2022 ■ Vol. 18 ■ no. 1

Review of the partially overlapping samples framework: Paired observations and independent observations in two samples

Ben Derrick [a] and Paul White [a]

[a]University of the West of England, Bristol

Abstract ■ A frequently asked question in quantitative research is how to compare two samples that include some combination of paired observations and unpaired observations. In our publications and R package, we refer to the scenario 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. In the 20th century, traditional approaches that discard either the paired observations or the independent observations were customary. In the 21st century approaches that make use of all available data are becoming more prominent. Traditional and modern approaches for the analyses for each of these research questions are reviewed. We conclude that tests that report a directly measurable difference between the two groups provide the best solutions.

Keywords ■ missing observations, paired samples, partially overlapping samples, partially paired data. **Tools** ■ R, Partiallyoverlapping.

paul.white@uwe.ac.uk

10.20982/tqmp.18.1.p055

Acting Editor ■ Roland Pfister (Universität Würzburg)

Reviewers ■ One anonymous reviewer

Introduction

In tests of differences, the most frequent desired comparison is a comparison of group means. In general, the comparisons of central location or distributions, is a source of much confusion in the literature, because these are often mistreated as having the same null hypothesis (Fagerland & Sandvik, 2009). Non-parametric tests designed to assess differences in central location are appropriate when the distributions are the same or with a simple location shift (Rietveld & van Hout, 2017). Non-parametric tests designed to be sensitive to location effects, may also be sensitive to other distributional differences, including those situations when the central location is the same but with different distributions (Hart, 2001) and as such clarity over the precise form of the statistical null hypothesis is needed along with clarity over the distributional assumptions.

Within the partially overlapping samples framework, the research question and related statistical hypotheses may relate to a comparison of central locations such as means or medians, or of distributions, or proportions or variances as appropriate.

The avoidance of partially overlapping samples is desirable through good study design. In general, paired samples designs can be advantageous relative to independent samples designs (Amro, Konietschke, & Pauly, 2019). However, real-world paired designs rarely result in only complete pairs being collected (Derrick, Dobson-Mckittrick, Toher, & White, 2015).

To demonstrate the breadth of the two partially overlapping samples design, the following list illustrates some situations which may be encountered:

A paired samples design, which inadvertently contains independent observations. The most frequent occurrence of partially overlapping samples is a paired samples design with missing observations (Martinez-Camblor, Corral, & De La Hera, 2013; Ramosaj, Amro, & Pauly, 2020). If observations are missing due to equipment failure or loss in transit, the data could potentially be regarded as being MCAR (Kang, 2013). A likely general scenario is missing data due to participant drop out leading to unpaired data in a paired design, and participant attrition could lead to

independent observations in one sample only (e.g. complete baseline data, but some data missing at follow-up).

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 on two occasions, 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 participants that are only available to provide a response on one of the two occasions, thus forming observations which cannot be paired.

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, and these participants were then given the active treatment and the response variable again recorded.

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 can arise in a situation where practices are different, for example a clinic taking measurements at baseline and follow-up, a clinic only taking measurements at baseline and a clinic only taking measurements at follow-up. In other experimental contexts, some participants may go into both conditions A and B (paired data), whereas others only go into A or B (independent data).

A matched pairs design where some participants cannot be paired. 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 (Cochran, 1953).

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, especially in matched case-control studies, the status for a response variable in some participants may be difficult to detect (Tian, Zhang, & Jiang, 2018).

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, and would want to avoid wastage by making use of the independent observations. An example given in cancer genomic experiments is where either the normal tissue or tumour tissue for an individual is not large enough for extraction (Qi, Yan, & Tian, 2018).

Partially overlapping samples by design due to a partially common element. Two groups with some common units in both, but without intrinsically 'missing data', are frequently confronted. For example, in comparing an assessment conducted in Spanish with an assessment conducted in English, bi-lingual participants may be assessed in both conditions but mono-lingual participants in just one condition. This category also includes scenarios for the incomplete block design where a factor has multiple levels and experimental units are each assigned randomly to two of those levels.

Pairwise comparisons in a comparison of more than two partially overlapping samples. In a repeated measures scenario, multiple pairwise comparisons can be considered within the two partially overlapping samples framework. Approaches to pairwise comparisons may not be without controversy due to the issues concerned with the control of the Type I error rate (e.g., Streiner & Norman, 2011; Saville, 1990).

