SOME SAMPLING SYSTEMS PROVIDING UNBIASED RATIO ESTIMATORS

By N. S. NANJAMMA, M. N. MURTHY, *and* V. K. SETHI
Indian Statistical Institute, Calcutta

SUMMARY. In this paper modifications of many of the selection procedures commonly adopted in practice, namely, equal probability sampling, varying probability sampling, stratified sampling and multi-stage sampling have been proposed, which, while retaining the form of the usual ratio estimators, make them unbiased. For many of the situations commonly met with in practice, this modification of a given sampling scheme consists essentially in first selecting one unit with probability proportional to its value of the characteristic occurring in the denominator of the ratio and then the remaining units in the sample according to the original scheme of sampling. The expressions for unbiased variance estimators of the unbiased ratio estimators have been given for some of the more important sampling schemes. Further the selection and estimation procedures which provide unbiased ratio estimators in the case of a certain general class of population parameters together with the expressions for its sampling variance and variance estimators have also been considered.

1. Introduction

As the relationship between two characteristics is usually of much interest, estimation of ratios of certain population parameters has become quite important in a large number of surveys. The method of ratio estimation is also being used to estimate population totals, since a ratio estimator is more efficient than the conventional unbiased estimator under certain circumstances not uncommon in actual practice. The usual procedure of using the ratio method in estimating any population ratio or total has been to take the ratio of unbiased estimators of the numerator and the denominator and in the latter case multiply it by the population total of the supplementary variate taken in the denominator. A disadvantage of this method is that the estimator so obtained is biased for many of the selection procedures commonly adopted in surveys. Further a completely satisfactory (at least to the present authors) treatment of the errors and biases of a ratio estimator is not yet available. For small samples, at least, the bias is not likely to be small.

In recent years attempts have been made to give selection and estimation procedures which provide unbiased ratio estimators. Lahiri (1951) has given a method of selecting a sample with probability proportional to its total size (pps) (sum of the sizes of the units in the sample) which is essentially similar to his method of selecting a unit with pps, namely, of selecting a unit with equal probability and including that unit in the sample if a number chosen at random from one to an upper bound of the units is less than or equal to the size of the selected unit. By 'size' here is meant the value of the supplementary variate under consideration. Obviously this method avoids the need for completely enumerating all possible samples and finding their total sizes and the cumulated sizes. Once a sample is chosen with pps it is easy to obtain an unbiased ratio estimator. The disadvantage of the selection procedure given by Lahiri is that it involves rejection of some draws.

Midzuno (1952) and Sen (1952) have independently given a simple procedure for obtaining a sample with pps. Their method consists in selecting one unit with pps and the rest with equal probability without replacement from the remaining units of the population. It may be observed that Lahiri's method of selecting one unit with pps could profitably be used in the selection procedure given by Midzuno and Sen.

In the case of stratified sampling Lahiri has pointed out that his method could be applied to select a sample with probability proportional to $\sum_{s=1}^{k} N_s \bar{x}_s$ where k is the number of strata, N_s the number of units in the s-th stratum and $\bar{x}_s$ the s-th stratum sample mean of the supplementary variate under consideration with a view to get an unbiased ratio estimator. Des Raj (1954) has given the expressions for the variance and an unbiased variance estimator of the ratio estimator in the case of a multi-stage design where the sample of first stage units is selected with pps.

So far the selection procedures providing unbiased ratio estimators have been given only for simple designs and that too for a very restricted class of parameters. In the next few sections, modifications of many of the selection procedures commonly adopted in practice, namely, equal probability sampling, varying probability sampling, stratified sampling and multi-stage sampling, have been given which, while retaining the form of the usual biased ratio estimators, make them unbiased. The expressions for the unbiased variance estimator for some of the more important cases of ratio estimators are also given. Further the selection and estimation procedures for obtaining unbiased ratio estimators in the case of a certain general class of population parameters together with the expressions for the sampling variance and the variance estimator are given in the last few sections.

For many of the situations commonly met with in practice, the modification of a given sampling scheme referred to above which provides unbiased ratio estimator consists essentially in first selecting one unit with probability proportional to its value of the variate occurring in the denominator of the ratio and then the remaining units in the sample according to the original scheme of sampling. For many of the sampling schemes considered in this paper it might be expected that in large samples the bias of the conventional ratio estimator is unlikely to be large, since the form of the ratio estimator is the same in the case of the biased and the unbiased ratio estimators and the sample based on the original sampling scheme and that on the modified scheme could be made the same but for a difference of one unit at the most.

In this paper the estimator and its variance estimator have been given in the case of estimating the ratio $R = \frac{Y}{X}$ where Y and X are the population totals for two characters. The ratio estimator and its variance estimator for estimating Y can be obtained by multiplying the corresponding estimators in the case of estimation of R by X and X^2 respectively.

