$, $t_{\text{new1}} = 2.208$, $t_{\text{new2}} = 2.194$, $v_{\text{new1}} = 17.733$, $v_{\text{new2}} = 17.148$. Assuming equal variances and using the test statistic t_{new1}, the p-value is 0.041. For completion, using the test statistic t_{new2}, the p-value is 0.042. H_0 is rejected at the 5% significance level and the statistical conclusions are as before.

Example 2: Pilkington (2017)

The difference between baseline and three month follow-up of a 'HeadStrong' service for breast cancer patients, with respect to the distress caused due to hair loss, was considered by Pilkington (2017). Six variables were recorded based on a summed Likert scale. The number of participants that completed the survey at both baseline and follow-up, n_c, was $8 \leq n_c \leq 9$ depending on the variable under consideration. The number of independent observations, n_a, was $7 \leq n_a \leq 8$. The independent observations were participants with observations recorded at baseline only. This is an example of Scenario (8) in Chapter 1. The drop-outs are assumed to be MCAR.

Pilkington decided to proceed with the Looney and Jones (2003) test, $Z_{\text{corrected}}$. At the time of the submission by Pilkington, the partially overlapping samples t-tests had not been published and therefore results for these tests were not included.

Under the conditions of MCAR, equal variances can be assumed between baseline and follow-up, and the partially overlapping samples t-test using pooled variances could be an appropriate alternative test. This work is now revisited with the additionally included test for illustrative purposes. Based on this test statistic, a 95% confidence interval for the true difference in means using all available data are formed. Table 3.2 shows the p-values calculated when performing each of the tests. Statistically significant mean differences at the 5% significance level are highlighted in bold.

Table 3.2: Pilkington Type I error rates and confidence intervals updated.

Variable	$Z_{\text{corrected}}$	T_{new1}	95% CI
Appearance distress	**<0.001**	**0.014**	**(-4.419, -15.762)**
Self esteem	0.583	0.570	(-3.480, 5.100)
Quality of life	**<0.001**	**0.01**	**(8.314, 27.811)**
Negative acceptance	0.154	0.161	(-10.576, 18.326)
Anxiety	0.611	0.597	(-1.785, 3.035)
Depression	**0.022**	**0.035**	**(-0.499, -0.139)**

$Z_{\text{corrected}}$ by Looney and Jones (2003) as per original analyses
T_{new1} partially overlapping samples t-test equal variances

In this instance, conclusions from the $Z_{\text{corrected}}$ test are the same as in this update incorporating the T_{new1} test.

The 95% confidence intervals for the true difference in means indicate arguably wide intervals due to the small sample size in the study.

Example 3: Fenton et al. (2018)

The Bystander intervention initiative at the University of the West of England, Bristol, is recognised as a promising strategy for the prevention of violence against women in university settings. Students from traditionally male orientated degrees were asked to fill in a questionnaire before and after participating in the bystander intervention initiative. This is an example of Scenario (6) in Chapter 1.

The partially overlapping samples t-test which imposes no assumption of equal variances, T_{new2}, and maximally uses all sample information, was used to compare student responses prior to and after taking part in the bystander intervention initiative. Responses were taken on several summed Likert scales as summarised in Table 3.3.

Table 3.3: Summed Likert scales of before and after bystander initiative.

Scale	Before ($n_1 = 22$)	After ($n_2 = 7$)	p-value
IRMA	1.88 (0.45)	1.75 (0.52)	0.489
DV Myth Acceptance	2.41 (0.51)	1.94 (0.51)	0.475
Intent to Help	3.51 (0.52)	3.88 (0.52)	0.110
Bystander Efficacy	27.84 (15.19)	13.20 (12.25)	**0.010**
Bystander Behaviour	1.14 (2.88)	1.71 (1.89)	0.499
AWS	60.91 (5.66)	63.71 (6.65)	0.382

p-value T_{new2} partially overlapping samples t-test unequal variances

For each scale the mean (and standard deviation) of the responses are given, with the result when performing T_{new2}. As shown in Table 3.3, the partially overlapping t-tests comparing before and after scores indicated that the Bystander Efficacy score significantly decreased, $t(9.37) = 3.19, p = 0.01$, i.e. the participants confidence to intervene significantly increased. For this contrast Cohen's d is estimated to be a very large effect size, $d = 2.08$. No other contrasts reported a significant effect, likely due to the small sample size.

Example 4: Derrick et al. (2017a)

In education, for credit towards an undergraduate statistics course, students may take optional modules in either 'Mathematical Statistics' or 'Operational Research' or both. Management is interested whether the exam marks for the two optional modules differ. This is an example of Scenario (3) in Chapter 1. The data and worked example can be found in Derrick et al. (2017a).

For the REML analysis, a mixed model is fitted with 'Module' as a repeated measures fixed effect with two factors, and 'Student' as a random effect. Results from performing tests for the comparisons of means are given in Table 3.4.

With the exception of REML, the estimates for the mean difference is simply the difference in the means of the two samples based on the observations used in the calculation. It can quickly be seen from Table 3.4 that the conclusions differ depending on the test used. It is of note that only the tests

Table 3.4: Two sample tests for the comparison of two optional modules.

test statistic	estimated mean difference	p-value
T_1	-13.375	0.056
T_2	2.167	0.739
T_3	2.167	0.579
T_{new1}	-12.486	**0.035**
T_{new2}	-12.486	**0.045**
$Z_{correted}$	-12.486	**0.023**
REML	-12.517	**0.027**

using all of the available data result in the rejection of the null hypothesis at the 5% significance level.

Example 5: Rempala and Looney (2006)

A classic example by Rempala and Looney (2006), was used by Guo and Yuan (2017) and Amro and Pauly (2017) to illustrate the partially overlapping samples problem. This is an example of Scenario (6) in Chapter 1. In this example, the outcome variable is not recorded on a continuous scale. This is not remarked upon by Guo and Yuan (2017) or Amro and Pauly (2017), who both tackle the problem using parametric methods. This example is henceforth extended to include non-parametric solutions.

The outcome variable is the Karnofsky performance status scale, which measures the functional status of a patient. The data is recorded on the last day of life and on the second to the last day. For parametric tests, the null hypothesis that the mean Karnofsky score is the same on the last two days of life is tested. For non-parametric tests, the null hypothesis that the distribution of the Karnofsky score is the same on the last two days is tested. Assuming the distributions differ only in central location, both the parametric and non-parametric tests are assessing the same research question.

For a total of 60 patients, 9 were recorded on both days, 28 were recorded only on the second to the last day, and 23 were recorded only on the last day, observations as per Table 3.5.

Table 3.5: Example from Rempala and Looney (2006).

Patients with scores on both days
(20,10), (30,20), (25,10), (20,20), (25,20), (10,10), (15,15), (20,20), (30,30)
Patients with scores only on the second to the last day
10,10,10,10,15,15,15,20,20,20,20,20,20,20,20,20,20,20, 25,25,25,25,30,30,30,30,30,30
Patients with scores only on the last day
10,10,10,10,10,10,10,10,10,15,15,20,20,20,20,20,20,20,25,25,30,30,30

The parametric partially overlapping samples t-tests provide evidence at the 5% significance level to suggest that there is a difference in the mean Karnofsky scores between the last two days of life ($t_{\text{new1}} = 2.522$, $v_{\text{new1}} = 51.609, p = 0.015$), ($t_{\text{new2}} = 2.522$, $v_{\text{new2}} = 49.341, p = 0.016$).

Turning attention to non-parametric proposals, there are many tied ranks. Using the midpoint for tied ranks, the ranks are allocated as per Table 3.6.

Table 3.6: Ranks applied to Rempala and Looney (2006) data.

Patients with scores on both days
(37,9), (63.5,37), (53.5,9), (37,37), (53.5,37), (9,9), (21,21), (37,37), (63.5,63.5)
Patients with scores only on the second to the last day
9,9,9,9,21,21,21,37,37,37,37,37,37,37,37,37,37,37, 53.5,53.5,53.5,53.5,63.5,63.5,63.5,63.5,63.5,63.5
Patients with scores only on the last day
9,9,9,9,9,9,9,9,9,21,21,37,37,37,37,37,37,37,53.5,53.5,63.5,63.5,63.5

The non-parametric partially overlapping samples t-tests provide evidence at the 5% significance level to suggest that there is a difference in the distribution of the Karnofsky scores between the last two days of life ($T_{\text{RNK1}} = 2.534, p = 0.014$), ($T_{\text{RNK2}} = 2.521, p = 0.015$).

As mentioned in Chapter 2.7, preliminary tests could potentially be performed to determine which of these four partially overlapping samples t-tests to perform. A preliminary test for equal variances on the independent observations gives no evidence to suggest that the variances are not equal, (Levene's test $p = 0.358$, Brown-Forsythe test $p = 0.397$). This suggests that statistics that do not relax the assumption of equal variances may be reasonable. Note that tests for equality of variances could be based on paired observations or independent observations, tests for equal variances using all of the available data appear in Chapter 10. A preliminary test for normality of the differences from the group means as per Totton and White (2011), using all available data, gives evidence to suggest that the normality assumption is violated (Shapiro-Wilk test $p = 0.004$; Kolmogorov-Smirnov test $p = 0.009$). Thus using this sample data to determine the appropriate test that uses all of the available data would lead to the decision to perform T_{RNK1}. However, as noted in Chapter 2.7 the selection of the test in this way is controversial. Instead the appropriate tests should be chosen based on the study design and existing knowledge of the behaviour of the response variable.

As a further alternative, performing the proposed distribution free tests also supplies evidence to reject the null hypothesis of equal means of the transformed data ($T_{\text{INT1}} = 2.15$, $p = 0.036$), ($T_{\text{INT2}} = 2.12$, $p = 0.039$).

The conclusions made for each of the proposed test statistics, T_{new1}, T_{new2}, T_{RNK1}, T_{RNK2}, T_{INT1} and T_{INT2} are consistent with conclusions in the context of this application made using the methods by Looney and Jones (2003), Samawi and Vogel (2011), and Samawi and Vogel (2014b) and using REML (Guo and Yuan, 2017). Amro and Pauly (2017) supply confirmation that these methods are also consistent with conclusions made by the Lin and Stivers (1974) method and their own permutation based proposal. In contrast to these methods making use of all available data, the naive tests, T_1, T_2, T_3, MW, W, all fail to reject H_0.

Example 6: Oliveira-Costa (2018)

Information was collected on the psychological well-being of parents with a child born with a congenital abnormality. The parents self-report well-being measures. However, as per Scenario (7) in Chapter 1, in some cases there is an absent father or the biological father is not known.

In this ongoing research, the paired sample size is large (approximately

500), the number of cases where there is an absent father is large (approximately 220), but the cases of an absent mother are very rare. Although it is likely that the paired samples t-tests on all of the paired data would pick up significant differences, it is considered appropriate to use the partially overlapping samples t-test to make use of all of the available collected data. The test statistic not restricted to equal variances, T_{new2}, is considered appropriate due to previous knowledge suggesting that males and females have greater variability in results of psychological self-reporting.

This example demonstrates the requirement for the proposed test statistic to be robust for extreme sample size imbalances that are not usually considered in the literature.

Chapter 4

Methodology

Procedures for the assessment of the robustness of test statistics are deliberated. This includes a discussion of the random number generation process, which is assessed for validity to illustrate the concept of Type I error robustness. Methodology used in this thesis is as put forward in this chapter (unless explicitly stated otherwise).

4.1 Monte-Carlo methods

Interpretations of ‘statistical robustness’ vary (Rasch and Guiard, 2004). The sensitivity of tests for equal variances to violations of the normality assumption was where the term ‘robustness’ was coined, or moreover lack of robustness (Hogg, 1979).

When a new test statistic is proposed, the robustness for validity (Type I error rate) and efficacy (power) can be explored using simulation techniques (Serlin, 2000).

Frequently, inappropriate emphasis is placed on the power of the test when like for like probabilities of Type I errors are not equal (Zimmerman, 1997). If Type I error rates are not equal it is not possible to correctly compare the power of tests, and thus the preferred test is the one with the Type I error rate closest to the nominal (Penfield, 1994). Box (1953) suggests that the assumptions of tests statistics can be taken too literally, and the important aspect of robustness is for a test to maximise power. The power of the test is the probability of rejecting the null hypothesis when it is false. There is no standard criterion for quantifying when a test can be deemed powerful. In a practical environment power of 80% is typically considered as desirable (Cohen, 1992). However, it is the power of test statistics relevant to each other that is of greater relevance for theoretical development.

Simulation techniques herein involve the random sampling of numbers, then observing the sample properties. These techniques are therefore an application of Monte-Carlo methods. Random samples are taken from various distributions, so that test statistics can be analysed for long run performance in various scenarios. The terminology Monte-Carlo simulation was coined by Von Neumann and Ulam (1951). The terminology is thought to reflect chance outcomes based on probabilities that arise at a casino. Some pictures of a recent trip to Monte-Carlo can be seen in Figure 4.1.

Null hypothesis significance testing (NHST) is most frequently performed with a nil-null hypothesis specifying that no difference between groups is present, with a two directional alternative hypothesis (Levine et al., 2008b).

There are many pitfalls of NHST (Levine et al., 2008a). A prominent limitation is sensitivity to sample size, it is known for many statistical tests that a larger sample size is more likely to conclude a significant result. In addition, the standard approach of testing the probability of observed data given the null hypothesis is true, is not equivalent to the probability that the null hypothesis is true given the observed data. Thus rejection of the null hypothesis, which is always strictly false, cannot be taken as evidence that some other alternative is true.

Whether or not the nil-null hypothesis can ever be true is open to debate (Sawilowsky, 2016), this is because researchers are forced to work with samples rather than the entire population. However, in a Monte-Carlo study the null hypothesis can be true (Rao and Lovric, 2016). In these Monte-Carlo simulations NHST with a nil-null hypothesis is performed at the $\alpha = 0.05$ significance level as standard, two-sided.

The issues associated with NHST can be minimised with understanding of what a p-value represents and using p-values in conjunction with descriptive

Figure 4.1: Sights of Monte-Carlo, August 2018.

statistics, effect size and confidence intervals.

4.1.1 Generating normally distributed random variates

For an independent vector of observations, the Mersenne-Twister algorithm by Matsumoto and Nishimura (1998) is used to generate pairs of random $U(0,1)$ variates, x_1 and x_2. The Mersenne-Twister algorithm is a well-established random number generator in statistical software, and is the default in SPSS. These Uniform variates are transformed into standard Normal variates using the transformation by Paley and Wiener (1934), now better known as the Box and Muller (1958) transformation so that:

$z_1 = \sqrt{-2lnx_1}\cos(x_2 2\pi)$ and $z_2 = \sqrt{-2lnx_1}\,\sin(x_2 2\pi)$.

The pairs are returned sequentially. Further pairs of Uniform variates are transformed into pairs of Normal variates until the required length of Normal variates have been returned. The generation of the n_a and n_b variates are independent to each other, thus the assumption of MCAR is implicit.

This process of transforming Uniform variates to Normal variates is frequently used in practice, and such practice is not discredited (Chay, Fardo, and Mazumdar, 1975). Due to its construction specific to the Normal distribution the Paley and Wiener (1934) approach for generating random numbers may result in more randomness than inverse transform sampling (Devroye, 1986).

4.1.2 Generating non-normal data

For the comparison of test statistics under non-normality, data are generated by transformation of bivariate standard Normal deviates, N as per Forbes et al. (2011). For a moderately skewed distribution, Gumbel variates, G, are generated using the transformation $G =$ -log (-log U), where U is the cumulative distribution function of N. Exponential variates, E, are generated using the transformation $E =$ -log (U) -1. To demonstrate the robustness of the test statistics for a more extreme skewed distribution, bivariate Normal variates, N, are transformed into Lognormal variates, L, using the transformation L = exponential (N). These transformations are used by Zimmerman (2005) in simulation exploration of the performance of the t-test. These transformations ensure that the distributions compared are of the same shape, and only differ in terms of central location.

4.1.3 Assessing randomness

To confirm the legitimacy of the process detailed for the generation of random Normal variates, 1,000 sets of $n = 1,000$ $N(0,1)$ variates are obtained. For each set, the test by Wald and Wolfowitz (1940) is performed. This is a comparison of the number of runs of sequential observations above or below the median against what would be expected using a Normal approximation.

As an aside, the Wald-Wolfowitz test was intended as an alternative non-parametric test for the comparison of two independent samples (Wald and Wolfowitz, 1940). The test is now lesser known in the context of two samples.

The corresponding p-values for $n = 1,000$ Wald-Wolfowitz runs tests are summarised in Figure 4.2.

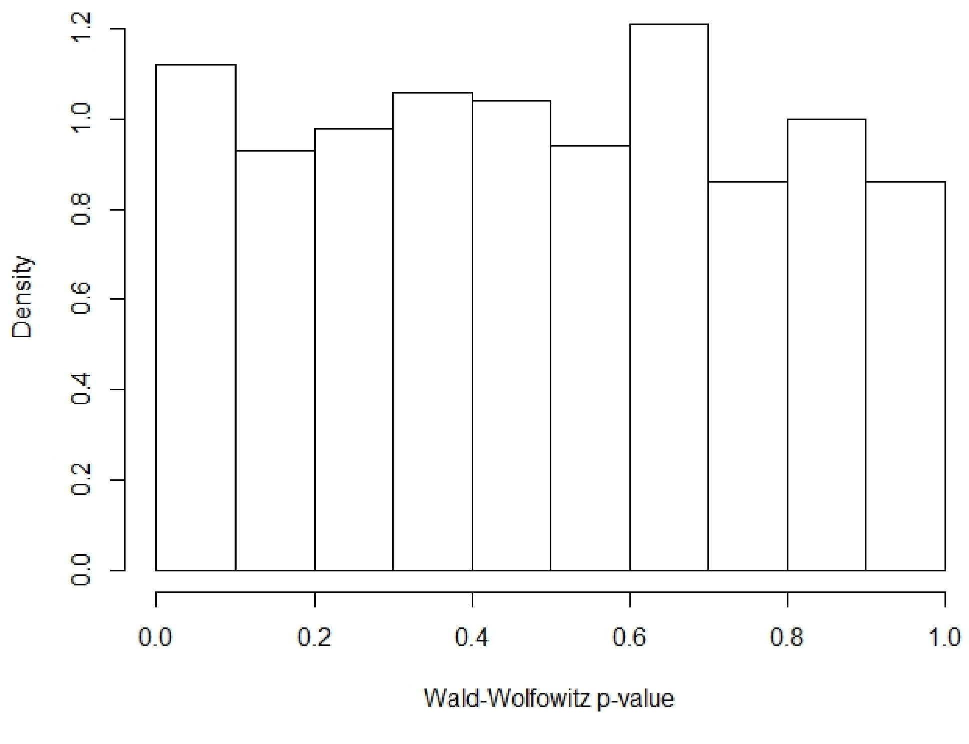


Figure 4.2: Distribution of the p-values for $n = 1,000$ Wald-Wolfowitz runs tests

Figure 4.2 shows approximately uniformly distributed p-values. Hence the variates generated have no detectable systematic pattern.

As an alternative perspective to conceptualise the process of checking the robustness of a test statistic, if we assume that the variates are random, then this indicates that the Wald-Wolfowitz test is valid for checking the assumption of randomness.

The proportion of the $n = 1,000$ Wald-Wolfowitz tests where the null hypothesis of randomness is rejected at the 5% significance level is 0.061, which is approximately equal to the nominal 5% significance level.

The simulations above are repeated for the first 10,000 seeds in R. It is anticipated that the null hypothesis rejection rates would be symmetrically distributed around the nominal significance level. The Type I error rates for 10,000 seeds is summarised in Figure 4.3.

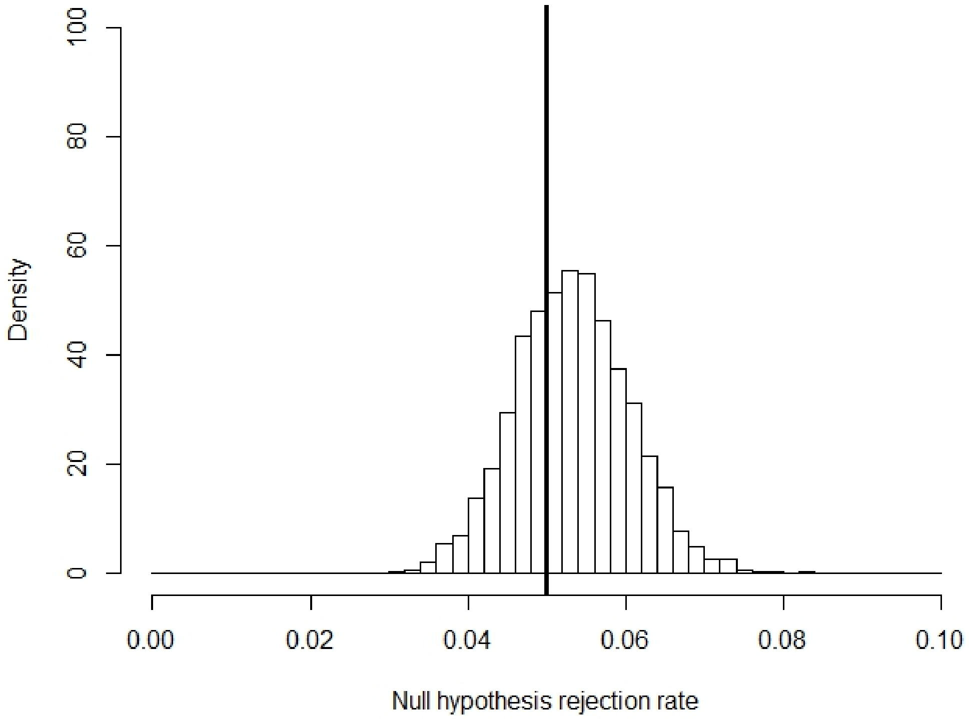


Figure 4.3: Wald-Wolfowitz Null hypothesis rejection rates for 10,000 seeds.

Inspection of Figure 4.3 indicates that the central location of the null hypothesis rejection rate for 10,000 seeds is slightly higher than the anticipated

$\alpha = 0.05$ reference line. This implies that either the data are not perfectly random using this method, or that the Wald-Wolfowitz test is not a perfect test for randomness. It appears that the Wald-Wolfowitz test is liberal, i.e. it slightly inflates Type I error rates (mean = 0.054).

Using the default Inversion method of random number generation in R, this results in the same conclusions as above (output not displayed). Given the widespread praise of both of these random number generators, this suggests that it may be the Wald-Wolfowitz test generating undue concern. To consider this further, the simulations described above are repeated using two alternative randomness tests; the test by Cox and Stuart (1955) and the test by Bartels (1982). The Cox-Stuart test sequentially places observations in pairs and then performs the sign test on the pairs. The Bartel's rank test, slightly more computationally intensive, is a non-parametric version of the test by Von Neumann (1951). For these alternative approaches of assessing randomness, null hypothesis rejection rates for $n = 1,000$ iterations across the same 10,000 seeds are shown in Figure 4.4.

Analyses from the Cox-Stuart test indicate the opposite conclusion to the Wald-Wolfowitz test (mean = 0.044), thus suggesting that if the randomness assumption is true the Cox-Stuart test is conservative. The Bartel's rank test apparently shows that the number generation process ensures random Normal variates (mean = 0.050). Assuming randomness then the Bartel's rank test is the most valid test for randomness.

The three randomness tests collectively give evidence to suggest that the generation process of $N(0,1)$ data is reasonable.

4.1.4 Correlated variates

Correlation has an impact on Type I error rate and power of the paired samples t-test (Fradette et al., 2003), hence a range of correlation coefficients are considered in a thorough simulation design.

For paired observations, Normal variates are generated as above. These are transformed to correlated Normal bivariates, $z_{ij\rho}$, as per Kenney and Keeping (1951) so that:

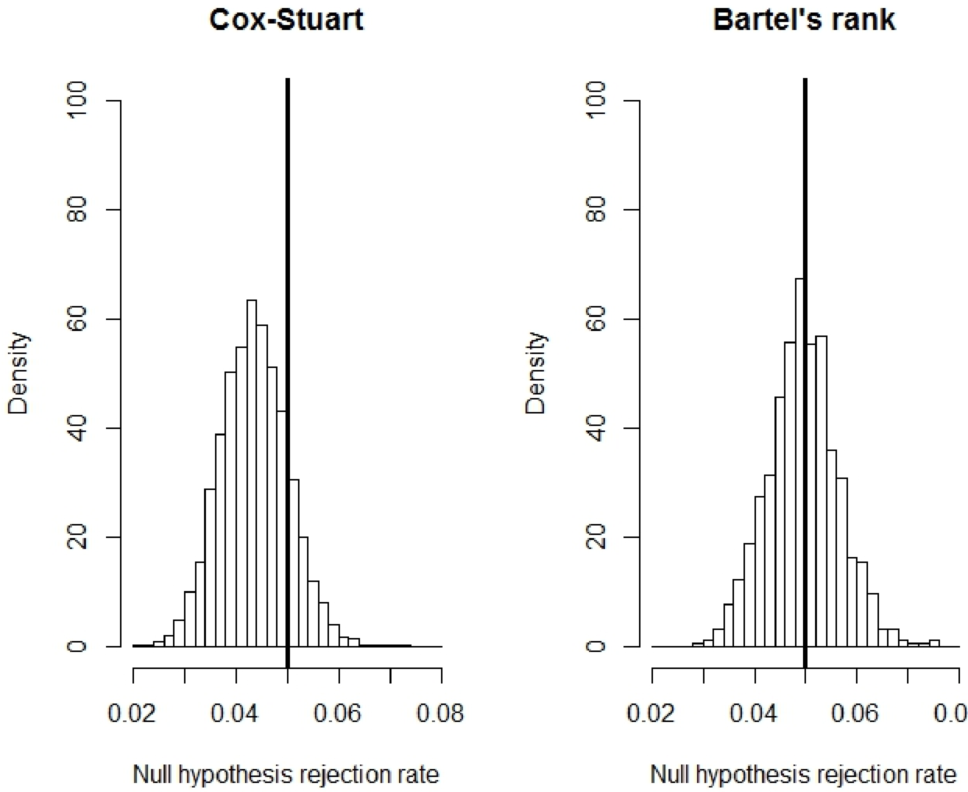


Figure 4.4: Null hypothesis rejection rates for 10,000 seeds.

$z_{1j\rho} = \sqrt{\frac{1+\rho}{2}} z_{1j} + \sqrt{\frac{1-\rho}{2}} z_{2j}$ and $z_{2j\rho} = \sqrt{\frac{1+\rho}{2}} z_{1j} - \sqrt{\frac{1-\rho}{2}} z_{2j}$

where $i =$ (Group 1, Group 2), $j = (1, 2, \ldots,\ n_{12})$ and ρ is the population correlation coefficient between Group 1 and Group2.

4.1.5 Sample size

Unbalanced designs are frequent in psychology (Sawilowsky and Hillman, 1992), thus a comprehensive range of values for n_a, n_b and n_c are simulated. Simulations are performed for large and small sample sizes.

Following the global strive towards metric measures of base ten, sample sizes of factor ten are common in research, with the exception of a smaller sample size of five. This also allows for procedures involving trimming to be considered without undue additional complication.

4.1.6 Number of iterations

The number of iterations (simulation runs) for each parameter combination varies within the literature. Examples from the literature reviewed in Chapter 2 where 10,000 iterations are performed include Fagerland and Sandvik (2009a), Rochon, Gondan, and Kieser (2012), Guiard and Rasch (2004) and Penfield (1994). Varying numbers of iterations from 20,000 to 100,000 depending on the sample size were used by Zimmerman (2011). However, small sample sizes are not necessarily slower at converging, and this decision could bias the results and give the false impression of stability of results for smaller sample sizes.

There is no standard criteria for reasonable precision and hence a judgment call must be made. It appears that 10,000 iterations is the most commonly performed in the literature and is therefore reasonable.

The Null Hypothesis Rejection Rate (NHRR) for each parameter combination is calculated as the proportion of iterations where the null hypothesis is rejected. When the underlying assumptions of the null hypothesis are true, the NHRR represents the Type I error rate of the test.

An indication of the precision of the calculated NHRR as the number of iterations increases is given in Figure 4.5 for selected test statistics. The *p*-values are calculated for a systematically increasing number of iterations up to 50,000 iterations.

It can be seen from Figure 4.5 that for each of the test statistics, the NHRR stabilises as the number of iterations increases. Most of the values are within 50% of the overall mean NHRR when the number of iterations reaches approximately 10,000.

A 95% confidence interval for the deviation from the theoretical value can be mathematically defined, for a given number of iterations n, and is equivalent to $\theta \pm 1.96\sqrt{\frac{\alpha(1-\alpha)}{n}}$. At the 5% significance level, for 10,000 iterations the calculated rejection rate is within 0.00427 of the true value, with 95% confidence.

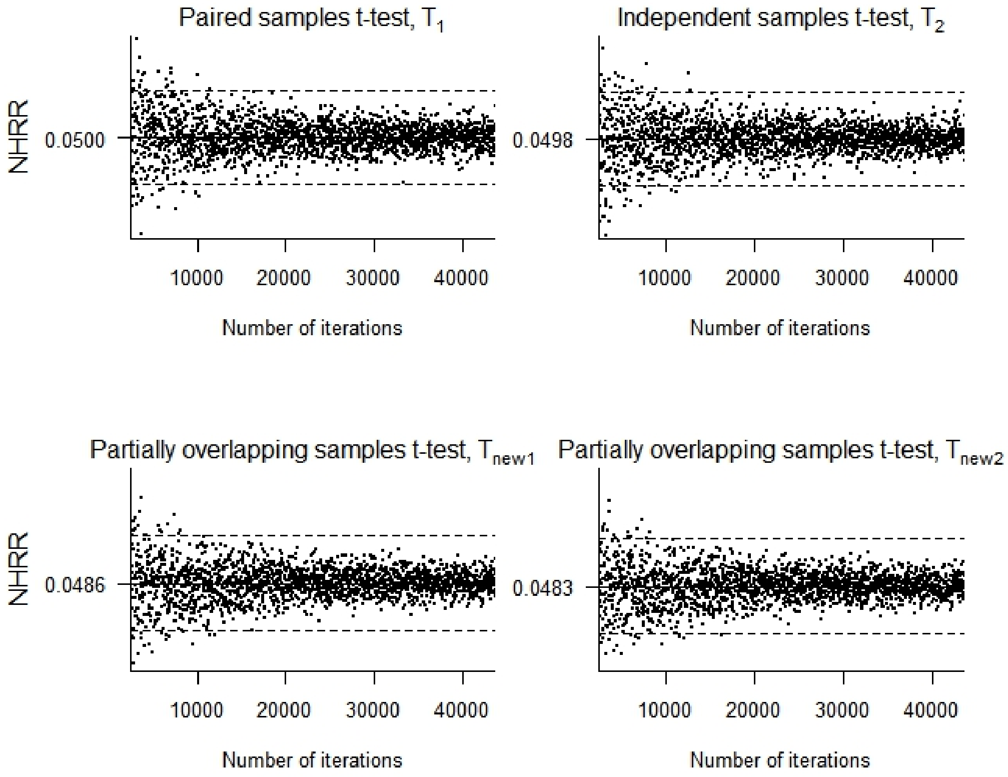


Figure 4.5: NHRR for selected test statistics, $N(0, 1)$, $n_a = n_b = n_c = 5, \rho = 0$. The overall mean across all p-values calculated is recorded. The reference lines represent the tolerance region of within 50% of this value.

Figure 4.5 gives some evidence to suggest that the proposed test statistics are valid for the single parameter combination simulated, thus an extended simulation design using 10,000 iterations per parameter combination may be justified.

For simulations under the alternative hypothesis, an arbitrary amount is added to each observation within Sample 2. In this case the NHRR represents the power of the test.

The simulation process is summarised in Figure 4.6 at the end of this chapter.

4.2 Analysis methodology

Assuming the null hypothesis is true, a test statistic is valid if ordered p-values from the simulation are uniformly distributed (Bland, 2013). Thus the Type I error robustness of a test statistic can be viewed graphically using a probability-probability, P-P plot. The theoretical cumulative distribution function of the uniform distribution, ‘expected’, is plotted on the y-axis. The cumulative distribution function of the empirical p-values, ‘observed’, is plotted on the x-axis. P-P plots are used to show that the two distributions are similar with points approximately on the line. Alternatively, a quantile-quantile, Q-Q plot, produced by plotting the observed and expected quantiles can be used to display deviations from the theoretical distribution. When comparing a distribution against $U(0, 1)$, a P-P plot is equivalent to a Q-Q plot (Wilk and Gnanadesikan, 1968).

When the observed values are consistently greater than the expected values, this suggests that the test statistic is conservative, it is failing to reject the null hypothesis at least as frequently as would be dictated by the nominal significance level. When the observed values are consistently less than the expected values, this suggests that the test statistic is liberal, it is rejecting the null hypothesis more frequently than would be dictated by the nominal significance level.

Given a nominal Type I error rate of $\alpha = 0.05$, a valid test should incorrectly reject the null hypothesis approximately 5% of the time. Bradley (1978) noted that in research reporting to assess ‘robustness’, there is little quantitative indication by authors what is meant by ‘robustness’. When the null hypothesis is true, Bradley’s robustness criteria states that the observed NHRR is Type I error robust if it is within $x\%$ of α, where $x = 50$ for Bradley’s liberal criteria, $x = 20$ for Bradley’s moderate criteria, and $x = 10$ for Bradley’s stringent criteria. For $\alpha = 0.05$, Sullivan and D’Agostino (1996) state that Type I error rates ≤ 0.055 are acceptable. Similarly, Guo and Luh (2000) state that Type I error rates ≤ 0.075 are acceptable. The proposals by Sullivan and D’Agostino (1996) and Guo and Luh (2000) are modifications to Bradley’s stringent and liberal criteria respectively, but only flag as a concern the Type I error rates above α. These modifications to Bradley's robustness criteria take into account that a conservative test statistic is of less concern than a liberal test statistic.

Bradley's robustness criteria does not specify what constitutes acceptable variability in Type I error robustness across multiple parameter combinations. Some subjective decisions therefore remain when some parameter combinations fail Bradley's liberal Type I error robustness criteria.

An alternative proposal by Derrick and White (2018) is to quantify robustness as $(1 - \pi)\%$ of Type I error rates within $\pi\%$ of α. For example a robustness score of 94% would mean that the calculated Type I error rates are within 6% of α.

Some authors simply report the difference between the nominal significance level and the observed Type I error rate, with brief reference to Bradley's robustness criteria e.g. Fagerland, Sandvik, and Mowinckel (2011). Other methods for quantifying robustness is to report the confidence interval coverage of the true difference (Fagerland, 2012). The latter is only practical for parametric tests.

Bradley's stringent criteria is more demanding than researchers are willing to use (Serlin, 2000). Bradley's liberal criteria has been used in many studies analysing the validity of t-tests and their adaptations. Examples of this approach from the literature in Chapter 2 include Fradette et al. (2003), Nguyen et al. (2012), and Kellermann et al. (2013).

For consistency, Bradley's liberal robustness criteria is used throughout this thesis. Test statistics that perform consistently within the interval are recommended for practical use. Test statistics are first assessed for Type I error robustness. Only test statistics that demonstrate liberal Type I error robustness are then assessed for power (Derrick, 2017b).

All simulations herein are performed in R, various versions (R Core Team, 2019), using the R Studio interface (R Studio Team, 2019). Extant t-tests are calculated using the 'stats' package. Degrees of freedom are used in calculations without rounding, and reported to 3 decimal places in the text. The Wilcoxon test is calculated using the Normal approximation corrected for ties with continuity correction factor using the 'stats' package. Pratt's

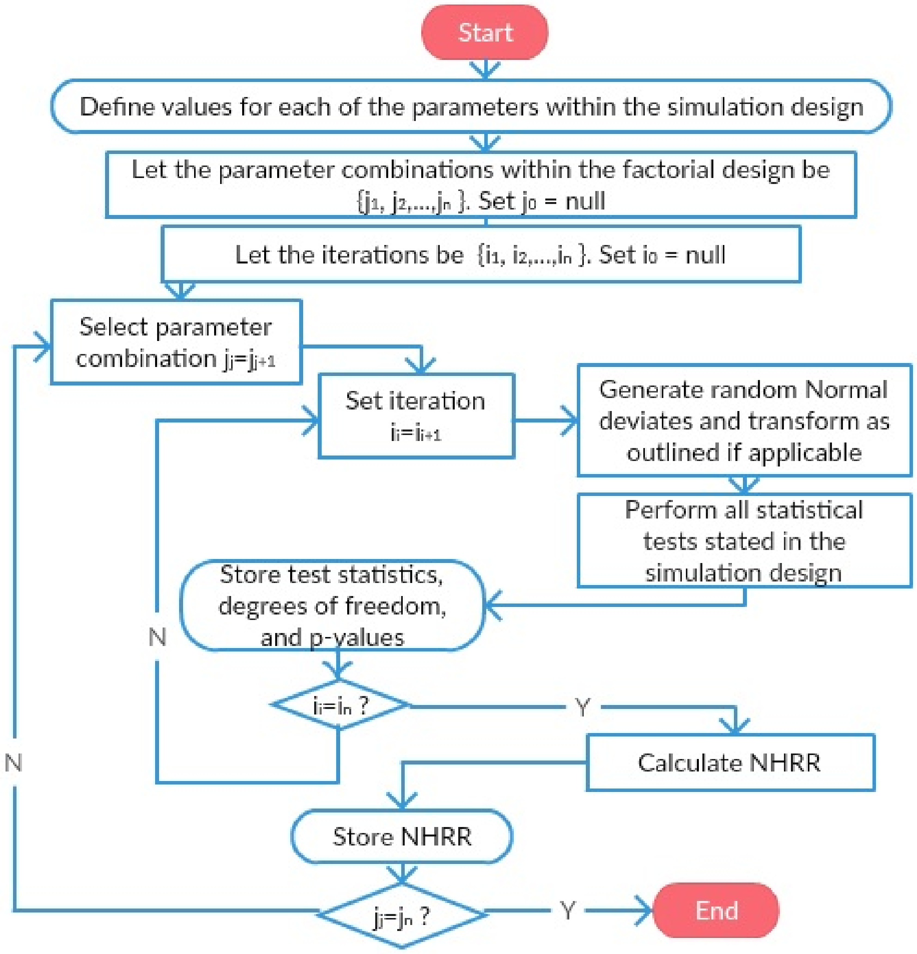

Start
Define values for each of the parameters within the simulation design
Let the parameter combinations within the factorial design be {j1, j2,...,jn }. Set j0 = null
Let the iterations be {i1, i2,...,in }. Set i0 = null
Select parameter combination jj=ji+1
Set iteration ii=ii+1
Generate random Normal deviates and transform as outlined if applicable
Perform all statistical tests stated in the simulation design
Store test statistics, degrees of freedom, and p-values
N
ii=in ?
Y
Calculate NHRR
N
Store NHRR
jj=jn ?
Y
End

Figure 4.6: Simulation methodology

test is calculated under the same conditions using the ‘coin’ package. For the mixed model approach utilising REML, the package ‘lme4’ is used and corresponding p-values are calculated using the Satterthwaite approximation adopted by SAS using the package ‘lmerTest’ (Barr et al., 1979). The partially overlapping samples t-tests in Chapters 3.1.1 and 3.1.2 are calculated using the ‘partiallyoverlapping’ package (Derrick, 2017a).

Chapter 5

The comparison of means, for normally distributed data

The Type I error robustness and power of the partially overlapping samples t-tests are investigated under normality. These test statistics are compared against standard tests that discard data, the Looney and Jones (2003) procedure and the REML procedure. Results within are summarised in Derrick et al. (2017a). An R package to facilitate application of the proposed tests is documented. The chapter concludes with an assessment of the partially overlapping samples t-tests in scenarios where elements of the test are at their extremes.

5.1 Simulation parameters and test statistics

Overall, the comparison of means for partially overlapping samples 'has been poorly treated in the literature' (Martinez-Camblor, Corral, and De La Hera, 2013, p.77). At the genesis of this thesis, and when preparing a simulation design to compare proposed test statistics with extant approaches, the $Z_{\text{corrected}}$ test statistic by Looney and Jones (2003) is the main competitor to naive tests that discard data or performing the independent samples t-test on all of the available data ignoring pairing (Samawi and Vogel, 2014a).

The simulation methodology outlined in Chapter 4, is used to assess the

Type I error robustness and power of the partially overlapping samples t-tests, namely T_{new1} and T_{new2} as defined in Chapter 3.1.1 and 3.1.2 respectively. These are compared against standard tests which discard observations; namely T_1, T_2, and T_3. Approaches that ignore any pairing and use all of the available data are also investigated, T_2^{all} and T_3^{all}. Techniques using all of the available data, The $Z_{\text{corrected}}$ statistic by Looney and Jones (2003) and the REML procedure outlined in Chapter 1, are also included in the comparison. The parameters used within the simulation design are given in Table 5.1.

Table 5.1: Simulation parameters.

Parameter	**Values**
μ_1	0
μ_2	0 (under H_0); 0.25, 0.50, 0.75, 1.00, 1.25, 1.50 (under H_1)
σ_1^2	1, 2, 4, 8
σ_2^2	1, 2, 4, 8
n_a	5, 10, 30, 50, 100, 500
n_b	5, 10, 30, 50, 100, 500
n_c	5, 10, 30, 50, 100, 500
ρ	-0.75, -0.50, -0.25, 0.00, 0.25, 0.50, 0.75

5.2 Type I error rates

Type I error robustness is firstly assessed under the condition of equal variances. Under H_0, 10,000 replicates are obtained for the 4 × 6 × 6 × 6 × 7 = 6,048 scenarios where $\sigma_1^2 = \sigma_2^2$. Figure 5.1 exhibits the Type I error rates for each of the test statistics under equal variances. Each point represents one parameter combination with the simulation design, reference lines for Bradley's liberal Type I error robustness criteria are included.

Figure 5.1 indicates that when $\sigma_1^2 = \sigma_2^2$, the statistics T_1, T_2, T_3, T_{new1} and T_{new2} remain within Bradley's liberal Type I error robustness criteria throughout the entire simulation design. T_2^{all} and T_3^{all} are not Type I error robust. This finding is compatible with the findings of Zimmerman (1997) when ignoring the pairing in a paired samples design. The statistic $Z_{\text{corrected}}$

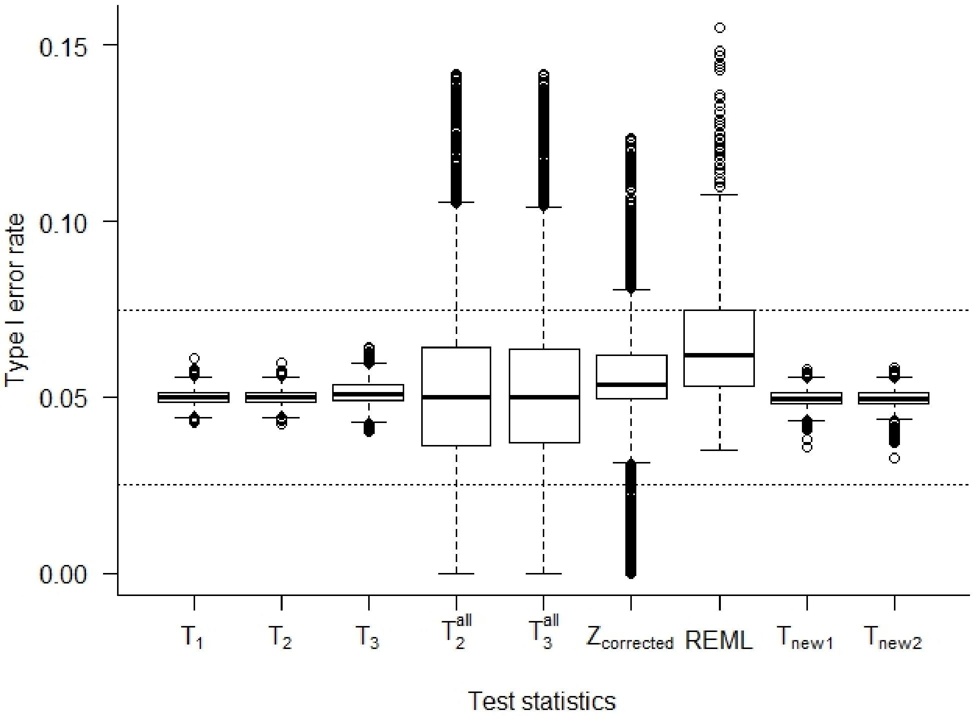


Figure 5.1: Type I error rates where $\sigma_1^2 = \sigma_2^2$, reference lines show Bradley's liberal criteria.

is not Type I error robust, confirming smaller scale simulation findings by Mehrotra (2004). Figure 5.1 also shows that REML is not Type I error robust throughout the entire simulation design.

Type I error robustness is assessed under the condition of unequal variances. Under the null hypothesis, 10,000 replicates were obtained for the $4 \times 3 \times 6 \times 6 \times 6 \times 7 = 18{,}144$ scenarios where $\sigma_1^2 \neq \sigma_2^2$. For assessment against Bradley's liberal robustness criteria, Figure 5.2 shows the Type I error rates.

It can be seen from Figure 5.2 that the statistics defined using a pooled standard deviation, T_2 and T_{new1}, do not provide Type I error robust solutions when variances are not equal. The statistics T_1, T_3 and T_{new2} retain their Type I error robustness under unequal variances and normality throughout

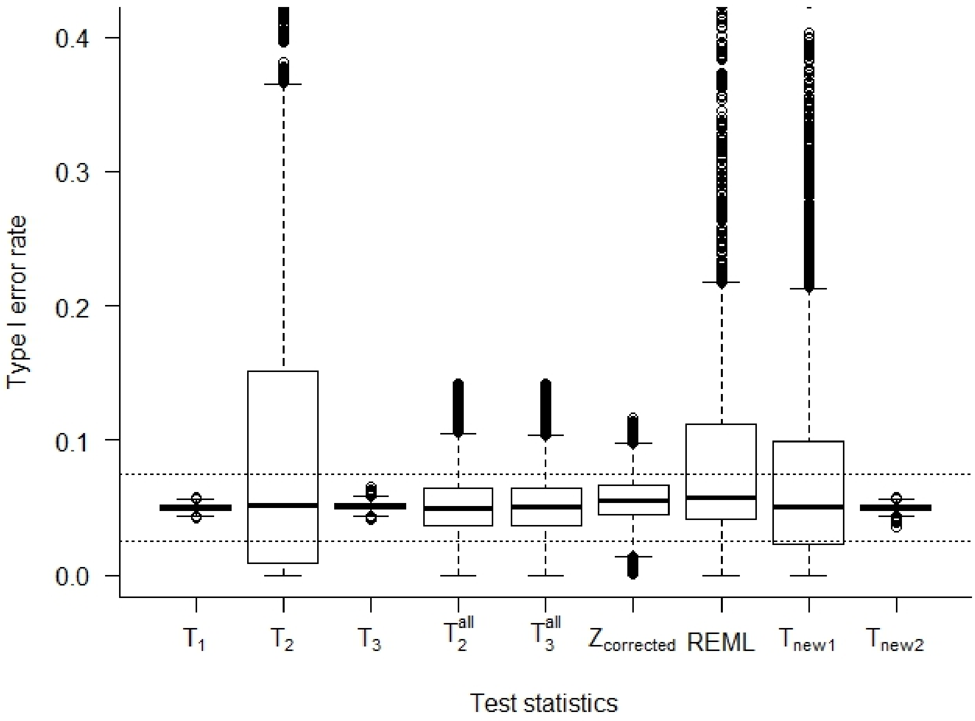


Figure 5.2: Type I error rates where $\sigma_1^2 \neq \sigma_2^2$, reference lines show Bradley's liberal criteria.

all conditions simulated.

The statistic $Z_{\text{corrected}}$ exhibits similar Type I error rates under unequal variances as it does when variances are equal. The statistic $Z_{\text{corrected}}$ results in an unacceptable amount of false positives when $\rho \leq 0.25$ or max $\{n_a,\, n_b,\, n_c\}$ - min $\{n_a,\, n_b,\, n_c\}$ is large. In addition, the statistic $Z_{\text{corrected}}$ is conservative when ρ is large and positive. The largest observed deviations from Type I error robustness for REML are when $\rho \leq 0$ or max $\{n_a,\, n_b,\, n_c\}$ - min $\{n_a,\, n_b,\, n_c\}$ is large. Further insight to the Type I error rates for REML can be seen in Figure 5.3, showing observed p-values against expected p-values from a uniform distribution.

It can be seen from Figure 5.3 that REML is not Type I error robust

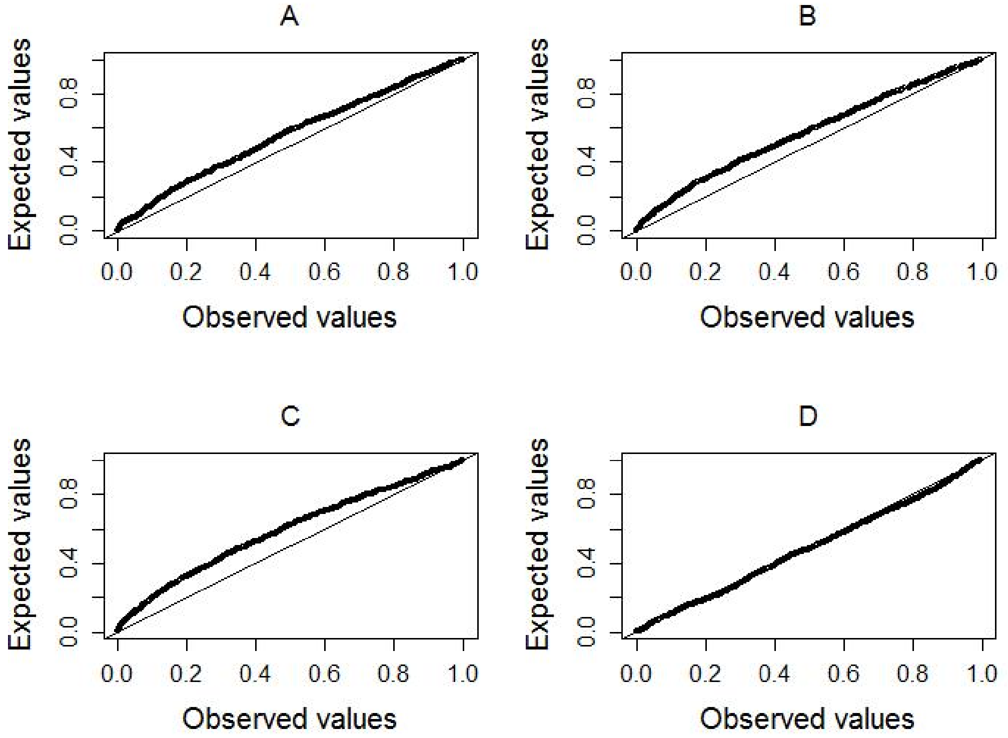


Figure 5.3: P-P plots for 10,000 simulated p-values using REML procedure. Selected parameter combinations (n_a, n_b, n_c, σ_1^2, σ_2^2, ρ) are as follows; A (5,5,5,1,1,-0.75), B (5,10,5,8,1,0), C (5,10,5,8,1,0.5), D (10,5,5,8,1,0.5).

when the correlation coefficient is negative. In addition, caution should be exercised if using REML when the larger variance is associated with the smaller sample size. REML maintains Type I error robustness for positive correlation and equal variances or when the larger sample size is associated with the larger variance.

In contrast, Figure 5.4 shows that T_{new2} maintains uniform p-values across the same set of scenarios. This indicates that although the 5% significance level has been used as standard, the test statistic T_{new2} remains valid at any significance level.

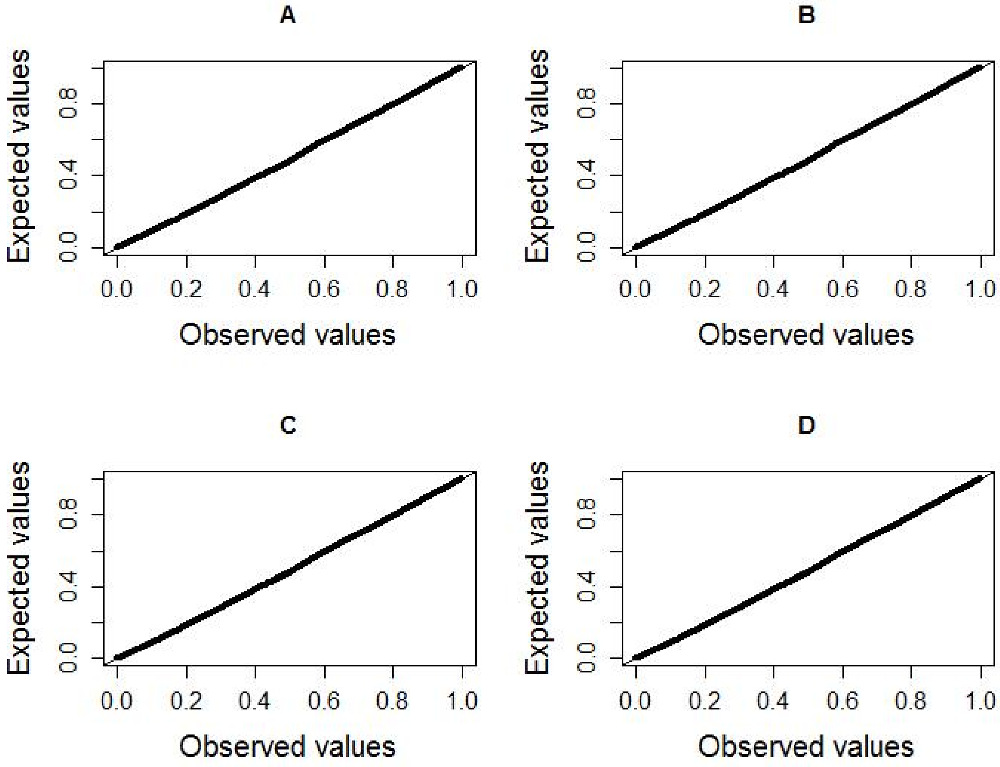


Figure 5.4: P-P plots for 10,000 simulated p-values using T_{new2}. Selected parameter combinations (n_a, n_b, n_c, σ_1^2, σ_2^2, ρ) are as follows; A (5,5,5,1,1,-0.75), B (5,10,5,8,1,0), C (5,10,5,8,1,0.5), D (10,5,5,8,1,0.5).

5.3 Power

Test statistics that do not fail to maintain Bradley's Type I error liberal robustness criteria are assessed under H_1. REML is included in the comparisons for $\rho \geq 0$. The power of the test statistics are assessed where $\sigma_1^2 = \sigma_2^2 = 1$, followed by an assessment of the power of the test statistics where $\sigma_1^2 > 1$ and $\sigma_2^2 = 1$.

Table 5.2 shows the power of T_1, T_2, T_3, T_{new1}, T_{new2} and REML, averaged over all sample size combinations where $\sigma_1^2 = \sigma_2^2 = 1$ and μ_1 - $\mu_2 = 0.5$.

Table 5.2 shows that REML, T_{new1} and T_{new2} are more powerful than naive approaches, T_1, T_2 and T_3, when variances are equal. Consistent with the paired samples t-test, T_1, the power of T_{new1} and T_{new2} is relatively lower when

Table 5.2: Power of Type I error robust test statistics, $\sigma_1^2 = \sigma_2^2$, μ_1 - μ_2 = 0.5.

	ρ	T_1	T_2	T_3	T_{new1}	T_{new2}	REML
$n_a = n_b$	0.75	0.785	0.567	0.565	0.887	0.886	0.922
	0.50	0.687	0.567	0.565	0.865	0.864	0.880
	0.25	0.614	0.567	0.565	0.842	0.841	0.851
	0	0.558	0.567	0.565	0.818	0.818	0.829
	<0	0.481	0.567	0.565	0.778	0.778	-
$n_a \neq n_b$	0.75	0.784	0.455	0.433	0.855	0.847	0.907
	0.50	0.687	0.455	0.433	0.840	0.832	0.861
	0.25	0.615	0.455	0.433	0.823	0.816	0.832
	0	0.559	0.455	0.433	0.806	0.799	0.816
	<0	0.482	0.455	0.433	0.774	0.766	-

there is zero or negative correlation between the two populations. Similar to contrasts of the independent samples t-test, T_2, with Welch's test, T_3, for equal variances but unequal sample sizes, T_{new1} is marginally more powerful than T_{new2}, but not to any practical extent. For each of the tests statistics making use of paired data the power increases as the correlation between the paired samples increases.

To investigate further the relationship and differences between T_{new1} and T_{new2}, Figure 5.5 depicts a scatterplot of the the p-values for the two tests, and a Bland-Altman plot of the mean and differences between the p-values for the two tests, where $\sigma_1^2 = \sigma_2^2 = 1$.

In Figure 5.5 there is an apparent very strong correlation between the two partially overlapping samples t-tests (Pearson's $r = 0.993$). Using correlation as a crude measure of agreement can mask some interesting differences (Bland and Altman, 1986). The Bland-Altman plot suggests that for the scenarios where the power may be low, the power gain of T_{new1} over T_{new2} is greater. In this simulation design these scenarios occur when max $\{n_a, n_b, n_c\}$ - min $\{n_a, n_b, n_c\}$ is large.

As the correlation between the paired observations increases, the power advantage of the proposed test statistics relative to the paired samples t-test becomes smaller. Therefore the proposed statistics T_{new1} and T_{new2} may be

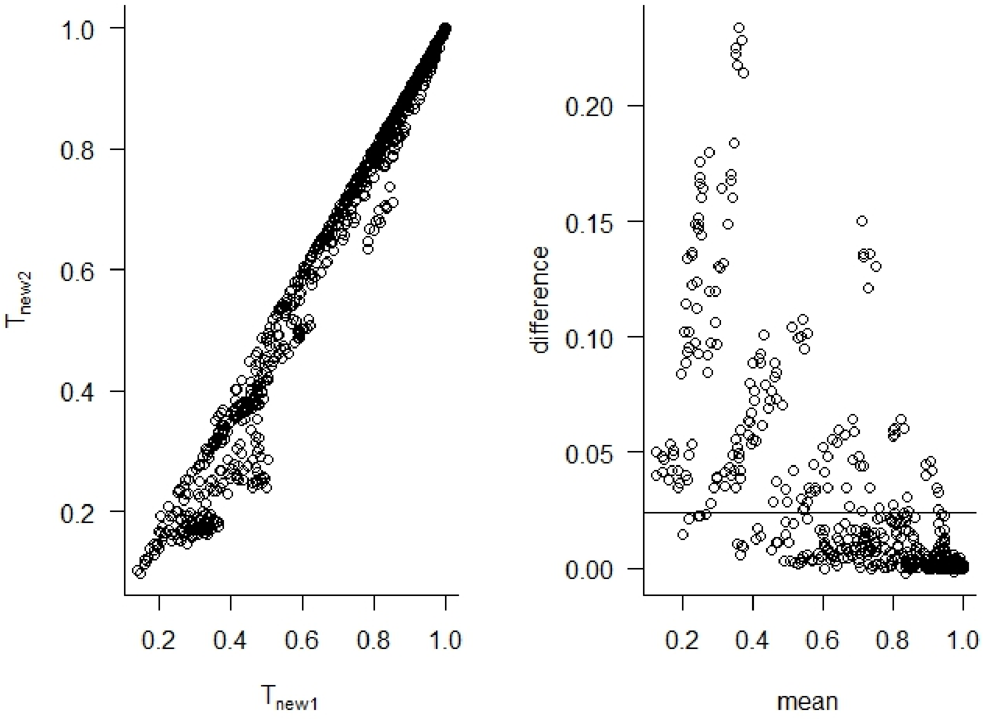


Figure 5.5: Relationship and differences between T_{new1} and T_{new2}, p-values where $\sigma_1^2 = \sigma_2^2 = 1$ and μ_1 - $\mu_2 = 0.5$. A reference line for the average difference in p-values between the two tests (0.024) is included

especially useful when the correlation between the two populations is small.

Figure 5.6 gives the power curves for the partially overlapping samples t-test, and for comparative purposes the paired samples t-test, for a total sample size of 30 observations. As anticipated the power increases as the true difference in population means increases. It is of note that the relative gain in power using the partially overlapping samples t-test over the paired samples t-test is related to the difference in population means.

Test statistics that do not violate Bradley's liberal robustness criteria when $\sigma_1^2 \neq \sigma_2^2$ are assessed for power. Table 5.3 gives power averaged over the simulation design for these parameter combinations.

Table 5.3 shows that T_{new2} has superior power properties to both T_1 and

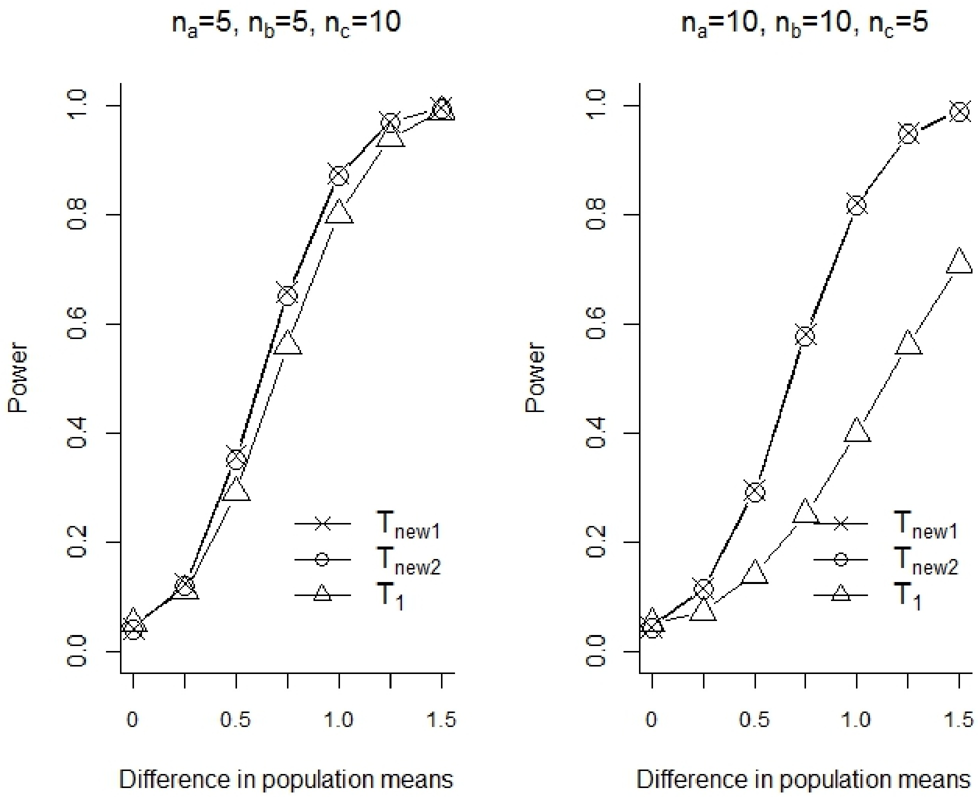


Figure 5.6: Power curves averaged over all values of ρ, $\sigma_1 = \sigma_2 = 1$. The left hand side has a greater number of paired observations, the right hand side has a greater number of independent observations.

T_2 when $\sigma_1^2 \neq \sigma_2^2$. In common with the performance of Welch's test for independent samples, T_3, the power of T_{new2} is higher when the larger variance is associated with the larger sample size. In common with the performance of the paired samples t-test, T_1, the power of T_{new2} is relatively lower when there is zero or negative correlation between the two populations.

The apparent power gain for REML when the larger variance is associated with the larger sample size, can be explained by the pattern in the Type I error rates. REML follows a similar pattern to the independent samples t-test, which is liberal when the larger variance is associated with the larger sample size, thus giving the perception of higher power.

To show the relative increase in power for varying sample sizes, Figure 5.7

Table 5.3: Power of Type I error robust test statistics, $\sigma_1^2 \neq \sigma_2^2$, μ_1 - μ_2 = 0.5.

	ρ	T_1	T_3	T_{new2}	REML
$n_a = n_b$	0.75	0.555	0.393	0.692	0.645
	0.50	0.481	0.393	0.665	0.588
	0.25	0.429	0.393	0.640	0.545
	0	0.391	0.393	0.619	0.515
	<0	0.341	0.393	0.582	-
$n_a > n_b$	0.75	0.555	0.351	0.715	0.589
	0.50	0.481	0.351	0.688	0.508
	0.25	0.429	0.351	0.665	0.459
	0	0.391	0.351	0.642	0.422
	<0	0.341	0.351	0.604	-
$n_a < n_b$	0.75	0.555	0.213	0.559	0.693
	0.5	0.481	0.213	0.539	0.649
	0.25	0.429	0.213	0.522	0.62
	0	0.391	0.213	0.507	0.603
	<0	0.341	0.213	0.480	-

shows the power for selected test statistics for small-medium sample sizes, averaged across the simulation design for unequal variances.

Figure 5.7 shows a relative power advantage when the larger variance is associated with the larger sample size, as per B_2 and D_2. Across the simulation design, power is adversely affected for all test statistics when variances are not equal. This is exacerbated for small-medium sample sizes.

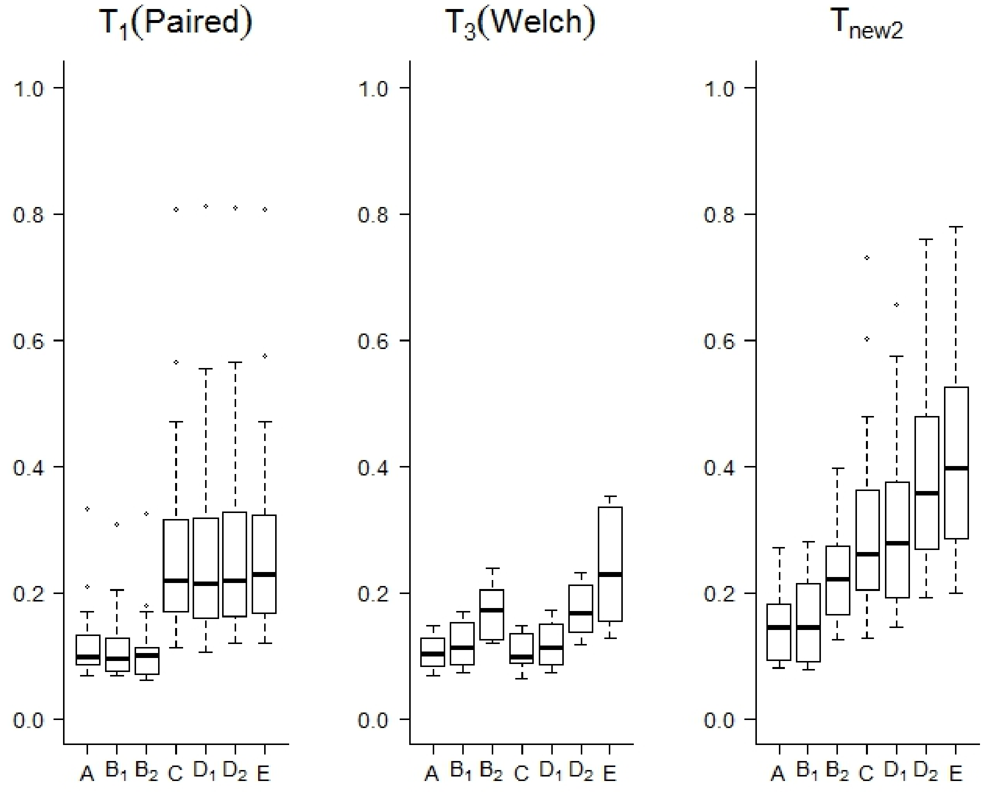


Figure 5.7: Power for Type I error robust test statistics, $\sigma_1^2 > \sigma_2^2$ and $\mu_1 - \mu_2 = 0.5$. The sample sizes (n_a, n_b, n_c) are as follows; A (10,10,10), B_1 (10,30,10), B_2 (30,10,10), C (10,10,30), D_1 (10,30,30), D_2 (30,10,30), E (30,30,30).

5.4 R package

To provide ease of calculation of T_{new1} and T_{new2}, an R package 'partiallyoverlapping' (Derrick, 2017a) is supplied, version control as per Table 5.4.

Table 5.4: Version control, 'partiallyoverlapping'.

Version	Date	Notes
1.0	1/1/2017	Partover.test introduced
1.1	11/11/2018	command mu= added to Partover.test
2.0	12/12/2018	Prop.test added (see Chapter 9)

To ease facilitation, the structure of the Partover.test function is the same as the ‘t.test’ function in the ‘stats’ package.

5.4.1 Help manual

Extracts from the supporting help pages are given below to demonstrate how the function Partover.test within the ‘partiallyoverlapping’ package is used.

Description

The formula is only applicable for the two sample partially overlapping samples *t*-test. The number of unpaired observations may be zero for up to one of the two samples. The number of paired observations must be of equal length of two or greater. Error messages are given when these conditions are not true.

Performs a comparison of means using the partially overlapping *t*-test, for two samples each with paired and unpaired observations. This functions calculates the test statistic, the degrees of freedom, and the *p*-value. Additionally calculates a confidence interval for the difference in means when requested. By default, four vectors are to be specified: unpaired observations in Sample 1, unpaired observations in Sample 2, paired observations in Sample 1, paired observations in Sample 2. If the structure of your data is of two vectors, one for each sample, then the option stacked = TRUE can be specified.

Usage and default options

Partover.test(x1 = NULL, x2 = NULL, x3 = NULL, x4 = NULL, var.equal = FALSE, mu = 0, alternative = "two.sided", conf.level = NULL, stacked = FALSE)

Arguments

x1 a vector of unpaired observations in Sample 1 (or all observations in Sample 1 if stacked = "TRUE")

x2 a vector of unpaired observations in Sample 2 (or all observations in Sample 2 if stacked = "TRUE")

x3 a vector of paired observations in Sample 1 (not applicable if stacked = "TRUE")

x4 a vector of paired observations in Sample 2 (not applicable if stacked = "TRUE")

var.equal a logical variable indicating whether to treat the two variances as being equal. If "TRUE" then the pooled variance is used to estimate the variance, otherwise the Welch approximation to the degrees of freedom is used.

mu difference in population means under the null hypothesis

alternative a character string specifying the alternative hypothesis, must be one of "two.sided" (default), "greater" or "less".

conf.level confidence level of the interval.

stacked indicator of whether paired and unpaired observations are stacked within one vector ("TRUE"), or if specified as four separate vectors (default). Corresponding pairs should be given on the same row when "TRUE" is selected.

Values

statistic The value of the *t*-statistic parameter The degrees of freedom for the test statistic

p.value The *p*-value for the test

estimate The estimated difference in the means

conf.int A confidence interval for the mean appropriate to the specified alternative hypothesis

Example
[This is Example 1 in Section 3.2, taken from Derrick, Toher, and White (2017)].
The sample means for two groups, "a" and "b" are compared for a two sided test assuming equal variances.
Approach 1: For each sample, unpaired observations and paired observations defined as separate vectors:

```
a.unpaired<-c(20,21,16,18,14,12,14,17)

a.paired<-c(14,15,18,20,11,19,14,15)

b.unpaired<-c(10,16,18,16,15,14,13,10)

b.paired<-c(15,10,15,17,13,19,12,13)

Partover.test(a.unpaired,b.unpaired,a.paired,b.paired,
            var.equal=TRUE)
```

Resulting output gives; p.value = 0.026.

Equivalently, Approach 2: Independent observations and the paired samples stacked for each sample:

```
a<-c(20,21,16,18,14,12,14,17,NA,NA,NA,NA,NA,NA,NA,NA,
     14,15,18,20,11,19,14,15)

b<-c(NA,NA,NA,NA,NA,NA,NA,NA,10,16,18,16,15,14,13,10,
     15,10,15,17,13,19,12,13)

Partover.test(a,b,var.equal=TRUE,stacked=TRUE)
```

Resulting output gives; p.value = 0.026, the samples from group "a" and group "b" have significantly different means.

5.4.2 Further application

In an application by Polster et al. (2019), irritable bowel syndrome sufferers in two cohorts are compared. Some respondents are within the ‘Rome III cohort’, some are in the ‘Rome IV cohort’ and some are in both. Their severity of symptoms as reported via a questionnaire is compared between the two cohorts using the ‘Partover.test’. The form of the test, equal variances assumed or not assumed, is not reported. All of the authors significant results have $p < 0.01$ suggesting that both tests will give the same conclusion. The reported variances in each group appear approximately equal, and thus either test selected would be a robust option.

In an application by Raymundo et al. (2019), the amount of coral reef cover following bleaching events in 2013 and 2017 is compared. Normality was considered using the Shapiro-Wilk test and equal variances was considered using the Brown-Forsythe test. Upon satisfying these two assumptions T_{new1} is performed, it is concluded that there is a difference in the impact of the bleaching events between the two years ($t(39.2) = 2.61, p = 0.013$).

5.5 The performance of the partially overlapping samples t-tests at the limits.

It is of interest whether the partially overlapping samples t-tests remain valid for circumstances which include parameters of the test at their limits. The following extreme or unusual conditions are explored:

1. $n_a = 0$ and $n_b = 0$
2. $n_a = 0$ or $n_b = 0$
3. $n_c = 0$ or $n_c = 1$
4. $\rho = 1$ or $\rho =$ -1
5. $\rho = 1$ because the paired observations are identical
6. $\sigma^2 = 0$
7. $H_0 : \mu_1 - \mu_2 = x$

Particular attention is given to the validity of the 'Partover.test' function in the R package 'partiallyoverlapping' by Derrick (2017a). Where appropriate, additional simulations are performed to assess the Type I error rate under theses conditions.

1. $n_a = 0$ **and** $n_b = 0$

If there are no independent observations, both forms of the partially overlapping samples t-test are equivalent to the paired samples t-test. The 'Partover.test' function in R can be performed and gives equivalent results to the paired samples t-test.

2. $n_a = 0$ **or** $n_b = 0$

If paired observations are present and only one sample has independent observations, this is equivalent to scenario (8) in Chapter 1, and equivalent to that outlined by Qi, Yan, and Tian (2018) where incompleteness is in a single response only. The partially overlapping samples t-tests have the advantage that this is not a restriction to performing the 'Partover.test' test.

The simulation design as per Table 5.1 is repeated under H_0 with a fixed value of $n_a = 0$. For both forms of the partially overlapping samples t-test, the proportion of parameter combinations that satisfy Bradley's liberal Type I error robustness criteria is given in Figure 5.8. Parameter combinations with a Type I error rate in excess of 0.075 are deemed liberal, and those parameter combinations with a Type I error rate less than of 0.025 are deemed conservative. See Table 5.5 for assessment of Type I error robustness for a small selection of parameter combinations under these conditions, with values that fulfill Bradley's liberal Type I error robustness criteria highlighted in bold.

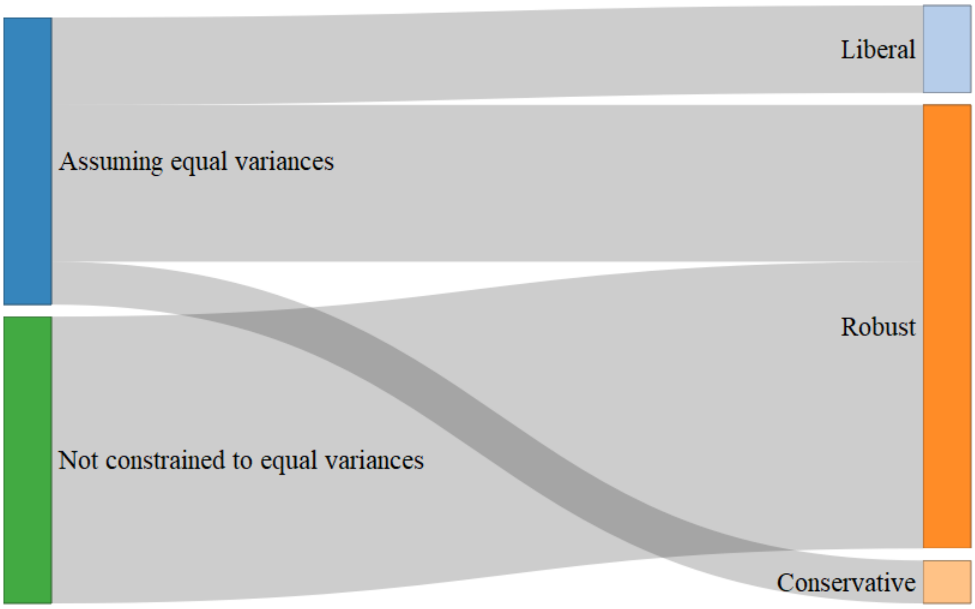


Figure 5.8: Sankey plot for extended simulation design with $n_a = 0$.

Table 5.5: Type I error rates, extended design, $n_a = 0$.

ρ	n_a	n_b	n_c	σ_1^2	σ_2^2	T_{new1}	T_{new2}
0.25	0	30	5	1	1	**0.045**	**0.053**
0.75	0	5	5	1	1	**0.041**	**0.040**
0.25	0	30	5	4	1	0.222	**0.053**
0.75	0	5	5	4	1	0.124	**0.048**
0.25	0	30	5	1	4	0.001	**0.051**
0.75	0	5	5	1	4	**0.032**	**0.041**

3. $n_c = 0$ **or** $n_c = 1$

If there are no paired samples, T_{new1} is mathematically equivalent to the independent samples t-test, and T_{new2} is mathematically equivalent to Welch's test. When using the 'Partover.test' in R, a minimum of $n_c = 2$ is required due to the required calculation of a correlation coefficient. If $n_c = 1$, it is a single pair that could be discarded and the independent samples t-test or Welch's test could be performed without greatly impacting power. However, caution should be exercised if the discarded pairing contains an extreme observation [see Chapter 7].

4. $\rho = 1$ **or** $\rho = -1$

It is theoretically possible to have perfect correlation, without the paired observations being identical.

The simulation design in Table 5.1 is again repeated under H_0, but with a fixed parameter of $\rho = 1$.

For both forms of the partially overlapping samples t-test, the proportion of parameter combinations that satisfy Bradley's liberal Type I error robustness criteria is given in Figure 5.9.

The similarity of Figure 5.8 and Figure 5.9 indicates that the partially overlapping samples t-tests exhibit similar robustness properties at the extremes of $n_a = 0$ or $\rho = 1$, and these Type I error rates are in agreement with the Type I error robustness observed for the original simulation design.

Selected results from this extension to the simulation design for selected parameter combinations with $\rho = 1$ are given in Table 5.6. This table includes

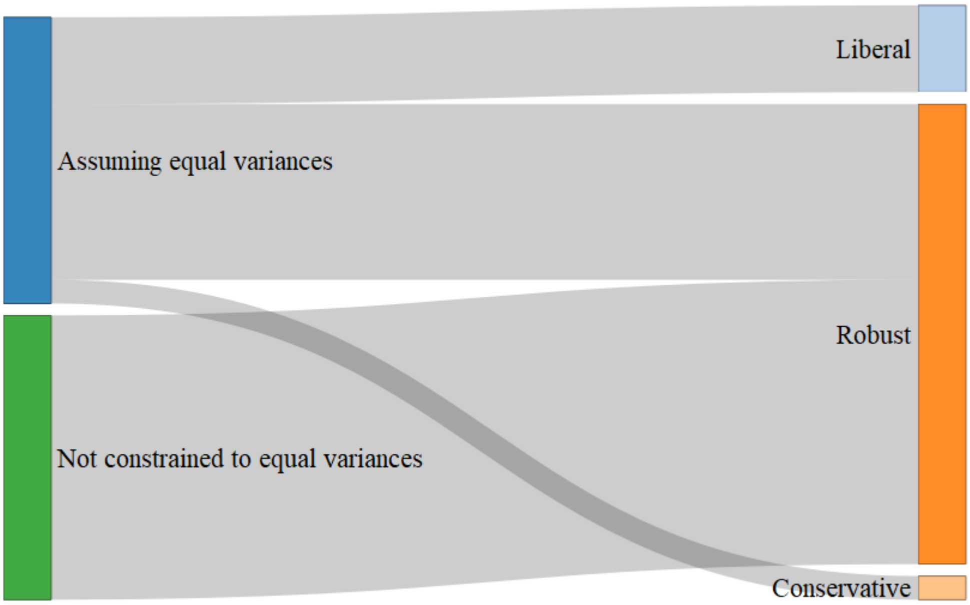


Figure 5.9: Sankey plot for extended simulation design with $\rho = 1$.

results of some additional simulations where both $n_a = 0$ and $\rho = 1$.

Table 5.6: Type I error rates, extended design, $\rho = 1$.

ρ	n_a	n_b	n_c	σ_1^2	σ_2^2	T_{new1}	T_{new2}
1	5	30	5	1	1	**0.044**	**0.047**
1	30	5	5	1	1	**0.042**	**0.047**
1	5	30	5	4	1	**0.050**	**0.055**
1	30	5	5	4	1	**0.044**	**0.050**
1	5	30	5	1	4	0.003	**0.042**
1	30	5	5	1	4	0.185	**0.050**
1	0	30	5	1	1	**0.039**	**0.058**
1	0	5	5	1	1	0.019	**0.044**
1	0	30	5	4	1	**0.037**	**0.066**
1	0	5	5	4	1	0.019	**0.043**
1	0	30	5	1	4	0.264	**0.061**
1	0	5	5	1	4	0.180	**0.064**

It can be seen from Table 5.5 and Table 5.6 that when variances are equal, both partially overlapping samples t-tests remain valid when one sample has no independent observations. Additionally when relaxing the equal variances assumption, T_{new2} remains valid when the incompleteness is in a single sample. Perfectly correlated data does not detract from the validity of the tests.

5. $\rho = 1$ because the paired observations are identical

There may be occasions by design where two subsets of the same group are compared, and some units are common to both subsets. For example, in education where the mean module score for 'Statistical Modelling' is to be compared for two groups, those taking the optional module 'Mathematical Statistics' and those taking the optional module 'Operational Research'. Students taking both 'Mathematical Statistics' and 'Operational Research' could be said to be 'paired' observations. For the students taking both optional modules, the score for the Statistical Modelling module has $\rho = 1$. The 'paired' observations could be discarded and the independent samples t-test, T_2, or Welch's test, T_3, could be performed. Instead, the situation could be viewed as a one-way ANOVA with three groups, consisting of the two sets of independent observations, and the one set of paired observations. Alternatively the partially overlapping samples t-tests could be applied.

To assess whether the proposed test statistics are valid under these conditions, an additional consideration is required in the simulation design with respect to the variance. The variance of the paired observations, σ_c^2, could be equal to the variance of the observations in Group 1, σ_a^2, or the variance of the observations in Group 2, σ_b^2, or both, or neither. Table 5.7 give results of an extension to the simulation design to take into account these properties. Parameter combinations which fulfill Bradley's liberal Type I error robustness are highlighted in bold.

Table 5.7 shows that T_{new1} and T_{new2} are only valid if the variance in the paired observations is equal to the variance of both sets of independent observations. In general terms if the variance of the paired observations is not equal to the variance of the independent observations, it is likely that the observations are actually from separate populations, therefore a one way

Table 5.7: Type I error rates, $n_a = 5$ $\rho = 1$ and identical paired observations.

n_b	n_c	σ_a^2	σ_b^2	σ_c^2	T_{new1}	T_{new2}	T_1	T_2	ANOVA
5	5	1	1	1	**0.035**	**0.031**	**0.048**	**0.043**	**0.049**
30	5	1	1	1	**0.047**	**0.048**	**0.051**	**0.056**	**0.053**
5	30	1	1	1	**0.048**	**0.046**	**0.054**	**0.047**	**0.050**
5	5	1	4	1	0.075	**0.058**	**0.059**	**0.051**	**0.070**
5	5	1	1	4	0.004	0.003	**0.052**	**0.045**	**0.042**
5	5	1	4	4	0.021	0.015	**0.055**	**0.046**	**0.057**
30	5	1	4	1	0.009	**0.063**	0.004	**0.049**	0.092
5	30	1	4	1	0.000	0.000	**0.054**	**0.049**	**0.071**

ANOVA may be more appropriate. The partially overlapping samples t-test may be valid, assuming the exact form of the research question is taken into consideration when selecting an appropriate test.

6. $\sigma^2 = 0$

If the variability of the differences in the paired observations is equal to zero, the paired samples t-test cannot be performed. If there is no variability in the differences in the independent observations, the independent samples t-test or Welch's test cannot be performed. The partially overlapping samples t-tests remain functional in both these instances, so long as there is variability in either the independent observations or the paired observations.