Tests for the comparison of two partially overlapping samples

Whether the research question be a comparison of means, proportions or variances, arguments for and against the application of competing approaches are similar. Some of the best-known approaches from the 20th century are given.

Amro and Pauly (2017) define four categories of solutions to the partially overlapping samples problem that use all available data. The categories are stated as; 'tests based on maximum likelihood estimators', 'weighted combination tests', 'solutions requiring resampling methods', and 'tests based on a simple mean difference'. To reflect the broader partially overlapping samples framework, this last category is changed to 'tests based on a parameter difference'.

Primitive approaches

Some simple approaches to analysis of partially overlapping samples data, simply ignore the paired nature of data and proceed to analyse all data assuming independence. However, if the number of pairs is relatively large, then an alternative common approach is to perform a paired samples test on only the paired observations and to not include unpaired observations in the analysis. When the sample comprises more independent observations than paired observations, some researchers may choose to perform an independent samples test on only the independent observations discarding paired data from the analysis. In the comparison of proportions for example, the

traditional two sample tests are the Chi-square test for association on the independent observations, or the McNemar test on the paired observations. In either case, the corresponding paired or independent observations are not utilised. Approaches that discard observations to perform a basic traditional test are referred to as naive approaches (Guo & Yuan, 2017; Mantilla & Terpstra, 2018). Naive approaches have reduced relative power because they do not make use of all of the available information. In addition, the discarding of data could induce bias.

For the comparison of means, some statistical software perform paired samples tests, discarding unpaired observations in the dataset. This is often done without any warning to the user. Examples of this include SPSS, SAS and Unistat (Derrick, Toher, & White, 2017). Caution should be exercised when using software because inexperienced users may not realise that they have discarded observations, and may be unaware of the consequences. The `scipy.stats` module within Python will not perform a related samples t-test with unequal length arrays. Likewise, Minitab and the default t-test in R present similar error messages when a paired samples test is selected for unequal sample sizes. This starts to make users aware that there are further considerations with the analyses they are attempting to perform.

Identifying suitable tests for equality of variances provide a challenge. The sensitivity of tests for equal variances to violations of the normality assumption is where the term 'statistical robustness' was coined, or moreover the lack of 'robustness' (Box, 1953; Hogg, 1979). For the comparisons of variances, tests which use deviations from the median are preferable (Conover, Johnson, & Johnson, 1981). Shoemaker (2003) offers two potential fixes to the F-test, and concludes that the Brown-Forsythe test is robust even for highly skewed distributions.

Tests for the comparison of variances with paired data are less well established in statistical software, for example, in Minitab, only independent tests are available as standard. This could encourage researchers to adopt an approach of ignoring the pairing, but this is not statistically valid (Derrick, Toher, & White, 2017; Zumbo, 2002). These naive and other ad-hoc approaches emphasise the need for statistically valid tests in the partially overlapping samples case that use all the available data.

Maximum likelihood estimators

Early literature for the partially overlapping samples framework focused on maximum likelihood estimators, assuming Normal distributions. Lin (1973) and Lin and Stivers (1974) use maximum likelihood estimates for the comparison of group means, but find no single solution is universally applicable.

Using simulation of normally distributed data, Ekbohm (1976) compared the Lin and Stivers (1974) approach with similar proposals based on maximum likelihood estimators and naive tests. He found the only solution to maintain the nominal 5% Type I error rate when the correlation is zero is the independent samples t-test. For positive correlation, the paired samples t-test has greater power than the other tests considered, except when the variances between two samples are not equal. Therefore, the proposed maximum likelihood statistics offer little further benefit relative to primitive approaches.

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

These maximum likelihood proposals are complex mathematical procedures, which could be a barrier to some analysts in a practical setting (Choi & Stablein, 1982; Derrick, Russ, Toher, & White, 2017). A more practical solution, easily performed 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 experimental units are declared as a random effect. Mehrotra (2004) indicates that in the comparison of means, REML is Type I error robust for positive correlation. However, REML is not Type I error robust when there is a large imbalance in the sample sizes of the two groups, or if there is negative correlation (Derrick, Russ, et al., 2017).

Weighted combination tests

Weighted combination tests are where separate tests 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 estimated 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.

