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, examples include Swinscow (2002) and OpenStax (2013). 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). An example 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. 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.

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). Procedures specifically attempting to act as a weighting between the paired samples t-test and the independent samples t-test under normality were 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 appropriate 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.

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:

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

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:

$$T_{\text{new2}} = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2} - 2r\left(\frac{S_1 S_2 n_c}{n_1 n_2}\right)}}$$

The test statistic T_{new2} is referenced against the t-distribution with degrees of freedom:

$$v_2 = (n_c - 1) + \left(\frac{\gamma - n_c + 1}{n_a + n_b + 2n_c}\right)(n_a + n_b) \text{ where } \gamma = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{\left(S_1^2 / n_1\right)^2}{n_1 - 1} + \frac{\left(S_2^2 / n_2\right)^2}{n_2 - 1}}$$

These solutions are easily performed using the R package ‘Partiallyoverllaping’ (Derrick, 2017) as introduced and explained 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 $\alpha_{\text{nominal}} = 0.05$. For each of the scenarios in Table 1, two sided tests are performed and the null hypothesis rejection rate is recorded as 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 above transformations outlined 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 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.

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 give Type I error rates for all 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 ensure 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 are added to represent Bradley's liberal Type I error robustness criteria.

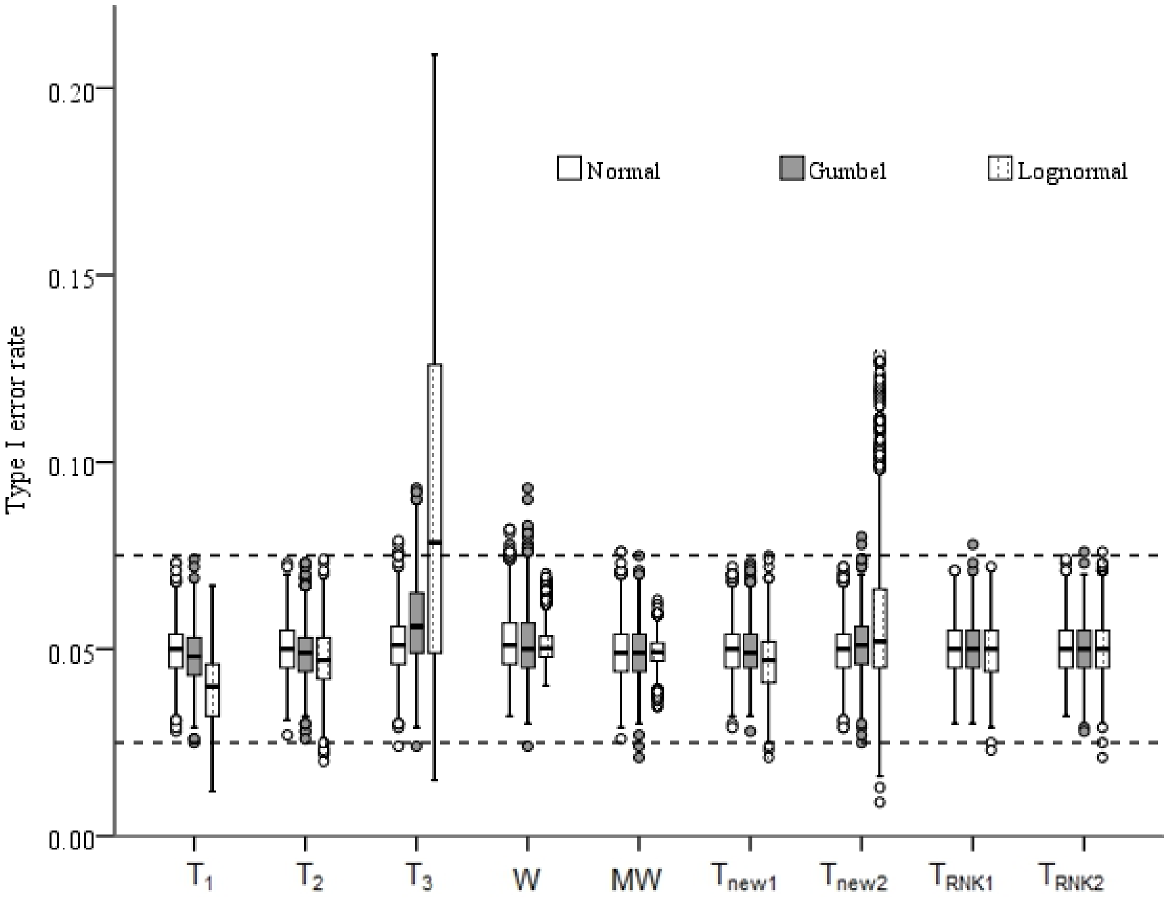

Figure 1. Type I error rates for when both samples are taken from the standard Normal distribution.

