SAMPLING DESIGN FROM EXPERIMENTAL DESIGN

BY

S. MOHANTY

Orissa University of Agriculture and Technology, Bhubaneswar.

(Received : August, 1980)

SUMMARY

Through the present investigation, the relationship between sampling design with that of experimental design has been shown the variance of the estimate and estimate of the variance in term of the characterising parameters of the sampling design for several sampling schemes usually adopted in practice, are given.

1. INTRODUCTION

It has been noticed through the present investigation that the basic problem in construction of incomplete block designs and evolving different sampling plans are basically the same. Incomplete block designs are mainly characterised by the parameters v, denoting the number of varities, b, the number of blocks, k_t $(<v)$, the block size of the t th block, r_i the number of times the i th varity occurs in the b blocks and λ_{ij} the number of times the ith and jth varieties occur together in b blocks. Using these parameters an incomplete block design is defined as follows.

DEFINITION 1.0

An incomplete block design D^* is the arrangement of v varities in b blocks of k_t plots each such that the ith variety occurs in r_i plots and ith and jth varities occur together in λ_{ij} plots of the design.

The design D^* is said to be balanced when $k_t=k$, $r_i=r$ and $\lambda_{ij}=\lambda$ for all i, j and t.

To define a sampling design, we first define a finite population of N numbers of distinguishable units given by

$$U=(u_1, u_2, \ldots\ldots\ldots\ldots u_N). \qquad \ldots(1.0)$$

A sample s of size n from U is a finite order sequence of n units from (1.0) given by

$$s=(u_{i1}, u_{i2}, \ldots\ldots\ldots\ldots, u_{in}), \qquad \ldots(1.1)$$

where i's which need not be destinct range from 1 to N. Let $n(s)$ be the number of distinct units in s, which take different values depending on the sampling plan. Let S be collection of all (or possible) countable samples s from U. Call this S as the basic sample space. This basic sample space is called a basic sample design S^*, only when a probability measure $P(s)$ is assigned to each of s of S, such that $P(s)\geqslant 0$ and $\sum_{s\in S} P(s)=1$. The basic sample space is also characterised by the following parameters.

N=number of distinguishable units in the population.

b=the number of samples s in S.

n=the number of units in s.

R_i=the number of times u_i occurs in b samples of S.

$\triangle_{ij}$=the number of times u_i and u_j occur together in b sample of S.

Using the above parameters we can define a sampling design as follows.

DEFINITION 1.1

A sample design S^* is defined as the collection of b finite order sequence $s=(u_{i1}, u_{i2}, \ldots, u_{in})$ of n units each from N number of distinguishable units of $U=(u_1, u_2, \ldots, u_N)$, such that u_i occurs together in R_i number of times and u_i and u_j occur together $\triangle_{ij}$ times in S^* and each order sequence s of S has an assigned probability measure of selection $P(s)\geqslant 0$ and $\sum_{s\in S} P(s)=1$.

The parameters of S^* are N, b, n, R_i and $\triangle_{ij}$ for $i, j=1, 2, \ldots N$ These parameters have several relationship as the parameters of the corresponding D^*. The relationships are

$$(i) \quad bn=\sum_{i=1}^{N} R_i \qquad \ldots(1.2)$$

and

$$(ii) \quad \sum_{j\neq i}^{N} \triangle_{ij}=(n-1)\, R_i \qquad \ldots(1.2)$$

Further, in an experimental design D^*, if a unit normally does not occur more than once in a block, it leads to a sampling design S^* without replacement. There is special group of designs, known as balanced n-ary design, where an unit occurs more than once in a block. This design corresponds to sampling designs with replacement. The main difference between the two designs D^* and S^* is that in the experimental design all the blocks are used for the experiment, where as in the sampling design one of the samples is drawn at random from the sample space S^* with a probability measure $P(s)$ for s, such that $\sum_{s \epsilon S} P(s)=1$. When R_i's are constant for all units, the scheme is called equal probability sampling scheme.

ESTIMATION OF THE POPULATION PARAMETERS.

Estimation of mean.

Let $(y_1, y_2, ..., y_n)$ denote a set of measurements of the units of a selected sample s_j, solected with equal probability $P(s_j)$ from the basic sample space S^* defined earlier with parameters N, b, n, R_i and $\triangle_{ij}$. The sample mean based on the tth selected sample is given by

$$\bar{y}_t = \frac{1}{n} \sum_{i} f_{it}\, y_i, \qquad ...(2.1)$$

where $\sum_{i}$ denotes the summation over $n(s_t)$ distinct units in the sample and $f_{ij}(\neq 0)$ is the frequency of the ith unit present in s_t. If we assume that f_{it} takes the value zero, when the ith unit is not present in s_t, we have

$$R(\bar{y}_t) = \frac{1}{n} \sum_{i=1}^{N} E(f_{it})\, y_i$$

$$= \frac{1}{n} \sum_{i=1}^{N} \frac{R_i}{b}\, y_i$$

$$= \frac{1}{nb} \sum_{i=1}^{N} R_i y_i = \bar{Y}_w, \text{ say} \qquad ...(2.2)$$

since $E(f_{it}) = \frac{1}{b} \sum_{t=1}^{b} f_{it} = R_i/b.$

(*i*) When we consider equal probability sampling, $R_i = R$ and hence

$$E(\bar{y}_t) = \bar{Y}. \qquad \ldots(2.3)$$

(*ii*) When we consider equal probability without replacement scheme, f_{ij} takes value either 1 or 0 and $n(s) = n$.

