Evaluation of Some Outlier Detection Methods based on Real Life Data Application

Obikee, A. C.[1] and Okoli, C. N.[2]
[1]Department of Statistics, Nnamdi Azikiwe University, Awka, Nigeria.
[2]Department of Statistics, Chukwuemeka Odumegwu Ojukwu University, Uli, Nigeria.
Email: pobikeeadaku@yahoo.com, cecilia.okoli@yahoo.com
Correspondence*: pobikeeadaku@yahoo.com

ABSTRACT
The research work evaluated some outlier detection techniques such as t-statistic (T), Modified Z-Statistic (τ), Cancer Outlier Profile Analysis (COPA), Outlier Sum-Statistic (OS), Outlier Robust T-Statistic (ORT), and the Truncated Outlier Robust T-Statistic (TORT) to verify which technique has the highest power of detecting outliers on the bases of their Rank Values, P-values, True Positives(Sensitivity), False Positives (Specificity) and False Discovery Rate (FDR) using real life data application. It was observed using the Rank Values that OS has the highest Rank Value followed by t-statistic, ORT, TORT, Z while COPA had the least rank. It was also observed using the P-values that COPA performed better than the other methods by having the highest number of False Positives (specificity) followed by OS with a better specificity (FP) and sensitivity (TP) while Z, T, ORT and TORT have no False Positive. In terms of their False Discovery Rate (FDR), the performance of OS is outstanding with a smaller FDR followed by COPA, T, ORT, TORT and Z.

Keywords: Real Data Application, P-Value, Sensitivity, Specificity, False Discovery Rate, Rank Value.

1.0 INTRODUCTION

Grubbs (1969) defined an outlier as an observation that appears to deviate markedly from other members of the sample in which it occurs. Hawkins (1980) rightly defined the concept of an outlier as "an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism". The key part of any successful data analysis is the examination, detection and proper management of outliers. It has been a difficult task on the side of researchers to identify or detect potential outliers in a given data set. Also, the argument in the literature regarding extreme or influential data points-outliers has been what to do and how to manage or handle outliers in a given data set (Barnett and Lewis, 1994; Orr *et al*, 1991). Even when outliers are identified, there is no particular tool or test in the literature that can best manage or handle significant outliers so that the result of our statistical estimates will be reliable. Hence, the intent is to compare six outlier techniques to determine which one among them that has the highest power of detecting potential outliers using real life data application. In this work, we considered the ages of women who delivered through normal (Spontaneous Vaginal) delivery and those who delivered through C/S (Caesarian Surgical). Those who delivered normally are regarded as the normal group while those who delivered through C/S are regarded as the disease group. The data were collected from the delivery register of Ebonyi State University Teaching Hospital Abakaliki Nigeria and Our Lady of Loudes Mission Hospital Ihiala, Anambra State Nigeria between the period of 2007 and 2011.

References can be made to the following authors: Dudoit *et al* (2002), Troyanskaya *et al* (2002), Tomlins etal (2005), Efron *et al* (2001),Iglewicz and Hoaglin (2010), Ghosh (2006), Lyons et al (2004), Tibshirani and Hastie (2006), Benjamini and Hochberg(1995),Wu (2007), June (2012), Fonseca(2004), MacDonald and Ghosh (2006), Jianhua (2008), Heng (2008),Ghosh (2009), Lin-An *et al* (2010), Ghosh (2010), Filmoser etal(2008), and Keita *et al* (2013)

2.0 METHOD OF ANALYSIS

This research work considered a two-group data for detecting outliers. The normal or control group data N_1 and the disease group data N_2. Let x_{ij} be the expression value for the normal group samples for $i = 1, 2, \ldots, n_1$ with the number of observations in the *jth* sample of the normal group $j = 1, 2, \ldots, p_1$ is the number of samples in the normal group. Let y_{ij} be the expression value for the disease group samples for $i = 1,2,\ldots,n_2$ and the number of observations in the *jth* sample of the disease group and $j = 1, 2, \ldots, p_2$ is the number of samples in the disease group.

The Modified Z-Statistic, COPA and OS utilize information from the disease group only while T-Statistic, ORT and TORT utilize information from both the normal group and the disease group.

