A Comparison between Ratio Estimation and Post-Stratification

Chang-Tai Chao and Tzu-Ching Chiang
Department of Statistics, National Cheng Kung University, Taiwan

Abstract

The information obtained from the auxiliary variables can be utilized in the design or/and inference stages of sampling survey. For better inference of the population quantity of interest, ratio estimator is often recommended when certain auxiliary variable is available. By taking advantage of the correlation between the variable of primary interest and the auxiliary variable, though design-biased, ratio estimator often can provide more efficient estimation result than the sample mean under simple random sampling. On the other hand, post-stratification is widely used in practice , especially when the population cannot be stratified beforehand. It is also able to provide more efficient estimate than the overall sample mean of the population variable of interest. In this research, these two estimation methods are compared in terms of their mean squared error via a design-based perspective. We also discuss the situations in which whether ratio estimator or post-stratification should be used in practice.

Key Words: Auxiliary variable, Ratio estimation, Post-stratification, Relative efficiency, Design-based sampling strategy.

1. Introduction

The primary objective of sampling survey is to collect data and then accordingly make inference for the population quantity of interest, which is a function of the population variable of interest, denoted as y. From the view pint of design-based sampling, no sampling strategy can be uniformly better than others because there is no complete sufficient statistic in a design-based approach (Godambe 1955). Hence, how to establish a better inference under different types of sampling survey situation is always an important issue in this field, and one can approach this via either different designs or inference methods, or both.

In a sampling survey situation, the investigators often collect observations from more than one variable, including the variable of interest y and some auxiliary variables x's. For example, to estimate the average household living expense, the variable of interest is the living expense of a household, and the auxiliary variable can be the total income, the number of household members, the social status or the residential area of the household. For obtaining a better inference, one would like to utilize the information provided by the auxiliary variable to make the best use of the survey data.

The utilization of the auxiliary information in a sampling survey can be roughly divided into design and inference stages. From the design prospective, stratified sampling is an family of designs that makes use of the auxiliary information in the design stage and it is often able to provide better inference results. Under a stratified sampling, the population is divided into several strata and then a sample is selected by some design within each stratum. The within-stratum designs are not necessarily to be the same or restricted by any other certain condition, but most importantly the selections of the within-strata sample have to be independent. Stratified sampling is known to be able to select sample which is more representative, and provide more precise estimation results. However, often in practice it is impossible to stratify the population before the sample is selected. For example, it is possible to stratify the population by different region beforehand in a telephone survey, but not by the age or gender. Generally speaking, a population could not be stratified before the survey by criteria which are unknown before the sampled units has been observed. For such a situation, a sample might be selected by a simple random sampling, and still one could stratify the sample into strata after the sample has been observed, and a stratified estimate will be used. Such a procedure, referred as post-stratification, is often used in practice to utilize the auxiliary information in the inference stage.

Instead of fixed constants, the within-stratum sample sizes are random variables under the post-stratification. Hence extra variability would be introduced via the random within-stratum sample sizes. However, the related stratified estimation result can be expected to be improved from the original non-stratified estimate. The main reason is that, on average the results provided by post-stratification is similar to what provided by a usual stratified design with proportional allocation, and it is an well-

known optimal allocation method when the within-stratum variances are unknown.

On the other hand, ratio estimator is a widely used estimator that utilizes the data from the y together with an auxiliary variable x. For example, the area of tillage can be considered as a useful auxiliary variable when the harvest is the population quantity of interest. Also, the amount of food resource can be used as an auxiliary variable when the number of certain species of animal is of primary interest. Although it is design-biased, ratio estimator is well-known for its ability to provide more efficient estimate in terms of giving lower mean-square error when certain correlation exists between y and x. The advantage becomes more considerable as the correlation between y and x increases (e.g. Lohr 1999 pp.71).

With a quantitative auxiliary variable, either one of post-stratification and ratio estimation could be used. The exact values of the auxiliary variable of each population units are not necessary in both methods. However, the population mean or population total of x is assumed given in ratio estimation, whereas it is not a necessary given condition in post-stratification. Furthermore, the x-value of each sampled unit usually has to be measured for ratio estimation, but as long as we could categorize the sampled units into different strata based on x, the exact x-values are not necessary for each sampled unit. Hence, compared to ratio estimation, seemingly post-stratification requires less population information and sampling cost. Hence, a question arises naturally is that which of these two methods, both are widely used in practice, should be suggested according to their regarding performances.

