MINIMISING TOTAL SURVEY ERROR ARISING FROM SAMPLING AND NON-SAMPLING ERRORS: A DESIGN BASED APPROACH

MUNYARADZI DAMSON

MASTER OF SCIENCE

(Mathematics - Statistics Option)

PAN AFRICAN UNIVERSITY INSTITUTE FOR BASIC SCIENCES TECHNOLOGY AND INNOVATION

2014

MINIMISING TOTAL SURVEY ERROR ARISING FROM SAMPLING AND NON-SAMPLING ERRORS: A DESIGN BASED APPROACH

MUNYARADZI DAMSON

MS300-0004/12

A Thesis submitted to Pan African University Institute for Basic Sciences, Technology and Innovation in partial fulfilment of the requirements for the degree of Master of Science in Mathematics (Statistics Option)

2014

DECLARATION

This thesis is my original work and has not been presented for a degree/diploma/certificate in any other University.

....

Signature Date

DAMSON MUNYARADZI

MS300-0004/12

This thesis has been submitted for examination with my approval as University Supervisor.

....

Signature Date

Professor R. Odhiambo

Jomo Kenyatta University of Agriculture and Technology

This thesis has been submitted for examination with my approval as University Supervisor.

....

Signature Date

Dr G.Orwa

Jomo Kenyatta University of Agriculture and Technology

DEDICATION

This thesis is dedicated to my beloved grand parents, parents, wife, children, sisters, brothers and friends who worked tirelessly for my success.

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to my supervisors, Professor R. O. Odhiambo and Dr G. O. Orwa with their constructive suggestions, guidance and criticisms throughout the thesis. Their valuable knowledge and contribution enabled me to go through the thesis with few hassles. Special thanks goes to the African Union, Jomo Kenyatta University of Agriculture and Technology, Pan African University Institute of Basic Sciences Technology and Innovation and all the lecturers for the academic support they gave me. Worth appreciating too are all students for their team spirit and cooperation that created a good learning atmosphere for the entire study period. My sincere gratitude also goes to the Almighty God for his guidance and protection during this study. If it was not the will of God I could not be who I am today.

All views, errors and omissions are my own and should not be directed to any organization or person mentioned above.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACRONYMS OR ABBREVIATIONS

TSE	:Total Survey Error
TSQ	:Total Survey Quality
MSE	:Mean Square Error
EA	:Enumeration Area
PSU	:Primary Sampling Units
SSU	:Secondary Sampling Units
SRS	:Simple Random Sampling
SRSWOR	:Simple Random Sampling Without Replacement

ABSTRACT

This study was based on the total survey error paradigm in which the purpose was to examine a variety of sources of errors in a survey. The study sought to quantify the total survey error for estimating population total in two stage cluster sampling and empirically compute the mean square error estimate using data from an actual survey. Furthermore, the study analysed a method of handling non-response in complex survey design. The methodology of the study was centred on the decomposition of the mean square error into sampling error, refusal error, non-interview error and response error. These were further simplified to come with a total survey error expression. An empirical study with data based on waiting time to conception in women in Zimbabwe was conducted to estimate the population total survey error using this model. Data collection was done through face-to-face interviews in the selected clusters. The results shows that both unit and item non-response contributed significantly (90.8%) to the total mean square error. Imputation was done in order to reduce the error and the results show that imputation should be done within each cluster since the non-response depends on a cluster level random effect.

CHAPTER ONE

INTRODUCTION

1.1 Background of the study

Measures of data quality are very essential for the evaluation and improvement of survey design and procedures. A comprehensive study of the sources, magnitude and impact of survey errors is necessary in the identification and use of appropriate survey design and sampling procedures. Much of the available research in survey methodology during the last 25 years has emphasized methods of reducing sampling errors rather than minimising total survey error (Montgomery, 2009). The total survey error (TSE) model offers a theoretical framework for optimizing surveys by maximizing the quality of data. Survey samplers have emphasised the need for a total survey error design approach by which available resources are distributed to those error sources where error reduction is most effective, hence leading to superior survey designs. This study examines a variety of error sources for estimates obtained from a survey and the perspective taken follows an error model based on a finite population model, in which the overall objective of the survey is to minimise total survey error in complex surveys.

The total survey error (TSE) paradigm encompasses the idea of optimal allocation of resources to minimize the total survey error (TSE) for key statistics. In order to fully implement the total survey error (TSE) paradigm, all the major error sources should be identified, evaluated, controlled and reduced to levels where their existence does not negatively affect the usefulness of the survey data. According to Biemer and Lyberg (2003) total survey error is defined as the accumulation of all errors that arise in the design, collection, processing, and analysis of survey data. Therefore in this context, a survey error is defined as the deviation of a survey response from its underlying true value.

Issues to do with data quality in sample surveys have received great attention in recent years, with an increasing number of inventiveness and journal publications focusing on this area. This also includes several international workshops, seminars and conferences (Dillman et al., 2009).

The concept of data quality appreciates the fact that users of data and producers of data frequently view survey quality from diverse perspectives. Data users take data accuracy for granted and place a higher premium on attributes such as the accessibility, timeliness and how usable the data is. On the other hand producers of data place a high premium on high response rate, large sample size, good coverage of the target population and consistent responses. These diverse perspectives to some extent imply that survey quality is a complex, multidimensional concept that goes even beyond total survey error (Biemer, 2011). The determination on the type of data required by users is an indispensable first step to good survey planning.

Although a sizeable number of studies on non-response bias have been done, relatively very little is known about other sources of non-sampling errors. In most studies, interviewer variance is rarely estimated in centralized telephone surveys, even though the cost of doing so routinely is relatively small. The credibility and authenticity of a survey depends on quality of the survey data (Thompson and Randy, 2007). Studies of frame bias or data-processing errors are seldom reported. Platek and Sarndal (2001) note a lack of progress over the last 50 years in integrating sampling and non-sampling errors as measures of uncertainty. The quality of survey data and results is determined by three elements which are relevance, timeliness and accuracy (Montgomery, 2009). Deming (1944) describes 13 factors that affect the usefulness of surveys and these factors include sampling errors as well as non-sampling errors. These survey errors may arise from survey frame deficiencies, sampling process, interviewing and interviewers, missing data, respondents, data coding, keying and editing processes. Survey errors are problematic because they distort and diminish the accuracy of inferences resulting from the survey data. A statistic will be accurate if it has a small bias and variance, which occurs only if the influence of Total Survey Error (TSE) on the estimate is small (Duane, 2007). Biemer (2011) went further and stated that estimators used for the population parameter are subject to both variable errors and biases. These variable errors arise from sources that can change over replications of the survey and biases are errors that remain the same over replications of the survey.

According to Kalton and Heeringa (2003), the accuracy of survey estimates greatly depends on the sampling and non-sampling errors of these estimates that result from the survey. These

sampling errors in survey estimates occur due to the fact that data are collected only for a subset of the target population, thus, the estimates from the sample may differ from the actual values obtained from a complete census. On the other hand non-sampling errors are defined as a residual category, that is, as all errors of estimation that are not the result of sampling. Unlike sampling errors these non-sampling errors have many different sources and can occur at different stage of a survey, and are often exceedingly hard to detect and control (Couper, 2008). Kalton and Heeringa (2003) consider total survey error as a conceptual framework that is used to scientifically consider types of survey error during the design process and in describing its quality when completed. Frederick (2005) went further and stated that total error in surveys can be conceptualized and categorized in many ways. One conventional approach is dividing total error into sources of sampling errors and sources of non-sampling errors. Another classification is by dividing it between coverage errors, sampling errors, non-response errors, and measurement errors. Lohr (2009) came up with a more modern approach in which he grouped various sources of error into the classes of representation and measurement.

1.2 Statement of the Problem

The estimation of the population total or mean is a persistent issue in both sampling theory and practice and numerous efforts have been designed in order to improve the precision of these estimates. Survey samplers have paid a lot of attention to sampling errors and survey estimates have been presented with only one source of error measured. This error is due to sampling resulting from the fact that survey estimates would have different values had another sample been drawn using the same sampling design. This has led to the development of numerous methods for optimal allocation of resources to minimize the sampling variance associated with an estimated total or mean. However they have given very little attention to total survey error arising from both sampling and non-sampling errors. Other variable errors like the response error are ignored, and biases are rarely mentioned. The presence of a total survey error model offers a rare opportunity to measure and quantify a large set of variable errors and biases that are normally assumed to be negligible in survey data analysis.

1.3 Objectives of the study

The overall objective of this study was to investigate the effect of imputation on total survey error arising from sampling and non-sampling errors.

The specific objectives of the study are:

(i) To quantify the total survey error for estimating population total in two stage cluster sampling.

(ii) To propose an estimator of non-response error and investigate it for unbiasedness.

(iii) To compute the mean square error estimate using data from an actual survey.

1.4 Justification

In an ideal survey situation, where respondents answer all questions honestly without error, researchers only need to worry about sampling errors. However this complex situation is not ideal and non-sampling errors occur. These contribute significantly to both bias and increased variability thereby affecting the quality of data produced. This justifies the need for a total survey error design in complex surveys so that errors are quantified and error sources identified. This total survey error design facilitates the optimal allocation of resources to those error sources where improvement is achievable. Bias represents a form of systematic error (e.g. inaccurate responses) that does not disappear as the sample size increases. Survey samplers regularly stress that these non-sampling errors contribute a lot more to the total mean squared error (MSE) which is regularly used as a quantitative measure of quality of estimators, and that a total survey error design approach is desirable.

1.5 Scope of the study

During past decade, incredible progress has been made towards improving the accuracy of survey data and results. The total survey error (TSE) paradigm is part of the much broader

concept of total survey quality (TSQ), which considers the fitness for use of a statistic. However the scope of this study was only limited to total survey error, in which we mainly focused on sampling error, refusal error, non-interview error and response error. These errors were measured and quantified by the mean square error in two stage cluster sampling. To date it has become very conventional to separate the total survey error into categories representing sampling and non-sampling errors.

1.6 Limitations

Survey researchers would always want to reduce errors in their surveys, however the extent to which these errors can be reduced is restricted by several practical constraints. The conventional exposition of the total survey error (TSE) approach is a trade-off between survey accuracy and important constraints such as costs and time. Some survey costs are fixed while others are variable. In this study the main limitations where high cost of traveling and inaccessibility of some respondents. The cost of conducting the household listing was very high to the extent that we could not cover the entire target population thus we ended up relying on the 2012 Zimbabwe census data. The updating operation consisted of listing all of the households residing in the selected enumeration areas (EAs) and recording for each household the required basic information. This listing procedure if successful would have provided a complete list of the households residing in the selected EAs, which would have served as the sampling frame for the second-stage sampling for household selection. Secondly, the data collection was conducted during the rainy season and due to flooding we could not reach some respondents thus we were forced to re-sample other respondents in the accessible areas to attain the desired sample size.