Iglewicz and Hoaglin (2010) recommended the modified Z-score
They suggested that modified Z-scores with absolute value greater than 3.5 should be labeled potential outliers, that is:

$$Modified\,Z_{ij} = \left| \left[\frac{y_{ij} - Med_j}{Mad_j} \right] \times 0.6745 \right| > 3.5 \qquad i = 1, 2, \ldots, n_2 \text{ and } j = 1, 2, \ldots, p_2 \qquad (1)$$

Where
y_{ij} are the observed values in the j^{th} sample of the disease group.
Med_j is the median of the j^{th} sample of the disease group
Mad_j is median absolute deviation of the j^{th} sample of the disease group

The t-statistic for a two sample test by Dudoit *et al* (2002) is given as:

$$t = \frac{\bar{Y} - \bar{X}}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \qquad (2)$$

Where $\bar{Y}$ is the mean of the sample of the disease group.
$\bar{X}$ is the mean of the sample from the normal group.
n_1 is the sample size for the normal group sample and n_2 is the sample size for the disease group.
The denominator S_p is the pooled standard deviation for the normal and disease group.

Tomlins *et al* (2005) defined the COPA statistic as the r^{th} percentile of standardized samples of the disease group. The author recommended using the 75th percentile of standardized samples which is 0.75 of the standardized samples. The COPA statistic by Tomlins *et al* (2005) has the formula:

$$C_{ij} = q_{75}\left[\frac{\left(y_{ij} - Med_j\right)}{Mad_j} \right] \qquad (3)$$

$$i = 1, 2, \ldots, n_2 \text{ and } j = 1, \ldots, p_2$$

Where q_{75} is the 75th percentile (0.75) of the standardized samples. Med_j is the median of all values in j, and Mad_j is the median absolute deviation of all expressions in j^{th} sample. C_{ij} is an outlier if $1 \le C_{ij} \le n_2$.

According to Tibshirirani and Hastie (2006),

$$OS_{ij} = \sum_{i \in O_i} \left[\frac{\left(y_{ij} - Med_j\right)}{Mad_j} \right] \qquad i = 1, 2, 3, \ldots, n_2 \text{ and } j = 1, \ldots, p_2 \qquad (4)$$

O_i is the set of "outlier observations in the *jth* sample of the disease group" defined in the closed bound: $O_i \notin \left[Q_1 - 1.5(IQR), Q_3 + 1.5(IQR)\right]$. Where Q_1, Q_3 and IQR are the first quartile, third quartile and the interquartile range of all values in the *jth* sample respectively. n_2 is the number of observations in the *jth* samples and p_2 is the number of *jth* samples of the disease group. The statistic OS sums over the outlier set O_i of the disease group samples.

Accordingly, Wu (2007) defined the Outlier Robust *t*-statistic ORT as:

$$ORT_{ij} = \frac{\sum_{i \in_{Oi}} \left(Y_{ij} - Med_j^c\right)}{Median\left\{X_{ij} - Med_j^c \middle\| Y_{ij} - Med_j^d\right\}} \qquad i = 1, 2, \ldots, n_2 \text{ and } j = 1, \ldots, p_2 \quad (5)$$

O_i is the set of "outlier observations in the *jth* sample of the disease group" defined in the closed bound: $O_i \notin [Q_1 - 1.5(IQR), Q_3 + 1.5(IQR)]$ Where Med_i^c is the median of the normal group and Med_i^d is the median of the disease group. The statistic ORT concentrates only on the outlier set O_i.

According to June (2012), TORT is given as:

$$TORT_{ij} = \frac{\sum_{i \in_{Oi}} \left(Y_{ij} - Med_j^c\right)}{Median\left\{X_{ij} - Med_j^c \middle\| Y_{ij} - Med_j^o\right\}} \qquad i = 1, 2, \ldots, n_2 \text{ and } j = 1, \ldots, p_2 \quad (6)$$

O_i is the set of "outlier observations in the *jth* sample of the disease group" defined in the closed bound: $O_i \notin [Q_1 - 1.5(IQR), Q_3 + 1.5(IQR)]$

Med_i^c is the median of the normal group and Med_i^o is the median of the outlier set. The statistic TORT concentrates only on the outlier set in the disease group.

2.1 Normality Test: The Anderson-Darling normality test will be carried out in this analysis to test for the normality assumption on the real life data and also to check for the existence of outliers.