2. Equal probability sampling

In the case of *unstratified unistage sampling with equal probability without replacement*, as has been mentioned earlier, Midzuno and Sen have suggested the method of selecting one unit with probability proportional to x(ppx), where x is the value of the variate occurring in the denominator of the ratio and the rest of the $n-1$ units in the sample from the remaining $N-1$ units in the population with equal probability without replacement. The probability of getting a particular sample s by this approach is given by

$$P(s) = \frac{1}{\binom{N}{n}} \frac{\bar{x}}{\bar{X}} \qquad \dots \quad (2.1)$$

where $\bar{x}$ and $\bar{X}$ are the sample and the population means respectively. Hence the estimator

$$\hat{R} = \frac{\bar{y}}{\bar{x}} \qquad \dots \quad (2.2)$$

where $\bar{y}$ is the sample mean of the variate y is an unbiased estimator of the ratio $R = Y/X$. Though this estimator resembles the usual ratio estimator in the case of equal probability sampling without replacement, this is unbiased while the latter is not. An unbiased estimator of the variance of $\hat{R}$ given in (2.2) is given by

$$\hat{V}(\hat{R}) = \hat{R}^2 - \frac{\sum_{i=1}^{n} y_i^2 + 2\frac{N-1}{n-1} \sum_{i>j}^{n} y_i y_j}{N n \bar{x} \bar{X}}. \qquad \dots \quad (2.3)$$

It can be seen that the efficiency of the unbiased estimator given in (2.2) will be greater than, equal to, or less than that of the corresponding biased estimator according as

$$\rho\left(\frac{\bar{y}^2}{\bar{x}}, \bar{x}\right) \lesseqgtr 0. \qquad \dots \quad (2.4)$$

The modification of the procedure of *sampling with equal probability with replacement* which provides an unbiased ratio estimator would be to select one unit with ppx, replace it and then select the rest of the $(n-1)$ units from the whole population with equal probability with replacement at each draw. With this selection procedure the ratio estimator given in (2.2) is unbiased for estimating the population ratio R, since the probability of getting a particular sample in this case is

$$P(s) = \frac{1}{N^n} \cdot \frac{n!}{\prod_{i=1}^{\nu} \lambda_i!} \cdot \frac{\bar{x}}{\bar{X}}, \qquad \dots \quad (2.5)$$

where λ_i is the number of repetitions of the i-th unit and ν is the number of distinct units in the sample. The sampling variance and an unbiased variance estimator of ratio estimator in this case would be different from those in the case of sampling with equal probability without replacement. The variance estimator is given by

$$\hat{V}(\hat{R}) = \hat{R}^2 - \frac{\sum_{i=1}^{\nu} \lambda_i(\lambda_i - 1)y_i^2 + 2\sum_{i>j}^{\nu} \lambda_i\lambda_j y_i y_j}{n(n-1)\bar{X}^2}. \quad \ldots \quad (2.6)$$

In the case of *sampling with equal probability systematically*, an unbiased ratio estimator could be obtained by considering each unit as made up of n sub-units with each sub-unit of the i-th unit having the size $\frac{X_i}{n}$ and selecting one sub-unit with ppx and the other sub-units in the sample systematically proceeding cyclically with the sub-unit selected first as the random start and N as the sampling interval. The probability of getting a particular sample s is given by

$$P(s) = \frac{\bar{x}}{\bar{X}}. \quad \ldots \quad (2.7)$$

With this probability scheme the estimator given in (2.2) is an unbiased estimator of R. The variance of the ratio estimator in this case is different from those of estimators based on equal probability selection with or without replacement. Since the selection is being done systematically in this sampling scheme, it would not be possible to get an unbiased variance estimator of the unbiased ratio estimator from a single sample.

It is to be noted that even if the values of the variate x coming in the denominator are not known for all the units in the population at the time of selection, it is possible to select one unit with ppx by adopting Lahiri's method provided an upper bound of the values of x which is not much greater than the maximum value is known. The population total of the variate x would be necessary only if the population total Y is to be estimated using x as the supplementary variate.

3. VARYING PROBABILITY SAMPLING

The ratio of an unbiased estimator of Y to that of X based on pps sampling scheme, the size being the value of a variate z related to both the characteristics x and y, is known to be biased for the population ratio $R = Y/X$. In this section are given the selection procedures which provide unbiased ratio estimators corresponding to the usual biased ratio estimators in the case of sampling with pps with replacement, pps without replacement and pps systematically.