It is possible that the paired observations within one sample could be constant (particularly for discrete data). This would give zero variability within the paired sample. Where the paired samples t-test in R would error, the 'Partover.test' function is set to give $r = 0$ in this circumstance so that the test can proceed. This is a valid approach because the partially overlapping samples t-test has been shown to be valid for $\rho = 0$. In this scenario the partially overlapping samples t-statistic is identical to the independent samples t-statistic performed on all of the available data, however the degrees of freedom differ because the partially overlapping samples t-tests incorporate the size of the paired sample. This means that the degrees of freedom are lower for the partially overlapping samples t-test than the independent samples t-test on all of the data. The partially overlapping samples t-test in this extreme scenario is therefore less powerful than performing the independent samples t-test on all of the available data. However, a researcher would not know in advance that the correlation is zero. If $\rho = 0$ is anticipated, a paired design may not be appropriate.

In the event of zero variability within both samples, 'Partover.test' is set to give p-value = 1 if $\bar{x}_1$ - $\bar{x}_2 = 0$, else 'Partover.test' gives p-value = 0.

7. $H_0 : \mu_1 - \mu_2 = x$

The above simulations are concerned with testing whether there is a difference between two groups, which is the same as testing if the difference between the two groups is zero. There may be occasions where researchers wish to test whether the difference between the two groups is equal to some other fixed value. Cao, Pauly, and Konietschke (2018) state Welch's test with the hypothesised difference in population means on the numerator. This extension to the numerator could be generalised to all forms of the t-test including the partially overlapping samples t-tests. Thus each t-test has the form:

$$T = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{stderr(\bar{X}_1 - \bar{X}_2)}$$

To demonstrate robustness of the t-tests when assessing against a null hypothesis of a defined difference x between the two populations, the simulation design in Table 5.1 is repeated for $H_0 : \mu_1 - \mu_2 = 10$. Correlated variates are generated per Chapter 4 with $\mu = 0$, then $x = 10$ is added to each variate in Group A. Type I error rates for selected parameter combinations, and averaged across the entire simulation design are given in Table 5.8.

Table 5.8 shows that the Type I error rates for both equal and unequal variances follow the same pattern as when testing for $H_0 : \mu_1 - \mu_2 = 0$ thus the conclusions regarding robustness for each test generalise to any $H_0 : \mu_1 - \mu_2 = x$

When performing the 'Partover.test', assessment against a $H_0 : \mu_1 - \mu_2 = x$ can be performed by adding the command 'mu = x'.

Table 5.8: Type I error rates, $H_0 : \mu_1 - \mu_2 = 10$, $\sigma_2^2 = 1$, $n_c = 5$.

ρ	n_a	n_b	σ_1^2	T_1	T_2	T_3	T_{new1}	T_{new2}
-0.75	5	10	1	**0.051**	**0.050**	**0.051**	**0.051**	**0.051**
-0.50	10	30	4	**0.051**	0.160	**0.053**	0.129	**0.048**
-0.25	30	5	1	**0.048**	**0.052**	**0.059**	**0.051**	**0.052**
0.00	5	5	4	**0.050**	**0.061**	**0.052**	**0.055**	**0.046**
0.25	10	5	1	**0.049**	**0.055**	**0.054**	**0.047**	**0.046**
0.50	30	10	4	**0.052**	0.009	**0.047**	0.013	**0.048**
0.75	5	30	1	**0.051**	**0.047**	**0.054**	**0.042**	**0.043**
Overall				**0.050**	0.101	**0.051**	0.079	**0.049**

5.6 Summary

The statistic T_{new2} is Type I error robust across all conditions simulated under normality and MCAR. The greater power observed for T_{new1} compared to T_{new2} under equal variances, is likely to be of negligible consequence in a practical environment. This is in line with empirical evidence for the performance of Welch's test, when only independent samples are present, which leads to many observers recommending the routine use of Welch's test under normality, e.g. Ruxton (2006).

The Type I error rates and power of T_{new2} follow the properties of its counterparts, T_1 and T_3. Thus T_{new2} can be seen as a trade-off between the paired sample *t*-test and Welch's test, with the advantage of increased power, due to using all available data.

A mixed model procedure using REML is not fully Type I error robust. In those scenarios in which this procedure is Type I error robust, the power is similar to that of T_{new1} and T_{new2}. The partially overlapping samples *t*-tests are less computationally intensive competitors to REML. The REML procedure does not directly calculate the difference between the two sample means, in a practical environment this makes its results hard to interpret.

In conclusion, for equal variances, T_{new1} and T_{new2} are Type I error robust. In addition they are more powerful than the traditional Type I error robust approaches. When variances are equal, there is a slight power advantage of using T_{new1} over T_{new2}, particularly when sample sizes are not equal. Under unequal variances, T_{new2} is the most powerful Type I error robust statistic considered.

When faced with a research problem involving two partially overlapping samples, if MCAR, normality and within sample independence can be reasonably assumed, the statistic T_{new1} can be used when variances are equal. Under the same conditions when equal variances cannot be assumed the statistic T_{new2} is recommended. The proposed test statistics for partially overlapping samples provide a competitive alternative method for analysis of normally distributed data. These methods remain valid when unpaired observations are in only one sample, and where there is perfect correlation.

Chapter 6

The comparison of two groups, for non-normally distributed data

The partially overlapping samples t-tests, and the additional proposed solutions in Chapter 3, are compared against existing non-parametric solutions. The comparison is extended to non-normal distributions. The findings presented within this chapter are published in Derrick, White, and Toher (2017) and Derrick, White, and Toher (in press).

6.1 Simulation parameters and test statistics

The simulation study in Chapter 5 is replicated to consider the performance of the test statistics for non-normal distributions. Inverse Normal Transformations (INTs) and non-parametric tests are also considered.

Type I error robustness and power are assessed for the partially overlapping samples t-test; namely T_{new1} and T_{new2} as defined in Chapter 3.1.1 and 3.1.2 respectively. These are compared against naive parametric and non-parametric tests which discard data; namely T_1, T_2, T_3, MW and W. Additional proposals that conceptually may be appropriate for non-normal data are also considered. These additional comparators are T_{RNK1}, T_{RNK2},

T_{INT1} and T_{INT2} as defined in Chapter 3.1.3 and Chapter 3.1.4.

Each of the test statistics are assessed under $N(0,1)$. In addition the test statistics are assessed under non-normality, for two samples from the Gumbel distribution, two samples from the Exponential distribution, and two samples from the Lognormal distribution. The data is generated as described in Chapter 4.1.1 and Chapter 4.1.2.

Fagerland and Sandvik (2009b) show that the Mann-Whitney test results in Type I errors that deviate from the nominal 5% significance level, if the two distributions do not have the same shape. Fagerland and Sandvik (2009b) note a common situation in medical research where the Mann-Whitney is incorrectly interpreted, a disparity in the variance and skewness between two distributions is often confounded with a difference in central location, so the assumption of a location shift model is unrealistic.

Additional analyses are performed when the samples are drawn from the Normal distribution with unequal variances, and then when samples are drawn from distributions with differing functional form, for example one sample taken from a Normal distribution and one sample taken from a Lognormal distribution. For assessing the Type I error robustness under normality with unequal variances, the n_1 observations are multiplied by σ_1 and the n_2 observations multiplied by σ_2. Standardising is performed when comparing samples from two distributions with differing functional form.

Type I error rates for each of the test statistics considered are reported, followed by power for each test statistic that controls Type I error rates. The scenario where samples are drawn from the same distribution is firstly considered (Section 6.2). This is followed by the scenario where samples are drawn from the Normal distribution with unequal variances (Section 6.3), and finally the scenario when the samples are drawn from distinctly differing distributions (Section 6.4).

6.2 Samples taken from distributions of the same shape

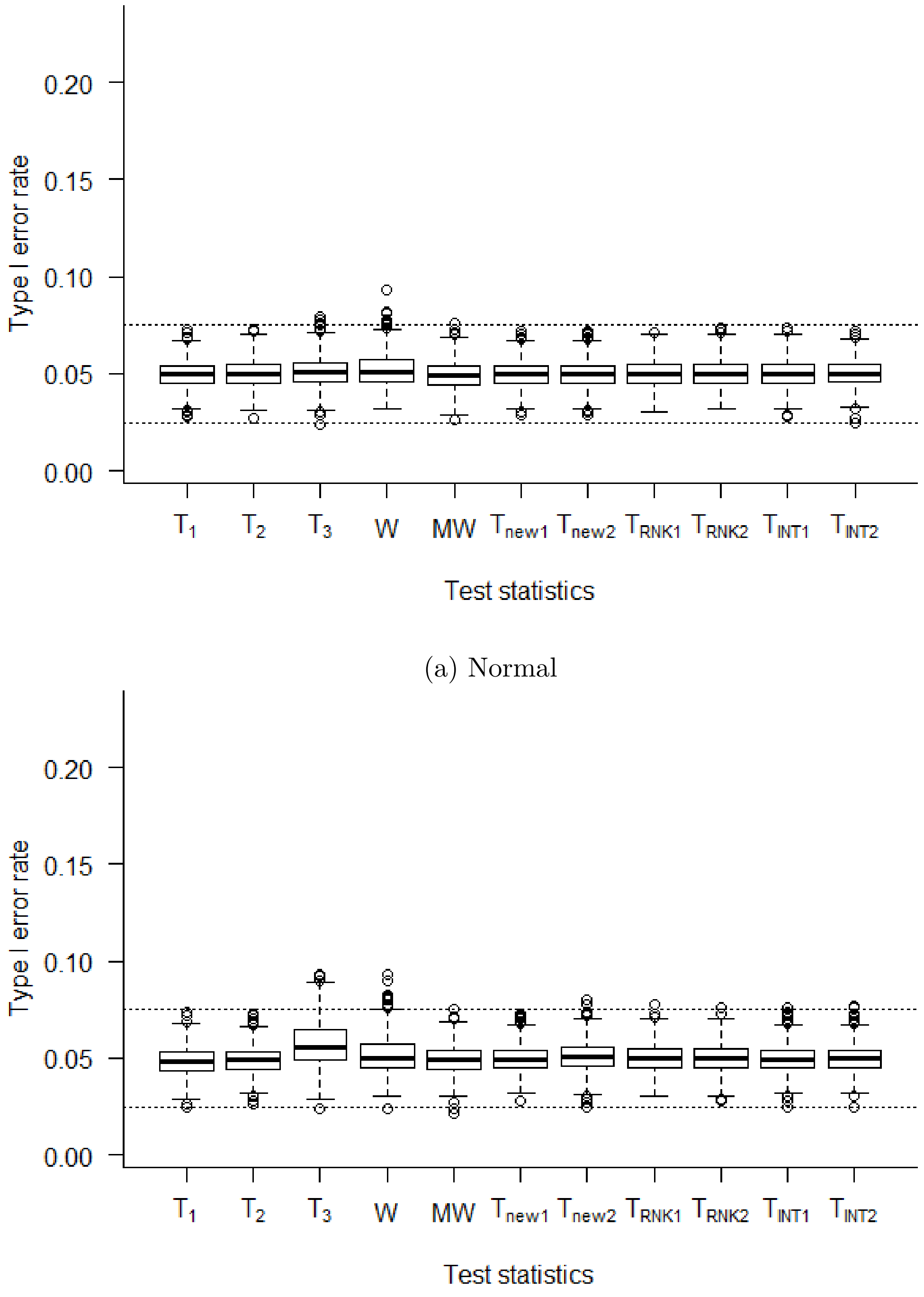
For parameter combinations where $\mu_1 = \mu_2$ and $\sigma_1^2 = \sigma_2^2$, the Type I error rates for each of the four distributions are summarised in Figure 6.1 and Figure 6.2. Each point represents one parameter combination within the simulation design, reference lines are for Bradley's liberal Type I error robustness criteria.

Figure 6.1(a) shows that each of the test statistics considered broadly maintain liberal Type I error robustness criteria when both samples are drawn from $N(0,1)$. However, there is some minor inflation of Type I error rates particularly for the Wilcoxon test based on only the paired observations, these occur predominantly when $n_c = 5$. Figure 6.1(b) suggests that the test statistics under consideration are not sensitive to relatively minor deviations from the Normal distribution. However, there is some minor inflation of Type I error rates particularly for Welch's test. The inflation of Type I error rates occurs when there is a sample size imbalance (e.g. $n_a = 500$, $n_b = 5$).

Figure 6.2 shows that for increasing skewness, the validity of some of the test statistics start to deteriorate. The degree of skewness for the Lognormal distribution in this simulation is larger than the degree of skewness considered by Fagerland and Sandvik (2009b). The Mann-Whitney test remains Type I error robustness even for the more extreme degree of skewness in this study. Most of the test statistics proposed also remain within Bradley's liberal Type I error robustness criteria. However, test statistics using separate variances, T_3 and T_{new2}, frequently exceed the upper threshold of Bradley's liberal Type I error robustness criteria.

For the statistics that use all of the available data, higher Type I error rates are associated with large sample size imbalances between n_a, n_b and n_c. For these statistics, lower Type I error rates are associated with small sample sizes and negative correlation.

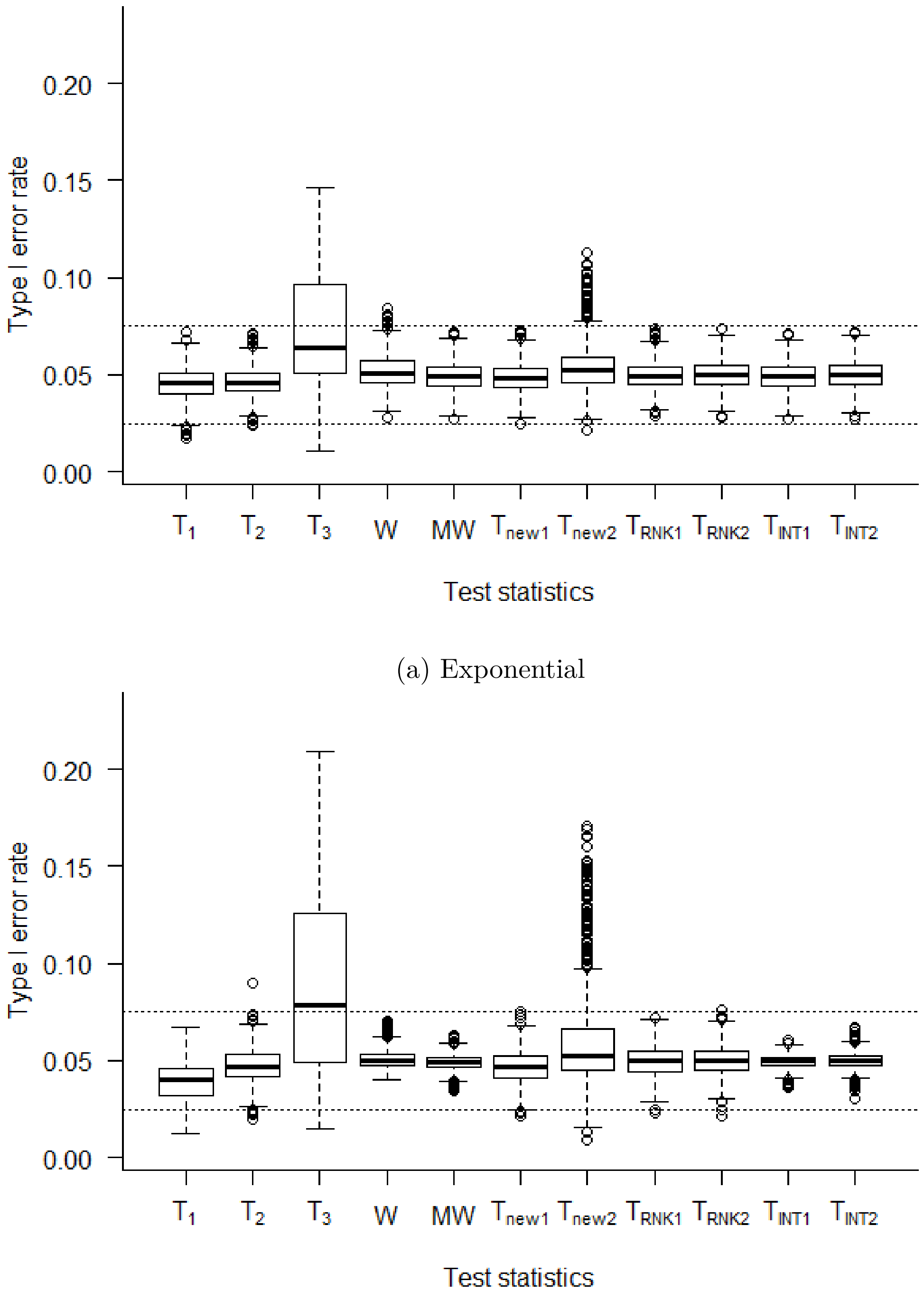
A summary of the power for the proposed test statistics is given in Table 6.1. Power is reported only for scenarios that exhibit Type I error robustness.



(a) Normal

(b) Gumbel

Figure 6.1: Type I error rates, Normal distribution and Gumbel distribution.



(a) Exponential

(b) Lognormal

Figure 6.2: Type I error rates, Exponential distribution and Lognormal distribution.

Table 6.1: Power, continuous data.

	ρ	T_{new1}	T_{new2}	T_{RNK1}	T_{RNK2}	T_{INT1}	T_{INT2}
Normal							
$n_1 = n_2$	>0	0.865	0.864	0.856	0.855	0.855	0.854
	0	0.819	0.819	0.811	0.811	0.811	0.811
	<0	0.779	0.779	0.772	0.771	0.770	0.769
$n_1 \neq n_2$	>0	0.839	0.832	0.829	0.824	0.827	0.824
	0	0.806	0.798	0.795	0.790	0.795	0.790
	<0	0.774	0.767	0.763	0.760	0.761	0.758
Gumbel							
$n_1 = n_2$	>0	0.783	0.782	0.815	0.814	0.824	0.823
	0	0.720	0.718	0.761	0.760	0.774	0.774
	<0	0.678	0.678	0.719	0.719	0.727	0.726
$n_1 \neq n_2$	>0	0.740	0.735	0.779	0.776	0.789	0.786
	0	0.693	0.689	0.740	0.736	0.749	0.747
	<0	0.655	0.651	0.702	0.699	0.712	0.710
Exponential							
$n_1 = n_2$	>0	0.867	0.864	0.938	0.937	0.946	0.944
	0	0.824	0.824	0.915	0.914	0.926	0.925
	<0	0.795	0.795	0.894	0.894	0.906	0.906
$n_1 \neq n_2$	>0	0.841	-	0.933	0.930	0.943	0.938
	0	0.811	-	0.919	0.917	0.930	0.926
	<0	0.786	-	0.904	0.903	0.918	0.915
Lognormal							
$n_1 = n_2$	>0	0.596	0.590	0.893	0.891	0.905	0.904
	0	0.535	0.533	0.857	0.856	0.911	0.912
	<0	0.506	0.506	0.826	0.826	0.918	0.925
$n_1 \neq n_2$	>0	0.514	-	0.874	0.873	0.879	0.876
	0	0.467	-	0.851	0.850	0.850	0.851
	<0	0.438	-	0.825	0.826	0.848	0.849

Averaged over the simulation design, Table 6.1 shows that the power difference between T_{new1} and T_{new2} is negligible, and both have higher power under normality than the other methods considered. If the normality assumption does not apply T_{new1} is recommended over T_{new2} because T_{new2} is not Type I error robust for the most skewed distributions when sample sizes are not equal. However, Table 6.1 shows that for the non-normal distributions in this simulation design, non-parametric methods are more powerful than their parametric counterparts where both samples are taken from the same distribution.

There are only a few scenarios where T_{RNK2} outperforms T_{RNK1}, or where T_{INT2} outperforms T_{INT1}. These are in rare situations when there is extreme skewness and negative correlation between the two groups. Furthermore the difference in power between the non-parametric methods and the distribution free methods is negligible, thus the more straightforward solution T_{RNK1} should suffice in most practical application.

In any event, the null hypothesis that 'the mean of the INT for Group A is equal to the mean of the INT for Group B' is likely to be poorly understood and impractical. Furthermore, such methods do not make a population normal, it makes a sample appear normal. The use of an INT imposes normality on the data, this is not the same as directly ensuring the assumption of normally distributed residuals (Servin and Stephens, 2007).

Figure 6.3 shows the power for each parameter combination within the simulation design for T_{new1} and T_{RNK1}.

Figure 6.3 illustrates that under normality, the proposed T_{RNK1} is virtually equivalent to the proposed T_{new1}. But for increasing degrees of skewness, the non-parametric test statistic T_{new1} exhibits an increasing power advantage over its parametric counterpart T_{RNK1}.

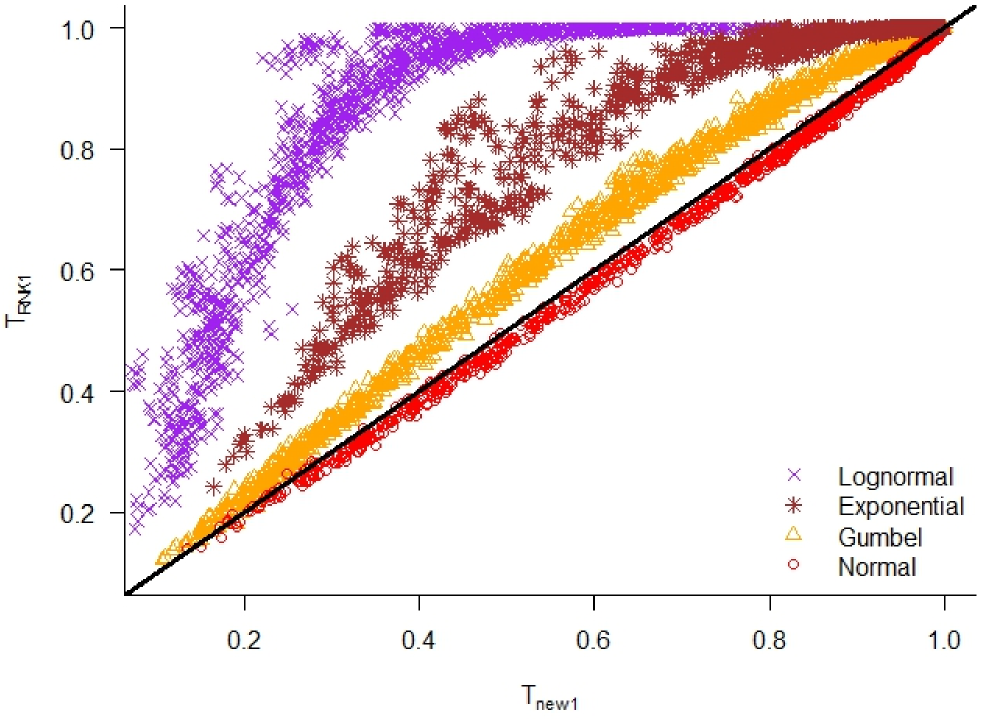


Figure 6.3: Power for each parameter combination, T_{new1} and T_{RNK1}.

6.3 Samples taken from Normal distributions with unequal variance

Null hypothesis rejection rates are obtained for each of the parameter combinations where $\mu_1 = \mu_2$ and $\sigma_1^2 \neq \sigma_2^2$. When the observations are sampled from two Normal distributions with equal means and unequal variances, the null hypothesis rejection rate represents the Type I error rate of the test. Type I error rates for each of the test statistics are given in Figure 6.4.

Figure 6.4 shows that Type I error robustness is maintained under normality for T_{new2}. Thus T_{new2} is the only test statistic making use of all available data to be Type I error robust under normality for both equal and unequal variances.

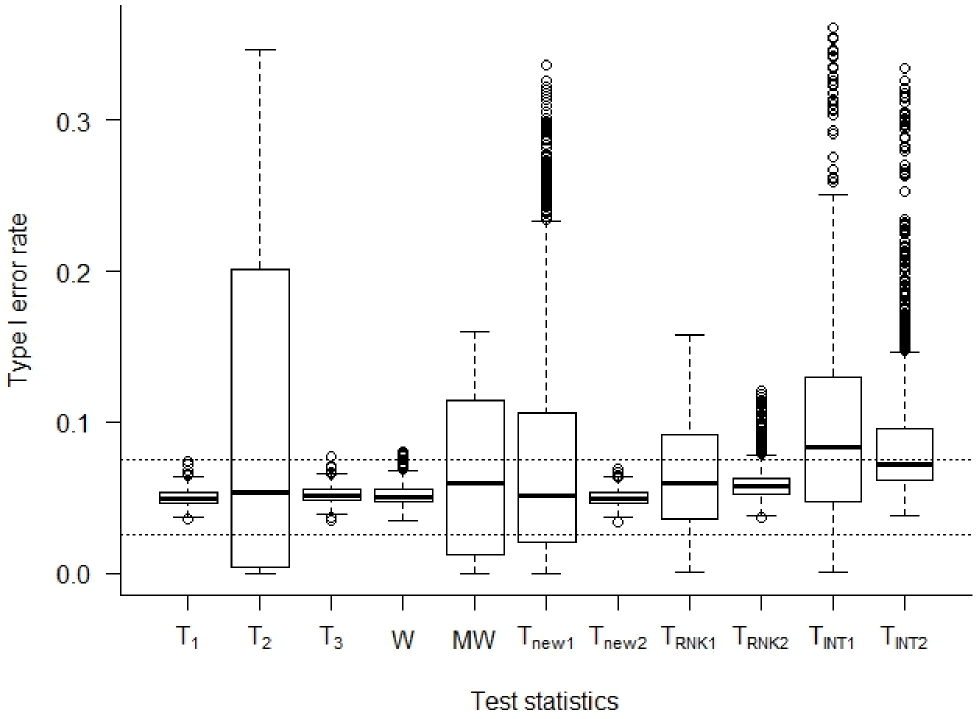


Figure 6.4: Type I error rates, Normal distribution.

For normally distributed data and unequal population variances, the test statistics not constrained to equal variances are more Type I error robust than the statistics that assume equal variances. Nevertheless, for T_{RNK2} and T_{INT2} the number of times the null hypothesis is rejected is probably greater than would be deemed acceptable. Closer inspection of the results shows these statistics are not robust when the number of paired observations is large relative to the total number of independent observations. This effect is exacerbated when ρ is large and positive.

6.4 Samples taken from distributions of differing functional form

To consider the behaviour of the test statistics when the two samples are drawn from distinctly different distributions (standardised to ensure equal means), Figure 6.5 shows the null hypothesis rejection rates when observations for Sample 1 are taken from the Standard Normal distribution, and observations for Sample 2 are taken from the Lognormal distribution.

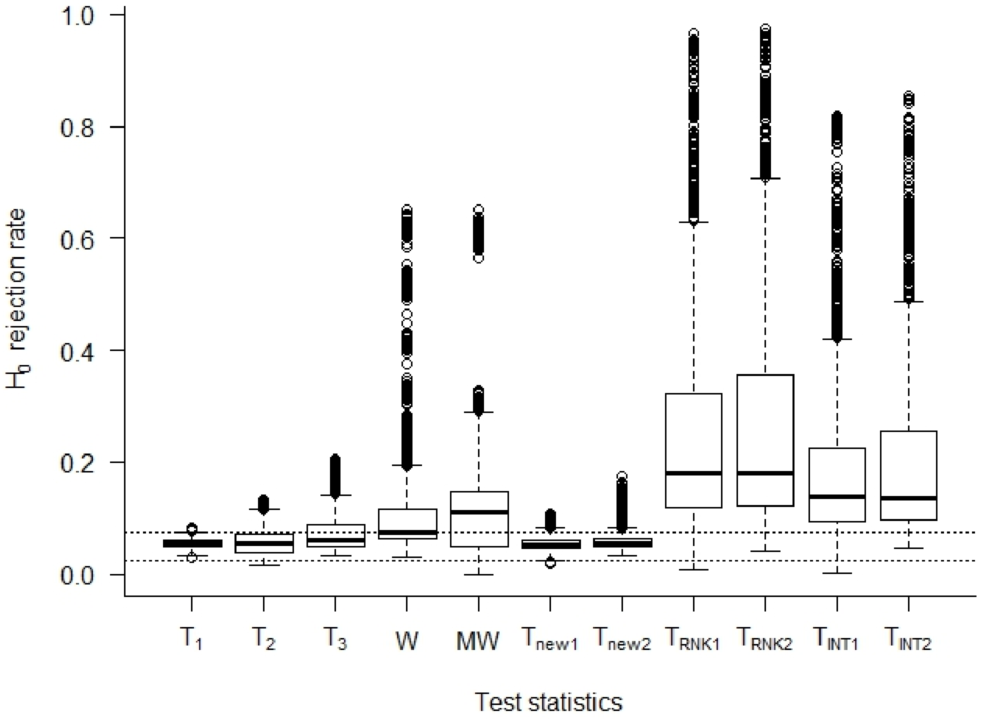


Figure 6.5: Sample 1 values taken from the Standard Normal distribution, Sample 2 observations are taken from a standardised Lognormal distribution.

Under the simulation design, standardising of the population ensures that the mean for both distributions is the same, but the shapes of the distributions are different. The null hypothesis rejection rate only represents the

Type I error rate if the null hypothesis is strictly that there is no difference in means. Figure 6.5 shows that the parametric tests are not sensitive to the different shapes of the distributions and remain valid for testing the hypothesis of equal means. Conversely, the null hypothesis rejection rate is well in excess of 5% for the non-parametric test statistics. The non-parametric statistics are sensitive to differences in the shape of the distribution, thus could be used to assess whether the distributions are equal. The null hypothesis rejection rates represent power under this latter form of H_0.

6.5 Summary

The test statistics not assuming equal variances, T_{new1}, T_{RNK1} and T_{INT1}, exhibit superior Type I error robustness for some parameter combinations, and across all scenarios have similar power properties to their counterparts where equal variances are not assumed T_{new2}, T_{RNK2} and T_{INT2} respectively. Therefore, when comparing two samples from the same non-normal distribution, the equal variances assumed forms of the test are the most appropriate.

When considering the performance of the Mann-Whitney test and other non-parametric tests, an increase in Type I error rate could be viewed as an increase in power. For example, when comparing two samples with different variances, the underlying distributions are not equal, therefore a null hypothesis of equal distributions is not true.

Under normality, the partially overlapping samples t-test proposed for equal variances, T_{new1}, is more powerful than the non-parametric equivalent T_{RNK1} and the inverse Normal transformation approach T_{INT1}.

Due to its Type I error robustness, power properties and relative simplicity, T_{RNK1} is recommended over T_{INT1} as the best solution for comparing partially overlapping samples from non-normal distributions.

Chapter 7

The impact of outliers in the comparison of two samples

This chapter begins with an introduction to the extant debate on how to detect and handle outliers. The main focus is to explore the robustness of two sample tests to outliers. This is first considered for a simulation design with a single aberrant observation. Simulations are performed for an independent samples design, then a paired samples design, then a partially overlapping samples design. This is followed by a simulation design representing multiple outliers in the partially overlapping samples framework. The concepts for an independent samples design and a partially overlapping samples design were presented at the Royal Statistical Society annual conference (Derrick, 2018a), following published results for a paired samples design in Derrick et al. (2017b).

7.1 Introduction to the outlier debate

Outliers increase the variability within a sample, this results in an increased probability of making a Type II error, an issue that is exacerbated for small sample sizes (Cousineau and Chartier, 2010). There is no approach to outliers that can be applied universally (Hodge and Austin, 2004). How to determine whether outliers are present, and the process for handling outliers, is the subject of differing opinion.

A major difficulty in empirical research can be detecting which type of outlier has occurred. Anscombe (1960) identify three types of outlier; (i) measurement error, (ii) execution error, (iii) inherent variability. The first two types specifically identify that an error was made. The decision to remove outliers that are a result of measurement error or execution error may not be unjustified. The approach for dealing with an outlier that is the result of inherent variability is of further debate, frequently authors argue against the removal of outliers in such circumstances (Grubbs, 1969; Orr, Sackett, and Dubois, 1991). If an outlier is removed prior to analyses, conclusions reported following the removal of the outlier, should be reported alongside conclusions prior to the removal of the outlier (Walfish, 2006).

Examples of outliers observed in popular culture given by Gladwell (2008) include the wealth of the individual Bill Gates or the success of the Beatles. These extreme observations can be seen as integral to an overall dataset and of interest in their own right. When outliers cannot be dismissed as data errors, studying their phenomenon can give useful novel insights (Osborne and Overbay, 2004; Aguinis, Gottfredson, and Joo, 2013).

Basic texts suggest that outliers can be identified by simple exploratory data analyses, for example boxplots. However, different statistical software adopt different methods for producing boxplots, which for the same dataset can result in inconsistency in the outliers detected (Frigge, Hoaglin, and Iglewicz, 1989). Due to the nature of random sampling, samples drawn from a Normal distribution can be expected to have outliers identified (Dawson, 2011).

The extremity of an observation is not easily identified using simple techniques. A frequently used approach to identify outliers is to define outliers as observations which are a given number of standard deviations away from the mean, or the mean difference. This number of standard deviations away from the mean varies in the literature. For instance, following ideas embedded in quality control, Ray et al. (2016) define two standard deviations beyond the mean as 'an alert', and three standard deviations beyond the mean is defined as 'an alarm'. However, the definition of arbitrary cut points such as these may be subjective and misleading. For small samples, z-scores do not offer an effective method for identifying outliers (Shiffler, 1988).

There is an abundance of formal tests to detect outliers. Grubbs's test is frequently praised for its robustness when testing for a single outlier (Baksalary and Puntanen, 1990). A comprehensive list of available tests is offered by Barnett and Lewis (1994). However, with reference to the paired samples t-test, Preece (1982) state that formal procedures for the detection and rejection of outliers are of negligible use for small sample sizes.

7.2 Robustness of tests in the presence of an aberrant observation

The simulation approach is to generate sample data under the normality assumption, then include one aberrant observation. This additional observation systematically changes in its observed value, from -8 to 8 in increments of 0.1, and is referred to as a 'marching observation'. Thus one observation is directly manipulated to create an extreme observation with otherwise normally distributed data (this may be compounded with outliers due to inherent variability within the other observations).

A systematically manipulated additional observation, demonstrates the impact on the test statistics when the aberrant observations is close to a mean difference of zero, as well as what happens when an extreme positive or negative observation is included in a sample with a non-negative mean.

For each parameter combination and each test statistic the proportion of 10,000 iterations where the null hypothesis is rejected is calculated at the 5% significance level, two sided. This gives the Null Hypothesis Rejection Rate (NHRR). Note that the terminology NHRR is used rather than Type I error rate, because the inclusion of the marching observation may invalidate the underpinning assumptions.

7.2.1 Independent samples simulation design

An independent samples design is firstly considered. The tests performed are; the independent samples t-test, Welch's test, the Mann-Whitney test, and the Yuen-Welch test. The Yuen-Welch test is performed using the R package 'PairedData' with 10% trimming per tail as outlined by Wilcox (2012).

Specifically, n_a and n_{b-1} Standard Normal deviates are generated using the Box and Muller (1958) transformation. A fixed aberrant observation, x_b, is appended to the $x_1, x_2 \cdots, x_{b-1}$ observations to give a total sample size of n_b. For each simulated sample, the value of x_b is systematically varied from -8 to 8 in increments of 0.1. It is this value, x_b, which is referred to as the 'marching observation'. The values of x_b approximately range between +/- 8 standard deviations from the mean and therefore cover limits likely encountered in a practical environment. Without loss of generality, if $\bar{x}_a - \bar{x}_{b-1} < 0$ then the observations $x_1, x_2 \cdots, x_{b-1}$ are multiplied by -1 to ensure a non-negative sample mean. This change of sign does not affect the validity of a two-sided test of a nil-null hypothesis for these data. This condition is to ensure that the concordance of effects $\bar{x}_a - \bar{x}_{b-1} > 0, x_b > 0$ or discordance of effects $\bar{x}_a - \bar{x}_{b-1} > 0, x_b < 0$ can be established.

The effect of gradually increasing the marching observation is to gradually violate the assumption of the nil-null hypothesis, therefore large positive values of the marching observation would increase the NHRR. Negative values of x_b would cancel out the overall positive difference observed within the sample differences and decrease the NHRR.

Interest is on relatively small sample sizes as these are situations in which potentially large observations may have the greatest practical effect. The sample sizes of n_a and n_b that are varied within a factorial design are {10, 15, 20}. The values of σ_1 and σ_2 that are varied within the factorial design are {1, 2}. The simulation is run 10,000 times for each parameter combination of n_a, n_b, x_b, σ_1, σ_2.

Results from a selection of scenarios from the independent samples simulation design are displayed. Each scenario consists of the same total sample size of 30, and are as follows:

1. $n_a = 15,\ n_{b-1} = 14,\ \sigma_1 = 1,\ \sigma_2 = 1$
2. $n_a = 10,\ n_{b-1} = 19,\ \sigma_1 = 1,\ \sigma_2 = 1$
3. $n_a = 20,\ n_{b-1} = 9,\ \sigma_1 = 1,\ \sigma_2 = 1$
4. $n_a = 15,\ n_{b-1} = 14,\ \sigma_1 = 1,\ \sigma_2 = 2$
5. $n_a = 10,\ n_{b-1} = 19,\ \sigma_1 = 1,\ \sigma_2 = 2$
6. $n_a = 20,\ n_{b-1} = 9,\ \sigma_1 = 1,\ \sigma_2 = 2$

For each of the six scenarios; Figure 7.1 gives the NHRR when performing the independent samples t-test; Figure 7.2 gives the NHRR when performing Welch's test; Figure 7.3 gives the NHRR when performing the Yuen-welch test; Figure 7.4 gives the NHRR when performing the Mann-Whitney test.

Figure 7.1 shows that for Scenarios 1-4, when $x_b = 0$, the NHRR is approximately equal to the nominal Type I error rate. However, an extreme observation paradox is apparent. The paradox is a contrariwise decrease in the NHRR as the value of an extreme observation increases in the direction of the overall effect. For positive sample means, as the value of x_b increases, the independent samples t-test has an increasingly higher NHRR, until a turning point is reached.

Figure 7.1 shows that for scenarios 5-6, when $x_b = 0$, the NHRR is not approximately equal to the nominal Type I error rate. This is further evidence of the non-robustness of the independent samples t-test under these conditions.

Figure 7.2 shows that for each of the scenarios, when $x_b = 0$ the NHRR is approximately equal to the nominal Type I error rate. This is anticipated given the Type I error robustness of Welch's test. However, Figure 7.2 indicates that for increasing values x_b of the paradox is also observed for Welch's test.

Figure 7.3 and Figure 7.4 show that the Mann-Whitney test and the Yuen-Welch test are liberal for positive values of the marching observation, and are conservative for negative values of the marching observation. The

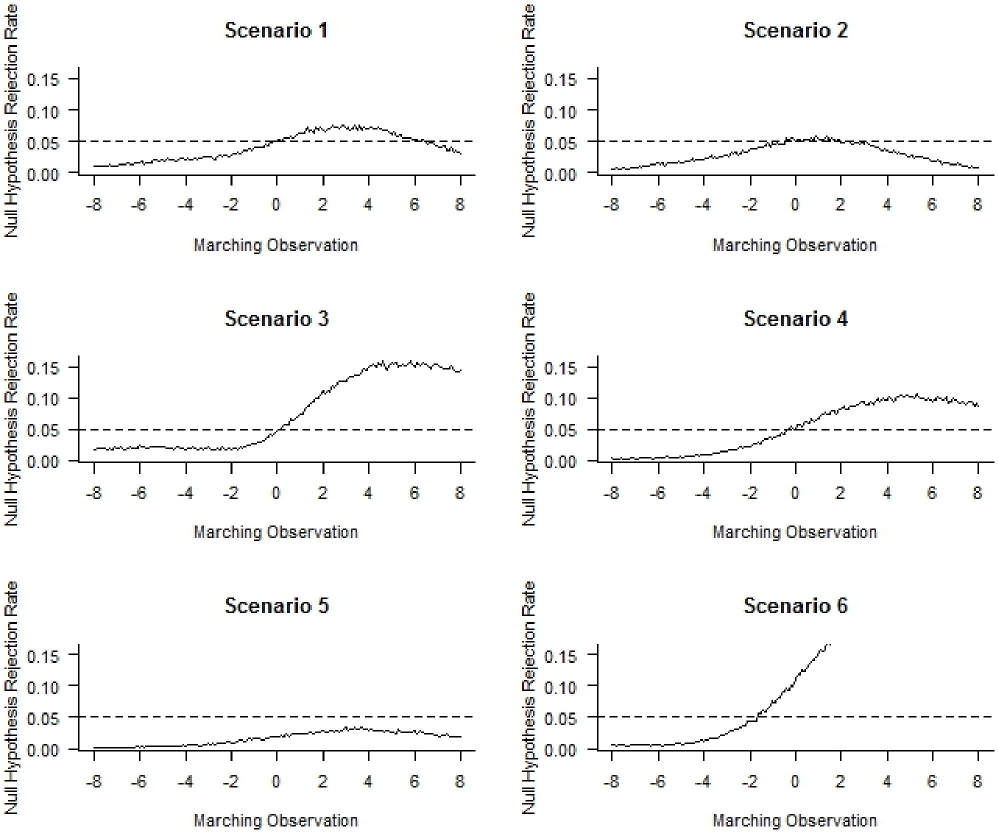


Figure 7.1: NHRR when performing the independent samples t-test.

Mann-Whitney test and Yuen-Welch test maintain NHRR close to the nominal significance level when sample sizes are equal or the larger sample size includes the marching observation. Both tests tend to a fixed value as $x_b \to \infty$, and both tests tend to a fixed value close to zero as $x_b \to -\infty$. For the Mann-Whitney test, due to the use of rank values, the test is not greatly affected by the magnitude of the extreme observation. Similarly due to the trimming, the Yuen-Welch test is not greatly affected by the magnitude of the extreme observation.

For the independent samples t-test, Bakker and Wicherts (2014) found inflated Type I error rates following the removal of outliers, and recommend proceeding with the Mann-Whitney test or the Yuen-Welch test without removal of outliers. However, Figure 7.3 and Figure 7.4 show that the fixed

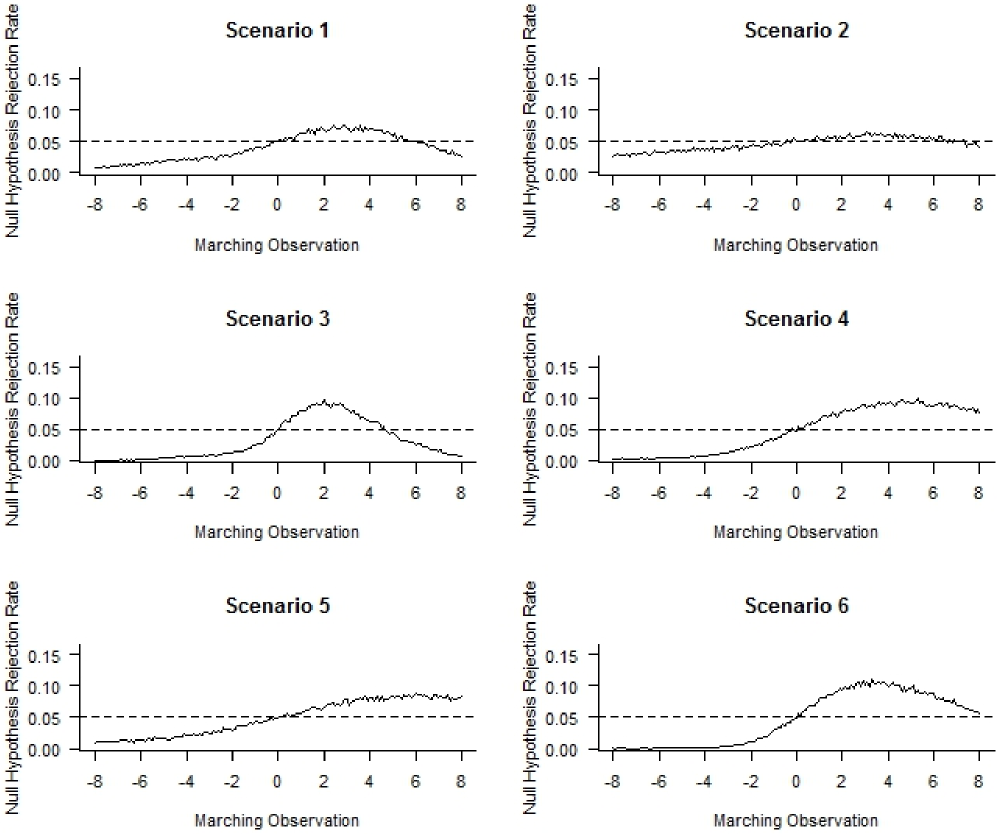


Figure 7.2: NHRR when performing Welch's test.

value that these tests tend to is dependent on sample size and variance. For Scenario 5, it can be seen that the NHRR when performing the Mann-Whitney test remains below the nominal significance level for all values of x_b. It can also be seen that the Mann-Whitney test only maintains the nominal significance level at $x_b = 0$ for Scenario 1. These results corroborate findings by Zimmerman (1998) that the Mann-Whitney test does not always provide a robust alternative approach when an outlier is present.

Fagerland (2012) suggest that the problem is not the t-test itself, moreover it may be that in the presence of an outlier, the mean may be a poor measure of central location, and other measures of location such as trimmed means or non-parametric tests may be more appropriate.

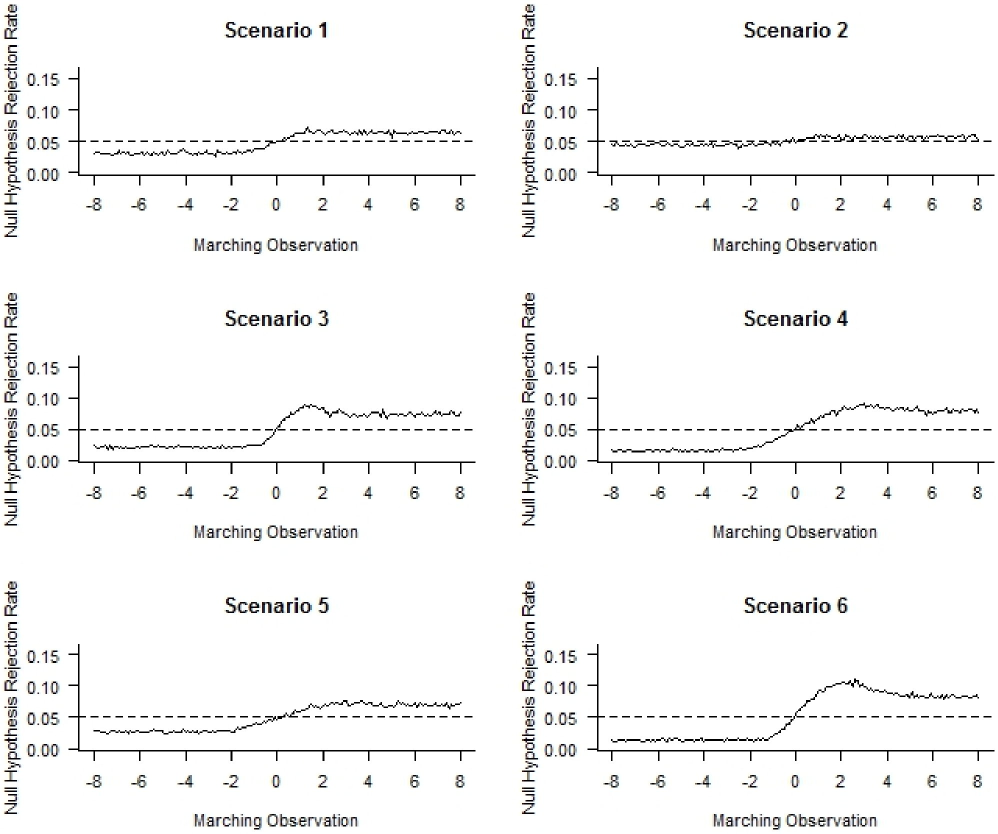


Figure 7.3: NHRR when performing the Yuen-Welch test.

7.2.2 Paired samples simulation design

Using methodology akin to the above, the effect of an aberrant observation is considered for a paired samples simulation design.

The paired samples t-test, the Wilcoxon test, and Yuen's paired sample t-test, are performed for a two-sided nil-null hypothesis. Yuen's paired samples t-test is performed using the R package 'PairedData' with 10% trimming per tail as outlined by Wilcox (2012). This uses the principles of trimmed means and windsorized variances in the Yuen-Welch test, applied to paired data.

The paired samples t-test is logically and numerically equivalent to the one sample t-test performed on paired differences. Within the simulation, differences are generated rather than the paired observations themselves.

Specifically, $n-1$ Standard Normal deviates are generated using the Box

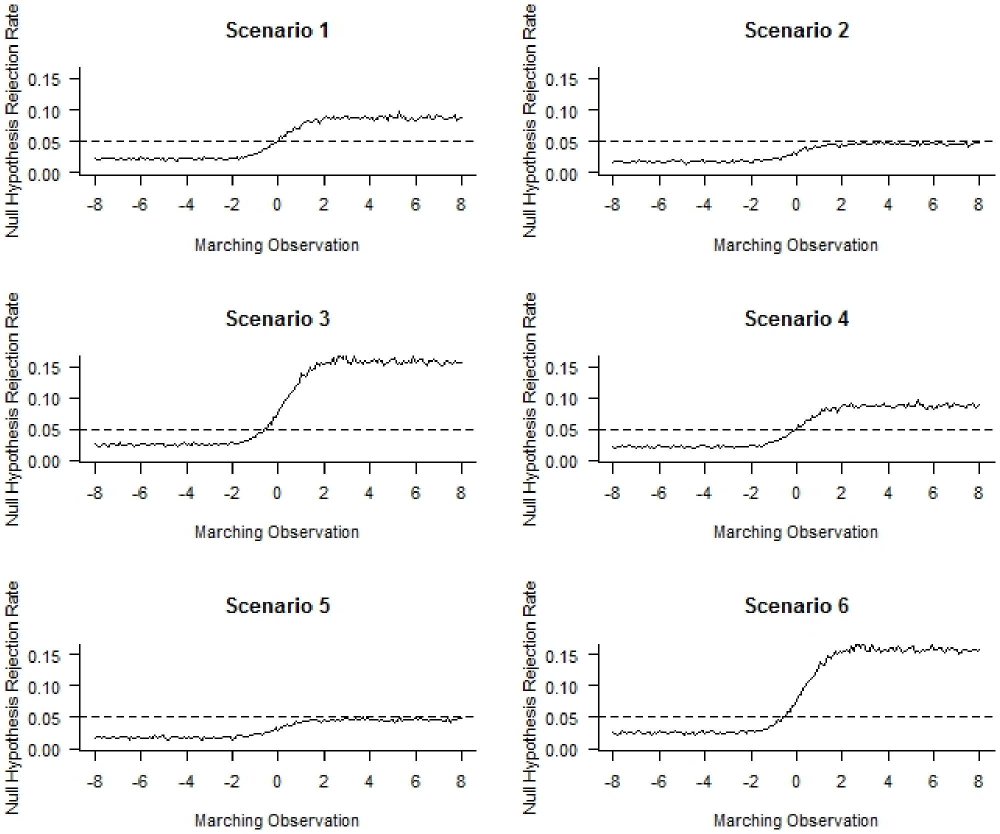


Figure 7.4: NHRR when performing the Mann-Whitney test.

and Muller (1958) transformation. A fixed aberrant observation, x_n, is appended to the $x_1, x_2 \cdots, x_{n-1}$. For each simulated sample, the value of x_n is systematically varied from -8 to 8 in increments of 0.1. It is this value, x_n, which is referred to as the 'marching observation'. To isolate the phenomenon of interest, if $\sum(x_1, x_2 \cdots, x_{n-1})/(n-1) < 0$ then the observations $x_1, x_2 \cdots, x_{n-1}$ are multiplied by -1 to ensure a non-negative sample mean.

The sample sizes, n, varied within a factorial design are {10, 15, 20, 25}. The simulation is run 10,000 times for each parameter combination of n and x_n, using the nominal significance level of 5%.

Figure 7.5 gives the NHRR of the paired samples t-test, Figure 7.6 gives the NHRR of Yuen's paired samples t-test and Figure 7.7 gives the NHRR of the Wilcoxon test.

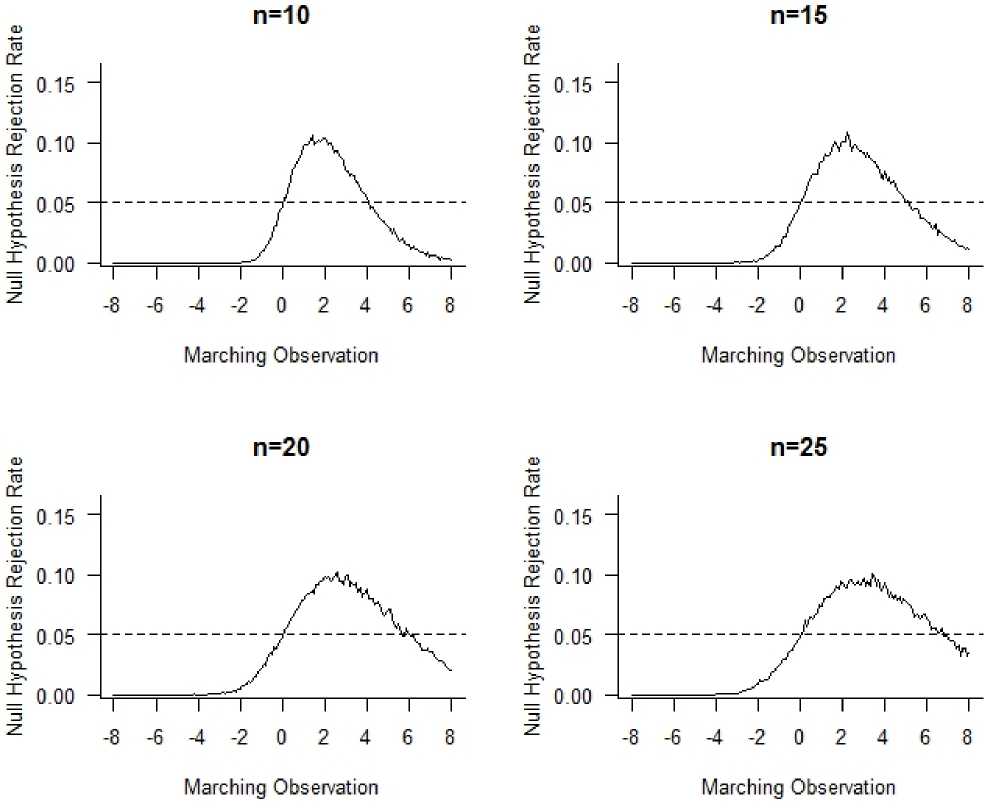


Figure 7.5: NHRR when performing the paired samples t-test.

Figure 7.5 shows that when the value of $x_n = 0$, the NHRR is approximately equal to the nominal Type I error rate. For positive sample differences, as the value of x_n increases, the paired samples t-test has an increasingly higher NHRR until a turning point is reached. Extreme and increasingly larger values of the marching observation in the direction of the sample effect results in a progressively lower NHRR, with values noticeably lower than the nominal Type I error rate. The paradox referred to for the independent samples t-test and Welch's test is also observed for the paired samples t-test. These effects are replicated for all four sample sizes, but are marginally less extreme with increasing sample size. Figure 7.5 also shows that a large value for the marching observation in the opposite direction to the mean of the first $n-1$ observations, effectively results in a zero value for

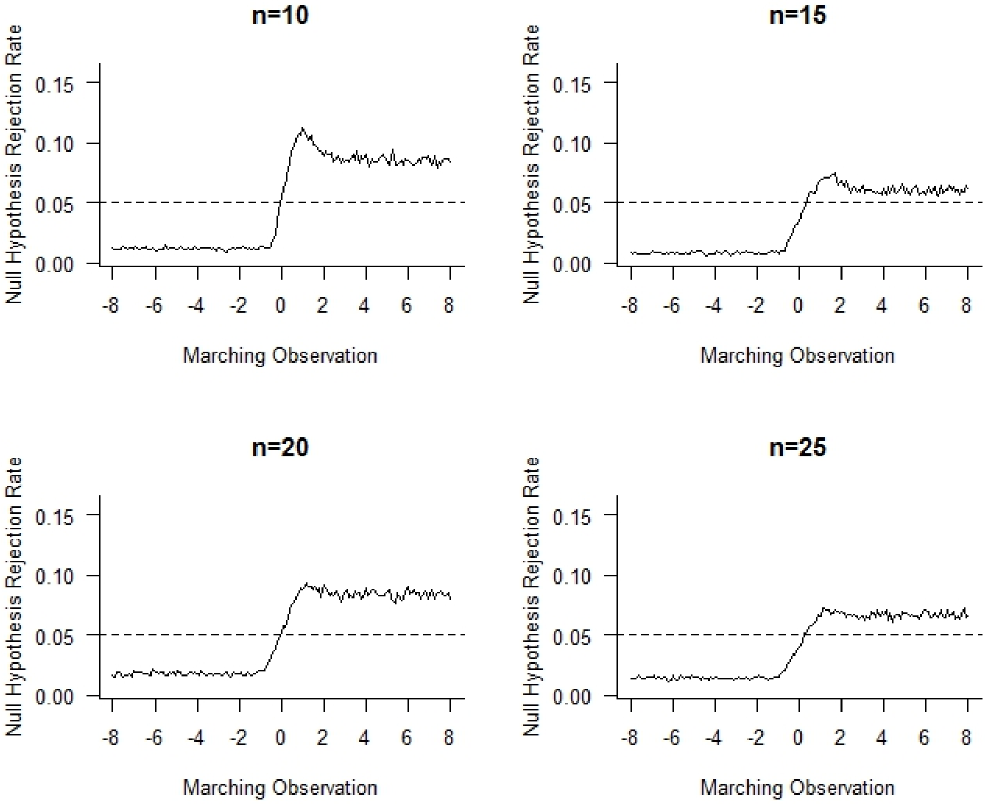


Figure 7.6: NHRR when performing Yuen's paired samples t-test.

the NHRR.

Zumbo and Jennings (2002), using a novel contamination model, conclude that the paired samples t-test has an inflated Type I error rate with increasing asymmetric contamination, the marching observation simulations above indicate that the effect of a single outlier is dependent on sample size, magnitude and direction of the outlier, and could lead to increases as well as decreases in the NHRR.

Figure 7.6 and Figure 7.7 show that when $x_n > 0$ and $\bar{x}_{n-1} > 0$, both Yuen's paired samples t-test and the Wilcoxon test result in the null hypothesis being rejected more frequently than the nominal significance level. Conversely, when $x_n < 0$ and $\bar{x}_{n-1} > 0$, both Yuen's paired samples t-test and the Wilcoxon test have a NHRR lower than the nominal significance

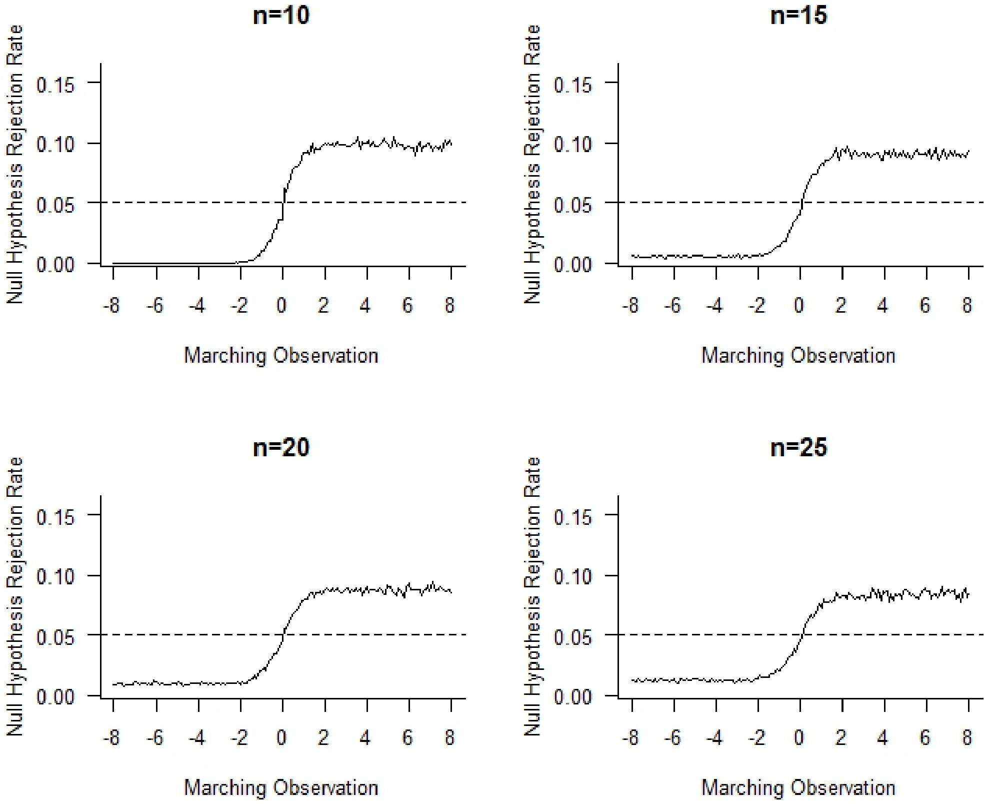


Figure 7.7: NHRR when performing the Wilcoxon test.

level. These findings are entirely consistent with expectation for a robust test given the design of the simulation.

For the Wilcoxon test, due to the use of rank values, the test is not greatly affected by the magnitude of the extreme observation. Similarly due to the trimming, Yuen's paired samples t-test is not greatly affected by the magnitude of the extreme observation. The phenomenon of a turning point when $x_n > 0$ is not observed, the NHRR tends to a fixed value as $x_n \to \infty$. Mathematical proof of this property is given in Derrick et al. (2017b)

Under a location shift model, the inclusion of genuinely large positive observation x_n into a sample with $\bar{x}_{n-1} > 0$ should lead to an increase in NHRR in a two-sided test of the nil-null hypothesis. This effect is observed with Yuen's paired samples t-test and with the Wilcoxon signed rank sum test, but it is not consistently observed with the paired samples t-test. Likewise, the inclusion of a large negative observation x_n into a sample with $\bar{x}_{n-1} > 0$ should lead to a relative decrease in NHRR. This effect is observed with Yuen's paired samples t-test and with the Wilcoxon test, but the effect is most evident, and is sample size dependent, for the paired samples t-test.

The simulations demonstrate the seemingly paradoxical effect of large outliers on the performance of the paired samples t-test. The simulations indicate that Yuen's paired samples t-test and the Wilcoxon signed rank sum test have robust behaviour in the presence of a single outlying observation, as found by Zimmerman (2011). However, there is evidence that rank based methods do not completely eliminate the influence of outliers.

7.2.3 Partially overlapping samples simulation design

The simulation design is extended so that the partially overlapping samples t-tests are performed for a two-sided nil-null hypothesis.

Under a nil-null hypothesis; the parametric partially overlapping samples t-tests, T_{new1} and T_{new2}, are used to test for a mean difference of zero. Under the same conditions, the non-parametric partially overlapping samples t-tests T_{RNK1} and T_{RNK2} are used to test for differences symmetrically distributed around zero.

The approach is to simulate two groups of Normal deviates for a paired design with $n = 15$ and $\rho = 0.5$. The samples are multiplied by fixed values of σ_1 and σ_2 as detailed in the six scenarios below. Without loss of generality, if $\bar{x}_a - \bar{x}_{b-1} < 0$ then the observations in Sample 2 are multiplied by -1 to ensure a non-negative sample mean. Observations are then deleted completely at random with the constraint that remaining sample sizes are as per the six scenarios below:

1. $n_a = 5,\ n_{b-1} = 4,\ n_c = 5,\ \sigma_1 = 1,\ \sigma_2 = 1$

2. $n_a = 5,\ n_{b-1} = 4,\ n_c = 5,\ \sigma_1 = 1,\ \sigma_2 = 2$

3. $n_a = 5,\ n_{b-1} = 4,\ n_c = 5,\ \sigma_1 = 2,\ \sigma_2 = 1$

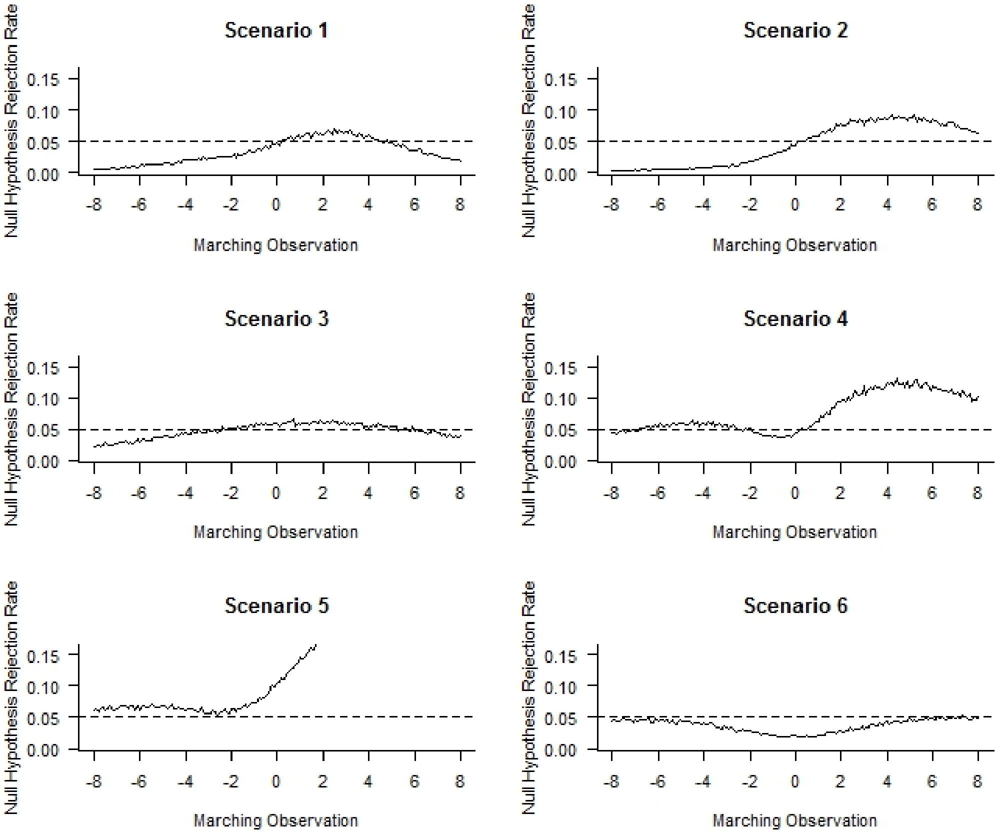


Figure 7.8: NHRR when performing T_{new1}.

4. $n_a = 10,\ n_{b-1} = 3,\ n_c = 2,\ \sigma_1 = 1,\ \sigma_2 = 1$

5. $n_a = 10,\ n_{b-1} = 3,\ n_c = 2,\ \sigma_1 = 1,\ \sigma_2 = 2$

6. $n_a = 10,\ n_{b-1} = 3,\ n_c = 2,\ \sigma_1 = 2,\ \sigma_2 = 1$

The six scenarios selected are for indicative purposes to show the behaviour of the partially overlapping samples t-tests, T_{new1}, T_{new2}, T_{RNK1} and T_{RNK2}.

An additional observation, x_b, is appended to the n_{b-1} observations. For each simulated sample, the value of x_b is systematically varied from -8 to 8 in increments of 0.1. Again, it is this value, x_b, which is referred to as the 'marching observation'.

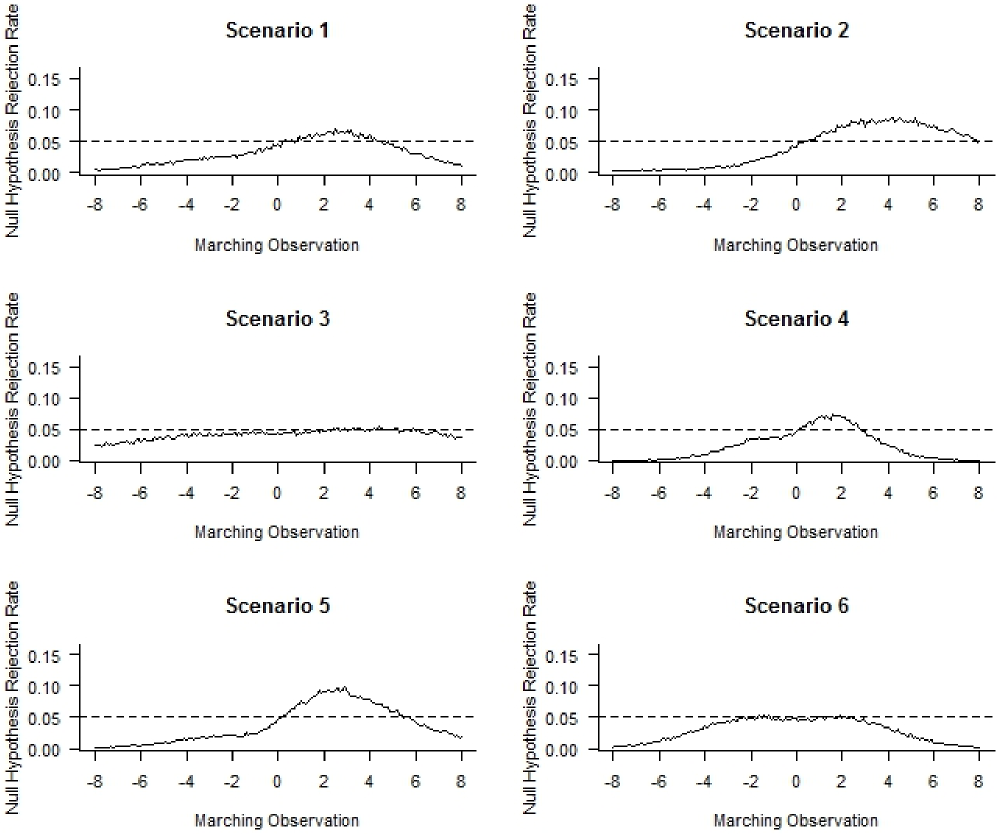


Figure 7.9: NHRR when performing T_{new2}.

Figure 7.8 - Figure 7.10 exhibits the NHRR for T_{new1}, T_{new2} and T_{RNK1} respectively.

Figure 7.8 shows that the extreme observation paradox identified for the independent samples t-test is also observed for T_{new1}. However, under unequal sample sizes and unequal variances, alternative undesirable patterns are also observed. This can be explained by, and add further support to, the previously established non-robustness of this test statistic in these conditions.

Figure 7.9 shows the extreme observation paradox observed for Welch's test is also observed when performing T_{new2}.

Figure 7.10 shows that T_{RNK1} tends towards a fixed value for the NHRR. However the fixed NHRR value is inflated when the smaller sample size is associated with the larger variance. Similar patterns are demonstrated for

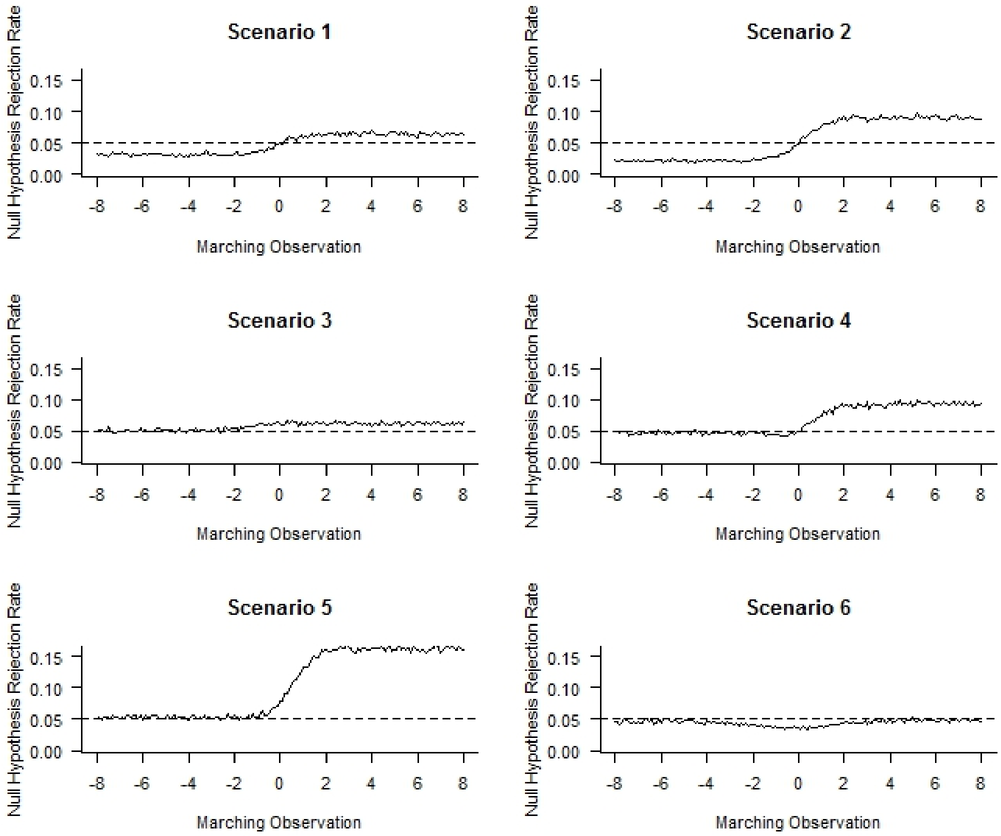


Figure 7.10: NHRR when performing T_{RNK1}.

T_{RNK2} (not displayed).

7.3 Robustness in the presence of multiple outliers

To directly assess the robustness of test statistics to multiple outliers, a mixed Normal distribution is considered.

Simulations are performed as per the design and test statistics in Chapter 6, with the following transformation applied to observations in each sample to induce outliers; X = 10N with probability 0.1 and X = N with probability 0.9.

This is first considered under the null hypothesis where $\mu_1 = \mu_2 = 0$, then under the alternative hypothesis where $\mu_2 - \mu_1 = 0.5$. The NHRR for the independent samples tests and the paired samples tests are calculated by discarding the paired observations or independent observations respectively.

The Type I error robustness, where $\mu_1 = \mu_2 = 0$, for the comparison of two samples from the mixed Normal distribution is given in Figure 7.11.

A power comparison for a selection of test statistics, where $\mu_2 - \mu_1 = 0.5$, is given in Table 7.1.

Table 7.1: Power, mixed Normal distribution.

ρ	W	MW	T_{new1}	T_{new2}	T_{RNK1}	T_{RNK2}	T_{INT1}	T_{INT2}
$n_a = n_b$								
>0	0.599		0.417	0.412	0.796	0.795	0.793	0.792
0	0.459	0.486	0.346	0.344	0.739	0.737	0.739	0.738
<0	0.391		0.304	0.304	0.697	0.696	0.694	0.694
$n_a \neq n_b$								
>0	0.597		0.329	0.351	0.752	0.753	0.751	0.752
0	0.462	0.348	0.271	0.292	0.710	0.712	0.710	0.712
<0	0.393		0.233	0.253	0.672	0.674	0.673	0.674

It can be seen from Figure 7.11 that each of the test statistics generally maintain Type I error robustness, or are conservative, across the simulation design. However there is some evidence that when samples are drawn from the same mixed Normal distribution, test statistics constrained to equal variances are more Type I error robust. The Type I error rate deficiencies for the proposed tests occur where there is a very small sample size and a very

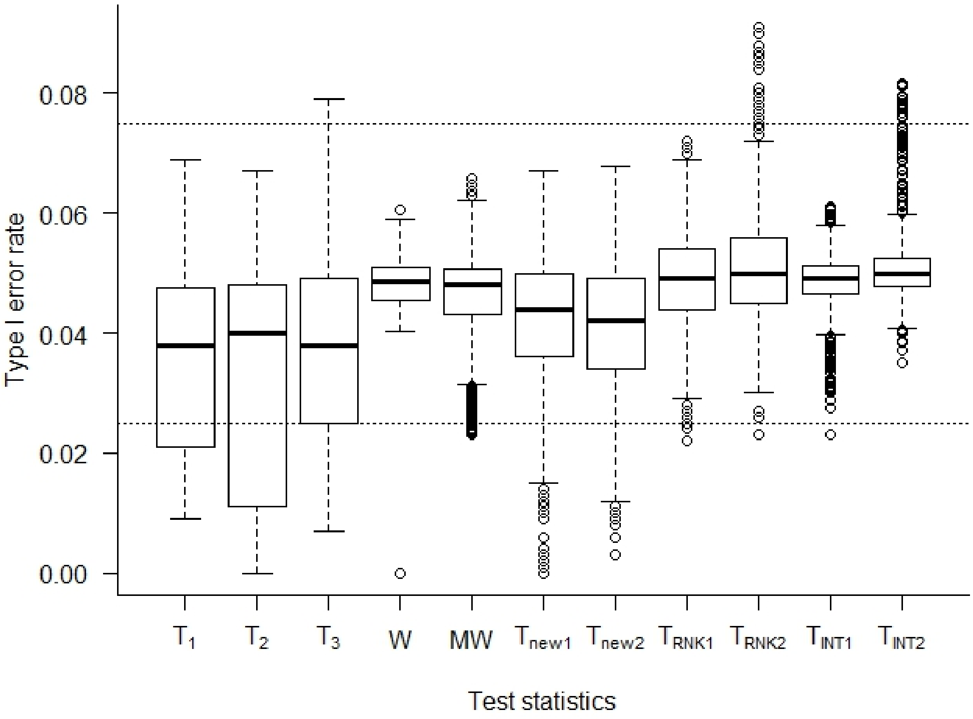


Figure 7.11: Type I error rates, mixed Normal distribution.

large sample size, i.e. where max $\{n_a, n_b, n_c\}$ - min $\{n_a, n_b, n_c\} = 495$. For moderate sample sizes that may be encountered in most practical settings, performing T_{new1} or T_{new2} may be reasonable.

Table 7.1 shows that there are clear power advantages for using the rank based methods over the parametric methods under these conditions.

When comparing two samples from the same mixed Normal distribution, T_{RNK1} demonstrates the tightest Type I error rate control, and superior power. This conclusion is in line with the conclusions in Chapter 6 for the comparison of two non-normal distributions.

7.4 Summary

Given the debate in the literature regarding the removal of an outlier, it is important to be aware of the impact of an outlier on the outcome of statistical tests. The results show that a single aberrant observation can potentially either mask true effects or show phantom significant effects. Typically the natural desire of a researcher is to prove significant effects, the researcher will often consider the removal of outliers in order to conclude a significant effect. It is of note that the removal of an outlier may in fact produce the opposite outcome. The decision not to remove an outlier could be taken so that a significant effect is observed. In this respect, a decision not to remove an observation should be considered with as much vigour as the decision to remove an observation. In addition, it should be considered that an observation that may appear to be an outlier may represent a location shift (Walfish, 2006).

The extreme observation paradox is the contrariwise decrease in the NHRR as the value of an extreme observation increases in the direction of the overall effect. This paradox is observed for the paired samples *t*-test, the independent samples *t*-test and Welch's test. As a consequence, this paradox is also observed for the parametric partially overlapping samples *t*-tests. These tests display behaviour strongly dependent on the magnitude of the outlier. In contrast, test statistics making use of rank values do not suffer from the extreme observation paradox.

In the presence of multiple outliers, as demonstrated by a mixed Normal distribution, non-parametric tests demonstrate superior Type I error robustness and power relative to the parametric tests.

In a paired samples design, outliers are identified based on the sample differences. In an independent samples design, outliers are identified based on the raw data in each of the two samples. Due to the presence of both paired and independent observations, definition of an outlier in the partially overlapping samples scenario is more complex. However, the use of a single marching observation, or multiple extreme observations are broadly in line with the definition of an outlier by Anscombe (1960). Thus the results herein give an indication of the partially overlapping samples t-tests in the presence of outliers. Under these conditions T_{RNK1} is the recommended test.

In textbooks listing the assumptions of the t-test, the assumption of no significant outliers is sometimes listed, but sometimes not. Given the results above, the assumption should be listed. The question of how to identify a 'significant outlier' has no answer that is applied universally (Hodge and Austin, 2004; Barnett and Lewis, 1994), and is therefore an area of debate that will continue.

Chapter 8

The comparison of two samples, for ordinal data

Naive tests and the partially overlapping samples t-tests are assessed for Type I error robustness and power when two samples are taken from an ordinal scale. For a five option Likert question, Derrick and White (2017) provide an overview of the Type I error robustness and power for test statistics that utilise either only pairs or only independent observations. Extension to a seven option Likert scale, and the partially overlapping samples scenario is summarised in Derrick and White (2018). The simulation in this section is designed as such that the partially overlapping samples scenario is explored, alongside traditional test statistics that discard observations. The recommended solution herein is compared to the partially overlapping samples t-test solutions, and further documentation from the R package to facilitate application is supplied.

8.1 Background

An application where partially overlapping samples can occur on an ordinal measurement scale is a comparison of responses of two Likert questions, where some participants did not complete both questions (Maisel and Fingerhut, 2011). A further application of two partially overlapping samples is a comparison of the responses of the same Likert question on two separate occasions, where some participants were not available at both measurement periods (Bradley, Waliczek, and Zajicek, 1999). In both of these applications, the authors discarded the unpaired observations and performed the paired samples t-test. Assuming that data are MCAR this approach is not unjustified, given the large sample sizes obtained. However, power may be adversely affected for studies with smaller sample sizes.

Due to their intuitive appeal and simple construction, Likert questions are popular for measuring attitudes of respondents (Nunnally, 1994). A Likert item is a forced choice ordinal question which captures the intensity of opinion or degree of assessment in survey respondents. Historically a Likert item comprises five options worded: ‘Strongly approve’, ‘Approve’, ‘Undecided’, ‘Disapprove’, ‘Strongly disapprove’ (Likert, 1932). Other alternative wording such as ‘Agree’ or ‘Neutral’ or ‘Neither agree nor disagree’ may be used depending on the context.

The literature is sometimes confused between the comparison of samples using summed Likert scales and the comparison of samples for individual Likert items (Boone and Boone, 2012). A summed Likert scale is formed by the summation of multiple Likert items that measure similar information. This summation process necessarily requires the assignment of scores to the Likert ordinal category labels. The summation of multiple Likert items to produce Likert scales is a well-established practice in scale construction, and is one which can produce psychometrically robust scales with interval-like properties. Such derived scales could potentially yield data amenable to analysis using parametric techniques (Carifio and Perla, 2007). Distinct from summed Likert scales, the comparison of two samples from a single Likert question is considered herein so as to compare test statistics for two samples when ordinal data is present.

In certain methodological and practical aspects, Likert question responses may approximate interval level data and can be analysed assuming an underlying continuous scale (Norman, 2010). Likert questions with five options are frequently used, and the ordinal codes {-2, -1, 0, 1, 2} could be applied to these options for a balanced question, with ‘0’ representing the neutral response. In addition, seven option Likert-type questions are often used, because the summation of responses is known to have high reliability (Cicchetti, Shoinralter, and Tyrer, 1985). Henceforth the ordinal codes {-3, -2, -1, 0, 1, 2, 3} are used as numerical scores. Balanced response options around the neutral option is typically assumed. Although the exact wording of the neutral response is not an issue (Armstrong, 1987), if the options either side of the neutral response are not balanced then the assumption that the responses approximate interval level data is not reasonable (Bishop and Herron, 2015). Other issues with Likert scales include respondents tendency to give positive responses, and the potential for differing interpretation of categorical options by both the responder and the analyst (Hodge and Gillespie, 2003). When the assumption of an underlying continuous distribution is not inappropriate and the questions are suitably formed, parametric tests for differences between the two sample means are reasonable (Jamieson, 2004; Allen and Seaman, 2007).

For two independent samples De Winter and Dodou (2010) found that both the independent samples *t*-test and the Mann-Whitney test are generally Type I error robust at the 5% significance level for a five option Likert item. This is true across a diverse range of distributions and sample sizes. Both tests suffer some exceptions to Type I error robustness when the distributions have extreme kurtosis and skew. The power is similar between the two tests, for both equal and unequal sample sizes. When the distribution is multimodal with responses split mainly between ‘Strongly approve’ and ‘Strongly disapprove’, the independent samples *t*-test is more powerful than the Mann-Whitney test. Rasch, Teuscher, and Guiard (2007) show that using the Mann-Whitney test using the Normal approximation with correction for ties is Type I error robust for two groups of independent observations on a five option Likert item.

