ROBUST M-ESTIMATORS OF THE POPULATION MEAN.

Mutuguta Bernard Githui
I56/CE/22016/2010
Bsc.(Mathematics)

A research project submitted in partial fullfillment of the requirement for the award of degree Master of Science (Statistics) in the School of Pure and Applied Sciences of Kenyatta University

JULY 2018

Declaration

This research project is my original work and has not been presented for a degree award in any other University.

Signed.....Date.....

Bernard Githui Mutuguta

This work has been presented with my approval as the University supervisor.

Signed.....Date.....

Dr. Kube Ananda

Department of Statistics and Actuarial Sciences

Kenyatta University.

Dedication

This work is dedicated to my beloved wife Esther Wairimu and our children, Regina, Michael and Joan.

Acknowledgements

I would want to most sincerely start by acknowledging Professor Leo Odongo, who introduced the concept of estimation to me till I finally got interested in pursuing it as a topic of my research. Secondly I would want to thank Dr. Kube Ananda for the assistance he accorded me as my supervisor and for devoting a lot of his precious time to guide me through this project.

I would like to additionally thank all the Lecturers of the Statistics and Actuarial Sciences department of Kenyatta University who enabled me go through the Course work successfully.

I take this golden opportunity to sincerely thank my family for the over- whelming moral and material support that they have provided to me during my study period.You have all been a deep inspiration to me and your prayers truly kept me going even during the difficult times. May the Almighty God bless you abundantly, each according to ones needs.

I cannot also forget to thank my classmates whose prayers and financial support saw me recover from a life threatening ailment that made me lose a whole academic year. Once again, I wholeheartedly thank you all.

Last but not the least, I thank the almighty God who has continued to make me rise up every morning to fend for my family.

Contents

Declaration ii

Dedication iii

Acknowledgements iv

List of Tables viii

List of Figures ix

Abstract x

1 Introduction 1

1.1 Background of the study 1

1.2 Problem Statement 7

1.3 Objectives of the study 7

1.3.1 Main Objective 7

1.3.2 Specific Objectives 8

1.4 Significance of the study 8

1.5 Justification of the study 9

1.6 Literature Review 9

2 Review of Robust M-estimators **13**

2.1 Introduction 13

2.1.1 Huber estimator for location parameter μ 15

2.1.2 Redescending M-estimators 17

2.2 Asymptotic Properties of the M-estimators 19

2.2.1 Limiting distribution of T_n 21

2.2.2 Minimax variance 22

2.2.3 Influence Curve of an M-estimate 25

2.2.4 Breakdown and Continuity Properties of M-estimates 26

2.3 M-estimators in Known and Unknown Scale Cases 29

3 Data Analysis and Discussion of Results **32**

3.1 Introduction 32

3.2 Simulated Normal Population. 33

3.3 Contaminated Simulated Population. 34

3.4 Comparison of Means 38

3.5 Standardization of the M-estimators 39

3.6 Application of M-estimators of Location on Real Data 40

3.6.1 Introduction 40

3.6.2 Comparison of the Means 43

3.7 Standardization of the M-estimators 44

4 Summary, Conclusions and Recommendations 46

4.1 Introduction 46

4.2 Summary from asymmetrical population 46

4.2.1 Standardized Z-values for the means of selected samples . . . 47

4.3 Summary from Real Data case 48

4.3.1 Standardized Z-values for the means of selected samples . . . 48

4.4 Conclusion 49

4.5 Further research 49

References 51

List of Tables

3.1 Means of 50 Samples obtained using Huber's method 36

3.2 Means of 50 Samples obtained using Hampel's method 37

3.3 Means of 50 Samples obtained using Tukey's biweight method 37

3.4 Means of Selected Samples 39

3.5 Standard Z-values representing the M-estimators for the various samples 40

3.6 Means of Samples from Huber's method 41

3.7 Means of Samples from Hampel's method 42

3.8 Means of Samples from Tukey's method 43

3.9 Means of Selected Samples 44

3.10 Standard Z-values representing the M-estimators for the various samples 45

List of Figures

2.1 Graphic Representation of the Huber's Function 16

2.2 Hampel three part Redescending function 18

2.3 Graphic Representation of the Tukey's Biweight 19

3.1 Graphical representation of the population data 33

3.2 Boxplot for the population data 34

3.3 Graph of Contaminated Population Data 35

3.4 Boxplot of the Contaminated population data 35

3.5 Histogram of the Real data 41

3.6 Boxplot of the Real Data 42

Abstract

Valuable information may be contained in outliers. Data collected from many medical equipment such as the ECG time-series may result in unusual patterns, a situation that may represent disease conditions. Some outliers may therefore provide an insight into crucial information that may lead to the discovery of new knowledge.Due to the difficulties involved in the study of outliers, majority of current research works have chosen not to out-rightly reject the outliers but have adopted methods that accommodates them but have their influence reduced to some degree. This has resulted in the development of estimation and testing techniques that have yielded robust estimation. Three robust M-estimators (the Huber, the Hampel and the Tukey's biweight) were adopted in this paper with an aim of comparing their efficiency in the estimation of population mean in an asymmetrical distribution. Simulation was used to generate a normal population which was then contaminated with a few outlying observations. A simple example illustrated how the comparison of the M-estimators was carried out on real data. On the basis of the findings from this research, the M-estimator that turned out to be the best, was the Tukey's biweight followed by the Hampel's and the Huber's was the weakest of the three. However all the three M-estimators yielded unbiased estimates of the population mean.

Chapter 1

Introduction

1.1 Background of the study

The problem of estimation of the mean has for long time been of great research interest. In Manufacturing industry for example, researchers are interested in computing the expected mean life time of components of the production process, for it enables the products sell competitively, when such mean lifetimes are stated to the consumer.

In the collection and analysis of data however, there are times when mistakes or errors arise during recording. Such errors may end up being outliers and may seriously affect the mean. Many researchers have devised various methods of estimating mean robustly, such that the estimate so obtained approaches the actual mean of the distribution under study.

In estimating a measure of location in a sample one has to search for some point say c,

which is part of the sample and which is close to most of the values of the sample, say $(x_1, x_2, \ldots, x_n)$. If we sum the squares of the differences between the point c and each of the n observations in the sample, the value of the derivative is found to approach zero, the point c is a said to be an unbiased estimate of the mean of the distribution where the sample was obtained from. The main disadvantage of the sample mean is that as compared to M-estimators of location, it's standard error turns out to be higher, thus making the sample mean a worse of estimator of location in the presence of outliers. The standard error of some M-estimators yield values that are very close to those obtained for the sample mean in the normal distribution case while still continues to perform much better than the mean when the sample contains some outlying observations. M-estimators are considered to posses properties that enable them to be robust estimators in the presence of asymmetry in distributions. The properties are, the influence function, gross error sensitivity, local shift sensitivity, rejection point and breakdown point.

The influence function determines the influence an outlying observation has on the magnitude of the estimator. If the influence function is not bounded, it means that the further away the outlying observation is from the rest of the observations, the greater the influence it will have on the estimator. The arithmetic mean has an influence function that is linear and is not therefore bounded. The gross error sensitivity determines the degree of influence the outlying observations in the data has on the value of the estimate. If the value obtained is finite, the estimator is said to be B-robust.

The local shift sensitivity is the one determined by small variations in the data, and it is best for it to be small and finite. For symmetrical distributions in the neighbourhood of zero, observations are rejected only if they are beyond a certain point referred to as the rejection point. The Estimator too, is expected to have a finite rejection point for it to be reliable. The breakdown point of an estimator is the percentage of outliers the estimator can withstand before it can be said to have become invalid.The breakdown point of an estimator is what defines it's quantitative robustness.

Researchers have over the years used several terms to define what outliers are. These are for example, "surprising values," "discordant observations,""contaminants," "rogue values," "mavericks," or "dirty data," just to mention but a few. A researcher has got to be concerned when such values occur hence the need to establish what may have caused them. These observations may be caused by a multiplicity of factors.

W.Osborne and Overbay (2004), noted that, the errors may be due to Man made causes- such as during data collection, recording, or even during entry. During interviews data may be at times be recorded incorrectly leading to discordant values. Intentional or motivated mis-reporting- Huck (2000) noted that due to various reasons a participant in a research may be unwilling to volunteer correct information or may have the intention of sabotaging a research. The motives for such acts may be driven by social desirability and self-presentation of the participant especially if the research was to be encroaching on such spheres. When data is sensitive misreporting can also happen for obvious reasons (e.g., teenagers may under-report on the use of drug or

alcohol,they may also misreport on sexual behavior). If majority of the teenagers under-report a behavior for example, the frequency at which they are involved in sexual intercourse, the few who may be giving honest responses might appear to be outliers when in fact they are the legitimate and correct values. Sampling is yet another cause of outliers . If some of the sample observations were erroneously drawn from a different population, this can also lead to the existence of some discordant elements in the sample.

