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 sub-sampling 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 \( \pi \), divide it into \( R \) strata and draw a sample of size \( n \) from \( \pi \) 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

\[
\sigma_s^2 = \sum_{i=1}^{R} \pi_i \left( \frac{\sigma_{wi}^2}{n_i} + \sigma_{bi}^2 \right)
\]

where \( n_i, \pi_i, \sigma_{wi}^2, \sigma_{bi}^2 \) 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

\[
\sigma_s^2 = \frac{1}{n} \sum_{i=1}^{R} \pi_i \sigma_{wi}^2 + \sum_{i=1}^{R} \pi_i^2 \sigma_{bi}^2
\]

where

\[
\sigma_{wi}^2 = \sum_{j=1}^{N_i} \frac{N_{ij}}{N_i} \sigma_{ij}^2
\]

\[
\sigma_{bi}^2 = \sum_{j=1}^{N_i} \frac{N_{ij}}{N_i} (\bar{X}_{ij} - \bar{X}_i)^2
\]

\[
\sigma_{ij} = \frac{1}{N_{ijk}} \sum_{k=1}^{N_{ijk}} (X_{ijk} - \bar{X}_{ij})^2
\]

\[
\bar{X}_i = \frac{1}{N_i} \sum_{j=1}^{N_i} X_{ij}
\]

\(^*\) 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.
\[ X_{ij} = \frac{1}{N_i} \sum_{k=1}^{N_i} X_{ikj} \]

\( X_{ikj} \): the attribute of the \( k \)-th secondary sampling unit of the
\( j \)-th primary sampling unit in the \( i \)-th stratum.

Since \( \frac{N_j}{N} = P_j \), \( \sigma^2 \) becomes

\[ \sigma^2 = \frac{1}{n} \sigma^2_{m^2} + \sum_{i=1}^{R} \left( \frac{N_i}{N} \right)^2 \sigma_{ij}^2 \]

(6)

where \( \sigma^2 \) denotes \( \sum_{i=1}^{R} \sum_{j=1}^{N_i} \frac{N_i}{N} \sigma_{ij}^2 \).

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 \( \sigma_{ij}^2 \)'s. Now, we introduce the distribution function \( F(x) \) of means \( \bar{X}_{ij} \) of the primary sampling units, which is given by considering for every \( \bar{X}_{ij} \) the weight \( \frac{N_j}{N} \). Now, for brevity we confine ourselves to the case where a ratio of individuals having some characteristic in \( \pi \) should be estimated. The general case will be similarly treated.

\[ \sigma^2_{m^2} = \int_{0}^{1} x(1-x)dF(x) = \bar{X}(1 - \bar{X}) - \sigma^2 \]

where

\[ \bar{X} = \int_{0}^{1} xdF(x) \]

is the population mean

and

\[ \sigma^2 = \int_{0}^{1} (x - \bar{X})^2dF(x) \]

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:

\[ \sigma_{ij}^2 = \int_{I_i} (x - \bar{X}_i)^2dF_i(x) \]

\[ = \int_{I_i} x^2dF_i(x) - \bar{X}_i^2 \]

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 \( \bar{X}_i = \int_{I_i} xdF_i(x) \) is the population mean of the \( i \)-th stratum.

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