The modification of the *pps with replacement scheme* consists in selecting first one unit with ppx, replacing it and then selecting the rest of the $(n-1)$ units from the whole population with ppz with replacement. The probability of getting a particular sample s by this procedure is given by

$$P(s) = \frac{n! \prod_{i=1}^{\nu} p_i^{\lambda_i}}{X \prod_{i=1}^{\nu} \lambda_i !} \cdot \left(\frac{1}{n} \sum_{i=1}^{n} \frac{x_i}{p_i} \right) \quad \ldots \quad (3.1)$$

where λ_i and N are as in (2.5) and $p_i = \frac{z_i}{Z}$ where $Z = \sum_{i=1}^{N} z_i$. It may be verified that in this case an unbiased estimator of R is given by

$$\hat{R} = \frac{\frac{1}{n}\sum_{i=1}^{n} \frac{y_i}{p_i}}{\frac{1}{n}\sum_{i=1}^{n} \frac{x_i}{p_i}}. \qquad \ldots \quad (3.2)$$

It may be noted that the expression for the unbiased ratio estimator is the same as that of the usual biased ratio estimator in the case of pps with replacement sampling. For in the latter case $\frac{1}{n}\sum_{i=1}^{n}\frac{y_i}{p_i}$ and $\frac{1}{n}\sum_{i=1}^{n}\frac{x_i}{p_i}$ are unbiased estimators of Y and X respectively. Of course, the variance and variance estimator would be different in the two cases.

If in the above selection procedure the units sampled are not replaced before the next and subsequent draws, we get the modified *pps without replacement scheme* which provides an unbiased ratio estimator. But in practice, it is difficult to compute the estimate as the computations involved are quite heavy except in some special cases. Two such special cases have been considered to illustrate the method.

(i) *ppx and ppz of the remaining* ($n = 2$). The modification of this procedure consists in selecting first one unit with ppx and another unit from the remaining $(N-1)$ units with ppz (z being a size other than x). The probability of getting a particular sample s $(x_1\, x_2)$ is given by

$$P(s) = \frac{x_1}{X} \cdot \frac{p_2}{1-p_1} + \frac{x_2}{X} \cdot \frac{p_1}{1-p_2}. \qquad \ldots \quad (3.3)$$

It can be seen that this procedure provides the following unbiased estimator of the ratio R.

$$\hat{R} = \frac{\frac{y_1}{p_1}(1-p_2)+\frac{y_2}{p_2}(1-p_1)}{\frac{x_1}{p_1}(1-p_2)+\frac{x_2}{p_2}(1-p_1)} \qquad \ldots \quad (3.4)$$

(ii) *ppx, ppz of the remaining and then equal probabilities.* The modification of this procedure consists in following up the procedure explained for case (i) above by selecting the rest of $(n-2)$ units with equal probability without replacement from the remaining $(N-2)$ units in the population. This selection procedure makes the following estimator unbiased for the ratio R

$$\hat{R} = \frac{\sum_{i=1}^{n} \frac{y_i}{1-p_i}\left(\sum_{j\neq i}^{n} p_j\right)}{\sum_{i=1}^{n} \frac{x_i}{1-p_i}\left(\sum_{j\neq i}^{n} p_j\right)}, \qquad \ldots \quad (3.5)$$

since in this case the probability of getting a particular sample s is given by

$$P(s) = \frac{\sum_{i=1}^{n} \frac{x_i}{1-p_i} \left(\sum_{j \neq i}^{n} p_j \right)}{X \binom{N-2}{n-2}} . \qquad \ldots \quad (3.6)$$

Before giving the selection procedure which provides an unbiased ratio estimator corresponding to the usual biased ratio estimator in the case of *pps systematic sampling*, the method of sampling with probability proportional to size (say the value of a character z) systematically will be briefly explained, since this has not become, as yet, well known. Let $Z_1\ Z_2 \ldots Z_N$ be the sizes of the units in the population. Suppose the i-th unit is made up of nZ_i sub-units, each having the value $\frac{Y_i}{nZ_i}$ for the variate y. The procedure of pps systematic sampling consists in selecting a random number from 1 to $Z\left(=\sum_{i=1}^{N} Z_i\right)$. The sub-unit having that number is selected in the sample together with every subsequent Z-th sub-unit. As the total number of sub-units is nZ there will be n sub-units in the sample. Let the sample be $(z_1, z_2 \ldots z_n)$. An unbiased estimator of the population total Y is given by

$$\hat{y} = \frac{1}{n} \sum_{i=1}^{n} \frac{y_i}{p_i} \qquad \ldots \quad (3.7)$$

where $p_i = \frac{z_i}{Z}$. It may be noted that this estimator resembles that used in the case of pps with replacement sampling. But the variances of these two estimators are different.

