Fjrs.futa.edu.ng
FUTA Journal of Research in Sciences
ISSN: 2315 – 8239 (Print); E-ISSN: 2489 - 0413

FUTA Journal of Research in Sciences, Vol. 13 (2), October, 2017: 350-363

A STUDY ON SENSITIVITY AND ROBUSTNESS OF MATCHED-PAIRS INFERENTIAL TEST STATISTICS TO OUTLIERS

[1*]Adegoke S. Ajiboye, [2]Taiwo Joel Adejumo and [1]Kayode Ayinde,

[1]Department of Statistics, Federal University of Technology, Akure, Nigeria.
[2]Department of Statistics, Ladoke Akintola University of Technology, Ogbomoso, Nigeria.
Correspondence: asajiboye@futa.edu.ng,

ABSTRACT

Outliers are data points that are different from others. Their presence may affect the robustness and the sensitivity of test statistics used for inferential purposes. Test statistics that are meant for the purpose of inference have been developed which include: Paired t-test, Distributional Wilcoxon Sign Test (DST), Asymptotic Wilcoxon Sign Test (AST), Distributional and Asymptotic Wilcoxon Signed rank Test (DWST and AWST), t-test for rank transformation (Rt-test) and Trimmed t-test statistics (Tt-test). Consequently, the effect of outliers on these test statistics needs to be investigated so as to determine the one that is sensitive and robust. The experiment was conducted five thousand (5000) times using Monte Carlo experiments at eight (8) levels of sample sizes namely: 10, 15, 20, 25, 30, 35, 40 and 50 by generating data from normal distribution with the aid of R- statistical programming codes. Also, in order to exhibit different degree of correlations between the paired samples, the levels of correlations reconsidered are; 0, 0.3, 0.6, 0.9, 0.95 and 0.99. At each sample size, 10% and 20% of the generated data were invoked with twenty-one (21) various magnitude (k) of outliers ($_{-10, -9, -8, ...8, 9, 10}$). The three (3) commonly used preselected levels of significance used were 0.1, 0.05 and 0.01. The Type 1 error rate of the inferential test statistics was determined when there was no outlier in the data sets. While, to assess the sensitivity and robustness of the test statistics hence, the Power rate was determined. A test is considered robust if its estimated error rate approximates the true error rate and has the highest number of times it approximates the error rate when counted over the levels of significance otherwise sensitive. With different levels of correlation, results revealed that the Type 1 error rate of the paired t- test and AWST are good; and that AST, Rt-test and DST, and Tt-test statistics are respectively robust to outliers at 0.1, 0.05 and 0.01 levels of significance.

Keywords: Outliers, Power rate, Sensitivity, Robustness, Inferential Test Statistics, Monte Carlo

INTRODUCTION

Brase and Brase (1999) defined an outlier as a score or case that is so low or so high that it stands apart from the rest of the data. Some descriptive statistics of a data set have been noted to be affected by outliers. While the median is not being affected, other measures especially the arithmetic mean is affected. Most parametric test statistics including the Student t-test (Gosset, 1908), z-test (Gauss, 1809), paired t-test etc. are directly meant for hypothesis testing about the mean while the non-parametric ones test hypothesis about the median. The non-parametric inferential test statistics for one sample and paired sample problem include the Sign test (John, 1710) and Wilcoxon signed rank test (Wilcoxon, 1945). The t-test for rank transformation (Rt-test) by Conover and Ronald (1981) developed new test statistics in rank form which bridged the gap between the parametric and the non-parametric while Yuen (1974) developed

another test statistic called the trimmed t-test (Tt-test) which excludes outliers in its test procedure. The menace of outliers on these various test statistics needs to be investigated as this inevitably affects inference. Consequently, this research work examines the effect of outliers on some test statistics so as to determine the one that is robust and sensitive to outliers. Therefore, this study identified, Distribution Wilcoxon Sign rank test (DWST) and the Paired t-test Statistics to be sensitive while the Sign test (Asymptotic Sign test (AST), Distribution Sign test (DST)) and Trimmed t-test (Tt-test) Statistics were also identified to be robust to outliers.

Review on some matched-pairs inferential test statistics

Paired (or related samples) t-test

A paired t-test can be used to compare means when there are two samples in which observations in one sample can be paired with observations in the other sample. In order words the two samples related or dependent. Its test statistic is given as:

$$T = \frac{\overline{D} - \mu_D}{SE(\overline{D})} \qquad (1)$$

Where $SE(\overline{D}) = \frac{S_D}{\sqrt{n}}$, $\overline{D}$ = the (sample) mean of the difference scores, μ_D = the mean difference in the population, given a true H_0 (often $\mu_D = 0$, but not always), S_D = the sample standard deviation of the difference scores (with division by n-1), $SE(\overline{D})$ = the standard error of the mean difference scores with df = n-1 and n is the number of matched pairs.

The sign test

The sign test discovered by John, (1710) is a non-parametric test that is used to test whether or not two groups are equally sized. It is used when dependent samples are ordered in pairs, where the bivariate random variables are mutually independent. It is based on the plus and minus sign of the observation and not on their numerical magnitude. It is also called the binomial sign test, with p=0.5. The sign test is considered a weaker test, because it tests the pair value below or above the median and it does not measure the pair difference (Wikipedia, 2016). The paired sample sign test is an alternative to the paired sample t-test. The test statistics are W^+or W^-. UnderH_0, binomial distribution is used and H_0 is rejected if P (W$\leq W^+$) <α or P (W$\leq W^-$) <α.$\alpha/2$is used instead of α when the test is two-tail. Asymptotically, the sign test has its distribution binomial (n, 0.5) formula as: ${}^nC_x q^{n-x}, p_x$ with p_x = 0.5.

Its asymptotic test statistic is:

$$Z = \frac{W^+ - \frac{n}{2}}{\sqrt{\frac{n}{4}}} \sim N(0,1) \qquad (2)$$

Wilcoxon signed rank test

Wilcoxon signed-rank test proposed by Wilcoxon, (1945) is a non-parametric statistical hypothesis test used when comparing two related samples, or repeated measurements on a so assess whether their population mean ranks differ (i.e it is a paired difference test). It can be used as an alternative to the paired student's t-test for dependent samples when the population cannot be assumed to be normally distributed. The test was further popularized by Sidney, (1956).The test in his influential text book on non-parametric he used the symbol T for value related to, but not the same.(Wikipedia, 2016).

