Effect of unequal variance on statistical tests for mixed paired and two-sample designs

By: Bailey Meche, Nathan McEntire, Melissa Hall, & Scott Richter

Meche, B., McEntire, N., Hall, M. and Richter, S. (2024). Effect of unequal variance on statistical tests for mixed paired and two-sample designs. *Journal of Statistical Theory and Practice* **18**, 40. (DOI: https://doi.org/10.1007/s42519-024-00394-3).

Abstract:
Observations from studies comparing two treatments can involve both paired and independent samples. While tests have been proposed to utilize both sample types, most of these either assume equal variances across groups or have not been tested under the effect of unequal variance. Two methods have been proposed to handle the unequal variance case, but their relative performance is unknown. In this paper, a new procedure for the unequal variance case is proposed and an extensive simulation study is conducted to compare its performance to previously proposed methods. The new method is shown to be the most powerful test in certain cases. The simulation study suggests that tests that do not assume equal variance should be recommended as the default procedures.

Keywords: Mixed paired-unpaired design | Unequal variance | means comparison

Article:

1 Introduction

In certain experiments, a combination of paired and independent observations may occur. For example, studies where two treatments must be applied to the same subject may not be able to enroll a sufficient number of subjects. Dubnicka et al. [7] describes an experiment to compare two methods of laser eye surgery in which both eyes were available for a small number of subjects, while a larger number had only one eye eligible. Thus, subjects with both eyes eligible received both methods, randomly assigned to either the left eye or right eye, while the remaining subjects were randomly assigned to receive just one of the treatments. The resulting data consists of mixed paired and unpaired observations. Lin and Stivers [10] and Bhoj [1,2,3] proposed several statistics for combining the paired and unpaired data into a single test. Most of these tests assumed sample variances between treatments were equal.

More recently, Derrick et al. [6] also proposed a combined test that assumed equal variance. Dubnicka et al. [7] and Magel and Fu [11] suggested nonparametric tests using weighted averages of the Wilcoxon signed-rank and rank-sum statistics, but neither considered the unequal variance case. Lin and Stivers [10] was the first to propose a test that did not assume homoscedasticity between groups for which they suggested referring their statistic to an approximate t-distribution. Bhoj [3] found that this test suffered from inflated Type I error for small samples (less than 10), but did not examine any heteroscedastic cases with larger sample sizes. Derrick et al. [6] also proposed a statistic using all observations whose distribution was approximated using a t-distribution with degrees of freedom found by interpolating between the

degrees of freedom of the corresponding paired and Welch *t*-tests. However, their simulation study did not compare the performance of this test to the Lin and Stivers [10] test.

In this paper, a new procedure for the unequal variance case is proposed. In Sect. 2.1.7, we demonstrate that this test is uniformly more powerful than the method introduced by Derrick et al. [6]. An extensive simulation study is conducted to compare the performance of the new test to the previous methods proposed for the unequal variance case [6, 10]. Previously proposed methods assuming equal variance are also studied to explore their robustness to the equal variance assumption and to compare their performance to the methods that allow heteroscedasticity.

2 Methods

In this section, the methods to be evaluated in this paper are discussed. Notation utilized throughout each approach are summarized in Table 1.

2.1 Previous tests

2.1.1 Z_{ls} statistic [10]

The Z_{ls} statistic [10] was based on a maximum likelihood estimation of the mean difference δ using estimators for $\delta\frac{2}{1}$, $\sigma\frac{2}{2}$, and ρ where homoscedasticity was not assumed. The Z_{ls} statistic as interpreted by Bhoj [3] is given by:

$$Z_{ls} = \frac{\sqrt{n_p - 1}\,(g\bar{X}_p + (1-g)\,\bar{X}_u - h\bar{Y}_p - (1-h)\bar{Y}_u)}{\sqrt{\left(\frac{g^2}{n_p} + \frac{(1-g)^2}{n_{Xu}}\right)(n_p-1)S^2_{Xp} + \left(\frac{h^2}{n_p} + \frac{(1-h)^2}{n_{Yu}}\right)(n_p-1)S^2_{Yp} - \frac{2hg(n_p-1)S^2_{XpYp}}{n_p}}}, \quad (1)$$

