Jour. Ind. Soc. Ag. Statistics
Vol. XXXVIII, No. 1 (1986), pp. 82-91

ON THE ESTIMATION OF FINITE POPULATION CORRELATION COEFFICIENT

S. K. SRIVASTAVA and H. S. JHAJJ
Punjabi University, Patiala

(Received : May, 1982)

SUMMARY

Gupta, Singh and Lal [1] have studied the bias and variance of the sample correlation coefficient r based on a simple random sample from a finite population. They obtained asymptotic expressions for bias and variance of r in terms of sums of powers of the two variables and sums of their products. In this paper asymptotic bias and asymptotic variance of r are obtained in simple expressions involving bivariate moments of the finite population. A class of estimators of the population correlation coefficient is suggested when information about the population mean and population variance of one of the variables is available.

Keywords : Auxiliary information; Correlation coefficient; Finite population; Simple random sampling.

Introduction

The correlation coefficient ρ for a finite population of N units with variate values (x_i, y_i), $i = 1, \ldots, N$ is defined by

$$\rho = S_{xy}/(S_x \cdot S_y) \tag{1.1}$$

where

$$S_{xy} = (N-1)^{-1} \sum_{i=1}^{N} (x_i - \bar{X})(y_i - \bar{Y}),$$

$$S_x^2 = (N-1)^{-1} \sum_{i=1}^{N} (x_i - \bar{X})^2, \; S_y^2 = (N-1)^{-1} \sum_{i=1}^{N} (y_i - \bar{Y})^2$$

and $\overline{X}$ and $\overline{Y}$ denote the population means of the variates x and y respectively.

Based on a simple random sample of size n, (x_i, y_i), $i = 1, \ldots, n$, the conventional estimator of ρ is the corresponding sample correlation coefficient

$$r = s_{xy}/(s_x \cdot s_y) \tag{1.2}$$

where

$$s_{xy} = (n-1)^{-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}),$$

$$s_x^2 = (n-1)^{-1} \sum_{i=1}^{n} (x_i - \bar{x})^2,\ s_y^2 = (n-1)^{-1} \sum_{i=1}^{n} (y_i - \bar{y})^2,$$

$$\bar{x} = n^{-1} \sum_{i=1}^{n} x_i \text{ and } \bar{y} = n^{-1} \sum_{i=1}^{n} y_i \cdot$$

Recently Gupta, Singh and Lal [1] studied the sample correlation coefficient r and obtained its bias and variance up to terms of order n^{-1}. These expressions are given in terms of sums of powers of x_i and y_i and sums of products of powers of x_i and y_i, and are quite lengthy.

In this paper we have obtained asymptotic expressions for the bias and variance of r in terms of the moments of the finite bivariate population, from which the infinite bivariate population case follows immediately.

In sample survey situations, many a times the information on one variable, say x (the auxiliary variable), for all the units in the population is available. In such situations a class of estimators of ρ has been suggested which utilize the known values of the population mean $\overline{X}$ and the population variance S_x^2 of the auxiliary variable x. Asymptotic expressions for the variance of estimators of this class and the variance of optimum estimators of the class are obtained.

2. The Bias and Variance of r

To find the bias and variance of r, we write

$$\epsilon_1 = (s_x^2/S_x^2) - 1,\ \epsilon_2 = (s_y^2/S_y^2) - 1$$

and

$$\epsilon_3 = (s_{xy}/S_{xy}) - 1,$$

then

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

and

$$E(\epsilon_1^2) = \frac{(N-n)}{(N-2)n}\left(K\,\frac{\mu_{40}}{\mu_{20}^2} - M\right) \tag{2.1}$$

$$E(\epsilon_2^2) = \frac{(N-n)}{(N-2)n}\left(K\,\frac{\mu_{04}}{\mu_{02}^2} - M\right) \tag{2.2}$$

$$E(\epsilon_3^2) = \frac{(N-n)}{(N-2)n}\left\{K\,\frac{\mu_{22}}{\mu_{11}} + \frac{(N-1)(N-n-1)}{(n-1)\,N(N-3)}\,\frac{\mu_{20}\mu_{02}}{\mu_{11}^2} - \frac{(N^2 n - Nn - 2N^2 + 4N - 2n - 2)}{(n-1)\,N(N-3)}\right\} \tag{2.3}$$

