IOSR Journal of Mathematics (IOSR-JM)
e-ISSN: 2278-5728, p-ISSN:2319-765X. Volume 10, Issue 1 Ver. IV. (Feb. 2014), PP 01-06
www.iosrjournals.org

On Consistency and Limitation of paired t-test, Sign and Wilcoxon Sign Rank Test

[1]Akeyede Imam, [2]Usman, Mohammed. and [3]Chiawa, Moses Abanyam.
[1]Department of Mathematics, Federal University Lafia, Nigeria.
[2]Department of Statistics, Federal Polytechnic Bali, Nigeria.
[3]Department of Mathematics and Computer Science, Benue State University Makurdi, Nigeria.
Correspondence: Akeyede Imam, Department of Mathematics, Federal University Lafia, PMB 106 Lafia, Nasarawa State, Nigeria.

Abstract: *This article investigates the strength and limitation of t-test and Wilcoxon Sign Rank test procedures on paired samples from related population. These tests are conducted under different scenario whether or not the basic parametric assumptions are met for different sample sizes. Hypothesis testing on equality of means require assumptions to be made about the format of the data to be employed. The test may depend on the assumption that a sample comes from a distribution in a particular family. Since there is doubt about the nature of data, non-parametric tests like Wilcoxon signed rank test and Sign test are employed. Random samples were simulated from normal, gamma, uniform and exponential distributions. The three tests procedures were applied on the simulated data sets at various sample sizes (small, moderate and large) and their Type I error and power of the test were studied in both situations under study.*

Keywords*: Parametric, t-test, Sign-test, Wilcoxon Sign Rank, Type I error, Power of a test.*

I. Introduction

Nonparametric tests are "distribution-free" methods because they do not rely on any underlying mathematical distribution, in other words they are distribution free statistics. The paired sample Wilcoxon signed rank test and sign-test are nonparametric methods used in the comparison of the equality of the medians of two populations especially when the normality assumption of the data is violated. The test makes use of data input from a matched pair. Unlike the paired-sample t-test, the paired-sample Wilcoxon signed rank and Sign test do not require the assumption that the populations are normally distributed. So when the normality is questionable, the paired sample Wilcoxon signed rank is one of the best tests to use to substitute the paired-sample t-test.

In this work we develop some hypothesis tests in situations where the data come from a probability distribution whose underlying distribution may be normal or non normal and different sample sizes are considered for the each case of a paired sample. If the observations from two samples are related, then we have paired observations. Examples of paired observations include:

(i) The same subjects measured for a characteristic on two occasion such as before and after receiving a treatment
(ii) Performance of a student in his year 1 and 2
(iii) Effectiveness of a method over control.

In non parametric tests, it will not be assumed that the underlying distribution is normal, or exponential, or any other given type. Because no particular parametric form for the underlying distribution is assumed and example of such tests for one or two sample locations are Wilcoxon sign ranked and sign tests. The strength of a nonparametric test resides in the fact that it can be applied without any assumption on the form of the underlying distribution. It is good for data with outliers and work well for ordinal data (data that have a defined order) because it based on ranks of data.

Of course, if there is justification for assuming a particular parametric form, such as normality, then the relevant parametric such as t-test should be employed. The focal point of parametric test is some population parameters for which the sampling statistics follows a known distribution, with measurements being made at the interval or ratio scale. When one or more of these requirements or assumptions are not satisfied, then non-parametric methods can be used, which focuses particularly on the fact that the distribution of the sampling statistics is not known ([1]).

In non-parametric tests very few assumptions are made about the distribution underlying the data and, in particular, it is not assumed to be a normal distribution. Some statisticians prefer to use the term distribution-free rather than non-parametric to describe these tests ([2]). Non-parametric statistical tests are concerned with the application of statistical data in nominal or ordinal scale to problems in pure science, social science, engineering and other related fields. Most of the present analysis carried out by non science and science

oriented researchers are based on parametric test, and it is often reasonable to assume that observations come from a particular family of distributions. Moreover, experience backed by theory, suggest that for measurements, inferences based on the assumption that observations form a random sample from some normal distribution may not be misleading even if the normality assumption is incorrect, but this is not always true ([3]).