In this article, we will study the performances of the inferences established by post-stratification and ratio estimation in terms of the relative efficiency. The general estimation procedures with an assumed SRSWOR (simple random sampling without replacement) design are given in Section 2. together with brief descriptions of their related properties. A design-based comparison of these two methods based on a closed form of the difference between their mean squared estimation errors is given in Section 3.. Although no closed form available at this moment, we will brief address their performances via a model-based perspective in Section 3. as well. We also studied the impacts of different factors, such as the number of strata, correlation coefficient between y and x, and sample size n on the relative efficiency of ratio estimation to post-stratification by a numerical simulation study. The related results are presented in Section 4.. We will conclude our findings in Section 5. with suggestions of how to properly utilize auxiliary variable in practice.

2. Estimation Procedures and Related Properties

As in the usual finite population sampling situation, the population is considered to consist of N units that labeled from 1 to N, denoted as $u = \{1, 2, \ldots, N\}$. Associated with each unit i, the values of population variable of interest and the auxiliary variable are denoted as y_i and x_i, respectively. In this article, the sample is selected by a SRSWOR design. The data d with sample size n is a collection of the labels of sampled units $s = \{i_1, i_2, \ldots, i_n\}, i_j \in u$, the associated y-values $\mathbf{y}_s$, and x-values $\mathbf{x}_s$. That is, $d = \{s, \mathbf{y}_s, \mathbf{x}_s\}$. The population quantity of interest to be estimated is the population mean of y

$$\mu = \frac{1}{N}\sum_{i=1}^{N} y_i.$$

Furthermore, the population mean of x

$$\mu_x = \frac{1}{N}\sum_{i=1}^{N} x_i$$

is assumed given when ratio estimation is used. Under an SRSWOR design, the sample mean of y

$$\bar{y} = \frac{1}{n}\cdot\sum_{i\in s} y_i$$

is an unbiased estimator of μ. $\bar{y}$ will be used as a baseline to examine the performances of ratio-estimation and post-stratification in the following discussion.

2.1 Ratio Estimation

The ratio estimator of the population mean is

$$\hat{\mu}_r = \frac{\sum_{i\in s} y_i}{\sum_{i\in s} x_i}\cdot\mu_x = \frac{\bar{y}}{\bar{x}}\cdot\mu_x = \hat{\beta}\mu_x \qquad (1)$$

where $\bar{y}$ and $\bar{x}$ are the sample means of y and x, respectively, and $\hat{\beta}$, referred as the sample ratio, is an estimator of the population ratio

$$\beta = \frac{\mu_y}{\mu_x}.$$

$\hat{\mu}_r$ is a design-biased estimator, the bias is

$$\mathrm{E}(\hat{\mu}_r - \mu_y) \le \mathrm{CV}(\bar{x})\sqrt{\mathrm{var}(\hat{\mu}_r)} = \mathrm{CV}(x)\sqrt{\frac{\mathrm{var}(\hat{\mu}_r)}{n}} \qquad (2)$$

It is clear that $\hat{\mu}_r$ is asymptotically unbiased. Furthermore, the bias decreases when $\mathrm{CV}(x)$, the coefficient of variance of x, is smaller. In addition, $\hat{\mu}_r$ is known to be able to provide more precise estimate than $\bar{y}$ when certain correlation between y and x exists. The MSE of $\hat{\mu}_r$ is

$$\begin{aligned}\mathrm{MSE}(\hat{\mu}_r) &= \mathrm{var}(\hat{\mu}_r) + [\mathrm{E}(\hat{\mu}_r) - \mu_y]^2 \\ &\leq \left[1 + \frac{\mathrm{CV}^2(x)}{n}\right] \mathrm{var}(\hat{\mu}_r) \\ &\approx \frac{N-n}{Nn\,(N-1)} SS_{tot} \left(\beta^2 \frac{\sigma_x^2}{\sigma_y^2} - 2\beta\rho\frac{\sigma_x}{\sigma_y} + 1\right)\end{aligned} \tag{3}$$

where

$$SS_{tot} = \sum_{i=1}^{N} (y_i - \mu)^2$$

is the total sum of squares of y, and ρ is the finite population correlation coefficient between y and x.