$$E(\epsilon_1\,\epsilon_2) = \frac{(N-n)}{(N-2)n}\left\{K\,\frac{\mu_{22}}{\mu_{20}\mu_{02}} + \frac{2(N-1)(N-n-1)}{(n-1)\,N(N-3)}\,\frac{\mu_{11}^2}{\mu_{20}\mu_{02}} - \frac{(N^2 n - 2Nn - N^2 + 2N - n - 1)}{(n-1)\,N(N-3)}\right\} \tag{2.4}$$

$$E(\epsilon_1\,\epsilon_3) = \frac{(N-n)}{(N-2)n}\left(K\,\frac{\mu_{31}}{\mu_{20}\mu_{11}} - M\right) \tag{2.5}$$

$$E(\epsilon_2\,\epsilon_3) = \frac{(N-n)}{(N-2)n}\left(K\,\frac{\mu_{13}}{\mu_{02}\mu_{11}} - M\right) \tag{2.6}$$

where

$$\mu_{ab} = N^{-1}\sum_{i=1}^{n}(x_i - \overline{X})^a\,(y_i - \overline{Y})^b,$$

$$K = \frac{(N-1)(Nn - N - n - 1)}{(n-1)(N(N-3))}$$

and

$$M = \frac{N^2 n - 3N^2 + 6N - 3n - 3}{(n-1)\,N(N-3)}.$$

The expressions in (2.1)-(2.6) could be obtained either directly or by the method given in Kendall and Stuart [2]. The expression in (2.1) is equivalent to the one given by Kendall and Stuart ([2], p. 326, Ex. 12.11).

From (1.2), we have

$$r = \rho(1 + \epsilon_3)\,\{(1 + \epsilon_1)(1 + \epsilon_2)\}^{-\frac{1}{2}}\,. \tag{2.7}$$

Expanding, retaining terms up to the second power of ϵ's and taking expectation, we have to the first degree of approximation

$$E(r) = \rho E\left\{1 - \frac{1}{2}\epsilon_1\epsilon_3 - \frac{1}{2}\epsilon_2\epsilon_3 + \frac{1}{4}\epsilon_1\epsilon_2 + \left(\frac{3}{8}\right)\epsilon_1^2 + \left(\frac{3}{8}\right)\epsilon_2^2\right\}$$

$$= \rho - \frac{(N-n)}{(N-2)} K \frac{\rho}{n}\left\{\frac{1}{2}\left(\frac{\mu_{31}}{\mu_{20}\mu_{11}} + \frac{\mu_{13}}{\mu_{02}\mu_{11}}\right) - \frac{1}{4}\frac{\mu_{22}}{\mu_{20}\mu_{02}} - \frac{3}{8}\left(\frac{\mu_{40}}{\mu_{20}^2} + \frac{\mu_{04}}{\mu_{02}^2}\right)\right\} \tag{2.8}$$

This shows that the bias of r is of the order of n^{-1}, and hence its contribution to the mean square error will be of the order of n^{-2}. Thus, up to terms of order n^{-1}, the expressions for the mean square error and the variance will be same. Onwards, we derive variances up to terms of order n^{-1} only.

Retaining terms up to second power of ϵ's, we obtain the variance of r

$$V(r) = E(r - \rho)^2$$

$$= \rho^2 E\left\{\epsilon_3^2 + \frac{1}{4}\left(\epsilon_1^2 + \epsilon_2^2 + 2\epsilon_1\epsilon_2\right) - (\epsilon_1\epsilon_3 + \epsilon_2\epsilon_3)\right\}$$

Substituting from (2.1)-(2.6), and retaining terms up to order n^{-1}, we have

$$V(r) = \frac{(N-n)}{(N-2)} K \frac{\rho^2}{n} \cdot A \tag{2.9}$$

where

$$A = \frac{\mu_{22}}{\mu_{11}^2} + \frac{1}{4}\left(\frac{\mu_{40}}{\mu_{20}^2} + \frac{\mu_{04}}{\mu_{02}^2} + \frac{2\mu_{22}}{\mu_{20}\mu_{02}}\right) - \left(\frac{\mu_{31}}{\mu_{11}\mu_{20}} + \frac{\mu_{13}}{\mu_{11}\mu_{02}}\right)$$

When N tends to infinity, $K(N-n)/(N-2)$ tends to unity and $V(r)$ is given by $(\rho^2/n)A$, the asymptotic expression for $V(r)$ in infinite populations; see for example, Kendall and Stuart ([2], p.251). The expression $K(N-n)/(N-2)$ could therefore, be considered as the finite population correction term for variance of r.

