




Permutation tests for mixed paired and two-sample designs

E. N. Johnson¹ · S. J. Richter² 

Received: 28 January 2021 / Accepted: 8 July 2021

© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2021

Abstract

Permutation tests based on previously developed statistics are proposed for the case of mixed paired and two-sample designs. Different weighting schemes of previous tests are explored to understand the strengths and weaknesses of each test. A simulation study compares the power and Type I error rates of the new tests with those previously developed. Rank-based statistics generally performed as well as or better than parametric statistics, particularly for nonnormal distributions.

Keywords Paired design · Missing data · Permutation test · Ranks

1 Introduction

In certain experiments and observational studies for paired data, missing data may arise in several ways. Longitudinal studies will often involve missing observations at one time point. Studies where two treatments are to be applied to the same subject may also suffer from incomplete pairs if some subjects are not eligible for one treatment. This leads to a mixed design with paired data and data that resemble independent random samples from each of the two groups. Various methods have been proposed to retain as much information as possible in the presence of incomplete pairs.

Lin and Stivers (1974) proposed several adaptations to the paired and unpaired t -tests, and the choice of test statistic depended upon the true correlation between the paired observations and sample size. Bhoj (1978, 1984, 1989) proposed several weighted averages of the paired and independent t statistics for distributions with equal variances. None of these tests had exact distributions, but were instead based on asymptotic approximations.

✉ S. J. Richter
sjricht2@uncg.edu

E. N. Johnson
ecnance27@gmail.com

¹ Wake Forest School of Medicine, Winston-Salem, NC, USA

² University of North Carolina at Greensboro, Greensboro, NC, USA

Maritz (1995) proposed an exact permutation test, however the incomplete pairs were permuted in a way that may not be considered standard. For all pairs, including those with a missing value, observations were randomly interchanged within each pair. This assumed that each variable was equally likely to have missing values for the incomplete pairs and included the extreme cases where all missing values may be for one variable, which would be unlikely in most practical situations. Einsporn and Habtzghi (2013) also proposed a permutation test for the mixed design. Their statistic was a weighted combination of the difference in means of the variables between the complete and incomplete pairs. However, unlike the permutation scheme proposed by Maritz (1995), they chose to permute the incomplete pairs such that the sample size for each variable remained fixed in all arrangements.

Dubnicka et al. (2002) combined nonparametric statistics, using the Wilcoxon on signed-rank statistic for complete pairs and Wilcoxon rank-sum statistic for incomplete pairs. They proposed both unweighted and weighted combinations of the statistics based on the asymptotic distributions of the nonparametric statistics. Magel and Fu (2014) modified the unweighted statistic of Dubnicka et al. (2002) by standardizing the signed-rank and rank-sum statistics before adding them, whereas Dubnicka had suggested standardizing the sum.

Bhoj (1989) found that their Z_b statistic performed best compared to earlier versions (1978, 1984), the statistics of Lin and Stivers (1974) and the paired t -test ignoring incomplete pairs, having superior power when the data had high correlation and the sample sizes of the incomplete pairs were relatively small compared to the complete pairs. Magel and Fu (2014) found their test to be superior to the unweighted R statistic of Dubnicka et al. (2002) when there were more incomplete than complete pairs for normal and exponential distributions as well as the exponential and t -distributions with unequal variances. However, when there were equal variances under the exponential distribution, the test usually had higher power when there were more complete pairs, and had higher power more consistently for the t -distributions with equal variances.

Einsporn and Habtzghi (2013) included a limited simulation that considered two sets of sample sizes with normal and exponential underlying distributions. Their permutation test T usually had marginally higher power than tests proposed by Bhoj (1989), Lin and Stivers (1974), Maritz (1995), and Dubnicka et al. (2002) for small sample sizes and low correlation and normally distributed data. Bhoj's Z_b test (1989) had the highest power of the parametric procedures considered and was often more powerful than T for high correlation and large sample sizes. The weighted nonparametric statistic proposed by Dubnicka et al. (2002) was almost always most powerful when the data were exponential. It was also apparent that the power of the Z_b test increased faster than that of the permutation test as the correlation increased and had higher power when the incomplete pairs' sample sizes became much larger than that of the complete pairs. Dubnicka's (2002) weighted statistic also had higher power than the unweighted statistic as the correlation increased. While Einsporn and Habtzghi's (2013) T statistic had power similar to or slightly higher than Dubnicka's (2002) weighted statistic for data from a normal distribution, when the data came from an exponential distribution, Dubnicka's (2002) weighted statistic always performed better. These limited simulated results suggest that using the combination of Wilcoxon

signed-rank and Wilcoxon rank-sum statistics may be a better choice in most practical situations since the true distribution will not be known.

