(2) A NOTE ON NESTED SAMPLING

By MOHONLAL GANGULI

In some of the random sampling surveys recently conducted by the Statistical Laboratory, Calcutta, the procedure that had been followed in selecting samples has been termed *nested sampling* by Professor P. C. Mahalanobis. This consists in dividing up the entire area under survey into a large number of zones, say Z_1. Each of these Z_1-zones are sub-divided into a large number of smaller zones called, say, Z_2. These Z_2's in turn are split up into still smaller zones Z_3's and so on until one arrives at the ultimate unit of sampling. Nested sampling is done, firstly, by selecting a number of Z_1-zones at random from the population; then from each of these selected Z_1-zones, selecting at random a number of Z_2-zones. From each of these selected Z_2-zones again a number of Z_3-zones are chosen at random and so on until the ultimate units of sampling are reached.

Nested sampling, so far, has been applied to agricultural surveys only. It is easily conceivable that this type of sampling is applicable to other kinds of large-scale economic and sociological surveys.

The sample mean, we know, will be an unbiassed estimate of the population mean. The error of the sample mean will be a linear function of all the variances between various zones. The variances are found, in practice, to be existent. The problem here is to determine, firstly, an exact expression of this error and, secondly, to estimate these various zone-variances. Estimation of these zone-variances are necessary not only for the purpose of estimating the error of the sample mean, but also, when the total number of sample to be taken is predetermined by the amount of fund available and the amount of accuracy desired, for the purpose of ascertaining the number of units to be selected from each zone.

Professor P. C. Mahalanobis suggested the form in which the error of the sample mean should come out, in the case when the numbers of zones selected at each successive stage are equal.

In this paper the problem has been solved for unequal cases and for w-fold nesting, that is Z going from Z_1 to Z_w. In practice, however, one hardly goes beyond four-fold nesting; and even if one does, the calculations of the constants become increasingly prohibitive due to the magnitude of the work involved. In this paper we have given the results up to four-fold nesting for ready reference. The general expression, both in the case of error of the mean and also in the case of variance between zones, can be derived readily from the particular ones or vice versa and are also given. These are expressed with slightly different notations for the sake of neatness.

The Set-up

We propose to have a linear set-up for our observations arising out of the particular kind of sampling detailed above and describe an individual observation as

$$x_{ijkl\ldots\ldots w} = A + b_i + c_{ij} + d_{ijk} + \ldots\ldots + z_{ijk\ldots\ldots w}$$

where $x_{ijkl\ldots\ldots w}$ is the observation of wth ultimate unit in the i-th Z_1-zone, j-th Z_2-zone, k-th Z_3-zone and so on. b_i, c_{ij}, d_{ijk} etc. are variable quantities with finite number of values. Each of these variables has expectation equal to zero and their variances are σ^2_i, σ^2_{ij}, σ^2_{ijk} etc. respectively. The material is assumed to be homoscedastic, because, this assumption is always made in problems of analysis of variance, without which the analysis has no meaning. Thus, we may say σ_i's are constant for all i's, σ_{ij}'s are constant for all i, j's and so on for all variances.

It will be noticed, the unknown zone-variances are estimated from the sum of squares obtained from usual analysis of variance table for testing null hypothesis, although our set-up is different from the usual one. It will also be noticed that the corrected means of i-th Z_1-zone i, j-th Z_2-zone etc. will not represent the maximum likelihood estimates of b_i, c_{ij} etc. This, however, will not prevent us from finding out the mathematical expectations of the different sums of squares in the analysis of variance table.

(1) In a particular case of six-fold nesting carried out in the Statistical Laboratory, Calcutta, with 320 ultimate sampling units, the time that was actually required for calculation of all the cons-constants was found to be 6 hours, 20 minutes.

Below are given the results in case of 2-fold, 3-fold and 4-fold nesting and the generalised form.

Two-fold Nesting

(a) Set up

The set up is $x_{ij}=A+b_i+z_{ij}$, where i goes from 1 to t and j goes from 1 to n_i

Let us call $\sum_i n_i=n$

We have, $E(b_i)=0$, $E(b_i)^2=\sigma_1^2$, $E(z_{ij})=0$, $E(z_{ij})^2=\sigma_2^2$

(b) Analysis of Variance

The analysis of variance table may be written up as

Due to	Sum of Squares	Degrees of Freedom	Variance
Between Z_1-zones	$S(x_{i.}-x_{..})^2$	$t-1$	V_1
Between Z_2-zones within Z_1-zones	$S(x_{ij}-x_{i.})^2$	$n-t$	V_2
Total	$S(x_{ij}-x_{..})^2$	$n-1$	

(c) Variance of Sample mean and its estimation

$$E(x_{..})=A \quad V(x_{..})=\frac{\sum_i n_i^2}{n^2}\sigma_1^2+\frac{1}{n}\sigma_2^2$$