Table 1. Commonly used notation

X, Y	Observation vectors
X_p, Y_p	Paired (X, Y) observations
X_u, Y_u	Independent X and Y observations
n_p	Number of total paired observations
nX_u	Number of independent observations from Sample X
nY_u	Number of independent observations from Sample Y
n_{Xp}	Number of paired observations from Sample X
n_{Yp}	Number of paired observations from Sample Y
nX	Total number of X observations ($nX_u + nX_p$)
nY	Total number of Y observations ($nY_u + nY_p$)
n	Total sample size ($n_X + n_Y$)
$\bar{X}$	Mean of all X observations
$\bar{Y}$	Mean of all Y observations
$S\frac{2}{X}$	Sample variance of all X observations
$S\frac{2}{Y}$	Sample variance of all Y observations

ρ	Pearson population correlation coefficient
δ	Population mean difference ($\mu_X - \mu_Y$)
γ	Population variance ratio (σ_X^2/σ_Y^2)

where *g, h* are given by

$$g = n\frac{n + n_{Yu} + \frac{n_{Xu}(n_p-1)S^2_{XpYp}}{(n_p-1)S^2_{Xp}}}{(n + n_{Xu})(n + n_{Yu}) - n_{Xu}n_{Yu}r^2},$$

$$h = n\frac{n + n_{Xu} + \frac{n_{Yu}(n_p-1)S^2_{XpYp}}{(n_p-1)S^2_{Yp}}}{(n + n_{Xu})(n + n_{Yu}) - n_{Xu}n_{Yu}r^2}.$$

The distribution of Z_{ls} is approximated by the Student's *t* distribution with n p degrees of freedom. In a simulation study including three proposed tests, the Z_{ls} test was the most powerful test for conditions of unequal variance provided that $n_p \geq 10$ [3].

2.1.2 T_b and Z_b statistics by Bhoj [1-3]

The T_b statistic [1] is a weighted combination of paired and independent *t*-statistics assuming the data are normally distributed and can be approximated by Student's *t*-distribution. The statistic is given by

$$T_b = \lambda\frac{\bar{X} - \bar{Y}}{\frac{s}{\sqrt{n}}} + (1 - \lambda)\frac{\bar{X}_u - \bar{Y}_u}{S_{X_u,Y_u}\sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}}. \quad (2)$$

Bhoj [3] further developed this approach by proposing two new test statistics for comparing correlated means in the presence of incomplete data, most notably the Z_b test. A simulation study revealed that the Z_b test consistently outperformed Bhoj's earlier proposed statistics [1, 2]. This statistic is given by

$$Z_b = \frac{\lambda_b U_1 + (1 - \lambda_b)U_3}{\sqrt{\lambda_b^2 + (1 - \lambda_b)^2}}, \quad (3)$$

where λ_b was a weighting constant for the transformed *t*-statistics $U_1$and U_3. See Bhoj [3] for transformation details. The Z_b statistic was shown in a simulation study to perform best for normally distributed data when sample variances were equal. This test is recommended when both the sample sizes were less than 10 and the larger sample was associated with the larger variance. Bhoj [3] noted that the Z_b statistic only performed well for this case and did not perform well when the assumption of equal variance was relaxed.

2.1.3 R and R_w statistics [7]

Two rank-based nonparametric statistics were proposed [7] that were weighted (R_w) and unweighted (R) sums of the Wilcoxon signed-rank statistic, S, for paired data and the Wilcoxon-Mann–Whitney statistic, U, for independent samples. The R statistic is given by (4) and R_w by (5):

$$R = S + U; \tag{4}$$

$$R_w = \frac{2(n_1 + n_2)}{n(n_1 + n_2) + 2n_1n_2} S + \frac{2}{n(n_1 + n_2) + 2n_1n_2} U. \tag{5}$$

These statistics were based on a location-shift model and thus equal population spread. In a simulation study, the weighted R_w rank statistic was generally more efficient than the unweighted R rank statistic. However, the authors recommended using the R statistic instead of R_w in cases when the number of paired observations is at least as large as the total number of independent observations ($n \geq n_1 + n_2$), when R is nearly as efficient as R_w.

2.1.4 *M* statistic [11]

[11] modified the weighted test statistic R_w. They proposed a statistic M by first using the standardized Wilcoxon signed-rank S^* and Wilcoxon-Mann–Whitney U^* statistics, then re-standardizing their sum as

$$M = \frac{S^* + U^*}{\sqrt{2}}. \tag{6}$$

A simulation study found that if the variance of the unpaired data was equal to the variance of the paired data, M had approximately the same power as R_w. However, when the sample sizes for the unpaired data were equal to or larger than that of the paired data, their statistic M exhibited higher power. It is important to note that their proposed test was only compared to R_w, and no comparisons were made to other previous methods. They also did not evaluate their statistic under unequal variances between groups.

