Ibn Al Haitham Journal for Pure and Applied Science

Journal homepage: http://jih.uobaghdad.edu.iq/index.php/j/index

Robust Tests for the Mean Difference in Paired Data by Using Bootstrap Resampling Technique

Ghufran A. Ghadhban **Huda A. Rasheed**

Department of Mathematics, College of Science, Mustansiriyah University, Baghdad, Iraq.

ghyfrang1994@gmail.com hudamath@uomustansiriyah.edu.iq

Article history: Received 28 September 2020, Accepted 8 November 2020 Published in July 2021.

Doi: 10.30526/34.3.2680

Abstract

The paired sample t-test for testing the difference between two means in paired data is not robust against the violation of the normality assumption. In this paper, some alternative robust tests have been suggested by using the bootstrap method in addition to combining the bootstrap method with the W.M test. Monte Carlo simulation experiments were employed to study the performance of the test statistics of each of these three tests depending on type one error rates and the power rates of the test statistics. The three tests have been applied on different sample sizes generated from three distributions represented by Bivariate normal distribution, Bivariate contaminated normal distribution, and the Bivariate Exponential distribution.

Keywords: Paired t-test, Robust, Bootstrap, Wilcoxon signed-rank test, Bivariate contaminated normal distribution, Bivariate Exponential.

1. Introduction

Comparing the 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 assumption. The robustness concept is introduced firstly by Box in 1953. There are many definitions of the concept of robustness, perhaps the most important one that was stipulated in the Huber (1981) definition that 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.[1]

Bradley (1978) defined what it means (Robust test), stipulates that the test is called robust against the violation of one or more of the test's assumptions if that violation does not 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 criterion of robustness and called it liberal criterion, that is the test could be regarded as robust only if its Type 1 error rate $\hat{\alpha}$ fall in the following interval:[2]

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

i.e.,

$$|\hat{\alpha} - \alpha| \leq \frac{\alpha}{10} \qquad (1)$$

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

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

Where R represents the replicated times.

This paper aims to study and investigate the effect 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 paper is to find a robust test that achieves the highest power of the test when the set of paired data violate the assumptions of the normality and the homogeneity of variances of the correlated variables. Therefore, a number of robust tests has been suggested represented by Wilcoxon–matched pairs signed-ranks using bootstrap (BWS), Wilcoxon–matched pairs signed-ranks when sample size n > 25 using bootstrap (BWL), also of bootstrapping the paired t-test (BT).

2. Test Statistics

2.1 Paired t-Test

The paired t-test is one of the most important tests employed to test the significance of the difference between the means of the two dependent variables, it is sometimes called the dependent sample t-test. Also, known as the repeated measurements, when we have them before and after the treatment.

Let the two-dimensional random variables (X, Y) have a bivariate normal distribution with parameters μ_X, μ_Y , σ_X , σ_Y and ρ if there joint pdf is defined as: [4, 5]

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

Where, $\mu_X, \mu_Y \in R$, σ_X , $\sigma_Y \in R^+$, $\rho \in (-1,1)$. And

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

The paired t-test aims to test the following null hypothesis:

$$H_0: \mu_X = \mu_Y \qquad (4)$$

Against the alternative hypothesis:

$$H_1: \mu_X \neq \mu_Y$$

Let

$D_i = X_i - Y_i$, i = 1, 2, …, n (5)

Then, D ~ Normal(μ_D, σ_D^2), where, $\mu_D = \mu_X - \mu_Y$ is the mean of the difference between two populations D and $\sigma_D^2 = \sigma_D^2 + \sigma_D^2 - cov\,(x, y)$ is the variance of D.

Therefore, the significance of the difference between μ_X and μ_Y can be tested using the paired t test by testing the following hypothesis

$H_0: \mu_D = 0$ (6)

Against alternative hypothesis:

$H_1: \mu_D \neq 0$

The paired t-test statistics is given by:[6, 7]

$t = \frac{\bar{D}-0}{\frac{S_D}{\sqrt{n}}} \sim t_{(n-1)}$ (7)

Where, $\bar{D}$ and S_D are presented respectively, the mean and the standard deviation of D_i in the matched sample.