3. A Class of Estimators of ρ

In sample survey situations, many a times the information on an auxiliary variable is available. In this section we define a class of estimators of ρ when the values of the population mean $\bar{X}$ and the population variance S_x^2 of the variable x, are known.

We write $u = \bar{x}/\bar{X}$ and $v = s_x^2/S_x^2$. Whatever be the sample chosen, let (u, v) assume values in a bounded closed convex subset, R, of the two-dimensional real space containing the point (1, 1), let $t(u, v)$ be a function of u and v such that

$$t(1, 1) = 1 \tag{3.1}$$

and such that it satisfies the following conditions.

1. The function $t(u, v)$ is continuous and bounded in R.
2. The first and second partial derivatives of $t(u, v)$ exist and are continuous and bounded in R.

We consider the class of estimators of ρ, defined by*

$$\tilde{r}_t = r \cdot t(u, v). \tag{3.2}$$

To find the bias and variance of $\tilde{r}_t$ we will need the following expressions which are easily derived

$$E(u - 1) = 0$$

$$E(u - 1)^2 = \frac{(N - n)}{(N - 1)n} \frac{\mu_{20}}{\bar{X}^2} \tag{3.3}$$

$$E\{\epsilon_3(u - 1)\} = \frac{(N - n)}{(N - 2)n} \frac{\mu_{21}}{\bar{X}\mu_{11}} \tag{3.4}$$

$$E\{\epsilon_1(u - 1)\} = \frac{(N - n)}{(N - 2)n} \frac{\mu_{30}}{\bar{X}\mu_{20}} \tag{3.5}$$

$$E\{\epsilon_2(u - 1)\} = \frac{(N - n)}{(N - 2)n} \frac{\mu_{12}}{\bar{X}\mu_{02}} \tag{3.6}$$

To find the expectation and variance of $\tilde{r}_t$, we expand $t(u, v)$ about the point (1, 1) in a second-order Taylor's series, put this value and the value of r in terms of ϵ_1, ϵ_2, ϵ_3 from (2.7) in (3.2). Expanding in powers of ϵ_1, ϵ_2, ϵ_3 and $(u - 1)$ and retaining terms up to second power, we have

$$E(\tilde{r}_t) = \rho + 0(n^{-1})$$

which shows that the bias of $\tilde{r}_t$ is of the order of n^{-1} and so up to order n^{-1}, the mean square error and the variance of $\tilde{r}_t$ are same.

*Srivastava and Jhajj [4], [5] defined classes of estimators of this form for population mean $\bar{Y}$ and population variance S_y^2.

Expanding $(\tilde{r}_t - \rho)^2$, retaining terms up to second power in $\epsilon_1, \epsilon_2, \epsilon_3$ and $(u - 1)$, taking expectation and using (2.1)-(2.6) and (3.3)-(3.6), after some algebra, we obtain

$$V(\tilde{r}_t) = \frac{(N-n)}{(N-2)} \frac{\rho^2}{n}\Big\{KA + B\,t_1(1,1) + KCt_2(1,1) + \left(\frac{N-2}{N-1}\right)\frac{\mu_{20}}{\bar{X}^2}\, t_1^2(1,1) + (K\beta_2 - M)\, t_2^2(1,1) + \frac{2\mu_{30}}{\bar{X}\mu_{20}}\, t_1(1,1)\, t_2(1,1)\Big\} \quad (3.7)$$

where

$$B = \frac{1}{\bar{X}}\left(\frac{2\,\mu_{21}}{\mu_{11}} - \frac{\mu_{12}}{\mu_{02}} - \frac{\mu_{30}}{\mu_{20}}\right),$$