V_2 will be the unbiased estimate of σ_2^2. V_1 will be the unbiased estimate of

$$\frac{n-\sum_i n_i^2/n}{t-1}\sigma_1^2+\sigma_2^2$$

In "equal" case, that is, when the number of zones selected from each of the next higher-order zones is equal, we shall have

$$V(x_{..})=\frac{\sigma_1^2}{t}+\frac{\sigma_2^2}{n}$$

V_2 will be the estimate of σ_2^2. V_1 will be the estimate of $\frac{n}{t}\sigma_1^2+\sigma_2^2$

Three-fold nesting

(a) Set up

The set up is $x_{ijk}=A+b_i+c_{ij}+Z_{ijk}$

where i goes from 1 to t, j goes from 1 to n_i, and k goes from 1 to m_{ij}

Let us call

$$\sum_i n_i=n \quad \sum_j m_{ij}=m_i, \text{ and } \sum_i m_i=m.$$

We have,

$$E(b_i)=0 \quad E(b_i)^2=\sigma_1^2, \quad E(c_{ij})=0 \quad E(c_{ij})^2=\sigma_2^2, \quad E(z_{ijk})=0 \quad E(z_{ijk})^2=\sigma_3^2,$$

(b) Analysis of Variance

The analysis of variance table may be written up as

Due to	Sum of Squares	Degrees of Freedom	Variance.
Between Z_1-zones.	$S(x_{i..}-x_{...})^2$	$t-1$	V_1
Between Z_2-zones within Z_1-zones.	$S(x_{ij.}-x_{i..})^2$	$n-t$	V_2
Between Z_3-zones within Z_2-zones.	$S(x_{ijk}-x_{ij.})^2$	$m-n$	V_3
Total	$S(x_{ijk}-x_{...})^2$	$m-1$	

(c) Variance of Sample mean and its estimation

$$E(x_{...}) = A$$

$$V(x_{...}) = \frac{\sum_i m_i^2}{m^2}\,\sigma_1^2 + \frac{\sum_i \sum_j m_{ij}^2}{m^2}\,\sigma_2^2 + \frac{1}{m}\,\sigma_3^2$$

V_3 will be the unbiased estimate of σ_3^2

V_2 will be the unbiased estimate of

$$\frac{m - \sum_i \frac{\sum_j m_{ij}^2}{m_i}}{n-t}\,\sigma_2^2 + \sigma_3^2$$

V_1 will be the unbiased estimate of

$$\frac{m - \frac{\sum_i m_i^2}{m}}{t-1}\,\sigma_1^2 + \frac{\sum_i \frac{\sum_j m_{ij}^2}{m_i} - \frac{\sum_i \sum_j m_{ij}^2}{m}}{t-1}\,\sigma_2^2 + \sigma_3^2$$

In "equal" case,

$$V(x_{...}) = \frac{\sigma_1^2}{t} + \frac{\sigma_2^2}{n} + \frac{\sigma_3^2}{m}$$

V_3 will estimate σ_3^2

V_2 will estimate $\frac{m}{n}\sigma_2^2 + \sigma_3^2$

V_1 will estimate $\frac{m}{t}\,\sigma_1^2 + \frac{m}{n}\,\sigma_2^2 + \sigma_3^2$

Four-fold nesting

(a) Set up

The set up is $x_{ijkl} = A + b_i + c_{ij} + d_{ijk} + Z_{ijkl}$

where i goes from 1 to t
j goes from 1 to n_i
k goes from 1 to m_{ij}
l goes from 1 to p_{ijk}

Let us call

$$\sum n_i = n$$

$$\sum_j m_{ij} = m_{i.} \quad \sum_i m_i = m$$

$$\sum_k p_{ijk} = p_{ij.} \quad \sum_j p_{ij} = p_{i.} \quad \sum_i p_i = p$$

We have

$E(b_i)=0$ $E(b_i)^2=\sigma_1^2$ $E(c_{ij})=0$ $E(c_{ij})^2=\sigma_2^2$ $E(d_{ijk})=0$ $E(d_{ijk})^2=\sigma_3^2$ $E(z_{ijkl})=0$ $E(z_{ijkl})^2=\sigma_4^2$

(b) Analysis of Variance

The analysis of variance table may be written up as

Due to	Sum of Squares.	Degrees of Freedom	Variance
Between Z_1-zones	$S(x_{i...} - x_{....})^2$	$t-1$	V_1
Between Z_2-zones within Z_1-zones	$S(x_{ij..} - x_{i...})^2$	$n-t$	V_2
Between Z_3-zones within Z_2-zones	$S(x_{ijk.} - x_{ij..})^2$	$m-n$	V_3
Between Z_4-zones within Z_3-zones	$S(x_{ijkl} - x_{ijk.})^2$	$p-m$	V_4
Total	$S(x_{ijkl} - x_{....})^2$	$p-1$	