For two independent samples, Nanna and Sawilowsky (1998) found that the independent samples *t*-test and the Mann-Whitney test are Type I error robust for seven option Likert item responses, with the Mann-Whitney test superior in power. This is likely observed because there is more scope to encounter greater skew with more options to choose from.

The Type I error robustness of the independent samples t-test is further supported by Heeren and D'Agostino (1987) for a four point ordinal scale.

The literature is much quieter on the analyses of Likert items in paired samples designs. When performing the Wilcoxon test, if the samples are from an underlying Normal distribution, the null hypothesis of equal distributions may not be unreasonable, but this is particularly sensitive to changes in location (Hollander, Wolfe, and Chicken, 2013). Thus if samples are from a bivariate Normal distribution, assessing for a location shift is reasonable.

For multiple independent groups with ordinal data, application of real data suggests that there is little practical difference whether parametric or non-parametric approaches are taken when sample sizes are large (Mircioiu and Atkinson, 2017). However, the correct choice of analysis depends on the exact form of the question of interest (Roberson et al., 1994). Non-parametric tests are not inappropriate when interval approximating data is assumed, if the only potential difference between the samples is their central location (Clason and Dormody, 1994; Sisson and Stocker, 1989).

Given the discrete nature of Likert scales, differences of zero between the two samples occur frequently. The Pratt (1959) test which incorporates these zero differences in its calculation may overcome this issue. In Pratt's test the absolute paired differences are ordered including the zero differences, ranks are applied to the non-zero differences as if the zero differences had received ranks, and these ranks used in the Wilcoxon test. Conover (1973) compared the Wilcoxon test dropping zero differences to Pratt's test incorporating zero differences and concluded that the relative performance of the two approaches depends on the underlying distribution. The comparison conducted by Conover (1973) did not include Likert items and did not extend to the inclusion of the paired samples t-test. A second method for handling zero differences also suggested by Pratt (1959) is to randomly allocate each of the zero differences to either positive or negative ranks. To achieve this, for every zero difference a deviate sampled from $U(-0.1, 0.1)$ is added before proceeding with the ranking. The range of values for this sampling distribution is arbitrary, but should be lower than the minimum distance between two units on the discrete scale. An issue with this test is that adding a randomly generated value to each score could mean that the result is different each time the test is performed on the same data. In preliminary investigations by Derrick and White (2017), the former method attributed to Pratt (1959) is found to be more Type I error robust than the latter. The former method also appears to be more widely known. Most reference in the literature refers to the 'Wilcoxon-Pratt' test without formal definition of which form of the test is used, and the 'dictionary of statistics and methodology' gives no further clue than 'correction for ties' (Vogt, 2018). For the avoidance of doubt, the former method where ranks are applied to the non-zero differences as if the zero differences had received ranks is considered in the simulation below and is referred to as Pratt's test.

There are occasions where a test for extremely small samples of $n \leq 5$ in each group is required (De Winter, 2013). However, six pairs is the minimum number required before a significant difference can be found when performing the Wilcoxon test or Pratt's test at the 5% significance level. If performing the paired samples t-test under the same conditions a minimum of three pairs is required.

Pratt's test, the Wilcoxon test and the paired samples t-test can be easily extended for use when partially overlapping samples are present, if the researcher is willing to discard any unpaired data. Similarly, the independent samples t-test and the Mann-Whitney test can be easily extended for the use when partially overlapping samples are present, if the researcher is willing to discard any paired data. A simulation design to consider each of the tests for data on an ordinal scale is given in Section 8.3. However, the discarding of data may introduce bias and reduce power. As an alternative, the partially overlapping samples t-tests T_{new1} and T_{new2} that make use of all of the available data are also considered in the simulation design. Both T_{RNK1} and T_{RNK2} are not considered in the simulation design due to the high volume of ties inherent in this type of data, which would result in relatively low power for these tests.

8.2 Example

The following example for illustrative purposes only is based on an undergraduate project conducted via an online survey of student respondents.

The number of people in industrialised countries that are following vegan diets is increasing (Janssen et al., 2016). Attitudes towards vegetarian and vegan dietary provisions with respect to a particular student catering facility are considered. Responses to the following two statements are on a Likert scale, with the frequency of responses given in Table 8.1.

Statement 1. I would like more vegetarian options on the menu.

Statement 2. I would like more vegan options on the menu.

Table 8.1: Responses to attitudes survey, Statement 1 (horizontal) and Statement 2 (vertical).

	Strongly disagree	Disagree	Neither	Agree	Strongly agree
Strongly disagree	4	2	4	1	0
Disagree	1	3	6	2	2
Neither	0	0	16	9	10
Agree	0	0	1	9	4
Strongly agree	0	0	1	6	12

In addition to the paired responses in Table 8.1, five participants only responded to Statement 1 (Strongly agree: 1, Agree: 2, Neither:2), and seven participants only responded to Statement 2 (Strongly disagree: 5, Disagree: 2).

Using codes {-2, -1, 0, 1, 2} for 'Strongly disagree' through to 'Strongly agree', the mean response for Statement 1 is 0.73, and the mean response for Statement 2 is 0.04.

Overall, it appears that the catering facility could improve the customer satisfaction by increasing the number of vegetarian options on the menu. Although slightly less clear, customer satisfaction may also be improved by increasing the number of vegan options on the menu.

Performing T_{new1} in a comparison of the responses to Statement 1 against Statement 2, suggests that students responded to each of the two differently.

It may suggest that prioritising increasing the vegetarian options over the vegan options may improve the customer satisfaction ($t_{\text{new1}} = 6.33, v_{\text{new1}} = 98.028, p < 0.001$). Further investigation would be required to explore the attitudes further.

8.3 Methodology

Monte-Carlo methods are used to compare test statistics for two samples which include both paired and unpaired observations, using methodology outlined in Chapter 4. The comparison is undertaken by discretising each of the x_{ij} Normal variates to a five point scale and a seven point scale, over a range of sample sizes and correlation coefficients.

Without loss of generality, for a five point Likert scale the options are numbered from -2 to 2. The Likert-style responses y_{ij} are generated using the cut-points as follows:

$$y_{ij} = \left\{ \begin{array}{ccc} 2 & \text{if} & x_{ij} > 0.8416 \\ 1 & \text{if} & 0.2533 \leq x_{ij} \leq 0.8416 \\ 0 & \text{if} & -0.2533 \leq x_{ij} \leq 0.2533 \\ -1 & \text{if} & -0.8416 \leq x_{ij} \leq -0.2533 \\ -2 & \text{if} & x_{ij} < -0.8416 \end{array} \right\}$$

For a seven-point Likert-like scale, the responses y_{ij} numbered from -3 to 3 are generated using the cut-points as follows:

$$y_{ij} = \left\{ \begin{array}{ccc} 3 & \text{if} & x_{ij} > 1.6757 \\ 2 & \text{if} & 0.5659 \leq x_{ij} \leq 1.6757 \\ 1 & \text{if} & 0.1800 \leq x_{ij} \leq 0.5659 \\ 0 & \text{if} & -0.1800 \leq x_{ij} \leq 0.1800 \\ -1 & \text{if} & -0.5659 \leq x_{ij} \leq -0.1800 \\ -2 & \text{if} & -1.6757 \leq x_{ij} \leq -0.5659 \\ -3 & \text{if} & x_{ij} < -1.6757 \end{array} \right\}$$

The cut-points are calculated so that under $N(0, 1)$ the theoretical distribution of the responses is uniform. The median of Group 1 and the median of Group 2 are represented by η_1 and η_2 respectively.

Figure 8.1 gives the marginal density for Group 1, the marginal density for Group 2, and the joint density for responses. These illustrate the approximate distribution of responses a seven point scale where $\eta_1 = -1$ and $\eta_2 = 2$.

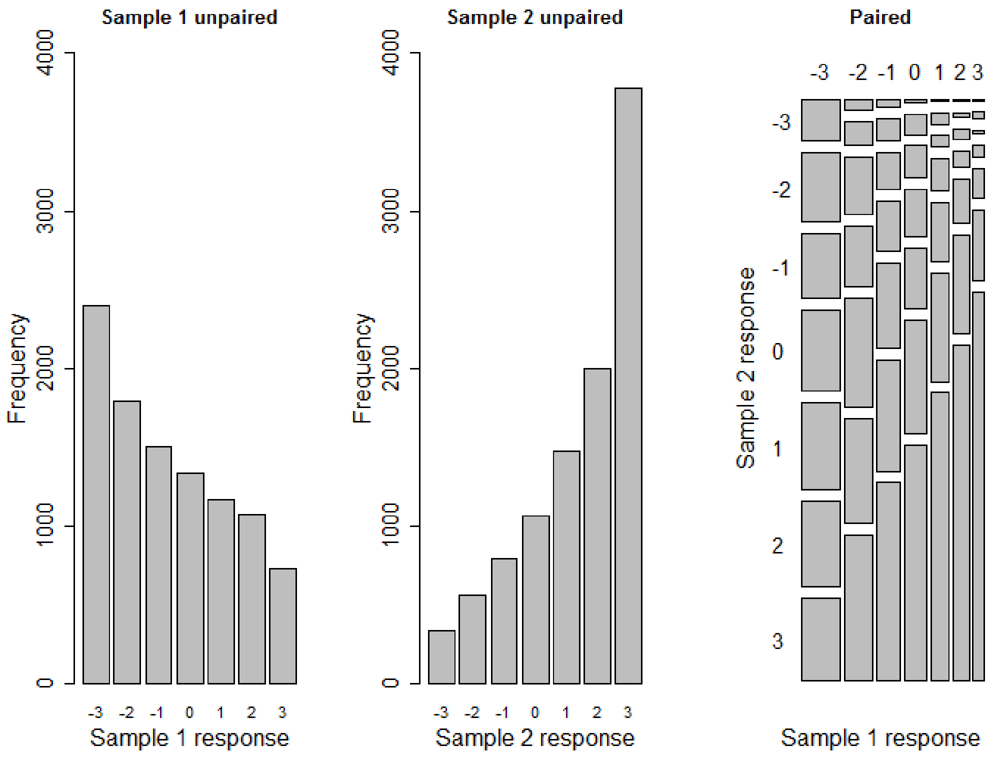


Figure 8.1: Theoretical distributions of the observed responses on a seven option Likert question, $\eta_1 = -1, \eta_2 = 2$, calculated based on $n_a = 10,000, n_b = 10,000, n_c = 5,000, \rho = 0.5$.

The complete list of the scenarios considered is given in Table 8.2. The scenarios considered encompass each integer combination of η_1 and η_2. For example, by symmetry the Type I error rate when $\eta_1 = \eta_2 = 1$ is equivalent to the Type I error rate when $\eta_1 = \eta_2 = -1$.

Simulations are performed for each scenario in a factorial design with $\rho = \{0, 0.25, 0.5, 0.75\}$ and sample size of $\{5, 10, 20, 30\}$ in each of n_a, n_b,

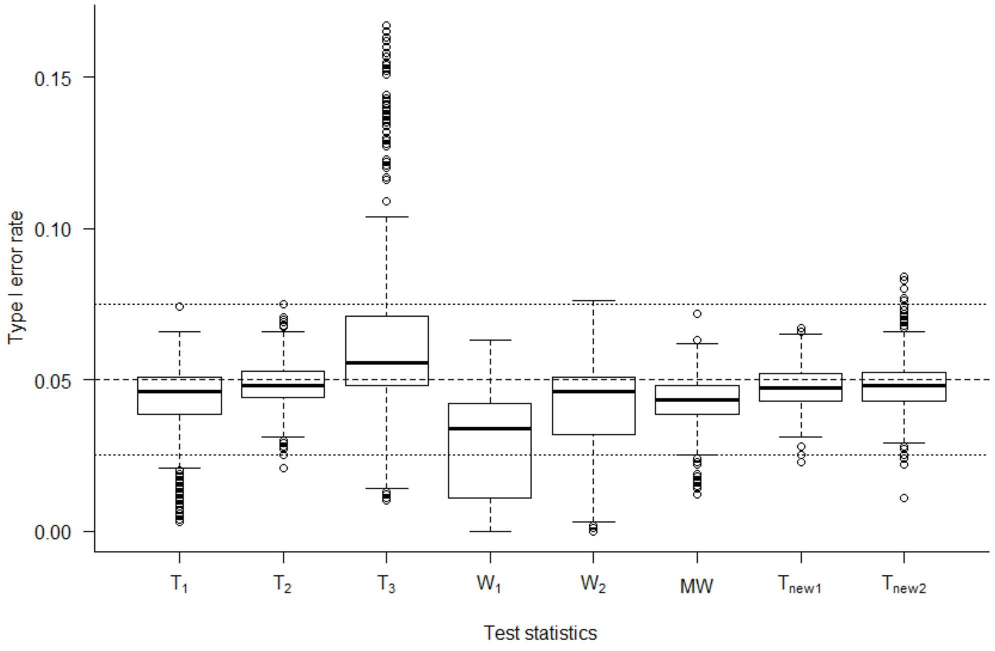
Table 8.2: Simulation scenarios

Scenario	Options	True	μ_1	μ_2	η_1	η_2
i	Five	H_0	0	0	0	0
ii	Five	H_0	0.5244	0.5244	1	1
iii	Five	H_0	1.2816	1.2816	2	2
iv	Five	H_1	0	0.5244	0	1
v	Five	H_1	0	1.2816	0	2
vi	Five	H_1	0.5244	1.2816	1	2
vii	Five	H_1	-0.5244	0.5244	-1	1
viii	Five	H_1	-0.5244	1.2816	-1	2
ix	Five	H_1	-1.2816	1.2816	-2	2
x	Seven	H_0	0	0	0	0
xi	Seven	H_0	0.3661	0.3661	1	1
xii	Seven	H_0	0.7916	0.7916	2	2
xiii	Seven	H_0	1.4652	1.4652	3	3
xiv	Seven	H_1	0	0.3661	0	1
xv	Seven	H_1	0	0.7916	0	2
xvi	Seven	H_1	0	1.4652	0	3
xvii	Seven	H_1	0.3661	0.7916	1	2
xviii	Seven	H_1	0.3661	1.4652	1	3
xix	Seven	H_1	0.7916	1.4652	2	3
xx	Seven	H_1	-0.3661	0.3661	-1	1
xxi	Seven	H_1	-0.3661	0.7916	-1	2
xxii	Seven	H_1	-0.3661	1.4652	-1	3
xxiii	Seven	H_1	-0.7916	0.7916	-2	2
xxiv	Seven	H_1	-0.7916	1.4652	-2	3
xxv	Seven	H_1	-1.4652	1.4652	-3	3

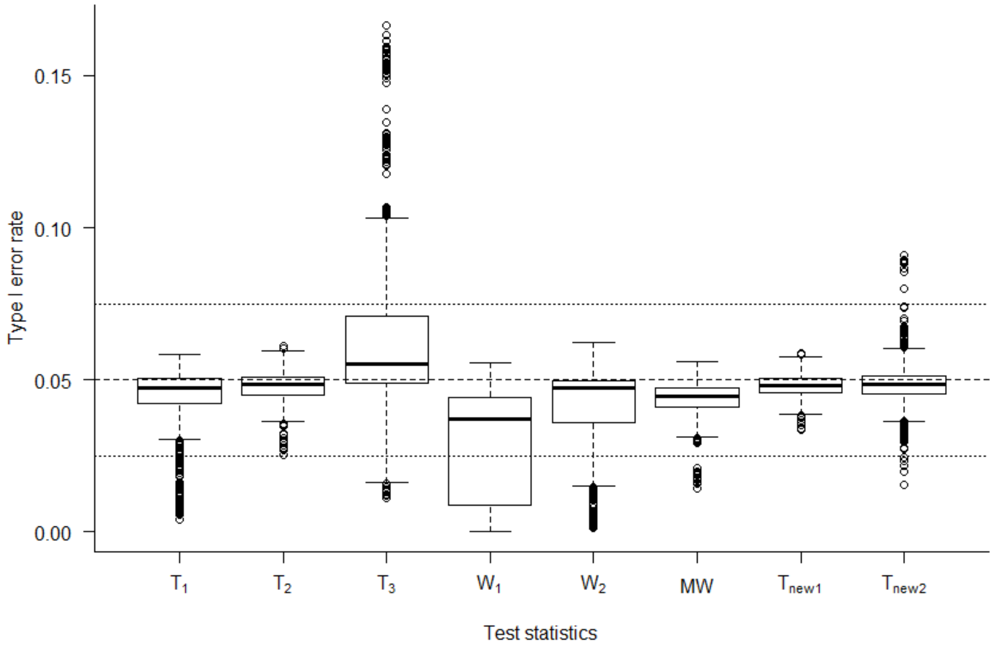
and n_c. For each scenario and parameter combination, 10,000 iterations are performed, and for each repetition the null hypothesis is assessed at the $\alpha =$ 5% significance level. Naive standard tests, T_1, T_2, T_3, W_1, and Pratt's test, W_2, are considered alongside the parametric partially overlapping samples t-tests, T_{new1} and T_{new2}.

8.4 Results: Type I error rates

Type I error rates where $\eta_1 = \eta_2$ are given in Figure 8.2.



(a) Five point scale



(b) Seven point scale

Figure 8.2: Type I error rates for each of the test statistics, ordinal data.

Figure 8.2 shows a similar pattern in Type I error rates whether a five point scale or a seven point scale is used. It can be seen that with the exception of T_3, and arguably T_{new2}, the test statistics are generally within Bradley's liberal robustness criteria, or are conservative. They are therefore consistent with Guo and Luh (2000) Type I error robustness criteria.

The results support the literature that the independent samples t-test assuming equal variances, T_2, is Type I error robust for ordinal data. However, there is evidence to suggest that this is not the case for the form of the independent samples t-test not assuming equal variances, T_3. When comparing two independent samples on a five option Likert question, there is little practical difference between the independent samples t-test and the Mann-Whitney test.

It should be noted that the test statistics making use of only the independent observations have paired samples discarded, thus have an approximate zero correlation structure in the simulations detailed here. Derrick and White (2017) considered a paired simulation design where observations were not discarded, and found that the test statistics assuming independent samples are biased. In other words, the tests T_2^{all}, T_3^{all} and the Mann-Whitney test using all of the available data ignoring the pairing are not Type I error robust.

The remainder of this chapter focuses on tests which make use of the paired information. Type I error rates for each of the parameter combinations within the simulation design can be seen in Figure 8.3 for the five point design, and Figure 8.4 for the seven point design.

Figure 8.3 shows that Pratt's test retains Type I error robustness better than the Wilcoxon test, however Pratt's test is not Type I error robust for the smallest sample size within the simulation design. It can also be seen that the paired samples t-test is not Type I error robust for the smallest sample size within the simulation design. Both partially overlapping samples t-tests appear to maintain reasonable Type I error robustness.

Figure 8.3 and Figure 8.4 shows that each of the test statistics echo similar Type I error robustness whether a five point scale or a seven point scale is used.

Closer inspection of the parameter combinations that exceed the upper

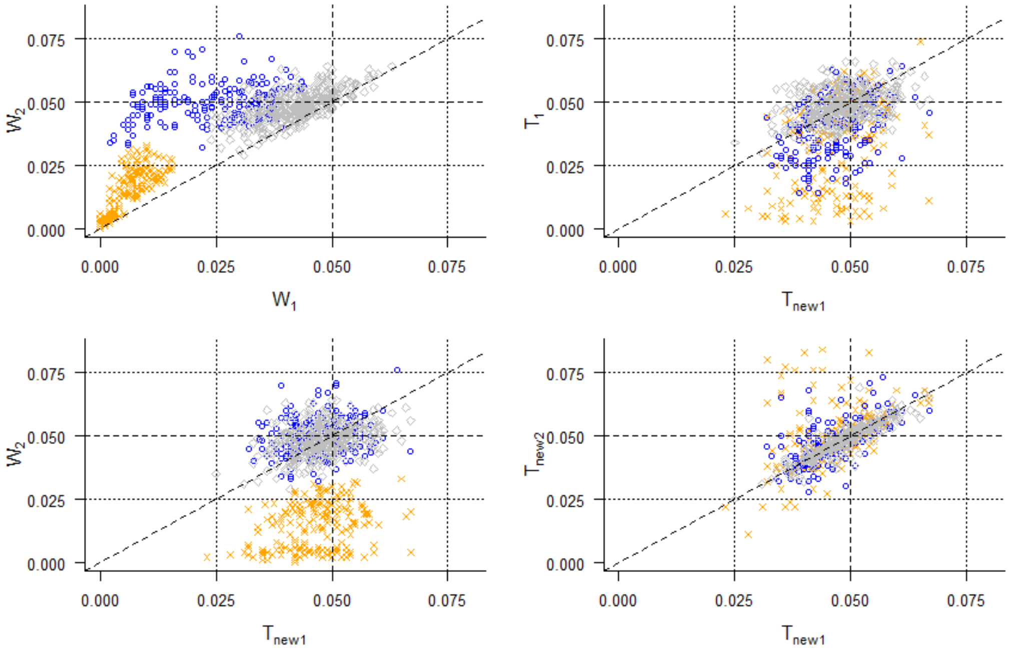


Figure 8.3: Type I error rates for each parameter combination of each of the scenarios under the five option Likert question simulation design. The symbols $\{\times, \circ, \diamond\}$ represent the sample sizes $\{n_c = 5,\ n_c = 10,\ n_c \geq 20\ \}$ respectively.

limit of Bradley's Type I error robustness, reveals that they are in the scenarios with the highest degree of skew, where $\eta_1 = \eta_2 = 3$, and the smaller sample sizes.

An alternative way of quantifying robustness put forward by Derrick and White (2018), is the value of π such that $(1-\pi)*100$ percent of Type I error rates are within $\pi \times 100$ percent of α. Across the five point scale simulation design the paired samples t-test is 74.7% robust. This means that 74.7% of the Type I error rates for parameter combinations within the simulation design are within 25.3% of the nominal Type I error rate. The Wilcoxon test is 62.3% robust, Pratt's test is 70.6% robust, T_{new1} is 82.1% robust and T_{new2} is 81.4% robust. This is an intuitive and simple way of quantifying robustness, however the robustness percentage depends on the parameters used within the simulation design. Across the seven point scale simulation design the pattern is the same, with T_{new2} being the most robust at 87.1%,

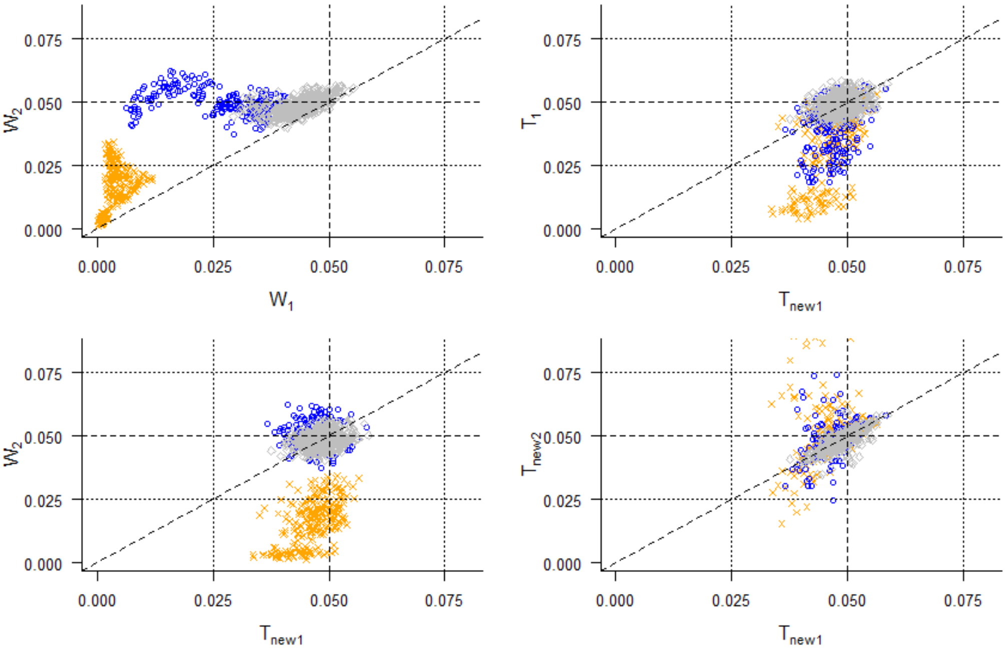


Figure 8.4: Type I error rates for each parameter combination of each of the scenarios under the seven option Likert question simulation design. The symbols $\{\times, \circ, \diamond\}$ represent the sample sizes $\{n_c = 5,\ n_c = 10,\ n_c \geq 20\ \}$ respectively.

closely followed by T_{new1} at 84.8%.

8.5 Results: Power

For all parameter combinations where $\eta_1 \neq \eta_2$, the percentage of iterations where the null hypothesis is rejected represents the power of the test under those parameter conditions. Table 8.3 summarises the power for each scenario using each test statistic, averaged over all parameter combinations.

Table 8.3: Power, ordinal data.

scenario	η_1	η_2	T_1	W_1	W_2	T_{new1}	T_{new2}
iv	0	1	0.380	0.327	0.358	0.530	0.524
v	0	2	0.788	0.679	0.740	0.972	0.966
vi	1	2	0.503	0.440	0.492	0.734	0.723
vii	-1	1	0.744	0.642	0.698	0.937	0.931
viii	-1	2	0.916	0.746	0.855	0.998	0.998
ix	2	-2	0.983	0.779	0.946	1.000	1.000
average			0.747	0.623	0.706	0.882	0.877
xiv	0	1	0.244	0.206	0.227	0.323	0.319
xv	0	2	0.611	0.536	0.579	0.821	0.812
xvi	0	3	0.840	0.713	0.794	0.989	0.986
xvii	1	2	0.285	0.244	0.269	0.392	0.387
xviii	1	3	0.713	0.628	0.682	0.933	0.923
xix	2	3	0.433	0.384	0.431	0.646	0.634
xx	-1	1	0.582	0.509	0.550	0.782	0.774
xxi	-1	2	0.794	0.677	0.752	0.964	0.959
xxii	-1	3	0.918	0.743	0.871	0.999	0.998
xxiii	-2	2	0.899	0.733	0.856	0.996	0.995
xxiv	-2	3	0.967	0.754	0.931	1.000	1.000
xxv	-3	3	0.994	0.773	0.977	1.000	1.000
average			0.690	0.575	0.660	0.820	0.816

Table 8.3 shows that the partially overlapping samples *t*-tests perform similarly to each other, and consistently out-perform the tests that discard data.

For a paired samples design, if a non-parametric test is favoured then Pratt's test should be used over the Wilcoxon test, because it has better Type I error robustness and power properties. However, the evidence presented in Figure 8.2 and Table 8.3 elicit no reason why a non-parametric test would be favoured over the paired samples t-test.

Table 8.3 shows that when the difference between the two groups is only one point on the ordinal scale, and one of the groups has a neutral median, none of the tests statistics considered have high power. For a difference of one point on the ordinal scale, a significant effect is more likely to be observed when both groups average responses are either positive or negative. Assuming that the scale represents interval data of equal difference between each point, it is a limitation that the power is not equal for each comparison between groups with a one point difference on an ordinal scale.

The power difference between T_{new1} and the paired samples t-test is represented visually in Figure 8.5. This demonstrates that the scenarios where the power gain of T_{new1} is greatest are when the group responses are relatively similar.

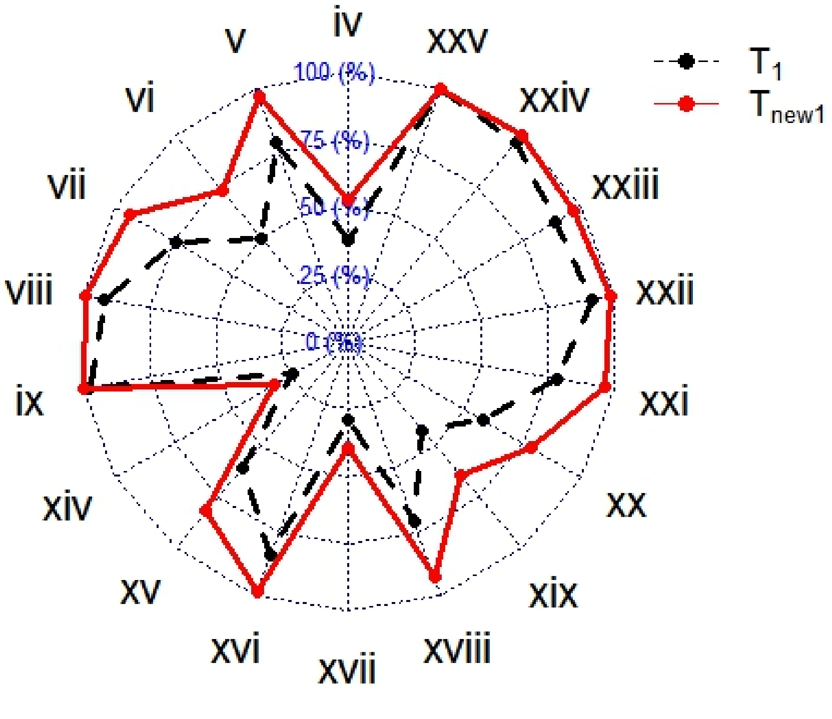


Figure 8.5: Radar chart showing the average power for each scenario, using the paired samples t-test and T_{new1}.

Figure 8.6 shows the power for each of the parameter combinations within the simulation design for a five point scale.

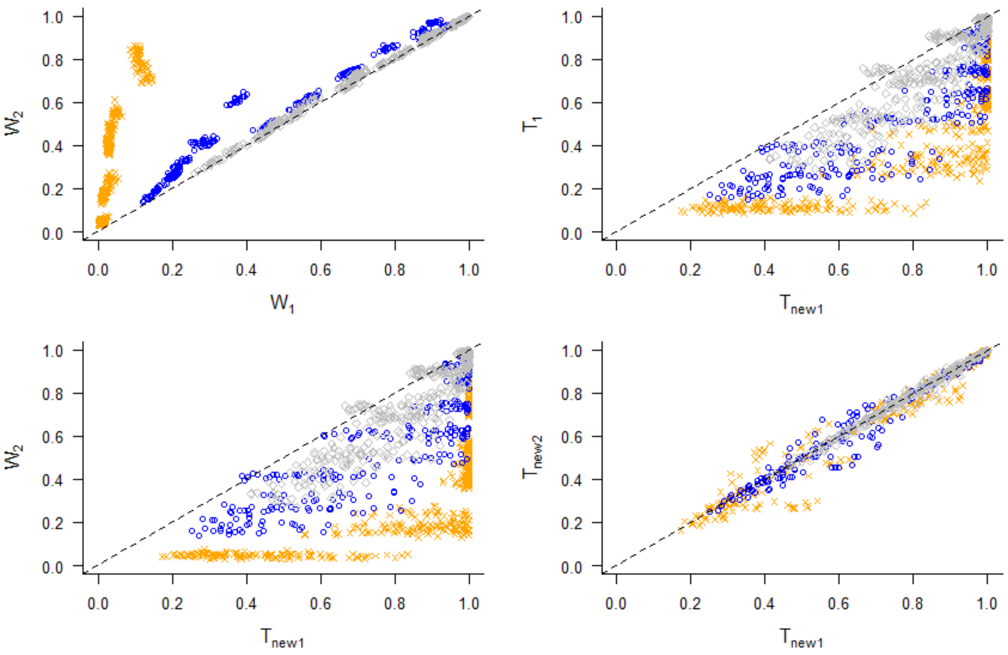


Figure 8.6: Power for each parameter combination of each of the scenarios under the five option Likert question simulation design. The symbols {×, ○, ◇} represent the sample sizes $\{n_c = 5,\, n_c = 10,\, n_c \geq 20\}$ respectively.

Figure 8.6 shows that T_{new1} is always at least as powerful as T_1, W_1, and W_2. Pratt's test should be chosen over the Wilcoxon test, however if $n_c \geq 20$ it makes the choice between the two largely academic. As sample size increases, the Wilcoxon test becomes large enough to compensate for discarded zeroes. Power comparisons where $n_c = 5$ suggest that partially overlapping samples t-tests may be particularly useful when sample sizes are small.

There is an apparent strong correlation between T_{new1} and T_{new2}, suggesting that they share similar power properties across the parameter combinations and scenarios simulated. Findings for a seven point scale (not displayed) are the same as for a five point scale. For each test statistic relative to each other, a similar pattern across the range of parameter combinations is observed whether a five point scale or a seven point ordinal scale is used.

8.6 Summary

A comparison of test statistics has been performed for ordinal data, specifically for responses from either a five point Likert question, or a seven point Likert style question. Assuming the responses represent interval data, standard approaches such as the paired samples t-test or Pratt's test may not be inappropriate. However, these standard approaches discard the independent observations and as such are less than ideal, particularly if the sample sizes are small.

Across a range of sample sizes and correlation coefficients, T_{new1} and T_{new2} offer Type I error robust alternatives for the analysis of two partially overlapping samples. The test statistic assuming equal variances, T_{new1}, maintains the nominal Type I error rate better than T_{new2}. In addition T_{new1} is more powerful than standard approaches that discard data, and also marginally more powerful than T_{new2}. For a small difference between the two groups, greater power is obtained when responses in both groups are in the same direction.

T_{new1} is the recommended test for analysing data recorded on an ordinal scale when partially overlapping samples are present, and is particularly useful when sample sizes are small.

Chapter 9

The comparison of two samples, with dichotomous data

In this chapter, approaches for comparing two proportions when partially overlapping samples are present are explored. Six new test statistics are presented and compared against standard test statistics and other alternatives in the literature. The main concepts and results of this chapter are summarised in Derrick et al. (2015).

9.1 Current strategies for comparing proportions

A dichotomous dependent variable is probably the most commonly occurring scenario studied in biological research (Gart, 1971). Tests for comparing two sample proportions of a dichotomous dependent variable with either two independent or two dependent samples are long established. Let π_1 and π_2 be the proportions of interest for two populations or distributions. The hypothesis being tested is $H_0 : \pi_1 = \pi_2$ against $H_1 : \pi_1 \neq \pi_2$.

Historically, when analysing partially overlapping samples, a practitioner will choose between discarding the paired observations, or discarding the independent observations, and proceeding to perform the corresponding 'standard' test. It is likely the decision will be based on the sample sizes of the independent and paired observations. Existing 'standard' approaches include; discarding all paired observations and performing Pearson's Chi square test of association on the unpaired data, or discarding all unpaired observations and performing McNemar's test on the paired data. Alternatively, techniques for combining p-values for separate tests for paired samples and unpaired samples could be applied. Fisher's inverse Chi-square method, Tippett's test and Stouffer's test are more powerful than techniques that discard data (Samawi and Vogel, 2011). However, it should be noted that Samawi and Vogel (2011) did not consider Type I error rates. The scenario outlined in this chapter is not immune to other ad-hoc approaches for using all available data such as randomly pairing any unpaired observations, or ignoring any pairing. This reiterates the need for research into statistically valid approaches.

The 'standard' approaches for tests of equal proportions are inappropriate for similar reasons as traditional tests for equal means are inappropriate. In fact it is more likely that these naive approaches result in violations of the assumptions of the tests if sample sizes are small. For the χ^2 test, an expected frequency ≥ 1 is required in every cell and an expected frequency ≥ 5 is required in at least 80% of the cells (Fisher, Marshall, and Mitchell, 2011). If the analysis were to be split between independent and dependent samples, at least 16 observations in each design would be advisable. As a 2×2 matrix is considered by the two standard approaches, a continuity correction is typically applied for small sample sizes.

These naive approaches are likely to have relatively low power for small samples when the number of discarded observations is large. A method of analysis for partially overlapping samples that takes into account the paired design but does not lose the unpaired information would therefore be beneficial.

Choi and Stablein (1982) performed a simulation study to assess approaches for the partially overlapping samples case, they ultimately recommended their test making use of all the available data as the best practical approach. Their proposal uses one combined test statistic, weighting the variance of the paired observations and the independent observations. The authors additionally considered an approach using maximum likelihood estimators for the proportions. The latter approach was found to be of little practical benefit in terms of Type I error rate or power. It was noted by Choi and Stablein (1982) that given the additional computation, the maximum likelihood solution is not a practical solution. Others have also considered maximum likelihood approaches. For example, Thomson (1995), using maximum likelihood estimators found their proposed procedure to perform similarly to that of Choi and Stablein (1982). However the solution by Thomson (1995) makes the assumption that the independent observations and the paired observations within a sample have equal proportions (Bland and Butland, 2011).

Samawi and Vogel (2011) proposed a further approach, using the theory from Tippett (1931). This is to test dependent data separately from the independent data and then compare the smallest p-value to the Bonferroni-Sidak lower bound i.e. $1-(1-\alpha)^{0.5}$. This method has the limitation of equal weights for both the independent and dependent samples.

Tang and Tang (2004) propose a test procedure which is a direct adaption of the approach proposed by Choi and Stablein (1982). But this adaption is found to violate Type I error robustness in scenarios considered when $n_a + n_b + 2n_c = 20$. The original test proposed by Choi and Stablein (1982) is found to be Type I error robust in this scenario.

Based on the existing literature, a solution to the partially overlapping samples case will have to outperform the best practical solution by Choi and Stablein (1982). Tang and Tang (2004, p. 81) concluded that 'there may exist other test statistics which give better asymptotic or unconditional exact performance'.

9.2 Definition of standard test statistics

Assuming a dichotomous dependent variable, where a comparison in proportions between two samples is required, the layout of frequencies for the paired observations and the independent observations is set out as per Table 9.1 and Table 9.2 respectively.

Table 9.1: Paired samples design for two samples with one dichotomous dependent variable.

	Response Sample 2		
Response Sample 1	Yes	No	Total
Yes	a	b	m
No	c	d	$n_c - m$
Total	k	$n_c - k$	n_c

Table 9.2: Independent samples design for two samples with one dichotomous dependent variable.

	Response		
	Yes	No	Total
Sample 1	e	f	n_a
Sample 2	g	h	n_b

9.2.1 Discarding paired observations

For two independent samples as per Table 9.2, a Chi-square test of association is often performed. This test is displayed in standard textbooks in terms of χ_1^2. A Chi-square distribution on one degree of freedom is equivalent to the square of the z-distribution. Therefore under the null hypothesis an asymptotically $N(0,1)$ equivalent statistic is defined as:

$$z_1 = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n_a} + \frac{\hat{p}(1-\hat{p})}{n_b}}} \tag{9.1}$$

where $\hat{p}_1 = \frac{e}{n_a}$, $\hat{p}_2 = \frac{g}{n_b}$ and $\hat{p} = \frac{e+g}{n_a+n_b}$.

For small samples, Yates's correction is often performed to reduce the error in approximation. Yates's correction is given by:

$$z_2 = \sqrt{\frac{(n_a + n_b)\left(|eh - fg| - 0.5\,(n_a + n_b)\right)^2}{(e+g)\,(f+h)\,n_a n_b}}. \tag{9.2}$$

The statistic z_2 is referenced against the upper tail of the standard Normal distribution.

An alternative to this approach is Fisher's exact test. This is computationally more difficult. Furthermore, Fisher's exact test is shown to deviate from Type I error robustness (Berkson, 1978).

9.2.2 Discarding unpaired observations

For two dependent samples as per Table 9.1, McNemar's test is often performed. Under the null hypothesis, the asymptotically $N(0,1)$ equivalent to McNemar's test is:

$$z_3 = \frac{b-c}{\sqrt{b+c}}. \tag{9.3}$$

When the number of discordant pairs is small, a continuity correction is often performed. McNemar's test with continuity correction is equivalent to:

$$z_4 = \sqrt{\frac{(|b-c|-1)^2}{b+c}}. \tag{9.4}$$

The statistic z_4 is referenced against the upper tail of the standard Normal distribution.

Several methods have been proposed in the literature for calculating confidence intervals for McNemar's test. The method by Fay (2011) is used here and is available in the R package 'exact2x2'. In any event, odds ratios make the results of McNemar's test more challenging to interpret.

9.2.3 Combination of the independent and paired tests using all of the available data

Given that a naive test for the paired observations and a separate naive test for the independent observations can be performed, an extension to these techniques which makes use of all of the available data would be a combination of the two tests.

In terms of power, Fisher's test and Tippett's test are comparable to a weighted approach using sample size as the weights (Samawi and Vogel, 2011). Tippett's method and Fisher's method are not as effective as Stouffer's weighted z-score test (Kim et al., 2005). Stouffer's weighted z-score, for combining z_1 and z_3 is defined as:

$$z_5 = \frac{wz_1 + (1-w)z_3}{\sqrt{w^2 + (1-w)^2}} \text{ where } w = \frac{n_a + n_b}{2n_c + n_a + n_b} \tag{9.5}$$

Under the null hypothesis, the test statistic z_5 is asymptotically $N(0, 1)$

Many other procedures for combining separate p-values are available, but these are less effective than Stouffer's test (Whitlock, 2005).

The drawbacks of Stouffer's test are that it has issues in the interpretation and confidence intervals for the true difference in population proportions cannot be easily formed. Also, if the unpaired observations were only in one of the two samples, the calculation of z_1 and therefore z_5 is not possible.

9.3 Definition of alternative test statistics using of all of the available data

The following proposals are designed to overcome the drawbacks identified of the tests above. In these proposals, observations are not discarded, a single test is performed, and the test statistics are readily considered for the formation of confidence intervals.

9.3.1 Proposals using the phi correlation coefficient or the tetrachoric correlation coefficient.

It is proposed that a test statistic for comparing the difference in two proportions with two partially overlapping samples can be formed so that the overall estimated difference in proportions is divided by its combined standard error, i.e.

$$\frac{\overline{p}_1 - \overline{p}_2}{\sqrt{Var(\overline{p}_1) + Var(\overline{p}_2) - 2Cov(\overline{p}_1, \overline{p}_2)}}$$

where $Var(p_1) = \dfrac{\overline{p}_1(1-\overline{p}_1)}{n_c + n_a}, \quad Var(p_2) = \dfrac{\overline{p}_2(1-\overline{p}_2)}{n_c + n_b},$

$$Cov(p_1, p_2) = \frac{\sqrt{\overline{p}_1(1-\overline{p}_1)}\sqrt{\overline{p}_2(1-\overline{p}_2)}n_c}{(n_c+n_a)(n_c+n_b)}.$$

Test statistics constructed in this manner facilitate the construction of confidence intervals, for example a 95% confidence interval θ is equivalent to:

$$\theta = (\overline{p}_1 - \overline{p}_2) \pm 1.96 \times \sqrt{Var(\overline{p}_1) + Var(\overline{p}_2) - 2Cov(\overline{p}_1, \overline{p}_2)}$$

Pearson's phi correlation coefficient or Pearson's tetrachoric correlation coefficient are often used for measuring the correlation r_x between dichotomous variables.

Pearson's phi correlation coefficient, r_1 is calculated as

$$r_1 = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}$$

The result of r_1 is numerically equivalent to Pearson's product-moment correlation coefficient and Spearman's rank correlation coefficient applied to Table 9.1, using binary outcomes '0' and '1' in the calculation (Rodgers and Nicewander, 1988). This suggests that r_1 is an appropriate correlation coefficient to consider.

Alternatively, assuming the underlying distribution is normal, a polychoric correlation coefficient may be considered. A special case of the polychoric correlation coefficient for two dichotomous samples is the tetrachoric correlation coefficient.

An approximation to the tetrachoric correlation coefficient, r_2 given by Digby (1983) is

$$r_2 = \frac{s-1}{s+1}, \text{ where } s = \left(\frac{ad}{bc}\right)^{0.7854}$$

Other approximations are available, however there is no conclusive evidence which is the most appropriate (Digby, 1983). In any event, r_1 is likely to be more practical than r_2, because if any of the observed paired frequencies are equal to zero then the calculation of r_2 is not possible.

Constructing a test statistic using correlation coefficients r_1 and r_2 respectively, the following test statistics are proposed:

$$z_6 = \frac{\overline{p}_1 - \overline{p}_2}{\sqrt{\frac{\overline{p}_1(1-\overline{p}_1)}{n_c+n_a} + \frac{\overline{p}_2(1-\overline{p}_2)}{n_c+n_b} - 2r_1\left(\frac{\sqrt{\overline{p}_1(1-\overline{p}_1)}\sqrt{\overline{p}_2(1-\overline{p}_2)}n_c}{(n_c+n_a)(n_c+n_b)}\right)}} \tag{9.6}$$

$$z_7 = \frac{\overline{p}_1 - \overline{p}_2}{\sqrt{\frac{\overline{p}_1(1-\overline{p}_1)}{n_c+n_a} + \frac{\overline{p}_2(1-\overline{p}_2)}{n_c+n_b} - 2r_2\left(\frac{\sqrt{\overline{p}_1(1-\overline{p}_1)}\sqrt{\overline{p}_2(1-\overline{p}_2)}n_c}{(n_c+n_a)(n_c+n_b)}\right)}} \tag{9.7}$$

where $\overline{p}_1 = \dfrac{a+b+e}{n_c+n_a}$ and $\overline{p}_2 = \dfrac{a+c+g}{n_c+n_b}$.

Under H_0, $\pi_1 = \pi_2 = \pi$, thus two additional test statistics considered are defined as:

$$z_8 = \frac{\overline{p}_1 - \overline{p}_2}{\sqrt{\frac{\overline{p}(1-\overline{p})}{n_c+n_a} + \frac{\overline{p}(1-\overline{p})}{n_c+n_b} - 2r_1\left(\frac{\sqrt{\overline{p}(1-\overline{p})}\sqrt{\overline{p}(1-\overline{p})}n_c}{(n_c+n_a)(n_c+n_b)}\right)}} \tag{9.8}$$

$$z_9 = \frac{\overline{p}_1 - \overline{p}_2}{\sqrt{\frac{\overline{p}(1-\overline{p})}{n_c+n_a} + \frac{\overline{p}(1-\overline{p})}{n_c+n_b} - 2r_2\left(\frac{\sqrt{\overline{p}(1-\overline{p})}\sqrt{\overline{p}(1-\overline{p})}n_c}{(n_c+n_a)(n_c+n_b)}\right)}} \tag{9.9}$$

where $\overline{p} = \dfrac{(n_a+n_c)\overline{p}_1 + (n_b+n_c)\overline{p}_2}{2n_c+n_a+n_b}$.

The test statistics z_6, z_7, z_8 and z_9 are referenced against the standard Normal distribution.

In the extreme scenario of $n_c = 0$, it is quickly verified that $z_8 = z_9 = z_1$.

Under H_0, in the extreme scenario of $n_a = n_b = 0$, as $n_c \to \infty$ then $z_8 \to z_3$. This is confirmed in the following proof:

If $n_a = n_b = 0$;

$$\overline{p}_1 = \frac{a+b}{n_c}, \overline{p}_2 = \frac{a+c}{n_c}, \overline{p}_1 - \overline{p}_2 = \frac{c-b}{n_c},$$

$$\overline{p} = \frac{(n_c)\overline{p}_1 + (n_c)\overline{p}_2}{2n_c} = \frac{2a+b+c}{2n_c},$$

$$1 - \overline{p} = \frac{(n_c)(1-\overline{p}_1) + (n_c)(1-\overline{p}_2)}{2n_c},$$

and so if $n_a = n_b = 0$;

$$z_8^2 = \frac{(\overline{p}_1 - \overline{p}_2)^2}{\frac{2\overline{p}(1-\overline{p})}{n_c}(1 - r_1)} = \frac{n_c(\overline{p}_1 - \overline{p}_2)^2}{2\overline{p}(1-\overline{p})(1 - r_1)}$$

Substituting $\overline{p}$ and $\overline{p}_1 - \overline{p}_2$ into z_8^2 gives:

$$z_8^2 = \frac{2n_c(c - \mathrm{b})^2}{(b + c + 2d)(b + c + 2\mathrm{a})(1 - r_1)}$$

Under H_0, as $n_c \to \infty$ then $c \to \mathrm{b}$ thus

$$z_8^2 = \frac{2n_c(c - \mathrm{b})^2}{(2a + 2b)(2b + 2d)(1 - r_1)}$$

Given $r_1 = \dfrac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}} = \dfrac{ad - b^2}{\sqrt{(a+b)(b+d)(a+b)(b+d)}}$
then as $n_c \to \infty, r_1 \to \dfrac{ad - b^2}{(b+d)(a+b)}$.

Substituting r_1 into z_8^2 it follows that:

$$z_8^2 = \frac{2n_c(c - \mathrm{b})^2}{4(b+d)(a+b)(1 - \frac{ad-b^2}{(b+d)(a+b)})}$$

$$= \frac{2n_c(c - \mathrm{b})^2}{4(b+d)(a+b)((b+d)(a+b) - \frac{ad-b^2}{(b+d)(a+b)})}$$

$$= \frac{2n_c(c - \mathrm{b})^2}{4(b(a + d + 2b))}$$

$$= \frac{2n_c(c - \mathrm{b})^2}{4bn_c}$$

$$= \frac{(c - \mathrm{b})^2}{2b} = \frac{(c - \mathrm{b})^2}{c + b}$$

$$\therefore z_8 = \frac{c - b}{\sqrt{c + b}} = \mathrm{z}_3$$

This property is not observed for z_9.

The properties of z_8 give support from a mathematical perspective as a valid test statistic due to the interpolation between two established statistical tests where overlapping samples are not present.

For paired data, the effect size for McNemar's test requires the odds ratio of discordant pairs. As the partially overlapping z-statistic itself includes no information on the number of discordant pairs, a standard calculation of the effect size for two independent groups rather than a standard calculation for effect size based on paired observations could act as a reasonable approximation. The recommended method for effect size for two independent proportions is as given in Cohen (1992), i.e. the absolute value of $\arcsin\sqrt{p_1} - \arcsin\sqrt{p_2}$.

9.3.2 Proposal by Choi and Stablein (1982)

Choi and Stablein (1982) proposed the following test statistic as the best practical solution for analysing partially overlapping samples:

$$z_{10} = \frac{\overline{p}_1 - \overline{p}_2}{\sqrt{\overline{p}(1-\overline{p})\left(\frac{\psi_1^2}{n_a} + \frac{(1-\psi_1)^2}{n_c} + \frac{\psi_2^2}{n_b} + \frac{(1-\psi_2)^2}{n_c}\right)}} \tag{9.10}$$

where $\psi_1 = \frac{n_a}{n_a+n_c}$, $\psi_2 = \frac{n_b}{n_b+n_c}$ and $D = \frac{(1-\psi_1)(1-\psi_2)((a/n_c)-\overline{p}^2)}{n_c}$.

The test statistic z_{10} is referenced against the standard Normal distribution.

The authors additionally offer an extension of how optimization of ψ_1 and ψ_2 could be achieved, but suggest that the additional complication is unnecessary and the difference in results is negligible. In common with the other statistics presented, z_{10} is computationally tractable, but it may be less easy to interpret, particularly if $\psi_1 + \psi_2 \neq 1$.

9.3.3 Proposals based on the formation of a new correlation coefficient.

Given the literature in the support of the Chi-square test and McNemar's test, a statistic which defaults to these in the extreme under both H_0 and

H_1 is derived so that:

$$z_{11} = \frac{\overline{p}_1 - \overline{p}_2}{\sqrt{\frac{\overline{p}_1(1-\overline{p}_1)}{n_c+n_a} + \frac{\overline{p}_2(1-\overline{p}_2)}{n_c+n_b} - 2r_3\left(\frac{\sqrt{\overline{p}_1(1-\overline{p}_1)}\sqrt{\overline{p}_2(1-\overline{p}_2)n_c}}{(n_c+n_a)(n_c+n_b)}\right)}} \tag{9.11}$$

$$z_{12} = \frac{\overline{p}_1 - \overline{p}_2}{\sqrt{\frac{\overline{p}(1-\overline{p})}{n_c+n_a} + \frac{\overline{p}(1-\overline{p})}{n_c+n_b} - 2r_3\left(\frac{\sqrt{\overline{p}(1-\overline{p})}\sqrt{\overline{p}(1-\overline{p})n_c}}{(n_c+n_a)(n_c+n_b)}\right)}} \tag{9.12}$$

Utilising the definitions of z_1 and z_3, here z_{12} is used to show the derivation of a new correlation coefficient r_3. A similar derivation is applied to achieve z_{11}, where z_1 and z_3 are defined in terms of $\overline{p}_1$ and $\overline{p}_2$, instead of $\overline{p}$.

Under H_0 and H_1, when $n_a = n_b = 0$, a statistic is required so that:

$$\frac{2n_c(\mathrm{c}-\mathrm{b})^2}{(b+c+2d)(b+c+2a)(1-r_3)} = \frac{(\mathrm{c}-\mathrm{b})^2}{(b+c)} = z_3$$

Therefore

$$(b+c) = \frac{(2a+2b)(2b+2d)(1-r_3)}{2n_c}$$

and so

$$(1-r_3) = \frac{2n_c(b+c)}{(b+c+2d)(b+c+2a)}$$

Hence

$$r_3 = \frac{(b+c+2d)(b+c+2a) - 2n_c(b+c)}{(b+c+2d)(b+c+2a)}$$

Under H_0 and H_1, when $n_c = 0$, it is confirmed that $r_3 = a+b+c+d = 0$ and so the test statistic z_{12} defaults to z_1.

9.4 Worked example

The objective of a seasonal affective disorder support group is to see if there is a difference in the quality of life for sufferers at two different times of the year. A binary response was required to the question whether sufferers were satisfied with life. Membership of the group remains fairly stable, but there is some natural turnover of membership over time. Responses were obtained for $n_c = 15$ paired observations and a further $n_a = 9$ and $n_b = 6$ independent observations. These are given in Table 9.3 and Table 9.4.

Table 9.3: Paired observations for worked example.

	Response Time 2		
Response Time 1	Yes	No	Total
Yes	8	1	9
No	3	3	6
Total	11	4	15

Table 9.4: Independent observations for worked example.

	Response		
	Yes	No	Total
Time 1	5	4	9
Time 2	6	0	6

The elements of the test statistics are calculated as: $\hat{p}_1 = 0.556, \hat{p}_2 = 1.000, \hat{p} = 0.733, \overline{p}_1 = 0.583, \overline{p}_2 = 0.810, \overline{p} = 0.689, r_1 = 0.431, r_2 = 0.673, r_3 = 0.400, w = 0.333, \psi_1 = 0.375, \psi_2 = 0.286, D = 0.002$.

If discarding unpaired observations, due to the small number of discordant pairs, McNemar's test with correction is performed. The odds ratio is 1/3. The 95% confidence interval for the true odds ratio is (0.006, 4.151), thus there is no significant change in quality of life between the two response times.

The results for the test statistics making use of all available data are given in Table 9.5. The results in Table 9.5 are presented in the format above for consistency with Derrick et al. (2015). However, the implicit researcher aim is to demonstrate if there is an improvement in quality of life over time following the intervention. For ease of interpretation in similar research, the analyses could be performed so that it is $\overline{p}_2 - \overline{p}_1$ that is calculated, thus an odds ratio greater than 1 and positive z-values are desirable.

At the 5% significance level, whether or not the null hypothesis is rejected depends on the test performed. It is of note that the significant differences

Table 9.5: Summarised results of worked example.

statistic	z-score	p-value	confidence interval
z_6	-2.023	0.043	(-0.445,-0.007)
z_7	-2.295	0.022	(-0.419,-0.033)
z_8	-1.937	0.053	(-0.455,0.003)
z_9	-2.202	0.028	(-0.427,-0.025)
z_{10}	-1.809	0.070	(-0.471,0.019)
z_{11}	-1.995	0.046	(-0.448,-0.004)
z_{12}	-1.909	0.056	(-0.458,0.006)

arise only with tests introduced in this chapter, z_6, z_7, z_9, and z_{11}. The effect size is 0.250, small effect, with a greater frequency of respondents reporting that they are satisfied with life after attending the support group.

Although the statistical conclusions differ for this particular example, the numeric difference between many of the tests is small. To consider further the situations where differences between the test statistics might arise, simulations are performed.

9.5 Simulation design and results

A comprehensive set of simulations with varying sample sizes, correlation coefficients, and population proportions is given in Section 9.5.1. In Section the empirical properties of the new correlation coefficient is assessed. In Section 9.5.3 the Type I error rates for each of the test statistics is considered. For the Type I error robust statistics, power and confidence interval coverage are considered in Section 9.5.4 and Section 9.5.5.

9.5.1 Simulation design

Normal deviates generated as per Box and Muller (1958) are transformed into binary outcomes using critical values from the Normal distribution to obtain the desired π. The simulation design is summarised in Table 9.6.

The values of π are restricted to ≤ 0.5 due to the proposed statistics being palindromic invariant with respect to π and $1 - \pi$. Negative ρ is considered

Table 9.6: Simulation parameters, dichotomous data.

Parameter	**Values**
π_1	0.15, 0.30, 0.45, 0.50
π_2	0.15, 0.30, 0.45, 0.50
n_a	5, 10, 30, 50, 100, 500
n_b	5, 10, 30, 50, 100, 500
n_c	5, 10, 30, 50, 100, 500
ρ	-0.75, -0.50, -0.25, 0.00, 0.25, 0.50, 0.75

for theoretical interest, although $\rho < 0$ is less likely to occur in practical applications.

9.5.2 Comparison of correlation coefficients

The value of the new correlation coefficient, r_3, relative to the corresponding value of r_1 is explored. Figure 9.1 illustrates the relationship for seven parameter combinations within the simulation design.

It can be seen that $-1 \leq r_3 \leq 1$, thus fulfills typical definitions of a correlation coefficient. Further exploration shows that if $b = c$ then $r_3 = r_1$. However if $b \neq c$ then $r_3 \leq r_1$. Another interesting property of r_3 is that its value does not change with differing b and c when the total $b + c$ remains constant. The implication of this is that r_3 can be seen as measure of the relationship between the number of concordant pairs and the number of discordant pairs. Therefore, r_3 is proposed as a competing correlation coefficient for measuring the relationship in two samples where the dependent variable is dichotomous.

9.5.3 Type I error rates

Under the null hypothesis, for a selected parameter combination, the distribution of the p-values is assessed for uniformity. P-P plots for each of the test statistics demonstrate that the p-values that are not uniformly distributed, particularly at the upper tail. Due to the nature of comparing dichotomous variables, the calculated z-values for small samples are often equivalent to

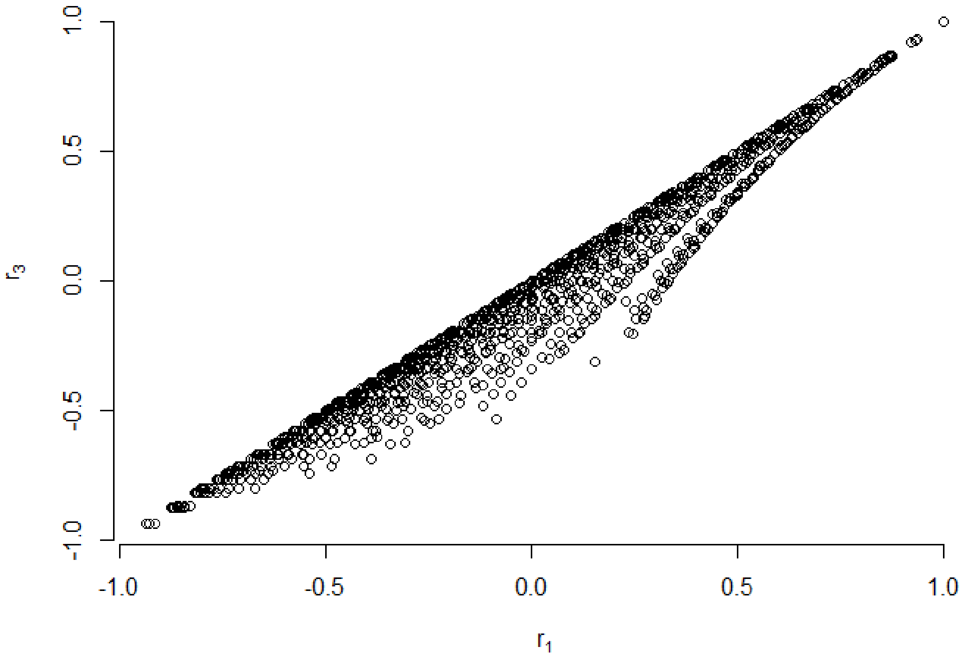


Figure 9.1: Comparison of new correlation coefficient r_3 against r_1 for $n_c = 30, \pi_1 = 0.5, \pi_2 = 0.3$ and $\rho = \{0.75, 0.5, 0.25, 0, -.025, -0.5, -0.75\}$.

zero. This is observed with the standard tests and the alternative proposed tests. This is what results in the phenomenon at the upper tail. This is demonstrated in Figure 9.2 for a selection of the test statistics.

Figure 9.2 suggests that z_8, z_{10} and z_{12} are more Type I error robust than the 'standard' options, because they appear to deviate less from uniformity in the lower tails also. For the assessment of robustness of Type I errors, the deviation in the upper tail is not majorly important because it is the lower tail that is of interest. The deviation in the upper tail may be problematic for clinicians performing equivalence and non-inferiority testing. These studies require large sample sizes (Da Silva, Logan, and Klein, 2009). Investigation of additional P-P plots from the simulation design (not displayed) demonstrate asymptotic Type I error robustness as sample sizes increase.

Under H_0, 10,000 replicates are obtained for $4 \times 5 \times 5 \times 5 \times 7 = 3500$

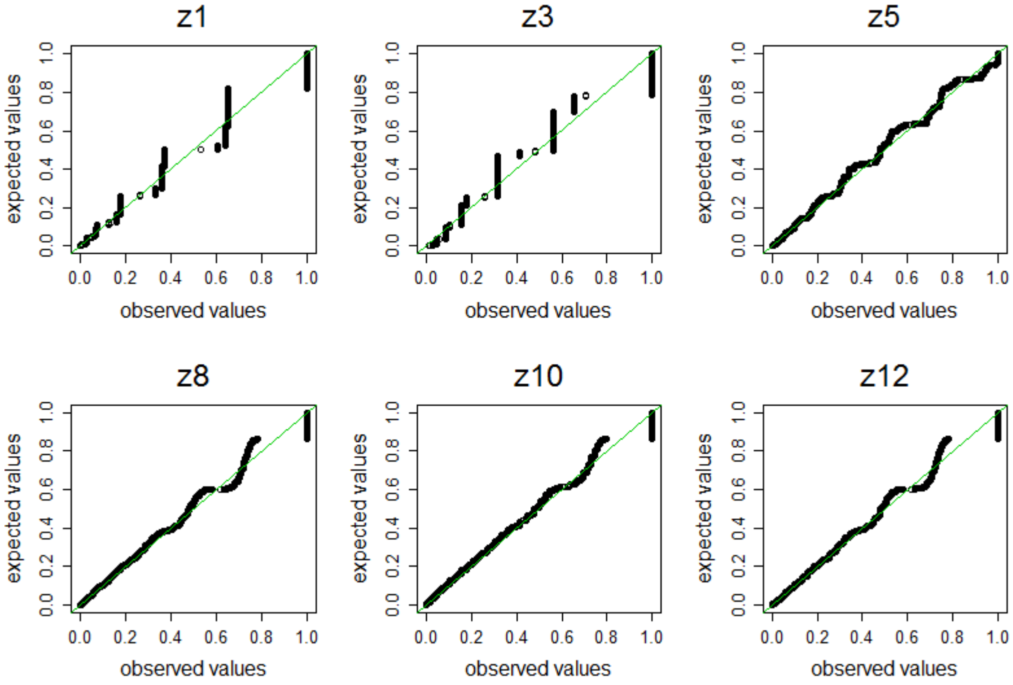


Figure 9.2: P-P plots for selected test statistics, where $\rho = 0.75$, $n_a = n_b = n_c = 10$, $\pi_1 = \pi_2 = 0.5$.

scenarios. For assessment against Bradley's liberal Type I error robustness criteria, Figure 9.3 shows the Type I error rates for all scenarios where $\pi_1 = \pi_2$ using $\alpha = 0.05$.

As may be anticipated, z_1 maintains Bradley's liberal Type I error robustness, because pairs are ignored. Similarly, z_3 performs as anticipated because the unpaired observations are ignored. The deviations from robustness for z_3 arise when the proportion of success is small, the paired sample size is small, and the correlation is high. Crucially, the deviations from Type I error robustness of z_3 are conservative, resulting in less false-positives, as such the test statistic may be acceptable in practical research.

The corrected statistics using naive approaches discarding observations, z_2 and z_4, give Type I error rates below the nominal α, particularly with small sample sizes. These findings echo the work by Ury and Fleiss (1980), it is concluded that the corrected statistics do not provide a robust solution

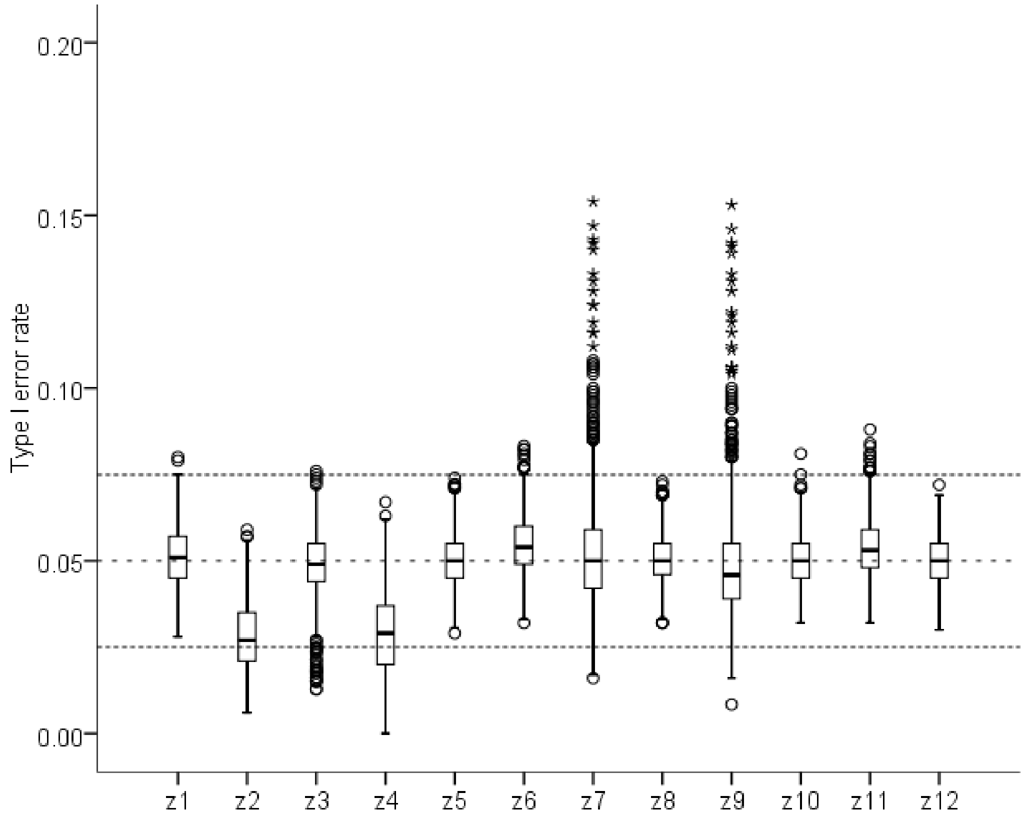


Figure 9.3: Type I error rates for each test statistic over all combinations.

to the partially overlapping samples problem.

The statistics using the phi correlation coefficient, z_6 and z_8, are generally Type I error robust. However, for z_6 there is some deviation from Bradley's liberal Type I error robustness criteria. Further inspection of the results shows that this deviation from liberal robustness occurs when $\min\{n_a,\ n_b,\ n_c\}$ is small, $\max\{n_a,\ n_b,\ n_c\} - \min\{n_a,\ n_b,\ n_c\}$ is large and $\rho < 0$. In these scenarios the impact of this is that z_6 is not Type I error robust and results in a high likelihood of false-positives. It is therefore concluded that z_6 does not universally provide a Type I error robust solution to the partially overlapping samples situation.

The statistics using r_2, namely z_7 and z_9, have more variability in Type I errors than the statistics that use r_1. These statistics using the tetrachoric correlation coefficient inflate the Type I error when $\rho > 0.25$ and n_c is large.

When $\min\{n_a,\ n_b,\ n_c\}$ is small, these test statistics are conservative. A test statistic that performs consistently is favoured for practical use, therefore z_7 and z_9 do not provide Type I error robust solutions to the partially overlapping samples situation.

The statistics using r_3 or r_1 and incorporating $\overline{p}$, i.e. z_8 and z_{12}, fulfill liberal robustness criteria throughout the entire simulation design.

The proposal by Choi and Stablein (1982), z_{10}, maintains liberal robustness except in the scenario where $\rho = 0.75$ and the smallest sample size simulated $n_a = n_b = n_c = 5$. Since this is only one combination it is likely due to the nature of simulated data. Additional ad-hoc runs of the simulation with small sample sizes indicates that z_{10} demonstrates liberal robustness.

A total of four statistics making use of all of the available data, z_5, z_8, z_{10} and z_{12}, are deemed to demonstrate liberal Type I error robustness. Further analysis of Type I error rates show near identical boxplots to Figure 9.3 when each of n_a, n_b, n_c, and ρ are fixed in turn. This means these statistics are robust across all sample sizes and correlations. These statistics are therefore considered for their power properties.

9.5.4 Power

The power of the test statistics that do not fail Type I error robustness criteria, is summarised in Table 9.7. For each of the test statistics, as the correlation increases from -0.75 through to 0.75, the power of the test increases. Similarly as sample sizes increase, the power of the test increases.

It can be seen from Table 9.7 that negative correlation results in lower power than the equivalent test statistic when positive correlation of the paired samples is present. Negative correlation therefore should be avoided where possible in the design stage of an experiment.

Clearly, z_5 is more powerful than the other standard tests, z_1 and z_3, but it is not as powerful as the alternative methods that make use of all the available data.

The power of z_8, z_{10} and z_{12} are similar. Inspection of the analyses indicate that the statistics are comparable across the various sample sizes and

Table 9.7: Power where $\pi_1 = 0.5$ averaged over all combinations sample sizes.

π_2	ρ	z_1	z_3	z_5	z_8	z_{10}	z_{12}
0.45	>0	0.095	0.173	0.208	0.221	0.221	0.218
0.45	0	0.095	0.133	0.168	0.186	0.186	0.184
0.45	<0	0.095	0.112	0.150	0.166	0.166	0.164
0.3	>0	0.509	0.653	0.807	0.856	0.855	0.854
0.3	0	0.509	0.569	0.772	0.828	0.827	0.826
0.3	<0	0.509	0.508	0.746	0.801	0.801	0.800
0.15	>0	0.843	0.874	0.975	0.989	0.989	0.986
0.15	0	0.843	0.834	0.970	0.985	0.986	0.986
0.15	<0	0.843	0.795	0.966	0.980	0.982	0.982

correlation, and power is maximised when $n_a = n_b$. Furthermore, increases in the number of pairs increases the power at a greater rate than equivalent increases in the number of independent observations.

9.5.5 Formation of confidence intervals

95% confidence intervals are generated for the Type I error robust, and most powerful statistics, z_8, z_{10} and z_{12}. The percentage of iterations where the true population difference appears within the confidence interval is the confidence interval coverage. In the case of a 95% confidence interval, the coverage should be approximately 95%. The confidence interval coverage is given in Figure 9.4.

All of the test statistics considered in Figure 9.4 demonstrate reasonable coverage for the true population difference $\pi_1 - \pi_2$. However, z_8 frequently performs closer to the desired 95% success rate relative to the other two test statistics. Taking this result into account, when the objective is to form a confidence interval, z_8 is recommended as the test statistic of choice in the partially overlapping samples case.

Caution should be expressed with these results because confidence intervals for proportions are known to be inadequate due to the discrete nature of proportions (Newcombe, 1998).

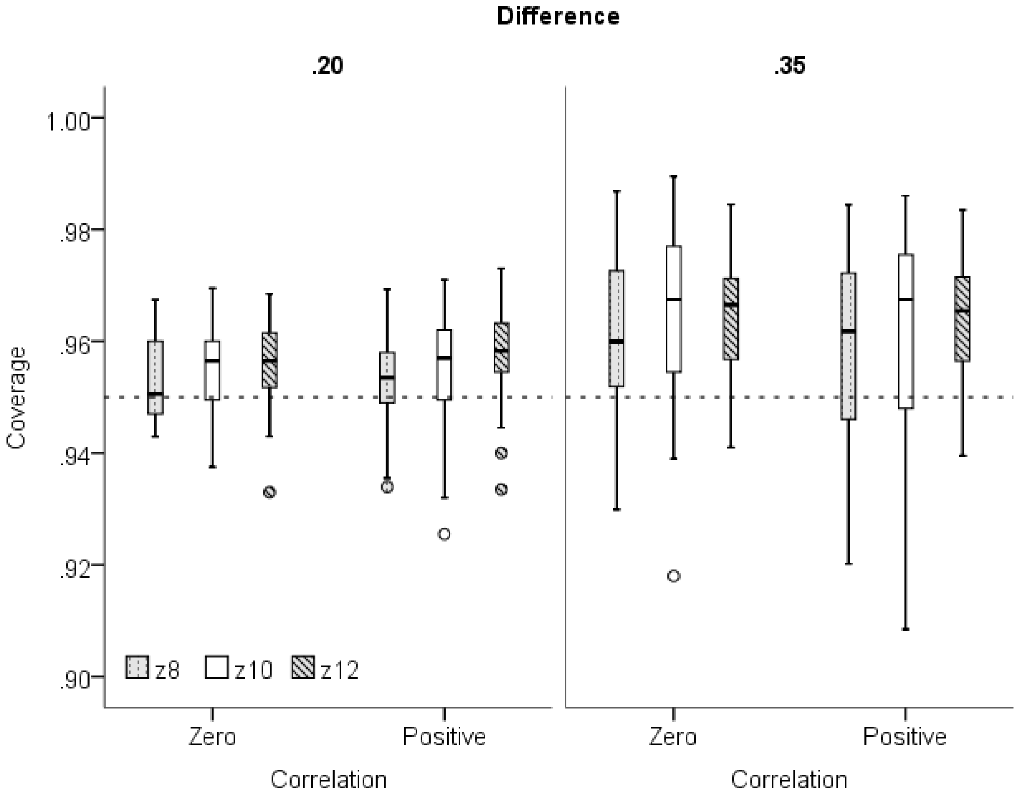


Figure 9.4: 95% confidence interval coverage of the population difference in proportions.

9.6 Comparison to *t*-test solution

The situation outlined could alternatively be viewed from the perspective of a comparison of the means for two partially overlapping samples. This approach may be reasonable, noting the slightly different form of the null hypothesis due to the explicit reference to means. Although it is standard procedure to report frequencies, observations could be coded with either a '1' for each 'success' or 'Yes', or a '0' for each 'failure' or 'No'. This approach can be considered as mathematically equivalent to observations recorded on a two-point ordinal scale.

Application to the seasonal affective disorder example, where the assumption of equal variances is reasonable, gives T_{new1} = -1.978, p-value = 0.060.

This represents weaker evidence of a treatment effect, relative to the other statistics proposed in this chapter making use of all the available data (with the exception of z_{10}).

The simulations as per Table 9.6 are repeated, performing the partially overlapping samples t-tests for comparative purposes.

Table 9.8 compares the Type I error rate and power for the partially overlapping samples t-tests and z_8, averaged over the simulation design for positive correlation. It can be see from this comparison that all three tests under consideration maintain good Type I error robustness, but there is apparent disparity in the power of the tests.

Table 9.8: Comparison of z_8 with T_{new1} and $T_{\text{new2}}, \rho > 0$.

	π_1	π_2	T_{new1}	T_{new2}	z_8
Type I error robustness	0.15	0.15	**0.047**	**0.060**	**0.048**
	0.3	0.3	**0.049**	**0.051**	**0.049**
	0.45	0.45	**0.049**	**0.050**	**0.050**
	0.5	0.5	**0.049**	**0.050**	**0.050**
Power	0.15	0.5	0.950	0.939	0.989
	0.3	0.5	0.757	0.748	0.855
	0.45	0.5	0.188	0.188	0.221

In conclusion, for the comparison of two dichotomous variables, performing the partially overlapping samples t-test is a valid solution, however the z_8 solution proposed in this chapter is recommended for its superior power.

9.7 R Package

Version 2.0 of the R package ‘partiallyoverlapping’ includes a function called ‘Prop.test’ to ease calculation of the test procedure $z8$.

9.7.1 Help manual

Extracts from the supporting help pages are given below to demonstrate how the function Prop.test within the ‘partiallyoverlapping’ package is used.

Description

The partially overlapping samples z-test for the comparison of two dichotomous samples.

Usage and default options

Prop.test(x1 = NULL, x2 = NULL, x3 = NULL, x4 = NULL, alternative = "two.sided", conf.level = NULL, stacked = FALSE)

Arguments

x1 a vector of unpaired observations in Sample 1 (or all observations in Sample 1 if stacked = "TRUE")

x2 a vector of unpaired observations in Sample 2 (or all observations in Sample 2 if stacked = "TRUE")

x3 a vector of paired observations in Sample 1 (not applicable if stacked = "TRUE")

x4 a vector of paired observations in Sample 2 (not applicable if stacked = "TRUE")

alternative a character string specifying the alternative hypothesis, must be one of "two.sided" (default), "greater" or "less".

conf.level confidence level of the interval.

stacked indicator of whether paired and unpaired observations are stacked within one vector ("TRUE"), or if specified as four separate vectors (default). Corresponding pairs should be given on the same row when "TRUE" is selected.

Values

statistic The value of the z-statistic

p.value The *p*-value for the test

estimate The estimated difference in proportions

conf.int A confidence interval for the difference in proportions appropriate to the specified alternative hypothesis

Example

[This is the example outlined in Section 9.4, taken from Derrick et al. (2015).]
The proportions for two groups, "a" and "b" are compared where the raw data "1", or "0" for each unit is recorded in a data frame.
15 paired observations are given first, followed by 9 independent observations in Sample 1, followed by 6 independent observations in Sample 2.
Independent observations and the paired samples stacked for each sample.

```
a<-c(1,1,1,1,1,1,1,1,1,0,0,0,0,0,0,
     1,1,1,1,1,0,0,0,0,NA,NA,NA,NA,NA,NA)

b<-c(1,1,1,1,1,1,1,1,0,1,1,1,0,0,0,
     NA,NA,NA,NA,NA,NA,NA,NA,NA,1,1,1,1,1,1)

Prop.test(a,b,stacked=TRUE,conf.level=.95)
```

Resulting output gives; p.value = 0.053, conf.int = (-0.455, 0.003).

9.7.2 Further application

Using the partially overlapping samples z-test, z_8, NeMoyer et al. (2018) were able to identify where the reasons given for raising competence to plead guilty differed between attorneys representing adult clients and juvenile clients. They found that mental illness of the client was statistically significant, it was given as a reason more frequently for juvenile clients than for adult clients.

9.8 Summary

Standard approaches for analysing the difference in proportions for a dichotomous variable with partially overlapping samples often discard some available information. If there is a large paired sample or a large unpaired sample, it may be reasonable in a practical environment to use the corresponding naive test. For small samples, the test statistics which discard data have inferior power properties to tests statistics that make use of all available data. These standard approaches and other ad-hoc approaches are less than desirable.

Combining the paired and independent samples z-scores using Stouffer's method is a more powerful standard approach, but leads to complications in interpretation, and does not readily extend to the creation of confidence intervals for differences in proportions. The tests introduced here, as well as the test outlined by Choi and Stablein (1982) are more powerful than the test statistics in common use.

The alternative tests proposed here, z_6, z_7, z_8, z_9, z_{11} and z_{12}, overcome the interpretation barrier, in addition confidence intervals can readily be formed.

Tests using the phi correlation coefficient, z_6 and z_8, are more robust than the equivalent tests introduced using the tetrachoric correlation coefficient, z_7 and z_9.

The most powerful valid tests are z_8, z_{10} and z_{12}. For ease of computation and intuitive construction, z_8 or z_{12} are recommended as the best practical solutions to the partially overlapping samples problem. The empirical evidence suggests that z_8 is marginally better suited for forming confidence intervals for the true population difference than z_{10} or z_{12}. Furthermore Pearson's correlation coefficient and associated direct descendants for different types of data are highly regarded and ingrained within statistics (Rodgers and Nicewander, 1988). Thus z_8 has relative simplicity in calculation, strong mathematical properties and provides ease of interpretation. In conclusion, z_8 is recommended as the best practical solution to the partially overlapping samples framework when comparing two proportions.

Chapter 10

The comparison of variances

Tests for equality of variances between two partially overlapping samples are explored. New solutions which make use of all of the available data are put forward. These new approaches are compared against approaches that discard either the paired observations or the independent observations. The approaches are assessed under equal variances and unequal variances for two samples taken from the same distribution, as seen in Derrick et al. (2018).

10.1 Background

Much of the literature regarding assessing variances evolves from the assumption of equal variances that is a requirement of many statistical tests (see Chapter 2.8). However, equality of variances may be of direct interest, for example to assess two treatments that have a similar mean efficacy, or a comparison of variances in human populations. Variance can be seen as an indicator of risk or uniformity (Parra-Frutos, 2009). Thus equality of variances can be of interest for a variety of reasons, from comparing stocks in the stock market, to comparing products in a quality control process. Tests for equal variances have wide ranging applications including areas in archaeology, environmental science, business and medical research (Gastwirth, Gel, and Miao, 2009). This chapter considers tests for equality of variances, where the comparison of two groups with respect to their variances is of direct interest.

In the two partially overlapping samples scenario, if the number of paired observations is relatively large and the number of independent observations is relatively small, a solution may be to discard independent observations and perform a test for equal variances based on the paired observations only. For the comparison of variances for paired data, the Pitman-Morgan test can be performed. The Pitman-Morgan test is widely regarded as the best test of equal variances for two paired samples under normality (Mudholkar, Wilding, and Mietlowski, 2003). However, the Pitman-Morgan test is not robust to violations of the normality assumption (Mudholkar, Wilding, and Mietlowski, 2003; Grambsch, 1994). For heavy tailed distributions, the Type I error rate of the Pitman-Morgan test is inflated (McCulloch, 1987; Wilcox, 2015).

Alternatively, if the number of independent observations is relatively large and the number of paired observations is relatively small, a solution may be to discard paired observations and perform a test for the comparison of variances with only the independent observations. Numerous tests for the comparison of variances of two independent samples have been documented (Conover, Johnson, and Johnson, 1981). For two independent samples the most common approach is to calculate $\frac{s_1^2}{s_2^2}$ and compare against the F-distribution, i.e. perform the F-test. Equivalently the numerator may be fixed to be the sample with the highest variance.

When the normality assumption is met, the Neyman-Pearson lemma can be used to show that the F-test is the uniformly most powerful test for comparing the variances of two independent samples. However, the F-test is not robust to deviations from normality (Marozzi, 2011; Markowski and Markowski, 1990; Nordstokke and Zumbo, 2007; Conover, Johnson, and Johnson, 1981).

Levene (1961) propose that for two independent groups, the differences between the absolute deviations from the group means could be used to assess equality of variances. In the two sample case, this test is equivalent to Student's t-test applied to absolute deviations from the group means. This version of Levene's test fails to control the Type I error rate when the population distribution is skewed (Nordstokke and Zumbo, 2007; Carroll and Schneider, 1985).

The classical Levene's test is performed using absolute deviations from each group mean. When considering variability, statisticians often consider square deviations from the mean. It is hypothesised that using squared deviations may offer a robust alternative. Some preliminary simulations are performed to explore this hypothesis. In a factorial simulation design, two samples of sizes $\{5, 10, 20, 30, 40\}$ from $N(0,1)$ are generated for 1,000 iterations. For each iteration Levene's test using absolute deviations (L) and Levene's test using squared deviations (LS) are performed at the 5% significance level. For reference, the F-test (F) is included in the comparison. This process is repeated for the Exponential distribution. The overall Type I error robustness for each test averaged over the simulation design is given in Table 10.1.

Table 10.1: Comparison of absolute deviations and squared deviations.

Distribution	F	L	LS
Normal	**0.050**	**0.056**	**0.039**
Exponential	0.220	**0.062**	0.151

Table 10.1 indicates that the use of absolute deviations, rather than squared deviations, better maintains Type I error robustness. This suggests that for skewed distributions the Type I error rate inflation is exacerbated when using squared deviations. Conclusions from this crude preliminary investigation match those by Smith and Cody (1997). Thus the remainder of this chapter proceeds on the basis of absolute deviations, as per standard definitions of Levene's test.