Notice that the test statistics T represents the one sample t test applied on the difference between two dependent variables D.

2.2 Wilcoxon –Matched Pairs Signed-Ranks (W.M)

This test is an extension of the Wilcoxon signed-rank test, proposed by Frank Wilcoxon in 1945. It is widely used as an alternative test of the paired t-test where the paired data violate the normality assumption which inflates Type I error rate [8]. In fact, this test requires that paired samples should be random and independent. It is used to compare the means of two dependent samples or repeated measurements on a single sample in case of non-normality data. 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, to test the following two sided null hypothesis:

$H_0: \theta_D = 0$ (8)

Against alternative hypothesis: $H_1: \theta_D \neq 0$

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 difference scores D_i , (i=1, 2, …, n) for each pair of data.
2. Ranke 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, and reduces 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 statistics, say W is given by:

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

6. Compare the test statistics W with the critical value W^* at a specific significant level, then reject H_0 if: [8] $W \leq W^*$

If the sample size is relatively, large, the normal approximation of the W.M statistics can be used for testing the null hypothesis (7) by using the following test statistics[8]

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

n^*: Represents the number of difference scores non-zero rank.

W: Represents the calculated value of W.M statistic defined in (9). If we use the continuity coefficient, the test statistic become

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

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. **Bivariate Contaminated Normal Distribution** 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 specific proportion, say λ and allowing a few of these sample observations to become 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, it can be expressed as follows:

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

whereas,

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 contaminated 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 distributions G, 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 symmetric contaminated normal distribution in the rate of λ i.e., $X\sim(1-\lambda)F + \lambda G$

The other type is known 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 G_2 and F have the same variance but they are differ from location (i.e. $G_2\sim N(\mu + a, \sigma^2) \quad , a > 0$

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]

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

Where, $\mu_X, \mu_Y \in R^+$ and, $\rho \in (-1,1)$. With

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

5. **The Bootstrap Resampling Tchnique** Bootstrap is the most popular resampling technique used in statistical analysis. It was first developed and introduced by Ephron in 1979. It is a computer-based resampling technique developed to make statistical inferences simpler [10] and has the potential to be used for precision-based data simulation problems for statistical reasoning. According to Ephron, the boot-up process differs from statistical inference, as the method is very simple and based on re-sampling procedures. The bootstrap statistics (BT) is given by

$$\text{BT} = \frac{\bar{\bar{x}}}{\frac{vb}{\sqrt{n}}} \qquad (14)$$

where, $\bar{\bar{x}} = \frac{\sum_{i=1}^{b} \bar{x}_i}{b}$ and $vb = \frac{\sum(\bar{x}_i - \bar{\bar{x}})^2}{b-1}$.

In this paper, we used the following procedure, which gave better results

$$\text{BT} = \frac{\sum_{\text{j}=1}^{\text{b}} \text{T}_j}{\text{b}}$$

Where, T_j (j = 1, 2, …, b) is the paired t-test statistics applied on D_{ij}^* (i = 1, 2, …n), Where, D_{ij}^* represents the difference of the i[th] resampling variables in the j[th] bootstrap resampling.

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

$$\text{BWS} = \frac{\sum_{\text{j}=1}^{\text{b}} \text{WS}_j}{\text{b}}$$

Where WS_j is j[th] Wilcoxon –matched pairs signed-ranks for small sampls

$$\text{BWL} = \frac{\sum_{\text{j}=1}^{\text{b}} \text{WL}_j}{\text{b}}$$

Where WL_j is j[th] Wilcoxon –matched pairs signed-ranks for large sampls

6. **Simulation Study**

A Monte-Carlo simulation study is conducted to examine and compare the behavior of different test statistics represented by Paired t-test (T), Paired t-test uses Bootstrap resampling (BT), W.M test for small sample sizes ($n \leq 30$) (WS), W.M test (WL) when $n > 30$, bootstrapping W.M test for small samples (BWS) and bootstrapping W.M test when sample size $n > 30$ (BWL). The distribution of matched pairs has been generated from the following joint pdf's:

