Improved Ratio Type Estimators using some Prior Information in Sample Surveys: A Case Study of Fine Particulate Matter in Thailand

NUANPAN LAWSON
Department of Applied Statistics, Faculty of Applied Science,
King Mongkut's University of Technology North Bangkok,
1518 Pracharat 1 Road, Wongsawang, Bangsue, Bangkok 10800,
THAILAND

Abstract: - Air pollution affects Thai people's health and social life nowadays as it exceeds the standards levels of both Thailand and the World Health Organization. Estimating air pollution data can benefit understanding and determining policies to help deal with this issue. Prior knowledge from past surveys or censuses could be useful for increasing the effect of the estimation. Improved ratio estimators utilizing prior knowledge in simple random sampling without replacement have been advocated. The property of the mean square error of the proposed class of estimators is obtained. We applied the proposed estimators to the fine particulate matter data in Dindang in 2019. The results from the air pollution data illustrate the improved ratio type estimators work better with respect to the existing estimator using some prior information. Existing knowledge of the quartile average and the median of the auxiliary variable gives rise to the best estimators with the lowest mean square errors for estimating fine particulate matter. Nevertheless, the proposed estimators are useful for small sampling fractions which can help in financial and time-consuming.

Key-Words: - ratio estimators, prior information, auxiliary variable, fine particulate matter, air pollution, sample surveys

Received: August 24, 2022. Revised: April 14, 2023. Accepted: May 2, 2023. Published: May 29, 2023.

1 Introduction

A sample survey is an essential aspect of statistics for the inference of the population based on the sample. In general, inferential statistics concerns the population parameters e.g. the mean, total, and proportion. Each value collected from the population concerns the interested parameter. Sometimes some prior information is available from previous sample surveys or conducting a small survey and can be supportive for the estimation of parameters in sample surveys. A sample mean ($\bar{y}$) of a study variable Y is employed to estimate the population mean $\bar{Y}$ based on the sample. Utilizing known prior information was shown by [1], who introduced utilizing the available coefficient of variation in the population of Y (C_y) in simple random sampling without replacement (SRSWOR). The, [1], estimator is

$$\hat{\bar{Y}}_S = K_S^* \bar{y}, \quad (1)$$

where $K_S^* = \left(1+\gamma C_y^2\right)^{-1}$, $\gamma = \frac{\left(1-n/N\right)}{n}$, n and N are the sample and population sizes. The mean square error (MSE) of $\hat{\bar{Y}}$ is

$$MSE\left(\hat{\bar{Y}}_S\right) = \left(\frac{1-f}{n}\right)\bar{Y}^2 \frac{C_y^2}{1+\gamma C_y^2}. \quad (2)$$

However, the positive correlation between the auxiliary variable X and Y (ρ) can be used in estimating a population mean. An example of the popular estimators using this is the ratio estimators proposed by [2]. The ratio estimator is

$$\hat{\bar{Y}}_R = \frac{\bar{y}}{\bar{x}}\bar{X}, \quad (3)$$

where $\bar{x}$ is a sample mean of X.

The MSE of $\hat{\bar{Y}}_R$ is

$$MSE\left(\hat{\bar{Y}}_R\right) = \left(\frac{1-f}{n}\right)\bar{Y}^2\left[C_y^2 + C_x^2 - 2\rho C_x C_y\right], \quad (4)$$

where C_x is the coefficient of variation of X.

In [3], the author introduced the estimators inspired by [1], employing a known C_y. Prasad's estimators are

$$\hat{\bar{Y}}_{P_1} = K_S^* \frac{\bar{y}}{\bar{x}}\bar{X}, \quad (5)$$

$$\hat{\bar{Y}}_{P_2} = K_P^* \frac{\bar{y}}{\bar{x}} \bar{X}, \quad (6)$$

where $K_P^* = \frac{1+\gamma\rho C_x C_y}{1+\gamma C_y^2} = K_S^* \left(1+\gamma\rho C_x C_y\right)$, $\bar{X}$ is the population mean of X.

