SOME CONTRIBUTIONS TO THE THEORY OF SAMPLING

SOME CONTRIBUTIONS TO THE THEORY OF SAMPLING

by

Pramod Kumar Pathak

A thesis submitted to the Indian Statistical Institute for the degree of Doctor of Philosophy

Research and Training School
Indian Statistical Institute
Calcutta.

April, 1961.

ACKNOWLEDGEMENT

I wish to express my sincere thanks to Dr. D. Basu for his inspiring guidance and constant encouragement throughout; and also to Professor C. R. Rao for his great interest in the work and for providing research facilities.

Pramod Kumar Pathak

TABLE OF CONTENTS

CHAPTER I

INTRODUCTION

The work presented here has originated from a pioneering paper by Basu[1] and is essentially an extension of his ideas to different problems in Sample Surveys. The whole work mainly consists in deriving estimators uniformly better than those usually adopted in with replacement sampling schemes.

In with replacement sampling schemes, the 'order-statistic' (distinct sample units arranged in an ascending order of their unit-indices) forms a sufficient statistic. Therefore, if any estimator (e.g., say of the population mean) does not depend on the 'order-statistic', it can be uniformly improved by the use of the Rao-Blackwell theorem. The author has not hesitated to use this powerful theorem to derive improved estimators - If T is a sufficient statistic, for any convex (downwards) loss function, an estimator uniformly better than $g(s)$ (where $g(s)$ is some estimator based on the sample s) is given by

$$E[g(s) \mid T] = \frac{\sum_{s \supset T} g(s) P(s)}{\sum_{s \supset T} P(s)},$$

1. Cf. Basu, D., 'On Sampling With and Without Replacement', Sankhya, Vol.20(1958), pp. 287-294.

where the summation is taken over all samples giving rise to the statistic T and $P(S)$ stands for the probability of selection of the sample S.

The whole thesis is divided into eight chapters and two appendices. Chapter II has been devoted to the problem of finding moments of distinct units that appear in a sample; this chapter has been very helpful in getting new results in subsequent chapters which would have been, otherwise, difficult to obtain. It may be pointed out that it was this chapter which ultimately led the author to write down the thesis. It has been the author's endeavour to present a self-contained treatment of the problems discussed herein. It is for this and for the purpose of completeness that some problems already considered by other authors, are also given in a simplified form.

The problems with which we have been mainly concerned, are the estimation of the population mean $\bar{Y}$ (or total), its square and the population variance. The problem of finding unbiased estimator of the square of the population mean arose while finding unbiased variance estimators of the estimators of $\bar{Y}$. The following technique for finding unbiased variance estimators has been used :- if t is an unbiased estimator of $\bar{Y}$, an unbiased estimator of $V(t)$ is given by[1]

1. $V(t)$ denotes the variance of t.

$$v(t) = t^2 - \text{est}\,(\bar{Y}^2),$$

where $\text{est}\,(\bar{Y}^2)$ is some unbiased estimator of $\bar{Y}^2$.

Variance expressions for the estimators of $\bar{Y}$ are derived. In those cases, where it was difficult to derive, we have given their unbiased estimators.

It has been observed that in sampling schemes with unequal probabilities of selection, the best estimators are unwieldy to compute in large samples. Consequently, other (though, of course, somewhat less efficient) improved estimators that are easy to compute in practice are given. Further, as the variance estimators of the improved estimators are also complicated, it is suggested to use the variance estimators of the original estimators from which the improved estimators were derived (which are mostly simple). As these estimators will over-estimate the actual variances, we will always be on the safe side.

We now start to consider in detail the contents of the thesis chapter-wise.

In Chapter II, moments of distinct units that appear in a sample, are derived under any sampling scheme. If we denote the number of distinct units by ν, it is proved that

$$E\left[\frac{1}{\nu}\right] = \frac{1}{N}\left[1 + \frac{1}{(N-1)}\Sigma_1 q_1 + \frac{1.2}{(N-1)(N-2)}\Sigma_1 q_{12} + \frac{1.2.3}{(N-1)(N-2)(N-3)}\Sigma_1 q_{123}\right.$$

$$\left. + \ldots\ldots + \frac{1.2.\ldots(N-1)}{(N-1)(N-2)\ldots 1}\Sigma_1 q_{12\cdots N-1}\right],$$

where $q_{12\ldots m}$ denotes the probability of exclusion of units 1, 2, ..., and m from the sample and the summation Σ_1 is taken over all possible combinations. This expression was required while deriving the variance expression for the average of distinct units observed in a simple random sample (with replacement). Basu (4), and Des Raj and Khamis (14) have shown that this average is a better estimator of the population mean than the usual overall average. Exact expressions for $E(\nu^t)$ $(t \neq 0)$ and $E[f(\nu)]$ (where $f(\nu)$ is a function of ν satisfying some regularity conditions) are also derived. Throughout this chapter, we have restricted ourselves to sampling schemes for which $P[\nu \geq 1] = 1$. For simple random sampling (with replacement), a table of values of $E\left(\frac{n}{\nu}\right)$, where n is the sample size, is provided for n upto fifty and sampling fractions 0.001 (0.001) 0.010 (0.005) 0.100.

In Chapter III, simple random sampling (with replacement) is considered in detail. In addition to showing that the average of distinct units ($\bar{\bar{y}}_\nu$) is better than the usual overall average ($\bar{\bar{y}}$), an exact expression for the variance of the average of distinct units is given. Several other estimators of the population mean are

suggested and their relative efficiencies are compared. A direct comparison between the variances of Horvitz — Thompson estimator $\left(\frac{\nu}{E(\nu)} \bar{y}_\nu \right)$, and $\bar{y}_\nu$ is made. It is found that $\bar{y}_\nu$ is better than Horvitz — Thompson estimator if the coefficient of variation of the population is less than a given quantity and worse otherwise. Some admissible properties of $\bar{y}_\nu$ are proved. Unbiased estimators of $V(\bar{y}_\nu)$ are given. A numerical example, using the populations of Yates and Grundy (46), is also given to study the relative performance of the variance estimators.

The usual estimator of the population variance, $\sigma^2 = \frac{1}{N} \sum_{j=1}^{N} (Y_j - \bar{Y})^2$, is given by the sample variance

$$s^2 = \frac{1}{(n-1)} \sum_{i=1}^{n} (y_i - \bar{y})^2.$$

It is proved that for any convex (downwards) loss function, an estimator uniformly better than s^2 is given by

$$s^2 = \begin{cases} \left[\dfrac{e_\nu(n) - e_\nu(n-1)}{e_\nu(n)} \right] s_d^2 & \text{if } \nu > 1, \\ 0 & \text{otherwise,} \end{cases}$$

where s_d^2 is the sample variance based on ν distinct units (with divisor $\nu - 1$) and

$$e_\nu(n) = \nu^n - \binom{\nu}{1} (\nu - 1)^n + \ldots (-)^{\nu-1} \binom{\nu}{\nu-1} 1^n.$$

A table of values of $\frac{c_\nu(n-1)}{c_\nu(n)}$ is given at the end of the chapter for all ν and $n-1$ to 50. Finally, for the purpose of estimation of $\bar{Y}$, a comparison between simple random sampling schemes (with and without replacement) is made.