The asymptotic distribution of Wilcoxon signed rank test is:

$$Z = \frac{T^+ - E_0(T^+)}{\sqrt{V_0(T^+)}} \sim N(0,1) \qquad (3)$$

where $E_0(T^+) = \frac{(n+1)}{4}$ and

$V_0(T^+) = \sqrt{\frac{n(n+1)(2n+1)}{24}}$

T-test for Rank Transformation in one and matched pairs sample

Conover and Ronald (1981) bridged the gap between parametric and non-parametric test by proposing new test statistic. The test statistic was defined as follows: Let D_1, D_2,…, D_n represent independent random variables with a common mean where in the case of matched pairs (X_i, Y_i); $D_i = X_i$, - Y_i. The test statistic is defined as:

$$T = \frac{\sum_{l=1}^{n} R_i}{\sqrt{\sum_{l=1}^{n} R_i^{\,2}}} \qquad (6)$$

Where R_i = (sign D_i) x (rank of / D_i/)

The alternative t-test statistic is computed on the signed ranks as;

$$t_R = \frac{\sum_{l=1}^{n} R_i}{\sqrt{\frac{n\sum_{l=1}^{n} R_i^{\,2} - (\sum_{l=1}^{n} R_i)^2}{n-1}}} \qquad (7)$$

which is compared with the t-distribution (n-1) degree of freedom.

t_R can be expressed as:

$$t_R = \frac{T}{\sqrt{\frac{n}{n-1} - \frac{T^2}{n-1}}} \qquad (8)$$

Which was defined to be a monotonic function of T.

The Trimmed t-test

Yuen proposed another test statistic for the independent two-sample case, under unequal population variances that excludes outlier in (1974) named the Trimmed t-test (Keselman, Wilcox, Algina, and Fradette, 2008).The trimmed mean is computed by removing g-observations from each tail of the distribution. The trimmed mean is computed as follows:

$$\overline{X}_t = \frac{X_{g+1} + X_{g+2} + \ldots + X_{n-g}}{n-2g} \qquad (9)$$

Where $x_1,\ldots,x_n$ are the ordered values in a sample, g = observations trimmed from each tail of the sample distribution, n – 2g = the number of observations in the trimmed sample.

In addition to the trimmed mean, the Winsorized mean is required to compute the appropriate variance estimate. Instead of "trimming" this method replaces the most extreme g- observations by the next-most-extreme value. The Winsorized mean is computed as:

$$\overline{X}_w = \frac{([g+1]X_{g+1} + X_{g+2} + \ldots + [g+1]X_{n-g})}{n} \qquad (10)$$

The Winsorized sum-of-squared derivation is computed as:

$$SSD_w = [g+1]\left[x_{g+1} - \overline{X}_w\right]^2 + \left[x_{g+2} - \overline{X}_w\right]^2 + \cdots + [g+1]\left[x_{n-g} - \overline{X}_w\right]^2 \qquad (11)$$

The Winsorized variance is obtained as:

$$S_w^2 = \frac{SSD_w}{n-2g-1} \qquad (12)$$

Methodology

Matched Pairs Sample Problem

Using Monte Carlo simulation procedures, data were generated from the univariate normal distribution with the aids of R-programming statistical codes.

Y_i was generated from the univariate normal distribution with means $(\mu_1 = \mu_2)$ =10 and Standard deviations $(\sigma_1 = \sigma_2)$ = 5 Magnitude of outlier (k). Percentages of the generated data to be

invoked with outliers are 10% and 20%. In the same vein, twenty one (21) magnitude of outliers (k) taken are: -10, -9, -8, ... 8, 9, 10. In order to exhibit different degrees of correlations between paired observations, the levels of correlation used are: $\rho_{12} = \rho = 0$, 0.3, 0.6, 0.9, 0.95 and 0.99. At eight sample sizes namely: 10, 15, 20, 25, 30, 35, 40 and 50 the experiment was replicated five thousand (5000) times.

The simulation procedures for invoking the outliers and estimation of the Type 1 error rate of the paired test statistics are as follows:

(i) Choose a percentage of the data to be replaced with outliers.

(ii) Choose a particular magnitude of outlier to invoke into the generated data

(iii) Choose a sample size to work with, say n.

(iv) Generate random sample with size n from a normal distribution with $\mu_1 = \mu_2 = 10$ and $\sigma_1 = \sigma_2 = 5$

$Y_i \sim N(10, 25)$

(v) Randomly select those observations making up the percentage of the generated data to be replaced with outliers.

(vi) Outliers are invoked as follows:

$Y(i)_{outlier} = k*Max(Y_i) + Y_i$ (13)

Where: Y_i = selected generated observation i

$Y(i)_{outlier}$ = Outlier to replace Y_i

k= Magnitude of outliers

$Max(Y_i)$ = Maximum of the generated data

(vii) Replace the outliers in the data originally generated in (iv)

(viii) Apply the various test statistics and keep their p-values.

(ix) For each test statistics in (viii), define:

$$A_i = \begin{cases} 1, if\ \ p-value < \alpha \\ 0, otherwise \end{cases} \quad (14)$$

Where α is a preselected level of significance, say 0.1.

(x) Repeat steps (v) to (ix) five thousand (5000) times, R= 5000.

(xi) For each of the test statistics, sum the results obtained in step (x), i.e

$$A = \sum_{i=1}^{R} A_i \quad (15)$$

(xii) For each of the test statistics, divide the result in step (xi) by the number of replications to estimate the Type 1 error of each test statistics, i.e

$$B_\alpha = \frac{\sum_{i=1}^{R} A_i}{R} = \frac{A}{R} \quad (16)$$

(xiii) Choose another magnitude of outlier to invoke into the generated data and repeat step (v) to (xii).