CHAPTER TWO

LITERATURE REVIEW

2.1 Introduction

This chapter reviews literature on total survey errors arising from both sampling and non sampling errors and seeks to establish whether solutions could be found that address the problem of limited research focused on total survey errors. The major concepts under discussion in this chapter are total survey error, mean square error, sampling error, non-sampling error and two-stage cluster sampling. In addition the chapter reviews literature on imputation and waiting time to conception in women. Throughout the chapter the researcher used the APA System of referencing.

2.2 Total Survey Error (TSE)

According to Duane (2007), the total survey error (TSE) paradigm is extensively acknowledged as a conceptual framework used for evaluating survey data quality and is measured by the mean square error (MSE). It is also used to systematically consider the various types of survey error during the design process and in describing its quality when completed. Biemer (2011) defines in quality total survey error as the estimation and reduction of the mean square error (MSE) of estimates of interest, which is the sum of random errors known as variance and squared systematic errors known as bias. Couper (2008) states that total survey error encompasses measurement construct validity, measurement error and processing error. This entails a clear understanding of how well survey questions measure the constructs of interest and representation i.e. coverage error, sampling error, non-response error and adjustment error.

Montgomery (2009) postulates that in the total survey error paradigm, there may be cost-error trade-offs resulting in tension between reducing such errors and the cost incurred in reducing them. To date, a lot is known about the statistical efficiencies of various sampling schemes,

however lack of adequate information on total survey costs is a major factor restraining extensive use of total survey error concept. Some significant progress has been achieved in the development of principles and methods of reducing non-sampling errors, but almost all the methodological studies have been aimed at minimizing sampling errors per dollar as compared to minimizing total survey errors (Arlene, 2003).

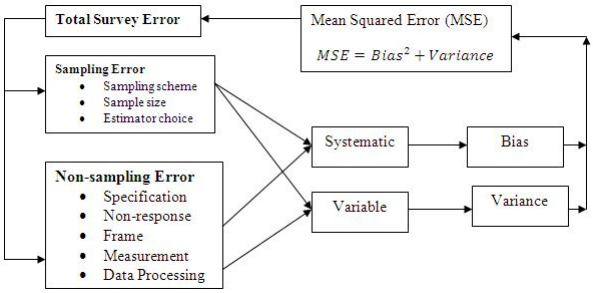

Figure 2.2.1: Total survey error conceptual framework.
(source Biemer (2011))

Linacre and Trewin [1993] used the total survey error approach in conducting a survey in the construction industry. They considered approximately 180 resource allocation options and acquired measures of the total cost and total survey error depending on the root mean squared error (RMSE) technique for each of the options. One of their results shows a good balance between cost and error.

2.3 Total Survey Quality

Sinclair and Gastwirth (1996) proposed seven different dimensions that are frequently used to evaluate the quality of official statistics in terms of both survey error and fitness for data use. These were comparability, accuracy, relevance, accessibility, timeliness and punctuality, coherence and interpretability. In this framework, total survey error was regarded as being covered by the accuracy dimension and the overall aim was to optimise costs and minimize burden on design constraints. In their paper, Platek and Sarndal (2001) proposed a sequential sample design for quality control that incorporated decisions in the shortest possible time and with the smallest possible sample size. Their results show that, for the cases in which the sequential procedure could not be adopted directly by the controller, a two phase selection

process with some permanent random numbers which permits a very efficient assessment of the quality of the data collection was essential. According to Saris and Irmtraud (2007), in any survey a significant part of resources should be committed to quality control of the data collected. A good number of projects include quality control particularly in activities such as data entry, elaboration of auxiliary variables like remote sensing data, characteristics and functionalities of the information system.

2.4 Mean Square Error (MSE)

According to Biemer (2011) a number of acceptable metrics for quantifying TSE have been proposed and accepted in statistical literature, the most important metric in survey statistics is the mean squared error (MSE). Each estimate that will be computed from the survey data has a corresponding MSE that summarizes the effects of all sources of error on the estimate (Biemer, 2011). In addition Wolter (2007) postulates that a small mean squared error (MSE) indicates that the Total Survey Error (TSE) is small and under control whilst a large MSE indicates that one or more sources of error are adversely affecting the accuracy of the estimate.

Lohr (2009) states that one of the primary uses of the MSE is as a measure of the accuracy of survey data and MSE is the expected squared difference between an estimate $\hat{\theta}$ and the parameter it is intended to estimate θ. According to Montgomery (2009), each source of error in a survey may significantly contribute a variable error, systematic error or even both. The variable errors are reflected in the variance of the statistic, whilst the systematic errors are reflected in the bias squared component. Zieba and Kordos (2010) went further and decomposed the variance and bias components into process-level and sub-process level components thereby identifying specific error sources and their root causes.

Biemer (2011) considered a simple model for decomposing the total Mean Square Error (MSE) of any particular characteristic in the survey say y . The survey errors arising from all the various sources of errors in the survey have a cumulative effect on the observed value of y. These errors may significantly cause the observed value of y to be higher or lower than its true value for an

individual.

2.5 Two stage cluster sampling

Carle (2009) defines a cluster as an aggregation of sampling units of interest for a particular survey that can be unambiguously defined and can be used as a sampling unit from which to select a smaller sub-sample. Heeringa et al. (2010) also state that in cluster sampling, a group of sampling elements constitutes the sampling unit, instead of a single element of the population. In two stage cluster sampling a specified number of elements is selected from each of the selected cluster. The clusters that form the units of the sampling at the first stage are called first stage units (FSU) or primary sampling units (PSU). Those elements or groups of elements within the clusters that form units of sampling at the second stage are called the secondary stage sampling units (SSU).

In their paper Sarndal and Lundstrom (2005) states that cluster sampling is often stimulated by cost efficiency, that is low cost of data collection per sample element. The use of cluster sampling has attracted great attention because travelling costs of interviewers can be significantly reduced thus the cost efficiency of cluster sampling can therefore be high. However Lumley (2010) argues that there are also certain shortcomings of cluster sampling that affect statistical efficiency. If each cluster closely reflects the population structure, we would definitely attain efficient sampling such that standard errors of estimates would not exceed those of simple random sampling. However, in practice, clusters tend to be internally homogeneous, and this intra-cluster homogeneity increases standard errors and thus decreases statistical efficiency (Demnati and Rao, 2004).

2.6 Sources of error

The ultimate goal of optimal survey design as stated by Biemer (2011) is to minimise total survey error subject to availability of funds and timeliness constraints that are always consistent

with other user-centric quality dimensions. A cautious approach is advisable when allocating resources to the different stages of the survey process so that the major sources of errors in the survey are controlled to tolerable limits. According to Groves and Tourangeau (2004) survey results are subject to different sources of error. The total survey error in an estimate is the actual difference between the estimate derived from the sample data collected and the true value for the population. The total survey error is made up of two main types and these are systematic and random error. Lohr (2009) also states survey estimates are subject to two broad types of errors and these are bias and variable errors. Bias refers to errors that affect the expected value of the survey estimate, taking it away from the true value of the target parameter. Variable errors affect the spread of the distribution of the survey estimates over potential repetitions of the survey process. Sampling errors vanish if observations cover the complete population (Brick et al., 2006).

2.6.1 Sampling Errors

According to Frederick (2005), sampling errors are statistical errors which survey researchers expose statistics simply because of working with sample data rather than whole population. In their paper (Biemer and Lyberg, 2003) states factors such as the sample design, sample size, the estimation method and the variability of the population characteristic under study affect the sampling error. Lohr and Rao (2006) also argued that if the population is very heterogeneous, a large sample size is required in order to obtain reliable statistics. This sampling error is measured by a statistical quantity known as the standard error. This quantity reflects the expected variability of the survey statistic of a particular population characteristic if repeated sampling is conducted Lohr (2009). In his paper Popinski (2006) states that if a statistic is determined from a specific sample then its value is most likely to be different within set of values if several additional sample estimates were measured in an identical manner. Thus the sampling error will be a function of the difference in the distribution of the values in the sample and the distribution of the values in the population.

2.6.2 Non-sampling Errors

Non-sampling errors arise as a result of failure to get accurate information about each and every sampling unit in the sample. The reasons are so diverse and may include imperfect memory on the part of the respondent, oversight by interviewer on some of the specifications on the data to be included in the interview guide, and deliberate misstatements by respondents Lohr and Rao (2006). Brick et al. (2006) states that non-sampling errors may occur even when the whole population is observed. According to Thompson and Randy (2007) it is general knowledge that non-sampling errors, or response errors, can be large enough to cause serious problems and the magnitude of non-sampling errors might increase with an increase in sample size. This could be due to administrative complexity of maintaining high standards in the interview process and keeping sufficient control over all phases of the work. Heeringa et al. (2010) divided non-sampling errors into two broad categories; non-sampling variance and non-sampling bias. Non-sampling variance is known to measure the variation in survey statistics that is if the same sample would be submitted to hypothetical repetitions of the survey process under the same essential conditions. On the other hand non-sampling bias refers to the survey errors that accrue from the survey process and survey conditions, and would lead to survey statistics with an expected value different from the true population parameter value.

2.6.3 Non-response errors

Non-response errors are those errors caused by failure to obtain data for sampling units selected for the survey Groves and Magilavy (1984). Non-response certainly affects every survey in one way or another despite it being a sample or a census. In their paper Groves and Tourangeau (2004) found that non-response negatively affected data from administrative sources that are used for statistical production. According to Sarndal and Lundstrom (2005), most survey samplers develop and employ operational procedures that guard against or at least reduce the incidence of these non-responses. Non-response is problematic when responses to the survey are not at random and response rates are low. If this non-response is purely at random, its main

effect is an increase in variance of the survey statistics due to sample size reduction. However, Lohr (2009) argues that in some cases if survey response is dependent on certain features and characteristics of respondents or interviewers, then bias is the main problem that we needs to worry about.

The major techniques for dealing with non-response in surveys are imputation and weighting adjustment. Imputation entails the substitution of good artificial values for the missing values while weighting adjustment entails increasing the weights functional in the estimation of the variable of interest of the respondents so as to compensate for the values that are lost because of non-response (Dillman et al., 2009). Among the known three categories of non-response, unit non-response has proved to be the most difficult to compensate. This is because in most cases there is insufficient information within survey frames and records that can be used for that purpose. The most common and reliable compensation method used to suppress the negative effects of unit non-response is the weighting adjustment, whereby those responding sampling units have their weights inflated so as to account for the loss of sample units due to non-response (Heeringa et al., 2010).

2.6.4 Interviewer effect

According to Mahalanobis (1946) estimating interviewer variance can be quite challenging from an operational perspective, particularly for face-to-face interviews. This is due to the fact that the estimation process requires that sampling elements be randomly assigned to interviewers, and this process is known as the interpenetration. Failure to interpenetrate interviewer assignments will result in biased estimators of interviewer variance. In face-to-face interviews, geographically proximate interviewer assignment areas may be combined so that the sampling elements in the combined area can be assigned at random to each interviewer working in that area. The interpenetration process is much simpler in centralized telephone surveys if the telephone numbers to be called during a particular shift are randomly assigned to all the interviewers working the shift (Duane, 2007).