Chapter IV is devoted to sampling with unequal probabilities of selection. Let us consider a population of N units. Let

$$z_j = Y_j / P_j$$

be the z-value of the j-th population unit, P_j being its probability of selection. From a sample of size n, the usual estimator of the population total

$$Y = \sum_{j=1}^{N} Y_j$$

is given by

$$\bar{z} = \frac{1}{n} \sum_{i=1}^{n} z_i ,$$

where z_i is the z-value of the i-th sample unit. If $x_{(1)},\ldots,x_{(\nu)}$ are the distinct units observed in the sample, a better estimator of Y is given by

$$\bar{z}_\nu = \sum_{i=1}^{\nu} c_{(i)}\, z_{(i)} ,$$

where

$$c_{(i)} = \frac{p_{(i)}\left[(p_{(1)}+\ldots+p_{(\nu)})^{n-1} - \Sigma_1^i (p_{(1)}+\ldots+p_{(\nu-1)})^{n-1} + \ldots (-)^{\nu-1} p_{(i)}^{n-1}\right.}{\left[(p_{(1)}+\ldots+p_{(\nu)})^{n} - \Sigma_1 (p_{(1)}+\ldots+p_{(\nu-1)})^{n} + \ldots (-)^{\nu-1} \Sigma_1 p_{(1)}^{n}\right]}$$

the summations Σ_1 and Σ_1^i stand for all combinations of p's and all combinations of p's containing $p_{(i)}$ (chosen out of $p_{(1)}, p_{(2)}, \ldots, p_{(N)}$) respectively.

A table of exact values of $e_{(i)}$ for n = 2, 3, 4 and 5 is given.

In order to estimate $V(\bar{y}_N)$, the usual unbiased estimator of Y^2 is considered and an estimator better than the usual estimator is given. As these better estimators of Y and Y^2 are unwieldy to compute in practice, other improved unbiased estimators of Y and Y^2 are given that are simple to compute in practice, but are somewhat less efficient than the former estimators.

In Chapter V, we discuss sampling schemes with unequal probabilities of selection but without replacement. As the 'order-statistic' is sufficient, any estimator (e.g., of the population total) which is not a function of the 'order-statistic', can be uniformly improved by the Rao-Blackwell theorem. In this chapter, certain results obtained by Murthy (31) are shown to be immediate consequences of this observation. It is shown that if we rely only on the 'order-statistic', sampling with unequal probabilities (with replacement) until we get a specified number of distinct units is equivalent to sampling with unequal probabilities (without replacement).

In Chapter VI, the following two-stage sampling schemes have been taken up:

i) The first-stage units are selected with unequal probabilities (with replacement).

ii) The second-stage units are selected by simple random sampling (a) with replacement (first case), (b) without replacement (second case).

Here again, estimators are obtained that are better than usually employed estimators. Two sets of improved estimators are evolved. The first set of estimators suggests the immediate necessity of employing these estimators in practice as they are as simple as the original estimators from which they were derived. The second set, though being better than the first set, is not of much use in practice.

Manjasma, Murthy and Sethi (33) gave the modifications of the usually employed sampling schemes which provide unbiased ratio estimators of the ratios of population totals of two characteristics. Their modification consists essentially in selecting first unit in the sample with probabilities proportional to W_j (W_j being the value of the auxiliary characteristic of the j-th population unit) from the whole population and the remaining sample units according to the original sampling scheme. In Chapter VII, we have given unbiased ratio estimators better than those given by them in cases of sampling with unequal probabilities, two-stage sampling (first-stage sampling with unequal probabilities with replacement and second-stage simple random sampling without replacement) and stratified

sampling with unequal probabilities.

The problem of estimating the total size of a population and its reciprocal is known to be of special interest in biological and other related problems, e.g., well-known problems of this type are the estimation of the total number of fish in a lake or the estimation of the total number of wild animals in a forest etc.. In Chapter VIII, independent simple random sampling (without replacement) at several stages is considered for this purpose. New unbiased estimators for the population size and its reciprocal are given. The estimator of the reciprocal of the population size is unbiased, but the estimator of the population size is unbiased only if the total number of fish caught (including the repetitions) is not less than the total number of fish in the lake. In addition, the problem of estimating the population mean of a characteristic and the ratio of the population means of two characteristics is also considered. A modification of the above sampling scheme on the lines of Manjamma, Murthy and Sethi (33) provides an unbiased ratio estimator for the ratio of the population means of two characteristics. This modification also provides simpler estimator of the population size. This simple estimator has the same bias as Bailey's estimator in case of direct sampling (simple random sampling with replacement), but has smaller risk function than Bailey's estimator for any convex (downwards) loss function.

Finally, in the two appendices given in the end we prove certain algebrical results that are relevant to the thesis.

CHAPTER II

EVALUATION OF MOMENTS OF DISTINCT UNITS IN A SAMPLE

Summary.

In this chapter, exact expressions for the moments of distinct units that appear in a sample, are derived under any sampling scheme. The importance of such a problem arises, e.g., when we select a simple random sample (with replacement) from a finite population, and require the variance of the average of distinct units selected. It has been shown by Basu [4] that this average is a better estimator of the population mean than the usual overall average.

2.1. Preliminary.

In this section, we give an important lemma which will be used repeatedly in this and subsequent chapters. The proof of this lemma is given in Appendix I.

Lemma 1: The coefficient, $C_m(n)$, of $Z_1^{\alpha_1} Z_2^{\alpha_2} \dots Z_m^{\alpha_m}$ (where $m \leq N$, α_i's > 0 and $\sum_{i=1}^{m} \alpha_i = n$) in the expansion of

$$(Z_1 + Z_2 + \dots \dots + Z_N)^n ,$$

is given by[1]

$$C_m(n) = m^n - \binom{m}{1}(m-1)^n + \binom{m}{2}(m-2)^n - \dots \dots (-)^{m-1}\binom{m}{m-1}1^n . \quad \dots \quad \dots \quad (2.1.1)$$

---

1. Note that $C_m(n) = 0$ for $m > n$.

In terms of the 'differences of zeros', $C_m(n)$ can be represented as:

$$C_m(n) = \Delta^m 0^n = \Delta^m x^n \big|_{x=0} , \quad \ldots \quad \ldots \quad (2.1.2)$$

where Δ is the difference operator with unit increments. We shall be using hereafter these two expressions (2.1.1) and (2.1.2) for $C_m(n)$, whichever will be convenient to us.

