International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064

Total Survey Error Model for Estimating Population Total in Two Stage Cluster Sampling

Damson Munyaradzi [1], Otieno Romanus Odhiambo [2], Orwa George Otieno [3]

[1]Pan African University Institute of Basic Science,
Technology and Innovations P.O Box 62000 00200 Nairobi-Kenya

[2, 3] Department of Statistics and Actuarial Sciences
Jomo Kenyatta University of Agriculture and Technology
P.O Box 62000-00200, Nairobi-Kenya

Abstract: ***This study is based on the total survey error paradigm in which the purpose is to examine a variety of sources of errors in a survey. The perspective taken follows an error model based on a finite population model, in which the main objective of this study is to propose a total survey error model in two stage cluster sampling. Our input is the demonstration that survey estimates have been presented with only one source of error measured, error due to sampling, resulting from the fact that survey estimates would have different values had another sample been drawn using the same sampling design. 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. The estimators used for the population parameter are seen to be subject to both variable errors and biases.***

Key words: Total Survey Error, Mean Square Error, Two-Stage Cluster Sampling

1. Introduction

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 15 years has emphasized methods of reducing sampling errors rather than minimising total survey error. 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 propose a total survey error model. 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 so that available resources can be appropriately allocated to reduce their errors as much as possible, at the same time satisfying the specified costs and timeliness objectives. According to [1] 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. The credibility and authenticity of a survey depends on quality of the survey data.

Although a sizable number of studies on nonresponse bias have been done, relatively very little is known about other sources of non-sampling error. In most studies, interviewer variance is rarely estimated in centralized telephone surveys, even though the cost of doing so routinely is relatively small. Studies of frame bias or data-processing errors are seldom reported. [9] note a lack of progress over the last 50 years in integrating sampling and non-sampling errors as measures of uncertainty.

2. Total Survey Error

According to [4], 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). Total Survey Error defines quality 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. [3] 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. [8] postulates that in the total survey error paradigm, there may be cost-error tradeoffs resulting in tension between reducing such errors and the cost incurred in reducing them.

[10] 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 , which in most cases may be written as

$$MSE(\hat{\theta}) = E(\hat{\theta} - \theta)^2$$

This can be decomposed into terms for the squared bias and the variance, as

$$MSE\left(\hat{\theta}\right) = B^2\left(\hat{\theta}\right) + Var(\hat{\theta})$$

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. According to [8], 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. [11] 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.

2.1 Interviewer effect

According to **[7]** 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 to 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.

2.2 Survey Errors

Variable errors in surveys can be measured when there exists in the design more than one unit over which errors vary and there is a randomisation step to ensure that the expected values achieved by the various units are equivalent, except for differences arising from the variable errors [6]. According to [5], sampling errors are statistical errors which survey researchers expose models simply because of working with sample data rather than whole population. These sampling errors are variable errors because the deviation of the sample mean from the true population mean will vary over replications, using the same design. Another variable error arises because errors are made in response to survey questions. Most of these errors arise because of inaccuracies on the part of the respondent. The response errors may also vary because different interviewers are assigned to administer the questionnaire to each respondent.

2.3 Total Survey Model and Its Estimators

Our model observes that different interviewers, through their peculiarity, question delivery and recording habits will obtain different data from the same respondent **[6]**. 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) [2]

Two stage sampling with equal FSU
Let **r** be elements that would refuse or yield item missing data
p would be non-interviews
q would be interviews
Such that

$$\boldsymbol{r + p + q = nm}$$

Also let

$$\overline{y} = \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} y_{ij} \ \ (full\ sample\ true\ mean)$$

$$\overline{x}_q = \frac{1}{nq}\sum_{i=1}^{n}\sum_{j=1}^{q} x_{ij} \ \ (the\ mean\ for\ responses\ for\ the\ interviewed\ cases)$$

$$\overline{y}_q = \frac{1}{nq}\sum_{i=1}^{n}\sum_{j=1}^{q} y_{ij} \ \ (the\ mean\ for\ true\ values\ of\ the\ interviewed\ cases)$$

$$\overline{y}_r = \frac{1}{nr}\sum_{i=1}^{n}\sum_{j=1}^{r} y_{ij} \ \ (the\ true\ mean\ for\ refused\ and\ item\ missing\ data)$$

$$\overline{y}_p = \frac{1}{np}\sum_{i=1}^{n}\sum_{j=1}^{p} y_{ij} \ \ (the\ true\ mean\ for\ other\ non-interviews)$$

