ISSN: 0067-2904

Robust Tests for the Mean Difference in Paired Data using Jackknife Resampling Technique

Ghufran A. Ghadhban*, Huda A. Rasheed
Department of Mathematics, College of Science, Mustansiriyah University, Baghdad, Iraq

Received: 20/10/2020 Accepted: 15/12/2020

Abstract

The paired sample t-test is a type of classical test statistics that is used to test the difference between two means in paired data, but it is not robust against the violation of the normality assumption. In this paper, some alternative robust tests are suggested by combining the Jackknife resampling with each of the Wilcoxon signed-rank test for small sample size and Wilcoxon signed-rank test for large sample size, using normal approximation. The Monte Carlo simulation experiments were employed to study the performance of the test statistics of each of these tests depending on the type one error rates and the power rates of the test statistics. All these tests were applied on different sample sizes generated from three distributions, represented by Bivariate normal distribution, contaminated Bivariate normal distribution, and Bivariate exponential distribution.

Keywords: Paired t-test, Robust, Jackknife, Wilcoxon signed-rank test, Contaminated Bivariate normal, Bivariate Exponential.

الاختبارات الحصينة للفرق بين المتوسطين في البيانات المترابطة بأستخدام تقنية اعادة التشكيل Jackknife

غفران علي غضبان* ، هدى عبد الله رشيد
قسم الرياضيات، كلية العلوم، الجامعة المستنصرية بغداد، العراق

الخلاصة

ان اختبار t للعينة المزدوجة هو احد الاختبارات التقليدية التي تستخدم لاختبار الفرق بين متوسطي متغيرين في البيانات المزدوجة، ولكنه ليس حصيناً ضد اختراق البيانات لشرط التوزيع الطبيعي. تم في هذا البحث اقتراح بعض الاختبارات الحصينة البديلة من خلال دمج تقنية اعادة المعاينة Jackknife مع كلا من اختبار Wilcoxon signed–rank test للعينات الصغيرة و احتبار Wilcoxon signed–rank test الذي تم تقريبه الى التوزيع الطبيعي والذي يستخدم للعينات الكبيرة. تم توظيف المحاكاة بطريقة مونت–كارلو لدراسة أداء احصاءة الاختبار لكل من هذه الإختبارات بالإعتماد على معدلات الخطأ من النوع الأول و معدلات قوة الاختبار لإحصاءات الإختبار. جميع هذه الإختبارات تم تطبيقها على عينات باحجام مختلفة تم توليدها من ثلاثة توزيعات تمثلت بالتوزيع الطبيعي الثنائي، التوزيع الطبيعي الثنائي الملوث، والتوزيع الثنائي الأسي.

1. Introduction

Comparing two means of correlated variables is often of interest to researchers in various fields, especially medical and biological. The paired t-test is one of the most important tests that are widely used for this purpose. However, the paired t-test is not robust against the departure of the normality

*Email: ghyfrang1994@gmail.com

assumption. The robustness concept was introduced initially by Box in 1953. There are many definitions of the concept of robustness, perhaps the most important of which is that stipulated in the Huber definition (1981) [1]. That definition implies that the robustness has many meanings and implications that may be inconsistent with each other, but robustness can be expressed as referring to insensitivity to slight departures from the assumptions of the test statistics.

Bradley (1978) defined the Robust test and stipulated that the test is called robust against the violation of one or more of the test's assumptions, if that violation has no effect on the distribution of the test statistic, due to tending the true probability of a Type I error to differ from the nominal α. He suggested the liberal criterion to represent the robustness; the test could be regarded as robust only if its Type 1 error rate $\hat{\alpha}$ falls in the following interval: [2]

$0.9\ \alpha < \hat{\alpha} < 1.1\ \alpha$

i.e.,

$$|\hat{\alpha} - \alpha| \ \leq \ \frac{\alpha}{10} \tag{1}$$

On the other hand, Salter and Fawcett (1985) proposed another criterion for the robustness of the test, which does require the Type I error values to lie within the following interval [3]

$$\alpha \ \pm 2\sqrt{[\alpha\ (1-\alpha)/R]} \tag{2}$$

where R represents the replicated times.

This research aims to investigate the effects of the violation of some assumptions of the hypothesis test equality of means of two correlated variables on the distribution of test statistics.

These violations are represented by the following points:

1. Violation of the normality assumption due to the existence of outliers.
2. The smallness of the sample size.
3. The paired data follow a distribution other than the normal distribution.
4. Heterogeneity of the variances of the two dependent variables.