2.1.5 $T_{new}1$ and $T_{new}2$ statistics [6]

Two new test statistics were introduced [6] for comparing means between two samples that include both paired and independent observations. In contrast to the methods described earlier, where separate paired and unpaired statistics were computed and then combined, the sample means and variances in the statistics below are calculated using all available observations, including both paired and independent samples from both X and Y treatment groups.

$T_{new}1$ statistic for equal variances

The first statistic Tnew1 was proposed as a combination of the pooled and paired t-tests. This test assumes $\sigma_X^2 = \sigma_Y^2$. The test statistic is given by

$$T_{new1} = \frac{\bar{X} - \bar{Y}}{S_p \sqrt{\frac{1}{n_X} + \frac{1}{n_Y} - 2\rho \frac{n_p}{n_X n_Y}}}, \tag{7}$$

where

$$S_p = \sqrt{\frac{(n_X - 1)S_X^2 + (n_Y - 1)S_Y^2}{(n_X - 1) + (n_Y - 1)}}.$$

The $T_{new}1$ statistic is referenced against the t-distribution with degrees of freedom, $v_{new}1$, derived by linear interpolation between the degrees of freedom of the pooled and paired t-tests:

$$v_{new1} = (n_p - 1) + \left(\frac{n_X + n_Y + n_p - 1}{n_X + n_Y + 2n_p}\right)(n_X + n_Y).$$

2.1.6 $T_{new}2$ statistic for unequal variance

The second statistic $T_{new}2$ was proposed as a combination of paired and Welch's t-tests, and is defined as

$$T_{new2} = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{S_X^2}{n_X} + \frac{S_Y^2}{n_Y} - 2\rho \frac{S_X S_Y n_p}{n_X n_Y}}}. \tag{8}$$

The $T_{new}2$ statistic is referenced against the t-distribution with degrees of freedom $v_{new}2$ derived by linear interpolation between the Welch-Satterthwaite approximate degrees of freedom and the degrees of freedom of the paired t-test:

$$v_{new2} = (n_p - 1) + \left(\frac{\gamma - n_p - 1}{n_{Xu} + n_{Yu} + 2n_p}\right)(n_{Xu} + n_{Yu}), \tag{9}$$

where

$$\gamma = \frac{\left(\frac{S_X^2}{n_X} + \frac{S_Y^2}{n_Y}\right)^2}{\frac{\left(\frac{S_X^2}{n_X}\right)^2}{n_X - 1} + \frac{\left(\frac{S_Y^2}{n_Y}\right)^2}{n_Y - 1}}.$$

In cases where there are no independent observations available, both $T_{new}1$ and $T_{new}2$ default to the paired samples *t*-statistic. Similarly, when there are no paired observations, $T_{new}1$ defaults to the two-sample *T* -statistic, and $T_{new}2$ defaults to Welch's test statistic.

2.1.7 Modification of $T_{new}2$ test

The test based on $T_{new}2$ used the Pearson product-moment estimator, r, of the population correlation, ρ. However, it is known that r is biased for estimating ρ [8]. Thus, a modified statistic is proposed:

$$T_{adj} = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{S_X^2}{n_X} + \frac{S_Y^2}{n_Y} - 2\rho^* \frac{S_X S_Y n_p}{n_X n_Y}}}, \tag{10}$$

where ρ^* is an unbiased estimator of ρ [12] given by

$$\rho^* = \left(\frac{\text{cov}(x, y)}{S_{Xp} S_{Yp}}\right)\left(1 - \frac{1 - \left(\frac{\text{cov}(x,y)}{S_{Xp} S_{Yp}}\right)^2}{2(n-3)}\right).$$

Further, the distribution of $T_{new}2$ was estimated as a t-distribution with degrees of freedom derived by linear interpolation between the Welch-Satterthwaite approximate degrees of freedom and the degrees of freedom of the paired *t*-test. However, linear interpolation is a conservative estimate between two degrees of freedom [9]. Thus, a less conservative approximation is proposed, which is the sum of the paired *t*-test and Welch-Satterthwaite degrees of freedom:

$$f_{adj} = (n_p + 1) + \frac{\left(\frac{S_Y^2}{n_x} + \frac{S_Y^2}{n_Y}\right)^2}{\frac{\left(\frac{S_X^2}{n_X}\right)^2}{n_X - 1} + \frac{\left(\frac{S_Y^2}{n_Y}\right)^2}{n_Y - 1}}. \tag{11}$$

This modified procedure will have uniformly higher power than the $T_{new}2$ test.

