# A Note on Estimation of Variance Components in a Multi-stage Sampling Design Based on Interpenetrating Sub-samples

By M. N. MURTHY, Calcutta[1])

*Summary:* In this note it is shown that unbiased estimators of the components of the sampling variance of an estimator in the case of a stratified multi-stage sampling design could easily be obtained by selecting the samples at the different stages in the form of two or more independent interpenetrating sub-samples.

## 1. Introduction

The need for estimating validly the sampling error of an estimator based on a sample survey to permit proper interpretation of the survey results is widely recognized. In fact, one of the main features of the sampling approach is that when properly used it enables us to estimate not only the population parameter under consideration, but also the sampling variance of the estimate from the sample observations. Further, a knowledge of the relative magnitudes of the components of the sampling variance in a multi-stage design is also important, as this would help in evolving a suitable sampling design with appropriate allocation of the total sample size among the different stages of the design.

The main difficulty in obtaining estimates of variances of the estimates of the population characteristics and of their components in large scale sample surveys is the elaborate and time-consuming calculations involved in their computation. For some sampling schemes, such as systematic sampling, sampling without replacement, etc., it is not even possible to obtain simple expressions for the variance estimator. However, it is well known that in such cases it is possible to estimate unbiasedly the total variance of an estimate in a simple manner, if the sample is drawn in the form of two or more independent interpenetrating sub-samples. In this note, it is shown that it is possible to obtain easily unbiased estimators

[1]) Dr. M. N. MURTHY, Professor and Head of Design Div., Dept. of National Sample Survey, Indian Statistical Institute, Calcutta-35, India.

of the components of the sampling variance of an estimator in the case of a stratified multi-stage sampling design by extending the principle of interpenetrating sub-samples to the different stages of the sampling design.

## 2. Two-Stage Sampling Design

Suppose in a stratified two-stage sampling design, the sample of first stage units ($f$ s $u$'s) in each stratum is drawn in the form of $n$ ($\geqslant 2$) independent sub-samples and the sample of second stage units ($s$ s $u$'s) from each selected $f$ s $u$ is selected in the form of $m$ ($\geqslant 2$) independent sub-samples according to any probability design permitting valid estimation of the population parameter under consideration. Then there are $n$ $m$ sub-samples, each capable of providing a valid estimate of the $s$th stratum total.

Let $y_{s,ij}$ be the estimate of the $s$th stratum total based on the $j$th sub-sample of the $s$ s $u$'s in the $i$th sub-sample of the $f$ s $u$'s. Then the overall estimator of the population total $Y$ is given by

$$\hat{Y} = \frac{1}{n\,m} \sum_{i=1}^{n} \sum_{j=1}^{m} y_{s,ij}\,, \tag{2.1}$$

where $K$ is the number of strata and its variance is of the form

$$V(\hat{Y}) = \sum_{s=1}^{K} \left\{ \frac{A_s}{n} + \frac{B_s}{n\,m} \right\} = \frac{A}{n} + \frac{B}{n\,m}\,, \tag{2.2}$$

where $A_s$ and $B_s$ are respectively the variation between the $f$ s $u$'s and that between the $s$ s $u$'s within the $f$ s $u$'s in the $s$th stratum for an estimator based on one of the $n$ $m$ sub-samples, and $A = \sum_{s=1}^{K} A_s$ and $B = \sum_{s=1}^{K} B_s$. At this stage it may be noted that $A$ and $B$ might be quite complicated expressions for some designs such as those involving systematic sampling, without replacement selection, etc.

It can be easily shown that unbiased estimators of $V(\hat{Y})$ based on the stratum level sub-sample estimates and on the over-all sub-sample estimates are respectively given by

$$\hat{V}_1(\hat{Y}) = \frac{1}{n(n-1)} \sum_{s=1}^{K} \sum_{i=1}^{n} (\bar{y}_{s\,i.} - \bar{y}_{s\,..})^2 \tag{2.3}$$

and

$$\hat{V}_2(\hat{Y}) = \frac{1}{n(n-1)} \sum_{i=1}^{n} (\bar{y}_{.\,i.} - \bar{y}_{...})^2, \tag{2.4}$$

where

$$\bar{y}_{si.} = \frac{1}{m}\sum_{j=1}^{m} y_{sij}, \qquad \bar{y}_{s..} = \frac{1}{n}\sum_{i=1}^{n} \bar{y}_{si.}, \qquad \bar{y}_{.i.} = \sum_{s=1}^{K} \bar{y}_{si.}$$

and

$$\bar{y}_{...} = \bar{Y}.$$

It can be shown that (2.3) is more efficient than (2.4).