The main goal of this research is to find a robust test that achieves the highest power of the test when the set of paired data violates the assumptions of the normality and the homogeneity of variances of the correlated variables. Therefore, a number of robust tests are suggested, represented by Wilcoxon–matched pairs signed-ranks using Jackknife (JWS), Wilcoxon–matched pairs signed-ranks when sample size n > 25 using Jackknife (JWL), in addition to Jackknifing paired t-test (JT).

2. Test statistics

2.1 Paired t-test

The paired t-test is one of the widely applied statistical procedures that is used to determine the difference in the mean values between two sets of observations (before and after the treatment). The paired t-test is based on the differences between the values of a single pair, whose variations are almost normally distributed. It sometimes called the correlated pairs t-test, also known as the repeated measurements. In general, the test can be performed through 4 steps.

Suppose that the two-dimensional random variables (X, Y) have a bivariate normal distribution with parameters $\mu_X, \mu_Y\ , \sigma_X\ , \sigma_Y$ and ρ, if their joint pdf is defined as: [4, 5]

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho\ 2}}\ exp\left\{\frac{-1}{2(1-\rho\ 2)}\left[(\frac{x-\mu_X}{\sigma_X})^2 - 2\rho(\frac{x-\mu_X}{\sigma_X})(\frac{y-\mu_Y}{\sigma_Y}) + (\frac{y-\mu_Y}{\sigma_Y})^2\right]\right\},\ -\infty < x, y < \infty \tag{3}$$

where, $\mu_X, \mu_Y\ \in R\ , \sigma_X\ , \sigma_Y \in +R$ and $\rho \in (-1,1)$ and

$$\rho = Corr(X,Y\ )\ = \frac{Cov(X,Y)}{\sigma_X\sigma_Y}$$

The test has two hypotheses; the null hypothesis assumes that the true mean difference between the paired samples is zero:

$$H_0: \mu_X = \mu_Y \tag{4}$$

Whereas the alternative hypothesis is: $H_1: \mu_X \neq \mu_Y$

Let

$$D_i = \ X_i - \ Y_i \quad , \text{i} = 1, 2, \ldots, \text{n} \tag{5}$$

Then, $D \sim \text{Normal}(\mu_D, \sigma_D^2)$, where $\mu_D = \mu_X - \mu_Y$ is the mean of D_i , and σ_D^2 is the variance of D_i. Therefore, the significance of the difference between μ_X and μ_Y can be tested using paired t-test by testing the following hypothesis:

$$H_0: \mu_D = 0 \tag{6}$$

against the alternative hypothesis: $H_1: \mu_D \neq 0$.
The formula for the paired t-test is given by [4, 6, 7]:

$$T = \frac{\overline{D}-0}{\frac{S_D}{\sqrt{n}}} \sim t_{(n-1)} \tag{7}$$

where $\overline{D}$ and S_D are, respectively, the mean and the standard deviation of D_i in a matched sample. Notice that the statistical t-test represents the one sample t-test applied on the difference between two dependent variables D.

2.2 Wilcoxon –matched pairs signed-ranks (W.M)

The Wilcoxon –matched pairs signed-ranks is a popular, nonparametric test that is often an alternative test to t-test for matched Pairs, but it is used when the data violate the assumption of a bivariate normal distribution or for ordinal data [8]. This test is an extension of the Wilcoxon signed-rank test, proposed by Frank Wilcoxon in 1945. In fact, this test requires that paired samples should be random and independent. It is used to compare the means of two dependent samples and it is very appropriate for a repeated measure design where the same subjects are evaluated under two different conditions. The W.M is used to test whether the matched random sample is drawn from a population in which the median of the differences is equal to a specific value; in other words, it tests the following two-sided null hypothesis:

$$\mathrm{H_0}: \theta_\mathrm{D} = 0 \tag{8}$$

against the alternative hypothesis: $\mathrm{H_1}: \theta_\mathrm{D} \neq 0$
where m is the median of the differences (D_i) between the two populations.
The W.M-test can be carried out using the following steps:

1. Compute the difference scores D_i, (i=1, 2, …, n) for each pair of data.
2. Rank the absolute value of difference scores $|D_i|$, from 1 through n. If two or more difference scores are the same, the mean of the ranks of these scores is given to each of the tied ranks.
3. When $D_i = 0$, the pair is not assigned a rank, with reducing n by the number of cases in which the difference score = 0.
4. Calculate the sum of the ranks of each of the positive signs (R^+) and negative signs (R^-), as follows:

$R^+ = \sum_{\forall D_i>0} \text{sign}\,(D_i)\text{Rank}|D_i|$, $R^- = \sum_{\forall D_i<0} \text{sign}\,(D_i)\text{Rank}|D_i|$

Notice that,

$$R^+ + R^- = \frac{n(n+1)}{2}$$

5. The test statistic, say W is given by:

$$W = \min(R^+, R^-) \tag{9}$$

6. Compare the test statistic W with the critical value W^* at a specific significant level, then reject H_0 if: $W \leq T$.