Let

$$\beta\frac{\sigma_x}{\sigma_y} = M\rho, \tag{4}$$

then Equation 3 can be rewritten as

$$\mathrm{MSE}(\hat{\mu}_r) \approx \frac{N-n}{Nn\,(N-1)} SS_{tot} \left[1 + \left(M^2 - 2M\right)\rho_r^2\right] \tag{5}$$

Notice that

$$\beta\frac{\sigma_x}{\sigma_y} = \frac{\sigma_x/\mu_x}{\sigma_y/\mu_y} = \frac{\mathrm{CV}(x)}{\mathrm{CV}(y)}$$

is in fact the ratio of the coefficient of variances of y and x.

According to Equation 5, we conclude the followings with n is large enough

1. When $0 < M < 2$, we have $\left(M^2 - 2M\right) < 0$ and $\mathrm{MSE}(\hat{\mu}_r)$ decreases as ρ increases.
2. The minimum of $\mathrm{MSE}(\hat{\mu}_r)$ happens when $M = 1$, and accordingly we have

   $$\rho = \frac{\mathrm{CV}(x)}{\mathrm{CV}(y)}$$

   The minimum value of $\mathrm{MSE}(\hat{\mu}_r)$ is

   $$\mathrm{MSE}_{M=1}(\hat{\mu}_r) \approx \frac{N-n}{Nn\,(N-1)} SS_{tot}(1-\rho_r)^2 \tag{6}$$

   Equation 6 also indicates that $\mathrm{MSE}(\hat{\mu}_r)$ decreases as ρ increases when $M = 1$.

2.2 Post-stratification

As the usual post-stratification situation, a sample is selected by SRSWOR and stratified into several sub-samples according to some criterion, usually the associated values of certain auxiliary variable. Suppose that the population can be divided into H strata accordingly, labeled by $h = 1, \ldots, H$. In each stratum h there are N_h units, labeled by $u_h = \{h_1, \ldots, h_{N_h}\}$, such that $\sum_{h=1}^{H} N_h = N$. y_{hi} is the value of population variable associated with the i_{th} units in the h_{th} strata. The within-stratum population mean, denoted as μ_h, is

$$\mu_h = \frac{\sum_{i=1}^{N_h} y_{hi}}{N_h}.$$

The overall population mean is a weighted average of μ_h

$$\mu = \frac{1}{N}\sum_{h=1}^{H}\sum_{i=1}^{N_h} y_{hi} = \sum_{h=1}^{H} \frac{N_h}{N} \cdot \mu_h.$$

The within-stratum sample size is denoted as n_h and

$$n = \sum_{h=1}^{H} n_h.$$

Furthermore, denote the within-stratum sample as

$$s_h = \{hi_1, \ldots, hi_{n_h}\}, hi_j \in \{1, \ldots, N_h\}$$

The post-stratification estimation of μ is

$$\hat{\mu}_{post} = \sum_{h=1}^{H}\sum_{i\in s_h} \frac{N_h}{N}\frac{y_{hi}}{n_h} = \sum_{h=1}^{H} \frac{N_h}{N}\bar{y_h} \tag{7}$$

where $\bar{y}_h = \sum_{i\in s_h} y_{hi}/n_h$ is the within-stratum sample mean. It is simply the weighted average of $\bar{y}_h$ weighted by the proportion of the h_{th} stratum with respect to the whole population. $\hat{\mu}_{post}$ is a design-unbiased estimator for μ, and its average performance is similar to what in the usual stratified random sampling with proportional allocation. Hence, it is expected to be better than $\bar{y}$ as long as the population is properly stratified.

Since $\hat{\mu}_{post}$ is a design-unbiased estimator (cf. Thompson 2002) of μ, therefore $\mathrm{MSE}(\hat{\mu}_{post})$ is equal to $\mathrm{Var}(\hat{\mu}_{post})$. Additionally, since under post-stratification the within-stratum sample sizes n_h's are random variables which are jointly distributed as a multivariate hypergeometric distribution, $\mathrm{Var}(\hat{\mu}_{post})$ can be derived by the expected value of its conditional variance together with Taylor's expansion. Also we assume N is large enough such

that $N \approx N-1$:

$$\begin{aligned}
&\mathrm{MSE}(\hat{\mu}_{post}) = \mathrm{E}[\mathrm{var}(\hat{\mu}_{st}|n_1,\ldots,n_L)] \\
&= \mathrm{E}\left[\sum_{h=1}^{H}\left(\frac{N_h}{N}\right)^2\left(\frac{N_h-n_h}{N_h}\right)\frac{\sigma_h^2}{n_h}\right] \\
&\approx \frac{N-n}{Nn(N-1)}\sum_{h=1}^{H}\sum_{i=1}^{N_h}\left[\frac{N-N_h+nN_h}{n\left(N_h-1\right)}\right](y_{hi}-\mu_h)^2
\end{aligned} \tag{8}$$