The nonparametric tests generally performed better in the limited simulations of Einsporn and Habtzghi (2013), however a wider variety of conditions, such as heavier-tailed distributions and different sample size combinations, need to be considered to better understand the relative performance of the different methods. In addition, exact permutation versions of the rank statistics may also perform better than the asymptotic versions, especially for small to moderate sample sizes. Thus, we propose permutation versions of the rank statistics developed by Dubnicka et al. (2002) and compare these new permutation tests to previously proposed tests over a wide range of conditions.

While both Magel and Fu (2014) and Einsporn and Habtzghi (2013) included simulation results, Magel and Fu (2014) only compared their M statistic to statistic to unweighted statistic R of Dubnicka et al. (2002), while Einsporn and Habtzghi (2013) did not include M , and also only considered two data generating distributions (normal, exponential) and two sample size cases. Neither simulation considered permutation tests based on the nonparametric statistics. Thus, in this paper we propose new permutation tests based on the nonparametric statistics of Dubnicka et al. (2002). A comprehensive simulation including all methods that have performed well in previous studies is conducted to investigate the performance of the new tests relative to previously proposed methods for a wide range of distributions and sample sizes.

2 Methods

2.1 Bhoj's Z_b statistic

Bhoj (1978) proposed a weighted sum of paired and independent t tests of the form $Z = \frac{wT_1 + (1-w)T_2}{D}$ where T_1 and T_2 were the usual independent and paired t -statistics respectively and D a multiplier to help achieve an approximate t -distribution. They derived two complex transformations to achieve a better approximation, which resulted in $Z_b = \frac{\lambda_b U_1 + (1-\lambda_b)U_3}{\sqrt{\lambda_b^2 + (1-\lambda_b)^2}}$ where U_1 and U_3 were the transformed t -statistics and λ_b the weight for the transformed statistics. See Bhoj (1989) for exact transformation details. In a simulation comparing several statistics, including that of Lin and Stivers (1974), Bhoj's Z_b test had higher power than that of Lin and Stivers for small sample sizes and high correlation and in general for larger sample sizes.

2.2 Dubnicka's R and R_w statistics

Dubnicka et al. (2002) developed nonparametric tests that were weighted sums of Wilcoxon signed-rank (U) and rank-sum (S) statistics. The signed-rank statistic has mean $E(S) = \frac{n(n+1)}{4}$ and variance $Var(S) = \frac{n(n+1)(2n+1)}{24}$, and the rank-sum statistic has mean $E(U) = \frac{n_1 n_2}{2}$ and variance $Var(U) = \frac{n_1 n_2 (n_1 + n_2 + 1)}{12}$. The unweighted statistic suggested by Dubnicka et al. (2002) was by $R = S + U$, which has mean $E(R) = \frac{n(n+1)}{4} + \frac{n_1 n_2}{2}$ and variance $Var(R) = \frac{n(n+1)(2n+1)}{24} + \frac{n_1 n_2 (n_1 + n_2 + 1)}{12}$. Then the

statistic $R_z = \frac{R - E(R)}{\sqrt{\text{Var}(R)}}$ is assumed to follow an approximate $\mathcal{N}(0, 1)$ distribution. They also considered a weighted statistic, $R_w = \frac{2(n_1 + n_2)}{(n(n_1 + n_2) + 2n_1n_2)(n+1)} S + \frac{2}{n(n_1 + n_2) + 2n_1n_2} U$, with mean $E(R_w) = \frac{1}{2}$ and variance equal to the weighted sum of the variances of the Wilcoxon signed-rank and rank-sum statistics. The statistic $R_{w,z} = \frac{R_w - \frac{1}{2}}{\sqrt{\text{Var}(R_w)}}$ is also assumed to follow an approximate $\mathcal{N}(0, 1)$ distribution. Dubnicka et al. (2002) found that the weighted rank statistic was generally more efficient than the unweighted statistic. If the number of complete pairs was more than the sum of the incomplete pairs, however, the unweighted statistic was nearly as efficient. Thus, they recommended the simpler unweighted statistic in those cases.

2.3 Magel and Fu's M statistic

Magel and Fu (2014) proposed a slight variation of Dubnicka et al. (2002) in which they standardized the signed-rank and rank-sum statistics before combining them. The resulting statistic was $M = \frac{S^* + U^*}{\sqrt{2}}$, where S^* and U^* were the standardized statistics respectively. They found that the power of M was generally higher than the unweighted statistic R of Dubnicka et al. (2002) when there were more incomplete than complete pairs.