Corollary 1: Coefficient of $Z_{i_1}^{\alpha_1} Z_{i_2}^{\alpha_2} \ldots \ldots Z_{i_m}^{\alpha_m}$ (where $i_1, i_2, \ldots, i_m$ are any m different integers chosen out of $1, 2, \ldots \ldots, N$; α_i's > 0 and $\sum_{i=1}^{m} \alpha_i = n$) in the expansion of $(Z_1 + Z_2 + \ldots \ldots + Z_N)^n$, is given by $C_m(n)$.

Corollary 2:

$$C_m(n) = m \left[C_m(n-1) + C_{m-1}(n-1) \right]. \quad \ldots \quad \ldots \quad (2.1.3)$$

Corollary 3: For all positive integral values of N, we have

$$N^n = \sum_{m=1}^{N} C_m(n) \binom{N}{m} . \quad \ldots \quad \ldots \quad (2.1.4)$$

2.2A. Moments of distinct units of any positive order.

Consider a population containing N units and a sampling scheme S for which

p_i = probability of inclusion of the ith unit in the sample; $(i = 1, 2, \ldots, N)$

q_i = probability of exclusion of the ith unit from the sample;

p_{ij} = probability of inclusion of the ith and jth units in the sample;

q_{ij} = probability of exclusion of the ith and jth units from the sample;

etc.

We shall denote by ν , the number of distinct units that appear in a sample.

It is obvious that

$$\nu = Z_1 + Z_2 + \dots \dots + Z_N, \quad \dots \quad \dots \qquad (2.2.1)$$

where

$$Z_i = \begin{cases} 1 & \text{if the ith unit is included in the sample;} \\ 0 & \text{otherwise.} \end{cases}$$

Now, by definition, if n is any positive integer, the nth order moment of ν is given by

$$E(\nu^n) = E(Z_1 + Z_2 + \dots \dots + Z_N)^n$$

$$= E\Big[\sum_{m=1}^{N} \Sigma_1 \Sigma_2 z_1^{\alpha_1} \dots \dots z_m^{\alpha_m}\Big] , \quad \dots \qquad (2.2.2)$$

where Σ_1 denotes the summation over $\binom{N}{m}$ combinations of m Z's chosen out of $Z_1, Z_2, \dots, Z_N$ and Σ_2 denotes the summation over all products of the type

$$z_1^{\alpha_1} z_2^{\alpha_2} \dots z_m^{\alpha_m} \quad (\alpha_i\text{'s} > 0 \; ; \; \sum_{i=1}^{m} \alpha_i = n) .$$

Obviously,

$$E\left(\sum_2 z_1^{\alpha_1} z_2^{\alpha_2} \ \dots \ z_m^{\alpha_m} \right) = p_{12 \dots m} \, C_m(n) \ ,$$

and therefore,

$$E(\nu^n) = \sum_{m=1}^{N} C_m(n) \sum_1 p_{12 \dots m} \ . \qquad \dots \qquad (2.2.3)$$

When $p_{i_1 i_2 \dots i_m} = p_{12 \dots m}$ for every set of m distinct units $i_1, i_2, \dots, i_m$, the expression (2.2.3) reduces to

$$E(\nu^n) = \sum_{m=1}^{N} \binom{N}{m} C_m(n) \, p_{12 \dots m} \ . \qquad \dots \qquad (2.2.3A)$$

2.2B. Moments of ν of any negative order.

To derive the negative moments of ν of any order under any sampling scheme, we assume[1] that $\nu \geq 1$, i.e.,

$$q_{12 \dots N} = 0 \ ,$$

and define

$$u_i = 1 - z_i \qquad (i = 1, 2, \dots, N) \ .$$

---

1. It is evident that this assumption is indeed necessary, otherwise no negative moment of ν exists. In this chapter, we restrict ourselves to those sampling schemes for which $P[\nu \geq 1] = 1$.

Therefore, if t is any positive integer, the negative moment of ν of order t is given by

$$E\left(\frac{1}{\nu^t}\right) = E\left[\frac{1}{(N - u_1 - u_2 - \ldots - u_N)^t}\right]$$

$$= \frac{1}{N^t} E\left[1 - \frac{\sum_{i=1}^{N} u_i}{N}\right]^{-t} . \quad \ldots \quad (2.2.4)$$

Since, $0 < \sum_{i=1}^{N} u_i \leq (N - 1)$, the infinite expansion is possible.

Now, let

$$(1 - x)^{-t} = 1 + \sum_{r=1}^{\infty} A_r x^r ,$$

so that

$$E\left(\frac{1}{\nu^t}\right) = \frac{1}{N^t} E\left[1 + \sum_{r=1}^{\infty} \frac{A_r}{N^r} (u_1 + u_2 + \ldots + u_N)^r\right] .$$

$$\ldots \quad \ldots \quad (2.2.5)$$

Since the infinite series, $1 + \sum_{r=1}^{\infty} \frac{A_r}{N^r} (u_1 + u_2 + \ldots + u_N)^r$, is bounded above by the absolutely convergent series

$$1 + \sum_{r=1}^{\infty} A_r \left(\frac{N-1}{N}\right)^r ,$$

it, therefore, follows that

$$E\left(\frac{1}{\nu^t}\right) = \frac{1}{N^t} \left[1 + \sum_{r=1}^{\infty} \frac{A_r}{N^r} E (u_1 + u_2 + \ldots + u_N)^r\right] .$$

$$\ldots \quad (2.2.6)$$

But it is apparent from (2.2.3) that

$$E(u_1 + u_2 + \dots + u_N)^r = E\left\{ \sum_{i=1}^{N} \Sigma_1 \Sigma_2 \; u_1^{\alpha_1} u_2^{\alpha_2} \dots u_m^{\alpha_m} \right\}$$

$$= \sum_{m=1}^{N-1} (\Sigma_1 q_{12 \dots m}) \Delta^m x^r \,|_{x=0} \; .$$

The N-th term vanishes by assumption $q_{12 \dots N} = 0$.