1. Bivariate normal distribution.
2. Bivariate contaminated normal distribution.
3. Bivariate exponential.

Different sample sizes (n = 10, 20, 30, 50, 100) have been 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 Bradley's liberal criterion, the test will be regarded robust if it's Type I error rate $\hat{\alpha}$ fall 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 [3], the test will be regarded as robust if it isType I error rate $\hat{\alpha}$ satisfies

$$0.05 \pm 2\sqrt{0.05\,(1-0.05)/10000}$$

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

The Algorithm of Simulation Experiments can be summaried in the following table:

Table 1. The Algorithm of Simulation Experiments

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

7. Simulation Results

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

1. Bivariate normal distribution with equality of variances

i) Type I Error Rates

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) and it shows that:

- It can be seen that the value of Type 1 error rate of the T , BWS and WL tests for all cases are within Bradley's liberal criterion (0.045-0.055)
- The performance of BT test is not good because the value of Type I error rate is outside the Bradley's liberal criterion for all cases except one case when $n = 100$ with different values of ρ.
- Generally, it can be seen that all of Type 1 error rates of the test statistics do well except that of the BWL when $n \leq 20$ and WS when $n \leq 10$ for all the different values of ρ.

Table 2. Type 1 error rates on different test statistics at α = 0.05 distributed according to the ρ and n with Bivariate normal distribution, $X \sim N(1,1)$ and $Y \sim N(1,1)$

ρ	n	T	WS	WL	BT	BWS	BWL
0	10	*0.0496	0.0369	*0.0495	0.1059	0.0506*	0.0589
	20	*0.0518	*0.0456	*0.0489	0.0744	0.0537*	0.0577
	30	*0.0559	*0.0518	*0.0547	0.0614	0.0489*	0.0500*
	50	*0.0523	-	*0.0504	0.0578	-	0.0522*
	100	*0.0518	-	*0.0536	0.0556	-	0.0550*
0.4	10	*0.0508	0.0380	*0.0505	0.1080	0.0484*	0.0569
	20	*0.0536	*0.0484	*0.0525	0.0752	0.0520*	0.0565
	30	*0.0550	*0.0515	*0.0538	0.0633	0.0499*	0.0510*
	50	*0.0504	-	*0.0499	0.0596	-	0.0524*
	100	*0.0525	-	*0.0505	0.0550*	-	0.0556
0.8	10	*0.0517	0.0383	*0.0516	0.1052	0.0517*	0.0601
	20	*0.0528	*0.0477	*0.0517	0.0743	0.0527*	0.0565
	30	*0.0546	*0.0510	*0.0538	0.0644	0.0517*	0.0528*
	50	*0.0504	-	*0.0489	0.0598	-	0.0529*
	100	*0.0543	-	*0.0531	0.0548*	-	0.0546*

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

ii) **Power rates**

The power rates of different tests at (α = 0.05) applied on samples taken from a Bivariate normal distribution have summarized in the **Table (3)** and show that:

- The T test is the most powerful when compared to other tests followed by the BWL test which can be used with large sample sizes (n ≥30).
- It is clear that, with the increasing sample size, the power rates for all tests are increasing and converged to 1, which corresponds to the central limit theory.
- The power rates are increasing with the increase in the correlation coefficient.
- It can be observed that the power rates for the BT test when the sample size (n=100) for all the different values of ρ are greater than power rates for the T test.

Table 3. The 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	BT	BWS	BWL
0	10	*0.1716	0.1375	*0.1680	0.2561	0.1535*	0.1705
	20	*0.3244	*0.2955	*0.3105	0.3752	0.3023*	0.3143
	30	*0.4662	*0.4432	*0.4526	0.4943	0.4412*	0.4443*
	50	*0.6881	-	*0.6665	0.7005	-	0.6633*
	100	*0.9399	-	*0.9295	0.9415*	-	0.9304*
0.4	10	*0.2567	0.2061	*0.2458	0.3539	0.2220*	0.2429
	20	*0.4953	*0.4581	*0.4748	0.5382	0.4568*	0.4689
	30	*0.6769	*0.6513	*0.6580	0.7024	0.6486*	0.6516*
	50	*0.8873	-	*0.8707	0.8914	-	0.8672*
	100	*0.9940	-	*0.9923	0.9947*	-	0.9923*
0.8	10	*0.6035	0.5305	*0.5835	0.6980	0.5369*	0.5713
	20	*0.9203	*0.9027	*0.9099	0.9229	0.8850*	0.8907
	30	*0.9857	*0.9805	*0.9819	0.9891	0.9830*	0.9838*
	50	*0.9998	-	*0.9998	0.9997	-	0.9995*
	100	*1.0000	-	*1.0000	1.0000*	-	1.0000*