2.4 Einsporn and Habtzghi's T statistic

Einsporn and Habtzghi (2013) proposed a permutation test based on mean differences, similar to the asymptotic test from Bhoj (1989), defined as $T = w\bar{d}_p + (1 - w)\bar{d}_u$ with weight $w = \left(\frac{1}{n_{ux}} + \frac{1}{n_{uy}}\right) / \left(\frac{2-2\rho}{n_p} + \frac{1}{n_{ux}} + \frac{1}{n_{uy}}\right)$ that minimized the variance of T . In a limited simulation that included Bhoj's Z_b statistic and the weighted (R_w) and unweighted (R) statistics of Dubnicka et al. (2002), they found T to generally have highest power for normally distributed data with low correlation and small sample sizes, but that R and R_w had highest power when data were exponential.

2.5 New permutation tests

Permutation versions of the rank tests of Dubnicka et al. (2002) are proposed. The data are permuted following the method of Einsporn and Habtzghi (2013). Both the unweighted $R_{perm} = S + U$ and weighted $R_{w,perm} = \frac{2(n_1 + n_2)}{(n(n_1 + n_2) + 2n_1n_2)(n+1)} S + \frac{2}{n(n_1 + n_2) + 2n_1n_2} U$ statistics are calculated on the observed values and for each permutation. Standardizing the statistic is not necessary because the mean and variance are based only on sample sizes and thus remain constant for all permutations. The one-sided p -values are then the proportion of permutations with statistics greater than or equal to the observed statistic.

While both Magel and Fu (2014) and Einsporn and Habtzghi (2013) included simulation results, Magel and Fu (2014) only compared their M statistic to statistic to unweighted statistic R of Dubnicka et al. (2002), while Einsporn and Habtzghi

(2013) did not include M , and also only considered two data generating distributions (normal, exponential) and two sample size cases. Neither simulation considered permutation tests based on the nonparametric statistics. Thus, in this paper we propose new permutation tests based on the nonparametric statistics of Dubnicka et al. (2002). A comprehensive simulation including all methods that have performed well in previous studies is conducted to investigate the performance of the new tests relative to previously proposed methods for a wide range of distributions and sample sizes.

3 Simulation

3.1 Design

The new permutation rank tests were compared to those of Einsporn and Habtzghi (2013), Magel and Fu (2014), Bhoj (1989), and both weighted and unweighted tests of Dubnicka et al. (2002). The standard paired t -test and Wilcoxon signed-rank tests were also included to compare to the extreme case of simply discarding the incomplete pairs.

g and h distributions (Hoaglin 2006) were used to generate data with different characteristics, where the g value controls the skewness of the distribution and the h value controls the elongation of the tails of the distribution (Hoaglin 2006). Four different settings were used: $g = 0.7$ and $h = 0$ for a skewed light-tailed distribution (similar to an exponential distribution); $g = 0.7$ and $h = 0.35$ for a skewed heavier-tailed distribution; $g = 0$ and $h = 0$ for a normal distribution, and $g = 0$ and $h = 0.4$ for a symmetric heavier-tailed distribution. The paired samples were generated with equal variance $\sigma^2 = 1$ and Pearson correlation values of $\rho = 0.1, 0.5$, and 0.9 . In order to generate correlated variables under all distributions we first generated random data from a bivariate normal distribution with means μ_X and μ_Y and then used their probabilities to generate data under the new distribution in such a way that the Spearman correlation between the normally distributed variables and transformed variables were equal. Once correlated pairs were generated, observations were randomly deleted from each variable separately to simulate incomplete pairs. The sample Spearman correlation from the remaining complete pairs was used when calculating the weights for test statistics. We estimated Type I error with a true difference of $\delta = \mu_X - \mu_Y = 0$ and estimated power for $\delta = 0.5$ and 1 . For a few cases power was 1 for all tests when $\delta = 0.5$, and thus $\delta = 0.25$ was also considered for those cases.

Total sample sizes of $10, 20$, and 40 with small, moderate, and large percentages of missing data were considered. The small sample size only allowed us to consider 50% missing, while we included $25\%, 50\%$, and 75% missing for the moderate sample size and $20\%, 50\%$, and 80% missing for the large sample size. We also included the two sets of sample sizes presented by Einsporn and Habtzghi (2013) for comparison. For each distribution, mean difference, sample size, and correlation, Type I error or power was estimated as the proportion of rejections out of 5000 data sets. For each permutation test 566 random permutations were used to estimate the p -value. Marozzi (2014) showed that power increases only slightly with more permutations, and the maximum root mean squared error of the estimation is 0.00849 compared to 0.005 if

all permutations are used. If estimates for two tests have a difference greater than at least $2 \times 0.00849 = 0.0168$, then we can be approximately 95% confident that the true probability of rejection for one test differs statistically from the another.

3.2 Results

Type 1 error and power estimates are given in the tables below.