For the comparisons of means, many authors have proposed solutions. A method by Bhoj (1978) demonstrates reasonable Type I error robustness, although they do not consider situations that violate the normality assumption (Derrick, White, & Toher, 2020). 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 (2014) and Martinez-Camblor et al. (2013) propose weighted averages of paired and independent samples t-tests. The principle is adding two t-tests together and treating the combination as a t-statistic. However, the sum of two distributions does not lead to a t-distribution, and the approximation used in this approach is particular problematic with small sample sizes. The weights used by Samawi and Vogel (2014) involve a square root weighting function meaning that the weights do not accurately reflect the ratio of the number of observations in the two samples. Uddin and Hasan (2017) optimised the weighting constants used by Bhoj (1978) so that the combined variance of the two elements is minimised. Samawi, Yu, and Vogel (2015) put forward a non-parametric solution attempting to combine the Wilcoxon signed rank test and the Mann Whitney test, which offers a Type I error robust alternative for non-normal distributions.

The comparison of variances in the partially overlapping samples case has been given very little consideration in the existing literature. Bhoj (1978) and Ekbohm (1981) separately consider a weighted combination of independent observations and paired observations to create a new test statistic. 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 these proposals (Ekbohm, 1982). No solution is comprehensively agreed upon for all scenarios. A different solution that uses all available data without a complex weighting structure may therefore be advantageous.

For the comparison of proportions, the approach by Choi and Stablein (1982) is found to maintain better Type I error robustness than its competitors (Bland & Butland, 2011; Tang & Tang, 2004).

A familiar weighted combination approach from meta-analysis is to obtain the p-values for a paired samples test (discarding unpaired observations) and an independent samples test (discarding paired observations). These are then combined using a weighted z-test (Stouffer et al., 1949). In general, it is usual that the weights 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 deciding upon 'optimal weights' and doing so could be computationally intensive. Practitioners may be more comfortable adopting the Stouffer et al. (1949) method if weights are based on sample sizes. There are many 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, this method is more powerful (Chen, 2011). Advantages for this type of approach are that; it can be performed without the requirement of resampling methods, 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. The key disadvantage of these techniques is that confidence intervals for the mean difference are not easy to obtain. Of note is that two separate tests are required before applying this third test. These drawbacks make the results harder to interpret.

These approaches are also limited by the robustness of each individual test. For example, despite the Pitman-Morgan test being the best-known paired samples test for equality of variances, it is not robust for skewed distributions (Mudholkar, Wilding, & Mietlowski, 2003).

Resampling based methods

Resampling methods are increasingly advocated techniques that do not make any prior assumptions about the distribution of the data (Odén & Wedel, 1975). Many resampling strategies are available and bootstrapping generally gives the most consistent results (Fan & Wang, 1996). However, resampling methods are not often used for paired samples designs (Rietveld & van Hout, 2017).

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 test statistic 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) demonstrate that a linear combination of randomization tests can be robust. However, it is not robust for non-positive correlation. Yu et al. (2012) applied permutation methods on statistics proposed by both Bhoj (1978) and Kim et al. (2005), and found that permutation based methods perform similarly to their counterparts. In all tests simulated, when applied to non-normal data, the Type I error rates are reasonably maintained, but with decreased power.

Amro et al. (2019) propose further non-parametric permutation approaches. 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 Log-normal, they show that show that these methods do not maintain Type I error robustness, unless permutation tests are adopted. However, Amro et al. (2019) only show the average Type I error rate at the Pearson's correlation coefficient level, so some sample size conditions where the test may be particularly liberal or conservative may be obscured.

These resampling strategies are computationally intensive to the point of being prohibitive if the sample size is large, and modifications such as randomization tests may be needed rather than complete enumeration of all permutations.

Tests based on a simple parameter difference

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

Looney and Jones (2003) proposed a statistic constructed as a linear interpolation between the paired samples z-test and the independent samples z-test, using the Normal distribution to calculate p-values based on the resulting statistic. Qin, Prentice, and Freeman (2018) demonstrated that application of this test can be extended to community survey research with cross-sectional data and praised the effectiveness of this test relative to the naive paired samples t-test or independent samples t-test. In the extremes, this test defaults to the independent samples z-test or the paired samples z-test. For example, if there were no paired observations this would result in the test statistic defaulting to the independent version of the test statistic. Looney and Jones (2003) demonstrate that when only 10% of the sample is paired, the naive independent samples t-test performs just as well as the proposed z-test. When the paired sample is 50% or 90% of the data, then the proposed z-test maintains the Type I error rate to a better extent relative to the independent samples t-test, however the error rate is generally slightly higher than the nominal significance level. The approach was designed for equal variances only and is not robust under unequal variances with a large sample size imbalance (Derrick, Russ, et al., 2017). Looney and Jones (2003) do not give guidance as 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 statistic by Samawi and Vogel (2014) is very conservative, but the principle that as the sample size increases, asymptotically the t-distribution approximates to the z-distribution, is useful to incorporate. 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.