If the sample size is relatively large, the normal approximation of the W.M statistic can be used for testing the null hypothesis (7), by using the following test statistics:

$$z = \frac{W - \frac{n^*(n^*+1)}{4}}{\sqrt{\frac{n^*(n^*+1)(2n^*+1)}{24}}} \tag{10}$$

n^*: the number of difference scores with non-zero rank.
W: the calculated value of W.M statistic defined in (9). If we use the continuity coefficient, the test statistic becomes:

$$Z = \frac{\left| W - \frac{n^*(n^*+1)}{4} \right| - 0.5}{\sqrt{\frac{n^*(n^*+1)(2n^*+1)}{24}}} \tag{11}$$

When a repeating state appears in the different observations, it is appropriate to use the following statistics:

$$z = \frac{W - \frac{n^*(n^*+1)}{4}}{\sqrt{\frac{n^*(n^*+1)(2n^*+1)}{24} - \frac{\sum t_i^3 - \sum t_i}{48}}} \quad (12)$$

For all cases, the null hypothesis will be rejected if $z \geq z^*$, where z^* represents the tabled critical value of the test at a specific level of significance.

3. Contaminated bivariate normal distribution

The contaminated bivariate normal distribution is a simple but useful distribution that can be used to simulate outliers. It was originally studied by John Tukey in the 1990s and 1950s. In order to study the robustness of the test's statistics against the departure of normality assumption, the bivariate normal distribution has been contaminated by outliers. The latter process was done by generating the random sample from the original distribution, denoted by F, with a specific proportion, say λ, and allowing a few of these sample observations to be coming from other distributions $G_1, G_2, \cdots, G_k$ that differ in their parameters from the original distribution. These observations are known as (Contaminated). Usually, they can be expressed as follows:

$(1 - \lambda_1 - \lambda_2 - \cdots - \lambda_k)F + \lambda_1 G_1 + \cdots + \lambda_k G_k$

where

λ_i: contamination rate by the distribution G_i where $i = 1, \cdots, k$

There are two types of contaminants: the first type is known as symmetric contaminant. The symmetric contaminant is obtained when generating a symmetric contaminated distribution G, around the original distribution center F, equal in the μ of both distributions and difference in σ^2 to make the variance of G bigger than the variance of F. If both of the distributions G and F are normal distribution, where

$F: N\sim(\mu, \sigma^2), G: N\sim(\mu, \sigma^2 b) \quad , b > 1,$

the continuous random variable X resulting from the mixture distribution will have a symmetric contaminated normal distribution in the rate of λ , i.e. $X\sim(1-\lambda)F + \lambda G$.

The other type is known as the asymmetric contamination. It is obtained when generating the contaminated distribution G_2 symmetrically about any point within the distribution F, if the center is not equal. That is, when G_2 is equal to the distribution F in variance and different from it in location, i.e.: $G_2\sim N(\mu + a, \sigma^2) \quad , \quad a > 0,$

in this case, the distribution of the random variable X can be expressed as follows:

$X\sim(1-\lambda)F + \lambda G_2$

4. Bivariate exponential distribution

There are several formulas for the bivariate exponential distributions. The Downton's bivariate exponential distribution is the most important of these distributions, which has the density: [9]

$$f_{x,y}(x,y) = \begin{bmatrix} \frac{\mu_x \mu_y}{1-\rho} exp\left\{-\frac{\mu_x x + \mu_y y}{1-\rho}\right\} \sum_{n=0}^{\infty} \left\{\frac{\rho \mu_x \mu_y xy}{(1-\rho)^2}\right\}^n \frac{1}{(n!)^2} & , \ x, y > 0 \\ 0 & , \quad o.w \end{bmatrix} \quad (13)$$