(xiv) Repeat step (v) to (xiii) until all the magnitudes of the outliers are exhausted.

(xv) Choose another sample size to work with and repeat step (v) to (ix)

(xvi) Repeat step (v) to (xv) until all the sample sizes are exhausted.

(xvii) Choose another percentage of the data to be replaced with outliers and repeat step (ii) to (xvi).

(xviii) Repeat step (ii) to (xvii) until all the levels of percentages are exhausted.

The random sample of observations in step (iv) are paired correlated observations generated using equations provided by Ayinde, (2007) given as:

$$Y_1 = \mu_1 + \sigma_1 Z_1 \quad (17)$$

$$Y_2 = \mu_2 + \rho_{12}\sigma_2 Z_1 + \sqrt{m_{22}} Z_2 \quad (18)$$

Where $Z_1 \sim N(0,1)$, $Z_2 \sim N(0,1)$, and $m_{22} = \sigma_2^2(1-\rho_{12}^2)$

It should be noted that in invoking the paired observations with outliers, the paired randomly selected observations are invoked with outliers separately. Moreover, all steps above were taken for each level of correlation.

Sensitivity and Robustness of the Test Statistics

The test statistics were considered robust if their estimated Type 1 error rates at different % and magnitude of outliers are within the preferred interval of levels of significance suggested by Kuranga (2015) and used by Ayinde, et.al (2016) as presented in Table 1.

On the other hand, the test statistics were considered sensitive if as percentage of outliers and magnitude of outliers increase, the Type 1 error rate of the test statistics also increases. i.e many do not fall into the preferred intervals (the null hypothesis is rejected often).

Table 1: The True level of significance and preferred interval

True level of significance	Preferred interval
0.1	0.095 - 0.14
0.05	0.045 - 0.054
0.01	0.005 - 0.014

Source: Kuranga, (2015) and Ayinde et.al (2016)

Examination under matched-pairs sample problem

The number of times Type 1 error rates were within the preferred interval was counted over the levels of correlation, percentage of outliers, magnitude of outliers and sample size. A test statistic that is robust is expected to have the highest number of counts, the mode and sensitive if the count is the smallest.

Results and Discussion

The results of Type 1 error rates of one and paired inferential test statistics as affected by outliers are presented and discussed as follows:

Matched Pairs Sample Problem

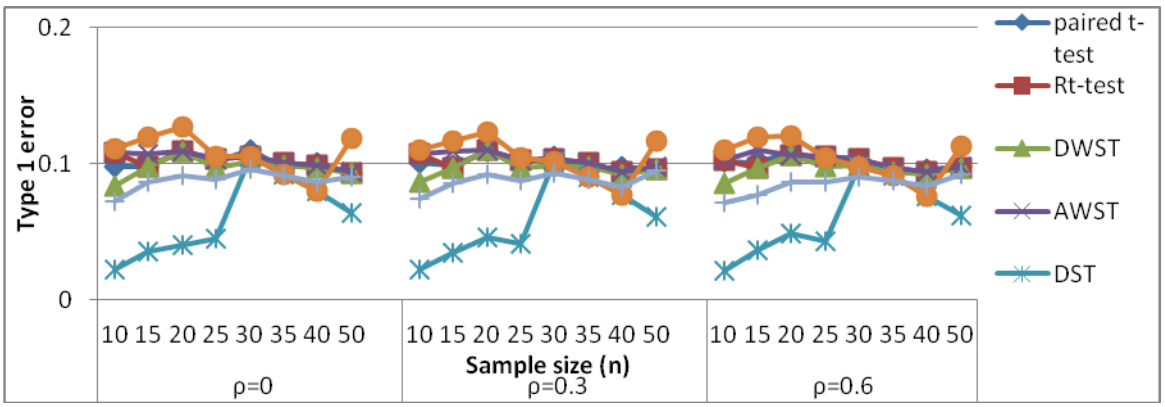
The simulation results for Type 1 error rate are available in (Adejumo, 2016) and the graphical representations of the test statistics are presented and discussed.

Type 1 Error Rates investigation of Matched-pairs test Statistics

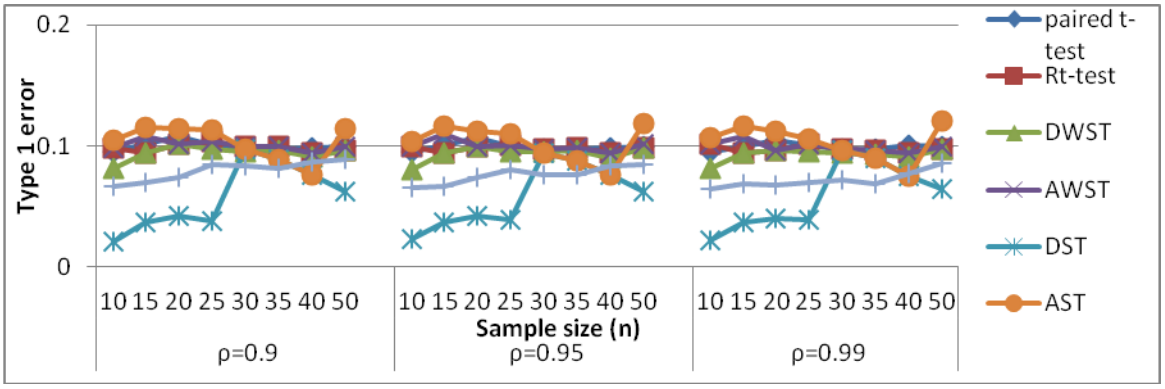
The Simulation results of Type 1 error rates of the inferential test statistics for the matched pairs sample problem at 0.1, 0.05 and 0.01 are provided in (Adejumo, 2016) while the sample graphs at differentlevels of significance are presented and discussed.

Results of Type 1 Error Rates for matched pairs sample at 0.1 level of significance

Figure 1 given below shows graphical representations of the Type 1 error rate of the matched-pairs test statistics at 0.1 level of significance. Having counted over levels of correlation, it was observed that the Type 1 Error rate of Paired t-test, AWST, Rt-test, AST and DWST in this order are good while, that of Tt-test and DST are not good.