For both heavy-tailed skewed and symmetric distributions the rank based statistics generally had the highest power. With 75% incomplete pairs (Table 1), R_z had the highest power for low correlation while M generally had the highest power for moderate to high correlation. For one case with $\delta = 0.5$ and $\rho = 0.9$, though, S had the highest power even ignoring the complete pairs. With 80% incomplete pairs (Table 2), $R_{w,z}$ and $R_{w,perm}$ had the highest power for low and moderate correlation, but M had highest power with high correlation. The power of both $R_{w,z}$ and $R_{w,perm}$ increased faster than R_z and R_{perm} as the correlation increased. For the heavy-tailed symmetric distribution with only 20% incomplete pairs (Table 3), $R_{w,perm}$ generally had the highest power for low and moderate correlation, while S was best when correlation was high. For moderate and large sample sizes with 50% missing for both the skewed and symmetric distributions, R_z had the highest power (Tables 4 and 5).

The rank-based statistics were also most powerful for the light-tailed skewed distribution. With 75% missing and a moderate sample size of 20, M generally had the highest power (Table 6). Table 7 contains another case with the same sample size but with 50% incomplete pairs where R_z had the highest power for all combinations of correlation and mean differences. When the sample size doubled to 40 (Table 8) $R_{w,perm}$ had the highest power for $\delta = 0.5$ and $\rho = 0.1$ and 0.5 , M had the highest power for $\delta = 1$ and $\rho = 0.1$, while $R_{w,z}$ had the highest power for the other cases. $R_{w,perm}$ also had the highest power for the unequal incomplete sample sizes (Einsporn and Habtzghi 2013) previously presented (Table 9).

For sample sizes of 20 and 40 with 50% incomplete pairs from a normal distribution (Tables 10 and 11), T had highest power in all but one case. With 75% missing

Table 1 Heavy-tailed skewed distribution with $n_p = 5$, $n_{ux} = 7$, $n_{uy} = 8$

δ	ρ	T	R_z	R_{perm}	$R_{w,z}$	$R_{w,perm}$	Z_b	M	t	S
0	0.1	0.0598	0.0588	0.0494	0.0506	0.0518	0.0254	0.053	0.0326	0.0596
0	0.5	0.063	0.058	0.049	0.0512	0.0528	0.0364	0.0536	0.0312	0.0588
0	0.9	0.0576	0.0606	0.0518	0.05	0.0526	0.0524	0.0542	0.0318	0.0576
0.5	0.1	0.195	0.25	0.2206	0.233	0.2326	0.0392	0.2412	0.0944	0.1632
0.5	0.5	0.224	0.2774	0.2468	0.2698	0.2712	0.0682	0.2786	0.1264	0.209
0.5	0.9	0.3762	0.3588	0.3188	0.429	0.4294	0.2822	0.4322	0.3438	0.4554
1.0	0.1	0.3896	0.5524	0.5092	0.5278	0.5262	0.0798	0.538	0.2152	0.3136
1.0	0.5	0.4406	0.5886	0.5462	0.6032	0.5998	0.16	0.6118	0.306	0.4112
1.0	0.9	0.6478	0.679	0.6354	0.7682	0.7668	0.5786	0.7716	0.6524	0.7322

Empirical size and power at $\alpha = 0.05$

Table 2 Heavy-tailed skewed distribution with $n_p = 8$, $n_{ux} = 16$, $n_{uy} = 16$

δ	ρ	T	R_z	R_{perm}	$R_{w,z}$	$R_{w,perm}$	Z_b	M	t	S
0	0.1	0.0584	0.0506	0.051	0.053	0.0524	0.0108	0.0528	0.032	0.051
0	0.5	0.0594	0.0534	0.0532	0.052	0.053	0.0188	0.051	0.0296	0.0512
0	0.9	0.056	0.049	0.0496	0.0508	0.0512	0.0304	0.0512	0.0304	0.0512
0.5	0.1	0.2246	0.3572	0.355	0.3748	0.3706	0.0228	0.3724	0.1238	0.178
0.5	0.5	0.2546	0.3732	0.3744	0.4264	0.4274	0.056	0.4244	0.166	0.2334
0.5	0.9	0.4324	0.4492	0.4436	0.6328	0.631	0.3366	0.6368	0.4336	0.512
1.0	0.1	0.4672	0.7762	0.776	0.7904	0.7908	0.068	0.7868	0.2866	0.3578
1.0	0.5	0.5144	0.7984	0.796	0.849	0.8486	0.1674	0.8464	0.3854	0.4638
1.0	0.9	0.7242	0.8482	0.8438	0.9508	0.949	0.6476	0.952	0.7248	0.7572

 Empirical size and power at $\alpha = 0.05$
Table 3 Heavy-tailed symmetric distribution with $n_p = 32$, $n_{ux} = 4$, $n_{uy} = 4$