where $\mu_X, \mu_Y \in +R$ and $\rho \in (-1,1)$, with

$E(x) = \frac{1}{\mu_x} \quad , \text{var}(x) = \frac{1}{{\mu_x}^2}$

$E(y) = \frac{1}{\mu_y} \quad , \text{var}(y) = \frac{1}{{\mu_y}^2}$

5. The Jackknife resampling technique

In statistics, the Jackknife is a resampling technique used when it is not viable to evaluate data with the parametric methods and is particularly well-suited for complex and non-parametric designs applications. The Jackknife method was first introduced by Maurice Henry Quinn in 1949 to reduce statistical biases. In 1958, John Wilder Tuckey expanded its use to include variance estimation. Quenouille used the Jackknife method to correct and estimate the bias of estimation its application, based on the deletion of some of the original observations of the sample. Next, Tukey used the method to create confidence intervals for data that have large variations of an estimator. It is similar to the bootstrap method, but with no replacement [10]. The Jackknife statistic (JT) is given by:

$$JT = \frac{\overline{D}_{jack}}{\frac{\hat{\sigma}_D}{\sqrt{n}}}$$

This method is computer-based for estimating biased standard error with the least possible bias. it is developed to minimize the sampling error and obtaining narrow confidence intervals in estimating population parameters. It is also called a pocket - knife method and is a hand tool that is to use on various problems.

In general, Jackknife samples can be obtained by accepting a random sample of size n of observations and through it, the estimate is calculated. The mean of these calculations is then found by deleting one observation at a time without returning, to make the number of statistics computed n of times. When the application is achieved, the Jackknife estimate appears by aggregating each estimate in the sample [11].

The method of Jackknife application has the following steps:

1- By using the Jackknife method, $\overline{D} = \frac{\sum_{j=1}^{n} D_j}{n}$

2- Estimation of the variance by the Jackknife resampling technique, as follows:

$$\hat{\sigma}_D^2 = \frac{\sum_{j=1}^{n} (Dj - \overline{D})^2}{n(n-1)}$$

3- Application of the paired t-test for data under the Jackknife resampling technique:

$JT = \frac{\overline{D}_{jack}}{\frac{\hat{\sigma}_D}{\sqrt{n}}}$.

Similarly, we are jackknifing the W.M test and the approximation of W.M to normal distribution, respectively, as follows:

$$\text{JWS} = \frac{\sum_{\text{i=1}}^{\text{n}} \text{WS}_j}{\text{n}}$$

where WS_j is j[th] Wilcoxon–matched pairs signed-ranks for small samples:

$$\text{JWL} = \frac{\sum_{\text{i=1}}^{\text{n}} \text{WL}_j}{\text{n}}$$

where WL_j is j[th] Wilcoxon–matched pairs signed-ranks for large samples.

6. Simulation Study

1. A Monte-Carlo simulation study was conducted to examine and compare the behavior of different test statistics represented by Paired t-test (T), Paired t-test using Jackknife resampling (JT), W.M-test for small sample sizes ($n \leq 30$) (WS), W.M-test (WL) when $n > 30$, jackknifing W.M-test for small samples (JWS), and jackknifing W.M-test when sample size $n > 30$ (JWL). The distribution of matched pairs was generated from the following joint PDFs: bivariate normal distribution, contaminated bivariate normal distribution, and bivariate exponential distribution.

Different sample sizes (n = 10, 20, 30, 50, 100) generated to represent small, moderate and large sample sizes with different values if correlation coefficient $\rho = 0, 0.4, 0.8$. The experiment was replicated (10000) times.

Based on the Bradley's liberal criterion, the test will be regarded robust if it's Type I error rate $\hat{\alpha}$ falls within the interval: $0.9\,\alpha < \hat{\alpha} < 1.1\,\alpha$.

In this paper, we use nominal $\alpha = 0.05$. Therefore, Bradley's liberal criterion is $0.045 < \hat{\alpha} < 0.055$.

According to the Salter and Fawcett criterion, the test will be regarded as robust if it's Type I error rate $\hat{\alpha}$ satisfies: $0.05 \pm 2\sqrt{0.05\,(1-0.05)/10000}$,

i.e., the test is robust if $\hat{\alpha}$ falls within the interval 0.0456 – 0.0543. Notice that, in this article, the two criterions of robustness are quite closed, Bradley's liberal criterion will be used because it is more popular.