2. Bivariate contaminated normal distribution with equality of variances

To study the influence of departures from normality on four test statistics, the tests have been applied on different paired samples that generated form the Bivariate contaminated 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 paired t-test statistics is extremely sensitive (not robust) to the contaminated data when $n \leq 30$ with different values of ρ which means it is not robust against the departure from normality assumption.
- The most robust tests are BWL and WL with all cases followed by BWS for small sample sizes.
- The BT test improves the robustness of the paired t-test because it robust in all cases except with n = 10 when $\rho = 0, 0.4$.
- All test statistics are robust against the normality assumption when $n \geq 50$.

Table 4. Type 1 Error Rates on different test statistics with Bivariate Contaminated Normal distribution, $X \sim N(1,1)$ and $Y \sim N(1,1)$

ρ	n	T	WS	WL	BT	BWS	BWL
0	10	0.0334	0.0384	*0.0503	0.0827	0.0488*	0.0543*
	20	0.0401	0.0429	*0.0470	0.0542*	0.0529*	0.0542*
	30	*0.0464	*0.0494	*0.0519	0.0543*	0.0517*	0.0526*
	50	*0.0468	-	*0.0500	0.0543*	-	0.0523*
	100	*0.0481	-	*0.0513	0.0466*	-	0.0519*
0.4	10	0.0270	0.0377	*0.0506	0.0744	0.0494*	0.0529*
	20	0.0368	0.0443	*0.0490	0.0537*	0.0522*	0.0535*
	30	0.0442	*0.0489	*0.0513	0.0508*	0.0524*	0.0532*
	50	*0.0460	-	*0.0493	0.0542*	-	0.0537*
	100	*0.0470	-	*0.0504	0.0471*	-	0.0541*
0.8	10	0.0154	0.0370	*0.0500	0.0539*	0.0494*	0.0542*
	20	0.0267	*0.0446	*0.0491	0.0.428	0.0502*	0.0535*
	30	0.0384	*0.0484	*0.0512	0.0447*	0.0526*	0.0540*
	50	*0.0448	-	*0.0516	0.0498*	-	0.0540*
	100	*0.0458	-	*0.0509	0.0470*	-	0.0542*

i) Power rates

The power rates of different tests applied on samples from a Bivariate contaminated normal distribution have been tabulated in table (5) and we have observed the following important points:

- The BWL test achieved the highest power rate, which means it has less Type II error (β) in comparison to other tests, followed by the BWS test which can be used with small sample sizes ($n \leq 30$).
- It is obvious that the power rates are increasing with the increase of the correlation coefficient.
- It is clear that, with the increasing sample size, the power rates of all tests are increasing and converged to each other.
- In general, for all test statistics, it is clear that all power rates of tests in the Bivariate contaminated normal distribution are lower than Bivariate normal distribution.

- Finally, it is clear that all power rates of tests are increasing with the increasing of Type I error because both of them represent the number of rejecting H_0.

