Estimating a Finite population Mean in Presence of Outliers under stratified random sampling

Wanjiru, Magdalyne Njeri

170435

Submitted in partial fulfilment of the requirements for the degree of Master of Science in Statistical Science of Strathmore University

Strathmore Institute of Mathematical Sciences
Strathmore University
Nairobi, Kenya

June 2025

Declaration

I declare that this work has not been previously submitted and approved for award of a degree by this or any other University. To the best of my knowledge and belief, the thesis contains no material previously published or written by another person except where due reference is made in the thesis itself.

Name: **Wanjiru Magdalyne Njeri**

Signature: .....

Date: May 20, 2025

Approval

The dissertation of Wanjiru Magdalyne Wanjiru was reviewed and approved by the following:

Dr. Christopher Ouma Onyango,
Lecturer, Strathmore Institute of Mathematical Sciences,
Strathmore University

Prof. Godfrey Madigu,
Dean, Strathmore Institute of Mathematical Sciences,
Strathmore University

Prof. Bernard Shibwabo,
Director of Graduate Studies,
Strathmore University

Abstract

Estimation of finite population parameters in the presence of outliers has been studied by many researchers in Sample Survey Theory. Outliers are known to skew the estimator of the population parameters resulting in bias of estimators. To improve on efficiency of the mean estimator, we propose a modified estimator of the population mean in stratified random sampling which incorporates both the sample mean and weighted mean of each stratum. Simulation studies are performed to compare the estimators obtained using the proposed approach and their rival estimators under stratified random sampling technique. The proposed estimator of the mean demonstrated improved coverage rates compared to rival estimators, by having lower variance and competitive Mean Squared Error in certain cases.

KEY WORDS: Outliers, Stratified random sampling, Skewed.

Table of contents

Acknowledgement

First and above all, I am grateful to the Almighty God for His provision, grace, and for granting me good health throughout the study period.

I am grateful to my husband, Mr. Eugene Alfred Orocho, and our child Dyne Blessing Orocho for their unwavering support and cheering me on throughout the study period.

I am grateful to Mr.Peter Mwangi,Director Alliance Realtors and my Grandmother for their immeasurable support throughout the study period.

I am indebted to my supervisor, Dr. Christopher Ouma Onyango for his availability, kind guidance and support without which this thesis would not have been a success.

I am grateful to my classmates for their support and encouragement.

Dedication

This proposal is dedicated to my husband, whose unwavering support and encouragement have shaped my journey.

Chapter 1

Introduction

1.1 Background to the Study

Stratified random sampling is a widely used technique that improves the precision of estimates by dividing the population into distinct strata that are internally homogeneous. However, the presence of outliers can severely affect the performance of estimators, leading to biased results and reduced accuracy. Outliers can distort the results of traditional estimation methods such as the sample mean. According to Tukey (1977), outliers can influence the mean to such an extent that it no longer represents the central tendency of the data. This distortion is particularly pronounced in finite populations, where a few extreme values can significantly skew the overall mean due to the limited number of observations.

In finite populations, each observation carries more weight compared to infinite populations, making the effect of outliers more substantial and consequential. A study by Rousseeuw and Leroy (1987) highlights that robust statistical methods are essential in the presence of outliers as standard estimators such as the sample mean can yield misleading results. Stratified sampling can mitigate some of the negative impacts of outliers as each stratum can be treated separately, allowing for more localized estimations. However, outliers can still exist within strata and can influence the stratum means which in turn affects the overall population mean.

Barnett and Lewis (1994) emphasize that even a single outlier in a stratum can disproportionately affect the pooled estimate of the population mean particularly when the sample size within the stratum is small. Several studies have explored the impact of outliers on different estimation methods. Chen and Chen (2006) examined the performance of various estimators under stratified sampling with a focus on outlier resistance. Their findings indicated that robust estimators such as the trimmed mean and Winsorized mean, significantly outperform

traditional methods when outliers are present. These robust methods minimize the influence of extreme values providing more reliable estimates of the population mean.

Moreover, studies have shown that employing techniques like data transformation or outlier detection and removal can also be effective strategies in improving estimator performance. Hodge and Bickel (2005) proposed a combination of stratification and robust statistical techniques, demonstrating that this approach could yield more accurate mean estimates in the presence of outliers. The application of machine learning methods to identify and manage outliers has gained traction in recent years. Techniques such as clustering and anomaly detection can aid in distinguishing between genuine observations and outliers allowing for more informed statistical decisions. In this context, Koller and Friedman (2009) emphasized the importance of integrating statistical and computational approaches to enhance the robustness of estimators, especially in finite population settings where each data point has a significant impact on estimates.