A legitimate member of a normal population may also be classified as an outlier. A sample observation drawn form either of the two extremes of a normal distribution if majority of the other sample elements, drawn randomly but happen to have been around the centre of the normal population. In any normal distribution there is always a 1% chance of observations lying beyond 3 standard deviations on either side of the mean of the distribution. Even though such observations are genuine members of the normal population they may appear like outliers if majority of the sample observations were drawn close to the centre of the distribution.

The presence of outliers can lead to over-estimation or under-estimation of population parameters. The various statistics used in the estimation of population parameters such as the sample mean and standard deviation, may thus be rendered invalid as estimators of their respective parameters due to the existence of outliers in samples. The existence of outliers end up increasing standard error and reduce the power of statistical tests; if they are not randomly distributed, outliers decrease normality and increases the possibility of making both Type I and Type II errors.

There are many cases in real life when a researcher encounters data that may have some outlying observations that cannot be done away with or that cannot simply be replaced with others for the sake of coming up with an inference about the population being studied. Examples may include but not limited to situations like when testing the mean length of life of components coming from a production process. A single outlying observation may substantially lower or increase the mean length of life of the components. This may be detrimental to the manufacturer or the consumer. The manufacturer may under price the components for having an unusually low mean lifespan thus make losses. The consumer may pay more for components expecting that since the components have been declared to have a long lifespan, they would last long; only to realize that the declared lifespan was not true. Such a consumer may lose confidence in that particular manufacturer. When researching on the correct dosage to be administered in the treatment of a particular disease. outlying dosages administered to special cases of patients (e.g. Alcoholics) may end up distorting the mean effective dosage to the detriment of the patients. An under dose may lead to the drug failing in its efficacy while an overdose may result in serious side effects or even death of a patient. Many more practical cases can be quoted. This therefore means that for particular cases of data, something has got to be done to reduce the effect of outliers on the estimation of population parameters being estimated, irrespective of how they arose. M-estimators are examples of robust estimators of population parameters that can be used in distributions that are non-normal.

The sample mean and the sample standard deviation are always good estimators of

the location and scale parameter of a statistical population; In the presence of outliers they are rendered unreliable . In such a situation other more reliable estimators of population parameters are required; these estimators are referred to as M-estimators. They are maximum-likelihood type of estimators and they are found to be a zero of an estimating function; the estimating function itself often being the derivative of another statistical function: For example, a maximum-likelihood estimate is often defined to be a zero of the derivative of the likelihood function with respect to the parameter: thus, a maximum-likelihood estimator is often a critical point of the score function. Assuming that data is obtained from the model distribution $F_{\mu,\sigma}$ then the log-likelihood can be written as

$$\sum_{i=1}^{n}\{\log f_o(\frac{x_i-\mu}{\sigma}) - log\sigma\} \tag{1.1}$$

The first order condition for the ML-estimator of μ is then given by

$$\frac{1}{n}\sum_{i=1}^{n}\psi_{ML}(\frac{x_i-\mu_{ML}}{\sigma}) = 0 \tag{1.2}$$

while the ML-estimator of scale verifies

$$\frac{1}{n}\sum_{=1}^{n}\rho_{ML}(\frac{x_i-\mu}{\sigma_{ML}}) = 1 \tag{1.3}$$

with $\psi_{ML}(\mu) = -f_o'(\mu)/f_o(\mu)$ the score function, and $\rho_{ML}(u) = \psi_{ML}(u)u$ Under regularity conditions the ML-estimators have a 100% efficiency, meaning that their

asymptotic variance equals the inverse of the Fisher information the lower bound of the Cramer-Rao inequality.

1.2 Problem Statement

There are many times anomalous/outlying observations occur in data. If such observations are used in their raw form, then population parameters may seriously be affected.

An overestimation or an underestimation of a parameter may have serious repercussions. If for example a study aimed at establishing the effective dosage of a new drug, then an overestimation of the dosage level may result in death.

Estimators that are insensitive to outliers in data should be used in the hope of obtaining unbiased estimates of population parameters.

1.3 Objectives of the study

1.3.1 Main Objective

This study aims to compare three robust M-estimators of population mean methods namely the Huber, Hampel and Tukey's biweight.

1.3.2 Specific Objectives

(i) To compare M-estimators of a location parameter under independent samples drawn from a given distribution function F(.).

(ii) To review the asymptotic properties of the M-estimators in (i) above.

(iii) To perform a simulation study of the estimators considered in (i), to verify the properties studied under (ii).

(iv) To compare the performance of the M-estimators in a real data case.

1.4 Significance of the study

In real life, majority of the observations that are encountered fall under the normally distribution. But there are times when some observations deviate from the norm. This study will find applications in very many sectors where a number of outlying observations arise. The outlying observations may not always occur as a result of errors and should not be disregarded in the estimation of the parameters under study. This study compares three different robust methods of estimation of population mean in asymmetrical distributions to establish which is more reliable.

The study also is expected to add to the body of knowledge in the field of robust estimation that have been undertaken in the past.

1.5 Justification of the study

Although outliers are often measurement or recording errors, some of them have been found to represent phenomena of interest. It may not therefore be advisable to reject all the outliers as doing this may lead to loss of crucial information that may have been very helpful in understanding the phenomenon being studied. The study of robust estimation thus helps a great deal in utilizing outliers to a reasonable degree hence may guard against total loss of useful information that may lead to the discovery of new knowledge.

1.6 Literature Review

Researchers have sometimes used various robust techniques to ensure that data does not get distorted by the presence of outliers instead of using transformations or truncation of the outliers. The techniques used accommodate the outliers without alot of distortion in the estimation of parameters and are hence said to be robust against the presence of outliers. The mean and Least Squares estimations, are found to seriously get distorted in their estimation of parameters,in the presence of a few outlying observations i.e they have low breakdown values. Researchers have therefore turned to robust or high breakdown estimation methods that provide alternative estimators of parameters for the populations under study.

Crow and Siddiqui (1967), found out that the problem of estimating a location parameter from a random sample when the form of distribution is unknown or there

is contamination, is dealt with by deriving estimators which are efficient over a class of two or more forms of continuous symmetric unimodal distributions. Thall (1979), applied Huber (1964) classical theory of robust M-estimation of a location parameter and using the results reformulated the scale parameter context and then applied to the problem of robust estimation of the parameter of the exponential distribution. Lenth (1981), illustrated that robust M-estimators of a location parameter are also used for directional data. He found out that a periodic version of any of the commonly used ψ-functions can be used to define a comparable estimator of angular location. The proposed estimators appear to perform at efficiency levels similar to those of ordinary M- estimators in the linear case.

Fung et al. (1985), while studying some robust test statistics for the 2- sample problem proposed three families of robust test statistics from M-estimators. These were the sine-wave t, the bisquare t and the Hampel t. After carrying out some comparisons, they found out that robust statistics should definitely be recommended when longer tails are expected while students t should be used for normality and for distributions with shorter tails.

Akritas. (1991), studied the M-estimators of location difference of pairwise samples drawn randomly from a given location-scale family. He found out that, when the same score function is used, the new and usual M-estimators of location shift have the same influence functions and asymptotic variance. He however observed that all members of the new class of optimal robust M-estimators of location shift have sample-wise breakdown point of 0.5 and are thus preferable over the usual M-estimate

in the asymmetric case.

Basak (1998), developed Robust M-estimation procedures for relevant parameters of discriminant analysis. The optimal robust M-estimators for discriminant function coefficients, Mahalanobis' Δ^2 and mis-classification probabilities were obtained but they failed to have a breakdown point of 0.5. He however proposed other robust estimators belonging to one special class of estimators. He was finally able to show that certain estimators belonging to this class attained breakdown points of 0.5. He however observed that the efficiency of the estimators was compromised as the optimality in breakdown point was achieved. Stefanski and Boos (2002), illustrated the breadth and generality of the M-estimation approach and justified its use in the study of large-sample inference. Wang et al. (2007), studied Robust estimation using the Huber function with a Data-Dependent Tuning Constant . They observed that Robust estimation is often desirable in the presence of outliers and that robustness against outliers is gained at the price of efficiency loss when the resistance is unnecessarily high. They concluded that efficient estimation is possible only if a dispersion function with an appropriate resistance level is chosen. Beran (1991), investigated the behavior of M-estimators of the location parameter for stochastic processes with long-range dependence. The processes he considered are Gaussian or one-dimensional transformations of Gaussian processes. It turned out that, up to a constant, all M estimators are asymptotically equivalent to the arithmetic mean. He found out that for Gaussian processes this constant is always equal to one, irrespective of the ψ-function. Birnbaum and Mike (1970), studied the problem of efficiency-robust

estimation of location, using Monte Carlo methods under the following distributions: normal, logistic, double-exponential, and contaminated normal (one percent, five percent, ten percent), for sample sizes n = 20, 30, 40, 50, and 100. They found out that over all these shapes the efficiencies obtained are approximately 88 percent or more for all n; they rise to approximately 91 percent or more for n = 100. Kagan (1966), in his paper, on the Estimation Theory of Location Parameter, came up with admissibility conditions for the sample mean as an estimator of the location parameter in various classes of unbiased estimators.