Putting this in (2.2.6), we obtain

$$E \left(\frac{1}{\mathcal{D}^t} \right) = \frac{1}{N^t} \left[1 + \sum_{r=1}^{\infty} \frac{A_r}{N^r} \sum_{m=1}^{N-1} (\Sigma_1 q_{12 \dots m}) \Delta^m x^r \,|_{x=0} \right]$$

$$= \frac{1}{N^t} \left[1 + \sum_{m=1}^{N-1} (\Sigma_1 q_{12 \dots m}) \Delta^m \sum_{r=1}^{\infty} \frac{A_r}{N^r} x^r \,|_{x=0} \right]$$

$$= \frac{1}{N^t} \left[1 + \sum_{m=1}^{N-1} (\Sigma_1 q_{12 \dots m}) \Delta^m (1 - \tfrac{x}{N})^{-t} \,|_{x=0} \right] ,$$

which on expansion gives

$$E \left(\frac{1}{\mathcal{D}^t} \right) = \frac{1}{N^t} + \sum_{m=1}^{N-1} (\Sigma_1 q_{12 \dots m}) \left[\frac{1}{(N-m)^t} - \frac{\binom{m}{1}}{(N-m+1)^t} + \dots \right.$$

$$\left. \dots \quad \dots + (-)^m \frac{\binom{m}{m}}{N^t} \right] . \qquad \dots \qquad (2.2.7)$$

In case, $q_{i_1 i_2 \dots i_m} = q_{12 \dots m}$ for every set of m distinct units $i_1, i_2, \dots, i_m$; then (2.2.7) reduces to

$$E\left(\frac{1}{\nu^t}\right) = \left[\frac{1}{N^t} + \sum_{m=1}^{N-1} \binom{N}{m} q_{12\ldots m} \frac{1}{(N-m)^t} - \frac{\binom{m}{1}}{(N-m+1)^t} + \ldots\right.$$

$$\left. \ldots \quad \ldots + (-)^m \frac{\binom{m}{m}}{N^t}\right] . \quad \ldots \quad \ldots \quad (2.2.7A)$$

Corollary 1: Putting $t = 1$ in the above result, we get

$$E\left(\frac{1}{\nu}\right) = \left[\frac{1}{N} + \sum_{m=1}^{N-1} \left(\Sigma_1 q_{12\ldots m}\right) \left\{\frac{1}{N-m} - \frac{\binom{m}{1}}{(N-m+1)} + \ldots\right.\right.$$

$$\left.\left. \ldots \quad \ldots (-)^m \frac{\binom{m}{m}}{N}\right\}\right] .$$

Since

$$\frac{1}{N-m} - \frac{\binom{m}{1}}{(N-m+1)} + \ldots (-)^m \frac{\binom{m}{m}}{N} = \frac{m!}{N(N-1)\ldots(N-m)} ,$$

$$\therefore \quad E\left(\frac{1}{\nu}\right) = \frac{1}{N}\left[1 + \frac{1}{(N-1)} \Sigma_1 q_1 + \frac{1.2}{(N-1)(N-2)} \Sigma_1 q_{12} + \right.$$

$$\left. \frac{1.2.3}{(N-1)(N-2)(N-3)} \Sigma_1 q_{123} + \cdots + \frac{1.2 \ldots (N-1)}{(N-1)(N-2)\ldots 2.1} \Sigma_1 q_{12\ldots N-1}\right] .$$

$$\ldots \quad (2.2.8)$$

In case, $q_{i_1 i_2 \ldots i_m}$ are all equal for every set of m distinct units $i_1, i_2, \ldots, i_m$;

$$E\left(\frac{1}{\nu}\right) = \frac{1}{N} + \sum_{m=1}^{N-1} \frac{q_{12\ldots m}}{(N-m)} \quad . \quad \ldots \quad \ldots \quad (2.2.8A)$$

Particular cases:

(a) Simple random sampling (with replacement).

In this sampling scheme

$$q_{12 \ldots m} = \frac{(N-m)^n}{N^n}, \qquad (m = 1, 2, \ldots, N)$$

where n is the sample size.

$$\therefore \quad E\left(\frac{1}{\nu}\right) = \frac{1}{N} + \sum_{m=1}^{N-1} \frac{(N-m)^n}{N^n} \cdot \frac{1}{(N-m)}$$

$$= \frac{1^{n-1} + 2^{n-1} + \ldots + N^{n-1}}{N^n} \cdot \quad \ldots \quad \ldots \qquad (2.2.8B)$$

In terms of Bernoulli's numbers, (2.2.8B) is given by, Davis [9] (page 188),

$$= \frac{1}{n} + \frac{1}{2N} + \frac{1}{N} \sum_{s=1}^{\leq \frac{n-1}{2}} (-)^{s-1} \binom{n}{2s} \frac{B_s}{N^{2s}}, \quad \ldots \qquad (2.2.9)$$

where B_i is the ith Bernoulli's number .

For large N, this gives us a very convenient method for computing $E\left(\frac{1}{\nu}\right)$. Table 2.1 gives values of $E\left(\frac{n}{\nu}\right)$ correct to five places of decimals for sample sizes upto fifty and for a reasonably wide range of the sampling fraction $\frac{n}{N}$.

(b) **Simple random sampling (without replacement)**

Here

$$q_{12\ldots m} = \begin{cases} \dfrac{\binom{N-m}{n}}{\binom{N}{n}} & \text{for } m \leq N-n\ ; \\ 0 & \text{otherwise.} \end{cases}$$

$$\therefore\ E\left(\frac{1}{\nu}\right) = \left[\frac{1}{N} + \frac{\binom{N-1}{n}}{(N-1)\binom{N}{n}} + \ldots\ldots + \frac{\binom{n}{n}}{n\binom{N}{n}}\right]$$

$$= \frac{1}{\binom{N}{n}}\left[\frac{1}{n} + 1 + \frac{(n+1)}{1.2} + \frac{(n+1).(n+2)}{1.2.3} + \ldots\right.$$

$$\left.\ldots + \frac{(n+1)(n+2).\ \ldots\ \ldots\ (N-1)}{1.2.3.\ \ldots\ \ldots\ (N-n)}\right].$$

Combining the terms one by one, we get

$$E\left(\frac{1}{\nu}\right) = \frac{1}{\binom{N}{n}}\ \frac{(n+1)(n+2).\ \ldots\ .\ N}{n.1.2.\ \ldots\ .\ (N-n)} = \frac{1}{n}\ ,$$

which is in agreement with the sampling scheme.

Corollary 2: For any integer t ($\neq 0$), it can be shown in a similar manner that

$$E(\nu^t) = N^t + \sum_{m=1}^{N-1} (\Sigma_1 q_{12\ldots m})\ \Delta^m\ (N-x)^t\ \Big|_{x=0}$$

$$= N^t + \sum_{m=1}^{N-1} (\Sigma_1 q_{12\ldots m})\left[(N-m)^t - \binom{m}{1}(N-m+1)^t + \ldots\right.$$

$$\left.\ldots (-)^m \binom{m}{m} N^t\right]. \qquad \ldots \qquad (2.2.10)$$

We conclude this chapter with an obvious generalisation of (2.2.10) for a certain class of functions.