Though the expected number of repetitions in a sample for the i-th unit is np_i in both the pps with replacement scheme and the pps systematic sampling, the numbers of possible repetitions in a sample are different. For instance, in pps with replacement scheme, the i-th unit may occur $0, 1, 2 \ldots n$ times in a sample of size n whereas in pps systematic sampling it occurs either $[np_i]$ or $[np_i]+1$ times. As the randomisation of the number of repetitions is over a smaller range in the case of pps systematic sampling than that in the case of pps with replacement, it is expected that the former method is more efficient than the latter. Further the efficiency of the estimator based on this method could be increased appreciably by effecting a suitable arrangement of the units in the population before selection. Being a systematically drawn sample it would not be possible to estimate the sampling variance unbiasedly from a single sample.

The modification of the above method to provide an unbiased estimator of $R = \frac{Y}{X}$ consists in selecting one sub-unit with probability proportional to x_i/p_i. Then with that as the random start a systematic sample of n sub-units is selected proceeding cyclically with Z as the sampling interval. The probability of getting a particular sample s is given by

$$P(s) = \frac{1}{ZX} \cdot \frac{1}{n} \cdot \sum_{i=1}^{n} \frac{x_i}{p_i}. \qquad \ldots \quad (3.8)$$

This procedure provides the following unbiased estimation of the ratio R

$$\hat{R} = \frac{\frac{1}{n}\sum_{i=1}^{n} \frac{y_i}{p_i}}{\frac{1}{n}\sum_{i=1}^{n} \frac{x_i}{p_i}}. \qquad \ldots \quad (3.9)$$

4. Stratified sampling

Let k be the number of strata and N_i and n_i be the number of units in the population and the sample respectively for the i-th stratum. For stratified simple random sampling without replacement the modification in the selection procedure for getting an unbiased ratio estimator consists of selecting one unit (say the j-th unit in the i-th stratum) from the whole population with ppx, (n_i-1) units from the remaining (N_i-1) units in the i-th stratum and $n_{i'}$ units from $N_{i'}$ units of the i'-th stratum $(i' \neq i)$ with equal probability without replacement. The probability of getting a particular sample s is given by

$$P(s) = \frac{\sum_{i=1}^{k} N_i \bar{x}_i}{X \prod_{i=1}^{k} \binom{N_i}{n_i}} \qquad \ldots \quad (4.1)$$

where $\bar{x}_i$ is the sample mean in the i-th stratum for the variate x. With this procedure an unbiased estimator of the ratio R is given by

$$\hat{R} = \frac{\sum_{i=1}^{k} N_i \bar{y}_i}{\sum_{i=1}^{k} N_i \bar{x}_i}. \qquad \ldots \quad (4.2)$$

An unbiased estimator of the variance of $\hat{R}$ is given by

$$\hat{V}(\hat{R}) = \hat{R}^2 - \frac{1}{X\left(\sum_{i=1}^{k} N_i \bar{x}_i\right)} \left[\sum_{i=1}^{k} \frac{N_i}{n_i} \sum_{j=1}^{n} y_{ij}^2 + \sum_{i=1}^{k} \frac{N_i(N_i-1)}{n_i(n_i-1)} \sum_{\substack{j,j' \\ j \neq j'}}^{n} y_{ij} y_{ij'} + \sum_{i \neq i'}^{n} \frac{N_i N_{i'}}{n_i n_{i'}} \sum_{j=1}^{n_i} \sum_{j'=1}^{n_{i'}} y_{ij} y_{i'j'} \right] \qquad \ldots \quad (4.3)$$

It may be noted that $\hat{R}$ resembles the biased combined ratio estimator of Y. The modifications of sampling schemes with other types of designs in the strata can be given on similar lines with a view to getting unbiased ratio estimators.

5. Two-stage sampling

In the case of a two-stage sampling design with equal probability selection without replacement at each stage, the selection procedure for providing an unbiased ratio estimator consists in selecting one second stage unit from the whole population of second stage units with ppx. If this second stage unit is from the i-th first stage unit, the rest of (n_i-1) second stage units to be sampled from the i-th first stage unit are selected from the remaining (N_i-1) units there with equal probability without replacement. The rest of the sample of $(n-1)$ first stage units is drawn from the remaining $(N-1)$ units with equal probability without replacement. From those selected first stage units the required number of second stage units are selected with equal probability without replacement. In this case the probability of getting a particular sample s is given by

$$P(s) = \frac{\sum_{i=1}^{n} N_i \bar{x}_i}{\binom{N-1}{n-1} \times \prod_{i=1}^{n} \binom{N_i}{n_i}} \qquad \dots \quad (5.1)$$

where $\bar{x}_i$ is the sample mean in the i-th selected first stage unit. This selection procedure provides an unbiased estimator of the ratio $R = \frac{Y}{X}$.

$$\hat{R} = \frac{\sum_{i=1}^{n} N_i \bar{y}_i}{\sum_{i=1}^{n} N_i \bar{x}_i}. \qquad \dots \quad (5.2)$$

The above procedure could easily be extended to the case of sampling designs with more than two stages, as is shown in the next section.

