Relative Efficiency of Estimator of the Population Mean in Sample Random Sampling with Post-Stratification

Varit Intrama
Graduate School of Applied Statistics
National Institute of Development Administration, Thailand
varit.i@acc.msu.ac.th

Prof. Prachoom Suwattee¹, Ph.D.
Graduate School of Applied Statistics
National Institute of Development Administration, Thailand
prachoom@as.nida.ac.th

Preecha Vonghirandecha
Computer Science Department, Faculty of Science
Prince of Songkla University, Thailand
preecha.v@psu.ac.th

Abstract - The population mean is one of the most popular parameters when one wants to study the population characteristics. The estimation of the population parameters depends largely on the sampling design applied in the survey. This research aimed to compare the estimates of the population mean when a simple random sampling (SRS) was drawn from the population and the population mean was estimated from the original sample without post-stratification (PS). The study was carried out by simulation. The values of an independent variable were simulated as chi-square distributions with degrees of freedom (DF) 5, 10, 20 and 50 and the values of the study variable were generated from the independent variable so that their correlation coefficients were at 0.10, 0.20, …, 0.90. From each simulated population of size 5,000, a simple random sample of certain size was drawn without replacement and the population mean was estimated unbiasedly by the sample mean. The sample was also post stratified by the frequency of the independent variable similarly to the population. Unbiased estimate of the population mean was also estimated from the post-stratified sample. The variances of the estimates were then compared by the sample size, the degrees of freedom of the independent variable and the levels of correlation of the two types of variables.

The result indicated that the PS estimate was more efficient than the SRS estimate when the DF of the independent variable as well as the correlation coefficient increased. However, the relative efficiency was not significantly changed for large samples.

From the research findings, post-stratification should be useful in survey research, if a simple random sampling is applied in the estimation of population characteristics.

Keywords- Efficiency of estimate of the mean, Interdisciplinary research, Post-Stratification, Sample design, Simple random sampling, Social science research.
I. INTRODUCTION

Sample surveys play important role in social science research and also in interdisciplinary research. One of the most popular sample designs used by survey researchers is simple random sampling (SRS) design which is often applied without consideration of the population random variable one wants to study. When the population distribution is highly skewed, to the right or to the left, like the Pereto distribution or the chi-square distributions with low degrees of freedom (DF), SRS might not be the most appropriate sampling design to study the nature of the population. This is because most estimators of parameters in SRS design are unweighted and hence treat the importance of all the sampled data equally. Some researchers know quite well that in certain situations some sampling units in the population may have more contributions to the values of parameters in the population than the others. In such cases, different weights should be applied to different sample data.

To obtain more precise estimates of some parameters the researcher should consider the importance of the sample units or the data from a sample drawn from a population differently. One of the sampling designs that applies different weights to the sampling units drawn from different subsets of the population is stratified sampling (SS) design. In stratified random sampling, the population is partitioned into a number of strata and a simple random sample is drawn from each stratum independently from the others. For the estimation of the population parameters in such a design, one uses different weights to the units drawn from different strata to obtain unbiased estimates. By this way, the estimates of the population characteristics from SS design are usually more precise than those from other designs (Suwattee, 2009).

Despite the fact that social science researchers pay some attention on the effects of sample designs to the precision of the estimates of population characteristics, sometimes there are some limitations that prevent them from using suitable sample designs. They might have to disregard proper sampling techniques and choose to use the most simple sampling design. Most of them still use to draw a random sample from the population under study. When SRS design is adopted in a sample survey, the researcher still can improve the estimates of the parameters by applying proper weights to the sample data. The weights of the data may be obtained by way of post stratification (PS). For PS an original sample may be drawn by SRS and the data are collected from the units drawn. However, the SRS sample is post stratified by some auxiliary variable highly related to the the study variable. The population characteristics are then estimated from the SRS sample data using the weights of the post strata. Sometimes, the weights are obtained from the population structure with respect to the study variable.

This research aims to study the usefulness of PS when a SRS design is applied to draw a sample from each of the skewed populations. The relative efficiencies of the PS estimates of the population means with respect to the SRS unbiased estimates. The effects of the sample sizes, the levels of correlation between the study variable and the auxiliary variable used as post stratification variable, as well as the skewness of the population will be investigated by simulation.

