https://doi.org/10.56093/JISAS.V79I2.8
Available online at http://isas.org.in/isa/jisas
JOURNAL OF THE INDIAN SOCIETY OF
AGRICULTURAL STATISTICS 79(2) 2025 127–139

On the Estimation of Fano's Dispersion Index using Auxiliary Information

Steven Chavez, Stephen A. Sedory and Sarjinder Singh

Department of Mathematics, Texas A&M University-Kingsville
Kingsville, TX 78363, USA

Received 10 July 2025; Revised 24 November; Accepted 27 November 2025

SUMMARY

This paper addresses the problem of estimating the Fano's dispersion index (also known as the Fano factor) in the presence of auxiliary information. The dispersion index, defined as the ratio of the variance to the mean, is a key measure in quantifying over dispersion and arises in numerous applied settings, including count data modeling in biology, epidemiology, and engineering. We begin by introducing the necessary notation and establishing the theoretical framework, including the distributional assumptions and expected values underpinning the estimators. The core of the paper involves a comparative study of three estimators: (i) the naïve estimator, which does not incorporate auxiliary information; (ii) a ratio-type estimator, and (iii) a regression-type estimator, which utilize the known auxiliary variable. For each estimator, we derive expressions for the expected value, bias, and mean squared error (MSE), facilitating a theoretical comparison of their performance. The analysis demonstrates the potential gains in efficiency when auxiliary information is properly utilized. A comprehensive simulation study, carried out in SAS, evaluates the finite-sample properties of the estimators across a range of scenarios. We conclude with an application to a real-world data set, highlighting the practical implications of the proposed methods. The performance of the estimators is assessed via relative bias and relative efficiency, offering guidance for their use in applied statistical practice.

Keywords: Estimation of dispersion; Relative bias; Relative eficiency; Ratio and regression Type estimators; Fano factor; Simulation; Applications.

Data Statement: The data sets used are available from the cited references or are reproducible through codes used in the simulation.

Conflict of Interest: There is no conflict of interest.

1. INTRODUCTION

The dispersion index referred to as variance mean ratio (VMR), it can measure whether a set of data is dispersed, over dispersed and under dispersed. It was introduced by the Italian American physicist Fano (1947) and called Fano factor. The importance of the Fano factor is well known in various fields of studies such as counting processes, theoretical physics etc. In addition, some distributions, such as binomial distribution have a dispersion index between 0 and 1, and are considered under-dispersed. The geometric and negative binomial distributions have dispersion index greater than 1 and are considered over-dispersed, while the Poisson distribution has a dispersion index equal to 1, so is neither over-dispersed nor under-dispersed. The coefficient of variation and the dispersion index are similar with the coefficient of variation being the standard deviation to the mean ratio; the coefficient of variation is dimensionless while the dispersion index has dimension. Also, the dispersion index is likely to have a chi-squared distribution unlike coefficient of variation. For detail one could refer to De Oliveira, de Oliveira Peres, Martinez and Achcar (2019).

Consider a sample, s_n, of n units selected by using simple random without replacement sampling (SRSWOR) from a finite population Ω of N units. Let

Corresponding author: Sarjinder Singh
E-mail address: sarjinder@yahoo.com

(y_i, x_i), $i = 1, 2, \ldots, N$ be the values of the study variable, *y*, and the auxiliary variable, *x*, respectively in the population. Let (y_i, x_i), *i*= 1, 2,…,n be the values of the study variable, y_i, and the auxiliary variable, x_i, respectively, in the sample.

Let

$\bar{y}_n = \frac{1}{n}\sum_{i=1}^{n} y_i$; be the sample mean of the study variable, y,

$\bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$; be the sample meanof the auxiliary variable, x,

$\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} y_i$; be the population mean of the study variable, y,

$\bar{X} = \frac{1}{N}\sum_{i=1}^{N} x_i$; be the population mean of the auxiliary variable, x,

The ratio estimator for estimating the population mean $\bar{Y}$ was first introduced by Cochran (1940) and is defined as

$$\bar{y}_R = \bar{y}\left(\frac{\bar{X}}{\bar{x}}\right) \tag{1.1}$$

The ratio estimator is found to be more efficient than the sample mean estimator $\bar{y}$ for a value of correlation coefficient 0.5 to 1.0 between a study variable *y* and an auxiliary variable *x*.

Let

$s_x^2 = \frac{1}{(n-1)}\sum_{i=1}^{n}\left(x_i - \bar{x}_n\right)^2$; be the sample variance of the auxiliary variable, *x*,

and

$s_{xy} = \frac{1}{(n-1)}\sum_{i=1}^{n}\left(y_i - \bar{y}\right)\left(x_i - \bar{x}\right)$; be the sample covariance of the study variable, *y*, and the auxiliary variable *x*.

The regression estimator for estimating the population mean $\bar{Y}$ was proposed by Hansen, Hurwitz, and Madow (1953) and is given by

$$\bar{y}_{LR} = \bar{y} + \hat{\beta}\left(\bar{X} - \bar{x}\right) \tag{1.2}$$

where $\hat{\beta} = s_{xy} / s_x^2$ is the sample regression coefficient. The regression estimator is always more efficient than the sample mean estimator $\bar{y}$ for any non-zero value of correlation coefficient between the study variable and the auxiliary variable.

Further let

$s_y^2 = \frac{1}{(n-1)}\sum_{i=1}^{n}\left(y_i - \bar{y}_n\right)^2$; be the sample variance of the study variable, y,

$S_y^2 = \frac{1}{(N-1)}\sum_{i=1}^{N}\left(y_i - \bar{Y}\right)^2$; be the population mean squared error of the study variable, y,

and

$S_x^2 = \frac{1}{(N-1)}\sum_{i=1}^{N}\left(x_i - \bar{X}\right)^2$; be the population mean squared error of the auxiliary variable, x,

Let us define

$\epsilon_0 = \frac{\bar{y}_n}{\bar{Y}} - 1$, $\epsilon_1 = \frac{\bar{x}_n}{\bar{X}} - 1$, $\epsilon_2 = \frac{s_y^2}{S_y^2} - 1$, and $\epsilon_3 = \frac{s_x^2}{S_x^2} - 1$

such that

$$E(\epsilon_0) = E(\epsilon_1) = E(\epsilon_2) = E(\epsilon_3) = 0$$

$$E(\epsilon_0^2) = \left(\frac{1-f}{n}\right)C_y^2,\ E(\epsilon_1^2) = \left(\frac{1-f}{n}\right)C_x^2,$$

$$E(\epsilon_0\epsilon_1) = \left(\frac{1-f}{n}\right)\rho_{xy}C_xC_y,\ E(\epsilon_2^2) = \left(\frac{1-f}{n}\right)(\lambda_{40} - 1),$$

$$E(\epsilon_3^2) = \left(\frac{1-f}{n}\right)(\lambda_{04} - 1), E(\epsilon_0\epsilon_2) = \left(\frac{1-f}{n}\right)C_y\lambda_{30},$$

$$E(\epsilon_0\epsilon_3) = \left(\frac{1-f}{n}\right)C_y\lambda_{12},\ E(\epsilon_1\epsilon_2) = \left(\frac{1-f}{n}\right)C_x\lambda_{21},$$