Nonparametric tests often are used in conjunction with small samples, because for such samples the central limit theorem cannot be invoked. Nonparametric tests can be directed toward hypothesis concerning the form, dispersion or location (median) of the population. In the majority of the applications, the hypothesis is concerned with the value of a median, the difference between medians or the differences among several medians. This contrasts with the parametric procedures that are focused principally on population means. If normal model cannot be assumed for the data then the tests of hypothesis on means are not applicable. Nonparametric tests were created to over come this difficulty. Nonparametric tests are often (but not always) based on the use of ranks; such as Wilcoxon rank test, Sign test, Wilcoxon rank sum test, Kruskal wallis test, Kolmogorov test, etc ([4], [5]).

The objectives of this paper are of two folds:

i. To examine the effect of non-normality on parametric t-test and the non-parametric tests of the Wilcoxon sign rank test and Sign test effect.
ii. To examine the effects of sample size on the three test procedures based on type I error and power of test.

II. Materials And Methods

The materials used for the analysis were generated data using simulation procedures from the required distributions. Since it is very difficult to get data that follows these distribution patterns, even if there is, it is very difficult to get the required number of replicates for the sample sizes of interest. The parametric (t-test) and nonparametric (Wilcoxon signed rank test and Sign-test), methods of analyzing paired sample were applied, to compare the performance of each test on the generated data from the Normal, Uniform, Exponential and Gamma distributions based on the underlying criteria for assessment

2.1 Simulation Procedures and Analysis

Random samples were simulated from Normal, Uniform, Gamma and Exponential distributions respectively for sample size of 5 10, 20, 25, 30 and 40 which considered as small moderate and large sample sizes respectively. Each test procedures were applied on the data sets at varying sample sizes and their Type I error and power of the tests were studied in each situation. At every replicate two samples were simulated simultaneously from each distribution using the same parameters to form the paired sample from the same family. The process was repeated 500 times for each sample size considered and results were displayed in Table 1-5.

2.2 Criteria for Assessment and Test of Significance

Some decision must often be made between significance of a test or not. Turning the p-value into a binary decision allows us to examine two questions about the comparative value of statistical tests:

1. What percent of significant results will a researcher mistakenly judge to be in significant?
2. What percent of reported significant results will actually be in significant?

Indeed the number of rejecting H_0 when it is true is counted for Type I error and number of times H_0 is accepted when it is true was recorded as power of the test from each statistic under study.

2.3 Student's Paired t-test

The t-test's null hypothesis is that systems A and B are random samples from the same normal distribution. The details of the paired t-test can be found in most statistics texts (such as [6]).

In this case, we use the differences between the individual pair says x_i and y_i on individual i such that: $d_i = x_i - y_i$ and

$$T = \frac{\sqrt{n}(\bar{d} - \mu_d)}{s_d} \sim t_{n-1} \qquad (1)$$

where $\bar{d}$ is the mean of the sample differences and s_d is the standard deviation of the sample differences.

Under H_o: $\mu_d = 0$ (i.e. hypothesis of no difference). Note that the null hypothesis may also be in the form $\mu_d = \mu_0$ when we wish to know if the difference is a given value μ_0. The t-test strictly assumes that the observations in the sample have come from a normally distributed population. The t-test also requires the observation be measured at least in an interval scale ([7]). Meanwhile, the Sign and Wilcoxon test provided a means to test "the

wider hypothesis in which no normality distribution is implied", our contention here was that if the p-value produced by the t-test in any distribution was close to the p-value produced by the sign and Wilcoxon tests, then the t-test could be trusted. In practice, the t-test has been found to be a good when the assumption of normality is found.

2.4 The Wilcoxon Signed Rank Test

The null hypothesis of the Wilcoxon signed rank is the same as the sign test ([8]), i.e. both tests test hypothesis about the median. Whereas the sign test does not take the magnitude of the observation into account the Wilcoxon signed rank test does. The Wilcoxon test statistic takes the paired score differences and ranks them in ascending order by absolute value. The sign of each difference is given to its rank as a label so that we will typically have a mix of "negative" and "positive" ranks. For a two-sided test, the minimum of the sums of the two sets of ranks is the test statistic. Differences of zero and tied differences require special handling ([8]).