$$C = \frac{1}{\mu_{20}}\left(\frac{2\,\mu_{31}}{\mu_{11}} - \frac{\mu_{40}}{\mu_{20}} - \frac{\mu_{22}}{\mu_{02}}\right),$$

$$\beta_2 = \frac{\mu_{40}}{\mu_{20}^2},\ \beta_1 = \frac{\mu_{30}^2}{\mu_{20}^3},$$

and $t_1(1, 1)$ and $t_2(1, 1)$ respectively denote the first partial derivatives of $t(u, v)$ at the point (1, 1).

Any parametric function $t(u, v)$ satisfying (3.1) and the conditions (1) and (2) can generate an estimator of the class (3.2). Four examples of such functions $t(u, v)$ are given in Srivastava and Jhajj [5].

The optimum values of the parameters in $t(u, v)$ which minimize the variance of $\tilde{r}_t$ are given by

$$\text{Opt. } t_1(1, 1) = \frac{1}{2}\left\{\frac{\bar{X}\mu_{30}}{\mu_{20}^2}\, KC - \frac{\bar{X}^2}{\mu_{20}}\,(K\beta_2 - M)\, B\right\}\Big/\left\{\left(\frac{N-2}{N-1}\right)(K\beta_2 - M) - \beta_1\right\} \quad (3.8)$$

$$\text{Opt. } t_2(1, 1) = \frac{1}{2}\left\{\frac{X\mu_{30}}{\mu_{20}^2}\, B - \left(\frac{N-2}{N-1}\right) KC\right\}\Big/\left\{\left(\frac{N-2}{N-1}\right)(K\beta_2 - M) - \beta_1\right\} \quad (3.9)$$

and the minimum variance is given by

$$\text{Min. } V(\tilde{r}_t) = \frac{(N-n)}{(N-2)}\frac{\rho^2}{n}\left[KA - \frac{1}{4}\left(\frac{N-1}{N-2}\right)\frac{\bar{X}^2}{\mu_{20}}\, B^2 - \frac{1}{4}\,\frac{(N-1)}{(N-2)}\,\frac{\{\bar{X}\mu_{30}/\mu_{20}^2\, B - (N-2)/(N-1)\, KC\}^2}{\{(N-2)/(N-1)(K\beta_2 - M) - \beta_1\}}\right] \quad (3.10)$$

For infinite populations the expressions (3.7)-(3.10) up to terms of order n^{-1} are obtained by taking limit as N tends to infinity. These are

$$V(\tilde{r}_t) = \frac{\rho^2}{n}\left\{ A + Bt_1(1,1) + Ct_2(1,1) + \frac{\mu_{20}}{\bar{X}^2} t_1^2(1,1) + (\beta_2 - 1) t_2^2(1,1) + \frac{2\mu_{30}}{\bar{X}\mu_{20}} t_1(1,1)\, t_2(1,1)\right\} \quad (3.11)$$

$$\text{Opt. } t_1(1,1) = \frac{1}{2}\left\{ \frac{\bar{X}\mu_{30}}{\mu_{20}^2} C - (\beta_2 - 1)\frac{\bar{X}^2}{\mu_{20}} B \right\} \Big/ (\beta_2 - \beta_1 - 1) \quad (3.12)$$

$$\text{Opt. } t_2(1,1) = \frac{1}{2}\left(\frac{\bar{X}\mu_{30}}{\mu_{20}^2} B - C \right) \Big/ (\beta_2 - \beta_1 - 1) \quad (3.13)$$

$$\text{Min. } V(\tilde{r}_t) = \frac{\rho^2}{n}\left\{ A - \frac{1}{4}\frac{\bar{X}^2}{\mu_{20}} B^2 - \frac{(\bar{X}\mu_{30}/\mu_{20}^2\, B - C)^2}{4(\beta_2 - \beta_1 - 1)} \right\} \quad (3.14)$$

It is clearly seen that if optimum values of parameters given by (3.12) and (3.13) are used, the variance of $\tilde{r}_t$ is not larger than that of r as the last two terms on the r.h.s. of (3.14) are non-negative.