Chapter 2

Review of Robust M-estimators

2.1 Introduction

The sample mean and the sample variance are known to be seriously affected by the presence of just a few outliers in a sample. In such a case, reliable i.e less sensitive estimators have been found and are referred to as M-estimators and their study called Robust estimation. They are Maximum-likelihood type estimators and are insensitive to outlying observations in a distribution. In the comparison of robustness properties of the various estimators several measures are used. The most commonly used measures are the Breakdown point and the influence function. The former gives the highest proportion of outliers that data may posses, before the estimator goes over all bounds. The breakdown point is termed as a global measure of robustness. The influence function locally measures robustness, hence it represents the effect of a single outlier, Hampel (1974).

Any estimate T_n defined by a minimum problem of the form

$$\sum_{1}^{n} \rho(x_i; T_n) = min! \tag{2.1}$$

or by an implicit equation

$$\sum_{1}^{n} \psi(x_i; T_n) = 0 \tag{2.2}$$

where ρ is an arbitrary function $\psi(x, \theta) = (\frac{\partial}{\partial \theta})\rho(x; \theta)$,is called an M-estimate (or maximum likelihood type estimate; $\rho(x; \theta) = -logf(x; \theta)$ gives the ordinary M.L.E). The M-estimates of location are given by :-

$$\sum \rho(x_i - T_n) = min! \tag{2.3}$$

or

$$\sum \psi(x_i - T_n) = 0 \tag{2.4}$$

if equation (2.4) is written as

$$\sum w_i(x_i - T_n) = 0 \tag{2.5}$$

with $w_i = \frac{\psi(x_i - T_n)}{x_i - T_n}$ we obtain a representation of T_n as a weighted mean

$$T_n = \frac{\sum w_i x_i}{\sum w_i} \tag{2.6}$$

With weights depending on the sample. The favourite choices will be of the form:-

$$\rho(x) = \begin{cases} \frac{x^2}{2} & \text{for} \quad |x| \leq c \\ c|x| - c^2/2 & \text{for} \quad |x| > c \end{cases} \tag{2.7}$$

$$\psi(x) = \begin{cases} [x]_{-c}^{c} = -c & \text{for} \quad x < -c \\ x & \text{for} \quad -c \leq x \leq c, \\ c & \text{for} \quad x > c. \end{cases} \tag{2.8}$$

leading to weights

$$W_i = \begin{cases} 1 & \text{for} \quad |x_i - T_n| \leq c \\ \frac{c}{|x_i - T_n|} & \text{for} \quad |x_i - T_n| > c \end{cases} \tag{2.9}$$

2.1.1 Huber estimator for location parameter μ

If F has a symmetric density such that,

$$f_{\mu,\sigma}(x) = \frac{1}{\sigma} f\left(\frac{x-\mu}{\sigma}\right) \tag{2.10}$$

The M-estimator of location parameter μ is defined as

$$\Sigma_{i=1}^{n} \psi\left(\frac{x_i - \theta}{\sigma}\right) = 0 \tag{2.11}$$

Where $\theta = \{\theta_1, \theta_2\}$ Huber M-estimator is defined by the function ψ in (2.11) above as and graphically representated by figure 2.1.

$$\psi_k(x) = \begin{cases} k, & x \geq k \\ x, & -k \leq x \leq k \\ -k, & x \leq -k \end{cases} \tag{2.12}$$

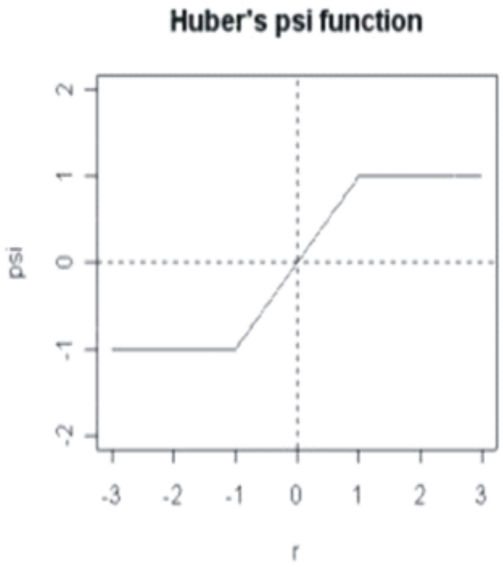

Figure 2.1: Graphic Representation of the Huber's Function

Huber's function is a parabola in the vicinity of zero, and increases linearly at a given level $|x| > k$, with k = 1.345 and referred to as the tuning constant. On the standard normal distribution a 95 % asymptotic efficiency is attained with this value of the tuning constant.

2.1.2 Redescending M-estimators

Redescending M-estimators are Ψ-type M-estimators which have ψ functions that are non-decreasing near the origin, but decreases towards zero far from the origin. Their ψ functions can be chosen to re-descend smoothly to zero, hence satisfying the equation $\psi(x) = 0$ for all x with $|x| > r$ where r is termed as the minimum rejection point. Due to these properties of the ψ function, these kinds of estimators are very efficient and have a high breakdown point. They are said to be efficient because they completely ignore moderately large outliers.

For completely ignoring gross outliers redescending M-estimators are slightly more efficient than the Huber's estimators which normally treat the gross outliers in the same as moderate outliers. It has however been found out that for these type of M-estimators their estimating equations may not always have a unique solution.

When choosing a redescending ψ function, care must be taken such that it does not descend too sharply, as this may cause it to have bad influence on the denominator in the expression for the asymptotic variance

$$\frac{\int \Psi^2 dF}{(\int \Psi dF)^2}$$

where F is the mixture model distribution.

This effect is particularly harmful when large negative values of $\psi'(x)$ combines with large positive values of $\psi^2(x)$ and there is a cluster of outliers near x.

Redescending M-estimator for location parameter μ

The Hampel's 25A estimator from Andrews et al. (1972) with redescending three-part function ψ for determining the location parameter is given by the equation below and graphically represented by figure 2.2.

$$\psi(x) = \begin{cases} |x|, & 0 \leq |x| \leq a, \\ a, & a \leq |x| \leq b, \\ a\left(\frac{c-|x|}{c-b}\right), & b \leq |x| \leq c \\ 0, & |x| > c \end{cases} \tag{2.13}$$

with

$$0 \leq a \leq b \leq c < \infty$$

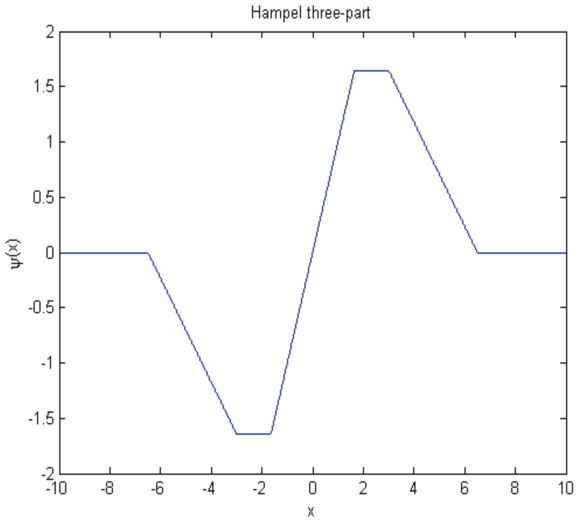

Figure 2.2: Hampel three part Redescending function

The Tukey's Biweight has a ψ-function

$$\psi(x) = \begin{cases} \frac{1}{6}(1-u^2)^2 & \text{if} \quad |c| < 1 \\ \frac{1}{6} & \text{if} \quad |c| > 1 \end{cases} \tag{2.14}$$

where $u = \frac{x-T_n}{cS_n}$, T_n and S_n being location and scale scalars respectively while c is the tuning constant. The ψ-function is shown in figure 2.3.

The ψ-function redescends to zero, that is if $|c|$ is large enough $\psi(c) = 0$. Since the ψ-function determines the weights assigned to the data points, the points with large values of c do not affect the calculation of the biweight estimate.

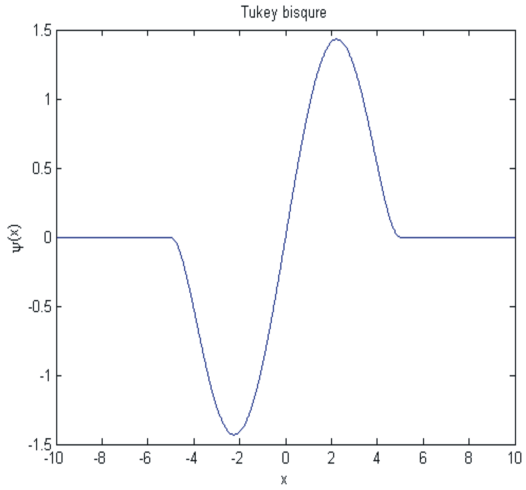

Figure 2.3: Graphic Representation of the Tukey's Biweight

2.2 Asymptotic Properties of the M-estimators

In order to compare the performance of different estimates, and also obtain confidence intervals within which the parameters being estimated are expected to lie, the

distributions from which the samples were obtained from, will be required. It will be necessary to resort to approximating their distributions for large n, the so called asymptotic distribution.