2.7 Auxiliary information

According to Bouza (2007) the utilisation of additional information on an auxiliary variable for estimating the mean of the variable of interest played an important role in survey sampling theory and practice. If the auxiliary information associated with the variable of interest is available on each and every unit of the population, then it would be beneficial to employ such supplementary information in survey sampling. Dryver and Thompson (2005) estimated a ratio estimate of the population mean $\bar{Y}$ in stratified random sampling using two different methods. In their first method they separated ratio estimate of the total of each stratum and in their second method they derived an estimate from a single combined ratio.

In his study Rao (1987) studied several variables and assumed that information in some cases was not obtained at the first attempt even after some call-backs. The estimates they obtained from such incomplete data were found to be misleading, particularly in cases where respondents differ from the non-respondents. Hansen and Hurwitz (1946) proposed a method for fine-tuning non-response to rectify the bias problem. Their idea was to take a sub sample from the non-respondents to obtain an estimate for the sub-population characterised by the non-respondents. By means of the Hansen and Hurwitz (1946) method, Cochran (1977) proposed the ratio estimators and regression estimators of the population mean of the variable of interest, in which the information on the auxiliary variable was obtained from all the units in the sample, and the population mean of the auxiliary variable was known, whereas some units failed to provide information on the variable of interest.

2.8 Variance Estimation

Sampling variance is a very important measure of the quality of estimates of finite population parameters. It measures the amount of sampling error in the estimate due to observing a sample instead of the whole population (Rabe-Hesketh and Skrondal, 2006). Lohr and Rao (2006) states that problems of variance estimation for estimators derived from sample surveys are two dimensional. Firstly they depend on the form of the estimator and secondly the complexity of

the sampling designs. Verma and Betti (2010) defines a linear estimator as one which can be expressed as a linear function of random variables. According to Zieba and Kordos (2010), sampling designs can be complex. Wolter (2007) defines a complex design as one which induces correlations between element values. An example of a complex design is a multi-stage sampling procedure where one first selects a sample of clusters and then selects a sample of elements within each of the chosen clusters. Verma and Betti (2010) argue that since elements within clusters tend to be homogeneous, clustering induces positive correlations between element values, which result in an increase in the variance.

According to Fay (2003) it may not be possible to obtain a closed-form algebraic expression for the estimated variance, thus the research literature on variance estimation for complex sample survey data contains several approximate methods from which sample survey data professionals can choose. Rabe-Hesketh and Skrondal (2006) state that there are two basic approaches to variance estimation which are the analytical approach (Taylor series linearisation method) and the re-sampling - replication method (jackknifing, balanced repeated replication, bootstrapping).

2.9 Optimal Allocation of resources

The process of designing a survey begins with the identification of feasible procedures taking into account practical constraints. Furthermore major alternatives that exist in survey design rarely present themselves as feasible alternatives within a fixed budget. Various approaches are associated with huge costs and require facilities which differ significantly in order of the magnitude. Lumley (2010) also acknowledges that the start-up cost of setting-up a new procedure or operation is quite different from the recurrent cost as an ongoing process. Moreover, a single method may affect a number of components of the error, in multifaceted ways that are often impossible to see or predict in advance. According to Platek and Sarndal (2001), many of household surveys are multi-purpose, meaning that they are intended to estimate several separate statistics on a variety of research topics. The optimal allocation of resources differs substantially for various estimates and various categories of estimates, so that at least some compromise is essential in the overall survey design.

Some compromise which is needed between conflicting requirements may have vital repercussions on the survey design. In support of this Kalton and Heeringa (2003) state that there are raging conflicts between the need to accumulate data over a specified period in order to get the best estimates of levels and to accurately measure differences and time trends. However, Heeringa et al. (2010) points out that despite all these constraints, the underlying principle of total survey error (TSE) design, optimal allocation of cost, balance and accuracy between different error sources, are of great value in coming up with complex decisions that face survey designers.

Time and again, the difficulty of improving the survey design presents itself in the form of the need to focus additional resources in the areas where the best improvement is attainable Mahalanobis (1946). In contrast, Couper (2008) postulates that being in a position to achieve the best improvement presupposes that we have prior knowledge of where the most good lies. Carle (2009) concludes by stating that it is essential to identify and categorise the components of error which are the most important in the sense that they are open to the significant reductions. It is in this regard that the development and implementation of the total survey error paradigm offers new, rare and cost-effective opportunities for improving the overall survey design.

2.10 Imputation

Naturally in large sample surveys, not all the sampled units take time to respond fully to the survey. Practically some sampled units do not respond at all, while others choose to respond only to certain items. Rubin (1996) developed a method known as the multiple imputation technique that can be used to handle such non-responses. Item non-response occurs more often in sample surveys with scores of items. It is more often than not, handled by several types of imputation methods to fill in the missing data. The principle of imputation for non-response data necessitates that multiple imputations be made dependent on the sampling design (Lumley, 2010). However, Raghunathan et al. (2003) argues that most known standard softwares that perform model-based multiple imputation assume in almost all cases simple random sampling, thereby leading many survey designers not to take into consideration variation for complex

sample design features, such as clustering and stratification in their multiple imputations.

According to Raghunathan et al. (2003), non-response on a particular variable in two stage cluster sampling greatly depends on a cluster level random effect and for this reason, it cannot be ignored. Estimators of the population total or population mean obtained by the mean imputation method or the re-weighting method under the ignorable non-response assumption becomes biased. Basing on this, Kaarik (2006) developed an unbiased estimator for the population mean by imputing or re-weighting within each cluster in the sample or a collection of sampled clusters having some common characteristics. Rubin (1987) states that the probability of getting a non-response on a survey item variable say y classically depends upon the unobserved value of y, which then creates an immense challenge in the manner in which these non-responses are handled. Frequently used procedures for handling these non-responses are all based on the supposition that the non-responses are ignorable conditional on the auxiliary variable.

In his paper Kaarik (2006) postulates that the main advantage of multiple imputations is that the same sampling weight can be adopted for all the items, as compared to the weight-adjustment technique which is naturally used for unit non-response. Frequently used imputation techniques are, but not limited to deterministic imputation, for example the mean imputation within imputation classes, and the stochastic imputation, for instance the random imputation within classes. Deterministic imputation is known to eliminate imputation variance of the estimator of a total or mean, however the distribution of item values will not be preserved.

Lavori et al. (2001) presented a simple adjusted random imputation criterion for eliminating the imputation variance of the estimator of a total or mean and simultaneously conserving the distribution of item values. In addition the estimates of the distribution function based on the imputed data were shown to remain consistent and asymptotically normal. However the criteria method did not completely eliminate the imputation variance from the estimated distribution function.

2.11 Summary

The total survey error (TSE) framework consists of those survey errors which differ over hypothetical trials of a survey known as variance and those errors which are constant for all implementations of a survey known as bias. Naturally, the types of errors survey samplers are most concerned with are related to the proposed use of the data collected. The two main uses of survey data are descriptive and analytic. To date the greatest threats to the use of descriptive data of a population are errors of non-observation such as non-response and non-coverage. On the other hand modellers are to a lesser extent concerned about these errors, but are however affected more by variables which are unreliable, have poor validity and are inconsistent. Total survey error (TSE) is used to methodically consider the various types of survey errors that occur during the design process and also in describing its quality properties when completed and also takes into consideration both measurement and representation of survey data. In the total survey error (TSE) perspective, there are cost-error trade-offs, that is, there is tension between reducing these errors and the cost of reducing them.

In the modern world characterized by advances in computerized interviewing software and electronic sample management systems, data related to quality increasingly is being collected with survey data, and being used to measure various components of error. These may include paradata , data from experimental designs and supplementary data, such as non-response, call back and follow-up questions. Each of these aid the evaluation of survey data in terms of total survey error. Biemer and Lyberg (2003) argue that the total survey error paradigm lacks a user perspective, and it requires enhancement by using a more contemporary quality paradigm i.e. one that is multidimensional and that address quality in terms of the degree to which survey data meet user specific requirements.

CHAPTER THREE

METHODOLOGY

3.1 Introduction

In recapitulation of the research problem outlined in chapter one, this methodology chapter focuses on the research design as it can influence the way in which the research is undertaken, from design through to conclusions. The focus was mainly on quantifying the total survey error using the mean square error in two stage cluster sampling. The chapter outlines the research design, study population, sampling and sample size determination, data collection, imputation and research ethics. This offers a platform of study that permits accurate assessment of the total survey error thereby facilitating the overall achievement of the research objectives.

3.2 Research Design

The research design of this study was structured around the total survey error (TSE) framework, with an attempt to identify and quantify errors that exist in sample surveys under two stage cluster sampling with equal first stage sampling units. While several up to standard metrics for measuring and quantifying total survey errors have been proposed in sample surveys, this study adopted the mean square error (MSE). We decomposed the variance into sampling error under two stage cluster sampling, using a hypothetical situation. It was vital to establish theoretical support for the practice of two-stage cluster sampling. Two-stage cluster design was preferred because it offers an opportunity for researchers to conduct analyses at more than one level of data aggregation. Furthermore, in most practical scenarios, sampling frames are seldom available and when they do, it is very expensive to obtain the sampling units directly from the target population. Therefore in such situations the two-stage cluster sampling becomes handy and its efficiency can be improved by the use of auxiliary variables.

We assumed that in any sample survey, sampling error will always exist since survey estimates

always have different values upon several replications of the sample using the same design. In addition, the bias component of the mean square error was decomposed into refusal error, non-interview error, and response error. Such decompositions were quite helpful in coming up with a mathematical expression for mean square error. We made use of auxiliary variable in the determination of the response error. Tackling the most serious errors by well thought-out survey design was made possible by decomposing these errors to smaller and smaller components. For evaluating the errors, the key components of the mean square error were estimated and combined in accordance with the decomposition formulas to form an estimate of the mean square error.

Subsequent to obtaining a mathematical expression of the total mean square error, an empirical study was conducted. The study was based on the waiting time to conception in women in Manicaland province of Zimbabwe. The province is sub-divided into ten districts and these formed the clusters.

3.3 Population

According to Easterby-Smith et al (2008), a research population is generally a large collection of individuals or objects that is the main focus of a scientific query. However, due to the large sizes of populations, researchers often cannot measure every individual in the population because it is too expensive and time-consuming. This is the reason why researchers rely on sampling techniques. In this study all women with at least one live birth residing within the province constituted the study population.