Table 5. Power rates on different test statistics with Bivariate Contaminated Normal Distribution, $X \sim N(1,1)$ and $Y \sim N(1,1)$

ρ	N	T	WS	WL	BT	BWS	BWL
0	10	0.0882	0.0970	*0.1172	0.193	0.1279*	0.1427*
	20	0.1487	0.1891	*0.1994	0.2111*	0.2141*	0.2214*
	30	*0.2036	*0.2876	*0.2952	0.2531	0.3046	0.3078
	50	*0.2850	-	*0.4356	0.3293*	-	0.4627*
	100	*0.4880	-	*0.7499	0.5247*	-	0.7586*
0.4	10	0.1058	0.1296	*0.1544	0.2456	0.1757*	0.1933*
	20	0.1861	0.2698	*0.2844	0.2672*	0.3026*	0.3120*
	30	0.2494	*0.4062	*0.4149	0.3216	0.4412	0.4441
	50	*0.3456	-	*0.6169	0.4143*	-	0.6462*
	100	*0.5848	-	*0.9068	0.6277*	-	0.9145*
0.8	10	0.1622	0.2500	*0.2790	0.4353*	0.3629*	0.3863*
	20	0.2622	*0.5515	*0.5693	0.4175	0.6236*	0.6358*
	30	0.3340	*0.7553	*0.7636	0.4557*	0.8165*	0.8185*
	50	0.4543	-	*0.9454	0.5514*	-	0.9585*
	100	*0.7134	-	*0.9991	0.7678*	-	0.9997*

3. Bivariate normal distribution without equality of variances

In this case the variances X and Y have been assumed unequal, where, $X \sim N(1,1)$ and $Y \sim N(1,25)$, when estimating Type I error, and
$X \sim N(1,1.5)$ and $Y \sim N(1,25)$, in case of estimating the power rate.

i) Type 1 error rates

In this case, the result in **Table (6)** show that:
When σ_Y^2 increases ($\sigma_Y^2 = 25$) and differs from σ_X^2 ($\sigma_X^2 = 1$), Type I error rates are very little different from Type I error rates when $\sigma_Y^2 = 1$, (see table 2 and its discussion)

Table 6. Type I error rates on different test statistics with Bivariate Normal distribution, $X \sim N(1,\ 1)$ and $Y \sim N(1,\ 25)$

ρ	n	T	WS	WL	BT	BWS	BWL
0	10	0.051	0.0371	0.0496	0.1059	0.0507*	0.0587
	20	0.0535	0.046	0.0508	0.0736	0.0529*	0.0568
	30	0.0538	0.05	0.0532	0.0647	0.0511*	0.0524*
	50	0.0516	-	0.0494	0.0593	-	0.0543*
	100	0.0546	-	0.0543	0.0565	-	0.0536*
0.4	10	0.0499	0.0374	0.0483	0.1025	0.0500*	0.0593
	20	0.0522	0.0476	0.0505	0.0762	0.0502*	0.0550
	30	0.051	0.0477	0.05	0.0661	0.0549*	0.0564
	50	0.0494	-	0.0474	0.0635	-	0.0539*
	100	0.0511	-	0.0511	0.0571	-	0.0547*
0.8	10	0.0513	0.0374	0.0503	0.1058	0.0487*	0.0585
	20	0.0499	0.0435	0.0469	0.0699	0.0473*	0.0513*
	30	0.0492	0.0445	0.0477	0.0648	0.0526*	0.0532*
	50	0.049	-	0.0488	0.0608	-	0.0549*
	100	0.048	-	0.047	0.0570	-	0.0555

ii) Power rates

From **Table (7),** results for the power rates show that:

- Generally, we can say that the power tests are decreased when the variance of Y increases when compared to the results with the corresponding results of the equality case of variance
- It is clear that, the power rates are increasing with the increasing of the correlation coefficient.

  It can be seen that with the increasing of the sample size, the power rates for all tests are increasing and converged to each other.
- Finally, for all test statistics, all power rates of tests are increasing with the increasing Type I error.

Table 7. Power rates on different test statistics with Bivariate Normal distribution, $X \sim N(1,\ 1)$ and $Y \sim N(1,\ 25)$