Assuming that $\psi(x;\theta)$ is measurable in x and decreasing (non-increasing)in θ, from strictly positive to strictly negative values.

Let

$$T_n^* = \sup\{t|\Sigma_{i=1}^n \psi(X_i;t) > 0\}$$

$$T_n^{**} = \inf\{t|\Sigma_{i=1}^n \psi(X_i;t) < 0\}.$$

Clearly $-\infty < T_n^* < T_n^{**} < \infty$ and any value T_n satisfying $T_n^* < T_n < T_n^{**}$ can serve as an estimate.

But

$$\{T_n^* < t\} \subset \{\Sigma\psi(x_i;t) \leq 0\} \subset \{T_n^* \leq t\}.$$

$$\{T_n^{**} < t\} \subset \{\Sigma\psi(x_i;t) < 0\} \subset \{T_n^{**} \leq t\}.$$

Hence

$$P\{T_n^* < t\} = P\{\Sigma\psi(x_i;t) \leq 0\}$$

$$P\{T_n^{**} < t\} = P\{\Sigma\psi(x_i;t) < 0\}$$

at the continuity points t of the left-hand side.

The distribution of the customary midpoint estimate $\frac{1}{2}(T_n^* + T_n^{**})$ is somewhat difficult to work out, but the randomized estimate T_n, which selects one of T_n^* or T_n^{**} at random

with equal probability, has an explicitly expressible distribution function

$$P\{T_n < t\} = \frac{1}{2}P\{\Sigma\psi(x_i;t) \leq 0\} + \frac{1}{2}P\{\Sigma\psi(x_i;t) < 0\}$$

It follows that the exact distributions of T_n^*, T_n^{**} and T_n can be calculated from the convolution powers of $\mathcal{L}\{\psi(x_i;t)\}$. Asymptotic approximations can be found by expanding $G_n = \mathcal{L}\{\Sigma_{i=1}^n \psi(x_i;t)\}$ into an asymptotic series.

Hampel (1973) noticed that the principal error term of the saddlepoint method seemed to reside in the normalizing constant (Standardizing the total mass of G_n to 1). He found out that it would be advantageous not to expand G_n or its density g_n, but rather expand $g_n^{'}/g_n$ and then determine the normalizing constant by numerical integration. This method he found out appears to give fantastically accurate approximations down to small sample sizes (n = 3 or 4).

2.2.1 Limiting distribution of T_n

Putting

$$\lambda(t) = \lambda(t, F) = E_F\psi(X, t) \tag{2.15}$$

if λ exist and is finite for at least one value of t, then it exists and is monotone (although not necessarily finite)for all t. this follows at once from the remark that

$\psi(X;t) - \psi(X;s)$ is positive for $t \leq s$ and hence has a well defined expectation (possibly $+\infty$).

If a proposition is made that there exists a t_0 such that $\lambda(t) > 0$ for $t < t_0$ and $\lambda(t) < 0$ for $t > t_0$, then T_n^* and T_n^{**} converge in probability and almost surely to t_0.

2.2.2 Minimax variance

Robust M-estimates are expected to have a relatively small and stable Asymptotic Variance (AV) of their score functions over the contaminated neighbourhood. Therefore an estimate with a small AV(ψ) should always be preferred, from a robustness point of view (other things being held constant).

Most of the estimators commonly studied are (under suitable regularity conditions) asymptotically normal about the center of symmetry, with asymptotic variance depending on the underlying distribution. The asymptotic variance therefore, becomes the criterion, for comparing the performance of different estimators for a given underlying distribution, and can also be used to compare the performance of a given estimator for different underlying distributions.

Huber (1964) formulated and solved some minimax problems, in which the estimators are judged by their asymptotic variance. He considered the case of contaminated neighbourhoods

$$\mathcal{F}_\epsilon^* = \{H : H(y) = (1-\epsilon)\Phi + \epsilon G, \} \tag{2.16}$$

where ϵ is fixed, Φ is a fixed symmetric, strongly unimodal distribution, H is a variable symmetric distribution and G is symmetric.

Huber's goal was to find the location M-estimate score function ψ that minimizes the maximum asymptotic variance over the symmetric contamination family (2.16). When the dispersion parameter is known (taken to be equal to one without loss of generality) the asymptotic distribution of location M-estimators is normal with mean equal to zero and variance $AV(\psi, H)$ given by

$$AV(\psi, H) = \frac{E_H\{\psi^2(x)\}}{[E_H\{\psi'(x)\}]^2} \tag{2.17}$$

If ψ is non decreasing then

$$\sup_{H\epsilon\mathcal{F}_\epsilon} AV(\psi, H) = \frac{1}{1-\epsilon} AV(\psi, \Phi) + \frac{\epsilon}{(1-\epsilon)^2}\gamma^2(\psi, \Phi) \tag{2.18}$$

where $\gamma(\psi, \Phi)$ is the gross-error sensitivity at the normal model.Therefore the maximum asymptotic variance under the normal ϵ-contamination family $\mathcal{F}_\epsilon$ is a linear combination of the normal asymptotic variance and gross-error sensitivity, with coefficients $\frac{1}{(1-\epsilon)}$ and $\frac{\epsilon}{(1-\epsilon)^2}$, respectively. This was arrived at as follows:-

Let

$$H = (1-\epsilon)N[0,1] + \epsilon H_1 \tag{2.19}$$

From equations (2.16) and (2.17) we have

$$AV(\psi, H) = \frac{(1-\epsilon)E_\Phi\{\psi^2(x)\} + \epsilon E_{H_1}\{\psi^2(x)\}}{[(1-\epsilon)E_\Phi\{\psi'(x)\} + \epsilon E_{H_1}\psi'(x)]^2} \tag{2.20}$$

Since ψ is non decreasing, it is clear that

$$(1-\epsilon)E_\Phi\{\psi^2(x)\} + \epsilon E_{H_1}\{\psi^2(x)\} \leq (1-\epsilon)E_\Phi\{\psi^2(x)\} + \epsilon\psi^2(\infty) \tag{2.21}$$

Moreover, since $\psi^{|}(x) \geq 0$ for all x, we can write

$$(1-\epsilon)E_\Phi\{\psi^{'}(x)\} + \epsilon E_{H_1}\{\psi^{'}(x)\} \geq (1-\epsilon)E_\Phi\{\psi^{'}(x)\} \tag{2.22}$$

Using eqn (2.19) and (2.20) we obtain

$$\begin{aligned} AV(\psi, H) &= \frac{(1-\epsilon)E_\Phi\{\psi^2(x)\} + \epsilon E_{H_1}\{\psi^2(x)\}}{[(1-\epsilon)E_\Phi\{\psi'(x)\} + \epsilon E_{H_1}\psi'(x)]^2} \\ &\leq \frac{(1-\epsilon)E_\Phi\{\psi^2(x)\} + \epsilon\psi^2(\infty)}{[(1-\epsilon)E_\Phi\{\psi'(x)\}]^2} \\ &= \frac{1}{(1-\epsilon)}\frac{E_\Phi\{\psi^2(x)\}}{[E_\Phi\{\psi'(x)\}]^2} + \frac{\epsilon}{(1-\epsilon)^2}\frac{\psi^2(\infty)}{[E_\Phi\{\psi'(x)\}]^2} \\ &= \frac{1}{1-\epsilon}AV(\psi, \Phi) + \frac{\epsilon}{(1-\epsilon)^2}\gamma(\psi, \Phi), \text{for all } H\epsilon\mathcal{F}_\epsilon \end{aligned}$$

Therefore

$$\sup_{H\epsilon\mathcal{F}_\epsilon} AV(\psi, H) \leq \frac{1}{1-\epsilon}AV(\psi, \Phi) + \frac{\epsilon}{(1-\epsilon)^2}\gamma^2(\psi, \Phi) \tag{2.23}$$

where $\gamma(\psi, \Phi)$ is the gross-error sensitivity of the M-estimate at the normal model.