$E(\epsilon_1\epsilon_3)=\left(\frac{1-f}{n}\right)C_x\lambda_{03}$, and $E(\epsilon_2\epsilon_3)=\left(\frac{1-f}{n}\right)(\lambda_{22}-1)$

where

$$\lambda_{rs}=\frac{\mu_{rs}}{\mu_{20}^{r/2}\mu_{02}^{s/2}}.\ \mu_{rs}=\frac{1}{N-1}\sum_{i=1}^{N}\left(y_i-\bar{Y}\right)^r\left(x_i-\bar{X}\right)^s$$

$C_y^2=\frac{S_y^2}{\bar{Y}^2}$, $C_x^2=\frac{S_x^2}{\bar{X}^2}$, $u_{20}=S_y^2$, $u_{02}=S_x^2$, $u_{11}=S_{xy}$,

$\rho_{xy}=\frac{S_{xy}}{S_xS_y}$ and $f=\frac{n}{N}$ is called the finite population correction factor (f. p. c.).

Let

$\hat{\lambda}_{rs}=\frac{\hat{\mu}_{rS}}{\hat{\mu}_{20}^{r/2}\hat{\mu}_{02}^{s/2}}$ be a consistent estimator of λ_{rs}

where

$$\hat{\mu}_{rs}=\frac{1}{(n-1)}\sum_{i=1}^{n}\left(y_i-\bar{y}_n\right)^r\left(x_i-\bar{x}_n\right)^s$$

so that $s_y^2=\hat{\mu}_{20}$, $s_x^2=\hat{\mu}_{02}$ and $s_{xy}=\hat{\mu}_{11}$, have their usual meanings.

Let $\hat{c}_y=\frac{s_y}{\bar{y}}$ and $\hat{c}_x=\frac{s_x}{\bar{x}}$ be the usual consistent estimators for the coefficient of variation of y and x respectively, given by $C_y=\frac{S_y}{\bar{Y}}$ and $C_x=\frac{S_x}{\bar{X}}$.

Note that all approximate results stated here are derived from the asymptotic expansions in probability rule due to Wu (1982). Let $(x_1,y_1),(x_2,y_2),...,(x_n,y_n)$ be i.i.d to a distribution with means $(\bar{X},\bar{Y})$. Let $\bar{x}$ and $\bar{y}$ are the sample means of the auxiliary variable and the study variable respectively, then

$$E\left(\bar{x}-\bar{X}\right)^r\left(\bar{y}-\bar{Y}\right)^s=\begin{cases}O\left(n^{-\frac{1}{2}(r+s)}\right) & (r+s\quad\text{even})\\ O\left(n^{-\frac{1}{2}(r+s+1)}\right) & (r+s\quad\text{odd})\end{cases}\tag{1.3}$$

and

$$\left(\bar{x}-\bar{X}\right)^r\left(\bar{y}-\bar{Y}\right)^s=O_p\left(n^{-\frac{1}{2}(r+s)}\right)\tag{1.4}$$

where r and s are non-negative integers. Here O_p represents big-O notation for a sequential random variable while O represent the big-O notation for sequence of constants, that is, in our case a sequence of expected values. One could also refer to David and Srivastava (1974), Sukhatme (1944), Sukhatme and Sukhatme (1970), Srivastava and Jhajj (1981) and Tracy (1984) for such asymptotic approximations. In the next section, we consider a naïve estimator of the Fano's Dispersion Index.

2. NAIVE ESTIMATOR OF THE FANO'S DISPERSION INDEX

Following De Oliveira *et al.* (2019), the Fano's Dispersion Index of the study variable, y, is defined as

$$DI_y=\frac{S_y^2}{\bar{Y}}\tag{2.1}$$

We consider a naive estimator of the Fano's Dispersion Index of the study variable y, as

$$\widehat{DI}_y=\frac{s_y^2}{\bar{y}_n}\tag{2.2}$$

Now, we have the following theorems; the proofs of the theorems are given in online documented Appendix-A.

Theorem 2.1. The bias in the naive estimator, $\widehat{DI}_y$, to the first order of approximation, is given by

$$B\left(\widehat{DI}_y\right)=\frac{(1-f)}{n}\left(DI_y\right)\left[C_y^2-C_y\lambda_{30}\right]\tag{2.3}$$

Note that if $n\to N$ then $B\left(\widehat{DI}_y\right)\to 0$, the naive estimator $\widehat{DI}_y$ is a consistent estimator of the Fano's Dispersion Index $\left(DI_y\right)$ (or Fano factor).

Theorem 2.2. The mean squared error, to the first order of approximation, of the naive estimator $\widehat{DI}_y$ is given by

$$MSE\left(\widehat{DI}_y\right)=\left(\frac{1-f}{n}\right)\left(DI_y\right)^2\left[\lambda_{40}-1+C_y^2-2C_y\lambda_{30}\right]\tag{2.4}$$

Clearly as $n \to N$, the $MSE\left(\widehat{DI}_y\right) \to 0$, which shows that $\left(\widehat{DI}_y\right)$ is a consistent estimator of the Fano's Dispersion Index $\left(DI_y\right)$.

Corollary 2.1. A consistent estimator of $B\left(\widehat{DI}_y\right)$ is given by

$$\hat{B}\left(\widehat{DI}_y\right) = \left(\frac{1-f}{n}\right)\left(\widehat{DI}_y\right)\left[\hat{c}_y^2 - \hat{c}_y\hat{\lambda}_{30}\right] \tag{2.5}$$

. and that of $MSE\left(\widehat{DI}_y\right)$ is given by

$$\widehat{MSE}\left(\widehat{DI}_y\right) = \left(\frac{1-f}{n}\right)\left(\widehat{DI}_y\right)^2\left[\hat{\lambda}_{40} - 1 + \hat{c}_y^2 - 2\hat{c}_y\hat{\lambda}_{30}\right] \tag{2.6}$$

In the next section, we introduce a new ratio type estimator of the dispersion index of the study variable by making use of the known value of the dispersion index of the auxiliary variable.