The sample mean for interviewed cases can simply be expressed as

$$\boldsymbol{\overline{x}_q = \overline{y} + \frac{r}{nm}\left[\overline{y}_q - \overline{y}_r\right] + \frac{p}{nm}\left[\overline{y}_q - \overline{y}_p\right] + \left[\overline{x}_q - \overline{y}_q\right]} \ldots\ldots (1)$$

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

$$MSE\left(\bar{x}_q\right) = E[\bar{y} - \bar{Y}]^2 (sampling\ error)$$

$$+E\left[\frac{r}{nm}(\bar{y}_q-\bar{y}_r)\right]^2\ (refusal\ error)$$

$$+E\left[\frac{p}{nm}(\bar{y}_q-\bar{y}_p)\right]^2\ (non-interview\ error)$$

$$+E\left[(\bar{x}_q-\bar{y}_q)\right]^2(response\ error)$$

$$+2E\left[(\bar{y}-\bar{Y})\frac{r}{nm}(\bar{y}_q-\bar{y}_r)\right]\ (covariance\ between\ sampling\ and\ refusal\ error)$$

$$+2E\left[(\bar{y}-\bar{Y})\frac{p}{nm}(\bar{y}_q-\bar{y}_p)\right](covariance\ between\ sampling\ \&\ noninterview\ error)$$

$$2E\left[(\bar{y}-\bar{Y})(\bar{x}_q-\bar{y}_q)\right](covariance\ between\ sampling\ \&\ response\ error)$$

$$+2E\left[\frac{r}{nm}(\bar{y}_q-\bar{y}_r)\frac{p}{nm}(\bar{y}_q-\bar{y}_p)\right](covariance\ between\ refusal\ \&\ noninterview\ error)$$

$$+2E\left[\frac{r}{nm}(\bar{y}_q-\bar{y}_r)(\bar{x}_q-\bar{y}_q)\right](covariance\ between\ refusal\ \&\ response\ error\ )$$

$$+2E\left[\frac{p}{nm}(\bar{y}_q-\bar{y}_p)(\bar{x}_q-\bar{y}_q)\right](covariance\ between\ noninterview\ \&\ response\ error\ )$$

The covariance terms present more complicated estimation problems. It is hoped that they represent much lower order magnitude of error thus they will be assumed negligible. Thus the model can be expressed as:

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

Sampling Error

$$E[\bar{y}-\bar{Y}]^2=Var(\bar{y})$$

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

Where

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

$$=\frac{1}{n}\sum_{i=1}^{n}E_2\left(\frac{1}{m}\sum_{j=1}^{m}y_{ij}\right)$$

$$=\frac{1}{n}\sum_{i=1}^{n}E(\bar{y}_i)$$

$$\frac{1}{n}\sum_{i=1}^{n}\bar{Y}_i$$

$$Var_2(\bar{y})=Var_2\left[\frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}y_{ij}\right]$$

$$=Var_2\left[\frac{1}{n}\sum_{i=1}^{n}\bar{y}_i\right]$$

$$=\frac{1}{n^2}\sum_{i=1}^{n}Var_2(\bar{y})$$

$$=\frac{1}{n^2}\sum_{i=1}^{n}\left(\frac{M-m}{M}\right)\frac{S_i^2}{m}$$

$$=\frac{M-m}{Mmn}\left(\frac{1}{n}\right)\sum_{i=1}^{n}S_i^2$$

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

Let

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

Thus

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

$$=Var_1(\bar{Z})$$

$$=\left(\frac{N-n}{N}\right)\frac{S_z^2}{n}$$

Where

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

$$E_1[Var_2(\bar{y})]=E_1\left[\left(\frac{M-m}{Mmn}\right)\frac{1}{n}\sum_{i=1}^{n}S_i^2\right]$$

$$=\left(\frac{M-m}{Mmn}\right)E\left[\frac{1}{n}\sum_{i=1}^{n}S_i^2\right]$$

$$=\left(\frac{M-m}{Mmn}\right)\frac{1}{N}\sum_{i=1}^{N}S_i^2$$

Thus

$$\boldsymbol{Var(\bar{y})=\left(\frac{N-n}{N}\right)\frac{S_z^2}{n}+\left(\frac{M-m}{Mmn}\right)\frac{1}{N}\sum_{i=1}^{N}S_i^2}\ \ldots\ldots(3)$$

