Comparison of Three Estimators through Simulation Technique

Jakkula Srinivas[1] and D.Vijaya Laxmi[2]

[1]*Department of Statistics, Kakatiya University, Warangal.*

[2]*Department of Statistics, TSWRDC (W), Warangal west, Warangal.*

Abstract: *The objective of this paper is to compare the three estimators namely, Mean per unit, Ratio and Regression estimators with respect to relative bias and relative efficiency using Monte carlo simulation for the four bivariate populations viz., Uniform, Exponential, Normal and Double exponential.*

Keywords: *Simple random sampling, Mean per unit, Ratio estimator, Regression estimator, Relative bias, Relative efficiency and Simulation.*

I. Introduction

A collection of objects under study is known as population. The number of objects in the population is known as population size. It may be finite or infinite. In sampling theory, we assume that population size is finite. **A part or subset of the population is known as sample. The method of selecting a sample is known as sampling method. A sample is small if its size is less than 30 and otherwise it is known as large.** A statistical constant of the population is known as parameter. Population mean and population variance are examples for parameters. A statistical constant of the sample is known as statistic, Sample mean and sample variance are examples for statistic. Estimation is the process of estimating the parameters of the population using statistic. A Statistic used to estimate a parameter is known as an estimator of the parameter. The value of an estimator in a particular sample is known as an estimate.

Unbiased Estimator: An estimator T is said to be unbiased estimator for the parameter θ if

$E(T) = \theta$. That is, T is an unbiased estimator of θ if T is equal to θ on the average over all possible samples. If $E(T) \neq \theta$, then T is known as biased estimator of θ and its bias is given by $B(\theta) = E(T) - \theta$. The relative measure of bias is $B(\theta)/\theta$.

The mean square error of an estimator T in estimating θ is defined by $MSE = E(T-\theta)^2$

Relative Efficiency: Given two estimators T_1 and T_2 of a parameter, then the relative efficiency of T_1 as compared to T_2 which differs in respect of sample size or sampling method or both is defined as

$$RE(T_1, T_2) = \frac{MSE(T_2)}{MSE(T_1)},$$ If T_1 and T_2 are unbiased estimators,

then $$RE(T_1, T_2) = \frac{V(T_2)}{V(T_1)}$$

If $RE(T_1, T_2) < 1$, then T_2 is more efficient than T_1

If $RE(T_1, T_2) > 1$, Then T_1 is more efficient than T_2

If $RE(T_1, T_2) = 1$, Then T_1 & T_2 are equally efficient.

Simple Random Sampling (S.R.S)

If the sample is drawn unit by unit with equal probability of selection for every unit of the population at each draw, then the sample is known as simple random sample. The procedure of selecting a simple random sample is known as simple random sampling (S.R.S) method. If a unit that has been selected in the simple random sample is removed from the population for all subsequent draws, then it is known as simple random sampling without replacement (SRSWOR). Otherwise, it is known as simple random sampling with replacement (SRSWR). If all the units in the population are equally important or if the population is homogeneous, then the simple random sampling method is adopted.

Let us consider a finite population of N units and the values of a characteristic y on these N units are denoted by $Y_1, Y_2, ..., Y_N$. Further, assume that a simple random sample of n units is selected from the population and the values of the characteristic y on these n units are denoted by $y_1, y_2, ..., y_n$.

Population mean $= \bar{\mathrm{Y}} = \frac{\mathrm{Y}}{\mathrm{N}} = \frac{1}{\mathrm{N}} \sum_{i=1}^{\mathrm{N}} \mathrm{Y}_i$ Sample mean $= \bar{y} = \frac{y}{n} = \frac{1}{n} \sum_{i=1}^{n} y_i$

Mean per unit estimator

In simple random sampling without replacement, the sample mean per unit is an unbiased estimator of population mean.

i.e., $E(\bar{y}) = \bar{\mathrm{Y}}$ (1.1)

$$V(\bar{y}_{srs}) = \frac{\mathrm{N}-n}{\mathrm{N}n} \mathrm{S}^2 = \frac{\mathrm{S}^2}{n}(1-f) \qquad \text{where } f = \frac{n}{\mathrm{N}} \tag{1.2}$$

Ratio estimator of population mean

In ratio method of estimation, an auxiliary variable x_i which is correlated with y_i is obtained for each unit in the sample. The population means $\bar{\mathrm{X}}$ of x_i must be known. The ratio estimate of population mean $\bar{\mathrm{Y}}$ is

$$\bar{y}_r = \frac{\bar{y}}{\bar{x}} \bar{X} \tag{1.3}$$

In a simple random sample of size n, (n large)

$$V(\bar{y}_r) = \frac{1-f}{n}\left(S_y^2 + R^2 S_x^2 - 2R\rho S_y S_x\right) \tag{1.4}$$

Where $\rho = \frac{s_{yx}}{s_y s_x}$ is the population correlation between y and x.