3. RATIO TYPE ESTIMATOR OF THE FANO'S DISPERSION INDEX

Assuming the Fano's Dispersion Index of an auxiliary variable, x, given by

$$DI_x = \frac{S_x^2}{\bar{X}} \tag{3.1}$$

. is known. A natural naive estimator of the DI_x is given by

$$\widehat{DI}_x = \frac{s_x^2}{\bar{x}_n} \tag{3.2}$$

Following Cochran (1940), we define a ratio type estimator of the Fano's Dispersion Index as:

$$\widehat{DI}_{Rat} = \left(\widehat{DI}_y\right)\frac{DI_x}{\left(\widehat{DI}_x\right)} \tag{3.3}$$

Now we have the following theorems.

Theorem 3.1. The bias, to the first order of approximation in the ratio estimator $\widehat{DI}_{Rat}$ is given by

$$B\left(\widehat{DI}_{Rat}\right) = \left(\frac{1-f}{n}\right)\left(DI_y\right)[C_y^2 + \lambda_{04} + C_x\lambda_{21} + C_y\lambda_{12}$$
$$-\rho_{xy}C_xC_y - C_y\lambda_{30} - C_x\lambda_{03} - \lambda_{22}] \tag{3.4}$$

Clearly $B\left[\widehat{DI}_{Rat}\right] \to 0$ as $n \to N$, which shows that the ratio type estimator $\left(\widehat{DI}_{Rat}\right)$ is a consistent estimator of DI_y.

Theorem 3.2. The mean squared error, to the first order of approximation, of the ratio estimator $\widehat{DI}_{Rat}$ is given by

$$\text{MSE}[\widehat{DI}_{Rat}] = \left(\frac{1-f}{n}\right)(DI_y)^2[\lambda_{40} + C_y^2 -$$
$$2C_y\lambda_{30} + \lambda_{04} + C_x^2 - 2C_x\lambda_{03}$$
$$-2\left(\lambda_{22} - C_y\lambda_{12} - C_x\lambda_{21} + \rho_{xy}C_xC_y\right)] \tag{3.5}$$

Theorem 3.3. A consistent estimator of $B\left[DI_{Rat}\right]$ is given by

$$\hat{B}\left[\widehat{DI}_{Rat}\right] = \left(\frac{1-f}{n}\right)\left(\widehat{DI}_y\right)[\hat{c}_y^2 + \hat{\lambda}_{04} + \hat{c}_x\hat{\lambda}_{21} +$$
$$\hat{c}_y\hat{\lambda}_{12} - \hat{\rho}_{xy}\hat{c}_x\hat{c}_y - \hat{c}_y\hat{\lambda}_{30} - \hat{c}_x\hat{\lambda}_{03} - \hat{\lambda}_{22}]$$

and a consistent estimator of the $MSE\left[\widehat{DI}_{Rat}\right]$ is given by

$$\widehat{MSE}\left[\widehat{DI}_{Rat}\right] = \left(\frac{1-f}{n}\right)\left(\widehat{DI}_y\right)^2[\hat{\lambda}_{40} + \hat{c}_y^2 - 2\hat{c}_y\hat{\lambda}_{30} +$$
$$\hat{\lambda}_{04} + \hat{c}_x^2 - 2\hat{c}_x\hat{\lambda}_{03} - 2\left(\hat{\lambda}_{22} - \hat{c}_y\hat{\lambda}_{12} - \hat{c}_x\hat{\lambda}_{21} + \hat{\rho}_{xy}\hat{c}_x\hat{c}_y\right)]$$

Theorem 3.4. The ratio estimator will be more efficient than the naive estimator if

$$\rho_{xy} > \frac{\lambda_{04} + 1 + C_x^2 + 2\left\{C_y\lambda_{12} + C_x\lambda_{21} - C_x\lambda_{03} - \lambda_{22}\right\}}{2C_yC_x} \tag{3.6}$$

In the next section, we consider a regression type estimator for estimating the Fano's Dispersion Index.

4. REGRESSION TYPE ESTIMATOR OF THE FANO'S DISPERSION INDEX

We define a regression type estimator of DI_y as

$$\widehat{DI}_{Reg} = \widehat{DI}_y + \beta\left(DI_x - \widehat{DI}_x\right) \tag{4.1}$$

when β is a constant to be determined such that the mean squared error of the estimator $\widehat{DI}_{Reg}$ is minimum. We have the following theorems:

Theorem 4.1. The bias, to the first order of approximation, of the regression type estimator $\widehat{DI}_{Reg}$ is given by

$$B\left[\widehat{DI}_{Reg}\right]=\left(\frac{1-f}{n}\right)\left[\begin{array}{l}(DI_y)C_y(C_y-\lambda_{30})-\\ \beta(DI_x)C_x(C_x-\lambda_{03})\end{array}\right] \quad (4.2)$$

Clearly $B\left[\widehat{DI}_{Reg}\right]\to 0$ as $n\to N$, so $\widehat{DI}_{Reg}$ is a consistent estimator of DI_y.

Theorem 4.2. For the optimal value of

$$\beta=\frac{(DI_y)(\lambda_{22}-1-C_y\lambda_{12}-C_x\lambda_{21}+\rho_{xy}C_xC_y)}{(DI_x)(\lambda_{04}-1+C_x^2-2C_x\lambda_{03})} \quad (4.3)$$

the minimum mean squared error, to the first order of approximation, of the regression type estimator $\widehat{DI}_{Reg}$ is given by

$$Min.MSE\ [\widehat{DI}_{Reg}]$$
$$=MSE\left(\widehat{DI}_y\right)\left[1-\frac{(\lambda_{22}-1-C_y\lambda_{12}-C_x\lambda_{21}+\rho_{xy}C_xC_y)^2}{(\lambda_{40}-1+C_y^2-2C_y\lambda_{30})(\lambda_{04}-1+C_x^2-2C_x\lambda_{03})}\right] \quad (4.4)$$

Note that the optimum value of β depends on unknown parameters, however a consistent estimator of β is given by

$$\hat{\beta}=\frac{(\widehat{DI}_y)(\hat{\lambda}_{22}-1-\hat{c}_y\hat{\lambda}_{12}-\hat{c}_x\hat{\lambda}_{21}+\hat{\rho}_{xy}\hat{c}_x\hat{c}_y)}{(\widehat{DI}_x)(\hat{\lambda}_{04}-1+\hat{c}_x^2-2\hat{c}_x\hat{\lambda}_{03})} \quad (4.5)$$

So, a practically useful regression type estimator of Fano's Dispersion Index, DI_y, is given by

$$\widehat{DI}^{p}_{Reg}=\widehat{DI}_y+\hat{\beta}\left(DI_x-\widehat{DI}_x\right) \quad (4.6)$$