Figure 1 shows that when two samples are drawn from a Normal distribution with equal variances, 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, there is some minor inflation of Type I error rates. 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 demonstrate the performance of the test under extreme scenarios, Table 2 shows 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.

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 $T_{\text{new1}}$and T_{RNK1}.

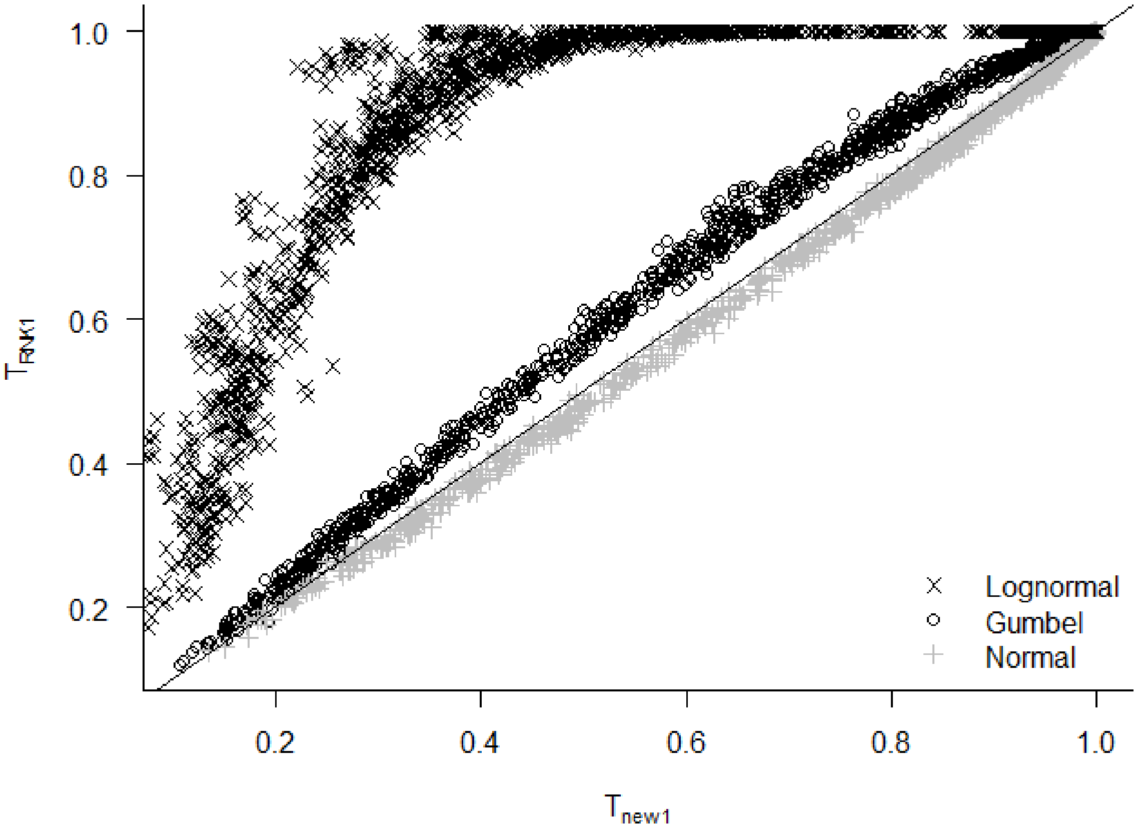

Figure 2. Power for each parameter combination, for T_{new1} and T_{RNK1}.

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 evident 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 gives 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.