Let

$$K = \frac{N-N_h+nN_h}{n\left(N_h-1\right)}$$

and assume that $N_h \approx N_{h'}$ so that $N \approx HN_h$, we have

$$K_{N_h \approx N_h'} \approx 1 + \frac{H-1}{n} \tag{9}$$

Suppose that N is large enough, and the size of each stratum are approximately equal, we can conclude

1. Since $0 < \frac{H-1}{n} < 1$, therefore $1 < K < 2$, and we have an approximate MSE

$$\mathrm{MSE}_{N_h \approx N_h'}\left(\hat{\mu}_{post}\right) = K\left[\frac{N-n}{Nn\left(N-1\right)}\right]SS_w \tag{10}$$

   where

$$SS_w = \sum_{h=1}^{H}\sum_{i=1}^{N_h}\left(y_{hi}-\mu_h\right)^2$$

   is the within-stratum sum of squares.

2. Since $\frac{H-1}{n}$ decreases as n increases, and accordingly $K \approx 1$ when n is large enough, therefore $\mathrm{MSE}(\hat{\mu}_{post})$ with large n can then be approximated as

$$\mathrm{MSE}_{N_h \approx N_h', K=1}\left(\hat{\mu}_{post}\right) = \frac{N-n}{Nn\left(N-1\right)}SS_w \tag{11}$$

Equation 10 and 11 indicate that $\mathrm{MSE}(\hat{\mu}_{post})$ is proportional to SS_w. That is, one should stratify the population in a way that the units within the same stratum should be as similar as possible for smaller SS_w, which accords to the principal of stratification.

3. Comparison between Ratio Estimation and Post-stratification

For comparing the performances between $\hat{\mu}_r$ and $\hat{\mu}_{post}$, one can evaluate the difference between $\mathrm{MSE}(\hat{\mu}_r)$ and $\mathrm{MSE}(\hat{\mu}_{post})$, and smaller MSE indicates better performance. For the purpose to simplify the discussion, we assume that n is large enough as well as $N_h \approx N_{h'}$ and then compare Equation 3 to Equation 11. The difference is

$$\begin{aligned}
&\mathrm{MSE}(\hat{\mu}_r) - \mathrm{MSE}\left(\hat{\mu}_{post}\right) \\
&= \frac{N-n}{Nn(N-1)}\left[SS_{tot}\left(\beta^2\frac{\sigma_x^2}{\sigma_y^2} - 2\beta\rho_r\frac{\sigma_x}{\sigma_y} + 1\right) - SS_w\right] \\
&= \frac{N-n}{Nn(N-1)}SS_{tot}\left[\left(\beta\frac{\sigma_x}{\sigma_y} - \rho_r\right)^2 + \left(\rho_{post}^2 - \rho_r^2\right)\right]
\end{aligned} \tag{12}$$

where $\rho^2 = 1 - \frac{SS_{res}}{SS_{tot}}$, which is in fact the coefficient of determination of a simple linear regression model without intercept, $y = \beta x$, and $\rho_{post}^2 = 1 - \frac{SS_w}{SS_{tot}}$, the proportion of between-stratum sum of squares out of the total sum of squares.

Recall that $\beta\frac{\sigma_x}{\sigma_y} = M\rho$, hence we conclude the followings based on Equation 12:

1. When $\rho_{post}^2 > \rho_r^2(2\mathrm{M} - \mathrm{M}^2)$, then we have $\mathrm{MSE}(\hat{\mu}_r) > \mathrm{MSE}(\hat{\mu}_{post})$. Consequently post-stratification is better than ratio estimation, and vice versa.

2. When $M = 1$, under which ratio estimation is the most advantageous over $\bar{y}$, then $\hat{\mu}_{post}$ is better when $\rho_{post}^2 > \rho_r^2$, and vise versa.