Using the result for the difference between two random variables, relatively simple tests for the comparison of means, proportions and variances are proposed by Derrick, Russ, et al. (2017); Derrick et al. (2015); Derrick, Ruck, Toher, and White (2018). For the comparison of means, in the extremes the solution defaults to the independent samples t-test or paired samples t-test.

Derrick, Russ, et al. (2017) show that the parametric partially overlapping samples t-tests for comparing means, T_{new1} and T_{new2}, are Type I error robust under conditions of normality. The latter is designed for when equal variances cannot be assumed. These solutions remain valid when only one sample contains independent observations (Derrick, Toher, & White, 2019). A worked example for the comparison of means under the normal distribution is given in Derrick, Toher, and White (2017).

When both samples are taken from the same skewed continuous distribution, T_{new1} maintains Type I error robustness, however T_{new2} is not robust for highly skewed distributions Derrick et al. (2020). T_{new1} maintains Type I error robustness for discrete observations on a five-point or seven-point ordinal scale (Derrick & White, 2018). On these discrete scales, T_{new2} is not robust when the distributions are skewed (Derrick & White, 2018).

Parametric tests dominate the literature due to their ease of interpretation and degree of robustness particularly with increasing sample size. Non-parametric tests still retain their use particularly with skewed data. Some practitioners may habitually test samples for normality to determine if a non-parametric approach may be more appropriate (Mahdizadeh, 2018) and this increases their relative frequency in use, although test selection based assumption testing is not universally endorsed. For a non-parametric solution to compare central location, Derrick et al. (2020) consider applying rank values to the statistics T_{new1} and T_{new2} to give T_{RNK1} and T_{RNK2} respectively. For the same distributions considered by Amro et al. (2019), when two samples are taken from the same distribution, T_{RNK1} provides a Type I error robust solution which maximises power (Derrick et al., 2020). T_{RNK1} also provides superior Type I error robustness relative to parametric tests when the distribution is inherently Normal but with outliers (Derrick, 2018; Derrick, White, & Toher, 2017). Thus, T_{RNK1} is recommended for a location shift model. For a less restrictive null hypothesis of equal distributions, the rank-based approaches offer high power.

Derrick, White, and Toher (2017) consider performing inverse normal transformations of the data and applying these to the statistics T_{new1} and T_{new2} to give T_{INT1} and T_{INT2} respectively. T_{INT1} maintains superior Type I error robustness relative to T_{INT2}. However, these tests do not sufficiently improve the Type I error rate or power relative to T_{RNK1}.

For a comparison of two groups with a dichoto-

mous response variable, the z_8 statistic given by Derrick et al. (2015) offers a powerful Type I error robust solution and can be performed using the R package `Partiallyoverlapping` (Derrick, 2017). This statistic is the difference in the two group proportions, divided by the combined standard error, i.e.

$$z_8 = \frac{p_1 - p_2}{\sqrt{\sigma_{p_1}^2 + \sigma_{p_2}^2 - 2\varphi\sigma_{p_1}\sigma_{p_2}}}$$

where φ is Pearson's phi correlation coefficient.

This test statistic construction allows for the formation of confidence intervals, and the original paper demonstrates that this approach, when used for calculating confidence intervals, is robust and superior to competing alternatives.

For fully paired designs, the effect size for McNemar's test requires the odds ratio of discordant pairs. The partially overlapping z_8 statistic itself includes no information on the number of discordant pairs, therefore a calculation of the effect size for two independent proportions could act as a reasonable approximation, for example $\arcsin\sqrt{p_1} - \arcsin\sqrt{p_2}$ (Cohen, 1992).