Comparison of degrees of freedom f_{adj} and $v_{new}2$

Let f_1 be the degrees of freedom for the paired *t*-test and f_2 for the Welch-Satterthwaite test. It is known that both f_1 and $f2$ are strictly greater than 0. The degrees of freedom $v_{new}2$ for Derrick's $T_{new}2$ statistic is derived using a linear interpolation between $f1$ and $f2$ (9) and T_{adj} is derived by the sum (11).

Proof We begin with the statement of f_{adj}:

$$f_{adj} = f1 + f2.$$

Assume that $n_p > 1$. Then, $f2 > f2 - n_p - 1$ and

$$f_{adj} > f1 + f2 - n_p - 1.$$

Next, since n_p, n_{Xu}, $n_{Yu} > 0$, we have $\frac{n_{Xu}+n_{Yu}}{n_{Xu}+n_{Yu}+2n_p} < 1$. Then,

$$f_{adj} > f_1 + f_2 - n_p - 1 > f_1 + (f_2 - n_p - 1)\left(\frac{n_{Xu} + n_{Yu}}{n_{Xu} + n_{Yu} + 2n_p}\right)$$

For which we have

$$\begin{aligned} f_1 + (f_2 - n_p - 1)\left(\frac{n_{Xu} + n_{Yu}}{n_{Xu} + n_{Yu} + 2n_p}\right) &= f_1 + \left(\frac{f_2 - n_p - 1}{n_{Xu} + n_{Yu} + 2n_p}\right)(n_{Xu} + n_{Yu}) \\ &= (n_p - 1) + \left(\frac{\gamma - n_p - 1}{n_{Xu} + n_{Yu} + 2n_p}\right)(n_{Xu} + n_{Yu}) \\ &= v_{new2}. \end{aligned}$$

Therefore, $f_{adj} > v_{new2}$ where n_p, n_{Xu}, $n_{Yu} > 0$ and $n_p > 1$.

3 Simulation study

Power and Type I error rates of all tests discussed in Sect. 2, including the new approach, were estimated. The Wilcoxon signed-rank test *S* and the standard paired *t*-test *t* were also included to compare against the extreme case of discarding the independent observations.

Table 2. Sample size settings

Total sample size n	Sample sizes (n_p, n_{Xu}, n_{Yu})	Paired proportion n_p/n	Independent proportion ($n_{Xu} + n_{Yu}$
$n = 20$	(5,5,10)	0.25	0.33
	(5,10,5)	0.25	0.66
	(5,7,8)	0.25	0.50
	(5,8,7)	0.25	0.50
	(10,5,5)	0.50	0.50
	(10,7,3)	0.50	0.70
	(10,3,7)	0.50	0.30
	(15,2,3)	0.75	0.40
	(15,3,2)	0.75	0.60
$n = 60$	(15,15,30)	0.25	0.33
	(15,30,15)	0.25	0.66
	(15,22,23)	0.25	0.50
	(30,15,15)	0.50	0.50
	(30,21,9)	0.50	0.70
	(30,9,21)	0.50	0.30
	(45,7,8)	0.75	0.50
	(45,8,7)	0.75	0.50

3.1 Design

Paired samples were generated with Pearson correlation values of $\rho = 0.1, 0.3, 0.5$, and 0.9. Samples were generated using either $\gamma = \sigma_X^2/\sigma_Y^2 = 1$, 2, or 4. Normally distributed data were generated as bivariate normal pairs (X, Y) with specified means, variances and correlation. Observations were randomly deleted from each variable separately to create unpaired observations. Type I error was estimated using a mean difference of $\delta = \mu_X - \mu_Y = 0$ at $\alpha = 0.05$, and power was estimated for $\delta = 0.5$ and 1. For all mean difference tests, the tested null and alternative hypotheses are given by:

$$H_0: \delta = 0 \qquad H_1: \delta > 0.$$

Overall sample sizes of $n = 20$ and $n = 60$ were chosen to represent small and moderate samples sizes, respectively. Each sample size n was tested with several proportions of paired and unpaired observations, $\frac{n_p}{n_X + n_Y + n_p}$. Also considered were different equal and unequal sets of unpaired X and Y. Simulation sample size settings are given in Table 2. A given sample size tuple is given by (n_p, n_{Xu}, n_{Yu}). Type I error and power were estimated as the proportion of rejections out of 5000 data sets for each variance, mean difference, sample size, and correlation configuration.