(a) Rural	Current Age of Women							
Age at First Live Birth	15 - 19	20 - 24	25 - 29	30 - 34	35 - 39	40 - 44	45 - 49	Total
15 - 19	10.3	23.5	20.1	16.9	13	8.9	7.3	100
20 - 24	-	17.4	24.9	20.2	17.5	12	8.1	100
25 - 29	-	-	21.9	25.7	22.1	18.2	12.1	100
30 - 34	-	-	-	26.2	31.1	22.4	20.4	100
35 - 39	-	-	-	-	36.6	36.0	27.5	100
40 - 44	-	-	-	-	-	40.9	59.1	100
45 - 49	-	-	-	-	-	-	100	100
Number	12664	44250	50307	43528	36016	25340	18862	230967
Median	17.5	18.8	20.1	20.3	20.7	20.8	20.3	19.7

(b) Urban	Current Age Of Women							
Age at First Live Birth	15 - 19	20 - 24	25 - 29	30 - 34	35 - 39	40 - 44	45 - 49	Total
15 - 19	8.4	23.9	22.8	18.1	12.6	8.2	6	100
20 - 24	-	18.3	27.6	21.2	15.9	10.4	6.6	100
25 - 29	-	-	27.4	27.5	20	14.9	10.1	100
30 - 34	-	-	-	30.2	31	23.5	15.3	100
35 - 39	-	-	-	-	37.3	40.7	22	100
40 - 44	-	-	-	-	-	35	65	100
45 - 49	-	-	-	-	-	-	-	-
Number	1894	10262	14387	11909	8805	5932	4018	57207
Median	17.5	19.8	21.4	21.7	21.8	22	21.8	21.1

Table 3.1: Extract of Women by Current Age, by Age with at One Live Birth for Rural and Urban, Manicaland Province, Zimbabwe 2012 Census

Table 3.1 is an extract from the Zimbabwe population census 2012 report. The table shows the total population of women with at least one live birth for both rural and urban areas as 288174. According to Cochran (1977) the variation of the cluster size is seldom important and thus for simplicity reasons, we assumed that the clusters are all of equal size $M = 20000$.

3.4 Sample size determination

Determining optimal sample size is a trade-off between the funds available and the required survey precision. In their paper, Rabe-Hesketh and Skrondal (2006) states that cluster sampling is often stimulated by cost efficiency, that is low cost of data collection per sample element. The use of cluster sampling has attracted great attention because travelling costs of interviewers can be significantly reduced thus the cost efficiency can therefore be high. However Lumley (2010) argues that there are also certain shortcomings of cluster sampling that affect statistical efficiency. If each cluster closely reflects the population structure, we would definitely attain

efficient sampling such that standard errors of estimates would not exceed those of simple random sampling. However, in practice, clusters tend to be internally homogeneous, and this intra-cluster homogeneity increases standard errors and thus decreases statistical efficiency. It is however important to determine, at the sampling design stage, the number of clusters to be selected and the respondents to be interviewed in each cluster, in order to achieve the required precision within the survey budget. Therefore the optimal sample size is a function of the cost ratio and the intra-cluster correlation.

The cost ratio of a survey represents the cost of interviewing a cluster compared to the cost of interviewing an individual (Arlene, 2003). This cost of interviewing a cluster mainly include the cost of household listing and of travelling between clusters for household listing and for individual interviews; the costs of individual interviews are mainly the interview cost and the travel cost within a cluster. In Zimbabwe the cost ratio varies from province to province depending on a number of factors which include level of urbanization, population density and the infrastructure in the province. When the cost ratio is high, it follows that travelling between clusters is expensive and it is advisable to select very few clusters and interview more individuals per cluster. Conversely better accuracy is obtained by selecting more clusters and interviewing fewer individuals per cluster. On the other hand the intra-cluster correlation is very critical in the determination of the optimal sample size at the second stage (Dillman et al., 2009).

The intra-cluster correlation measures the similarity of the individuals on the survey characteristic within a cluster. A high intra-cluster correlation means that there are strong similarities between the individuals within the same cluster and therefore a large sample taken per cluster will decrease the survey's precision. Similarly a low intra-cluster correlation means weak similarities between the individuals within the same cluster and therefore a large sample taken will decrease the survey cost Couper (2008).

Now consider a simple cost function

$$C = c_1 n + c_2 nm$$

Where C is the total cost of the survey (excluding fixed costs), c_1 is the unit cost per cluster for household listing and interviews c_2 is the unit cost per individual interview, n is the total number of clusters and m is the number of units selected in each cluster. Using the assumption that the clusters are of equal size M , the variance of the sample mean is given by (Cochran, 1977) as:

$$\begin{aligned} Var(\bar{\bar{y}}) &= \left(\frac{N-n}{N}\right)\frac{S_1^2}{n} + \left(\frac{M-m}{M}\right)\frac{S_2^2}{nm} \\ &= \left(\frac{1}{n}-\frac{1}{N}\right)S_1^2 + \left(\frac{1}{mn}-\frac{1}{Mn}\right)S_2^2 \\ &= \left(\frac{1}{n}\right)S_1^2 - \left(\frac{1}{N}\right)S_1^2 + \left(\frac{1}{mn}\right)S_2^2 - \left(\frac{1}{Mn}\right)S_2^2 \\ &= \frac{1}{n}\left[S_1^2 - \frac{1}{M}S_2^2\right] + \left(\frac{1}{mn}\right)S_2^2 - \left(\frac{1}{N}\right)S_1^2 \\ &= \frac{1}{n}\left[S_u^2\right] + \left(\frac{1}{mn}\right)S_2^2 - \left(\frac{1}{N}\right)S_1^2 \end{aligned} \tag{3.4.1}$$

where $S_u^2 = S_1^2 - \frac{1}{M}S_2^2$ also $S_1^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(\bar{Y}_i - \bar{Y}\right)^2$ and $S_2^2 = \frac{1}{N(M-1)}\sum_{i=1}^{N}\sum_{j=1}^{M}\left(Y_{ij} - \bar{Y}\right)^2$. We observe that the last term on the right of equation 3.4.1 does not depend on the choice of n and m (Cochran, 1977). Minimising the variance V for fixed C; or C for fixed V is equivalent to minimising the product;

$$\begin{aligned} \left[V + \frac{1}{N}S_1^2\right]C &= \frac{1}{n}\left[S_u^2 + \frac{S_2^2}{m}\right](c_1 + c_2 m)\,n \\ &= \left[S_u^2 + \frac{S_2^2}{m}\right](c_1 + c_2 m) \end{aligned}$$

By the Cauchy-Schwarz Inequality (Cochran, 1977)

$$m_{opt} = \frac{S_2}{S_u}\sqrt{\frac{c_1}{c_2}} \tag{3.4.2}$$

$$n_{opt} = \frac{C}{c_1 + c_2 m_{opt}} \tag{3.4.3}$$

The value of $\frac{c_1}{c_2}$ is known but the value of $\frac{S_2}{S_u}$ is not known. To calculate the optimal value of m_{opt} , we must find a way to estimate this variance ratio (Cochran, 1977). The intra-cluster correlation reflects the homogeneity of the sample. The variance can be decomposed as follows:

$$\begin{aligned} Total\,variance &= Variance\,within + Variance\,between \\ \sigma^2 &= \sigma_w^2 + \sigma_b^2 \end{aligned}$$

Thus the intra-cluster correlation is defined as:

$$\begin{aligned} \rho &= 1 - \frac{\sigma_w^2}{\sigma^2} \\ &= \frac{\sigma_b^2}{\sigma_w^2 - \sigma_b^2} \end{aligned}$$

According to (Cochran, 1977) it can be deduced that:

$$\begin{aligned} S_1^2 &\cong \frac{1}{M} S^2 \left[1 + (M-1)\rho\right] \\ S_2^2 &\cong S^2(1-\rho) \\ S_u^2 &\cong S^2\rho \end{aligned}$$

Thus the variance ratio is $\frac{S_2^2}{S_u^2}$ is given by

$$\frac{S_2^2}{S_u^2} \cong \frac{1-\rho}{\rho} \tag{3.4.4}$$

Using equation (3.4.4) in equation (3.4.2) and (3.4.3) the approximate optimal sample size is given by

$$m_{opt} = \sqrt{\left(\frac{1-\rho}{\rho}\right)\left(\frac{c_1}{c_2}\right)} \tag{3.4.5}$$

$$n_{opt} = \frac{C}{c_1 + c_2 m_{opt}} \tag{3.4.6}$$

The optimal sample size depends explicitly on the cost ratio $\frac{c_1}{c_2}$ and the intra-cluster correlation ρ , but not on the cluster size (the number of second-stage sampling units in the cluster) (Cochran, 1977). The cluster size has very little effect on the sampling error if the second-stage sample size is fixed. The optimal sample taken is an increasing function of $\frac{c_1}{c_2}$ and a decreasing function of ρ. This therefore means that if the sampling cost of drawing a cluster is high, we draw fewer clusters and more sub-sampling units within each cluster. If $\rho > 0$, it means that there is a strong intra-cluster homogeneity, and we draw fewer secondary sampling units and more clusters. On the other hand if $\rho < 0$, it means there is a strong intra-cluster heterogeneity, and we take many of the secondary sampling units in the selected cluster and use fewer clusters to decrease the sampling cost.

The data on waiting time to conception (WTC) in months was collected. The total cost of the survey was $C = US\$1467.50$, unit cost per cluster for household listing and interviews $c_1 = US\$533.17$, the unit cost per individual interview $c_2 = US\$1$ and the intra-cluster correlation was computed using Stata version 12 and the output is shown below in Table 3.2.

```
. loneway WTC Area

                    One-way Analysis of Variance for WTC:

                                                Number of obs =       239
                                                    R-squared =    0.0384

    Source                SS         df       MS            F     Prob > F
-------------------------------------------------------------------------
Between Area           370.94984       2     185.47492       4.71     0.0099
Within Area            9300.3406     236     39.408223
-------------------------------------------------------------------------
Total                  9671.2905     238     40.635674

         Intraclass       Asy.
         correlation      S.E.       [95% Conf. Interval]
    ---------------------------------------------------------
            0.04452      0.05428        0.00000      0.15090
```

Table 3.2: Intra-cluster correlation

From the output it can be seen that the intra-cluster correlation is $\rho = 0.04452$. Substituting the intra-cluster correlation in equation 3.4.5 we get:

$$\begin{aligned} m_{opt} &= \sqrt{\left(\frac{1-0.04452}{0.04452}\right)\left(\frac{523.17}{1}\right)} \\ &= 105.96 \end{aligned}$$

and

$$\begin{aligned} n_{opt} &= \frac{1467.50}{523.17+102.46} \\ &= 2.35 \end{aligned}$$

Based on these finding we concluded that the optimal sample size was $n = 3$ $m = 103$. These three clusters were selected using the simple random sampling, the same way the second stage sampling were selected.

3.5 Data collection

According to Biemer and Lyberg (2003), the research topic determines what type of data is necessary for the study. There are two main sources of data, secondary and primary data. Secondary sources are data that have already been collected. Saunders et al. (2007) distinguish three types of secondary data used in research; documentary, survey, and multiple sources. It is very common that researchers collect data from a variety of sources. In this study primary data collection was done through face-to-face interviews in the three selected clusters. The enumerators used structured interview guides to collect data from the respondents in the selected areas. Call backs were arranged where possible and some data on waiting time to conception was also obtained from local clinics in the cases of non-response. This enabled the computation of the total mean square error (MSE) of the study.