For variances, the widely acclaimed Brown-Forsythe test for independent samples (Mirtagiouglu, Yiugit, Mendecs, & Mendecs, 2017; Nordstokke & Zumbo, 2007), is equivalent to the independent samples t-test applied to absolute deviations from group medians. It follows that in the partially overlapping samples case, the statistic T_{new1} could be used in a similar manner. This solution, T_{VAR1} is a valid solution, and more powerful than performing the Brown-Forsythe test on the independent observations only (Derrick et al., 2018).

Revisiting a classic example

An example by Rempala and Looney (2006), considered by Guo and Yuan (2017), Amro and Pauly (2017) and Derrick et al. (2020) is revisited. The outcome variable is the Karnofsky performance score, which measures the functional status of a cancer patient (bigger scores represent a better health status). 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 of life 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 of life, and 23 were recorded only on the last day of life. Observations are as per Table 1.

Using the `Partiallyoverlapping` R package (Derrick, 2017), the results are ($T_{\text{new1}} = 2.522$, $v = 51.609$, $p = .015$), ($T_{\text{new2}} = 2.522$, $v = 49.341$, $p = .016$) which in both cases indicates a statistically significant effect with mean scores on the last day of life lower than the second to last day.

The partially overlapping R package was introduced in this journal with an explanation of the mathematics and usage for the comparison of means (Derrick, Toher, & White, 2017).

The `Partover.test` function has usage as per the `t.test` function in base R. An additional feature is that data can either be passed into the function as two variables (including the missing observations) using the argument `stacked=TRUE`, or as four variables separating the paired and independent observations for each group using the argument `stacked=FALSE`.

The empirical test statistic T_{new1} or T_{new2} is compared to the theoretical t-distribution with v degrees of freedom to obtain the p-value. The choice to report either T_{new1} or T_{new2} is based on underlying knowledge of the variance of the two groups under consideration; T_{new1} if equal variances can be reasonably assumed, T_{new2} otherwise.

Application of 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 performance scores between the last two days of life. Using T_{new1}, a 95% confidence interval for the mean difference calculated using the argument `conf.level = 0.95` gives (0.763 , 6.711). The effect size is calculated (manually) as $d = 2t/\sqrt{v} = 0.698$, indicating a medium sized effect.

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. A principled view might lead an analysis to use a non-parametric test. Table 2 provides rank values for the non-parametric proposals given by Derrick et al. (2020).

Using the `Partiallyoverlapping` R package (Derrick, 2017), applied to the rank values, the non-parametric partially overlapping samples t-tests provide evidence at the 5% significance level to suggest that there is a difference in the distributions of the Karnofsky scores between the last two days of life ($T_{\text{RNK1}} = 2.534$, $p = .014$), ($T_{\text{RNK2}} = 2.521$, $p = .015$).

Performing the tests on a Van der Waerden (1952) inverse normal transformation of the ranks gives evidence to reject the null hypothesis of equal means of the transformed data ($T_{\text{INT1}} = 2.15$, $p = .036$), ($T_{\text{INT2}} = 2.12$, $p = .039$).

A preliminary test for normality of the differences

Table 1 ■ Example from Rempala and Looney (2006)

Patients with scores on both days (ordered pairs, second to last day, last day)
(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, 20, 25, 25, 30, 30, 30

Table 2 ■ 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, 37, 53.5, 53.5, 63.5, 63.5, 63.5

from the group means , using all available data, gives evidence to suggest that the normality assumption is violated (Shapiro-Wilk $p = .004$; Kolmogorov-Smirnov $p = .009$). These preliminary tests of normality might lead a practitioner to use a non-parametric test.

Using the sample data alone to determine the appropriate test that uses all of the available data, would lead to the decision to perform T_{RNK1}. However, the selection of the appropriate test in this way may be controversial and is not straightforward given the vast number of tests for normality and equal variances available (Razali, Wah, et al., 2011). Instead, the appropriate test should be chosen based on the study design and existing knowledge of the behaviour of the response variable (Zumbo & Coulombe, 1997). In this specific example context, the very nature of the outcome measure (ordered data) may be a better principled guide to using a non-parametric approach.

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 methods by Looney and Jones (2003); Samawi and Vogel (2011, 2014) and using REML (Guo & Yuan, 2017). Amro and Pauly (2017) supply confirmation that these methods are also consistent with conclusions made by the method by Lin and Stivers (1974) and their own permutation proposal. In contrast to these methods, naive tests fail to reject H_0 (Derrick et al., 2020).