6. Multi-stage design

The principle involved in giving a selection procedure which provides an unbiased ratio estimator in the case of multi-stage sampling is the same as in the cases illustrated earlier. That is, one final stage unit is to be selected with ppx and the rest of the units according to some probability scheme. For the sake of simplicity only the probability scheme where the units are selected with equal probability without replacement at each stage is considered here.

One final stage unit is to be selected first from the whole population with ppx and then the rest of the sample units are to be selected from the remaining units in the universe with equal probability without replacement at each stage. This can be achieved as follows. Suppose there are m stages. One first stage unit (i_1-th) is to be selected with ppx and the other $(n-1)$ units with equal probability without replacement from the remaining $(N-1)$ units. From the first stage unit selected with ppx, one second stage unit is to be selected with ppx and the rest of $(n_{i_1}-1)$ units are to be selected from the remaining $(N_{i_1}-1)$ units with equal probability without replacement. Similarly from the j-th stage unit selected with ppx one $(j+1)$th stage unit is to be selected with ppx and the other $(n_{i_1 i_2 \ldots i_j}-1)$ units are to be selected with equal probability without replacement from the remaining $(N_{i_1 i_2 \ldots i_j}-1)$ units. $(j = 0, 1, 2, (m-1))$. From the first and the subsequent stage units selected with equal probability, the required number of higher stage units are to be selected with equal probability without replacement. The probability of getting a particular sample s is given by

$$P(s) = \frac{\frac{N}{n}\sum_{i_1=1}^{n} N_{i_1}\, \bar{x}_{i_1(m)}}{X \prod_{j=0}^{m-1}\left[\prod_{i_1=1}^{n}\prod_{i_2=1}^{n_{i_1}}\cdots\prod_{i_j=1}^{n_{i_1 i_2 \ldots i_{j-1}}}\binom{N_{i_1 i_2 \ldots i_j}}{n_{i_1 i_2 \ldots i_j}}\right]} \qquad \ldots \quad (6.1)$$

$$\text{where,}\quad \bar{x}_{i_1(m)} = \frac{1}{n_{i_1}}\sum_{i_2=1}^{n_{i_1}}\frac{N_{i_1 i_2}}{n_{i_1 i_2}}\cdots\sum_{i_{m-1}=1}^{n_{i_1 i_2 \ldots i_{m-2}}}\frac{N_{i_1 \ldots i_{m-1}}}{n_{i_1 \ldots i_{m-1}}}\sum_{i_m=1}^{n_{i_1 i_2 \ldots i_{m-1}}} x_{i_1 i_2 \ldots i_m},$$

$x_{i_1 i_2 \ldots i_m}$ being the value of a typical final stage unit. In this case an unbiased estimator of

$$\hat{R} = \frac{\sum_{i_1=1}^{n} N_{i_1}\, \bar{y}_{i_1(m)}}{\sum_{i_1=1}^{n} N_{i_1}\cdot \bar{x}_{i_1(m)}}, \qquad \ldots \quad (6.2)$$

where, $\bar{y}_{i_1(m)}$ has an interpretation similar to that of $\bar{x}_{i_1(m)}$.

7. A generalised estimation procedure

In this section a generalised procedure for estimating unbiasedly certain types of parameters applicable to a large number of sampling designs is given. For the sake of generality it has become necessary to use some notations which are explained below with suitable examples.

Let $\mathscr{X}$ denote a population of finite number of units, say, a universe of N units $u_1\, u_2 \ldots u_N$ and A the class of sets α whose elements belong to $\mathscr{X}$. In such a set the same unit may or may not occur more than once. The class of all point sets and the class of all pairs of units belonging to $\mathscr{X}$ are examples of the class A.

Let the population parameter F be expressible as

$$F = \sum_{\alpha \epsilon A} f(\alpha) \quad \dots \quad (7.1)$$

where $f(\alpha)$ is a single-valued set function defined over the class A and $\sum_{\alpha \epsilon A}$ stands for summation over all sets α belonging to the class A. For example, the population total Y can be expressed as F in (7.1) with α as a point set (u_i) and $f(\alpha)$ as y_i the value of the i-th unit for the character y, and Y^2 can be expressed as F in (7.1) with α as a set with two units $\{u_i, u_j\}$* and with $f(\alpha)$ defined as

$$\begin{aligned} f(\alpha) &= 2y_iy_j \qquad \alpha = \{u_i,\ u_j\}, \qquad i \neq j = 1, 2, \dots N \\ &= y_i^2 \qquad\quad \alpha = \{u_i,\ u_i\}, \qquad i = 1, 2, \dots N \end{aligned}$$

Let a sample ω be drawn from the population $\mathscr{X}$ with probability $P(\omega)$. This ω again is a set whose elements belong to $\mathscr{X}$. It may be noted that the same unit may or may not occur more than once in ω. The class of all such sets will be denoted by Ω which will be the total sample space.