1.2 Statement of the Problem and justification

Estimating the mean of a finite population is crucial in various fields including economics, healthcare and social sciences. Traditional methods often assume normally distributed data and do not adequately address the influence of outliers in the proposed estimator developed. According to Osborne and Gunter (2001),many researchers rarely report checking for outliers in their work .The mean estimator is sensitive to outlier values as it leads to biased estimators. Studies by Chen and Chen (2006) and Wang and Wu (2010) have been seen to focus on robust statistical methods in the presence of outliers especially on homogeneous simple random sampling without replacement (SRSWOR). However in practice heterogeneous population are also commonly encountered and in such situation stratified random sampling is recommended. Stratified random sampling not only provides more representative samples in such populations but also allows for localized handling of outliers within strata.

The need to focus on finite populations is crucial because estimators behave differently under finite population models compared to infinite population models, particularly when dealing

with outliers. In finite populations, the distortion caused by a single outlier cannot be diluted by increasing the population size or sample size indefinitely. Therefore, this study is essential in developing methodologies tailored to finite population settings, ensuring more accurate population mean estimates and improving the overall quality of statistical inference.

1.3 Research Objectives

1.3.1 General Objective

To develop an estimate for finite population mean in the presence of outliers under stratified random sampling.

1.3.2 Specific Objectives

1. To derive an estimate of finite population mean using weighted mean of each stratum under stratified random sampling.
2. To determine the asymptotic properties of the developed estimator of the finite population mean.
3. To perform simulation study to compare the performance of the proposed estimator with rival estimators of the population mean.

1.4 Significance of the Study

This study sought to improve the precision of the estimate of the population mean in the presence of outlier values.

1.5 Scope of the study

The study was limited to determining the estimate of the mean of a finite population in the presence of outliers under stratified random sampling. This estimator of the mean was based on the sample mean and weighted mean of each stratum.

Chapter 2

Literature Review

2.1 Introduction

Precision of estimators for population mean is crucial in statistics especially when working with finite populations. Stratified random sampling is a popular method that improves estimates by dividing a population into subgroups (strata) that share similar characteristics. This technique enhances representation and precision in estimates. However, the presence of outliers in stratified samples can introduce significant bias, potentially compromising the accuracy of the population mean estimation. Therefore, researchers have explored various methods to mitigate the effects of outliers while maintaining the precision of estimators.This review provides relevant literature on estimation of finite population parameters in the presence of outliers

2.2 Population mean estimate under stratified random sampling in the presence of outliers

The theoretical basis for estimating the population mean using stratified random sampling is primarily attributed to Neyman (1934), whose work introduced stratified sampling as an efficient alternative to simple random sampling design in heterogeneous population. The work showed that by partitioning the population into strata that are internally homogeneous and then sampling from each stratum, one could obtain more precise estimates of population parameters including the population mean. However, the work by Neyman (1934) did not factor in the presence of outlier values. According to Cochran (1977) foundational work,

"Sampling Techniques," provides a comprehensive overview of stratified random sampling methods. The study discussed how stratification can improve the accuracy of population estimates by ensuring that different segments of the population are adequately represented. However, Cochran (1977) also acknowledged that outliers within strata can disproportionately affect the estimates. These insights set the stage for further research into how outlier handling can be integrated into stratified sampling designs.

The issue of outliers and their impact on statistical estimations has been widely studied. In "Robust Regression and Outlier Detection," Rousseeuw and Leroy (1987) explored various robust techniques for dealing with outliers, focusing on regression analysis but providing insights relevant to stratified sampling. The findings showed that traditional estimators are highly sensitive to outliers, leading to biased results. Robust methods such as the M-estimator could be applied to stratified sampling to improve the precision of estimates derived from different strata. Building on these findings,Tukey (1977)proposed exploratory data analysis techniques that could help detect and manage outliers before estimating population parameters. These early contributions paved the way for contemporary studies on robust estimation in stratified sampling frameworks. More study by Huber (1981)explored alternative robust estimation techniques in stratified sampling. Robust statistical approaches that minimize the influence of extreme values on estimation procedures was introduced.

Further study by Chen and Chen (2004) on outlier detection in stratified samples emphasized on the challenges posed by extreme values. Several methods were examined for identifying outliers within strata, including robust statistical tests and graphical techniques. The findings indicated that failing to address outliers could result in significant bias in the overall population mean estimate. The use of robust methods alongside stratified sampling was recommended to enhance the accuracy and reliability of estimates.

In the empirical study, Chen and Chen (2006) evaluated the performance of different estimators in stratified sampling with outliers. Simulations are conducted to compare traditional methods such as the sample mean with robust alternatives like the trimmed mean. The findings revealed that robust methods provided more reliable estimates when outliers were present, particularly in small strata where the influence of outliers was magnified. This

research underscores the necessity of employing robust techniques in stratified sampling designs.

Wang and Wu (2010) examined robust statistical methods in the context of stratified sampling focusing on the influence of outliers. A modified estimators of the population mean specifically designed to mitigate the impact of outliers within strata was developed. The simulations demonstrated that these modified methods significantly improved the accuracy of population mean estimates compared to conventional techniques. The work highlighted the importance of adapting statistical methods to account for outliers in stratified sampling frameworks.