(a)

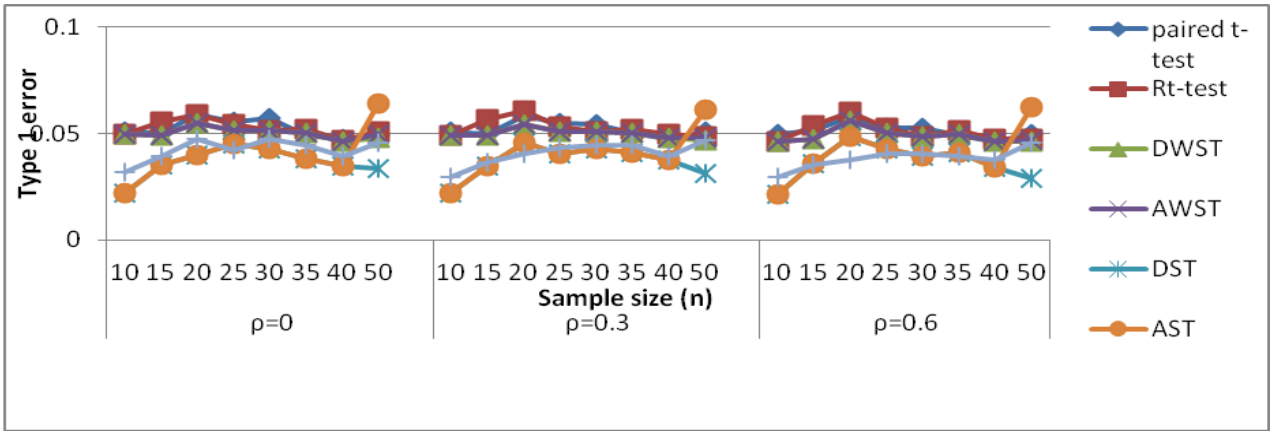


(b)

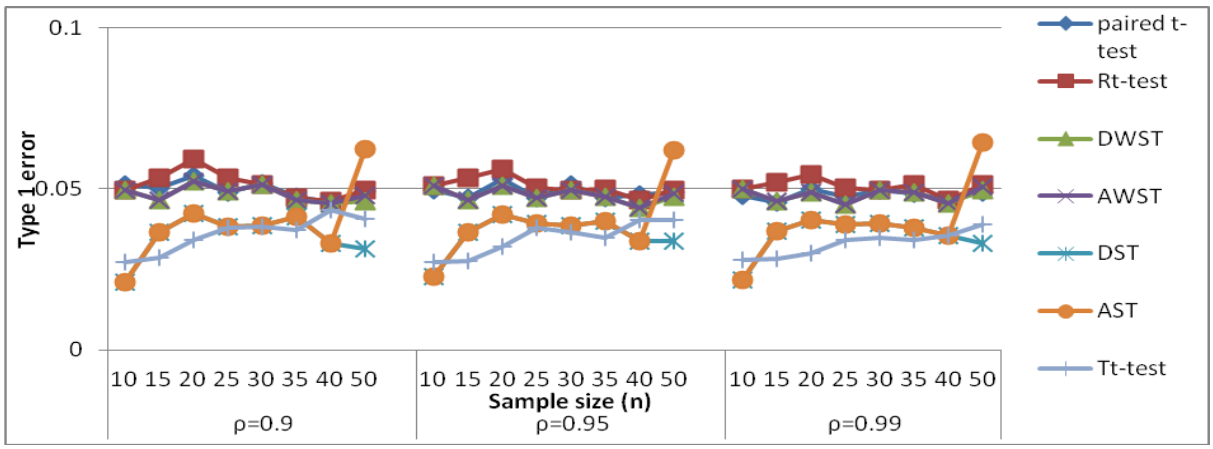
Figure 1: Graphical representation of Type 1 error rate of Matched–Pairs sample test statistics at 0.1 level of significance.

Results of Type 1 Error Rates for matched pairs sample at 0.05 level of significance.

Figure 2 shows the graphical representation of Type 1 error rate of the matched-pairs test statistics at 0.05 level of significance. Also, having counted over levels of correlation, it can be seen that the Type 1 Error rate of DWST, AWST, Paired t-test and Rt-test, in this order are good while, that of Tt-test, DST and AST are not good.



(a)

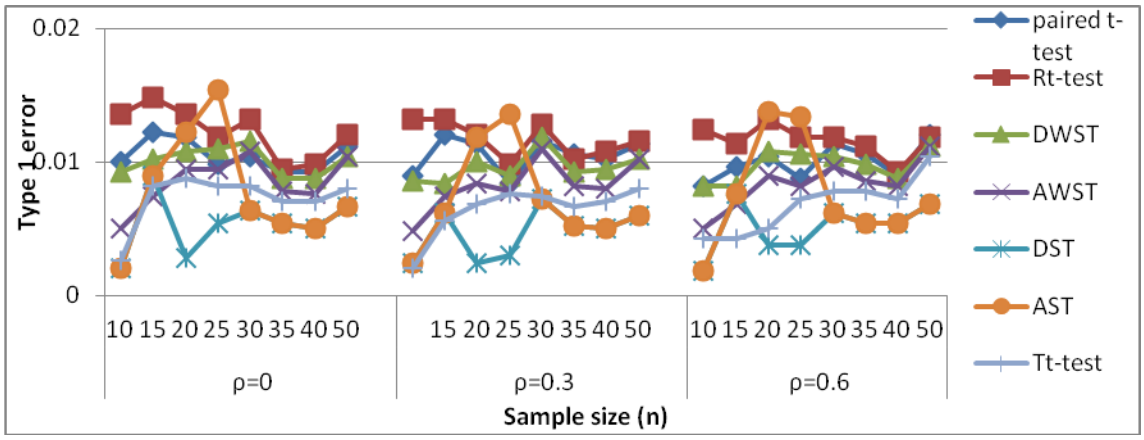


(b)