Linear regression estimate of population mean

As in the ratio method of estimation, linear regression estimate uses an auxiliary variable x_i that is correlated with y_i. The linear regression estimate of $\bar{\mathrm{Y}}$ is

$$\bar{y}_{lr} = \bar{y} + b(\bar{X} - \bar{x}) \quad (1.5)$$

Where b is a least square estimate of the change in y when x is increased by unity. In simple random sample of size n large [2][4][7]

$$V(\bar{y}_{lr}) = \frac{1-f}{n} S_y^2 (1-\rho^2) \quad (1.6)$$

II. Comparison of estimators using simple random sampling

The approximate formulae for the variance of ratio and regression estimates are valid only when sample size n is large. So, these comparisons are made for the sample size n large.

Therefore, the three comparable variances for the estimated population mean $\bar{Y}$ *as* given in (1.2), (1.4) and (1.6) are as follows:

$$V(\bar{y}) = \frac{N-n}{Nn} S^2 = \frac{S^2}{n}(1-f) \quad \text{(Mean per unit)}$$

$$V(\bar{y}_r) = \frac{1-f}{n}\left(S_y^2 + R^2 S_x^2 - 2R\rho S_y S_x\right) \quad \text{(Ratio)}$$

$$V(\bar{y}_{lr}) = \frac{1-f}{n} S_y^2 (1-\rho^2) \quad \text{(Regression)}$$

It is clear that variance of regression estimate is smaller than that of mean per unit only when $\rho \neq 0$, if $\rho = 0$ then the two variances are equal. And also, the variance of the regression estimate is smaller than that of the ratio estimate if $\left(\rho s_y - R s_x\right)^2 > 0$ or $(B - R)^2 > 0$. Thus the regression estimate is more precise than the ratio estimate if $B \neq R$, this happens only when the relation between y_i and x_i is a straight line through the origin.

In large samples, with simple random sampling, the ratio estimator has a smaller variance than the mean per unit estimator, if $\rho > \frac{c_x}{c_y}$.[4][7]

There is no theoretical expression for the relation among mean per unit, ratio and regression estimates. So, we made an attempt in this paper to derive the relation among these three estimators with respect to relative bias and relative efficiency by taking the samples from the bivariate populations generated from Uniform, Exponential, Normal and Double exponential distributions. [2][4][7]

III. Generation of random samples using simulation technique

Simulation is a technique that generates a large number of simulated samples of data based on an assumed data generating process that characterizes the population from which the simulated samples are drawn. Monte carlo simulation is mainly used when there is a difficulty to solve analytically or when there are too many particles in the system to solve and may be having complex interactions among the particles.

Given a random sample from standard uniform distribution U(0,1), a random sample for any distribution can be obtained by transformation. For some distributions, the transformation from uniform distribution is simple and can be made exactly, but for some distributions more complicated transformations must be approximated.[6]

However, firstly we must consider the generation of independent variate from U(0,1). The most useful source of pseudo random integers is linear congruential sequence. A congruential sequence can take many

forms, but the most commonly used form is $x_i = (ax_{i-1} + c)(mod m)$ for $i = 1,2,\ldots$ where

x_i, a, c & m are the integers and $0 \le x_i \le m$ (3.1)

Integers produced in (3.1) are in the interval $(0, m)$. They are transformed by $\frac{x_i}{m}$ onto (0, 1) over which they approximate a U (0, 1) process.[6]

Inverse method of transformation

Let us consider a continuous random variable with cumulative distribution function $F_X(x)$ i.e., $F_X(x) = P(X \le x)$ then the inverse of $F_X(x)$ denoted by $F^{-1}(x)$ if well defined for $0 \le X \le 1$.

If U is a standard uniform variate U(0,1), then $X = F^{-1}(U)$ is the required distribution function.