Theorem:

If $f(z)$ be a function of z for which the infinite expansion in powers of z is possible in the domain $0 < z \leq N$, and if the expectation can be taken term by term, then

$$E[f(\nu)] = f(N) + \sum_{m=1}^{N-1} (\Sigma_1 q_{12\ldots m})[f(N-m) - \binom{m}{1} f(N-m+1) + \ldots$$

$$\ldots (-)^m \binom{m}{m} f(N)] . \qquad \ldots \qquad (2.2.11)$$

Proof:

Express $f(z)$ in the form

$$f(z) = \sum_{r=-\infty}^{\infty} A_r z^r .$$

By assumption

$$E[f(\nu)] = E[\sum_{r=-\infty}^{\infty} A_r \nu^r] = \sum_{r=-\infty}^{\infty} A_r E(\nu^r).$$

Putting the value of $E(\nu^r)$ from (2.2.10), we get

$$E[f(\nu)] = \sum_{r=-\infty}^{\infty} A_r [N^r + \sum_{m=1}^{N-1} (\Sigma_1 q_{12\ldots m}) \Delta^m (N-x)^r \big|_{x=0}]$$

$$= f(N) + \sum_{m=1}^{N-1} (\Sigma_1 q_{12\ldots m}) \Delta^m f(N-x) \big|_{x=0} ,$$

which on expansion gives (2.2.11).

Table 2.1. Values of $E\left(\frac{n}{5}\right)$.

$\frac{n}{N}$ →	2	3	4	5	6
.001	1.00050*				
.002	1.00100*				
.003	1.00150*				
.004	1.00200*				
.005	1.00250*				
.006	1.00300*				
.007	1.00350*				
.008	1.00400	1.00400	1.00400	1.00400	1.00400
.009	1.00450	1.00450	1.00451*		
.010	1.00500	1.00501	1.00501*		
.015	1.00750	1.00751	1.00751	1.00751$_5$	1.00752*
.020	1.01000	1.01002	1.01002$_5$	1.01003*	
.025	1.01250	1.01253	1.01254	1.01254	1.01254
.030	1.01500	1.01505	1.01506	1.01506	1.01506
.035	1.01750	1.01757	1.01758	1.01758	1.01759
.040	1.02000	1.02009	1.02010	1.02011	1.02011
.045	1.02250	1.02261	1.02263	1.02263$_5$	1.02264
.050	1.02500	1.02514	1.02516	1.02517	1.02517
.055	1.02750	1.02767	1.02769	1.02770	1.02771
.060	1.03000	1.03020	1.03022$_5$	1.03024	1.03025
.065	1.03250	1.03273	1.03276	1.03278	1.03279
.070	1.03500	1.03527	1.03531	1.03533	1.03534
.075	1.03750	1.03781	1.03785	1.03787$_5$	1.03789
.080	1.04000	1.04036	1.04040	1.04043	1.04044
.085	1.04250	1.04290	1.04295	1.04298	1.04300
.090	1.04500	1.04545	1.04551	1.04554	1.04556
.095	1.04750	1.04800	1.04806	1.04810	1.04813
.100	1.05000	1.05056	1.05062$_5$	1.05067	1.05069

* Use this value for subsequent n.

Table 2.1. Values of $E\left(\frac{n}{N}\right)$ (contd.)

n → / n/N	7	8	9	10	12
.008	1.00400	1.00401*			
.009					
.010					
.015					
.020					
.025	1.01254	1.01255*			
.030	1.01506	1.01507*			
.035	1.01759	1.01759	1.01759	1.01759	1.01759
.040	1.02011	1.02012	1.02012	1.02012	1.02012
.045	1.02264	1.02265	1.02265	1.02265	1.02265
.050	1.02518	1.02518	1.02519	1.02519	1.02519
.055	1.02772	1.02772	1.02772	1.02773	1.02773
.060	1.03026	1.03026	1.03027	1.03027	1.03028
.065	1.03280	1.03281	1.03281	1.03282	1.03282
.070	1.03535	1.03536	1.03536	1.03537	1.03537
.075	1.03790	1.03791	1.03792	1.03792	1.03793
.080	1.04046	1.04047	1.04047	1.04048	1.04049
.085	1.04302	1.04303	1.04304	1.04304	1.04305
.090	1.04558	1.04559	1.04560	1.04561	1.04562
.095	1.04814	1.04816	1.04817	1.04818	1.04819
.100	1.05071	1.05073	1.05074	1.05075	1.05076

Table 2.1. Values of $E\left(\frac{n}{N}\right)$ (contd.)

n → / n/N	14	16	18	20	25
.035	1.01759	1.01760*			
.040	1.02012	1.02012 5	1.02013*		
.045	1.02266	1.02266	1.02266	1.02266	1.02266
.050	1.02519	1.02520	1.02520	1.02520	1.02520
.055	1.02773	1.02774	1.02774	1.02774	1.02774
.060	1.03028	1.03028	1.03028	1.03029	1.03029
.065	1.03283	1.03283	1.03283	1.03283	1.03284
.070	1.03538	1.03538	1.03539	1.03539	1.03539
.075	1.03794	1.03794	1.03794	1.03795	1.03795
.080	1.04050	1.04050	1.04050	1.04051	1.04051
.085	1.04306	1.04306	1.04307	1.04307	1.04308
.090	1.04563	1.04563	1.04564	1.04564	1.04565
.095	1.04820	1.04821	1.04821	1.04821	1.04822
.100	1.05077	1.05078	1.05079	1.05079	1.05080

Table 2.1. Values of $\Sigma\left(\frac{n}{\Sigma}\right)$ (contd.)

n → / $\frac{n}{\Sigma}$	30	35	40	45	50	∞
.045	1.02266	1.02266	1.02266	1.02266	1.02267*	1.02267
.050	1.02520	1.02520	1.02520	1.02520	1.02520	1.02521
.055	1.02774	1.02774	1.02775	1.02775	1.02775	1.02775
.060	1.03029	1.03029	1.03029	1.03029	1.03029	1.03030
.065	1.03284	1.03284	1.03284	1.03284	1.03285	1.03285
.070	1.03539	1.03540	1.03540	1.03540	1.03540	1.03541
.075	1.03795	1.03796	1.03796	1.03796	1.03796	1.03797
.080	1.04052	1.04052	1.04052	1.04052	1.04052	1.04053
.085	1.04308	1.04308	1.04309	1.04309	1.04309	1.04310
.090	1.04565	1.04566	1.04566	1.04566	1.04566	1.04567
.095	1.04823	1.04823	1.04823	1.04824	1.04824	1.04825
.100	1.05081	1.05081	1.05081	1.05081	1.05082	1.05083

CHAPTER III

ON SIMPLE RANDOM SAMPLING WITH REPLACEMENT

Summary.

The problem of improving estimators in with replacement sampling schemes has been considered by Basu [4], and Des Raj and Khamis [14]. Basu showed by an ingenious method that for estimating the population mean from a simple random sample (with replacement), the average of distinct sample units is more efficient than the overall sample mean. He, however, stated that it was not possible to give a simple expression for the variance of the above estimator. In this chapter, a detailed treatment of the above problem is given, and the exact expression for the variance of the above estimator is derived. The relative efficiency of the above estimator with other estimators is also considered. Some comparisons between the two simple random sampling schemes (with and without replacement) are made here. An improved estimator of the population variance is also obtained in simple random sampling with replacement.