It will be possible to estimate the population parameter F from the sample ω only if each ω contains at least one set α and each set α is contained in at least one ω.

An estimator of the parameter F is given by

$$\hat{F} = \frac{\sum_{\alpha \subset \omega} f(\alpha)\phi(\omega, \alpha)}{P(\omega)} \quad \dots \quad (7.2)$$

where, $\sum_{\alpha \subset \omega}$ stands for the summation over all sets α contained in the sample ω and $\phi(\omega, \alpha)$ is a function of ω and α. This estimator will be unbiased if

$$\sum_{\omega \supset \alpha, \alpha'} \phi(\omega, \alpha) = 1$$

where, $\sum_{\omega \supset \alpha, \alpha'}$ stands for the summation over all samples ω which contain α, since

$$\begin{aligned} E(\hat{F}) &= \sum_{\omega \epsilon \Omega} \sum_{\alpha \subset \omega} f(\alpha)\phi(\omega, \alpha) \\ &= \sum_{\alpha, \alpha' \epsilon A} f(\alpha) f(\alpha') \{ \sum_{\omega \supset \alpha, \alpha'} \psi(\omega, \alpha, \alpha')\} \end{aligned}$$

An unbiased estimator of the variance of F is given by

$$\hat{V}(\hat{F}) = \hat{F}^2 - \frac{\sum_{\alpha, \alpha' \subset \omega} f(\alpha) f(\alpha')\psi(\omega, \alpha, \alpha')}{P(\omega)} \quad \dots \quad (7.3)$$

* The curled brackets { } are used to denote unordered sets, that is, $\{u_i, u_j\}$ and $\{u_j, u_i\}$ are the same.

where, $\sum\limits_{\alpha,\alpha'\subset\omega}$ stands for the summation over all pairs $\{\alpha, \alpha'\}$ contained in the sample ω and $\psi(\omega, , \alpha, \alpha')$ is a function of ω and the pair (α, α') such that

$$\sum_{\omega\supset\alpha,\alpha'} \psi(\omega, \alpha, \alpha') = 1$$

where, $\sum\limits_{\omega\supset\alpha\alpha'}$ stands for the summation over all the samples containing the pair of sets (α, α'), since

$$E\left[\frac{\sum\limits_{\alpha,\alpha'\subset\omega} f(\alpha)f(\alpha')\psi(\omega, \alpha, \alpha')}{P(\omega)}\right]$$

$$= \sum_{\alpha,\alpha'\epsilon A} f(\alpha)f(\alpha') \sum_{\omega\supset\alpha,\alpha'} \psi(w, \alpha, \alpha') = F^2$$

The case where $\phi(\omega, \alpha)$ is taken as $P(\omega/\alpha)$, the conditional probability of getting the sample ω given that the set α has been selected first is of interest as in that case it is possible to verify that for many of the designs in general use, this estimator

$$\hat{F} = \frac{\sum\limits_{\alpha\subset\omega} f(a)P(\omega/\alpha)}{P(\omega)} \qquad \ldots \quad (7.4)$$

reduces to the usual estimators of the parameter. An unbiased estimator of its variance is given by

$$\hat{V}(\hat{F}) = \hat{F}^2 - \frac{\sum\limits_{\alpha,\alpha'\subset\omega} f(\alpha)f(\alpha')P(\omega/\alpha \cup \alpha')}{P(\omega)} \qquad \ldots \quad (7.5)$$

where, $P(\omega/\alpha \cup \alpha')$ is the conditional probability of getting the sample ω given that the units in the union of the two sets α and α' have been selected first. The above variance estimator may take negative values.

An estimator of the variance is possible only if every set $(\alpha \cup \alpha')$ is contained in at least one ω and every ω contains at least one set $(\alpha \cup \alpha')$.

8. Unbiased ratio estimator

The above estimation procedure, an estimator for the ratio R of two parameters F and G which can be expressed as

$$F = \sum_{\alpha\epsilon A} f(\alpha)$$

$$G(\alpha) = \sum_{\alpha\epsilon A} g(\alpha)$$

where $g(\alpha)$ is another single valued set function defined over the class A is given by

$$\hat{R} = \frac{\sum\limits_{\alpha\subset\omega} f(\alpha)\phi(\omega, \alpha)}{\sum\limits_{\alpha\subset\omega} g(\alpha)\phi(\omega, \alpha)}. \qquad \ldots \quad (8.1)$$