3.5.1 Pilot study

A pilot study was conducted before the start of the actual data collection. A pilot study is a mini survey which finally tests the the enumerator skills, instrument, data entry and data analysis techniques (Heeringa et al, 2010). Approximately ten respondents were randomly selected from each cluster and the intra cluster correlation was computed. This intra-cluster correlation together with the cost ratio were used to come with a representative sample.

CHAPTER FOUR

RESULTS AND DISCUSSION

4.1 Introduction

This chapter presents the research findings and discussion. These were discussed in relation to the literature of the study. The most important steps in a survey are survey design, collection and processing of data, data analysis and estimation. The main focus of this chapter is on decomposition of the total mean square error into sampling error, refusal error, non-interview error and response error. Subsequent to that, results of an empirical study using actual survey data are presented. Finally, in an attempt to improve the quality of the data, a method of handling non response errors is also shown.

4.2 Total survey error model and its estimators

In this study we observed that different interviewers, through their peculiarity, question delivery and recording habits obtained different data from the same respondent. In two stage cluster sampling with equal first stage sampling units (FSU) we assume that a population consists of N clusters each of size M. Then n clusters are selected from the N clusters by simple random sampling without replacement (SRSWOR). Furthermore we randomly select m elements from within the clusters which form units of sampling at the second stage and these are called second stage sampling units or secondary stage sample units (SSU).

We let r be respondents that refused or yield item missing data, p be non-interviews and q be the interviews such that:

$$r + p + q = mn \tag{4.2.1}$$

Also we let the full sample true mean be given by:

$$\bar{y} = \frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} y_{ij}$$

Let the mean for responses for the interviewed cases is given by:

$$\bar{x}_q = \frac{1}{nq} \sum_{i=1}^{n} \sum_{j=1}^{q} x_{ij}$$

Let the mean for true values of the interviewed cases is given by:

$$\bar{y}_q = \frac{1}{nq} \sum_{i=1}^{n} \sum_{j=1}^{q} y_{ij}$$

Let the true mean for refused and item missing data is given by:

$$\bar{y}_r = \frac{1}{nr} \sum_{i=1}^{n} \sum_{j=1}^{r} y_{ij}$$

Lastly let the true mean for non-interviews be given by:

$$\bar{y}_p = \frac{1}{np} \sum_{i=1}^{n} \sum_{j=1}^{p} y_{ij}$$

Thus the sample mean for interviewed cases can simply be expressed as

$$\bar{x}_q = \bar{y} + \frac{r}{nm} \left[\bar{y}_q - \bar{y}_r\right] + \frac{p}{nm} \left[\bar{y}_q - \bar{y}_p\right] + \left[\bar{x}_q - \bar{y}_q\right] \tag{4.2.2}$$

Therefore the Mean Square Error (MSE) of $\bar{x}_q$ is given as follows:

$$\begin{aligned}
MSE(\bar{x}_q) &= E\left[\bar{y}-\bar{Y}\right]^2 sampling\,error \\
&+ E\left[\frac{r}{nm}(\bar{y}_q-\bar{y}_r)\right]^2 refusal\,error \\
&+ E\left[\frac{p}{nm}\left(\bar{y}_q-\bar{y}_p\right)\right]^2 non-interview\,error \\
&+ E\left[\bar{x}_q-\bar{y}_q\right]^2 response\,error \\
&+ 2E\left[\left(\bar{y}-\bar{Y}\right)\frac{r}{nm}\left(\bar{y}_q-\bar{y}_r\right)\right] covariance\,between\,sampling\,\&\,refusal\,error \\
&+ 2E\left[\left(\bar{y}-\bar{Y}\right)\frac{p}{nm}\left(\bar{y}_q-\bar{y}_p\right)\right] covariance\,between\,sampling\,\&\,non-interview\,error \\
&+ 2E\left[\left(\bar{y}-\bar{Y}\right)\left(\bar{x}_q-\bar{y}_q\right)\right] covariance\,between\,sampling\,\&\,response\,error \\
&+ 2E\left[\frac{r}{nm}\left(\bar{y}_q-\bar{y}_r\right)\frac{p}{nm}\left(\bar{y}_q-\bar{y}_p\right)\right] covariance\,between\,refusal\,\&\,non-interview\,error \\
&+ 2E\left[\frac{r}{nm}\left(\bar{y}_q-\bar{y}_r\right)\left(\bar{x}_q-\bar{y}_q\right)\right] covarianve\,between\,refusal\,\&\,response\,error \\
&+ 2E\left[\frac{p}{nm}\left(\bar{y}_q-\bar{y}_p\right)\left(\bar{x}_q-\bar{y}_q\right)\right] covariance\,between\,non-interview\,\&\,response\,error
\end{aligned}$$

We assume the covariance terms are very negligible. Thus the MSE can be expressed as:

$$MSE(\bar{x}_q) = E\left[\bar{y}-\bar{Y}\right]^2 + E\left[\frac{r}{nm}(\bar{y}_q-\bar{y}_r)\right]^2 + E\left[\frac{p}{nm}\left(\bar{y}_q-\bar{y}_p\right)\right]^2 + E\left[\bar{x}_q-\bar{y}_q\right]^2 \quad (4.2.3)$$

Sampling error

The sampling error is given by:

$$E\left(\bar{y}-\bar{Y}\right)^2 = Var\left(\bar{y}\right)$$

Where in two stage cluster sampling $Var(\bar{y})$ is given by:

$$Var(\bar{y}) = Var_1\left[E_2(\bar{y})\right] + E_1\left[Var_2(\bar{y})\right] \quad (4.2.4)$$

Consider $E_2(\bar{y})$ in the equation (4.2.4) above

$$\begin{aligned} E_2(\bar{y}) &= E_2\left[\frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} y_{ij}\right] \\ &= \frac{1}{n}\sum_{i=1}^{n} E_2\left(\frac{1}{m}\sum_{j=1}^{m} y_{ij}\right) \\ &= \frac{1}{n}\sum_{i=1}^{n} E_2(\bar{y}_i) \\ &= \frac{1}{n}\sum_{i=1}^{n} \bar{Y}_i \end{aligned}$$

Now we consider $Var_2(\bar{y})$ in equation (4.2.4) above

$$\begin{aligned} Var_2(\bar{y}) &= Var_2\left[\frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} y_{ij}\right] \\ &= Var_2\left[\frac{1}{n}\sum_{i=1}^{n} \bar{y}_i\right] \\ &= \frac{1}{n^2}\sum_{i=1}^{n} Var_2(\bar{y}_i) \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\left(\frac{M-m}{M}\right)\frac{S_i^2}{m} \\ &= \frac{M-m}{Mmn}\left(\frac{1}{n}\right)\sum_{i=1}^{n} S_i^2 \end{aligned}$$

Now we consider $Var_1\left[E_2(\bar{y})\right]$ of equation (4.2.4) above

$$Var_1\left[E_2(\bar{y})\right] = Var_1\left[\frac{1}{n}\sum_{i=1}^{n}\bar{Y}_i\right]$$

For simplicity reasons we let

$$\bar{Y}_i = Z_i$$

So that

$$\begin{aligned} Var_1\left[E_2(\bar{y})\right] &= Var_1\left[\frac{1}{n}\sum_{i=1}^{n} Z_i\right] \\ &= Var_1(\bar{Z}) \\ &= \left(\frac{N-n}{N}\right)\frac{S_z^2}{n} \end{aligned}$$

Where $S_z^2 = \frac{1}{N}\sum_{i=1}^{N}(\bar{Y}_i - Y)^2$is the inter-cluster variance.

Now we consider $E_1[Var_2(\bar{y})]$in equation (4.2.4) above

$$\begin{aligned} E_1[Var_2(\bar{y})] &= E_1\left[\frac{M-m}{Mmn}\left(\frac{1}{n}\right)\sum_{i=1}^{n}S_i^2\right] \\ &= \frac{M-m}{Mmn}E\left[\frac{1}{n}\sum_{i=1}^{n}S_i^2\right] \\ &= \frac{M-m}{Mmn}\left[\frac{1}{N}\sum_{i=1}^{N}S_i^2\right] \\ &= \frac{M-m}{Mmn}S_2^2 \end{aligned}$$

Where when we average over the first stage samples $\frac{1}{n}\sum_{i=1}^{n}S_i^2$ averages to $\frac{1}{N}\sum_{i=1}^{N}S_i^2 = S_2^2$ (Cochran, 1977)

Now substituting all these expressions into equation (4.2.4) we get the sampling error as follows:

$$E(\bar{y}-\bar{Y})^2 = \left(\frac{N-n}{N}\right)\frac{S_z^2}{n} + \frac{M-m}{Mmn}S_2^2 \tag{4.2.5}$$

Refusal Error

The refusal error is given by:

$$\begin{aligned} E\left[\frac{r}{nm}(\bar{y}_q - \bar{y}_r)\right]^2 &= \frac{r^2}{n^2m^2}\left[E\left(\bar{y}_q - \bar{y}_r\right)^2\right] \\ &= \frac{r^2}{n^2m^2}\left[E\left(\bar{y}_q^2 - 2\bar{y}_q\bar{y}_r + \bar{y}_r^2\right)\right] \\ &= \frac{r^2}{n^2m^2}\left[E(\bar{y}_q^2) - 2E(\bar{y}_q\bar{y}_r) + E(\bar{y}_r^2)\right] \end{aligned}$$

Assuming the covariance term $E(\bar{y}_q\bar{y}_r) = 0$ we get the refusal error as follows:

$$\begin{aligned} E\left[\frac{r}{nm}(\bar{y}_q - \bar{y}_r)\right]^2 &= \frac{r^2}{n^2m^2}\left[E(\bar{y}_q^2) + E(\bar{y}_r^2)\right] \\ &\quad (taking\,the\,expectation\,at\,the\,second\,stage) \\ &= \frac{r^2}{n^2m^2}\left[\bar{y}_q^2 + \bar{y}_r^2\right] \end{aligned} \tag{4.2.6}$$

Non-interview error

The non-interview error is given by:

$$\begin{aligned} E\left[\frac{p}{nm}\left(\bar{y}_q - \bar{y}_p\right)\right]^2 &= \frac{p^2}{n^2m^2}\left[E\left(\bar{y}_q - \bar{y}_p\right)^2\right] \\ &= \frac{p^2}{n^2m^2}\left[E(\bar{y}_q^2 - 2y_q\bar{y}_p + \bar{y}_p^2)\right] \\ &= \frac{p^2}{n^2m^2}\left[E(\bar{y}_q^2) - 2E(y_q\bar{y}_p) + E(\bar{y}_p^2)\right] \end{aligned}$$