3.1. Introduction.

Let us index the N population units as $1, 2, \ldots, N$ and let Y_j be some real valued characteristic (in which we are interested) of the jth population unit[1]. Here, we consider the problem of estimating the population mean

$$\bar{Y} = N^{-1} \sum Y_j,$$

1. Throughout this chapter, with whatever accents, j runs from 1 to N; i runs from 1 to n; and (i) runs from (1) to (ν).

and the population variance

$$\sigma^2 = N^{-1} \Sigma (Y_j - \bar{Y})^2 \quad .$$

For simplicity of future discussion, we shall always refer population units by capital letters and sample units by small letters, e.g., u_i and y_i will denote the unit index and the variate value associated with the ith sample unit respectively.

3.2. Estimation of $\bar{Y}$ from a simple random sample (with replacement).

Basu [4], in simple random sampling (with replacement), considered two estimators of the population mean:

(i) $\bar{y} = \frac{1}{n} \Sigma y_i \equiv$ Average over all n units of the sample;

(ii) $\bar{y}_\nu = \frac{1}{\nu} \Sigma y_{(i)} \equiv$ Average of ν distinct units observed in the sample.

If we record the sample of observation as

$$s = (x_1, x_2, \ldots, x_n),$$

where $x_i = (y_i, u_i)$ and if ν be the number of distinct units observed in the sample, he showed that the 'order-statistic' (sample units arranged in ascending order of their unit indices)

$$T = [x_{(1)}, x_{(2)}, \ldots \ldots, x_{(\nu)}] \quad ,$$

{where $x_{(i)} = (y_{(i)}, u_{(i)})$, and $y_{(i)}$ is the variate value of the sample unit with unit index $u_{(i)}$} forms a sufficient statistic and therefore, for any convex (downwards) loss function

$$E(\bar{y}' \mid T) = E(y_1 \mid T) \qquad \ldots \qquad (3.2.1)$$

has uniformly smaller risk than $\bar{y}$.

Now , for given T, the conditional distribution of x_1 is concentrated at ν points $x_{(1)}, x_{(2)}, \ldots, x_{(\nu)}$; and

$$P[x_1 = x_{(i)} \mid T] = \frac{\frac{1}{N} \Sigma'' \frac{(n-1)!}{\alpha_{(1)}! \cdots \alpha_{(\nu)}!} \left(\frac{1}{N}\right)^{\alpha_{(1)}} \cdots \left(\frac{1}{N}\right)^{\alpha_{(\nu)}}}{\Sigma' \frac{n!}{\alpha_{(1)}! \cdots \alpha_{(\nu)}!} \left(\frac{1}{N}\right)^{\alpha_{(1)}} \cdots \left(\frac{1}{N}\right)^{\alpha_{(\nu)}}},$$

$$\ldots \qquad \ldots \qquad (3.2.2)$$

where Σ' means summation over all integral α's such that

$$\alpha_{(i)} > 0 \quad \text{and} \quad \alpha_{(1)} + \alpha_{(2)} + \cdots + \alpha_{(\nu)} = n ;$$

and Σ'' means summation over all integral α's such that

$$\alpha_{(i)} \geq 0, \quad \alpha_{(i')} > 0 \quad \text{for } i' \neq i = 1, \ldots, \nu$$

and $\alpha_{(1)} + \alpha_{(2)} + \cdots + \alpha_{(\nu)} = n-1.$

Next, from (2.1.1) and (2.1.3), we have

$$\Sigma' \frac{n!}{\alpha_{(1)}! \cdots \alpha_{(\nu)}!} = C_{\nu}(n) \; ; \qquad \ldots \qquad \ldots \quad (3.2.3)$$

and

$$\Sigma'' \frac{(n-1)!}{\alpha_{(1)}! \cdots \alpha_{(\nu)}!} = C_{\nu}(n-1) + C_{\nu-1}(n-1) = \frac{C_{\nu}(n)}{\nu} .$$

$$\ldots \quad \ldots \quad (3.2.4)$$

$$\therefore \quad P[x_1 = x_{(i)} \mid T] = \frac{C_{\nu}(n)}{\nu C_{\nu}(n)} = \frac{1}{\nu} \quad . \quad \ldots \quad (3.2.5)$$

Thus,

$$E(\bar{y} \mid T) = E(y_1 \mid T) = \Sigma y_{(i)} P[x_1 = x_{(i)} \mid T]$$

$$= \frac{1}{\nu} \Sigma y_{(i)} \equiv \bar{y}_{\nu} \qquad \ldots \qquad (3.2.6)$$

is a better estimator than $\bar{y}$.

Since, for any $c_1, c_2, \ldots, c_n$ such that $\Sigma c_i = 1$,

$$E[\Sigma c_i y_i \mid T] = E[y_1 \mid T] ,$$

it follows that $\bar{y}_{\nu}$ is indeed better than any unbiased estimator

$$\Sigma c_i y_i , \qquad (\Sigma c_i = 1)$$

of $\bar{Y}$.

3.3. Variance of $\bar{y}_{\nu}$.

Since the probability of selecting the 'order-statistic'

$$T = (x_{(1)}, x_{(2)}, \ldots, x_{(\nu)})$$

is given by

$$P(T) = \Sigma' \frac{n!}{\alpha_{(1)}! \cdots \alpha_{(\nu)}!} \left(\frac{1}{N}\right)^{\alpha_{(1)}} \cdots \left(\frac{1}{N}\right)^{\alpha_{(\nu)}}$$

{where Σ' has a meaning similar to (3.2.3)},

therefore, by (3.2.3)

$$P(T) = \frac{C_\nu(n)}{N^n} \quad \ldots \quad \ldots \quad (3.3.1)$$

Similarly, the probability of selecting any 'order-statistic' with ν distinct units is given by

$$P(\nu) = \binom{N}{\nu} \frac{C_\nu(n)}{N^n} \quad \ldots \quad (3.3.2)$$

Thus, for given ν, the conditional probability of selecting T is given by

$$P(T \mid \nu) = \frac{1}{\binom{N}{\nu}} \quad \ldots \quad (3.3.3)$$

Therefore, by theorems of simple random sampling (without replacement), we have

$$V(\bar{y}_\nu) = E\, V(\bar{y}_\nu \mid \nu) = E\left(\frac{1}{\nu} - \frac{1}{N}\right) S^2, \quad \ldots \quad (3.3.4)$$

where $S^2 = [N/(N-1)]\,\sigma^2$.

Substituting for $E\left(\frac{1}{\nu}\right)$ from (2.2.88) in (3.3.4), we get

$$V(\bar{y}_{\nu}) = \left[\frac{1^{n-1} + 2^{n-1} + \ldots + N^{n-1}}{N^n} - \frac{1}{N}\right] S^2$$

$$= \left[\frac{1^{n-1} + 2^{n-1} + \ldots + (N-1)^{n-1}}{N^n}\right] S^2 \quad \ldots (3.3.5)$$

Direct proof that $V(\bar{y}_{\nu}) \leq V(\bar{y})$.