Refusal error

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

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

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

Assuming the covariance term is equal to zero

$$E\left(\overline{y}_q \overline{y}_r\right) = 0$$

Thus

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

$$= \underline{\frac{\boldsymbol{r^2}}{\boldsymbol{n^2m^2}}\left[\left(\overline{\boldsymbol{y}}_{\boldsymbol{q}}^{\boldsymbol{2}}\right) + \left(\overline{\boldsymbol{y}}_{\boldsymbol{r}}^{\boldsymbol{2}}\right)\right]} \ldots\ldots\ldots (4)$$

Non-interview error

$$E\left[\frac{p}{nm}\left(\overline{y}_q - \overline{y}_p\right)\right]^2 = \frac{p^2}{n^2m^2}\left[E\left(\overline{y}_q - \overline{y}_p\right)^2\right]$$

$$= \frac{p^2}{n^2m^2}\left[E\left(\overline{y}_q^2 - 2\overline{y}_q\overline{y}_p + \overline{y}_p^2\right)\right]$$

$$= \frac{p^2}{n^2m^2}\left[E\left(\overline{y}_q^2\right) - 2E\left(\overline{y}_q\overline{y}_p\right) + E\left(\overline{y}_p^2\right)\right]$$

Assuming the covariance term is equal to zero

$$E\left(\overline{y}_q \overline{y}_p\right) = 0$$

Thus

$$E\left[\frac{p}{nm}\left(\overline{y}_q - \overline{y}_p\right)\right]^2 = \frac{p^2}{n^2m^2}\left[E\left(\overline{y}_q^2\right) + E\left(\overline{y}_p^2\right)\right]$$

$$= \underline{\frac{\boldsymbol{p^2}}{\boldsymbol{n^2m^2}}\left[\left(\overline{\boldsymbol{y}}_{\boldsymbol{q}}^{\boldsymbol{2}}\right) + \left(\overline{\boldsymbol{y}}_{\boldsymbol{p}}^{\boldsymbol{2}}\right)\right]} \ldots\ldots\ldots (5)$$

Response error

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

In developing the theory of sample surveys, most cases have considered only estimates based on simple averages of sample values. There are other methods however which make use of auxiliary information and which under certain situations give more reliable 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 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 $\boldsymbol{\tau}$ be the ratio estimator such that:

$$\tau = \frac{\overline{y}_q}{\overline{x}_q}$$

$$\xrightarrow{yields} \overline{x}_q = \frac{\overline{y}_q}{\tau}$$

Thus

$$E\left[\overline{x}_q - \overline{y}_q\right]^2 = E\left(\frac{\overline{y}_q}{\tau}\right)^2 - 2E\left(\frac{\overline{y}_q^2}{\tau}\right) + E\left(\overline{y}_q^2\right)$$

$$= \frac{1}{\tau^2}E\left(\overline{y}_q^2\right) - \frac{2}{\tau}E\left(\overline{y}_q^2\right) + E\left(\overline{y}_q^2\right)$$

$$= \left[\frac{1}{\tau^2} - \frac{2}{\tau} + 1\right]E\left(\overline{y}_q^2\right)$$

$$= \underline{\left[\frac{\boldsymbol{1}}{\boldsymbol{\tau^2}} - \frac{\boldsymbol{2}}{\boldsymbol{\tau}} + \boldsymbol{1}\right]\left(\overline{\boldsymbol{y}}_{\boldsymbol{q}}^{\boldsymbol{2}}\right)} \ldots\ldots (6)$$

