Comparison Between Robust Trimmed and Winsorized Mean: Based on Asymptotic Variance of the Influence Functions

Mahfuzur Rahman Khokan
Department of Statistics, University of Dhaka, Dhaka- 1000, Bangladesh.

Abstract

A robust trimmed mean and winsorized mean has been compared in terms of influence function under the situation when a small change occur in the underlying symmetric distribution. The behavior of the two robust estimators have been compared through the asymptotic variance of the influence functions of the corresponding estimators. A Monte Carlo simulation studies has also been conducted to examine how asymptotic variance the influence function of the two robust estimators behave with the variation of the amount of trimming as well as with various the sample sizes. The simulated result revealed that the asymptotic variance of the influence function for both robust estimators increases when the amount of trimming increases but having lower trend for the estimator winsorized mean. That is, the estimator winsorized mean provides more efficient as well as robust result compared to the estimator trimming mean.

Keywords: Trimmed Mean, Winsorized Mean, Influence Function, Monte Carlo Simulation

1. Introduction

In the robust literature, several robust methods of estimation have been proposed (Hampel 1974, Huber 1981, Hampel et al 1986) to reduce the influence of outliers in the data, on the estimates. An outlying observation or 'outlier' is one that appears to deviate markedly from other members of the sample in which it occurs [Grubbs (1969)]. It may arise because of generating from different mechanism or assumption [Hawkins (1980), Johnson (1992)]. Once the observations arise as an outlier, then the estimation procedure may fail to produce an efficient as well as robust estimator. Then one remedy can be removing the contaminated observations from the sample or replaced by the corrected observations. In statistical data analysis, the rejection of outliers from the data may have serious consequences on further analysis for the sample being reduced. If the outliers get rejected from the data then the data is no more complete but censored. In practice, replacing the rejected outliers by statistical equivalents i.e, by simulated random observations from the assumed underlying distribution may also have similar consequences. In this situation, the robust method

of estimation aims to produce statistical procedure which do not directly examine the outliers but seeks to accommodate them such that their influence on the estimation procedure become less serious. The robust methods which are usually used in this situation to characterize the underlying distribution defined as 'Winsorization' and 'Trimming'.

The main purpose of this paper is to discuss one of the robustness properties (influence function) of the location estimators such as trimmed mean and winsorized mean for the underlying symmetric distribution and to compare which estimator provides more efficient as well as robust result in terms of influence function with the variation of amount of trimming proportion for the underlying symmetric distribution. The influence function of an estimator measures the amount of change in an estimator that can be influenced by the change of an individual observation. This appealing idea introduced by [Hampel (1974)] defining the term as influence function or influence curve (IC). Suppose we have a basic symmetric model F and a random contamination model $(1-F)\lambda+\lambda G$. Then the influence function of an estimator $T(x_1, x_2, \ldots\ldots, x_n)$ for the basic distribution function F is defined as,

$$IC_{T,F}(\xi) = \lim_{\lambda \to 0} \frac{T(\,(1-\lambda)F + \lambda G) - T(F)}{\lambda}$$

Where, λ is the proportion of contamination and G be the distribution function that puts all probability mass in the point ξ, or

$$G_{\xi}(x) = \begin{cases} 0, & if\ x < \xi \\ 1, & 1\ if\ x \geq \xi \end{cases}$$

The influence function measures the effect of an infinitesimal contamination at the point x on the estimate. If the argument ξ regarded as a random quantity distributed according to the basic model F, then it can be shown [(Huber, 1981, p.14)] that the expectation of the influence function with respect to this variation in ξ is zero i.e,

$$\int [IC_{T,F}\ (\xi)]dF(\xi) = 0$$

and that the mean squared value of the influence function is equal to the asymptotic variance of T defined as

$$\int \left[IC_{T,F}(\xi)\right]^2 dF(\xi)$$

which is also the asymptotic variance of $\sqrt{n}(T(\hat{F}) - T(F)), where\ T(\hat{F})$ is the empirical influence function. A finite-sample influence function depends on the argument ξ, on the estimator T, and in general on the basic distribution F. It is also can be used the asymptotic equivalent to the influence function as

$$\lim_{n\to\infty} IC_{T,\widehat{F_n}}(\xi) = IC_{T,F}(\xi) \quad as\ \widehat{F_n} \to F$$