An estimator of $B$ in (2.2) based on the stratum level sub-sample estimates is given by

$$\hat{B}_1 = \frac{1}{n(m-1)}\sum_{s=1}^{K}\sum_{i=1}^{n}\sum_{j=1}^{m}(y_{sij} - \bar{y}_{si.})^2, \tag{2.5}$$

for

$$\begin{aligned} E\{n(m-1)\hat{B}_1\} &= \sum_{s=1}^{K} E\left\{\sum_{i=1}^{n}\sum_{j=1}^{m} y_{sij}^2 - m\sum_{i=1}^{n}\bar{y}_{si.}^2\right\} \\ &= \sum_{s=1}^{K} n\,m\,\{V(y_{sij}) - V(\bar{y}_{si.})\} \\ &= n\,m\sum_{s=1}^{K}\left\{(A_s + B_s) - \left(A_s + \frac{B_s}{m}\right)\right\} \\ &= n(m-1)\sum_{s=1}^{K} B_s = n(m-1)\,B. \end{aligned}$$

Another estimator of $B$, which is easier to calculate, but less efficient, than $\hat{B}_1$ can be obtained on the basis of the overall sub-sample estimates and it is given by

$$\hat{B}_2 = \frac{1}{n(m-1)}\sum_{i=1}^{n}\sum_{j=1}^{m}(y_{.ij} - \bar{y}_{.i.})^2, \tag{2.6}$$

where

$$y_{.ij} = \sum_{s=1}^{K} y_{sij}.$$

Using the estimators of $V(\bar{Y})$ and $B$ considered here, we can get the following two estimators of $A$, one based on $(\hat{V}_1, \hat{B}_1)$ and the other on $(\hat{V}_2, \hat{B}_2)$, that is,

$$\hat{A}_1 = n\,\hat{V}_1 - \frac{\hat{B}_1}{m} \tag{2.7}$$

and

$$\hat{A}_2 = n\,\hat{V}_2 - \frac{\hat{B}_2}{m}, \tag{2.8}$$

where $A_1$ is based on the stratum-wise sub-sample estimates and $A_2$ is based on the overall sub-sample estimates.

This procedure can also be applied to the case of estimating the population ratios. Incidentally it may be mentioned that for some simpler sampling schemes, such as sampling with replacement, it is also possible to obtain estimators of the inherent between- and within-variations by this procedure. For instance, in the case of using simple random sampling with replacement at both the stages of sampling, the sampling variance is of the form

$$V(\bar{Y}) = \frac{\sigma_b^2}{n'} + \frac{\sigma_w^2}{n' m'}, \tag{2.9}$$

where $n'$ and $m'$ are the number of sample $f$ s $u$'s and the number of sample $s$ s $u$'s per sample $f$ s $u$, and hence in this case

$$A = \frac{\sigma_b^2}{(n'/n)} \quad \text{and} \quad B = \frac{\sigma_w^2}{(n'/n)(m'/m)}. \tag{2.10}$$

Thus we see that it is possible to estimate $\sigma_b^2$ and $\sigma_w^2$ unbiasedly using the estimators of $A$ and $B$ obtained in this section.

## 3. Three-Stage Sampling Design

The procedure given in Section 2 for two-stage sampling design can easily be extended to three-stage and higher stage designs for estimating the variance components of the sampling error. In a three-stage design, suppose the samples at the three stages are drawn in the form of $n$, $m$ and $l$ independent sub-samples in each stratum, and let $y_{sijk}$ be the estimate of the $s$th stratum total based on the $k$th sub-sample of the third stage units ($t$ s $u$'s) in the $j$th sub-sample of $s$ s $u$'s of the $i$th sub-sample of $f$ s $u$'s in the $s$th stratum. Then the combined estimator of the population total $Y$ is given by

$$\bar{Y} = \frac{1}{n\,m\,l} \sum_{s=1}^{K} \sum_{i=1}^{n} \sum_{j=1}^{m} \sum_{k=1}^{l} y_{sijk}. \tag{3.1}$$