Anderson-Darling Normality Test

According to D'Agostino and Stephens (1986), the Anderson□Darling statistics (A^2) measures the area between the fitted line (based on chosen distribution) and the nonparametric step function (based on the plot points). The statistic is a squared distance that is weighted more heavily in the tails of the distribution. A smaller Anderson-Darling value indicates that the distribution fits the data better.

The Anderson□Darling normality test procedure is as follows:

H_0: The data follow a normal distribution.

H_A: The data do not follow a normal distribution.

Test Statistic: The Anderson□Darling test statistic is defined as

$$A^2 = -N - \left(\frac{1}{N}\right)\sum_{i=1}^{n}(2i-1)\left[InF\left(Y_i\right) + In\left(1 - F\left(Y_{N+1-i}\right)\right)\right] \quad (7)$$

Where Y_i are the ordered data. $Y_i = Y_1, Y_2, Y_3, \ldots, Y_N$

$F(Y_i) = \Phi\left(\frac{Y_i - \overline{X}}{S}\right)$ which is the cumulative distribution function of the standard normal distribution and

$F\left(Y_{N+1-i}\right)$ is the ordered cumulative distribution function of the standard normal distribution

The P-Value: Another quantitative measure for reporting the result of the Anderson□Darling normality test is the p-value.

A small p-value is an indication that the null hypothesis is false.

2.2 Bartlett test for Equality of Variance (Snedecor and Cochran, 1989)

In statistics, **Bartlett's test** is used to test if *k* samples are from populations with equal variances.

The Bartlett test is defined as:

H_0: $\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2$

H_a: $\sigma_i^2 \neq \sigma_j^2$ for at least one pair (i,j).

Test Statistic: The Bartlett test statistic is designed to test for equality of variances across groups against the alternative that variances are unequal for at least two groups.

$$T=(N-k)\ln s2p-\sum ki=1(Ni-1)\ln s2i1+(1/(3(k-1)))((\sum ki=11/(Ni-1))-1/(N-k)) \quad (8)$$

In the above, s_i^2 is the variance of the ith group, N is the total sample size, N_i is the sample size of the ith group, k is the number of groups, and s_p^2 is the pooled variance. The pooled variance is a weighted average of the group variances and is defined as:

$$s2p=\sum i=1k(Ni-1)s2i/(N-k)$$

3.0 REAL LIFE DATA APPLICATION

Here, analysis was carried out on the real life data from both hospitals mentioned above to compare the performance of the various outlier methods. We have two groups of data for the analysis. The first group data are the ages of women who delivered through normal delivery regarded as the normal or control group N_1. The second group data are the ages of women who delivered through Caesarian Section C/S regarded as the disease group N_2. Sets of eight (8) samples were randomly drawn with a sample size of 27 each from the hospitals considered for both the normal group X_j and the disease group Y_j

Table 1: Observed Ages for the Normal Group (Women who have a normal delivery)

X_1	X_2	X_3	X_4	X_5
20	45	24	37	31
23	35	18	42	20
30	34	30	30	22
19	28	33	30	18
26	30	40	36	25
26	36	21	30	19
23	24	20	35	26
28	21	23	26	30
23	29	31	30	29
26	40	20	24	20
26	31	30	27	33
20	25	34	30	21
30	36	37	34	25
29	28	21	26	35
20	24	23	35	18
35	30	49	26	31
40	32	23	31	39
23	38	28	26	23
27	27	34	23	26
24	38	20	37	18
30	20	30	27	27
41	26	23	46	40

Where, $x_1, x_2, \ldots, x_8$ are the observed ages for the various sets for the normal group.

Table 4.2: Observed Ages for the Disease group (Women who delivered through Caesarian Section)

y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8
20	27	27	32	40	17	22	31
40	30	38	31	31	21	18	19
40	31	23	18	20	19	25	25
35	27	31	30	18	25	47	18
16	39	18	27	29	45	27	23
17	30	31	49	36	30	31	29
45	17	31	21	20	22	19	33
40	35	30	30	19	15	25	39
45	18	25	31	25	29	18	40
20	27	35	30	32	31	23	16
40	31	30	35	15	35	29	45
13	30	27	30	33	42	33	27
30	25	29	16	22	18	39	29
40	35	20	23	20	20	40	20
16	16	23	28	24	16	16	36
45	30	35	36	40	19	45	22
45	40	30	20	31	26	27	43
31	35	33	25	45	33	29	37
24	16	31	19	19	29	20	31
31	20	30	27	23	35	36	49
34	40	37	35	28	21	22	30
24	45	31	39	33	25	43	16
23	36	26	30	21	49	37	25
26	23	39	20	41	24	31	35
30	26	36	29	27	29	49	28
21	40	38	37	26	18	30	26
45	35	26	45	24	37	16	20

Here, $y_1, y_2, y_3, \ldots, y_8$ are the observed ages for the various sets for the disease group

Underlying Distribution Assumptions of the Six Outlier Methods

1. Normality: It is assumed that samples are drawn from the normal distribution.
2. Independence: It is expected that samples drawn are independent of each other.
3. Homogeneity: Equality of variance.