2.2.3 Influence Curve of an M-estimate

If we put

$$F_{t,\varepsilon} = (1-\varepsilon)F_0 + \varepsilon F_1 \qquad\qquad 0 \leq \varepsilon \leq 1$$

then the Influence Curve IC $(x; F_0, T)$ is the ordinary derivative

$$\dot{T} = [\frac{d}{dt}T(F_t)]_{t=0} \qquad \text{with} \qquad F_1 = \delta_x$$

In particular, for an M-estimate, i.e. for functional $T(F)$ defined by

$$\int \psi(x; T(F))F(dx) = 0$$

we obtain by inserting F_t for F and taking the derivative (with $\psi'(x,\theta) = (\frac{\partial}{\partial\theta})\psi(x;\theta)$)

$$\int \psi(x; T(F_0)d(F_1 - F_0) + \dot{T}\int \psi'(x; T(F_0))F_0(dx) = 0$$

or

$$\dot{T} = \frac{\psi(x; T(F_0))F_1(dx)}{-\int \psi'(x; T(F_0))F_0(dx)}$$

after putting $F_1 = \delta_x$ we obtain

$$IC(x; F_0, T) = \frac{\psi(x; T(F_0))}{-\int \psi'(x; T(F_0))F_0(dx)}$$

therefore the influence curve of an M-estimate is simply proportional to ψ

2.2.4 Breakdown and Continuity Properties of M-estimates

Taking the location case, with $T(F)$ defined by

$$\int \psi(x - T(F))F(dx) = 0$$

we assume that ψ is nondecreasing, but not necessarily continous. Then

$$\lambda(t, F) = \int \psi(x - t)F(dx)$$

is decreasing in t, and increasing in F. $T(F)$ is not necessarily unique. we have

$T^* \leq T(F) \leq T^{**}$ with

$$T^* = \sup\{t|\lambda(t, F) > 0\}$$

$$T^{**} = \inf\{t|\lambda(t, F) < 0\}$$

Letting F range over all densities with $d_L(F_0, F) \leq \varepsilon$

The stochastically largest member of this set is the improper distribution F, (it puts mass ε at $+\infty$)

$F_1(x) = F_0(x - \varepsilon) - \varepsilon$ for $x > x_0 + \varepsilon$

$= 0$ for $x \leq x_0 + \varepsilon$ when x_0 is defined by

$$F_0(x_0) = \varepsilon$$

Thus

$$\lambda(t, F) \leq \lambda(t, F_1) = \int_{x_0}^{\infty} \psi(x - t + \varepsilon) F_0(dx) + \varepsilon\psi(\infty)$$

if we define

$$b_+(\varepsilon) = \sup\{T(F) | d_L(F_0, F) \leq \varepsilon\} = \inf\{t | \lambda(t, \lambda(t, F_1) < 0\},$$

$$b_-(\varepsilon) = \inf\{T(F) | d_L(F_0, F) \leq \varepsilon\} = \sup\{t | \lambda(t, \lambda(t, F_1) > 0\},$$

then the maximum asymptotic bias is

$$b_1(\varepsilon) = max\{b_+(\varepsilon) - T(F_0), T(F_0) - b_-(\varepsilon)\}$$

Breakdown. $b_+(\varepsilon) < b_+(1) = \infty$ holds iff

$$\psi(\infty) < \infty$$

and

$$\lim_{t \to \infty} \lambda(t, F_1) = (1 - \varepsilon)\psi(-\infty) + \varepsilon\psi(+\infty) < 0$$

For $b_{-}(\varepsilon)$ a similar relation holds, with the roles of $+\infty$ and $-\infty$ interchanged. It follows that the breakdown point is

$$\varepsilon^* = \frac{\eta}{1+\eta}$$

with

$$\eta = min\left\{-\frac{\psi(-\infty)}{\psi(\infty)} - \frac{\psi(\infty)}{\psi(-\infty)}\right\}$$

if $\psi(\infty) = -\psi(-\infty)$, we have $\varepsilon^* = \frac{1}{2}$

Continuity Properties

Putting $k = \varpi\varphi\psi(\infty) - \psi(-\infty)$. Then

$$\lambda(t+\varepsilon, F_0) \leq k\varepsilon \leq \lambda(t, F)\lambda(t-\varepsilon, F_0) + k\varepsilon$$

Hence if

(a) ψ is bounded, and

(b) $\lambda(t, F_0)$ has a unique zero at $t = T(F_0)$

then $T(F) \to T(F_0)$ as $\varepsilon \to 0$. It follows that T is weakly continous at F_0.The conditions are also necessary.

The following examples can be quoted:-

(i) The median corresponds to $\psi(x) = Sign(x)$, is a continous function at every F_0, whose Median is uniquely defined.

(ii) if ψ is bounded and strictly monotone, the corresponding M-estimate is everywhere continous.

2.3 M-estimators in Known and Unknown Scale Cases

The robust M-estimators are to be obtained from the three methods by minimizing

$$\sum_{j=1}^{n}\sum_{i=2}^{4}\rho(\theta_{i_{n_j}}-\theta) \tag{2.24}$$

where $\rho(\cdot)$ are some real valued non-constant functions. As special cases we note that $\rho(\theta)=\theta^2$ yields the sample mean, $\rho(\theta)=|\theta|$ yields the sample median, and $\rho(\theta)=-logf(\theta)$ yields the maximum likelihood estimator (where $f(x_n)$ and $f(x_{2_{n_j}})$ are density functions under the basic model when $\theta=0$). If $\rho(\cdot)$,is continous with derivative $\psi(\cdot)$ equivalently we estimate $\theta_{2_{n_j}}$, $\theta_{3_{n_j}}$ or $\theta_{4_{n_j}}$,$(j=1,2,\cdots,n)$ satisfying

$$\sum_{j=1}^{n}\sum_{i=2}^{4}\psi(\theta_{i_{n_j}}-\theta)=0 \tag{2.25}$$

Such estimators are called Maximum likelihood type estimators or M-estimators. The M-estimators of population mean will be worked out for two cases of the scale parameter in this research. These are:-

(a) $\sigma_2=\sigma$ and σ being a known constant for the non-contaminated population and

(b) $\sigma_2 \neq \sigma$, σ_2 being unknown and to be estimated from the Median Absolute Deviation (MAD)in the case of contaminated population.

When the scale parameter σ in the model is unknown, one proposal for defining the M-estimator is to solve

$$\sum_{i=1}^{n} \psi\left(\frac{X_i - \hat{\theta}_n}{\hat{\sigma}_n}\right) = 0$$

for $\hat{\theta_n}$, where $\hat{\sigma_n} = \hat{\sigma_n}(X_1, X_2, ..., X_n)$ is an estimator of σ that is unbiased when $H = \Phi$ under regularity conditions, $\hat{\theta_n}$ is consistent and $n^{\frac{1}{2}}(\theta_n - \theta)$ is asymptotically normal with mean 0 and variance $AV(\psi, H\psi)$ given by

$$AV(\psi, H) = \frac{E_H\{\psi^2(x)\}}{[E_H\{\psi'(x)\}]^2}$$

where $H = (1-\varepsilon)\Phi+\varepsilon G$ for the class of symmetric G and Φ being Normal distribution with fraction of contamination equal to ε.

The second proposal is that the Standard deviation σ can be estimated by the Median Absolute Deviation about the Median(MAD)

Defined:-

$$MAD(x) = MAD(x_1, x_2, \ldots, x_n)$$

$$= Med\{|x - Med(x)|\}$$

This estimator uses the sample median twice. First to get an estimate of the centre of the data in order to form the set of absolute residues about the sample median,$Med\{|x - Med(x)|\}$, and then compute the sample median of the absolute residuals. To make the MAD comparable to the standard deviation, we define the normalized MAD ("MADN") as

$$MADN(x) = \frac{MAD(X)}{0.6745} \tag{2.26}$$

The reason for this defination is that 0.6745 is the MAD of a Standard Normal random variable hence $N[\mu, \sigma^2]$ variable has $MADN = \sigma$

Chapter 3

Data Analysis and Discussion of Results

3.1 Introduction

A normal population made up of 1,000 observations was simulated. The mean θ and standard deviation σ were worked out to be 57.12 and 7.07 respectively.

The normal population simulated above was then contaminated by substituting some observations with some outlying ones. 10 samples made up of between 20 and 30 observations were randomly drawn from this population. Additionally 40 samples made up of between 31 and 100 observations were also randomly drawn. 100 (Monte-carlo) simulations were carried out for each sample size and the mean of each sample computed by the method of M-estimators of location using the Hubers', Hampels' and the Tukeys' biweight methods.

The means of samples of sizes 20, 30, 40, 80 and 100 were selected for standardization against the mean of the normal population.

3.2 Simulated Normal Population.

The graph of the normal population of 1,000 observations that was simulated, is displayed in Figure 3.1 showing the symmetrical nature of the data. The Boxplot graph of the same data is displayed in Figure 3.2. It is clear from both the Histogram and the Boxplot graphs that the data represents a population that is normally (symmetrical) distributed.