The Wilcoxon test statistic throws away the true differences and replaces them with ranks that crudely approximate the magnitudes of the differences. This loss of information gained computational ease and allowed the tabulation of an analytical solution to the distribution of possible rank sums. One refers the test statistic to this table to determine the p-value of the Wilcoxon test statistic. For sample sizes greater than 25, a normal approximation to this distribution exists ([8]).

It denotes S_+ to be the sum of positive ranks and S_- the sum of negative ranks.

$$S_+ = \sum_i^n \Psi_i r|Z_i| \quad \text{where } \Psi_i = \begin{cases} 1 & \text{if } Z_i > 0 \\ 0 & \text{if } Z < 0 \end{cases} \tag{2}$$

or

$$S_- = \sum_i^n \Psi_i r|Z_i| \quad \text{where } \Psi_i = \begin{cases} 0 & \text{if } Z_i > 0 \\ 1 & \text{if } Z_i < 0 \end{cases} \quad i = 1,\ldots,n \tag{3}$$

and $r|Z_i|$ is the rank of absolute value of Z_i's, where $Z_i = x_i - y_i$ and x_i represent the individual observations. When testing H_0 against $H_1 : \mu_d \neq 0$, we reject H_0 if $S_+ \geq \frac{n(n+1)}{2} \rightarrow t_{\alpha/2}$. For exact p-value, that is $\Pr(S \leq S_+) = p$, it rejects H_0 if $p \leq \alpha$.

2.5 Sign-Test

Like Wilcoxon tests, the sign test has a null hypothesis that systems A and B have the same distribution ([8]). The test statistic for the sign test is the number of pairs for which system A is different from system B. Under the null hypothesis, the test statistic has the binomial distribution with the number of trials being the total number of pairs. The number of trials is reduced for each tied pair. [9] proposed that a tie should be determined based on some set absolute difference between two scores ([9]).

The sign test allocates a sign to each observation according to whether it lies above or below some hypothesized value, and does not take the magnitude of the observation into account**.** The sign-test does not specify any underlined distribution and therefore it is a distribution free statistics. The observations are continuous variable with atleast an ordinal scale. When testing H_0: $m_d = 0$ against H_0: $m_d \neq 0$, we let x and y to be number of first and second observations, respectively observed from the same population and we obtain $d_i = x_i - y_i$. Then, we count the number of positive d_i and represent it by T^+, if $d_i = 0$, we remove the observation from the sample and reduce the sample by one. We reject H_0 if T^+ lies outside the confidence interval i.e critical value from Binomial table, otherwise we do not reject H_0.

III. Data Analysis

A fixed significance level of 5% was selected for H_0: $\mu_d = 0$. In other words, the mean was the same from the paired sample against $H_1 : \mu_d \neq 0$, the mean was not the same for the paired sample, where μ_d represents the value of the average of the deviation between the two sets of the paired sample from each distribution of the sample size of interest. The test was carried on the 500 paired samples generated for each parameter of the distributions. We count the number of times we correctly accept H_0 for fixed H_0's to provide the power of the test and wrongly reject the fixed value of H_0 to determine the Type I error. These are recorded as probabilities for the four statistical tests, under the Normal, Uniform, Exponential distributions and Gamma.

For example, for the normal distribution, $H_0:\mu_d = 0$ against $H_1:\mu_d \neq 0$ for different values of generated μ, for the Uniform distribution $H_0:\mu_d = 0$ against $H_1:\mu_d \neq 0$ and for the exponential distribution $H_0:\mu_d = 0.5$ against $H_1:\mu_d \neq 0.5$, all for different values of generated μ. The test was carried out on the 500 samples generated for each μ_d and each distribution and were recorded as probabilities for the two statistical tests under each of the given distributions. Averages of power of the test for the levels of the population means were calculated and recorded for each of the sample size under the four given distributions for the three statistical tests. These averages presented in Tables 1, 2, 3 and 4.