II. RESEARCH METHODOLOGY

This study s carried out by simulation. Four populations, each with 5,000 data points, are generated. In the first place, generate 4 sets of 5,000 values of an independent variable (X) distributed as chi-square with degrees of freedom equal to 5, 10, 20 and 50. From each set of the X values, construct the values of the dependent variable (Y) so that the correlation coefficients between X and Y are equal to (X,Y) = 0.10, 0.20, ..., 0.90. So 40 populations of (X,Y) pairs, each of size
5,000, are obtained. To get such a population, 5,000 values of another variable (W) are generated as chi-square independently from X with specified DF. Then the Y values are obtained from X and W by \( Y = X + (1-\alpha) W \), where \( \alpha \) comes from the equation \( \rho(X,Y) = \frac{\alpha}{\sqrt{\alpha^2 + (1-\alpha)^2}} \), (Worrapak, 2007). Each population is also stratified into \( L = 2, 3 \) and 4 strata using X as the stratification variable by the cumulative square root frequency rule \( X_i = X + \sqrt{\sum_{j=1}^{i} f_j / L}, \ L = 1, 2, \ldots, L-1 \) (Suwattee, 2009), so that the stratum weights \( W_i \) are obtained. The population mean of Y is to be estimated unbiasedly by the sample mean \( \overline{Y}_{SRS} = \frac{1}{n} \sum_{i=1}^{n} Y_i \), from SRS samples of sizes \( n = 50, 100 \) and 200. This process is repeated \( k = 1,000 \) times and the variance of the sample mean \( \overline{Y} \) is estimated by \( S^2(\overline{Y}) = \frac{1}{k(k-1)} \sum_{i=1}^{k} (\overline{Y}_i - \overline{Y})^2 \). The data from each sample drawn as SRS are post stratified into \( L \) strata by the X variable as the population and the population mean of Y is estimated by \( \overline{Y}_{PS} = \sum_{i=1}^{L} W_i \overline{Y}_i \), where \( \overline{Y}_i \) is the sample mean in stratum \( i, i = 1, 2, \ldots, L \). The variance of the estimated mean \( \overline{Y}_{PS} \) from post stratification is then computed from the 1,000 replications as before. The estimates of the population mean \( \overline{Y} \) by SRS and by PS are then compared by considering the relative efficiency of \( \overline{Y}_{PS} \) with respect to \( \overline{Y}_{SRS} \), i.e. \( RE(\overline{Y}_{PS} / \overline{Y}_{SRS}) = S^2(\overline{Y}_{SRS}) / S^2(\overline{Y}_{PS}) \).

III. RESULTS AND CONCLUSIONS

The study indicates that most of the Fisher’s information of the sample mean by PS, \( 1 / V(\overline{Y}_{PS}) \), are larger than those of the sample mean by SRS, \( 1 / V(\overline{Y}_{SRS}) \). For all levels of DF’s of the auxiliary variable X (Figure 1). This means \( \overline{Y}_{PS} \) is more efficient than \( \overline{Y}_{SRS} \) when the population skewness is large which is shown by the small values of DF’s. The two estimates are not much different in terms of their variances when the DF’s of X are larger.

![Fig. 1 Fisher’s Information of the Estimates by DF of the Chi Square Distribution of X](image-url)
Varit Intrama, Prof. Prachoom Suwattee, and Preecha Vonghirandecha

For low DF (5 and 10), Figures 2.1 and 2.2 indicate that the RE of $\bar{Y}_{PS}$ with respect to $\bar{Y}_{SRS}$ increases as the values of the correlation coefficients increase when the distribution of the auxiliary variable and hence the study variable is highly skewed. For large DF (20 and 50), however, the RE dose not change much as the correlation coefficient increases. This observation is true for any sample size. It should be noted also that, for large sample with high DF, the RE’s are not much different.

![Diagram](image1)

**Fig.2 Changes in Relative Efficiencies by DF and Correlation Coefficients**

**IV. PRACTICAL APPLICATION**

The results of this research indicate that when SRS sample design is applied in a sample survey, the researcher should be cautioned by the fact that SRS design is not suitable for highly skewed probability distribution populations. The weakness of this sample design may be altered by the application of PS. In other words, the distribution of the population should be taken into consideration, at least for the shape of the distribution. The analysis of the sample data, especially the estimations as parameters can be improved using the proper weights corresponding to the importance of the sample observations.

**V. LIMITATIONS AND SUGGESTIONS**

This study may not give clear and definite conclusions because the study applies simulation technique to show the trend or changes of the precision of the estimates involved. There is no mathematical proof of the problem. We considered only the chi-square distribution of the population as a representative of the skewed distributions in order to indicate the effects on the precision of the estimates of the population mean when SRS design is adopted in a survey research. This is done through the post stratification. One might try to use some other probability distribution to study the changes in precision of the sample estimates of the population parameters.
Another suggestion is that if survey research is carried out with a sample drawn from a population by simple random sampling without consideration of the distribution, then better results might be obtained using post stratification at the analysis of data stage. It is well known that researchers usually draw random samples from a population by simple random sampling or even by some other sample designs but the sample data should be analyzed properly.

REFERENCES

1Corresponding Author