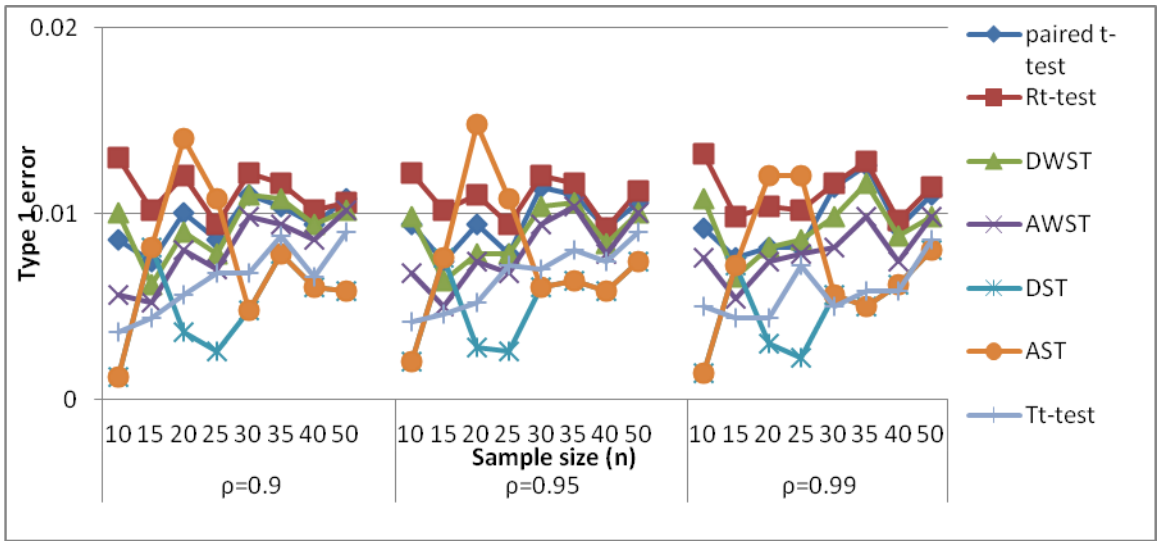
Figure 2: Graphical representation of Type 1 error rate of Matched–Pairs sample test statistics at 0.05 level of significance.

Results of Type 1 Error Rates for matched pairs sample at 0.01 level of significance.

Figure 3 shows the graphical representation of Type 1 error rate of the matched-pairs test statistics at 0.01 level of significance. Also, having counted from Simulation results available in (Adejumo, 2016) over levels of correlation, it can be observed that the Type 1 Error rate of all the test statistics are good, except DST, AST and Tt-test test statistics.



(a)



(b)

Figure 3: Graphical representation of Type 1 error rate of Matched–Pairs sample test statistics at 0.01 level of significance.

Results of Type 1 Error Rates for matched pairs sample when counted over all levels of significance.

Results from Table 2 and Figure 4, it can be seen that the Type 1 error rate of the Paired t-test and AWST is very good while that of DST and Tt-test statistics is not good.

Table 2: Overall number of times Type 1 Error Rate Approximate true level of significance when counted over levels of correlation and levels of significance

	Sample size									
Test statistics	10	15	20	25	30	35	40	50	Total	Rank
Paired t-test	17	18	14	16	16	17	18	17	**133**	**1.5**
Rt-test	18	12	12	17	18	18	13	17	125	3
DWST	12	15	15	18	16	16	12	17	121	4
AWST	17	18	15	18	18	18	12	17	**133**	**1.5**
DST	0	6	2	1	10	6	6	6	37	7
AST	6	12	13	11	10	6	6	12	76	5
Tt-test	1	2	6	6	8	6	6	9	44	6

Source: Counted from Simulation results available in (Adejumo, 2016)

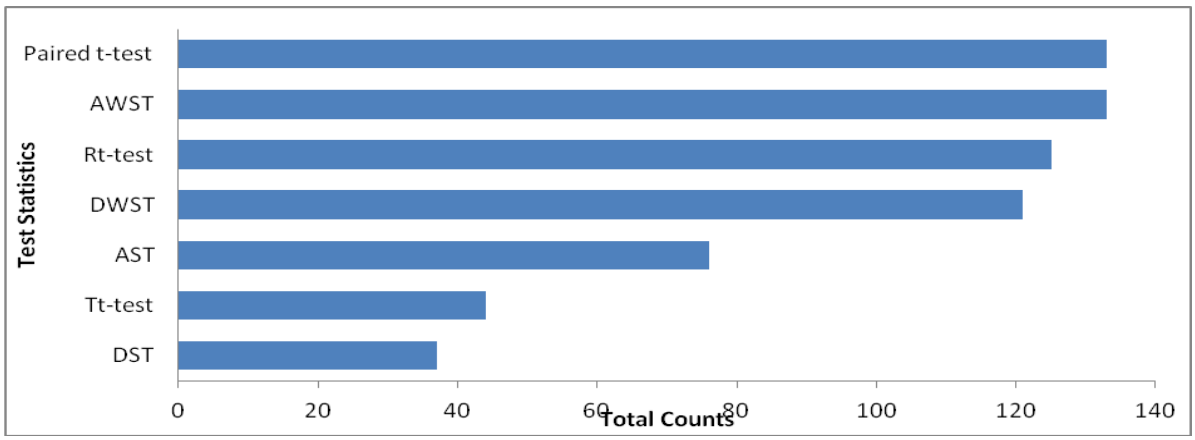


Figure 4: Bar chart showing overall total number of times Type 1 error Rate approximate true levels of significance over all levels of correlations and significance

Sensitivity and Robustness Investigation of Matched-pairs test Statistic

In order to assess the sensitivity and robustness of the test statistics, displayed and discussed are the samples of graphical representations of power rate of the inferential test statistics at 0.1 level of significance under different values of correlation, 10%outliers and magnitude of outliers.

The results obtained about the test statistics having counted the number of times the error rate fall into the preferred interval over the levels of correlation, percentage and magnitude of outliers are presented in Table 3.