Assuming the covariance term $E(\bar{y}_q\bar{y}_p) = 0$ we get the non-interview error as follows:

$$\begin{aligned} E\left[\frac{p}{nm}\left(\bar{y}_q - \bar{y}_p\right)\right]^2 &= \frac{p^2}{n^2m^2}\left[E(\bar{y}_q^2) + E(\bar{y}_p^2)\right] \\ &\quad (taking\,the\,expectation\,at\,the\,second\,stage) \\ &= \frac{p^2}{n^2m^2}\left[\bar{y}_q^2 + \bar{y}_p^2\right] \end{aligned} \tag{4.2.7}$$

Response Error

The response error is given by:

$$E\left[\bar{x}_q - \bar{y}_q\right]^2 = E\left[\bar{x}_q^2 - 2\bar{x}_q\bar{y}_q + \bar{y}_q^2\right]$$

In developing the theory of sample surveys, most cases have considered only estimates based on simple averages of sample values. However there are other methods which make use of auxiliary

information and which under certain situations give more precise estimates of the population parameters. One of such methods is the ratio method of estimation which forms a basis for all other methods that use auxiliary information. Let Y_i be the survey measurement for the i^{th} unit of the population. Also let X_i be the value of the auxiliary information or measurement for the i^{th} unit. We assume that X_i are known for all the units in the population. Thus using the ratio method of estimation we let τ be the ratio estimator such that:

$$\begin{aligned} \tau &= \frac{\bar{y}_q}{\bar{x}_q} \\ \Rightarrow \bar{x}_q &= \frac{\bar{y}_q}{\tau} \end{aligned}$$

Therefore

$$\begin{aligned} E\left[\bar{x}_q - \bar{y}_q\right]^2 &= E\left[\bar{x}_q^2 - 2\bar{x}_q\bar{y}_q + \bar{y}_q^2\right] \\ &= E\left(\bar{x}_q^2\right) - 2E\left(\bar{x}_q\bar{y}_q\right) + E\left(\bar{y}_q^2\right) \\ &= \frac{1}{\tau^2}E(\bar{y}_q^2) - \frac{2}{\tau}E\left(\bar{y}_q^2\right) + E\left(\bar{y}_q^2\right) \\ &= \left[\frac{1}{\tau^2} - \frac{2}{\tau} + 1\right] E\left(\bar{y}_q^2\right) \\ &\quad \textit{taking the expectation at the second stage} \\ &= \left[\frac{1}{\tau^2} - \frac{2}{\tau} + 1\right] \left(\bar{y}_q^2\right) \end{aligned} \tag{4.2.8}$$

Finally by substituting equations (4.2.5), (4.2.6), (4.2.7) and (4.2.8) into (4.2.3) we get

$$MSE(\bar{x}_q) = \left(\frac{N-n}{N}\right)\frac{S_z^2}{n} + \frac{M-m}{Mmn}S_2^2 + \frac{r^2}{n^2m^2}\left[\bar{y}_q^2 + \bar{y}_r^2\right] + \frac{p^2}{n^2m^2}\left[\bar{y}_q^2 + \bar{y}_p^2\right] + \left[\frac{1}{\tau^2} - \frac{2}{\tau} + 1\right]\left(\bar{y}_q^2\right) \tag{4.2.9}$$

4.3 Estimator for imputation

We considered a basic model in the decomposition of the total mean square error of the variable of interest y. In both theory and practice, survey errors that arise from several sources have a

cumulative effect on the observed value y. These errors cause the observed value of y to be higher than its true value or to be lower than its true value for each observation. Mathematically we can write:

$$y_{ij} = \mu_i + \varepsilon_{ij} \tag{4.3.1}$$

Where y_{ij} is the observed value of the j^{th} unit in the i^{th} cluster and μ_i is the corresponding unknown parameter. ε_{ij} represents the cumulative effects of all error sources for the j^{th} unit in the i^{th} cluster. The error(s) will be positive for some individuals and negative for others. If the net effect of all these errors over the sample is close to zero then the estimate $\bar{y}$ will be close to the population parameter $\bar{Y}$, apart from the sampling error. Now we let $E\left(\varepsilon_{ij}\right) = B_i$ where B_i is an unobserved cluster level random effect with zero mean and finite variance. Thus equation 4.3.1 above can be written as:

$$y_{ij} = \mu_i + B_i + e_{ij} \tag{4.3.2}$$

where $\varepsilon_{ij} = B_i + e_{ij}$ is an unobserved within cluster random effect with $E(e_{ij}) = 0$ and finite variance. We further assume that the errors between any two units are uncorrelated and $B_i s$ and $e_{ij} s$ are independent. Let γ_{ij} be the response indicator defined as follows:

$$\begin{aligned} \gamma_{ij} &= 1 \; if \; y_{ij} \; is \; a \; respondent \\ \gamma_{ij} &= 0 \; if \; y_{ij} \; is \; a \; non-respondent \end{aligned}$$

According to Kaarik (2006) γ_{ij} is defined for every unit in the population and non-response is part of the model. From the specification of the sampling design, we construct a survey weight $w = \frac{1}{nm}$ so that when there is no non-response $\bar{\bar{y}} = \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} y_{ij}$ is an unbiased estimator of the population mean $\bar{Y}$. Now we impute each non-respondent by the mean:

$$\begin{aligned} \bar{y}_{imp} &= \frac{\frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\gamma_{ij}y_{ij}}{\frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\gamma_{ij}} \\ &= \frac{\sum_{i=1}^{n}\sum_{j=1}^{m}\gamma_{ij}y_{ij}}{\sum_{i=1}^{n}\sum_{j=1}^{m}\gamma_{ij}} \end{aligned}$$

Now we have

$$\begin{aligned}
E(\bar{y}_{imp}) &= E_1E_2(\bar{y}_{imp}) \\
&= E_1E_2\left[\frac{\sum_{i=1}^{n}\sum_{j=1}^{m}\gamma_{ij}y_{ij}}{\sum_{i=1}^{n}\sum_{j=1}^{m}\gamma_{ij}}\right] \\
&= E_1E_2\left[\frac{\sum_{i=1}^{n}\sum_{j=1}^{m}\gamma_{ij}\left(\mu_i+B_i+e_{ij}\right)}{\sum_{i=1}^{n}\sum_{j=1}^{m}\gamma_{ij}}\right] \\
&= E_1E_2\left[\frac{\sum_{i=1}^{n}\sum_{j=1}^{m}\gamma_{ij}\mu_i}{\sum_{i=1}^{n}\sum_{j=1}^{m}\gamma_{ij}}+\frac{\sum_{i=1}^{n}\sum_{j=1}^{m}\gamma_{ij}B_i}{\sum_{i=1}^{n}\sum_{j=1}^{m}\gamma_{ij}}+\frac{\sum_{i=1}^{n}\sum_{j=1}^{m}\gamma_{ij}e_{ij}}{\sum_{i=1}^{n}\sum_{j=1}^{m}\gamma_{ij}}\right] \\
&= E_1E_2\left[\frac{\sum_{i=1}^{n}\sum_{j=1}^{m}\gamma_{ij}\mu_i}{\sum_{i=1}^{n}\sum_{j=1}^{m}\gamma_{ij}}\right]+E_1E_2\left[\frac{\sum_{i=1}^{n}\sum_{j=1}^{m}\gamma_{ij}B_i}{\sum_{i=1}^{n}\sum_{j=1}^{m}\gamma_{ij}}\right]+E_1E_2\left[\frac{\sum_{i=1}^{n}\sum_{j=1}^{m}\gamma_{ij}e_{ij}}{\sum_{i=1}^{n}\sum_{j=1}^{m}\gamma_{ij}}\right]
\end{aligned}$$

But

$$\begin{aligned}
E_1E_2\left(\gamma_{ij}e_{ij}\right) &= E_1\left[E_2\left(\gamma_{ij}e_{ij}|B_i\right)\right] \\
&= E_1\left[E_2\left(\gamma_{ij}|B_i\right)E_2\left(e_{ij}|B_i\right)\right] \\
&= 0
\end{aligned}$$

Thus

$$E\left(\bar{y}_{imp}\right) = E_1E_2\left[\frac{\sum_{i=1}^{n}\sum_{j=1}^{m}\gamma_{ij}\mu_i}{\sum_{i=1}^{n}\sum_{j=1}^{m}\gamma_{ij}}\right]+E_1E_2\left[\frac{\sum_{i=1}^{n}\sum_{j=1}^{m}\gamma_{ij}B_i}{\sum_{i=1}^{n}\sum_{j=1}^{m}\gamma_{ij}}\right] \tag{4.3.3}$$

Now we consider the first term of equation (4.3.3) above

$$
\begin{aligned}
E_1E_2\left[\frac{\sum_{i=1}^{n}\sum_{j=1}^{m}\gamma_{ij}\mu_i}{\sum_{i=1}^{n}\sum_{j=1}^{m}\gamma_{ij}}\right] &= E_1\left\{E_2\left[\frac{\sum_{i=1}^{n}\sum_{j=1}^{m}\gamma_{ij}\mu_i}{\sum_{i=1}^{n}\sum_{j=1}^{m}\gamma_{ij}}\right]\right\} \\
When\ n\ is\ large\quad this\quad expression\ is\ approximately\ (Shao,\ 2007) & \\
&= \frac{E_1E_2\left(\sum_{i=1}^{n}\sum_{j=1}^{m}\gamma_{ij}\mu_i\right)}{E_1E_2\left(\sum_{i=1}^{n}\sum_{j=1}^{m}\gamma_{ij}\right)} \\
&= \frac{E_1\left[\sum_{i=1}^{n}\sum_{j=1}^{m}\mu_i E_2\left(\gamma_{ij}\right)\right]}{E_1\left[\sum_{i=1}^{n}\sum_{j=1}^{m}E_2\left(\gamma_{ij}\right)\right]} \\
&= \frac{E_1\left[\sum_{i=1}^{n}\sum_{j=1}^{m}\mu_i\right]}{E_1\left[\sum_{i=1}^{n}\sum_{j=1}^{m}1\right]} \\
&= \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}\mu_i \\
&\simeq \bar{Y}
\end{aligned}
$$

We note that $E_2\left(\gamma_{ij}\right)$ does not depend on (i,j)thus $E_1E_2\left(\frac{\sum_{i=1}^{n}\sum_{j=1}^{m}\gamma_{ij}\mu_i}{\sum_{i=1}^{n}\sum_{j=1}^{m}\gamma_{ij}}\right) = \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}\mu_i$. Therefore the expectation of first term in equation (4.3.3) is approximately equal to the population mean $\bar{Y}$. Now we consider the second term of equation (4.3.3)

$$
\begin{aligned}
E_1E_2\left[\frac{\sum_{i=1}^{n}\sum_{j=1}^{m}\gamma_{ij}B_i}{\sum_{i=1}^{n}\sum_{j=1}^{m}\gamma_{ij}}\right] &= E_1\left\{E_2\left[\frac{\sum_{i=1}^{n}\sum_{j=1}^{m}\gamma_{ij}B_i}{\sum_{i=1}^{n}\sum_{j=1}^{m}\gamma_{ij}}\right]\right\} \\
When\ n\ is\ large\quad this\quad expression\ is\ approximately\ (Shao,\ 2007) & \\
&= \frac{E_1E_2\left(\sum_{i=1}^{n}\sum_{j=1}^{m}\gamma_{ij}B_i\right)}{E_1E_2\left(\sum_{i=1}^{n}\sum_{j=1}^{m}\gamma_{ij}\right)} \\
&\neq 0
\end{aligned}
$$

We notice that, because γ_{ij} and B_i are dependent $E_2\left(\gamma_{ij}B_i\right) \neq 0$. Thus the second term in equation (4.3.3) is not zero and hence it is biased. This bias does not disappear asymptotically as $n \to \infty$ or as $m \to \infty\ \forall_i$. We conclude that the problem with $\bar{y}_{imp}$ is that, imputation is done over the entire sample whereas the non-response depends on a cluster level random effect.