Brown and Forsythe (1974) proposed alternatives to Levene's test when data are not normally distributed. These alternatives use absolute deviations from the median or the trimmed mean. These variations are also often referred to as 'Levene's test' (Gastwirth, Gel, and Miao, 2009; Carroll and Schneider, 1985). For the avoidance of doubt, the convention followed herein is that the process of assessing equality of variances using absolute deviations from the group means is referred to as 'Levene's test'. Assessing equality of variances using absolute deviations from the group medians is referred to as the ‘Brown-Forsythe test’.

Shoemaker (2003) offer two potential fixes to the F-test. Despite simulations showing some good results for the adjustments, Shoemaker (2003) concludes that the Brown-Forsythe test is superior for highly skewed distributions.

Many modifications of Levene’s test have been proposed, but none performing well are found by Parra-Frutos (2009), the proposals by Brown and Forsythe (1974) were not considered.

Conover, Johnson, and Johnson (1981) explored 56 tests for equal variances for two independent groups and noted five tests that are Type I error robust over a large range of conditions, and each use deviations from the median rather than deviations from the mean. Conover, Johnson, and Johnson (1981) found that the only test that consistently meets Bradley’s liberal Type I error robustness criteria is the Brown-Forsythe test. However, it should be noted that this test can be conservative with small sample sizes (Loh, 1987; Lim and Loh, 1996).

Nordstokke and Zumbo (2010) propose a non-parametric alternative using Levenes’s test based on the observation ranks. It is Type I error robust and more powerful than the Brown-Forsythe test when population means are equal. However, Shear, Nordstokke, and Zumbo (2018) show that this alternative test is not robust when the assumption of equal population means is relaxed. Given that the population mean may be an unknown nuisance parameter, Shear, Nordstokke, and Zumbo (2018) direct researchers to the use of the Brown-Forsythe test.

The general consensus regarding two independent samples, is using deviations from the median, particularly the Brown-Forsythe test (Nordstokke and Zumbo, 2007; Mirtagiouglu et al., 2017; Carroll and Schneider, 1985).

The comparison of variances in the partially overlapping samples case has been given very little attention by recent authors that consider the comparison of means in the partially overlapping samples case. Bhoj (1979) and Ekbohm (1981) independently considered a weighted combination of independent and paired observations to create a new test statistic. Other solutions such as ignoring the pairing and performing the F-test on all of the available data were considered by Ekbohm (1982). Bhoj (1984) concluded that his test statistic is a powerful approach if the correlation is negative or small. Otherwise, performing the F-test on all of the available data is more powerful than the solutions put forward by either of the authors (Ekbohm, 1981; Ekbohm, 1982). The simulations performed by these authors were on a relatively small scale. No solution was comprehensively agreed upon for all scenarios. Furthermore, the non-robustness of the Pitman-Morgan test has a detrimental impact on the weighted tests. A different solution that uses all available data without a complex weighting structure, or the discarding of valuable information about the pairing, may therefore be advantageous.

It is proposed that as an alternative test for equal variances in the two sample case, the partially overlapping samples t-test can be performed, using absolute deviations from the sample group medians, as outlined below.

Let X_{ji} denote the i-th observation in group j, for $j = \{\text{Sample 1, Sample 2}\}$, and $\tilde{X}_j$ denote the sample median, so that $Y_{ji} = X_{ji} - \tilde{X}_j$, then

$$T_{\text{var1}} = \frac{\bar{Y}_1 - \bar{Y}_2}{S_{py}\sqrt{\frac{1}{n_1} + \frac{1}{n_2} - 2\rho\frac{n_c}{n_1 n_2}}} \text{ where } S_{py} = \sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{(n_1-1)+(n_2-1)}} \tag{10.1}$$

The test statistic T_{var1} is referenced against the t-distribution with v_1 degrees of freedom.

For the comparison of variances, Loh (1987) suggest adopting the form of the t-test unconstrained to equal variances, using absolute deviations from the sample group medians. The partially overlapping samples test statistic unconstrained to equal variances can be similarly modified to provide a test for equality of variances so that:

$$T_{\text{var2}} = \frac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2} - 2\rho\frac{S_1 S_2 n_c}{n_1 n_2}}} \tag{10.2}$$

The test statistic T_{var2} is referenced against the t-distribution with v_2 degrees of freedom.

10.2 Example

For illustrative purposes, returning to the example by Rempala and Looney (2006), with a research question of whether there is a difference in the variability of patients scores between the last day of life and the second to last day.

The deviations from the group medians are calculated, then the tests are performed using the R package by Derrick (2017a).

The results are: $\tilde{X}_1 = 20$, $\tilde{X}_2 = 20$, $T_{\text{var1}} = -0.886$ (p-value $= 0.380$), $T_{\text{var2}} = -0.882$ (p-value $= 0.380$).

Thus there is no evidence to suggest that the variability is not equal between the two days.

10.3 Methodology

Approaches for the comparison of variances in the two sample case are assessed using simulation. The tests considered are the Brown-Forsythe test, the Pitman-Morgan test, and the proposed T_{var1} and T_{var2}.

Within the simulation design, the sample sizes $\{n_a, n_b, n_c\}$ are $\{5, 10, 30, 50\}$. The correlation coefficients, ρ, are $\{0.00, 0.25, 0.50, 1.00\}$. Simulations for each possible parameter combination of n_a, n_b, n_c and ρ are performed in a factorial design.

Firstly the comparison of variances is performed for normally distributed data. Under the null hypothesis, $X_1 \sim N(0, 1)$ and $X_2 \sim N(0, 1)$. Under the alternative hypothesis, the observations in Sample 2 are multiplied by $\sigma = 2$, thus $X_1 \sim N(0, 1)$ and $X_2 \sim N(0, 4)$.

Secondly the comparison of variances is performed for a skewed distribution. Under the null hypothesis, Normal deviates are first generated as above, and then the exponent of each value is calculated to create a Lognormal distribution as per Chapter 4.1.2. Under the alternative hypothesis this process is repeated, each of the observations in Sample 2 multiplied by $\sigma = 2$ to create unequal variances.

For each parameter combination, the data generating process is repeated

10,000 times, and each of the statistical tests to be evaluated is performed on each replicate. Under the null hypothesis, the proportion of the 10,000 replicates where H_0 is rejected represents the Type I error rate. Under the alternative hypothesis, the proportion of the replicates where the H_0 is rejected represents the power of the test. The simulations and test procedures are performed in R, at the 5% significance level, two-sided. The simulation design allows that the conditions of MCAR are met.

10.4 Results

Type I error rates and power are investigated for each of; the Brown-Forsythe test, BF, the Pitman-Morgan test, PM, and the partially overlapping samples tests, T_{var1} and T_{var2}.

Firstly, for a comparison of variances for two samples from the Normal distribution, each of the test statistics are assessed under the null hypothesis where $X_1 \sim N(0, 1)$ and $X_2 \sim N(0, 1)$. The Type I error robustness for each of the parameter combinations within the simulation design are summarised in Figure 10.1.

Figure 10.1 shows that the Pitman-Morgan test and the proposed test statistics are Type I error robust throughout the simulation design, with T_{var1} being on average more conservative relative to T_{var2}. The Brown-Forsythe test has a Type I error rate below the nominal significance level for every parameter combination within the simulation design. For the smallest sample sizes within the design, the Brown-Forsythe test is very conservative. This result makes the Brown-Forsythe test appealing for researchers that like to err on the side of caution during their analyses.

Next, each of the test statistics are assessed when both samples are taken from skewed but identical distributions. The Type I error robustness for each of the parameter combinations within the simulation design are summarised in Figure 10.2.

Figure 10.2 shows that the Pitman-Morgan test is not Type I error robust when the samples are taken from the Lognormal distribution. This supports the findings by McCulloch (1987) and Wilcox (2015). In addition it can be

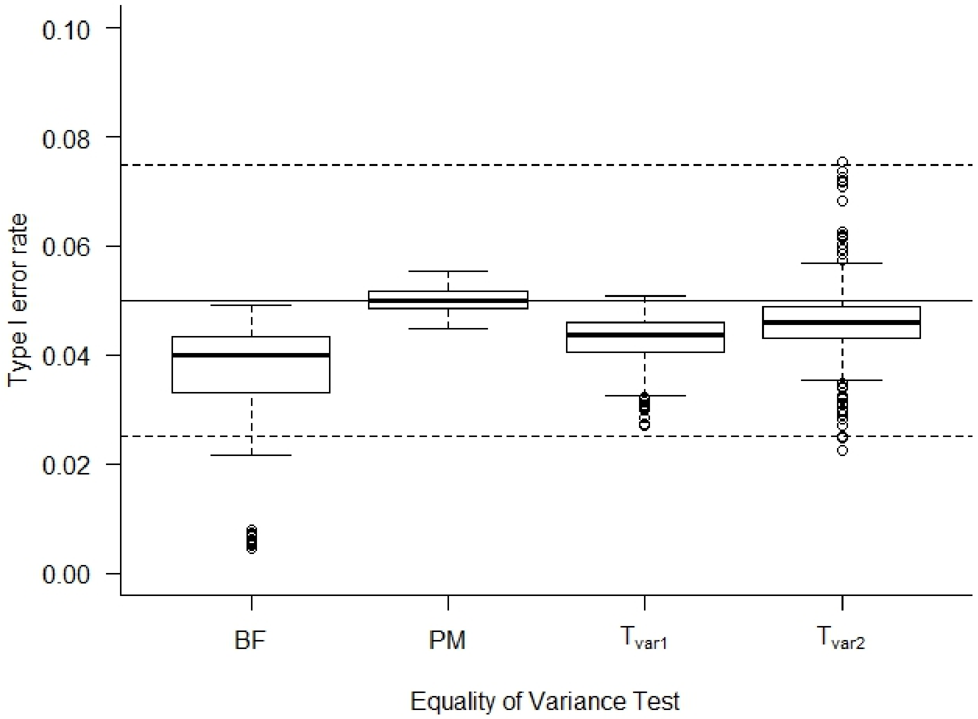


Figure 10.1: Type I error robustness for each parameter combination, samples from Standard Normal distribution.

seen that T_{var2} does not fully maintain Type I error robustness across all scenarios within the simulation design.

To identify where the deviations from Type I error robustness materialise, a selection of parameter combinations from the Lognormal distribution and their Type I error rates are given in Table 10.2.

Studying the Type I error rates for different parameter combinations in Table 10.2 shows that both BF and T_{var1} are similarly conservative, for small and large sample sizes. T_{var2} is liberal when one of the samples is more dominant in terms of size, and when there is a large imbalance between the number of independent observations and the number of pairs. These conclusions are replicated regardless of the extent of the correlation between

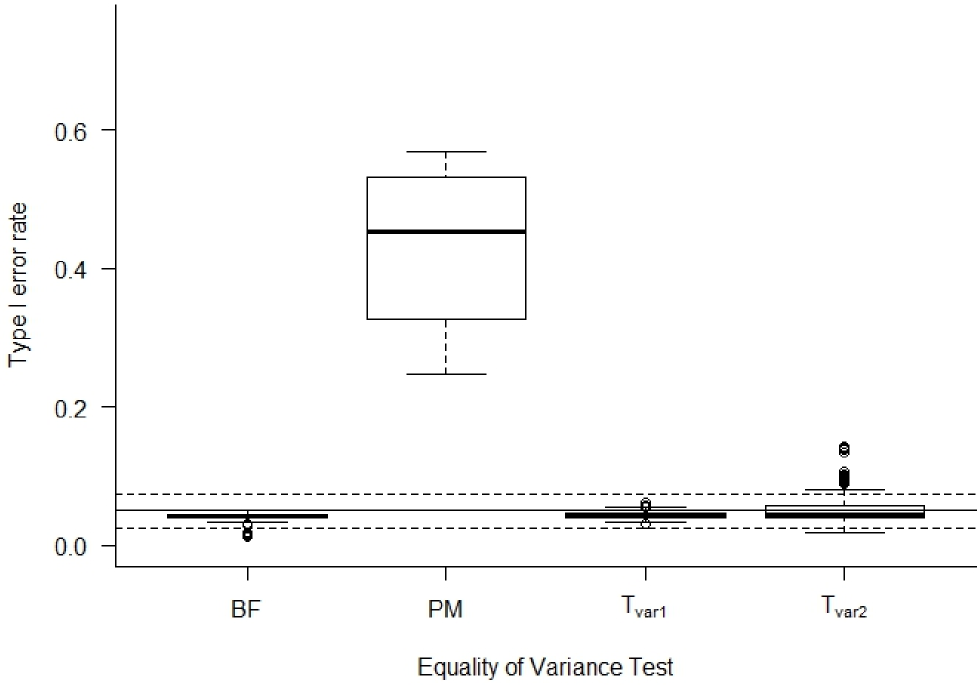


Figure 10.2: Type I error robustness for each parameter combination, samples from Lognormal distribution.

the two populations.

These findings coincide with findings in Chapter 6 with respect to the non-robustness of a test statistic modified to take into account unequal variances, when distributions are skewed.

Relative power comparisons for each of the test statistics are assessed where $X_1 \sim N(0,1)$ and $X_2 \sim N(0,4)$, followed by power for samples from distributions with differing skew. The power averaged across the simulation design for increasing ρ is given in Table 10.3.

Table 10.3 shows that the proposed solution, T_{var1}, is more powerful than the Brown-Forsythe test. It also indicates that both the Brown-Forsythe test and the newly proposed test, T_{var1}, are less powerful when samples are taken from a heavy-tailed distribution.

Table 10.2: Type I error robustness, tests for equal variances, selected parameter combinations, Lognormal distribution.

ρ	n_a	n_b	n_c	BF	PM	T_{var1}	T_{var2}
0.5	30	50	30	**0.046**	0.500	**0.047**	**0.043**
0.5	30	50	5	**0.039**	0.261	**0.036**	**0.043**
0.5	30	50	50	**0.041**	0.547	**0.045**	**0.040**
0.5	5	50	30	**0.045**	0.507	**0.045**	**0.059**
0.5	5	50	5	**0.045**	0.259	**0.042**	0.133
0.5	5	50	50	**0.039**	0.551	**0.052**	**0.053**
0.5	50	50	30	**0.040**	0.507	**0.045**	**0.041**
0.5	50	50	5	**0.040**	0.262	**0.037**	**0.036**
0.5	50	50	50	**0.041**	0.541	**0.045**	**0.040**
0	30	50	30	**0.039**	0.524	**0.043**	**0.044**
0	30	50	5	**0.041**	0.274	**0.039**	**0.046**
0	30	50	50	**0.041**	0.553	**0.043**	**0.042**
0	5	50	30	**0.043**	0.515	**0.043**	**0.063**
0	5	50	5	**0.043**	0.268	**0.044**	0.142
0	5	50	50	**0.042**	0.552	**0.044**	**0.053**
0	50	50	30	**0.041**	0.510	**0.043**	**0.043**
0	50	50	5	**0.043**	0.267	**0.040**	**0.038**
0	50	50	50	**0.040**	0.560	**0.043**	**0.043**

Table 10.3: Power, tests for equal variances.

Distribution	ρ	BF	PM	T_{var1}	T_{var2}
Normal	0	0.495	0.655	0.887	0.867
Normal	0.25	0.495	0.667	0.891	0.871
Normal	0.5	0.495	0.700	0.900	0.880
Normal	0.75	0.496	0.778	0.916	0.898
Skewed	0	0.198	0.621	0.623	0.442
Skewed	0.25	0.198	0.627	0.633	0.466
Skewed	0.5	0.198	0.651	0.657	0.507
Skewed	0.75	0.198	0.715	0.704	0.577

Power is given for PM for intrigue and completion purposes only. For some individual parameter combinations the power of T_{var2} exceeds the power of T_{var1} due to the liberal nature of T_{var2} for those parameter combinations. However it can be seen from Table 10.3 that averaged across the simulation design, T_{var2} is less powerful.

The superior validity and power of T_{var1} gives this test great credentials as a useful test within the partially overlapping samples framework.

10.5 Summary

A common research question in psychology, education, medical sciences, business and manufacturing, is whether or not the variances between two groups are equal (Gastwirth, Gel, and Miao, 2009).

There has been little previous research into techniques for the comparison of variances for samples that contain both independent observations and paired observations. Standard solutions that involve discarding data are less than desirable.

Two solutions that make use of the partially overlapping samples t-tests introduced in Chapter 3 are proposed in this chapter. Simulations across a range of sample sizes show that these solutions are Type I error robust under normality and the assumption of MCAR. These solutions are more powerful than established solutions that discard data, namely the Pitman-Morgan test and the Brown-Forsythe test.

The equal variances form of the partially overlapping samples variances test, T_{var1}, is marginally more powerful than the unconstrained form of the test T_{var2}.

The proposed test statistic T_{var1} further maintains Type I error robustness for skewed distributions where T_{var2} does not. T_{var1} is therefore recommended as a powerful test for the equality of variances between two samples when there is a combination of paired observations and independent observations in two samples.

Chapter 11

Conclusions and Further Work

The partially overlapping samples framework has been researched. Common issues identified with existing methods have driven this investigation. The focus of the thesis is on the robustness and application of newly proposed solutions. In this final chapter, recommendations are summarised, and future avenues of exploration are put forward. Finally, reflections on the journey are given.

11.1 Conclusion

There are many situations where the presence of both paired observations and independent observations cannot be avoided. Partially overlapping samples may occur due to missing data in a paired samples design, and other occasions where paired samples tests alone might not adequately reflect the structure of the data being collected.

For extreme sample size imbalances, where the number of independent observations in one sample greatly differs from the number of independent observations in the other sample or the number of paired observations, tests which use all of the available data are not always ideal. For large sample sizes, naive tests which discard observations are likely to be powerful enough for most practical applications. In this situation it is important to ensure that the discarded observations would not give different information to the included observations.

The extant tests considered and the newly proposed tests assume within sample independence. Simulations in this thesis have been performed under the assumption of MCAR. Partially overlapping samples that occur by design are regarded as being MCAR. Partially overlapping samples can also occur through missing data in a paired samples design. The assumption of MCAR is often reasonable in scenarios including equipment failure or loss in transit (Kang, 2013). A common occurrence of partially overlapping samples is a paired samples design where participants drop out of the study. In this case the researcher needs to make a judgment whether analyses on the response variable will be impacted by the missing data. The proposed partially overlapping samples tests are not recommended for data that is MNAR.

Controversial practices for comparing two samples of paired observations and independent observations are frequently performed, from discarding observations, to imputing observations (Choi and Stablein, 1982). Other poor practices observed range from treating all observations as independent and ignoring the pairing, to randomly pairing unpaired observations (Bedeian and Feild, 2002).

Particularly for smaller sample sizes, partially overlapping samples tests using all of the available data can be advantageous. Tests which act as a simple parameter difference ease interpretation, and confidence intervals can readily be formed. Solutions using all of the available data facilitated by the 'Partiallyoverlapping' R package (Derrick, 2017a) offer intuitive, valid and powerful solutions for the analyses of partially overlapping samples.

For the comparison of central location, if the assumptions of normality and equal variances can be made, the Type I error robustness of T_{new1} suggests that T_{new1} can be used as default. Little power is lost relative to T_{new2}. If the normality assumption is reasonable but the equal variances assumption is not reasonable, T_{new2} is the recommended test of choice. However, where the assumption of normality is not reasonable, including in the presence of outliers, the non-parametric test T_{RNK1} is the recommended test. The poor outcomes for these parametric tests with respect to outliers reflect the t-tests upon which these tests are based. The findings are in contrast to suggestions by Ruxton (2006) and Delacre, Lakens, and Leys (2017) who suggests the routine use of Welch's test. While the use of Welch's test under the condition of normality is reasonable, tests not constrained to equal variances perform badly under non-normality. Tests not constrained to equal variances should only be performed if there is no reason to doubt the normality assumption.

Non-parametric tests are often over used due to obsession with testing the normality assumption (Rasch and Guiard, 2004). For cases where extreme violation of the normality assumption is anticipated, the non-parametric T_{RNK1} offers a robust alternative to naive non-parametric tests. The form of the null hypothesis should be given consideration, the non-parametric tests can be viewed as a test of central location when the distributions are equal (Rietveld and Van Hout, 2017).

The recommended approach is to consider the reasonableness of normality first and foremost. Choose between T_{new1} and T_{new2} under normality, otherwise perform T_{RNK1}. This philosophy is in agreement with those that perform more formal preliminary testing because the default positions are normality and equal variances.

Similarly developed solutions which are robust, with easy to interpret results, are recommended for the comparison of proportions, z_8, and the comparison of variances, T_{var1}. Performing z_8 can be recommended for all applications where there is a dichotomous response for two samples. The use of T_{var1} is not directly recommended as a preliminary test, but is recommended for applications where equality of variances are of primary interest.

The new solutions provided are applicable in numerous disciplines, these include applications in; medicine (Alhouayek et al., 2019; Polster et al., 2019), environmental geography (Raymundo et al., 2019), psychology (NeMoyer et al., 2018; Cummins, Hussey, and Hughes, 2019), education (Guerrero Segura, 2019), business and finance (Kimotho, 2018), and technology and human relationships (Bourrelly et al., 2018; Carolus et al., 2019).

11.2 Further work

Alongside continual exploration into this area as the literaure develops, some specific areas of further work are given below.

11.2.1 Further R releases

Future releases of the 'Partiallyoverlapping' R package could include specific functions for T_{var1} and T_{RNK1} so that the preliminary ranking is performed for the user.

In communication with a Data Scientist from Highmark Health, Pennsylvania, I have been alerted to 'a minor issue with the Prop.test. I was using the function on a relatively large dataset (about a million observations), and got the warning message *In* $u \times v$ *: NAs produced by integer overflow.* This resulted in an NA value'. Changing the source code of ($u \times v$) to (as.numeric(u) $\times$ as.numeric(v)) resolves the issue for the user. This correction may not be worthy of a new release in its own right, but will be added to any future release.

11.2.2 Extension to Parallel Randomised Controlled Trials

Randomised Controlled Trials (RCTs) are widely regarded as the 'gold standard' method for estimating treatment effectiveness (Hewitt et al., 2010). Parallel RCTs are the most frequently used (Walsh et al., 2014). Any differences between two groups at baseline should be reported, however missing data at baseline complicates the issue of attrition (Hewitt et al., 2010). Attrition reduces the sample size, which reduces statistical power and limits the extent to which results can be generalised (Leon et al., 2006). The discarding of information to perform complete case analysis requires the implicit assumption of MCAR. An analysis that uses all of the available information without imputation is a mixed effect model (Leon et al., 2006). The volume of attrition that becomes problematic is unclear (Schulz and Grimes, 2002). A large volume of data MCAR results in less bias than a small amount of non-random attrition (Hewitt et al., 2010). In a RCT the sample size calculation frequently results in an impractical sample size requirement (Pocock, 1985). This can greatly impact the power of the test. It is usual for an RCT to have missing data on the dependent variable in excess of 10%, and a larger degree of missing data is not uncommon (Holis and Campbell, 1999).

As an extension to the partially overlapping samples framework, missing data may be present for both groups at both time points. The comparison of the differences between two groups over time can be performed using the t-test, T_D, being a comparison of the differences between Group A and Group B at follow-up less the differences between Group A and Group B at baseline. Partially overlapping samples naturally inhibit the power of this test, because four observations all need to be present to form any one comparison within T_D.

An alternative approach that uses all the available data is introduced. Z_D is constructed so that the differences in differences are divided by the standard error of the differences in differences. An approach that directly assesses a difference in differences, equivalent to the assessment of an interaction effect, is considered here.

The two group scenario can be given as per Equation 11.1 where n_{ij} = number of subjects recorded at time i only in group j only, $n_{\lambda j}$ = number of subjects recorded at both Time 1 and Time 2 in group j, S_{ij} = standard deviation of subjects recorded at time i in group j. Although time is used in this notation, the scenario described is not restricted to time and could be used for the comparison between any two factors.

$$Z_D = \frac{(\bar{X}_{1A} - \bar{X}_{2A}) - (\bar{X}_{1B} - \bar{X}_{2B})}{\sqrt{\frac{S_{1A}^2}{n_{1A}} + \frac{S_{2A}^2}{n_{2A}} - 2\rho\frac{S_{1A}S_{2A}n_{\lambda A}}{n_{1A}n_{2A}} + \frac{S_{1B}^2}{n_{1B}} + \frac{S_{2B}^2}{n_{2B}} - 2\rho\frac{S_{1B}S_{2B}n_{\lambda B}}{n_{1B}n_{2B}}}} \tag{11.1}$$

Some preliminary simulations are outlined here. Normally distributed deviates are sampled. Paired observations and independent observations are generated separately for each group at each time period. For the paired observations the correlation between Group A and Group B, is derived as per

Kenney and Keeping (1951). The implicit correlation between Time 1 and Time 2 is zero, and the differences are normally distributed. The simulation is based on Group A having the same variance for both time periods, and Group B having the same variance for both time periods, but Group A and B are not restricted to both having the same variance. Simulations are performed for a factorial design under H_0 with $n_{ij} = \{5,\ 10,\ 20,\ 30\}$, $\rho = \{0, 0.25, 0.5, 0.75\}$ and $\sigma_i^2 = \{1, 4\}$. For each parameter combination, 1,000 iterations are performed. The test statistic Z_D is compared against T_D. Both tests are performed at the 5% significance level for each iteration.

Initial simulations take place under H_0 where there are no differences in differences. Figure 11.1 shows the overall Type I error rates across the simulation design for each of Z_D and T_D. In some additional preliminary simulations, a test statistic devised with the restriction to equal variances was also attempted, however the results did not exhibit reasonable Type I error robustness.

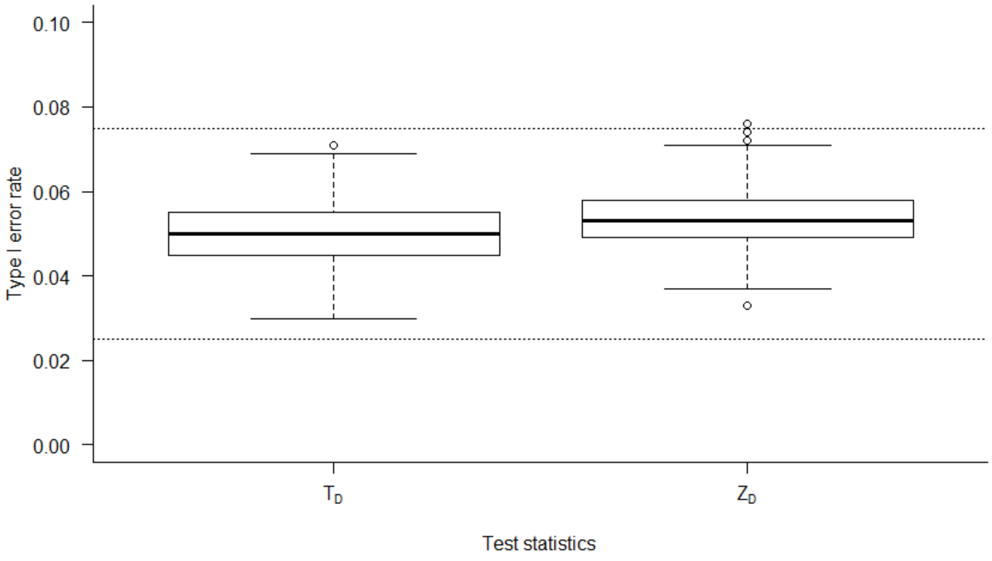


Figure 11.1: Type I error rates, T_D and Z_D.

Figure 11.1 demonstrates that both Z_D and T_D maintain Type I error robustness across the simulation design. The proposed approach could therefore be valuable for clinical trials performed under intent to treat, because all observations are included in the analysis.

This could be further extended to a cross-over trial. The numerator of Equation 11.1 can be modified to test for a period effect, i.e. $(\bar{X}_{1A} - \bar{X}_{1B}) - (\bar{X}_{2A} - \bar{X}_{2B})$. Likewise the numerator of Equation 11.1 can be modified to test for a carry-over effect, i.e. $(\bar{X}_{1A} - \bar{X}_{2B}) - (\bar{X}_{2A} - \bar{X}_{1B})$.

11.2.3 Multivariate scenario

There has been much less attention given to the multivariate situation, where partially overlapping samples are in more than two groups (Mantilla and Terpstra, 2018). Under these conditions, the issues with regards to discarding observations or imputing observations are exacerbated. This adds an area for further research. One potential avenue for exploration would be to construct a test making use of the generalized t-test, Hotelling's t-squared statistic (Lu and Yuan, 2010).

The Freidman test is a test for equal distributions in a repeated measures design, with three or more samples. However it can only be performed when the study design is completely balanced. The Skillings-Mack test is a known alternative to the Freidman test when some observations are missing, i.e. when the design is not balanced. Further alternative approaches to both the Freidman test and the Skillings-Mack test is proposed here. The solution is to calculate the mean of each block and deduct from each original score. This means that the repeated measures information is used, and thus the Kruskal-Wallis test, or the one-way ANOVA, can then be performed on the modified data. The mathematical form of this solution and the robustness of the solution is part of final year undergraduate student projects I am currently supervising.

11.2.4 Recent advances

Concurrently with this research, some authors are starting to look into non-parametric solutions, and solutions using resampling methods.

Samawi, Yu, and Vogel (2015) put forward a non-parametric solution attempting to combine the Wilcoxon test and the Mann-Whitney test, they suggest that this is among the most Type I error robust powerful procedures.

Resampling methods are increasingly advocated analyses methods that do not make any prior assumptions about the distribution of the data (Odén and Wedel, 1975). There are several resampling strategies, the common theme is that repeatedly sampling from the observed data gives a distribution based on the sample. Resampling methods are particularly of use for small sample sizes when tests would traditionally yield low power. Overviews of basic resampling methods are given by Yu et al. (2012), and Good (2013).

For paired differences, bootstrapping is not Type I error robust for small samples, whereas permutation tests do maintain Type I error robustness (Konietschke and Pauly, 2014). Permutation based approaches control the Type I error rate across many distributions and ρ (Konietschke and Pauly, 2014). Permutation based methods are computationally intensive to the point of being prohibitive if sample size is large. Randomization tests involve taking a random sample of permutations from the complete set of possible permutations, and as such have good intuitive and mathematical properties (Odén and Wedel, 1975; Edgington and Onghena, 2007).

Yu et al. (2012) show that permutation based methods based on Bhoj (1978) and Kim et al. (2005) perform similarly to their counterparts. When applied to non-normality, Type I error rates are reasonably maintained, but associated power values are approximately halved.

Amro and Pauly (2017) propose a permutation solution based on the solution by Bhoj (1978). This test statistic involves a complex weighting structure of the paired samples t-test and the independent samples t-test. Unlike the solution the permutation test is based on, the solution by Amro and Pauly (2017) is Type I error robust across a range of distributions. Rempala and Looney (2006) found that a linear combination of randomization tests can be robust. However, it is not robust for non-positive correlation.

Amro, Konietschke, and Pauly (2018) propose further non-parametric permutation approaches for performing weighted combination tests so that the combined Type I error rate is 5%. This incorporates weighted tests for the paired samples t-test and Welch's test, as well as for the Wilcoxon test and the Mann-Whitney test. For the three distributions they consider, Normal, Exponential and Lognormal, they show that these methods do not maintain Type I error robustness, unless permutation tests are adopted. However, Amro, Konietschke, and Pauly (2018) only show the average Type I error rate for each value of ρ, some sample size conditions where the test may be particularly liberal or conservative may be obscured.

11.3 Reflection

The Monte-Carlo simulation process has worked well for developing new solutions to an enduring problem. Collaboration on these solutions has led to outputs which have been of interest both to academics and those in more practical fields, as can be seen with the interactions in Appendix 1. The well-known problem of analysing partially overlapping samples, in addition to intuitive solutions, attracts interest from those in attendance at conference presentations and in other general discussions with those familiar with statistics. Conferences attended are listed in Appendix 2.

A seed was not set in early simulations, in hindsight making it difficult for precise replication of the results. Upon reflection, using a reference filing system from the outset would have been beneficial for recalling correct attribution during the write-up.

In many of the papers where the proposed methods have been applied, the form of the test performed is not stated e.g. Raymundo et al. (2019) and Carolus et al. (2019). In the published outputs, more explicitly stated decision rules regarding choice of test would encourage researcher transparency.

Statistics is often perceived as an 'art', not a 'science'. While the aims have been met with recommendations put forward, a degree of subjectivity is inevitable to remain. In order for the solutions to maintain broad appeal, the recommendations are likely to remain open to interpretation.

Many outputs have been published from this research, as detailed in Appendix 2. Publications include further exploration of existing methods (Appendices P2, P5, P8, P9), and technical detail of the newly proposed solutions

(Appendices P1, P3, P6, P7, P10). Outputs targeting user accessibility are well received, particularly the R package and the tutorial paper (Appendix P4).

11.4 Summary

This thesis makes a contribution to methodology, the principles of robustness and the process of using simulation to explore the robustness of test statistics. The criteria applied given by Bradley (1978) for identification of Type I error robustness is open to some subjectivity when comparing across multiple parameter combinations within a simulation design. An alternative way of measuring Type I error robustness is also introduced, but requires simulation designs to be comparable for the technique to be relevant across multiple studies.

The content contributes to ongoing debate regarding when to use non-parametric tests, and when to use preliminary testing. The recommendation is that if the comparison of central location is of primary interest, parametric tests should be used unless there is strong pre-existing evidence that distributions are not-normal. In the unlikely event of absence of any pre-existing evidence regarding the distribution, preliminary tests defined in advance can be carried out. Likewise the equality of variances assumption should be determined based on prior knowledge where available. If study units are randomly allocated to groups, the equal variances assumption should be applied.

Within this research users are provided with new tools for overcoming the common problems of partially overlapping samples. These solutions can be applied in any discipline and are recommended to be used when comparing two groups if there is a combination of paired observations and independent observations. An exception is if the sample size of either the independent observations or paired observations is large (≥ 100) and the sample size of either the independent observations or paired observations is small (≤ 5), here traditional methods that discard a small amount of data will typically suffice.

The R package by Derrick (2017a) has been validated by several statisticians, downloaded approximately 10,000 times, and can be used with confidence by researchers. When reporting results, as a minimum users should specifically state which form of the test they have used, the value of the test statistic, the degrees of freedom, and the p-value. More complete reporting would also include a confidence interval.

The solutions published are subject to the same limitations underlying all frequentist statistical tests, including good quality research design, any missing data does not arise due to any systematic pattern, and that there is independence of observations within groups.

The further work section introduces a solution that is of potentially high impact for trials performed under intention to treat principles. Bayesian techniques could also be considered for alternative solutions.

References

Aguinis, H., Gottfredson, R. K., and Joo, H. (2013). “Best-practice recommendations for defining, identifying, and handling outliers”. *Organizational Research Methods* 16 (2), pp. 270–301.

Alcala-Quintana, R. and Garcia-Perez, M. A. (2004). “The role of parametric assumptions in adaptive Bayesian estimation.” *Psychological Methods* 9 (2), pp. 250–271.

Alhouayek, M., Sorti, R., Gilthorpe, J. D., and Fowler, C. J. (2019). “Role of pannexin-1 in the cellular uptake, release and hydrolysis of anandamide by T84 colon cancer cells”. *Scientific reports* 9 (1), p. 7622.

Allen, I. E. and Seaman, C. A. (2007). “Likert scales and data analyses”. *Quality Progress* 40 (7), pp. 64–65.

Amro, L., Konietschke, F., and Pauly, M. (2018). “Multiplication combination tests for incomplete paired data”. *arXiv:1801.08821*.

Amro, L. and Pauly, M. (2017). “Permuting incomplete paired data: a novel exact and asymptotic correct randomization test”. *Journal of Statistical Computation and Simulation* 87 (6), pp. 1148–1159.

Anderson, G. (2014). *Flow chart for selecting commonly used tests*. URL: `http://abacus.bates.edu/~ganderso/biology/resources/stats_flow_chart_v2014.pdf` (visited on June 6, 2018).

Anscombe, F. J. (1960). “Rejection of outliers”. *Technometrics* 2 (2), pp. 123–146.

APA (2018). *American Psychological Association Author Instructions*. URL: `http://www.apa.org/pubs/authors/instructions.aspx` (visited on Jan. 1, 2018).

Armstrong, R. L. (1987). “The midpoint on a five-point Likert-type scale”. *Perceptual and Motor Skills* 64 (2), pp. 359–362.

Baker, M. (2016). “Is there a reproducibility crisis? A Nature survey lifts the lid on how researchers view the’crisis rocking science and what they think will help”. *Nature* 533 (7604), pp. 452–455.

Bakker, M. and Wicherts, J. M. (2014). “Outlier removal and the relation with reporting errors and quality of psychological research”. *PLoS One* 9 (7), e103360.

Baksalary, J. K. and Puntanen, S. (1990). “A complete solution to the problem of robustness of Grubbs’s test”. *Canadian Journal of Statistics* 18 (3), pp. 285–287.

Banks, P., Waugh, A., Henderson, J., Sharp, B., Brown, M., Oliver, J., and Marland, G. (2014). “Enriching the care of patients with dementia in acute settings”. *Dementia* 13 (6), pp. 717–736.

Barnett, V and Lewis, T (1994). *Outliers in statistical data.* Wiley, 3rd edition.

Barr, A., Goodnight, J., Sall, J., Blair, W., and Chilko, D. (1979). “General linear models procedure of the statistical analysis system”. *SAS User’s Guide. Cary, NC: Statistical Analysis System Institute.*

Bartels, R. (1982). “The rank version of von Neumann’s ratio test for randomness”. *Journal of the American Statistical Association* 77 (377), pp. 40–46.

Beasley, T. M., Erickson, S., and Allison, D. B. (2009). “Rank-based inverse normal transformations are increasingly used, but are they merited?” *Behavior Genetics* 39 (5), pp. 580–595.

Bedeian, A. G. and Feild, H. S. (2002). “Assessing group change under conditions of anonymity and overlapping samples”. *Nursing Research* 51 (1), pp. 63–65.

Behrens, W. U. (1928). *Ein beitrag zur fehlerberechnung bei wenigen beobachtungen.* Institut für Pflanzenbau.

Bell, C. and Doksum, K. (1965). “Some new distribution-free statistics”. *The Annals of Mathematical Statistics*, pp. 203–214.

Berkson, J. (1978). “In dispraise of the exact test: Do the marginal totals of the 2x2 table contain relevant information respecting the table proportions?” *Journal of Statistical Planning and Inference* 2 (1), pp. 27–42.

Bhoj, D. S. (1978). “Testing equality of means of correlated variates with missing observations on both responses”. *Biometrika* 65 (1), pp. 225–228.

Bhoj, D. S. (1979). “Testing equality of variances of correlated variates with incomplete data on both responses”. *Biometrika* 66 (3), pp. 681–683.

Bhoj, D. S. (1984). “On testing equality of variances of correlated variates with incomplete data”. *Biometrika* 71 (3), pp. 639–641.

Binder, A. (1984). “Restrictions on statistics imposed by method of measurement: Some reality, much mythology”. *Journal of Criminal Justice* 12 (5), pp. 467–481.

Bishara, A. J. and Hittner, J. B. (2012). "Testing the significance of a correlation with nonnormal data: comparison of Pearson, Spearman, transformation, and resampling approaches." *Psychological Methods* 17 (3), pp. 399–417.

Bishop, P. A. and Herron, R. L. (2015). "Use and misuse of the Likert item responses and other ordinal measures". *International Journal of Exercise Science* 8 (3), pp. 297–302.

Bland, J. M. and Altman, D. G. (1986). "Statistical methods for assessing agreement between two methods of clinical measurement". *The Lancet* 327 (8476), pp. 307–310.

Bland, J. M. and Butland, B. K. (2011). *Comparing proportions in overlapping samples.* URL: `http://www-users.york.ac.uk/~mb55/overlap.pdf` (visited on Apr. 2, 2018).

Bland, M. (2013). "Do baseline p-values follow a uniform distribution in randomised trials?" *PLoS One* 8 (10), e76010.

Blom, G. (1958). *Statistical estimates and transformed beta-variables.* Almqvist & Wiksell,Sweden.

Boneau, C. A. (1960). "The effects of violations of assumptions underlying the t test". *Psychological Bulletin* 57 (1), pp. 49–64.

Boone, H. N. and Boone, D. A. (2012). "Analyzing Likert data". *Journal of Extension* 50 (2), pp. 1–5.

Bourrelly, A, Naurois, C. J. de, Zran, A, Rampillon, F, Vercher, J., and Bourdin, C (2018). "Impact of a long autonomous driving phase on takeover performance". *6th Driver Distraction and Inattention Conference.*

Box, G. E. (1953). "Non-normality and tests on variances". *Biometrika* 40 (3), pp. 318–335.

Box, G. E. and Muller, M. E. (1958). "A note on the generation of random Normal deviates". *The Annals of Mathematical Statistics* 29 (2), pp. 610–611.

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

Bradley, J. V. (1982). "The insidious L-shaped distribution". *Bulletin of the Psychonomic Society* 20 (2), pp. 85–88.

Bradley, J. C., Waliczek, T. M., and Zajicek, J. M. (1999). "Relationship between environmental knowledge and environmental attitude of high school students". *The Journal of Environmental Education* 30 (3), pp. 17–21.

British Standards Institution (1975). *BS ISO 3301:1975. Statistical interpretation of data - Comparison of two means in the case of two samples.*

British Standards Institution (1997). *BS ISO 5479:1997. Statistical interpretation of data - Tests for departure from normality. British Standards Institution.*

Brown, M. B. and Forsythe, A. B. (1974). “Robust tests for the equality of variances”. *Journal of the American Statistical Association* 69 (346), pp. 364–367.

Bunner, J. and Sawilowsky, S. S. (2002). “Alternatives to Sw in the bracketed interval of the trimmed mean”. *Journal of Modern Applied Statistical Methods* 1 (1), p. 24.

Cao, C., Pauly, M., and Konietschke, F. (2018). “The Behrens-Fisher Problem with Covariates and Baseline Adjustments”. *arXiv:1808.08986*.

Carifio, J. and Perla, R. J. (2007). “Ten common misunderstandings, misconceptions, persistent myths and urban legends about Likert scales and Likert response formats and their antidotes”. *Journal of Social Sciences* 3 (3), pp. 106–116.

Carolus, A., Binder, J. F., Muench, R., Schmidt, C., Schneider, F., and Buglass, S. L. (2019). “Smartphones as digital companions: Characterizing the relationship between users and their phones”. *New Media & Society* 21 (4), pp. 914–938.

Carroll, R. J. and Schneider, H. (1985). “A note on Levene’s tests for equality of variances”. *Statistics & probability letters* 3 (4), pp. 191–194.

Chaffin, W. W. and Rhiel, S. G. (1993). “The effect of skewness and kurtosis on the one-sample t-test and the impact of knowledge of the population standard deviation”. *Journal of Statistical Computation and Simulation* 46 (1-2), pp. 79–90.

Chatfield, M. and Mander, A. (2009). “The Skillings–Mack test (Friedman test when there are missing data)”. *The Stata Journal* 9 (2), pp. 299–305.

Chay, S., Fardo, R., and Mazumdar, M (1975). “On using the Box-Muller transformation with multiplicative congruential pseudo-random number generators”. *Applied Statistics* 22 (1), pp. 132–135.

Chen, Z. (2011). “Is the weighted z-test the best method for combining probabilities from independent tests?” *Journal of Evolutionary Biology* 24 (4), pp. 926–930.

Choi, S. and Stablein, D. (1982). “Practical tests for comparing two proportions with incomplete data”. *Applied Statistics* 31 (3), pp. 256–262.

Cicchetti, D. V., Shoinralter, D., and Tyrer, P. J. (1985). “The effect of number of rating scale categories on levels of interrater reliability: A Monte Carlo investigation”. *Applied Psychological Measurement* 9 (1), pp. 31–36.

Clason, D. L. and Dormody, T. J. (1994). “Analyzing data measured by individual Likert-type items”. *Journal of Agricultural Education* 35 (4), pp. 31–35.

Cochran, W. G. (1953). “Matching in analytical studies”. *American Journal of Public Health and the Nations Health* 43 (6_Pt_1), pp. 684–691.

Cohen, J. (1992). “A power primer”. *Psychological Bulletin* 112 (1), p. 155.

Cohen, M. and Arthur, J. (1991). “Randomization analysis of dental data characterized by skew and variance heterogeneity”. *Community Dentistry and Oral Epidemiology* 19 (4), pp. 185–189.

Conover, W. J. and Iman, R. L. (1981). “Rank transformations as a bridge between parametric and nonparametric statistics”. *The American Statistician* 35 (3), pp. 124–129.

Conover, W. J., Johnson, M. E., and Johnson, M. M. (1981). “A comparative study of tests for homogeneity of variances, with applications to the outer continental shelf bidding data”. *Technometrics* 23 (4), pp. 351–361.

Conover, W. J. (1973). “On methods of handling ties in the Wilcoxon signed-rank test”. *Journal of the American Statistical Association* 68 (344), pp. 985–988.

Coombs, W. T., Algina, J., and 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), pp. 137–179.

Cousineau, D. and Chartier, S. (2010). “Outliers detection and treatment: a review.” *International Journal of Psychological Research* 3 (1), pp. 58–67.

Cox, D. R. and Stuart, A. (1955). “Some quick sign tests for trend in location and dispersion”. *Biometrika* 42 (1), pp. 80–95.

Creech, S. (2018). *t-test.* URL: `statisticallysignificantconsulting.com/ttest` (visited on June 21, 2018).

Cummins, J., Hussey, I., and Hughes, S. (2019). “The AMPeror’s New Clothes: Performance on the Affect Misattribution Procedure is Mainly Driven by Awareness of Influence of the Primes”. *PsyArXiv: 10.31234/osf.io/d5zn8.*

Da Silva, G. T., Logan, B. R., and Klein, J. P. (2009). “Methods for equivalence and noninferiority testing”. *Biology of Blood and Marrow Transplantation* 15 (1), pp. 120–127.

Dawson, R. (2011). “How significant is a boxplot outlier”. *Journal of Statistics Education* 19 (2), pp. 1–12.

De Winter, J. C. (2013). “Using the Student’s t-test with extremely small sample sizes”. *Practical Assessment, Research & Evaluation* 18 (10), pp. 1–12.

De Winter, J. C. and Dodou, D. (2010). “Five-point Likert items: t-test versus Mann-Whitney-Wilcoxon”. *Practical Assessment, Research & Evaluation* 15 (11), pp. 1–12.

Delacre, M., Lakens, D., and Leys, C. (2017). "Why psychologists should by default use Welch's test instead of Student's test". *International Review of Social Psychology* 30 (1).

Delaney, H. D. and Vargha, A. (2000). "The Effect of Nonnormality on Student's Two-Sample T Test." *Meeting of the American Educational Research Association.*

Derrick, B. (2017a). *Partiallyoverlapping: Partially overlapping samples t-tests.* R package version 2.

Derrick, B. (2017b). "Statistics: New t-tests for the comparison of two partially overlapping samples". *Faculty of Environment and Technology Degree Show, Frenchay Campus, UWE, Bristol, 1 June 2017.* URL: `http://eprints.uwe.ac.uk/31766`.

Derrick, B. (2018a). "An outlier in an independent samples design". *Royal Statistical Society Conference, 2018, Cardiff City Hall, Cardiff, Wales, 3-7 September 2018.* URL: `http://eprints.uwe.ac.uk/37137`.

Derrick, B. (2018b). "To preliminary test or not to preliminary test, that is the question". *Research Students Conference in Probability and Statistics, Sheffield University, 24-27 July 2018.* URL: `http://eprints.uwe.ac.uk/37005`.

Derrick, B. and Toher, D. (2016). *Eurovision 2016: Was Australia Robbed? (2018 repost).* URL: `https://dtoher.wordpress.com/2018/05/11/eurovision-2016-was-australia-robbed-2` (visited on June 6, 2018).

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

Derrick, B., Toher, D., and White, P. (2017). "How to compare the means of two samples that include paired observations and independent observations: A companion to Derrick, Russ, Toher and White (2017)". *The Quantitative Methods in Psychology* 13 (2), pp. 120–126.

Derrick, B., Toher, D., and White, P. (2019). "The performance of the partially overlapping samples t-tests at the limits". *arXiv:1906.01006.*

Derrick, B. and White, P. (2017). "Comparing two samples from an individual Likert question". *International Journal of Mathematics and Statistics* 18 (3), pp. 1–13.

Derrick, B. and White, P. (2018). "Methods for comparing the responses from a Likert question, with paired observations and independent observations in each of two samples". *International Journal of Mathematics and Statistics* 19 (3), pp. 84–93.

Derrick, B., White, P., and Toher, D. (2017). "An Inverse Normal Transformation solution for the comparison of two samples that contain both paired observations and independent observations". *arXiv:1708.00347.*

Derrick, B., White, P., and Toher, D. (in press). “Parametric and non-parametric tests for the comparison of two samples which both include paired and unpaired observations”. *Journal of Modern Applied Statistical Methods*.

Derrick, B., Dobson-Mckittrick, A., Toher, D., and White, P. (2015). “Test statistics for comparing two proportions with partially overlapping samples”. *Journal of Applied Quantitative Methods* 10 (3), pp. 1–14.

Derrick, B., Russ, B., Toher, D., and White, P. (2017a). “Test statistics for the comparison of means for two samples which include both paired observations and independent observations.” *Journal of Modern Applied Statistical Methods* 16 (1), pp. 137–157.

Derrick, B., Broad, A., Toher, D., and White, P. (2017b). “The impact of an extreme observation in a paired samples design”. *metodološki zvezki-Advances in Methodology and Statistics* 14.

Derrick, B., Ruck, A., Toher, D., and White, P. (2018). “Tests for equality of variances between two samples which contain both paired observations and independent observations”. *Journal of Applied Quantitative Methods* 13 (2), pp. 36–47.

Devroye, L. (1986). “Sample-based non-uniform random variate generation”. *Proceedings of the 18th conference on Winter simulation.* ACM, pp. 260–265.

Digby, P. G. (1983). “Approximating the tetrachoric correlation coefficient”. *Biometrics*, pp. 753–757.

Donders, A. R. T., Van der Heijden, G. J., Stijnen, T., and Moons, K. G. (2006). “A gentle introduction to imputation of missing values”. *Journal of Clinical Epidemiology* 59 (10), pp. 1087–1091.

Dong, Y. and Peng, C. J. (2013). “Principled missing data methods for researchers”. *SpringerPlus* 2 (1), p. 222.

Duncan, G. and Layard, M. (1973). “A Monte-Carlo study of asymptotically robust tests for correlation coefficients”. *Biometrika* 60 (3), pp. 551–558.

Edgington, E. and Onghena, P. (2007). *Randomization tests.* CRC Press.

Eisenhart, C. (1947). “The assumptions underlying the analysis of variance”. *Biometrics* 3 (1), pp. 1–21.

Ekbohm, G (1981). “A test for the equality of variances in the paired case with incomplete data”. *Biometrical Journal* 23 (3), pp. 261–265.

Ekbohm, G. (1976). “On comparing means in the paired case with incomplete data on both responses”. *Biometrika* 63 (2), pp. 299–304.

Ekbohm, G. (1982). “On comparing variances in the paired case with incomplete data”. *Biometrika* 69 (3), pp. 670–673.

Fagerland, M. W. (2012). "t-tests, non-parametric tests, and large studies, a paradox of statistical practice?" *Medical Research Methodology* 12 (78), pp. 1–7.

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

Fagerland, M. W. and Sandvik, L. (2009b). "The Wilcoxon-Mann-Whitney test under scrutiny". *Statistics in Medicine* 28 (10), pp. 1487–1497.

Fagerland, M. W., Sandvik, L., and Mowinckel, P. (2011). "Parametric methods outperformed non-parametric methods in comparisons of discrete numerical variables". *Medical Research Methodology* 11 (44), pp. 1–8.

Fairfield-Smith, H. (1936). "The problem of comparing the result of two experiments with unequal errors". *Journal Council for Scientific and Industrial Research* 9, pp. 211–212.

Fay, M. P. (2011). *Exact McNemar's Test and Matching Confidence Intervals.* URL: `http://www2.uaem.mx/r-mirror/web/packages/exact2x2/vignettes/exactMcNemar.pdf` (visited on July 7, 2019).

Fay, M. P. and Proschan, M. A. (2010). "Wilcoxon-Mann-Whitney or t-test? On assumptions for hypothesis tests and multiple interpretations of decision rules". *Statistics Surveys* 4, pp. 1–39.

Fenton, R., Jones, C., White, P., and Derrick, B. (2018). "They didn't really see the point: Getting students through the door. A mixed-methods exploratory evaluation of a bystander program for the prevention of violence against women in male-dominated university environments". *unpublished manuscript.*

Fisher, M. J., Marshall, A. P., and Mitchell, M. (2011). "Testing differences in proportions". *Australian Critical Care* 24 (2), pp. 133–138.

Fisher, R. A. (1935). "The fiducial argument in statistical inference". *Annals of Human Genetics* 6 (4), pp. 391–398.

Fisher, R. A. (1925). *Statistical methods for research workers.* Genesis Publishing Pvt Ltd.

Fisher, R. A. (1941). "The asymptotic approach to Behrens's integral". *Annals of Human Genetics* 11 (1), pp. 141–172.

Fisher, R. A., Yates, F., et al. (1938). *Statistical tables for biological, agricultural and medical research.* Oliver and Boyd Ltd, London.

Forbes, C., Evans, M., Hastings, N., and Peacock, B. (2011). *Statistical distributions.* John Wiley & Sons.

Fowler, R. L. (1987). "Power and robustness in product-moment correlation". *Applied Psychological Measurement* 11 (4), pp. 419–428.

Fradette, K., Keselman, H., Lix, L., Algina, J., and Wilcox, R. R. (2003). "Conventional and robust paired and independent-samples t tests: Type

I error and power rates". *Journal of Modern Applied Statistical Methods* 2 (2), pp. 1–39.

Frigge, M., Hoaglin, D. C., and Iglewicz, B. (1989). "Some implementations of the boxplot". *The American Statistician* 43 (1), pp. 50–54.

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

Gardner, M. J. and Altman, D. G. (1986). "Confidence intervals rather than p-values: estimation rather than hypothesis testing." *British Medical Journal* 292 (6522), pp. 746–750.

Gart, J. J. (1971). "The comparison of proportions: a review of significance tests, confidence intervals and adjustments for stratification". *Revue de l'Institut International de Statistique*, pp. 148–169.

Gastwirth, J. L., Gel, Y. R., and Miao, W. (2009). "The impact of Levene's test of equality of variances on statistical theory and practice". *Statistical Science*, pp. 343–360.

Gebski, V. J. and Keech, A. C. (2003). "Statistical methods in clinical trials". *The Medical Journal of Australia* 178 (4), pp. 182–184.

Ghasemi, A. and Zahediasl, S. (2012). "Normality tests for statistical analysis: a guide for non-statisticians". *International Journal of Endocrinology and Metabolism* 10 (2), pp. 486–489.

Gladwell, M. (2008). *Outliers: The story of success.* Hachette UK.

Good, P. (2013). *Permutation tests: a practical guide to resampling methods for testing hypotheses.* Springer Science & Business Media.

Graham, J. W. (2009). "Missing data analysis: Making it work in the real world". *Annual Review of Psychology* 60, pp. 549–576.

Grambsch, P. M. (1994). "Simple robust tests for scale differences in paired data". *Biometrika* 81 (2), pp. 359–372.

Graybill, F. A. F. A. (1976). *Theory and application of the linear model.* Duxbury Press.

Grimes, B. A., Federer, W. T., et al. (1982). *BU-762-M, Revised edition: comparison of means from populations with unequal variances, Biometrics unit technical reports.*

Grubbs, F. E. (1969). "Procedures for detecting outlying observations in samples". *Technometrics* 11 (1), pp. 1–21.

Guerrero Segura, R. A. (2019). "The use of online video-based forums for the improvement of suprasegmentals". MA thesis. Universidad Casa Grande. Departamento de Posgrado.

Guiard, V. and Rasch, D. (2004). "The robustness of two sample tests for means. A reply on von Eye's comment". *Psychology Science* 46 (4), pp. 549–554.

Guo, B. and Yuan, Y. (2017). “A comparative review of methods for comparing means using partially paired data”. *Statistical Methods in Medical Research* 26 (3), pp. 1323–1340.

Guo, J.-H. and Luh, W.-M. (2000). “An invertible transformation two-sample trimmed t-statistic under heterogeneity and nonnormality”. *Statistics and Probability Letters* 49 (1), pp. 1–7.

Gurland, J. and McCullough, R. S. (1962). “Testing equality of means after a preliminary test of equality of variances”. *Biometrika* 49 (3-4), pp. 403–417.

Halperin, M., Hartley, H. O., and Hoel, P. G. (1965). “Recommended standards for statistical symbols and notation: COPSS committee on symbols and notation”. *The American Statistician* 19 (3), pp. 12–14.

Hart, A. (2001). “Mann-Whitney test is not just a test of medians: differences in spread can be important”. *British Medical Journal* 323 (7309), pp. 391–393.

Havlicek, L. L. and Peterson, N. L. (1977). “Effect of the violation of assumptions upon significance levels of the Pearson r”. *Psychological Bulletin* 84 (2), pp. 373–377.

Heeren, T. and D’Agostino, R. (1987). “Robustness of the two independent samples t-test when applied to ordinal scaled data”. *Statistics in Medicine* 6 (1), pp. 79–90.

Hewitt, C. E., Kumaravel, B., Dumville, J. C., Torgerson, D. J., Group, T. A. S., et al. (2010). “Assessing the impact of attrition in randomized controlled trials”. *Journal of Clinical Epidemiology* 63 (11), pp. 1264–1270.

Hodge, D. R. and Gillespie, D. (2003). “Phrase completions: An alternative to Likert scales”. *Social Work Research* 27 (1), pp. 45–55.

Hodge, V. J. and Austin, J. (2004). “A survey of outlier detection methodologies”. *Artificial Intelligence Review* 22 (2), pp. 85–126.

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

Hogg, R. V. (1977). “An introduction to robust procedures”. *Communications in Statistics Theory and Methods* 6 (9), pp. 789–794.

Hogg, R. V. (1979). “Statistical robustness: one view of its use in applications today”. *The American Statistician* 33 (3), pp. 108–115.

Holis, S. and Campbell, F. (1999). “What is meant by intention to treat analysis. Survey of published randomised controlled trials.” *British Medical Journal* 19, pp. 670–674.

Hollander, M., Wolfe, D. A., and Chicken, E. (2013). *Nonparametric statistical methods.* John Wiley & Sons.

Hosgood, S. A., Saeb-Parsy, K., Wilson, C., Callaghan, C., Collett, D., and Nicholson, M. L. (2017). “Protocol of a randomised controlled, open-label trial of ex vivo normothermic perfusion versus static cold storage in donation after circulatory death renal transplantation”. *British Medical Journal open* 7 (1), e012237.

Howell, D. C. (2012). *Statistical methods for psychology.* Cengage Learning.

Huber, P. J. (2011). “Robust statistics”. *International Encyclopedia of Statistical Science.* Springer, pp. 1248–1251.

Jamieson, S. et al. (2004). “Likert scales: how to (ab)use them”. *Medical Education* 38 (12), pp. 1217–1218.

Janssen, M., Busch, C., Rödiger, M., and Hamm, U. (2016). “Motives of consumers following a vegan diet and their attitudes towards animal agriculture”. *Appetite* 105, pp. 643–651.

Jennings, M. J., Zumbo, B. D., and Joula, J. F. (2002). “The robustness of validity and efficiency of the related samples t-test in the presence of outliers”. *Psicologica* 23 (2), pp. 415–450.

Johnson, R. A. and Bhattacharyya, G. K. (1996). *Statistics: Principles and methods.* John Wiley & Sons.

Kang, H. (2013). “The prevention and handling of the missing data”. *Korean Journal of Anesthesiology* 64 (5), pp. 402–406.

Keller, G. (2014). *Statistics for management and economics.* Nelson Education.

Kellermann, A., Bellara, A. P., De Gil, P. R., Nguyen, D., Kim, E. S., Chen, Y.-H., and Kromey, J. (2013). “Variance heterogeneity and Non-Normality: How SAS PROC TTEST® can keep us honest”. *Proceedings of the Annual SAS Global Forum Conference, Cary, NC: SAS Institute Inc.* Citeseer.

Kenney, J. F. J. F. and Keeping, E. S. (1951). *Mathematics of statistics; part two.* D. Van Nostrand Company Inc; Toronto; Princeton; New Jersey; London; New York,; Affiliated East-West Press Pvt-Ltd; New Delhi.

Keselman, H., Wilcox, R. R., Kowalchuk, R. K., and Olejnik, S. (2002). “Comparing trimmed or least squares means of two independent skewed populations”. *Biometrical Journal* 44 (4), pp. 478–489.

Keselman, H., Othman, A. R., Wilcox, R. R., and Fradette, K. (2004). “The new and improved two-sample t-test”. *Psychological Science* 15 (1), pp. 47–51.

Kim, B. S., Kim, I., Lee, S., Kim, S., Rha, S. Y., and Chung, H. C. (2005). “Statistical methods of translating microarray data into clinically relevant diagnostic information in colorectal cancer”. *Bioinformatics* 21 (4), pp. 517–528.

Kim, H.-Y. (2013). “Statistical notes for clinical researchers: assessing normal distribution (2) using skewness and kurtosis”. *Restorative dentistry & endodontics* 38 (1), pp. 52–54.

Kimotho, T. N. (2018). “Influence of mergers and acquisitions on financial performance of firms listed in Nairobi securities exchange”. PhD thesis. Strathmore University.

Konietschke, F. and Pauly, M. (2014). “Bootstrapping and permuting paired t-test type statistics”. *Statistics and Computing* 24 (3), pp. 283–296.

Kornbrot, D. (2005). “Point biserial correlation”. *StatsRef: Statistics Reference Online.*

Laerd (2018). *independent t test using stata.* URL: `https://statistics.laerd.com/stata-tutorials/independent-t-test-using-stata.php` (visited on June 6, 2018).

Lancaster, H. (1961). “The combination of probabilities: an application of orthonormal functions”. *Australian & New Zealand Journal of Statistics* 3 (1), pp. 20–33.

Lee, A. F. and Gurland, J. (1975). “Size and power of tests for equality of means of two normal populations with unequal variances”. *Journal of the American Statistical Association* 70 (352), pp. 933–941.

Leon, A. C., Mallinckrodt, C. H., Chuang-Stein, C., Archibald, D. G., Archer, G. E., and Chartier, K. (2006). “Attrition in randomized controlled clinical trials: methodological issues in psychopharmacology”. *Biological Psychiatry* 59 (11), pp. 1001–1005.

Levene, H. (1961). “Robust tests for equality of variances”. *Contributions to probability and statistics. Essays in honor of Harold Hotelling*, pp. 279–292.

Levine, T. R., Weber, R., Park, H. S., and Hullett, C. R. (2008a). “A communication researchers’ guide to null hypothesis significance testing and alternatives”. *Human Communication Research* 34 (2), pp. 188–209.

Levine, T. R., Weber, R., Hullett, C., Park, H. S., and Lindsey, L. L. M. (2008b). “A critical assessment of null hypothesis significance testing in quantitative communication research”. *Human Communication Research* 34 (2), pp. 171–187.

Likert, R. (1932). “A technique for the measurement of attitudes”. *Archives of Psychology* 22 (140).

Lim, T.-S. and Loh, W (1996). “A comparison of tests of equality of variances”. *Computational Statistics and Data Analysis* 22 (3), pp. 287–301.

Lin, P.-E. (1973). “Procedures for testing the difference of means with incomplete data”. *Journal of the American Statistical Association* 68 (343), pp. 699–703.

Lin, P.-E. and Stivers, L. E. (1974). "On difference of means with incomplete data". *Biometrika* 61 (2), pp. 325–334.

Lissitz, R. W. and Chardos, S. (1975). "A study of the effect of the violation of the assumption of independent sampling upon the type I error rate of the two-group t-test". *Educational and Psychological Measurement* 35 (2), pp. 353–359.

Lix, L. M. and Keselman, H. (1998). "To trim or not to trim: tests of location equality under heteroscedasticity and nonnormality". *Educational and Psychological Measurement* 58 (3), pp. 409–429.

Loh, W.-y. (1987). "Some modifications of Levene's test of variance homogeneity". *Journal of Statistical Computation and Simulation* 28 (3), pp. 213–226.

Looney, S. W. and Jones, P. W. (2003). "A method for comparing two normal means using combined samples of correlated and uncorrelated data". *Statistics in Medicine* 22 (9), pp. 1601–1610.

Looney, S. W. and McCracken, C. E. (2018). "Much ado about almost nothing: methods for dealing with limited data". *Journal of Mathematics and System Science* (8), pp. 44–58.

Lu, Z. and Yuan, K.-H. (2010). *Encyclopedia of research design: Welch's t test*. Thousand Oaks, Sage, pp. 1620–1623.

Lumley, T., Diehr, P., Emerson, S., and Chen, L. (2002). "The importance of the normality assumption in large public health data sets". *Annual Review of Public Health* 23 (1), pp. 151–169.

Mahdizadeh, M. (2018). "Testing normality based on sample information content". *International Journal of Mathematics and Statistics* 19 (1), pp. 1–18.

Maisel, N. C. and Fingerhut, A. W. (2011). "California's ban on same-sex marriage: The campaign and its effects on gay, lesbian, and bisexual individuals". *Journal of Social Issues* 67 (2), pp. 242–263.

Mantilla, L. B. and Terpstra, J. T. (2018). "Means, Medians, and Multivariate Mixed Design Data". *American Journal of Mathematical and Management Sciences* 37 (1), pp. 56–65.

Markowski, C. A. and Markowski, E. P. (1990). "Conditions for the effectiveness of a preliminary test of variance". *The American Statistician* 44 (4), pp. 322–326.

Marozzi, M. (2011). "Levene type tests for the ratio of two scales". *Journal of Statistical Computation and Simulation* 81 (7), pp. 815–826.

Martinez-Camblor, P., Corral, N., and De La Hera, J. (2013). "Hypothesis test for paired samples in the presence of missing data". *Journal of Applied Statistics* 40 (1), pp. 76–87.

Matsumoto, M. and Nishimura, T. (1998). "Mersenne twister: a 623 dimension equidistributed uniform pseudo-random number generator". *ACM Transactions on Modeling and Computer Simulation (TOMACS)* 8 (1), pp. 3–30.

Mayfield, P. (2013). *Beyond the t-Test and F-Test*. URL: `http://www.sigmazone.com/Articles_BeyondthetandFTest.htm` (visited on June 6, 2018).

McCulloch, C. E. (1987). "Tests for equality of variances with paired data". *Communications in statistics-theory and methods* 16 (5), pp. 1377–1391.

McSweeney, M. and Penfield, D. (1969). "The normal scores test for the two sample problem". *British Journal of Mathematical and Statistical Psychology* 22 (2), pp. 177–192.

Mehrotra, D (2004). "Letter to the editor, a method for comparing two normal means using combined samples of correlated and uncorrelated data". *Statistics in Medicine* 23, pp. 1179–1180.

Mendenhall, W., Beaver, R. J., and Beaver, B. M. (2012). *Introduction to probability and statistics*. Cengage Learning.

Mendes, M. and Pala, A. (2003). "Type I error rate and power of three normality tests". *Pakistan Journal of Information and Technology* 2 (2), pp. 135–139.

Micceri, T. (1989). "The unicorn, the normal curve, and other improbable creatures." *Psychological Bulletin* 105 (1), p. 156.

Mickelson, W. T. (2013). "A Monte Carlo simulation of the robust rank-order test under various population symmetry conditions". *Journal of Modern Applied Statistical Methods* 12 (1), p. 7.

Mircioiu, C. and Atkinson, J. (2017). "A comparison of parametric and nonparametric methods applied to a Likert scale". *Pharmacy* 5 (2).

Mirtagiouglu, H., Yiugit, S., Mendecs, E., and Mendecs, M. (2017). "A Monte Carlo Simulation Study for Comparing Performances of Some Homogeneity of Variances Tests". *Journal of Applied Quantitative Methods* 12 (1), pp. 1–11.

Moser, B. K. and Stevens, G. R. (1992). "Homogeneity of variance in the two-sample means test". *The American Statistician* 46 (1), pp. 19–21.

Moser, B. K., Stevens, G. R., and Watts, C. L. (1989). "The two-sample t test versus Satterthwaite's approximate F test". *Communications in Statistics-Theory and Methods* 18 (11), pp. 3963–3975.

Mudholkar, G. S., Wilding, G. E., and Mietlowski, W. L. (2003). "Robustness properties of the Pitman–Morgan test". *Communications in Statistics-Theory and Methods* 32 (9), pp. 1801–1816.

Murphy, B. (1976). “Comparison of some two sample means tests by simulation”. *Communication in Statistics-Simulation and Computation* 5 (1), pp. 23–32.

Musil, C. M., Warner, C. B., Yobas, P. K., and Jones, S. L. (2002). “A comparison of imputation techniques for handling missing data”. *Western Journal of Nursing Research* 24 (7), pp. 815–829.

Nanna, M. J. and Sawilowsky, S. S. (1998). “Analysis of Likert scale data in disability and medical rehabilitation research.” *Psychological Methods* 3 (1), p. 55.

NeMoyer, A., Kelley, S., Zelle, H., and Goldstein, N. E. (2018). “Attorney perspectives on juvenile and adult clients’ competence to plead guilty”. *Psychology, Public Policy, and Law* 24 (2), p. 171.

Newcombe, R. G. (1998). “Two-sided confidence intervals for the single proportion: comparison of seven methods”. *Statistics in Medicine* 17 (8), pp. 857–872.

Nguyen, D., Rodriguez, 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, Cary, NC.*

Nordstokke, D. W. and Zumbo, B. D. (2007). “A cautionary tale about Levene’s tests for equal variances”. *Journal of Educational Research & Policy Studies* 7 (1), pp. 1–14.

Nordstokke, D. W. and Zumbo, B. D. (2010). “A new nonparametric Levene test for equal variances.” *Psicologica: International Journal of Methodology and Experimental Psychology* 31 (2), pp. 401–430.

Norman, G. (2010). “Likert scales, levels of measurement and the ‘laws’ of statistics”. *Advances in health sciences education* 15 (5), pp. 625–632.

Nunnally, J. C. (1994). *Psychometric theory.* Tata McGraw Hill Education.

Odén, A., Wedel, H., et al. (1975). “Arguments for Fisher’s permutation test”. *The Annals of Statistics* 3 (2), pp. 518–520.

Orr, J. M., Sackett, P. R., and Dubois, C. L. (1991). “Outlier detection and treatment in I/O psychology: A survey of researcher beliefs and an empirical illustration”. *Personnel Psychology* 44 (3), pp. 473–486.

Osborne, J. W. and Overbay, A. (2004). “The power of outliers (and why researchers should always check for them)”. *Practical Assessment, Research & Evaluation* 9 (6), pp. 1–12.

Paley, R. E. A. C. and Wiener, N. (1934). *Fourier transforms in the complex domain.* Vol. 19. American Mathematical Society.

Parra-Frutos, I. (2009). “The behaviour of the modified Levene’s test when data not normally distributed”. *Computational Statistics* 24 (4), pp. 671–693.

Pearce, J. and Derrick, B. (2019). “Preliminary testing: the devil of statistics?” *Reinvention: an International Journal of Undergraduate Research* 12.

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

Peró-Cebollero, M. and Guàrdia-Olmos, J. (2013). “The Adequacy of Different Robust Statistical Tests in Comparing Two Independent Groups.” *Psicologica: International Journal of Methodology and Experimental Psychology* 34 (2), pp. 407–424.

Pilkington, M. (2017). “Psychosocial interventions to support breast cancer patients affected by treatment-related hair loss”. PhD thesis. University of the West of England.

Pocock, S. J. (1985). “Current issues in the design and interpretation of clinical trials.” *British Medical Journal* 290 (6461), pp. 39–42.

Polster, A. V., Palsson, O. S., Törnblom, H., Öhman, L., Sperber, A. D., Whitehead, W. E., and Simrén, M. (2019). “Subgroups of IBS patients are characterized by specific, reproducible profiles of GI and non-GI symptoms and report differences in healthcare utilization: A population-based study”. *Neurogastroenterology and Motility* 31 (1), e13483.

Pratt, J. W. (1959). “Remarks on zeros and ties in the Wilcoxon signed rank procedures”. *Journal of the American Statistical Association* 54 (287), pp. 655–667.

Preece, D. (1982). “T is for trouble (and textbooks): a critique of some examples of the paired-samples t-test”. *The Statistician* 31 (2), pp. 169–195.

Puth, M.-T., Neuhäuser, M., and Ruxton, G. D. (2014). “Effective use of Pearson’s product moment correlation coefficient”. *Animal Behaviour* 93, pp. 183–189.

Qi, Q., Yan, L., and Tian, L. (2018). “Testing equality of means in partially paired data with incompleteness in single response”. *Statistical methods in medical research*, p. 0962280218765007.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: `http://www.R-project.org/`.

R Studio Team (2019). *RStudio: Integrated Development Environment for R*. RStudio, Inc. Boston, MA. URL: `http://www.rstudio.com/`.

Ramosaj, B., Amro, L., and Pauly, M. (2018). “A cautionary tale on using imputation methods for matched pairs design”. *arXiv:1806.06551*.

Ramosaj, B. and Pauly, M. (2017). “Who wins the Miss Contest for Imputation Methods? Our Vote for Miss BooPF”. *arXiv:1711.11394*.

Rao, C. R. and Lovric, M. M. (2016). “Testing Point Null Hypothesis of a Normal Mean and the Truth: 21st Century Perspective”. *Journal of Modern Applied Statistical Methods* 15 (2), pp. 2–21.

Rasch, D. and Guiard, V. (2004). “The robustness of parametric statistical methods”. *Psychology Science* 46, pp. 175–208.

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), pp. 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), pp. 2706–2720.

Ray, S., Nishimura, R., Clarke, S., and Simpson, I. (2016). “Maintaining good clinical practice: publication of outcomes and handling of outliers”. *Heart* 102 (19), pp. 1518–1519.

Raymundo, L., Burdick, D, Hoot, W., Miller, R., Brown, V, Reynolds, T, Gault, J, Idechong, J, Fifer, J, and Williams, A (2019). “Successive bleaching events cause mass coral mortality in Guam, Micronesia”. *Coral Reefs* 38 (4), pp. 1–24.

Razali, N. M., Wah, Y. B., et al. (2011). “Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests”. *Journal of Statistical Modeling and Analytics* 2 (1), pp. 21–33.

Rempala, G. A. and Looney, S. W. (2006). “Asymptotic properties of a two sample randomized test for partially dependent data”. *Journal of Statistical Planning and Inference* 136 (1), pp. 68–89.

Rietveld, T. and Van Hout, R. (2017). “The paired t test and beyond: Recommendations for testing the central tendencies of two paired samples in research on speech, language and hearing pathology”. *Journal of Communication Disorders* 69, pp. 44–57.

Roberson, P. K., Shema, S., Mundfrom, D., and Holmes, T. (1994). “Analysis of paired Likert data: how to evaluate change and preference questions”. *Family Medicine* 27 (10), pp. 671–675.

Rochon, J., Gondan, M., and Kieser, M. (2012). “To test or not to test: Preliminary assessment of normality when comparing two independent samples”. *Medical Research Methodology* 12 (81), pp. 1–11.

Rodgers, J. L. and Nicewander, W. A. (1988). “Thirteen ways to look at the correlation coefficient”. *The American Statistician* 42 (1), pp. 59–66.

Rosenthal, R. and Rosnow, R. L. (1991). *Essentials of behavioral research: Methods and data analysis.* Vol. 2. McGraw-Hill New York.

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), pp. 688–690.

Ruxton, G. and Neuhauser, M. (2018). "Striving for simple but effective advice for comparing the central tendency of two populations". *Jounal of Modern Applied Statistical Methods* 17 (2).

Sakia, R. (1992). "The Box-Cox transformation technique: a review". *The Statistician*, pp. 169–178.

Samawi, H., Yu, L., and Vogel, R. L. (2015). "On some nonparametric tests for partially observed correlated data: proposing new tests". *Journal of Statistical Theory and Applications* 14 (2), p. 131.

Samawi, H. M. and Vogel, R. (2011). "Tests of homogeneity for partially matched-pairs data". *Statistical Methodology* 8 (3), pp. 304–313.

Samawi, H. M. and Vogel, R. (2014a). "Notes on two sample tests for partially correlated (paired) data". *Journal of Applied Statistics* 41 (1), pp. 109–117.

Samawi, H. M. and Vogel, R. (2014b). "Notes on two sample tests for partially correlated (paired) data". *Journal of Applied Statistics* 41 (1), pp. 109–117.

Sawilowsky, S. S. (2005). "Misconceptions leading to choosing the t-test over the Wilcoxon-Mann-Whitney test for shift in location parameter". *Journal of Modern Applied Statistical Methods* 4 (2), pp. 598–600.

Sawilowsky, S. S. and Hillman, S. B. (1992). "Power of the independent samples t-test under a prevalent psychometric measure distribution." *Journal of Consulting and Clinical Psychology* 60 (2), p. 240.

Sawilowsky, S. S. (2016). "Rao-Lovric and the Triwizard Point Null Hypothesis Tournament". *Journal of Modern Applied Statistical Methods* 15 (2), p. 4.

Sawilowsky, S. S. and 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 (2), p. 352.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. CRC press.

Scheffé, H. (1970). "Practical solutions of the Behrens-Fisher problem". *Journal of the American Statistical Association* 65 (332), pp. 1501–1508.

Schlomer, G. L., Bauman, S., and Card, N. A. (2010). "Best practices for missing data management in counseling psychology." *Journal of Counseling Psychology* 57 (1), p. 1.

Schulz, K. F. and Grimes, D. A. (2002). "Sample size slippages in randomised trials: exclusions and the lost and wayward". *The Lancet* 359 (9308), pp. 781–785.

Serlin, R. C. (2000). "Testing for robustness in Monte Carlo studies." *Psychological Methods* 5 (2), pp. 230–240.

Servin, B. and Stephens, M. (2007). “Imputation-based analysis of association studies: candidate regions and quantitative traits”. *PLoS Genet* 3 (7), e114.

Shapiro, S. S., Wilk, M. B., and Chen, H. J. (1968). “A comparative study of various tests for normality”. *Journal of the American Statistical Association* 63 (324), pp. 1343–1372.

Shear, B. R., Nordstokke, D. W., and Zumbo, B. D. (2018). “A Note on Using the Nonparametric Levene Test When Population Means Are Unequal”. *Practical Assessment, Research & Evaluation* 23 (13), pp. 1–11.

Shiffler, R. E. (1988). “Maximum z-scores and outliers”. *The American Statistician* 42 (1), pp. 79–80.

Shoemaker, L. H. (2003). “Fixing the F-test for equal variances”. *The American Statistician* 57 (2), pp. 105–114.

Sisson, D. V. and Stocker, H. R. (1989). “Research corner: analyzing and interpreting Likert-type survey data”. *Delta Pi Epsilon Journal* 31 (2), pp. 81–92.

Skovlund, E. and Fenstad, G. U. (2001). “Should we always choose a nonparametric test when comparing two apparently nonnormal distributions?” *Journal of Clinical Epidemiology* 54 (1), pp. 86–92.

Smith, J. and Cody, R. (1997). *Applied statistics and the SAS programming language*. Prentice Hall, Upper Saddle River, NJ.

Stigler, R. G., Becker, K., Kloss, F. R., Gassner, R., and Lepperdinger, G. (2018). “Long-lived murine osteocytes are embodied by craniofacial skeleton in young and old animals whereas they decrease in number in postcranial skeletons at older ages”. *Gerodontology* 35 (4), pp. 391–397.

Stouffer, S. A., Lumsdaine, A. A., Lumsdaine, M. H., Williams Jr, R. M., Smith, M. B., Janis, I. L., Star, S. A., and Cottrell Jr, L. S. (1949). “The American soldier: combat and its aftermath”. *Studies in Social Psychology in World War II* 2.

‘Student’ (1908). “The probable error of a mean”. *Biometrika* 6 (1), pp. 1–25.

‘Student’ (1931). “The Lanarkshire milk experiment”. *Biometrika* 23 (3), pp. 398–406.

Sullivan, L. and D’Agostino, R. (1992). “Robustness of the t-test applied to data distorted from normality by floor effects”. *Journal of Dental Research* 71 (12), pp. 1938–1943.

Sullivan, L. and D’Agostino, R. (1996). “Robustness and power of analysis of covariance applied to data distorted from normality by floor effects: homogeneous regression slopes”. *Statistics in Medicine* 15 (5), pp. 477–496.

Tang, M. and Tang, N. (2004). “Exact tests for comparing two paired proportions with incomplete data”. *Biometrical journal* 46 (1), pp. 72–82.

Thomson, P. (1995). “A hybrid paired and unpaired analysis for the comparison of proportions”. *Statistics in Medicine* 14 (13), pp. 1463–1470.

Tian, G.-L., Zhang, C., and Jiang, X. (2018). “Valid statistical inference methods for a case–control study with missing data”. *Statistical methods in medical research* 27 (4), pp. 1001–1023.

Tippett, L. H. C. et al. (1931). “The Methods of Statistics”. *The methods of Statistics*.

Totton, N. and White, P. (2011). “The ubiquitous mythical normal distribution”. *Applied Statistics Group UWE Bristol*.

Tukey, J. W. (1962). “The future of data analysis”. *The Annals of Mathematical Statistics* 33 (1), pp. 1–67.

Uddin, N. and Hasan, M. (2017). “Testing equality of two normal means using combined samples of paired and unpaired data”. *Communications in Statistics-Simulation and Computation* 46 (3), pp. 2430–2446.

Ury, H. K. and Fleiss, J. L. (1980). “On approximate sample sizes for comparing two independent proportions with the use of Yates’ correction”. *Biometrics*, pp. 347–351.

Van der Wearden, B. (1952). “Order tests for the two-sample problem and their power”. *Indagationes Mathematicae (Proceedings)*. Vol. 55. Elsevier, pp. 453–458.

Van Noorden, R., Maher, B., and Nuzzo, R. (2014). “The top 100 papers”. *Nature News* 514 (7524), pp. 550–551.

Vogt, W. (2018). *Wilcoxon-Pratt test*. URL: `http://methods.sagepub.com/base/download/ReferenceEntry/dictionary-of-statistics-methodology/n2090.xml` (visited on June 20, 2018).

Von Neumann, J. (1951). “Various techniques used in connection with random digits”. *Applied Mathematics Series* 12 (36-38), p. 5.

Von Neumann, J. and Ulam, S. (1951). “Monte Carlo method”. *National Bureau of Standards Applied Mathematics Series* 12, p. 36.

Vonesh, E. F. (1983). “Efficiency of repeated measures designs versus completely randomized designs based on multiple comparisons”. *Communications in Statistics Theory and Methods* 12 (3), pp. 289–301.

Wald, A. and Wolfowitz, J. (1940). “On a test whether two samples are from the same population”. *The Annals of Mathematical Statistics* 11 (2), pp. 147–162.

Walfish, S. (2006). “A review of statistical outlier methods”. *Pharmaceutical Technology* 30 (11), pp. 1–5.

Walsh, M., Srinathan, S. K., McAuley, D. F., Mrkobrada, M., Levine, O., Ribic, C., Molnar, A. O., Dattani, N. D., Burke, A., Guyatt, G., et al. (2014). “The statistical significance of randomized controlled trial results is frequently fragile: a case for a Fragility Index". *Journal of clinical epidemiology* 67 (6), pp. 622–628.

Wang, Y. Y. (1971). "Probabilities of the type I errors of the Welch tests for the Behrens-Fisher problem". *Journal of the American Statistical Association* 66 (335), pp. 605–608.

Weir, I., Gwynllyw, R., and Henderson, K. (2015). *One sample t-test location.* URL: `http://www.statstutor.ac.uk/topics/t-tests/one-samplet-test/?audience=students` (visited on Jan. 3, 2017).

Weir, I. (2018). *Business Statistics.* Applied Statistics Group UWE Bristol.

Welch, B. L. (1938). "The significance of the difference between means when population variances are unequal". *Biometrika* 29 (3/4), pp. 350–362.

Welch, B. L. (1947). "The generalization of Student's' problem when several different population variances are involved". *Biometrika* 34 (1/2), pp. 28–35.

Welch, B. (1951). "On the comparison of several mean values: an alternative approach". *Biometrika* 38 (3/4), pp. 330–336.

