On Optimum Balancing Between Sample Size and Number of Strata in Sub-Sampling

By Yasushi TAGA

(Received January 5, 1953)

1. As to the effects of stratification in sampling procedure there are various arguments made to the single-stage sampling procedure, but it seems to me, that there are few to the sub-sampling procedure. In this article we shall give the limits of effects of stratifications in subsampling to some special cases and show how to determine optimally sample-size and number of strata. The results will be of service to the treatment of the general case.

2. Given a population π, divide it into R strata and draw a sample of size n from π by the sampling method with probabilities proportionate to sizes of the primary sampling units. Then the variance of the sample mean $\bar{x}$ is represented approximately as follows

$$\sigma_{\bar{x}}^2 = \sum_{i=1}^{R} p_i^2 \left(\frac{\sigma_{wi}^2}{n_i} + \sigma_{bi}^2 \right) \tag{1}$$

where n_i, p_i, σ_{wi}^2, σ_{bi}^2 are the sample-size, the weight of the i-th stratum, the within-and between-variance in the i-th stratum respectively. Further, assume that the sample is allocated to every stratum proportionately to its size.[*] Then the variance of $\bar{x}$ becomes

$$\sigma_{\bar{x}}^2 = \frac{1}{n} \sum_{i=1}^{R} p_i \sigma_{wi}^2 + \sum_{i=1}^{R} p_i^2 \sigma_{bi}^2 \tag{2}$$

where

$$\sigma_{wi}^2 = \sum_{j=1}^{Mi} \frac{N_{ij}}{N_i} \sigma_{ij}^2 \tag{3}$$

$$\sigma_{bi}^2 = \sum_{j=1}^{Mi} \frac{N_{ij}}{N_i} (\bar{X}_{ij} - \bar{X}_i)^2 \tag{4}$$

$$\sigma_{ij} = \frac{1}{N_{ij}} \sum_{k=1}^{Nij} (X_{ijk} - \bar{X}_{ij})^2 \tag{5}$$

$$\bar{X}_i = \sum_{j=1}^{Mi} \frac{N_{ij}}{N_i} \bar{X}_{ij}$$

[*] This method (size proportionate allocation) is often used in practical surveys, for the benefit of counting and analysis. Therefore this limitation will not be so serious.

$$\overline{X}_{ij} = \frac{1}{N_{ij}}\sum_{k=1}^{N_{ij}} X_{ijk}$$

X_{ijk}: the attribute of the k-th secondary sampling unit of the j-th primary sampling unit in the i-th stratum.

Since $\frac{N_i}{N} = p_i$, $\sigma_{\bar{x}}^2$ becomes

$$\sigma_{\bar{x}}^2 = \frac{1}{n}\sigma_w^2 + \sum_{i=1}^{R}\left(\frac{N_i}{N}\right)^2\sigma_{bi}^2 \tag{6}$$

where σ_w^2 denotes $\sum_{i=1}^{R}\sum_{j=1}^{M_i}\frac{N_{ij}}{N}\sigma_{ij}^2$.

Once primary sampling units are determined, the first term of this expression depends only on the sample size n and the second term only on the method of stratification. Therefore, stratification has the effects only to the control of between-variances σ_{bi}^2's. Now, we introduce the distribution function $F(x)$ of means $\overline{X}_{ij}$ of the primary sampling units, which is given by considering for every $\overline{X}_{ij}$ the weight $\frac{N_{ij}}{N}$. Now, for brevity we confine ourselves to the case where a ratio of individuals having some characteristic in π should be estimated. The general case will be similarly treated.

$$\sigma_w^2 = \int_0^1 x(1-x)dF(x) = \overline{X}(1-\overline{X}) - \sigma^2$$

where $\overline{X} = \int_0^1 xdF(x)$ is the population mean

and $\sigma^2 = \int_0^1 (x-\overline{X})^2dF(x)$ is the variance between primary sampling units in the whole population.

And the variance between the primary sampling units in the i-th stratum is:

$$\begin{aligned}\sigma_{bi}^2 &= \int_{I_i}(x-\overline{X}_i)^2dF_i(x)\\ &= \int_{I_i}x^2dF_i(x) - \overline{X}_i^2\end{aligned}$$

where I_i is the interval or the set of intervals, representing the i-th stratum in the line of real numbers, $F_i(x) = F(x)/p_i$, and $\overline{X}_i = \int_{I_i}xdF_i(x)$ is the population mean of the i-th stratum.

