# SOME FURTHER RESULTS ON ERRORS IN DOUBLE SAMPLING TECHNIQUE*

*By* CHAMELI BOSE
*Statistical Laboratory, Calcutta*

1. The author (1942, 1943) studied some types of double sampling technique which consists in obtaining an expression for a character $y$, sometimes difficult or uneconomic to measure directly, in terms of an appreciably correlated character $x$, easier to obtain. Such problems were studied also by Neyman (1938), and Cochran (1939); but their methods of approach were different. Snedecor and King (1942) obtained a formula for the variance of the forecasting equation for a situation which is closely similar to a case that has been called Type (1) double sampling technique by the present author (1943). In this method of sampling, of which extensive use has been made in U.S.A., repeated samples are considered in both the stages. Type (2) double-sampling technique consists in a single sampling in the first stage, and repeated sampling in the second stage whereas, Type (3) constitutes repeated sampling in the first stage and a unique sample in the second stage.

The adequate error formulae for determining the accuracy of the estimates in all these types of sampling were derived by the present author in the above mentioned works. In the present paper some additional formulae for the error of the forecasting equations have been investigated for the cases where the character $x$, the concomitant variate, is kept constant (*i*) in both the stages of all the types of sampling and (*ii*) in the first stage only, but will vary in the second stage according to the conditions of the sampling type.

2. Let $x'_1, x'_2, \ldots, x'_N$ be the values of the character $x$ in the second stage of sampling and $a$ and $b$ be the regression constants estimated from the $n$ sets

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$

of the first stage. Then

$$\bar{Y}' = a + b\bar{x}' = \bar{y} + b(\bar{x}' - \bar{x})$$

where $\bar{x}' = (x'_1 + x'_2 + \ldots + x'_N)/N$, is the graduated mean of the character $y$.

The variance of $\bar{Y}'$ in the three different types for the cases (i) and (ii) are given below:

*Case* (*i*):

Type (1) : Varying over both the stages, *i.e.* at both the stages of sampling $x$'s are fixed but $y$'s are allowed to vary.

$$\sigma_{\bar{Y}'}^2 = \sigma^2_{y.x}\left[\frac{1}{n} + \frac{(\bar{x}' - \bar{x})^2}{\Sigma(x - \bar{x})^2}\right]$$

with $E(\bar{Y}') = \alpha + \beta\bar{x}'$.

*The results on the errors of all the estimating functions given in this paper were presented before the thirty-third session on the Indian Science Congress at Bangalore in 1940.

Type (2) : Varying over the second stage only, *i.e.* unknown values of $y$ in the second stage are only allowed to vary and $E\ (\hat{Y}')$ becomes same as $\hat{Y}'$ which is $a+b\bar{x}'$ and naturally variance becomes nil.

Type (3) : Varying over the first stage only, *i.e.* $x$'s are fixed at both the stages of sampling but $y$'s vary only at the first stage of sampling.

$$\sigma_{\hat{Y}'}{}^{2} = \sigma^{2}{}_{y\cdot x}\left[\frac{1}{n}+\frac{(\bar{x}'-\bar{x})}{\Sigma(x-\bar{x})^{2}}\right]$$

with $E\ (\hat{Y}') = \alpha+\beta\bar{x}'$. The variance of Type (3) of case (i) is same as that of Type (1) of case (i), since unknown values of $y$ in the second stage of sampling do not enter in the variance formula.

*Case (ii)*:

Type (1) : Varying over both the stages, *i.e.* $x$'s in the first stage only are fixed and others are allowed to vary.

$$\sigma^{2}_{\hat{Y}'} = \sigma^{2}{}_{y\cdot x}\left[\frac{1}{n}+\frac{(\bar{x}-\xi)^{2}}{\Sigma(x-\bar{x})^{2}}\right]+\left[\frac{\sigma^{2}{}_{y\cdot x}}{\Sigma(x-\bar{x})^{2}}+\beta^{2}\right]\ \sigma^{2}_{\bar{x}'}$$

with $E\ (\hat{Y}') = \alpha+\beta\xi$.

Type (2) : Varying over the second stage only, *i.e.* $x$'s and $y$'s in the second stage of sampling only are allowed to vary

$$\sigma_{\hat{Y}'}{}^{2} = \left(\frac{s_y}{s_x}\ r\right)^{2}\ \sigma^{2}_{\bar{x}'}$$

with $E\ (\hat{Y}')=a+b\xi$. The variance in this case is same as that of the original Type (2) technique since the resultant effect on $x$'s and $y$'s due to restrictions becomes same.

Type (3) : Varying over the first stage only, *i.e.* $x$'s are fixed at both stages and $y$'s are allowed to vary at the first stage only.

$$\sigma^{2}_{\hat{Y}'} = \sigma^{2}{}_{y\cdot x}\left[\frac{1}{n}+\frac{(\bar{x}'-\bar{x})^{2}}{\Sigma(x-\bar{x})^{2}}\right]$$

with $E\ (\hat{Y}') = \alpha+\beta\bar{x}'$. The variance in this case is the same as that of Type (1) and Type (3) of case (i).

3. To measure the goodness of the forecasting equation, the expectation of the square of the discrepancy between the true sample mean $\bar{y}'$ (with the notation of the author's previous works), which is unknown, corresponding to the sample mean value of the character $x$ in the second stage of sampling and the estimated mean value $\hat{Y}'$ for three different types of sampling technique have been worked out (Bose, 1943). The expectation of the square of the discrepancy will serve the purpose of deciding which of the concomitant variates or which particular combination of them should be chosen for furnishing the best forecasting formula. Usually that variate or combination of variates will be chosen for which discrepancy will be minimum at certain cost.

