ISSN 2756-4444
E-ISSN 2756-4452

EFFICIENCY OF ROBUST ESTIMATORS AGAINST THE SAMPLE MEAN UNDER OUTLIER CONTAMINATION: A COMPARATIVE STUDY

BY

[1]S.A. Ibrahim & [2]K.O. Badmus
[1]Department of Physical Sciences, Al-Hikmah University, Ilorin, Nigeria
[2]Department of Statistics, Federal Polytechnic, Offa, Kwara State, Nigeria
Corresponding Author: adeshinas2010@alhikmah.edu.ng

Abstract
In this study, the efficiency of some selected robust estimators such as the sample mean, sample median, 10% trimmed mean and 10% winsorized mean are compared in the presence of outliers through simulation study. Data used for this study was generated from R programming language under different distributions assumptions such as normal, laplace and mixture of normal distribution with 10% identical outliers, in order to evaluate the robustness of estimator under study. Mean square error (MSE) was used as a performance metric for each estimator. It was found that the sample mean has the best performance in the presence of clean data (i.e. normal distribution). In the study of outliers, it was found that the sample median has the best performance when data was generated from the Laplace distribution as well as normal distribution with 10% identical outliers. In conclusion, the robust estimators are more efficient and an appropriate alternative to the sample mean in the presence of outliers
Keywords: Outliers, Population parameter mean, Sample mean, Robust estimators, MSE

Introduction

Outliers are extreme observations that deviate markedly from the majority of data points and can significantly distort statistical analysis and inference (Alma, 2011; David & Vikki, 2008), among others. In classical statistical practice, the sample mean is widely used as an estimator of the population parameter mean μ due to its desirable properties under normality assumptions. However, the presence of outliers can severely affect the efficiency and reliability of the sample mean, often leading to biased or misleading estimates. Consequently, robust estimators have been developed to reduce the influence of extreme observations and provide more reliable estimates of population parameters. Also, a robust estimation procedure provides protection against outlier data value as well as minimizing the effect of outliers on estimate of population parameter. Al-Athari & Al-Amleh (2016) and Montgomery, Peck & Vining (2012) note that the properties of efficiency and breakdown points are used to define robust techniques performance in a theoretical sense. Besides, Montgomery, Peck & Vining, (2012) argue that robust estimators are those estimators that are more efficient in the presence of outliers and their performances are similar to sample mean,$\bar{x}$ when data follow normal distribution.

Previous studies have investigated the performance of various robust estimators under outlier contamination. For instance, Ibrahim, Ibrahim, Jaiyeola, and Oni (2017) conducted a comparative study of some univariate estimators in the presence of outliers and demonstrated that certain robust estimators perform better than the classical sample mean when the data contain extreme observations. Their study provided important empirical evidence on the relative efficiency of some commonly used estimators under contaminated data conditions. Despite the valuable contributions of Ibrahim et al. (2017), their study was limited to a specific set of univariate estimators and did not extensively explore additional robust estimation techniques or broader comparative frameworks. Therefore, there remains a need for further investigation into the performance of robust estimators relative to the sample mean under

VOL. 5 (2), DECEMBER, 2025
ISSN 2756-4444
E-ISSN 2756-4452

varying levels of outlier contamination. Building on the foundation established by Ibrahim *et al*. (2017), therefore, this study aims to assess the efficiency of some selected robust estimators against the sample mean in the presence of outliers, using mean square error (MSE) as a performance metric and boxplots to examine the behaviour and stability of the estimators under contaminated data. All analyses were done using R programming language.

Methodology

A brief description of the theoretical background of population mean, sample mean, sample median, trimmed and winsorized mean are described as follow:

Population mean

Consider a population with parameter mean μ (1)

Thus the population mean is defined as $\mu = \frac{\sum x}{N}$ (2)

and, population variance $\sigma^2 = \frac{\sum(x-\mu)^2}{N}$ (3)

where X represents the individual observations values in the population variable, and N is the total number of observations values in the population of equation (2) and (3) respectively.

In estimating population meanμ, the estimate of $\hat{\mu}$ is the sample mean$\bar{x}$.