4. Simulation Study

Relative Efficiency (R.E.) is often used to compare the performances of two estimators. The definition of R.E. of estimator $\hat{t}_1$ to $\hat{t}_2$ (under the sample sampling design) is

$$\mathrm{R.E.} = \frac{\mathrm{MSE}(\hat{t}_1)}{\mathrm{MSE}(\hat{t}_2)},$$

hence $\hat{t}_2$ is better when $R.E. > 1$. Nevertheless, the closed form of the R.E. of $\hat{\mu}_r$ to $\hat{\mu}_{post}$ is not currently available, hence a simulation study is conducted in order to evaluate R.E. of $\hat{\mu}_r$ to $\hat{\mu}_{post}$ empirically under different conditions, such as different ρ, H, and/or n. For each condition, a pseudo population with size $N = 1000$ was generated, and 1000 random sample were selected by SRSWOR design for each population. For each sample selected, we calculated $\hat{\mu}_r$ and $\hat{\mu}_{post}$ and the empirical MSE is defined as the average of the squared errors of the 1000 random sample. Furthermore, the Empirical Relative Efficiency (E.R.E) is defined as the ratio of the empirical MSE's of $\hat{\mu}_r$ and $\hat{\mu}_{post}$.

The simulation process can be summarized as

1. Population size $N = 1000$.

2. Generate $\mathbf{x} = (x_1, \ldots, x_N)$ as the fixed values of auxiliary variable.

3. Generate the fixed values of population variable $\mathbf{y} = (y_1, \ldots, y_N)$ based on a bivariate normal population model with a given correlation coefficient ρ between x and y.

4. Select a random sample of (x, y) with sample size n.

   - Calculate $\hat{\mu}_r$.
   - Stratified the sample into H strata based on a given condition of x, and calculate $\hat{\mu}_{post}$.

5. For each case of correlation coefficient or H, 1000 different random sample were selected to calculate the empirical MSE for $\hat{\mu}_r$ and $\hat{\mu}_{post}$, denoted as $\widehat{\text{MSE}}(\hat{\mu}_r)$ and $\widehat{\text{MSE}}(\hat{\mu}_{post})$, respectively.

6. Calculate the E.R.E. of $\hat{\mu}_r$ to $\hat{\mu}_{post}$

$$\text{E.R.E.} = \frac{\widehat{\text{MSE}}(\hat{\mu}_r)}{\widehat{\text{MSE}}(\hat{\mu}_{post})}$$

In this simulation, we would like to study the impacts of the population correlation coefficient ρ, number of strata H, and sample size n on E.R.E.. First we fixed the sample size as $n = 100$, and then simulate E.R.E. under different H and ρ. The results is summarized in Figure 1,

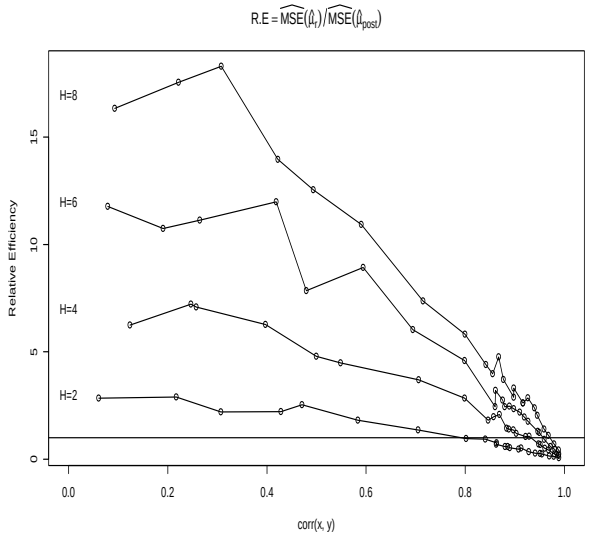

Figure 1: E.R.E. of $\hat{\mu}_r$ to $\hat{\mu}_{post}$ under different H and ρ with $n = 100$.

From Figure 1, it can be seen that the performance of $\hat{\mu}_{post}$ is superior to $\hat{\mu}_r$ most of the time. E.R.E. of $\hat{\mu}_r$ to $\hat{\mu}_{post}$ increases as H increases, that is, the performance of $\hat{\mu}_{post}$ is more preferable as the number of strata increases. On the other hand, E.R.E. decreases as ρ increases, which is expectable since the performance of $\hat{\mu}_r$ highly depends on ρ. However, $\hat{\mu}_r$ is better only when ρ is greater than 0.8 when $H = 2$, and ρ has to be at least 0.95 for $\hat{\mu}_r$ performs better than $\hat{\mu}_{post}$ when $H \geq 4$. The E.R.E. can be as high as more than 15 when the population correlation coefficient is low and the number of strata is 8. That is, $\widehat{\text{MSE}}(\hat{\mu}_{post})$ can be as low as less than 6% of $\widehat{\text{MSE}}(\hat{\mu}_{post})$. Even with moderate values of $\rho \doteq .5$ and $H = 4$, $\widehat{\text{MSE}}(\hat{\mu}_{post})$ sill is as low as about 20% of $\widehat{\text{MSE}}(\hat{\mu}_{post})$.

