Estimation from Super-Population in case of Finite Population with Incomplete Sampling Frame

Dr. Bhawna Agarwal[1], Dr. Sumer Singh[2]

[1]*Associate Professor, Amity College of Commerce and Finance, Amity University, Noida, India*
[2]*Vice-Chancellor, Sangam University, Bhilwara, Rajasthan, India*

***Abstract*: - Sampling theory involves various sampling schemes like simple random sampling, stratified sampling, systematic sampling, pps sampling, etc. which mostly considered a population of finite size. Also the workers thought that a complete sampling frame for the population to be sampled is at hand. But this approach ignored two important aspects of sampling theory namely,(i) every finite population is a constituent part of some superpopulation , (ii) complete sampling frame is often not available for finite population under consideration. In the present paper, the author has taken into consideration both these aspects and the estimators for total of some variable (characteristics) and its variance have been delineated under various situations. It is shown that these estimators provide better estimates than those in vogue.**

***Keywords*: Super-population, Incomplete sampling frame, Non-included units, Estimator, Infinite population, Finite correction factor.**

I. INTRODUCTION

All statistical studies are conducted to study a population. In a few cases, complete enumeration process is adopted and otherwise mostly sampling studies are conducted. Truly speaking, a sample is a replica of a population. These samples are drawn from various types of populations. So it seems germane to distinctly specify some unusual types of populations. In general, one comes across two types of population i.e. finite and infinite populations. Finite population is one which consists of a fixed countable number of sampling units whatever may be their origin. Infinite population consists of those sampling units which are generated through an unending process, may it be experimental or natural. Also the sampling units are not enumerable. But for studies, one restricts to a limited number of sampling units. Actually we study a target population which is a part of a bigger population that we call Super-population. Let us consider the population of India as per census 2011. Population is increasing every year. So we are drawing conclusion for much bigger population than that we have included in our study. In marketing research, one selects a sample of some known or likely buyers from a section of the society. But the population of buyers is unlimited that is a super-population about which we are drawing the conclusions. Let us consider another example of eggs laid by hens in a country or world. The population of eggs is non-enumerable and a continuing process. Hence, it amounts to super-population. For any study, we draw a sample from poultry farm(s) and by the time a sample is selected, hens lay more eggs which are not the part of our sampling frame. This example clearly shows the need of estimation from super population having finite population with incomplete frame. Any number of such examples can be cited. Paper is fully concern with the estimation of parameters, therefore it is relevant to express estimator in simple words. "An estimator is a composite formula of sample variate values of a variable which equivalently stands for a parameter of the parent population. When the actual values are substituted in the formula, it results into a single value which is known as estimate i.e., a value to represent the corresponding parametric value".

In this paper the author has developed formulae for estimating the total, mean and variance for super - population in various situations.

II. REVIEW OF LITERATURE

The origin of super-population can be found in the works of Cochran (1946) and later in 1977, he used two variances namely S^2, for finite population and σ^2 for large populations. Super-population models need not be Bayesian in the sense of expressing personal subjective belief. In this paper, the author has kept the Bayesian approach out of context. Besides Cochran, many other workers like Deming and Stephan (1941), Madow and Madow (1944) can be cited in the literature.

Hartley and Sielken (1975) gave a super - population view point for finite population sampling. They explained that stochastic process generating a sample of n units follows basically two steps:

Step 1: Draw a 'large sample' of size N from an infinite population. This step is an imaginary step where it is assumed that the resulting sample elements are independent and identically distributed.

Step 2: Draw a sample of size $n < N$ from the large sample of size N drawn in the first step. This step is real with consideration of some sample design.

Therefore, it was explained by them that super-population theory is concerned with repeated implementation of two stochastic process (Imaginary step and real step as mentioned above).

Singh (1977) proposed suitable estimation procedure for making use of incomplete auxiliary information available for several characters.

Cho and Eltinge (2001) published a paper on Diagnostics for Evaluation of Superpopulation Models for Variance Estimation under Systematic Sampling. In that, they discussed that under systematic sampling with multiple random starts, one may use variance estimators based, respectively, on (1) a relatively simple design based approach; or (2) specific superpopulation models. Variance estimators derived from (1) generally will be approximately design unbiased, but may be somewhat unstable if the number of random starts is small or moderate. In addition, the performance of estimators based on (2) will depend on the extent to which the underlying finite population is consistent with the assumed superpopulation model. This paper considers diagnostics for the comparison of estimators from (1) and (2), with special emphasis on (a) exploratory analysis of the underlying finite population; (b) variance estimator bias; (c) variance estimator stability; and (d) coverage rates and widths of associated confidence intervals. Some of the proposed methods are applied to sample data from the U.S. Bureau of Labor Statistics also.