The setting values of experiments of simulation experiments can be summarizing by the following table:

Table 1-The Algorithm of Simulation Experiments

n	ρ	Bivariate Normal distribution		Contaminated Bivariate Normal distribution		Bivariate Exponential distribution	
		$X \sim N(\mu, \sigma$	$Y \sim N(\mu, \sigma^2$	$X \sim N(\mu, \sigma^2)$	$Y \sim N(\mu, \sigma^2$	$X \sim EXP(\lambda)$	$Y \sim EXP(\lambda$
10, 20, 30, 50, 100	0, 0.4, 0.8	$X \sim N(1,1$	$Y \sim N(1,1)$	$80\% X \sim N(1,1) + 20\% X \sim N(1,25)$	$Y \sim N(1,1)$	$X \sim EXP(1)$	$Y \sim EXP(1$
		$X \sim N(1.5,$	$Y \sim N(1,1)$	$80\% X \sim N(1.5,1) + 20\% X \sim N(1.5,2$	$Y \sim N(1,1)$		
		$X \sim N(1,1$	$Y \sim N(1,25$	$80\% X \sim N(1,1) + 20\% X \sim N(1,25)$	$Y \sim N(1,25$	$X \sim EXP(1/1.5)$	$Y \sim EXP(1$
		$X \sim N(1.5,$	$Y \sim N(1,25$	$80\% X \sim N(1.5,1) + 20\% X \sim N(1.5,2$	$Y \sim N(1,25$		

7. Simulation Results

To examine and compare the behaviors of test statistics under different cases, the simulation experiment's results, represented by Type I error rates and power rates, are summarized in Tables- 2 to 11. In this paper, the behavior of different tests is discussed briefly, according to the distribution of the population that the matched sample is drown from, as follows.

1. Bivariate normal distribution with equality of variances

i) Type I Error Rates

The Type I error rates for different tests at (α = 0.05), applied on matched data from a bivariate normal distribution, are tabulated in the Table-2. Generally, it can be seen that all of the type 1 error rates of the test statistics were good and within Bradley's liberal criterion (0.045-0.055), except WS and JWL tests, when the sample size n $\leq$ 10 for all the different values of ρ. The JWS test performance was not good because the value of the Type I error rate is outside Bradley's liberal criterion for all cases with different values of ρ. It is worth mentioning here that the T and JT tests are equaled when n=100.

Table 2- Type 1 error rates on different test statistics with bivariate normal distribution, $X \sim N(1,1)$ and $Y \sim N(1,1)$

ρ	n	T	WS	WL	JT	JWS	JWL
0.0	10	*0.0496	0.0369	*0.0495	*0.0526	0.0628	0.0388
	20	*0.0518	*0.0456	*0.0489	*0.0522	0.0758	*0.0509
	30	*0.0559	*0.0518	*0.0547	*0.0559	0.0792	*0.0504
	50	*0.0523	-	*0.0504	*0.0524	-	*0.0501
	100	*0.0518	-	*0.0536	*0.0518	-	*0.0518
0.4	10	*0.0508	0.0380	*0.0505	*0.0530	0.0683	0.0388
	20	*0.0536	*0.0484	*0.0525	*0.0541	0.0768	*0.0534
	30	*0.0550	*0.0515	*0.0538	*0.0554	0.0805	*0.0512
	50	*0.0504	-	*0.0499	*0.0504	-	*0.0479
	100	*0.0525	-	*0.0505	*0.0525	-	*0.0503
0.8	10	*0.0517	0.0383	*0.0516	*0.0548	0.0668	0.0402
	20	*0.0528	*0.0477	*0.0517	*0.0532	0.0805	*0.0514
	30	*0.0546	*0.0510	*0.0538	*0.0551	0.0783	*0.0522
	50	*0.0504	-	*0.0489	*0.0504	-	*0.0477
	100	*0.0543	-	*0.0531	*0.0543	-	*0.0530

***: Type 1 error rate is within the Bradley's liberal criterion (0.045-0.055).**

ii) Power rates

The power rates for different tests at (α = 0.05), applied on samples from a Bivariate normal distribution, are summarized in the Table-3.

Generally, it can be seen that the JT-test and T-test are the most powerful, with greater power rates for all sample sizes with the different values of ρ. It is clear that, with increasing sample size and the correlation coefficient, the power rates for all tests are increasing and converged to 1, which corresponds to the central limit theory.