Now consider 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. Using the `Partiallyoverlapping` R package (Derrick, 2017), applied to the deviations from group medians, the results are $T_{\text{VAR1}} = 0.886$ (p-value $= .380$). Thus, there is no evidence to suggest that the variability in Karnofsky performance differs between the last and second last day of life.

Conclusion

We have identified 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 correctly reflect the structure of the data being collected.

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.

Particularly for smaller sample sizes, partially overlapping samples tests using all of the available data could be advantageous. Tests which act as a simple parameter difference have ease of interpretation, and confidence intervals can readily be formed.

The choice of which test to perform should be based on prior knowledge of the context. For example, if random allocation has been used in a controlled experiment, the assumption of equal variances may be reasonable, whereas this assumption may not be tenable when comparing two pre-existing groups in an observational study. The practice of preliminary testing assumptions may not be advisable when the preliminary test is less robust than the statistical test of interest. 'To make the preliminary test on variances is rather like putting to sea in a rowing boat to find out whether conditions are sufficiently calm for an ocean liner to leave port' (Box, 1953, p. 22). Preliminary testing of assumptions should be avoided unless prior consideration is given to this in an analysis plan (Pearce & Derrick, 2019)

and preferably with a robust testing strategy.

For the comparison of central location, if the assumption of equal distributions and 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 our recommended test of choice.

Non-parametric tests are relatively over used by some researchers that have an obsession with testing the normality assumption (Rasch & Guiard, 2004) whereas some researchers will routinely use parametric tests. 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 it is reasonable to postulate a location shift problem (Rietveld & van Hout, 2017).

The statistic z_8 (Derrick et al., 2015) is recommended for the comparison of proportions and the statistic T_{VAR1} (Derrick et al., 2018) is recommended for the comparison of variances. These solutions are robust with easy to interpret results.

The above solutions do not easily extend to the partially overlapping problem with more than two samples, which is an area requiring further attention in the literature (Mantilla & Terpstra, 2018).

Controversial practices for comparing two samples of paired observations and independent observations are frequently performed, from discarding observations, to imputing observations (Choi & Stablein, 1982). Other poor practices observed in a similar context range from treating all observations as independent and ignoring the pairing, to randomly pairing unpaired observations (Bedeian & Feild, 2002). Solutions discussed using all of the available data facilitated by the `Partiallyoverlapping` R package (Derrick, 2017) offer intuitive, valid and powerful solutions for partially overlapping samples.

References

Amro, L., Konietschke, F., & Pauly, M. (2019). Multiplication-combination tests for incomplete paired data. *Statistics in Medicine*, *38*(17), 3243–3255. doi:10.1002/sim.8178

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. doi:10.1080/00949655.2016.1249871

Bedeian, A. G., & Feild, H. S. (2002). Assessing group change under conditions of anonymity and overlapping samples. *Nursing Research*, *51*(1), 63–65. doi:10.1097/00006199-200201000-00010

Bhoj, D. S. (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

Bhoj, D. S. (1984). On testing equality of variances of correlated variates with incomplete data. *Biometrika*, *71*(3), 639–641. doi:10.1093/biomet/71.3.639

Bland, J. M., & Butland, B. K. (2011). Comparing proportions in overlapping samples. Retrieved April 2, 2018, from http://www-users.york.ac.uk/mb55/overlap.pdf

Box, G. E. (1953). Non-normality and tests on variances. *Biometrika*, *40*(3/4), 318–335. doi:10.2307/2333350

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. doi:10.1111/j.1420-9101.2010.02226.x

Choi, S., & Stablein, D. (1982). Practical tests for comparing two proportions with incomplete data. *Applied Statistics*, *31*(3), 256–262. doi:10.2307/2347999

Cochran, W. G. (1953). Matching in analytical studies. *American Journal of Public Health and the Nations Health*, *43*(6), 684–691. doi:10.2105/AJPH.43.6_Pt_1.684

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155–159.

Conover, W. J., Johnson, M. E., & 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), 351–361. doi:10.1080/00401706.1981.10487680

Derrick, B. (2017). Partiallyoverlapping: Partially overlapping samples tests [r package] (Version 2.0). Retrieved from https://CRAN.R-project.org/package=Partiallyoverlapping

Derrick, B. (2018). An outlier in an independent samples design. In *In royal statistical society conference* (pp. 3–7). Cardiff City Hall, Cardiff, Wales.