The variance of $\bar{Y}$ is of the form

$$V(\bar{Y}) = \frac{A}{n} + \frac{B}{n\,m} + \frac{C}{n\,m\,l}, \tag{3.2}$$

where $A$, $B$ and $C$ are respectively the variations (i) between $f$ s $u$'s, (ii) between $s$ s $u$'s within $f$ s $u$'s and (iii) between $t$ s $u$'s within $s$ s $u$'s for an estimate based on any one of the $n\,m\,l$ possible sub-samples. An unbiased estimator of the variance of $\bar{Y}$ based on the stratum-wise sub-sample estimates is given by

$$V_1(\hat{Y}) = \frac{1}{n(n-1)} \sum_{s=1}^{K} \sum_{i=1}^{n} (\bar{y}_{si..} - \bar{y}_{s...})^2, \tag{3.3}$$

where

$$\bar{y}_{si..} = \frac{1}{ml} \sum_{k=1}^{l} y_{sijk} \quad \text{and} \quad \bar{y}_{s...} = \frac{1}{n} \sum_{i=1}^{n} \bar{y}_{si..},$$

since $\bar{y}_{si..}$, $i = 1, 2, \ldots, n$, are $n$ independent estimates of the $s$th stratum total. A less efficient variance estimator, which is easier to compute than (3.3), can be obtained on the basis of the overall sub-sample estimates and it is given by

$$V_2(\hat{Y}) = \frac{1}{n(n-1)} \sum_{i=1}^{n} (\bar{y}_{si..} - \bar{y}_{s...})^2, \tag{3.4}$$

where

$$\bar{y}_{.i..} = \sum_{s=1}^{K} \bar{y}_{si..} \quad \text{and} \quad \bar{y}_{....} = \hat{Y}.$$

Unbiased estimators of $A$, $B$ and $C$ based on the stratum level sub-sample estimates and those based on the overall sub-sample estimates, which are easier to calculate, but less efficient, than the former, are given by

$$\hat{A}_1 = n V_1 - \frac{\hat{B}_1}{m} - \frac{C_1}{ml}, \tag{3.5}$$

$$B_1 = \frac{1}{n} \frac{1}{m-1} \sum_{s=1}^{K} \sum_{i=1}^{n} \sum_{j=1}^{m} (\bar{y}_{sij.} - \bar{y}_{si..})^2 - \frac{C_1}{l}, \tag{3.6}$$

$$C_1 = \frac{1}{n} \frac{1}{m} \frac{1}{l-1} \sum_{s=1}^{K} \sum_{i=1}^{n} \sum_{j=1}^{m} \sum_{k=1}^{l} (\bar{y}_{sijk} - \bar{y}_{sij.})^2, \tag{3.7}$$

and

$$\hat{A}_2 = n V_2 - \frac{B_2}{m} - \frac{C_2}{ml}, \tag{3.8}$$

$$B_2 = \frac{1}{n} \frac{1}{m-1} \sum_{i=1}^{n} \sum_{j=1}^{m} (\bar{y}_{.ij.} - \bar{y}_{.i..})^2 \frac{C_2}{l}, \tag{3.9}$$

$$C_2 = \frac{1}{n} \frac{1}{m} \frac{1}{l-1} \sum_{i=1}^{n} \sum_{j=1}^{m} \sum_{k=1}^{l} (y_{sijk} - \bar{y}_{.ij.})^2, \tag{3.10}$$

where

$$y_{.ijk} = \sum_{s=1}^{K} y_{sijk},$$

other notations being similar to those used earlier.

## 4. Practical Implications

By selecting a multi-stage sample in the form of independent sub-samples at each stage and tabulating each of the sub-samples separately at the stratum level or at an overall level, it is possible to obtain quick estimates of the sampling variance and of its components. For instance, in the case of a two-stage sampling design with two or more sub-samples at each stage, the presentation of the results separately for the four or more sub-samples would help in easy computation of the estimates of the variance and its components. Similarly in the case of a three-stage design, the data are to be presented separately for eight or more sub-samples to permit simple computation of the variance estimate and the estimates of its components.

Prior classification of the ultimate sampling units into the various sub-samples would be of considerable help in large scale survey work, as it would be convenient to obtain the sub-sample estimates as a part of the routine tabulation operation itself. Thus, if the procedure outlined in this note is used, there would not be any need for elaborate separate tabulations for getting the estimates of the variance components.

It may, however, be noted that the procedure of selecting independent sub-samples at the different stages may make the design less efficient for some sampling schemes. If this loss in efficiency is found to be substantial for a specified sampling design, the above procedure can be applied by splitting the entire sample at the different stages into sub-samples at the tabulation stage for getting approximations to variance components.