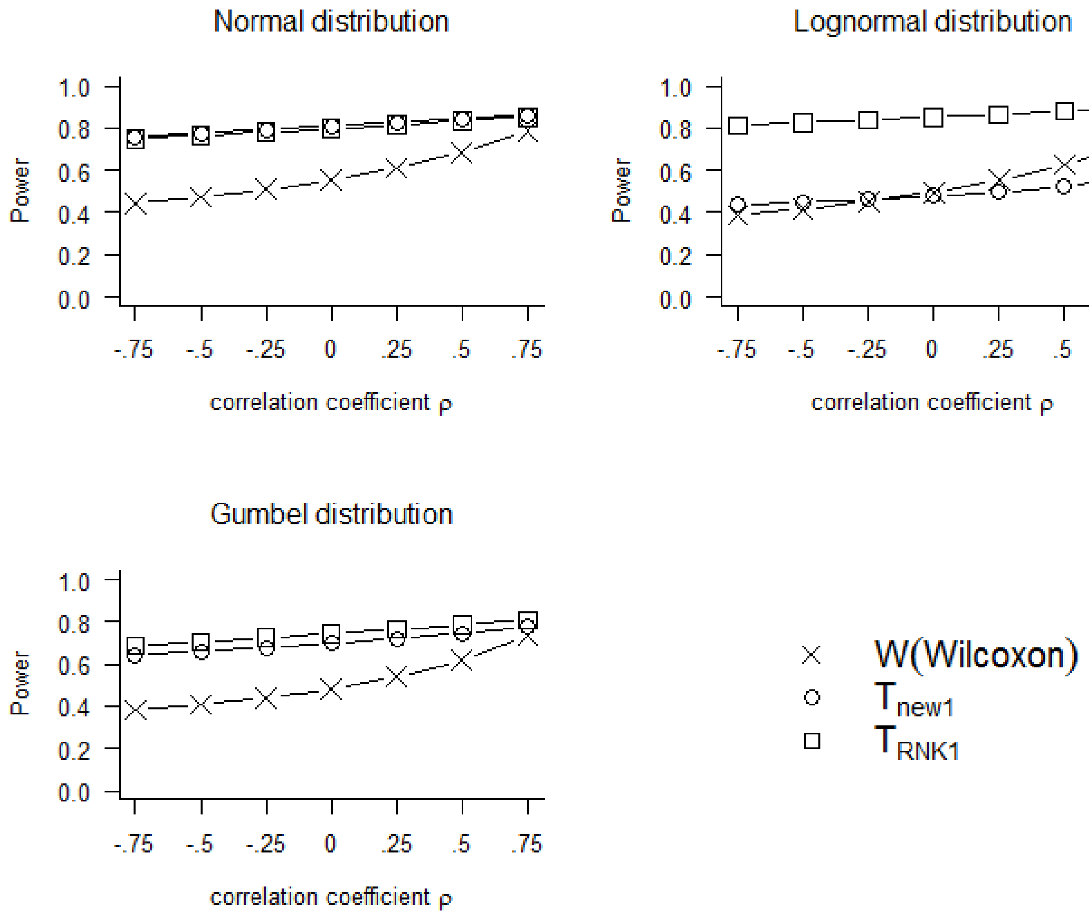

Figure 3. Power of selected test statistics making use of paired data, for two N(0,1) samples.

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.