Zhengdong (2011) discussed the sampling frame errors as non-sampling errors. First a brief review of the sampling frame, together with the type and structure of the sampling frame, is given. Next the distinction between sampling frame errors and sampling errors is made theoretically in general. Then through the analysis of a series of non-random impact factors and the application of corresponding improvements or solutions, the sampling frame errors are reduced or controlled within a certain range. Finally, this paper summed up and sorted out the influencing factors based on the sample units or elements for the sampling frame, and also discussed the problems and solutions.

Fahimi et al.(2015) proposed Scientific Surveys Based on Incomplete Sampling Frames and High Rates of Nonresponse. They proposed a robust weighting methodology that can reduce the inherent biases associated with nonprobability samples, as well as probability-based sample surveys that suffer from incomplete frames and high rates of nonresponse. The efficacy of the proposed methodology is assessed in light of comparisons of survey estimates to external benchmarks, relying on parallel surveys that were conducted in two states using both probability-based and non-probability samples.

III. OBJECTIVE

Not much work has been done on super-population concept. Whatever work has been done is by taking finite population having known complete sampling frame. Korn and Graubard (1994) investigated variance estimation for superpopulation parameters under some general without replacement sampling designs and if the finite population correction factor is being ignored, with replacement variance estimators can be used. Also, Agarwal and Gupta (2008) expatiated estimators for population parameters in case of finite population with incomplete sampling frame. Now the present paper pertains to estimation of certain parameters where super-population having finite population is with incomplete sampling frame. An attempt is also made to show their superiority as well.

IV. ESTIMATION OF PARAMETERS

As the super-population is considered as hypothetical or imaginary or theoretical, it is required to refer the 'repeated sampling variance' of a statistic. When we discuss about super-population, actually we are supposed to consider the following steps:

Step 1: Consider an imaginary infinite population that is called super-population.

Step 2: Select a sample of N units which is actually a finite population of variables Y_1, Y_2, Y_3……Y_N which are assumed to be independent, with Y_i being a realization of a random variable with mean μ_i and variance σ_i^2. The (μ_i, σ_i^2) are assumed to be independent and identically distributed from a distribution $F(\mu,\sigma^2)$ and the super-population mean is defined as $\mu_{SP} = E(\mu)$

Now, the sampling frame which is considered here in incomplete having two types of units (i) included units (ii) non-included units with two kinds of situations:

(1) When non-included units behave in the same manner of included units.
(2) When non-included units behave differently than included units.

Let there be N units in a sampling frame and no. of non-included units between sampling units and at the end be U_1……..M_1; U_2…….M_2; U_3……..M_3;……..;U_N……..M_N where M_i is the no. of non-included units in the i^{th} gap i.e. in between i^{th} and $(i+1)^{th}$ unit.

Suppose $\sum_{i=1}^{N} M_i = M \quad ; \quad 0 \le M_i \le M$

A sample of size n from N units is selected by SRSWOR. Suppose that the character under study is Y and the sample values be y_j (j=1,2,3…….n). All the units in the gaps between selected units which are not in the frame are automatically selected in the sample. The non-included units in between are M_j (j=1,2,3……n). Thus the average no. of non-included units between two included units is :

$$\overline{m} = \frac{1}{n}\sum_{j=1}^{n} M_j$$

$\overline{m}$ provides an unbiased estimator of $\overline{M} = \frac{1}{N}\sum_{j=1}^{N} M_j$

Whereas $\overline{M}$, the average no. of non-included units between any two included units in the population. Thus, an estimate of the total number of non-included units of the target population is given by N $\overline{m}$ and the variance of $\overline{m}$ is

$$v(\overline{m}) = (\frac{1}{n} - \frac{1}{N})S_m^2$$

Theorem 4.1 Under this sampling scheme, if we consider the first situation when the non-included units behave similar to the included units of the frame, the unbiased estimator of population total is given by

$$T_1 = N\overline{y}_n + N\overline{m}\overline{y}_n = N(1+\overline{m})\overline{y}_n$$

And its variance and consistent estimator of this variance is given by

$$v(T_1) = N^2[(1+\overline{M})^2 v(\overline{y}_n) + v(\overline{m})(\overline{y}_n^2 + v(\overline{y}_n)]$$

$$\hat{v}(T_1) = N^2[(1+\overline{m})^2 \hat{v}(\overline{y}_n) + \hat{v}(\overline{m})(\overline{y}_n^2 + \hat{v}(\overline{y}_n)]$$

Here, when we discuss about SRSWOR, the estimator of variance is given by

$$\hat{v}(\overline{m}) = \frac{1}{n}\sum_{j=1}^{n}(M_j - \overline{m})^2$$

And if we look at the super-population concept, and the unbiased estimator of the repeated sampling variance of $\overline{y}_n$ is given by

$$\hat{v}(\overline{y}_n) = \hat{\text{va}}\, r_{wor}(\overline{y}_n) = \frac{(1-f)}{n}\frac{1}{n-1}\sum_{i=1}^{n}(yi - \overline{y}_n)^2$$

where (1-f) is a finite correction factor.