Generation of Random Samples from Uniform Distribution

A continuous random variable X is said to follow U (a,b) distribution, its pdf is

$$f(x) = \begin{cases} \dfrac{1}{b-a} & , \quad a < x < b \\ 0 & , \quad otherwise \end{cases}$$, then by inverse transform method,

$$x_i = a + (b-a)u_i , \quad \text{where } u_i \sim U(0,1) \tag{3.2}$$

Generation of Random Samples from Exponential Distribution

A continuous random variable X is said to follow exponential distribution with location parameter *a* (any real number) and scale parameter *b*>0, if its pdf is given by

$$f(x) = \begin{cases} \dfrac{1}{b} e^{\frac{-(x-a)}{b}} & ; x \ge a \\ 0 & ; \ otherwise \end{cases}$$

By inverse transform method,

$$x_i = a - b\ln(u_i) \qquad \text{where } u_i \sim U(0,1). \tag{3.3}$$

Generation of Random Samples from Double Exponential Distribution

A continuous random variable X is said to follow double exponential (Laplace) distribution with location parameter *a* (any real number) and scale parameter *b*>0, if its pdf is given by

$$f(x) = \frac{1}{2b} e^{\frac{-|x-a|}{b}}; \qquad -\infty < x < \infty$$

By inverse transform method,

$$x_i = \begin{cases} a + b\ln(u_2)\ ;\ if\ \ u_1 \geq 1/2 \\ a - b\ln(u_2)\ ;\ if\ \ u_1 < 1/2 \end{cases} \qquad \text{Where } u_1\ and\ u_2 \sim U(0,1) \tag{3.4}$$

Generation of Random Samples from Normal Distribution (Box-Muller Method)

Another method that is also very easy to implement was introduced by Box and Muller (1958). It is a direct transformation of two independent U(0,1) variates U_1 and U_2 two independent N (0,1) variates X_1 and X_2,

$$x_1 = \sqrt{(-2\ln(u_1))}\ \cos(2\pi u_2)$$
$$x_2 = \sqrt{(-2\ln(u_1))}\ \sin(2\pi u_2) \ \text{ where } u_1\ and\ u_2 \sim U(0,1). \tag{3.5}$$

This Method is adopted here for the generation of normal random variable.

If we want to generate bivariate distributions when the variates are independent then we simply generate the distribution for each dimension separately. However there may be known correlations between the variates. To generate correlated random variates in two dimensions, the basic idea is that, we first generate independent variates and then perform a rotation of the coordinate system to bring about the desired correlation.[6]

Thus, the algorithm for generating correlated random variables (X,Y) with the correlation coefficient ρ is as follows.

(1) Independently generate X and X^1 from the same distribution.

(2) Set $Y = \rho X + X^1\sqrt{1-\rho^2}$ } (3.6)

(3) Return the correlated pair (X, Y).

IV. Computation of Relative Bias, Standard error and Relative efficiency

A bivariate population (X, Y) of size N = 2000 is generated as in (3.6) with correlation coefficient ρ (X, Y) = 0.8 when each of X and Y follows uniform, exponential, normal and double exponential distributions. Let $\overline{X}$ and $\overline{Y}$ be the population means of X and Y respectively.

Generation of bivariate populations is given below:

Bivariate uniform distribution: If x and x' follows $U(0,1)$ and Y= $\rho x + x'\sqrt{1-\rho^2}$ then (X,Y) follows Bivariate uniform with correlation coefficient is $\rho = 0.8$

Bivariate exponential distribution: If x and x' follows exp (1),and Y= $\rho x + x'\sqrt{1-\rho^2}$ then (X,Y) follows Bivariate exponential with correlation coefficient is $\rho = 0.8$

Bivariate Normal distribution: If x and x' follows normal (1,1) and Y= $\rho x + x'\sqrt{1-\rho^2}$ then (X,Y) follows Bivariate Normal distribution with correlation coefficient is $\rho = 0.8$

Bivariate Double exponential distribution:If x and x' follows double exponential (1,2) and Y= $\rho x + x'\sqrt{1-\rho^2}$ then (X,Y) follows Bivariate Double exponential distribution with correction coefficient is $\rho = 0.8$.

SRSWOR of size 'n' (n = 10, 30, 50, 70, 90, 110) are selected from each population. Let $\bar{y}_i$ be the mean per unit estimator of $\bar{Y}$ based on the i th sample for i = 1, 2… 1000 (iterations) then the estimator of $E(\bar{y}) = \frac{1}{1000}\sum_{i=1}^{1000}\bar{y}_i = \bar{Y}_{srs}$ (say).