ρ	n	T	WS	WL	BT	BWS	BWL
0	10	0.0618	0.0451	0.0574	0.1194	0.0586*	0.0687
	20	0.0758	0.0671	0.0732	0.0969	0.0692*	0.0737
	30	0.0869	0.0814	0.0841	0.1022	0.0840*	0.0853*
	50	0.1036	-	0.1013	0.1144	-	0.1033*
	100	0.1611	-	0.1568	0.168	-	0.1592*
0.4	10	0.0621	0.0447	0.0604	0.1186	0.0605*	0.0703
	20	0.0795	0.0709	0.0757	0.0996	0.0704*	0.0755
	30	0.0902	0.086	0.0891	0.1087	0.0916*	0.0929
	50	0.113	-	0.1089	0.1291	-	0.1164*
	100	0.1819	-	0.1752	0.1909	-	0.1792*
0.8	10	0.0636	0.0476	0.0633	0.1203	0.0616*	0.0723
	20	0.0816	0.0749	0.0807	0.1079	0.0778*	0.0831*
	30	0.0966	0.091	0.0947	0.1164	0.0954*	0.0971*
	50	0.1301	-	0.1252	0.1417	-	0.1275*
	100	0.2161	-	0.2067	0.2171	-	0.2051

4. Bivariate contaminated normal distribution without equality of variances

In this case, different tests have been applied on the paired data from the Bivariate contaminated normal distribution with assuming the variances X and Y which have been assumed unequal

i) **Type 1 error rates**

Table (8) includes the results of Type I error rates in case of Bivariate contaminated normal distribution, with the following

$80\%X\sim N(1,1) + 20\%X\sim N(1,25)$ and $Y\sim N(1,25)$.

The important points of the results can be summarized as follows

- The results of the non-parametric tests (BWS, BWL, WS, WL) nearly are the same as the results of
- Bivariate contaminated normal distribution, with the equal variances, assumed (homogeneity of variances, see **Table 4).**
- In general, Type I error rates of T test become better than the corresponding values of Bivariate contaminated normal distribution, with the equal variances assumes (homogeneity of variances, see **Table 4**), due to 20% of the matched samples contaminated by paired data that have the same variances of the two correlated variables, i.e. $\sigma_X^2 = \sigma_Y^2 = 25$

Table 8. Type I error rates on different test statistics with Bivariate contaminated normal distribution, $X \sim N(1,\ 1)$ and $X \sim N(1,\ 25)$

ρ	n	T	WS	WL	BT	BWS	BWL
0	10	0.0515	0.0395	0.0526	0.1040	0.0520*	0.0593
	20	0.051	0.0462	0.0502	0.0730	0.0505*	0.0540*
	30	0.0542	0.0547	0.0571	0.0620	0.0509*	0.0522*
	50	0.0517	-	0.0503	0.0637	-	0.0547*
	100	0.0504	-	0.0494	0.0565	-	0.0568
0.4	10	0.0499	0.0368	0.0497	0.1025	0.0500*	0.0593
	20	0.0528	0.0443	0.0498	0.0762	0.0502*	0.0550
	30	0.0539	0.051	0.0538	0.0661	0.0549*	0.0564
	50	0.0534	-	0.0505	0.0635	-	0.0539*
	100	0.0489	-	0.0496	0.0579	-	0.0570
0.8	10	0.0508	0.0393	0.0509	0.1048	0.0502*	0.0592
	20	0.0517	0.0455	0.05	0.0710	0.0470*	0.0512*
	30	0.0525	0.0504	0.0519	0.0623	0.0530*	0.0535*
	50	0.0516	-	0.0495	0.0625	-	0.0587
	100	0.0472	-	0.049	0.0591	-	0.0587

ii) Power rates

From table (9), the results for the power rates show that:

- Generally, we can say that the power tests are decreased when the variance of Y increases in comparison to e results with the corresponding results of the case of an equality of variance
- It is clear that, the power rates are increasing with the increasing correlation coefficient.
- It can be seen that, with the increasing sample size, the power rates for all tests are increasing and converged to each other.
- Finally, for all test statistics, all power rates of tests are increasing with the increasing Type I error.

