ON ESTIMATION OF POPULATION PARAMETERS IN SAMPLING THEORY USING A SENSITIVE VARIABLE

ABSTRACT

of

THESIS

SUBMITTED TO

BABASAHEB BHIMRAO AMBEDKAR UNIVERSITY
(A CENTRAL UNIVERSITY)
LUCKNOW

BABASAHEB BHIMRAO AMBEDKAR UNIVERSITY
LUCKNOW
प्रज्ञा शील करुणा
ESTABLISHED 1996

FOR THE AWARD OF DEGREE OF

Doctor of Philosophy

IN

STATISTICS

Submitted by

Tarushree Bari

Enrolment No: 437/14

Co-Supervisor
Dr. Subhash Kumar Yadav
Associate Professor

Supervisor
Dr. Amit Kumar Misra
Assistant Professor

DEPARTMENT OF STATISTICS
SCHOOL OF PHYSICAL & DECISION SCIENCES
BABASAHEB BHIMRAO AMBEDKAR UNIVERSITY
(A CENTRAL UNIVERSITY)
Vidya Vihar, Raebareli Road, Lucknow-226025 (U.P.), India

2022

Abstract

One of the main purposes of statistical research is to determine the true values of population parameters. However, gathering data from every person of a huge population would be too expensive and time consuming. Rather than conducting a census, we can collect data from a sample and make inferences about the population of interest using sample statistics. A sample may not adequately represent the population due to several sampling or non sampling errors. Sampling error occurs when we operate with a subset of the population rather than the entire population. It is often possible to reduce it by increasing the sample size. While on the contrary, non-sampling errors can be a result of a variety of factors, including respondent error, measurement error, and non-response. As a result, meaningful inferences may be drawn only if the sample accurately represents the population, otherwise, the sample is skewed, and the conclusions drawn from the study are not reliable. Based on the circumstances, we could employ a variety of sampling procedures such as simple random sampling, cluster sampling, and stratified random sampling, in order to obtain a representative sample.

We mostly add information on a variable that is highly linked with the study variable. This additional information is referred to as auxiliary, ancillary, or previous information, and the variable from which it is collected is referred to as auxiliary or ancillary variable. The information on auxiliary variable may be known in advance, based on previous data, a pilot survey, or the observer's experience. When a sample is drawn, one can use auxiliary information to enhance the accuracy of estimation, regardless of the sampling design used. The ratio and regression techniques of sampling include auxiliary information into the estimation procedure to increase the precision of estimates, i.e., to provide estimates that are close to the corresponding population values, while also increasing the estimator's efficiency. It has been statistically demonstrated that the incorporation of auxiliary information in probability sampling reduces the Mean Square Error (MSE) of the estimator

of the population parameter by a significant amount. However, it is highly dependent on the manner in which the estimator has been proposed, that is, on the way that the role of auxiliary information has been taken into consideration.

Many survey practitioners have introduced and improved ratio, product, difference, exponential, and linear regression estimators in their search for the most accurate and efficient estimator of population mean. Cochran (1940, 1942) recommended the use of ratio estimator if the study and auxiliary variable are closely related and have a positive correlation. An extensive literature review has been carried out on mean estimation when the primary variable is highly sensitive and there are no auxiliary variables. Some examples are Eichhron and Hayre (1983), Gupta and Shabbir (2004) and Perri (2008). Later, a number of researchers including Gupta *et al.* (2012), Koyuncu *et al.* (2014) attempted to estimate the population mean of the sensitive study variable in presence of non-sensitive study variable.

Randomized Response Technique (RRT) is a methodology for gathering sensitive information from persons in which survey interviewers and data processors are ignorant of which of two alternative the respondents actually replied. RRT models can be broadly classified into three types: Full RRT models, in which all respondents provide scrambled responses; Partial RRT models, in which a randomly selected sub-group of respondents is asked to provide truthful responses; and Optional RRT models, in which the respondent has the option of providing either a scrambled response or a truthful response, depending on whether the respondent considers the question sensitive or non-sensitive.