3.2 Simulation results

A comparison of several statistical tests for their effectiveness in handling unequal variance is included in this section. Many of these, as seen in the literature review, have not been tested for this effect of unequal variance nor compared under the same conditions. The statistics considered in this study include Z_{ls} (1) by Lin and Stivers [10], T_b (2) and Z_b (3) by [3], R (4) and R_w (5) by [7], M (6) by [11], $T_{new}1$ (7) and $T_{new}2$ (8) by Derrick et al. [6], the Wilcoxon signed-rank statistic S, the paired t-statistic, and our proposed T_{adj} (10). Seven simulations in this section selected from Table 2 are discussed as representative of all results. All simulation results are listed in the Appendix.[1]

Table 3 Type I error estimates for samples (5, 5, 10) for $\gamma = 2, 4$

γ	ρ	T_{adj}	T_{new1}	T_{new2}	Z_{ls}	Z_b	T_b	M	R	R_w	t	S
2	0.1	0.0594	0.0812	0.0526	0.0734	0.0776	0.0792	0.0690	0.0768	0.0638	0.0422	0.0522
2	0.3	0.0588	0.0806	0.0504	0.0792	0.0772	0.0788	0.0678	0.0754	0.0628	0.0430	0.0528
2	0.5	0.0594	0.0846	0.0506	0.0824	0.0778	0.0764	0.0680	0.0776	0.0634	0.0430	0.0544
2	0.9	0.0612	0.0956	0.0530	0.0772	0.0682	0.0738	0.0664	0.0754	0.0608	0.0424	0.0524
4	0.1	0.0604	0.0950	0.0508	0.0754	0.0924	0.0962	0.0794	0.0940	0.0740	0.0394	0.0536
4	0.3	0.0598	0.1012	0.0520	0.0776	0.0922	0.0972	0.0776	0.0952	0.0726	0.0396	0.0532
4	0.5	0.0616	0.1084	0.0516	0.0772	0.0922	0.0974	0.0776	0.0928	0.0726	0.0396	0.0530
4	0.9	0.0638	0.1270	0.0520	0.0660	0.0908	0.0958	0.0794	0.0938	0.0758	0.0406	0.0548

3.2.1 Type I error

To define which estimates are admissible, a confidence interval for 5000 observations is given by $[\alpha - 0.015, \alpha + 0.015]$. For our simulations at $\alpha = 0.05$, this choice is a moderate choice between [4]'s criteria for liberal and stringent meanings of robust Type I error rate, specifically $0.7\alpha \le \pi \le 1.3\alpha$. All of the tests derived under the assumption of equal population variances (T_b, Z_b, R, M, and $T_{new}1$) showed inflated Type I error estimates under heteroscedasticity and were thus

determined to be inadmissible (See Table 3). In contrast, $T_{new}2$ and Tadj did not suffer from this problem and were always admissible. Z_{ls} occupied a middle ground, showing inflated Type I error for $n_p < 10$ (See Table 3) but acceptable error estimates for all cases where $n_p \geq 10$ (See Table 4).

Table 4 Type I error estimates for samples (10, 7, 3) for γ = 2, 4

γ	ρ	T_{adj}	T_{new1}	T_{new2}	Z_{ls}	Z_b	T_b	*M*	*R*	R_w	*t*	*S*
2	0.1	0.0524	0.0362	0.0458	0.0512	0.0414	0.0428	0.0402	0.0476	0.0452	0.0438	0.0496
2	0.3	0.0524	0.0382	0.0462	0.0508	0.0406	0.0414	0.0412	0.0464	0.0446	0.0448	0.0488
2	0.5	0.0514	0.0406	0.0448	0.0504	0.0384	0.0420	0.0412	0.0450	0.0438	0.0450	0.0480
2	0.9	0.0544	0.0496	0.0458	0.0502	0.0366	0.0526	0.0410	0.0480	0.0456	0.0436	0.0488
4	0.1	0.0508	0.0342	0.0436	0.0526	0.0346	0.0376	0.0374	0.0454	0.0440	0.0462	0.0494
4	0.3	0.0518	0.0396	0.0444	0.0508	0.0340	0.0378	0.0362	0.0452	0.0426	0.0466	0.0486
4	0.5	0.0530	0.0460	0.0446	0.0502	0.0338	0.0398	0.0370	0.0456	0.0434	0.0466	0.0488
4	0.9	0.0566	0.0714	0.0470	0.0520	0.0326	0.0420	0.0350	0.0446	0.0416	0.0466	0.0474