The MSE of $\hat{\bar{Y}}_{P_1}$ and $\hat{\bar{Y}}_{P_2}$ are

$$MSE\left(\hat{\bar{Y}}_{P_1}\right) = \left(\frac{1-f}{n}\right)\bar{Y}^2\left[\frac{1-2\rho C_x / C_y}{1+\gamma C_y^2} C_y^2 + C_x^2\right], \quad (7)$$

$$MSE\left(\hat{\bar{Y}}_{P_2}\right) = \left(\frac{1-f}{n}\right)\bar{Y}^2\left[\frac{1-\rho^2\gamma C_x^2}{1+\gamma C_y^2} C_y^2 + C_x^2 - \frac{2}{1+\gamma C_y^2}\rho C_x C_y\right]. \quad (8)$$

The estimator $\hat{\bar{Y}}_{P_2}$ is more efficient than $\hat{\bar{Y}}_{P_1}$. Many researchers also suggested to include some known parameters in their studies to make more efficient estimators [4], [5], [6], [7], [8].

Air pollution issues have become detrimental to the point where they affect Thai people's health. It is an increasing concern due to the amount of fine particulate matter with a diameter smaller than 2.5 microns (PM 2.5), that exceeds standards both in Bangkok and the northern region which have exceeded the standard value in Thailand and also the world's standard. According to the World Health Organization's criteria, the 24-hour average should not be more than 25 micrograms per cubic meter while the average of Thailand, the 24-hour average is less than 50 micrograms per cubic meter. The continuous levels of dust affect health in the long term causing chronic respiratory problems, lung cancer, and chronic cardiovascular disease. In [9], the authors studied how to estimate PM2.5 in Bangkok, Thailand through the density of ozone when the data are missing using the new estimator under SRSWOR. In, [10], the authors also estimated PM2.5 in Bangkok, Thailand using carbon monoxide by utilizing the number of respondents, sample size, and constant for estimating PM2.5. In, [11], the author suggested utilizing the transformation of an auxiliary variable to estimate carbon monoxide by PM2.5 in Nan, Thailand. In, [12], the author studied how to estimate PM2.5 through nitrogen dioxide in Chiang Rai using the transformed combined estimators under double sampling.

Utilizing known prior information, a class of ratio type estimators under SRSWOR is suggested. The MSE of the new estimators is obtained using the Taylor series. The pollution data from Dindang in 2019 is applied in the study using the MSE as a criterion.

2 Proposed Estimator

Using the idea of [3], a class of ratio type estimators utilizing the known population coefficient variation Y has been proposed. The proposed estimator is

$$\hat{\bar{Y}}_{NP} = K_P^* \bar{y}\left(\frac{A\bar{X}+D}{A\bar{x}+D}\right), \quad (9)$$

where $K_P^* = \frac{1+\gamma\rho C_x C_y}{1+\gamma C_y^2} = K_S^*\left(1+\gamma\rho C_x C_y\right)$, A and D are the available information for instance the coefficient of variation of X (C_x), the coefficient of the kurtosis of X $\left(\beta_{2(x)}\right)$, the coefficient of the skewness of X $\left(\beta_{2(x)}\right)$, the correlation coefficient between X and Y (ρ), the inter-quartile range of X (Q_r), the semi-quartile range of X (Q_d), the quartile average of X (Q_a), median of X (M) or others.

The Taylor series approach is used to study the MSE of the estimator to get Theorem 1.

Theorem 1. The approximated MSE of the proposed estimator $\hat{\bar{Y}}_{NP}$ in equation (9) for population mean $\bar{Y}$ is

$$MSE\left(\hat{\bar{Y}}_{NP}\right) = \left(\frac{1-f}{n}\right)\bar{Y}^2\left[\frac{1-W^2\rho^2\gamma C_x^2}{1+\gamma C_y^2} C_y^2 + W^2 C_x^2 - \frac{2}{1+\gamma C_y^2} W\rho C_x C_y\right],$$

where $W = \frac{A\bar{X}}{A\bar{X}+D}$, $\gamma = \frac{(1-n/N)}{n}$.