Since,

$$V(\bar{y}) = \left(\frac{N-1}{n\,N}\right) S^2, \qquad \ldots \qquad (3.3.6)$$

by comparing (3.3.5) and (3.3.6), we see that

$$V(\bar{y}_{\nu}) \leq V(\bar{y}),$$

if and only if

$$\frac{1^{n-1} + 2^{n-1} + \ldots + (N-1)^{n-1}}{N^n} \leq \frac{N-1}{n\,N},$$

or,

$$n\,S_{n-1}(N-1) \leq N^n - N^{n-1}, \qquad \ldots \qquad (3.3.7)$$

where

$$S_n(N) = 1^n + 2^n + \ldots \ldots + N^n .$$

It can be easily shown that

$$N^n - N^{n-1} = n\, S_{n-1}(N-1) + \binom{n-1}{2} S_{n-2}(N-1) + \ldots + \binom{n-1}{n-1} S_1(N-1) \, .$$

$$\ldots \qquad \ldots \qquad (3.3.8)$$

$$\therefore \quad n S_{n-1}(N-1) \leq N^n - N^{n-1} \, .$$

The equality holds only for $n = 1$ and $n = 2$.

Thus $\bar{y}_\nu$ has a smaller variance than $\bar{y}$ except for $n = 1$ and $n = 2$. For $n = 1$ and $n = 2$, $\bar{y}_\nu$ and $\bar{y}$ are identical.

Gain in efficiency: The gain in efficiency of $\bar{y}_\nu$ over $\bar{y}$ is given by

$$\frac{V(\bar{y}) - V(\bar{y}_\nu)}{V(\bar{y})} = \frac{\binom{n-1}{2} S_{n-2}(N-1) + \ldots + \binom{n-1}{n-1} S_1(N-1)}{N^n - N^{n-1}} \, .$$

$$\ldots \qquad \ldots \qquad (3.3.9)$$

An approximate expression for $V(\bar{y}_\nu)$:

For large samples, it is rather cumbersome to compute $V(\bar{y}_\nu)$ owing to the difficulty of computing $S_{n-1}(N-1)$. An approximate expression for $V(\bar{y}_\nu)$ valid for terms upto order N^{-2} is given by

$$V(\bar{y}_\nu) = \left[\frac{1}{n} - \frac{1}{2N} + \frac{n-1}{12N^2} \right] S^2 \, . \qquad \ldots \qquad \ldots \qquad (3.3.10)$$

3.4. Admissibility properties of certain estimators.

Let $\mathcal{C}$ denote a certain class of estimators of $\bar{Y}$. For a given loss function, $R(t)$ will represent the risk (or expected loss) associated with the estimator t of $\bar{Y}$.

Of two estimators t_1 and t_2 of $\bar{Y}$, t_1 will be said to be uniformly better than t_2 if, for a given loss function,

$$R(t_1) \leq R(t_2) \qquad \ldots \qquad \ldots \qquad (3.4.1)$$

holds for all possible values $(Y_1, Y_2, \ldots, Y_N)$ with the strict sign of inequality holding for at least one $(Y_1, Y_2, \ldots, Y_N)$.

An estimator t belonging to $\mathcal{C}$ is said to be admissible in $\mathcal{C}$ if there exists no estimator in $\mathcal{C}$ which is better than t.

Let us now consider the problem of finding admissible estimators of $\bar{Y}$. As the 'order-statistic' T is sufficient, we have to restrict ourselves to functions of T only. Moreover the distribution of T is not complete, therefore, many different estimators of $\bar{Y}$ can be suggested. For simplicity, we shall consider the following class of unbiased linear estimators of $\bar{Y}$ [1].

$$\bar{y}_s = f_1(\nu)\, \bar{y}_\nu + f_2(\nu). \qquad \ldots \qquad (3.4.2)$$

In view of the fact that

$$E[\, \bar{y}_s \mid \nu \,] = f_1(\nu)\, \bar{Y} + f_2(\nu),$$

---

1. In fact, it has been conjectured by Dr. Basu that any unbiased estimator of $\bar{Y}$, which is a linear function of T, must be of the form (3.4.2).

it is obvious that necessary and sufficient conditions for $\bar{y}_s$ to be an unbiased estimator of $\bar{Y}$, are

$$E[f_1(\nu)] = 1 \quad \text{and} \quad E[f_2(\nu)] = 0 . \quad \ldots \quad (3.4.3)$$

Consider, now, the class $\mathcal{C}$ of estimators $\bar{y}_s$ which satisfy the conditions of (3.4.3).

Now,

$$V(\bar{y}_s) = E[f_1^2(\nu)(\frac{1}{\nu} - \frac{1}{N})S^2] + V[f_1(\nu)\bar{Y} + f_2(\nu)] .$$

$$\ldots \quad \ldots \quad (3.4.4)$$

In order to choose a good estimator from $\mathcal{C}$, we are to minimise (3.4.4) by proper choices of $f_1(\nu)$ and $f_2(\nu)$. The first expression on the right hand side of (3.4.4) is independent of $f_2(\nu)$; so for a proper choice of $f_2(\nu)$, we are to minimise

$$V[f_1(\nu)\bar{Y} + f_2(\nu)],$$

which is minimum if $\bar{Y} f_1(\nu) + f_2(\nu)$ is constant for all values of ν, i.e.,

$$\bar{Y} f_1(\nu) + f_2(\nu) = E[f_1(\nu)\bar{Y} + f_2(\nu)] = \bar{Y} ,$$

or,

$$f_2(\nu) = \bar{Y}[1 - f_1(\nu)] . \quad \ldots \quad \ldots \quad (3.4.5)$$

Since the above solution of $f_2(\nu)$ contains the unknown $\bar{Y}$, the exact value of $f_2(\nu)$ is not known unless $f_1(\nu) = 1$. Thus, if we choose $f_1(\nu) = 1$, the best estimator of $\bar{Y}$ would be $\bar{y}_\nu$. However, in practical situations, when some a priori knowledge about $\bar{Y}$ is available, it seems appropriate to approximate $f_2(\nu)$ by

$$f_2(\nu) = \bar{X}\,[\,1 - f_1(\nu)], \qquad \ldots \qquad (3.4.6)$$

where $\bar{X}$ is some a priori estimate of $\bar{Y}$. For example, $\bar{X}$ may be taken as the estimate of the population mean of the same variate obtained from some previous survey etc.. On the other hand if no such information about $\bar{Y}$ is available, it would be safe to take $f_2(\nu) = 0$.

From (3.4.4), we see that to choose the optimum value of $f_1(\nu)$, we have to minimise

$$E[f_1^2(\nu)\,(\tfrac{1}{\nu} - \tfrac{1}{N})], \qquad \ldots \qquad (3.4.7)$$

subject to the condition that $E[\,f_1(\nu)] = 1$.