In addition to these contributions, other researchers such as Singh and Sedory (2015), explored the use of weighted means and Winsorized estimators in stratified sampling. Their research demonstrated that weighting observations differently based on their potential influence could yield more stable estimates of the population mean. This approach aligns with the idea that outliers should not be entirely discarded but rather down-weighted in calculations. More recent studies,have investigated machine learning and computational approaches in stratified sampling, particularly in handling outliers. For example, Ghosh (2018) examined how machine learning-based anomaly detection techniques could be integrated with stratified sampling frameworks to improve robustness. The study highlighted how algorithms such as support vector machines (SVM) and random forests could assist in identifying and mitigating the influence of extreme values. Their findings suggested that combining statistical and machine learning methods could offer a more comprehensive approach to dealing with outliers in stratified sampling.

Lately, Aloma and Al-Sharif (2020) conducted a study on the effects of outliers in stratified random sampling and proposed an outlier detection framework tailored for this context. The methodology involved using robust statistical techniques to identify and manage outliers effectively. The results indicated that applying the proposed framework improved estimation accuracy significantly particularly in populations with a high proportion of outliers.

Moreover, recent advancements in Bayesian statistics have also been applied to robust estimation in stratified sampling. Johnson (2021) introduced a Bayesian hierarchical model

that accounts for outliers within stratified samples. The model assigns probabilistic weights to observations based on their likelihood of being outliers, thus reducing their influence on the final population mean estimate. The results showed that Bayesian approaches could provide flexible and adaptive estimation methods, particularly when dealing with highly variable data.

The study by Wang and Xu (2023)) focused on improving the estimation of the population mean in stratified random sampling when the population contains outliers or exhibits non-normal distributions. The study proposed a hybrid estimator that combines the sample mean and the trimmed mean in stratified sampling to provide more robust population mean estimates in the presence of outliers. Through simulations, the study demonstrated that the hybrid estimator improves the accuracy of population mean estimates compared to traditional methods, particularly when data exhibits skewness, heavy tails or significant outliers. However by trimming the data outliers are removed, this could be suitable when the outliers could be as a result of errors but not when the outliers belong to the distribution. It may also not be appropriate for small sample sizes since removing a percentage of the data points potentially resulting in a loss of information and biased estimation.

Motivated by the study of Wang and Xu (2023), this study sought to estimate the mean of the study population under stratified sampling in presence of outliers using sample mean and weighted mean.By incorporating these methods, the study aimed at improving the accuracy of population mean estimation, particularly in datasets characterized by significant variability and the presence of extreme values.

Chapter 3

Methodology

3.1 Introduction

This chapter presents a hybrid estimator for estimating the population mean in stratified random sampling.The estimator incorporates sample and weighted means, with an adjustment factor (α) based on outliers.

The asymptotic properties of the estimator are also examined to check if it is unbiased and whether it have better precision.

3.2 Proposed Estimator

According to Wang and Xu (2023), the population is divided into H strata. Each stratum h has a sample mean $\bar{y}_h$ and a trimmed mean T_h , which is calculated by removing a fixed percentage.The hybrid estimator for the population mean is expressed as:

$$\bar{Y} = \sum_{h=1}^{H} P_h[w\bar{Y}_h + (1-w)T_h]$$

where; P_h is the proportion of the population in stratum h

$P_h = \frac{N_h}{N}$

w is the weight that determines the relative contribution of the sample mean and the trimmed mean within each stratum.

The study by Wang and Xu (2023)) focused on trimming the data to remove outliers, this is not suitable when the outliers belong to the distribution and for small sample sizes since this resulted to loss of information.Considering the shortcomings of the work by Wang and Xu (2023), this study proposed a Hybrid estimator that combines the sample mean and weighted mean in stratified random sampling, given by:

$$\hat{Y} = \sum_{h=1}^{H} P_h[\alpha \bar{Y}_h + (1-\alpha)W_h]$$

where; W_h is the weighted mean of stratum h with outlier

$$W_h = \frac{\sum_{i \text{ is in stratum with outlier}} X_i \cdot \beta_i}{\sum_{i \text{ is in stratum with outlier}} \beta_i}$$

$$\beta_i = \frac{1}{1 + |\frac{X_i - \mu_h}{\sigma}|}$$

$$\alpha = \begin{cases} 1 & \text{if the stratum has no outlier,} \\ 0 & \text{otherwise} \end{cases}$$

α is the weight that determines the relative contribution of the sample mean and the weighted mean within each stratum.

X_i is the data point in the stratum that have outlier,μ_h is sample mean exclusive of outliers,σ is the standard deviation.