The primary goal of this thesis is to strengthen the parameter estimation of a sensitive variable by making use of non-sensitive auxiliary information in the estimation process. In order to accomplish this, we introduce some population mean estimators that are based on the additive randomized response technique model. Based on the data, it is possible to derive expression for the bias and mean square error for each of the proposed estimators. Additionally, for each of the study estimators, a comprehensive simulation study is conducted, followed by an application to a real-world dataset. All of the application domains in this collection are developed using the statistical software R.

In Chapter 1, we have provided an introduction about sampling theory and sampling designs related to our study. We have presented an introduction about randomized response technique and reviewed the literature related to these problems. This chapter contains some basic notation, models and definitions relevant to the thesis.

In Chapter 2, we present a searls type regression estimator for elevated estimation of the population mean of a sensitive study variable in presence of known non-sensitive additional variable under Simple Random Sampling Scheme. The expressions for the bias and the MSE are obtained using the first order of approximation.

In Chapter 3, we generalize the Sousa *et al.* (2010) family of estimators using some new population parameters of auxiliary information based on a RRT. Further, we introduce a new efficient family of estimators for estimating the population mean of sensitivity variable using the approach given in Sousa *et al.* (2010) in the presence of the auxiliary information. The optimal value of Searl's constant is obtained using Lagrange's method of maxima-minima. Theoretical results are supported with a numerical illustration based on real data sets. In addition, a simulation study is carried out to compare the performances of the suggested and competing families of estimators.

Chapter 4 presents a new class of regression type estimator using different robust techniques like Least Trimmed Square Method (LTS), Least Median of Square (LMS), Least Absolute Deviation (LAD), Tukey M, Hampel M, Huber M and Huber MM. The study also includes a new robust method, that is, the S method of estimation. The theory is supported by real and simulated data.

In Chapter 5, under a two-phase sampling method, we introduce an enhanced double sampling generalize type estimator for the population mean of a sensitive research variable utilizing information from a non-sensitive supplementary variate. Some special cases of the suggested family of estimators are also discussed. The expressions for the bias and mean squared error of the proposed generalized estimators are derived and theoretical comparisons are made with competing estimators. Theoretical conclusions are supported with a numerical illustration based on real-world data. In addition, a simulation analysis is conducted to compare the efficiencies of the suggested and competing family of estimators.

Chapter 6 provides a broad discussion of the major findings and conclusions, as well as some suggestions for further research.

Bibliography

Cochran, W. G. (1940). The estimation of the yields of the cereal experiments by sampling for the ratio of grain to total produce. *Journal of Agricultural Science* 30, 262–275.

Cochran, W. G. (1942). Sampling theory when the sampling units are of unequal size. *Journal of the American Statistical Association* 37 (218), 199–212.

Eichhron, B. H. and Hayre, L. S. (1983). Scrambled randomized response methods for obtaining sensitive quantitative data. *Journal of Statistical Planning and Inference* 7 (4), 307–316.

Gupta, S. and Shabbir, J. (2004). Sensitivity estimation for personal interview survey questions. *Statistica* 64 (4), 643–653.

Gupta, S., Shabbir, J., Sousa, R., and Real, P. C. (2012). Estimation of the mean of a sensitive variable in the presence of auxiliary information. *Communications in Statistics-Theory and Methods* 41 (13), 2394–2404.

Koyuncu, N., S., Gupta, and Sousa, R. (2014). Exponential-type estimators of the mean of a sensitive variable in the presence of non sensitive auxiliary information. *J Communications in Statistics-Simulation and Computation* 43 (7), 1583–1594.

Perri, P. F. (2008). Modified randomized devices for Simmons model. *Model Assisted Statistics and Applications* 3 (3), 233–239.

Sousa, R., Shabbir, J., Real, P. C., and S., Gupta (2010). Ratio estimation of the mean of a sensitive variable in the presence of auxiliary information. *Journal of Statistical Theory and Practice* 4 (3), 495–507.