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

Whitlock, M. C. (2005). "Combining probability from independent tests: the weighted Z-method is superior to Fisher's approach". *Journal of Evolutionary Biology* 18 (5), pp. 1368–1373.

Wiedermann, W. T. and Alexandrowicz, R. W. (2007). "A plea for more general tests than those for location only: Further considerations on Rasch & Guiard's 'The robustness of parametric statistical methods'". *Psychology Science* 49 (1), pp. 2–12.

Wiedermann, W. T. and von Eye, A. (2013). "Robustness and power of the parametric t-test and the nonparametric Wilcoxon test under non-independence of observations". *Psychological Test and Assessment Modeling* 55 (1), pp. 39–61.

Wilcox, R. R. (1990). "Comparing the means of two independent groups". *Biometrical Journal* 32 (7), pp. 771–780.

Wilcox, R. R. and Charlin, V. L. (1986). "Comparing medians: A Monte Carlo study". *Journal of Educational and Behavioral Statistics* 11 (4), pp. 263–274.

Wilcox, R. R. (2012). *Introduction to robust estimation and hypothesis testing.* Academic Press.

Wilcox, R. R. (2015). "Comparing the variances of two dependent variables". *Journal of Statistical Distributions and Applications* 2 (1), p. 1.

Wilk, M. B. and Gnanadesikan, R. (1968). "Probability plotting methods for the analysis for the analysis of data". *Biometrika* 55 (1), pp. 1–17.

Yu, D., Lim, J., Liang, F., Kim, K., Kim, B. S., and Jang, W. (2012). "Permutation test for incomplete paired data with application to cDNA microarray data". *Computational Statistics & Data Analysis* 56 (3), pp. 510–521.

Yuen, K. K. (1974). "The two-sample trimmed t for unequal population variances". *Biometrika* 61 (1), pp. 165–170.

Zarembka, P. (1990). "Transformation of variables in econometrics". *Econometrics*. Springer, pp. 261–264.

Zhu, H., Xu, X., and Ahn, C. (2019). "Sample size for paired experimental design with incomplete observations of continuous outcomes". *Statistical Methods in Medical Research* 28 (2), pp. 589–598.

Zimmerman, D. W. (1987). "Comparative power of Student t test and Mann-Whitney U test for unequal sample sizes and variances". *The Journal of Experimental Education* 55 (3), pp. 171–174.

Zimmerman, D. W. (1997). "Teacher's corner: A note on interpretation of the paired-samples t test". *Journal of Educational and Behavioral Statistics* 22 (3), pp. 349–360.

Zimmerman, D. W. (1998). "Invalidation of parametric and nonparametric statistical tests by concurrent violation of two assumptions". *The Journal of Experimental Education* 67 (1), pp. 55–68.

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

Zimmerman, D. W. (2005). "Increasing power in paired-samples designs by correcting the Student t statistic for correlation." *Interstat* 28, p. 2008.

Zimmerman, D. W. (2011). "Inheritance of Properties of Normal and Non-Normal Distributions after Transformation of Scores to Ranks." *Psicologica: International Journal of Methodology and Experimental Psychology* 32 (1), pp. 65–85.

Zimmerman, D. W. and Zumbo, B. D. (1993). "Significance testing of correlation using scores, ranks, and modified ranks". *Educational and Psychological Measurement* 53 (4), pp. 897–904.

Zimmerman, D. W. and Zumbo, B. D. (2009). "Hazards in choosing between pooled and separate-variances t-tests". *Psicologica* 30 (2), pp. 371–390.

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/Revue canadienne de psychologie expérimentale* 51 (2), pp. 139–150.

Zumbo, B. D. and Jennings, M. J. (2002). "The robustness of validity and efficiency of the related samples t-test in the presence of outliers". *Psicologica* 23 (2), pp. 415–450.

Appendix 1: Testimonials

For confidentiality all names have been removed in the following list of testimonials received via e-mail, the institution of each person is listed.

‘I’m currently looking at some of the methods we employ for diversity analysis within the Department. One area I am looking at is year on year changes in the percentage of staff declaring themselves as BAME within the Department. This seems to fit the criteria of partially overlapping data (involving staff who leave the department, staff who join the department and staff who remain in place year on year) so I thought I could apply one of the statistics you suggest in the paper (probably z_8)... Excellent.’

Operational Research Analyst, Department for Transport, UK

‘I’ve planned to try to implement (we’ll see) your test... my wife works in neuroscience and experimental psychology, and she is also interested in such method because she sometimes uses mixed experiences combining paired and unpaired participants regarding the complexity of her protocols and the limited time available for each participant. I’m not a statistician, the Derrick, Toher, and White (2017) paper is easier to use for programming and check computing.’

R&D engineer, Decathlon Sports Lab, France

‘We work at the New York City Department of Health and are evaluating the results of a participant survey for a hepatitis C care coordination program. Some of our respondents completed both our baseline and our post survey, but a good number completed only one or the other. We read with great interest your papers on adapted t-tests and z-tests for partially overlapping samples and intend to use your methodology in our evaluation. You’ve very helpfully addressed a problem that has been largely overlooked by the literature, although it must arise often.’

‘This is very helpful. Thank you so much, Ben!’

Bureau of Communicable Diseases, New York, USA

‘I am a post-doctoral research fellow at Massachusetts General Hospital in the United States. I recently received feedback on a journal submission suggesting that I utilize the methods you described in 2015 to compare proportions with partially overlapping samples. In familiarizing myself with that paper, I also came across your 2017 paper discussing a t-test for use with partially overlapping samples, which I also found very helpful. Thank you so much for your time and for your research.’

‘Thank you so much for your response, Ben – I really appreciate it!’

Research fellow, Massachusetts General Hospital/Harvard Medical School, Boston, USA

‘I am an MS student major in Statistics at Portland State University. Now we get a project to deal with a big data set with both paired observations and independent observations and luckily find that your research is a perfect match and you also derived an R package which is such a wonderful work and makes us excited.’

‘Thank you Ben, for sending us your newly published paper which is very impressive. And since the real data would not be allowed to send to us, we only summarized the scenario...which is a perfect match with your new test statistics.’

MS student, Portland State University, USA

‘I am reading your paper Why Welch’s test Type I error robust which was published in the journal of TQMP in 2016. This is a nice paper because it really teaches me how to make a sound simulation design.’

‘Great thanks... The code is very helpful!’

Research student, Florida State University, USA

‘I’m a PhD student in Epidemiology at McGill University in Canada and am hoping to use your methods for overlapping samples in my PhD. I’m very interested in your 2017 paper in the Journal of Modern Applies Statistical Methods, Test statistics for the comparison of means for two samples which include both paired observations and independent observations. I would like to cite your novel t-test equations in a PhD research proposal I am submitting to my academic department. Out of interest, are you also developing other statistical methods for overlapping samples? As a last note, thanks for creating an R package for your methods!'

PhD candidate, McGill University, Canada

[Re: How to compare the means of two samples that include paired observations and independent observations: A companion to Derrick, Russ, Toher and White (2017)]

'I must say I found these new tests very exciting. I also checked the R package. I found everything in good order. Thank you so much for sharing this excellent work with the readers of TQMP.'

Editor, The Quantitative Methods for Psychology.

[Re: Comparing two samples from an individual question on a five point Likert scale]

'We feel that your manuscript would be well-suited to our Cogent Series, a multidisciplinary, open journal platform for the rapid dissemination of peer-reviewed research across all disciplines.'

Editor, Journal of Statistical Computation and Simulation

'Recently I came across your paper Test Statistics for the Comparison of Means for Two Samples at Include Both Paired and Independent Observations. I'm glad that I found your paper since it is one of the rare references with clear guidance on how to address samples which are only partially independent.'

PhD candidate, Department of Land Economy, University of Cambridge, UK

'I was so glad to find an R package for comparing partially overlapping samples, thanks for your work on this!'

Data Scientist, Highmark Health, Pennsylvania

Appendix 2: Outputs

The cumulative number of downloads for the R package ‘Partiallyoverlapping’ is given in Figure A1. The gradient change at the end of 2018 coincides with the release of version 2 featuring the z_8 test for comparing two dichotomous samples.

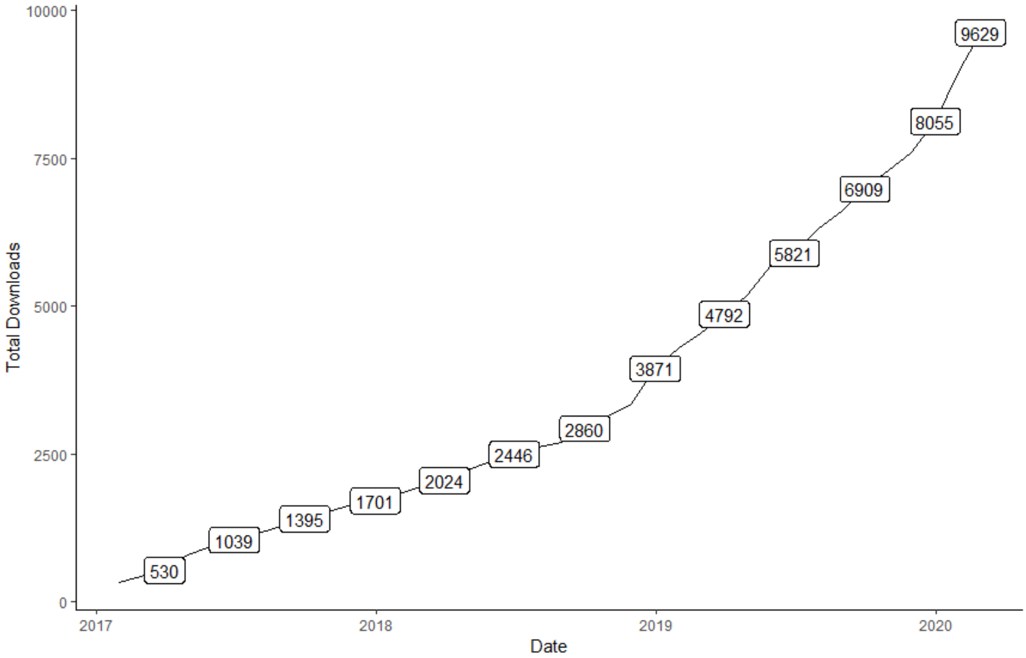


Figure A1: Cumulative downloads of ‘Partiallyoverlapping’ R package, recorded monthly, quarterly values displayed

Table A1 lists open access papers authored towards this thesis, with summary statistics of their impact. Table A2 lists restricted access publications authored towards this thesis with summary statistics of their impact. Each of the publications listed in Table A1 and Table A2 have been peer reviewed, and are available as Appendices P1-P10. Additional unpublished research is listed in Table A3. Table A4 lists relevant conferences attended and presentations.

Table A1: Publication metrics: Open access papers.

Publication	Citations	Appendix
Derrick, B. et al. (2015). "Test statistics for comparing two proportions with partially overlapping samples". *Journal of Applied Quantitative Methods* 10 (3), pp. 1–14	14	P1
Derrick, B., Toher, D., and White, P. (2016). "Why Welch's test is Type I error robust". *The Quantitative Methods in Psychology* 12 (1), pp. 30–38	67	P2
Derrick, B. et al. (2017a). "Test statistics for the comparison of means for two samples which include both paired observations and independent observations." *Journal of Modern Applied Statistical Methods* 16 (1), pp. 137–157	22	P3
Derrick, B., Toher, D., and White, P. (2017). "How to compare the means of two samples that include paired observations and independent observations: A companion to Derrick, Russ, Toher and White (2017)". *The Quantitative Methods in Psychology* 13 (2), pp. 120–126	27	P4
Derrick, B. et al. (2017b). "The impact of an extreme observation in a paired samples design". *metodološki zvezki-Advances in Methodology and Statistics* 14	7	P5
Derrick, B. et al. (2018). "Tests for equality of variances between two samples which contain both paired observations and independent observations". *Journal of Applied Quantitative Methods* 13 (2), pp. 36–47	6	P6
Derrick, B., White, P., and Toher, D. (in press). "Parametric and non-parametric tests for the comparison of two samples which both include paired and unpaired observations". *Journal of Modern Applied Statistical Methods*	2	P7
Pearce, J. and Derrick, B. (2019). "Preliminary testing: the devil of statistics?" *Reinvention: an International Journal of Undergraduate Research* 12	1	P8

Citations from Google Scholar, correct up to 02/2020

Table A2: Publication metrics: Papers not open access.

Publication	Citations	Appendix
Derrick, B. and White, P. (2017). "Comparing two samples from an individual Likert question". *International Journal of Mathematics and Statistics* 18 (3), pp. 1–13	51	P9
Derrick, B. and White, P. (2018). "Methods for comparing the responses from a Likert question, with paired observations and independent observations in each of two samples". *International Journal of Mathematics and Statistics* 19 (3), pp. 84–93	1	P10

Citations from Google Scholar, correct up to 02/2020

Table A3: Unpublished contributions.

Derrick, B., White, P., and Toher, D. (2017). "An Inverse Normal Transformation solution for the comparison of two samples that contain both paired observations and independent observations". *arXiv:1708.00347*
Fenton, R. et al. (2018). "They didn't really see the point: Getting students through the door. A mixed-methods exploratory evaluation of a bystander program for the prevention of violence against women in male-dominated university environments". *unpublished manuscript*
Derrick, B., Toher, D., and White, P. (2019). "The performance of the partially overlapping samples t-tests at the limits". *arXiv:1906.01006*

Table A4: Conference proceedings.

Title	Location	Date
Robust methods for the analysis of partially overlapping samples	Young Statisticians Meeting, University College London	Aug 2016
Statistics: New t-tests for the comparison of two partially overlapping samples [poster]	Faculty of Environment and Technology Degree Show, UWE Bristol	Jun 2017
To preliminary test or not to preliminary test, that is the question	Research Students Conference in Probability and Statistics, University of Sheffield	Jul 2018
An outlier in an independent samples design	Royal Statistical Society Conference, Cardiff City Hall	Sep 2018
For the establishment	Faculty of Environment and Technology Postgraduate Conference, UWE Bristol	Jun 2019

Appendix P1

Derrick, B. et al. (2015). “Test statistics for comparing two proportions with partially overlapping samples”. *Journal of Applied Quantitative Methods* 10 (3), pp. 1–14

Published Version

TEST STATISTICS FOR COMPARING TWO PROPORTIONS WITH PARTIALLY OVERLAPPING SAMPLES

Ben DERRICK[1]
PhD Candidate
University of the West of England, Bristol

E-mail:

Anselma DOBSON-MCKITTRICK[2]
University of the West of England, Bristol

E-mail:

Deirdre TOHER[3]
PhD, Senior Lecturer
University of the West of England, Bristol

E-mail:

Paul WHITE[4]
PhD, Associate professor
University of the West of England, Bristol

E-mail: paul.white@uwe.ac.uk

Abstract

Standard tests for comparing two sample proportions of a dichotomous dependent variable where there is a combination of paired and unpaired samples are considered. Four new tests are introduced and compared against standard tests and an alternative proposal by Choi and Stablein (1982). The Type I error robustness is considered for each of the test statistics. The results show that Type I error robust tests that make use of all the available data are more powerful than Type I error robust tests that do not. The Type I error robustness and the power among tests introduced in this paper using the phi correlation coefficient is comparable to that of Choi and Stablein (1982). The use of the test statistics to form confidence intervals is considered. A general recommendation of the best test statistic for practical use is made.

Key words: Partially overlapping samples, Partially matched pairs, Partially correlated data, Equality of proportions

1. Introduction

Tests for comparing two sample proportions of a dichotomous dependent variable with either two independent or two dependent samples are long established. Let π_1 and π_2 be the proportions of interest for two populations or distributions. The hypothesis being tested is $H_0 : \pi_1 = \pi_2$ against $H_1 : \pi_1 \neq \pi_2$. However, situations arise where a data set comprises a combination of both paired and unpaired observations. In these cases, within a sample there are, say a total of 'n_{12}' observations from both populations, a total of 'n_1' observations only from population one, and a total of 'n_2' observations only from population two. The hypothesis being tested is the same as when either two complete independent samples or two complete dependent samples are present. This situation with respect to comparing two means has been treated poorly in the literature (Martinez-Camblor et al, 2012). This situation with respect to comparing proportions has similarly been poorly treated.

Early literature in this area with respect to comparing proportions, refers to paired samples studies in the presence of incomplete data (Choi and Stablein, 1982; Ekbohlm, 1982), or missing data (Bhoj, 1978). These definitions have connotations suggesting that observations are missing only by accident. Recent literature for this scenario refers to partially matched pairs (Samawi and Vogel, 2011), however this terminology may be construed as the pairs themselves not being directly matched. Alternatively, the situation outlined can be referred to as part of the 'partially overlapping samples framework' (Martinez-Camblor et al, 2012). This terminology is more appropriate to cover scenarios where paired and independent samples may be present by accident or design. Illustrative scenarios where partially overlapping samples may arise by design include:

i) Where the samples are taken from two groups with some common element. For example, in education, when comparing the pass rate for two optional modules, where a student may take one or both modules.

ii) Where the samples are taken at two points in time. For example, an annual survey of employee satisfaction will include new employees that were not employed at time point one, employees that left after time point one and employees that remained in employment throughout.

iii) When some natural pairing occurs. For example, a survey taken comparing views of males and females, there may be some matched pairs 'couples' and some independent samples 'single'.

Repeated measures designs can have compromised internal validity through familiarity (e.g. learning, memory or practise effects). Likewise, a matched design can have compromised internal validity through poor matching. However, if a dependent design can avoid extraneous systematic bias, then paired designs can be advantageous when contrasted with between subjects or independent designs. The advantages of paired designs arise by each pair acting as its own control helping to have a fair comparison. This allows differences or changes between the two samples to be directly examined (i.e. focusing directly on the phenomenon of interest). This has the result of removing systematic effects between pairs. This leads to increased power or a reduction in the sample size required to retain power compared with the alternative independent design. Accordingly, a method of analysis for partially overlapping samples that takes into account any pairing, but does not lose the unpaired information, would be beneficial.

Historically, when analysing partially overlapping samples, a practitioner will choose between discarding the paired observations or discarding the independent observations and proceeding to perform the corresponding 'standard' test. It is likely the decision will be based on the sample sizes of the independent and paired observations. Existing 'standard' approaches include:

Option 1: Discarding all paired observations and performing Pearson's Chi square test of association on the unpaired data.

Option 2: Discarding all unpaired observations and performing McNemar's test on the paired data.

Option 3: Combining p-values of independent tests for paired and unpaired data. This can be done by applying Fisher's inverse Chi square method or Tippett's test. These approaches make use of all of the available data. These techniques were considered by Samawi and Vogel (2011) and are shown to be more powerful than techniques that discard data. However, it should be noted that the authors did not consider Type I error rates.

Other ad-hoc approaches for using all available data include randomly pairing any unpaired observations, or treating all observations as unpaired ignoring any pairing. These ad-hoc approaches are clearly incorrect practice and further emphasise the need for research into statistically valid approaches.

Choi and Stablein (1982) performed a small simulation study to consider standard approaches and ultimately recommended an alternative test making use of all the available data as the best practical approach. This alternative proposal uses one combined test statistic weighting the variance of the paired and independent samples, see Section 3.2 for definition. The authors additionally considered an approach using maximum likelihood estimators for the proportions. This approach was found to be of little practical benefit in terms of Type I error rate or power. Others have also considered maximum likelihood approaches. For example Thomson (1995) considered a similar procedure, using maximum likelihood estimators, and found the proposed procedure to perform similarly to that of Choi and Stablein (1982). It was noted by Choi and Stablein (1982) that given the additional computation, the maximum likelihood solution would not be a practical solution.

Tang and Tang (2004) proposed a test procedure which is a direct adaption of the best practical approach proposed by Choi and Stablein (1982). This adaption is found to be not Type I error robust in scenarios considered when $n_1 + n_2 + 2n_{12} = 20$. The test proposed by Choi and Stablein (1982) is found to be Type I error robust in this scenario. The literature reviewed suggests that a solution to the partially overlapping samples case will have to outperform the best practical solution by Choi and Stablein (1982). Tang and Tang (2004, p.81) concluded that, 'there may exist other test statistics which give better asymptotic or unconditional exact performance'.

In this paper, we introduce four test statistics for comparing the difference between two proportions with partially overlapping samples. These test statistics are formed so that no observations are discarded. The statistics represent the overall difference in proportions, divided by the combined standard error for the difference.

This paper will explore test statistics for testing H_0, in the presence of partially overlapping samples. In Section 2, existing 'standard' approaches and variants of are defined. In Section 3, our alternative proposals making use of all the available data are then introduced, followed by the most practical proposal of Choi and Stablein (1982).

In Section 4, a worked example applying all of the test statistics is given, followed by the simulation design in Section 5.

In Section 6.1, for all of the test statistics, the Type I error robustness is assessed when H_0 is true. This is measured using Bradley's (1978) liberal criteria. This criteria states that the Type I error rate should be between $\alpha_{\text{nominal}} \pm 0.5\,\alpha_{\text{nominal}}$.

There is no standard criteria for quantifying when a statistical test can be deemed powerful. The objective is to maximise the power of the test subject to preserving the Type I error rate α_{nominal}. If Type I error rates are not equal it is not possible to correctly compare the power of tests. The preferred test where Type I error rates are not equal should be the one with the Type I error rate closest to α_{nominal} (Penfield 1994). In Section 6.2, power will be considered under H_1 for the test statistics that meet Bradley's liberal criteria.

There is frequently too much focus on hypothesis testing. Confidence intervals may be of more practical interest (Gardner and Altman 1986). Confidence intervals allow insight into the estimation of a difference and the precision of the estimate. In Section 6.3, the coverage of the true difference under H_1 within 95% confidence intervals is considered. This is considered only for the most powerful test statistics that are Type I error robust.

2. Definition of standard test statistics

Assuming a dichotomous dependent variable, where a comparison in proportions between two samples is required, the layout of frequencies for the paired and the independent samples would be as per Table 1 and Table 2 respectively.

Table 1. Paired samples design for two samples and one dichotomous dependent variable

	Response Sample 2		
Response Sample 1	Yes	No	Total
Yes	*a*	*b*	*m*
No	*c*	*d*	n_{12} - *m*
Total	*k*	n_{12} -*k*	n_{12}

Table 2. Independent samples design for two samples and one dichotomous dependent variable

	Response		
	Yes	No	Total
Sample 1	*e*	*f*	n_1
Sample 2	*g*	*h*	n_2

2.1. Option 1: Discarding all paired observations

For two independent samples in terms of a dichotomous variable, as per Table 2, a Chi-square test of association is typically performed. This test will be displayed in standard textbooks in terms of χ_1^2. A chi square distribution on one degree of freedom is equivalent to the square of the z-distribution. Therefore under the null hypothesis an asymptotically N(0,1) equivalent statistic is defined as:

$$z_1 = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n_1} + \frac{\hat{p}(1-\hat{p})}{n_2}}} \text{ where } \hat{p}_1 = \frac{e}{n_1},\ \hat{p}_2 = \frac{g}{n_2} \text{ and } \hat{p} = \frac{e+g}{n_1+n_2}.$$

For small samples, Yates's correction is often performed to reduce the error in approximation. Yate's correction is given by:

$$z_2 = \sqrt{\frac{(n_1+n_2)((|eh-fg|-0.5(n_1+n_2))^2}{(e+g)(f+h)n_1n_2}}\ .$$

The statistic z_2 is referenced against the upper tail of the standard normal distribution.

An alternative to the Chi square approach is Fisher's exact test. This is computationally more difficult. Furthermore, Fisher's exact test is shown to deviate from Type I error robustness (Berkson, 1978). Fisher's exact test will not be considered for the analysis of the partially overlapping samples design in this paper.

2.2. Option 2: Discarding all unpaired observations

For two dependent samples in terms of a dichotomous variable, as per Table 1, McNemar's test is typically performed. Under the null hypothesis, the asymptotically N(0,1) equivalent to McNemar's test is:

$$z_3 = \frac{b-c}{\sqrt{b+c}}\ .$$

When the number of discordant pairs is small, a continuity correction is often performed. McNemar's test with continuity correction is the equivalent to:

$$z_4 = \sqrt{\frac{(|b-c|-1)^2}{b+c}}\ .$$

The statistic z_4 is referenced against the upper tail of the standard normal distribution.

Test statistics based on Option 1 and Option 2 are likely to have relatively low power for small samples when the number of discarded observations is large. A method of analysis for partially overlapping samples that takes into account the paired design but does not lose the unpaired information could therefore be beneficial.

2.3. Option 3: Applying an appropriate combination of the independent and paired tests using all of the available data

Given that test statistics for the paired samples and dependent samples can be calculated independently, an extension to these techniques which makes use of all of the available data would be some combination of the two tests.

In terms of power, Fisher's test and Tippett's test are comparable to a weighted approach using sample size as the weights (Samawi and Vogel, 2011). Tippett's method and Fisher's method are not as effective as Stouffer's weighted z-score test (Kim et al, 2013). Stouffer's weighted z-score, for combining z_1 and z_3 is defined as:

$$z_5 = \frac{wz_1+(1-w)z_3}{\sqrt{w^2+(1-w)^2}} \text{ where } w = \frac{n_1+n_2}{2n_{12}+n_1+n_2}\ .$$

Under the null hypothesis, the test statistic z_5 is asymptotically N(0,1).

Many other procedures for combining independent p-values are available, but these are less effective than Stouffer's test (Whitlock, 2005).

The drawbacks of Stouffer's test are that it has issues in the interpretation and confidence intervals for the true difference in population proportions cannot be easily formed.

3. Definition of alternative test statistics making use of all of the available data

The following proposals are designed to overcome the drawbacks identified of the standard tests. In these proposals observations are not discarded and the test statistics may be considered for the formation of confidence intervals.

3.1. Proposals using the phi correlation or the tetrachoric correlation coefficient

It is proposed that a test statistic for comparing the difference in two proportions with two partially overlapping samples can be formed so that the overall estimated difference in proportions is divided by its combined standard error, i.e.

$$\frac{\overline{\mathrm{p}}_1-\overline{\mathrm{p}}_2}{\sqrt{Var(\overline{\mathrm{p}}_1)+Var(\overline{\mathrm{p}}_2)-2r_xCov(\overline{\mathrm{p}}_1,\overline{\mathrm{p}}_2)}}$$

where $Var(\overline{\mathrm{p}}_1)=\dfrac{\overline{\mathrm{p}}_1(1-\overline{\mathrm{p}}_1)}{n_{12}+n_1}$, $Var(\overline{\mathrm{p}}_2)=\dfrac{\overline{\mathrm{p}}_2(1-\overline{\mathrm{p}}_2)}{n_{12}+n_2}$,

$$Cov(\overline{\mathrm{p}}_1,\overline{\mathrm{p}}_2)=\frac{\sqrt{\overline{\mathrm{p}}_1(1-\overline{\mathrm{p}}_1)}\sqrt{\overline{\mathrm{p}}_2(1-\overline{\mathrm{p}}_2)}n_{12}}{(n_{12}+n_1)(n_{12}+n_2)}$$

and r_x is a correlation coefficient.

Test statistics constructed in this manner will facilitate the construction of confidence intervals, for example a 95% confidence interval θ would be equivalent to:

$$\theta=(\overline{\mathrm{p}}_1-\overline{\mathrm{p}}_2)\pm1.96\times\sqrt{Var(\overline{\mathrm{p}}_1)+Var(\overline{\mathrm{p}}_2)-2r_xCov(\overline{\mathrm{p}}_1,\overline{\mathrm{p}}_2)}\ .$$

Pearson's phi correlation coefficient or Pearson's tetrachoric correlation coefficient are often used for measuring the correlation r_x between dichotomous variables.

Pearson's phi correlation coefficient is calculated as

$$r_1=\frac{ad-bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}\ .$$

The result of r_1 is numerically equivalent to Pearson's product-moment correlation coefficient and Spearman's rank correlation coefficient applied to Table 1, using binary outcomes '0' and '1' in the calculation. In this 2×2 case, r_1 is also numerically equivalent to Kendall's Tau-a and Kendall's Tau-b as well as Cramér's V and Somer's *d* (symmetrical). This suggests that r_1 would be an appropriate correlation coefficient to use.

Alternatively, assuming the underlying distribution is normal, a polychoric correlation coefficient may be considered. A special case of the polychoric correlation coefficient for two dichotomous samples is the tetrachoric correlation coefficient.

An approximation to the tetrachoric correlation coefficient as defined by Edwards and Edward (1984) is:

$$r_2=\frac{s-1}{s+1}\text{ where } s=\left(\frac{ad}{bc}\right)^{0.7854}\ .$$

Other approximations are available, however there is no conclusive evidence which is the most appropriate (Digby, 1983). In any event, r_1 is likely to be more practical than r_2 because if any of the observed paired frequencies are equal to zero then the calculation of r_2 is not possible.

Constructing a test statistic using correlation coefficients r_1 and r_2 respectively, the following test statistics are proposed:

$$z_6 = \frac{\overline{p}_1 - \overline{p}_2}{\sqrt{\frac{\overline{p}_1(1-\overline{p}_1)}{n_{12}+n_1} + \frac{\overline{p}_2(1-\overline{p}_2)}{n_{12}+n_2} - 2r_1\left(\frac{\sqrt{\overline{p}_1(1-\overline{p}_1)}\sqrt{\overline{p}_2(1-\overline{p}_2)}n_{12}}{(n_{12}+n_1)(n_{12}+n_2)}\right)}}$$

$$z_7 = \frac{\overline{p}_1 - \overline{p}_2}{\sqrt{\frac{\overline{p}_1(1-\overline{p}_1)}{n_{12}+n_1} + \frac{\overline{p}_2(1-\overline{p}_2)}{n_{12}+n_2} - 2r_2\left(\frac{\sqrt{\overline{p}_1(1-\overline{p}_1)}\sqrt{\overline{p}_2(1-\overline{p}_2)}n_{12}}{(n_{12}+n_1)(n_{12}+n_2)}\right)}}$$

where: $\overline{p}_1 = \frac{a+b+e}{n_{12}+n_1}$ and $\overline{p}_2 = \frac{a+c+g}{n_{12}+n_2}$.

Under H_0, $\pi_1 = \pi_2 = \pi$, therefore two additional test statistics that may be considered are defined as:

$$z_8 = \frac{\overline{p}_1 - \overline{p}_2}{\sqrt{\frac{\overline{p}(1-\overline{p})}{n_{12}+n_1} + \frac{\overline{p}(1-\overline{p})}{n_{12}+n_2} - 2r_1\left(\frac{\sqrt{\overline{p}(1-\overline{p})}\sqrt{\overline{p}(1-\overline{p})}n_{12}}{(n_{12}+n_1)(n_{12}+n_2)}\right)}}$$

$$z_9 = \frac{\overline{p}_1 - \overline{p}_2}{\sqrt{\frac{\overline{p}(1-\overline{p})}{n_{12}+n_1} + \frac{\overline{p}(1-\overline{p})}{n_{12}+n_2} - 2r_2\left(\frac{\sqrt{\overline{p}(1-\overline{p})}\sqrt{\overline{p}(1-\overline{p})}n_{12}}{(n_{12}+n_1)(n_{12}+n_2)}\right)}}$$

where $\overline{p} = \frac{(n_1+n_{12})\overline{p}_1 + (n_2+n_{12})\overline{p}_2}{2n_{12}+n_1+n_2}$.

The test statistics z_6, z_7, z_8 and z_9 are referenced against the standard normal distribution.

In the extreme scenario of $n_{12} = 0$, it is quickly verified that $z_8 = z_9 = z_1$. Under H_0 in the extreme scenario of $n_1 = n_2 = 0$, as $n_{12} \to \infty$ then $z_8 \to z_3$. This property is not observed for z_9. The properties of z_8 give support from a mathematical perspective as a valid test statistic to interpolate between the two established statistical tests where overlapping samples are not present.

3.2. Test statistic proposed by Choi and Stablein (1982)

Choi and Stablein (1982) proposed the following test statistic as the best practical solution for analysing partially overlapping sample:

$$z_{10} = \frac{\overline{\mathrm{p}}_1 - \overline{\mathrm{p}}_2}{\sqrt{\overline{\mathrm{p}}(1-\overline{\mathrm{p}})\left\{\frac{\psi_1^2}{n_1} + \frac{(1-\psi_1)^2}{n_{12}} + \frac{\psi_2^2}{n_2} + \frac{(1-\psi_2)^2}{n_{12}}\right\} - 2D}}$$

where $\psi_1 = \frac{n_1}{n_1 + n_{12}}$, $\psi_2 = \frac{n_2}{n_2 + n_{12}}$ and $D = \frac{(1-\psi_1)(1-\psi_2)(p_a - \overline{\mathrm{p}}^2)}{n_{12}}$.

The test statistic z_{10} is referenced against the standard normal distribution.

The authors additionally offer an extension of how optimization of w_1 and w_2 could be achieved, but suggest that the additional complication is unnecessary and the difference in results is negligible.

In common with the other statistics presented, z_{10} is computationally tractable but it may be less easy to interpret, particularly if $\psi_1 + \psi_2 \neq 1$.

4. Worked example

The objective of a Seasonal Affective Disorder (SAD) support group was to see if there is a difference in the quality of life for sufferers at two different times of the year. A binary response, 'Yes' or 'No' was required to the question whether they were satisfied with life. Membership of the group remains fairly stable, but there is some natural turnover of membership over time. Responses were obtained for $n_{12} = 15$ paired observations and a further $n_1 = 9$ and $n_2 = 6$ independent observations. The responses are given in Table 3.

Table 3. Responses to quality of life assessment.

	Response Time 2		
Response Time 1	Yes	No	Total
Yes	8	1	9
No	3	3	6
Total	11	4	15
	Response		
	Yes	No	Total
Time 1	5	4	9
Time 2	6	0	6

The elements of the test statistics (rounded to 3 decimal places for display purposes), are calculated as: $\hat{\mathrm{p}}_1 = 0.556$, $\hat{\mathrm{p}}_2 = 1.000$, $\hat{\mathrm{p}} = 0.733$, $\overline{\mathrm{p}}_1 = 0.583$, $\overline{\mathrm{p}}_2 = 0.810$, $\overline{\mathrm{p}} = 0.689$, $r_1 = 0.431$, $r_2 = 0.673$, $w = 0.333$, $\psi_1 = 0.375$, $\psi_2 = 0.286$, $D = 0.002$. The resulting test statistics are given in Table 4.

Table 4. Calculated value of test statistics (with corresponding p-values)

-score	1.907	.311	1.000	.500	1.747	2.023	2.295	1.937	2.202	1.809
-value	.057	.190	.317	.617	.081	.043	.022	.053	.028	.070

At the 5% significance level, whether H_0 is rejected depends on the test performed. It is of note that the significant differences arise only with tests introduced in this paper, z_6, z_7 and z_9.

Although the statistical conclusions differ for this particular example, the numeric difference between many of the tests is small. To consider further the situations where differences between the test statistics might arise, simulations are performed.

5. Simulation design

For the independent observations, a total of n_1 and n_2 unpaired standard normal deviates are generated. For the n_{12} paired observations, additional unpaired standard normal deviates X_{ij} are generated where i = (1,2) and j = (1,2,...., n_{12}). These are converted to correlated normal bivariates Y_{ij} so that:

$$Y_{1j} = \sqrt{\frac{1+\rho}{2}}X_{1j} + \sqrt{\frac{1-\rho}{2}}X_{2j} \text{ and } Y_{2j} = \sqrt{\frac{1+\rho}{2}}X_{2j} - \sqrt{\frac{1-\rho}{2}}X_{1j}$$

where ρ = correlation between population one and population two.

The normal deviates for both the unpaired and correlated paired observations are transformed into binary outcomes using critical values $C_{\pi i}$ of the normal distribution. If $X_{ij} < C_{\pi i}, Y_{ij} = 1$, otherwise $Y_{ij} = 0$

10,000 iterations of each scenario in Table 5 are performed in a 4×4×5×5×5×7=14000 factorial design.

Table 5. Values of parameters simulated for all test statistics.

Parameter	Values
π_1	0.15, 0.30, 0.45, 0.50
π_2	0.15, 0.30, 0.45, 0.50
n_1	10, 30, 50, 100, 500
n_2	10, 30, 50, 100, 500
n_{12}	10, 30, 50, 100, 500
ρ	-0.75, -0.50, -0.25, 0.00, 0.25, 0.50, 0.75

A range of values for n_1, n_2 and n_{12} likely to be encountered in practical applications are considered which offers an extension to the work done by Choi and Stablein (1982). Simulations are conducted over the range π from 0.15 to 0.5 both under H_0 and H_1. The values of π have been restricted to π <= 0.5 due to the proposed statistics being palindromic invariant with respect to π and $1-\pi$. Varying ρ is considered as it is known that ρ has an impact on paired samples tests. Negative ρ has been considered so as to provide a comprehensive overview and for theoretical interest, although ρ < 0 is less likely to occur in practical applications.

Two sided tests with $\alpha_{\text{nominal}} = 0.05$ is used in this study. For each combination of 10,000 iterations, the percentage of p-values below 0.05 is calculated to give the Type I error rate α . The Type I error rate under H_0 , for each combination considered in the simulation design, should be between 0.025 and 0.075 to meet Bradley's liberal criteria and to be Type I error robust.

All simulations are performed in R.

6. Simulation Results

A comprehensive set of results with varying independent and paired sample sizes, correlation, and proportions was obtained as outlined in Section 5.

6.1. Type I error rates

Under H_0, 10,000 replicates were obtained for 4×5×5×5×7=3500 scenarios. For assessment against Bradley's (1978) liberal criteria, Figure 1 shows the Type I error rates for all scenarios where $\pi_1 = \pi_2$ using $\alpha_{\text{nominal}} = 0.05$.

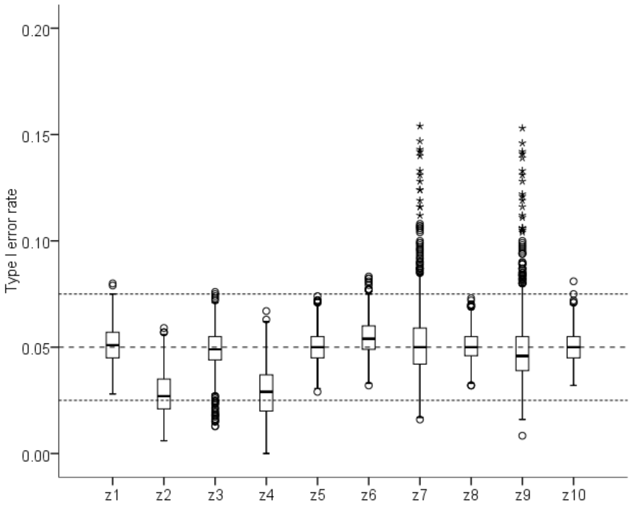


Figure 1: Type I error rates for each test statistic.

As may be anticipated, z_1 is Type I error robust because matched pairs are simply ignored. Similarly, z_3 performs as anticipated because the unpaired observations are ignored. Deviations from robustness for z_3 appear when n_{12} is small and ρ is large. Although deviations from stringent robustness are noted for z_3, this is not surprising since the cross product ratio is likely to be small when the proportion of success is low and the sample size is low. Crucially, the deviations from Type I error robustness of z_3 are conservative and will result in less false-positives, as such the tests statistic may not be considered unacceptable.

The corrected statistics, z_2 and z_4, generally give Type I error rates below the nominal alpha, particularly with small sample sizes. Ury and Fleiss (1980) found that z_1 is Type I error robust even with small samples, however applying Yate's correction is not Type I error robust and gives Type I error rates less than the nominal alpha. It is therefore concluded that z_2 and z_4 do not provide a Type I error robust solution.

The statistics using the phi correlation coefficient, z_6 and z_8, are generally liberal robust. For z_6 there is some deviation from the nominal Type I error rate. The deviations occur when min{n_1, n_2, n_{12}} is small, max{n_1, n_2, n_{12}} $-$ min{n_1, n_2, n_{12}} is large and $\rho < 0$. In these scenarios the effect of this is that z_6 is not liberal robust and results in a high likelihood of false-positives. It is therefore concluded that z_6 does not universally provide a Type I error robust solution to the partially overlapping samples situation.

The statistics using the tetrachoric correlation coefficient, z_7 and z_9, have more variability in Type I errors than the statistics that use the phi correlation coefficient. The statistics using the tetrachoric correlation coefficient inflate the Type I error when $\rho > 0.25$ and n_{12} is large. When min{n_1, n_2, n_{12}} is small the test statistic is conservative. A test statistic that performs consistently would be favoured for practical use. It is therefore concluded that z_7 and z_9 do not provide a Type I error robust solution to the partially overlapping samples situation.

Three statistics making use of all of the available data, z_5, z_8 and z_{10}, demonstrate liberal robustness across all scenarios. Analysis of Type I error rates show near identical boxplots to Figure 1 when each of the parameters are considered separately. This means these statistics are Type I error robust across all combinations of sample sizes and correlation considered.

6.2. Power

The test statistics z_2, z_4, z_6, z_7 and z_9 are not Type I error robust. Therefore only z_1, z_3, z_5, z_8 and z_{10} are considered for their power properties (where H_1 is true). Table 6 summarises the power properties where $\pi_1 = 0.5$.

Table 6. Power averaged over all sample sizes.

π_1	π_2	ρ	z_1	z_3	z_5	z_8	z_{10}
0.5	0.45	>0	0.095	0.173	0.208	0.221	0.221
		0		0.133	0.168	0.186	0.186
		<0		0.112	0.150	0.166	0.166
0.5	0.3	>0	0.509	0.653	0.807	0.856	0.855
		0		0.569	0.772	0.828	0.827
		<0		0.508	0.746	0.801	0.801
0.5	0.15	>0	0.843	0.874	0.975	0.989	0.989
		0		0.834	0.970	0.985	0.986
		<0		0.795	0.966	0.980	0.982

For each of the test statistics, as the correlation increases from -0.75 through to 0.75 the power of the tests increase. Similarly, as sample sizes increase the power of the test increases.

Clearly, z_5 is more powerful than the other standard tests z_1 and z_3, but it is not as powerful as the alternative methods that make use of all the available data.

The power of z_8 and z_{10} are comparable. Separate comparisons of z_8 and z_{10} indicates that the two statistics are comparable across the factorial combinations in the simulation design. Either test statistic could reasonably be used for hypothesis testing in the partially overlapping samples case.

6.3. Confidence interval coverage

For z_8 and z_{10}, the coverage of the true difference of population proportions within 95% confidence intervals has been calculated as per the simulation design in Table 5 where $\pi_1 \neq \pi_2$. The results are summarised in Figure 2.

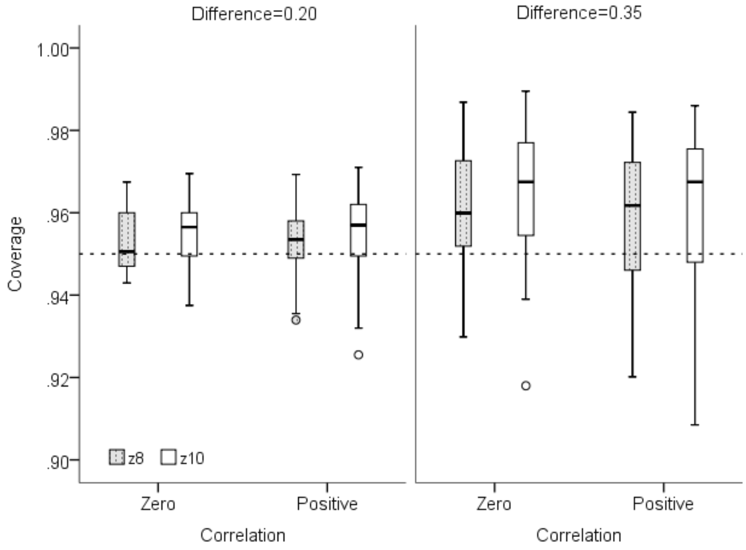


Figure 2: Percentage of iterations where the true difference is within the confidence interval.

Both z_8 and z_{10} demonstrate reasonable coverage of the true population difference $\pi_1 - \pi_2$. However, Figure 2 shows that z_8 more frequently performs closer to the desired 95% success rate. Taking this result into account, when the objective is to form a confidence interval, z_8 is recommended as the test statistic of choice in the partially overlapping samples case.

7. Conclusion

Partially overlapping samples may occur by accident or design. Standard approaches for analysing the difference in proportions for a dichotomous variable with partially overlapping samples often discard some available data. If there is a large paired sample or a large unpaired sample, it may be reasonable in a practical environment to use the corresponding standard test. For small samples, the test statistics which discard data have inferior power properties to tests statistics that make use of all the available data. These standard approaches and other ad-hoc approaches identified in this paper are less than desirable.

Combining the paired and independent samples z-scores using Stouffer's method is a more powerful standard approach, but leads to complications in interpretation, and does not readily extend to the creation of confidence intervals for differences in proportions. The tests introduced in this paper, as well as the test outlined by Choi and Stablein (1982) are more powerful than the test statistics in 'standard' use.

The alternative tests introduced in this paper, z_6, z_7, z_8 and z_9, overcome the interpretation barrier, in addition confidence intervals can readily be formed.

Tests introduced using the phi correlation coefficient, z_6 and z_8, are more robust than the equivalent tests introduced using the tetrachoric correlation coefficient, z_7 and z_9.

The most powerful tests that are Type I error robust are z_8 and z_{10}. The empirical evidence suggests that z_8 is better suited for forming confidence intervals for the true population difference than z_{10}. Additionally, z_8 has relative simplicity in calculation, strong mathematical properties and provides ease of interpretation. In conclusion, z_8 is recommended as the best practical solution to the partially overlapping samples framework when comparing two proportions.

References

1. Berkson J. **In dispraise of the exact test,** Journal of Statistic Planning and Inference. 2, 1978, pp. 27–42.
2. Bhoj D. **Testing equality of means of correlated variates with missing observations on both responses**, Biometrika. 1978; 65:225-228.
3. Bradley JV. **Robustness?** British Journal of Mathematical and Statistical Psychology. 31(2), 1978, pp.144-152.
4. Choi SC, Stablein DM. Practical tests for comparing two proportions with incomplete data. Applied Statistics. 1982; 31:256-262.
5. Digby PG. **Approximating the tetrachoric correlation coefficient**. Biometrics. 1983; pp. 753-757.
6. Edwards JH, Edwards AWF, **Approximating the tetrachoric correlation coefficient**. Biometrics. 40(2), 1984, 563.
7. Ekbohm G., **On testing the equality of proportions in the paired case with incomplete data.** Psychometrika. 47(1), 1982, pp. 115-118.
8. Gardner MJ, Altman DG, **Confidence intervals rather than p values: estimation rather than hypothesis testing**. BMJ.292(6522), 1986, pp. 746-750.

9. Kim SC, Lee SJ, Lee WJ, Yum YN, Kim JH, Sohn S, Park JH, Jeongmi L, Johan Lim, Kwon SW., **Stouffer's test in a large scale simultaneous hypothesis testing**. Plos one. 8(5):e63290, 2013

10. Martinez-Camblor P, Corral N, De la Hera JM., **Hypothesis test for paired samples in the presence of missing data**. Journal of Applied Statistics. 40(1), 2012, pp. 76-87.

11. Penfield DA. **Choosing a two-sample location test**. Journal of Experimental Education. 62(4), 1994, 343-360.

12. R Core Team, **R: A language and environment for statistical computing**. R Foundation for Statistical Computing, Vienna, Austria. www.R-project.org. 2014; version 3.1.2.

13. Samawi HM, Vogel R., **Tests of homogeneity for partially matched-pairs data**. Statistical Methodology. 8(3), 2011, pp. 304-313.

14. Tang ML, Tang NS., **Exact tests for comparing two paired proportions with incomplete data**. Biometrical Journal. 46(1), 2004, pp. 72-82.

15. Thomson PC., **A hybrid paired and unpaired analysis for the comparison of proportions**. Statistics in Medicine. 14, 1995, pp. 1463-1470.

16. Ury HK, Fleiss JL., **On approximate sample sizes for comparing two independent proportions with the use of Yates' correction**. Biometrics. 1980, pp. 347-351.

17. Whitlock MC., **Combining probability from independent tests: the weighted z-method is superior to Fisher's approach**. Journal of Evolutionary Biology. 18(5), 2005, pp. 1368-1373.

[1] Ben holds a first class honours degree in accounting and statistics and completed his masters in biometry with distinction at the University of Reading. He has previously worked in clinical research, business management information, training, and quality control. Ben is a full-time lecturer at the University of the West of England, Bristol with a teaching portfolio in statistical modelling, times series analysis, and in applications in the business sciences. He is undertaking doctoral work on the analysis of partially overlapping samples and is an active member of the consultancies arranged through the Applied Statistics group.

[2] Anselma has graduated from the University of the West of England, 2015, with first class honours in mathematics and statistics and was a recipient of the IMA (Institute of Mathematics and Its Applications) student prize. She optionally completed a full-year sandwich placement at the Office for National Statistics, where she worked collaboratively in methodological research most notably publishing papers in support of the Johnson Review of Consumer Price Indices

[3] Dr Deirdre Toher holds at PhD in applied statistics from Trinity College Dublin and is a senior lecturer in statistics at the University of the West of England. Deirdre is an active member of the Applied Statistics group working collaboratively in a wide range of diverse multi-disciplinary teams such as in disease profiling using high dimensional sensor data, providing methodological support to medical researchers, design and analysis of large scale national surveys, and the analysis of complex data such as experimental interventions in domestic water consumption. Deirdre is a Fellow of the Higher Education Academy, a fellow of the Royal Statistical Society (RSS), a Statistical Ambassador for the RSS, and trainer for the RSS Science Journalism Programme.

[4] Paul holds a PhD in pure and applied statistics. He is an associate professor in applied statistics and is the academic lead for the Applied Statistics Group at the University of the West of England. He has a wide range of undergraduate and postgraduate teaching across the University as well as providing PhD supervisory support in quantitative investigations. He has worked on over 100 projects in economics, psychology, health, and the bio- and medical sciences, as well as being a member of research ethics committees, data monitoring committees, and reviewer of grant applications.

Appendix P2

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

Published Version

2016 ■ Vol. 12 ■ no. 1

Why Welch's test is Type I error robust.

Ben Derrick[a], Deirdre Toher[a] & Paul White[a, ✉]

[a] University of the West of England, Bristol, England

Abstract ■ The comparison of two means is one of the most commonly applied statistical procedures in psychology. The independent samples t-test corrected for unequal variances is commonly known as Welch's test, and is widely considered to be a robust alternative to the independent samples t-test. The properties of Welch's test that make it Type I error robust are examined. The degrees of freedom used in Welch's test are a random variable, the distributions of which are examined using simulation. It is shown how the distribution for the degrees of freedom is dependent on the sample sizes and the variances of the samples. The impact of sample variances on the degrees of freedom, the resultant critical value and the test statistic is considered, and hence gives an insight into why Welch's test is Type I error robust under normality.

Keywords ■ Independent samples t-test; Welch's test; Welch's approximation; Behrens-Fisher problem; Equality of means.

✉ Paul.White@uwe.ac.uk

Introduction

One of the most commonly applied hypothesis test procedures in applied research is the comparison of two population means (Wilcox, 1992). For theoretical development purposes, assume two normally distributed populations $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$ are to be compared based upon n_1 and n_2 mutually independent observations. Let $\overline{X}_i$ and S_i^2 denote random variables for sample means and variances respectively $(i = 1,2)$.[1] If the population variances, σ_1^2 and σ_2^2, are assumed to be equal, then an appropriate test statistic is the independent samples t-test, based on (1) and (2).

$$T_1 = \frac{\overline{X}_1 - \overline{X}_2}{\text{StandardError}(\overline{X}_1 - \overline{X}_2)} \tag{1}$$

In the independent samples t-test, the standard error of $(\overline{X}_1 - \overline{X}_2)$, say SE_1, is given by:

$$SE_1 = S_p\sqrt{\frac{2}{\tilde{n}}} \tag{2}$$

where $S_p = \sqrt{\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{(n_1-1)+(n_2-1)}}$ and $\tilde{n}$ is the harmonic mean of n_1 and n_2. T_1 is referenced against the t-distribution with degrees of freedom equal to $\upsilon_1 = n_1 + n_2 - 2$.

It is known that, when the assumptions of the independent samples t-test are met, the independent samples t-test is an exact test and is the most uniformly powerful test (Sawilowsky & Blair, 1992). The independent samples t-test is an approximate test when population variances are unequal. If sample sizes are unequal and variances are unequal, the probability of rejecting the null hypothesis when it is true deviates from the nominal Type I error rate. This is particularly problematic when the smaller sample size is associated with the larger variance (Zimmerman & Zumbo, 2009; Coombs, Algina, & Oltman, 1996). This gives rise to the dilemma of how to compare means in the presence of unequal variances. This question, applied to two independent random samples from normal populations, is known as the Behrens-Fisher problem. Behrens (1929) and Fisher (1935, 1941) suggested a solution for the problem. It is proposed that the t-test when equal variances cannot be assumed is defined as per (3) and (4).

$$T_2 = \frac{\overline{\mathrm{X}}_1 - \overline{\mathrm{X}}_2}{\text{StandardError}(\overline{\mathrm{X}}_1 - \overline{\mathrm{X}}_2)} \tag{3}$$

In the unequal variances case, the standard error of $(\overline{X}_1 - \overline{X}_2)$, say SE_2 is estimated by:

$$SE_2 = \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \tag{4}$$

The formula developed for the degrees of freedom is complex, but it is proposed that an approximation for the degrees of freedom could be given by (5). This is given in most textbooks (e. g., Alfassi, Boger, & Ronen, 2005; Miles & Banyard, 2007).

$$\upsilon_2 = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\left(\frac{S_1^2}{n_1}\right)^2 / (n_1-1) + \left(\frac{S_2^2}{n_2}\right)^2 / (n_2-1)} \tag{5}$$

[1] As standard notation, random variables are shown in upper case, and derived sample values are shown are in lower case.

A numerically equivalent expression for the approximation υ_2 is given in (6). This is shown in some textbooks (e. g., Ott & Longnecker, 2001).

$$\upsilon_2 = \frac{(n_1-1)(n_2-1)}{(n_2-1)\,c^2+(n_1-1)\,(1-c)^2} \tag{6}$$

where

$$c = \frac{S_1^2/n_1}{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}$$

The approximation υ_2 dates back to a series of papers by Welch (1938, 1947, 1951), independent work by Satterthwaite (1946), works by Fairfield-Smith (1936), and Aspin (1948, 1949). The independent samples t-test corrected for unequal variances is sometimes referred to as the Satterthwaite-Smith-Welch test, the Welch-Aspin-Satterthwaite test, or other interchangeable variations. This may be referred to generically as the unequal variances t-test, or as the separate variances t-test. Usually the unequal variances t-test with the degrees of freedom approximated as above is simply known as Welch's test.

Originally, an alternative approximation for the degrees of freedom given by Welch, is given in (7):

$$\upsilon_3 = \frac{\left(\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}\right)^2}{\left(\frac{S_1^2}{n_1}\right)^2/\,(n_1+1)+\left(\frac{S_2^2}{n_2}\right)^2/\,(n_2+1)} - 2 \tag{7}$$

The approximation is given in some textbooks (e. g. Frank & Althoen, 1994), rounded down to the nearest integer. However, υ_3 is not generally used, and is not numerically equivalent to υ_2.

Textbooks frequently recommend the calculation of υ_2, rounded down to the nearest integer (e. g. Frank & Althoen, 1994; Ott & Longnecker, 2001). Rounding down tends to produce a conservative test. More generally, some textbooks recommend rounding to the nearest integer (e. g. Alfassi et al., 2005). The rounding requirements appear in textbooks for the purposes of manual calculations. There is a need to use integer degrees of freedom when using statistical tables for critical values. However, the calculation of Welch's test is easy in statistical software such as R and SPSS (Rasch, Kubinger, & Yanagida, 2011). These statistical software would ordinarily conduct the test with non-integer degrees of freedom.

Welch's test better approximates nominal significance levels, and has greater power than the Behrens-Fisher solution (Lee & Gurland, 1975; Best & Rayner, 1987). Fay and Proschan (2010, p. 14) confirm that Welch's solution “is approximately valid for the Behrens-Fisher perspective”.

When sample sizes are equal and variances are equal, both the independent samples t-test and Welch's test perform similarly (Zimmerman & Zumbo, 1993; Moser, Stevens, & Watts, 1989). For unequal sample sizes and unequal variances, Welch's test has superior Type I error robustness (Fagerland & Sandvik, 2009). Ruxton (2006) advocates the routine use of Welch's test.

Grimes and Federer (1982, p.10) state that, "In the case of comparing two sample means, the consensus in the literature seems to be the approval of Welch's approximate solution". Thus the most commonly used solution to the Behrens-Fisher problem, is Welch's test with the degrees of freedom calculated by approximation. In a practical environment, Welch's approximation can be used with little loss of accuracy (Wang, 1971; Scheffe, 1970).

It can be seen from (5) that Welch's degrees of freedom, υ_2, is a random variable and therefore has its own sampling distribution. Consequently the critical value used in hypothesis testing is also a random variable. In addition, it can be seen from (4) that the sample variances affect both the value of T_2 and the value of υ_2.

In this paper; worked examples of the independent samples t-test and Welch's test are provided. The distributions of the degrees of freedom for Welch's test are explored, and the two methods of estimating the standard error of are considered. Simulation is used to identify how the estimated standard error facilitates the Type I error robustness of Welch's test, and provides insight into why the Welch test works in a practical environment.

Worked examples

As part of an investigation into sensitivity when exposed to evidence of "White Privilege", Phillips and Lowery (2015) randomly allocated U.S. participants who self-identified as White/European-American into two groups. The participants completed a survey about equality and their childhood memories ("Experiment 1a"). Prior to completing the survey, Group 1 ($n_1 = 54$) were given a paragraph to read about "White Privilege", whereas Group 2 ($n_2 = 40$) were not. Questions on the survey measured participants perceived "life hardship" on a Likert type scale, between 1 = "strongly disagree" and 7 = "strongly agree". The authors performed the independent samples t-test using each participant's mean score.[2] This implies that equality of variance between groups is assumed; this is a seemingly reasonable assumption due to the random assignment of participants. For demonstration purposes, both the independent samples t-test and Welch's test are provided in the present paper. For "Experiment 1a", the published data are as follows; the average participant score for Group 1 is 4.41, (standard deviation of 1.20). The average participant

[2] The published results differ slightly from the calculations given here, due to the use of the published (rounded) sample data in the present paper.

score for Group 2 is 3.82 (standard deviation of 1.20). Thus, $\bar{x}_1 = 4.410$, $s_1^2 = 1.440$, $\bar{x}_2 = 3.820$ and $s_2^2 = 1.440$. Calculations for the independent samples t-test give: $s_p = 1.200$, $se_1 = 0.250$, $t_1 = 2.357$, $v_1 = 92.000$, the p-value using the independent samples t-test is 0.021. Calculations for Welch's test give: $se_2 = 0.250$, $t_2 = 2.357$, $v_2 = 84.186$, the p-value using Welch's test is 0.021. It can be seen that because the two sample variances are equal, $t_1 = t_2$. The degrees of freedom applicable for each test are different, but the impact of this on the critical values of the tests is small. Thus the p-values for both tests are the same to three decimal places. The statistical conclusion made at the 5% significance level, is that the sample mean for Group 1 is significantly greater than the sample mean for Group 2. The authors conclude that perceived "life hardship" is greater when participants are subjected to evidence of "White Privilege".

Phillips and Lowery (2015) replicated this experiment with $n_1 = 49$ and $n_2 = 42$ participants ("Experiment 1b"). The published data shows that the average participant score for Group 1 is 4.53, (standard deviation of 1.52). The average participant score for Group 2 is 3.96, (standard deviation of 1.28). Thus, $\bar{x}_1 = 4.530$, $s_1^2 = 2.310$, $\bar{x}_2 = 3.960$ and $s_2^2 = 1.638$. Calculations for the independent samples t-test give: $s_p = 1.415$, $se_1 = 0.297$, $t_1 = 1.916$, $v_1 = 89.000$, the p-value using the independent samples t-test is 0.059. Calculations for Welch's test give: $se_2 = 0.294$, $t_2 = 1.942$, $v_2 = 88.978$, the p-value using Welch's test is 0.055. In this experiment, the p-values for the two tests are different due to the unequal sample sizes and unequal variances of the two samples. With reference to Experiment 1b, the authors state that participants in Group 1 claim more "life hardship" than participants in Group 2. However, for either test, at the 5% significance level, Experiment 1b alone represents insufficient statistical evidence that there is a difference between Group 1 and Group 2.

Methodology

Simulation is used to investigate Welch's test for Type I error robustness, and the distributional properties of v_2. For both the independent samples t-test and Welch's test, two sided tests are performed with nominal Type I error rate of $\alpha = 0.05$. The aim is to demonstrate deviations from Type I error robustness for the independent samples t-test for unequal variances. The standard error of the independent samples t-test and Welch's test are explored to assess the impact of the standard error on the result of the tests. To achieve these goals, simulations under H_0 for two normally distributed samples are performed as per the layout in Table 1; with n_1 at two levels, n_2 at two levels and σ_2 at two levels. Parameters are selected to cover both "large" and "small" samples and equal and unequal variances. The sample sizes represent extreme scenarios in order to assist in the illustration of the effects.

For each scenario in the simulation design, 10,000 iterations are performed under the condition where H_0 is true.

Results

Welch's degrees of freedom.

The investigation of the distribution of v_2, gives insight into when the degrees of freedom used in Welch's test differ from the degrees of freedom used in the independent samples t-test.

Figure 1 shows the distribution of the degrees of freedom for each of the 8 scenarios simulated (10,000 observations per scenario).

Inspection of Figure 1 shows the greatest discrepancy between v_1 and v_2 to occur when $n_1 \neq n_2$. The simulations demonstrate that $[min\{n_1, n_2\} - 1] \leq v_2 \leq v_1$. This can be proven mathematically using (6). By differentiation, the maximum value of v_2 is found when $s_1^2/s_2^2 = \{(n_1 - 1)n_1\}/\{(n_2 - 1)n_2\}$. The minimum value of v_2 is fixed by the sample with the larger variance. If Sample 1 has the larger variance, then the lower bound is $n_1 - 1$. If Sample 2 has the larger variance, then the lower bound is $n_2 - 1$. Hence, $min\{n_1, n_2\} - 1$ is a very conservative approximation to the degrees of freedom when the smaller sample size is associated with the larger variance. To illustrate these points, see Figure 2 with a fixed variance for Sample 1.

From Figure 2 it can be seen that as s_2^2/s_1^2 tends to zero, the degrees of freedom tends to $n_1 - 1$. As s_2^2/s_1^2 becomes increasingly large, the degrees of freedom asymptotically tends to $n_2 - 1$. The maximum value occurs when $s_1^2/s_2^2 = \{(n_1 - 1)n_1\}/\{(n_2 - 1)n_2\}$. The examples have a total sample size of 30, thus the maximum value of v_2 is 28.

Type I error robustness for the independent samples t-test and Welch's test.

In this section, p-values calculated from performing both the independent samples t-test and Welch's test are considered, as per the simulation design in Table 1. If H_0 is true and if underlying assumptions hold, then the p-values from a valid test procedure are expected to be uniformly distributed (Bland, 2013). Deviations from uniformity give evidence that the test is not Type I error robust. If p-values are consistently less than expected under a uniform distribution, the test gives too many false positives, and is said to be "liberal". If p-values are consistently greater than expected under a uniform distribution, the test is "conservative".

There is negligible difference between the p-values when performing the independent samples t-test or Welch's test under equal variances, regardless of sample size. In this case, p-values are approximately uniformly dis-

Table 1 ■ Summary of the simulation design.

Test statistics	T_1, T_2
Degrees of freedom	υ_1, υ_2
Sample sizes (n_1, n_2)	(5,5), (5,100), (100,5), (100,100)
Standard deviations (σ_1, σ_2)	(1,1), (1,2)
Programming language	R version 3.1.2 (R Development Core Team, 2013)

tributed for both tests (results not shown).

When variances are unequal, Welch's test is not a linear function of the independent samples t-test. Figure 3 is a P-P plot (percentile-percentile plot), for p-values for both the independent samples t-test (T_1) and Welch's test (T_2), with unequal variances. This shows ordered expected p-values from a uniform distribution plotted against ordered observed p-values. Given that for a valid test procedure, observed p-values should be approximately uniformly distributed on (0, 1) then an approximate diagonal would demonstrate Type I error robustness.

Both panels of Figure 3 show that when sample sizes are unequal and variances are unequal, the independent samples t-test is not Type I error robust. When the smaller sample size is associated with the larger variance (left panel, Figure 3), the observed p-values under the independent samples t-test are smaller than expected, and the test is liberal. Conversely, when the larger sample size is associated with the larger variance (right panel, Figure 3), the p-values are larger than expected and the independent samples t-test is conservative, (i.e. the expected Type I error rate is less than the pre-chosen nominal level of significance, α).

The p-values for Welch's test are also given in Figure 3. The simulated p-values for Welch's test, are approximately uniformly distributed. This results in the approximate line of equality observed. Welch's test therefore "corrects" for the fact that the independent samples t-test gives p-values that are not Type I error robust.

To demonstrate the impact of the degrees of freedom, for insight only, the independent samples t-test T_1 but with υ_2 degrees of freedom is considered. Likewise, for insight only, Welch's test using statistic T_2 but with υ_1 degrees of freedom is considered. These are compared against the standard approaches for the independent samples t-test and Welch's test. Table 2 summarises the Type I error rates observed ($\alpha = .05$, two-sided) for each combination. Bradley's (1978) liberal robustness criteria states that the Type I error rate when the nominal α is .05 should be in the range {0.025, 0.075}.

Table 2 shows that Welch's test (test statistic and degrees of freedom) is Type I error robust across all scenarios simulated. For unequal sample sizes and unequal variances, T_1 used in conjunction with υ_1 or υ_2, and T_2 used in conjunction with υ_1, do not meet liberal robustness criteria. Welch's degrees of freedom therefore represent an important property for controlling Type I error rates. However, clearly the calculation of the test statistic, which takes into account the two separate sample variances, is also important.

Impact of the standard error on the properties of Welch's test.

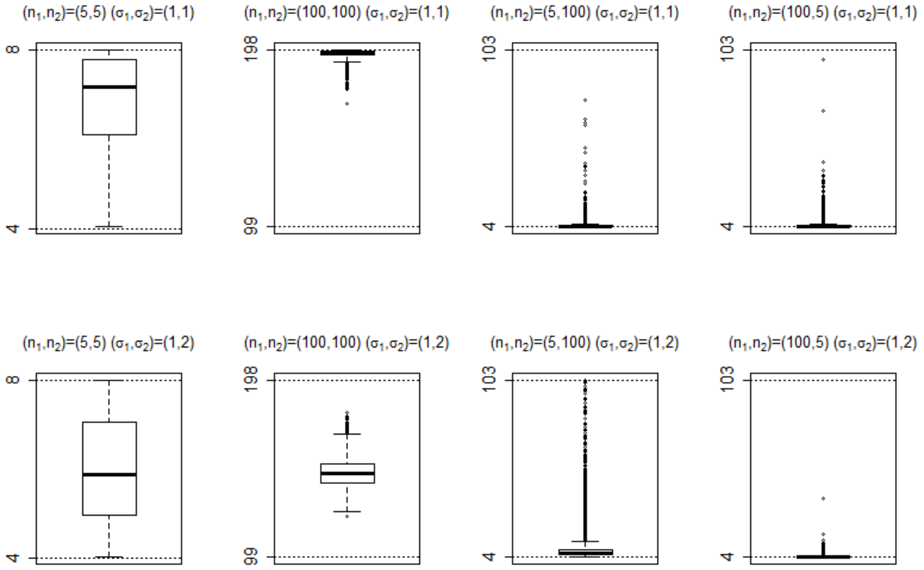
In this section, the impact of the standard error of the test statistics for the independent samples t-test and Welch's test is considered. The corrective properties of Welch's test are, in part, due to the impact of the sample variances on the degrees of freedom, which in turn affects the critical value used in the test. However, Type I error robustness could also be due to the impact of the estimated standard error on the magnitude of the test statistic. Figure 4 and Figure 5 demonstrate how the standard error, SE_1 and SE_2, relate to the critical value and to the absolute values of the test statistic for the independent samples t-test, T_1, and Welch's test, T_2, respectively.

Both panels of Figure 4 suggest that, when performing the independent samples t-test, the estimated standard error, SE_1, has no apparent relationship with the value of the test statistic, T_1. When the smaller sample size is associated with the larger population variance (left panel, Figure 4), the absolute value of the test statistic has a larger mean and a larger variability. When the larger sample size is associated with the larger population variance (right panel, Figure 4), the absolute value of the test statistic has a smaller mean and a smaller variability. This has the result that more false positives are observed when the smaller sample size is associated with the larger variance.

Both panels of Figure 5 demonstrate the impact of the degrees of freedom on the critical value. In the simulated scenario; the theoretical minimum degrees of freedom is $min(n_1, n_2) = 4$, accordingly the upper bound of the critical value is 2.776; the theoretical maximum degrees of freedom is $\upsilon_1 = 98$, accordingly the lower bound of the critical value is 1.984.

It can be seen from both panels of Figure 5 that as Welch's estimate of standard error, SE_2, increases, the absolute value of T_2 decreases. As the estimated standard error becomes large, the impact is far greater on the absolute value of T_2 relative to the critical value. This combination results in fewer false positives being observed as the esti-

Figure 1 ■ Distribution of υ_2 for each scenario. The references lines represent the theoretical maximum and minimum values that υ_2 can take. The upper reference line is equivalent to υ_1.



mated standard error increases.

Discussion

For additional clarity of the above findings, Table 3 summarises theoretical values for each of the combinations in the simulation design. For illustration purposes differences in means are fixed at 1.000, s_1 and s_2 are fixed as σ_1 and σ_2 respectively.

From Table 3, it can be seen that when sample sizes are equal or variances are equal, the test statistics for the independent samples t-test and Welch's test are equivalent. Therefore, the difference in p-values are a direct result of the degrees of freedom used to calculate the critical value.

When variances are not equal, Welch's estimated standard error impacts the critical value, but this effect is smaller than the effect on the value on the test statistic. When the smaller sample size is associated with the larger variance, the effect on the value of the test statistic is exacerbated.

Conclusion

The literature favours Welch's test for a comparison of two means. This paper adds further support to the findings in the literature with respect to the Type I error robustness of Welch's test. The degrees of freedom of Welch's test are a random variable based on the sample size and variance of each sample. The degrees of freedom used in Welch's test are always less than or equal to the degrees of freedom used in the independent samples t-test. The degrees of freedom used in the independent samples t-test and Welch's test are equivalent when $s_1^2/s_2^2 = \{(n_1-1)n_1\}/\{(n_2-1)n_2\}$. The minimum value of Welch's degrees of freedom is $min\{n_1, n_2\} - 1$, this minimum is determined by the sample with the larger variance. Therefore Welch's approximate degrees of freedom are more conservative than the degrees of freedom used in the independent samples t-test, particularly when

Figure 2 ■ Value of v_2 with varying s_2^2, and fixed value $s_1^2 = 1$. Values to the left of $s_2^2 = 1$ have the larger variance associated with Sample 1. Values to the right of $s_2^2 = 1$ have the larger variance associated with Sample 2.

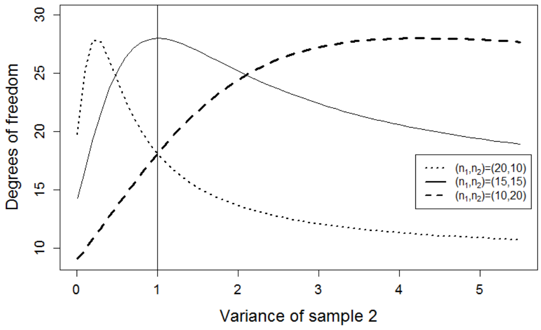


Table 2 ■ Type I error rates for each combination of test statistic with degrees of freedom. Type I error robust combinations are highlighted in bold.

(n_1, n_2)	(σ_1, σ_2)	T_1 with v_1	T_1 with v_2	T_2 with v_1	T_2 with v_2
5,5	1,1	**0.050**	**0.045**	**0.050**	**0.045**
	1,2	**0.056**	**0.047**	**0.056**	**0.047**
5,100	1,1	**0.053**	0.012	0.110	**0.056**
	1,2	0.001	0.000	0.093	**0.060**
100,5	1,1	**0.050**	0.011	0.108	**0.055**
	1,2	0.295	0.153	0.118	**0.052**
100,100	1,1	**0.049**	**0.049**	**0.049**	**0.049**
	1,2	**0.050**	**0.049**	**0.050**	**0.049**

the smaller sample size is associated with the larger variance. When performing Welch's test, the estimated standard error impacts the magnitude of the test statistic. Under the null hypothesis, it is the estimated standard error when performing Welch's test, which is the most influential factor on the result of the test. For Welch's test, the probability of making a Type I error decreases as the standard error increases. This paper gives insight in to why Welch's test is Type I error robust for normally distributed data, in scenarios when the independent samples t-test is not. Additionally, it is shown that in situations when the independent samples t-test is Type I error robust, Welch's test is also. In a practical environment for the comparisons of two means from assumed normal populations, a general rule to preserve Type I error robustness is, if in doubt use Welch's test.

References

Alfassi, Z. B., Boger, Z., & Ronen, Y. (2005). *Statistical treatment of analytical data*. CRC Press.

Aspin, A. A. (1948). An examination and further development of a formula arising in the problem of comparing two mean values. *Biometrika. 2*, 88–96. doi:10.2307/2332631

Aspin, A. A. (1949). Tables for use in comparisons whose accuracy involves two variances, separately estimated. *Biometrika. 36*, 290–296. doi:10.2307/2332668

Behrens, W. U. (1929). Ein beitrag zur fehlerberechnung bei wenigen beobachtungen. *Landwirtschaftliche Jahrbucher. 68*, 807–837.

Best, D. J. & Rayner, J. C. W. (1987). Welch's approximate solution for the behrens-fisher problem. *Technometrics. 29*(2), 205–210. doi:10.2307/1269775

Figure 3 ■ P-values for the independent samples t-test, T_1, and Welch's test, T_2. The left panel shows the smaller sample size associated with the larger variance. The right panel shows the larger sample size associated with the larger variance.

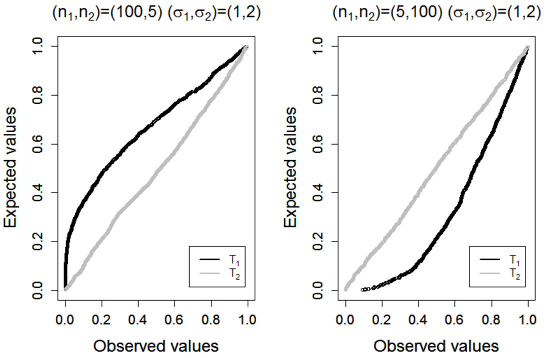


Table 3 ■ Components of the tests for each scenario in the simulation design.

		Independent samples t-test			Welch's test		
(n_1, n_2)	(s_1, s_2)	test statistic	critical value	p-value	test statistic	critical value	p-value
5,5	1,1	1.581	2.306	0.153	1.581	2.306	0.153
	1,2	1.000	2.306	0.347	1.000	2.571	0.363
5,100	1,1	2.182	1.983	0.031	2.182	2.776	0.095
	1,2	1.107	1.983	0.271	2.041	2.571	0.097
100,5	1,1	2.182	1.983	0.031	2.182	2.776	0.095
	1,2	2.065	1.983	0.041	1.111	2.776	0.329
100,100	1,1	7.071	1.972	<0.001	7.071	1.972	< 0.001
	1,2	4.472	1.972	< 0.001	4.472	1.976	< 0.001

Bland, M. (2013). Do baseline p-values follow a uniform distribution in randomised trials? *PloS one*, *8*(10), e76010. doi:10.1371/journal.pone.0076010

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology. 31*(2), 144–152. doi:10.1111/j.2044-8317.1978.tb00581.x

Coombs, W. T., Algina, J., & Oltman, D. (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), 137–79. doi:10.3102/00346543066002137

Fagerland, M. W. & Sandvik, L. (2009). Performance of five two-sample location tests for skewed distributions with unequal variances. *Contemporary Clinical Trials. 30*(5), 490–496. doi:10.1016/j.cct.2009.06.007

Fairfield-Smith, H. (1936). The problem of comparing the results of two experiments with unequal errors. *Journal of the Council for Scientific and Industrial Research*, *9*, 211–212.

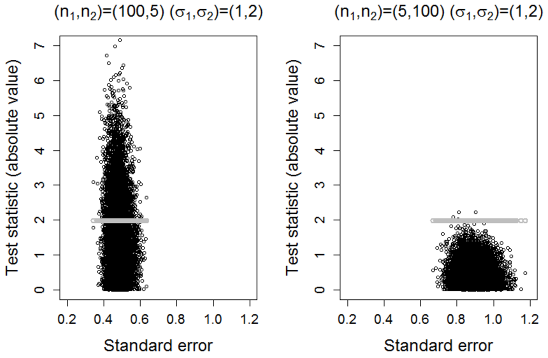
Fay, M. P. & Proschan, M. A. (2010). Wilcoxon-mann-whitney or t-test? *On assumptions for hypothesis tests and multiple interpretations of decision rules. Statistics Surveys. 4*(1), 1. doi:10.1214/09-SS051

Fisher, R. A. (1935). The fiducial argument in statistical inference. *Annals of Eugenics*. 391–398. doi:10.1111/j.1469-1809.1935.tb02120.x

Fisher, R. A. (1941). The asymptotic approach to behrens' integral, with further tables for the d test of significance. *Annals of Eugenics. 11*, 141–172. doi:10.1111/j.1469-1809.1941.tb02281.x

Frank, H. & Althoen, S. C. (1994). *Statistics: concepts and applications.* Cambridge University Press.

Figure 4 ■ Simulated values of the standard error, SE_1, against the absolute value of the test statistic, T_1, for the independent samples t-test. The critical value, a constant at 1.984, has been superimposed. The left panel shows the smaller sample size associated with the larger variance. The right panel shows the larger sample size associated with the larger variance.



Grimes, B. A. & Federer, W. T. (1982). *Comparison of means from populations with unequal variances.* Biometrics Unit Technical Reports: Number BU-762-M.

Lee, A. F. S. & Gurland, J. (1975). Size and power of tests for equality of means of two normal populations with unequal variances. *Journal of the American Statistical Association. 70*(352), 933–941. doi:10.1080/01621459.1975.10480326

Miles, J. & Banyard, P. (2007). *Understanding and using statistics in psychology: a practical introduction.* Sage.

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

Ott, R. L. & Longnecker, M. (2001). *An introduction to statistical methods and data analysis.* Pacific Grove, CA: Duxbury.

Phillips, L. T. & Lowery, B. S. (2015). The hard-knock life? whites claim hardships in response to racial inequity. *Journal of Experimental Social Psychology*, *61*, 12–18. doi:10.1016/j.jesp.2015.06.008

R Development Core Team. (2013). *R: a language and environment for statistical computing.* ISBN 3-900051-07-0. R Foundation for Statistical Computing. Vienna, Austria. Retrieved from http://www.R-project.org

Rasch, D., Kubinger, K., & Yanagida, T. (2011). *Statistics in psychology using r and spss.* John Wiley and Sons.

Ruxton, G. (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

Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin. 2*, 110–114. doi:10.2307/3002019

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. *American Psychological Association. 111*(2), 352–360. doi:10.1037/0033-2909.111.2.352

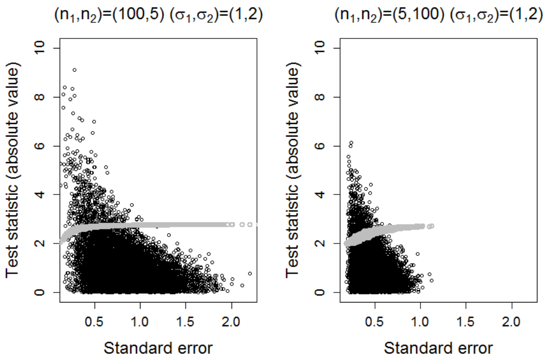
Scheffe, H. (1970). Practical solutions of the behrens-fisher problem. *Journal of the American Statistical Association. 65*, 1501–1508. doi:10.1080/01621459.1970.10481179

Wang, Y. Y. (1971). Probabilities or the type l errors of the welch tests for the behrens-fisher problem. *Journal of the American Statistical Association. 66*, 605–608. doi:10.1080/01621459.1971.10482315

Welch, B. L. (1938). The significance or the difference between two means when the population variances are unequal. *Biometrika. 29*, 350–362. doi:10.2307/2332010

Welch, B. L. (1947). The generalization of 'student's' problem when several different population variances are involved. *Biometrika. 34*, 28–35. doi:10.2307/2332510

Figure 5 ■ Properties of Welch's test. The critical values have been superimposed. The left panel shows the smaller sample size associated with the larger variance. The right panel shows the larger sample size associated with the larger variance.



Welch, B. L. (1951). On the comparison of several mean values: an alternative approach. *Biometrika. 38*, 330–336. doi:10.2307/2332579

Wilcox, R. R. (1992). Why can methods for comparing means have relatively low power, and what can you do to correct the problem? *Current Directions in Psychological Science. 1*(3), 101–105.

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. *Canadian Journal of Experimental Psychology. 47*(3), 523–39. doi:10.1037/h0078850

Zimmerman, D. W. & Zumbo, B. D. (2009). Hazards in choosing between pooled and separate-variances t-tests. *Psicológica: Revista de metodología y psicología experimental, 30*(2), 371–390.

Citation

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

Received: 28/08/2015 ~ Accepted: 12/10/2015

Appendix P3

Derrick, B. et al. (2017a). “Test statistics for the comparison of means for two samples which include both paired observations and independent observations.” *Journal of Modern Applied Statistical Methods* 16 (1), pp. 137–157

Accepted Version

Test Statistics for the Comparison of Means for Two Samples Which Include Both Paired Observations and Independent Observations.

Introduction

Hypothesis tests for the comparison of two population means, μ_1 and μ_2, with two samples of either independent observations or paired observations are well established. When the assumptions of the test are met, the independent samples t-test is the most powerful test for comparing means between two independent samples (Sawilowsky & Blair, 1992). Similarly, when the assumptions of the test are met, the paired samples t-test is the most powerful test for the comparison of means between two dependent samples (Zimmerman, 1997). If a paired design can avoid extraneous systematic bias, then paired designs are generally considered to be advantageous when contrasted with independent designs.

There are scenarios where, in a paired design, some observations may be missing. In the literature, this scenario is referred to as paired samples that are either “incomplete” (Ekbohm, 1976) or with “missing observations” (Bhoj, 1978). There are designs that do not have completely balanced pairings. Occasions where there may be two samples with both paired observations and independent observations include:

i) Two groups with some common element between both groups. For example, in education when comparing the average exam marks for two optional subjects, where some students take one of the two subjects and some students take both.

ii) Observations taken at two points in time, where the population membership changes over time but retains some common members. For example, an annual survey of employee satisfaction may include new employees that were unable to respond at time point one, employees that left after time point one, and employees that remained in employment throughout.

iii) When some natural pairing occurs. For example, in a survey taken comparing views of males and females, there will be some matched pairs “couples” and some independent samples “single”.

The examples given above can be seen as part of the wider missing data framework. There is much literature on methods for dealing with missing data and the proposals in this paper do not detract from extensive research into the area. The simulations and discussion in this paper are done in the context of data missing completely at random (MCAR).

Two samples which include both paired and independent observations is referred to using varied terminology in the literature. The example scenarios outlined can be referred to as “partially paired data” (Samawi & Vogel, 2011). However, this terminology has connotations suggesting that the pairs themselves are not directly matched. Derrick et.al. (2015) suggest that appropriate terminology for the scenarios outlined gives reference to “partially overlapping samples”. For work that has previously been done on a comparison of means when partially overlapping samples are present, “the partially overlapping samples framework….has been treated poorly in the literature” (Martínez-Camblor, Corral, & María de la Hera, 2012, p.77). In this paper, the term partially overlapping samples will be used to refer to scenarios where there are two samples with both paired and independent observations.