Substituting these relations into the formula (6), we have

$$\sigma_{\bar{x}}^2 = \frac{1}{n}\{\bar{X}(1-\bar{X}) - \sigma^2\} + \sum_{i=1}^{R} p^2{}_i \left\{\int_{I_i} x^2 dF_i(x) - \bar{X}_i^2\right\}$$

$$= \frac{1}{n}\{\bar{X}(1-\bar{X}) - \sigma^2\} + \left\{\sum_{i=1}^{R} p_i \int_{I_i} x^2 dF(x) - \sum_{i=1}^{R} p_i^2 \bar{X}_i^2\right\} \qquad (7)$$

For the convenience of administration in surveys, we often divide of whole population into the strata, each being of the same size. In our case it means $p_1 = p_2 = \ldots = p_R = \frac{1}{R}$, and the formula becomes

$$\sigma_{\bar{x}}^2 = \frac{1}{n}\{\bar{X}(1-\bar{X}) - \sigma^2\} + \frac{1}{R}\left\{(\bar{X}^2 + \sigma^2) - \frac{1}{R}\sum_{i=1}^{R} \bar{X}_i^2\right\} \qquad (8)$$

Putting

$$\sigma_b^2 = \sigma^2 - \left(\frac{1}{R}\sum_{i=1}^{R} \bar{X}_i^2 - \bar{X}^2\right)$$

then $\sigma_{\bar{x}}^2$ is transformed into

$$\sigma_{\bar{x}}^2 = \frac{\sigma_w^2}{n} + \frac{\sigma_b^2}{R} \qquad (9)$$

Now σ_w^2, σ_b^2 mean the within-, between-variance, respectively, as to the primary sampling units in the whole population. The former is determined uniquely by selection of the primary and secondary sampling units which is independent of the method of the stratification, and the latter depends on both, especially on the method of stratification. The decreasing amount of the between-variance is represented by $\frac{1}{R}\sum_{i=1}^{R}\bar{X}_i^2 - X^2$ as $\sigma_b^2 = \sigma^2$ when there is no stratification. Therefore, we define the effect of stratification by the ratio

$$e = \frac{\sigma^2 - \sigma_b^2}{\sigma^2} = \frac{\frac{1}{R}\sum_{i=1}^{R}\bar{X}_i^2 - \bar{X}^2}{\sigma^2}. \qquad (10)$$

3. In the following we shall illustrate effects of stratification by several examples. (We suppose that the distribution function $F(x)$ is differentiable and $F'(x) = f(x)$, and sizes of primary sampling units are sufficiently small compared with N/R.)

[Example 1] Uniform distribution:

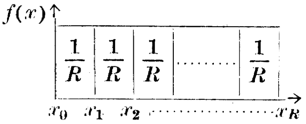


Fig. 1

Obviously, we have

$$\bar{X} = \frac{1}{2}, \quad \sigma^2 = \frac{1}{12}$$

$$\sigma_w^2 = \frac{1}{4} - \frac{1}{12} = \frac{1}{6} \qquad (11)$$

When the stratification is performed most effectively (in the best case), we see from the Fig. 1

$$x_i = \frac{i}{R}$$

and

$$\overline{X_i} = \frac{i}{R} - \frac{1}{2R} \quad {}^{**)}$$

(Here $I_i = (x_{i-1}, x_i)$ represents the interval correspoding to the i-th stratum) Hence we have

$$\sigma_b^2 = \left(\frac{1}{4} + \frac{1}{12}\right) - \frac{1}{R}\sum_{i=1}^{R}\left(\frac{i}{R} - \frac{1}{2R}\right)^2 = \frac{1}{12R^2} \tag{12}$$

then consequently,

$$\sigma_{\bar{x}}^2 = \frac{1}{6n} + \frac{1}{12R^3} \tag{13}$$

If the stratification is done in the worst method, we have

$$\overline{X_i} = \frac{1}{2}$$

Then

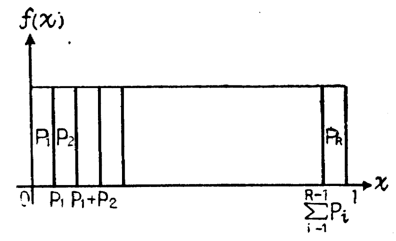
$$\sigma_b^2 = \left(\frac{1}{4} + \frac{1}{12}\right) - \frac{1}{4} = \frac{1}{12}$$

hence

$$\sigma_{\bar{x}}^2 = \frac{1}{6n} + \frac{1}{12R} \quad {}^{**)} \tag{14}$$

Comparing the second term of (13) with that of (14), we see that their orders are R^{-3} and R^{-1}, respectively. Therefore, it follows that in the usual stratifications the degree of R in $\sigma_{\bar{x}}^2$ lies between -1 and -3, for example, -2, and in this case, we get,