2. Influence Function for the Estimators

2.1.1 Trimmed Mean: Population Version

Suppose, F is continuous with density f and mean μ , then the α-trimmed mean for the basic model F can be defined as,

$$T_{Trim}(F) = \frac{1}{1-2\alpha}\int_{y_\alpha}^{y_{1-\alpha}} xdF$$

where, x_α denotes the α −quantile of F such that $F(x_\alpha) = \alpha$ and the α −trimmed mean for the contaminated model F_λ can be defined as,

$$T(F_\lambda) = \frac{1-\lambda}{1-2\alpha}\int_{y_\alpha}^{y_{1-\alpha}} xdF + \frac{\lambda}{1-2\alpha}\int_{y_\alpha}^{y_{1-\alpha}} xdG_\xi(x)$$

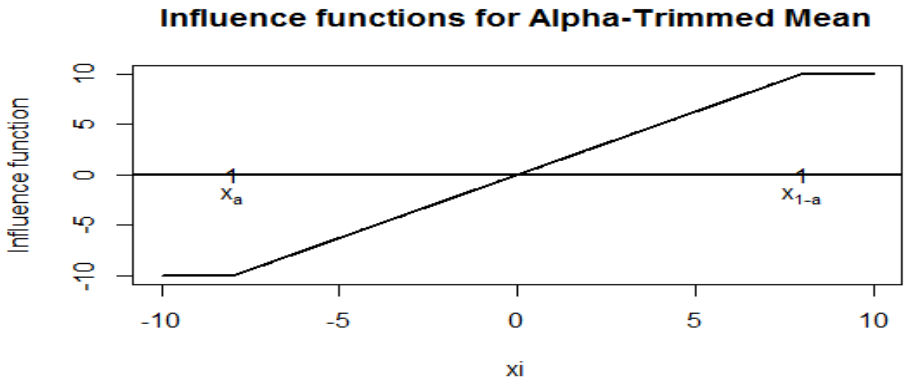
where, y_α is determined from $(1-\lambda)F(y_\alpha) + \lambda = \alpha \sim (when \sim \xi < x_\alpha)$and $(1-\lambda)F(y_\alpha) = \alpha \sim (when \sim \xi > x_\alpha)$.The influence function for α-trimmed mean can be written as follow

$$IC_{T_{\{trim\}}}(F,\xi) = \begin{cases} \dfrac{-\left(x_{\{1-\alpha\}} - \mu\right)}{(1-2\alpha)} & if\ \xi < F\text{^}(-1)(\alpha) \\ \dfrac{(\xi-\mu)}{(1-2\alpha)} & if\ F^{-1}(\alpha) < \xi < F^{-1}\{(1-\alpha)\} \\ \dfrac{\left(x_{\{1-\alpha\}} - \mu\right)}{(1-2\alpha)} & if\ \xi \geq F^{-1}(1-\alpha) \end{cases}$$

where, $x_\alpha\ and\ x_{\{1-\alpha\}}$ indicate quantile values at α and $(1-\alpha)$ respectively.