The following notations shall be adopted throughout the study:

H = the number of strata

N_h = number of population units in stratum h, h = 1, 2, . . . , H

$N = \sum_{h=1}^{H} N_h$ = the number of units in the population

n_h = number of sampled units in stratum h, h = 1, 2, . . . , H

$n = \sum_{h=1}^{H} n_h$ = the total number of units sampled

y_{hi} = the y-value associated with unit i in stratum h

$\bar{y}_h$ = the sample mean for stratum h

$\bar{y}_h = \sum_{i=1}^{N_h} xi/N_h$

$\bar{Y}$ = the sample mean under stratified random sampling

$P_h = \frac{N_h}{N}$

3.3 Asymptotic Properties of the Hybrid Estimator

3.3.1 Expectation of Proposed Estimator of Population Mean

The hybrid estimator combines the sample mean and weighted mean in a stratified sampling. To calculate its expected value of;

$$\hat{Y} = \sum_{h=1}^{H} P_h \left[\alpha \bar{Y} + (1-\alpha) W_h\right], \tag{3.1}$$

where:

P_h is the proportion of the population in stratum h,

$\bar{Y}_h$ is the sample mean of stratum h without outlier,

$W_h = \frac{\sum_{i \text{is in stratum with outlier}} X_i \cdot \beta_i}{\sum_{i \text{is in stratum with outlier}} \beta_i}$ is the weighted mean of stratum h with outlier,

$\beta_i = \frac{1}{1+|\frac{X_i - \mu_h}{\sigma}|}$,

where X_i is a data point in the stratum with outlier and μ_h is stratum mean exclusive of outliers.

$$\alpha = \begin{cases} 1 & \text{if the stratum has no outlier,} \\ 0 & \text{otherwise} \end{cases}$$

The expected value of the estimator is:

$$\mathbb{E}[\hat{Y}] = \mathbb{E}\left[\sum_{h=1}^{H} P_h(\alpha\bar{Y}_h + (1-\alpha)W_h)\right]. \tag{3.2}$$

$$\mathbb{E}[\hat{Y}] = \sum_{h=1}^{H} P_h\mathbb{E}[\alpha\bar{Y}_h + (1-\alpha)W_h]. \tag{3.3}$$

$$\mathbb{E}[\hat{Y}] = \sum_{h=1}^{H} P_h(\alpha\mathbb{E}[\bar{Y}_h] + (1-\alpha)\mathbb{E}[W_h]) \tag{3.4}$$

Handling each component of the summation in equation (3.4) at a time we have;

(a) Expected Value of $\bar{Y}_h$

The sample mean $\bar{Y}_h$ of stratum h is an unbiased estimator of the population mean in stratum h, denoted by μ_h:

$$\mathbb{E}[\bar{Y}_h] = \mu_h. \tag{3.5}$$

(b) Expected Value of W_h

The weighted mean W_h involves the weights β. Assuming the weights β are random variables dependent on X, the expected value is:

$$\mathbb{E}[W_h] = \mathbb{E}\left[\frac{\sum_{x_i \text{ is in stratum with outlier}} X_i \cdot \beta}{\sum_{x_i \text{ is in stratum with outlier}} \beta}\right] \approx \mu_{ih}. \tag{3.6}$$

Substituting the components in equation(3.4) we have:

$$\mathbb{E}[\hat{Y}] = \sum_{h=1}^{H} P_h \left[\alpha[\mu_h] + (1-\alpha)\mu_{ih}]\right]. \tag{3.7}$$

$$\mathbb{E}[\hat{Y}] = \sum_{h=1}^{H} P_h \left[\alpha\mu_h + \mu_{i_h} - \alpha\mu_{i_h}\right]. \tag{3.8}$$

Suppose $\mu_{ih} = \mu_h, then$;

$$\mathbb{E}[\hat{Y}] = \sum_{h=1}^{H} P_h \mu_h. \tag{3.9}$$

Therefore,the expected value of the hybrid estimator $\hat{Y}$ equals the true population mean under the assumption that the weighted mean W_h and sample mean $\bar{Y}_h$ are unbiased, and α correctly accounts for outliers. Thus, the estimator is unbiased.

3.3.2 Mean Squared Error

The Mean Squared Error (MSE) of an estimator measures the average squared difference between the estimator and the true value of the parameter. For the hybrid estimator, the MSE can be calculated as:

$$\text{MSE}(\hat{Y}) = \mathbb{E}[(\hat{Y} - \bar{Y})^2], \tag{3.10}$$

where:

$\hat{Y}$ is the hybrid estimator,

$\bar{Y}$ is the true population mean.