δ	ρ	T	R_z	R_{perm}	$R_{w,z}$	$R_{w,perm}$	Z_b	M	t	S
0	0.1	0.052	0.0504	0.0496	0.0498	0.0524	0.08	0.0512	0.0456	0.0534
0	0.5	0.0548	0.0514	0.0514	0.0472	0.0498	0.0834	0.0502	0.0454	0.0536
0	0.9	0.0524	0.0504	0.0532	0.0506	0.053	0.0852	0.05	0.0416	0.0526
0.5	0.1	0.309	0.3992	0.399	0.4224	0.4258	0.2716	0.3724	0.2798	0.3912
0.5	0.5	0.4042	0.539	0.5368	0.5618	0.5586	0.3768	0.4762	0.377	0.533
0.5	0.9	0.7858	0.9698	0.9694	0.9692	0.9692	0.7686	0.8476	0.7672	0.9704
1.0	0.1	0.6502	0.8312	0.8296	0.8544	0.857	0.5994	0.7964	0.6082	0.82
1.0	0.5	0.7742	0.9506	0.9492	0.9594	0.96	0.7408	0.8924	0.7434	0.944
1.0	0.9	0.949	1	1	0.9998	0.9998	0.9398	0.9954	0.9398	1

 Empirical size and power at $\alpha = 0.05$
Table 4 Heavy-tailed skewed distribution with $n_p = 10$, $n_{ux} = 5$, $n_{uy} = 5$

δ	ρ	T	R_z	R_{perm}	$R_{w,z}$	$R_{w,perm}$	Z_b	M	t	S
0	0.1	0.0556	0.0546	0.0492	0.0516	0.051	0.0458	0.0504	0.0368	0.0506
0	0.5	0.056	0.0546	0.0466	0.051	0.051	0.0542	0.0512	0.0344	0.0518
0	0.9	0.0544	0.0546	0.0452	0.0494	0.0498	0.0708	0.0522	0.0334	0.05
0.5	0.1	0.1974	0.2566	0.2312	0.245	0.2448	0.101	0.252	0.14	0.1968
0.5	0.5	0.235	0.3296	0.292	0.3142	0.3102	0.1622	0.303	0.1854	0.256
0.5	0.9	0.4872	0.6402	0.6034	0.6166	0.6142	0.4562	0.522	0.4644	0.6122
1.0	0.1	0.3982	0.5678	0.5308	0.555	0.5548	0.2406	0.5618	0.3058	0.416
1.0	0.5	0.4896	0.691	0.6566	0.6766	0.6762	0.3694	0.657	0.4244	0.557
1.0	0.9	0.7572	0.9156	0.8974	0.9022	0.902	0.7354	0.8424	0.7428	0.893

 Empirical size and power at $\alpha = 0.05$

Table 5 Heavy-tailed symmetric distribution with $n_p = 10$, $n_{ux} = 5$, $n_{uy} = 5$

δ	ρ	T	R_z	R_{perm}	$R_{w,z}$	$R_{w,perm}$	Z_b	M	t	S
0	0.1	0.056	0.0564	0.0476	0.0528	0.0538	0.0604	0.0502	0.0416	0.0542
0	0.5	0.0542	0.0562	0.0474	0.0512	0.0532	0.069	0.05	0.0416	0.0542
0	0.9	0.0512	0.0542	0.0466	0.0514	0.0522	0.085	0.05	0.0398	0.0534
0.5	0.1	0.2042	0.2502	0.2206	0.2386	0.2366	0.114	0.237	0.1498	0.1868
0.5	0.5	0.2482	0.3108	0.2784	0.2958	0.2924	0.1714	0.273	0.1988	0.247
0.5	0.9	0.5236	0.6226	0.5842	0.5976	0.596	0.5022	0.5032	0.5068	0.5962
1.0	0.1	0.4492	0.5636	0.5188	0.5482	0.5412	0.272	0.5478	0.3448	0.4142
1.0	0.5	0.5344	0.6716	0.639	0.6594	0.6582	0.4134	0.636	0.465	0.541
1.0	0.9	0.8448	0.9348	0.9194	0.9246	0.9236	0.829	0.8522	0.8342	0.9188

Empirical size and power at $\alpha = 0.05$ **Table 6** Light-tailed skewed distribution with $n_p = 5$, $n_{ux} = 7$, $n_{uy} = 8$