2.2 Sample mean

The sample mean $\bar{X}$ is given as follow:

$$\bar{x} = \frac{\sum x}{n} \tag{4}$$

and the sample variance S^2 the estimate of $\hat{\sigma}^2$ is given by

$$S^2 = \frac{\sum(x-\bar{x})^2}{n-1} \tag{5}$$

where x represents the individual observation values in the sample, and n is the total number of observation values in the sample of equation (4) and (5) respectively.

If a sample observation follows normal distribution, sample mean$\bar{x}$ has least variance, that is, sample mean is more efficient than any other estimators within the class of unbiased estimators. However, in a situation whereby the distribution of data is non normal or follow heavy tail distribution like Laplace distribution, sample mean is inefficient, that is, it posses large variance, consequently lead to poor performance. Oyejola (2014) noted that alternative estimations procedures called robust estimations have been suggested in literature that less sensitive to outlier as well as minimizing the effect of outlier are desire. The robust estimators includes sample median, trimmed mean, winsorize mean etc.

Sample median ($\tilde{x}$)

Sample median is the middle value, once the observations are ordered from smallest to largest, for $x_1 \leq x_2 \leq \cdots \leq x_n$. Thus, the sample median $\tilde{x}$ is given as follow.

$$\tilde{x} = \frac{(n+1)^{th}}{2} \quad \text{if observations are odd number} \quad (6)$$

$$\tilde{x} = \left(\left(\frac{n}{2}\right)^{th} + \left(\frac{n+2}{2}\right)^{th} \right)/2 \quad \text{if observations are even number} \quad (7)$$

(Warrant & Fred, 1997)

Trimmed mean ($\bar{X}_i$)

A trimmed mean is a compromise between sample mean $\bar{x}$ and sample median $\tilde{x}$. A 10% trimmed mean, for example, would be computed by eliminating the smallest 10% and the largest 10% of the sample data and then averaging what remains. Thus, a α trimmed mean is given as:

$$\bar{X}_i = \frac{1}{n-2g}(X_{g+1}+X_{g+2}+\ldots+X_{n-2g}) \quad \text{(Oyejola, 2014)} \quad (8)$$

Where $0 \leq \alpha \leq 0.5$, $g = \alpha n$; g is grounded up to nearest integer, and X_i are ordered values. The value of $\alpha = 0.2$ is often used because under normality the trimmed mean performed as the mean. When $= 0.5$, the trimmed mean is the median.

Winsorized mean:

The winsorized mean is similar to the trimmed mean, except that rather than deleting the extreme values, they are set equal to the next largest (or smallest) value. Rutgers, (n.d) noted that a α −Windsorized mean replace a proportion

of α from both ends of the data set by the next closest observation and then take the mean.

Simulation study

Following Ibrahim, Ibrahim, Jaiyeola, & Oni (2017) a simulation study was carried out, whereby the sample size n =[10, 30, 100], and population parameter mean $\mu = 9$ and $\sigma = 1$ were set arbitrarily. Thus, data set was generated through R console version 3.3.3 under different distributions assumptions, in order to evaluate the robustness of estimators under study. The assumptions are as follow:

Assumption 1: $X \sim N(9, 1)$ (i.e. normal distribution) (9)

Assumption II: $X \sim LA(9, 1)$ (i.e. Laplace distribution) (10)

Assumption III: $X \sim N(9, 1)$ with 10% identical outliers X=45 (where the first 10% observations at extreme end equal 45). (11)

Thus, effect of outliers on the dataset were investigated based on the following estimators:

- Sample mean;
- Sample median;
- 10% trimmed mean; and
- 10% winsorized mean.

For comparison purposes, MSE was used as a performance metric for each estimator, and boxplots for parameter estimate visualization. The formula used for MSE was given in equation (12).

$$\text{MSE} = \frac{1}{R}\sum(\hat{\mu}_i - \mu)^2 \qquad i = 1,2,\dots,R \qquad (12)$$

where $\hat{\mu}_i$ are the estimates for various estimators used in this study, μ is the true population parameters mean and R is called the number of iterations, where R is set at 1000.