The influence function for α -trimmed mean shows that the influence curve is continuous and bounded. The influence function for the α −trimmed can be figured as below:

Fig-1: Influence function for Trimmed Mean



2.1.2 Trimmed Mean: Sample Version

In the sample case, a robust trimmed mean can be calculated after discarding the given parts of a distribution function or sample at the upper and lower end, and typically ignoring the equal amount of both ends. This number of points to be ignored is given as a percentage of the total number of points or may also be given as fixed number of points. If the amount of trimming in both ends are equal then the symmetric trimmed mean can be defined as

$$x_{r,r}^{T} = \frac{\left(x_{(r+1)} + \ \dots\dots\dots\dots\dots\dots + x_{(n-r)}\right)}{(n-2r)}$$

If the amount of trimming specified as α%, then the number of αn observations supposed to be trimmed from both ends which may not be an integer. Suppose the integer part is r , so that $\alpha n = r + f(0 < \ f \ < 1)$.We then ignore r observations at each end and include the nearest retained observations, $x_{(r+1)} \ and \ x_{(n-r)}$ each with reduced weight $(1-f)$:

$$x_{\alpha,\alpha}^{T} = \frac{\left((1-f)x_{(r+1)} + \ \dots\dots\dots\dots\dots\dots + (1-f)x_{(n-r)}\right)}{(n-2r)}$$

2.2.1 Winsorized Mean: Population Version

Suppose, F is continuous with density f and mean μ , then the winsorized mean for the basic model F can be defined as,

$$T_{win}(F) = \alpha x_{\alpha} + \int_{x_{\alpha}}^{x_{1-\alpha}} x dF + \alpha x_{1-\alpha}$$

where,x_α denotes the α-quantile of F such that $F(x_\alpha) = \alpha$ and the winsorize mean for the contaminated model F_λ can be defined as,

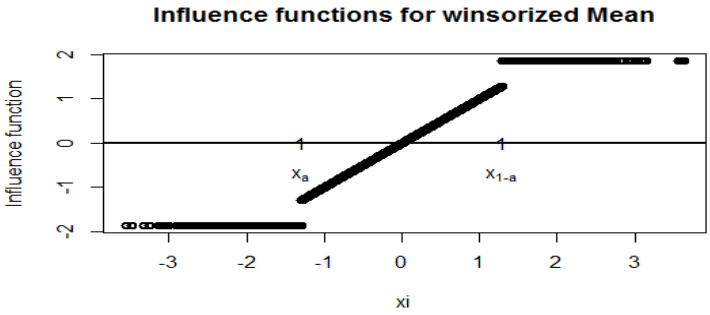
$$T(F_\lambda) = \alpha y_\alpha + \int_{y_\alpha}^{y_{1-\alpha}} x dF_\lambda + \alpha y_{1-\alpha}$$

where, y_α is determined from $(1-\lambda)F(y_\alpha) + \lambda = \alpha\ (when\ \xi < x_\alpha)$ and $(1-\lambda)F(y_\alpha) = \alpha\ (when\ \xi > x_\alpha)$. The influence function for winsorized mean can be written as follow

$$IC_{T_{\{win\}}}(F,\xi) = \begin{cases} -[(x_{\{1-\alpha\}} - \mu) + \dfrac{\alpha}{f(x_\alpha)}] & if\ \xi < F^{-1}(\alpha) \\ (\xi - \mu) \quad if\ F^{-1}(\alpha) < \xi < F^{-1}\{(1-\alpha)\} & \\ [(x_{\{1-\alpha\}} - \mu) + \dfrac{\alpha}{f(x_\alpha)}] & if\ \xi \geq F^{-1}(1-\alpha) \end{cases}$$

where, x_α and $x_{\{1-\alpha\}}$ indicate quantile values at α and respectively. The influence curve shows that the influence function is indeed bounded, the outliers "brought in", but there is a jump at$[x_\alpha\ and\ x_{\{1-\alpha\}}]$. The influence curve is also discontinuous and very sensitive to the local behavior of the true underlying distribution at two of is quantiles [(Hampel,1974)].The influence function for the winsorized mean can be drawn as bellow where observations has been taken form standard normal distribution. The following graph shows a bounded function but having two little jump at α and $(1-\alpha)$ quantile respectively.