δ	ρ	T	R_z	R_{perm}	$R_{w,z}$	$R_{w,perm}$	Z_b	M	t	S
0	0.1	0.061	0.0588	0.0502	0.048	0.0484	0.0292	0.0508	0.0366	0.0604
0	0.5	0.0634	0.0576	0.0478	0.0512	0.0516	0.0438	0.054	0.0352	0.0606
0	0.9	0.0608	0.06	0.0516	0.0516	0.0536	0.0638	0.0546	0.0342	0.0592
0.5	0.1	0.285	0.3412	0.3012	0.3264	0.3268	0.0596	0.3356	0.1482	0.2112
0.5	0.5	0.3346	0.3766	0.3378	0.3868	0.387	0.1192	0.3972	0.2132	0.2888
0.5	0.9	0.557	0.4758	0.4292	0.5888	0.589	0.4668	0.5932	0.5372	0.6118
1.0	0.1	0.6132	0.7252	0.6814	0.7152	0.7124	0.135	0.7258	0.342	0.4204
1.0	0.5	0.6826	0.7672	0.7288	0.7862	0.785	0.2818	0.7934	0.4782	0.555
1.0	0.9	0.8762	0.8276	0.8002	0.9106	0.9088	0.8018	0.9118	0.8638	0.8916

Empirical size and power at $\alpha = 0.05$ **Table 7** Light-tailed skewed distribution with $n_p = 10$, $n_{ux} = 5$, $n_{uy} = 5$

δ	ρ	T	R_z	R_{perm}	$R_{w,z}$	$R_{w,perm}$	Z_b	M	t	S
0	0.1	0.054	0.054	0.0466	0.0498	0.05	0.0624	0.0522	0.0448	0.0526
0	0.5	0.0548	0.0548	0.0444	0.0512	0.0494	0.0728	0.0506	0.043	0.0526
0	0.9	0.0524	0.0554	0.0446	0.0498	0.0502	0.0856	0.0512	0.0408	0.0528
0.5	0.1	0.2988	0.3694	0.3312	0.3538	0.3564	0.1756	0.3578	0.2306	0.2718
0.5	0.5	0.3816	0.4792	0.4384	0.463	0.4576	0.2846	0.438	0.3222	0.3864
0.5	0.9	0.747	0.8232	0.7968	0.8034	0.8026	0.7306	0.6988	0.7352	0.808
1.0	0.1	0.667	0.766	0.7382	0.757	0.7548	0.4372	0.756	0.531	0.595
1.0	0.5	0.7626	0.871	0.8472	0.8612	0.8572	0.6418	0.837	0.688	0.7558
1.0	0.9	0.9662	0.986	0.9818	0.9822	0.9822	0.9654	0.9492	0.9662	0.9842

Empirical size and power at $\alpha = 0.05$

Table 8 Light-tailed skewed distribution with $n_p = 20$, $n_{ux} = 10$, $n_{uy} = 10$

δ	ρ	T	R_z	R_{perm}	$R_{w,z}$	$R_{w,perm}$	Z_b	M	t	S
0	0.1	0.0524	0.049	0.049	0.05	0.0508	0.0512	0.0538	0.0452	0.0472
0	0.5	0.0524	0.0488	0.0502	0.0496	0.0516	0.0638	0.0532	0.046	0.0466
0	0.9	0.052	0.0496	0.0516	0.0508	0.0532	0.0878	0.0526	0.046	0.047
0.5	0.1	0.4408	0.5628	0.5634	0.5682	0.5726	0.2544	0.5718	0.3566	0.4146
0.5	0.5	0.5548	0.7072	0.7042	0.7118	0.7122	0.4234	0.6894	0.4848	0.5958
0.5	0.9	0.914	0.9796	0.9792	0.9798	0.979	0.8992	0.94	0.9032	0.9696
1.0	0.1	0.8546	0.9542	0.9532	0.956	0.956	0.6516	0.9576	0.7458	0.8422
1.0	0.5	0.9286	0.9888	0.9872	0.989	0.988	0.8428	0.9844	0.881	0.9516
1.0	0.9	0.9978	1	1	1	1	0.9958	0.9992	0.996	1

Empirical size and power at $\alpha = 0.05$ **Table 9** Light-tailed skewed distribution with $n_p = 20$, $n_{ux} = 30$, $n_{uy} = 10$

δ	ρ	T	R_z	R_{perm}	$R_{w,z}$	$R_{w,perm}$	Z_b	M	t	S
0	0.1	0.0524	0.05	0.0496	0.0522	0.0524	0.0276	0.0492	0.0474	0.0506
0	0.5	0.0522	0.0506	0.0514	0.0534	0.0528	0.0442	0.0518	0.0446	0.051
0	0.9	0.0528	0.053	0.0512	0.0524	0.0526	0.0768	0.0516	0.0458	0.0504
0.5	0.1	0.4896	0.6472	0.6396	0.6306	0.631	0.1798	0.6438	0.338	0.4112
0.5	0.5	0.5912	0.7412	0.734	0.7642	0.7608	0.343	0.7552	0.4686	0.5862
0.5	0.9	0.9224	0.9334	0.9292	0.9796	0.9802	0.892	0.9542	0.9052	0.9746
1.0	0.1	0.9004	0.9752	0.9736	0.978	0.9784	0.5504	0.9788	0.7466	0.8492
1.0	0.5	0.948	0.9908	0.9902	0.9952	0.9954	0.8034	0.9926	0.8806	0.9526
1.0	0.9	0.9978	0.999	0.999	1	1	0.9954	1	0.9958	1

Empirical size and power at $\alpha = 0.05$

(Table 12), T also had the highest power for low to moderate correlation, while S achieved the highest power with high correlation.