It is straightforward to show that the asymptotic mean squared error, to the first order of approximation, of the proposed estimator $\widehat{DI}^{p}_{Reg}$ is equivalent to that of the estimator $\widehat{DI}_{Reg}$ i.e.

$$MSE\left[\widehat{DI}^{p}_{Reg}\right]=MSE\left[\widehat{DI}_{Reg}\right] \quad (4.7)$$

It is worth pointing out that, following Srivatava (1971), various estimators of the form:

$$e_1=\widehat{DI}_y exp\left(DI_x/\widehat{DI}_x\right) \quad (4.8)$$

$$e_2=\widehat{DI}_y log\left(DI_x/\widehat{DI}_x\right) \quad (4.9)$$

and

$$e_3=\widehat{DI}_y\left(DI_x/\widehat{DI}_x\right)^{\alpha} \quad (4.10)$$

where α is any real value, etc., have variance greater than or equal to the variance of the regression type estimator. No doubt a new papermill can be created to estimate Fano's Dispersion Index based on such extensions, but all those estimators would belong to Srivastava (1971) class of estimators.

In the next section, we provide empirical evidence for the performance of various estimators through a simulation study.

5. EMPIRICAL EVIDENCE: SIMULATION

Using Statistical Analysis Software (SAS), a simulation study was conducted to compare the relative biases of the new proposed naive estimator, the proposed ratio estimator, and the proposed regression estimator, and to study the percent relative efficiencies of the proposed ratio and the regression estimators with respect to the naïve estimator. Following Singh and Horn (1998), we let the study variable, y, be given by

$$y_i=\mu_y+\left(\sqrt{1-\rho^2}\right)\left(y_i^*-\mu_y^*\right)+\rho\left(\frac{\sigma_y}{\sigma_x}\right)\left(x_i^*-\mu_x^*\right) \quad (5.1)$$

and the auxiliary variable, x, be given by

$$x_i=\mu_x+\left(x_i^*-\mu_x^*\right) \quad (5.2)$$

for $i=1,2,\ldots..N$ with N =5000and various values of ρ. The variables y^* and x^* are independent gammavariables generated in SAS.

In particular, we have $y^*\sim G(a_y,b_y)$ and $x^*\sim G(a_x,b_x)$, where a_y, b_y are the shape and scale parameters for the study variable, and a_x, b_x are those for the auxiliary variable. Note that $\mu_y^*=a_yb_y$, $\sigma_y^*=\sqrt{a_yb_y^2}$, $\mu_x^*=a_xb_x$, and $\sigma_x^*=\sqrt{a_xb_x^2}$. Also note that ρ_{xy} ,the observed correlation between the study variable y_i in (5.1) and the auxiliary variable x_i in (5.2), should be close to the value of ρ chosen to generate our population. The means of the transformed variables y_i and x_i will be the chosen values, μ_y and μ_x, respectively. Also, the standard deviations of the transformed variables y_i and x_i will be $\sigma_y = \sigma_y^*$ and $\sigma_x = \sigma_x^*$ respectively.

We computed the relative bias of the naïve estimator by

$$RB\left(\widehat{DI}_y\right) = \frac{B\left(\widehat{DI}_y\right)\times 100\%}{\left(DI_y\right)} \tag{5.3}$$

We also computed the relative bias of the ratio type estimator by

$$RB\left(\widehat{DI}_{Rat}\right) = \frac{B\left(\widehat{DI}_{Rat}\right)\times 100\%}{\left(DI_y\right)} \tag{5.4}$$

followed by the relative bias of the regression type estimator given by

$$RB\left(\widehat{DI}_{Reg}\right) = \frac{B\left(\widehat{DI}_{Reg}\right)\times 100\%}{\left(DI_y\right)} \tag{5.5}$$

To compare the results of the naïve estimator we also calculated $MSE\left(\widehat{DI}_{Rat}\right)$ for the ratio estimator and the $MSE\left(\widehat{DI}_{Reg}\right)$ for the regression estimator. Using these we then computed the percent relative efficiency of the ratio estimator and the regression estimator with respect to the naïve estimator given by

$$RE_1 = \frac{MSE\left(\widehat{DI}_y\right)\times 100\%}{MSE\left(\widehat{DI}_{Rat}\right)} \tag{5.6}$$

and

$$RE_2 = \frac{MSE\left(\widehat{DI}_y\right)\times 100\%}{Min.MSE\left(\widehat{DI}_{Reg}\right)} \tag{5.7}$$

Table 1. Summary relative efficiency, sample size, and relative bias.

ρ	ρ_{xy}	RE_1	RE_2	Sample Size	RB Naive	RB Ratio	RB Regression
0.6	0.60050	76.776	126.042	250	-0.048462	1.11732	-0.001518
				500	-0.022956	0.52926	-0.000719
				750	-0.014454	0.33324	-0.000453
				1000	-0.010203	0.23522	-0.000320
0.7	0.70017	106.374	158.487	250	-0.048604	0.87029	0.014383
				500	-0.023023	0.41224	0.006813
				750	-0.014496	0.25956	0.004290
				1000	-0.010232	0.18322	0.003028
0.8	0.79993	180.061	237.695	250	-0.052037	0.58484	0.029354
				500	-0.024649	0.27703	0.013904
				750	-0.015520	0.17443	0.008755
				1000	-0.010955	0.12312	0.006180
0.9	0.89984	447.760	508.348	250	-0.060313	0.26098	0.041743
				500	-0.028569	0.12362	0.019773
				750	-0.017988	0.07784	0.012450
				1000	-0.012697	0.05494	0.008788

In Table 1, from the population size of 5000, we set ρ to be 0.6, 0.7, 0.8 and 0.9 and ρ_{xy} is the observed correlation between x and y in the population being used at the time of simulation. The sample size is determined by 5%, 10%, 15% and 20% of the population size. The relative efficiency of the regression estimator with respect to the naive estimator is always more than 100%but the relative efficiency of the ratio estimator with respect to the naïve estimator depends on the value of the correlation coefficient, ρ_{xy}. We note that the relative efficiency of the ratio estimator is less than 100% with respect to the naïve estimator when ρ is less than 0.70. The absolute value of percent relative bias for every sample size generated was less than 10% for the ratio, regression and the naïve estimator. In every case of the simulation study, the absolute value of relative bias for the naive estimator is always less than the absolute value of relative bias for the ratio estimator, while the absolute value of relative bias in the regression estimator is always least among the three estimators. In Table 1, the negative values of relative bias indicate that the estimators are underestimating in those situations. SAS codes used in the simulation study are documented in Appendix B.

In the next section, we examine the performance of the proposed estimators by considering real data applications.