Fig-I1: Influence function for Winsorized Mean



2.2.2 Winsorized Mean: Sample Version

In the sample case, a robust winsorized mean can be calculated after replacing the given parts of a distribution function at the upper and lower ends with the most extreme remaining values. It is a

robust estimator because of its less sensitiveness to outliers. If the amount of trimming at lower-tail and upper-tail are same then the symmetric winsorized mean defined as

$$x_{r,r}^{W} = \frac{\left(rx_{(r+1)} + \ \ldots\ldots\ldots\ldots\ldots\ldots + rx_{(n-s)}\right)}{n}$$

3. Comparison between Robust Trimmed Mean and Robust Winsorized Mean

The robust method of estimation such as trimmed mean and winsorized mean reduces the influence of contamination on the estimation procedure. The influence function of the estimators identify how the estimators behave in the presence of outliers. It can also be examined which estimator provide more efficient estimator with the help of asymptotic variance of the influence functions. In the following sections, the asymptotic variance of the influence functions for the estimators trimmed mean and winsorized mean has been derived and their performances has been compared by Monte Carlo simulation.

3.1.Comparison through Asymptotic Variance of Influence Function

The asymptotic variance of the influence function of the robust estimator trimmed mean can be derived as follow:

Asymptotic variance $= \int_{-\infty}^{\infty}\left[IC_{T_{trim}}(\xi, F)\right]^2 dF(\xi)$

$$= \left(\int_{-\infty}^{x_\alpha} + \int_{x_\alpha}^{x_{1-\alpha}} + \int_{x_{1-\alpha}}^{\infty}\right)\left[IC_{T_{trim}}(\xi, F)\right]^2 dF(\xi)$$

$$= \int_{-\infty}^{x_\alpha}\left[\frac{(x_{1-\alpha}-\mu)}{1-2\alpha}\right]^2 dF(\xi) + \int_{x_\alpha}^{x_{1-\alpha}}\left[\frac{(\xi-\mu)}{1-2\alpha}\right]^2 dF(\xi) + \int_{x_{1-\alpha}}^{\infty}\left[\frac{(x_{1-\alpha}-\mu)}{1-2\alpha}\right]^2 dF(\xi)$$

Suppose, the basic model is a Gaussian distribution defined as

$$dF(\xi) = \frac{1}{\sqrt{(2\pi\sigma^2)}}\exp\left[-\frac{1}{2}\left(\frac{\xi-\mu}{\sigma}\right)^2\right]d\xi$$

We can write that

$$\int_{x_\alpha}^{x_{1-\alpha}} \frac{1}{\sqrt{2\pi\sigma^2}}\, exp\left[-\frac{1}{2}\left(\frac{\xi-\mu}{\sigma}\right)^2\right]d\xi = (1-2\alpha)$$

Considering the underlying basic model is Gaussian, the asymptotic variance of Influence function for α-trimmed mean can be formulated as

$$\text{Asymptotic Variance} = 2\alpha\sigma^2 \frac{z_{1-\alpha}^2}{(1-2\alpha)^2} + \sigma^2\left[\frac{1}{1-2\alpha} + \frac{z_\alpha}{(1-2\alpha)^2} f(z_\alpha) - \frac{z_{1-\alpha}}{(1-2\alpha)^2} f(z_{1-\alpha})\right]$$

$$= \sigma^2\left[\frac{2\alpha z_\alpha^2}{(1-2\alpha)^2} + \frac{1}{1-2\alpha} + \frac{2z_\alpha}{(1-2\alpha)^2} f(z_\alpha)\right] \; ; [since, z_\alpha = -z_{1-\alpha} \; and \; f(z_\alpha) = f(z_{1-\alpha})]$$