Table 9. Power rates on different test statistics Bivariate contaminated normal distribution, $X \sim N(1,\ 1)$ and $X \sim N(1,\ 25)$

ρ	n	T	WS	WL	BT	BWS	BWL
0	10	0.0583	0.0443	0.0563	0.1156	0.0584*	0.0685
	20	0.0694	0.0642	0.0703	0.0912	0.0658*	0.0700*
	30	0.0831	0.0779	0.0817	0.0962	0.0809*	0.0819*
	50	0.0982	-	0.0950	0.1064	-	0.0970*
	100	0.1434	-	0.1400	0.1512	-	0.1477
0.4	10	0.0616	0.0463	0.0602	0.1195	0.0586*	0.0703
	20	0.0755	0.0676	0.0747	0.0971	0.0695*	0.0754
	30	0.0913	0.0859	0.0902	0.1061	0.0905*	0.0920
	50	0.1123	-	0.1079	0.1268	-	0.1143*
	100	0.1747	-	0.1681	0.1761	-	0.1705
0.8	10	0.0681	0.0526	0.0673	0.1192	0.0625*	0.0723
	20	0.0833	0.0747	0.0810	0.1059	0.0778*	0.0823*
	30	0.1020	0.0958	0.0991	0.1196	0.0990*	0.1007*
	50	0.1452	-	0.1378	0.1546	-	0.1376
	100	0.2336	-	0.2251	0.2377	-	0.2268

5. Bivariate Exponential distribution

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

i) Type 1 error rates

Table (10) shows Type 1 error rates for each test, under the Bivariate exponential distribution assumption.

- It can be noted that WL is the most robustness in different cases due to its Type I error rates which are within the Bradley's liberal criterion (0.045-0.055), followed by BWS.
- We can say that T-test is insensitive to non-normality assumption for all cases except for one case when the sample sizes ($n = 10$) with the different values of ρ.
- The BWS is the most robust test for all sample sizes compared to other tests.
- It can be observed 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 noticed that the T test is insensitive to non-normality assumption for all cases except for one case when the sample sizes (n=10) with the different values of ρ.

Table 10. Type I error rates on different test statistics with Bivariate Exponential distribution

ρ	n	T	BT	BWS	BWL	WS	WL
0	10	0.0452*	0.0894	0.0519*	0.0597	0.0377	0.0494*
	20	0.0495*	0.0671	0.0524*	0.0565	0.0447	0.0483*
	30	0.0485*	0.0590	0.0514*	0.0521*	0.0456*	0.0469*
	50	0.0453*	0.0543*	-	0.0512*	-	0.0458*
	100	0.0486*	0.0518*	-	0.0565	-	0.0532*
0.4	10	0.0432	0.0913	0.0543*	0.0634	0.0423	0.0530*
	20	0.0471*	0.0652	0.0508*	0.0542*	0.0431	0.0479*
	30	0.0474*	0.0588	0.0543*	0.0552	0.0501*	0.0518*
	50	0.0489*	0.0543*	-	0.0572	-	0.0543*
	100	0.0513*	0.0541*	-	0.0518*	-	0.0503*
0.8	10	0.0411	0.0920	0.0507*	0.0575	0.0388	0.0469*
	20	0.0433	0.0633	0.0494*	0.0525	0.0412	0.0450*
	30	0.0463*	0.0583	0.0506*	0.0519*	0.0451*	0.0468*
	50	0.0484*	0.0536*	-	0.0508*	-	0.0483*
	100	0.0456*	0.0488*	-	0.0501*	-	0.0461*

ii) Power rates

The results of the simulation study of the power rates can be summarized in **Table (11).**

- The power rates for all methods are increasing with the increase of the sample size and the correlation coefficient may approach 1 when n=30,50 and ρ=0.8.
- The t-test has the most powerful rate for all sample sizes compared to other tests when ρ=0.4,0.8.
- Generally, the power rates of bivariate exponential distribution and bivariate normal are the most powerful than the bivariate contaminated normal distribution.
- When $\rho = 0 \; and \; n = 30,50$, it can be seen that BWL has higher power rates compared to other tests.
- The WL test when the sample size is equal to 10 and ρ≤0.4 have higher power rates.
- It is clear that the power rates of the non-parametric tests (WS, WL, BWS, BWL, BT) when the data follow the bivariate exponential distribution are the most

powerful when compared with the power rates when data follow the bivariate contaminated normal distribution with different values of ρ.