Let $y'_1, y'_2, \ldots, y'_N$ be the unknown values of $y$ (in the second stage of sampling) corresponding to $x'_1, x'_2, \ldots, x'_N$ and $\bar{y}' = (y'_1+y'_2+\ldots+y'_N)/N$ then the discrepancy between $\bar{y}'$ (the true sample mean of the character $y$) and $\Upsilon'$ (the graduated mean value of the character $y$) is given by $\{\bar{y}'-(a+b\bar{x}')\}$. The expectation of the square of the discrepancy, *i.e.* $E\{\bar{y}'-(a+b\bar{x}')\}^2$ for the three different types of sampling with the two special restrictions (i) and (ii) are given below :

*Case* (*i*):

Type (1) : Varying over both the stages

$$E(\bar{y}'-\Upsilon')^2 = \sigma^2_{y\cdot x}\left[\frac{1}{N}+\frac{1}{n}+\frac{(\bar{x}'-\bar{x})^2}{\Sigma(x-\bar{x})^2}\right]$$

Type (2) : Varying over the second stage only

$$E(\bar{y}'-\Upsilon')^2 = \frac{\sigma^2_{y\cdot x}}{N}+\left[(\bar{y}-\alpha-\beta\bar{x}')+(\bar{x}'-\bar{x})\frac{s_y}{s_x}\,r\right]^2$$

Type (3) : Varying over the first stage only

$$E(\bar{y}'-\Upsilon')^2 = \sigma^2_{y\cdot x}\left[\frac{1}{n}+\frac{(\bar{x}'-\bar{x})^2}{\Sigma(x-\bar{x})^2}\right]+\{\bar{y}'-E(\Upsilon')\}^2$$

For the above three cases, it is assumed that the conditional distribution of $y$ on $x$ will remain same at both the stages of sampling.

*Case* (*ii*):

Type (1) : Varying over both the stages

$$E(\bar{y}'-\Upsilon')^2 = \sigma^2_{y\cdot x}\left[\frac{1}{n}+\frac{(\bar{x}-\xi)^2}{\Sigma(x-\bar{x})^2}\right]+\left[\frac{\sigma^2_{y\cdot x}}{\Sigma(x-\bar{x})^2}+\beta^2\right]\sigma^2_{\bar{x}'}+\sigma^2_{\bar{y}'}(1-2\rho^2)$$

Type (2) : Varying over the second stage only

$$E(\bar{y}'-\Upsilon')^2 = \sigma^2_{\bar{y}'}+[(\bar{y}-\alpha-\beta\xi)-\frac{s_y}{s_x}\,r\,(\bar{x}-\xi)]^2+\left(\frac{s_y}{s_x}\,r\right)^2\sigma^2_{\bar{x}'}-2\frac{s_y}{s_x}\,r\,\frac{\rho\sigma_y\sigma_x}{N}$$

Type (3) : Varying over the first stage only

$$E(\bar{y}'-\Upsilon')^2 = \sigma^2_{y\cdot x}\left[\frac{1}{n}+\frac{(\bar{x}'-\bar{x})^2}{\Sigma(x-\bar{x})^2}\right]+[\bar{y}'-E(\Upsilon')]^2$$

Expressions for discrepancies in the cases (i) and (ii) for Type (3) are same, as the sampling procedures become identical.

4. The instructive feature of the application of Type (1) sampling technique (Bose, 1943) is that it does not add any extra information about the population mean of $y$ unless the information about the concomitant variate $x$ obtained in the second stage of sampling is greater than that of the first stage of sampling where the

measurements of both the characters are taken. Original Type (2) sampling technique can be applied with advantages in those situations where second stage sampling is done on a different population or on a set of populations having the same physical relationship between characters $y$ (difficult to measure) and $x$ (easier to measure) as that of the first stage of sampling but different mean levels at different stages of sampling. Such a situation arose in connection with estimating cinchona bark from a knowledge of physical measurements such as height, girth, number of stems, surface area, thickness of the bark etc. of the living cinchona plants at different places in the same year as well as in different years for the same places where in actual practice it was found that though the mean values of physical measurements as well as yield of bark varying considerably, the physical relationship between the characters remained remarkably stable at different points of possible variations, thus suggesting the possibility of constructing a single forecasting function which is convenient as well as economic to apply. If, however, there is any doubt about the stability of regression equation at different points of possible variations, then Type (1) sampling technique, *i.e.* repetitions of both the stages of sampling on every occasion should be used for checking the continued stability of the physical relationship between the characters $x$ and $y$.

### References

Bose, Chameli (1942): The variance of the forecasted mean value subjecting to two-way fluctuations. *Science and Culture*, 7, 514.

————(1943): Note on the sampling error in the method of double sampling. *Sankhyā*, 6 (3), 329-330.

————(1946): On the sampling error in the method of double sampling. Proc. Indian Science Congress, thirty-third session, Bangalore, Part 3, 11-12.

Cochran, W. G. (1939): The use of the analysis of variance in enumeration by sampling. *J. Amer. Stat. Assocn.* 34, 492-510.

Neyman, J. (1938): Contribution to the theory of sampling human population. *J. Amer. Stat. Assocn.* 33, 101-116.

Snedecor, G.W. and King, Arnold K. (1942): Recent developments in sampling for agricultural statistics. *J. Amer. Stat. Assocn.* 37, 95-102.

*Paper received: September, 1950.*