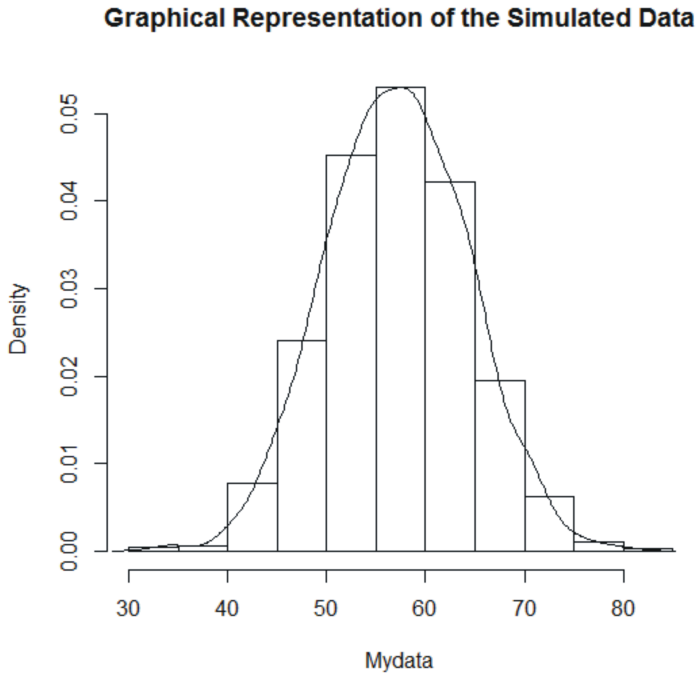

Figure 3.1: Graphical representation of the population data

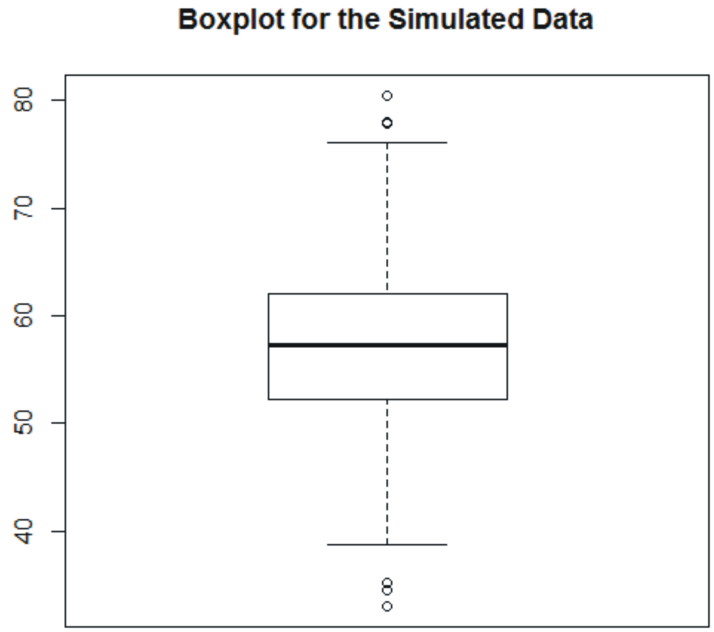

Figure 3.2: Boxplot for the population data

3.3 Contaminated Simulated Population.

The graph of the contaminated population made up of 1,000 observations that was simulated, is displayed in Figure 3.3 showing the asymmetrical nature of the data. The Boxplot graph of the same data is displayed in Figure 3.4. The two graphs (the Histogram and the Boxplot) reflect a distribution that is asymmetrical.

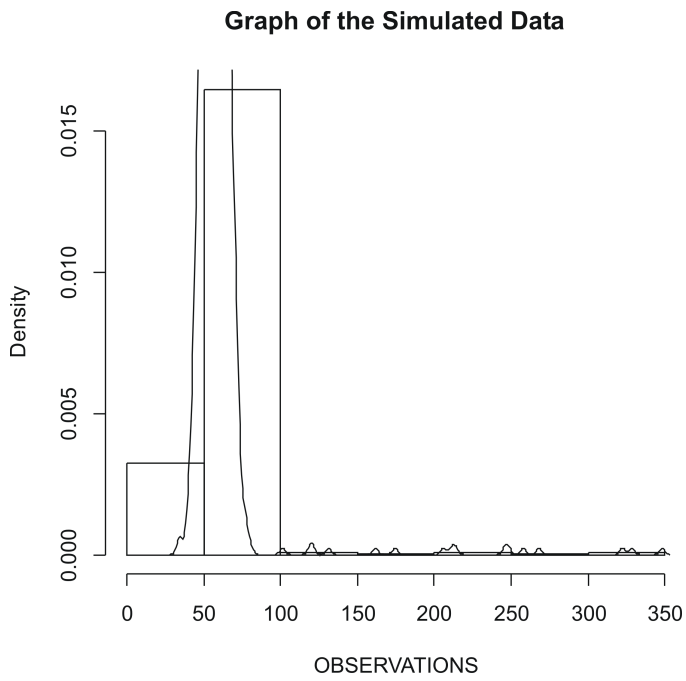

Figure 3.3: Graph of Contaminated Population Data

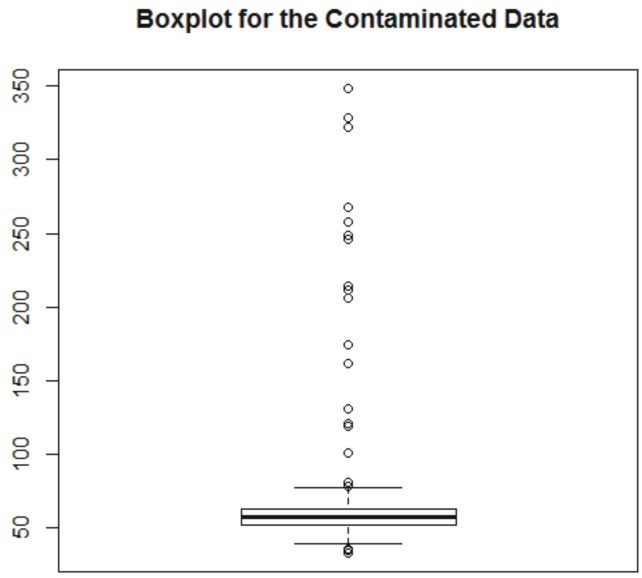

Figure 3.4: Boxplot of the Contaminated population data

Table 3.1: Means of 50 Samples obtained using Huber's method

57.11950	57.34220	57.42449	57.24741	57.26112
57.06087	57.50461	57.38623	57.31015	57.27117
57.24995	57.29495	57.35397	57.15724	57.51416
57.36389	57.46010	57.33477	57.15379	57.34676
57.21985	57.24996	57.29616	57.28087	57.19312
57.11424	57.26858	57.40136	57.32748	57.40146
57.22236	57.14820	57.40407	57.37232	57.34629
57.32588	57.34949	57.38325	57.08764	57.32397
57.27030	57.38862	57.31607	57.36645	57.20041
57.40205	57.23606	57.15831	57.28658	57.39675

Table 3.1 represents the means of all the 50 samples, with the first 10 entries (first 2 rows) being the means of samples with less than 30 observations while the remaining 40 entries (next 8 rows) represent the means of samples with 30 or more observations as computed using the Huber's M-estimator method as in the equation 3.1, with the default tuning constant k = 1.345 and represented by Figure 2.1.

$$\psi_k(x) = \begin{cases} k, & x \geq k \\ x, & -k \leq x \leq k \\ -k, & x \leq -k \end{cases} \tag{3.1}$$

Table 3.2 represents the means of all the 50 samples, with the first 10 entries (first 2 rows) being the means of samples with less than 30 observations while the remaining 40 entries (next 8 rows) represent the means of samples with 30 or more observations as computed using the Hampel's M-estimator method as in the equation 3.2 and Figure 2.2.

Table 3.2: Means of 50 Samples obtained using Hampel's method

57.07623	57.24936	57.23065	57.38398	57.34727
57.15379	57.41361	57.15234	56.96512	56.93930
57.18036	57.14235	57.23534	57.51623	57.19795
57.09058	57.24078	57.12886	57.03276	57.10090
57.14323	57.23542	57.12173	57.03192	57.15265
57.19998	57.12904	57.17906	57.13869	57.08432
57.21856	57.09883	57.01409	57.08487	57.27995
57.07450	57.12762	57.19370	57.19297	57.15244
57.08407	57.13912	57.38526	57.20075	57.03786
57.26777	57.05068	57.38188	57.02781	57.20859

Table 3.3: Means of 50 Samples obtained using Tukey's biweight method

57.03297	56.97048	57.26459	57.03896	56.91646
57.07780	57.28612	56.92406	57.16335	57.06043
57.09956	57.00824	57.13636	57.15974	56.96838
57.15823	57.34433	57.17320	56.93437	57.10593
57.30278	57.17985	57.16886	57.14057	57.20387
57.23310	57.14745	57.11665	57.03047	57.14319
57.02305	57.01375	57.06834	57.18964	57.10401
57.19894	57.14543	57.04326	57.23976	57.10610
56.89671	57.13506	57.22114	57.10015	57.13348
57.14856	57.24906	57.18234	57.03891	56.95155

$$\psi(x) = \begin{cases} |x|, & 0 \leq |x| \leq a, \\ a, & a \leq |x| \leq b, \\ a\left(\frac{c-|x|}{c-b}\right), & b \leq |x| \leq c \\ 0, & |x| > c \end{cases} \tag{3.2}$$

with the default tuning constant values being a = 1.7, b = 3.4 and c = 8.5. Table 3.3 represents the means of all the 50 samples, with the first 10 entries (first 2 rows) being the means of samples with less than 30 observations while the remaining 40 entries (next 8 rows) represent the means of samples with 30 or more observations as

computed using the Tukey's biweight M-estimator method as in the equation 3.3 and Figure 2.3.

$$\psi(x) = \begin{cases} \frac{1}{6}(1-u^2)^2 & \text{if} \quad |c| < 1 \\ \frac{1}{6} & \text{if} \quad |c| > 1 \end{cases} \tag{3.3}$$

where $u = \frac{x-T_n}{cS_n}$, T_n and S_n being location and scale scalars respectively while c = 4.685 is the default tuning constant.