Table 11. Power rates on different test statistics with Bivariate Exponential D istribution

ρ	n	T	BT	BWS	BWL	WS	WL
0	10	0.1504*	0.2344	0.1626*	0.1815	0.1367	0.1667*
	20	0.2659*	0.3200	0.3074*	0.3171	0.2964	0.3098*
	30	0.3545*	0.3986	0.4410*	0.4445*	0.4321*	0.4405*
	50	0.5218*	0.5518*	-	0.6493*	-	0.6438*
	100	0.8150*	0.8277*	-	0.9161	-	0.9137*
0.4	10	0.1281	0.2087	0.1427*	0.1594	0.1244	0.1502*
	20	0.2993*	0.3371	0.2782*	0.2903*	0.2682	0.2820*
	30	0.4753*	0.5004	0.4416*	0.4444	0.4361*	0.4442*
	50	0.7090*	0.7124*	-	0.6408	-	0.6417*
	100	0.9510*	0.9528*	-	0.9144*	-	0.9139*
0.8	10	0.2987	0.4101	0.3099*	0.3402	0.2866	0.3401*
	20	0.7096*	0.7466	0.6619*	0.6711	0.6610	0.6761*
	30	0.8798*	0.8917	0.8447*	0.8481*	0.8456*	0.8514*
	50	0.9895*	0.9886*	-	0.9795*	-	0.9790*
	100	1.0000*	1.0000*	-	0.9998*	-	1.0000*

8. Conclusion

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 the two paired populations. Based on the theoretical part and the results of the Type one error rates and the power rates of the tests, the most important conclusions have been reached:

- It is obvious that the power rates are increasing with the increase of the correlation coefficient and the sample size.
- The presence of outliers leads to a decrease of the type I error rates for the paired t-test statistics.
- In case of the existence of outliers in 20% and the homogeneity of variances of correlated variables, the bootstrapping of the Wilcoxon signed rank test for large sample sizes is best in comparison to other tests for all cases and different value of ρ. Followed by the bootstrapping of the Wilcoxon signed rank test for small sample sizes when $n \leq 10$.
- When the paired data follow the bivariate exponential distribution, the Wilcoxon signed-rank test for large sample sizes is the most powerful compared to the other tests in case of small sample sizes, with different values of ρ, while the bootstrapping of the paired t-test is the best in comparing to other tests when $n \geq 30$ and $\rho > 0$.
- In case of the existence of outliers (with homogeneity of variance), we recommend apply the Wilcoxon signed rank test for large sample sizes test when the sample size $n \leq 50$ and $\rho > 0$.

References

1. Huber, P.J. *Robust Statistics*; John Wiley & Sons, Inc.: New York, USA, **1981**.
2. Bradley, J. V. A common situation conducive to bizarre distribution shapes. *The American Statistician*. **1977**, *31*, 147-150.
3. Salter, K.C. ; Fawcett, R. F. A robust and powerful rank test of treatment effects in balanced incomplete block designs. *Communications in Statistics: Simulation and Computation*. **1985**, *14, 4*, 807-828.
4. Kim, H. Park. C. ; Wang, M. Paired t-test based on robustified statistics, Conference Paper. **1985**.
5. Ganji, M.; Bevrani, H. ; Golzar, N. H. A New Method for Generating Continuous Bivariate Distribution Families. *JIRSS*. **2018**, *17, 1*, 109-129, doi: 10.29252/jirss.17.1.109.
6. Park, C.;Wang, M. ; Hwang, W. Y. A Study on Robustness of the Paired Sample Tests. *Industrial Engineering & Management Systems*. **2020**, *19, 2*, 386-397.
7. Zimmerman, D. W. A. Note on the interpretation of the paired samples, *Journal of Educational and Behavioral Statistics*.**1997**, *22, 3*, 349 – 360.
8. Sheskin, D. J. Handbook of parametric and nonparametric statistical procedures. second ed. Boca Raton, FL: Chapman & Hall/CRC; **2000**, [Google Scholar].
9. Al-Saadi. S. D.; Young. D. H. Estimators for the correlation coefficient in a bivariate exponential distribution. *Journal of Statistical Computation and Simulation*. **1980**, *11*,13-20.
10. Scheffe, H. The Analysis of Variance. John Wiley and Sons: New York. **1959**.