Results

A 1000 repeated random samples of size n = [10, 30, 100] were simulated through R code from normal, Laplace and normal distributions with 10% identical outliers$X = 45$. Thus, the sample mean, sample median, 10% trimmed mean and 10% winsorized mean were calculated for each sample and the results were analysed. The summary statistic obtained for each estimator under study is presented in Table 1 through 3, while the box plot was used to visualized and compared estimates of population parameter computed from each estimator to true population parameter mean, at different sample size n are as presented in Figure 1 through 3.

Table 1: Shows the parameter estimates, MSE and 95% confidence interval when data was generated from normal distribution for different sample size n.

Sample size	Estimators	Estimates$\hat{\mu}$	MSE	95% Confidence Interval
	sample mean	9.0061	0.0918	8.4121, 9.6002
n =10	sample median	9.0161	0.1279	8.3155, 9.7068
	10%trimmed mean	9.0083	0.0967	8.3986, 9.6179
	10%winsorized mean	9.0071	0.0956	8.401, 9.6131
	sample mean	9.0008	0.0320	8.6499, 9.3517
n =30	sample median	9.0051	0.0496	8.5685, 9.4374
	10%trimmed mean	8.9998	0.0337	8.6399, 9.3597
	10%winsorized mean	8.9992	0.0328	8.6442, 9.3542
	sample mean	9.0029	0.0101	8.806, 9.1998
n =100	sample median	9.0038	0.0153	8.7615, 9.2452
	10%trimmed mean	9.0021	0.0106	8.8005, 9.2038
	10%winsor ized mean	9.0017	0.0103	8.8025, 9.2010

VOL. 5 (2), DECEMBER, 2025
ISSN 2756-4444
E-ISSN 2756-4452

Table 2: Shows the parameter estimates, MSE and 95% confidence interval when data was generated from laplace distribution for different sample size n.

Sample size	Estimators	Estimates$\hat{\mu}$	MSE	95% 95% Confidence Interval
	sample mean	9.0002	0.1920	8.1408, 9.8595
n =10	sample median	8.9948	0.1387	8.2645, 9.7306
	10%trimmed mean	8.9999	0.1546	8.2289, 9.7708
	10%winsorized mean	9.0001	0.1733	8.1838, 9.8164
	sample mean	9.0137	0.0642	8.5174, 9.5100
n =30	sample median	9.0105	0.0399	8.6196, 9.4046
	10%trimmed mean	9.0118	0.0481	8.5822, 9.4414
	10%winsorized mean	9.0121	0.0575	8.5426, 9.4816
	sample mean	9.0022	0.0211	8.7175, 9.2869
n =100	sample median	9.0020	0.0119	8.7882, 9.2161
	10%trimmed mean	9.0015	0.0152	8.7597, 9.2433
	10%winsorized mean	9.0019	0.0184	8.7358, 9.2679

Table 3: Shows the parameter estimates, MSE and 95% confidence interval when data was generated from mixture of normal with 10% identical outliers x=45 for different sample size n.

Sample size	Estimators	Estimates$\hat{\mu}$	MSE	95% Confidence Interval
	sample mean	12.6014	13.0554	12.0287, 13.1742
n =10	sample median	9.1501	0.1629	8.4155, 13.3361
	10%trimmed mean	9.1872	0.1455	8.5354, 9.8390
	10%winsorized mean	9.5426	0.4034	8.8952, 10.1900
	sample mean	12.5983	12.9769	12.2639, 12.9327
n =30	sample median	9.1395	0.0742	8.6807, 13.0571
	10%trimmed mean	9.2000	0.0771	8.8220, 9.5779
	10%winsorized mean	9.5882	0.3841	9.2051, 9.9712
	sample mean	12.6022	12.9851	12.4154, 12.7891
n =100	sample median	9.1453	0.0380	8.8906, 12.8568
	10%trimmed mean	9.2124	0.0568	9.0009, 9.4240
	10%winsor ized mean	9.6345	0.4156	9.4104, 9.8585