6. APPLICATIONS

The real-world data used to assess the relative efficiency and relative bias in the new proposed naïve estimator, ratio type estimator and regression type estimator, can be found in Särndal, Swensson and Wretman (1992) titled "the CO124 Population." The population set contains 124 countries, the variables used were IMP, EXP, GNP, and MEX. The variable IMP stands for the 1983 imports (in millions of the U.S. dollars), EXP stands for the 1983 exports (in millions of U.S. dollars), GNP stands for the 1982 gross national product (in tens of millions of U.S. dollars), and MEX stands for 1981 military expenditure (in millions of U.S. dollars). Table 2 gives a summary of descriptive parameters for these variables.

Table 2. Summary of the descriptive parameters for each variable

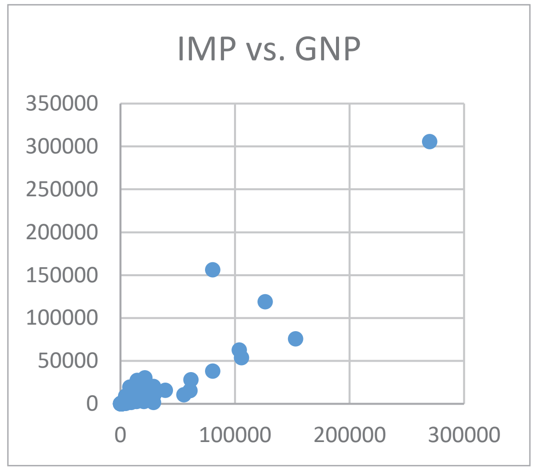
Variables	Mean	Median	St. Dev	Range	Min	Max
IMP	14604.49	2892	34064.88	269769	109	269878
EXP	14276.90	2291	31431.48	200505	33	200538
GNP	10240.84	1288	33638.20	305658	32	305690
MEX	4150.13	455	16626.12	134388	2	134390

In the simulation study, we noted that the ratio type estimator is not more efficient than the naïve estimator in the case of two variables having correlation of 0.60 (or low), so the correlation is critical in deciding which pair to use. The values of the correlation coefficient between the various pairs of the above variables is given in Table 3.

Table 3. The correlation between various variables

Variables	IMP	EXP	GNP	MEX
IMP	1	.976	.906	.751
EXP	.976	1	.859	.697
GNP	.906	.859	1	.912
MEX	.751	.697	.912	1

We ran four scenarios with two variables per case. In the first two cases, we used study variable IMP and the auxiliary variables GNP and MEX. The next two cases we needed to use a logarithmic transformation to optimize the relative efficiency and relative bias. In Figure 1, we display the scatterplots of the variables IMP versus GNP and IMP versus MEX. Table 4 displays the performance of the first case, IMP vs. GNP, followed by a summary.



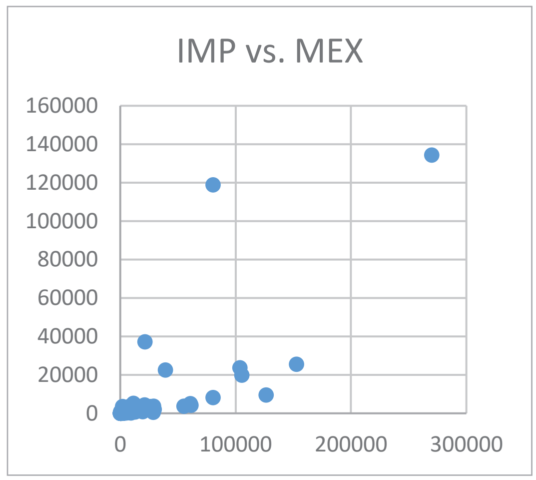


Fig. 1. Scatterplots with dependent variable, IMP.

Table 4. Summary of the performance of the proposed estimators for IMP vs. GNP.

f.p.c. f	DI_y	DI_x	ρ_{xy}	RB Naive	RB Ratio	RB Reg	RE Ratio	RE Reg
0.05	79456.1	110491.8	0.906	-84.40	-48.09	41.90	567.56	1067.73
0.10	79456.1	110491.8	0.906	-39.98	-22.78	19.85	567.56	1067.73
0.15	79456.1	110491.8	0.906	-25.17	-14.34	12.50	567.56	1067.73
0.20	79456.1	110491.8	0.906	-17.77	-10.12	8.82	567.56	1067.73
0.25	79456.11	110491.8	0.906	-13.33	-7.59	6.62	567.56	1067.73
0.30	79456.1	110491.8	0.906	-10.36	-5.91	5.15	567.56	1067.73
0.35	79456.1	110491.8	0.906	-8.25	-4.70	4.10	567.56	1067.73

The dispersion index for the y variable, DI_y is 79456.1 and for the variable x, DI_x is 110491.8. The ρ_{xy} is 0.906 which falls into the acceptable range for the ratio and the regression estimators. The absolute value of relative bias should be under 10%, so the sample size less than 30% of the population size would be too small for the absolute value of relative bias for the naïve estimator to meet the criteria. For the absolute value of relative bias of the ratio estimator to be less than 10%, a minimum 25% of the population size would be required to meet the criteria. Finally, for the regression estimator a minimum of 20% sample size of the population would be sufficient meet the criteria of 10% absolute relative bias. The negative values of the relative biases in Table 4 indicate the naïve and the ratio estimators are underestimating, whereas positive value of the relative bias for the regression estimator indicates it is overestimating. We note the relative efficiency for the ratio estimator with respect to the naïve estimator is 567.56%and is free from the sample size. The relative efficiency for the regression estimator with respect to the naïve estimator is 1067.73%,and which is also free from the sample size. For these variables IMP and GNP, the relative efficiency for the regression type estimator is always more than that of the ratio type estimator.

Next, we are considering the performance of the proposed estimators with the variables IMP and MEX.

Table 5. Summary of the performance of the proposed estimators for IMP vs. MEX

f.p.c. f	DI_y	DI_x	ρ_{xy}	RB Naive	RB Ratio	RB Reg	RE Ratio	RE Reg
0.05	79456.1	66607.0	0.751	-84.40	-53.88	48.96	217.63	232.86
0.10	79456.1	66607.0	0.751	-39.98	-25.52	23.19	217.63	232.86
0.15	79456.1	66607.0	0.751	-25.17	-16.07	14.60	217.63	232.86
0.20	79456.1	66607.0	0.751	-17.77	-11.34	10.31	217.63	232.86
0.25	79456.11	66607.0	0.751	-13.33	-8.85	7.73	217.63	232.86
0.30	79456.1	66607.0	0.751	-10.36	-6.62	6.01	217.63	232.86
0.35	79456.1	66607.0	0.751	-8.25	-5.27	4.79	217.63	232.86

The dispersion index of y, DI_y is the same as before since it is the same variable IMP, while the dispersion index of x, DI_x, is 66607.0 and the ρ_{xy} is 0.751. A sample size less than 30% of the population would be too small of a sample to have an absolute value of relative bias less than 10% in the naïve estimator. A sample size less than 25% of the population would be too small for the naïve estimator to have an absolute value of relative bias less than 10%. A sample size less than 20% of the population would be too small to have an absolute value of relative bias less than 10% in both the ratio and the regression type estimators. The negative values of the relative biases in Table 5 indicate underestimation and positive values indicate overestimation. The relative efficiency for the ratio estimator with respect to the naïve estimator is 217.63% and the relative efficiency for the regression estimator with respect to the naïve estimator is 232.86%both being independent of sample size. In Figure 2 we display the scatterplot of the EXP vs. GNP and the EXP vs. MEX. Table 6 shows the performance of the proposed estimators for the variables EXP vs. x=log(MEX).