The above assumptions must be satisfied before the six outlier detection methods shall be applied on both the real life data and the simulated data.

4.0 TEST ANALYSIS

4.1 ANDERSON-DARLING'S TEST FOR NORMALITY OF THE FIRST SAMPLE FROM THE DISEASE GROUP

Anderson-Darling normality test was carried out on the first sample from the disease group $n_1$with 27 observations. Here are the samples: 20, 40, 40, 35, 16, 17, 45, 40, 45, 20, 40, 13, 30, 40, 16, 45, 45, 31, 24, 31, 34, 24, 23, 26, 30, 21 and 45.

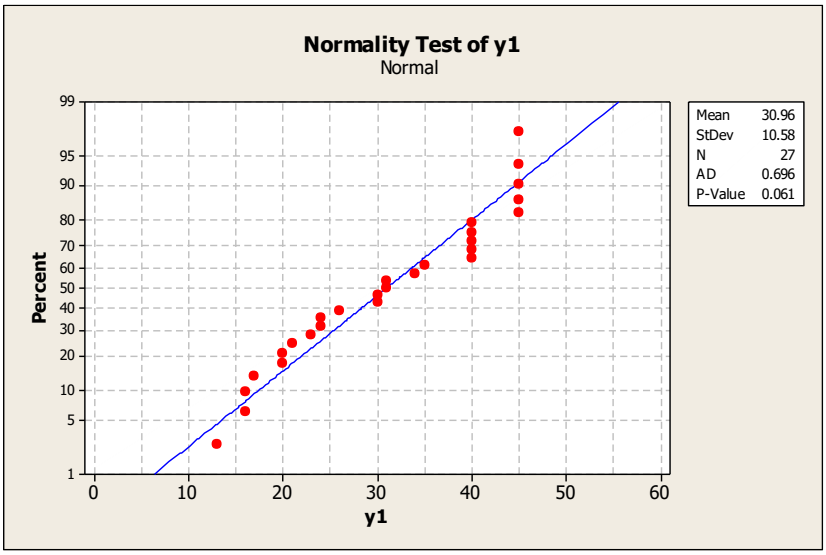

Figure 1: Normality test of y_1.

Figure 1 shows the Anderson-Darling normality test for y_1. From the plot, y_1 has the mean of 30.96, StDev =10.58, N = 27, Anderson Darling value AD = 0.696 and P-Value=0.061.

Test hypothesis

H_0: The data follow a normal distribution.
H_A: The data do not follow a normal distribution.
Since the P-Value = 0.061 > α = 0.05, we accept the null hypothesis that the data follows a normal distribution. Also the extreme points and deviants in the plot indicate the existence of outliers.

Table 5: Anderson-Darling's Test for Normality of Samples from the Normal Group

Sample	AD Value	P-Value	Ho
x1	0.601	0.107	Accepted
x2	0.157	0.947	Accepted
x3	0.509	0.181	Accepted
x4	0.557	0.136	Accepted
x5	0.412	0.317	Accepted
x6	0.366	0.41	Accepted
x7	0.594	0.111	Accepted
x8	0.236	0.768	Accepted

Table 5 shows the Anderson's Value and the P-Value of all the samples from the normal group. H_o being accepted implies that the data follow a normal distribution.

Table 6: Anderson-Darling's Test for Normality of Samples from the Disease Group

Sample	AD Value	P-Value	Ho
y1	0.696	0.061	Accepted
y2	0.351	0.443	Accepted
y3	0.339	0.475	Accepted
y4	0.392	0.355	Accepted
y5	0.49	0.202	Accepted
y6	0.551	0.141	Accepted
y7	0.389	0.36	Accepted
y8	0.227	0.796	Accepted

Table 6 shows the Anderson's Value and the P-Value of all the samples from the disease group. H_o being accepted implies that the data follow a normal distribution.