Another simulation study was also conducted to examine the impact of sample size n as well as ρ on E.R.E.. The number of strata is chosen to be 4, which is not a large number of strata when $N = 1000$, so that the simulation would be fair to both methods. The results are summarized in Figure 2,

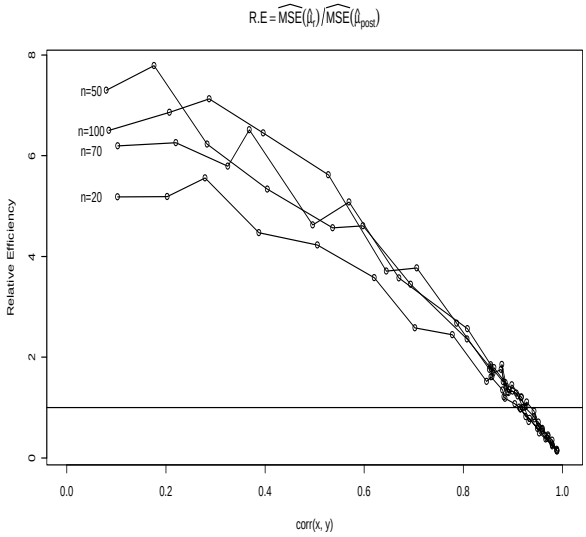

Figure 2: E.R.E. of $\hat{\mu}_r$ to $\hat{\mu}_{post}$ under different n and ρ with $H = 4$.

It is clear that $\hat{\mu}_{post}$ is superior to $\hat{\mu}_r$ as what appeared in Figure 1. Additionally, the sample size does not seem to be as decisive as ρ and H. With a smaller sample size $n = 20$, the E.R.E. is less than the other cases when $n = 50$, 70, and 100. One of the reason is that some of n_h has better chance to be zero with smaller sample size.

For model-based perspective, simulation study shows the similar results as Figure 1 and 2. It is not surprising since in the model-based approach,

the simulation results can be considered as an average of different populations that generated by the same population stochastic model.

5. Final Comments

With a appropriate auxiliary variable in a sampling survey research, both post-stratification and ratio estimator can provide significant improvement on the inference of the population quantity of interest. They also help the investigators to better understand other related phenomena other than the population quantity of primary interest. For example, ratio estimator also estimates the ratio between the population variable of interest and the auxiliary, and it is often considered as an helpful information in many sampling survey researches. Post-stratification is able to provide information in certain sub-population in addition to the whole population. Hence, both have been widely used in practice by field researchers.

In this article, we did not consider the different extra information provided by post-stratification or ratio estimation but only the population quantity of primary interest in order to compare their performances. IN the simulation study, the populations were generated by a linear model which can be considered as the preferable model for ratio estimator. On the other hand, the post-stratification is done according to the order of x, and such stratification follows the principal of stratification according to the linear relationship between x and y. Therefore, the simulation is in general fair for both methods.

The simulation results indicates that post-stratification perform considerably better that ratio estimator. In addition to that, ratio estimation require more population information such as the population mean of the auxiliary variable. More sampling effort/cost is necessary in order to measure the x-value of the sampled units. On the other hand, one would be able to use post-stratification would be able to use as long as the information from x is enough to categorize the sampled units into different strata. Nevertheless, The stratum sizes N_h's are necessary for post-stratification and this might be its main disadvantage. In any of it, post-stratification is recommended since it is worse than ratio estimator only when y and x are highly correlated, but it is doubtful if it is reasonable to expect such population in practice. Also, if the relation between y and x is not a ratio model as expected, then ratio estimator might provide poor results, but post-stratification would still give fair estimator. If one could have all the population information such as μ_x/τ_x, N_h, and sampled x-values, a post-stratification together with the within-stratum ratio estimation should be the best alternative.

6. References

Godambe VP. 1955. A unified theory of sampling from finite population. *Journal of the Royal Statistical Society* **B17**:269-278

Lohr, S.L. (1999). *Sampling: Design and Analysis* Pacific Grove, CA: Duxbury Press.

Thompson, S.K.(2002). *Sampling.* 2nd ed. New York: Wiley.