Proof of theorem 1

Let $e_0 = \left(\bar{y}-\bar{Y}\right)/\bar{Y}$ and $e_1 = \left(\bar{x}-\bar{X}\right)/\bar{X}$ then $E\left(e_0\right) = E\left(e_1\right) = 0$, $E\left(e_0^2\right) = \frac{1-f}{n}C_y^2$, $E\left(e_1^2\right) = \frac{1-f}{n}C_x^2$ and $E\left(e_0 e_1\right) = \frac{1-f}{n}C_{xy} = \frac{1-f}{n}\rho C_y C_x$. To find the MSE of $\hat{\bar{Y}}_{NP}$, we write equation (9) in terms of e's then the $E(\hat{\bar{Y}}_{NS} - \bar{Y})^2$ of the $\hat{\bar{Y}}_{NP}$ is shown as below.

The MSE of $\hat{\bar{Y}}_{NP}$ is

$$MSE\left(\hat{\bar{Y}}_{NP}\right) \approx E(\hat{\bar{Y}}_{NP} - \bar{Y})^2$$
$$= \left(\frac{1-f}{n}\right)\bar{Y}^2\left[\frac{1-W^2\rho^2\gamma C_x^2}{1+\gamma C_y^2}C_y^2 + W^2C_x^2 - \frac{2}{1+\gamma C_y^2}W\rho C_x C_y\right],$$

Some possible proposed estimators that we considered in this study $\hat{\bar{Y}}_{NP_i}$, i = 1, 2, ..., 10 with A and D are in Table 1.

Table 1. The estimators $\hat{\bar{Y}}_{NP_i}$, i = 1, 2, ..., 10 .

Estimator	A	D
$\hat{\bar{Y}}_{NP_1} = K_P^*\bar{y}\left(\frac{\bar{X}+C_x}{\bar{x}+C_x}\right)$	1	C_x
$\hat{\bar{Y}}_{NP_2} = K_P^*\bar{y}\left(\frac{\bar{X}+\beta_{2(x)}}{\bar{x}+\beta_{2(x)}}\right)$	1	$\beta_{2(x)}$
$\hat{\bar{Y}}_{NP_3} = K_P^*\bar{y}\left(\frac{\beta_{2(x)}\bar{X}+C_x}{\beta_{2(x)}\bar{x}+C_x}\right)$	$\beta_{2(x)}$	C_x
$\hat{\bar{Y}}_{NP_4} = K_P^*\bar{y}\left(\frac{C_x\bar{X}+\beta_{2(x)}}{C_x\bar{x}+\beta_{2(x)}}\right)$	C_x	$\beta_{2(x)}$
$\hat{\bar{Y}}_{NP_5} = K_P^*\bar{y}\left(\frac{\beta_{2(x)}\bar{X}+\beta_{2(x)}}{\beta_{2(x)}\bar{x}+\beta_{2(x)}}\right)$	$\beta_{1(x)}$	$\beta_{2(x)}$
$\hat{\bar{Y}}_{NP_6} = K_P^*\bar{y}\left(\frac{\bar{X}+\rho}{\bar{x}+\rho}\right)$	1	ρ
$\hat{\bar{Y}}_{NP_7} = K_P^*\bar{y}\left(\frac{\bar{X}+Q_r}{\bar{x}+Q_r}\right)$	1	Q_r
$\hat{\bar{Y}}_{NP_8} = K_P^*\bar{y}\left(\frac{\bar{X}+Q_d}{\bar{x}+Q_d}\right)$	1	Q_d
$\hat{\bar{Y}}_{NP_9} = K_P^*\bar{y}\left(\frac{\bar{X}+Q_a}{\bar{x}+Q_a}\right)$	1	Q_a
$\hat{\bar{Y}}_{NP_{10}} = K_P^*\bar{y}\left(\frac{\bar{X}+M}{\bar{x}+M}\right)$	1	M

3 Efficiency Comparison

The MSE of the estimator is considered to compare $\hat{\bar{Y}}_{NP}$ with the, [3], estimator, $\hat{\bar{Y}}_{P_2}$, because the estimator $\hat{\bar{Y}}_{P_2}$ is more efficient than $\hat{\bar{Y}}_{P_1}$.