This estimator will be unbiased if

$$P(\omega) = \frac{\sum\limits_{\alpha \subset \omega} g(\alpha)\phi(\omega, \alpha)}{\sum\limits_{\alpha \subset A} g(\alpha)} \quad \dots \quad (8.2)$$

since
$$E(\hat{R}) = \sum_{\omega \in \Omega} \frac{\sum\limits_{\alpha \subset \omega} f(\alpha)\phi(\omega, \alpha)}{\sum\limits_{\alpha \subset \omega} g(\alpha)\phi(\omega, \alpha)} P(\omega).$$

If $g(\alpha)$'s are either all positive or all negative the above form of $P(\omega)$ can be obtained by first selecting a set α with probability proportional to $g(\alpha)$ and then drawing the rest of the units with some probability scheme. In this case the probability of getting ω is

$$P(\omega) = \frac{\sum\limits_{\alpha \subset \omega} g(\alpha)P(\omega/\alpha)}{\sum\limits_{\alpha \in A} g(\alpha)}$$

This shows that if in the general case $\phi(\omega,\alpha)$ is taken as $P(\omega/\alpha)$, the estimator given in (8.1) becomes unbiased for the ratio

$$\hat{R} = \frac{\sum\limits_{\alpha \subset \omega} f(\alpha)P(\omega/\alpha)}{\sum\limits_{\alpha \subset \omega} g(\alpha)P(\omega/\alpha)}. \quad \dots \quad (8.3)$$

An unbiased ratio estimator of F is given by $\hat{F} = \hat{R}.\ G.$ ... (8.4)

If $P(\omega/\alpha)$ is independent of the set α, the estimator becomes

$$\hat{R} = \frac{\sum\limits_{\alpha \subset \omega} f(\alpha)}{\sum\limits_{\alpha \subset \omega} g(\alpha)} \quad \dots \quad (8.5)$$

and if $\dfrac{P(\omega/\alpha)}{h(\alpha)}$ is independent of the set α,

$$\hat{R} = \frac{\sum\limits_{\alpha \subset \omega} f(\alpha)h(\alpha)}{\sum\limits_{\alpha \subset \omega} g(\alpha)h(\alpha)}. \quad \dots \quad (8.6)$$

An unbiased estimator of the variance of $\hat{R}$ is given by

$$\hat{V}(\hat{R}) = \hat{R}^2 - \frac{\sum\limits_{\alpha, \alpha' \subset \omega} f(\alpha)f(\alpha')P(\omega/\alpha \cup \alpha')}{G \sum\limits_{\alpha \subset \omega} g(\alpha)P(\omega/\alpha)} \quad \dots \quad (8.7)$$

It is possible that F and G can be expressed as sums of set functions defined over more than one class of sets, that is,

$$F = \sum_{x \epsilon A} f_1(x) = \sum_{x' \epsilon A'} f_2(x') = \cdots\cdots$$

$$G = \sum_{x \epsilon A} f_1(x) = \sum_{x' \epsilon A} f_2(x') = \cdots\cdots$$

For each such expression we can give a sampling procedure providing an unbiased ratio estimator. From the point of view of operational convenience, it is preferable to take that class of sets which contains the smaller sets. The size of a set is judged by the number of units it contains. Two examples are given to illustrate the point.

(a) The population total X can be expressed in the following two ways

$$X = \sum_{i=1}^{N} X_i = S \frac{\sum_{i=1}^{n} x_i}{\binom{N-1}{n-1}}$$

where S stands for the summation over all sets of n distinct units. In this case the former is to be preferred to the latter because in the former only one unit is to be selected with ppx whereas in the latter case n units are to be drawn with probability proportional to their total size.

(b) The population variance $\sigma^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{X})^2$ where $\bar{X} = \frac{1}{N} \sum_{i=1}^{N} X_i$ can

again be expressed in the following two ways

$$\sigma_2 = \frac{1}{N} \cdot \frac{N-1}{n-1} \cdot \frac{1}{\binom{N}{n}} \cdot S \sum_{i=1}^{n} (x_i - \bar{x})^2$$

$$= \frac{1}{N^2} \cdot S'(x_i - x_j)^2.$$

Where S stands for the summation over all sets of n units, $\bar{x}$ is the sample mean and S' stands for the summation over all sets of two units. Here the latter is to be preferred to the former because in the latter case only two units as compared to n units in the former case are to be selected with probability proportional to their measure of size.

It may be verified that all the cases discussed in earlier sections are particular cases of the generalised unbiased ratio estimator considered here. The procedure explained above will be illustrated by applying it to the question of getting an unbiased estimator of the regression coefficient.