3.2.2 Power

Power was estimated for all tests found to be admissible. T_{adj} and Z_{ls} consistently had the highest power. Among these two tests, T_{adj} demonstrated the higher power for scenarios with low correlation ($\rho = 0.1$), Z_{ls} higher power for high correlation ($\rho = 0.9$), and the two had similar power with moderate correlation ($\rho = 0.5$). Tables 5 and 6 demonstrate these findings for small samples, and Tables 7 and 8 show similar relationships with larger samples. These relationships were consistent across all sample sizes and configurations of paired and unpaired observations. For cases with $\rho = 0.9$, where Z_{ls} enjoyed a power advantage over T_{adj}, the disparity tended to be larger than when T_{adj} was more powerful ($\rho = 0.1$). However, these disparities became smaller as the true effect (δ) became larger. (See Tables 5 and 6). An interesting finding was that one of T_{adj} or Z_{ls} was also most powerful even when variances were equal (See Table 9).

Table 5 Power estimates for samples (10, 7, 3) for δ = 0.5, 1.0

$\delta = 0.5$							$\delta = 1.0$						
γ	ρ	T_{adj}	Z_{ls}	T_{new2}	*t*	*S*	γ	ρ	T_{adj}	Z_{ls}	T_{new2}	*t*	*S*
1	0.1	0.3972	0.3706	0.3668	0.2896	0.2832	1	0.1	0.8624	0.8410	0.8464	0.7062	0.6998
1	0.3	0.4454	0.4240	0.4090	0.3440	0.3422	1	0.3	0.9046	0.8890	0.8862	0.7958	0.7884
1	0.5	0.5032	0.5102	0.4654	0.4292	0.4248	1	0.5	0.9426	0.9456	0.9334	0.8966	0.8892
1	0.9	0.6724	0.9540	0.6394	0.9426	0.9392	1	0.9	0.9894	1.0000	0.9866	1.0000	1.0000
2	0.1	0.2346	0.2178	0.2154	0.1586	0.1646	2	0.1	0.5616	0.5246	0.5296	0.3702	0.3686
2	0.3	0.2550	0.2414	0.2302	0.1764	0.1816	2	0.3	0.6140	0.5820	0.5836	0.4248	0.4266
2	0.5	0.2812	0.2770	0.2526	0.2028	0.2070	2	0.5	0.6730	0.6610	0.6416	0.4970	0.4956
2	0.9	0.3596	0.4716	0.3330	0.3274	0.3284	2	0.9	0.8196	0.9312	0.7986	0.7818	0.7722
4	0.1	0.1276	0.1270	0.1100	0.0912	0.0970	4	0.1	0.2606	0.2480	0.2386	0.1712	0.1734
4	0.3	0.1310	0.1268	0.1154	0.0940	0.1018	4	0.3	0.2744	0.2636	0.2476	0.1802	0.1840
4	0.5	0.1370	0.1348	0.1210	0.0984	0.1050	4	0.5	0.2862	0.2836	0.2604	0.1902	0.1984
4	0.9	0.1506	0.1620	0.1322	0.1120	0.1182	4	0.9	0.3188	0.3568	0.2920	0.2250	0.2310

Table 6 Power estimates for samples (10, 3, 7) for δ = 0.5, 1.0

$\delta = 0.5$							$\delta = 1.0$						
γ	ρ	T_{adj}	Z_{ls}	T_{new2}	t	s	γ	ρ	T_{adj}	Z_{ls}	T_{new2}	t	s
1	0.1	0.3970	0.3650	0.3744	0.2894	0.2872	1	0.1	0.8644	0.8356	0.8518	0.6980	0.6962
1	0.3	0.4390	0.4160	0.4120	0.3350	0.3380	1	0.3	0.9024	0.8936	0.8890	0.7924	0.7862
1	0.5	0.4944	0.5056	0.4636	0.4208	0.4230	1	0.5	0.9414	0.9452	0.9326	0.8976	0.8884
1	0.9	0.6772	0.9550	0.6500	0.9458	0.9404	1	0.9	0.9938	1.0000	0.9912	1.0000	1.0000
2	0.1	0.2098	0.1992	0.1868	0.1610	0.1660	2	0.1	0.5102	0.4802	0.4782	0.3766	0.3804
2	0.3	0.2246	0.2174	0.1974	0.1770	0.1842	2	0.3	0.5454	0.5300	0.5170	0.4282	0.4308
2	0.5	0.2428	0.2476	0.2154	0.2058	0.2076	2	0.5	0.5848	0.5978	0.5554	0.5028	0.5062
2	0.9	0.2990	0.4384	0.2728	0.3320	0.3320	2	0.9	0.7030	0.9096	0.6678	0.7918	0.7852
4	0.1	0.1180	0.1190	0.1038	0.0988	0.1032	4	0.1	0.2220	0.2202	0.2010	0.1784	0.1824
4	0.3	0.1214	0.1226	0.1044	0.1030	0.1084	4	0.3	0.2302	0.2372	0.2070	0.1874	0.1918
4	0.5	0.1242	0.1290	0.1072	0.1084	0.1110	4	0.5	0.2414	0.2504	0.2154	0.1990	0.2058
4	0.9	0.1310	0.1526	0.1146	0.1208	0.1234	4	0.9	0.2654	0.3336	0.2404	0.2346	0.2370