Table 1: Relative Frequencies of Acceptance of the Null Hypothesis $H_0: \mu_d = 0$ from Data Generated from Normal Distribution

	Test Statistic		
Sample size(n)	t-test	Sign-test	Wilcoxon Sign Rank Test
5	0.9560	0.8900	0.9020
10	0.9760	0.8600	0.8980
15	0.982	0.8500	0.8840
20	0.9660	0.7980	0.8020
30	0.9980	0.6640	0.7780
40	0.9540	0.5100	0.6040
Average	**0.9720**	**0.762**	**0.8113**

Table 2: Relative Frequencies of Acceptance of the Null Hypothesis $H_0: \mu_d = 0.2$ from Data Generated from Gamma Distribution

	Test Statistic		
Sample size(n)	t-test	Sign-test	Wilcoxon Sign Rank Test
5	0.7940	0.9960	0.9960
10	0.8120	0.9860	0.9960
15	0.8220	0.9860	0.9980
20	0.8200	0.9800	0.9840
30	0.9000	0.8160	0.8040
40	0.7600	0.760	0.7740
Total	**0.818**	**0.9467**	**0.9281**

Table 3: Relative Frequencies of Acceptance of the Null Hypothesis $H_0: \mu_d = 0.5$ from data generated from Exponential Distribution

	Test Statistic		
Sample size(n)	t-test	Sign-test	Wilcoxon Sign Rank Test
5	0.7980	0.9940	0.9960
10	0.8200	0.9880	0.9980
15	0.8220	0.9860	0.9980
20	0.8200	0.9800	0.9980
30	0.9060	0.8000	0.9980
40	0.9080	0.7440	0.8040
Total	**0.8457**	**0.9467**	**0.9660**

Table 4: Relative frequencies of acceptance of the null hypothesis $H_0: \mu_d = 0$ from data generated from Uniform Distribution

	Test Statistic		
Sample size(n)	t-test	Sign-test	Wilcoxon Sign Rank Test
5	0.8000	0.9800	0.9840
10	0.9020	0.9960	0.9880
15	0.9340	0.9860	0.9940
20	0.9560	0.8060	1.000
30	0.7980	0.8080	0.8400
40	0.7600	0.760	0.8160
Total	**0.8583**	**0.8893**	**0.9437**

Table 5: Relative frequencies of Rejection of the null hypothesis $H_0: \mu_d = 0$ from data generated from Normal Distribution

	Test Statistic		
Sample size(n)	t-test	Sign-test	Wilcoxon Sign Rank Test
5	0.0400	0.0280	0.0480
10	0.0440	0.0320	0.0480
15	0.0440	0.0460	0.0460
20	0.0435	0.0420	0.040
30	0.0380	0.0360	0.0360
40	0.0340	0.0380	0.0340
Average	**0.0406**	**0.0370**	**0.0420**

Table 6: Relative frequencies of Rejection of the null hypothesis H_0: $\mu_d = 0.2$ from data generated from Gamma Distribution

	Test Statistic		
Sample size(n)	t-test	Sign-test	Wilcoxon Sign Rank Test
5	0.1140	0.0520	0.0520
10	0.1080	0.0620	0.0580
15	0.0960	0.0820	0.0760
20	0.018	0.1100	0.0780
30	0.078	0.1040	0.0840
40	0.0820	0.138	0.1240
Total	**0.0827**	**0.0913**	**0.0787**

Table 7: Relative frequencies of Rejection of the null hypothesis H_0: $\mu_d = 0.5$ from data generated from Exponential Distribution

	Test Statistic		
Sample size(n)	t-test	Sign-test	Wilcoxon Sign Rank Test
5	0.0820	0.0360	0.0460
10	0.0685	0.040	0.0500
15	0.057	0.0605	0.0760
20	0.0533	0.055	0.0740
30	0.0635	0.0585	0.0860
40	0.055	0.0675	0.1200
Total	**0.0632**	**0.0529**	**0.0753**

Table 8: Relative frequencies of Rejection of the null hypothesis H_0: $\mu_d = 0$ from data generated from Uniform Distribution