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., Ruck, A., Toher, D., & 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), 9.

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 Statis-*

tical Methods, *16*(1), 137–157. doi:10.22237/jmasm/1493597280

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 in Psychology*, *13*(2), 120–126. doi:10.20982/tqmp.13.2.p120

Derrick, B., & 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), 31–99.

Derrick, B., White, P., & Toher, D. (2017). An inverse normal transformation solution for the comparison of two samples that contain both paired observations and independent observations. *arXiv preprit arXiv:1708.00347*.

Derrick, B., White, P., & Toher, D. (2020). Parametric and non-parametric tests for the comparison of two samples which both include paired and unpaired observations. *Journal of Modern Applied Statistical Methods*, *18*(1), 1–99. doi:10.22237/jmasm/1556669520

Derrick, B., Toher, D., & White, P. (2019). The performance of the partially overlapping samples t-tests at the limits. *arXiv preprint arXiv:1906.01006*, 1–9.

Ekbohm, G. (1976). On comparing means in the paired case with incomplete data on both responses. *Biometrika*, *63*(2), 299–304. doi:10.1093/biomet/63.2.299

Ekbohm, G. (1981). A test for the equality of variances in the paired case with incomplete data. *Biometrical Journal*, *23*(3), 261–265. doi:10.1002/bimj.4710230306

Ekbohm, G. (1982). On comparing variances in the paired case with incomplete data. *Biometrika*, *69*(3), 670–673. doi:10.1093/biomet/69.3.670

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

Fan, X., & Wang, L. (1996). Comparability of jackknife and bootstrap results: An investigation for a case of canonical correlation analysis. *The Journal of Experimental Education*, *64*(2), 173–189. doi:10.1080/00220973.1996.9943802

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. doi:10.1177/0962280215577111

Hart, A. (2001). Mann-whitney test is not just a test of medians: Differences in spread can be important. *British Medical Journal*, *323*(7309), 391–399. doi:10.1136/bmj.323.7309.391

Hogg, R. V. (1979). Statistical robustness: One view of its use in applications today. *The American Statistician*, *33*(3), 108–115. doi:10.1080/00031305.1979.10482673

Kang, H. (2013). The prevention and handling of the missing data. *Korean Journal of Anesthesiology*, *64*(5), 402–406. doi:10.4097/kjae.2013.64.5.402

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

Lancaster, H. (1961). The combination of probabilities: An application of orthonormal functions. *Australian & New Zealand Journal of Statistics*, *3*(1), 20–33. doi:10.1111/j.1467-842X.1961.tb00058.x

Lin, P. E. (1973). Procedures for testing the difference of means with incomplete data. *Journal of the American Statistical Association*, *68*(343), 699–703. doi:10.1080/01621459.1973.10481407

Lin, P. E., & Stivers, L. E. (1974). On difference of means with incomplete data. *Biometrika*, *61*(2), 325–334. doi:10.1093/biomet/61.2.325

Looney, S. W., & Jones, P. W. (2003). A method for comparing two normal means using combined samples of correlated and uncorrelated data. *Statistics in Medicine*, *22*(9), 1601–1610. doi:10.1002/sim.1514

Mahdizadeh, M. (2018). Testing normality based on sample information content. *International Journal of Mathematics and Statistics*, *19*(1), 1–18.

Mantilla, L. B., & Terpstra, J. T. (2018). Means, medians, and multivariate mixed design data. *American Journal of Mathematical and Management Sciences*, *37*(1), 56–65. doi:10.1080/01966324.2017.1379917

Martinez-Camblor, P., Corral, N., & 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*, 1179–1180.

Mirtagiouglu, H., Yiugit, S., Mendecs, E., & Mendecs, M. (2017). A monte carlo simulation study for comparing performances of some homogeneity of variances tests. *Journal of Applied Quantitative Methods*, *12*(1), 1–99.

Mudholkar, G. S., Wilding, G. E., & Mietlowski, W. L. (2003). Robustness properties of the pitman-morgan test. *Communications in Statistics-Theory and Methods*, *32*(9), 1801–1816. doi:10.1081/STA-120022710

Nordstokke, D. W., & Zumbo, B. D. (2007). A cautionary tale about levene's tests for equal variances. *Journal of Educational Research & Policy Studies*, *7*(1), 1–14.