3.4 Comparison of Means

The mean and standard deviation of the Normal population were 57.12 and 7.07 respectively while those of the contaminated population were 59.70 and 23.24 respectively. For ease of comparison of the means, samples of sizes 20,30,40,60,80 and 100 were picked.

From the results displayed in Table 3.4, it is clear that all the M-estimators of the means obtained for the various samples by using the Huber's, Hampel's and Tukey's Biweight methods are very close to the mean of the Normal population which was computed to be 57.12. The M-estimators were found to be in the range:-

(i) 57.060 - 57.327 for the Huber's,

(ii) 57.032 - 57.235 for the Hampel's,

(iii) 57.023 - 57.18 for the Tukey's Biweight.

Table 3.4: Means of Selected Samples

Sample size (n)	$\theta_{Huber's} = \theta_2$	$\theta_{Hampel's} = \theta_3$	$\theta_{Tukey's} = \theta_4$
20	57.120	57.076	57.033
30	57.060	57.154	57.078
40	57.250	57.235	57.180
60	57.222	57.219	57.023
80	57.281	57.032	57.140
100	57.327	57.139	57.030

3.5 Standardization of the M-estimators

The M-estimators of the means from the samples of various sizes are standardized against the mean of the normal population at 95% level of significance to check whether they are significantly different from the mean of the normal population by using the following equation:-

$$Z = \frac{\theta_i - \theta}{\frac{\sigma}{\sqrt{n}}} \qquad \text{for i = 2,3 and 4} \tag{3.4}$$

The results of the standardization of the three different types of M-estimators are displayed in Table 3.5 above.

All the computed Z-values are found to be within the 95% band hence the estimators are unbiased in their estimation of the population mean.

Table 3.5: Standard Z-values representing the M-estimators for the various samples

Sample size (H)	$\theta_{Huber's} = \theta_2$	$\theta_{Hampel's} = \theta_3$	$\theta_{Tukey's} = \theta_4$
20	0	-0.028	-0.055
30	-0.046	0.026	-0.033
40	0.145	0.129	0.067
60	0.112	0.108	-0.106
80	0.204	-0.111	0.025
100	0.293	0.027	-0.127

3.6 Application of M-estimators of Location on Real Data

3.6.1 Introduction

From the site www.OpenMV.net datasets, the researcher obtained real data that was asymmetrical in nature. This data was obtained from snapshot measurements on 27 variables from a distillation column measured over 2.5 years. The data was in 253 rows and 27 columns. On analyzing the data, a column was found to have some outlying observations and the researcher settled on it, for the study on real data with outliers. To obtain an estimate of the mean of this distribution, the 10% trimmed mean method was used.

From this population, 5 small samples of sizes of between 20 and 30 observations and 45 big samples of sizes of between 31 and 100 were randomly drawn. Each sample was then simulated 10 times and the means obtained by the three M-estimator of location methods i.e Huber's, Hampel's and the Tukey;s biweght methods.

Table 3.6: Means of Samples from Huber's method

177.4951	175.4355	176.2758	174.7935	175.8405
175.8631	176.3851	175.8908	176.3140	175.5947
176.0795	176.6220	175.6716	175.7515	175.4025
175.9103	176.4670	176.9966	176.0200	175.3716
176.3678	176.0458	175.9509	176.9138	176.1327
176.1367	175.7818	176.8226	176.0031	176.2237
176.2381	175.7063	176.5173	176.1051	176.3488
175.2528	175.9868	175.8395	176.1663	176.3270
177.4753	176.1817	176.2974	176.0293	177.1053
175.4414	175.5031	175.3231	176.0888	175.9964

The asymmetry of the data is clearly displayed in the Histogram and the Boxplot in Figures 3.5 and 3.6 respectively.

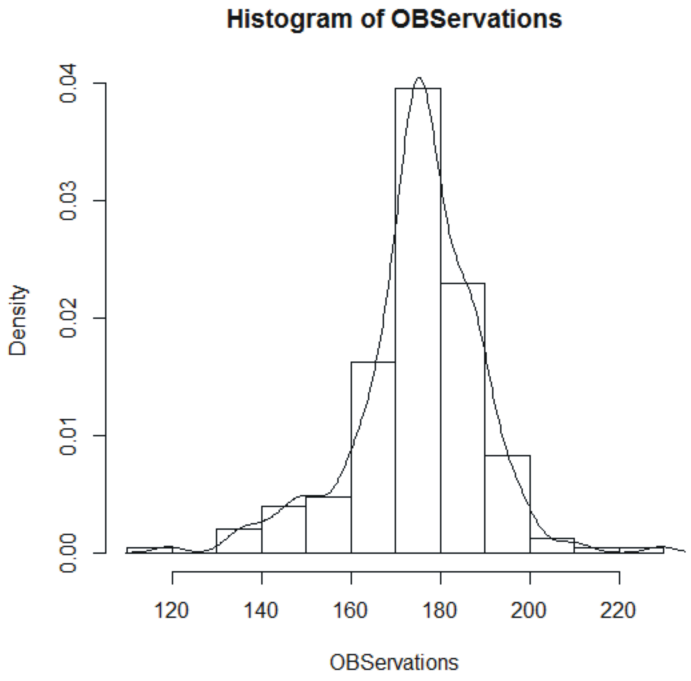

Figure 3.5: Histogram of the Real data

Table 3.6 represents the means of 50 samples, with the first 5 entries (first row) being the means of the samples of size less than 30 observations while the remaining

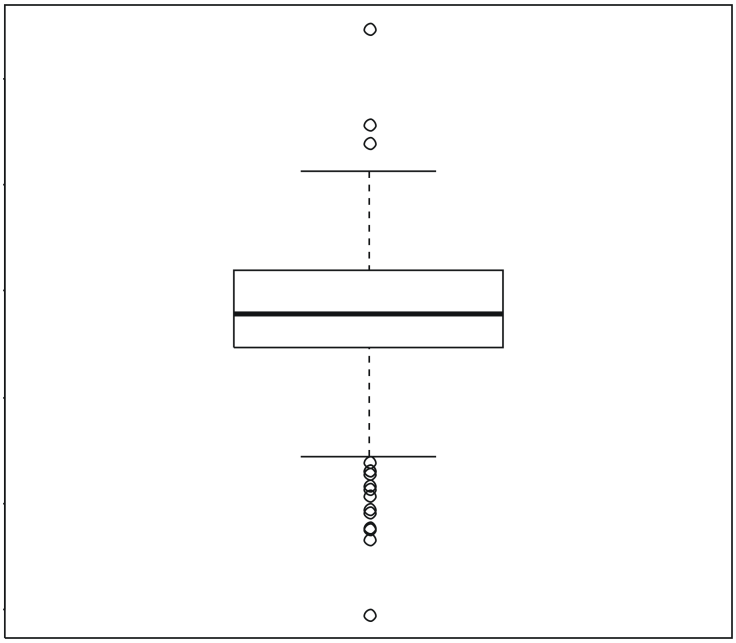

Figure 3.6: Boxplot of the Real Data

45 entries(next 9 rows) represent the means of samples of size 30 or more observations, as computed using the Huber's M-estimator method as in equation 2.12 and Figure 2.1.

Table 3.7 represents the means of 50 samples, with the first 5 entries (first row) being the means of the samples of size less than 30 observations while the remaining

Table 3.7: Means of Samples from Hampel's method

177.0243	175.9064	175.2922	175.9605	175.9837
175.6586	175.3788	175.7766	176.4889	177.0499
175.9026	175.8441	175.7062	175.3348	175.2027
175.4008	176.0106	175.6328	175.2916	176.3392
175.2638	175.5187	175.7842	175.6309	176.0309
175.5366	175.5896	175.5131	175.8841	176.1617
175.8574	175.3184	176.3015	176.1865	175.6418
175.9774	176.3089	176.6908	175.3701	175.4683
174.6539	175.5683	175.5702	175.8037	174.7534
176.1824	175.6663	177.1889	175.4363	176.1903

Table 3.8: Means of Samples from Tukey's method

176.9638	176.6007	175.4793	175.0982	176.5607
176.3381	177.1205	176.0661	176.3760	176.7265
176.1653	176.9066	176.4512	175.5637	176.6014
175.9070	176.4279	176.3051	176.3925	176.3015
176.5895	176.2543	175.7179	177.0554	175.7841
175.7496	175.9522	176.4447	176.4838	176.7159
176.1426	177.1279	176.4747	176.1180	175.7919
177.5410	176.6132	176.2412	176.0652	175.9008
176.3684	176.5097	176.3720	176.1844	175.8167
175.1696	176.0641	176.4008	176.2391	176.4580

45 entries(next 9 rows) represent the means of samples of size 30 or more observations as computed using the Hampel's M-estimator method as in equation 2.13 and Figure 2.2.