For all of the cases with 10 or fewer incomplete pairs on each variable, the estimated Type I error rates of Z_b were inflated and generally increased with correlation. This may be due to a poor asymptotic approximation for these cases, and further investigation into this statistic is needed to understand and address this issue.

4 Example

The methods discussed in this paper are illustrated using an example reported by Dubnicka et al. (2002). They selected a subset of data from a study to compare the efficacy of red krypton versus blue-green argon laser photocoagulation for the management of high-risk proliferative diabetic retinopathy. Patients with both eyes eligible were assigned both treatments, one in each eye, while those with only one eye eligible were randomly assigned to one of the two treatments. The result was both

Table 10 Normal distribution with $n_p = 10$, $n_{ux} = 5$, $n_{uy} = 5$

δ	ρ	T	R_z	R_{perm}	$R_{w,z}$	$R_{w,perm}$	Z_b	M	t	S
0	0.1	0.054	0.0554	0.0468	0.052	0.0512	0.0746	0.0494	0.049	0.0514
0	0.5	0.0554	0.0544	0.046	0.0492	0.0504	0.0884	0.0496	0.049	0.0514
0	0.9	0.0522	0.0536	0.0466	0.0494	0.0504	0.1032	0.0498	0.049	0.0514
0.5	0.1	0.3966	0.3788	0.345	0.3658	0.3648	0.2304	0.3516	0.291	0.2908
0.5	0.5	0.5132	0.5068	0.466	0.489	0.4846	0.3844	0.4434	0.4254	0.4282
0.5	0.9	0.9436	0.9288	0.9086	0.9114	0.9068	0.9416	0.7712	0.9436	0.938
1.0	0.1	0.8566	0.84	0.8134	0.8298	0.829	0.593	0.8076	0.6956	0.689
1.0	0.5	0.951	0.9482	0.9332	0.9414	0.9372	0.8564	0.899	0.8962	0.8872
1.0	0.9	0.9998	1	0.9992	0.9994	0.999	1	0.97	1	1

Empirical size and power at $\alpha = 0.05$ **Table 11** Normal distribution with $n_p = 20$, $n_{ux} = 10$, $n_{uy} = 10$

δ	ρ	T	R_z	R_{perm}	$R_{w,z}$	$R_{w,perm}$	Z_b	M	t	S
0	0.1	0.0562	0.051	0.0496	0.0528	0.0514	0.06	0.0508	0.0454	0.0464
0	0.5	0.0538	0.0472	0.0458	0.0486	0.0486	0.0762	0.0518	0.0454	0.0464
0	0.9	0.049	0.0502	0.0494	0.0518	0.052	0.0964	0.0514	0.0454	0.0464
0.5	0.1	0.6316	0.5936	0.5942	0.5996	0.6016	0.379	0.5854	0.4952	0.4722
0.5	0.5	0.7858	0.761	0.759	0.7664	0.7668	0.6434	0.7184	0.702	0.6776
0.5	0.9	0.999	0.9976	0.9972	0.9978	0.9972	0.9988	0.976	0.999	0.9978
1.0	0.1	0.9892	0.984	0.9832	0.9846	0.9842	0.887	0.9806	0.9384	0.926
1.0	0.5	0.999	0.9984	0.9988	0.9988	0.9988	0.9916	0.9966	0.9956	0.9928
1.0	0.9	1	1	1	1	1	1	0.9998	1	1

Empirical size and power at $\alpha = 0.05$ **Table 12** Normal distribution with $n_p = 5$, $n_{ux} = 7$, $n_{uy} = 8$

δ	ρ	T	R_z	R_{perm}	$R_{w,z}$	$R_{w,perm}$	Z_b	M	t	S
0	0.1	0.062	0.06	0.0494	0.0504	0.0512	0.0434	0.0532	0.0472	0.0604
0	0.5	0.0648	0.0598	0.0488	0.0542	0.0546	0.0638	0.057	0.0472	0.0604
0	0.9	0.0616	0.0612	0.0502	0.0518	0.0546	0.0884	0.0556	0.0472	0.0604
0.5	0.1	0.346	0.3188	0.2826	0.3158	0.3096	0.0806	0.3222	0.1686	0.1996
0.5	0.5	0.406	0.352	0.3144	0.3728	0.3714	0.1434	0.3826	0.2358	0.2798
0.5	0.9	0.7044	0.4688	0.4272	0.62	0.6174	0.5932	0.623	0.6608	0.713
1.0	0.1	0.7818	0.7338	0.6954	0.7446	0.7412	0.164	0.7554	0.4024	0.4484
1.0	0.5	0.8538	0.7786	0.7392	0.8366	0.8348	0.3554	0.8418	0.5776	0.6354
1.0	0.9	0.9878	0.832	0.7998	0.9352	0.9362	0.9634	0.9352	0.9906	0.9942