9. REGRESSION COEFFICIENT

$$\beta = \frac{\sum\limits_{i=1}^{N} (Y_i - \bar{Y})(X_i - \bar{X})}{\sum\limits_{i=1}^{N} (X_i - \bar{X})^2}. \quad \ldots \quad (9.1)$$

The numerator and the denominator can be expressed as follows

$$F = \frac{1}{N} \sum_{i=1}^{N} \sum_{j>i} (Y_i - Y_j)(X_i - X_j)$$

$$G = \frac{1}{N} \sum_{i=1}^{N} \sum_{j>i} (X_i - X_j)^2$$

Thus the parameters F and G are sums of set functions defined over the class of sets containing only two elements. In the terminology of section 8,

$$f(\alpha) = \frac{1}{N} (Y_i - Y_j)(X_i - X_j)$$

$$g(\alpha) = \frac{1}{N} (X_i - X_j)^2$$

where α is a set containing two elements.

The selection procedure consists in selecting a pair of units with probability proportional to $(X_i - X_j)^2$ and the rest $(n-2)$ units with equal probability without replacement from the remaining $(N-2)$ units. The conditional probability of getting the sample ω given that the pair (i, j) is selected first is

$$P(\omega/ij) = \frac{1}{\binom{N-2}{n-2}}. \quad \ldots \quad (9.2)$$

This is independent of the pair of units selected first.

Hence an unbiased estimator of β is given by

$$\hat{\beta} = \frac{\sum\limits_{i=1}^{n} \sum\limits_{j>i} (y_i - y_j)(x_i - x_j)}{\sum\limits_{i=1}^{n} \sum\limits_{j>i}^{n} (x_i - x_j)^2}$$

i.e.

$$\hat{\beta} = \frac{\sum\limits_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})}{\sum\limits_{i=1}^{n} (x_i - \bar{x})^2}$$

where $\bar{x}$ and $\bar{y}$ are the sample means. The variance and the variance estimator can be got by referring to section 8.

For selecting one pair of units, with probability proportional to $(X_i - X_j)^2$, the following procedure may be adopted:

(i) two random numbers should be selected from 1 to N. (say, ij);

(ii) a pair of random numbers should be selected from 1 to Max. $|X_i - X_j|$ (= range of x);

(iii) if both these numbers are less than or equal to $|X_i - X_j|$ the pair (i, j) is accepted, otherwise it is rejected;

(iv) if a pair is rejected, the operation is to be repeated starting from (1).

It may be noted that unbiased estimators of ρ^2 (square of the correlation coefficient of x and y) and $\beta_2 \left(= \frac{\mu_4}{\mu_2^2} \right)$ can be got by selecting a set of four elements first with suitable probabilities and the rest with any probability scheme.

10. Two-phase sampling

For estimating the parameter F unbiasedly using a ratio estimator with $g(\alpha)$ as supplementary information, it is necessary to know the value of G. If the value of G is not known in advance and if it is easier and less costly to observe $g(\alpha)$ than $f(\alpha)$, then a two-phase design may be used to get an unbiased ratio estimator of F.

The procedure consists in selecting a large sample S from the whole population with some probability scheme and observing the value of $g(\alpha)$ for all sets $\alpha \epsilon A$ which are contained in S. A sample ω is drawn from S by first selecting a set $\alpha \epsilon A$ with probability proportional to $g(\alpha)P(S/\alpha)$ and then selecting the rest of the sample with some probability scheme. In this case an unbiased ratio estimator of F is given by

$$\hat{F} = \frac{\sum\limits_{\alpha \subset \omega} f(\alpha)P(S/\alpha).P(\omega/S,\alpha)}{\sum\limits_{\alpha \subset \omega} g(\alpha)P(S/\alpha)P(\omega/S,\alpha)} \times \frac{\sum\limits_{\alpha \subset \omega} g(\alpha)P(S/\alpha)}{P(S)}$$

where $P(\omega/S,\alpha)$ denotes the probability of selecting the sample ω from S given that α was selected first in the process. This probability refers to the second-phase sampling.

The authors wish to thank Prof. D. B. Lahiri and Dr. D. Basu for their constant encouragement and advice.

REFERENCES

DES RAJ (1954): Ratio estimation in sampling with equal and unequal probabilities. *Jour. Ind. Soc. Agric. Stat.*, 6, No. 2, 127-138.

LAHIRI, D. B. (1951): A method of sample selection providing unbiased ratio estimates. *Bull. Int. Stat. Inst.*, 33, part 2, 133-140.

MIDZUNO, H. (1950): An outline of the theory of sampling systems. *Ann. Inst. Stat. Math.*, 1, 149-156.

——— (1952): On the sampling system with probability proportional to sum sizes. (*Ibid*) 3, 99-107.

SEN, A. R. (1952): Present status of probability sampling and its use in the estimation of characteristics, (abstract). *Econometrica*, 20, 103.

Paper received : October, 1956.

Revised : March, 1959.