The estimator $\hat{\bar{Y}}_{NP}$ is more efficient than $\hat{\bar{Y}}_{P_2}$ if

$$MSE(\hat{\bar{Y}}_{NP}) < MSE(\hat{\bar{Y}}_{P_2})$$
$$\left(\frac{1-f}{n}\right)\bar{Y}^2\left[\frac{1-W^2\rho^2\gamma C_x^2}{1+\gamma C_y^2}C_y^2 + W^2C_x^2 - \frac{2}{1+\gamma C_y^2}W\rho C_x C_y\right] <$$
$$\left(\frac{1-f}{n}\right)\bar{Y}^2\left[\frac{1-\rho^2\gamma C_x^2}{1+\gamma C_y^2}C_y^2 + C_x^2 - \frac{2}{1+\gamma C_y^2}\rho C_x C_y\right]$$
$$\frac{1-W^2\rho^2\gamma C_x^2}{1+\gamma C_y^2}C_y^2 + W^2C_x^2 - \frac{2}{1+\gamma C_y^2}W\rho C_x C_y <$$
$$\frac{1-\rho^2\gamma C_x^2}{1+\gamma C_y^2}C_y^2 + C_x^2 - \frac{2}{1+\gamma C_y^2}\rho C_x C_y$$
$$\rho < \frac{(W+1)C_x}{\gamma\left(K_S^*C_y - 2\right)}$$

4 An Empirical Study

The pollution data from Dindang, Thailand in January 2019, [13], are used to see how the proposed estimators work with the real world data over the existing estimators. PM2.5 (mpcm) is considered as Y and carbon monoxide CO (ppm) is considered as X in this study to see how the estimators perform. The details of the data are as follows:

$\bar{Y}$ = 51.78, C_y = 0.45, $\bar{X}$ = 1.91, C_x = 0.33, and ρ = 0.4. Samples, using SRSWOR, of sizes n = 15, 30, 60, 150, and 300 are taken from the population size N = 900 using the R program, [14]. Figure 1 shows the plot between PM2.5 and CO data. We can see that there is a positive relationship between PM2.5 and CO. The results are presented in Table 2 and Figure 2 respectively.