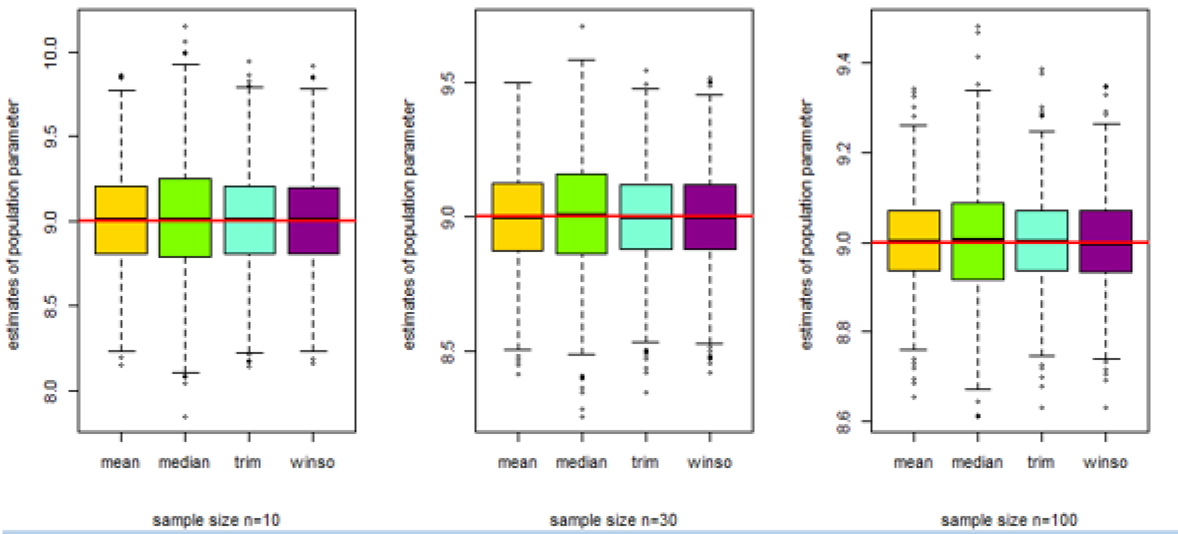

Figure 1: A comparison of behaviour of estimators under study to true population parameter mean $\mu = 9$ when data follows normal distribution at different sample size n =[10, 30, 100].

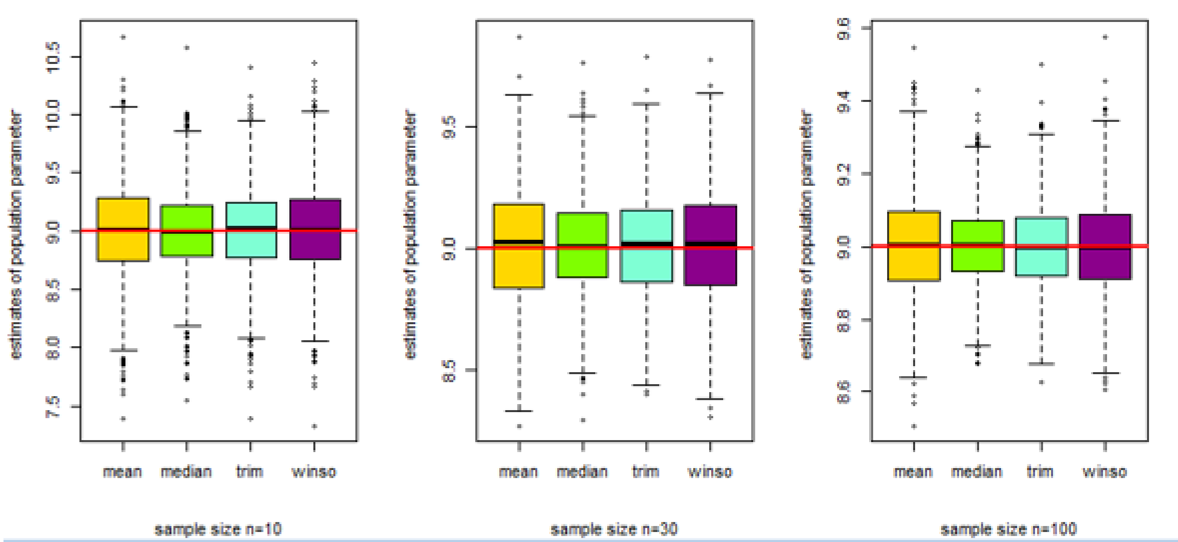

Figure 2: A comparison of behaviour of estimators under study to true population parameter mean $\mu = 9$ when data follows laplace distribution at different sample size n =[10, 30, 100].