(c) Variance of sample mean and its estimation

$$E(x\ldots.)=A$$

$$V(x\ldots.)=\frac{\sum\limits_{i} p^2_i}{p^2}\sigma^2_1+\frac{\sum\limits_{i}\sum\limits_{j} p^2_{ij}}{p^2}\sigma^2_2+\frac{\sum\limits_{i}\sum\limits_{j}\sum\limits_{k} p^2_{ijk}}{p^2}\sigma^2_3+\frac{1}{p}\sigma^2_4$$

V_4 will be the unbiased estimate of σ^2_4

$(m-n)\ \hat{V}_3$ will be the unbiased estimate of

$$\left\{p-\sum_{i}\sum_{j}\frac{\sum\limits_{k} p^2_{ijk}}{p_{ij}}\right\}\sigma^2_3+\sigma^2_4$$

$(n-t)\ V_2$ will be the unbiased estimate of

$$\left\{p-\sum_{i}\frac{\sum\limits_{j} p^2_{ij}}{p_i}\right\}\sigma^2_2+\left\{\sum_{i}\sum_{j}\frac{\sum\limits_{k} p^2_{ijk}}{p_{ij}}-\sum_{i}\frac{\sum\limits_{j}\sum\limits_{k} p^2_{ijk}}{p_i}\right\}\sigma^2_3+\sigma^2_4$$

$(t-1)\ V_1$ will be the unbiased estimate of

$$\left\{p-\frac{\sum\limits_{i} p^2_i}{p}\right\}\sigma^2_1+\left\{\sum_{i}\frac{\sum\limits_{j} p^2_{ij}}{p_i}-\frac{\sum\limits_{i}\sum\limits_{j} p^2_{ij}}{p}\right\}\sigma^2_2+\left\{\sum_{i}\frac{\sum\limits_{j}\sum\limits_{k} p^2_{ijk}}{p_i}-\frac{\sum\limits_{i}\sum\limits_{j}\sum\limits_{k} p^2_{ijk}}{p}\right\}\sigma^2_3+\sigma^2_4$$

In "equal" case,

$$V(x\ldots.)=\frac{\sigma^2_1}{t}+\frac{\sigma^2_2}{n}+\frac{\sigma^2_3}{m}+\frac{\sigma^2_4}{p}$$

V_4 will estimate σ^2_4

V_3 will estimate $\frac{p}{m}\sigma^2_3+\sigma^2_4$

V_2 will estimate $\frac{p}{n}\sigma^2_2+\frac{p}{m}\sigma^2_3+\sigma^2_4$

V_1 will estimate $\frac{p}{t}\sigma^2_1+\frac{p}{n}\sigma^2_2+\frac{p}{m}\sigma^2_3+\sigma^2_4$

The generalised form.

Let us represent any sample in w-fold nesting by $X_{i_1 i_2 \ldots i_w}=A+{}^1B_{i_1}+{}^2B_{i_1 i_2}+\ldots{}^wB_{i_1 i_2\ldots i_w}$ where A is a constant and for $1<K<w$

(1) i_k goes from 1 to ${}^kN_{i_1 i_2\ldots i_{k-1}}$

(2) $E({}^kB_{i_1 i_2\ldots i_k})=0$

(3) $E({}^kB_{i_1 i_2\ldots i_k})^2=\sigma^2_k$

(4) $\sum\limits_{i_p}\sum\limits_{i_{p+1}}\ldots\ldots\sum\limits_{i_{k-1}} {}^kN_{i_1 i_2\ldots i_{k-1}}={}^kN_{i_1 i_2\ldots i_{p-1}}\ (1<p<k-1)$

and

${}^kN_{i_0}={}^kN$ (say).

The typical term in the Analysis of Variance will be $S(X_{i_1 i_2\ldots i_k}-X_{i_1 i_2\ldots i_{k-1}})^2$ which is "sum of squares between Z_k-zones within Z_{k-1} zones" and its corresponding degrees of freedom and variance being f_k and V_k respectively.

Then for $1<K<w-1$ and $0<j<w-k-1$ the variance of the sample-grand-mean will be

$$\sum_{k=1}^{w-1}\left[\frac{\sigma^2_k}{{}^wN^2}\sum_{i_1}\sum_{i_2}\ldots\sum_{i_k}({}^wN_{i_1 i_2\ldots i_k})^2\right]+\frac{\sigma^2_w}{{}^wN}$$

The mathematical expectation of $V_k f_k$ will be

$$\sum_{j=0}^{w-K-1}\left[\sum_{i_1}\sum_{i_2}\ldots\sum_{i_{k+j}}({}^wN_{i_1 i_2\ldots i_{k+j}})^2\left\{\frac{1}{{}^wN_{i_1 i_2\ldots i_k}}-\frac{1}{{}^wN_{i_1 i_2\ldots i_{k-1}}}\right\}\times\sigma^2_{k+j}\right]+\sigma^2_w$$

It is important to note that the question of estimating the values of σ's arises only when the null-hypothesis is found to be false. We may, for example, find V_1 to be smaller than V_2. This will mean that the difference between the means of Z_1-zones is not significant. So, if we have to give an estimate of σ_1 it must be given as zero.

[*Paper received : 30 May, 1941*].