4.2 BARTLETT'S TEST FOR EQUALITY OF VARIANCE FOR THE NORMAL AND THE DISEASE GROUP SAMPLES

Here a homogeneity test was carried out on the first sample from the disease group y_1 with 27 observations- 20, 40, 40, 35, 16, 17, 45, 40, 45, 20, 40, 13, 30, 40, 16, 45, 45, 31, 24, 31, 34, 24, 23, 26, 30, 21, 45 and the first sample from the normal group with 27 observations- 20, 23, 30, 19, 26, 26, 23, 28, 23, 26, 26, 20, 30, 29, 20, 35, 40, 23, 27, 24, 30, 41, 30, 29, 21, 31, 27

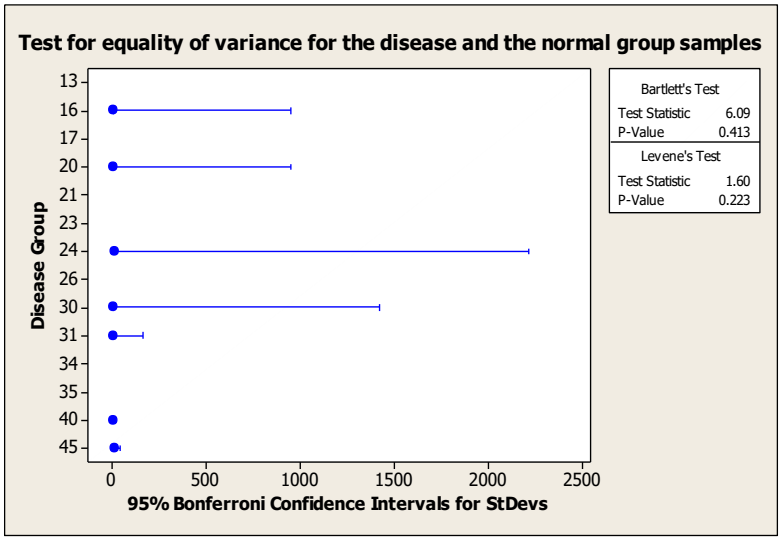

Figure 2: Homogeneity test

H_o: The samples have equal variances
H_1: The samples do not have equal variances

Decision: Since the p-values 0.413 and 0.223 are greater than 0.05, we accept the null hypothesis that the samples have equal variances.

Table7: Summary of Bartlett's Test for Equality of Variance for the Normal and the Disease Group Samples

Sample	Battlet's Value	P-Value	Ho
x1 and y1	6.09	0.413	Accepted
x2 and y2	4.19	0.757	Accepted
x3 and y3	2.99	0.559	Accepted
x4 and y4	4.62	0.329	Accepted
x5 and y5	7.86	0.164	Accepted
x6 and y6	6.28	0.392	Accepted
x7 and y7	10.81	0.212	Accepted
x8 and y8	4.91	0.427	Accepted

Table 7 above summarizes the Bartlett's Test for Equality of Variance for the Normal and the Disease Group Samples. Ho is accepted in all the samples signifying that both samples have equal variances.

*Since the individual observed values are selected independent of each other and both the normality and the homogeneity assumptions are satisfied, hence, the six outlier detection techniques can be applied on the above data for analysis.

Table 8: Summary of calculated values by the outlier methods

Samples	Modified Z-Statistic	T-Statistic	COPA	OS	ORT	TORT
Y1	0	-0.44	-0.916665	0	0	0
y2	0	-0.3	-1.95	0	0	0
y3	0	0.9	-2.25	0	0	0
y4	0	-0.46	-1.8	3.8	0.95	0
Y5	0	0.64	-0.9999975	0	0	0
Y6	0	-0.76	-1.2500025	4	0.875	0
Y7	0	1.21	-1.2	0	0	0
Y8	0	0.26	-1.1785725	0	0	0

Table 9: Rank values obtained by the outlier methods.