From Table 3 the following can be observed:

(i) At 0.1 level of significance, the AST is very robust while Tt- test is sensitive

(ii) At 0.05 level of significance, the Rt- test and DST are robust while Tt- test is still sensitive

(iii) At 0.01 level of significance, the Tt- test is most robust while Rt- test and DWST are sensitive.

Further counting over all the levels of significance resulted into Table 4and Figure 5.

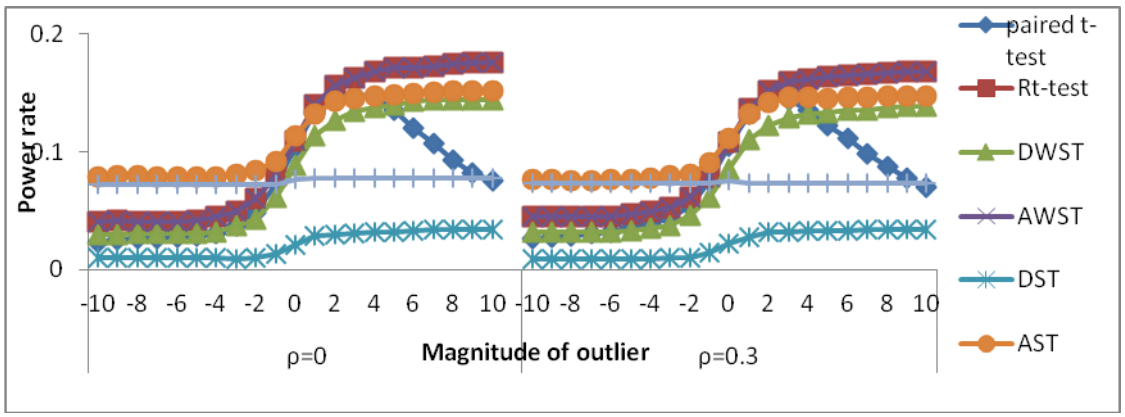
Summarily, it can be concluded that the Sign test (AST and DST) and Tt-test Statistics are robust while DWST and the paired t-test are sensitive.

Table 3: Number of times Power rate approximate true levels of significance when counted over levels of correlation, percentage and magnitude of outliers at the different levels of significance

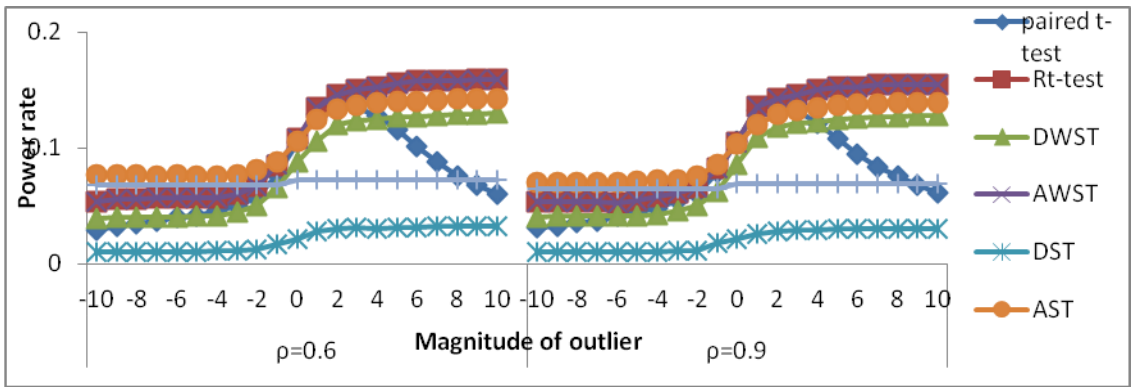
		Sample size									
Alpha level	Test statistics	10	15	20	25	30	35	40	50	Total	Rank
	Paired t-test	43	64	51	54	27	29	22	16	306	6
	Rt-test	33	97	26	32	62	55	39	24	368	3
α=0.1	DWST	57	96	26	26	61	47	38	24	375	2
	AWST	33	64	26	32	62	55	39	27	338	5
	DST	0	30	45	18	69	68	58	67	355	4
	AST	39	159	75	68	70	68	58	33	570	1
	Tt-test	0	0	0	0	1	0	42	0	43	7

		Sample size									
Alpha level	Test statistics	10	15	20	25	30	35	40	50	Total	Rank
	Paired t-test	24	22	16	21	19	16	7	8	133	5
	Rt-test	12	12	35	19	31	33	7	16	165	1
α=0.5	DWST	12	18	17	16	31	32	7	7	140	4
	AWST	12	18	17	16	31	32	7	9	142	3
	DST	21	42	21	16	10	7	1	30	148	2
	AST	21	42	21	18	9	6	1	0	118	6
	Tt-test	0	0	1	0	1	1	0	10	13	7
		Sample size									
Alpha level	Test statistics	10	15	20	25	30	35	40	50	Total	Rank
	Paired t-test	78	127	83	76	72	75	33	24	568	4
	Rt-test	87	75	79	77	47	59	37	25	486	6
α=0.01	DWST	50	105	72	66	54	67	40	30	484	7
	AWST	71	125	70	61	62	74	47	30	540	5
	DST	56	158	81	112	114	88	83	86	778	2
	AST	56	160	90	66	112	88	83	87	742	3
	Tt-test	63	144	176	107	139	147	157	161	1094	1

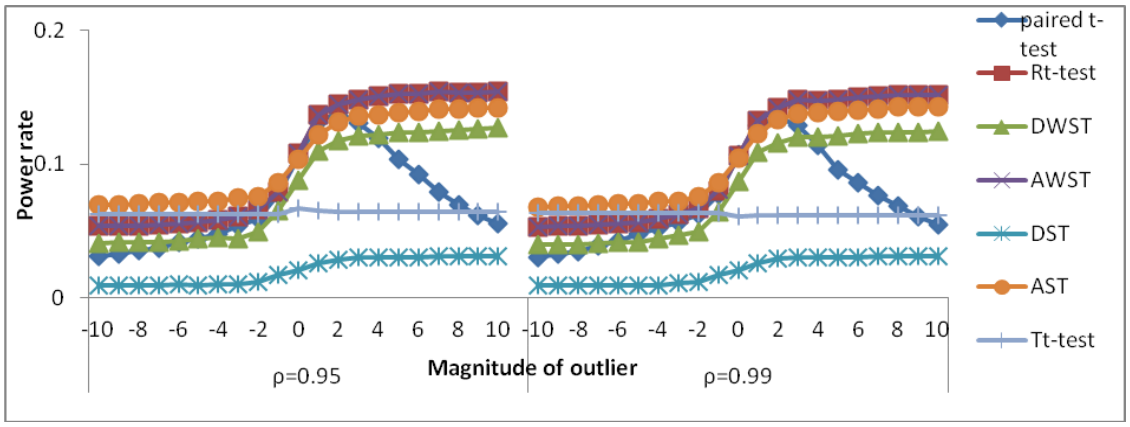
Source: Counted from Simulation results available in (Adejumo, 2016)