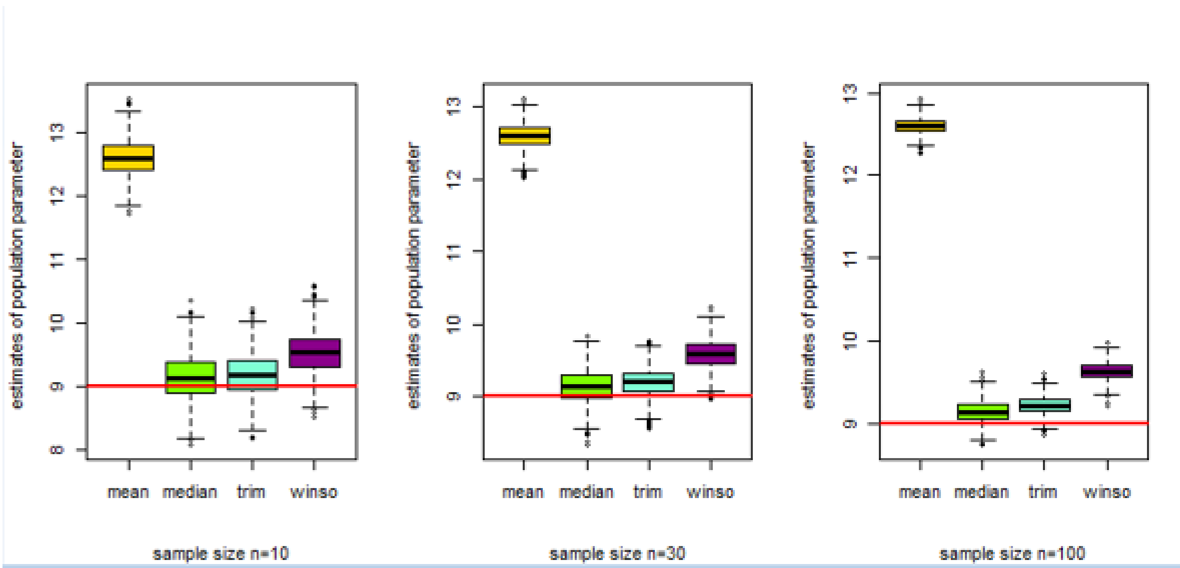

Figure 3: A comparison of behaviour of estimators under study to true population parameter mean $\mu = 9$ when data follows mixture of normal with 10% identical outliers $X = 45$ at different sample size $n = [10, 30, 100]$.

Discussions

In the study of clean data (i.e. normal distribution), as expected, it was found that the sample mean, $\bar{x}$ has the least MSE for all sample sizes n considered, and therefore is considered as the best estimator, followed by 10% winsorized mean and 10% trimmed mean, as confirmed from Table 1. In the study of outliers, when data was generated from Laplace distributions, it was found that the sample mean, $\bar{x}$ has the largest MSE for all sample sizes n considered, while the sample median has the least MSE, for all sample sizes n considered. The MSE of the sample mean, $\bar{x}$ increases as sample size n increases, as confirmed from Table 2. Also, when data was generated from mixture of normal distribution with 10% identical outliers. It was found also that the sample mean, $\bar{x}$ has the highest MSE, while 10% trimmed mean has the least MSE for sample size $n = 10$. Furthermore, the sample median has the least MSE for sample size $n = [30, 100]$ respectively, as confirmed from Table 3. However, when estimates of the sample mean $\bar{x}$ obtained from normal distribution was used to construct 95% confidence interval, the interval truly contains the population parameter mean$\mu = 9$, as confirm from Table 1. Similarly, robust estimators also contain the true population parameter mean, $\mu = 9$, as confirmed from Table 1 through 3, except 10% winsorize at $n = [30, 100]$ in Table 3.