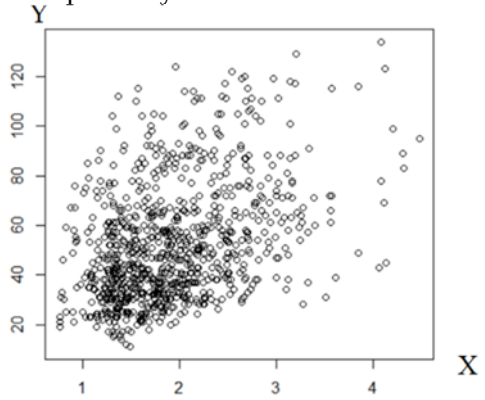

Fig. 1: A scatter plot between PM2.5 and CO in Dindang

Table 2. The MSE of the estimators for the pollution Data

Estimator	Sample size				
	15	30	60	150	300
$\bar{y}$	35.16	17.28	8.34	2.98	1.19
$\hat{\bar{Y}}_R$	33.40	16.42	7.93	2.83	1.13
$\hat{\bar{Y}}_S$	34.70	17.17	8.32	2.98	1.19
$\hat{\bar{Y}}_{P_1}$	33.22	16.37	7.91	2.83	1.13
$\hat{\bar{Y}}_{P_2}$	33.18	16.36	7.91	2.83	1.13
$\hat{\bar{Y}}_{NP_1}$	30.98	15.29	7.40	2.64	1.06
$\hat{\bar{Y}}_{NP_2}$	30.01	14.84	7.19	2.57	1.03
$\hat{\bar{Y}}_{NP_3}$	32.48	16.02	7.75	2.77	1.11
$\hat{\bar{Y}}_{NP_4}$	32.23	15.95	7.72	2.76	1.11
$\hat{\bar{Y}}_{NP_5}$	30.10	14.89	7.21	2.58	1.03
$\hat{\bar{Y}}_{NP_6}$	30.66	15.14	7.32	2.62	1.05
$\hat{\bar{Y}}_{NP_7}$	29.55	14.60	7.06	2.53	1.01
$\hat{\bar{Y}}_{NP_8}$	30.61	15.11	7.31	2.61	1.05
$\hat{\bar{Y}}_{NP_9}$	29.11	14.39	6.97	2.49	1.00
$\hat{\bar{Y}}_{NP_{10}}$	29.10	14.39	6.97	2.49	1.00

From Table 2, the results illustrated that the proposed estimators work well in this situation and performed superior to the existing estimators. A small sample size can lead to more errors compared to a big sample size in general. The proposed estimator $\hat{\bar{Y}}_{NP_9}$ utilizing the known quartile average of X and the proposed estimator $\hat{\bar{Y}}_{NP_{10}}$ utilizing the median of X performed the best among other proposed estimators. Using some available auxiliary information can assist in increasing the accuracy and gives fewer errors.

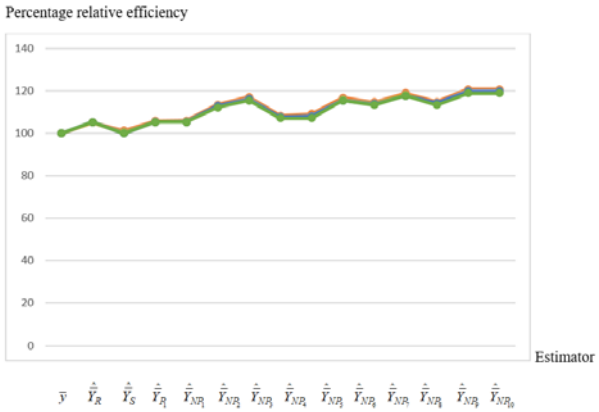

Fig. 2: Percentage relative efficiency of the estimators with respect to the sample mean estimator

Figure 2 showed that the percentage relative efficiency of the proposed estimators that perform the best concerning the sample mean was at least around twenty percent more efficient at all levels of sampling fractions. Increasing sample sizes do not affect the efficiency of the estimators although bigger sizes lead to a smaller MSE as shown in Table 2 but it gave the same rate of efficiency for the estimators. We can imply that using the proposed estimators to conduct a small survey can increase the performance of the estimators and reduce time consumption and save on budget.

5 Conclusion

A general form of estimator is introduced in this study based on prior knowledge of variables for estimating population mean. The MSE of the new estimator is expressed and showed the performance via an application to air pollution data in Dindang, Thailand. We can see that prior information helps to increase the performance of the estimators, yielding the least MSE. In this scenario, the best proposed estimators that gave the lowest MSE are the ones that use prior knowledge of the quartile average and the median of X for estimating fine particulate matter. Large sample sizes give more accurate results when compared to smaller sample sizes in terms of MSE. However, in terms of efficiency, we can see that the proposed estimators can result in the same efficiency at all levels of sampling fractions and therefore it benefits small surveys to study the variable of interest and can save time and financial costs. Utilizing some known prior information could benefit by reducing the MSE of the population mean estimator which ultimately results in greater efficiency. We can apply the proposed estimators using other available prior information and it can

also be useful in other survey designs. Nevertheless, the application to real data can be used in future studies by applying the new estimator so that the variable of interest can be estimated.