Table 3-Power rates on different test statistics with bivariate normal distribution, $X \sim N(1,1)$ and $Y \sim N(1,1)$

ρ	n	T	WS	WL	JT	JWS	JWL
0.0	10	*0.1716	0.1375	*0.1680	*0.1766	0.2027	0.1303
	20	*0.3244	*0.2955	*0.3105	*0.3256	0.3889	*0.3007
	30	*0.4662	*0.4432	*0.4526	*0.4671	0.5308	*0.4284
	50	*0.6881	-	*0.6665	*0.6882	-	*0.6560
	100	*0.9399	-	*0.9295	*0.9399	-	*0.9279
0.4	10	*0.2567	0.2061	*0.2458	*0.2634	0.2901	0.1905
	20	*0.4953	*0.4581	*0.4748	*0.4973	0.5623	*0.4568
	30	*0.6769	*0.6513	*0.6580	*0.6777	0.7268	*0.6381
	50	*0.8873	-	*0.8707	*0.8875	-	*0.8638
	100	*0.9940	-	*0.9923	*0.9940	-	*0.9914
0.8	10	*0.6035	0.5305	*0.5835	*0.6130	0.6421	0.4811
	20	*0.9203	*0.9027	*0.9099	*0.9211	0.9374	0.8949
	30	*0.9857	*0.9805	*0.9819	*0.9859	0.9892	0.9796
	50	*0.9998	-	*0.9998	*0.9998	-	0.9999
	100	*1.0000	-	*1.0000	*1.0000	-	1.0000

2. Contaminated bivariate normal distribution with equality of variances

To study the influence of departures from normality on four test statistics, the tests were applied on different paired samples generated form the contaminated bivariate normal distribution, represented by: $80\%X \sim N(1,1) + 20\%X \sim N(1,25)$ and $Y \sim N(1,25)$.

i) Type 1 Error Rates

Results of Type 1 error rates on different test statistics at 0.05 level of significance with contaminated data by outliers are summarized in Table-4 and show that:

The T statistic is extremely sensitive (not robust) to the contaminated data when n $\leq$ 30 with different values of ρ, which means that it is not robust against the departure from normality assumption. However, the JT-test improves the test robustness of the paired t-test but still not robust in all cases, except with n ≤ 20 when ρ=0, n ≤ 30 when ρ=0, 4, and n ≤ 50 when ρ=0, 8. In general, the most robust tests are WL with all cases followed by JWL for large sample sizes with all cases, except one case when n=10.

Table 4-Type 1 error rates on different test statistics with contaminated bivariate normal distribution, $X \sim N(1,1)$ and $Y \sim N(1,1)$

ρ	n	T	WS	WL	JT	JWS	JWL
0.0	10	0.0334	0.0384	*0.0503	0.0354	0.0641	0.0386
	20	0.0401	0.0429	*0.0470	0.0402	0.0760	*0.0534
	30	*0.0464	*0.0494	*0.0519	*0.0466	0.0788	*0.0504
	50	*0.0468	-	*0.0500	*0.0468	-	*0.0503
	100	*0.0481	-	*0.0513	*0.0481	-	*0.0515
0.4	10	0.0270	0.0377	*0.0506	0.0295	0.0670	0.0396
	20	0.0368	0.0443	*0.0490	0.0372	0.0760	*0.0527

	30	0.0442	*0.0489	*0.0513	0.0443	0.0802	*0.0497
	50	*0.0460	-	*0.0493	*0.0458	-	*0.0497
	100	*0.0470	-	*0.0504	*0.0470	-	*0.0522
0.8	10	0.0154	0.0370	*0.0500	0.0163	0.0634	0.0384
	20	0.0267	*0.0446	*0.0491	0.0271	0.0754	*0.0512
	30	0.0384	*0.0484	*0.0512	0.0382	0.0797	*0.0507
	50	*0.0448	-	*0.0516	*0.0447	-	*0.0502
	100	*0.0458	-	*0.0509	*0.0457	-	*0.0511

***: Type 1 error rate is within the Bradley's liberal criterion (0.045-0.055).**

ii) Power rates

The power rates for different tests applied on samples from a contaminated bivariate normal distribution are tabulated in Table-5. We observed the following important points:

- In general, when n ≥ 20, JWL is more powerful test for all cases compared with the other tests, while WL is the best test when n=10. The JT-test achieved the best power rate when n ≥ 30, followed by the t-test, which means that it has a highest type I error. Finally, it is clear that all power rates of tests are increasing with the increase of sample size and ρ and they converge to each other.