When partially overlapping samples exist, the goal remains to test the null hypothesis $H_0 : \mu_1 = \mu_2$. Standard approaches when faced with such a situation, are to perform the paired samples t-test, discarding the unpaired data, or alternatively perform the independent samples t-test, discarding the paired data (Looney & Jones, 2003). These approaches are wasteful and can result in a loss of power. The bias created with these approaches may be of concern. Other solutions proposed in a similar context are to perform the independent samples t-test on all observations ignoring the fact that there may be some pairs, or alternatively randomly pairing unpaired observations and performing the paired samples t-test (Bedeian & Feild, 2002). These methods distort Type I error rates (Zumbo, 2002) and fail to adequately reflect the design. This emphasises the need for research into a statistically valid approach. A method of analysis which takes into account any pairing but does not lose the unpaired information would be beneficial.

One analytical approach is to separately perform both the paired samples t-test on the paired observations and the independent samples t-test on the independent observations. The results are then combined using Fisher's (1925) Chi-square method, or Stouffer's (1949) weighted z-test. These methods have issues with respect to the interpretation of the results. Other procedures weighting the paired and independent samples t-tests, for the partially overlapping samples scenario, have been proposed by Bhoj, (1978), Kim et. al. (2005), Martínez-Camblor, Corral, & María de la Hera (2012), and Samawi & Vogel (2011).

Looney & Jones (2003) proposed a statistic making reference to the z-distribution that uses all of the available data, without a complex weighting structure. Their corrected z-statistic is simple to compute and it directly tests the hypothesis $H_0 : \mu_1 = \mu_2$. They suggest that their test statistic is generally Type I error robust across the scenarios that they simulated. However, they only consider normally distributed data with a common variance of 1 and a total sample size of 50 observations. Therefore their simulation results are relatively limited, simulations across a wider range of parameters would help provide stronger conclusions. Mehrotra (2004) indicates that the solution provided by Looney & Jones (2003) may not be Type I error robust for small sample sizes.

Early literature for the partially overlapping samples framework focused on maximum likelihood estimates, when data are missing by accident rather than by design. Lin (1973) use maximum likelihood estimates for the specific case where data is missing from one of the two groups. Lin (1973) uses assumptions such as the variance ratio is known. Lin & Strivers (1974) apply maximum likelihood solutions to the more general case, but find that no single solution is applicable.

For normally distributed data, Ekbohm (1976) compared Lin & Strivers (1974) tests with similar proposals based on maximum likelihood estimators. Ekbohm (1976) found that maximum likelihood solutions do not always maintain Bradley's liberal Type I error robustness criteria. The results suggest that the maximum likelihood approaches are of little added value compared to standard methods. Furthermore the proposals by Ekbohm (1976) are complex mathematical procedures and are unlikely to be considered as a first choice solution in a practical environment.

A solution available in most standard software is to perform a mixed model using all of the available data. In a mixed model, effects are assessed using Restricted Maximum Likelihood estimators

“REML”. Mehrotra (2004) indicates that for positive correlation, REML is Type I error robust and more powerful approach than that proposed by Looney & Jones (2003).

For small sample sizes, an intuitive solution to the comparison of means with partially overlapping samples, would be a test statistic derived using concepts similar to that of Zumbo (2002) so that all available data are used making reference to the t-distribution.

In this paper, two test statistics are proposed. The proposed solution for equal variances acts as a linear interpolation between the paired samples t-test and the independent samples t-test. The consensus in the literature is that Welch’s test is more Type I error robust than the independent samples t-test, particularly with unequal variances and unequal samples sizes (Derrick, Toher & White, 2016; Fay & Proschan, 2010; Zimmerman & Zumbo, 2009). The proposed solution for unequal variances is a test which acts as a linear interpolation between the paired samples t-test and Welch’s test.

Standard tests and the proposal by Looney & Jones (2003) are given below. This is followed by the definition of the presently proposed test statistics. A worked example provided using each of these test statistics and REML is provided. The Type I error rate and power for the test statistics and REML is then explored using simulation, for partially overlapping samples simulated from a Normal distribution.

Notation

Notation used in the definition of the test statistics is given in Table 1.

Table 1. Notation used in this paper.

$n_a =$	number of observations exclusive to Sample 1
$n_b =$	number of observations exclusive to Sample 2
$n_c =$	number of pairs
$n_1 =$	total number of observations in Sample 1 (i.e. $n_1 = n_a + n_c$)
$n_2 =$	total number of observations in Sample 2 (i.e. $n_2 = n_b + n_c$)
$\bar{X}_1 =$	mean of all observations in Sample 1
$\bar{X}_2 =$	mean of all observations in Sample 2
$\bar{X}_{\mathrm{a}} =$	mean of the independent observations in Sample 1
$\bar{X}_{\mathrm{b}} =$	mean of the independent observations in Sample 2
$\bar{X}_{1\mathrm{c}} =$	mean of the paired observations in Sample 1
$\bar{X}_{2\mathrm{c}} =$	mean of the paired observations in Sample 2
$S_1^2 =$	variance of all observations in Sample 1
$S_2^2 =$	variance of all observations in Sample 2
$S_a^2 =$	variance of the independent observations in Sample 1
$S_b^2 =$	variance of the independent observations in Sample 2
$S_{1c}^2 =$	variance of the paired observations in Sample 1
$S_{2c}^2 =$	variance of the paired observations in Sample 2
$S_{12} =$	covariance between the paired observations
$r =$	Pearson's correlation coefficient for the paired observations

All variances above are calculated using Bessel's correction, i.e. the sample variance with $n_i - 1$ degrees of freedom (see Kenney & Keeping 1951, p.161).

As standard notation, random variables are shown in upper case, and derived sample values are shown are in lower case.

Definition of Existing Test Statistics

Standard approaches for comparing two means making reference to the t-distribution are given below. These definitions follow the structural form given by Fradette et.al. (2003), adapted to the context of partially overlapping samples.

To perform the paired samples t-test, the independent observations are discarded so that

$$T_1 = \frac{\bar{X}_{1c} - \bar{X}_{2c}}{\sqrt{\frac{S_{1c}^2}{n_c} + \frac{S_{2c}^2}{n_c} - 2r\left(\frac{S_{1c}S_{2c}}{n_c}\right)}}$$

The statistic T_1 is referenced against the t-distribution with $\nu_1 = n_c - 1$ degrees of freedom.

To perform the independent samples t-test, the paired observations are discarded so that

$$T_2 = \frac{\bar{X}_a - \bar{X}_b}{S_p\sqrt{\frac{1}{n_a} + \frac{1}{n_b}}} \text{ where } S_p = \sqrt{\frac{(n_a - 1)S_a^2 + (n_b - 1)S_b^2}{(n_a - 1) + (n_b - 1)}}$$

The statistic T_2 is referenced against the t-distribution with $\nu_2 = n_a + n_b - 2$ degrees of freedom.

To perform Welch's test, the paired observations are discarded so that

$$T_3 = \frac{\bar{X}_a - \bar{X}_b}{\sqrt{\frac{S_a^2}{n_a} + \frac{S_b^2}{n_b}}}$$

The statistic T_3 is referenced against the t-distribution with degrees of freedom approximated by

$$\nu_3 = \frac{\left(\frac{S_a^2}{n_a} + \frac{S_b^2}{n_b}\right)^2}{\left(\frac{S_a^2}{n_a}\right)^2 / (n_a - 1) + \left(\frac{S_b^2}{n_b}\right)^2 / (n_b - 1)}$$

For large sample sizes, the test statistic for partially overlapping samples proposed by Looney & Jones (2003) is

$$Z_{\text{corrected}} = \frac{\overline{\mathrm{X}}_1 - \overline{\mathrm{X}}_2}{\sqrt{\dfrac{S_1^2}{n_a + n_c} + \dfrac{S_2^2}{n_b + n_c} - \dfrac{(2n_c)S_{12}}{(n_a + n_c)(n_b + n_c)}}}$$

The statistic $Z_{\text{corrected}}$ is referenced against the standard Normal distribution. In the extremes of $n_a = n_b = 0$, or $n_c = 0$, $Z_{\text{corrected}}$ defaults to the paired samples z-statistic and the independent samples z-statistic respectively.

Definition of Proposed Test Statistics

Two new t-statistics are proposed; T_{new1}, assuming equal variances, and T_{new2}, when equal variances cannot be assumed. The test statistics are constructed as the difference between two means taking into account the covariance structure. The numerator is the difference between the means of the two samples and the denominator is a measure of the standard error of this difference. Thus the test statistics proposed here are directly testing the hypothesis $H_0 : \mu_1 = \mu_2$.

The test statistic T_{new1} is derived so that in the extremes of $n_a = n_b = 0$ or $n_c = 0$, T_{new1} defaults to T_1 or T_2 respectively, thus

$$T_{\text{new1}} = \frac{\overline{\mathrm{X}}_1 - \overline{\mathrm{X}}_2}{S_p \sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2} - 2r\left(\dfrac{n_c}{n_1 n_2}\right)}} \quad \text{where } S_p = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{(n_1 - 1) + (n_2 - 1)}}$$

The test statistic T_{new1} is referenced against the t-distribution with degrees of freedom derived by linear interpolation between v_1 and v_2 so that: $v_{\text{new1}} = (n_c - 1) + \left(\dfrac{n_a + n_b + n_c - 1}{n_a + n_b + 2n_c}\right)\left(n_a + n_b\right)$.

In the extremes, when $n_a = n_b = 0$, v_{new1} defaults to v_1; or when $n_c = 0$, v_{new1} defaults to v_2.

Given the superior Type I error robustness of Welch's test when variances are not equal, a test statistic is derived making reference to Welch's approximate degrees of freedom. This test statistic makes use of the sample variances, S_1^2 and S_2^2. The test statistic T_{new2} is derived so that in the extremes of $n_a = n_b = 0$ or $n_c = 0$, T_{new2} defaults to T_1 or T_3 respectively, thus

$$T_{\text{new2}} = \frac{\overline{\mathrm{X}}_1 - \overline{\mathrm{X}}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2} - 2r\left(\frac{S_1 S_2 n_c}{n_1 n_2}\right)}}$$

The test statistic T_{new2} is referenced against the t-distribution with degrees of freedom derived as a linear interpolation between v_1 and v_3 so that

$$v_{\text{new2}} = (n_c - 1) + \left(\frac{\gamma - n_c + 1}{n_a + n_b + 2n_c}\right)(n_a + n_b) \text{ where } \gamma = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\left(\frac{S_1^2}{n_1}\right)^2 / (n_1 - 1) + \left(\frac{S_2^2}{n_2}\right)^2 / (n_2 - 1)}$$

In the extremes, when $n_a = n_b = 0$, v_{new2} defaults to v_1; or when $n_c = 0$, v_{new2} defaults to v_3.

Note that the proposed statistics, T_{new1} and T_{new2}, use all available observations in the respective variance calculations. The statistic $Z_{\text{corrected}}$ only uses the paired observations in the calculation of covariance.

Worked Example

An applied example is given to demonstrate the calculation of each of the test statistics defined. In education, for credit towards an undergraduate Statistics course, students may take optional modules in either Mathematical Statistics, or Operational Research, or both. The programme leader is interested whether the exam marks for the two optional modules differ. The exam marks attained for a single semester are given in Table 2.

Table 2. Exam marks for Students studying on an undergraduate Statistics course.

Student	Mathematical Statistics	Operational Research
1	73	72
2	82	
3	74	89
4	59	78
5	49	64
6		83
7	42	42
8	71	76
9		79
10	39	89
11		67
12		82
13		85
14		92
15	59	63
16	85	

As per standard notion, the derived sample values are given in lower case. In the calculation of the test statistics, $\bar{x}_1 = 63.300$, $\bar{x}_2 = 75.786$, $s_1^2 = 263.789$, $s_2^2 = 179.874$, $n_a = 2$, $n_b = 6$, $n_c = 8$, $n_1 = 10$, $n_2 = 14$, $v_1 = 7$, $v_2 = 6$, $v_3 = 6$, $\gamma = 17.095$, $v_{\text{new1}} = 12$, $v_{\text{new2}} = 10.365$, $r = 0.366$, $s_{12} = 78.679$.

For the REML analysis, a mixed model is performed with “Module” as a repeated measures fixed effect and “Student” as a random effect. Table 3 gives the calculated test statistics, degrees of freedom and corresponding p-values.

Table 3. Test statistic values and resulting p-values (two-sided test).

	T_1	T_2	T_3	$Z_{\text{corrected}}$	REML	T_{new1}	T_{new2}
estimate of mean difference	-13.375	2.167	2.167	-12.486	-12.517	-12.486	-12.486
t-value	-2.283	0.350	0.582	-2.271	-2.520	-2.370	-2.276
degrees of freedom	7.000	6.000	6.000		11.765	12.000	10.365
p-value	0.056	0.739	0.579	0.023	0.027	0.035	0.045

With the exception of REML, the estimates of the mean difference are simply the difference in the means of the two samples, based on the observations used in the calculation. It can quickly be seen that the conclusions differ depending on the test used. It is of note that only the tests using all of the available data result in the rejection of the null hypothesis at $\alpha_{\text{nominal}} = 0.05$. Also note that the results of the paired samples t-test and the independent samples t-test have sample effects in different directions. This is only one specific example given for illustrative purposes, investigation is required into the power of the test statistics over a wide range of scenarios. Conclusions based on the proposed tests cannot be made without a thorough investigation into their Type I error robustness.

Simulation Design

Under normality, Monte-Carlo methods are used to investigate the Type I error robustness of the defined test statistics and REML. Power should only be used to compare tests when their Type I error rates are equal (Zimmerman & Zumbo, 1993). Monte-Carlo methods are used to explore the power for the tests that are Type I error robust under normality.

Unbalanced designs are frequent in psychology (Sawilowski & Hillman, 1982), thus a comprehensive range of values for n_a, n_b and n_c are simulated. These values offer an extension to the work done by Looney & Jones (2003). Given the identification of separate test statistics for equal and unequal variances, multiple population variance parameters $\{\sigma_1^2,\ \sigma_2^2\}$ are considered. Correlation has an impact on Type I error and power for the paired samples t-test (Fradette et. al., 2003), hence a range of correlations $\{\rho\}$ between two normal populations are considered. Correlated normal variates are obtained as per Kenney & Keeping (1951). A total of 10,000 replicates of each of the scenarios in Table 4 are performed in a factorial design.

All simulations are performed in R version 3.1.2. For the mixed model approach utilising REML, the R package lme4 is used. Corresponding p-values are calculated using the Satterthwaite approximation adopted by SAS using the R package lmerTest (Goodnight, 1976).

For each set of 10,000 p-values, the proportion of times the null hypothesis is rejected, for a two sided test with $\alpha_{\text{nominal}} = 0.05$ is calculated.

Table 4. Summary of simulation parameters

Parameter	Values
μ_1	0
μ_2	0 (under H_0)
	0.5 (under H_1)
σ_1^2	1, 2, 4, 8
σ_2^2	1, 2, 4, 8
n_a	5, 10, 30, 50, 100, 500
n_b	5, 10, 30, 50, 100, 500
n_c	5, 10, 30, 50, 100, 500
ρ	-0.75, -0.50, -0.25, 0.00, 0.25, 0.50, 0.75

Type I Error Robustness

For each of the test statistics, Type I error robustness is assessed against Bradley's (1978) liberal criteria. This criteria it is widely used in many studies analysing the validity of t-tests and their adaptions. Bradley's (1978) liberal criteria states that the Type I error rate α should be within $\alpha_{\text{nominal}} \pm 0.5\,\alpha_{\text{nominal}}$. For $\alpha_{\text{nominal}} =$ 0.05, Bradley's liberal interval is [0.025, 0.075].

Type I error robustness is firstly assessed under the condition of equal variances. Under the null hypothesis, 10,000 replicates are obtained for the 4 × 6 × 6 × 6 × 7 = 6,048 scenarios where $\sigma_1^2 = \sigma_2^2$. Figure 1 shows the Type I error rates for each of the test statistics under equal variances for normally distributed data.

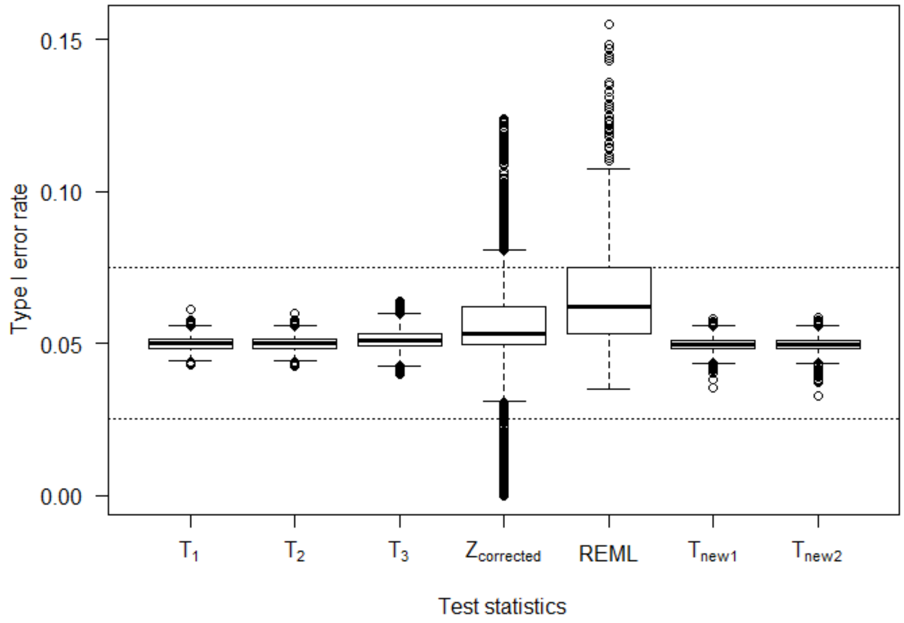


Figure 1. Type I error rates where $\sigma_1^2 = \sigma_2^2$, reference lines show Bradley's (1978) liberal criteria.

Figure 1 indicates that when variances are equal, the statistics T_1, T_2, T_3, T_{new1} and T_{new2} remain within Bradley's liberal Type I error robustness criteria throughout the entire simulation design. The statistic $Z_{\text{corrected}}$ is not Type I error robust, thus confirming the smaller simulation findings of Mehotra (2004). Figure 1 also shows that REML is not Type I error robust throughout the entire simulation design. A review of our results shows that for REML the scenarios that are outside the range of liberal Type I error robustness are predominantly those that have negative correlation, and some where zero correlation is specified. Given that negative correlation is rare in a practical environment, the REML procedure is not necessarily unjustified.

Type I error robustness is assessed under the condition of unequal variances. Under the null hypothesis, 10,000 replicates were obtained for the 4 × 3 × 6 × 6 × 7 = 18,144 scenarios where

$\sigma_1^2 \neq \sigma_2^2$. For assessment against Bradley's (1978) liberal criteria, Figure 2 shows the Type I error rates for unequal variances for normally distributed data.

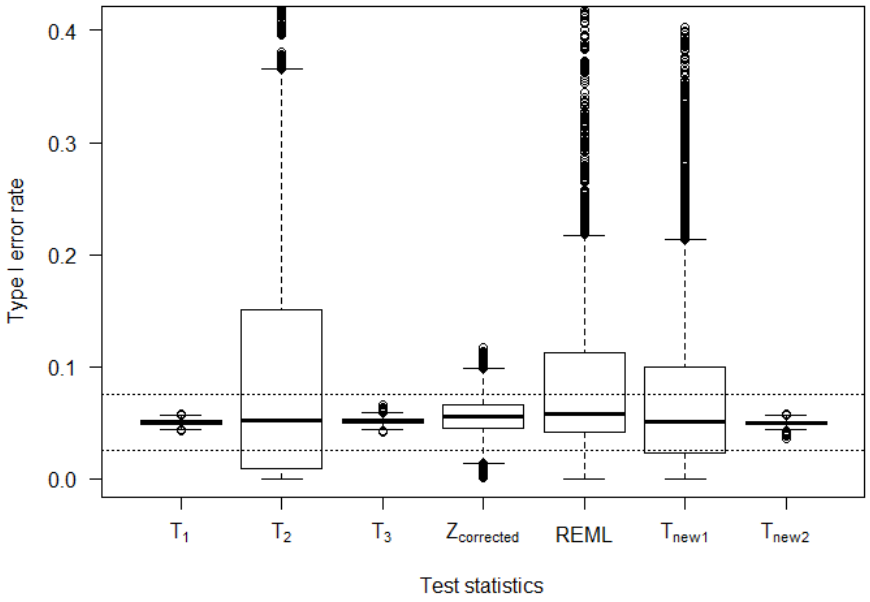


Figure 2. Type I error rates when $\sigma_1^2 \neq \sigma_2^2$, reference lines show Bradley's (1978) liberal criteria.

It can be seen from Figure 2 that the statistics defined using a pooled standard deviation T_2 and T_{new1}, do not provide Type I error robust solutions when equal variances cannot be assumed. The statistics T_1, T_3 and T_{new2} retain their Type I error robustness under unequal variances throughout all conditions simulated.

The statistic $Z_{\text{corrected}}$ maintains similar Type I error rates under equal and unequal variances. The statistic $Z_{\text{corrected}}$ was only designed to be used in the case of equal variances. For unequal variances, we observe that the statistic $Z_{\text{corrected}}$ results in an unacceptable amount of false positives when $\rho \leq 0.25$ or max $\{n_a, n_b, n_c\}$ - min$\{n_a, n_b, n_c\}$ is large. In addition, the statistic $Z_{\text{corrected}}$ is conservative when ρ is large and positive. The largest observed deviations from Type I error robustness for REML are when $\rho \leq 0$ or max {n_a, n_b, n_c} - min{n_a, n_b, n_c} is large. Further insight to the Type I error rates for REML can be seen in Figure 3 showing observed p-values against expected p-values from a uniform distribution.

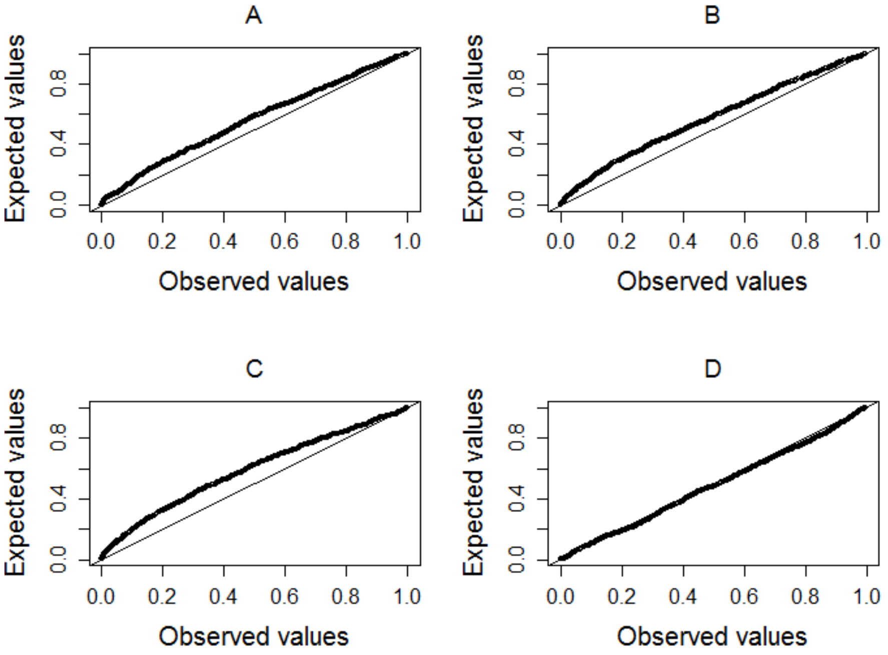


Figure 3. P-P plots for simulated p-values using REML procedure. Selected parameter combinations ($n_a, n_b, n_c, \sigma_1^2, \sigma_2^2, \rho$) are as follows; A$=$(5,5,5,1,1,-0.75), B$=$(5,10,5,8,1,0), C$=$ (5,10,5,8,1,0.5), D $=$(10,5,5,8,1,0.5).

If the null hypothesis is true, for any given set of parameters the p-values should be uniformly distributed. Figure 3 gives indicative parameter combinations where the p-values are not uniformly distributed when applying a mixed model assessed using REML. It can be seen that REML is not Type I error robust when the correlation is negative. In addition, caution should be exercised if using REML when the larger variance is associated with the smaller sample size. REML maintains Type I error robustness for positive correlation and equal variances or when the larger sample size is associated with the larger variance.

Power of Type I Error Robust Tests under Equal Variances

The test statistics that do not fail to maintain Bradley's Type I error liberal robustness criteria are assessed under H_1. REML is included in the comparisons for $\rho \geq 0$. The power of the test statistics are assessed where $\sigma_1^2 = \sigma_2^2 = 1$, followed by an assessment of the power of the test statistics where $\sigma_1^2 > 1$ and $\sigma_2^2 = 1$.

Table 5 shows the power of T_1, T_2, T_3, T_{new1}, T_{new2} and REML, averaged over all sample size combinations where $\sigma_1^2 = \sigma_2^2 = 1$.

Table 5. Power of Type I error robust test statistics, $\sigma_1^2 = \sigma_2^2 = 1$, $\alpha = 0.05$, $\mu_2 - \mu_1 = 0.5$.

	ρ	T_1	T_2	T_3	T_{new1}	T_{new2}	REML
$n_a = n_b$	0.75	0.785	0.567	0.565	0.887	0.886	0.922
	0.50	0.687	0.567	0.565	0.865	0.864	0.880
	0.25	0.614	0.567	0.565	0.842	0.841	0.851
	0	0.558	0.567	0.565	0.818	0.818	0.829
	< 0	0.481	0.567	0.565	0.778	0.778	-
$n_a \neq n_b$	0.75	0.784	0.455	0.433	0.855	0.847	0.907
	0.50	0.687	0.455	0.433	0.840	0.832	0.861
	0.25	0.615	0.455	0.433	0.823	0.816	0.832
	0	0.559	0.455	0.433	0.806	0.799	0.816
	< 0	0.482	0.455	0.433	0.774	0.766	-

Table 5 shows that REML and the test statistics proposed in this paper, T_{new1} and T_{new2}, are more powerful than standard approaches, T_1, T_2 and T_3, when variances are equal. Consistent with the paired samples t-test, T_1, the power of T_{new1} and T_{new2} is relatively lower when there is zero or negative correlation between the two populations. Similar to contrasts of the independent samples t-test, T_2, with Welch's test, T_3, for equal variances but unequal sample sizes, T_{new1} is marginally more powerful than T_{new2}, but not to any practical extent. For each of the tests statistics making use of paired data, as the correlation between the paired samples increases, the power increases.

As the correlation between the paired samples increases, the power advantage of the proposed test statistics relative to the paired samples t-test becomes smaller. Therefore the proposed statistics T_{new1} and T_{new2} may be especially useful when the correlation between the two populations is small.

To show the relative increase in power for varying sample sizes, Figure 4 shows the power for selected test statistics for small-medium sample sizes, averaged across the simulation design for equal variances.

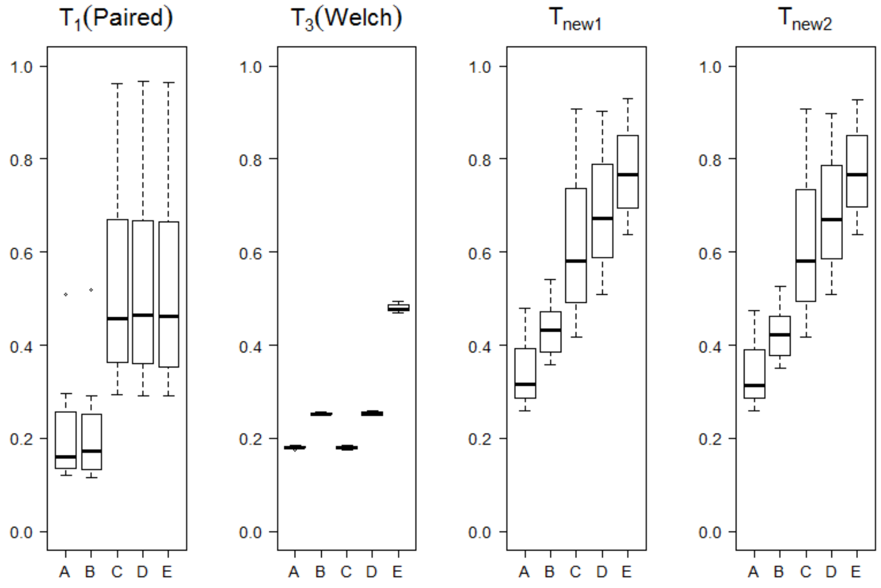


Figure 4. Power for Type I error robust test statistics, averaged across all values of ρ where $\sigma_1^2 = \sigma_2^2$ and $\mu_2 - \mu_1 = 0.5$. The sample sizes (n_a, n_b, n_c) are as follows; $A = (10,10,10)$, $B = (10,30,10)$, $C = (10,10,30)$, $D = (10,30,30)$, $E = (30,30,30)$.

From Figure 4 it can be seen that for small–medium sample sizes, the power of the proposed test statistics T_{new1} and T_{new2} is superior to standard test statistics.

Power of Type I Error Robust Rests under Unequal Variances

For the Type I error robust test statistics under unequal variances, Table 6 shows the power of T_1, T_3, T_{new2} and REML, averaged over the simulation design where $\mu_2 - \mu_1 = 0.5$.

Table 6. Power of Type I error robust test statistics where $\sigma_1^2 > 1$, $\sigma_2^2 = 1, \alpha = 0.05$, $\mu_2 - \mu_1 = 0.5$. Within this table, $n_a > n_b$ represents the larger variance associated with the larger sample size, and $n_a < n_b$ represents the larger variance associated with the smaller sample size.

	ρ	T_1	T_3	T_{new2}	REML
$n_a = n_b$	0.75	0.555	0.393	0.692	0.645
	0.50	0.481	0.393	0.665	0.588
	0.25	0.429	0.393	0.640	0.545
	0	0.391	0.393	0.619	0.515
	< 0	0.341	0.393	0.582	-
$n_a > n_b$	0.75	0.555	0.351	0.715	0.589
	0.50	0.481	0.351	0.688	0.508
	0.25	0.429	0.351	0.665	0.459
	0	0.391	0.351	0.642	0.422
	< 0	0.341	0.351	0.604	-
$n_a < n_b$	0.75	0.555	0.213	0.559	0.693
	0.50	0.481	0.213	0.539	0.649
	0.25	0.429	0.213	0.522	0.620
	0	0.391	0.213	0.507	0.603
	< 0	0.341	0.213	0.480	-

Table 6 shows that T_{new2} has superior power properties to both T_1 and T_3 when variances are not equal. In common with the performance of Welch's test for independent samples, T_3, the power of T_{new2} is higher when the larger variance is associated with the larger sample size. In common with the performance of the paired samples t-test, T_1, the power of T_{new2} is relatively lower when there is zero or negative correlation between the two populations.

The apparent power gain for REML when the larger variance is associated with the larger sample size, can be explained by the pattern in the Type I error rates. REML follows a similar pattern to the independent samples t-test, which is liberal when the larger variance is associated with the larger sample size, thus giving the perception of higher power.

To show the relative increase in power for varying sample sizes, Figure 5 shows the power for selected test statistics for small-medium sample sizes, averaged across the simulation design for unequal variances.

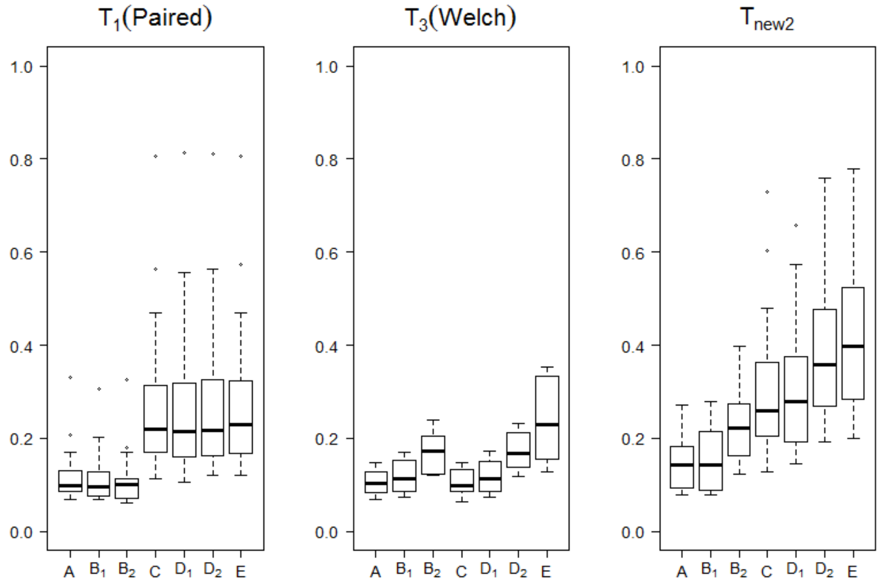


Figure 5. Power for Type I error robust test statistics, $\sigma_1^2 > \sigma_2^2$ and $\mu_2 - \mu_1 = 0.5$. The sample sizes (n_a, n_b, n_c) are as follows; $A =$(10,10,10), $B_1 =$(10,30,10), $B_2 =$(30,10,10), $C =$(10,10,30), $D_1 =$ (10,30,30), $D_2 =$(30,10,30), $E =$(30,30,30).

Figure 5 shows a relative power advantage when the larger variance is associated with the larger sample size, as per B_2 and D_2. A comparison of Figure 4 and Figure 5 shows that for small-medium sample sizes, power is adversely effected for all test statistics when variances are not equal.

Discussion

The statistic T_{new2} is Type I error robust across all conditions simulated under normality. The greater power observed for T_{new1}, compared to T_{new2}, under equal variances, is likely to be of negligible consequence in a practical environment. This is in line with empirical evidence for the performance of Welch's test, when only independent samples are present, which leads to many observers recommending the routine use of Welch's test under normality (e.g. Ruxton, 2006).

The Type I error rates and power of T_{new2} follow the properties of its counterparts, T_1 and T_3. Thus T_{new2} can be seen as a trade-off between the paired sample t-test and Welch's test, with the advantage of increased power across all conditions, due to using all available data.

The partially overlapping samples scenarios identified in this paper could be considered as part of the missing data framework and all simulations have been performed under the assumption of MCAR.

The statistics proposed in this paper form less computationally intensive competitors to REML. The REML procedure does not directly calculate the difference between the two sample means, in a practical environment this makes its results hard to interpret. The statistics proposed in this paper will also far more easily lend themselves to the development of non-parametric tests.

Conclusion

A commonly occurring scenario when comparing two means is a combination of paired observations and independent observations in both samples, this scenario is referred to as partially overlapping samples. Standard procedures for analysing partially overlapping samples involve discarding observations and performing either the paired samples t-test, or the independent samples t-test, or Welch's test. These approaches are less than desirable. In this paper, two new test statistics making reference to the t-distribution are introduced and explored under a comprehensive set of parameters, for normally distributed data. Under equal variances, T_{new1} and T_{new2} are Type I error robust. In addition they are more powerful than standard Type I error robust approaches considered in this paper. When variances are equal, there is a slight power advantage of using T_{new1} over T_{new2}, particularly when sample sizes are not equal. Under unequal variances, T_{new2} is the most powerful Type I error robust statistic considered in this paper. We recommend that when faced with a research problem involving partially overlapping samples and MCAR can be reasonably assumed, the statistic T_{new1} could be used when it is known that variances are equal. Otherwise under the same conditions when equal variances cannot be assumed the statistic T_{new2} could be used.

A mixed model procedure using REML is not fully Type I error robust. In those scenarios in which this procedure is Type I error robust, the power is similar to that of T_{new1} and T_{new2}.

The proposed test statistics for partially overlapping samples provide a real alternative method for analysis for normally distributed data, which could also be used for the formation of confidence intervals for the true difference in two means.

References

Bedeian, A. G., & Feild, H. S. (2002). Assessing group change under conditions of anonymity and overlapping samples. *Nursing research, 51*(1), 63-65.

Bhoj, D. (1978). Testing equality of means of correlated variates with missing observations on both responses. *Biometrika, 65*(1), 225-228. doi: 10.1093/biomet/65.1.225

Bradley, J. V. (1978). Robustness?. *British Journal of Mathematical and Statistical Psychology, 31*(2), 144-152. doi: 10.1111/j.2044-8317.1978.tb00581.x

Derrick, B., Dobson-McKittrick, A., Toher, D., & White P. (2015). Test statistics for comparing two proportions with partially overlapping samples. *Journal of Applied Quantitative Methods, 10*(3)

Derrick, B., Toher, D., & White, P. (2016). Why Welch's test is Type I error robust. *The Quantitative Methods for Psychology, 12*(1), 30-38. doi: 10.20982/tqmp.12.1.p030

Ekbohm, G. (1976). Comparing Means in the Paired Case with Incomplete Data on Both Responses, *Biometrika, 63*(2), 299-304. doi: 10.1093/biomet/63.2.299

Fay, M. P., & Proschan, M. A. (2010). Wilcoxon-Mann-Whitney or t-test? On assumptions for hypothesis tests and multiple interpretations of decision rules. *Statistics surveys, 4,* 1. doi: 10.1214/09-SS051

Fisher, R. A. (1925). Statistical methods for research workers. Genesis Publishing Pvt Ltd.

Fradette, K., Keselman, H. J., Lix, L., Algina, J., & Wilcox, R. (2003). Conventional and Robust Paired and Independent Samples t-tests: Type I Error and Power Rates, *Journal of Modern Applied Statistical Methods, 2*(2), 481-496.

Kenney, J. F., & Keeping, E. S. (1951). *Mathematics of statistics*, Pt. 2, 2nd ed. Princeton, NJ: Van Nostrand.

Kim, B. S., Kim, I., Lee, S., Kim, S., Rha, S. Y., & Chung, H. C. (2005). Statistical methods of translating microarray data into clinically relevant diagnostic information in colorectal cancer. *Bioinformatics, 21*(4), 517-528. doi: 10.1093/bioinformatics/bti029

Lin, P. E. (1973). Procedures for testing the difference of means with incomplete data. *Journal of the American Statistical Association, 68*(343), 699-703.

Lin, P. E., & Strivers L. (1974). Difference of Means with Incomplete Data. *Biometrika, 61*(2), 325-334. doi: 10.1093/biomet/61.2.325

Looney, S., & Jones, P. (2003). A method for comparing two normal means using combined samples of correlated and uncorrelated data. *Statistics in medicine, 22*, 1601-1610. doi:10.1002/sim.1514

Martínez-Camblor, P., Corral, N., & María de la Hera, J. (2013). Hypothesis test for paired samples in the presence of missing data. *Journal of Applied Statistics*, *40*(1), 76-87. doi: 10.1080/02664763.2012.734795

Mehrotra, D. (2004). Letter to the editor, a method for comparing two normal means using combined samples of correlated and uncorrelated data. *Statistics in medicine*, *23*(7), 1179–1180. doi: 10.1002/sim.1693

R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. www.R-project.org. 2014; version 3.1.2.

Ruxton., G. (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. doi: 10.1093/beheco/ark016

Goodnight, J. H. (1976) General Linear Models Procedure. S.A.S. Institute. Inc.

Samawi, H. M., & Vogel, R. (2011). Tests of homogeneity for partially matched-pairs data. *Statistical Methodology, 8*(3), 304-313. doi: 10.1016/j.stamet.2011.01.002

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*(2), 352. doi: 10.1037/0033-2909.111.2.352

Sawilowsky, S. S., & Hillman, S. B. (1992). Power of the independent samples t-test under a prevalent psychometric measure distribution, *Journal of Consulting and Clinical Psychology, 60*(2), 240-243. doi: 10.1037/0022-006X.60.2.240

Stouffer, S. A., Suchman, E. A., DeVinney, L. C., Star, S. A., & Williams Jr, R. M. (1949). The American soldier: adjustment during army life. *Studies in social psychology in World War II*, 1.

Zimmerman, D . W. (1997). A note on the interpretation of the paired samples, *Journal of educational and behavioral statistics, 22*(3), 349 – 360. doi:10.3102/10769986022003349

Zimmerman, D. W., & Zumbo, B. D. (1993). Significance testing of correlation using scores, ranks, and modified ranks. *Educational and psychological measurement, 53*(4), 897-904.

Zimmerman, D. W., & Zumbo, B. D. (2009). Hazards in choosing between pooled and separate-variances t tests. *Psicológica: Revista de metodología y psicología experimental, 30*(2), 371-390.

Zumbo, B. D. (2002). An adaptive inference strategy: The case of auditory data. *Journal of Modern Applied Statistical Methods, 1*(1), 60-68. doi: 10.22237/jmasm/1020255000

Appendix P4

Derrick, B., Toher, D., and White, P. (2017). “How to compare the means of two samples that include paired observations and independent observations: A companion to Derrick, Russ, Toher and White (2017)”. *The Quantitative Methods in Psychology* 13 (2), pp. 120–126
Published Version

2017 ■ Vol. 13 ■ no. 2

How to compare the means of two samples that include paired observations and independent observations: A companion to Derrick, Russ, Toher and White (2017)

Ben Derrick[a], Deirdre Toher[a] & Paul White[a, ✉]

[a]University of the West Of England, Bristol, England, UK

Abstract ■ Standard approaches for comparing the means of two samples, comprising both paired observations and independent observations, involve the discarding of valuable information. An alternative test which uses all of the available data, is the partially overlapping samples t-test. Two variations of the test are available, one assuming equal variances, and one assuming separate variances. Issues with standard procedures, and considerations for choosing appropriate tests in the partially overlapping scenario are discussed. An example with details of how to apply the partially overlapping samples t-test is given along with an R package that implement these new tests.

Acting Editor ■ Denis Cousineau (Université d'Ottawa)

Keywords ■ Partially overlapping samples; incomplete observations; Welch's test; independent samples; paired samples; equality of means . **Tools** ■ R.

✉ Paul.White@uwe.ac.uk

iD *BD*: 0000-0002-4064-1780; *DT*: 0000-0003-0788-6142; *PW*: 0000-0002-7503-9896

doi 10.20982/tqmp.13.2.p120

Introduction

It is well established that the paired samples t-test can be used for comparing means between two dependent samples (Zimmerman, 1997; Fradette, Keselman, Lix, Algina, & Wilcox, 2003). The assumptions of the paired samples t-test are that data are randomly sampled from two related populations, and that the differences between the paired observations are approximately normally distributed. It is also well established that the independent samples t-test can be used for comparing means between two independent samples with equal variances (Rasch, Teuscher, & Guiard, 2007; Fradette et al., 2003). When variances are not equal, the independent samples t-test is not Type I error robust, particularly when the sample sizes are not equal (Ramsey, 1980). When equal variances cannot be assumed, a Type I error robust alternative to the independent samples t-test is Welch's test (Derrick, Toher, & White, 2016; Fradette et al., 2003). For the avoidance of doubt, here the independent samples t-test assuming equal variances is referred to as the independent samples t-test, and the independent samples t-test not assuming equal variances is referred to as Welch's test. The assumptions of the independent samples t-test and Welch's test are that data are randomly sampled from two unrelated populations, which are approximately normally distributed. Welch's test is considered Type I error robust for all but the most extreme deviations from the normality assumption (Ruxton, 2006). Extensive testing of these assumptions is not recommended (Rasch, Kubinger, & Moder, 2011; Rochon, Gondan, & Kieser, 2012). A further assumption of these tests is that observations within a sample are independent of each other. This assumption is critical, violations of the independence of observations assumption make hypothesis testing invalid (Lissitz & Chardos, 1975).

Conventional teaching of statistics usually assumes a perfect world with completely dependent samples or completely independent samples (for example, Magel, 1998). However, a question that is often asked in research is how to compare means between two samples that include both paired observations and unpaired observations. These scenarios are referred to as 'partially overlapping samples'

(Derrick, Russ, Toher, & White, 2017; Derrick, Dobson-McKittrick, Toher, & White, 2015; Martinez-Camblor, Corral, & de la Hera, 2012). Paired samples designs are often advantageous relative to independent samples designs, because paired samples designs allow differences between two samples to be directly compared. However, partially overlapping samples designs are often required due to the limited resource of paired samples, where a number of independent observations are available to compensate. This could also occur in a matched pairs design, when pairing individuals on certain characteristics, there may be some additional independent observations that cannot be reasonably paired on any characteristics. In addition there are occasions where it is desired that observations from a paired samples design, and a separate independent samples design, may be combined, resulting in a partially overlapping samples design.

The approach for analysing partially overlapping samples by design has received relatively little attention within the literature. Consider the scenarios in Figure 1, which demonstrates eight scenarios where there are two samples, each with a different number of paired observations and independent observations.

These alternative approaches emphasise the need for statistically valid tests in the partially overlapping samples case.

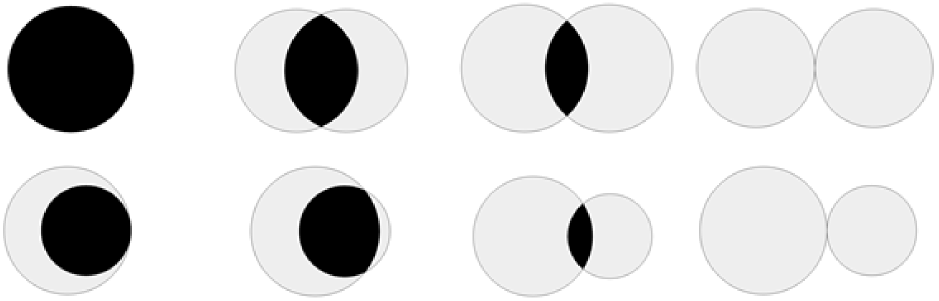
It is not well established how to proceed for the scenarios represented by Figure 1 where there is partial overlap. One 'standard' approach if the number of pairs is large, is to perform the paired samples t-test on only the paired observations. Conversely, if the number of independent samples is large a 'standard' approach is to perform the independent samples t-test or Welch's test, on only the independent samples (Looney & Jones, 2003). These standard methods discard data which adversely impacts the power of the test. Approaches that discard data are likely to maintain adequate power if the number of discarded observations is relatively 'small' and the sample sizes are relatively 'large'. One alternative approach that is commonly applied, is to perform the independent samples t-test on all of the available data. However, this is less powerful than a paired samples approach and ignores the fact that there are matched pairs. Alternative ad hoc approaches using all of the available data, but not mimicking the design structure, will not be considered further in this paper. These alternative approaches emphasise the need for statistically valid tests in the partially overlapping samples case.

A frequent occurrence of partially overlapping samples is a paired samples design with missing observations (Martinez-Camblor et al., 2012). In this situation partially overlapping samples do not occur by design, and so it is necessary to consider why the samples are incomplete. If data are missing completely at random (MCAR), the reason for missing data is not related to the value of the observation itself, or other variables recorded. An example of data that is MCAR is a question in a survey that is accidentally missed, or data that is accidentally lost. If incomplete observations are MCAR, it is reasonable to discard the corresponding paired observations without causing bias (Donders, van der Heijden, Stijnen, & Moons, 2006). If data are missing at random (MAR), data are missing based on characteristics not directly measured by the missing observation itself. However, the missing data is related to another variable in the dataset. The discarding of information that are MAR is likely to cause bias, therefore the standard approach of pairwise or listwise deletion is not recommended (Schafer, 1997; Donders et al., 2006). If data are missing not at random (MNAR), the probability of an observation being missing, directly depends on the value of the observation being recorded. When data are MNAR, there is no statistical procedure that can eliminate potential bias (Musil, Warner, Yobas, & Jones, 2002). This is particularly of concern for analyses with missing data because it is difficult to distinguish between data that is MAR and data that is MNAR. Nevertheless if the amount of missing data is small, the bias is likely to be inconsequential. The literature suggests that up to 5% of observations missing is acceptable (Graham, 2009; Schafer, 1997). Some take a more liberal stance suggesting that up to 20% of data missing may be acceptable (Schlomer, Bauman, & Card, 2010).

For a paired samples design with incomplete observations, researchers often attempt to impute the missing data. Ad hoc basic imputation approaches for imputing missing data are biased solutions (Schafer, 1997). Mean imputation reduces the variation in the data set. Regression imputation inflates the correlation between variables. More sophisticated techniques, Expected Maximisation and Multiple Imputation, minimise the bias of the parameter estimates (Musil et al., 2002; Dong & Peng, 2013).

Standard statistical software will perform the paired samples t-test, the independent samples t-test or Welch's test upon command. In SAS the standard 'proc ttest' performs the paired samples t-test, omitting cases pairwise from calculations when any observation from a declared paired variable is missing. Likewise in Unistat, a paired samples t-test is performed, excluding any 'missing values' pairwise. Performing the paired samples t-test in SPSS gives the options of excluding cases pairwise or excluding cases listwise, which are equivalent in the two sample

Figure 1 ■ Examples of 'partially overlapping samples'. In each scenario each of two samples are represented by a circle. The paired observations are represented by the overlap and shaded black. From left to right the graphic shows a decreasing number of paired observations. The relative sample sizes are represented by the size of the circle.



case. In all of these approaches, the unpaired observations are excluded and the analysis is done only on the paired data. Caution should be exercised when using SAS, SPSS or Unistat, because users may be tempted to analyse only the complete pairs when readily presented with the opportunity, and not realise the consequences of not using all of the data. Both Minitab and the standard 't.test' in R present an error message when a paired samples t-test is selected with unequal sample sizes, these software at the very least make users aware there are considerations to take into account with the analysis they are trying to perform.

Derrick, Russ, et al. (2017) developed two partially overlapping samples t-tests that make use of all of the available data, that are valid under MCAR and robust under the assumptions of normality. These test statistics act as a straightforward interpolation between the paired samples t-test, and either the independent samples t-test, or Welch's test. Using these tests for comparing two sample means represents a more powerful alternative to discarding information. In the case of a paired samples design with incomplete observations, these test statistics also represent an alternative to the need to perform complicated imputation techniques.

In this paper, the partially overlapping samples test statistics that make use of all of the available data, accounting for the fact that there is a combination of paired observations and independent observations, are demonstrated by use of example. It also shows how to perform these new tests using an R package, `partiallyoverlapping`. The paper concludes with a discussion on comparing the use of traditional tests against the partially overlapping samples t-tests.

Worked Example

In this section, an example of the partially overlapping t-test in application is given, with a summary of the calculations and the hypothesis test procedure.

The sleep fragmentation index measures the quality of sleep for an individual over one night. A lower sleep fragmentation score represents less disrupted sleep. The research question is whether the genre of a movie watched before bedtime impacts the quality of sleep. The data are plausible fictional data used for illustrative purposes only.

Study participants are randomly allocated to either a between subjects design (stage 1) or a repeated measures (stage 2) part of the investigation. In the first stage of the study, the sleep fragmentation score is taken over one night, for two groups of individuals. A sample of $n_a = 8$ individuals watch a 'horror' movie before bedtime. A separate sample of $n_b = 8$ individuals watch a 'feel good' movie before bedtime. This first stage is an independent samples design. In a second stage of the study, the sleep fragmentation index is recorded over two separate nights, for a sample of $n_c = 8$ individuals watching a 'feel good' movie and a 'horror' movie on two alternate nights before bedtime (with order counterbalanced). This second stage is a paired samples design. When the two stages of the study are combined, the total number of individuals who watched a 'horror' movie is $n_1 = n_a + n_c = 16$. The total number of individuals who watched a 'feel good' movie is $n_2 = n_b + n_c = 16$. The hypothesis being tested is whether the mean sleep fragmentation scores are the same between individuals watching a 'horror' movie and individuals watching a 'feel good' movie. Thus the null hypothesis is : $H_0 : \mu_1 = \mu_2$. The alternative hypothesis, assuming a two-sided test is : $H_1 : \mu_1 \neq \mu_2$. The sleep fragmentation scores are given in Table 1.

In this scenario, from a missing data perspective it would be reasonable to assume MCAR. There are no missing data per se; it is the design of the study that results in partially overlapping samples. Therefore standard approaches of discarding either the paired or independent samples are unbiased. However, performing either the paired samples t-test or the independent samples t-test requires discarding exactly half of the observations, and the power of the test is reduced. This therefore is a good example of where a test statistic that makes use of all available data, taking into account both paired and independent observations could be useful.

Assuming normality and MCAR, the partially overlapping samples t-test is a Type I error robust method for comparing means between the two samples (Derrick, Russ, et al., 2017). To calculate elements for the partially overlapping samples t-test let: $\bar{x}_1$ be the mean of all observations in Sample 1 (i.e. the mean for the n_1 observations for individuals watching a 'horror' movie), $\bar{x}_2$ be the mean of all observations in Sample 2 (i.e. the mean for the n_2 observations for individuals watching a 'feel good' movie), s_1 be the standard deviation of all observations in Sample 1, s_2 be the standard deviation of all observations in Sample 2, and r be the Pearson's correlation coefficient for the paired observations only (i.e. in n_c). There are two forms of the partially overlapping samples t-test; t_1 for when equal variances between the two samples can be assumed, and t_2 for when equal variances between the two samples cannot be assumed.

The partially overlapping samples t-test assuming equal variances acts as an interpolation between the independent samples t-test and the paired samples t-test, and is defined by Derrick, Russ, et al. (2017) as:

$$t_1 = \frac{\bar{x}_1 - \bar{x}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2} - 2r\frac{n_c}{n_1 n_2}}} \tag{1}$$

where

$$s_p = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{(n_1-1)+(n_2-1)}} \tag{2}$$

If the null hypothesis is true, the test statistic t_1 follows a t-distribution with approximate degrees of freedom given as:

$$\nu_1 = (n_c - 1) + \frac{n_a + n_b + n_c - 1}{n_a + n_b + 2n_c}(n_a + n_b). \tag{3}$$

If equal variances cannot be assumed, the partially overlapping samples t-test which acts as an interpolation between Welch's test and the paired samples t-test is defined by Derrick, Russ, et al. (2017) as:

$$t_2 = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} - 2r\frac{s_1 s_2 n_c}{n_1 n_2}}} \tag{4}$$

If the null hypothesis is true, the test statistic t_2 follows a t-distribution with degrees of freedom approximated by:

$$\nu_2 = (n_c - 1) + \frac{\gamma - n_c + 1}{n_a + n_b + 2n_c}(n_a + n_b) \tag{5}$$

and where

$$\gamma = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{(s_1^2/n_1)^2}{n_1-1} + \frac{(s_2^2/n_2)^2}{n_2-1}} \tag{6}$$

These test statistics can be viewed as a generalised form of the two sample t-tests. When there are no independent observations, t_1 and t_2 default to the paired samples t-test. When there are no paired observations, t_1 defaults to the independent samples t-test, and t_2 defaults to Welch's test.

For either version of the partially overlapping samples t-test, if $\mu_1 > \mu_2$ (i.e. the population mean score for 'horror' movie is greater than the population mean score for 'feel good' movie), then it is anticipated that this will be reflected in the sample values above, and the expectation is to observe a large positive value of the test statistic. Conversely if $\mu_1 < \mu_2$, the expectation would be for a large but negative value of the test statistic to be observed. In absolute terms it is anticipated that large values of the test statistic will be observed if the null hypothesis is not true. The null hypothesis is rejected if the observed value of the test statistic is greater than the critical value from a t-distribution with the degrees of freedom as defined by ν_1 or ν_2.

The elements of the calculation of the test statistics are[1] : $n_1 = 16$, $n_2 = 16$, $n_a = 8$, $n_b = 8$, $n_c = 8$, $\bar{x}_1 = 16.125$, $\bar{x}_2 = 14.125$, $s_1 = 2.986$, $s_2 = 2.778$, $r = 0.687$, $s_p = 2.884$, $\gamma = 29.845$, $t_1 = 2.421$, $t_2 = 2.419$, $\nu_1 = 18.500$, and $\nu_2 = 18.422$.

The calculated value of the test statistic t_1 is 2.421. The calculated value of the test statistic t_2 is 2.419. Using the degrees of freedom $\nu_1 = 18.500$ or $\nu_2 = 18.422$, from the t-distribution at the 5% significance level the critical value is 2.097. The calculated value of the test statistic is greater than the critical value, therefore the null hypothesis is rejected (p=0.026).

Instead of performing the above calculations manually, the partially overlapping samples t-tests can be easily performed in R, using the package 'Partiallyoverlapping' (Derrick, 2017). In the following, let 'a' represent 'horror' movie

[1] Unrounded values are used in each part of the calculation, each element displayed to 3 decimal places.

Table 1 ■ Sleep fragmentation scores obtained for each individual (ID)

Independent Samples (Stage 1)				Paired Samples (Stage 2)		
ID	Horror	ID	Feel good	ID	Horror	Feel good
I1	20	I9	10	P1	14	15
I2	21	I10	16	P2	15	10
I3	16	I11	18	P3	18	15
I4	18	I12	16	P4	20	17
I5	14	I13	15	P5	11	13
I6	12	I14	14	P6	19	19
I7	14	I15	13	P7	14	12
I8	17	I16	10	P8	15	13

and ‘b’ represent ‘feel good’ movie. Example R code to enter the data and perform the analyses assuming equal variances is given below:

```
install.packages('Partiallyoverlapping')
library(Partiallyoverlapping)
a.unpaired <- c(20,21,16,18,14,12,14,17)
b.unpaired <- c(10,16,18,16,15,14,13,10)
a.paired   <- c(14,15,18,20,11,19,14,15)
b.paired   <- c(15,10,15,17,13,19,12,13)
Partover.test(a.unpaired, b.unpaired,
   a.paired, b.paired, var.equal=TRUE)
#Output:  statistic =2.421,  parameter=18.500,
#            p.value=0.026.
```

Alternatively, to perform the test when equal variances are not assumed, the `var.equal=TRUE` option can be dropped or replaced by `var.equal= FALSE`. The results from either test performed replicate their respective manual calculation and show that the samples from group ‘a’ (Horror movie) and group ‘b’ (Feel good movie) have significantly different means at the 5% significance level.

When using the partially overlapping samples t-test at the 5% significance level, there is a statistically significant difference in the mean sleep fragmentation index between individuals watching a ‘horror’ movie prior to bedtime, and individuals watching a ‘feel good’ movie prior to bedtime. The results suggest that individuals watching a ‘feel good’ movie before bedtime, have less disrupted sleep compared to individuals watching a ‘horror’ movie before bedtime.

Discussion

Further consideration is given to the choice between traditional tests that discard information, and the partially overlapping samples t-tests. Table 2 gives a summary of results obtained from the example in Table 1. This shows results when performing ‘standard’ tests and results from performing the partially overlapping samples t-tests, with their respective statistical decisions at the 5% significance level.

It can be seen from Table 2 that the choice of test to apply is important because the statistical decision is not the same. This example emphasises the lower power for the traditional approaches. In general, the more observations used in the calculation of a test statistic, the greater the power of the test will be. However, rare situations may arise where the independent observations and the paired observations have mean differences in opposing directions. In these situations the partially overlapping samples t-test may cancel out these differences, but to ignore either the paired observations or independent observations could create bias.

In the worked example, the two samples are partially overlapping by design. It is also possible to encounter a partially overlapping samples design, with incomplete observations. In these situations, the partially overlapping samples t-test can similarly be performed on all available observations, when the missing observations are MCAR. To demonstrate this, consider the situation where there are occasional errors with the machine recording sleep fragmentation. As a result of errors, let the ‘horror’ observations for individuals ‘I1’ and ‘P1’ be missing. There is now one missing independent ‘horror’ observation and one missing paired observation. The resulting reduction in sample size is further to the detriment of the paired samples t-test, the independent samples t-test and Welch’s test. Using the partially overlapping samples t-test, the ‘feel good’ observation for individual ‘P1’ is not discarded, it is treated as an independent observation. Revised elements of the partially overlapping samples t-test are; $n_1 = 14$, $n_2 = 16$, $n_a = 7$, $n_b = 9$, $n_c = 7$, $\bar{x}_1 = 16.000$, $\bar{x}_2 = 14.125$, $s_1 = 2.961$, $s_2 = 2.778$, $r = 0.736$, $s_p = 2.864$, $\gamma = 26.903$, $t_1 = 2.208$, $t_2 = 2.194$, $\nu_1 = 17.733$, ν_2 =17.148. Assuming equal variances and using the test statistic t_1, the p-value is 0.041. For completion, using the test statistic t_2, the p-value is 0.042. The null hypothesis is

Table 2 ■ Summary of results for the worked example, including the calculated value of each test statistic (t), the degrees of freedom (df), the p-value (p) and the statistical decision.

Test	t	df	p	Decision
Paired samples t-test	1.821	7.000	0.111	Fail to reject H_0
Independent samples t-test	1.667	14.000	0.118	Fail to reject H_0
Welch's test	1.667	13.912	0.118	Fail to reject H_0
Partially overlapping samples t-test (t_1)	2.421	18.500	0.026	Reject H_0
Partially overlapping samples t-test with Welch's df (t_2)	2.419	18.422	0.026	Reject H_0

rejected at the 5% significance level and the statistical conclusions are as before.

The assumptions of the partially overlapping samples t-test (t_1) match the assumptions of the independent samples t-test. The assumptions are that observations within a sample are independent of each other, observations are sampled from normally distributed populations and equal variances between the two groups. The assumptions of the partially overlapping samples t-test with Welch's degrees of freedom (t_2), match the assumptions of Welch's test. This assumes that observations within a sample are independent of each other and observations are sampled from normally distributed populations. Similarly as stated for the standard tests that discard data, extensive testing of these assumptions is not recommended. The partially overlapping samples t-test with Welch's degrees of freedom is Type I error robust with equal and unequal variances, and the power difference relative to the independent samples t-test is negligible. Many authors advocate the routine use of Welch's test in the two independent samples case, (for example, Ruxton, 2006; Rasch et al., 2011). Therefore, if in doubt and normality and MCAR can be assumed, the partially overlapping samples t-test with Welch's degrees of freedom can be used routinely in the two partially overlapping samples case.

Conclusion

A common issue in psychology is a paired samples design with incomplete observations, or a study that otherwise results in both paired observations and independent observations being observed. These scenarios are referred to in the literature as partially overlapping samples.

In these scenarios, the discarding of observations is common practice. However, discarding observations may cause bias, and has a substantial impact on power when sample sizes are small and/or if the number of discarded observations is large. The partially overlapping samples approach uses all available data and has appeal when the assumption of normality has not been grossly violated, and the MCAR assumption is reasonable. These solutions do not detract from other analytical strategies but do provide a simple generalisation of the standard two sample t-tests.

References

Derrick, B. (2017). Partiallyoverlapping: Partially Overlapping Samples t-Tests [R package] (Version 1.0).

Derrick, B., Dobson-McKittrick, A., Toher, D., & White, P. (2015). Test statistics for comparing two proportions with partially overlapping samples. *Journal of Applied Quantitative Methods*, *10*(3), 1–14.

Derrick, B., Russ, B., Toher, D., & White, P. (2017). Test statistics for the comparison of means for two samples which include both paired observations and independent observations. *Journal of Modern Applied Statistical Methods*, *16*(1). doi:10.22237/jmasm/1493597280

Derrick, B., Toher, D., & White, P. (2016). Why Welch's test is Type I error robust. *The Quantitative Methods for Psychology*, *12*(1), 30–38. doi:10.20982/tqmp.12.1.p030

Donders, A. R. T., van der Heijden, G. J., Stijnen, T., & Moons, K. G. (2006). Review: A gentle introduction to imputation of missing values. *Journal of Clinical Epidemiology*, *59*(10), 1087–1091. doi:10.1016/j.jclinepi.2006.01.014

Dong, Y. & Peng, C. Y. J. (2013). *Principled missing data methods for researchers*. 2(1), 1-17: SpringerPlus. doi:10.1186/2193-1801-2-222

Fradette, K., Keselman, H. J., Lix, L., Algina, J., & Wilcox, R. (2003). Conventional and robust paired and independent samples t-tests: Type I error and power rates. *Journal of Modern Applied Statistical Methods*, *2*(2), 481–496. doi:10.22237/jmasm/1067646120

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology*, *60*, 549–576. doi:10.1146/annurev.psych.58.110405.085530

Lissitz, W. & Chardos, S. (1975). A study of the effect of the violations of the assumption of independent sampling upon the type one error rate of the two-sample t-test. *Educational and Psychological Measurement*, *35*, 353–359. doi:10.1177/001316447503500213

Looney, S. & Jones, P. (2003). A method for comparing two normal means using combined samples of correlated and uncorrelated data. *Statistics in Medicine*, *22*, 1601–1610. doi:10.1002/sim.1514

Magel, R. C. (1998). Testing for differences between two brands of cookies. *Teaching Statistics*, *20*(3), 81–83. doi:10.1111/j.1467-9639.1998.tb00775.x

Martinez-Camblor, P., Corral, N., & de la Hera, J. (2012). Hypothesis test for paired samples in the presence of missing data. *Journal of Applied Statistics*, *40*, 76–87. doi:10.1080/02664763.2012.734795

Musil, C. M., Warner, C. B., Yobas, P. K., & Jones, S. L. (2002). A comparison of imputation techniques for handling missing data. *Western Journal of Nursing Research*, *24*(7), 815–829. doi:10.1177/019394502762477004

Ramsey, P. H. (1980). Exact Type I error rates for robustness of Student's t test with unequal variances. *Journal of Educational and Behavioral Statistics*, *5*(4), 337–349. doi:10.2307/1164906

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

Rasch, D., Teuscher, F., & Guiard, V. (2007). How robust are tests for two independent samples? *Journal of Statistical Planning and Inference*, *137*(8), 2706–2720. doi:10.1016/j.jspi.2006.04.011

Rochon, J., Gondan, M., & Kieser, M. (2012). To test or not to test: preliminary assessment of normality when comparing two independent samples. *BMC Medical Research Methodology*, *12*(1), 1–11. doi:10.1186/1471-2288-12-81

Ruxton, G. (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

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. Roxton: CRC press.

Schlomer, G. L., Bauman, S., & Card, N. A. (2010). Best practices for missing data management in counseling psychology. *Journal of Counseling Psychology*, *57*(1), 1–11. doi:10.1037/a0018082

Zimmerman, D. (1997). A note on the interpretation of the paired samples t-test. *Journal of Educational and Behavioral Statistics*, *22*(3), 349–360. doi:10.3102/10769986022003349

Citation

Derrick, B., Toher, D., & White, P. (2017). How to compare the means of two samples that include paired observations and independent observations: A companion to Derrick, Russ, Toher and White (2017). *The Quantitative Methods for Psychology*, *13*(2), 120–126. doi:10.20982/tqmp.13.2.p120

Received: 25/04/2017 ~ Accepted: 29/04/2017

Appendix P5

Derrick, B. et al. (2017b). “The impact of an extreme observation in a paired samples design”. *metodološki zvezki-Advances in Methodology and Statistics* 14

Published Version

Metodološki zvezki, Vol. 14, No. 2, 2017, 1–17

The Impact of an Extreme Observation in a Paired Samples Design

Ben Derrick[1] Antonia Broad Deirdre Toher Paul White

Abstract

The effect of systematically altering the value of a single observation within a paired differences design is considered. A paradox is observed for the paired samples t-test, where increasing the value of an observation in the direction of the true mean difference results in a higher p-value. Using simulation, deviations from robustness of the paired samples t-test is demonstrated, and is contrasted with Yuen's paired samples t-test and the Wilcoxon signed rank sum test.

1 Introduction

The paired samples t-test is logically and numerically equivalent to the one sample t-test performed on paired differences, and it is one of the most well-established and commonly performed statistical tests. Zimmerman (1997) demonstrated that the type I error rate of the paired samples t-test remains close to the nominal significance level for varying correlation and sample sizes under normality. Under less idealised conditions, Posten (1979), Herrendörfer et al, (1983), Rasch and Guiard (2004), and Fradette et al. (2003) found that the paired samples t-test maintains type I error robustness for a range of non-normal distributions. However, Blair and Higgins (1985) found the Wilcoxon signed rank sum test to also be type I error robust and to have some power advantages over the paired samples t-test for a range of non-normal distributions. Chaffin and Rhiel (1993) demonstrated that the tails of the sampling distribution of the paired samples test statistic are skewness dependent, particularly with relatively small sample sizes.

Zumbo and Jennings (2002), using a novel contamination model, determined the effect of outliers on the validity and power of the paired samples t-test. They found the paired samples t-test to have robust validity for symmetric contamination, but with increasing inflation of the type I error rate with increasing asymmetric contamination. This is coupled with degradation in power in the presence of outliers when the true effect is small and sample sizes are small. In their work the number of outliers in the sample is considered to be a random variable.

One of the assumptions of the paired samples t-test is that the differences between the two samples are normally distributed, or alternatively and in a practical sense, that the mean difference has a distribution which can reasonably be approximated by a normal

[1] Engineering, Design and Mathematics, University of the West England, Bristol, United Kingdom; ben.derrick@uwe.ac.uk

distribution. A closely related assumption is that there are no large outliers in the differences. When performing the paired samples t-test, there may be competition between the magnitude of the mean difference and the standard deviation of the differences. In particular, extreme observations within a dataset can distort the balance between these two elements of the test. To illustrate this, consider the example data in Table 1.

Table 1: Example data for six units within a paired design

Pair	Sample 1	Sample 2	Difference
1	30	22	8
2	28	18	10
3	45	45	0
4	57	54	3
5	38	32	6
6	37	37 - X	X

For the first five pairs, the mean of Sample 1 is greater than or equal to the mean of Sample 2. For the sixth pair, let the difference between the Sample 1 observation and the Sample 2 observation be denoted as X. Intuition might suggest that a positive value of X may contribute towards an overall significant difference in means being observed. If this were the case, a large positive value of X should seemingly contribute towards a significant effect. In the following, the value of X is systematically altered in order to demonstrate its impact on the paired samples t-test. The observation X will "march" through the data set and will be colloquially referred to as a marching observation. Table 2 shows the results of a two-sided paired samples t-test for negative values of X through to large positive values of X.

Table 2: Paired samples t-test on five degrees of freedom for increasing values of X

X	t	p-value	X	t	p-value	X	t	p-value
-3	1.984	0.104	11	3.670	0.014	25	2.425	0.060
-1	2.406	0.061	13	3.461	0.018	27	2.319	0.068
1	2.870	0.035	15	3.240	0.023	29	2.226	0.077
3	3.321	0.021	17	3.033	0.029	31	2.145	0.085
5	3.671	0.014	19	2.848	0.036	33	2.073	0.093
7	3.840	0.012	21	2.687	0.043	35	2.009	0.101
9	3.820	0.012	23	2.546	0.052	37	1.953	0.108

The values of X for which the null hypothesis of equal means is rejected at the 5% significance level are highlighted in Table 2. For low values of X it can be seen that as the value of X increases, the p-value decreases. In this example, as the value of X increases beyond approximately 8, the p-value increases. As the value of the observed difference in the sixth pair increases (and hence as the mean difference increases), the p-value also increases. Observing an extreme value of X in the direction of the seemingly observed effect can increase the sample variance to such an extent that it impedes the test from giving a significant result. The extreme observation paradox is the contrariwise p-value increase as the value of an extreme observation increases in the direction of the overall effect.

As the absolute value of the marching observation increases, the assumptions of the paired samples t-test are increasingly violated. When the sample size is small or the assumptions of the paired samples t-test are violated, researchers often choose to perform the Wilcoxon signed rank sum test. Aguinis et al., (2013) summarise a comprehensive list of techniques for dealing with outliers and state that non-parametric tests give results that are robust in the presence of outliers. However, Zimmerman (2011) indicates that rank based methods do not necessarily eliminate the influence of outliers. Another alternative approach when outliers are present is to use Yuen's paired samples t-test. In this test, the principles of trimmed means outlined by Yuen (1974), are applied to the paired differences (Wilcox, 2005).

In this paper, simulation is used to explore the scenarios in which the extreme observation paradox is observed in a paired samples design. We are particularly interested in isolating those situations when two-sided hypothesis testing is undertaken (e.g. see Ringwalt et al., 2011), when sample sizes are relatively small (i.e. when outliers may have a greater effect on the paired samples t-test). The concept of a systematically marching observation similar to the demonstration in Table 2, is used to investigate the effects of an aberrant observation. In the simulation design, this aberrant observation is a forced additional observation not fitting with the simulated data, and is not due to inherent variability. Simulations are performed for an aberrant observation in the direction of the effect suggested by the rest of the sample, and secondly where an aberrant observation is in the opposing direction of the effect suggested by the rest of the sample. Thus situations where the sign of the marching observation is concordant or discordant with the mean of the other observations are considered. For comparative purposes, the paired samples t-test, the Wilcoxon signed rank sum test, and Yuen's paired samples t-test are included.

Null hypothesis significance testing is most frequently performed with a nil-null hypothesis specifying that no difference between groups is present, and a two directional alternative (Levine et al., 2008). Therefore the impact of an extreme observation for a two-sided test is the main emphasis of this paper. However, one-sided tests retain some practical utility, and the simulations are extended to a one-sided test.

We hypothesise that the seemingly paradoxical behaviour exhibited in Table 2 will be a feature of the paired samples t-test in general. In contrast, we hypothesise that Yuen's paired samples t-test and the Wilcoxon signed rank sum test will be robust to a single aberrant observation.

In order to gain insight, we firstly investigate the mathematical limiting forms of each of the three test statistics under consideration as a single marching observation becomes increasingly large compared with the rest of the sample, and then proceed to a simulation investigation.

2 An Unbounded Marching Observation

For development purposes consider a random sample $X_1, X_2, \ldots, X_{n-1}, X_n$, and let $X_{(1)} < X_{(2)} < \cdots X_{(n-1)} < X_{(n)}$ denote the order statistics. Further, let $X_k = Y_k$

for ($k = 1, 2, \dots, n-1$), let $Y_{(1)} < Y_{(2)} < \cdots Y_{(n-1)}$ be the corresponding order statistics, and let $X_n = \xi$ be the marching observation. In this notation, Y_k ($k = 1, \dots, n-1$) denotes the observations prior to the inclusion of the marching observation.

The following analytical exposition investigates the behaviour of the one sample t-test, Yuen's paired samples t-test, and the Wilcoxon signed rank sum test, as the marching observation $X_n = \xi$ becomes relatively large compared with the rest of the sample.

2.1 The t-test

Consider the single sample t-test test statistic on the paired differences, used to test $H_0\colon \mu_X = 0$, defined by

$$T\colon = \frac{\bar{X}}{\hat{\sigma}_{X+}}\sqrt{n}$$

where

$$\bar{X} := \frac{X_1 + X_2 + \cdots + X_n}{n}$$

and

$$\hat{\sigma}_{X+} := \sqrt{\frac{(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + \cdots + (X_n - \bar{X})^2}{n-1}}.$$

Observe that

$$\bar{X} = \frac{(n-1)\bar{Y} + \xi}{n}, \; \bar{X} - \bar{Y} = \frac{\xi - \bar{Y}}{n}, \text{ and } \xi - \bar{X} = \frac{(n-1)(\xi - \bar{Y})}{n}.$$

Thus

$$\hat{\sigma}_{X+} = \sqrt{\frac{(Y_1 - \bar{X})^2 + (Y_2 - \bar{X})^2 + \cdots + (Y_{n-1} - \bar{X})^2 + (\xi - \bar{X})^2}{n-1}}.$$

Note that

$$\begin{aligned}\sum_{j=1}^{n-1}(Y_j - \bar{X})^2 &= \sum_{j=1}^{n-1}(Y_j - \bar{Y} + \bar{Y} - \bar{X})^2 \\ &= \sum_{j=1}^{n-1}(Y_j - \bar{Y})^2 + (n-1)(\bar{Y} - \bar{X})^2 + 2(\bar{Y} - \bar{X})\sum_{j=1}^{n-1}(Y_j - \bar{Y}) \\ &= (Y_1 - \bar{Y})^2 + (Y_2 - \bar{Y})^2 + \cdots + (Y_{n-1} - \bar{Y})^2 + (n-1)(\bar{Y} - \bar{X})^2 + 0.\end{aligned}$$

Hence

$$\hat{\sigma}_{X+} = \sqrt{\frac{(Y_1 - \bar{Y})^2 + (Y_2 - \bar{Y})^2 + \cdots + (Y_{n-1} - \bar{Y})^2 + (\xi - \bar{X})^2}{n-1} + (\bar{X} - \bar{Y})^2}.$$

For the $n-1$ values, define

$$\hat{\sigma}_Y := \sqrt{\frac{(Y_1 - \bar{Y})^2 + (Y_2 - \bar{Y})^2 + \cdots + (Y_{n-1} - \bar{Y})^2}{n-1}}.$$

Note that $\hat{\sigma}$ does not have the "+" symbol, i.e. that the marching observation is not included. An alternative definition for $\hat{\sigma}$ could have $n-2$ in the denominator and so

$$\begin{aligned}
\hat{\sigma}_{X+} &= \sqrt{\hat{\sigma}_Y^2 + \frac{(\xi-\bar{X})^2}{n-1} + (\bar{X}-\bar{Y})^2} \\
&= \sqrt{\hat{\sigma}_Y^2 + \frac{(n-1)^2}{n^2}\frac{(\xi-\bar{Y})^2}{n-1} + \frac{(\xi-\bar{X})^2}{n^2}} \\
&= \sqrt{\hat{\sigma}_Y^2 + \frac{(\xi-\bar{Y})^2}{n}}
\end{aligned}$$

and hence

$$T = \frac{(n-1)\bar{Y}+\xi}{\sqrt{n\hat{\sigma}_Y^2 + (\xi-\bar{Y})^2}}$$

It can be seen that as $\xi \to \infty, T \to 1$, and similarly as $\xi \to -\infty, T \to -1$. Accordingly, for any value of significance level likely to be encountered in practice the results $\xi \to \pm\infty$, $T \to \pm 1$ indicate that the null hypothesis would not be rejected under the stated conditions.

2.2 Yuen's Paired Samples t-test

Let γ denote the per tail proportion of trimming, let $e := \lfloor \gamma n \rfloor$ and let $f := n - 2e$. Define the trimmed sample $X_{t1}, X_{t2}, \ldots, X_{tf-1}, X_{tf}$ as $X_{tk} := X_{(k+e)}$ $(k = 1, 2, \ldots, f)$ and define the winsorised sample $X_{w1}, X_{w2}, \ldots, X_{wf}$ as

$$X_{wk} := \begin{cases} X_{(e+1)} & k = 1, 2, \ldots, e \\ X_{(k)} & k = e+1, e+2, \ldots, n-e \\ X_{(n-e)} & k = n-e+1, n-e+2, \ldots, n \end{cases}$$

Let $\bar{X}_t = \sum_{k=1}^{f} X_{tk}/f$, and $\bar{X}_w = \sum_{k=1}^{n} X_{wk}/n$ define the trimmed mean and winsorised mean respectively and let, $\hat{\sigma}^2_{Xw+} = \sum_{k=1}^{n}(X_{wk}-\bar{X}_w)^2/(n-1)$ denote the winsorised variance. In this notation, Yuen's test statistic is given by $T_Y := \frac{\bar{X}_t}{\hat{\sigma}_{Xw+}}\sqrt{n}(1-2\gamma)$.

For $\xi < Y_{(e)}$

$$\bar{X}_t = \frac{Y_{(e)} + Y_{(e+1)} + Y_{e+2} + \cdots + Y_{(n-e-1)}}{f},$$
$$\bar{X}_w = \frac{eY_{(e)} + Y_{(e)} + Y_{(e+1)} + Y_{(e+2)} + \cdots + Y_{n-e-1} + eY_{n-e-1}}{n}$$

and

$$\hat{\sigma}^2_{Xw+} = \frac{e(Y_{(e)}-\bar{X}_w)^2 + \sum_{k=e}^{n-e-1}(Y_{(k)}-\bar{X}_w)^2 + e(Y_{(n-e-1)}-\bar{X}_w)^2}{n-1}$$

For fixed values, $Y_1, Y_2, \ldots, Y_{n-1}$, as $\xi \to -\infty$, $T_Y := \frac{\bar{X}_t}{\hat{\sigma}_{Xw+}}\sqrt{n}(1-2\gamma)$ stabilises to some limiting value.

Similarly, for $\xi > Y_{(n-e)}$

$$\bar{X}_t = \frac{Y_{(e+1)} + Y_{(e+2)} + Y_{(e+2)} + \cdots + Y_{(n-e)}}{f},$$

$$\bar{X}_w = \frac{eY_{(e+1)} + Y_{(e+1)} + Y_{(e+2)} + Y_{(e+2)} + \cdots + Y_{(n-e)} + eY_{n-e}}{n}$$

and

$$\hat{\sigma}^2_{Xw+} = \frac{e(Y_{(e+1)} - \bar{X}_w)^2 + \sum_{k=e+1}^{n-e}(Y_{(k)} - \bar{X}_w)^2 + e(Y_{(n-e)} - \bar{X}_w)^2}{n-1}$$

For fixed values, $Y_1, Y_2, \ldots, Y_{n-1}$ as $\xi \to -\infty$, $T_Y := \frac{\bar{X}_t}{\hat{\sigma}_{Xw+}}\sqrt{n}(1-2\gamma)$ stabilises to some limiting value. Moreover, for a sufficiently large sample, the limit values for both directions of the marching observation should be close to each other. Hence the properties displayed as $\xi \to -\infty$ or $\xi \to -\infty$ are consistent with T_Y being a robust test statistic.

2.3 The Wilcoxon signed rank sum test

Assuming no ties and no zero observations, then the test statistic for the Wilcoxon signed rank sum test, W, is defined as

$$W = R_1^X \operatorname{sgn}(X_1) + R_2^X \operatorname{sgn}(X_2) + \cdots + R_n^X \operatorname{sgn}(X_n)$$

where R_k^X is the rank of $|X_k|$ among $|X_1|, |X_2|, \ldots, |X_n|$. If $X_1, X_2, \ldots, X_n$ are independent and follow the same symmetric continuous distribution, then W follows a distribution with mean 0 and variance $n(n+1)(2n+1)/6$.

Denote by R_k^Y the rank of $|Y_k|$ among $|Y_1|, |Y_2|, \ldots, |Y_{n-1}|$. For

$$|\xi| > \max\{|Y_1|, |Y_2|, \ldots, |Y_{n-1}|\},$$

$$W = R_1^Y \operatorname{sgn}(Y_1) + R_2^Y \operatorname{sgn}(Y_2) + \cdots + R_n^Y \operatorname{sgn}(Y_n) + n \operatorname{sgn}(\xi).$$

Hence under the stated conditions, for fixed values $Y_1, Y_2, \ldots, Y_{n-1}$, the Wilcoxon signed rank sum statistic stabilises to some situation dependent limit value as $\xi \to +\infty$, and to some situation dependent limit value as $\xi \to -\infty$. The difference between these two values is $n - (-n) = 2n$, and the standardised values differ by $\sqrt{24n/\{(n+1)(2n+1)\}}$. These are close to each other for sufficiently large n.

3 Simulation Methodology

The approach is to generate sample data meeting the assumptions of the paired samples t-test, and to then include an additional observation in the sample. This additional observation systematically changes in its observed value. The paired samples t-test, the Wilcoxon signed rank test, and Yuen's paired samples t-test, are performed for a two-sided nil-null hypothesis. Under a two-sided nil-null hypothesis; the paired samples t-test is used to test a distribution mean difference of zero; and Yuen's paired samples t-test is used to test the distribution of the trimmed mean equal to zero. Historically, the derivation of the Wilcoxon rank sum distribution has been made for continuous random variables under a null hypothesis of no distributional differences, and is sensitive to changes in central location (Gibbons and Chakraborti, 2011).

Within the simulation, the differences are generated rather than the paired observations themselves. Specifically, $n - 1$ random normal deviates $n_1, x_2, \ldots, x_{n-1}$ are generated using the Box-Muller (1958) transformation, where n represents the sample size of the paired differences. Under H_0 , the $n - 1$ random normal deviates have a population mean of zero ($\mu = 0$) and a standard deviation of one ($\sigma = 1$).

To isolate the phenomenon and behaviour of interest, if $\bar{x}_{n-1} = \sum_{i=1}^{n-1} x_i/(n-1) < 0$ then $x_1, x_2, \ldots, x_{n-1}$ are multiplied by -1 to ensure a non-negative sample mean. (This change of sign does not affect the validity of a two-sided test of a nil-null hypothesis for these data.)

Under H_1, for each of the $n - 1$ deviates, a constant d is added to each of the values. The simulations are performed under normality so that the data fulfil the assumptions of the test with the exception of an aberrant observation.

An additional observation, x_n, is added to the $n-1$ observations to give a total sample size of n. For any simulated sample, the value of x_n is systematically varied from -8 to 8 in increments of 0.1. It is this value, x_n, which is referred to as the ‘marching observation’. The values of x_n approximately range between ± 8 standard deviations from the mean and would therefore cover limits likely encountered in a practical environment. Note that the condition of $\bar{x}_{n-1} > 0$ is to ensure that the concordance of effects ($\bar{x}_{n-1} > 0$, $x_n > 0$) or discordance of effects ($\bar{x}_{n-1} > 0$, $x_n < 0$) can be established.