The asymptotic variance of the influence function of the robust estimator winsorized mean can be derived as follow:

$$\text{Asymptotic variance} = \int_{-\infty}^{\infty} \left[IC_{T_{win}}(\xi, F)\right]^2 dF(\xi)$$

$$= \left(\int_{-\infty}^{x_\alpha} + \int_{x_\alpha}^{x_{1-\alpha}} + \int_{x_{1-\alpha}}^{\infty}\right) \left[IC_{T_{win}}(\xi, F)\right]^2 dF(\xi)$$

$$= \int_{-\infty}^{x_\alpha} \left[(x_{1-\alpha} - \mu) + \frac{\alpha}{f(x_\alpha)}\right]^2 dF(\xi) + \int_{x_\alpha}^{x_{1-\alpha}} [(\xi - \mu)]^2 dF(\xi) + \int_{x_{1-\alpha}}^{\infty} \left[(x_{1-\alpha} - \mu) + \frac{\alpha}{f(x_\alpha)}\right]^2 dF(\xi)$$

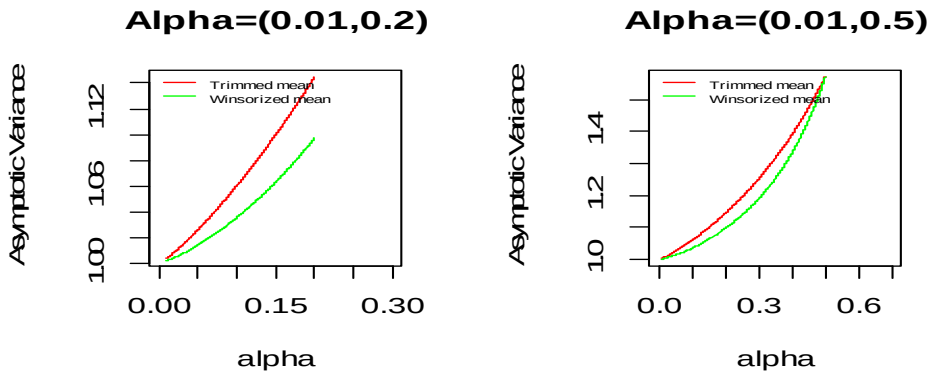
Considering the underlying basic model is Gaussian, the asymptotic variance of Influence function for winsorized mean can be written as

$$\text{Asymptotic Variance} = 2\alpha\sigma^2 \left[z_{1-\alpha} + \frac{\alpha}{f(z_\alpha)}\right]^2 + \sigma^2[(1-2\alpha) + z_\alpha f(z_\alpha) - z_{1-\alpha} f(z_{1-\alpha})]$$

$$= \sigma^2\left[2\alpha\left[\frac{\alpha}{f(z_\alpha)} - z_\alpha\right]^2 + (1-2\alpha) + 2z_\alpha f(z_\alpha)\right] ; [since, z_\alpha = -z_{1-\alpha} \; and \; f(z_\alpha) = f(z_{1-\alpha}) \;]$$

The asymptotic variance of the estimators α-trimmed mean and winsorized mean in terms of influence function can be drawn as follows:

Fig-III: Asymptotic Variance of Influence functions for Trimmed and Winsorized Mean



It is known that the variance of influence function for the median is $\pi/2$ and the breakdown point for median is 0.5 or 50%. So it is noticed from the graph that both trimmed mean and winsorized mean are more efficient than median when the amount of trimming less than 50%. The graph also

shows that the asymptotic variance of the influence function for both the trimmed mean and winsorized mean increases if the amount of trimming increases. Initially the variance for both of them are very close to the variance of the underlying basic model. But the variances are increasing due to increasing amount of trimming showing winsorized mean has lower trend than trimmed mean. That indicates the robust winsorized mean provide better result than robust trimmed mean when the amount of trimming increases.

3.2.Comparison through Monte Carlo Simulation

A Monte Carlo simulation study has been performed to compare the performance of trimmed and winsorized mean with respect to different level of trimming (α) to confirm the theoretical findings. First we generated 100 sample from a standard normal distribution $N(0,1)$ and computed trimmed mean, winsorized mean for these sample at the different level of trimming. We repeated this procedure for 1000 times and calculated the Monte Carlo variance of influence function for trimmed mean, winsorized mean and continued the above task for three different sample sizes($n = 100,200,300\ and\ 500$) taken from standard normal distribution at the different values of trimming(α). Firstly, we considered the variation of trimming proportion (α) within $(0.01,0.2)$ and again the same procedure for the variation of trimming proportion (α) within $(0.01,0.5)$ to observe how the simulated variance of influence function for the three estimators behave at different level of trimming (α).