By substituting equations 3, 4, 5 and 6 into equation 2 our final model can be expressed as:

$$\boldsymbol{MSE}\left(\overline{\boldsymbol{x}}_{\boldsymbol{q}}\right) = \left\{\left(\frac{\boldsymbol{N-n}}{\boldsymbol{N}}\right)\frac{\boldsymbol{S_z^2}}{\boldsymbol{n}} + \left(\frac{\boldsymbol{M-m}}{\boldsymbol{Mmn}}\right)\frac{\boldsymbol{1}}{\boldsymbol{N}}\sum_{\boldsymbol{i=1}}^{\boldsymbol{N}}\boldsymbol{S_i^2}\right\}$$
$$+ \left\{\frac{\boldsymbol{r^2}}{\boldsymbol{n^2m^2}}\left[\left(\overline{\boldsymbol{y}}_{\boldsymbol{q}}^{\boldsymbol{2}}\right) + \left(\overline{\boldsymbol{y}}_{\boldsymbol{r}}^{\boldsymbol{2}}\right)\right]\right\}$$
$$+ \left\{\frac{\boldsymbol{p^2}}{\boldsymbol{n^2m^2}}\left[\left(\overline{\boldsymbol{y}}_{\boldsymbol{q}}^{\boldsymbol{2}}\right) + \left(\overline{\boldsymbol{y}}_{\boldsymbol{p}}^{\boldsymbol{2}}\right)\right]\right\}$$
$$+ \left\{\left[\frac{\boldsymbol{1}}{\boldsymbol{\tau^2}} - \frac{\boldsymbol{2}}{\boldsymbol{\tau}} + \boldsymbol{1}\right]\left(\overline{\boldsymbol{y}}_{\boldsymbol{q}}^{\boldsymbol{2}}\right)\right\} \ldots\ldots (7)$$

3. Conclusion and Recommendation

Each and every estimate that will be computed from the survey data based on the above model has a corresponding mean square error that summarizes the effects of all sources of error on the estimate. A small MSE indicates that the TSE is small and under control. A large MSE indicates that one or more sources of error are adversely affecting the accuracy of the estimate. One of the most important uses of the MSE is as a measure of the accuracy of survey data. The MSE concept is quite useful for understanding how the combined effects of survey errors reduce estimation accuracy

References

[1] Biemer, P.P (2011), 'Total survey error design, implementation and evaluation' *Public Opinion Quarterly*, Vol. 74, No. 5, 2010, pp. 817–848.
[2] Cochran W.C. (1997), Sampling techniques, 3rd edition, Wiley, New York.
[3] Couper, MP. (2008). Designing Effective Web Surveys. New York: Cambridge University Press.
[4] Duane, A (2007), 'Margins of Error: A Study of Reliability in Survey Measurement. New York: Wiley.
[5] Frederick W.H. (2005), Survey as a Source of Statistics and Factors Affecting the Quality of Survey Statistics, *International Statistical Review*, Vol. 73, No. 2, pp. 245 248.
[6] Groves R.M and Magilavy, L.J (1984). An experimental measurement of total survey error. New York: Academic press
[7] Mahalanobis, P.C. (1946). ''Recent Experiments in Statistical Sampling in the Indian Statistical Institute.'' *Journal of the Royal Statistical Society* 109:325–78.
[8] Montgomery, DC (2009). Introduction to Statistical Quality Control. 6th ed. Hoboken, NJ: John Wiley & Sons.
[9] Platek, R. and Sarndal, C. (2001). ''Can a Statistician Deliver?'' *Journal of Official Statistics* 17(1):1–20.
[10] Walter, K.M. (2007). Introduction to Variance Estimation, Second Edition. Springer Verlag.
[11] Zięba, A. and Kordos, J. (2010), Comparing Three Methods of Standard Error Estimation for Poverty Measures. In: J. Wywiał, W. Gamrot (Eds), *Research in Social and Economic Surveys,* Katowice, University of Economics.

Author Profile

Munyaradzi Damson received a BSc Statistics and MBA degree from the University of Zimbabwe in 2006 and 2012 respectively. Currently he is studying towards a MSc Statistics degree at Pan African University Institute of Basic Sciences, Technology and Innovation (PAUISTI) at Jomo Kenyatta University of Agriculture and Technology (JKUAT) in Kenya.

Professor Romanus Otieno Odhiambo is a professor in statistics. He received Bachelor of Education, MSc (Statistics), PhD (Statistics) from Kenyatta University. He has published over 25 papers and he is the current Deputy Vice Chancellor (Academic Affairs) at Jomo Kenyatta University of Agriculture and Technology.

Dr George Otieno Orwa is the current chairman of the Department of Statistics and Actuarial Science at Jomo Kenyatta University of Agriculture and Technology. He has over 11 publications in statistics.