Table 5-Power rates on different test statistics with contaminated bivariate normal distribution, $X \sim N(1,1)$ and $Y \sim N(1,1)$

ρ	n	T	WS	WL	JT	JWS	JWL
0.0	10	0.0882	0.0970	*0.1172	0.0925	0.1451	0.1089
	20	0.1487	0.1891	*0.1994	0.1501	0.2624	*0.2126
	30	*0.2036	*0.2876	*0.2952	*0.2049	0.3623	*0.2898
	50	*0.2850	-	*0.4356	*0.2855	-	*0.4509
	100	*0.4880	-	*0.7499	*0.4882	-	*0.7549
0.4	10	0.1058	0.1296	*0.1544	0.1134	0.1863	0.1501
	20	0.1861	0.2698	*0.2844	0.1880	0.3562	*0.3065
	30	0.2494	*0.4062	*0.4149	0.2506	0.4911	*0.4228
	50	*0.3456	-	*0.6169	*0.3464	-	*0.6345
	100	*0.5848	-	*0.9068	*0.5853	-	*0.9116
0.8	10	0.1622	0.2500	*0.2790	0.1747	0.3232	0.3279
	20	0.2622	*0.5515	*0.5693	0.2666	0.6498	*0.6225
	30	0.3340	*0.7553	*0.7636	0.3356	0.8265	*0.7903
	50	0.4543	-	*0.9454	*0.4557	-	*0.9532
	100	*0.7134	-	*0.9991	*0.7140	-	*0.9993

3. Bivariate exponential distribution

In this case, the paired samples were drown from the bivariate exponential distribution, i.e. $X \sim \exp(1)$ and $Y \sim \exp(1)$ in the case of estimating Type I error rate and $X \sim \exp(0.6667)$ and $Y \sim \exp(1)$ in the case of estimating power rate for different tests.

i) Type 1 error rates

Table-6 shows the Type 1 error rates for each test, under the Bivariate exponential distribution assumption.

We noticed that WL has the highest robustness in the different cases, because the Type I error rates are within Bradley's liberal criterion (0.045-0.055), followed by JWL, except for one case when n=10. The type 1 error rates for the WS test lie outside the expectable range, except for one case when n=30 for all values of ρ. It is clear that the T-test is insensitive to non-normality assumptions when the sample size n ≥ 30 with the different values of ρ. When the data violate the normality, the assumption revealed that the JWL was the most robust test for all sample sizes compared to other tests.

Table 6-Type 1 error rates on different test statistics with bivariate exponential distribution, $X \sim \exp(1)$ and $Y \sim \exp(1)$

ρ	n	T	WS	WL	JT	JWS	JWL
0.0	10	0.0424	0.0370	*0.0473	0.0434	0.0643	0.0381
	20	0.0436	0.0428	*0.0480	0.0439	0.0762	*0.0475
	30	*0.0466	*0.0457	*0.0474	*0.0467	0.0745	*0.0472
	50	*0.0526	-	*0.0502	*0.0527	-	*0.0504
	100	*0.0452	-	*0.0465	*0.0453	-	*0.0470
0.4	10	0.0399	0.0364	*0.0497	0.0420	0.0648	0.0396
	20	0.0430	0.0407	*0.0451	0.0432	0.0753	*0.0466
	30	*0.0490	*0.0506	*0.0524	*0.0490	0.0788	*0.0521
	50	*0.0492	-	*0.0484	*0.0494	-	*0.0476
	100	*0.0500	-	*0.0500	*0.0500	-	*0.0516
0.8	10	0.0419	0.0386	*0.0490	0.0431	0.0642	0.0402
	20	*0.0458	0.0438	*0.0481	*0.0461	0.0760	*0.0509
	30	*0.0485	*0.0456	*0.0479	*0.0485	0.0770	*0.0457
	50	*0.0453	-	*0.0482	*0.0454	-	*0.0480
	100	*0.0462	-	*0.0467	*0.0462	-	*0.0491

***: Type 1 error rate is within the Bradley's liberal criterion (0.045-0.055).**

Power rates

The results of the simulation study of the power rates are summarized in the Table-7.

Generally, it can be seen that the power rates of bivariate exponential distribution and bivariate normal distribution are higher than that of the contaminated bivariate normal distribution. The power rates for all tests are increasing with the increasing sample size and correlation coefficient; the value may approach 1 when n=50,100 and $\rho = 0.8$. It can be seen that WL, when $\rho=0$ and n = 10, 20, has the higher power rates compared to other tests.