Using the bias-variance decomposition, we can express the MSE as:

$$\text{MSE}(\hat{Y}) = \text{Var}(\hat{Y}) + \text{Bias}(\hat{Y})^2. \tag{3.11}$$

We breakdown the Variance and Bias

Variance of $\hat{Y}$

The hybrid estimator is given as:

$$\hat{Y} = \sum_{h=1}^{H} P_h \left(\alpha \bar{Y}_h + (1-\alpha) W_h\right). \tag{3.12}$$

The variance of $\hat{Y}$ can be computed as:

$$\text{Var}(\hat{Y}) = \text{Var}\left(\sum_{h=1}^{H} P_h \left(\alpha \bar{Y}_h + (1-\alpha) W_h\right)\right). \tag{3.13}$$

Then,

$$\text{Var}(\hat{Y}) = \sum_{h=1}^{H} P_h^2 \cdot \text{Var}\left(\alpha \bar{Y}_h + (1-\alpha) W_h\right) \tag{3.14}$$

Using the variance of a linear combination:

$$\text{Var}(\hat{Y}) = \sum_{h=1}^{H} P_h^2 \left[\alpha^2 \text{Var}(\bar{Y}_h) + (1-\alpha)^2 \text{Var}(W_h) + 2\alpha(1-\alpha)\text{Cov}(\bar{Y}_h, W_h)\right] \tag{3.15}$$

Assuming $\bar{Y}_h$ and W_h are uncorrelated, that is;

$$\mathbb{E}[\bar{Y}, W_h] = \mathbb{E}[\bar{Y}_h] \cdot [\mathbb{E} W_h] \tag{3.16}$$

Therefore;

$$Cov[\bar{Y}_h, W_h] = \mathbb{E}[\bar{Y}_h, W_h] - [\mathbb{E}\bar{Y}_h][\mathbb{E} W_h] = 0 \tag{3.17}$$

$$\text{Var}(\hat{Y}) = \sum_{h=1}^{H} P_h^2 \left[\alpha^2 \text{Var}(\bar{Y}_h) + (1-\alpha)^2 \text{Var}(W_h)\right] \tag{3.18}$$

where:

$\text{Var}(\bar{Y}_h) = \frac{\sigma_h^2}{n_h}$, assuming the sample mean is unbiased with sample size n_h,

$\text{Var}(W_h)$ depends on the variability of the weights β and X. If the weights β are treated as deterministic(non-random), $\text{Var}(W_h) = \frac{\sigma_h^2}{n_h}$ as well.

Thus:

$$\text{Var}(\hat{Y}) = \sum_{h=1}^{H} P_h^2 \left[\alpha^2 \frac{\sigma_h^2}{n_h} + (1-\alpha)^2 \frac{\sigma_h^2}{n_h} \right]. \tag{3.19}$$

But for each stratum $\alpha + (1-\alpha) = 1$

Similarly for each stratum

$\alpha^2 + (1-\alpha)^2 = 1$

$(1-\alpha)^2 = 1 - \alpha^2$

Thus,

$$\text{Var}(\hat{Y}) = \sum_{h=1}^{H} P_h^2 \left[\alpha^2 \frac{\sigma_h^2}{n_h} + \frac{\sigma_h^2}{n_h} - \alpha^2 \frac{\sigma_h^2}{n_h} \right]. \tag{3.20}$$

$$\text{Var}(\hat{Y}) = \sum_{h=1}^{H} P_h^2 \frac{\sigma_h^2}{n_h}. \tag{3.21}$$

Bias of $\hat{Y}$

The bias of $\hat{Y}$ is:

$$\text{Bias}(\hat{Y}) = \mathbb{E}[\hat{Y}] - \bar{Y}. \tag{3.22}$$

Recall:

$$\mathbb{E}[\hat{Y}] = \sum_{h=1}^{H} P_h \mu_h = \bar{Y} \tag{3.23}$$

hence

$$\text{Bias}(\hat{Y}) = \bar{Y} - \bar{Y} = 0. \tag{3.24}$$

Combining Variance and Bias we have;

$$\text{MSE}(\hat{Y}) = \text{Var}(\hat{Y}) + \text{Bias}(\hat{Y})^2. \tag{3.25}$$

$$\text{MSE}(\hat{Y}) = \sum_{h=1}^{H} P_h^2 \frac{\sigma_h^2}{n_h} + 0^2 = \sum_{h=1}^{H} P_h^2 \frac{\sigma_h^2}{n_h}. \tag{3.26}$$

It follows that since the proposed hybird estimator is unbiased the MSE is equal to the variance and it decreases as the sample size n_h increases.

Chapter 4

Simulation, Results and Discussion

4.1 Introduction

In this chapter, simulation using R statistical tool was done to evaluate the performance of the proposed hybrid estimator in terms of bias, variance, and Mean Squared Error (MSE) under various stratification and outlier scenarios. The simulation compared the proposed estimator with existing estimators such as the Neyman Sample Mean Estimator and Wang Xu hybrid estimator.

4.2 Simulation Study

A finite population consisting of 100,000 units was generated and divided into H strata.The data within each stratum was drawn from a normal distribution $N(\mu_h, \sigma_h^2)$,where μ_h and σ_h^2 vary across strata to simulate heterogeneity.

A fixed proportion of the population was replaced with extreme values (outliers) generated from a heavy-tailed t-distribution with 1 degree of freedom, shifted by the mean of the respective stratum. Outliers were introduced randomly across strata to assess their impact on the mean estimator. Data was simulated with 5 and 10 percent outliers and across 3,5 and 8 strata resulting in six different simulation cases based on the proportion of outliers and the number of strata.Each case was simulated 100 times to generate a sufficient number of sample statistics for each estimator.The study aimed to compute the Mean Squared Error(MSE),bias and Variance for each case,allowing a comparison of the following estimators under the influence of outliers and evaluating how stratum sizes affect their estimates.