(*iii*) When we follow unequal probability sampling scheme, $E(\bar{y}_t) \neq \bar{Y}$. In such cases the estimate of $\bar{Y}$ is modified and is given by $\bar{z}_t$, where $z_i = (n\, by_i)/(NR_i)$.

VARIANCE OF THE ESTIMATE

By definition the variance of $\bar{y}_t$ is given by

$$V(\bar{y}_t) = E\,[y_t - E(y_t)]^2$$

$$= (1/n^2)\, E\,[\sum_{i=1}^{N} f_{it}\,(y_i - \bar{Y}_w)]^2,$$

where f_{it} takes value 0 when not present in the sample s_t.

$$= n^{-2}[\sum_{i}^{N} E(f_{it})^2\,(y_i - \bar{Y}_w)^2$$

$$+ \sum_{i}^{N} \sum_{j \neq i}^{N} f_{it} f_{jt}\,(y_i - \bar{Y}_w)\,(y_j - \bar{Y}_w)]$$

From the definition of the sampling design

$$E(f_{it} f_{jt}) = \sum_{t=1}^{b} f_{it} f_{jt} P(s_t) = b^{-1} \sum_{t=1}^{b} f_{it} f_{jt} = b^{-1} \triangle_{ij}$$

and similarly

$$E(f_{it}^2) = b^{-1} \sum_{t} f_{it}^2$$

$$= \frac{P_i}{b} \text{ (say)}$$

Using (2.4), We have

$$V(\bar{y}_t) = b^{-1} n^{-2} [\sum_{i=1}^{N} P_i\,(y_i - \bar{Y}_w)^2 + \sum_{i}^{N} \sum_{j \neq 1}^{N} \triangle_{ij}\,(y_s - \bar{Y}_w)\,(y_j - \bar{Y}_w)\,]$$

$$= n^{-2} b^{-1}\,(Y - \bar{Y}_w)'\, D_p\,(Y - \bar{Y}_w) + (Y - \bar{Y}_w)'\, D_o\,(Y - \bar{Y}_w)] \quad \ldots(2.5)$$

where D_P is a $N \times N$ diagonal matrix with elements $P_1, P_2, \ldots P_N$, D_O is $N \times N$ symmetric matrix with $O's$ in diagonal and $\triangle_{ij}$ $(i \neq j)$ else where and $(Y-\bar{Y}_w)' = [(y_1-\bar{y}_w), (y_2-\bar{Y}_w), \ldots, (y_N-\bar{Y}_w)]$.

(*a*) When we consider without replacement sampling scheme, f_{ij} takes value either 1 or 0 and hence $E(f_{it}^2)=E(f_{it})=R_i$ and we have

$$V(\bar{y}_t)_{wor}=n^{-2}b^{-1}\,[(Y-\bar{Y}_w)'D_R\,(Y-\bar{Y}_w)+(Y-\bar{Y}_w)'D_o(Y-\bar{Y}_w)] \quad \ldots(2.6)$$

(*b*) When we consider equal probability sampling without replacement,

$$R_i=R, \text{ for all } i \text{ and } \bar{Y}_w=\bar{Y} \text{ and hence}$$

$$V(\bar{y}_t)_{wor\ equp}=n^{-2}b^{-1}\,[R\,(Y-\bar{Y})'\,I\,(Y-\bar{Y})+(Y+\bar{Y})'\,D_o\,(Y-\bar{Y})]. \quad \ldots(2.7)$$

Assuming further that a pair of units occur together constant number of times in the sample design S^*, *i.e.*, $\triangle_{ij}=\triangle$ for all i and j, we have

$$V(\bar{y}_t)_{wor\ equp}=n^{-2}b^{-1}\,[(R-\triangle)\,(Y-\bar{Y})'I(Y-\bar{Y})+\triangle(Y-\bar{Y})'\,E(Y-\bar{Y})],$$

where

E is a $N \times N$ matrix with one as elements.

$$=n^{-2}b^{-1}\,(R-\triangle)\,[Y-\bar{Y})'I\,(Y-\bar{Y})] \quad \ldots(2.8)$$

This is same as the variance of the sample mean of a simple random sampling without replacement scheme.

