Note on the double sampling method

Kono, Kazumasa
Mathematical Institute, Kyushu University

https://doi.org/10.5109/12949

出版情報：統計数理研究. 4 (1/2), pp.36-38, 1950-12. Research Association of Statistical Sciences
バージョン：
権利関係：

NOTE ON THE DOUBLE SAMPLING METHOD

By K. KÔNO.

(Mathematical Institute, Kyushu University, Hukuoka)

1. Introduction. In 1938, the theory of double sampling method applicable for the case where the second character may be used for stratification was first given by. J. NEYMAN[1]. However, in spite of his theorems on the stratified double sampling method, we have still a problem. The problem is to ascertain the condition under which the estimate based on the stratified double sampling method, for a given total expenditure, may be more accurate than that based on the simple random sampling method. The purpose of this paper is to give a sufficient condition under the assumption that the original character and the second one are distributed in a normal bivariate population.

2. Notation and Neyman's results. Denote by X the original character the average of which, say $\bar{X}$, is to be estimated, and by Y the second one, on which the collection of data is cheap, and which is assumed to be correlated with X. The procedure of the double stratified sampling method is as follows. We wish to stratify the population into a number of the classes according to the value of Y. The first sample is a random sample of size N, and $\hat{p}_i=N_i/N$ is the proportion of Y values found in the i-th stratum. Thus $\hat{p}_i$ is an estimate of the true proportion p_i. The second sample is a stratified random sample in which Y is measured: m_i units are drawn at random from the i-th stratum, where m_i is such that $m_i/m_0=\hat{p}_i$, $(i=1, 2,\ldots, S)$, hence

$$m_0=\sum_{i=1}^{S} m_i, \tag{1}$$

where S is the number of strata. Finally, let us denote by A and B the costs per a sampling unit for enumerating the values of X and those of Y respectively. Let us denote by C the total amount of expenditure available for the collection of data. Then the number m_0 and N must be subject to the restriction

$$Am_0 + BN = C. \tag{2}$$

Let X_{ij} denote the *j-th* value of X drawn from the i-th stratum, and let put

$$X_i=\frac{1}{m_i}\sum_{j=1}^{m_i} X_{ij}\,. \tag{3}$$

We shall consider a linear function F of the observation, with coefficients $\hat{p}_i=N_i/N$, that is,

$$F=\sum_{i=1}^{S}\hat{p}_iX_i. \tag{4}$$

J. NEYMAN[1] proved the following

THEOREM 1. *F is the best unbiased estimate of $\bar{X}$, whose variance V_d is given by*

$$V_d=\sum_{i=1}^{S}[\{p_i^2+(p_iq_i)/N\}(\sigma_i^2/m)+\{(p_iq_i)/N\}X_i^2]-(2/N)\sum_{i=1}^{S-1}\sum_{j=i+1}^{S}p_ip_jX_iX_j, \tag{5}$$

where σ_i means the population variance of X in the i-th stratum, and p_i is the ratio of the i-th stratum per whole population, and $q_i=1-p_i$.

THEOREM 2. *When we determine m_i such that,*

$$m_i=m_0\,\sigma_i\sqrt{p_i^2+(p_iq_i)/N}\,/\sum_{i=1}^{S}\sigma_i\sqrt{p_i^2+(p_iq_i)/N}\quad(i=1,2,\ldots,S). \tag{6}$$

then we have the best unbiased estimate, and its variance V_d, becomes

$$V_d=\frac{1}{m_0}[\sum_{i=1}^{S}\sigma_i\sqrt{p_i^2+(p_iq_i)/N}\,]^2(1/N)\sum_{i=1}^{S}p_i(X_i-\bar{X})^2. \tag{7}$$

This is the minimum value of V_d with respect m_i $(i=1,2,\ldots,s)$. Moreover, J. NEYMAN found the system of m_i m_0 and N under a given total cost C that assures the greatest accuracy in estimating $\bar{X}$. His study has finished at this step. Now, we shall seek at following section the condition under which the estimate based on this method is more accurate than that depends upon the simple random sampling method under a given cost.

3. The comparison with the simple random sampling method.

In order to make clear the contribution to an accuracy by the use of the correlation of X and Y and that of the method of stratification, we shall assume that (X,Y) is distributed in a normal bivariate distribution. Let us denote σ_2^2 and $\sigma_2^2(i)$ the population variance of Y and that of Y in the *i-th* stratum respectively, by σ_1^2 the population variance of X, and by ξ_2 and τ_i the population mean of Y and that of Y in the i-th stratum. respectively.

THEOREM 3. *Under the hypothesis to Theorem* 1, *let us assume that (X,Y) distributed in a normal bivariate distribution and that N is sufficiently large. Then we have the following approximation,*

$$V_d\doteqdot\frac{\sigma_1^2}{m_0}[1-\{(N-m_0)/N\}\cdot(\rho^2/\sigma_2^2)\sum_{i=1}^{S}p_i(\tau_i-\xi_2)^2 \tag{8}$$
$$+(S/N)(1-\rho^2)+\sum_{i=1}^{S}\rho^2\sigma_2^2(i)/N\sigma_1^2],$$

where ρ is the correlation coefficient of X and Y. Moreover, if $S\ll N$, $1/N\ll p_i$, $(i=1,2,\ldots,S)$ then we shall have

$$V_d\doteqdot(\sigma_1^2/m_0)[1-\{(N-m_0)/N\}(\rho^2D)] \tag{9}$$

where $D=(1/\sigma_2^2)\sum_{i=1}^{S} p_i(\tau_i-\xi_i)^2$.

Here we shall notice that on each of the following three trivial cases where the double stratified random sampling method has no advantage over the simple random one: (i) $\rho=0$; (ii) $D=0$; (iii) $m_0 \geqq N$.

We shall further observe

THEOREM 4. *Unless* $\rho>0$, $D>0$, $N>m_0$, *the stratified double sampling method has no advantages over the simple random sampling one. If* $\rho>0$, $D>0$, *and* $N>m_0$, *then the following* m_0 *and* N *assure the greatest accuracy in estimating* X *under a given cost;*

$$N=\frac{C\sqrt{\rho^2 D}}{B\sqrt{\rho^2 D}+\sqrt{1-\rho^2}\sqrt{AB}}$$

$$m_0=\frac{C\sqrt{1-\rho^2 D}}{A\sqrt{1-\rho^2 D}+\sqrt{\rho^2 DAB}}.$$

Examples. The values of D for a normal distribution $N(0,1)$ corresponding to the following various divisions will be of some interest in choosing the strata for y. It is to be noted that, (i) the more the division points are dense, generally the greater the values of D become; but (ii) for a given size of N, the estimates $\hat{p}_i$ of p_i becomes more inaccurate.

(1°) When the interval $-3\leqq y\leqq 3$ is divided into six (twelve) subintervals of equal length, then the division yields us, with the outside intervals $(-\infty, -3)$ and $(3, \infty)$, that $D=0.92$ (0.98).

(2°) Let the whole real y-axis is divided into eight (fourteen) subintervals each of which contains the same probability for the normal distribution $N(0,1)$, then $D=0.96(0.99)$.

Reference

(1) J. NEYMAN; "Contribution to the theory of sampling human population" *Journ. Amer. Stat. Assoc.* vol. 34. (1938).