Neyman Sample Mean Estimator: $\hat{Y} = \sum_{h=1}^{H} P_h \bar{y_h}$

Wang Xu hybrid estimator: $Y_{trimmed} = \sum_{h=1}^{H} P_h [w\bar{y_h} + (1-w)T_h$

Proposed Hybrid Estimator: $Y_{weighted} = \sum_{h=1}^{H} P_h [\alpha\bar{y_h} + (1-\alpha)W_h$

Sampling cases:

Case 1: Outliers proportion = 5 percent Number of strata = 3 Number of simulations = 100

Case 2: Outliers proportion = 5 percent Number of strata = 5 Number of simulations = 100

Case 3: Outliers proportion = 5 percent Number of strata = 8 Number of simulations = 100

Case 4: Outliers proportion = 10 percent Number of strata = 3 Number of simulations = 100

Case 5: Outliers proportion = 10 percent Number of strata = 5 Number of simulations = 100

Case 6: Outliers proportion = 10 percent Number of strata = 8 Number of simulations = 100

4.3 Simulation Results

Below is a tabular representation of estimated parameters from simulation.

Table 4.1: Estimated Population Means

	Case-1	**Case-2**	**Case-3**	**Case-4**	**Case-5**	**Case-6**
Population Means	1.53488	5.391395	7.65506	1.53488	5.391395	7.65506
Neyman estimates	1.824697	5.419464	7.769884	1.901161	5.483444	7.642358
Wang Xu Hybrid estimates	1.517549	5.343821	7.657510	1.537458	5.393288	7.636107
Proposed Hybrid estimates	1.517691	5.342931	7.660371	1.538058	5.391720	7.634343

After conducting the simulations, Mean Squared Error (MSE), Variance, and Bias were computed for each estimator across all listed cases. The results are presented below, providing a comparative analysis of the estimators' performance under different conditions.

Table 4.2: Neyman Estimator

	Case-1	**Case-2**	**Case-3**	**Case-4**	**Case-5**	**Case-6**
Variance	0.706642	0.041951	0.186614	0.518832	7.223978	4.787308
Bias	0.011761	0.005231	0.011569	0.030684	-0.113399	-0.244653
MS Error	0.706780	0.041978	0.186748	0.519774	7.236837	4.847163

Under Neyman estimator there was high variance and Mean Squared Error,especially in Case 5 and 6, suggesting that this estimator is highly sensitive to outliers. Extreme values inflate the variability of the estimator.Thus,the Neyman Estimator is not suitable for datasets with outliers.

Table 4.3: Wang Xu Hybrid Estimator

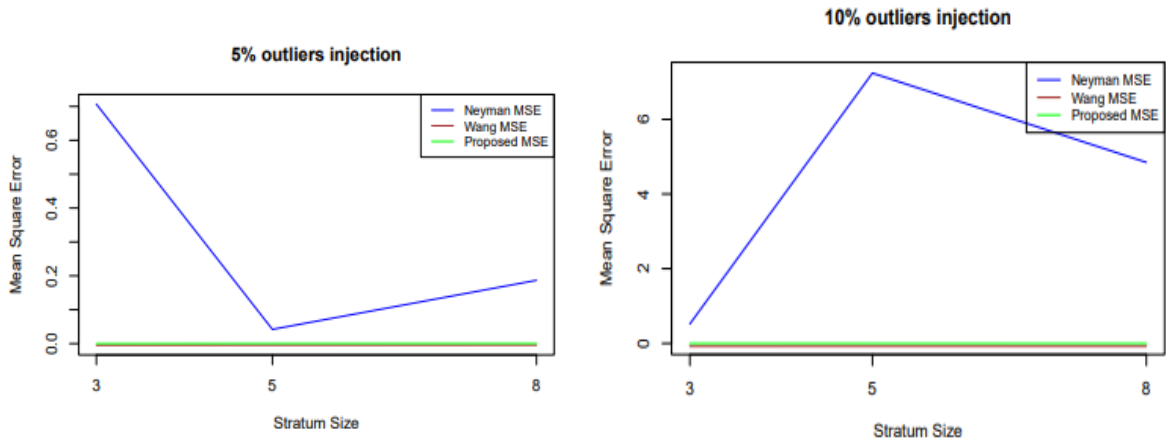
	Case-1	**Case-2**	**Case-3**	**Case-4**	**Case-5**	**Case-6**
Variance	0.000110	0.000320	0.000423	0.000086	0.000287	0.000464
Bias	0.000303	0.001567	0.000372	-0.001094	0.001685	0.000965
MS Error	0.000110	0.000323	0.000424	0.000087	0.000290	0.000465

Wang Xu hybrid estimator had low Mean Squared Error,Bias and Variance indicating that even with outliers, the estimator remains more efficient.