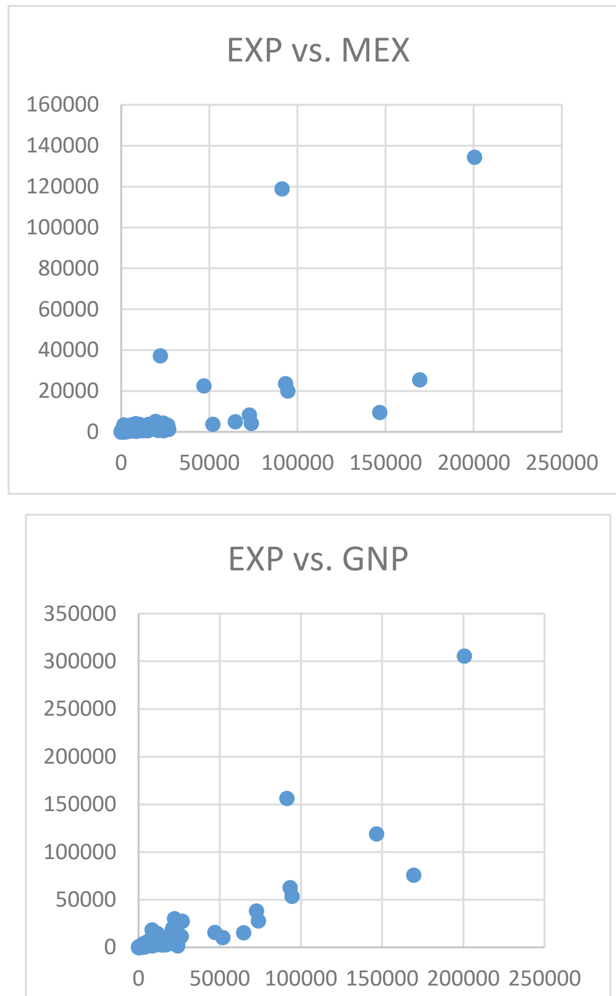


Fig. 2. Scatterplots with dependent variable, EXP.

To optimize performance of the estimators, we did a logarithmic transformation of the x variable. From Table 6, we see the dispersion index of the y variable (EXP) is 69198.4 and the dispersion index of the x variable log(MEX) is 0.815, the correlation between the variables is 0.632 which is within the

Table 6. Summary of the performance of the proposed estimators for EXP vs. log(MEX).

f.p.c. f	DI_y	DI_x	ρ_{xy}	RB Naive	RB Ratio	RB Reg	RE Ratio	RE Reg
0.05	69198.4	0.815	0.632	-53.33	-73.58	-54.33	118.53	120.25
0.10	69198.4	0.815	0.632	-25.26	-34.85	-25.73	118.53	120.25
0.15	69198.4	0.815	0.632	-15.90	-21.94	-16.20	118.53	120.25
0.20	69198.4	0.815	0.632	-11.22	-15.49	-11.43	118.53	120.25
0.25	69198.4	0.815	0.632	-8.42	-11.62	-8.58	118.53	120.25
0.30	69198.4	0.815	0.632	-6.55	-9.04	-6.67	118.53	120.25
0.35	69198.4	0.815	0.632	-5.21	-7.19	-5.31	118.53	120.25

range for the ratio estimator. For the naïve estimator, the sample size needs to be 25% of the population to have absolute value of the relative bias less than 10%. The ratio type estimator needs a sample size of 30% of the population to have absolute value of the relative bias less than 10%. The regression type estimator needs a sample size 25% of the population to have absolute value of the relative bias less than 10%. In Table 6 all negative values of relative biases indicate that all the three estimators viz. naïve, ratio and regression type estimators are underestimating. The relative efficiency of the ratio estimator with respect to the naïve estimator is 118.53% and the relative efficiency of the regression estimator with respect to the naïve estimator is 120.25%. The performance of the proposed estimators for the variables EXP vs. log(GNP) can be seen in Table 7.

Table 7. Summary of the performance of the proposed estimators for EXP vs. log(GNP).

f.p.c. f	DI_y	DI_x	ρ_{xy}	RB Naïve	RB Ratio	RB Reg	RE Ratio	RE Reg
0.05	69198.4	0.529	0.691	-53.33	-85.94	-53.06	141.44	142.39
0.10	69198.4	0.529	0.691	-25.26	-40.71	-25.13	141.44	142.39
0.15	69198.4	0.529	0.691	-15.91	-25.63	-15.82	141.44	142.39
0.20	69198.4	0.529	0.691	-11.23	-18.09	-11.17	141.44	142.39
0.25	69198.4	0.529	0.691	-8.42	-13.57	-8.38	141.44	142.39
0.30	69198.4	0.529	0.691	-6.55	-10.55	-6.52	141.44	142.39
0.35	69198.4	0.529	0.691	-5.21	-8.40	-5.19	141.44	142.39

Again, the use of a logarithmic transformation of the auxiliary variable x=log(GNP) is needed to optimize the performance. The dispersion index of y is 69198.4 and the dispersion index of the x variable is 0.529. The correlation is 0.691 between EXP and log(GNP). The absolute value of relative bias for the naïve estimator is less than 10% when the sample size is greater than or equal to 25% of the population size. The absolute value of the relative bias in the ratio estimator is less than 10% when the sample size is greater than 30% of the population size. For the regression estimator to have anabsolute value of the relative bias be less than 10%, the sample size needs to be greater than or equal to 25% of the population size. Again, the negative values of the relative biases in Table 7 indicate that all the three estimators are underestimating. The relative efficiency for the ratio estimator is 141.44% efficient and the relative efficiency for regression estimator with respect to the naïve estimator is 142.39%. The real-world data shows that the regression estimator is more efficient than the ratio and the naïve estimator.The SAS codes for the application section will be in Appendix C. The original version of this work is available in Chavez (2020).