Modified Z-Statistic	T-Statistic	COPA	OS	ORT	TORT
26.5	11	8	26.5	26.5	26.5
26.5	12	2	26.5	26.5	26.5
26.5	44	1	26.5	26.5	26.5
26.5	10	3	47	45	26.5
26.5	42	7	26.5	26.5	26.5
26.5	9	4	48	43	26.5
26.5	46	5	26.5	26.5	26.5
26.5	41	6	26.5	26.5	26.5

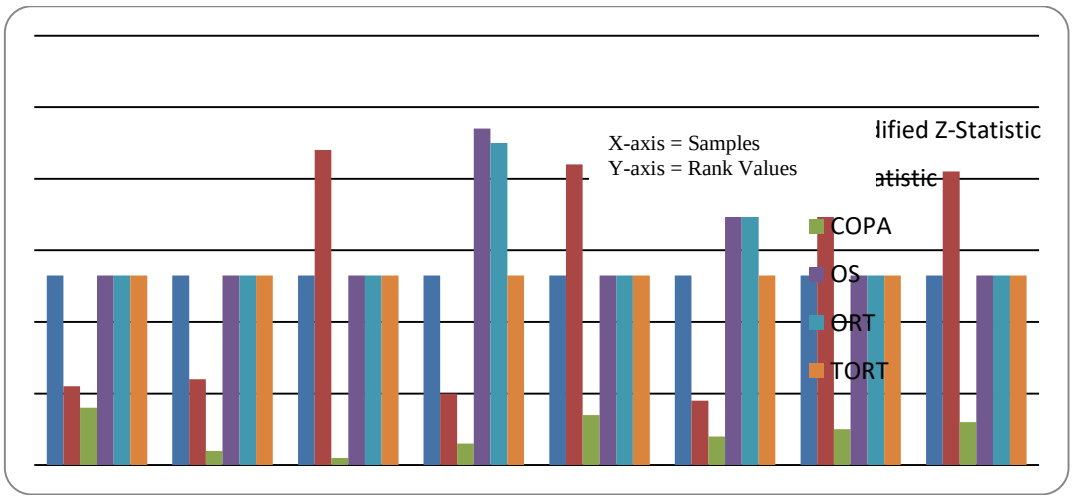

Figure 3: A Multiple bar chart of the outlier methods by their Rank Values

Table 10: Calculated P-Values

	Modified Z	t-statistic	COPA	OS	ORT	TORT
	1	0.6618	0.2216	1	1	1
	1	0.7654	0.0093	1	1	1
	1	0.3723	0.0027	1	1	1
	1	0.6474	0.0164	0.0001	0.3465	1
	1	0.5250	0.1824	1	1	1
	1	0.4507	0.0956	0.0001	0.3856	1
	1	0.2318	0.1096	1	1	1
	1	0.7959	0.1161	1	1	1
Maximum P-Value	1	0.7959	0.2216	1	1	1
Minimum P-Value	1	0.2318	0.0027	0.0001	0.347	1
True Positives	8	8	5	6	8	8
False Positives	0	0	3	2	0	0
Mean P-Value	1	0.5563	0.0942	0.714	0.819	1

Table 11: FDR of the outlier methods

	Modified Z-Statistic	T-Statistic	COPA	OS	ORT	TORT
	239.245	381.437	175.618	239.245	239.245	239.245
	239.245	404.386	29.481	239.245	239.245	239.245
	239.245	53.645	17.118	239.245	239.245	239.245
	239.245	410.452	34.659	0.013	48.818	239.245
	239.245	79.25	165.202	239.245	239.245	239.245
	239.245	317.493	151.526	0.013	56.854	239.245
	239.245	31.948	138.973	239.245	239.245	239.245
	239.245	123.073	122.679	239.245	239.245	239.245
Maximum FDR	239.25	410.5	175.600	239.2	239.2	239.25
Minimum FDR	239.25	31.9	17.100	0.013	48.8	239.25
Mean FDR	239.25	225.2	104.400	179.4	192.6	239.25