Box plots were used to compare the estimates of each estimator under study simultaneously with true population parameter mean, $\mu = 9$. When estimators under study were applied to normal distribution dataset, it was observed that each estimator lies on true population parameter mean$\mu = 9$, for all sample sizes n, with equally spread, except the sample median, as confirmed from Figure 1. In the study of outliers, when estimators were applied to data generated from Laplace, it was observed also, that each estimator approximately lies on the true population parameter value $\mu = 9$ with highly spread from the sample mean, as confirmed from Figure 2. Moreover, when estimators under study were applied to data generated from mixture of normal distribution with 10% identical outliers, it was observed that the sample mean,$\bar{x}$ and 10% winsorize mean were far away from the true population parameter mean$\mu = 9$, especially the sample mean, as confirmed from Figure 3. This study supports the literature that a single outlier can cause the breakdown of sample mean, $\bar{x}$ (Alima, 2011; Draper& Smith, 1998; Hogg, 1979; Montgomery, Peck & Vining, 2012), among others. Also, this study agrees quite well with the earlier study of (Ibrahim, *et al*., 2017), among others.

Conclusion

In this study, the effect of outliers on some robust estimators against the sample mean, $\bar{x}$ was investigated through simulation. When a dataset contains outliers, robust estimators are more efficient than the sample mean, $\bar{x}$. However, in the study of clean data, the sample mean, $\bar{x}$ performed better than robust estimators, yet robust estimators still compete favourably with the sample mean, $\bar{x}$ .Finally, this study suggests using robust estimators when the dataset contains outliers as against sample mean, $\bar{x}$. In future studies, live data will be used to compare the efficiency of estimators under study for generalization.

References

Al-Athari, F. M., & Al-Amleh, M. (2016). A Comparison between Least Trimmed of Squares and MM-Estimation in Linear Regression Using Simulation Technique.

Alma, O. G. (2011). Comparison of Robust Regression Methods in Linear Regression. *Int. J. Contemp. Math. Sciences*, Vol. 6 No. 9, 409 - 421.

Barnett, V., & Lewis, T. (1994). Outliers in Statistical Data. *John Wiley & Sons, Colchester, England 3rd Edition*.

David, M. E., & Vikki, M. M. (2008). Modern Robust Statistical Methods: An Easy Way to Maximize the Accuracy and Power of Your Research. *Vol.63 No.7,591-601* doi: 10.1037/0003-066X.63.7.591

Draper, N. R., & Smith, H. (1998). Aplied Regression Analysis. *Johon Wiley and sons, Inc., Third edition*.

Hogg, R. V. (1979). Statistical Robustness: One View of its use in Applications Today. *The American Statistician, 33*(3), 1537-2731. doi: 10.1080/00031305.1979.10482673

Ibrahim, S. A., Ibrahim, M. W., Jaiyeola, S. B., & Oni, A. A. (2017). A Comparison of Some Univariate Estimators in the Presence of Outliers. *Journal of the Nigerian Association of Mathematical Physics, 41*, 291-296.

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). Introduction to Linear Regression Analysis. *John Wiley & Sons, Inc., Hoboken, New Jersey, Fifth edition*.

Ord, K. (1996). Outliers in statistical data: V. Barnett and T. Lewis, 1994, (John Wiley & Sons, Chichester), 584 pp.,£ 55.00, ISBN 0-471-93094-6: Elsevier.

Oyejola, B. A. (2014). An Introduction to Robust Statistical Methods. *Pre-Conference Workshop*.

Rutgers, D. E. T. (n.d). A Short Course on Robust Statistics. *www.rci.rutgers.edu/ dtyler/ShortCourse.pdf*.

Saculinggan, M., & Balase, E. A. (2013). Empirical Power Comparison Of Goodness of Fit Tests for Normality In The Presence of Outliers. *Journal of Physics Conference Series 435*. doi: 10.1088/1742-6596/435/1/012041

Warren, C., & Fred, B. (1997). General Statistics *John Wiley & Sons, Inc., 3rd Edition*.