**)



In the case of unequal weights, we get

$$\overline{X_i} = \sum_{j=1}^{i} p_j - \frac{1}{2}p_i$$

$$\sigma_{\bar{x}}^2 = \frac{1}{6n} + \frac{1}{12}\sum_{i=1}^{R} p_i^4$$

Hence, $\sigma_{\bar{x}}^2$ attains its minimum value when $p_1 = p_2 = \ldots\ldots = pR$, namely in the case of equal weights. Therefore, it is enough in this example to discuss about the case of equal weights,

$$\sigma_{\bar{x}}^2 = \frac{1}{6n} + \frac{1}{12R^2} \quad (15)$$

Next, the effects of the stratifications represented by (13), (14) and (15) are respectively,

$$e_1 = 1 - \frac{1}{R^2}$$

$$e_2 = 0$$

$$e_3 = 1 - \frac{1}{R}$$

(See Fig. 2.)

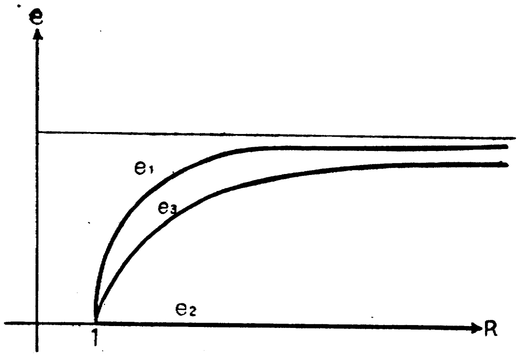


Fig. 2

[Example 2] Linear distribution: $f(x) = 2x$, $0 \leqq x \leqq 1$.

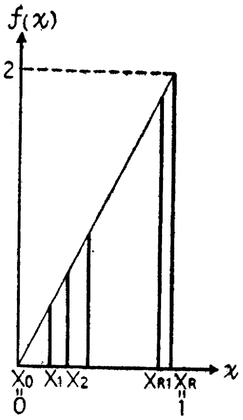


Fig. 3

In this case, we get easily

$$\bar{X} = \frac{2}{3}, \quad \sigma^2 = \frac{1}{18}, \quad \sigma_w^2 = \frac{1}{6}.$$

In the best case,

$$x_i^2 = x_{i-1}^2 + \frac{1}{R}$$

$$x_i = \sqrt{\frac{i}{R}}, \quad \bar{X}_i = \frac{2}{3}R(x_i^3 - x_{i-1}^3)$$

hence we have

$$\sum_{i=1}^{R} \bar{X}_i^2 = \frac{4}{9R}\left\{\sum_{i=1}^{R} i^3 + \sum_{i=1}^{R}(i-1)^3 - 2\sum_{i=1}^{R}(i(i-1))^{3/2}\right\}$$

$$\fallingdotseq \frac{1}{2}R - \frac{1}{24R}\log R.$$

Therefore

$$\sigma_b^2 = \left(\frac{4}{9} + \frac{1}{18}\right) - \frac{1}{R}\left(\frac{1}{2}R - \frac{1}{24R}\log R\right)$$

$$\doteq \frac{1}{24R^2}\log R$$

Hence

$$\sigma_{\bar{x}}^2 = \frac{1}{6n} + \frac{1}{24R^3}\log R \quad (16)$$

In the worst case, we have $\bar{X}_i = \frac{2}{3}$.

$$\sigma_b^2 = \frac{1}{18}$$

$$\sigma_{\bar{x}}^2 = \frac{1}{6n} + \frac{1}{16R} \tag{17}$$

Then we get the results like Example 1.

[Example 3] Triangle distribution:

$$f(x) = 4x \quad \left(0 \leqq x \leqq \frac{1}{2}\right)$$

$$= -4(x-1) \quad \left(\frac{1}{2} \leqq x \leqq 1\right)$$

$$\bar{X} = \frac{1}{2}, \quad \sigma^2 = \frac{1}{24}, \quad \sigma_w^2 = \frac{5}{24}$$