A summary of the values of n, x_n and d used in the full factorial simulation design is given in Table 3. The simulation is run $10\,000$ times for each combination of sample size and mean difference.

In a second set of simulations, the impact of the marching observation is similarly assessed, removing the condition that the mean sample difference is positive, and performing a one-sided test. This is done as per the parameter combinations in Table 3 using upper tail critical values.

Table 3: Summary of simulation design

Sample size	10, 15, 20, 25
Marching observation	-8:8 (0.1)
Mean difference	0, 0.5
Significance level	5%
Number of Iterations	10 000
Programming Language	R version 3.1.3

For the paired samples t-test and the Wilcoxon signed rank sum test, the default `stats` package in R is used. Yuen’s paired samples t-test is performed using the R package `PairedData` as outlined by Wilcox (2005). 10% trimming per tail is performed.

The proportion of the $10\,000$ iterations where the null hypothesis is rejected is calculated at the nominal significance level of 5%. This gives the Null Hypothesis Rejection Rate (NHRR). Note that the terminology NHRR is used and not type I error rate, because the inclusion of the marching observation would strictly invalidate the underpinning assumptions of the resultant test. The effect of gradually increasing the marching observation is to gradually violate the assumption of the nil-null hypothesis.

The research question being asked is "How is the performance of the paired samples t-test, Yuen's paired samples t-test, and the Wilcoxon signed rank sum test affected by the presence of an aberrant observation?"

4 Results

The Null Hypothesis Rejection Rate (NHRR) is assessed for each of the three statistical tests under consideration for a two-sided test, firstly when $d = 0$ and secondly in the presence of a systematic effect size ($d = 0.5$).

Figure 1 gives the NHRR of the paired samples t-test when $d = 0$, using the nominal significance level of 5%.

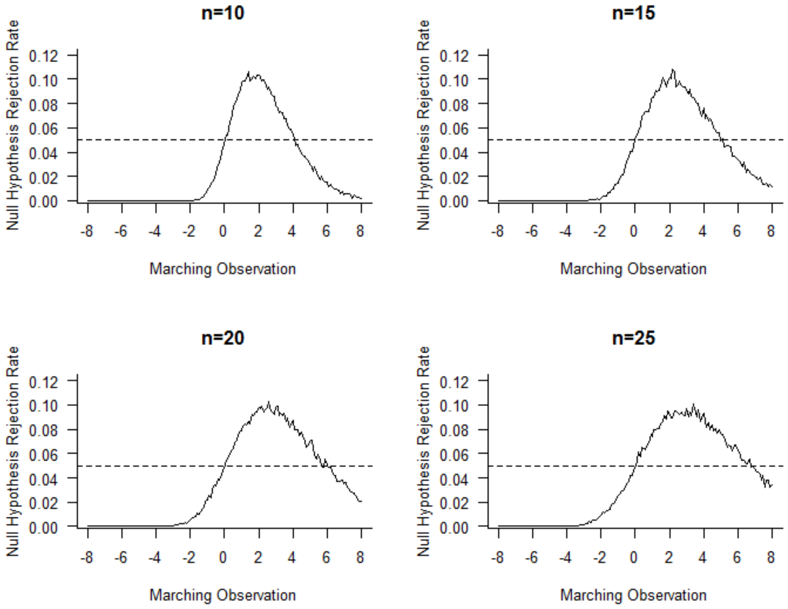


Figure 1: NHRR of the paired samples t-test, $d = 0$, two-sided

Figure 1 shows that when the value of $x_n = d = 0$, the NHRR is approximately equal to the nominal type I error rate of 5%. For positive sample means, as the value of x_n starts to increase above zero, the paired samples t-test has an increasingly higher NHRR until a turning point is reached and with a subsequent return to the nominal type I error rate. Extreme and increasingly larger values of the marching observation, x_n, in the direction of the sample effect results in a progressively lower NHRR, with values noticeably lower than the nominal type I error rate. These effects are replicated in all four sample sizes, but the effects are marginally less noticeable with increasing sample size. Figure 1 also shows that a large value for the marching observation in the opposite direction to the mean of the first $n - 1$ observations, effectively results in a zero value for the NHRR. This effect is consistent with the asymptotic behaviour given in Section 2 and the findings alluded to in the example given in Table 2.

Figure 2 gives the NHRR of Yuen's paired samples t-test and Figure 3 gives the NHRR of the Wilcoxon signed rank sum test, both when $d = 0$.

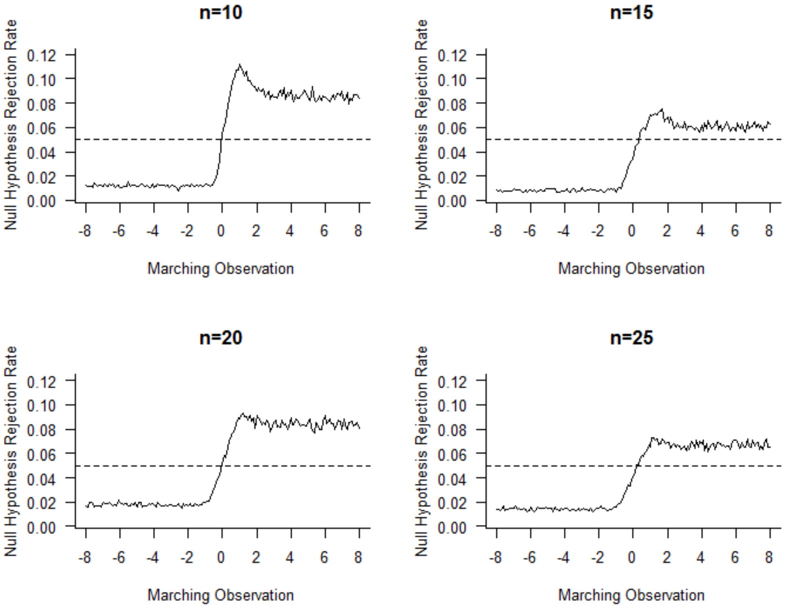


Figure 2: NHRR of Yuen's paired samples t-test, $d = 0$, two-sided.

Figure 2 and Figure 3 show that when $x_n > 0$ and $\bar{x}_{n-1} > 0$, both Yuen's paired samples t-test and the Wilcoxon signed rank sum test result in the null hypothesis being rejected more frequently than the nominal significance level. Conversely, when $x_n < 0$ and $\bar{x}_{n-1} > 0$, both Yuen's paired samples t-test and the Wilcoxon signed rank sum test have a NHRR lower than the nominal significance level. These findings are entirely consistent with expectation for a robust test given the design of the simulation.

For the Wilcoxon signed rank sum test, due to the use of rank values, the test is not greatly affected by the magnitude of the extreme observation. Similarly due to the trimming, Yuen's paired samples t-test is not greatly affected by the magnitude of the extreme observation. The phenomenon of a turning point when $x_n > 0$ is not observed for either the Wilcoxon signed rank sum test or Yuen's paired samples t-test.

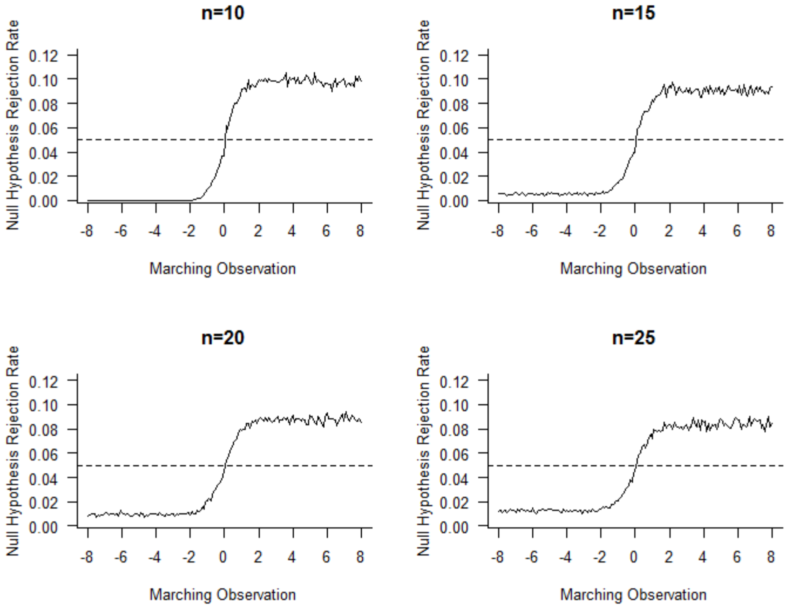


Figure 3: NHRR of the Wilcoxon signed rank sum test, $d = 0$, two-sided.

Figure 4 gives indicative power of the paired samples t-test, where $d = 0.5$. For a sample of size $n = 10$ independent Normal deviates with $\mu = 0$ and $\sigma = 1$, the power of the test for the paired samples t-test for testing $H_0\colon \mu = 0$ is 0.293. Under the same conditions, the power of the paired samples t-test for $n = 15$, 20 and 25 is 0.438, 0.565, and 0.670 respectively. These reference lines are added to the graphics for comparative purposes.

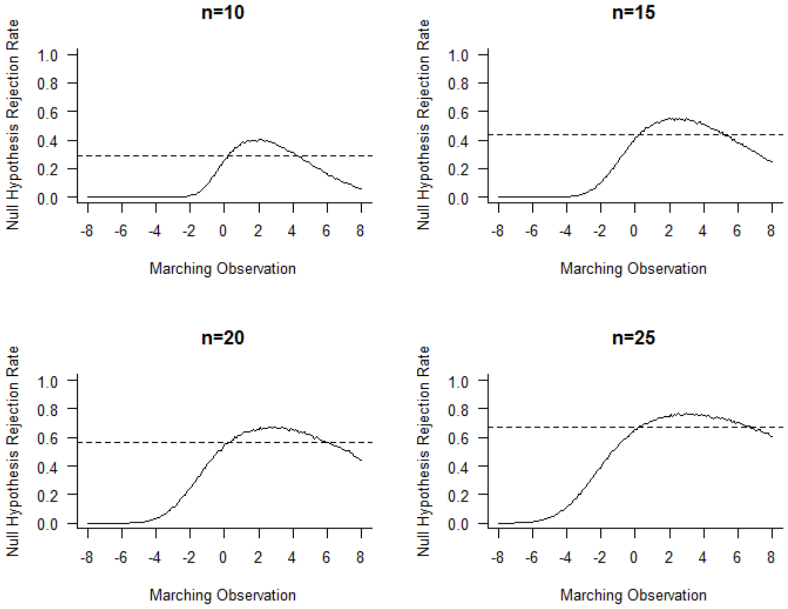


Figure 4: NHRR of the paired samples t-test, $d = 0.5$, two-sided

Figure 4 shows that for $x_n > d = 0.5$, increases in x_n are initially associated with an increase in power. This power increase relative to the expected power for each of the sample sizes is clear to see but might not be of great practical consequence. In addition, there is a noticeable turning point at which the power decreases as x_n further increases. For larger sample sizes, the paired samples t-test is relatively more robust to the presence of an extreme observation. For smaller sample sizes, the power reduction when an extreme observation is present is exacerbated. When the marching observation is in the opposite direction to the true effect, an increasingly large negative difference eliminates the effect under the stated conditions.

Figure 5 gives the NHRR of Yuen's paired samples t-test and Figure 6 gives the NHRR of the Wilcoxon signed rank sum test, both when $d = 0.5$. Under the same normality conditions, for $n = 10$, 15, 20 and 25, the corresponding power for the Wilcoxon signed rank sum test is 0.279, 0.419, 0.543, and 0.648 respectively, and the corresponding power for the Yuen paired samples t-test is 0.263, 0.356, 0.528, and 0.613 respectively. These reference lines are added to the graphic for comparative purposes.

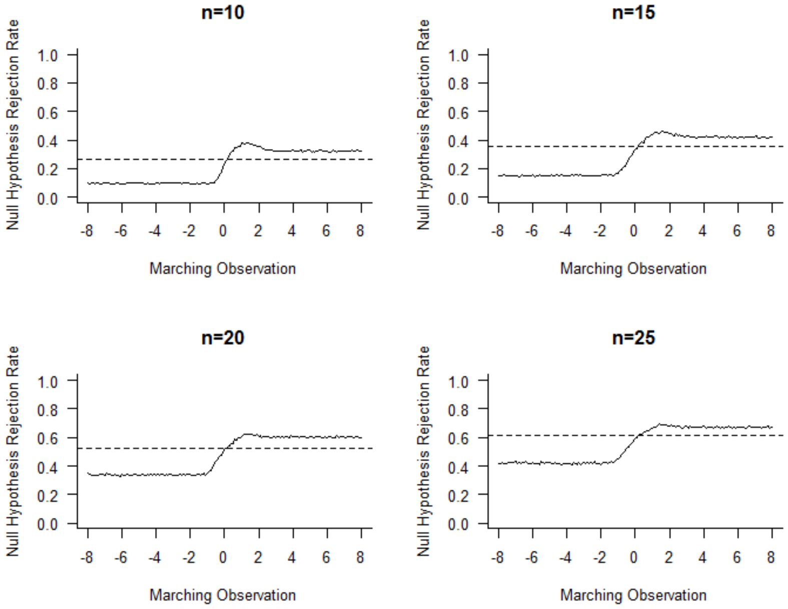


Figure 5: NHRR of Yuen's paired samples t-test, $d = 0.5$, two-sided

Figure 5 and 6 show that for $x_n > d = 0.5$, increases in x_n are associated with an increase in power relative to the expected power for each of the sample sizes, but the increase might not be of great practical consequence. For small samples, when the marching observation is in the opposite direction to the true effect, an increasingly large negative marching observation reduces the effect and this is seen in the reduced power.

The second simulation set-up is now considered. The condition that the sample mean differences are positive is removed, and a one-sided test using the upper tail of the distribution is performed. Figure 7 shows the impact of the marching observation for each of the three tests when the null hypothesis is true.

Figure 7 demonstrates that the patterns observed and identifiable conclusions for the two-sided tests are the same under these conditions. In fact, the impact of the marching observation in the second simulation set-up is qualitatively similar to the first simulation set-up. For brevity, the remaining graphics under this condition are not displayed.

5 Discussion

We have used a systematically increasing marching observation to demonstrate the impact on the Null Hypothesis Rejection Rate (NHRR) for the paired samples t-test, Yuen's

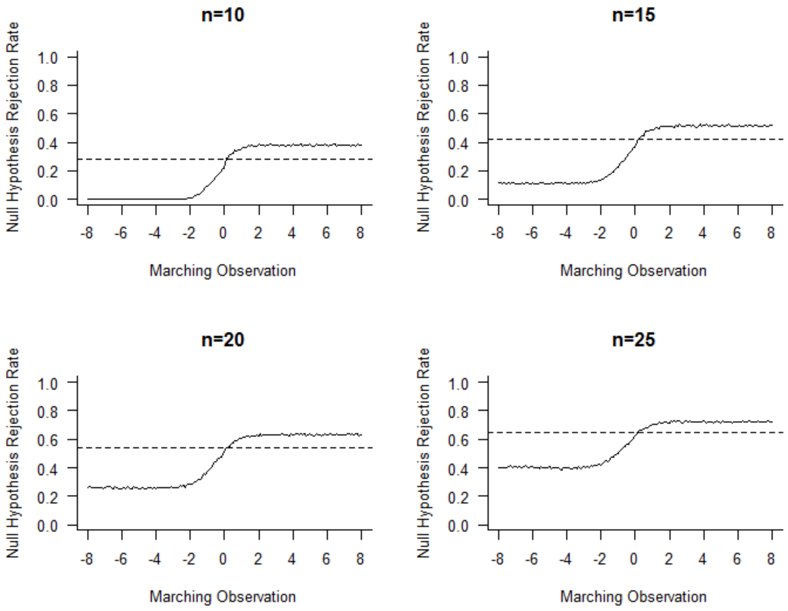


Figure 6: NHRR of the Wilcoxon test, $d = 0.5$, two-sided.

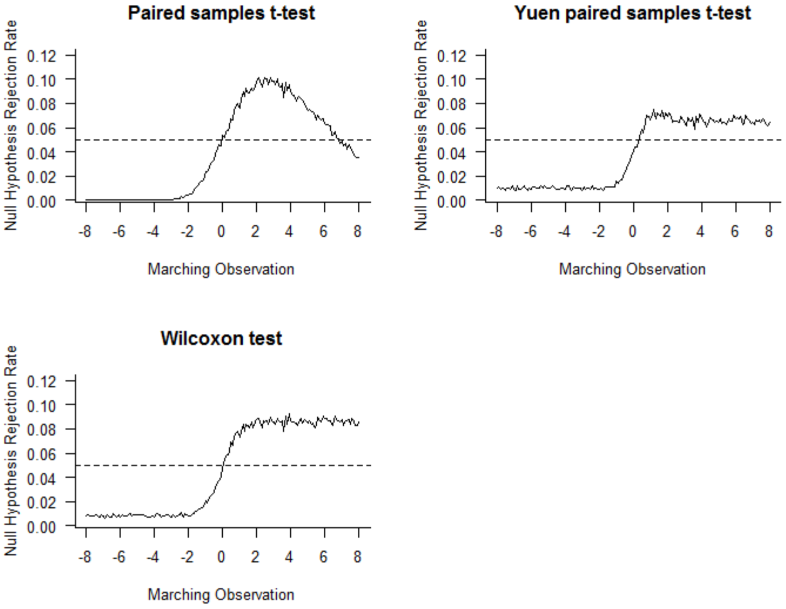


Figure 7: NHRR for each of the three tests when $n = 15$, $d = 0$, one sided

paired samples t-test, and the Wilcoxon signed rank sum test. This systematic approach, similar to one-factor at a time experimentation, would lend itself to other similar investigations e.g. two independent samples design, or to other single sample tests such as the single sample variance test, or be extended to investigations involving multiple marching observations. In practice, x_n and the condition $\bar{x}_{n-1} > 0$ may be independent and the condition $\bar{x}_{n-1} > 0$ is imposed to separate potential different behaviours of the tests statistics.

The mathematical exposition in Section 2 indicates that for a two sided paired samples t-test, a large observation either concordant or discordant with the rest of the sample will lead to a non-rejection of the null hypothesis. With the paired samples t-test the inclusion of a very large positive observation x_n into a sample with $\bar{x}_{n-1} > 0$ may in fact severely reduce the probability of rejecting the null hypothesis.

Simulations comprising normal deviates and in testing a nil-null hypothesis of no location effects have been performed. Stipulation of the condition $\bar{x}_{n-1} > 0$ does not invalidate the two-sided test procedure. However, the inclusion of a single, but often large discrepant observation, does imply that the nil-null hypothesis is not strictly true, hence our use of the terminology of the NHRR (the null hypothesis rejection rate), rather than using the terminology type I error rate.

For small sample sizes there is a paradox when performing the paired samples t-test that more extreme values of the marching observation in the direction of the sample mean difference result in a greater p-value than a less extreme value of the marching observation.

Under a location shift model, the inclusion of genuinely large positive observation x_n into a sample with $\bar{x}_{n-1}$ should lead to an increase in statistical power in a two-sided test of the nil-null hypothesis. This effect is observed with Yuen's paired samples t-test and with the Wilcoxon signed rank sum test, but it is not consistently observed with the paired samples t-test.

Under a location shift model, the inclusion of a large negative observation x_n into a sample with $\bar{x}_{n-1} > 0$ should lead to a relative decrease in statistical power. This effect is observed with Yuen's paired samples t-test and with the Wilcoxon signed rank sum test, but the effect is most evident, and is sample size dependent, for the paired samples t-test.

In summary, Yuen's paired samples t-test and the Wilcoxon signed rank sum test broadly display properties consistent with being robust statistical tests in the presence of a large outlier. In contrast the paired samples t-test displays behaviour strongly dependent on the magnitude of the outlier. Specifically, for small sample sizes the more extreme the values of the marching observation in the direction of the sample mean difference the greater the p-value compared to a less extreme value of the marching observation.

Zumbo and Jennings (2002), using their novel contamination model, concluded that the paired samples t-test had an inflated type I error rate with increasing asymmetric contamination, however our marching observation simulations indicate that the effect of a single outlier on this test is dependent on sample size, magnitude and direction of the outlier, and could lead to increases and decreases in the NHRR. It should be noted that the simulations of Zumbo and Jennings (2002) consisted of situations in which the underlying distributions were contaminated with outliers and simultaneously a true null hypothesis is maintained. In contrast our simulations are based on the fulfilment of correct assumptions prior to the inclusion of the marching observation.

Our simulations demonstrate the seemingly paradoxical effect of large outliers on the performance of the paired samples t-test, and although we concur with Zimmerman (2011) that rank based methods do not necessarily eliminate the influence of outliers, the simulations indicate that Yuen's paired samples t-test and the Wilcoxon signed rank sum test have robust behaviour in the presence of a single outlying observation.

In the preparation of this paper, methods for outlier detection in the conditions above were attempted, but we were unable to identify a suitable method. With reference to paired samples, Preece (1982) states that formal procedures for the detection and rejection of outliers are of negligible use for small sample sizes. Further debate and investigation into outlier detection methods offers an area for further research.

Acknowledgements

The authors thank the reviewers, and the editor, for their generous and insightful comments. Their valuable contributions have resulted in a significantly improved manuscript.

References

[1] Aguinis, H., Gottfredson, R. K., and Joo, H. (2013): Best-practice recommendations for defining, identifying, and handling outliers. *Organizational Research Methods*, **16**(2), 1–32.

[2] Blair, R. C., and Higgins, J. J. (1985): A comparison of the power of the paired samples rank transform statistic to that of Wilcoxon's signed ranks statistic. *Journal of Educational Statistics*, **10**(4), 368–383.

[3] Box, G. E., and Muller, M. E. (1958): A note on the generation of random normal deviates. *The Annals of Mathematical Statistics*, **29**(2), 610–611.

[4] Chaffin, W. W., and Rhiel, G. S. (1993): The effect of skewness and kurtosis on the one-sample t test and the impact of knowledge of the population standard deviation. *Journal of Computation and Simulation*, **46**, 79–90.

[5] Fradette, K., Keselman, H. J., Lix, L., Algina, J., and Wilcox, R. (2003): Conventional and robust paired and independent samples t-tests: Type I rrror and power rates. *Journal of Modern Applied Statistical Methods*, **2**(2), 481–496.

[6] Gibbons, J. D., and Chakraborti, S. (2011): Nonparametric statistical inference. In: M. Lovric (Ed): *International Encyclopedia of Statistical Science*, 977–979. Berlin: Springer.

[7] Herrendörfer, G., Rasch, D., and Feige, K. D. (1983): Robustness of statistical methods II. Methods of the one-sample problem. *Biometrical Journal*, **25**, 327–343.

[8] Levine, T. R., Weber, R., Hullett, C., Park, H. S., and Lindsey, L. L. M. (2008): A critical assessment of null hypothesis significance testing in quantitative communication research. *Human Communication Research*, **34**(2), 171–187.

[9] Posten, H. O. (1979): The robustness of the one-sample t-test over the Pearson system. *Journal of Statistical Computation and Simulation*, **6**, 133–149.

[10] Preece, D. A. (1982): T is for trouble (and textbooks): A critique of some examples of the paired-samples t-test. *The Statistician*, **31**(2), 169–195.

[11] R Core Team (2014): *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. https://www.r-project.org/.

[12] Rasch, D., and Guiard, V. (2004): The robustness of parametric statistical methods. *Psychology Science*, **46**, 175–208.

[13] Ringwalt, C., Paschall, M. J., Gorman, D., Derzon, J., and Kinlaw, A. (2010): The use of one-versus two-tailed tests to evaluate prevention programs. *Evaluation & the Health Professions* **34**(2), 135–150.

[14] Wilcox, R. R. (2005): *Introduction to robust estimation and hypothesis testing*. San diego, CA: Academic Press.

[15] Yuen, K. K. (1974): The two-sample trimmed t for unequal population variances. *Biometrika*, **61**, 165–170.

[16] Zimmerman, D. W. (1997): A note on the interpretation of the paired samples. *Journal of Educational and Behavioral Statistics*, **22**(3), 349–360.

[17] Zimmerman, D. W. (2011): Inheritance of properties of normal and non-normal distributions after transformation of scores to ranks. *Psicologica*, **32**(1), 65–85.

[18] Zumbo B. D., and Jennings, M. J. (2002): The robustness of validity and efficiency of the related samples t-test in the presence of outliers. *Psicologica*, **23**(2), 415–450.

Appendix P6

Derrick, B. et al. (2018). "Tests for equality of variances between two samples which contain both paired observations and independent observations". *Journal of Applied Quantitative Methods* 13 (2), pp. 36–47
Published Version

TESTS FOR EQUALITY OF VARIANCES BETWEEN TWO SAMPLES WHICH CONTAIN BOTH PAIRED OBSERVATIONS AND INDEPENDENT OBSERVATIONS

Ben DERRICK[1]
PhD Candidate
University of the West of England, Bristol

E-mail: Ben.Derrick@uwe.ac.uk

Annalise RUCK[2]
University of the West of England, Bristol

E-mail: Annalise.Ruck@officeforstudents.org.uk

Deirdre TOHER[3]
PhD, Senior Lecturer
University of the West of England, Bristol

E-mail: Deirdre.Toher@uwe.ac.uk

Paul WHITE[4]
PhD, Associate professor
University of the West of England, Bristol

E-mail: Paul.White@uwe.ac.uk

Abstract

Tests for equality of variances between two samples which contain both paired observations and independent observations are explored using simulation. New solutions which make use of all of the available data are put forward. These new approaches are compared against standard approaches that discard either the paired observations or the independent observations. The approaches are assessed under equal variances and unequal variances, for two samples taken from the same distribution. The results show that the newly proposed solutions offer Type I error robust alternatives for the comparison of variances, when both samples are taken from the same distribution.

Key words: Brown-Forsythe test; Equal variances; Partially overlapping samples; Pitman-Morgan test; Simulation; Robustness

1. Introduction

An equality of variances test is often performed as a preliminary test to inform the most appropriate statistical test for a comparison of means (Mirtagioğlu *et al.* 2017). The pitfalls of this process are well documented (Zimmerman, 2004; Zimmerman and Zumbo, 2009; Rasch *et al.*, 2011; Rochon *et al.*, 2012). This paper considers tests for equality of variances where it is the equality of variances that is of importance in their own right. Examples include a comparison of two treatments that have a similar mean efficacy, or a comparison of products in quality control, or a comparison of variances in human populations. Tests for equal variances have wide ranging applications including areas in archaeology, environmental science, business and medical research (Gastwirth *et al.*, 2009).

Numerous tests for the comparisons of variances for two independent samples have been documented (Conover, *et al.,* 1981). The Pitman-Morgan test is widely regarded as the optimum test of equal variances with two paired samples under normality (Mudholkar *et al.*, 2003). However, situations may arise where there are two samples which contain both independent observations and paired observations (Derrick *et al.,* 2015). For example, when some experimental data in a paired samples design is missing due to an error or accident.

This paper is concerned with the direct comparison of variances between two samples, which contain both paired observations and independent observations. For simplicity, these scenarios are referred to as partially overlapping samples (Martinez-Camblor *et al.,* 2013; Derrick *et al.*, 2017). The conditions of Missing Completely at Random (MCAR) are assumed.

In the two partially overlapping samples scenario, if the number of paired observations is relatively large and the number of independent observations is relatively small, a solution may be to discard independent observations and perform a test for equal variances on the paired observations. The standard F-test is not appropriate for paired samples (Kenny, 1953). For the comparison of variances for paired data, the Pitman-Morgan test can be performed (Pitman 1938; Morgan 1939). However, the Pitman-Morgan test is not robust to violations of the assumption of normality (Mudholkar *et al.*, 2003; Grambsch, 2015). For heavy tailed distributions the Type I error rate of the Pitman-Morgan test is larger than nominal Type I error rate (McCulloch, 1987; Wilcox, 2015).

Alternatively, if the number of independent observations is relatively large and the number of paired observations is relatively small, a solution may be to discard paired observations and perform one of numerous established tests for the comparison of variances with independent observations.

When the normality assumption is met, the standard F-test is the uniformly most powerful test for two independent samples. However, the standard F-test is not robust to deviations from normality (Marozzi, 2011).

Levene (1960) proposed that for two independent groups, the differences between the absolute deviations from the group means could be used to assess equality of variances. In the two sample case, this test is equivalent to Student's t-test applied to absolute deviations from the group means. This version of Levene's test, fails to control the Type I error rate when the population distribution is skewed (Carroll and Schneider, 1985; Nordstokke and Zumbo, 2007).

Brown and Forsythe (1974) proposed alternatives to Levene's test when data are not normally distributed. These alternatives use deviations from the median or trimmed mean. These variations are also often referred to as "Levene's test" (Carroll and Schneider, 1985; Gastwirth *et al.*, 2009). For the avoidance of doubt, in this paper the convention followed is that assessing equality of variances using deviations from the mean is referred to as Levene's test. Assessing equality of variances using deviations from the median is referred to as the Brown-Forsythe test.

Conover *et al.* (1981) explored 56 tests for equal variances for two independent groups and noted that the five tests that are Type I error robust use deviations from the median rather than deviations from the mean. Conover *et al.* (1981) found that the only test that consistently meets Bradley's (1978) liberal Type I error robustness criteria is the Brown-Forsythe test, using absolute deviations from the median. There is no uniformly robust and most powerful test applicable for all distributions and sample sizes. The general consensus is praise of the Brown-Forsythe test using deviations from the median (Carroll and Schneider, 1985; Nordstokke and Zumbo, 2007; Mirtagioğlu *et al.,* 2017). However, it should be noted that this test can be conservative with small sample sizes (Loh, 1987; Lim and Loh, 1995). The use of absolute deviations rather than squared deviations better maintains Type I error robustness (Cody and Smith, 1997).

Performing a test using either only the independent observations or only the paired observations may result in loss of power. The discarding of data is particularly problematic if the overall total sample size is small. In addition, if the assumption of MCAR is not reasonable, the discarding of data is likely to cause bias.

Bhoj (1979, 1984) and Ekbohm (1981, 1982) debated methods using all of the available data for testing the equality of variances in scenarios that they refer to as "incomplete data". In this debate the authors do not recognise that a combination of independent observations and paired observations may occur by design and not only by accident. Bhoj (1979) and Ekbohm (1981, 1982) independently considered a weighted combination of existing independent sum of squares techniques to create a new test statistic. Other solutions such as ignoring the pairing and performing the F-test on all of the available data were considered by Ekbohm (1982). Bhoj (1984) concluded that his test statistic is the most powerful if the correlation is negative or small. Otherwise, performing the F-test on all of the available data is more powerful than the solutions put forward by either of the authors (Ekbolm, 1982; Bhoj 1984). The simulations performed by these authors were on a relatively small scale, with only 1,000 replicates at each point in their design space. No solution was comprehensively agreed upon for all scenarios, and this is likely to contribute to them not being well established. Furthermore the non-robustness of the Pitman-Morgan test has a detrimental impact on their weighted tests. A solution that uses all available data without a complex weighting structure, or the discarding of valuable information about the pairing, may therefore be advantageous.

For the comparison of means when both independent observations and paired observations are present, partially overlapping samples t-tests are given by Derrick, *et al.* (2017). These solutions are generalised forms of the t-test and are Type I error robust under normality. These solutions are also robust in the comparison of two ordinal samples where the scale represents interval data (Derrick and White, 2018).

We propose that as an alternative test of equal variances when there is a combination of paired observations and independent observations, the partially overlapping samples t-test can be performed, using deviations from the group medians, as outlined below.

Let X_{ji} denote the *i*-th observation in group *j* for *j* = {Sample1, Sample 2}, and $\tilde{X}_j$ denote the sample median, so that $Y_{ji} = \left| X_{ji} - \tilde{X}_j \right|$, then

$$T_{\text{var}1} = \frac{\bar{Y}_1 - \bar{Y}_2}{S_{p\_y}\sqrt{\frac{1}{n_1} + \frac{1}{n_2} - 2r\left(\frac{n_c}{n_1 n_2}\right)}} \quad \text{and} \quad S_{p\_y} = \sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{(n_1-1)+(n_2-1)}}$$

The test statistic $T_{\text{var}1}$ is referenced against the t-distribution with degrees of freedom:

$$v_1 = (n_c - 1) + \left(\frac{n_a + n_b + n_c - 1}{n_a + n_b + 2n_c}\right)(n_a + n_b).$$

where $n_a =$ number of unpaired observations exclusive to Sample 1, $n_b =$ number of unpaired observations exclusive to Sample 2, $n_c =$ number of pairs, $n_j =$ total number of observations in Sample *j*, $S_j^2 =$ variance of Sample *j* based on the Y_{ji} observations.

For the comparison of variances, Loh (1987) suggested adapting the unequal variances t-test using deviations from the medians. For the comparison of means, Student's t-test is sensitive to deviations from the equal variances assumption (Ruxton, 2006; Derrick, Toher and White, 2016). As a result of this Derrick *et al.* (2017) additionally proposed the partially overlapping samples t-test for unequal variances. We propose that the partially overlapping samples test statistic unconstrained to equal variances can be similarly modified to provide a test for equality of variances so that:

$$T_{\text{var}2} = \frac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2} - 2r\left(\frac{S_1 S_2 n_c}{n_1 n_2}\right)}}$$

The test statistic $T_{\text{var}2}$ is referenced against the t-distribution with degrees of freedom:

$$v_2 = (n_c - 1) + \left(\frac{\gamma - n_c + 1}{n_a + n_b + 2n_c}\right)(n_a + n_b) \text{ where } \gamma = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{\left(S_1^2/n_1\right)^2}{n_1 - 1} + \frac{\left(S_2^2/n_2\right)^2}{n_2 - 1}}$$

Methodology for assessing the Type I error rate of these proposals is given in Section 2, with an example application given in Section 3.

2. Methodology

For two samples containing both independent observations and paired observations, approaches for the comparison of variances are assessed using simulation. The approaches considered are the Brown-Forsythe test, the Pitman-Morgan test, and the proposed $T_{\text{var}1}$ and $T_{\text{var}2}$. Type I error robustness is assessed using Bradley's (1978) liberal robustness criteria. Power is assessed for test statistics that do not violate Bradley's liberal criteria.

Within the simulation design, the sizes of n_a, n_b, n_c are {5, 10, 30, 50}. The correlation coefficients ρ are {0.00, 0.25, 0.50, 0.75}. Simulations for each possible parameter combination of n_a, n_b, n_c, ρ are performed in a factorial design. Standard Normal deviates are calculated using the Box-Muller (1958) transformation. For the n_c observations, correlated Standard Normal deviates are obtained as per Kenney and Keeping (1951)

In Section 4.1, the comparison of variances is performed for normally distributed data. Under the null hypothesis, X_1 ~ N(0,1) and X_2 ~ N(0,1). Under the alternative hypothesis, the observations in Sample 2 are multiplied by two, thus X_1 ~ N(0,1) and X_2 ~ N(0,4).

In Section 4.2, the comparison of variances is performed for skewed distributions. Under the null hypothesis, Normal deviates are first generated as above, and then the exponential of each value is calculated. Under the alternative hypothesis this process is repeated, and each of the observations in Sample 2 are multiplied by two to create unequal variances.

For each parameter combination, the data generating process is repeated 10,000 times, and each of the statistical tests to be evaluated is performed on each replicate. Under the null hypothesis, the proportion of the replicates where the null hypothesis is rejected represents the Type I error rate. Under the alternative hypothesis, the proportion of the replicates where the null hypothesis is rejected, represents the power of the test, assuming Type I error rates can be reasonably compared. The simulations and tests are performed in R, at the 5% significance level, two-sided.

The simulation design allows that the conditions of MCAR can be assumed.

3. Example

In the assessment of an undergraduate university module, two lecturers share the marking of 32 student submissions. As part of the marking regulations, at random six of the submissions are independently assessed by both lecturers. The remaining submissions are randomly split between the two lecturers, ensuring that both have an equal number to assess. Thus Lecturer 1 has one sample comprising of six paired observations and 13 independent observations. Likewise, Lecturer 2 has a sample of equal size. The samples are partially overlapping by design, thus MCAR can be reasonably assumed.

There is concern that the lecturers do not allocate marks at the top end and the bottom end of the marking scale in the same way. Tests for equal variances are performed on the independent observations (Table 1), the paired observations (Table 2), and all observations.

Table 1. Marks awarded to the 26 students randomly allocated to the lecturers.

Lecturer 1	55	56	58	60	60	60	61	61	62	62	64	65	67
Lecturer 2	40	50	51	60	60	60	60	60	61	66	69	72	82

Table 2. Marks awarded by each lecturer for the six students that are marked by both.

Student	A	B	C	D	E	F
Lecturer 1	54	55	60	63	65	70
Lecturer 2	50	56	60	61	67	73

The Brown-Forsythe test is performed on the data in Table 1 using the R package “lawstat” (Gastwirth *et al.*, 2015). This shows no evidence to reject the null hypothesis of equal variances (t = -1.9673, v = 24, p = 0.061).

The Pitman-Morgan test is performed on the data in Table 2 using the R package “PairedData” (Champely, 2013). This shows no evidence to reject the null hypothesis of equal variances (t = -2.352, v = 4, p = 0.078).

In order to perform the tests for equal variances using all of the available data, for each submission marked my Lecturer 1 the absolute deviation from the median mark given by Lecturer 1 is calculated. Similarly, the absolute deviations for Lecturer 2 are calculated.

The partially overlapping samples t-test is performed on the absolute deviations using the R package "Partiallyoverlapping" (Derrick, 2017). The null hypothesis of equal variances is rejected at the 5% significance level for both the equal variances assumed variant (t_{var1} = -2.324, v_1 = 26.211, p = 0.028) and the equal variances not assumed variant (t_{var2} = -2.183, v_2 = 17.488, p = 0.043). It would appear that Lecturer 2 is making greater use of the full range of potential marks relative to Lecturer 1.

3.1. Comparison of variances for two samples from the Normal distribution

Type I error rates and power are summarised for each of; the Brown-Forsythe test, BF, the Pitman-Morgan test, PM, and the partially overlapping samples tests, T_{var1} and T_{var2}. Each of the test statistics are assessed under the null hypothesis where X_1 ~ N (0,1) and X_2 ~ N (0,1). The Type I error robustness for each of the parameter combinations within the simulation design are summarised in Figure 1.

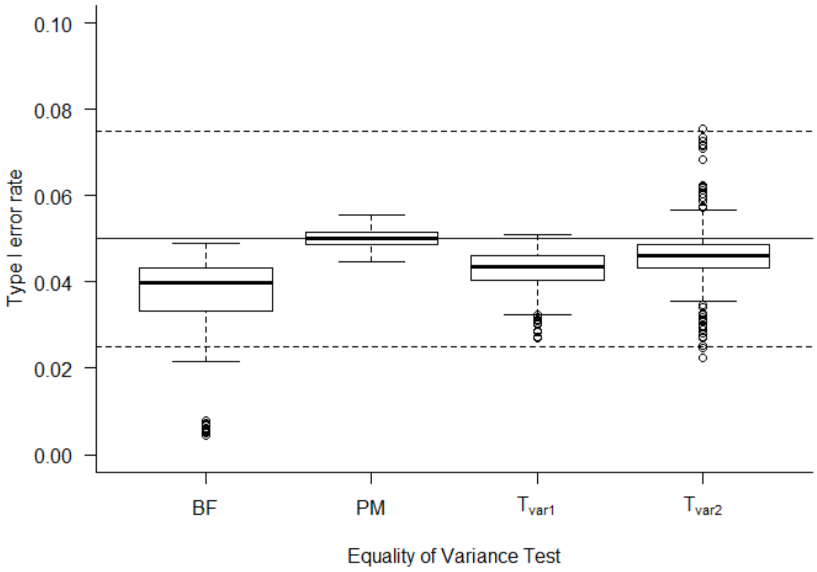


Figure 1. Type I error robustness for each parameter combination, assessed against Bradley's liberal criteria, samples from Standard Normal distribution

Figure 1 shows that the Pitman-Morgan test and the proposed test statistics are Type I error robust throughout the simulation design, with T_{var1} being more conservative than T_{var2}. For the smallest sample sizes within the design, the Brown-Forsyth test is very conservative.

Relative power comparisons for each of the test statistics are assessed where X_1 ~ N (0,1) and X_2 ~ N (0,4). The power averaged across the simulation design for increasing ρ is given in Figure 2.

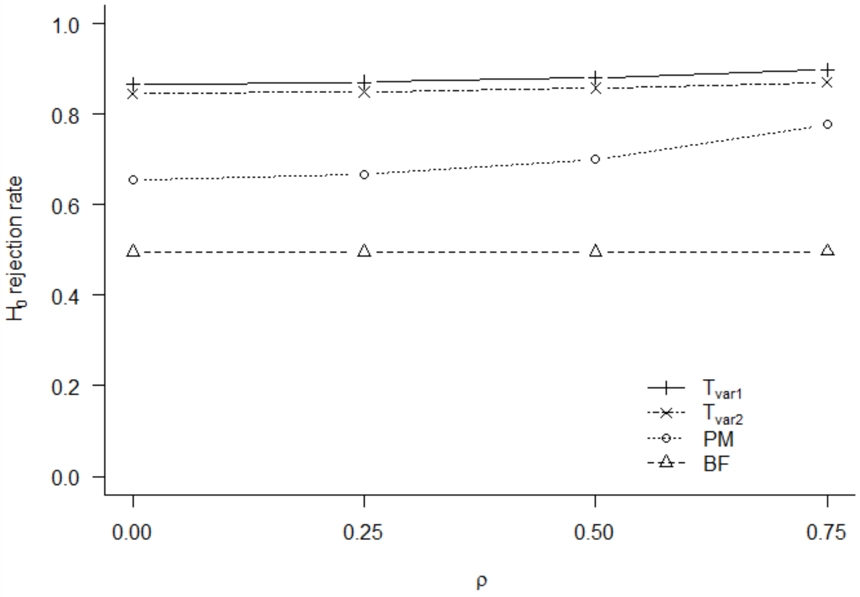


Figure 2. Relative power, averaged across the simulation design for increasing ρ, samples from Normal distributions.

Figure 2 shows that the proposed test statistics T_{var1} and T_{var2} perform similarly to each other under normality, and they have superior power qualities to the standard tests which discard data.

3.2. Comparison of variances for two samples from skewed distributions

Each of the test statistics are assessed when both samples are taken from skewed but identical distributions. The Type I error robustness for each of the parameter combinations within the simulation design are summarised in Figure 3.

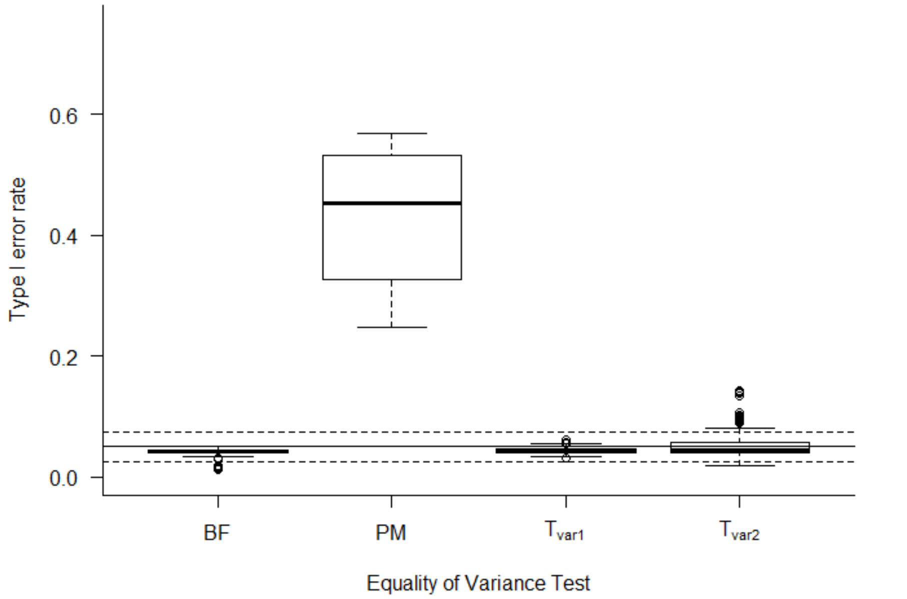


Figure 3. Type I error robustness for each parameter combination, assessed against Bradley's liberal criteria, samples from skewed distribution.

Figure 3 shows that the Pitman-Morgan test is not Type I error robust when the samples are taken from identical heavy tailed distributions. This supports the findings by McCulloch (1987) and Wilcox (2015). In addition it can be seen that T_{var2} does not fully maintain Type I error robustness. Further investigation shows that T_{var2} is liberal when one of the samples is more dominant in terms of size, and when there is a large imbalance between the number of independent observations and the number of pairs.

Relative power comparisons for each of the test statistics are assessed where the samples are taken from different skewed distributions. Due to the poor Type I error robustness of the Pitman-Morgan test and T_{var2}, this comparison is done only for the Brown-Forsythe test and T_{var1}. The power averaged across the simulation design for increasing ρ is given in Figure 4.

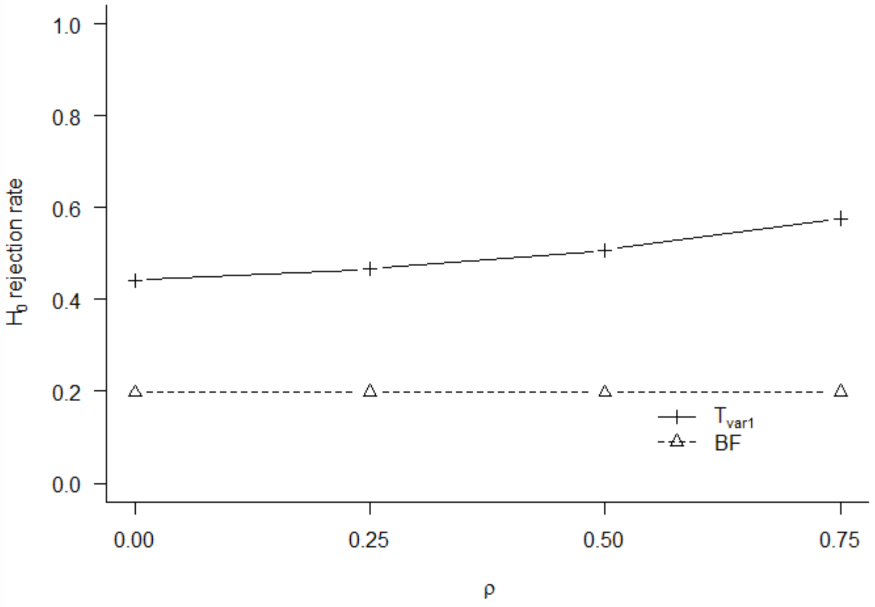


Figure 4. Relative power, averaged across the simulation design for increasing ρ, samples from skewed distributions.

Figure 4 shows that the proposed solution, T_{var1}, is more powerful than the Brown-Forsythe test. A comparison of Figure 4 against Figure 2 also indicates that both the Brown-Forsythe test and the newly proposed test, T_{var1}, are less powerful when samples are taken from a heavy-tailed distribution.

4. Conclusion

A common research question in psychology, education, medical sciences, business and manufacturing, is whether or not the variances are equal (Gastwirth, Gel and Miao, 2009).

There has been little research into techniques for the comparison of variances for samples that contain both independent observations and paired observations. Standard solutions that involve discarding data are less than desirable. Two solutions that make use of the tests statistics by Derrick *et al.* (2017) are proposed in this paper. Simulations across a range of sample sizes show that these solutions are Type I error robust under normality and the assumption of MCAR. These solutions are more powerful than established solutions that discard data, namely the Pitman-Morgan test and the Brown-Forsythe test.

The equal variances form of the partially overlapping samples variances test, T_{var1}, is marginally more powerful than the unconstrained form of the test T_{var2}.

The proposed test statistic T_{var1} further maintains Type I error robustness for skewed distributions where T_{var2} does not. T_{var1} is therefore recommended as a powerful alternative to test for the equality of variances between two samples when there is a combination of paired observations and independent observations in two samples.

References

1. Bhoj, D. S. **Testing equality of variances of correlated variates with incomplete data on both responses.** Biometrika. Vol. *66,* No. 3, 1979, pp. 681–683. doi: 10.1093/biomet/66.3.681
2. Bhoj, D. S. **On testing equality of variances of correlated variates with incomplete data.** Biometrika. Vol. 71, No. 3, 1984, pp. 639–641. doi: 10.1093/biomet/71.3.639
3. Bradley, J. V. **Robustness?**. British Journal of Mathematical and Statistical Psychology, Vol. 31, No. 2, 1978, pp. 144-152. doi: 10.1111/j.2044-8317.1978.tb00581.x
4. Brown, M. B., and Forsythe, A. B. **Robust tests for the equality of variances.** Journal of the American Statistical Association, Vol. 69, No. 346, 1974, pp. 364-367. doi: 10.1080/01621459.1974.10482955
5. Carroll, R. J., and Schneider, H. **A note on Levene's tests for equality of variances.** Statistics and probability letters, Vol. 3, No. 4, 1985, pp. 191-194. doi: 10.1016/0167-7152(85)90016-1
6. Champely, S. **PairedData: Paired Data Analysis.** R package version 1.0.1., 2013
7. Conover, W. J., Johnson, M. E., and Johnson, M. M. **A comparative study of tests for homogeneity of variances, with applications to the outer continental shelf bidding data.** Technometrics, Vol. 23, No. 4, 1981, pp. 351-361. doi: 10.1080/00401706.1981.10487680
8. Cody, R. P., and Smith, J. K. **Applied statistics and the SAS programming language.** Elsevier North-Holland Inc, 1986
9. Derrick, B. **Partiallyoverlapping: Partially Overlapping Samples t-Tests.** R package version 1.0., 2017
10. Derrick, B., Dobson-McKittrick, A., Toher, D., and White P. **Test statistics for comparing two proportions with partially overlapping samples.** Journal of Applied Quantitative Methods, Vol. 10, No. 3, 2015, pp. 1-14.
11. Derrick, B., Russ, B., Toher, D., and White P. **Test statistics for the comparison of means for two samples which include both paired observations and independent observations.** Journal of Modern Applied Statistical Methods, Vol. 16, No. 1, 2017, pp. 137-157. doi: 10.22237/jmasm/1493597280
12. Derrick, B., Toher, D., and White, P. **Why Welch's test is Type I error robust.** The Quantitative Methods for Psychology, Vol. *12,* No. 1, 2016, pp. 30-38. doi: 10.20982/tqmp.12.1.p030

13. Derrick, B., and White, P. **Methods for comparing the responses from a Likert question, with paired observations and independent observations in each of two samples.** International Journal of Mathematics and Statistics, Vol. 19, No. 3, 2018, pp. 71-83.
14. Ekbohm, G. **A test for the equality of variances in the paired case with incomplete data.** Biometrical Journal. Vol. 23, No. 3, 1981, pp. 261–265. doi: 10.1002/bimj.4710230306
15. Ekbohm, G. **On comparing variances in the paired case with incomplete data.** Biometrika. Vol. 69, No. 3, 1982, pp. 670–673. doi: 10.1093/biomet/69.3.670
16. Gatwirth, J. L., Gel, Y. R., and Miao, W. **The Impact of Levene's Test of Equality of Variances on Statistical Theory and Practice.** Statistical Science, Vol. 24, No. 3, 2009, pp. 343-360.
17. Gatwirth, J. L., Gel, Y. R., Wallace-Hui, W. L., Lyubchich, V., Miao, W., and Noguchi, K. **lawstat: Tools for Biostatistics, Public Policy, and Law.** R package version 3.0., 2015
18. Grambsch, P. M. **Simple robust tests for scale differences in paired data.** Biometrika, Vol. 81, No. 2, 1994, pp. 359-372.
19. Kenney, J. F., and Keeping, E. S. **Mathematics of statistics,** Princeton, NJ: Van Nostrand, Part. 2, 2nd edition., 1951
20. Kenny, D. T. **Testing of differences between variances based on correlated variates.** Canadian Journal of Experimental Psychology, Vol. 7, No. 1, 1953, pp. 25. doi: 10.1037/h0083569
21. Levene, H. **Robust tests for equality of variances.** Contributions to Probability and Statistics, Vol. 1, 1960, pp. 278-292.
22. Lim, T. S., and Loh, W. Y. **A comparison of tests of equality of variances.** Computational Statistics and Data Analysis, Vol. 22, No. 3, 1996, pp. 287-301.
23. Loh, W. Y. **Some modifications of Levene's test of variance homogeneity.** Journal of Statistical Computation and Simulation, Vol. 28, No. 3, 1987, pp. 213-226.
24. Marozzi, M. **Levene type tests for the ratio of two scales.** Journal of Statistical Computation and Simulation, Vol. 81, No. 7, 2011, pp. 815-826. doi: 10.1080/00949650903499321
25. Martínez-Camblor, P., Corral, N., and María de la Hera, J. **Hypothesis test for paired samples in the presence of missing data.** Journal of Applied Statistics, Vol. 40, No. 1, 2013, pp. 76-87. doi: 10.1080/02664763.2012.734795
26. McCulloch, C. E. **Tests for equality of variances with paired data.** Communications in Sstatistics - Theory and Methods, Vol. 16, No. 5, 1987, pp. 1377-1391. doi: 10.1080/03610928708829445
27. Mirtagioğlu, H., Yiğit, S., Mendeş, E., and Mendeş, M. **Monte Carlo Simulation Study for Comparing Performances of Some Homogeneity of Variances Tests.** Journal of Applied Quantitative Methods, Vol. 12, No. 1, 2017, pp. 1-11.
28. Morgan, W. A. **A test for the significance of the difference between the two variances in a sample from a normal bivariate population.** Biometrika, Vol. 31, No. 1/2, 1939, pp. 13-19.

29. Mudholkar, G. S., Wilding, G. E., and Mietlowski, W. L. **Robustness properties of the Pitman–Morgan test.** Communications in Statistics - Theory and Methods, Vol. 32, No. 9, 2003, pp. 1801-1816. doi: 10.1081/STA-120022710
30. Pitman, E. J. G. **Significance tests which may be applied to samples from any populations: III. The analysis of variance test.** Biometrika, Vol. 29, No. 3/4, 1938, pp. 322-335.
31. Rasch, D., Kubinger, K. D., and Moder, K. **The two sample t-test: pre-testing its assumptions does not pay off.** Statistical Papers, Vol. 52, No. 1, 2011, pp. 219-231. doi: 10.1007/s00362-009-0224-x
32. Rochon, J., Gondan, M., and Kieser, M. **To test or not to test: Preliminary assessment of normality when comparing two independent samples.** BMC Medical Research Methodology, Vol. 12, 2012, pp. 81 doi: 10.1186/1471-2288-12-81
33. Wilcox, R. **Comparing the variances of two dependent variables.** Journal of Statistical Distributions and Applications, Vol. 2, No. 1, 2015, pp. 1-8. doi: 10.1186/s40488-015-0030-z
34. Zimmerman, D. W. **A note on preliminary tests of equality of variances.** British Journal of Mathematical and Statistical Psychology, Vol. 57, No. 1, 2004, pp. 173-181. doi: 10.1348/000711004849222
35. Zimmerman, D. W., and Zumbo, B. D. **Hazards in choosing between pooled and separate-variances t tests.** Psicológica: Revista de metodología y psicología experimental, Vol. 30, No. 2, 2009, pp. 371-390.
36. Zimmerman, D. W., and Zumbo, B. D. **The relative power of parametric and non-parametric statistical methods.** in Keren, G. and Lewis, C. (eds.) "A handbook for data analysis in the behavioral sciences: Methodological issues", Hillsdale, NJ: Erlbaum, 1993, pp. 481-517.

[1] Ben is a lecturer in the Applied Statistics Group at the University of the West of England, Bristol. He is a fellow of the Higher Education Academy; and is currently undertaking doctoral work on the robustness of the test statistics.

[2] Annalise is an analyst at the Office for Students; an associate member of the Applied Statistics Group at the University of the West of England, Bristol; and a member of the Royal Statistical Society. Annalise graduated from the University of the West of England, Bristol with a first class honours degree in Mathematics, winning the Richard Margetts Memorial prize for the best final year mathematics project.

[3] Deirdre is a senior lecturer in Applied Statistics Group at the University of the West of England, Bristol; and is a statistical ambassador for the Royal Statistical Society.

[4] Paul is an associate professor and director of the Applied Statistics Group at University of the West of England, Bristol. He has provided statistical analyses in over 100 publications including areas in statistics, economics, psychology, health, and the bio- and medical sciences.

Appendix P7

Derrick, B., White, P., and Toher, D. (in press). “Parametric and non-parametric tests for the comparison of two samples which both include paired and unpaired observations”. *Journal of Modern Applied Statistical Methods*

Accepted Version

Parametric and non-parametric tests for the comparison of two samples which both include paired and unpaired observations.

Introduction

Basic teaching of statistics usually assumes a perfect world with completely independent samples or completely dependent samples. Real world study designs and associated analyses are often far from these simplistic ideals. There are occasions where there are a combination of paired observations and independent observations within a sample. These scenarios are referred to as ‘partially overlapping samples’ (Martinez-Camblor *et.al.*, 2012, Derrick *et.al.*, 2015; Derrick *et.al.*, 2017). Other terminology for the described scenario is ‘partially paired data’ (Samawi & Vogel, 2011; Guo & Yuan, 2017). However, this terminology can be misconstrued as referring to pairs that are not directly matched (Derrick *et.al.*, 2015).

A typical partially overlapping samples scenario is a design which includes both paired observations and unpaired observations due to limited resource of paired samples. When a resource is scarce, researchers may only be able to obtain a limited number of paired observations, but would want to avoid wastage and also make use of the independent observations. For example, in a clinical trial by Hosgood *et.al.*, (2017) assessing the performance of kidneys following transplantation, one group incorporates a new technique that reconditions the kidney prior to the transplant, and one group is the control group of standard cold storage. When the kidneys arrive at the transplanting centre in pairs, one is randomly allocated to each of the two groups. When a single kidney arrives at the transplanting centre, this is randomly allocated to one of the two groups in a 1:1 ratio.

A commonly encountered partially overlapping samples problem is a paired samples design which inadvertently contains independent observations (Martinez-Camblor *et.al.*, 2012; Guo & Yuan, 2017). In these circumstances the reason for the missing data should be considered carefully. Solutions proposed within the current paper do not detract from extensive literature on missing data and solutions herein are assessed under the assumption of data missing completely at random (MCAR).

A naive approach often taken when confronted with scenarios similar to the above is to discard observations and perform a basic parametric test (Guo & Yuan, 2017). Naive parametric methods for the analysis of partially overlapping samples used as standard include; i) Discard the unpaired observations and perform the paired samples t-test, T_1; ii) Discard the paired observations and perform the independent samples t-test assuming equal variances, T_2; iii) Discard the paired observations and perform the independent samples t-test not assuming equal variances, T_3.

When the omission of the paired observations or independent observations does not result in a small sample size, traditional methods may maintain adequate power (Derrick *et.al.*, 2015). However, the discarding of observations is particularly problematic when the available sample size is small (Derrick, Toher and White, 2017). Other naive approaches include treating all the observations as unpaired, or randomly pairing data (Guo & Yuan, 2017). These approaches fail to maintain the structure of the original data and introduce bias (Derrick *et.al.*, 2017).

Amro and Pauly (2017) define three categories of solution to the partially overlapping samples problem that use all available data and do not rely on resampling methods. The categories are; tests based on maximum likelihood estimators, weighted combination tests, and tests based on a simple mean difference. Early literature on the partially overlapping samples framework focused on maximum likelihood estimators when data are missing by accident. Guo and Yuan (2017) reviewed parametric solutions under the condition of normality, and recommend the Lin and Strivers (1974) maximum likelihood approach when the normality assumption is met. However, Amro and Pauly (2017) demonstrate that this maximum likelihood estimator approach has an inflated Type I error rate under normality and non-normality. Furthermore, maximum likelihood proposals are complex mathematical procedures, which would be a barrier to some analysts in a practical setting. Thus these are not considered further in this paper.

A weighted combination based approach is to obtain the p-values for T_1 and T_2 as defined above, then combine them using the weighted z-test (Stouffer *et.al.*, 1949), or the generalised Fisher test proposed by Lancaster (1961). When used to combine p-values from independent tests, the latter method is more powerful (Chen, 2011). A procedure specifically attempting to act as a weighting between the paired samples t-test and the independent samples t-test under normality was proposed by Bhoj (1978). Uddin and Hasan (2017) optimised the weighting constants used by Bhoj (1978) so that the combined variance of the two elements minimized. Further weighted combination tests are proposed by Kim *et.al.* (2005), Samawi and Vogel (2011), and Martinez-Camblor *et.al.* (2012). All of these weighting based approaches have issues with respect to the interpretation of the results. The mathematical formulation of the statistics does not have a numerator that is equivalent to the difference in the two means. Neither do these proposals have a denominator that represents the standard error of the difference in two sample means, therefore confidence intervals for mean differences are not easily formed. Thus these are not considered further in this paper.

Looney and Jones (2003) put forward a parametric solution using all of the available data that does not rely on a complex weighting structure and is regarded as a simple mean difference estimator. However, several issues with the test have been identified and their solution is not Type I error robust under normality (Mehrotra, 2004; Derrick *et.al.*, 2017). A correction to the test by Looney and Jones (2003) is provided by Uddin and Hasan (2017), however the test statistic is a minor adjustment, and also makes reference to the z-distribution.

For the partially overlapping two group situation, two parametric solutions that are Type I error robust under the assumptions of normality and MCAR are given by Derrick *et.al.* (2017). These solutions are simple mean difference estimators and act as an interpolation between, firstly T_1 and T_2, or secondly between T_1 and T_3. These solutions are referred to as the partially overlapping samples t-tests. The authors noted that their parametric partially overlapping samples t-tests can be readily developed to obtain non-parametric alternatives.

Naive non-parametric tests for the analysis of partially overlapping samples include; i) Discard the paired observations and perform the Mann-Whitney-Wilcoxon test, MW; ii) Discard the unpaired observations and perform the Wilcoxon Signed Rank test, W.

In a comparison of samples from two identical non-normal distributions, non-parametric tests are often more Type I error robust than their parametric equivalents (Zimmerman, 2004). For skewed distributions with equal variances, the MW test is the most powerful Type I error robust test when compared against T_2 and T_3 (Fagerland & Sandvik, 2009a).

These traditional non-parametric tests provide low power when the discarding of observations result in a small sample size. For very small samples MW will only detect differences when a very large effect size is present (Fay & Proschan, 2010). The normality assumption is often hard to ascertain for small samples, thus non-parametric solutions that take into account all of the available data would be beneficial.

In textbooks by Mendenhall, Beaver and Beaver (2008) and Howell (2012), the null hypothesis of the MW test is reported as the distributions are equal. Fagerland and Sandvik (2009b) assert that the null hypothesis is more correctly reported as Prob(X > Y) = 0.5. For a comparison of two distributions, it is possible that the latter null hypothesis is true, but for the samples to be from distributions of different shape. When the distributions are equal other than in central location, the MW test can be considered as a comparison of central location (Skovlund & Fenstad, 2001). The MW test is not recommended as a test for location shift when variances are not equal (Zimmerman 1987; Penfield, 1994; Moser & Stevens, 1989). Ultimately, the MW test can detect differences in the shape of the two sample distributions, or their medians, or their means (Hart, 2001).

When there are three or more groups with both paired observations and independent observations, a possible non-parametric approach is the Skillings-Mack test (Skillings & Mack, 1981). This test is equivalent to the Freidman test when data are balanced (Chatfield & Mander, 2009). For an unbalanced design the Skillings-Mack test requires that any block with only one observation is removed. The Skillings-Mack test therefore cannot be used in the two group situation. This gives further motivation for the development of non-parametric tests for the two sample scenario.

In this paper, non-parametric solutions to the partially overlapping samples problem are considered, under normality and non-normality. This comparison includes a recent parametric solution proposed by Derrick *et.al.* (2017) for comparative purposes. The parametric solutions by Derrick *et.al.* (2017) and newly proposed non-parametric solution are defined, and methodology for comparing the Type I error robustness and power of the solutions is given. Results of the simulations for Normal and non-normal distributions are then considered, followed by a practical example incorporating the techniques explored.

Solutions to the partially overlapping samples problem

Parametric test statistics for the comparison of equal means in the presence of partially overlapping samples are taken from Derrick *et.al.* (2017). Proposed non-parametric solutions derived using the ranks of the actual values within the partially overlapping samples t-test procedure are then introduced. In line with Derrick *et.al.* (2015) who derived solutions for two partially overlapping samples of a dichotomous variable, the standard error of the partially overlapping samples tests is derived as the difference between two random variables.

Parametric solutions

Without loss of generality let $\overline{X}_1$ = mean of Sample 1, $\overline{X}_2 =$ mean of Sample 2, $n_a =$ number of unpaired observations exclusive to Sample 1, $n_b =$ number of unpaired observations exclusive to Sample 2, $n_c =$ number of pairs, $n_1 =$ number of observations in Sample 1 (i.e. $n_1 = n_a + n_c$), $n_2 =$ number of observations in Sample 2 (i.e. $n_2 = n_b + n_c$), $S_1^2 =$ variance of Sample 1, $S_2^2 =$ variance of Sample 2, r = Pearson's correlation coefficient for the n_c observations. All variances above are calculated using Bessel's correction as per Kenney & Keeping (1951).

The parametric partially overlapping samples test statistic, T_{new1}, is an interpolation between the paired samples t-test, T_1, and the independent samples t-test assuming equal variances, T_2, defined as:

$$T_{\text{new1}} = \frac{\overline{\mathrm{X}}_1 - \overline{\mathrm{X}}_2}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2} - 2r\left(\frac{n_c}{n_1 n_2}\right)}} \quad \text{where } S_p = \sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{(n_1-1)+(n_2-1)}}$$

The test statistic T_{new1} is referenced against the t-distribution with degrees of freedom:

$$v_1 = (n_c - 1) + \left(\frac{n_a + n_b + n_c - 1}{n_a + n_b + 2n_c}\right)(n_a + n_b).$$

For normally distributed data, the independent samples t-test is sensitive to deviations from the equal variances assumption. If equal variances cannot be assumed then Welch's test is a Type I error robust alternative under normality (Ruxton, 2006; Derrick, Toher & White, 2016). It follows that T_{new1} is also sensitive to deviations from the equal variances assumption (Derrick *et.al.*, 2017). The partially overlapping samples test statistic when the comparison is not constrained to equal variances, T_{new2}, is an interpolation between the paired samples t-test, T_1, and Welch's test, T_3, defined as:

$$T_{\text{new2}} = \frac{\overline{\text{X}}_1 - \overline{\text{X}}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2} - 2r\left(\frac{S_1 S_2 n_c}{n_1 n_2}\right)}}$$

The test statistic T_{new2} is referenced against the t-distribution with degrees of freedom:

$$v_2 = (n_c - 1) + \left(\frac{\gamma - n_c + 1}{n_a + n_b + 2n_c}\right)(n_a + n_b) \text{ where } \gamma = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{\left(S_1^2 / n_1\right)^2}{n_1 - 1} + \frac{\left(S_2^2 / n_2\right)^2}{n_2 - 1}}$$

These solutions are easily applied using the R package 'Partiallyoverllaping' (Derrick, 2017) as demonstrated by Derrick, Toher & White (2017)

Non-parametric solutions

For the proposed non-parametric solutions, all observations are pooled into one data set and assigned rank values in ascending order. This is equivalent to an RT-1 (Conover & Iman, 1981) ranking procedure. The rank values are substituted into the elements of the calculation for T_{new1} and T_{new2} in place of the observed values. Tied ranks are each given the median of the tied ranks. This gives the test statistics T_{RNK1} and T_{RNK2} respectively. The degrees of freedom are υ_1 and υ_2 respectively, calculated using the pooled rank values. The calculation of *r* uses an RT-2 (Conover & Iman, 1981) ranking procedure, so that *r* represents Spearman's rank correlation coefficient between the paired observations. For the two sample situation, the means, variances, skewness and kurtosis maintain similar characteristics for a distribution transformed to ranks, as are observed in the original distribution (Zimmerman, 2011).

Simulation methodology

The robustness of existing test statistics and proposed test statistics for two samples containing both independent observations and paired observations is assessed using simulation. Monte-Carlo studies are long established techniques for identifying appropriate test statistics in a given scenario (Serlin, 2000). Firstly, Type I error robustness is assessed using liberal robustness criteria (Bradley, 1978). Power is only calculated for Type I error robust statistics, so that fair power comparisons can be made (Zimmerman, 1987; Penfield, 1994).

The values n_a, n_b, n_c, ρ, σ_1^2 and σ_2^2 are defined as part of a factorial design as given in Table 1. Normal deviates for n_a and n_b observations are calculated using methodology outlined by Box and Muller (1958). Similarly, two sets of n_c observations are generated, and are converted to correlated Normal variates using methodology outlined by Kenney and Keeping (1951).

Each of the test statistics given in Table 1 are assessed firstly under the standard Normal distribution. For the comparison of test statistics under non-normality, random numbers are generated by transformation of bivariate standard Normal deviates, N (Forbes *et.al.*, 2011). For a moderately skewed distribution, Gumbel deviates, G, are generated using the transformation G = $-$log($-$log U), where U is the cumulative distribution function of N. To demonstrate the robustness of the test statistics for a more extreme skewed distribution, bivariate Normal deviates, N, are transformed into Lognormal deviates, L, using the transformation L = exponential (N).

In this Monte-Carlo study, the nominal Type I error rate is α_{nominal} =0.05. For each of the parameter combinations in Table 1, two sided tests are performed and the null hypothesis rejection rate is the proportion of the 10 000 replicates where the null hypothesis is rejected.

The alternative hypothesis is generated by adding 0.5 to the n_2 observations so that $\mu_2 - \mu_1 = 0.5$. The difference applied is arbitrary for the purposes of comparing which test statistics are more powerful relative to each other for otherwise equivalent simulation parameters.

The transformations outlined above ensure that the distributions compared are of the same shape, and only differ in terms of central location. Additional analyses are then performed when the samples are drawn from the Normal distribution with unequal variances, and then when samples are drawn from distributions with differing functional form. For the latter one sample is taken from a Normal distribution and one sample taken from a Lognormal distribution. For assessing the Type I error robustness under normality with unequal variances, the n_1 observations are multiplied by σ_1 and the n_2 observations multiplied by σ_2. Standardising is performed when comparing samples from two distributions with differing functional form.

Table 1. Summary of the simulation design.

Parameter	Values	
n_a	5, 10, 30, 50, 100, 500	
n_b	5, 10, 30, 50, 100, 500	
n_c	5, 10, 30, 50, 100, 500	
ρ	-0.75, -0.50, -0.25, 0.00, 0.25, 0.50, 0.75	
(σ_1^2, σ_2^2)	(1,1) , (1,4) , (4,1)	
(μ_1, μ_2)	(0,0) , (0,0.5)	
Distributions	Normal, Lognormal, Gumbel.	
Test statistics	T_1	Paired Samples t-test (discard unpaired observations)
	T_2	Equal variances assumed Independent samples t-test (discard paired observations)
	T_3	Welch’s unequal variances independent samples t-test (discard paired observations)
	MW	Mann-Whitney test (discard paired observations)
	W	Wilcoxon test (discard unpaired observations)
	T_{new1}	Partially overlapping samples t-test, equal variances assumed
	T_{new2}	Partially overlapping samples t-test, equal variances not assumed
	T_{RNK1}	Non-parametric partially overlapping samples t-test, equal variances assumed
	T_{RNK2}	Non-parametric partially overlapping samples t-test, equal variances not assumed
Iterations	10,000	
α_{nominal}	0.05	
Language	R version 3.1.3	

Results

In general, Type I errors are more serious than Type II errors (Wells & Hintze, 2007). The results therefore show Type I error rates for each of the test statistics considered, followed by power only for test statistics that control Type I error. The scenario where samples are drawn from the same distribution is firstly considered. This is followed by the scenario where samples are drawn from the Normal distribution with unequal variances, and finally the scenario when the samples are drawn from distinctly differing distributions.

Samples taken from distributions of the same shape

Null hypothesis rejection rates are obtained for each of the parameter combinations where $\mu_1 = \mu_2$ and $\sigma_1^2 = \sigma_2^2$. Sampling from identical distributions with equal underlying population variances ensures that a difference in central location is directly assessed. For each parameter combination, the null hypothesis rejection rate represents the Type I error rate of the test. The Type I error rates for each of the distributions are given in Figure 1. Reference lines added represent Bradley's liberal Type I error robustness criteria.

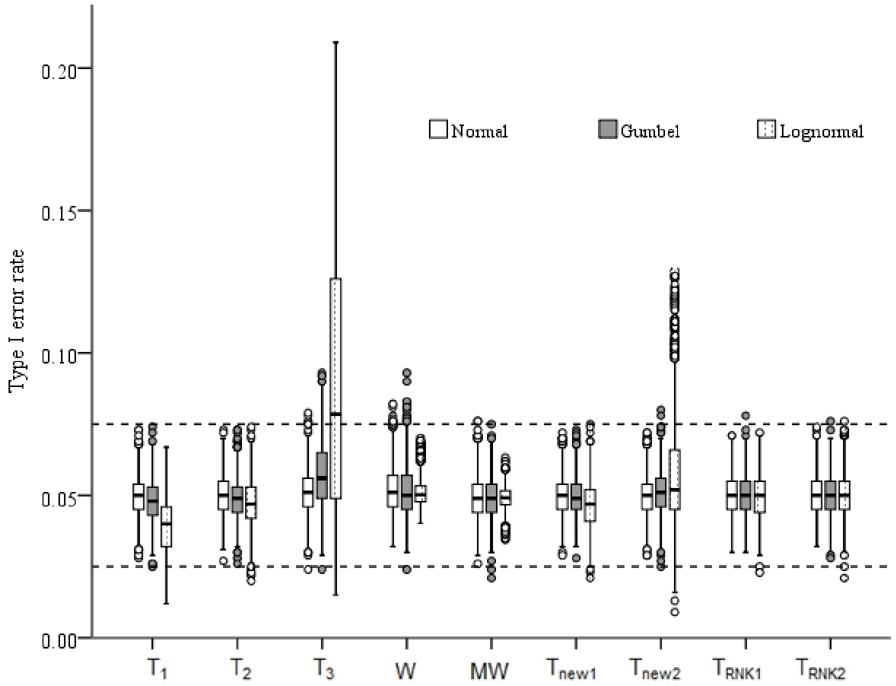


Figure 1. Type I error rates for when both samples are taken from the same distribution.

Figure 1 provides evidence that when two samples are drawn from the Standard Normal distribution, traditional test statistics that discard data, T_1, T_2, T_3, MW, W, MW, remain within Bradley's liberal Type I error robustness criteria. This coincides with findings by Fradette *et.al.*, (2003). Figure 1 also shows that the statistics T_{new1} and T_{new2} are Type I error robust under normality and equal variances. For normally distributed data, the proposed non-parametric statistics, T_{RNK1} and T_{RNK2}, have similar Type I error robustness to T_{new1} and T_{new2}.

Figure 1 suggests that the test statistics under consideration are not sensitive to relatively minor deviations from the Normal distribution. However, it can be seen that only the following test statistics maintain Bradley's liberal criteria when both samples are drawn from a Lognormal distribution; T_2, MW, W, T_{new1}, T_{RNK1}, and T_{RNK2}. The paired samples t-test, T_1, is slightly conservative relative to the other tests statistics.

The degree of skewness for the Lognormal distribution in this paper is larger than the degree of skewness considered by Fagerland and Sandvik (2009a). Figure 3 shows that the MW test remains Type I error robustness for the more extreme degree of skewness in this paper. However, test statistics using separate variances, T_3 and T_{new2}, frequently exceed the upper limit of Bradley's liberal Type I error robustness criteria.

To explore in more detail the performance of the tests under extreme scenarios, Table 2 gives Type I error rates under the Lognormal distribution for small sample size combinations and combinations where max { n_a , n_b , n_c } - min{ n_a , n_b , n_c } is large.

Table 2. Type I error rates for selected sample size combinations under the Lognormal distribution, $\rho = 0.5$.

n_{a}	n_{b}	n_{c}	T_1	T_2	T_3	W	MW	T_{new1}	T_{new2}	T_{RNK1}	T_{RNK2}
5	5	5	.029	.027	.020	.056	.062	.044	.018	.051	.042
10	5	5	.024	.042	.047	.046	.059	.046	.028	.044	.041
10	10	5	.022	.038	.033	.050	.064	.032	.020	.049	.046
10	10	10	.027	.040	.038	.051	.042	.045	.032	.048	.048
5	5	10	.030	.030	.020	.057	.049	.044	.013	.043	.042
30	5	5	.031	.058	.120	.048	.067	.046	.080	.047	.052
30	10	5	.026	.056	.070	.049	.067	.038	.060	.045	.045
50	5	5	.022	.053	.135	.052	.059	.055	.098	.040	.043
100	5	5	.019	.055	.176	.048	.061	.038	.130	.043	.065
500	5	5	.022	.044	.173	.047	.063	.042	.150	.049	.053
5	5	30	.032	.036	.025	.050	.053	.053	.036	.053	.051
5	10	30	.047	.044	.048	.040	.053	.072	.052	.050	.051
5	5	50	.049	.025	.016	.053	.048	.057	.046	.040	.039
5	5	100	.050	.028	.017	.053	.046	.056	.043	.056	.056
5	5	500	.062	.033	.018	.053	.056	.066	.059	.055	.055

The range of the sample sizes in this simulation design is large, Table 2 shows that the inflation in the Type I error rate of T_3 and T_{new2} increases as max { n_a , n_b , n_c } - min{ n_a , n_b , n_c } increases. In the scenario of partially overlapping samples, a large overall sample size does not necessarily result in a robust test. Simply increasing the number of independent observations does not compensate for a small number of paired observations, and vice-versa. When sample sizes are balanced, the non-parametric tests maintain Type I error robustness for the smallest sample size combinations in the simulation design. For a balanced design with increasing sample size the parametric test statistics improve their Type I error robustness as per the central limit theorem, the sampling distribution of the mean differences approaches normality as sample size increases.

Under the alternative hypothesis, when $\mu_2 - \mu_1 = 0.5$, the null hypothesis rejection rate represents the power of the test. For test statistics that do not clearly violate Bradley's liberal robustness criteria, the power of the test statistics for each of the distributions is given in Table 3.

Table 3. Power when $\mu_2 - \mu_1 = 0.5$. Calculated at $\alpha = 0.05$, two sided, averaged over all values of n_c. N = Normal, L = Lognormal, G = Gumbel. For test statistics using only independent observations, the value for $\rho = 0$ is displayed. NR is displayed if not Type I error robust.

		ρ	T_1	T_2	T_3	W	MW	T_{new1}	T_{new2}	T_{RNK1}	T_{RNK2}
N	$n_a = n_b$	> 0	.695			.693		.865	.864	.856	.855
		0	.558	.567	.565	.556	.563	.819	.819	.811	.811
		< 0	.481			.474		.779	.779	.772	.771
	$n_a \neq n_b$	> 0	.695			.692		.839	.832	.829	.824
		0	.559	.455	.433	.553	.438	.806	.798	.795	.790
		< 0	.482			.476		.774	.767	.763	.760
G	$n_a = n_b$	> 0	.611			.630		.783	.782	.815	.814
		0	.464	.472	.470	.483	.510	.720	.718	.761	.760
		< 0	.398			.407		.678	.678	.719	.719
	$n_a \neq n_b$	> 0	.612			.629		.740	.735	.779	.776
		0	.466	.345	.340	.481	.380	.693	.689	.740	.736
		< 0	.398			.410		.655	.651	.702	.699
L	$n_a = n_b$	> 0	.455			.727		.596	NR	.893	.891
		0	.334	.340	NR	.729	.533	.535	NR	.857	.856
		< 0	.297			.693		.506	NR	.826	.826
	$n_a \neq n_b$	> 0	.453			.562		.514	NR	.874	.873
		0	.336	.194	NR	.430	.518	.467	NR	.851	.850
		< 0	.296			.423		.438	NR	.825	.826

When population variances are equal, Table 3 shows that test statistics not assuming equal variances, T_{new2} and T_{RNK2}, perform similarly to their counterparts where equal variances are assumed T_{new1} and T_{RNK1} respectively.

From Table 3 it can be seen that for normally distributed data, traditional parametric methods, T_1 , T_2 and T_3 , are more powerful than their non-parametric counterparts, W and MW. Similarly when the normality assumption is true, the parametric statistics T_{new1} and T_{new2} are marginally more powerful than their non-parametric counterparts T_{RNK1} and T_{RNK2} , but not to any meaningful extent. Figure 2 shows the power for each parameter combination within the simulation design for T_{new1} and T_{RNK1} .

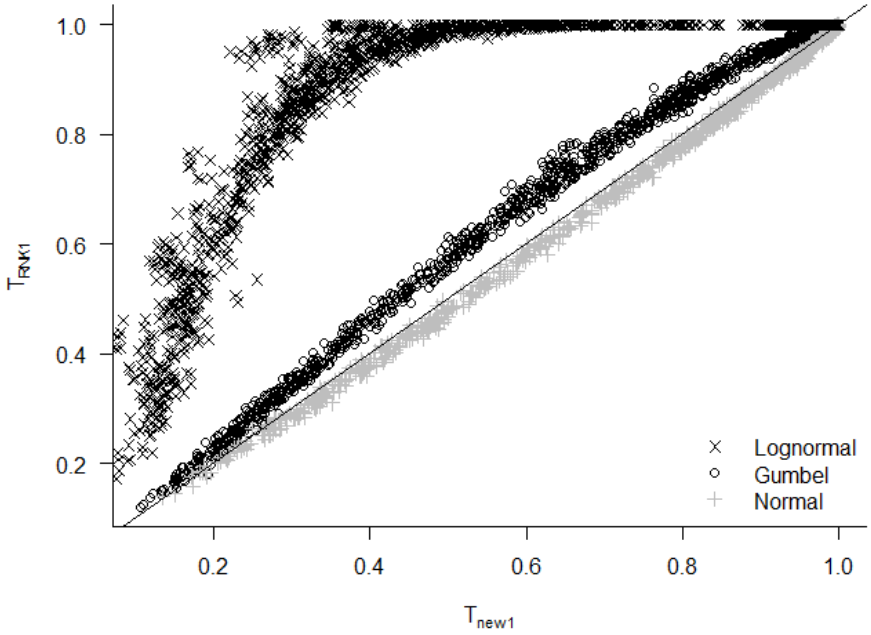


Figure 2. Power for each parameter combination, for T_{new1} and T_{RNK1} .

For the non-normal distributions in this simulation, non-parametric methods are more powerful than their parametric counterparts when both samples are taken from the same distribution. For increasing degrees of skewness, the proposed non-parametric test statistic, T_{RNK1}, exhibits an increasing power advantage over its parametric counterpart, T_{new1} .

From Table 3 it is apparent that for all of the test statistics making use of some paired element, a negative correlation between two samples is problematic. A large positive correlation results in more powerful results. This is true for each of the distributions in the simulation design. For selected tests making use of the paired data, Figure 3 shows the power for each parameter combination within the simulation design.

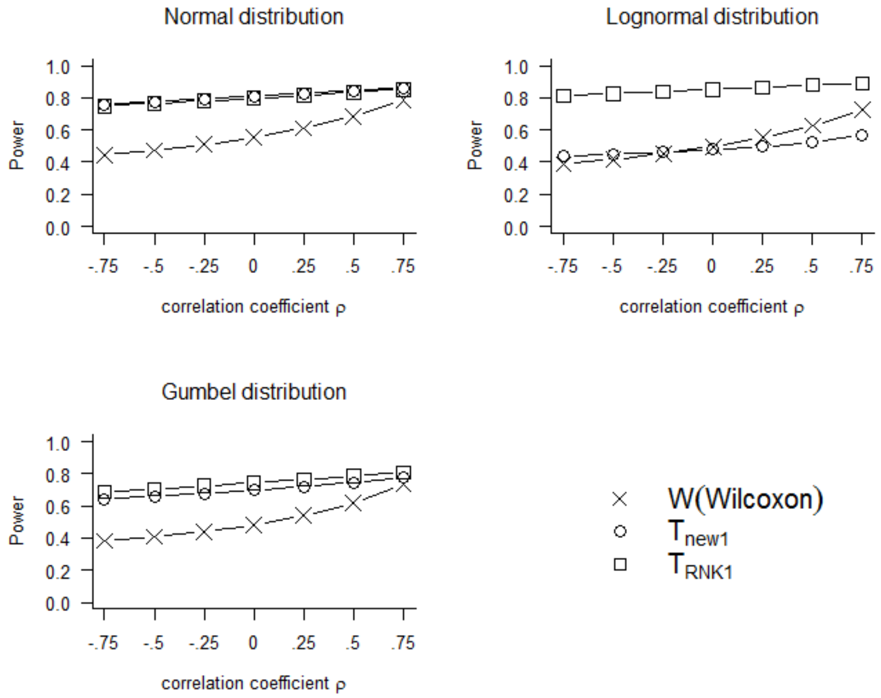


Figure 3. Power of selected test statistics making use of paired data, for two N(0,1) samples.