Empirical size and power at $\alpha = 0.05$

Table 13 Visual acuity measurements for 20 patients receiving both treatments and 20 patients who received only one treatment

Paired data										
X_i	4	69	87	35	39	79	31	79	65	95
Y_i	62	80	82	83	0	81	28	69	48	90
X_i	68	62	70	80	84	79	66	75	59	77
Y_i	63	77	0	55	83	85	54	72	58	68
Unpaired data										
T_i	36	86	39	85	74	72	69	85	85	72
C_i	88	83	78	30	58	45	78	64	87	65

Table 14 Test results for treatment difference, visual acuity measurements for 20 patients receiving both treatments and 20 patients who received only one treatment

Test	Statistic	p-value
t	0.543	0.297
T	3.147	0.279
Z_b	0.522	0.301
S	1.120	0.131
R_{perm}	189.0	0.134
$R_{w,perm}$	0.609	0.133
R_z	1.138	0.128
$R_{w,z}$	1.133	0.129
M	1.006	0.157

paired and unpaired data. Visual acuity as measured by the number of letters correctly read from a visual acuity chart was the outcome of interest. Table 13 shows the data. Table 14 reports the results of each of the tests, where the permutation tests were based on 10,000 random permutations. The tests based on means (t , T , Z_b) had similar p-values, although the T permutation test had a slightly lower p-value. Similarly, all of the tests based on ranks had similar p-values, with all versions of the tests based on the Dubnicka et al. (2002) statistics (R_{perm} , $R_{w,perm}$, D , D_w) with p-values less than the Wilcoxon signed-rank test (S), while the M test p-value was slightly higher than that of S .

5 Discussion

From the results of the simulations, several different tests stood out for certain cases, and it is hard to suggest a single best statistic. For normally distributed data, T is likely the best choice since it had the highest power more often than any of the other tests. However, since a researcher will not likely know the true distribution in practice, and since the rank based statistics were most powerful for nonnormal distributions and nearly as powerful as T for normal distributions, one of the rank based statistics will generally be the best choice. The light-tailed and heavy-tailed skewed distributions revealed that the weight on both $R_{w,z}$ and $R_{w,perm}$ was more important as the total

sample sizes and percent missing increased. R_z often had the highest power in the heavy-tailed symmetric case, along with $R_{w,z}$ and $R_{w,perm}$ as the sample size increased. Those statistics were also generally more powerful for low correlation, while M usually had higher power for high correlation. With about 50% or fewer incomplete pairs and high correlation, S was very competitive and often had the highest power for the nonnormal distributions. This suggests that we may not lose too much information by discarding the incomplete pairs in those cases. For small sample sizes, R_z and $R_{w,z}$ often had higher power than the permutation versions, but this may be due to poor approximations for the asymptotic statistics as higher estimated Type I error rate estimates were also observed in these cases. Thus, the permutation statistics R_{perm} and $R_{w,perm}$, which are guaranteed to provide unbiased power and Type I error rate estimates, should be preferred for small sample sizes.

References

- Bhoj DS (1978) Testing equality of means of correlated variates with missing data on both responses. *Biometrika* 65:225–228
- Bhoj DS (1984) On testing equality of variances of correlated variates with incomplete data. *Biometrika* 71:639–641
- Bhoj DS (1989) On comparing correlated means in the presence of incomplete data. *Biom J* 31:279–288
- Dubnicka SR, Blair RC, Hettmansperger TP (2002) Rank-based procedures for mixed paired and two-sample designs. *J Mod Appl Stat Methods* 1:32–41
- Einsporn RL, Habtzghi D (2013) Combining paired and two-sample data using a permutation test. *J Data Sci* 11:767–779
- Hoaglin DC (2006) Summarizing shape numerically: the g-and-h distributions. In: Hoaglin DC, Mosteller F, Tukey JW (eds) *Exploring data tables, trends, and shapes*. John Wiley & Sons, Inc., Hoboken, NJ
- Lin PE, Stivers LE (1974) On the difference of means with incomplete data. *Biometrika* 61:325–334
- Magel RC, Fu R (2014) Proposed nonparametric test for the mixed two-sample design. *J Stat Theory Pract* 8:221–237
- Maritz JS (1995) A permutation paired test allowing for missing values. *Aust J Stat* 37:153–159
- Marozzi M (2014) Multivariate tests based on interpoint distances with application to magnetic resonance imaging. *Stat Methods Med Res* 25(6)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.