Fig-IV: Monte Carlo Simulations for Asymptotic Variance of Influence functions for Trimmed mean, Winsorized mean and Mean: When n=100

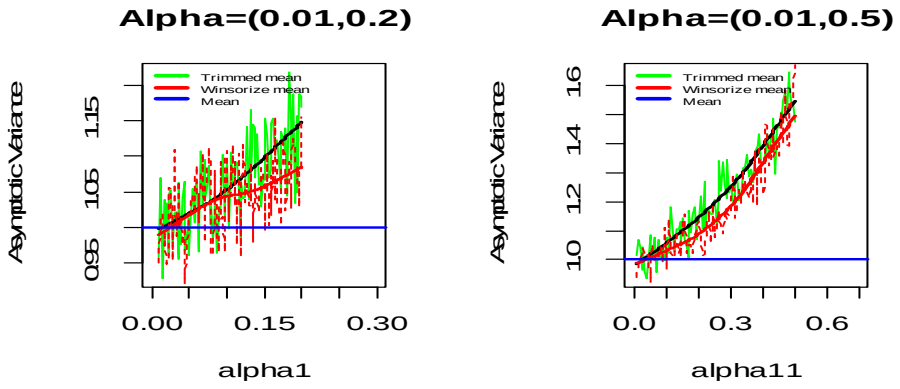


Fig-V: Monte Carlo Simulations for Asymptotic Variance of Influence functions for Trimmed mean, Winsorized mean and Mean: When n=200

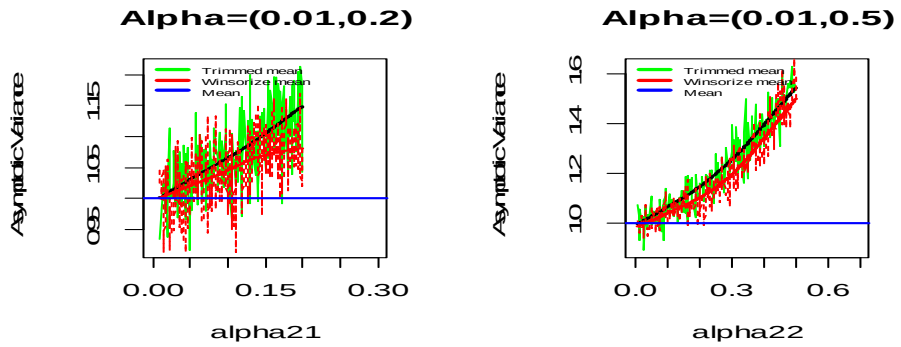


Fig-VI: Monte Carlo Simulations for Asymptotic Variance of Influence functions for Trimmed mean, Winsorized mean and Mean: When n=300