ACKNOWLEDGEMENTS

The authors are grateful to the Guest Editor, Dr. Rajender Parsad, Dr. Manish Sharma, and the reviewer for their valuable comments on the original version of this manuscript.

REFERENCES

Chavez, S. (2020).*Estimation of dispersion index using auxiliary information*. Unpublished thesis submitted to the Department of Mathematics, Texas A&M University-Kingsville, Kingsville, TX.

Cochran, W. G. (1940). The estimation of the yields of cereal experiment by sampling for the ratio of grain to total produce. *The Journal of Agriculture Science*, **30(2)**, 262-275.

David, I.P. and Sukhatme, B.V. (1974). On the bias and mean square error of the ratio estimator. *Journal of the American Statistical Association*, **69**, 464-466.

De Oliveira, R.P., de Oliveira Peres, M.V, Martinez, E.Z. and Achcar, J.A. (2019). Use of a discreteSushila distribution in the analysis of right-censored lifetime data. *Model Assisted Statistics and Applications,* **14(3)**, 255-268.

Fano, U. (1947). Ionization Yield of Radiations. II. The Fluctuations of the Number of Ions. *Physical Review*. **72(1)**, 26-29

Hansen, M.H., Hurwitz, W. N. and Madow, W. G. (1953) Sample Survey Methods and Theory Vol. 1, *Wiley*, New York.

Särndal C.E, Swenson B., Wretman J. (1992) *Model Assisted Survey Sampling*. Springer-Verlag.

Singh, S. and Horn, S. (1998). An alternative estimator for multi-character surveys. *Metrika,* **48**, 99-107.

Srivastava, S.K. (1971). A generalized estimator for the mean of a finite population using mutii-auxiliary information. *J. Amer. Statist. Assoc.*, **66(334)**, 404-407.

Srivastava, S.K. and Jhajj, H.S. (1981). A class of estimators of the population mean in survey sampling using auxiliary information. *Biometrika*, **68**, 341-343.

Sukhatme, P.V. (1954). *Sampling theory of surveys with applications*. Indian Society of Agricultural Statistics, New Delhi.

Sukhatme, P.V. and Sukhatme, B.V. (1970). *Sampling theory of surveys with applications.* Second Edition, Asia Publishing House, Bombay, India.

Tracy, D.S. (1984). Moments of sample moments. *Commun. Statist.-Theory Meth*, **3(5)**, 553-562.

Wu, C.-F. (1982). Estimation of Variance of the Ratio Estimator. *Biometrika*, **69(1)**, 183-189.

APPENDIX-A: PROOFS OF THEOREMS

Online documented Appendix-A or refer to Chavez (2020).

APPENDIX-B:SAS CODES FOR SIMULATION

```
*SAS codes used for paper;
data data1;
np=5000;
ay=2.0;
by=1.5;
ax=1.6;
bx=1.9;
callstreaminit(12345);
do I=1to np by1;
ystar= rand('gamma',ay,by);
xstar= rand('gamma',ax,bx);
sigmays= sqrt(ay*by**2);
meanys= ay*by;
sigmaxs= sqrt(ax*bx**2);
meanxs= ax*bx;
output;
end;
data data2;
set data1;
do rho=0.6to0.91by0.1;
meany= 15;
meanx= 10;
yp= meany + (ystar-meanys)*sqrt(1-rho**2) + sigmays*rho*(xstar-meanxs)/sigmaxs;
xp= meanx+(xstar-meanxs);
output;
end;
keep rho ypxp;
proc sort data=data2;
by rho;
proc gplot data=data2;
plot yp*xp;
by rho;
run;
proc sort data=data2;
by rho;
proc means data=data2 noprint;
variable ypxp;
by rho;
output out= data21 mean=meanypmeanxp;
data data22;
set data21;
drop _type_ _freq_;
run;
proc print data=data22;
run;
proc sort data=data22;
by rho;
data data3;
merge data2 (in=ABC) data22 (in=ABC);
by rho;
*proc print data=data3;
run;
data data4;
set data3;
mu20= (yp-meanyp)**2;
mu30= (yp-meanyp)**3;
mu40= (yp-meanyp)**4;
mu02= (xp-meanxp)**2;
mu03= (xp-meanxp)**3;
mu04= (xp-meanxp)**4;
mu11= (yp - meanyp)*(xp-meanxp);
mu12= (yp-meanyp)*(xp-meanxp)**2;
mu21= (yp-meanyp)**2*(xp-meanxp);
mu22= (yp-meanyp)**2*(xp-meanxp)**2;
*proc print data=data4;
run;
proc means data=data4 noprint;
var mu20 mu30 mu40 mu02 mu03 mu04 mu11 mu22 mu21 mu12;
by rho;
output out= data41 sum=smu20 smu30 smu40 smu02 smu03 smu04 smu11 smu22 smu21 smu12;
proc print data=data41;
run;
proc sort data=data21;
by rho;
```

```
proc sort data=data41;
by rho;
data data5;
merge data21 (in=abc) data41 (in=abc);
by rho;
proc print data=data5;
run;
data data6;
set data5;
np= _freq_;
mu20f=smu20/ (np-1);
mu30f=smu30/ (np-1);
mu40f=smu40/ (np-1);
mu02f=smu02/ (np-1);
mu03f=smu03/ (np-1);
mu04f=smu04/ (np-1);
mu11f=smu11/ (np-1);
mu22f=smu22/ (np-1);
mu21f=smu21/ (np-1);
mu12f=smu12/ (np-1);
drop smu20 smu30 smu40 smu02 smu03 smu04 smu11 smu22 smu21
    smu12 _freq_ _type_;
proc print data=data6;
run;
data data7;
set data6;
do frac = 0.05to.20by.05;
ns=frac*np;
f=ns/np;
diy= mu20f/ meanyp;
cy=sqrt(mu20f)/meanyp;
lam40= mu40f/mu20f**2;
lam30= mu30f/ mu20f**(3/2);
**********naive estimator**************;
b_naive= ((1-f)/ns)*diy*(cy**2-cy*lam30);
rb_naive= b_naive*100/diy;
mse_naive= (((1-f)/ns)*diy**2*(lam40-1+cy**2-2*cy*lam30));
**********ratio estimator***************;
lam04= mu04f/mu02f**2;
lam03= mu03f/ mu02f**(3/2);
lam21= mu21f/ (mu20f*mu02f**(1/2));
lam12= mu12f/ (mu20f**(1/2)*mu02f);
lam22= mu22f/(mu20f*mu02f);
rhoxy= mu11f/ sqrt(mu20f*mu02f);
cx=sqrt(mu02f)/meanxp;
b_ratio= ((1-f)/ns)*diy*(cy**2+lam04+cx*lam21-rhoxy*cx*cy-cy*
    lam30-cx*lam03-lam22);
rb_ratio=b_ratio*100/diy;
mse_ratio= ((1-f)/ns)*(diy**2)*(lam40+cy**2-2*cy*lam30+
    lam04+cx**2-2*cx*lam03-2*(lam22-cy*lam12-cx*lam21
    +rhoxy*cx*cy));
re_ratio= mse_naive*100/mse_ratio;
******************regression estimator******************;
dix= mu02f/meanxp;
beta_num=diy*(lam22-1-cy*lam12-cx*lam21+rhoxy*cx*cy);
beta_dem= dix*(lam04-1+cx**2-2*cx*lam03);
beta= beta_num/beta_dem;
b_reg= ((1-f)/ns)*(diy*(cy**2-cy*lam30)-beta*dix*(cx**2-cx*
    lam03));
rb_reg= b_reg*100/diy;
t11=(lam22-1-cy*lam12-cx*lam21+rhoxy*cx*cy);
t10=lam40-1+cy**2-2*cy*lam30;
t01=lam04-1+cx**2-2*cx*lam03;
mse_reg= mse_naive*(1-t11**2/(t10*t01));
re_reg= mse_naive*100/mse_reg;
output;
end;
keep np ns f rho rhoxy rb_naive rb_ratio rb_reg re_ratio re_reg;
proc print data= data7;
var np ns f rho rhoxy rb_naive rb_ratio rb_regre_ratio re_reg;
run;
```