Then the estimate of relative bias = RB ($\bar{Y}_{srs}$) = $\left|\frac{\bar{Y}_{srs} - \bar{Y}}{\bar{Y}}\right|$.

The standard error of mean per unit is given by SE ($\bar{Y}_{srs}$) = $\sqrt{\frac{1}{1000}\sum_{i=1}^{1000}(\bar{y}_i - \bar{Y}_{srs})^2}$.

Similarly, the relative biases and standard errors of ratio $(\bar{Y}_r)$ and regression $(\bar{Y}_{lr})$ estimators are defined below.

$$RB(\bar{Y}_r) = \left|\frac{\bar{Y}_r - \bar{Y}}{\bar{Y}}\right| \qquad RB(\bar{Y}_{lr}) = \left|\frac{\bar{Y}_{lr} - \bar{Y}}{\bar{Y}}\right|$$

$$SE(\bar{Y}_r) = \sqrt{\frac{1}{1000}\sum_{i=1}^{1000}(\bar{y}_{ri} - \bar{Y}_r)^2} \qquad SE(\bar{Y}_{lr}) = \sqrt{\frac{1}{1000}\sum_{i=1}^{1000}(\bar{y}_{lri} - \bar{Y}_{lr})^2}$$

The relative efficiency of an estimator with respect to some other estimator is computed using definition given in section (1).

V. Empirical results

Empirical results from bivariate uniform, bivariate exponential, bivariate normal, bivariate double exponential distributions are shown in the following tables.

Table 5.1: Shows relative bias of the estimators

Population	n	$\bar{Y}_{srs}$	$\bar{Y}_r$	$\bar{Y}_{lr}$	$RB(\bar{Y}_{srs})$	$RB(\bar{Y}_r)$	$RB(\bar{Y}_{lr})$
Bivariate Uniform distribution $\bar{Y}=0.6967$ $\bar{X}=0.4959$	**10**	0.693	0.707	0.694	0.002	0.009	0.008
	30	0.702	0.698	0.697	0.003	0.003	0.004
	50	0.700	0.696	0.697	0.001	0.005	0.005
	70	0.701	0.696	0.697	0.001	0.006	0.004
	90	0.700	0.695	0.697	0.000	0.007	0.005
	110	0.699	0.695	0.696	0.001	0.007	0.005
Bivariate Exponential distribution $\bar{Y}=1.4200$ $\bar{X}=1.0090$	**10**	1.401	1.480	1.412	0.001	0.057	0.009
	30	1.398	1.434	1.406	0.001	0.025	0.004
	50	1.403	1.425	1.408	0.002	0.018	0.005
	70	1.403	1.421	1.408	0.002	0.015	0.006
	90	1.398	1.420	1.407	0.001	0.014	0.005
	110	1.403	1.419	1.408	0.002	0.013	0.006
Bivariate Normal distribution $\bar{Y}=1.4046$ $\bar{X}=0.9916$	**10**	1.376	1.522	1.369	0.017	0.019	0.022
	30	1.408	1.411	1.392	0.006	0.008	0.006
	50	1.402	1.401	1.391	0.002	0.001	0.006
	70	1.400	1.398	1.391	0.000	0.001	0.006
	90	1.398	1.395	1.391	0.001	0.003	0.006
	110	1.400	1.393	1.391	0.000	0.005	0.006
Bivariate Double exponential distribution $\bar{Y}=1.3726$ $\bar{X}=0.9742$	**10**	1.394	1.451	1.360	0.004	0.027	0.028
	30	1.399	1.425	1.378	0.001	0.018	0.016
	50	1.395	1.389	1.378	0.004	0.008	0.016
	70	1.386	1.384	1.376	0.010	0.012	0.017
	90	1.395	1.379	1.378	0.004	0.015	0.016
	110	1.405	1.376	1.383	0.003	0.011	0.012