4.4 Empirical study

In order to compute the mean square error before and after imputation, actual data from a survey was used. The survey was based on the waiting time to conception (in months) in women in Zimbabwe. Singh et al. (2007) define waiting time to conception to be time interval between the resumption of menses after a pregnancy until the beginning of the next pregnancy and is highly influenced by socio-economic, cultural, demographic and behavioural factors. The study was conducted within Manicaland which is one of the densely populated province of Zimbabwe inhabited mainly by the Shona – Manyika people. The province is sub-divided into ten districts which we took to be our clusters. From the data set before imputation we observe that $r = 27$, $p = 52$, $q = 239$. The true mean for the interviewed cases was:

$$\bar{y}_q = \frac{1}{nq}\sum_{i=1}^{n}\sum_{j=1}^{q} y_{ij} = 19.33$$

The true mean for the refused and item missing data was:

$$\bar{y}_r = \frac{1}{nr}\sum_{i=1}^{n}\sum_{j=1}^{r} y_{ij} = 20.79$$

The true mean for non-interviews was:

$$\bar{y}_p = \frac{1}{np}\sum_{i=1}^{n}\sum_{j=1}^{p} y_{ij} = 20.90$$

It is known that s_z^2 is an unbiased estimator of S_z^2 (inter-cluster variance) thus

$$\begin{aligned} E(s_z^2) &= S_z^2 \\ s_z^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 \\ &= \frac{1}{2}\left[(17.84 - 19.33)^2 + (19.33 - 19.33)^2 + (20.84 - 19.33)^2\right] \\ &= 2.25 \end{aligned}$$

Also the sum of the intra-cluster variance was

$$\sum_{i=1}^{n} S_i^2 = (17.53 + 30.88 + 70.55) = 118.96$$

In order to compute the response error $\left[\frac{1}{\tau^2} - \frac{2}{\tau} + 1\right]\left(\bar{y}_q^2\right)$ we used the ratio estimation technique. Ratio estimation is the use of known population totals for auxiliary variables so as to improve the weighting from sample values to population estimates. It functions by comparing the survey sample estimate for an auxiliary variable with the known population total for the same variable on the sampling frame. In this study the auxiliary variable X_i was the duration of breast feeding (DBF) and the results showed that the $X_i s$ and $Y_i s$ were positively correlated. The mean for responses for the interviewed cases (auxiliary variable) was given as:

$$\begin{aligned} \bar{x}_q &= \frac{1}{nq} \sum_{i=1}^{n} \sum_{j=1}^{q} x_{ij} \\ &= 20.88 \end{aligned}$$

Therefore

$$\tau = \frac{\bar{y}_q}{\bar{x}_q} = \frac{19.33}{20.88} = 0.93$$

Sampling error

$$\left(\frac{N-n}{N}\right)\frac{S_z^2}{n} + \frac{M-m}{Mmn} S_2^2 = \left(\frac{10-3}{10}\right)\frac{2.25}{3} + \left(\frac{20000-106}{20000 \times 106 \times 3}\right)\frac{1}{3} \times 118.96 = 0.649$$

Refusal error

$$\frac{r^2}{n^2 m^2}\left[\bar{y}_q^2 + \bar{y}_r^2\right] = \left(\frac{27^2}{3^2 \times 106^2}\right)\left[19.33^2 + 20.79^2\right] = 5.810$$

Non interview error

$$\frac{p^2}{n^2 m^2}\left[\bar{y}_q^2 + \bar{y}_p^2\right] = \left(\frac{52^2}{3^2 \times 106^2}\right)\left[19.33^2 + 20.90^2\right] = 21.671$$

Response error

$$\left[\frac{1}{\tau^2}-\frac{2}{\tau}+1\right]\left(\bar{y}_q^2\right)=\left[\frac{1}{0.93^2}-\frac{2}{0.93}+1\right]\left(19.33^2\right)=2.117$$

Therefore

$$MSE\left(\bar{x}_q\right)=0.649+5.810+21.671+2.117=30.247$$

Table 4.1: Contribution towards MSE

Contribution towards MSE	**Quantity**	**Percentage**
Sampling error	0.649	2.15%
Refusal error	5.810	19.21%
Non interview error	21.671	71.65%
Response error	2.117	7.00%
Total	**30.247**	**100.00%**

Table 4.1 above shows that non-response errors contributed 90.8% to the total survey error. These non-response errors are part of the total survey error that can arise from respondents refusal to answer to some specific questions (item non-response) or respondents not being at home during the time of the interview (unit non-response). Unlike sampling error, increasing the sample size will not have any effect on reducing these non-response errors. Regrettably, it is practically impossible to completely eliminate these non-sampling errors.

4.4.1 MSE after imputation

The results show that can obtain an unbiased estimator by performing imputation within each cluster. When we impute within each cluster we observed that $r=0$, $p=0$ and $q=318$ and our model becomes:

$$MSE(\bar{x}_q)=\left(\frac{N-n}{N}\right)\frac{S_z^2}{n}+\frac{M-m}{Mmn}S_2^2+\left[\frac{1}{\tau^2}-\frac{2}{\tau}+1\right]\left(\bar{y}_q^2\right) \tag{4.4.1}$$

Now from the new data set we observed that $\bar{y}_q=19.01$, $s_z^2=2.2955$, $\sum_{i=1}^{n}S_i^2=88.57$ and the

sampling error becomes

$$\begin{aligned}\left(\frac{N-n}{N}\right)\frac{S_z^2}{n}+\frac{M-m}{Mmn}S_2^2 &= \left(\frac{10-3}{10}\right)\frac{2.295}{3}+\left(\frac{20000-106}{20000\times 106\times 3}\right)\left[\frac{1}{3}(88.57)\right]\\ &= 0.627\end{aligned}$$

The response error was

$$\begin{aligned}\left[\frac{1}{\tau^2}-\frac{2}{\tau}+1\right](\bar{y}_q^2) &= \left[\frac{1}{0.91^2}-\frac{2}{0.91}+1\right](19.01^2)\\ &= 3.53\end{aligned}$$

Therefore

$$\begin{aligned}MSE\left(\bar{x}_q\right) &= 0.627+3.53\\ &= 4.157\end{aligned}$$

Contribution to MSE after imputation	**Quantity**	**Percentage**
Sampling Error	0.627	15.1%
Response Error	3.53	84.9%
Total	**4.157**	**100.0%**

Table 4.2: MSE after imputation

Table 4.2 above is a representation of the mean square error after imputation. The mean square error reduced from a total of 30.247 before imputation to a total of 4.157 after imputation. This is clear evidence of the negative effect of non-response error. Non-sampling errors, particularly item and unit non-response should be given due attention in complex surveys since they can cause enormous bias in the survey data if not controlled. It is hoped that the best way to manage these non-sampling errors is to follow the right procedures of all survey activities from planning, sampling up to the final data analysis stage.

4.5 Discussion

Time and again, non-sampling errors pose serious estimation problems in surveys research. In addition, they are very difficult to quantify and yet they receive very little attention than sampling errors. Non-sampling error can hardly be controlled by increasing the sample size. Indeed, larger samples are more difficult to handle and increasing the sample size seems very beneficial from the viewpoint of sampling errors, however it may be counter-productive from the viewpoint of non-sampling errors.

This study proved just how high the MSE can be in the presence of non-response. The chief component of bias is non-response, that is, the bias resulting from failure of some selected persons in the sample to respond to the survey. An estimate for non-response bias, assuming that non-response is the only source of bias, is expressed in (Cochran, 1977) as:

$$Bias\left(\bar{y}_R\right) = \left(1 - W_R\right)\left(\bar{Y}_R - \bar{Y}_N\right) \tag{4.5.1}$$

where W_R is the response rate, $\bar{Y}_R$ is the mean value of those who responded and $\bar{Y}_N$ is the mean value of non-respondents. Therefore the survey estimates from any particular survey are subjected to some bias when some sampled units fail to take part in the survey and when respondents are found to be different from non-respondents. Non-response bias can be significant when the response rate is pretty low and also when difference between the characteristics of respondents and non-respondents is comparatively large. This response rate is extensively used as a measure of survey quality. However response rates alone cannot be good indicators of non-response bias. We must seriously consider the difference between $\bar{Y}_R$ and $\bar{Y}_N$ as this will assist in making decisions about how to curb this non-response bias.

4.6 Summary

The total survey error (TSE) paradigm allows useful focus on errors of non-observation versus errors of observation. These errors of non-observation are comprised of non-response, coverage and sampling errors. Errors of observation encompass those of measurement emanating from the mode of data collection, Interviewer effects, measurement instrumentation and respondents themselves. Non-response and non-coverage errors, in spite of the magnificent work in imputation, weighing and adjustment, are largely ignored by many survey practitioners. Surveys try hard to increase both unit and item response so as to reduce the possibility of bias being introduced into survey estimates. Bias emanate from systematic variation between the non-respondents and the respondents. Without sufficient data on the non-respondents, surveys hardly measure differences between these two groups.

CHAPTER FIVE

CONCLUSIONS AND RECOMMENDATIONS

5.1 Introduction

This chapter the covers the conclusion and recomendations of the study in line with the objectives. Although it might seem difficult to build up the principle of minimising total survey error in terms of mathematical equations, it is a sounder practical basis for making operating decisions. It is not feasible to get rid of error sources completely, but however continuous efforts in understanding and managing variability and errors ensure that survey practitioners put into effect a high level of control on every known source of error through efficient and effective resource allocation.

5.2 Summary

This study has managed to decomposed the mean square error of an estimator into various components that constitute the major sources of errors. We obtained the closed form of the sampling error, the refusal error, the non-interview error and the response error. We also proposed an estimator for imputation and tested it for unbiasedness. Finally we did an empirical study and come up with values of the mean square error before and after imputation. We found out that both sampling and non-sampling errors have to be identified, quantified, controlled and reduced to levels where their existence does not negatively affect the usefulness of the survey data. The various types of errors that exist and the several levels of these errors, will allow survey practitioners to assess the limitations encountered in the usage of survey data and how work may be affected.