Table 7 Power estimates for samples (15, 22, 23) for δ = 0.5

γ	ρ	T_{adj}	Z_{ls}	T_{new2}	t	S
1	0.1	0.6888	0.6688	0.6842	0.3900	0.3978
1	0.3	0.7192	0.7134	0.7130	0.4692	0.4682
1	0.5	0.7572	0.7746	0.7502	0.5826	0.5772
1	0.9	0.8338	0.9960	0.8288	0.9906	0.9890
2	0.1	0.4072	0.3918	0.3974	0.2156	0.2186
2	0.3	0.4262	0.4228	0.4190	0.2436	0.2492
2	0.5	0.4472	0.4658	0.4404	0.2864	0.2916
2	0.9	0.4898	0.7504	0.4794	0.4778	0.4796
4	0.1	0.1880	0.1858	0.1818	0.1162	0.1184
4	0.3	0.1896	0.1928	0.1840	0.1214	0.1224
4	0.5	0.1966	0.2046	0.1872	0.1270	0.1284
4	0.9	0.2070	0.2686	0.2000	0.1414	0.1458

Table 8: Power estimates for samples (45, 8, 7) for δ = 0.5

γ	ρ	T_{adj}	Z_{ls}	T_{new2}	t	S
1	0.1	0.8356	0.8288	0.8318	0.7802	0.7636
1	0.3	0.9008	0.8988	0.8962	0.8658	0.8480
1	0.5	0.9526	0.9596	0.9508	0.9462	0.9362
1	0.9	1.0000	1.0000	1.0000	1.0000	1.0000
2	0.1	0.5142	0.5096	0.5058	0.4568	0.4410
2	0.3	0.5730	0.5666	0.5634	0.5198	0.5008
2	0.5	0.6414	0.6490	0.6330	0.6044	0.5796
2	0.9	0.8312	0.9226	0.8266	0.8714	0.8562
4	0.1	0.2240	0.2190	0.2204	0.1988	0.1892
4	0.3	0.2350	0.2308	0.2298	0.2130	0.2022
4	0.5	0.2474	0.2468	0.2416	0.2282	0.2174
4	0.9	0.2868	0.3074	0.2804	0.2714	0.2628

Table 9 Power estimates for samples (10, 5, 5) for γ = 1

δ	ρ	T_{adj}	T_{new1}	T_{new2}	Z_{ls}	Z_b	T_b	M	R	R_w	t	S
0.5	0.1	0.4036	0.3768	0.3730	0.3738	0.3744	0.3814	0.3388	0.3770	0.3596	0.2800	0.2842
0.5	0.3	0.4490	0.4182	0.4142	0.4270	0.4248	0.4196	0.3814	0.4306	0.4100	0.3352	0.3388
0.5	0.5	0.5042	0.4740	0.4660	0.5100	0.5086	0.4750	0.4348	0.5072	0.4866	0.4262	0.4200
0.5	0.9	0.6798	0.6566	0.6496	0.9562	0.9506	0.6532	0.7602	0.9234	0.9004	0.9430	0.9370
1.0	0.1	0.8682	0.8538	0.8520	0.8422	0.8488	0.8538	0.8070	0.8390	0.8314	0.6996	0.6920
1.0	0.3	0.9046	0.8960	0.8950	0.8906	0.8974	0.8954	0.8502	0.8924	0.8850	0.7910	0.7816
1.0	0.5	0.9422	0.9350	0.9328	0.9480	0.9480	0.9342	0.8956	0.9422	0.9346	0.8934	0.8860
1.0	0.9	0.9922	0.9900	0.9896	1.0000	1.0000	0.9920	0.9644	0.9998	0.9984	1.0000	1.0000

4 Example

The tests discussed in the paper are applied to experimental data on the effect of movie genre on sleep fragmentation score: the quality of sleep for an individual for a given night [5]. The research question was whether watching a horror movie leads to higher sleep fragmentation compared to watching a 'feel-good' movie. Study participants were randomly assigned to either a between-subjects design (stage 1), representing independent observations, or a repeated measures (stage 2) part of the investigation, representing paired observations. In the first stage of the study, the sleep fragmentation score is taken over one night, for two groups of individuals. Participants in the single movie design watch either a 'horror' or a 'feel-good' movie before bedtime. Participants in the repeated measures group watch a 'feel good' movie and a 'horror' movie on two alternate nights before bedtime, where the order of the viewing is randomized for each participant. The sleep fragmentation scores are given in Table 10.