Acknowledgment:
Thank you to the referees for the helpful comments.

References:

[1] Searls, D.T., The Utilization of a Known Coefficient of Variation in the Estimation Procedure, *Journal of the American Statistical Association*, Vol. 59, 1964, pp. 1225-1226.

[2] Cochran, W.G., The Estimation of the Yields of the Cereal Experiments by Sampling for the Ratio of Grain to Total Produce, *The Journal of Agricultural Science*, Vol.30, No.2, 1940, pp. 262 – 275.

[3] Prasad, B., Some Improved Ratio Type Estimators of Population Mean and Ratio in Finite Population Sample Surveys. *Communication in Statistics – Theory and Methods*, Vol.18, No.1, 1989, pp. 379–392.

[4] Soponviwatkul, K. and Lawson, N., New Ratio Estimators for Estimating Population Mean in Simple Random Sampling Using a Coefficient of Variation, Correlation Coefficient and a Regression Coefficient. *Gazi University Journal of Science*, Vol. 30, No.4, 2017, pp. 610-621.

[5] Lawson, N., Ratio Estimators of Population Means Using Quartile Function of Auxiliary Variable Using Double Sampling, *Songklanakarin Journal of Science and Technology*, Vol.41, No.1, 2019, pp. 117-122.

[6] Lawson, N., An Alternative Family of Combined Estimators for Estimating Population Mean in Finite Populations. *Lobachevskii Journal of Mathematics*, Vol.42, No.13, 2021, pp. 3150-3157.

[7] Ponkaew, C. and Lawson, N., New Generalized Regression Estimators Using a Ratio Method and its Variance Estimation for Unequal Probability Sampling Without Replacement in the Presence of Nonresponse. *Current Applied Science and Technology*, Vol.23,No.2, 2023, pp. 1-27.

[8] Lawson, N., An Improved Class of Population Mean Estimators by Utilizing Some Prior Information in Simple Random Sampling Using Searl's approach, *Lobachevskii Journal of Mathematics*, Vol.43, No.11, 2022, pp. 3376–3383.

[9] Chodjuntug, K. and Lawson, N., Imputation for Estimating the Population Mean in the Presence of Nonresponse, With Application to Fine Particle Density in Bangkok, *Mathematical Population Studies*, Vol.29. No.4, 2022, pp. 204 – 22.

[10] Chodjuntug, K. and Lawson, N., The Chain Regression Exponential Type Imputation Method for Mean Estimation in the Presence of Missing Data, *Songklanakarin Journal of Science and Technology*, Vol.44, No.4, 2022, pp. 1109-1118.

[11] Thongsak, N. and Lawson, N., Bias and Mean Square Error Reduction by Changing the Shape of the Distribution of an Auxiliary Variable: Application to Air Pollution Data in Nan, Thailand, *Mathematical Population Studies*, 2022.

[12] Thongsak, N. and Lawson, N. Classes of Combined Population Mean Estimators Utilizing Transformed Variables Under Double Sampling: An Application to Air Pollution in Chiang Rai, Thailand, *Songklanakarin Journal of Science and Technology*, 44(5), 1390-1398.

[13] Pollution Control Department, Thailand's air quality and situation reports. Bangkok, Thailand, http://air4thai.pcd.go.th/webV2/history/, 2019.

[14] R Core Team, R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/, 2021.

Contribution of Individual Authors to the Creation of a Scientific Article (Ghostwriting Policy)

The author contributed in the present research, at all stages from the formulation of the problem to the final findings and solution.

Sources of Funding for Research Presented in a Scientific Article or Scientific Article Itself

This research was funded by the National Science, Research and Innovation Fund (NSRF), and King Mongkut's University of Technology North Bangkok Contract no. KMUTNB-FF-66-56.

Conflict of Interest

The author has no conflicts of interest to declare that are relevant to the content of this article.