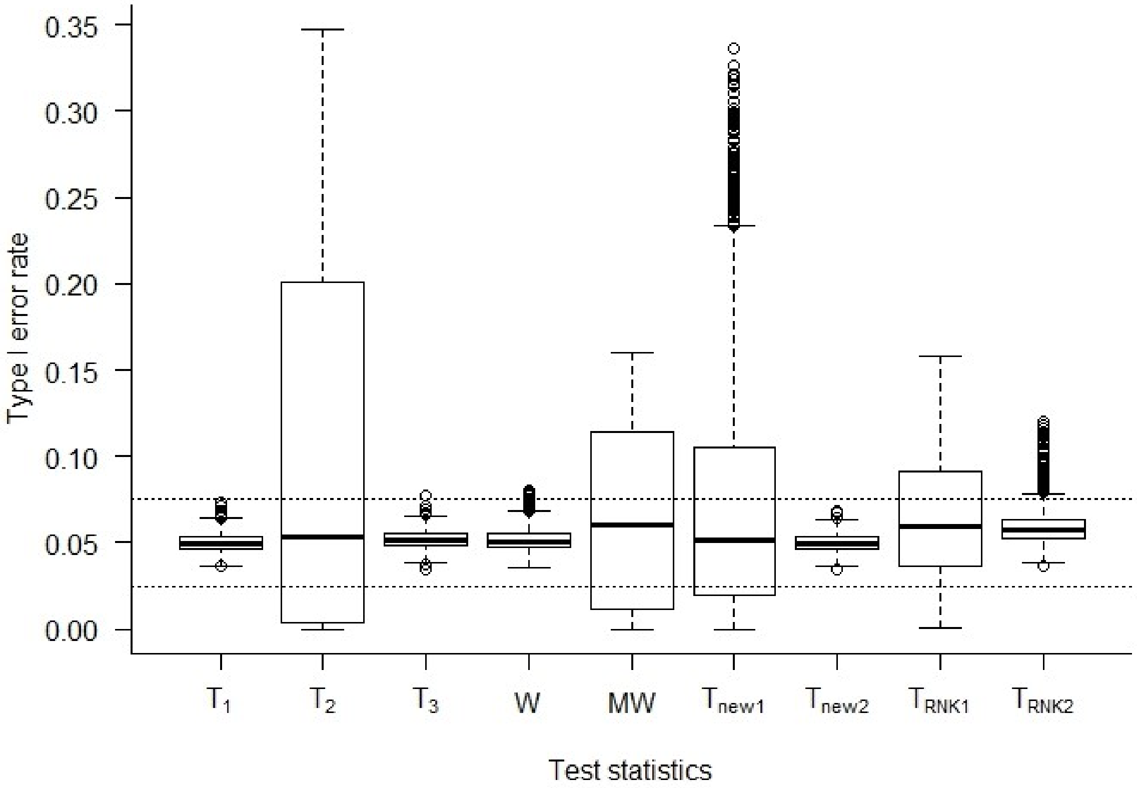

Figure 4. Type I error rates for samples from the Normal distribution with $\sigma_1^2 = 1$, $\sigma_2^2 = 4$.

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.

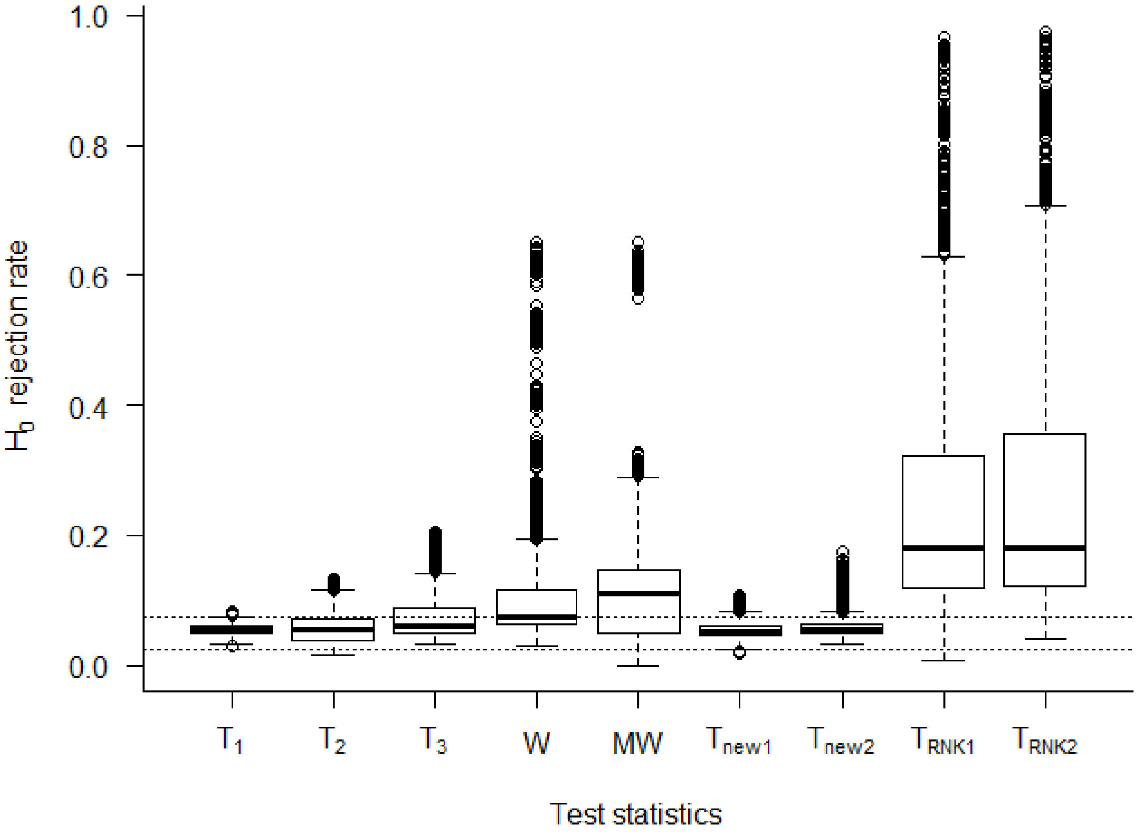

Figure 5. Sample 1 values taken from the standard Normal distribution, Sample 2 observations are taken from a standardised Lognormal distribution.

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 whether the distributions are equal. 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:
10,20,25,30,20,30,15,20,30,15,15,20,10,25,30,20,20,30,25,30,20,20,10,25,20,10,20,20
Patients with scores only on the last day:
15,25,30,20,10,20,10,30,10,10,10,25,15,20,20,20,20,10,10,10,20,30,10

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 that MCAR is true, 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 known 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

$T_{\text{new}1}$ and $T_{\text{new}2}$ 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.

OpenStax (2013), *Introductory Statistics*. OpenStax, Chapter 10.

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).

Swinscow T. D. V., & Campbell, M. J. (2002). *Statistics at square one*. London: BMJ, Chapter 7.

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.