Table 4.4: Proposed Hybrid Estimator

	Case-1	**Case-2**	**Case-3**	**Case-4**	**Case-5**	**Case-6**
Variance	0.000109	0.000311	0.000436	0.000085	0.000297	0.000511
Bias	0.000251	0.001532	-0.000004	-0.001339	0.001592	0.000712
MS Error	0.000109	0.000313	0.000436	0.000087	0.000299	0.000511

The proposed estimators shows slightly high variance and Mean Squared Error compared to the Wang Xu Hybrid Estimator,though, it rivals Wang Xu Hybrid Estimator some cases such as case 2.



(a) (b)

Figure 4.1: Panel (a) shows the MSE at 5 percent outlier injection. Panel (b) shows the MSE at 10 percent outlier injection.

From Figure 4.1,the Wang-Xu hybrid estimator have the smallest MSE compared to Neyman and the Proposed Estimator ,thus, it is clear that the Wang-Xu estimator is significantly more stable and resistant to outliers compared to the rival estimators.

4.4 Conclusion, Recommendation and Suggestion

4.4.1 Conclusion and Recommendation

Both the Wang Xu Hybrid Estimator and the Proposed Hybrid Estimator are more efficient estimators compared to the Neyman Estimator.However since Wang-Xu Hybrid Estimator trims the data to remove outliers,it is suitable when the outliers are as a result of errors and for large sample sizes.

When the outliers belong to the distribution and when dealing with small sample sizes, the proposed hybrid estimator is recommended since it utilizes all the data points with no loss of information.

4.4.2 Suggestions for further research

We suggest more study to be done on Estimation of population mean in the presence of outliers in two phase sampling and in other complex survey sampling schemes

References

Aloma, r. N. and Al-Sharif, A. (2020). Effects of outliers in stratified random sampling: An outlier detection framework. *Journal of Statistical Theory and Practice*, 14(1):23–45.

Barnett, V. and Lewis, T. (1994). *Outliers in statistical data (3rd ed.)*.

Chen, J. and Chen, L. (2004). Outlier detection in stratified samples: A comparison of methods. *Journal of Statistics in Medicine*, 23(14):2339–2356.

Chen, J. and Chen, L. (2006). Robust estimation in stratified sampling with outliers. *Journal of Statistical Planning and Inference*, 136(6):2071–2082.

Cochran, W. G. (1977). *Sampling techniques (3rd ed.)*.

Ghosh, S.and Chatterjee, P. (2018). Machine learning-based anomaly detection in stratified sampling. *Journal of Computational Statistics*, 42(1):56–78.

Hodge, R. J. and Bickel, P. J. (2005). Combining stratification and robust statistical techniques. *Journal of Statistical Science*, 20(3):360–372.

Huber, P. J. (1981). Robust statistics. Technical report, Wiley Sons.

Johnson, T., . B. M. (2021). Bayesian hierarchical modeling for outliers in stratified sampling. *Bayesian Analysis Journal*, 12(2):89–112.

Koller, D. and Friedman, N. (2009). *Probabilistic graphical models: Principles and techniques*. MIT press.

Neyman, J. (1934). *On the Two Different Approaches to the Application of Probability Theory to Agricultural Experiments*.

Osborne, J. W., C. R. A. and Gunter, L. (2001). Researchers rarely check for outliers: A review of the literature: A review of the literature. *Journal of Educational Psychology*, 93(4):769–780.

Rousseeuw, P. J. and Leroy, A. M. (1987). Robust regression and outlier detection. Technical report, Wiley.

Singh, S. and Sedory, S. (2015). Weighted means and winsorized estimators in stratified sampling. *Statistical Journal*, 50(1):45–60.