Table 7-Power rates on different test statistics with bivariate exponential distribution, $X \sim \exp(1)$ and $Y \sim \exp(1)$

ρ	n	T	WS	WL	JT	JWS	JWL
0.0	10	0.1529	0.1406	*0.1711	0.1577	0.2080	0.1427
	20	0.2623	0.2947	*0.3074	0.2640	0.3798	*0.3026
	30	*0.3476	*0.4305	*0.4385	*0.3488	0.5186	*0.4199
	50	*0.5162	-	*0.6469	*0.5167	-	*0.6391
	100	*0.8116	-	*0.9177	*0.8119	-	*0.9155
0.4	10	0.1328	0.1242	*0.1511	0.1363	0.1855	0.1192
	20	0.3113	0.2856	*0.2986	0.3120	0.3718	*0.2829
	30	*0.4582	*0.4121	*0.4205	*0.4587	0.5068	*0.3994
	50	*0.7072	-	*0.6466	*0.7079	-	*0.6374
	100	*0.9513	-	*0.9151	*0.9513	-	*0.9161
0.8	10	0.3125	0.3045	*0.3519	0.3187	0.4053	0.2740
	20	*0.6913	0.6524	*0.6671	*0.6923	0.7415	*0.6438
	30	*0.8850	*0.8418	*0.8482	*0.8853	0.8917	*0.8305
	50	*0.9877	-	*0.9764	*0.9877	-	*0.9746
	100	*1.0000	-	*0.9999	*1.0000	-	0.9999

8. Conclusions

The Monte-Carlo simulation was employed to study the behavior of different test statistics that are used for comparing the equality of the means of two paired populations. Based on the Type one error and the power of the test, it is shown that:

- The presence of outliers leads to a decrease of the Type I error for the paired t-test statistics.

- The power rates for all the tests are increasing with the increase of sample size and the correlation coefficient.
- Generally, all Type I error rates and the power of the tests are convergent to each other with the increasing of sample sizes.
- In case of the existence of outliers, the jackknifing of the Wilcoxon signed rank test for large sample sizes is the most powerful compared to the others when $n \geq 20$, while the Wilcoxon signed rank test for large sample sizes is the best compared to the others when n = 10.
- When the paired data follow the exponential distribution, the Wilcoxon signed rank test for large sample sizes is the most powerful compared to the others when $\rho = 0$ with all sample sizes, in addition to the case of $n \leq 20$ with different values of ρ. The jackknifing paired t-test is the best compared with other tests when $n \geq 30$ and $\rho > 0$.

References

1. Huber, P.J. **1981**. *Robust Statistics*, John Wiley, New York.
2. Bradley, J.V. **1977**. A common situation conducive to bizarre distribution shapes. *The American Statistician*, **31**: 147-150.
3. Salter, K.C. and Fawcett, R.F. **1985**. A robust and powerful rank test of treatment effects in balanced incomplete block designs. *Communications in Statistics: Simulation and Computation*, **14**(4): 807-828.
4. Kim, H., Park, C. and Wang, M. **2018**. Paired t_test based on robustified statistics, Conference Paper.
5. Ganji, M., Bevrani, H. and Golzar, N., H. **2018**. A New Method for Generating Continuous Bivariate Distribution Families. *JIRSS*, **17**(1): 109-129, DOI: 10.29252/jirss.17.1.109
6. Park. c, Wang and yean Hwange. W. **2020**. " A study on robustness of the paired sample tests", *Industrial engineering and management systems*, **19**(2): 386-397.
7. Zimmerman, D. W. **1997**. A note on the interpretation of the paired samples, *Journal of Educational and Behavioral Statistics*. **22**(3): 349 – 360.
8. Sheskin, D.J. **2000**. *Handbook of parametric and nonparametric statistical procedures*. second ed., chapman& hall/cRc.
9. Al-Saadi, S.D., Young, D.H. **1980**. Estimators for the correlation coefficient in a bivariate exponential distribution. *Journal of Statistical Computation and Simulation*, **11**: 13– 20.
10. Zaman. T and Alakus. K.," comparison of resampling methods in multiple linear regression ", *the Journal of science and arts*, **1**(46): 91-104,2 Statistics, 5, 645, 2015.
11. Fradette. K, keselman . H.J. lix. L, Algina. J and wilcox. R. **2003**. "Conventional and Robust Paired and Independent-Samples t-tests: Type I Error and Power Rates", *journal of modern applied statistical methods*.