	Test Statistic		
Sample size(n)	t-test	Sign-test	Wilcoxon Sign Rank Test
5	0.0700	0.0260	0.0300
10	0.0580	0.0280	0.0320
15	0.0520	0.0600	0.0300
20	0.0440	0.0600	0.0640
30	0.0240	0.0820	0.0760
40	0.0550	0.0980	0.1020
Total	**0.0505**	**0.0427**	**0.0557**

IV. Discussion Of Results

Tables 1 – 8 indicate results of analyses using the paired t-test, Sign-test and Wilcoxon signed rank test on how the tests perform based on the type I error and power of the test, both being compared at the 5% level of significance for two tailed test in each case. The average of each value of the type I error and power of the test were calculated and recorded under each statistical test for easy comparison.

The power of the t-test increases as sample size increases from data generated from the three distributions with value from 0.9560 to 0.9880 from normal of sample size of 5 to 30 respectively but started to decreases when sample size of 40 was used. However, the power of the paired sample Wilcoxon signed rank test and the sign test decreases as the sample size increases from 5 to 40 with the lowest values at sample sizes 40 as shown in Table 1-4. More so, the type I error of the t-test decreases from the three distributions and started to increase at sample size of 40 as we can see in the Table 5 – 8. However, the two nonparametric tests increase in the Type I error from the lowest power to the highest from each distribution under study.

The t-test test has the highest power from the data generated from normal distribution as shown in Table 1 followed by Wilcoxon Sign test. In the data generated from gamma distribution Sign test has the highest power followed by Wilcoxon test while the Wilcoxon Sign test has the highest power from exponential and uniform distribution especially from sample size of 5 to 40. The Wilcoxon Sign test has the lowest Type I error from normal while the t-test has the highest. The sign test has the lowest Type I error from other distributions followed by the Wilcoxon Sign test especially at the sample size less than 40 (see Table 5-8)

4.1 Conclusion

It was observed that the Sign test closes to '(1- β)' from the data generated from uniform distribution and therefore consider as the most powerful test in that respect while Wilcoxon Sign Rank test is the closest to '(1- β)' from the data generated from exponential and uniform especially for small sample sizes and considered as the most powerful test for that distributions. However, there is no significant differences in the power of the two tests when rounded to two decimal places, if they are compared based on the simulated data from the four selected distributions, using small sample sizes at the 5% levels of significance. Meanwhile the t-test is the most suitable test when the underline distribution is normal and when sample sizes are large for any distributions as

reported in the Table 1-8. However the two nonparametric tests are indeed alternative tests to t-test when the assumption of normality is not met.

References

[1] L. J. Kazmier, *Schaum's Outline of Business Statistics* 3rd Edition Mc Graw-Hill New York, 1996.

[2] G. M. Clarke *and D.* Cooke, *The Nature of Parametric and Non-P*arametric tests. *A Basic Course in* Statistics, 4th edn. Arnold, 1998.

[3] P. Sprent, *Applied Non-parametric Statistical Methods*. 2nd Edition. Chapman and Hall, 1993.

[4] M. Usman and A. I. Maksha, An Efficient Alternative To t-Test For One Sample Nonnormal Data, *Journal of Applied Science & technology*, Auchi Polytechnic, Nigeria, 2010.

[5] I. Akeyede and S. G. Akinyemi, Power Comparison of Sign and Wilcoxon Sign Rank Test Under Non Normal, *African Journal of Physical Sciences*, Devon Science Publication, 2010.

[6] B. A. Oyejola and S. B. Adebayo, *Basic Statistics for Biology and Agriculture Students*, Olad Publishers, Nigeria, 2003.

[7] S. Siegel and N. J. Castellan, *Non-Parametric Statistics for the Behavioural Science*, 2nd Edition New York: McGraw. Hill, 1988.

[8] W, Mendenhall, D. D. Wackerly and R. L Scheaffer, *Mathematical Statistics with Applications*, PWS-KENT Publishing Company, 1990.

[9] C. J. van Rijsbergen, *Information Retrieval. Butterworths*, 2nd Edition, 1979.