Table 10 Sleep fragmentation scores obtained for each individual (ID)

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

Several sample statistics may be observed from this example. This example is fitting for our discussion since the ratio of sample variances is $S_X^2/S_Y^2 = 2$. These data also have a moderately high correlation estimate given by the sample Pearson correlation of 0.74. Table 11 shows the *p*-values for our discussed tests.

Table 11 p-values of tests for sleep fragmentation study

Test	p-value

T_{adj}	0.010
Z_{IS}	0.021
$T_{new}1$	0.026
$T_{new}2$	0.026
S	0.088
Paired t	0.111
Independent t	0.118

Recall that the Z_{IS} test is inadmissible for fewer than 10 paired observations. This data set only includes 8 paired observations. With this exception, we observe the same order of power as from the power simulations in Sect. 3.

5 Discussion

A comprehensive study was conducted of the relative performance of several tests to compare means from the data that contain both paired and unpaired observations. While no single procedure emerged as the overall best choice, the simulations suggest the following recommendations:

1. The proposed test based on T_{adj} is recommended where the number of paired observations is less than 10 for all correlation levels, and for $n_p \geq 10$ where the correlation between paired observations is low.
2. The Z_{IS} test is recommended for $n_p \geq 10$ where the correlation between paired observation is moderate to high.

The previously proposed tests derived under the assumption of equal population variance - T_b [1], Z_b [3], R/R_w [7], M [11], and $T_{new}1$ [6]—are not recommended if equal variances cannot be assumed as all showed inflated Type I error rates for these cases. Further, the tests based on T_{adj} and Z_{IS} showed comparable power to these tests for the equal variance cases, and thus T_{adj} and Z_{IS} should be regarded as the default tests whenever the true variance ratio, σ_X^2/σ_Y^2, is unknown.

Declarations

Conflict of interest. On behalf of all authors, the corresponding author states that there is no Conflict of interest.

References

1. Bhoj DS (1978) Testing equality of means of correlated variates with missing observations on both responses. Biometrika 65(1):225–228. https://doi.org/10.2307/2335301
2. Bhoj DS (1984) On testing equality of variances of correlated variates with incomplete data. Biometrika 71(3):639–641. https://doi.org/10.1093/biomet/71.3.639
3. Bhoj DS (1989) On comparing correlated means in the presence of incomplete data. Biometrika 31(3):279–288. https://doi.org/10.1002/bimj.4710310304
4. Bradley JV (1978) Robustness? Br J Math Stat Psychol 31(2):144–152. https://doi.org/10.1111/j.2044-8317.1978.tb00581.x
5. Derrick B (2017) How to compare the means of two samples that include paired observations and independent observations: a companion to derrick, russ, toher and white. Quant Methods Psychol 13(2):120–126. https://doi.org/10.20982/tqmp.13.2.p120

6. Derrick B, Russ B, Toher D et al (2017) Test statistics for the comparison of means for two samples that include both paired and independent observations. J Modern Appl Stat Methods 16(1):137–157. https://doi.org/10.22237/jmasm/1493597280
7. Dubnicka S, Blair RC, Hettmansperger TP (2002) Rank-based procedures for mixed paired and two-sample designs. J Modern Appl Stat Methods 1(1):6. https://doi.org/10.22237/jmasm/1020254460
8. Fisher RA (1915) Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. Biometrika 10:507–521. https://doi.org/10.2307/2331838
9. Knafl G, Sacks J, Ylvisaker D (1985) Confidence bands for regression functions. J Am Stat Assoc 80(391):683–691. https://doi.org/10.2307/2288485
10. Lin P, Stivers LE (1974) On difference of means with incomplete data. Biometrika 61(2):325–334. https://doi.org/10.2307/2334361
11. Magel RC, Fu R (2014) Proposed nonparametric test for the mixed two-sample design. J Stat Theory Pract 8(2):221–237. https://doi.org/10.1080/15598608.2014.847768
12. Olkin I, Pratt JW (1958) Unbiased estimation of certain correlation coefficients. Ann Math Statist 29(1):203. https://doi.org/10.1214/aoms/1177706717