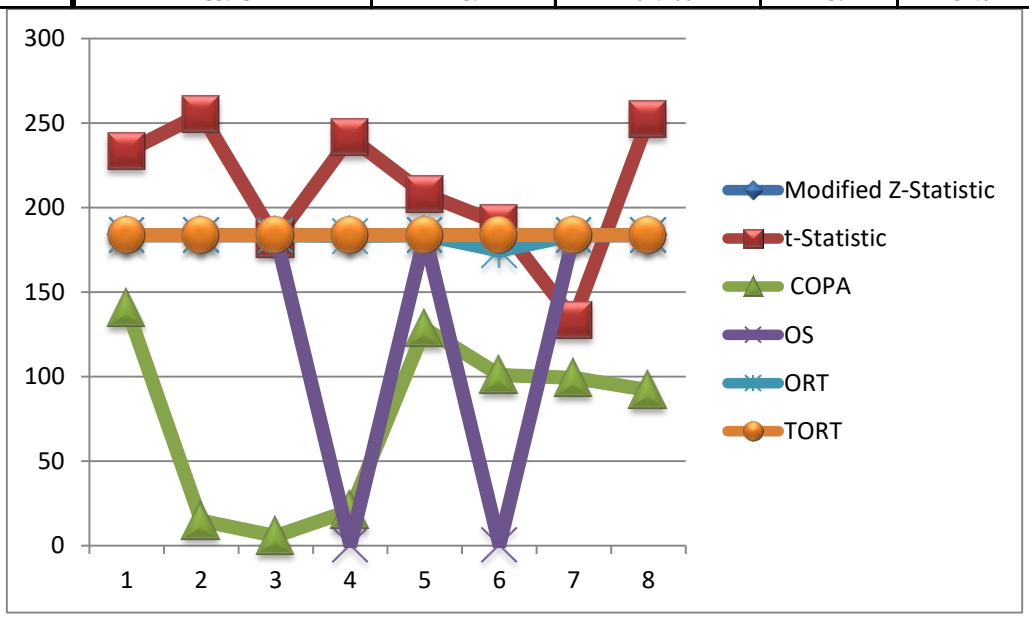

Figure 4: A Plot of the FDRs of the different outlier methods

4.3 Summary

Real data application has been employed to compare the performance of the various outlier methods and to evaluate which of the methods has a higher power of detecting outliers in terms of their Rank Values, P-Values, True Positives (Sensitivity), False Positives (Specificity), a smaller False Discovery Rate (FDR). We observed from Table 9 that OS performed better with Rank Value 48 followed by T-Statistic, ORT, TORT, Z and COPA with rank values 46, 45, 26.5, 26.5 and 8 in that order. The differences in ranking between the outlier methods were also plotted in Figure 3 as

multiple bar charts. The bar charts showed that the performance of OS is better than the other outlier methods as shown by the heavy long bar in the bar followed by T, ORT, TORT, Z and COPA. For the P-Values, we observed that OS, ORT, Z and TORT has a maximum P-Value of (1) while T= 0.7959 and COPA= 0.2216Z and TORT has a minimum P-Value of (1), T=0.2318, COPA= 0.0027, OS= 0.0001 and ORT=0.347. Z, T, ORT and TORT have a true positive (Sensitivity) of eight (8) while COPA= 5 and OS= 6. Z, T, ORT and TORT have a false positive (Specificity) of zero (0) while COPA =3 and OS = 2. Z and TORT has a mean P-Value of one (1), T=0.5563, COPA= 0.0942 OS= 0.714 and ORT= 0.8

We also observed that T-statistic, Z, TORT, OS, ORT and COPA has a maximum FDR of 410.5, 239.25, 239.25, 239.2, 239.2, 175.600 and a minimum FDR of 31.9, 239.25, 0.013, 48.8 and 17.100 in that order. Z and TORT has equal mean FDR of 239.25 while T, ORT, OS and COPA has mean FDR of 225.2, 192.6, 179.4 and 104.40

The different values of the FDRs were also plotted in figure 4 in a line plot and it was observed that FDRs for OS clustered at the base of the line plot indicating that OS has smaller FDR followed by COPA and others.

CONCLUSION

The performance of the various outlier methods- Z, T, COPA, OS, ORT and TORT has been statistically studied using real data to identify which method has the highest power of detecting outliers in terms of their Ranks, P-Values, True Positives(Sensitivity), False Positives(Specificity), False Discovery Rate FDR. We observed that OS has the highest Rank Value followed by T-Statistic, ORT, TORT, Z and COPA in that order. We observed using the P-values that COPA performed better than the other methods by having the highest number of False Positives(Specificity) followed by OS, Z, T, ORT while TORT have no False Positive. In terms of their False Discovery Rate (FDR), OS also performs better with a smaller FDR followed by COPA, t, ORT, TORT and Z.

Contribution to Knowledge

The Modified – Z, OS and COPA has a Normal Distribution while T, ORT and TORT has a T- Distribution. The Underlying Distribution Assumptions of the various outlier methods which includes: Normality, Independence and Equality of Variance were first checked before the six outlier detection methods were applied to the real life data. If these distribution assumptions are not satisfied, the result of the statistical estimates can be biased and unreliable.