Therefore, the estimator of variance of population total is:

$$\hat{v}(T_1) = N^2[(1+\overline{m})^2\frac{(1-f)}{n}\frac{1}{n-1}\sum_{i=1}^{n}(yi - \overline{y}_n)^2 + \frac{1}{n}\sum_{j=1}^{n}(M_j - \overline{m})^2(\overline{y}_n^2 + \frac{(1-f)}{n}\frac{1}{n-1}\sum_{i=1}^{n}(yi - \overline{y}_n)^2)]$$

The above formula for the variance of the population total clearly reveals that as the value of (1 − f) decreases, the variance of total decreases exorbitantly. In case of super-population having finite population with incomplete sampling frame, keeping the value of f large is quite feasible.

Theorem 4.2 Under the scheme of SRSWR, if we consider the first situation when the non-included units in the frame behave same as the included units

$$T_2 = N\overline{y}_n + N\overline{m}\overline{y}_n = N(1+\overline{m})\overline{y}_n$$

and

$$v(T_2) = N^2[(v(\overline{y}_n) + v(\overline{m}\overline{y}_n) + 2\,\text{cov}(\overline{y}_n, \overline{m}\overline{y}_n)]$$

Where

$$\text{cov}(\overline{y}_n, \overline{m}\overline{y}_n) = \overline{M}v(\overline{y}_n)$$

And estimator of variance is given by

$$\hat{v}(T_2) = N^2[(1+\overline{M})\hat{v}(\overline{y}_n) + \hat{v}(\overline{m})(\overline{Y}_N^2 + \hat{v}(\overline{y}_n))]$$

Now, when we consider the superpopulation model, if finite correction factor is one in SRSWOR, it behaves as SRSWR as mentioned below:

$$\hat{v}(\overline{y}_n) = \hat{\text{va}}\, r_{wr}(\overline{y}_n) = \frac{1}{n}\frac{1}{n-1}\sum_{i=1}^{n}(yi - \overline{y}_n)^2$$

Therefore the estimator of population total with incomplete frame in case of simple random sampling with replacement under superpopulation model will be:

$$\hat{v}(T_2) = N^2[(1+\overline{M})\frac{1}{n}\frac{1}{n-1}\sum_{i=1}^{n}(yi - \overline{y}_n)^2 + \frac{1}{n}\sum_{j=1}^{n}(M_j - \overline{m})^2(\overline{Y}_N^2 + \frac{1}{n}\frac{1}{n-1}\sum_{i=1}^{n}(yi - \overline{y}_n)^2)]$$

The above formula for the variance of the total depicts that the variance is larger when than SRSWOR when f is negligible i.e.,$1 - f \approx 1$.

VI. CONCLUSION

This paper pertains for estimation from super-population in case of finite population with incomplete sampling frame which provides estimators in more realistic situation as compared to finite population concept. The formulae derived for the estimation of variance of population total reveal the real situation which will be usable in surveys and research in times to come.

REFERENCES

[1]. Agarwal, B. & Gupta, P.C. (2008). Estimation from incomplete sampling frames in case of simple random sampling. Model Assisted Statistics and Applications, 3, 113-117.

[2]. Cochran, W.G. (1946). Relative accuracy of systematic and stratified random samples for a certain class of populations. The Annals of Mathematical Statistics, 17, 164-177.

[3]. Cochran, W.G. (1977). Sampling Techniques, Third Edition. Wiley, New York.

[4]. Cho M.J. & Eltinge, J.L. (2001). Diagnostics for evaluation of superpopulation models for variance estimation under systematic sampling. Proceedings of the Survey Research Methods Section, American Statistical Association.

[5]. Fahimi, M., Barlas, F.M., Thomas, R.K. & Buttermore, N. (2015). Scientific surveys based on incomplete sampling frames and high rates of nonresponse. Survey Practice, 8(5), 1-11.

[6]. Hartley, H.O. & Sielken, R.L. (1975). A super-population view point for finite population sampling. Biometics, 31, 411-422.

[7]. Korn, E.L. & Graubard, B.I. (1994). Variance estimation for superpopulation parameters: should one use with replacement estimators? Proceedings of the Survey Research Methods Section, American Statistical Association, 124-132.

[8]. Madow, W.G. & Madow, L.H. (1944). On theory of systematic sampling. The Annals of Mathematical Statistics, 15(1), 1-24.
[9]. Singh, R. (1977). A note on the use of incomplete multi-auxiliary information in sample surveys. Australian Journal of Statistics, 19(2), 105-107.
[10]. Zhengdong, Li (2011). Error analysis of sampling frame in sample survey. Studies in Sociology of Science, 2(1), 14-21.