Table 3.8 represents the means of 50 samples, with the first 5 entries (first row) being the means of the samples of size less than 30 observations while the remaining 45 entries(next 9 rows) represent the means of samples of size 30 or more observations as computed using the Tukey's M-estimator method as in equation 2.14 and Figure 2.3.

3.6.2 Comparison of the Means

The 10% trimmed mean θ of the real data composed of the 253 observations was 176.064 while the mean and the standard deviation σ of the asymmetrical data were 175.2 and 13.94 respectively; The Normalized Median Absolute Deviation (MADN) as in equation 2.26, was computed to give a better estimate of the standard deviation for the data and was obtained as 10.86. For ease of comparison of the means, samples

Table 3.9: Means of Selected Samples

Sample size (n)	$\theta_{Huber's} = \theta_2$	$\theta_{Hampel's} = \theta_3$	$\theta_{Tukey's} = \theta_4$
20	177.495	177.024	176.964
30	175.863	175.659	176.338
40	176.046	175.519	176.254
60	176.238	175.857	176.143
80	176.914	175.631	176.055
100	176.003	175.884	176.484

of sizes 20,30,40,60,80.and 100 were adopted.

From the results displayed in Table 3.9 it is clear that all the M-estimators of the mean obtained for the various samples by using the Huber's, Hampel's and Tukey's Biweight methods yields unbiased estimates of the mean obtained by 10% trimmed mean method for the real data which was computed as 176.064. For the asymmetrical real data the M-estimators were found to be in the range:-

(i) 175.863 - 177.495 for the Huber's,

(ii) 175.519 - 177.024 for the Hampel's,

(iii) 176.055 - 176.964 for the Tukey's Biweight.

3.7 Standardization of the M-estimators

The M-estimators of the means from the samples of various sizes were standardized against the 10% trimmed mean of the real data and tested at 95% level of significance

Table 3.10: Standard Z-values representing the M-estimators for the various samples

Sample size (n)	$\theta_{Huber's} = \theta_2$	$\theta_{Hampel's} = \theta_3$	$\theta_{Tukey's} = \theta_4$
20	0.589	0,395	0.371
30	-0.101	-0.204	0.138
40	-0.010	-0.317	0.111
60	0.124	-0.148	0.056
80	0.700	-0.357	0.816
100	0.056	-0,166	0.387

to check whether they are significantly different from the 10% trimmed mean of the real data by using the following equation:-

$$Z = \frac{\theta_i - \theta}{\frac{\sigma_{mad}}{\sqrt{n}}} \qquad \text{for i = 2,3 and 4} \tag{3.5}$$

where σ_{mad} is substituted with the value of MADN in the above standardization formula. The results of the standardization of the three different types of M-estimators are displayed in Table 3.10.

All the computed Z-values are found to be within the 95% band hence the estimators are unbiased in their estimation of the population mean.

Chapter 4

Summary, Conclusions and Recommendations

4.1 Introduction

In this chapter we outline the general achievements and conclusions of the project. we also suggest areas for further research which have emerged during the course of our study.

4.2 Summary from asymmetrical population

From the results of table 3.4 the range of the means for the samples considered under each of the three methods were found to enclose the mean of the non-contaminated (Normal) population. Even though they were each obtained from the contaminated

population it became clear from this study that any of the three methods can be relied upon to yield an unbiased estimate of a population mean.

For this study, the M-estimator of the mean by the Tukey's Biweight method yielded the narrowest range and hence was the best. The Hampel's Method yielded the second best range while the range from the Huber's method was the widest and hence considered the weakest.

4.2.1 Standardized Z-values for the means of selected samples

The M-estimators of samples of various sizes, were standardized at 95% level of significance to test whether they would be significantly different from the mean of the Normal population. The M-estimators would be considered significantly different from the mean of the normal population if they yielded a Z-value outside the range -1.96 to +1.96 (The lower and upper limits at the 5% level of significance).

From Table 3.5 no value of Z (Standard normal value) was found to be outside the range stated above, for all the three methods of estimating the population mean from contaminated data. All the three methods yielded unbiased estimates of the population mean and can therefore be relied upon in the estimation of the population mean.

4.3 Summary from Real Data case

From the results of table 3.9 the range of the means of the samples from all the three methods were found to enclose the trimmed mean of the asymmetrical population. The M-estimator of the mean by the Tukey's biweight method yielded the narrowest range and hence was the best. The Hampel's Method yielded the second best range while the range from the huber's method was the widest and hence turned out as the weakest.

From these results, the three M-estimation methods of the mean for the real data can be relied upon to yield an unbiased estimate of a population mean.

4.3.1 Standardized Z-values for the means of selected samples

When the M-estimators obtained using the three estimation methods are standardized at 95% the Z-values obtained were all within the range -1.96 to +1.96 as shown in Table 3.10. It can therefore be concluded that the M-estimation methods can be relied upon to yield unbiased estimates of the population mean when a few outliers are present in a distribution.

4.4 Conclusion

Unlike in the normal distribution where the estimation of a parameter improves as the sample size increases in this study the estimate of the mean of an asymmetrical population using the M-estimators is irregular in that it oscillates back and forth and does not seem to approach the mean of the population as the sample size increases. However the means obtained in this study by the three M-estimators, the Huber's the Hampel's and the Tukey's biweight have all been found to be robust in the estimation of the population mean.

4.5 Further research

The methods compared in the estimation of the mean of a population for this research, relied on allocating the same weight for outliers beyond a certain value k (the tuning constant) for the Huber's method(as reflected in the $\psi-$functions as in Figure 2.1) and secondly, allocating a weight of zero for outliers beyond a certain tuning constant (as reflected in the $\psi-$functions as in Figures 2.2 and 2.3) for the Hampel's and Tukey's biweight methods respectively.

Further research should be carried out with an aim of allocating decreasing weights for observations as their distance increases from the majority of the observations, rather than allocating them the same weight as the Huber's method does or ignoring the observations totally like the Hampel's and the Tukey's biweight do.

Further work may also be carried out using one of the heavy tailed distribution such as the log normal distribution to establish the behaviour of robust M-estimation of location in the estimation of population mean in the presence of outliers.

References

Akritas., M. G. (1991). *Random Sampling from non-homogeneous populations.*, volume 86. American Statistical Association. 10

Andrews, D., Bickel, P., Hampel, F., Huber, P., Rogers, W., and Tukey, J. (1972). Robust estimates of location: Survey and advances. 18

Basak, I. (1998). *Robust M-Estimation in Discriminant AnalysisAuthor*, volume 60. Springer on behalf of Indian Statistical Institute. 11

Beran, J. (1991). *M Estimators of Location for Gaussian and Related Processes With Slowly Decaying Serial Correlations*, volume 86. American Statistical Association. 11

Birnbaum, A. and Mike, V. (1970). *Asymptotically Robust Estimators of Location*, volume 65. American Statistical Association. 11

Crow, E. L. and Siddiqui, M. (1967). *Robust Estimation of Location*, volume 62. Journal of the American Statistical Association. 9

Fung, K. Y., Lee, H., and Tajuddin, I. (1985). *Some Robust Test Statistics for the Two-Sample Location Problem*, volume 34. wiley for the Royal Statistical Society. 10

Hampel, F. (1973). Some small sample asymptotics. hajek, j.(ed.), proceedings of the prague symposium on asymptotic statistics. 21

Hampel, F. (1974). The influence curve and its role in robust estimation. 13

Huber (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, pages 35, 73 101. 10

Huck, S. (2000). Reading statistics and research. *3rd Edition, New York*. 3

Kagan, A. (1966). *On the Estimation Theory of Location Parameter*, volume 28. Springer on behalf of the Indian Statistical Institute. 12

Lenth, R. V. (1981). *Measures of Location for Directional Data*, volume 23. Taylor and Francis ltd. 10

Stefanski, L. A. and Boos, D. D. (2002). *The Calculus of M-Estimation*, volume 56. American Statistical Association. 11

Thall, P. F. (1979). *Huber-Sense Robust M-Estimation of a Scale Parameter, with Applications to the Exponential Distribution*, volume 74. American Statistical Association. 10

Wang, Y.-G., Lin, X., Zhu, M., and Bai, Z. (2007). *Robust Estimation Using the Huber Function with a Data-Dependent Tuning Constant*, volume 16. American Statistical Association, Institute of mathematical Statistics and Interface Foundation of America. 11

W.Osborne, J. and Overbay, A. (2004). Practical assesment, research and evaluation. *North calorina state University*. 3