REFERENCES

Aggarwal, C. C. (2005); "*On Abnormality Detection in Spuriously Populated Data Streams"*, SIAM Conference on Data Mining. Kluwer Academic Publishers Boston/Dordrech/London.

Barnett, V. and Lewis, T. (1994); *"Outliers in Statistical Data"*, John Wiley & Sons, 3rd Edition. Kluwer Academic Publishers Boston/Dordrecht/London.

Benjamini, Y. and Hochberg Y. (1995); ' *Controlling the false discovery rate: A practical and powerful approach to multiple testing",* Journal of Research Stat. Soc. B.57:289–300.

D'Agostino, R. B. and Stephens, M. A. (1986); "*Goodness-of-Fit Techniques"*, Eds. Marcel Dekker.

Dudoit, S., Yang Y., Callow M. and Speed T. (2002); *"Statistical Methods for Identifying Differentially Expressed Genes in Replicated DNA Microarray Experiments,"* Statistica Sinica, Vol. 12, No. 1, 111- 139.

Ghosh, D. (2009); "*Genomic outlier profile analysis: mixture models, null hypotheses, and nonparametric estimation''* Biostatistics 10 (1): 60–69.

Ghosh, D. and Li, Y. (2014); *"Unsupervised Outlier Profile Analysis"*, Cancer Informatics 2014:13(S4) 25–33 doi: 10.4137/CIN.S13969.

Grubbs, F. E. (1969); "*Procedures for detecting outlying observations in samples".* Technometrics 11, 1–21.

Hawkins D. (1980); "*Identification of Outliers",* Chapman and Hall, Kluwer Academic Publishers Boston/Dordrecht/London

Heng Lian (2008); "*MOST: detecting cancer differential gene expression" Biostat 9 (3): 411-418. doi: 10.1093/biostatistics/kxm042*.

Iglewicz B. and Hoaglin D.C. (2010); "*Detection of outliers"* Engineering Statistical Handbook *1.3.5.17.* Database Systems Group.

Jianhua Hu (2008); '*Cancer outlier detection based on likelihood ratio test'* Bioinformatics. 24(19): 2193–2199. doi: 10.1093/bioinformatics/btn372

June Luo (2012); '*Truncated Outlier Roburst T-Statistic for outlier detection'.* Open Journal of Statistics, 120-123 doi:10.4236/ojs.2012.21013.

Keita Mori,Tomonori Oura, Hisashi Noma and Shigeyuki Matsui (2013); "*Cancer Outlier Analysis Based on Mixture Modeling of Gene Expression Data"* Computing and Math Methods Med. 693901. doi: 10.1155/2013/693901.

Lin-An Chen1, Dung-Tsa Chen, Wenyaw Chan (2010); *"The distribution-based p-value for the outlier sum in differential gene expression analysis"* Biometrika 97 (1): 246-253.
doi: 10.1093/biomet/asp075

MacDonald, J. W. and Ghosh, D. (2006); '*Copa-cancer outlier profile analysis''.* Bioinformatics. 22:2950–2951.

Snedecor, G. W. and Cochran, W. G. (1989), *Statistical Methods*, Eighth Edition, Iowa State University Press.

Tibshirani, R. and Hastie, R. (2006); *"Outlier Sums for Differen-tial Gene Expression Analysis"*, Biostatistics, Vol. 8, No. 1, 2-8. doi:10.1093/biostatistics/kxl005

Orr, J. M., Sackett, P. R., and Dubois, C. L. Z. (1991); "*Outlier detection and treatment in I/O Psychology"*. A survey of researcher beliefs and an empirical illustration, *Personnel Psychology*, *44*, 473-486.

Tomlins, S., Rhodes, D., Perner, S., Dhanasekaran, S., Mehra, R., Sun, X, Varambally, S., Cao, X.,
Tchinda, J. and Kuefer, R. (2005); *"Recurrent Fusion of Tmprss2 and ETS Transcription Factor Genes in Prostate Cancer,"* Science, Vol. 310, No. 5748, 644- 648. doi:10.1126/science.1117679

Wu, B. (2007); "*Cancer Outlier Differential Gene Expression Detection"*, Biostatistics, Vol.8, No. 3, 566-575. Doi:10.1093/biostatistics/kxl029.