If we further assume that the infinite population is bivariate normal, the above four expressions reduce to

$$V(\tilde{r}_t) = \frac{\rho^2}{n}\left\{ \frac{(1-\rho^2)^2}{\rho^2} + 2(1-\rho^2)\, t_2(1,1) + \frac{\mu_{20}}{\bar{X}^2} t_1^2(1,1) + 2t_2^2(1,1) \right\} \quad (3.15)$$

$$\text{Opt. } t_1(1,1) = 0 \quad (3.16)$$

$$\text{Opt. } t_2(1,1) = -\frac{1}{2}(1-\rho^2) \quad (3.17)$$

$$\text{Min. } V(\tilde{r}_t) = \frac{1}{n}(1-\rho^2)^2\left(1 - \frac{1}{2}\rho^2\right) \quad (3.18)$$

The expression on the right hand side of (3.18) clearly gives the amount by which the asymptotic variance $(1-\rho^2)^2/n$ of r is larger than the variance of $\tilde{r}_t$ with optimum values of the parameters. The relative decrease in variance is more for larger values of ρ.

Two simple functions $t(u, v)$ satisfying the required conditions are

$$t(u, v) = 1 + \alpha(u-1) + \beta(v-1) \quad (3.19)$$

$$t(u, v) = u^{\alpha}\, v^{\beta} \quad (3.20)$$

and for both these functions $t_1(1, 1) = \alpha$ and $t_2(1, 1) = \beta$. Thus one should use optimum values of α and β in $\tilde{r}_t$ to get the minimum variance. However, these optimum values of α and β depend upon unknown population parameters and will not be known, so some guessed values for optimum values of α and β may be used.

In the case of bivariate normal population, from (3.16) optimum value of α is zero. And with this value of α, the variance of $\tilde{r}_t$ is

$$V(\tilde{r}_t) = \frac{(1-\rho^2)^2}{n} + 2\beta(\beta + 1 - \rho^2)\,\frac{\rho^2}{n}. \tag{3.21}$$

The variance of $\tilde{r}_t$ at (3.21) is smaller than the variance of r for $-(1-\rho^2) < \beta < 0$. Thus if in the estimator $\tilde{r}_t$, $t(u, v)$ as given in (3.19) or (3.20), is used with $\alpha = 0$ and a value of β in the interval $(-(1-\rho^2), 0)$, it will have a smaller variance than the variance of r. For example this estimator $\tilde{r}_t$ will have a smaller variance than that of r for β lying between $-.19$ and 0 if $\rho = 0.9$, for β lying between $-.51$ and 0 if $\rho = 0.7$ and for β lying between $-.75$ and 0 if $\rho = 0.5$.

At this point it may be interesting to note that if $\bar{X}$ and S_x^2 are known, one may think that the most natural estimator for ρ will be of the form

$$r' = n^{-1} \sum_{i=1}^{n} (x_i - \bar{X})(y_i - \bar{y})/(S_x, s_y).$$

However it is not so as r' has a variance which is not necessarily smaller than that of r. In fact the estimator r' is a member of the class (3.2) with $t(u, v) = v^{\frac{1}{2}}$, and so $t_1(1, 1) = 0$ and $t_2(1, 1) = \frac{1}{2}$. The asymptotic variance of r' is

$$V(r') = \frac{(N-n)}{(N-2)}\frac{\rho^2}{n}\left\{KA + \frac{1}{2}KC + \frac{1}{4}(K\beta_2 - M)\right\},$$

which in the case of bivariate normal population reduces to

$$V(r') = \frac{(1-\rho^2)^2}{n} + \left(\frac{3}{2} - \rho^2\right)\frac{\rho^2}{n}.$$

The variance of r' thus is always larger than that of r in the case of bivariate normal populations.