(*c*) The variance given in (2.6) for without replacement sampling and unequal R_i is the M.S.E. of a biased estimate of $\bar{Y}$. Instead, when we use $\bar{z}_t$ as an unbiased estimate of $\bar{Y}$, the $V(\bar{z}_t)$ is given by

$$V(\bar{z}_t)=n^{-2}b^{-1}\,[(Z-\bar{Y})'D_R(Z-\bar{Y})+(Z-\bar{Y})'D_O\,(Z-\bar{Y})]. \quad \ldots(2.9)$$

The variance given above is same as the variance of $H-T$ estimate for pps sampling without replacement.

ESTIMATE OF THE VARIANCES.

Estimate of the variance for worequp scheme with constant $\triangle_{ij}\ (=\triangle)$

From (2.8) we have $V(\bar{y}_t)=n^{-2}b\,(R-\triangle)\,[Y'IY-N\bar{Y}^2]$. Now, since the inclusion probability of an unit is equal to R/b, the unbiased estimate of $Y'IY$ and $N\bar{Y}^2$ are provided by $(bR^{-1})\ y'Iy$ and $N\,[\bar{y}^2-v(\bar{y}_t)]$ respectively, where $y'=(y_1, y_2, \ldots, y_n)$, the measurements of the units in the selected sample s_t. Hence, the estimate measurements of the units in the selected sample s_t. Hence, the estimate of $V(\bar{y}_t)$ is given by

$$v(\bar{y}_t)=(nN\triangle)^{-1}\,(R-\triangle)\,(y-\bar{y})'\,I(y-\bar{y}). \qquad ...(3.1)$$

Estimates of the variance for unequal probability without replacement sampling scheme.

The $V(\bar{z}_t)$ given in (2.9) can be written as

$$V(\bar{z}_t)=(n^{-2}b^{-1})\,(Z-\bar{Y}')\,T(Z-\bar{Y}),$$

where T is a $N\times N$ symmetric matrix of element $t_{ij}=R_i$ or $\triangle_{ij}$ for $i=j$ or $i\neq j$ respectively.

$$=(n^{-2}b^{-1})\,(Z'TZ)-\bar{Y}^2.$$

Since $\bar{Y}=(nb)^{-1}\left[\sum_{i=}^{N} R_iZ_i\right]$, we have $\bar{Y}^2=(nb)^{-2}\left[(RZ)'\,E(RZ)\right]$.

Hence,

$$V(\bar{z}_t)=b^{-1}n^{-2}\,[z'TZ-b^{-1}\,(RZ)'E\,(RZ)]$$

As the inclusion probability of ith unit and ith and jth in a sample are R_i/b and $\triangle_{ij}/b$ respectively, the estimate of $V(\bar{z}_t)$ is given by

$$v(\bar{z}_t)=n^{-2}\,[z'E_{nxn}z-\,b^{-1}(Rz)'L_{nxn}(Rz)], \qquad ...(3.2)$$

where $z'=(z_1, z_2, ...z_n)$, $(Rz)'=(R_1z_1, R_2z_2, ..., R_nz_n)$ and L_{nxn} is $anxn$ symmetric matrix of elements $1_{ij}=R^{-1}$ or $\triangle^{-1}_{ij}$ for $i=j$ and $i\neq j$ respectively. The $v(\bar{z}_t)$ given above is same as the estimate of variance of H.T estimate for pps sampling without replacement.

CONCLUSION

The variance of the estimates of the population means of usual equal probability and unequal probability sampling schemes can very easily be obtained by the method described above after identifying the incomplete block design which corresponds to the sample space for the sampling method.

If a specific set of samples be taken from the sample space S^* to form a subsample space s^*, such that $R'_i/\triangle'_{ij}$ of s^* are equal to $R_i/\triangle_{ij}$ of the sample space S^*, then also the above result holds true without any change. This shows that for a sampling scheme without replacement the sample space may suitably be constructed so as to reduce the total number of samples in it and still as a result of sampling from such a sub-space, the population mean can be estimated with the same variance as in case of simple random sampling without replacement. As a matter of fact blocks of balanced incomplete design provides such a sample space, when the number of treatment v is taken as N, the block size k is taken on n and b, r and λ are taken as b, R and $\triangle$ respectively.

Naturally no one will like to construct or use the incidence matrix of a bibd for drawing a sample instead of using the simple random sampling mechanism without replacement. But the object of showing this is to bring forth the relationship between experimental design with that of sampling design plan, which will simplify the procedure, of teaching sampling methods. Moreover a general formulation of various results needed from samples has been attempted here so that the results for any specific scheme can be obtained as a particular case from the general results presented here.

REFERENCES

Das, M.N. and Mohanty, S. (1973): On pps sampling without replacement ensuring selection probabilities exactly proportional to size. *Aust. Jour. Stat.* Vol. 15.

Mohanty, S. (1977): Sampling with repeated units. *Sankhya*, Vol. 39 series C, pt. 1.