APPENDIX-C: SAS CODES FOR APPLICATIONS

```
*SAS Codes used in the application;
Proc import datafile =”E:\trial3.xls”out=data1 dbms=xls replace;
sheet=”trial3”;
proc print data=data1;
run;
data data2;
set data1;
xp= mex;
yp= imp;
;
Keep yp xp rho;
*proc print data=data2;
run;
*********plot******;
Proc gplot data=data2;
Plot yp*xp;
```

```
run;
proc summary data=data2;
var yp xp;
output out=out1;
data out1;
set out1;
drop _type_ _freq_;
proc print data=out1;
run;
proc means data=data2 noprint;
variables yp xp;
output out= data21 mean=meanyp meanxp;
data data22;
set data21;
drop _type_ _freq_;
run;
*proc print data=data22;
run;
*proc sort data=data22;
*by rho;
data data3;
set data2;
if _n_=1 then set data22;
*proc print data=data3;
run;
data data4;
set data3;
mu20= (yp-meanyp)**2;
mu30= (yp-meanyp)**3;
mu40= (yp-meanyp)**4;
mu02= (xp-meanxp)**2;
mu03= (xp-meanxp)**3;
mu04= (xp-meanxp)**4;
mu11= (yp - meanyp)*(xp-meanxp);
mu12= (yp-meanyp)*(xp-meanxp)**2;
mu21= (yp-meanyp)**2*(xp-meanxp);
mu22= (yp-meanyp)**2*(xp-meanxp)**2;
*proc print data=data4;
run;
proc means data=data4 noprint;
var mu20 mu30 mu40 mu02 mu03 mu04 mu11 mu22 mu21 mu12;
output out= data41 sum=smu20 smu30 smu40 smu02 smu03 smu04
    smu11 smu22 smu21 smu12;
*proc print data=data41;
run;
data data5;
set data21;
if _n_=1 then set data41;
*proc print data=data5;
run;
data data6;
set data5;
np= _freq_;
mu20f=smu20/ (np-1);
mu30f=smu30/ (np-1);
mu40f=smu40/ (np-1);
mu02f=smu02/ (np-1);
mu03f=smu03/ (np-1);
mu04f=smu04/ (np-1);
mu11f=smu11/ (np-1);
mu22f=smu22/ (np-1);
mu21f=smu21/ (np-1);
mu12f=smu12/ (np-1);
drop smu20 smu30 smu40 smu02 smu03 smu04 smu11 smu22 smu21
    smu12 _freq_ _type_;
*proc print data=data6;
run;
data data7;
set data6;
do frac = 0.05 to 0.35 by .05;
ns=frac*np;
f=ns/np;
diy= mu20f/ meanyp;
cy=sqrt(mu20f)/meanyp;
lam40= mu40f/mu20f**2;
lam30= mu30f/ mu20f**(3/2);
**********naive estimator**************;
b_naive= ((1-f)/ns)*diy*(cy**2-cy*lam30);
rb_naive= b_naive*100/diy;
mse_naive= (((1-f)/ns)*diy**2*(lam40-1+cy**2-2*cy*lam30));
**********ratio estimator***************;
lam04= mu04f/mu02f**2;
lam03= mu03f/ mu02f**(3/2);
lam21= mu21f/ (mu20f*mu02f**(1/2));
lam12= mu12f/ (mu20f**(1/2)*mu02f);
lam22= mu22f/(mu20f*mu02f);
rhoxy= mu11f/ sqrt(mu20f*mu02f);
cx=sqrt(mu02f)/meanxp;
```

```
b_ratio= ((1-f)/ns)*diy*(cy**2+lam04+cx*lam21-rhoxy*cx*cy-cy*
    lam30-cx*lam03-lam22);
rb_ratio=b_ratio*100/diy;
mse_ratio= ((1-f)/ns)*(diy**2)*(lam40+cy**2-2*cy*lam30+lam04+
    cx**2-2*cx*lam03-2*(lam22-cy*lam12-cx*lam21+rhoxy*cx*
    cy));
re_ratio= mse_naive*100/mse_ratio;
******************regression estimator******************;
dix= mu02f/meanxp;
beta_num=diy*(lam22-1-cy*lam12-cx*lam21+rhoxy*cx*cy);
beta_dem= dix*(lam04-1+cx**2-2*cx*lam03);
beta= beta_num/beta_dem;
b_reg= ((1-f)/ns)*(diy*(cy**2-cy*lam30)-beta*dix*(cx**2-cx*
    lam03));
rb_reg= b_reg*100/diy;
t11=(lam22-1-cy*lam12-cx*lam21+rhoxy*cx*cy);
t10=lam40-1+cy**2-2*cy*lam30;
t01=lam04-1+cx**2-2*cx*lam03;
mse_reg= mse_naive*(1-t11**2/(t10*t01));
re_reg= mse_naive*100/mse_reg;
output;
end;
keep np ns f diy dix rhoxy rb_naive rb_ratio rb_reg re_ratio re_reg;
proc print data= data7;
var np ns f diy dix rhoxy rb_naive rb_ratio rb_reg re_ratio re_reg;
run;
```