4. A Wider Class of Estimators

In this section we consider a class of estimators of ρ wider than (3.2) defined by

$$\tilde{r}_g = g(r, u, v), \tag{4.1}$$

where $g(r, u, v)$ is a function of r, u and v such that

$$g(\rho, 1, 1) = \rho.$$

Proceeding as in section 3 it is easily seen that the bias of $\tilde{r}_g$ is of the order n^{-1} and up to this order of terms the variance of $\tilde{r}_g$ is minimized for the following optimum values.

$$\text{Opt. } g_2(\rho, 1, 1) = \frac{\rho}{2}\left\{\frac{\bar{X}\mu_{30}}{\mu_{20}^2} KC - \frac{\bar{X}^2}{\mu_{20}} (K\beta_2 - M)B\right\}\Big/\left\{\left(\frac{N-2}{N-1}\right)(K\beta_2 - M) - \beta_1\right\}$$

and

$$\text{Opt. } g_3(\rho, 1, 1) = \frac{\rho}{2}\left\{\frac{\bar{X}\mu_{30}}{\mu_{20}^2} B - \left(\frac{N-2}{N-1}\right) KC\right\}\Big/\left\{\left(\frac{N-2}{N-1}\right)(K\beta_2 - M) - \beta_1\right\}$$

where $g_2(\rho, 1, 1)$ and $g_3(\rho, 1, 1)$ respectively denote the first partial derivatives of $g(r, u, v)$ with respect to u and v at the point $(\rho, 1, 1)$. The minimum variance of $\tilde{r}_g$ is equal to the minimum variance of $\tilde{r}_t$ as given by (3.10). Thus it is seen that the asymptotic variance for an optimum estimator of class (4.1) is the same as the asymptotic variance for an optimum estimator of class (3.2) and is not reduced.

For the case of bivariate normal population the optimum values of $g_2(\rho, 1, 1)$ and $g_3(\rho, 1, 1)$ reduce to

$$\text{Opt. } g_2(\rho, 1, 1) = 0 \tag{4.2}$$

and

$$\text{Opt. } g_3(\rho, 1, 1) = -\frac{\rho}{2}(1 - \rho^2) \tag{4.3}$$

and the minimum variance of $\tilde{r}_g$ is given by (3.18).

An estimator of ρ of the class (3.2) with $t(u, v)$ as given in (3.19) with optimum values of $t_1(1, 1)$ and $t_2(1, 1)$ for bivariate normal populations is given by

$$\tilde{r}_1 = r\left\{1 - \frac{1}{2}(1 - \rho^2)\left(\frac{s_x^2}{S_x^2} - 1\right)\right\}.$$

This estimator $\tilde{r}_1$ involves ρ which is unknown. If we replace ρ by r in $\tilde{r}_1$, it reduces to

$$\tilde{r}_2 = r\left\{1 - \frac{1}{2}(1 - r^2)\left(\frac{s_x^2}{S_x^2} - 1\right)\right\}.$$

The estimator $\tilde{r}_2$ so obtained is a member of the class (4.1) and satisfies the conditions (4.1) and (4.3), and hence is an optimum estimator of the class with asymptotic variance equal to (3.18).

The maximum likelihood estimator of ρ when $\bar{X}$ and S_x^2 are known is given by

$$\hat{\rho} = r\left\{(1 - r^2)\frac{s_x^2}{S_x^2} + r^2\right\}^{-\frac{1}{2}}$$

which too is a member of the class (4.1) and satisfies conditions (4.2) and (4.3). Hence $\hat{\rho}$ is also an optimum estimator of the class (4.1) with the same asymptotic variance.

REFERENCES

[1] Gupta, J. P., Singh, R. and Lal, B. (1978) : On the Estimation of the finite population correlation coefficient-I, *Sankhya* C, **40** : 38-59.

[2] Kendall, M. and Stuart, A. (1977) : *The Advanced Theory of Statistics*, Vol. I, Charles Griffin and Co., London.

[3] Srivastava, S. K. (1971) : A generalized estimator for the mean of a finite population using multi-auxiliary information, *J. Am. Statis. Assoc.*, **66** : 404-407.

[4] Srivastava, S. K. and Jhajj, H. S. (1980) : A class of estimators using auxiliary information for estimating finite population variance, *Sankhya* C, **42** : 87-96.

[5] 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.

[6] Sukhatme, P. V. and Sukhatme, B. V. (1970) : *Sampling Theory of Surveys with Applications*. New Delhi : Asia Publishing House.