NOTE ON THE SAMPLING ERROR IN THE METHOD OF DOUBLE SAMPLING

BY CHAMELI BOSE

In problems of sampling, the population mean value $\bar{y}$ of a character y can be estimated with a certain order of precision from a sample drawn at random from a population (say σ). It may be sometimes difficult or uneconomical to measure directly the mean character y from a large sample. However, in this situation the sample mean value of the character y as an estimate of population mean value $\bar{y}$ can be estimated with some margin of error from the knowledge of a second character x provided there is significant correlation between the two characters x and y.

This requires the evaluation of the regression equation of y on x in a preliminary or exploratory experiment, here called the first stage; and the ultimate use of the same relationship in estimating the sample mean value of the character y from a second sample consisting of the measurements of the character x only.

From the procedure outlined above, it is quite clear, that the error of estimate of the sample mean value of the character y will arise from two sources. There is error in the predicting equation and mean value of the character x. Firstly the parameters of the predicting equation and secondly the observed mean value of the character x (on which is based the estimated or predicted sample mean value of character y) are subject to sampling fluctuations; and any measure of reliability of the predicted value must therefore take note of these two errors.

The present problem may therefore be looked upon as a procedure of double sampling involving (1) the estimation of the regression equation from the first or 'exploratory stage' of sampling, and (2) the estimation of the mean value of the character x from the second or 'survey stage'. The question has already been discussed in an earlier note (Science & Culture, 7, 1941-42, 514). Three methods immediately suggest themselves in which this double sampling technique may be carried out.

Type (1) : Repeated sampling in the 'exploratory stage' would give us fresh sets of regression equations which can be ultimately used with different estimated mean values of the character x determined from several samples in the 'survey stage'.

Type (2) : A single sampling in the 'exploratory stage' gives us a uniquely estimated regression equation, which will be used with the different values of the estimated mean value of the character x obtained from different samples in the 'survey stage'.

Type (3) : Finally the estimated mean value of the character x may be determined once for all, and different estimates of the mean value of the character y obtained by using several regression equations calculated from several samples in the 'exploratory stage'. This is probably more of academic than practical interest.

The variance of the estimate $\bar{Y}'$ of the sample mean value of the character y made from two stages of sampling procedure (the first stage giving the estimates of the parameters and the second stage the estimate of the population mean value ξ of the character x) has been worked out for different types of sampling.

Let $x'_1, x'_2 \ldots, x'_N$ be the values of the character x in the 'survey stage' and 'a' and 'b' the constants estimated from the 'exploratory stage'. Then the estimate $\bar{Y}'$ of the sample mean value of the character y will be $a + b\bar{x}'$ where

$$\bar{x}' = (x'_1 + x'_2 + \ldots + x'_N)/N$$

The variance of the estimate $\bar{Y}'$ for different types of sampling is given below.

Type (1) : Varying over both the stages

$$\sigma^2_{\bar{y}'-\bar{Y}'} = \sigma^2_y \left[\frac{1}{n}(1-2\rho^2) + \left\{ \frac{1+(n-4)\rho^2}{n-3} \right\} \left(\frac{1}{N} + \frac{1}{n} \right) \right] \quad .. \quad (4\cdot1^*)$$

$$= \frac{\sigma^2_y}{n} \frac{(n-2)}{(n-3)}(1-\rho^2) + \frac{\sigma^2_y}{N(n-3)}\{1+\rho^2(n-4)\} \quad .. \quad (4\cdot2)$$

Type (2) : Varying over the second stage only

$$\sigma^2_{\bar{y}'} = \frac{1}{N}\left(r \frac{\sigma_x}{\sigma_y} \sigma_y \right)^2 \quad .. \quad (5\cdot0)$$

Type (3) : Varying over the first stage only

$$\sigma^2_{\bar{y}'-\bar{Y}'} = \frac{\sigma^2_y}{(n-3)}(1-\rho^2)\left[\frac{n-2}{n} + \frac{(\bar{x}'-\xi)^2}{\sigma^2_x} \right] \quad .. \quad (3\cdot0)$$

where n is the size of the sample in the 'exploratory stage' and N the size of the sample in the 'survey stage'.

If $y'_1, y'_2, \ldots, y'_N$ be the unknown values of y (in the survey stage) 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 $\bar{Y}'$ (the estimated sample mean of the character y) will be $(\bar{y}' - (\alpha + b\bar{x}'))$. The expectation of the square of the discrepancy *i.e.* $\varepsilon\{\bar{y}' - (\alpha + b\bar{x}')\}^2$ for the different types of sampling are given below.

Type (1) : Varying over both the stages

$$\varepsilon(\bar{y}' - \bar{Y}')^2 = \sigma^2_y \left(\frac{1}{N} + \frac{1}{n} \right) \left\{ \frac{(n-2)+(n-4)\rho^2}{n-3} - 2\rho^2 \right\} \quad .. \quad (4\cdot1^*)$$

$$= \sigma^2_y \left(\frac{1}{N} + \frac{1}{n} \right) \frac{(n-2)}{(n-3)}(1-\rho^2) \quad .. \quad (4\cdot2)$$

Type (2) : Varying over the second stage only

$$\varepsilon(\bar{y}' - \bar{Y}')^2 = \frac{1}{N}\{\sigma^2_y + (r\frac{\sigma_x}{\sigma_y}\sigma_y)^2 - 2r\frac{\sigma_x}{\sigma_y}(\rho\,\sigma_y\,\sigma_y)\} + \{(\bar{y}-\eta) - r\frac{\sigma_y}{\sigma_x}(\bar{x}'-\xi)\}^2 \quad .. \quad (5\cdot0)$$

Type (3) : Varying over the first stage only

$$\varepsilon(\bar{y}' - \bar{Y}')^2 = \{(\bar{y}-\eta) - \rho\frac{\sigma_y}{\sigma_x}(\bar{x}'-\xi)\}^2 + \frac{\sigma^2_y}{\sigma^2_x} \cdot \frac{(\bar{x}'-\xi)^2(1-\rho^2)}{n-3} + \frac{(n-2)(1-\rho^2)}{n(n-3)}\sigma^2_y \quad .. \quad (6\cdot0)$$

It is often convenient to estimate the yield of a crop indirectly from observations on more easily measured but correlated characters. For example, in the case of cinchona bark from a knowledge of physical measurements (such as height, girth etc) of living plants ; or in the case of jute fibre from weight of green plants etc. In such cases if the estimating equation is first determined on a sufficiently broad basis then the mean yield can be forecasted in different years (or seasons) or for different places in the same year (if the same relation between the characters holds for different places) by measuring the character x only in the 'survey stage'. How far the predicted values differ from the population values may be then obtained from equation (3). On the other hand if experiments on a sufficiently wide scale can not be carried out in the exploratory stage or if the estimating equation itself is likely to vary from season to season or from place to place then both the stages of sampling would have to be repeated on every occasion and the appropriate estimate of the discrepancy of the predicted values would be given by equation (4). In these problems the discrepancy, naturally, has more practical importance as the variance of y (given by equations (2) and (1) respectively) merely supplies an indirect estimate of the variance of x, and has no special significance.

My thanks are due to Prof. P. C. Mahalanobis, who used the method of double sampling in 1940-41 in estimating the yield of cinchona bark in the Government Cinchona Plantation at Mungpoo, Bengal for allotting the statistical analysis to me, and to Mr. S. N. Roy for his kind help and interest in my work.

* These are the revised forms of equations given in the earlier note in *Science & Culture*, 7, 1941-42, 514