(a)

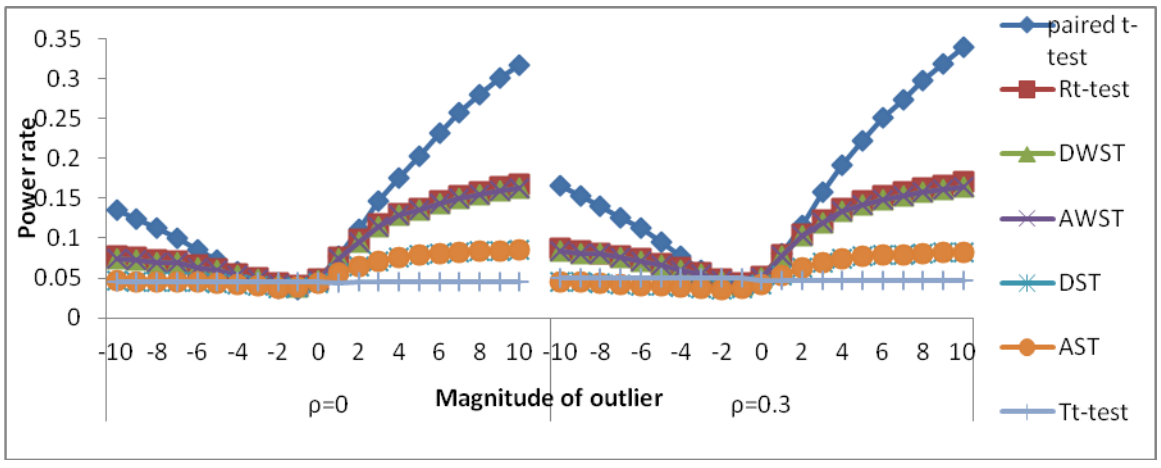


(b)

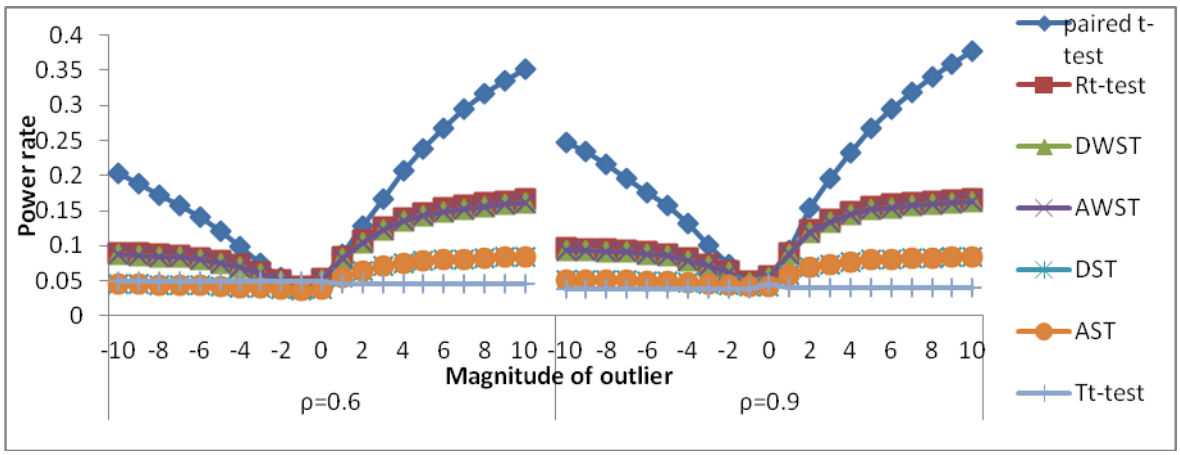


(c)

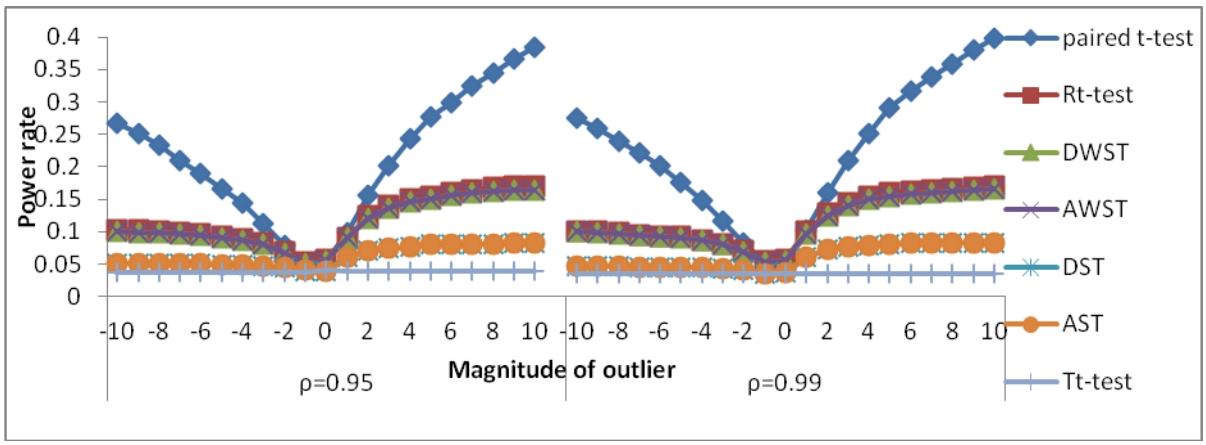
Figure 5: Graphical representation of Power Rates of test statistics when n=10 with 10% outlier at 0.1 level of significance and all levels of correlation.