5.3 Conclusion

Basing on the results of the study we conclude that it is very important to quantify the total survey error by decomposing the mean square error into its key components for easy analysis. This decomposition enabled us to isolate the different sources of errors and quantify their contribution on total survey error using actual data. Upon proposing an estimator and testing for the unbiasedness property we concluded that non-response depends on a cluster level random effect and an estimator based on the entire sample was biased. In two stage cluster sampling an estimator based on each cluster was found to be unbiased and hence will go a long way in the minimisation of the of the total survey error. The empirical study showed that the major source of error was due to non-response (both item and unit). This alone suggests that efforts to additionally improve the quality of the survey data for this complex design should focus on reducing the effects of non-response on the data. Designing a complex survey requires many decisions as to the procedures, personnel and instruments that are used in the collection and processing of data. Since it is complicated to measure and determine numerically the effect of non-sampling errors on survey estimates, a clear consideration of the reasons why these errors occur is of paramount importance in minimizing total survey errors.

5.4 Recommendations

In order for us to be able to improve survey quality, we need to have a way of quantifying the total survey error. This way, different survey designs that satisfy the specified constraints can be measured and compared using their total survey error (TSE) as a criterion for determining the best design. This approach will likely move the design nearer to optimality if the overall effect is a reduction in the total survey error. In the decomposition of the mean square error, we assumed the covariance terms where negliglible, however we recomend further studies focusing on obtaining their closed form so that their impact can be measured. In addition, since non-response depends on a cluster level random effect we recomend that imputation should be done within each cluster as opposed to the entire sample. More work is required in the general

field of curbing non-sampling errors as we have seen in this study that the contribution of non-sampling errors to the total survey error exceeds that of sampling errors. Strategies like callbacks and the use of several data collection methods should be put in place so as to monitor the major error sources. There is great need to regularly assess the joint effects of survey errors on analysis and estimation so that continuous progress and future design optimisations are possible. Data analysis should also appropriately consider the complex sampling design like the two stage cluster sampling and the effects of non-sampling errors on the analytical results. More often than not, non-response errors have been dealt with only in the post-collection data approach, through imputation, adjustment in the weighting process and post-survey evaluations. We therefore recommend that survey practitioners should create awareness of non-response bias all the way through the total survey error process. This is achieved by making non-response part of the initial planning, and this will lead to the identification, development and implementation of procedures that minimizes non-response bias. For example they should plan community outreach programmes in order to alert the community to data collection in their area. This will boast their interest, awareness and trust as the interviewers approach the selected respondents for the interview.

REFERENCES

Arlene, F. (2003). *How to sample in surveys*. Thousand Oaks: Sage., **7**.

Biemer, P. (2011). *Total survey error design, implementation and evaluation*. Oxford University Press, 74(5):817–848.

Biemer, P. & Lyberg, L. (2003). *Introduction to survey quality*. Hoboken,NJ: JohnWiley&Sons.

Bouza, C. (2007). Ranked set subsampling the non response strata for estimating the difference of means. *Biometrical Journal, (44):903–915*.

Brick, J., Dipko, S., Presser, S., Tucker, C. & Yuan, Y. (2006). Nonresponse bias in a dual frame sample of cell and landline numbers. *Public Opinion Quarterly, (70):780–793.*

Bridewell, W. & Langley, P. (2010). *Two kinds of knowledge in scientific discovery.* SAGE Publications Ltd., London.

Carle, A. (2009). *Fitting multilevel models in complex survey data with design weights.* BMC Medical Research Methodology.

Cochran, W. G. (1977). *Sampling Techniques*. John Wiley & Sons, New York.

Couper, M. P. (2008). *Designing effective web surveys.* New York: Cambridge University Press.

Deming, W. E. (1944). On errors in surveys. *American Sociological Review, 9(4):359–369.*

Demnati, A. & Rao, J. (2004). Linearization variance estimators for survey data. *Survey Methodology, (30):17–26.*

Dillman, D., Smyth, J. & Christian, L. (2009). *Internet, mail, and mixed-mode surveys: The tailored-design method.* 3rd ed. Hoboken, NJ: John Wiley & Sons.

Dryver, A. & Thompson, S. (2005). Improved unbiased estimators in adaptive cluster sampling. *Journal of the Royal Statistical Society, (67):157–166.*

Duane, A. (2007). *Margins of error: A study of reliability in survey measurement.* New York:

Wiley.

Easterby-Smith, M., Thorpe, R. & Jackson, P. (2008). *Management Research*, 3rd ed, SAGE Publications Ltd., London.

Fay, R. (2003). Theory and application of replicate weighting for variance calculations. *American Statistical Association, pages 212–217.*

Frederick, W. H. (2005). Survey as a source of statistics and factors affecting the quality of survey statistics. *International Statistical Review, 73(2):245–248.*

Groves, R. & Magilavy, L. (1984). *An experimental measurement of total survey error.* New York: Academic press.

Groves, R. & Tourangeau, R. (2004). *Survey methodology. Hoboken,* NJ: John Wiley and Sons.

Hansen, M. H & Hurwitz, W. N (1946). The problem of non-response in sample surveys. *Journal of American Statistics Association, 41:517–529.*

Heeringa, S. West, B. & Berglund, P. (2010). *Applied survey data analysis.* Chapman and Hall CRC Press.

Kaarik, E. (2006). *Imputation algorithm using copulas.* Metodoloski ki zvezki, 3(1):109–120.

Kalton, G. & Heeringa, S. (2003). *Survey methodology.* Wiley Series Hoboken.

Lavori, P. Dawson, R. & Shera, D. (2001). *A multiple imputation strategy for clinical trials with truncation of patient data.* Statistics in Medicine.

Linacre, S. J. & Trewin, D. J. (1993). Total survey design-application to a collection of the construction industry. *Journal of Official Statistics, 9(2):611–621.*

Lohr, S. (2009). *In handbook of statistics: Sample surveys design, methods and applications.* Elsevier Amsterdam, 29A.

Lohr, S. & Rao, J. (2006). Estimation in multiple-frame surveys. *Journal of the American*

Statistical Association, (101):1019–1030.

Lumley, T. (2010). *Complex surveys. A Guide to Survey Analysis:* Wiley.

Mahalanobis, P. (1946). Recent experiments in statistical sampling in the indian statistical institute. *Journal of the Royal Statistical Society, (109):325–378.*

Montgomery, D. C. (2009). *Introduction to statistical quality control*. 6th ed. Hoboken, NJ: John Wiley & Sons.

Platek, R. & Sarndal, C. (2001). Can a statistician deliver. *Journal of Official Statistics, 17(1):1–20.*

Popinski, W. (2006). Development of the polish labour force survey. *Statistics in Transition, 7(5):1009– 1030.*

Rabe-Hesketh, S & Skrondal, A. (2006). Multilevel modeling of complex survey data. *Journal of the Royal Statistical Society-Series, (169):805–827.*

Raghunathan, T., Reiter, J. & Rubin, D. (2003). Multiple imputation for statistical disclosure limitation. *Journal of Official Statistics, (19):1–16.*

Rao, P. (1987). Ratio and regression estimates with sub-sampling the non respondents. *Journal of International Statistical Association, pages 2–16.*

Rubin, B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association, (91):473–489.*

Rubin, D. (1987). *Multiple imputation for nonresponse insurveys.* New York: John Wiley & Sons, Inc.

Saris, W. & Irmtraud, G. (2007). *Design, Evaluation and Analysis of Questionnaires for Survey Research.* New York: John Wiley & Sons.

Sarndal C & Lundstrom, S. (2005). *Estimation in surveys with nonresponse.* Chichester: John Wiley & Sons.

Saunders, M., Lewis, P. & Thornhill, A. (2007). *Research Methods for Business Students.* Prentice Hall Financial Times, Harlow, 4 edition.

Shao, J. (2007) Handling survey nonresponse in cluster sampling, *Journal of Survey Methodology Statistics Canada Vol. 33, No. 1, pp. 8185*

Sinclair, M. & Gastwirth, J. (1996). On procedures for evaluating the effectiveness of reinterview survey methods: Application to labor force data. *Journal of the American Statistical Association, (91): 961–969.*

Singh, N., Narendra, R. & Hemochandra, L. (2007). Determinants of waiting time to conception (wtc) in manipuri women 43. *Kuwait Medical Journal, 39(1):39–43.*

Sukhatme, P. V. & Seth, G. R. (1952). Nonsampling errors in surveys. *Journal of Indian Society of Agricultural Statistics, pages 5–51.*

Thompson, J. R. & Randy, R. L. (2007). An overview of normal theory structural measurement error models. *International Statistical Review, 75(2):183–198.*

Verma, V. & Betti, G. (2010). Taylor linearization sampling errors and design effects for poverty measures and other complex statistics. *Journal of Applied Statistics.*

Wolter, K. (2007). *Introduction to Variance Estimation*. New York: Springer-Verlag.

Zieba, A. & Kordos, J. (2010). *Comparing three methods of standard error estimation for poverty measures.* Research in Social and Economic Surveys.

APPENDIX 1 QUESTIONNAIRE

QUESTIONNAIRE

This study is based on waiting time to conception in women in Zimbabwe. We are kindly requesting you to spare your time to complete this questionnaire. Please note that information collected from the questionnaire will be used purely for academic purposes only and the information will be kept confidential. Please in the tick were appropriate.

SECTION A: DEMOGRAPHIC

1. What was your age at marriage .....years
2. What was your age at first pregnancy (at delivery) .....years
3. Sex of previous child Male ☐ Female ☐
4. Desired number of sons .....
5. Infant mortality .....
6. Number of wives your husband has .....
7. Parity .....
8. Regularity of menstruation

   Irregular ☐ Regular ☐ Absent ☐

SECTION B: SOCIO-ECONOMIC

9. **Place of residence** Urban ☐ Rural ☐
10. **What is your employment status?**

    Unemployed ☐ Self-Employed ☐ Formally Employed ☐
11. **Employment status of husband**

    Unemployed ☐ Self-Employed ☐ Formally Employed ☐
12. **What is your highest educational level?**

Ordinary Level	
Advanced Level	
Certificate	
Diploma	
Degree	
Did not complete secondary school	
Did not complete primary school	

Never went to school	

13. **What is the highest educational level of your husband?**

Ordinary Level	
Advanced Level	
Certificate	
Diploma	
Degree	
Did not complete secondary school	
Did not complete primary school	
Never went to school	

14. **Family income per month**

Less than US$100	
Between US$100 - US$300	
Between US$301 - US$500	
More than US$500	
No income at all	

15. **Religious sect**

Indigenous Apostolic churches	
Pentecostal Churches	
Traditional denominational churches	
None	

SECTION C: BEHAVIORAL FACTORS

16. **Do you use contraceptives** Yes ☐ No ☐

17. **Duration of breastfeeding in months** .....

18. **What is your waiting time to conception** .....**months**