Odén, A., & Wedel, H. (1975). Arguments for fisher's permutation test. *The Annals of Statistics*, *3*(2), 518–520. doi:10.1214/aos/1176343082

Pearce, J., & Derrick, B. (2019). Preliminary testing: The devil of statistics? *Reinvention: an International Journal of Undergraduate Research*, *12*(2), 2–99. doi:10.31273/reinvention.v12i2.339

Qi, Q., Yan, L., & Tian, L. (2018). Testing equality of means in partially paired data with incompleteness in single response. *Statistical Methods in Medical Research, 0*, *28*, 1508–1522.

Qin, H., Prentice, E., & Freeman, K. (2018). Analyzing partially correlated longitudinal data in community survey research. *Society & Natural Resources*, *31*(1), 142–149.

Ramosaj, B., Amro, L., & Pauly, M. (2020). A cautionary tale on using imputation methods for inference in matched-pairs design. *Bioinformatics*, *36*(10), 3099–3106.

Rasch, D., & Guiard, V. (2004). The robustness of parametric statistical methods. *Psychology Science*, *46*, 175–208.

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), 21–33.

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. doi:10.1016/j.jspi.2004.06.002

Rietveld, T., & 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*, 44–57. doi:10.1016/j.jcomdis.2017.07.002

Samawi, H., & 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

Samawi, H., & Vogel, R. (2014). Notes on two sample tests for partially correlated (paired) data. *Journal of Applied Statistics*, *41*(1), 109–117. doi:10.1080/02664763.2013.830285

Samawi, H., Yu, L., & Vogel, R. (2015). On some nonparametric tests for partially observed correlated data: Proposing new tests. *Journal of Statistical Theory and Applications*, *14*(2), 131–199. doi:10.2991/jsta.2015.14.2.3

Saville, D. J. (1990). Multiple comparison procedures: The practical solution. *The American Statistician*, *44*(2), 174–180. doi:10.2307/2684163

Shoemaker, L. H. (2003). Fixing the f test for equal variances. *The American Statistician*, *57*(2), 105–114. Retrieved from https://doi.org/10.1198/0003130031441

Stouffer, S. A., Lumsdaine, A. A., Lumsdaine, M. H., Jr, W., M., R., Smith, M. B., ... Cottrell Jr, L. S. (1949). The american soldier: Combat and its aftermath. *Studies in Social Psychology in World War II*, *36*(3), 550–551.

Streiner, D. L., & Norman, G. R. (2011). Correction for multiple testing: Is there a resolution? *Chest*, *140*(1), 16–18.

Tang, M., & Tang, N. (2004). Exact tests for comparing two paired proportions with incomplete data. *Biometrical journal*, *46*(1), 72–82. doi:10.1002/bimj.200210003

Tian, G.-L., Zhang, C., & Jiang, X. (2018). Valid statistical inference methods for a case-control study with missing data. *Statistical Methods in Medical Research*, *27*(4), 1001–1023. doi:10.1177/0962280216649619

Uddin, N., & 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), 2430–2446. doi:10.1080/03610918.2015.1047527

Van der Waerden, B. (1952). Order tests for the two-sample problem and their power. *In Indagationes mathematicae (proceedings)*, *55*, 453–458. doi:10.1016/S1385-7258(53)50012-5

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), 1368–1373. doi:10.1111/j.1420-9101.2005.00917.x

Yu, D., Lim, J., Liang, F., Kim, K., Kim, B. S., & Jang, W. (2012). Permutation test for incomplete paired data with application to cdna microarray data. *Computational Statistics & Data Analysis*, *56*(3), 510–521. doi:10.1016/j.csda.2011.08.012

Zumbo, B. D. (2002). An adaptive inference strategy: The case of auditory data. *Journal of Modern Applied Statistical Methods*, *1*(1), 1–99. doi:10.22237/jmasm/1020255000

Zumbo, B. D., & 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), 139–150. doi:10.1037/1196-1961.51.2.139

Citation

Derrick, B., & White, P. (2022). Review of the partially overlapping samples framework: Paired observations and independent observations in two samples. *The Quantitative Methods for Psychology*, *18*(1), 55–65. doi:10.20982/tqmp.18.1.p055

Received: 10/08/2021 ~ Accepted: 27/01/2022