(a)



(b)



(c)

Figure 6: Graphical representation of Power Rates of test statistics when n=40 with 10% outlier at 0.05 level of significance and all levels of correlation.

Table 4: Overall total number of times Power Rates Approximates True levels of significance when counted over all levels of correlation, percentage of outliers and levels of significance for paired sample problem.

	Sample size									
Test statistics	10	15	20	25	30	35	40	50	Total	Rank
Paired t-test	**145**	213	150	151	118	120	62	48	1007	6
Rt-test	132	184	140	128	140	147	83	65	1019	5
DWST	119	219	115	108	146	146	85	61	999	7
AWST	116	207	113	109	155	161	93	66	1020	4
DST	77	230	147	146	**193**	**163**	142	**183**	1281	2
AST	116	**361**	**186**	**152**	191	162	142	120	**1430**	1
Tt-test	63	144	177	107	141	148	**199**	171	1150	3

Source: Counted from Simulation results

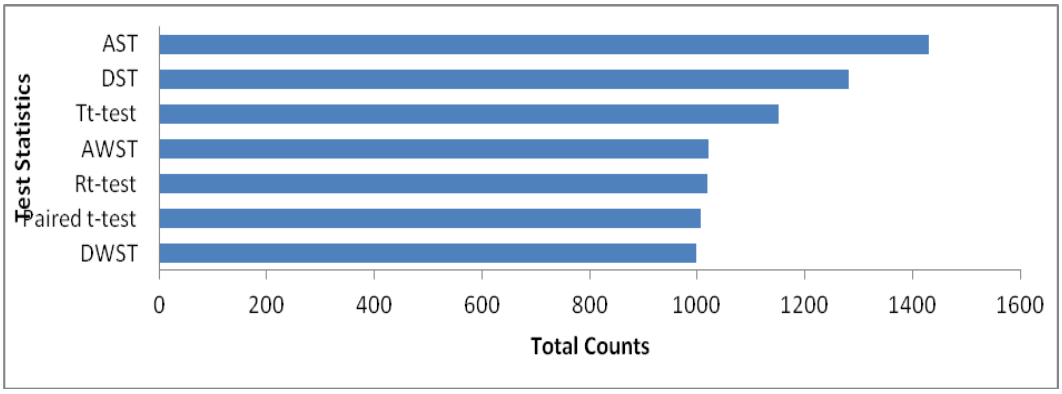


Figure 7: Bar chart showing overall total number of times Power Rates approximate true levels of significance when counted over levels of correlation, percentage of outliers and all levels of significance.

Table 5 gives the overall summary and conclusion of one and matched pairs sample inferential test statistics to outliers. They are outlined as follows:

Table 5: Summary of findings for Matched-Pairs Sample Test Statistics

α	**Matched Pairs Sample Investigation**		
	Type 1 Error	**Robustness**	**Sensitivity**
0.1	Paired t-test, AWST, Rt-test	AST	Tt-test
0.05	DWST, AWST, Paired t-test, Rt-test,	Rt-test	Tt-test
0.01	Paired t-test, DWST, AWST, Rt-test,	Tt-test	DWST, Rt-test
Overall	Paired t-test, AWST	AST,DST, Tt-test	DWST, Paired t-test

CONCLUSION

From Table 5, the following can be observed:

At 0.1 level of significance, paired t-test, AWST and Rt-test were identified to have better Type 1 error rate, the AST is very robust while Tt- test is sensitive to outlier. At 0.05 level of significance, the DWST, AWST, Paired t-test and Rt- test in this order have good Type I error rate and DST are robust while Tt- test is still sensitive. At 0.01 level of significance, the Paired t-test, DWST, AWST, and Rt-test have better Type 1 error rate, Tt- test is most robust test statistic to outlier while Rt- test and DWST are sensitive. Hence, over all levels of significance it can be concluded that the Paired t-test and AWST have good type 1 error rate, the Sign test (AST and DST) and Tt-test Statistics are robust while DWST and the paired t-test are the most statistics sensitive to outliers.

REFERENCES

Ayinde, K. (2007).Equations to Generate Normal Variates with Desired Intercorrelation Matrix" *International Journal Of Statistics And System*. Vol. 2(2), 99 – 111.

Brase, C.H. and Brase C.P. (1999). Understanding Statistics: Concepts and methods (6th ed.). Boston: Houghton Mifflin.

Conover, W.J. and Ronald, L.I. (1981): Rank transformations as a bridge between parametric and nonparametric statistics. *The American Statisticians*, 35, 125 – 129

Gauss,C. F. (1809)."Theoriamotus corporum coelestium in sectionibus conicis solem ambientium" Normal distribution-wikipedia,(2016); the free encyclopedia, <http://www.encyclopedia.com.

Gosset, W. S. (1908). The probable error of a mean. *Biometrika*-wikipedia (2016), the free encyclopedia, <http://www.encyclopedia.com.

John A., (1710). Sign test-the free encyclopedia, <http://www.encyclopedia.com (2016).

Keselman, H.J., Wilcox, Algina R.R. and Fradette, K. (2008): A comparative study of robust tests for spread: Asymmetric trimming strategies. *British Journal of Mathematical and Statistical Psychology*, 61, 235-253.

Kuranga, J. O. (2015): Type 1 Error Rate and Power comparison of some Normality test statistics. Unpublished M.Phil. Thesis, Department of Statistics, Ladoke Akintola University of Technonogy Ogbomoso, Nigeria

Siegel Sidney, (1956). Non-parametric Statistics for the behavioral sciences. New York. McGraw-Hill. pp. 75-83.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. Biometrics Bulletin 1 (6): 80-83.

Yuen, K. K. (1974). The two-sample trimmed *t* for unequal population variances. *Biometrika, 61*(1), 165-170.