Note on the use of order statistics

Kono, Kazumasa
Mathematical Institute, Kyushu University

NOTE ON THE USE OF ORDER STATISTICS

By K. KÔNO

(Mathematical Institute, Kyushu University)

1. Introduction. It is the purpose of this paper to give sampling procedure for a mean value of x of a population, which makes use of order statistics of an auxiliary variable y in the same population and which is not only more accurate than the simple random sampling procedure, but also more simple than the other sampling ones. On the following sections, we shall state its procedure under the assumption that (x, y) is distributed in normal bivariate population.

2. Sampling procedure. Let the correlation of x and y be sufficiently high and let us assume that a cost per unit for y is cheaper than that for x. Furthermore let us suppose that the regression curve of x and y is linear, and that the population distribution has a symmetric form with respect to the origin $(0, 0)$. In this case, our procedure consists of the following steps:

(i) *Draw a random sample of N for y from the population.*

(ii) *Divide this sample at random into λ samples of size n, and in each class arrange n values by the order of the magnitude of the values of y.*

(iii) *Enumerate two values $x_{i(l)}^{(n)}$ and $x_{(n-i+1)(l)}^{(n)}$ corresponding to the i-th and $(n-i+1)$-th of the order statistics of y in the l-th sample.*

The estimator is now defined by

$$(1) \qquad F = \frac{1}{\lambda} \sum_{l=1}^{\lambda} \frac{1}{2} [x_{i(l)}^{(n)} + x_{(n-i+1)(l)}^{(n)}].$$

Our estimator will yield us unbiased estimate, since (x, y) is distributed symmetrically. Now let us denote by $\hat{x}_y = by$ the linear regression equation and put $x = by + \varepsilon$. Next let $\bar{y}_{i(l)}^{(n)}$ and $\bar{y}_{(n-i+1)(l)}^{(n)}$ be the population means of the *i-th* and the $(n-i+1)$-*th* order statistics from the n units in the *l*-th samples and let $\tilde{x}$ be the population mean which is to be estimated. By this definition, it will be readily observed that the variance of this estimator is given by

$$(2) \; \text{Var.}\{ F \} = \frac{1}{4\lambda} \left\{ b^2 [\sigma^2\{\bar{y}_i^{(n)}\} + \sigma^2\{\bar{y}_{n-i+1}^{(n)}\}] + \sigma^2\{\varepsilon_i\} + \sigma^2\{\varepsilon_{n-i+1}\} \right\}.$$

3. The variance of our estimator, under the assumption that (x, y) is distributed in normal bivariate distribution.

Let us now assume that (x, y) is distributed in a bivariate normal distribution. By the assumption, we shall have that

(3) $\left\{\begin{array}{l} (1°) \;\; y \text{ is distributed according to } N(0,\ \sigma_y{}^2). \\ (2°) \;\; \varepsilon \text{ is distributed according to } N(0,\ \sigma_x{}^2(1-\rho^2)). \\ (3°) \;\; b = \rho\,\dfrac{\sigma_x}{\sigma_y}\,. \end{array}\right.$

Let us denote by c_1 and c_2 the costs per a sampling unit for the value of x and that of y respectively, and by C the total cost. Then n and λ must be subject to the restriction

(4) $$C = n\lambda c_2 + 2\lambda c_1.$$

From (1), the variance of F, which is denoted by V_0, becomes approximately

(5) $$V_0 = \frac{1}{2\lambda}\,\sigma_x{}^2\left[1 - \rho^2\Big(1 - \frac{1}{n}D\Big)\right],$$

where D is defined by

$$D = \frac{(n-i+1)}{(n+1)}\,\frac{1}{[\varphi(t_i^{(n)})]^2},$$

with $t_i^{(n)}$ which means the population mean of the *i-th* order statistics of samples of n drawn from $N(0,1)$ and with $\varphi(t)$ which denotes the density function of $N(0,1)$.

Now we shall consider what values of n and D, satisfying (4), yield the greatest accuracy in estimating the mean value of X.

Let us substitute (4) in (6)., then we have

(6) $$V_0 = \frac{c_1}{C}\,\sigma_x{}^2\Big(1 + n\frac{c_2}{2c_1}\Big)\Big\{1 - \rho^2\Big(1 - \frac{1}{n}D\Big)\Big\}.$$

For fixed values of D, the optimum value of n, which assure the greatest accuracy, becomes approximately

(7) $$n \doteqdot \sqrt{\frac{\rho^2 D}{1 - \rho^2}\cdot\frac{2c_1}{c_2}}\,,$$

(Note) If $0.22 < (n-i-1)/(n-1) < 0.3$, the minimum value of D is less than 2.5.

Let us substitute (8) in (7), then we have the following approximate formulae

(8) $$V_0 = \frac{c_1}{C}\,\sigma_x{}^2\Big(1 + \sqrt{\frac{\rho^2 D}{1-\rho^2}\cdot\frac{2c_1}{c_2}}\cdot\frac{c_2}{2c_1}\Big)\Big[1 - \rho^2\Big(1 - \sqrt{\frac{1-\rho^2}{\rho^2 D}\cdot\frac{c_2}{2c_1}}\cdot D\Big)\Big],$$

which is the variance of our optimum procedure.

4. The comparison of our method with a simpe random sampling procedure.

For this purpose, we shall consider a simple random sampling procedure

under a given total cost C. Its sample size must be equal to C/c_1 and its variance will become

$$(9) \qquad\qquad V_r = \frac{c_1}{C}\,\sigma_x{}^2.$$

Moreover, we shall seek the condition under which the estimate based on our procedure is more accurate than that depending upon the simple random sampling one.

$$(10) \qquad \left(1+\sqrt{\frac{\rho^2 D}{1-\rho^2}\cdot\frac{2c_1}{c_2}}\cdot\frac{c_2}{2c_1}\right)\left[1-\rho^2\left(1-\sqrt{\frac{1-\rho^2}{\rho^2 D}\cdot\frac{c_2}{2c_1}}\cdot D\right)\right]<1,$$

which will reduce to the following inequality:

$$(11) \qquad\qquad \frac{c_2}{c_1}<2\,\frac{(1-\sqrt{1-\rho^2})^2}{\rho^2 D}.$$

EXAMPLE. (i) when $\rho=1$, the condition (12) becomes that $c_2/c_1<2/2.5$.

(ii) when $\rho=1/2$, the condition (12) becomes that $c_2/c_1<1/15$.

5. Summary. Our procedure appealing to the use of the order statistics is very useful, under the circumstances where the following conditions are satisfied;

(i) The regression curve is linear.

(ii) There are the available imformation of distribution form.

(iii) The correlation of x and y is near 1.

(iv) A cost for y is less than 1/10 of a cost for x.

Our method may be perhaps cheaper than the ordinary double sampling such as make use of ratio, and regression estimate.

Reference

(1) F. MOSTELLER, "On some useful 'inefficient' statistics". *Ann. Math. Stat.* 17. (1946), p. 377.