Figure 3 illustrates that as the correlation between the paired observations increases, the power of the tests statistics making use of paired information increases. For the Normal distribution and the Gumbel distribution, when the correlation coefficient is negative or small, the power advantage when using all of the available data is large. For the Gumbel distribution, T_{new1} is only slightly less powerful than T_{RNK1}, however for the Lognormal distribution there is a clear power advantage of T_{RNK1} over T_{new1}. This suggests that the proposed T_{RNK1} is particularly useful for comparing two samples from a distribution with a clear deviation from normality, and a negative or small correlation between the two groups.

Samples taken from the Normal distributions with unequal variance

Null hypothesis rejection rates are obtained for each of the parameter combinations where $\mu_1 = \mu_2$ and $\sigma_1^2 \neq \sigma_2^2$. When the observations are sampled from two Normal distributions with equal means and unequal variances, the null hypothesis rejection rate represents the Type I error rate of the test. Type I error rates for each of the test statistics across the simulation design are given in Figure 4.

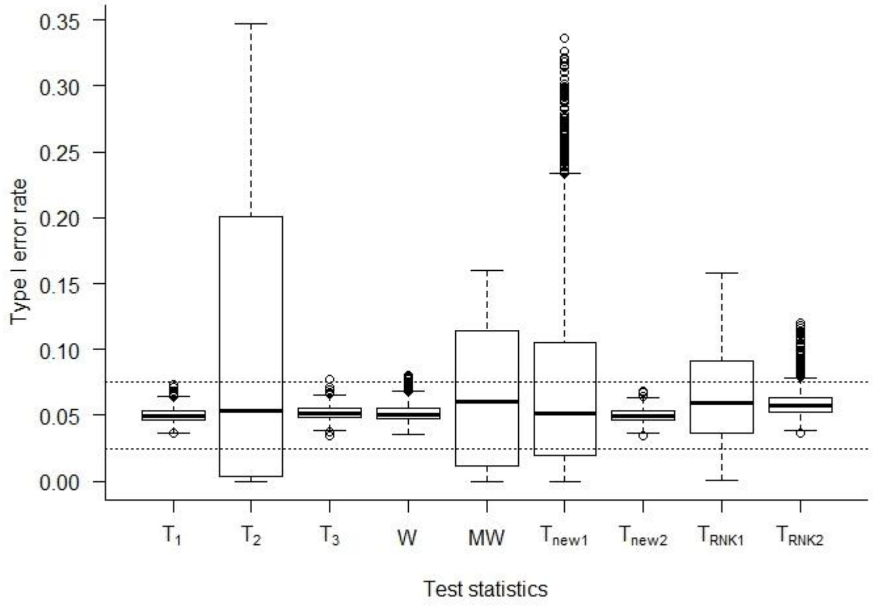


Figure 4. Type I error rates for samples from the Normal distribution with $\sigma_1^2 = 1,\ \sigma_2^2 = 4$.

Figure 4 shows that Type I error robustness is maintained under normality for T_{new2}. Thus T_{new2} is the only test statistic making use of all available data to be Type I error robust under normality for both equal and unequal variances.

For normally distributed data and unequal population variances, the test statistics not assuming equal variances are more Type I error robust than the statistics that do assume equal variances. Nevertheless, for T_{RNK2} the number of times the null hypothesis is rejected is in excess of acceptable levels. Closer inspection of our results shows these statistics are not robust when the number of paired observations is large relative to the total number of independent observations. This effect is exacerbated when ρ is large and positive. To a lesser extent, the rejection rates for T_{RNK2} are inflated when the total number of independent observations are very large relative to the number of paired observations.

Samples taken from distributions of unequal shape

To consider the behaviour of the test statistics when the two samples are drawn from distinctly different distributions (standardised to ensure equal means), Figure 5 shows the null hypothesis rejection rates when observations for Sample 1 are taken from the standard Normal distribution, and observations for Sample 2 are taken from the Lognormal distribution.

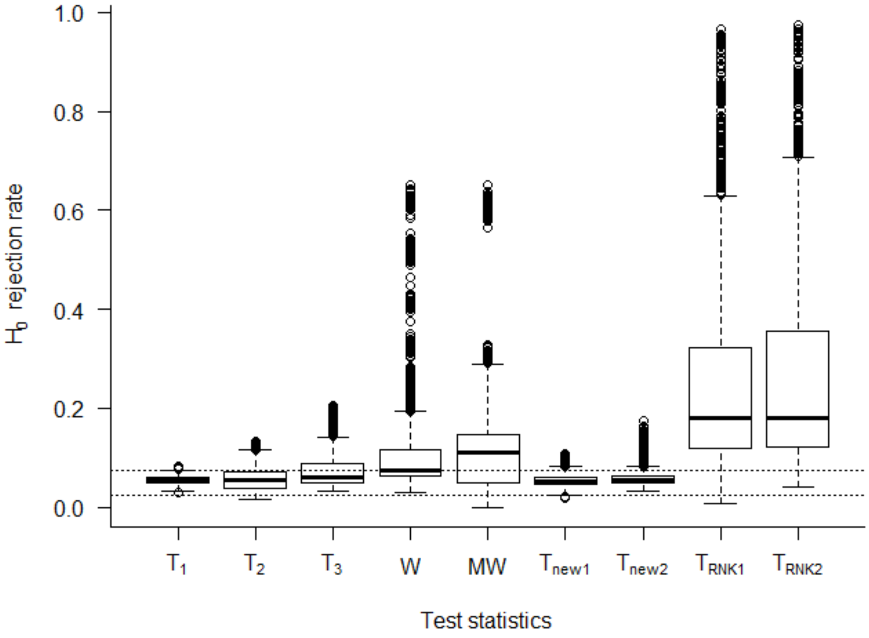


Figure 5. Sample 1 values taken from the standard Normal distribution, Sample 2 observations are taken from a standardised Lognormal distribution.

Under the simulation design, standardising of the population ensures that the mean for both distributions is the same, but the shapes of the distributions are different. The null hypothesis rejection rate only represents the Type I error rate if the null hypothesis is strictly that there is no difference in means. Figure 5 shows that the parametric tests are not sensitive to the different shapes of the distributions and remain valid for testing the hypothesis of equal means. Conversely, the null hypothesis rejection rate is well in excess of 5% for the non-parametric test statistics. The non-parametric statistics are sensitive to differences in the shape of the distribution, thus could be used to assess the null hypothesis of equal distributions. The null hypothesis rejection rates represent power under the latter form of the null hypothesis.

Example

The following is a classic example by Rempala and Looney (2006), used by Guo and Yuan (2017) and Amro and Pauly (2017) to illustrate the partially overlapping samples problem. The outcome variable is the Karnofsky performance status scale, which measures functional status of a patient. The data is recorded on the last day of life and on the second to the last day. For the parametric tests, the null hypothesis that the mean Karnofsky score is the same on the last two days of life is tested. For the non-parametric tests, the null hypothesis that the distribution of the Karnofsky score is the same on the last two days is tested. Assuming the distributions differ only in central location, both the parametric and nonparametric tests are assessing the same research question.

For a total of 60 patients, 9 were recorded on both days, 28 were recorded only on the second to the last day, and 23 were recorded only on the last day. The test statistic and p-value for each of the approaches considered are given in Table 4, based on the data below:

Patients with scores on both days:
(20, 10), (30, 20), (25, 10), (20, 20), (25, 20), (10, 10), (15, 15), (20, 20), (30, 30)
Patients with scores only on the second to the last day:
10,10,10,10,15,15,15,20,20,20,20,20,20,20,20,20,20,20,20,25,25,25,25,30,30,30,30,30,30
Patients with scores only on the last day:
10,10,10,10,10,10,10,10,10,15,15,20,20,20,20,20,20,20,20,25,25,30,30,30

Using the midpoint of tied ranks to calculate T_{RNK1} and T_{RNK2}; all scores of 10 have rank of 9, all scores of 15 have rank of 21, all scores of 20 have rank of 37, all scores of 25 have rank of 53.5, all scores of 30 have rank of 63.5.

Table 4. Results from Rempala and Looney example

Method	T_1	T_2	T_3	MW	W	T_{new1}	T_{new2}	T_{RNK1}	T_{RNK2}
Test statistic	1.818	1.800	2.286	412.5	10	2.522	2.507	2.534	2.521
p-value	0.075	0.079	0.052	0.078	0.098	0.015	0.016	0.014	0.015

Table 4 shows that the parametric partially overlapping samples t-tests provide evidence at the 5% significance level to suggest that there is a difference in the mean Karnofsky scores between the last two days of life. Similarly, the non-parametric partially overlapping samples t-tests provide evidence at the 5% significance level to suggest that there is a difference in the distribution of the Karnofsky scores between the last two days of life.

Conclusion

There are many scenarios which gives rise to partially overlapping samples. Traditional methods of analyses which discard data are less than desirable. The partially overlapping samples t-tests by Derrick *et.al.*, (2017) offer robust parametric solutions, assuming MCAR, using all of the available data.

Under normality, parametric solutions T_{new1} and T_{new2} are Type I error robust and have greater power than other tests statistics considered in this paper. When the normality assumption is true, T_{new1} is recommended for equal variances, and T_{new2} is recommended for unequal variances. For the non-normal distributions considered here, T_{new1} is Type I error robust when comparing two samples taken from the same distribution, whereas T_{new2} is not fully Type I error robust.

Non-parametric approaches developed in this paper, T_{RNK1} and T_{RNK2} are Type I error robust when comparing two samples taken from the same distribution with equal means and equal variances. When observations for two groups are sampled from the same non-normal distribution, there is a power advantage of using the non-parametric approaches T_{RNK1} and T_{RNK2}.

When comparing samples from two distinctly different distributions, the correct form of the null hypothesis for the non-parametric methods is open to interpretation. If performing parametric tests, the null hypothesis of equal means is valid. Results show that as with traditional non-parametric tests, the proposed non-parametric test statistics are sensitive to differences in location, but are simultaneously sensitive to differences in the shape of the distribution. If the sampling distributions are not thought to be identical, the proposed non-parametric tests are not appropriate when the primary goal is to assess for differences in location. If the research question is whether the distributions are equal, T_{RNK1} and T_{RNK2} offer valid and more powerful alternatives to their parametric counterparts T_{new1} and T_{new2} respectively, as well as more powerful alternatives to standard non-parametric methods which discard data.

References

Amro, L., & Pauly, M. (2017). Permuting incomplete paired data: a novel exact and asymptotic correct randomization test. *Journal of Statistical Computation and Simulation*, 87(6), 1148-1159.

Bhoj, D. (1978). Testing equality of means of correlated variates with missing observations on both responses. *Biometrika*, 65(1), 225-228.

Box, G. E. P., & Muller, M. (1958). A note on the generation of random normal deviates. *Annals of Mathematical Statistics,* 29, 610-611.

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

Conover, W. J., & Iman, R. L. (1981). Rank transformations as a bridge between parametric and nonparametric statistics. *The American Statistician*, 35(3), 124-129.

Chatfield, M., & Mander, A. (2009). The Skillings–Mack test (Friedman test when there are missing data). *The Stata Journal*, 9(2), 299-305.

Chen, Z. (2011). Is the weighted z-test the best method for combining probabilities from independent tests? *Journal of Evolutionary Biology*, 24(4), 926-930

Derrick, B. (2017) Partiallyoverlapping: Partially overlapping samples t-tests. CRAN [R-package].

Derrick, B., Dobson-McKittrick, A., Toher, D. & White P. (2015). Test statistics for comparing two proportions with partially overlapping samples. *Journal of Applied Quantitative Methods*, 10(3), 1-14.

Derrick, B., Russ, B., Toher, D. & White P. (2017). Test statistics for the comparison of means for two samples which include both paired observations and independent observations. *Journal of Modern Applied Statistical Methods*, 16(1), 137-157.

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

Derrick, B., Toher, D. & White, P. (2017). How to compare the means of two samples that include paired observations and independent observations: A companion to Derrick, Russ, Toher and White (2017). *The Quantitative Methods for Psychology*, 13(2), 120-126.

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

Fagerland. M., & Sandvik, L. (2009b) The Wilcoxon-Mann-Whitney test under scrutiny. *Statistics in Medicine*, 28(10), 1487-1497.

Fay, M. P., & Proschan, M. A. (2010). Signed Rank Sum Wilcoxon-Mann-Whitney or t-test? On assumptions for hypothesis tests and multiple interpretations of decision rules. *Statistics Surveys*, 4, 1-39.

Forbes, C., Evans, M., Hastings, N., & Peacock, B. (2011). *Statistical distributions*. John Wiley & Sons.

Fradette, K., Keselman, H.J., Lix, L., Algina, J., & Wilcox, R. (2003) Conventional and Robust Paired and Independent Samples t-tests: Type I Error and Power Rates, *Journal of Modern Applied Statistical Methods*, 2(2), 481-496.

Guo, B., & Yuan, Y. (2017). A comparative review of methods for comparing means using partially paired data. *Statistical methods in medical research*, 26(3), 1323-1340.

Hart, A. (2001). Mann-Whitney test is not just a test of medians: differences in spread can be important. *British Medical Journal*, 323(7309), 391.

Hosgood, S.A., Saeb-Parsy, K., Wilson, C., Callaghan, C., Collett, D. and Nicholson, M.L. (2017). Protocol of a randomised controlled, open-label trial of ex vivo normothermic perfusion versus static cold storage in donation after circulatory death renal transplantation. *BMJ open*, 7(1), p.e012237.

Howell, D. (2012). Statistical Methods for Psychology. Cengage Learning.

Kim, B. S., Kim, I., Lee, S., Kim, S., Rha, S. Y., & Chung, H. C. (2005). Statistical methods of translating microarray data into clinically relevant diagnostic information in colorectal cancer. *Bioinformatics*, 21(4), 517-528.

Kenney, J. F. & Keeping, E. S. (1951) *Mathematics of Statistics*, Pt. 2, 2nd ed. Princeton, NJ: Van Nostrand.

Lancaster, H. O. (1961). The combination of probabilities: an application of orthonormal functions. *Australian Journal of Statistics*, 3(1), 20-33.

Lin, P., & Strivers L. (1974) Difference of Means with Incomplete Data, *Biometrika*, 61(2), 325-334.

Looney, S. & Jones, P. (2003) A method for comparing two normal means using combined samples of correlated and uncorrelated data, *Statistics in Medicine*, 22, 1601-1610.

Mehrotra, D (2004). Letter to the editor, A method for comparing two normal means using combined samples of correlated and uncorrelated data. *Statistics in Medicine*, 23, 1179–1180.

Mendenhall, W., Beaver, R., & Beaver, B. (2008). *Introduction to Probability and Statistics*. Cengage Learning.

Martinez-Camblor, P., Corral, N., & De La Hera,. J. M. (2012) Hypothesis test for paired samples in the presence of missing data, *Journal of Applied Statistics*, 40(1), 76-87.

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.

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

R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. www.R-project.org. 2014; version 3.1.3.

Rempala, G. A., & Looney, S. W. (2006). Asymptotic properties of a two sample randomized test for partially dependent data. *Journal of statistical planning and inference*, 136(1), 68-89.

Ruxton, G. (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.

Samawi, H. M., & Vogel, R. (2011). Tests of homogeneity for partially matched-pairs data. *Statistical Methodology*, 8(3), 304-313.

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

Skillings, J. H., & Mack, G. A. (1981). On the use of a Friedman-type statistic in balanced and unbalanced block designs, *Technometrics*, 23(2), 171-177.

Skovlund, E., & Fenstad, G. U. (2001). Should we always choose a nonparametric test when comparing two apparently non-normal distributions?, *Journal of Clinical Epidemiology*, 54(1), 86-92.

Stouffer, S. A., Suchman, E. A., DeVinney, L. C., Star, S. A., & Williams Jr, R. M. (1949). The American soldier: combat and its aftermath. *Studies in Social Psychology in World War II* (2).

Uddin, N., & Hasan, M. S. (2017). Testing equality of two normal means using combined samples of paired and unpaired data. *Communications in Statistics-Simulation and Computation*, 46(3), 2430-2446.

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

Zimmerman, D. W. (1987). Comparative power of Student t-test and Mann-Whitney U test for unequal sample sizes and variances. *The Journal of Experimental Education*, 55, 171-174.

Zimmerman, D. W. (2004). Inflation of type I error rates by unequal variances associated with parametric, nonparametric, and rank-transformation tests. *Psicologica: International Journal of Methodology and Experimental Psychology*, 25(1), 103-133.

Zimmerman, D. W. (2011). Inheritance of Properties of Normal and Non-Normal Distributions after Transformation of Scores to Ranks. Psicologica: International Journal of Methodology and Experimental Psychology, 32(1), 65-85.

Appendix P8

Pearce, J. and Derrick, B. (2019). “Preliminary testing: the devil of statistics?” *Reinvention: an International Journal of Undergraduate Research* 12

Accepted Version

Preliminary Testing: The Devil of Statistics?

Jack Pearce

Applied Statistics Group

University of the West of England, Bristol

Frenchay Campus

BS16 1QY

jrepearce@outlook.com

Jack 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

Applied Statistics Group

University of the West of England, Bristol

Frenchay Campus

BS16 1QY

ben.derrick@uwe.ac.uk

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 outcome of 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 to compare two-independent samples.

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), others 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 Reject null hypothesis of equal variances.	p = .071 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 Fail to reject the null hypothesis that the two samples means do not differ.	p = .046 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.

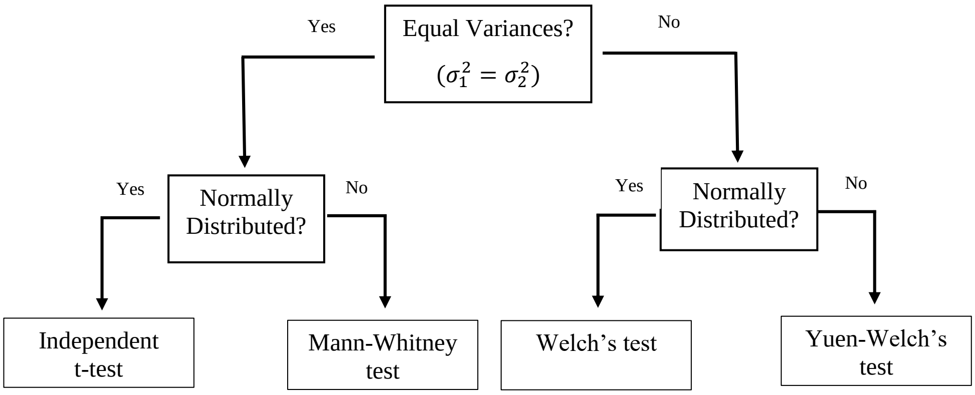


Figure 1: Two-Step test procedure, with both equal variance and normality preliminary tests.

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) emphasised 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) have shown 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) recommends 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). This highlights the amount of literature on the subject without a clear consensus.

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 means of zero. Firstly, where both groups have variances equal to one. Secondly, groups sampled from the Normal distributions with unequal variances {1, 2, 4} are considered. Observations from the exponential distribution are generated with a mean and variance of 1. Observations from the lognormal distribution 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.

	Overall Type I Error Rate			
Distribution	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 shows the distribution of the 16 Type I Error Rates for the different combinations of sample sizes (96 for the Normal Distribution under unequal variances). ‘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.

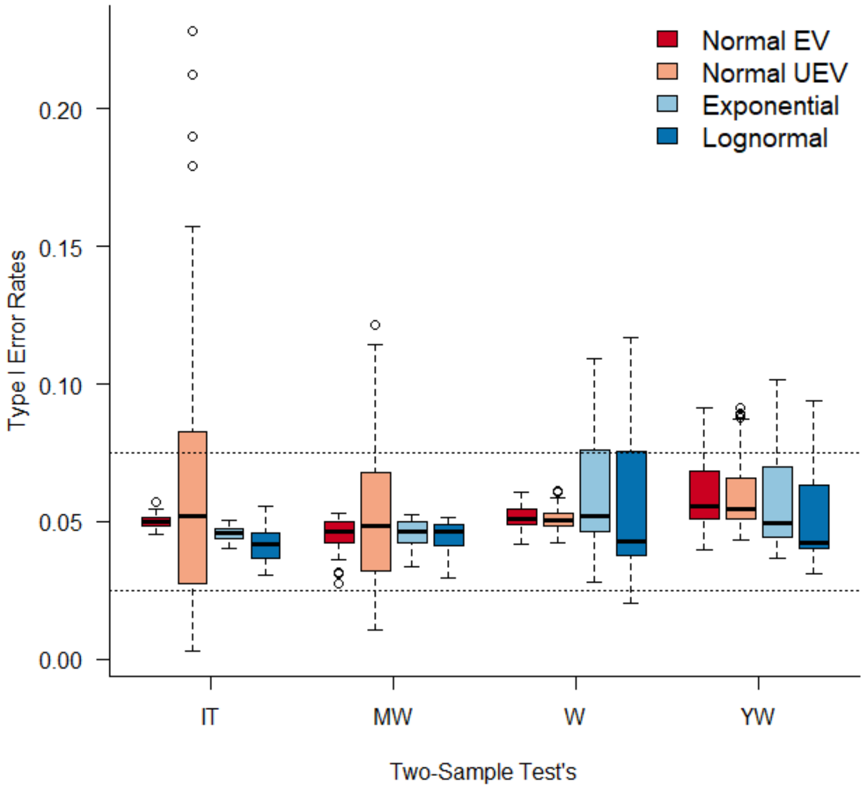


Figure 2: Two-Sample tests unconditional Type I error rates.

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 best preliminary test combinations for each distribution was assessed. For each of these preliminary test combinations, the average absolute deviation from 5% across the four distributions is given in Table 4.

	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.

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.

Acknowledgments

I would like to thank my Mum and Dad first and foremost for their support throughout my degree. I would also like to show my appreciation towards my A-Level Maths teacher Mr Gray, who did more for me in six months than any other teacher; as well as Mrs Doran who helped me throughout the whole seven years at school. Acknowledgement is given to three anonymous reviewers for their thorough discussion based on this paper. Thanks also goes to Dr Paul White for his invaluable advice on this paper and all things stats……. Jack

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.

Appendix P9

Derrick, B. and White, P. (2017). "Comparing two samples from an individual Likert question". *International Journal of Mathematics and Statistics* 18 (3), pp. 1–13

Published Version

International Journal of Mathematics and Statistics,
Year 2017; Volume 18; Issue Number 3;
ISSN 0974-7117 (Print); ISSN 0973-8347 (Online);

Comparing Two Samples from an Individual Likert Question

Ben Derrick and **Paul White**

Faculty of Environment and Technology,
University of the West of England, Bristol, BS16 1QY (UK)
Email: ben.derrick@uwe.ac.uk; paul.white@uwe.ac.uk

ABSTRACT

For two independent samples there is much debate in the literature whether parametric or non-parametric methods should be used for the comparison of Likert question responses. The comparison of paired responses has received less attention in the literature. In this paper, parametric and non-parametric tests are assessed in the comparison of two samples from a paired design on a five point Likert question. The tests considered are the independent samples t-test, the Mann-Whitney test, the paired samples t-test and the Wilcoxon test. Pratt's modified Wilcoxon test for dealing with zero differences is also included. The Type I error rate and power of the test statistics are assessed using Monte-Carlo methods. The parameters varied are; sample size, correlation between paired observations, and the distribution of the responses. The results show that the independent samples t-test and the Mann-Whitney test are not Type I error robust when there is correlation between the two groups compared. Pratt's test more closely maintains the Type I error rate than the standard Wilcoxon test does. The paired samples t-test is Type I error robust across the simulation design. As the correlation between the paired samples increases, the power of the test statistics making use of the paired information increases. The paired samples t-test is more powerful than Pratt's test when the correlation is weak. The power differential between the test statistics is exacerbated when sample sizes are small. Assuming equally spaced categories on a five point Likert item, the paired samples t-test is not inappropriate.

Keywords: Likert item; Likert scale; Wilcoxon test; Pratt's test; Paired samples t-test

Mathematics Subject Classification: 60 62

1. INTRODUCTION

A Likert item is a forced choice ordinal question which captures the intensity of opinion or degree of assessment in survey respondents. Historically a Likert item comprises five points worded: Strongly approve, Approve, Undecided, Disapprove, Strongly Disapprove (Likert, 1932). Other alternative wording, such as "agree" or "neutral" or "neither agree nor disagree" may be used depending on the context.

The literature is sometimes confused between the comparison of samples using summed Likert scales and the comparison of samples for individual Likert items (Boone and Boone, 2012). A summed Likert scale is formed by the summation of multiple Likert items that measure similar information. This summation process necessarily requires the assignment of scores to the Likert ordinal category labels. The summation of multiple Likert items to produce Likert scales has not been without controversy but it is a well-established practice in scale construction, and is one which may produce psychometrically robust scales with interval-like properties. Such derived scales, could potentially yield data amenable to analysis using parametric techniques (Carifo and Perla, 2007). Distinct from Likert scales, the comparison of two samples on an individual Likert question is the subject of this paper.

The response categories of a five point Likert item may be coded 1 to 5 and the item responses viewed as being ordinal under Stevens (1946) classification scheme. Extant literature acknowledges that in certain practical and methodological aspects, the Likert-item responses may approximate interval level data(Norman, 2010). The ordinal codes 1, 2, 3, 4, and 5 or alternatively -2, -1, 0, 1, 2 could be used as numerical scores in robust tests for differences. This change from codes to numeric scores is used in the creation of summated Likert scales and is at the heart of the controversy. Proponents in favour of such practice advance an argument that the Likert question is accessing some information from an underlying scale and the resultant score is a non-linear realisation from this scale (Norman, 2010). Thus, although the scored item may not perfectly have the required properties to be classed as interval level data under Stevens classification scheme, the scored item might, in practice, approximate interval level data and be amenable to analysis using parametric techniques.

When comparing two independent sets of responses from a Likert question, the independent samples t-test is frequently performed. The corresponding non-parametric test for independent samples is the Mann-Whitney-Wilcoxon test (Wilcoxon, 1945). This test may also be referred to as the Wilcoxon-Mann-Whitney test, or as is the case in this paper, simply referred to as the Mann-Whitney test.

For two independent samples, whether the correct approach for analysis should be a parametric t-test or the non-parametric Mann-Whitney test is much debated in the literature (Sullivan and Artino, 2013). The choice between parametric and non-parametric tests for the analysis of single Likert items depends on the assumptions that researchers are willing to make and the hypotheses that they are testing (Jamieson, 2004). Some practitioners are uncomfortable with a comparison of means using a parametric test, arguing that response categories cannot be justifiably assumed to be equally spaced and consequently the use of equally spaced scores is unwarranted. In contrast, Allen and Seaman (2007) suggests that Likert items measure an underlying continuous measure and suggests the use of the independent samples t-test as a pilot test, prior to obtaining a continuous measure. If the assumption that the underlying distribution is continuous can be deemed reasonable, Likert responses approximate interval data. For interval data, the use of parametric tests may not be inappropriate. When the assumption of interval data applies, consideration should be given to the sample size and distribution of the responses before applying the independent samples t-test (Jamieson, 2004).

If sample sizes are large, both parametric and non-parametric test statistics are likely to have adequate power. However, in research there is a trade-off between increasing sample size and reducing collection costs. When resource is scarce, the most powerful test statistic for small samples is of interest.

For two independent samples, De Winter and Dodou(2010) found that both the independent samples t-test and the Mann-Whitney test are generally Type I error robust at the 5% significance level for a five point Likert item. This is true across a diverse range of distributions and sample sizes. Both tests suffer some exceptions to Type I error robustness when the distributions have extreme kurtosis and skew. The power is similar between the two tests, for both equal and unequal sample sizes. When the distribution is multimodal with responses split mainly between strongly approve and strongly disapprove, the independent samples t-test is more powerful than the Mann-Whitney test. Rasch, Teuscher and Guiard(2007) show that using the Mann-Whitney test using the Normal approximation with correction for ties is Type I error robust for two groups of independent observations on a five point Likert item.

For two independent samples, Nanna and Sawilowski(1998) found that the independent samples t-test and the Mann-Whitney test are Type I error robust for seven point Likert item responses, with the Mann-Whitney test superior in power. This is likely observed because there is more scope to apply greater skew on a higher point Likert-style scale.

The literature is much quieter on the analysis of Likert items in paired samples designs. A non-parametric test for paired samples is the Wilcoxon rank sum test (Wilcoxon, 1945). This is often referred to as the Wilcoxon signed rank test, or as is the case in this paper, simply referred to as the Wilcoxon test. When the samples are from an underlying Normal distribution, the null hypothesis is of equal distributions, but this is particularly sensitive to changes in location (Hollander, Wolfe and Chicken, 2013). Thus if samples are from a bivariate Normal distribution, assessing for a location shift is reasonable.

When comparing two groups of paired samples on a five point Likert item, the paired samples t-test is often used in preference to the Wilcoxon test (Clason and Dormody, 1994). This choice of test is not inappropriate when interval approximating data is assumed, and when the null hypothesis is one of no difference in central location (Sisson and Stocker, 1989).

The degree of correlation between two samples is likely to impact the choice of test. The correlation between two sets of responses on a Likert scale is typically hard to quantify. With respect to bivariate Normal distributions, Fradette et.al. (2003) suggest that if the correlation is small then the independent samples t-test could be used. However, under the same conditions, Zimmerman (1997) argues that using the independent samples t-test for even a small a degree of correlation violates the independence assumption and can distort the Type I error rate. For bivariate normality, Vonesh(1983) demonstrates that the paired samples t-test is more powerful than the independent samples test when $\rho \geq 0.25$.

In general, the Wilcoxon test with a correction for ties, may be used to test for a location shift between two discrete groups. The Wilcoxon test discards observations where there is a zero difference between the two groups. Given the discrete nature of Likert item data, it would not be unusual to observe a large proportion of zero differences in a sample. The discarding of many data pairs with a zero difference may be problematic. Pratt (1959)proposed a modification of the Wilcoxon test to overcome potential problems caused by discarding zero differences. In Pratt's test, the absolute paired differences are ordered including the zero differences, ranks are applied to the non-zero differences as if the zero differences had received ranks, and these ranks used in the Wilcoxon test. Conover (1973) compared the Wilcoxon test dropping zero differences to Pratt's test incorporating zero differences and concluded that the relative performance of the two approaches depends on the underlying distribution. The comparison conducted by Conover (1973) did not include Likert items and did not extend to the inclusion of the paired samples t-test.

A further alternative method for handling zero differences suggested by Pratt (1959) is to randomly allocate zero differences to either positive or negative ranks. To achieve this for every zero difference add a random uniform deviate $\varepsilon \sim U(-0.1,\ 0.1)$ and then proceed with the ranking. This approach is referred to as the random epsilon method in the following.

For paired five point Likert data we seek to compare the relative behaviour of the Wilcoxon test, Pratt's test, the random epsilon method and the paired samples t-test. The comparison is undertaken by discretising realisations from bivariate Normal distributions on to a five point scale over a range of correlation coefficients, ρ, including ρ = 0. For this latter reason we additionally include the Mann-Whitney test and the independent samples t-test in the comparison. Mindful that differences in location are likely to be accompanied with differences in variances, we additionally include the separate variances t-test i.e. Welch's test in the comparison.It is known that for independent samples, Welch's test is Type I error robust under normality for both equal and unequal variances (Derrick, Toher and White, 2016).

Below we give the simulation study, key results and a discussion of the findings.

2. METHODOLOGY

Random Normal deviatesfor two groups of sample size n are generated usingthe Box–Muller (1958) transformation. These deviates are transformed into n pairs with Pearson's correlation coefficient ρ using methodology outlined by Kenney and Keeping (1951).

For each combination of n and ρ, correlated bivariate Normal deviates x_{ij} are generated, where *i*= {1:*n*} and *j* = {Group 1, Group 2}. The mean of the sample is varied by adding μ_j to each deviate so that x_{ij} ~N(μ_j ,1). The values of each of the parameters simulated are given in Table 1.

Table 1. Summary of the simulation design.

Sample size, n	10, 20, 30, 50
Correlation coefficient, ρ	0.00, 0.25, 0.50, 0.75
Scenarios	$\begin{array}{lcccc} & \mu_1 & \mu_2 & \eta_1 & \eta_2 \end{array}$ $\left.\begin{array}{lcccc} A) & 0 & 0 & 0 & 0 \\ B) & 0.5244 & 0.5244 & 1 & 1 \\ C) & 1.2816 & 1.2816 & 2 & 2 \end{array}\right\} H_0$ $\begin{array}{lcccc} & \mu_1 & \mu_2 & \eta_1 & \eta_2 \end{array}$ $\left.\begin{array}{lcccc} D) & 0 & 0.5244 & 0 & 1 \\ E) & 0 & 1.2816 & 0 & 2 \\ F) & 0.5244 & 1.2816 & 1 & 2 \\ G) & -0.5244 & 0.5244 & -1 & 1 \\ H) & -0.5244 & 1.2816 & -1 & 2 \\ I) & -1.2816 & 1.2816 & -2 & 2 \end{array}\right\} H_1$
Test Statistics	T_1 Paired samples t-test T_2 Independent samples t-test T_3 Welch's t-test W_1 Wilcoxon test (Traditional method, discarding zeroes) W_2 Pratt's test (Wilcoxon test, Pratt's zeroes modification) W_3 Random ε (Wilcoxon test, $\varepsilon \sim U(-0.1, 0.1)$ added to zeroes) MW Mann-Whitney test.
Number of iterations	10,000
Nominal significance level	5% (two-sided test)
Programming language	R version 3.1.3
	Complete tables of all results available on request.

Without loss of generality the five points on the Likert scale are numbered from -2 to 2, the "neutral" response is 0. The Likert-style responses y_{ij} are calculated using the cut-points as follows:

$$y_{ij} = \left\{ \begin{array}{lll} 2 & \text{if} & x_{ij} > 0.8416 \\ 1 & \text{if} & 0.2533 \le x_{ij} \le 0.8416 \\ 0 & \text{if} & -0.2533 \le x_{ij} \le 0.2533 \\ -1 & \text{if} & -0.8416 \le x_{ij} \le -0.2533 \\ -2 & \text{if} & x_{ij} < \text{-}0.8416 \end{array} \right\}$$

The cut-points are calculated so that under N(0,1) the theoretical distribution of the Likert-style responses is uniform. The median of Group 1 and the median of Group 2 are represented by η_1 and η_2 respectively. Scenarios A) to I) in Table 1 give an example of each of the possible bivariate pairings of η_1 and η_2 within a five point Likert design. For example, scenario D) $\eta_1 = 0, \eta_2 = 1$, is equivalent to $\eta_1 = 1, \eta_2 = 0$; $\eta_1 = 0, \eta_2 = -1$; and $\eta_1 = -1, \eta_2 = 0$.

For selected parameter combinations within the factorial simulation design, theoretical observed proportions of y_{ij} are illustrated in Figure 1. These showcase the range of distributions in the simulation design.

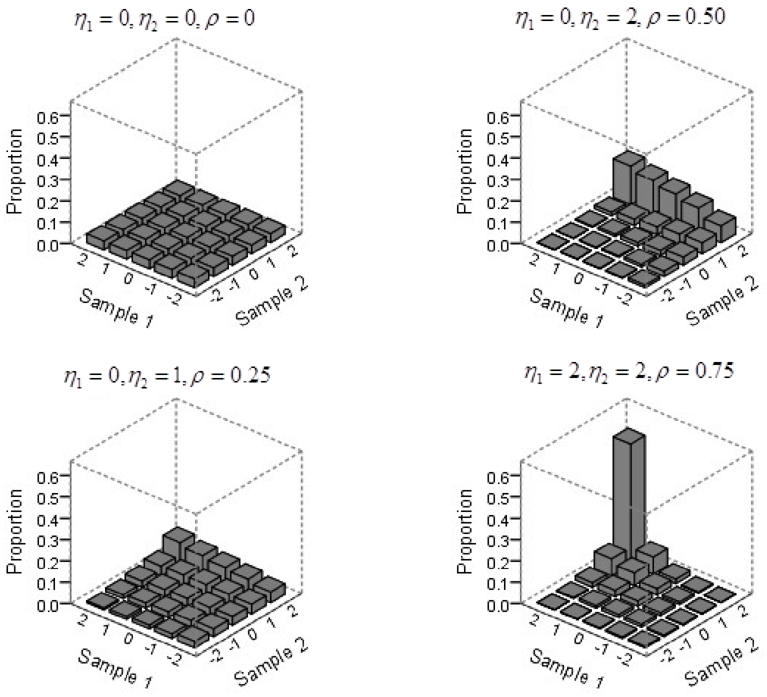


Figure 1. Theoretical distributions of the proportion of observed responses, for selected parameter combinations.

For non-parametric tests, exact p-values are difficult to obtain due to the frequent occurrence of ties for Likert data. When there are ties, the Normal approximation corrected for ties can be used to calculate p-values(Hollander, Wolfe and Chicken, 2013).The Normal approximations for both the Mann-Whitney test and the Wilcoxon test arevery accurate even for small sample sizes(Bellera, Julien and Hanley, 2010).The continuity correction factor is often used when approximating discrete distributions using the Normal distribution. The correction factor has little impact when $n \geq 10$ (Emerson and Moses, 1985). The non-parametric tests are performed using the Normal approximation with correction for ties. A continuity correction factor is also applied.Two-sided tests are performed at the nominal 5% significance level.

For each of the parametercombinations within the simulation design the process outlined above is repeated 10,000 times. The proportion of the 10,000 iterations where H_0 is rejectedis calculated. For the three scenarios in the simulation design where H_0 is true, the proportion of iterations where H_0 is rejected represents the Type I error rate. For the six scenarios in the simulation design where H_1 is true, the proportion of iterations where H_0 is rejected represents the power of the test.

3. RESULTS

Type I error rates for each of the test statistics are considered. This is followed by a summary of the power of the test statistics.

When the null hypothesis is true, H_0 rejection rates within the interval [0.025 , 0.075] are within Bradley's(1978) liberal limits. This Type I error robustness criteria is often used by researchers, although there is no consensus on the most appropriate criteria(Serlin, 2000). For the three scenarios in the simulation design where H_0 is true, the Type I error rates for each of the test statistics is given in Table 2.

Table 2. Type I error rates for selected combinations. For each parameter combination the test statistics within Bradley's (1978) liberal robustness criteria is highlighted in bold.

ρ	η_1	η_2	*n*	T_1	T_2	T_3	W_1	W_2	W_3	MW
0	0	0	10	**0.0528**	**0.0510**	**0.0498**	**0.0375**	**0.0487**	**0.0385**	**0.0441**
0	0	0	20	**0.0523**	**0.0513**	**0.0511**	**0.0466**	**0.0484**	**0.0464**	**0.0486**
0	0	0	30	**0.0494**	**0.0508**	**0.0508**	**0.0464**	**0.0501**	**0.0463**	**0.0494**
0	1	1	10	**0.0484**	**0.0498**	**0.0472**	**0.0344**	**0.0466**	**0.0356**	**0.0426**
0	1	1	20	**0.0549**	**0.0527**	**0.0524**	**0.0489**	**0.0506**	**0.0488**	**0.0509**
0	1	1	30	**0.0471**	**0.0486**	**0.0481**	**0.0447**	**0.0473**	**0.0446**	**0.0455**
0	2	2	10	**0.0352**	**0.0461**	**0.0313**	0.0168	**0.0570**	0.0185	**0.0441**
0	2	2	20	**0.0450**	**0.0460**	**0.0450**	**0.0350**	**0.0520**	**0.0400**	**0.0480**
0	2	2	30	**0.0410**	**0.0500**	**0.0500**	**0.0400**	**0.0440**	**0.0410**	**0.0490**
0.75	0	0	10	**0.0438**	0.0018	0.0018	0.0243	**0.0546**	**0.0268**	0.0014
0.75	0	0	20	**0.0498**	0.0005	0.0005	**0.0392**	**0.0482**	**0.0405**	0.0007
0.75	0	0	30	**0.0463**	0.0006	0.0006	**0.0432**	**0.0459**	**0.0438**	0.0005
0.75	1	1	10	**0.0381**	0.0014	0.0012	0.0207	**0.0514**	**0.0219**	0.0012
0.75	1	1	20	**0.0514**	0.0006	0.0006	**0.0398**	**0.0485**	**0.0406**	0.0004
0.75	1	1	30	**0.0468**	0.0009	0.0009	**0.0404**	**0.0439**	**0.0410**	0.0008
0.75	2	2	10	0.0221	0.0036	0.0025	0.0077	**0.0402**	0.0103	0.0036
0.75	2	2	20	**0.0460**	0.0050	0.0050	**0.0270**	**0.0470**	**0.0310**	0.0080
0.75	2	2	30	**0.0470**	0.0040	0.0040	**0.0380**	**0.0520**	**0.0400**	0.0050

Table 2 shows that all of the test statistics under consideration fulfil Bradley's Type I error robustness criteria when $\rho = 0$. As the correlation increases, test statistics assuming independent samples (T_2, T_3 and MW) do not maintain Type I error robustness. Test statistics assuming independent samples are valid when the structure of the data is unpaired, but appear biased when the structure of the data is paired.

Test statistics making use of paired information are robust across the range of ρ. Pratt's test (W_2) is Type I error robust for every combination of parameters under the simulation design. T_1, W_1 and W_3 are also generally Type I error robust, with minordeviations when the sample size is small ($n = 10$) and both samples are heavily skewed ($\eta_1 = 2, \eta_2 = 2$).

Figure 2 summarises for each test statistic the Type I error rates for all of the sample size and correlation coefficient combinations within the design. It can be seen from Figure 2 that the paired samples t-test (T_1) and Pratt's test (W_2) perform closest to the nominal Type I error rate of 5% across the simulation design.

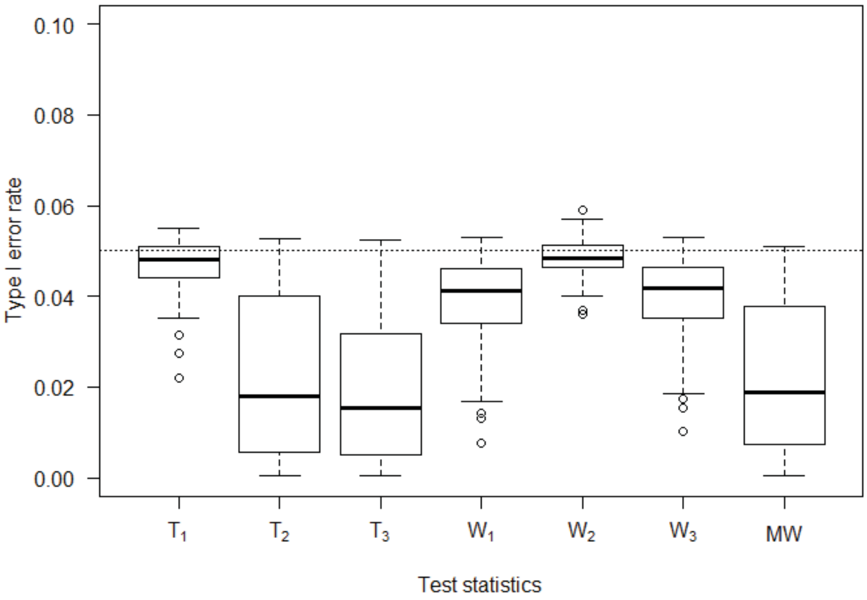


Figure 2. Type I error rates for each test statistic, averaged over each combination of parameters. The dotted line represents significance level of 5%.

Figure 2 demonstrates that each of the test statistics are generally conservative. A conservative test statistic is of less concern than a liberal test statistic (Mehta and Srinivasan, 1970). Alternative Type I error robustness criteria states that Type I error rates within the interval [0 , 0.055] are acceptable (Sullivan and D'Agostino, 1996). For all of the test statistics, each of the parameter combinations within the factorial design are within at least one of the mentioned Type I error robustness criteria. Hence, the power of each of the test statistics can be reasonably compared.

The power for each of the test statistics is given in Table 3, for the six scenarios in the simulation design where H_1 is true.

Table 3. Power for selected conditions. For each parameter combination the most powerful test is highlighted in bold.

ρ	η_1	η_2	n	T_1	T_2	T_3	W_1	W_2	W_3	MW
0	-1	1	10	0.5261	**0.5682**	0.5651	0.4587	0.4943	0.4594	0.5366
0	-1	1	20	0.8496	**0.8640**	0.8637	0.8283	0.8350	0.8292	0.8593
0	-1	1	30	0.9594	**0.9650**	**0.9650**	0.9534	0.9544	0.9535	0.9629
0	0	1	10	0.1742	**0.1905**	0.1873	0.1381	0.1631	0.1386	0.1707
0	0	1	20	0.3200	**0.3300**	0.3292	0.2956	0.3057	0.2965	0.3202
0	0	1	30	0.4522	**0.4631**	0.4627	0.4354	0.4404	0.4354	0.4573
0	0	2	10	0.6502	**0.7044**	0.6970	0.5745	0.6275	0.5791	0.6805
0	0	2	20	0.9415	**0.9497**	0.9494	0.9291	0.9338	0.9292	0.9488
0	0	2	30	0.9913	**0.9935**	0.9934	0.9897	0.9910	0.9900	0.9929
0.75	-1	1	10	0.9299	0.5935	0.5878	0.8741	**0.9353**	0.8808	0.5618
0.75	-1	1	20	**0.9989**	0.9508	0.9508	0.9988	0.9989	0.9988	0.9501
0.75	-1	1	30	**1.0000**	0.9967	0.9967	**1.0000**	**1.0000**	1.0000	0.9970
0.75	0	1	10	0.3974	0.0842	0.0819	0.3025	**0.4186**	0.3112	0.0764
0.75	0	1	20	**0.7496**	0.2342	0.2328	0.7119	0.7381	0.7164	0.2341
0.75	0	1	30	**0.9079**	0.4336	0.4334	0.8976	0.9000	0.8990	0.4292
0.75	0	2	10	0.9585	0.7548	0.7413	0.8986	**0.9706**	0.9114	0.7345
0.75	0	2	20	**0.9998**	0.9890	0.9888	0.9997	**0.9998**	0.9997	0.9896
0.75	0	2	30	**1.0000**	0.9996	0.9996	**1.0000**	**1.0000**	1.0000	0.9997

Table 3 shows that the power difference between the independent samples t-test (T_2) and Welch's test (T_3) is negligible. Additionally, there is little power differential between the traditional Wilcoxon test (W_1) and the Random ε method (W_3).

To summarise the power across the parameters within the simulation design, Figure 3 depicts how the test statistics T_1, T_2, W_1, W_2 and MW perform with increasing ρ for a small sample size of $n = 10$. Figure 4depicts how the test statistics T_1, T_2, W_1, W_2 and MW perform with increasing ρ for a larger sample size of $n = 20$.

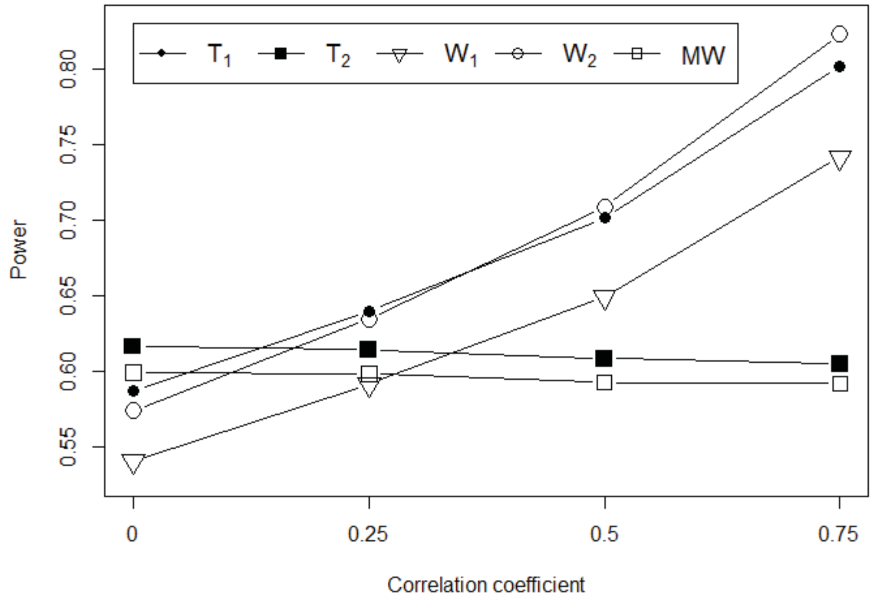


Figure 3. Power of the test statistics T_1, T_2, W_1, W_2 and MW where $n = 10$, averaged across each scenario within the simulation design.

Figure 3 shows that the independent samples t-test consistently out performs the Mann-Whitney test. When $\rho = 0$ the independent samples t-test is the recommended test of choice. When ρ >0 the paired samples t-test is more powerful than the independent samples t-test. These findings are consistent with the paired samples t-test and the independent samples t-test for continuous data (Fradette et. al. 2003; Zimmerman, 1997; Vonesh, 1983).

It can also be seen from Figure 3 that the standard Wilcoxon test consistently lacks power compared to Pratt's test and the paired samples t-test. When $\rho = 0.25$ the paired samples t-test is the most powerful test.As the correlation increases, Pratt's method becomes the test of choice.

As $\rho \to 1$ the power of both T_1 and W_2 increases. Given that both the paired samples t-test and Pratt's test havehigh power when the correlation is strong, the decision between the two tests is not of any major practical consequence in these circumstances.

Figure 4 shows that as sample size increases, thechoice between the Wilcoxon test, Pratt's test and the paired samples t-test becomes less important. Thesample size is large enough to compensate for discarded zeroes in the Wilcoxon test for $n \geq$ 20.

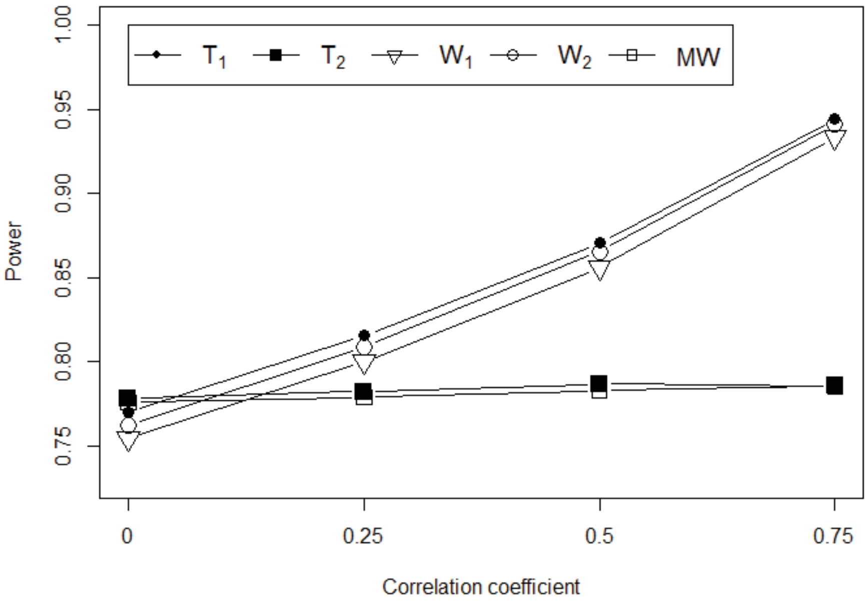


Figure 4. Power of the test statistics T_1, T_2, W_1, W_2 and MW where $n = 20$, averaged across each scenario within the simulation design.

4. CONCLUSION

Simulations have been performed based on an underlying continuum with a nonlinear transformation mapping to a five point equally spaced scoring scheme. The results indicate that parametric statistical procedures maintain good statistical properties for these data, i.e. the scores seemingly have interval like properties. This tends to suggest that if any real world application has a five point Likert scale designed to have perceived equally spaced categories, then the analyst may proceed with parametric approaches.

When comparing two independent samples on a five point Likert question, the independent samples t-test, Welch’s test and the Mann-Whitney test are Type I error robust. There is little practical difference between the power of these three tests. These findings support those in the literature (De Winter and Dodou, 2010; Rasch, Teuscher and Guiard, 2007).

When the structure of the experimental design includespaired observations, the independent samples t-test, Welch’s test and the Mann-Whitney test do not fulfil allType I error robustness definitions.

Nevertheless, these testsare conservative in natureandso their use may not be completely unjustified. However, these tests lack power in a paired design and are therefore not recommended, unless it is considered that the relationship between the two groups being compared is extremely small.

When sample sizes are large, there is little practical difference in the conclusions made from the paired samples t-test, the Wilcoxon test, or Pratt's test. When the sample size is large the choice becomes a more theoretical question about the exact form of the hypothesis being tested and the assumptions made.

When sample sizes are small and the correlation between two paired groups is strong, Pratt's test outperforms the paired samples t-test and the Wilcoxon test.When the correlation between the two groups is weak, the paired samples t-test outperforms the Wilcoxon test and Pratt's test.

5. REFERENCES

Allen, I. E., Seaman, C. A. 2007. Likert scales and data analyses. Quality Progress, **40(7)**, 64.

Bellera, C. A., Julien, M., Hanley, J. A. 2010. Normal approximations to the distributions of the wilcoxon statistics: Accurate to what N? graphical insights. Journal of Statistics Education, **18(2)**, 1-17.

Boone, H. N., Boone, D. A. 2012. Analyzinglikert data. Journal of Extension, **50(2)**, 1-5.

Box, G. E., Muller, M. E. 1958. A note on the generation of random normal deviates. The Annals of Mathematical Statistics, 29(2), 610-611.

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

Carifio, J., Perla, R. J. 2007. Ten common misunderstandings, misconceptions, persistent myths and urban legends about likert scales and likert response formats and their antidotes. Journal of Social Sciences, **3(3)**, 106-116.

Clason, D. L., Dormody, T. J. 1994. Analyzing data measured by individual likert-type items. Journal of Agricultural Education, **35**, 4.

Conover, W. J. 1973. On methods of handling ties in the Wilcoxon signed-rank test. *Journal of the* American Statistical Association,**68(344)**, 985-988.

De Winter, J. C., Dodou, D. 2010. Five-point likert items: T test versus mann-whitney-wilcoxon. Practical Assessment, Research Evaluation, **15(11)**, 1-12.

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

Emerson, J. D., Moses, L. E. 1985. A note on the wilcoxon-mann-whitney test for 2 xk ordered tables. Biometrics,**41(1)**, 303-309.

Fradette, K., Keselman, H., Lix, L., Algina, J., Wilcox, R. R. 2003. Conventional and robust paired and independent-samples t tests: Type I error and power rates. Journal of Modern Applied Statistical Methods, **2(2)**, 22.

Hollander, M., Wolfe, D. A., Chicken, E. 2013. Nonparametric statistical methods. John Wiley Sons.

Jamieson, S. 2004. Likert scales: How to (ab) use them. Medical Education, **38(12)**, 1217-1218.

Kenney, J. F., Keeping, E. S. 1951. Mathematics of Statistics; Part Two, Princeton, NJ: Van Nostrand.

Likert, R. 1932. A technique for the measurement of attitudes. Archives of Psychology.

Mehta, J., Srinivasan, R. 1970. On the Behrens—Fisher problem. Biometrika, **57(3)**, 649-655.

Nanna, M. J., Sawilowsky, S. S. 1998. Analysis of likert scale data in disability and medical rehabilitation research. Psychological Methods, **3(1)**, 55.

Norman, G. 2010. Likert scales, levels of measurement and the “laws” of statistics. Advances in Health Sciences Education, **15(5)**, 625-632.

Pratt, J. W. 1959. Remarks on zeros and ties in the wilcoxon signed rank procedures. Journal of the American Statistical Association, **54(287)**, 655-667.

R Core Team. 2014. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. www.R-project.org. version 3.1.3.

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

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

Sisson, D. V., Stocker, H. R. 1989. Research corner: Analyzing and interpreting likert-type survey data. Delta Pi Epsilon Journal, 31(2), 81.

Stevens, S. S. 1946. On the theory of scales of measurement. American Association for the Advancement of Science. **103(2684)**, 667-680.

Sullivan, G. M., Artino Jr, A. R. 2013. Analyzing and interpreting data from likert-type scales. Journal of Graduate Medical Education, **5(4)**, 541-542.

Sullivan, L. M., D'Agostino, R. B. 1992. Robustness of the t test applied to data distorted from normality by floor effects. Journal of Dental Research, **71(12)**, 1938-1943.

Vonesh, E. F. 1983. Efficiency of repeated measures designs versus completely randomized designs based on multiple comparisons. Communications in Statistics-Theory and Methods, **12(3)**, 289-301.

Wilcoxon, F. 1945. Individual comparisons by ranking methods. Biometrics Bulletin, **1(6)**, 80-83.

Zimmerman, D. W. 1997. Teacher’s corner: A note on interpretation of the paired-samples t test. Journal of Educational and Behavioral Statistics, **22(3)**, 349-360.

Appendix P10

Derrick, B. and White, P. (2018). “Methods for comparing the responses from a Likert question, with paired observations and independent observations in each of two samples”. *International Journal of Mathematics and Statistics* 19 (3), pp. 84–93
Published Version

International Journal of Mathematics and Statistics,
Year 2018; Volume 19; Issue Number 3;
ISSN 0974–7117 (Print); ISSN 0973-8347 (Online);

Methods for comparing the responses from a Likert question, with paired observations and independent observations in each of two samples

Ben Derrick, Paul White

Faculty of Environment and Technology, University of the West of England, Bristol, BS16 1QY, UK

Email: ben.derrick@uwe.ac.uk; paul.white@uwe.ac.uk

ABSTRACT

Researchers often encounter two samples of Likert data, which contain both independent observations and paired observations. Standard analyses in this scenario typically involve discarding the independent observations and performing the paired samples t-test, the Wilcoxon signed-rank test or the Pratt test. These naive approaches are examined alongside recently developed partially overlapping samples t-tests that make use of all of the available data in the two sample scenario. For two samples of observations from a Likert question with five categories or seven categories, test statistics are assessed for their Type I error robustness and power. A summary measure of Type I error robustness across the simulation design is quantified as that value of π such that $(1-\pi)\times100$ percent of Type I error rates are within $\pi\times100$ percent of the nominal significance level. Across a range of sample sizes and correlation coefficients, the partially overlapping samples t-tests are Type I error robust, and offer a more powerful alternative for the analysis of two samples including both paired observations and independent observations. In these scenarios, when the assumption of an underlying continuous distribution is not inappropriate, the partially overlapping samples t-test is recommended.

Keywords: Likert item; ordinal; partially overlapping samples; simulation; Type I error robustness

Mathematics Subject Classification: 60 62

1. INTRODUCTION

Situations arise when both paired observations and independent observations are present in a sample. This is referred to as a partially overlapping sample (Derrick *et al.*, 2015; Derrick, Toher and White, 2017). This paper evaluates tests used in the comparison of two partially overlapping samples. Previous literature in this area has focused on normally distributed data. The focus of this paper is for responses from an ordinal scale assuming equal spacing between the categories. The ordinal scales are represented by a five point Likert question, and a seven point Likert style question.

An example of two partially overlapping samples is a comparison of the responses of two Likert questions in the same survey, where some participants did not complete both questions, as obtained by Maisel and Fingerhut (2011). A further example of two partially overlapping samples for an individual Likert question is a comparison of the responses between pre-test and post-test, where some participants were not available at both times, as obtained by Bradley, Waliczek, and Zajicek

(1999). In both of these examples the authors discarded the unpaired observations and performed the paired samples *t*-test. Assuming data are missing completely at random (MCAR) this approach is not unjustified, given the large sample sizes obtained. However, power may be adversely affected for studies with smaller sample sizes.

Due to their intuitive appeal and simple construction, Likert questions are popular when measuring attitudes of respondents (Nunally, 1978). In certain methodological and practical aspects, Likert question responses may approximate interval level data, and can be analysed assuming an underlying continuous scale (Norman, 2010). Historically a Likert question consists of five options (Likert, 1932). The ordinal codes -2, -1, 0, 1, 2 could be applied to these options, with “0” representing the neutral response. In addition, seven point Likert style questions are commonly used, with ordinal codes -3, -2, -1, 0, 1, 2, 3.

Balanced and equally spaced response options around a neutral option are assumed for a valid Likert question (Uebersax, 2006). The exact wording of the neutral response is not an issue (Armstrong, 1987). If the options either side of the neutral response are not perceived to be balanced, then the assumption that responses approximate interval level data may not be reasonable (Bishop and Herron, 2015). Other issues with Likert questions include the responder tendency to give positive responses, and the potential for differing interpretation of the categorical options by both the responder and the analyst (Hodge and Gillespie, 2003). However, when the assumption of an underlying continuous distribution is not inappropriate and the questions are suitably formed, parametric tests for differences between the two sample means may be reasonable (Jamieson, 2004; Allen and Seaman, 2007; Derrick and White, 2017).

When comparing paired samples of ordinal data, the Wilcoxon signed-rank test can give dissimilar results to the paired samples *t*-test, and the correct choice of analysis depends on the exact form of the question of interest (Roberson *et al.*, 1994). Non-parametric tests are not inappropriate when interval approximating data is assumed, if the only potential difference between the samples is their central location (Clason and Dormody, 1994; Sisson and Stocker, 1989). Given the discrete nature of Likert questions, zero differences between pairs occur frequently. The Pratt (1959) test, which incorporates zero-differences in its calculation, can overcome the issues of the Wilcoxon signed-rank test which discards zero differences (Conover, 1973; Derrick and White, 2017).

The Pratt test, the Wilcoxon signed-rank test and the paired samples *t*-test are easily extended for the use when two partially overlapping samples are present, if the researcher is willing to discard any unpaired data. However, the discarding of data may introduce bias and reduce power and as is therefore a naïve approach (Guo and Yuan, 2017). Researchers should ensure that the analyses correctly reflect the design of the study. An alternative approach is the partially overlapping samples *t*-tests proposed by Derrick *et al.* (2017). These solutions are generalized forms of the *t*-test and have the advantage of making use of all of the available data. These solutions are Type I error robust under the assumptions of normality and MCAR, and are more powerful than alternatives that discard data (Derrick, Toher and White, 2017). The partially overlapping samples *t*-tests were previously considered for normally distributed data, the properties for ordinal data were not discussed.

The partially overlapping samples *t*-tests are given (Section 2) and demonstrated by example (Section 3). Methodology for comparing these proposals, the paired samples *t*-test, the Wilcoxon signed-rank test, and the Pratt test, is outlined for a five point Likert question and a seven point Likert style question (Section 4). Type I error robustness and power of the test statistics are assessed for scenarios where there are two partially overlapping samples (Section 5).

2. THE PARTIALLY OVERLAPPING SAMPLES *T*-TESTS

For situations comprising of a combination of both paired observations and unpaired observations for two samples, let 'n_a' represent the number of observations only in Group 1, and 'n_b' represent the number of observations only in Group 2 and 'n_c' represent the number of paired observations. It follows that the total sample size in Group 1 is $n_1 = n_a + n_c$, and the total sample size in Group 2 is $n_2 = n_b + n_c$.

There are two versions of the partially overlapping samples *t*-test, T_{new1} assumes equal variances between the two groups, and T_{new2} makes use of separate variances. For equal variances assumed the partially overlapping samples *t*-test by Derrick *et al.* (2017) is defined as:

$$T_{\text{new1}} = \frac{\overline{X}_1 - \overline{X}_2}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2} - 2r\left(\frac{n_c}{n_1 n_2}\right)}}$$

where $S_p = \sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{(n_1-1)+(n_2-1)}}$ and *r* is Pearson's correlation coefficient.

The test statistic T_{new1} is referenced against the t-distribution with degrees of freedom:

$$v_1 = (n_c - 1) + \left(\frac{n_a + n_b + n_c - 1}{n_a + n_b + 2n_c}\right)(n_a + n_b).$$

For the comparison of samples on a continuous scale, when variances are unequal, Welch's test has superior Type I error robustness properties (Derrick, Toher and White, 2016). The statistic T_{new2} uses Welch's approximation to degrees of freedom and is defined by Derrick *et al.* (2017) as:

$$T_{\text{new2}} = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2} - 2r\left(\frac{S_1 S_2 n_c}{n_1 n_2}\right)}}$$

The test statistic T_{new2} is referenced against the *t*-distribution with degrees of freedom:

$$v_2 = (n_c - 1) + \left(\frac{\gamma - n_c + 1}{n_a + n_b + 2n_c} \right) (n_a + n_b) \text{ where } \gamma = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2} \right)^2}{\left(\frac{S_1^2}{n_1} \right)^2 / (n_1 - 1) + \left(\frac{S_2^2}{n_2} \right)^2 / (n_2 - 1)} .$$

3. WORKED EXAMPLE

For a university undergraduate course, student satisfaction for a Mathematics module (Group 1) is compared against that of a Statistics module (Group 2). Most students under consideration study both modules, however some study only one. For each group, an online module evaluation is given to the students with the question “I am overall satisfied with the module”. The answers obtained are given in Table 1.

Table 1. Responses to the question “I am overall satisfied with the module”. Results coded as; Strongly Agree = 2, Agree = 1, Neither agree nor disagree = 0, Disagree = -1, Strongly Disagree = -2.

Unique ID	1	2	3	4	5	6	7	8	9	10	11	12	13
Group 1	0	0	2	0	-1		1	-1		0		1	2
Group 2	0		2	0	1	2	1	2	2	2	1		2

Using the convention of lower case for calculated sample values; $\bar{x}_1 - \bar{x}_2 =$ -0.964, $n_a = 2$, $n_b = 3$, $n_c = 8$, $T_{\text{new1}} =$ -2.666, $v_1 = 9.857$, $T_{\text{new2}} =$ -2.609, $v_2 = 9.304$. Using critical *t*-values at the 5% significance level (two sided) gives evidence to suggest than the module means are different. Students appear to be more satisfied with the Statistics module relative to the Mathematics module. Or more simply, using the R ‘partiallyoverlapping’ package (Derrick, 2017; Derrick, Toher and White, 2017), the p-values for T_{new1} and T_{new2} are 0.024 and 0.028 respectively.

Both forms of the partially overlapping samples t-test provide evidence to suggest that there is a significant difference between the two groups. This is in contrast to the conclusions that are made if performing standard tests that ignore the unpaired observations [paired samples *t*-test, *p*=0.088; Wilcoxon signed-rank test, *p*=0.174; Pratt test *p*=0.085].

4. METHODOLOGY

Monte-Carlo simulation methods are used to compare test statistics for the comparison of two samples which include both paired observations and unpaired observations. The standard tests considered are the paired samples *t*-test T_{paired}, the Wilcoxon signed-rank test W_1, and the Pratt test W_2. For each of these standard tests, randomly generated independent observations are ignored. The standard tests are compared against the proposals by Derrick *et al.* (2017), T_{new1} and T_{new2}. Small sample sizes are of particular interest.

The simulation is undertaken by discretizing realizations from bivariate Normal distributions to a five point scale and a seven point scale. This is done over a range of sample sizes {n_a, n_b, n_c} and non-negative Pearson's correlation coefficients {ρ}.

For the n_a independent observations in Group 1, the Mersenne-Twister algorithm (Matsumoto and Nishimura, 1998) generates pairs of random *U*(0,1) deviates and are transformed into Standard Normal deviates using the Box and Muller (1958) transformation. This process is repeated to generate the n_b independent observations in Group 2. For the n_c paired observations, additional Standard Normal deviates are generated, and these are transformed into correlated Standard Normal bivariates using methodology outlined by Kenney and Keeping (1951).

Let Standard Normal deviates be y_{ij} to denote the *i*-th observation in group *j*. Without loss of generality, for a five point scale the points are numbered from -2 to 2. The responses x_{ij} are calculated using the cut-points as follows:

$$x_{ij} = \left\{ \begin{array}{lll} 2 & \text{if} & y_{ij} > 0.8416 \\ 1 & \text{if} & 0.2533 \le y_{ij} \le 0.8416 \\ 0 & \text{if} & -0.2533 \le y_{ij} \le 0.2533 \\ -1 & \text{if} & -0.8416 \le y_{ij} \le -0.2533 \\ -2 & \text{if} & y_{ij} < \text{-}0.8416 \end{array} \right\}$$

The cut-points are calculated so that under the Standard Normal distribution the theoretical distribution of the responses is uniform. Similarly, for a seven point scale, x_{ij} are calculated using the cut-points as follows:

$$x_{ij} = \left\{ \begin{array}{lll} 3 & \text{if} & y_{ij} > 1.6757 \\ 2 & \text{if} & 0.5659 \le y_{ij} \le 1.6757 \\ 1 & \text{if} & 0.1800 \le y_{ij} \le 0.5659 \\ 0 & \text{if} & -0.1800 \le y_{ij} \le 0.1800 \\ \text{-}1 & \text{if} & -0.5659 \le y_{ij} \le -0.1800 \\ \text{-}2 & \text{if} & -1.6757 \le y_{ij} \le -0.5659 \\ \text{-}3 & \text{if} & y_{ij} < -1.6757 \end{array} \right\}$$

The median of Group 1 and the median of Group 2 are represented by η_1 and η_2 respectively. The scenarios compared encompass each integer value of η_1 and η_2. For example, by symmetry the

Type I error robustness when $\eta_1 = \eta_2 = 1$ is equivalent to Type I error robustness where $\eta_1 = \eta_2 =$-1. The complete list of scenarios, parameters and the test statistics compared, can be found in Table 2. The scenarios and parameter combination are considered as part of a factorial design. For each scenario and parameter combination, the number generating process is repeated 10,000 times. For each repetition the null hypothesis is assessed at the $\alpha = 5\%$ significance level, two sided.

Table 2. Simulation design

Sample size	$n_a =$ (5, 10, 20, 30), $n_b =$(5, 10, 20, 30), $n_c =$ (5, 10, 20, 30)									
ρ	0.00, 0.25, 0.50, 0.75									
Scenarios under $\eta_1 = \eta_2$	Five point scale					Seven point scale				
		μ_1	μ_2	η_1	η_2		μ_1	μ_2	η_1	η_2
	i)	0	0	0	0	*x)*	0	0	0	0
	ii)	0.5244	0.5244	1	1	*xi)*	0.3661	0.3661	1	1
	iii)	1.2816	1.2816	2	2	*xii)*	0.7916	0.7916	2	2
						xiii)	1.4652	1.4652	3	3
Scenarios under $\eta_1 \neq \eta_2$	Five point scale					Seven point scale				
		μ_1	μ_2	η_1	η_2		μ_1	μ_2	η_1	η_2
	iv)	0	0.5244	0	1	*xiv)*	0	0.3661	0	1
	v)	0	1.2816	0	2	*xv)*	0	0.7916	0	2
	vi)	0.5244	1.2816	1	2	*xvi)*	0	1.4652	0	3
	vii)	−0.5244	0.5244	−1	1	*xvii)*	0.3661	0.7916	1	2
	viii)	−0.5244	1.2816	−1	2	*xviii)*	0.3661	1.4652	1	3
	ix)	−1.2816	1.2816	−2	2	*xix)*	0.7916	1.4652	2	3
						xx)	−0.3661	0.3661	−1	1
						xxi)	−0.3661	0.7916	−1	2
						xxii)	−0.3661	1.4652	−1	3
						xxiii)	−0.7916	0.7916	−2	2
						xxiv)	−0.7916	1.4652	−2	3
						xxv)	−1.4652	1.4652	−3	3
Test Statistics	T_{paired}	Paired samples *t*-test								
	W_1	Wilcoxon signed-rank test (Standard method, discarding zeroes)								
	W_2	Pratt test (Wilcoxon signed-rank test, with Pratt's zeroes modification)								
	T_{new1}	Partially overlapping samples *t*-test with equal variances								
	T_{new2}	Partially overlapping samples *t*-test with unequal variances.								

Number of iterations: 10,000
Significance level: $\alpha = 0.05$

All calculations are performed in R. The paired samples *t*-test is calculated using the 'stats' package (R core team, 2015). The Wilcoxon signed-rank test is calculated using the Normal approximation corrected for ties with continuity correction factor, using the 'stats' package (R core team, 2015). The Pratt test is calculated under the same conditions using the 'coin' package (Hothorn, 2017). The partially overlapping samples *t*-tests are calculated using the 'partiallyoverlapping' package (Derrick, 2017).

5. RESULTS

For each parameter combination where $\eta_1 = \eta_2$, the proportion of the 10,000 iterations where the null hypothesis is rejected, represents the Type I error rate of the test under those conditions.

For selected parameter combinations, Type I error rates are given in Table 3. Liberal robustness criteria by Bradley (1978), offers guidance for assessing the Type I error rate for a given parameter combination. Under this criteria, Type I error robust statistics are within 50% of the nominal Type I error rate. For each parameter combination given in the table, the Type I error rates where $0.025 \leq \alpha \leq 0.075$ are highlighted in bold.

A summary measure of Type I error robustness across the entire simulation design for each of the test statistics is additionally put forward. The overall Type I error robustness is quantified as that value of π such that (1-π)×100 percent of Type I error rates are within π×100 percent of α. Large values of (1-π) are desirable. Table 3 shows the overall Type I error robustness of each of the test statistics.

Table 3. Type I error rates for selected parameter combinations, and overall robustness (1-π) across the simulation design, where $\eta_1 = \eta_2$.

	η_1	η_2	n_a	n_b	n_c	ρ	T_{paired}	W_1	W_2	T_{new1}	T_{new2}
Five point scale	0	0	5	5	5	0.5	**.042**	.010	.019	**.041**	**.038**
	0	0	5	20	10	0.5	**.052**	**.033**	**.051**	**.041**	**.044**
	1	1	5	5	5	0.5	**.040**	.006	.013	**.048**	**.041**
	1	1	5	20	10	0.5	**.045**	**.028**	**.049**	**.046**	**.050**
	2	2	5	5	5	0.5	.010	.001	.003	**.037**	.024
	2	2	5	20	10	0.5	**.027**	.001	**.044**	**.041**	**.052**
Value of (1-π) over all parameter combinations							.747	.584	.721	.821	.814
Seven point scale	0	0	5	5	5	0.5	**.046**	.008	.023	**.042**	**.048**
	0	0	5	20	10	0.5	**.049**	**.034**	**.048**	**.050**	**.050**
	1	1	5	5	5	0.5	**.044**	.005	.020	**.046**	**.041**
	1	1	5	20	10	0.5	**.049**	**.035**	**.051**	**.047**	**.047**
	2	2	5	5	5	0.5	**.026**	.001	.008	**.426**	**.032**
	2	2	5	20	10	0.5	**.041**	**.027**	**.046**	**.042**	**.045**
	3	3	5	5	5	0.5	.006	.000	.002	**.041**	.022
	3	3	5	20	10	0.5	**.026**	.001	**.055**	**.048**	**.053**
Value of (1-π) over all parameter combinations							.793	.622	.749	.871	.848

Table 3 shows that T_{new1} performs within Bradley's liberal Type I error robustness criteria, this remains true for the smallest sample size combination within the simulation design. For each of the other test statistics considered, Type I error robustness is not always maintained when both groups are heavily skewed. The Pratt test better controls for Type I error rates than the standard Wilcoxon signed-rank test.

In summary, for a five point Likert scale, the paired samples *t*-test is 74.7% robust. This means that 74.7% of the Type I error rates for parameter combinations within the simulation design are within

25.3% of the nominal Type I error rate. In this design, the paired samples *t*-test therefore maintains greater Type I error robustness than the Wilcoxon test (62.3%) or the Pratt test (70.6%). Both T_{new1} and T_{new2} are over 80% robust, thus maintain Type I error robustness better than the other tests considered.

The Type I error rates follow a similar pattern whether a five point scale or a seven point scale is used.

For all parameter combinations where $\eta_1 \neq \eta_2$, the percentage of iterations where the null hypothesis is rejected, represents the power of the test. Table 4 summarizes the power for each scenario using each test statistic, averaged over all parameter combinations.

Table 4. Power for each test statistic averaged over all scenarios where $\eta_1 \neq \eta_2$.

	Scenario	η_1	η_2	T_{paired}	W_1	W_2	T_{new1}	T_{new2}
Five point scale	*iv*	0	1	.380	.327	.358	.530	.524
	v	0	2	.788	.679	.740	.972	.966
	vi	1	2	.503	.440	.492	.734	.723
	vii	-1	1	.744	.642	.698	.937	.931
	viii	-1	2	.916	.746	.855	.998	.998
	ix	2	-2	.983	.779	.946	1.000	1.000
	Average			**.747**	**.623**	**.706**	**.882**	**.877**
Seven point scale	*xiv*	0	1	.244	.206	.227	.323	.319
	xv	0	2	.611	.536	.579	.821	.812
	xvi	0	3	.840	.713	.794	.989	.986
	xvii	1	2	.285	.244	.269	.392	.387
	xviii	1	3	.713	.628	.682	.933	.923
	xix	2	3	.433	.384	.431	.646	.634
	xx	-1	1	.582	.509	.550	.782	.774
	xxi	-1	2	.794	.677	.752	.964	.959
	xxii	-1	3	.918	.743	.871	.999	.998
	xxiii	-2	2	.899	.733	.856	.996	.995
	xxiv	-2	3	.967	.754	.931	1.000	1.000
	xxv	-3	3	.994	.773	.977	1.000	1.000
	Average			**.690**	**.575**	**.660**	**.820**	**.816**

Table 4 shows that T_{new1} and T_{new2} both consistently out-perform the standard tests which discard data. T_{new1} demonstrates marginally superior Type I error robustness and power properties relative to T_{new2}.

6. CONCLUSION

This paper has used simulation to compare the performance of test statistics where there are two samples, each sample with both paired observations and independent observations. This comparison has been performed for ordinal data, specifically for responses from either a five point Likert question, or a seven point Likert style question. Assuming the responses represent interval data, standard approaches such as the paired samples *t*-test or the Pratt test may not be inappropriate. However, these standard approaches discard the independent observations and as such are less than ideal, particularly if the sample sizes are small.

The partially overlapping samples *t*-tests proposed by Derrick *et al.* (2017) overcome the issue of discarding data. It is demonstrated that T_{new1} exhibits superior Type I error robustness relative to the other test statistics considered, and also has greater power. Therefore when the underlying assumptions of interval data are met, T_{new1} is recommended as the test of choice when comparing the responses from a Likert question with paired observations and independent observations in each of two samples.

7. REFERENCES

Allen, I. E., & Seaman, C. A. (2007). Likert scales and data analyses. *Quality Progress*, **40(7)**, 64-65.

Armstrong, R. L. (1987). The midpoint on a five-point Likert-type scale. *Perceptual and Motor Skills*, **64(2)**, 359-362. doi.org/10.2466/pms.1987.64.2.359

Bishop, P. A., & Herron, R. L. (2015). Use and misuse of the Likert item responses and other ordinal measures. *International Journal of Exercise Science*, **8(3)**, 297-302.

Box, G. E. P., & Muller, M. (1958). A note on the generation of random normal deviates. *Annals of Mathematical Statistics*, **29**, 610-611.

Bradley, J. C., Waliczek, T. M., & Zajicek, J. M. (1999). Relationship between environmental knowledge and environmental attitude of high school students. *The Journal of Environmental Education*, **30(3)**, 17-21. doi.org/10.1080/00958969909601873

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

Clason, D. L., & Dormody, T. J. (1994). Analyzing data measured by individual Likert-type items. *Journal of Agricultural Education*, **35(4)**, 32-35.

Conover, W. J. (1973). On methods of handling ties in the Wilcoxon signed-rank test. *Journal of the American Statistical Association*, **68(344)**, 985-988.

Derrick, B. (2017). Package Partiallyoverlapping: Partially overlapping samples t-tests. R package version 1.0.

Derrick, B., Dobson-McKittrick, A., Toher, D., & White P. (2015). Test statistics for comparing two proportions with partially overlapping samples. *Journal of Applied Quantitative Methods,* **10(3)**, 1-14.

Derrick, B., Russ, B., Toher, D., & White P. (2017). Test statistics for the comparison of means for two samples which include both paired observations and independent observations. *Journal of Modern Applied Statistical Methods*, **16(1)**, 137-157. doi.org/10.22237/jmasm/1493597280

Derrick, B., Toher, D., & White, P. (2016). Why Welch's test is Type I error robust. *The Quantitative Methods for Psychology*, **12(1)**, 30-38. doi.org/10.20982/tqmp.12.1.p030

Derrick, B., Toher, D., & White, P. (2017). How to compare the means of two samples that include paired observations and independent observations: A companion to Derrick, Russ, Toher and White (2017). *The Quantitative Methods for Psychology,* **13(2)**, 120-126. doi.org/10.20982/tqmp.13.2.p120

Derrick, B., & White, P. (2017) Comparing two samples from an individual Likert question. *International Journal of Mathematics and Statistics*, **18(3)**, 1-13.

Guo, B., & Yuan, Y. (2017). A comparative review of methods for comparing means using partially paired data. *Statistical methods in medical research*, **26(3)**, 1323-1340.

Hodge, D. R., & Gillespie, D. (2003). Phrase completions: An alternative to Likert scales. *Social Work Research*, **27(1)**, 45-55.

Hothorn, K. (2006). A Lego System for Conditional Inference. *The American Statistician*, **60(3)**, 257-263. doi.org/10.1198/000313006X118430

Jamieson, S. (2004). Likert scales: how to (ab)use them. *Medical Education*, **38(12)**, 1217-1218. doi.org/10.1111/j.1365-2929.2004.02012.x

Kenney, J. F., & Keeping E.S. (1951) *Mathematics of Statistics*, (2nd ed.), Princeton, NJ: Van Nostrand.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of psychology*, **22**.

Maisel, N. C., & Fingerhut, A. W. (2011). California's ban on same-sex marriage: The campaign and its effects on gay, lesbian, and bisexual individuals. *Journal of Social Issues*, **67(2)**, 242-263. doi.org/10.1111/j.1540-4560.2011.01696.x

Matsumoto, M., & Nishimura, T. (1998). Mersenne twister: a 623-dimensionally equidistributed uniform pseudo-random number generator. *ACM Transactions on Modeling and Computer Simulation*, **8(1)**, 3-30. doi.org/10.1145/272991.272995

Norman G (2010). Likert scales, levels of measurement and the “laws” of statistics. *Advances in Health Science Education Theory and Practice,* **15(5)**, 625–632. doi.org/10.1007/s10459-010-9222-y

Nunnally, J. C. (1978). *Psychometric theory,* (2nd ed.), New York: McGraw-Hill.

Pratt, J. W. (1959). Remarks on zeros and ties in the Wilcoxon signed rank procedures. *Journal of the American Statistical Association*, **54(287)**, 655-667. doi.org/10.1080/01621459.1959.10501526

R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Roberson, P. K., Shema, S. J., Mundfrom, D. J., & Holmes, T. M. (1994). Analysis of paired Likert data: how to evaluate change and preference questions. *Family Medicine*, **27(10)**, 671-675.

Sisson, D. V., & Stocker, H. R. (1989). Research corner: analyzing and interpreting Likert-type survey data. *Delta Pi Epsilon Journal*, **31(2)**, 81-85.

Uebersax, J. S. (2006). Likert scales: dispelling the confusion. Statistical Methods for Rater Agreement. Available at: http://john-uebersax.com/stat/likert.htm. Accessed: 08082017.

Glossary

ANOVA	Analysis of Variance
Brown-Forsythe test (BF)	Brown-Forsythe test for equal variances using absolute deviations from the median
Cov	Covariance
CI	Confidence Interval
INT	Inverse Normal Transformation
KS	Kolmogorov-Smirnov test for normality
Levene's test (L)	Levene's test for equal variances using absolute deviations from the mean
MAR	Missing at random
MCAR	Missing completely at random
MNAR	Missing not at random
NA	Not applicable or missing
NHRR	Null hypothesis rejection rate
NHST	Null hypothesis significance testing
P-P	Probability-Probability
PM	Pitman-Morgan test for equal variances
REML	Restricted Maximum Likelihood
SW	Shapiro-Wilk test for normality
Var	Variance