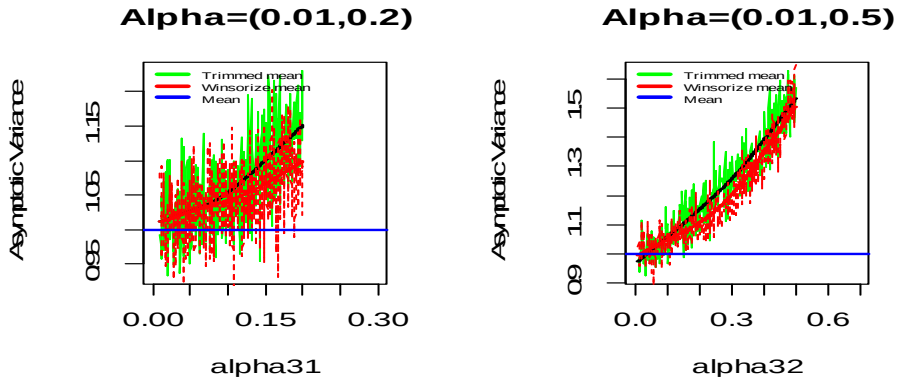
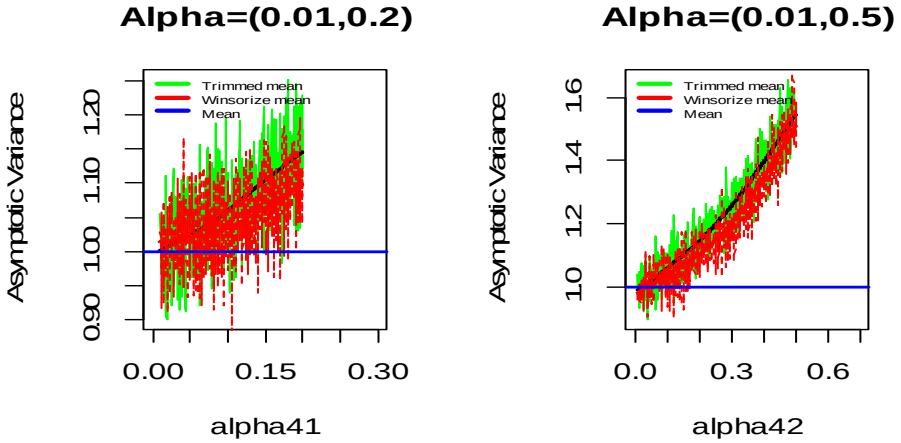


Fig-VII: Monte Carlo Simulations for Asymptotic Variance of Influence functions for Trimmed mean, Winsorized mean and Mean: When n=500



The graph plotted by simulation shows that when the proportion of trimming (α) increases then the variance of influence function for both estimators (trimmed mean and winsorized mean) increases. It is observed that when α is small then the asymptotic variance of influence function for trimming mean is very close to winsorized mean. The graph also shows that the variation of influence function for winsorized mean goes close to trimmed mean if the mount of trimming close to 0.5. It is noticed that when the amount of trimming increases than the asymptotic variance for both the trimmed mean and winsorized mean deviate from the variance of the basic standard normal distribution but lower the variance of influence function for median which is $\pi/2$. It is also observed that for one unit change in trimming (α) the efficiency for winsorized mean increase almost 10% than the trimmed mean for different sample sizes.

4. Conclusion

Robust estimation method always provide stable estimators for unknown parameters, when the underlying distribution contains a small departures from the parametric distributions. When a small departures occurs in the underlying parametric distribution then the classical location estimators loss their robustness. In that situation the robust location estimators such as trimmed mean, winsorized mean provide robust estimate for the corresponding location parameters. In this project, the performance of these robust location estimators has been discussed in terms of the efficiency of their corresponding influence functions for the variation of trimming proportions. The simulation study indicates that influence for the both trimming and winsorized mean increases when the trimming proportion increases and the winsorized mean has lower increasing trend than the trimming mean. The simulation result also revealed that both the location estimators are robust but the winsorized mean is more robust as well as efficient than the trimmed mean.

In future, we can check the performance of robust location estimators in term of influence functions when the underlying distributions are not symmetric. The influence function for the robust trimmed and winsorized mean can also be checked for multivariate data.

5. References

[1] Hampel, F. (1974). The influence curve and its role in robust estimation, J. Am. Statist. Assoc. 69: 383393.

[2] Huber, P.J. (1981). Robust Statistics. Wiley, New York.

[3] Hampel et al. (1986). Robust Statistics: The Approach Based on Influence Functions. Wiley, New York.

[4] Grubs, F.E. (1969).Procedures for detecting outlying observations in samples.Tecnometrics, 11, 1-21.

[5] Hawkins, D., 1980. Identification of Outliers, Chapman Hall.

[6] Johnson, R., 1992. Applied Multivariate Statistical Analysis, Prentice Hall.