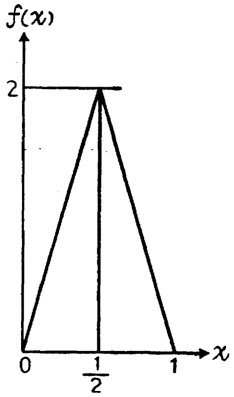


Fig. 4

On this case we devide the whole population into $2R$ strata, and in the best case we get

$$x_i = \sqrt{\frac{i}{4R}}, \quad \bar{X}_i = \frac{8R}{3}(x_i^3 - x^3_{i-1})$$

$$\frac{1}{2R}\sum_{i=1}^{2R}\bar{X}_i^2 = \frac{7}{24} - \frac{\log R}{24R^2}$$

$$\sigma_b^2 \doteqdot \left(\frac{1}{4} + \frac{1}{24}\right) - \left(\frac{7}{24} - \frac{\log R}{48R^2}\right) = \frac{\log R}{48R^2}$$

consequently

$$\sigma_{\bar{x}}^2 = \frac{5}{24n} + \frac{\log R}{48R^3} \tag{18}$$

In the worst case

$$\sigma_{\bar{x}}^2 = \frac{5}{24n} + \frac{1}{24R} \tag{19}$$

4. In conclusion, as to the above examples, when the variance of a sample is expressed in the form

$$\sigma_{\bar{x}}^2 = \frac{\sigma_w^2}{n} + \frac{\sigma_b^2}{R}$$

the second term of the right-hand side becomes smaller at the rate of R^{-3} in the best case, or at the rate of R^{-1} even in the worst as the number of strata R increases $\left(R \leqq \dfrac{N}{\max\limits_{(i,j)} N_{ij}}\right)$. Hence, to any stratification it will become at some rate between them.

Now, let us study optimum balancing between n and R. At first, we put the 1-st term equal to the 2-nd $\frac{\sigma_w^2}{n} = \frac{\sigma_b^2}{R}$. When we apply this result to example 1, we have

$$\frac{1}{6n} = \frac{1}{12R^3}$$

Since

$$\frac{\sigma_w^2}{n} = \frac{1}{6n}, \quad \sigma_b^2 = \frac{1}{12R^3}$$

we get

$$R = \sqrt[3]{\frac{n}{2}}$$

But generally in practice, the notion of optimum balancing between n and R cannot be considered without regard to cost for the survey. The cost function C may, as a first approximation, be expressed as a linear form of n and R:

$$C = c_0 + c_1 n + c_2 R.$$

Therefore, in order to obtain optimum balancing, we have only to find the solutions for n and R, which minimize the variance $\sigma_{\bar{x}}^2$ under the condition that the cost C be constant.

As to example 1, we have

(i) In the best case,

$$n = cR^2 \qquad \left(\text{where} \quad c = \sqrt{\frac{2c_2}{3c_1}}\right)$$

and R will be obtained as a solution of the quadratic equation

$$Cc_1R^2 + c_2R - (C - c_0) = 0.$$

(ii) In the worst case,

$$n = cR \qquad \left(\text{where} \quad C = \sqrt{\frac{2c_2}{c_1}}\right)$$

$$R = \frac{C - c_0}{Cc_1 + c_2}, \quad n = \frac{c(C - c_0)}{Cc_1 + c_2}.$$

5. The above results are not widely applicable because they are obtained to the special cases, using probability-proportionate sampling procedure giving equal weights to all strata and adopting the size proportionate allocation. Moreover, the above estimation by $\bar{x}$ concerns an estimate of a proportion of the population. However, our arguments is applicable to the cases as opinion surveys.

As for the distribution function $F(x)$, it is in general unknown before the survey. This has the result that we can not optimally determine sample size n and number of strata R, but using the above results we shall be able to design a perspective plan. After the survey we can learn pretty well about the shape of the distribution $F(x)$.

Several problems we want to solve in future are

(1) How to decide sizes of the primary sampling units optimally. (It should be noted that a stratum can be taken for a large unit.)

(2) How to get general solution for n and R without limitation about sampling method and sample allocation.

(3) Studies of cost functions in practice.

(4) Estimates of effects of stratifications in practical surveys.

Before long we shall be able to publish our studies on these points.

Institute of Statistical Mathematics

REFERENCES

1) Cochran, W. G.: "*Sample Survey Techniques.*" Inst. of Stat. North Carolina State College and Bureaw of Agr. Ec. U. S. Dep. of Agriculture Coop., 1948 (Mimeo.)

2) Hansen, M. H. and Hurwitz, W. N.: "On the Theory of Sampling from Finite Populations." *A. M. S.*, 1943, 332–362.

3) Hendricks, W. A.: "The Relative Efficiencies of Groups of Farms as Sampling Units." *J. A. S. A.*, 1944, 367–376.

4) Madow, L. H.: "On the Use of the County as the Primary Sampling Unit for State Estimates." *J. A. S. A.*, 1950, 30–47.

5) Stone, M.: "*Efficiency of National Samples Having the County as Sampling Unit.*" Iowa State College, 1946.