Table 5.2: Shows relative efficiency of the estimators

Population	n	$S.E.(\bar{Y}_{srs})$	$S.E.(\bar{Y}_r)$	$S.E.(\bar{Y}_{lr})$	$R.E.(\bar{Y}_{srs},\bar{Y}_r)$	$R.E.(\bar{Y}_{srs},\bar{Y}_{lr})$	$R.E.(\bar{Y}_r,\bar{Y}_{lr})$
Bivariate Uniform distribution $\bar{Y}=0.6967$ $\bar{X}=0.4959$	**10**	0.094	0.095	0.061	1.101	0.649	0.677
	30	0.054	0.047	0.033	0.087	0.611	0.702
	50	0.040	0.034	0.024	0.852	0.601	0.706
	70	0.035	0.029	0.021	0.828	0.600	0.724
	90	0.031	0.026	0.019	0.838	0.612	0.731
	110	0.028	0.024	0.017	0.857	0.607	0.708
Bivariate Exponential distribution	**10**	0.317	0.318	0.220	1.003	0.694	0.698
	30	0.190	0.170	0.114	0.895	0.600	0.671
	50	0.138	0.125	0.084	0.906	0.609	0.672
	70	0.116	0.101	0.070	0.871	0.603	0.693

$\bar{Y}=1.4200$ $\bar{X}=1.0090$	**90**	0.108	0.094	0.064	0.870	0.593	0.681
	110	0.096	0.087	0.059	0.906	0.615	0.678
Bivariate Normal distribution $\bar{Y}=1.4046$ $\bar{X}=0.9916$	**10**	0.323	0.358	0.207	1.108	0.641	0.578
	30	0.185	0.173	0.111	0.935	0.600	0.642
	50	0.146	0.135	0.087	0.925	0.596	0.644
	70	0.120	0.116	0.072	0.967	0.601	0.621
	90	0.106	0.104	0.065	0.981	0.613	0.607
	110	0.095	0.092	0.057	0.968	0.603	0.576
Bivariate Double exponential distribution $\bar{Y}=1.3726$ $\bar{X}=0.9742$	**10**	0.351	0.359	0.280	1.022	0.798	0.850
	30	0.325	0.332	0.159	1.021	0.489	0.478
	50	0.247	0.246	0.121	0.995	0.487	0.488
	70	0.219	0.209	0.104	0.954	0.475	0.498
	90	0.193	0.190	0.092	0.984	0.477	0.474
	110	0.177	0.173	0.085	0.977	0.480	0.464

VI. Final Conclusions

(i) The mean per unit estimator has lowest relative bias than that of ratio and linear regression estimators for all four bivariate distributions. The linear regression estimator has less relative bias than that of ratio estimator in bivariate uniform and bivariate exponential distributions. But the ratio estimator has less relative bias than that of linear regression estimator in bivariate normal and bivariate double exponential distributions.

(ii) The linear regression estimator has lowest standard error than that of mean per unit and ratio estimators for all four bivariate distributions. The ratio estimator has less standard error than that of mean per unit in bivariate uniform and bivariate exponential distributions. But the ratio and mean per unit has approximately same standard error in bivariate normal and bivariate double exponential distributions.

(iii) The linear regression estimator is more efficient than that of mean per unit and ratio estimators for all the four bivariate distributions irrespective of sample size. The mean per unit is more efficient than that of ratio estimator if the sample size is small whereas ratio estimator is more efficient than that of mean per unit if the sample size is large.

VII. References

[1] **Basu, D (1958)**, "On Sampling with and without Replacement", Sankhya, 20, pp 287-294.

[2] **Cochran, W.G. (1977)**,"Sampling Techniques", 3rd edition, Wiley Eastern, New Delhi.

[3] **Deng, L.Y. and Wu, C.F.J., (1987)**, "Estimators of Variance of the Regression Estimator", Jour. Amer. Stat. Assoc., 82, pp 568-576.

[4] **Des Raj (1968)**, "Sampling Theory", McGraw-Hill, New York.

[5] **Fuller, W.A. (1975)**,"Regression Analysis of Sample Surveys", Sankhya, C, 37, pp 117-132.

[6] **Kennedy, William J, (1980),** "Statistical Computing", Statistics, text books and monographs, Vol, 33, pp 133-245.

[7] **Mukhopadhyay, P (1998)**, "Theory and Methods of Survey Sampling", Prentice Hall of India, Pvt. Ltd., New Delhi.

[8] **Murthy, M.N., (1967)**, "Sampling Theory and Methods", Statistical Publishing Society, Culcutta.

[9] **Rao, J.N.K. and Beegle, L.D., (1967)**, "A Monte Carlo Study of Some Ratio Estimators", Sankhya, B29, pp 47-56.

[10] **Rao, J.N.K., (1968)**, "Some Small Sample Results in Ratio and Regression Estimation", J. Ind. Stat. Assoc., 6, pp 160-168.