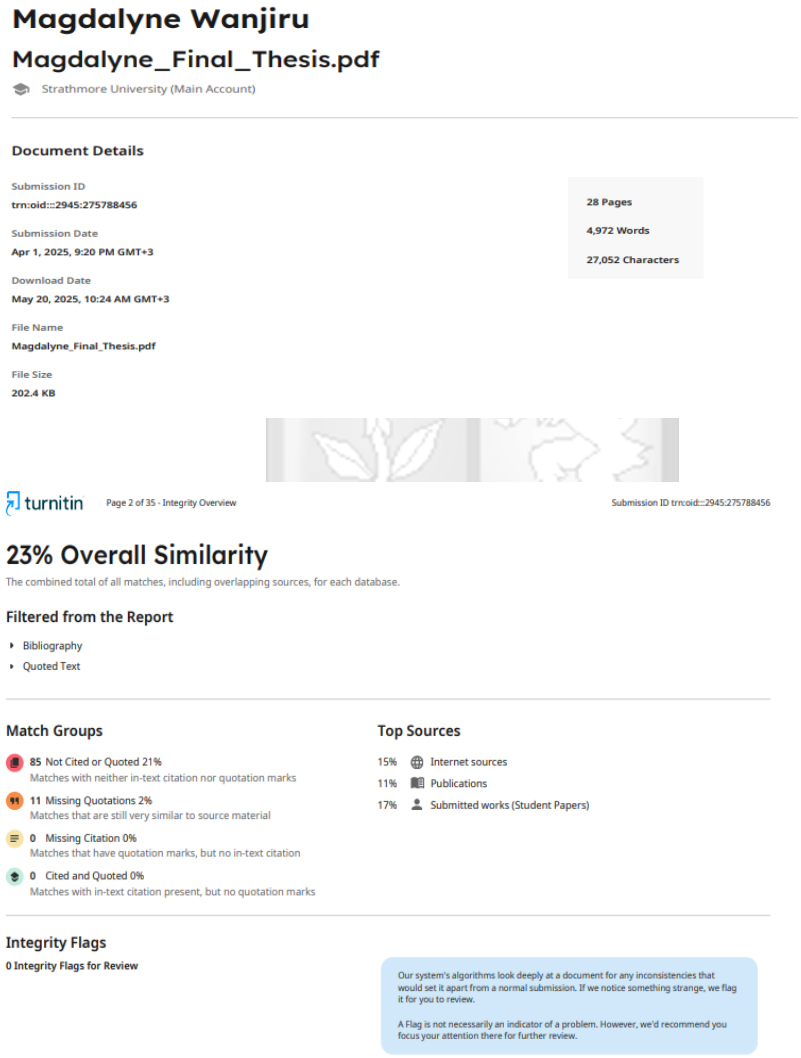
Tukey, J. W. (1977). Exploratory data analysis. Technical report, Addison-Wesley.

Wang, H. and Xu, G. (2023). A hybrid estimator using trimmed means in stratified sampling. *Journal of Statistical Research*, 45(3):215–232.

Wang, Q. and Wu, L. (2010). Modified estimators for outliers in stratified sampling. computational statistics data analysis. *Journal of Statistical Science*, 54(2):388–399.

Appendices

Appendix A: Similarity Report



Magdalyne Wanjiru

Magdalyne_Final_Thesis.pdf

Strathmore University (Main Account)

Document Details

Submission ID
trn:oid:::2945:275788456

Submission Date
Apr 1, 2025, 9:20 PM GMT+3

Download Date
May 20, 2025, 10:24 AM GMT+3

File Name
Magdalyne_Final_Thesis.pdf

File Size
202.4 KB

28 Pages

4,972 Words

27,052 Characters

turnitin Page 2 of 35 - Integrity Overview Submission ID trn:oid:::2945:275788456

23% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Filtered from the Report

- Bibliography
- Quoted Text

Match Groups

85 Not Cited or Quoted 21%
Matches with neither in-text citation nor quotation marks

11 Missing Quotations 2%
Matches that are still very similar to source material

0 Missing Citation 0%
Matches that have quotation marks, but no in-text citation

0 Cited and Quoted 0%
Matches with in-text citation present, but no quotation marks

Top Sources

15% Internet sources
11% Publications
17% Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

Appendix B: Ethical Clearance Confirmation

18th March 2025

Ms Wanjiru Magdalyne,
magdalyne.wanjiru@strathmore.edu

Dear Ms Wanjiru,

RE: Estimating a Finite Population Mean in Presence of Outliers Under Stratified Random Sampling

This is to inform you that SU-ISERC has reviewed and **approved** your above **SU-masters** proposal. Your application reference number is **SU-ISERC2663/25.** The approval period is from **18th March 2025 to 17th March 2026.**

This approval is subject to compliance with the following requirements:

i. Only approved documents including (informed consents, study instruments, MTA) will be used.
ii. All changes including (amendments, deviations, and violations) are submitted for review and approval by SU-ISERC.
iii. Death and life-threatening problems and serious adverse events or unexpected adverse events whether related or unrelated to the study must be reported to SU-ISERC within 72 hours of notification.
iv. Any changes anticipated or otherwise that may increase the risks or affected safety or welfare of study participants and others or affect the integrity of the research must be reported to SU-ISERC within 72 hours.
v. Clearance for the export of biological specimens must be obtained from relevant institutions.
vi. Submission of a request for renewal of approval at least 60 days prior to the expiry of the approval period. Attach a comprehensive progress report to support the renewal.
vii. Submission of an executive summary report within 90 days of completion of the study to SU-ISERC.

Before commencing your study, you will be expected to obtain a research license from National Commission for Science, Technology, and Innovation (NACOSTI) https://research-portal.nacosti.go.ke/ and obtain other clearances needed.

Yours sincerely,

Mr Ambrose Rachier,
Chairperson; SU-ISERC

Ole Sangale Rd, Madaraka Estate. PO Box 59857-00200, Nairobi, Kenya. Tel +254 (0)703 034000
Email admissions@strathmore.edu www.strathmore.edu