By Schwartz inequality, we have

$$E\{f_1^2(\nu)\,(\tfrac{1}{\nu} - \tfrac{1}{N})\}.\,E(\tfrac{1}{\nu} - \tfrac{1}{N})^{-1} \geq 1. \qquad \ldots \qquad (3.4.8)$$

The equality holds if and only if

$$f_1(\nu) = \frac{(\frac{1}{\nu} - \frac{1}{N})^{-1}}{E(\frac{1}{\nu} - \frac{1}{N})^{-1}} = \frac{\frac{N\nu}{N-\nu}}{E(\frac{N\nu}{N-\nu})} . \quad \ldots \quad (3.4.9)$$

Thus, when some a priori estimate $\bar{X}$ of $\bar{Y}$ is available, the reasonable estimator of $\bar{Y}$ is given by

$$\bar{y}_{\nu(1)} = \frac{[N\nu/(N-\nu)]}{E[N\nu/(N-\nu)]} \bar{y}_\nu + \bar{X}\left[1 - \frac{[N\nu/(N-\nu)]}{E[N\nu/(N-\nu)]}\right] .$$

$$\ldots \quad \ldots \quad (3.4.10)$$

Further, if no such information about $\bar{Y}$ is available, the following estimator is recommended:

$$\bar{y}_{\nu(2)} = \frac{[N\nu/(N-\nu)]}{E[N\nu/(N-\nu)]} \bar{y}_\nu . \quad \ldots \quad (3.4.11)$$

These two estimators are admissible in $\mathcal{C}$ in the sense that they minimise the first component of (3.4.4). Any estimator $\bar{y}_s$ different from either of them cannot be uniformly better than $\bar{y}_{\nu(1)}$ or $\bar{y}_{\nu(2)}$, because

$V(\bar{y}_{\nu(1)}) < V(\bar{y}_s)$ for all populations where $\bar{Y} = \bar{X}$;

$V(\bar{y}_{\nu(2)}) < V(\bar{y}_s)$ for all populations where $\bar{Y} = 0$.

The estimator $\bar{y}_\nu$ is also admissible in $\mathcal{C}$ in the sense that it minimises the second component of (3.4.4) and therefore has least variance for all populations where the first component is zero, i.e., $s^2 = 0$.

Expression for $E\left(\frac{N^{\nu}}{N-\nu}\right)$.

To evaluate $E\left(\frac{N^{\nu}}{N-\nu}\right)$, we shall follow the same method as used in Chapter II.

Now, we have

$$E\left(\frac{N^{\nu}}{N-\nu}\right) = E\,\nu\left(1-\frac{\nu}{N}\right)^{-1} = E\left[\sum_{t=1}^{\infty}\frac{\nu^{t}}{N^{t-1}}\right] \quad \ldots \quad (3.4.12)$$

The infinite expansion is valid since $1 \leq \nu \leq n < N$. Further, the series $\sum_{t=1}^{\infty}\frac{\nu^{t}}{N^{t-1}}$ is bounded above by the absolutely convergent series $\sum_{t=1}^{\infty}\frac{n^{t}}{N^{t-1}}$.

$$\therefore \quad E\left[\frac{N^{\nu}}{(N-\nu)}\right] = \sum_{t=1}^{\infty}\frac{E(\nu^{t})}{N^{t-1}} \quad \ldots \quad (3.4.13)$$

Putting the value of $E(\nu^{t})$ in (3.4.13) from (2.2.3A), we get

$$E\left[\frac{N^{\nu}}{(N-\nu)}\right] = \sum_{t=1}^{\infty}\sum_{m=1}^{n}\binom{N}{m} P_{12\ldots m}\,\Delta^{m}\,\frac{x^{t}}{N^{t-1}}\Big|_{x=0},$$

where

$$P_{12\ldots m} = \begin{cases} 1-\binom{m}{1}\left(\frac{N-1}{N}\right)^{n}+\ldots(-)^{m}\binom{m}{m}\left(\frac{N-m}{N}\right)^{n} & \text{for } m \leq n; \\ 0 & \text{otherwise.} \end{cases}$$

Thus,

$$E\left(\frac{N\nu}{N-\nu}\right) = \sum_{m=1}^{n} \binom{N}{m} p_{12\ldots m}\, \Delta^m \sum_{t=1}^{\infty} \frac{x^t}{N^{t-1}} \Big|_{x=0}$$

$$= \sum_{m=1}^{n} \binom{N}{m} p_{12\ldots m}\, \Delta^m x\left(1 - \frac{x}{N}\right)^{-1} \Big|_{x=0},$$

which on simplification gives

$$E\left(\frac{N\nu}{N-\nu}\right) = N^2 \sum_{m=1}^{n} \frac{p_{12\ldots m}}{(N-m)} \quad \ldots \quad \ldots \qquad (3.4.14)$$

We, thus, see that it may be cumbersome to compute the estimators (3.4.10) and (3.4.11) in case of large samples owing to the difficulty of computing $E\left(\frac{N\nu}{N-\nu}\right)$. If, however, the sampling fraction $\frac{n}{N}$ can be ignored, the estimators reduce to

$$\bar{y}^*_{\nu(1)} = \frac{\nu}{E(\nu)}\, \bar{y}_\nu + \bar{X}\left[1 - \frac{\nu}{E(\nu)}\right]; \quad \ldots \qquad (3.4.15)$$

$$\bar{y}^*_{\nu(2)} = \frac{\nu}{E(\nu)}\, \bar{y}_\nu \,. \qquad \ldots \quad \ldots \qquad (3.4.16)$$

It is easy to see that (3.4.16) is the well known Horvitz — Thompson estimator in case of equal probability sampling [23]. Because of the importance of $\bar{y}^*_{\nu(2)}$, we shall compare $\bar{y}^*_{\nu(2)}$ and $\bar{y}_\nu$.

3.5. Comparison of $\bar{y}_{\nu}$ and $\bar{y}^{*}_{\nu(2)}$.

We have shown that

$$V(\bar{y}_{\nu}) = \frac{1^{n-1} + 2^{n-1} + \ldots + (N-1)^{n-1}}{N^n} S^2 ,$$

and

$$V(\bar{y}^{*}_{\nu(2)}) = E\left[V\left[\frac{\nu}{N(\nu)}\ \bar{y}_{\nu} \mid \nu\right]\right] + V\left\{E\left[\frac{\nu}{N(\nu)}\ \bar{y}_{\nu} \mid \nu\right]\right\}$$

$$= E\ \frac{\nu^2}{N^2(\nu)}\left(\frac{1}{\nu} - \frac{1}{N}\right)S^2 + \frac{\bar{Y}^2}{N^2(\nu)}\, V(\nu) \quad . \quad \ldots (3.5.1)$$

It is not difficult to show that

$$E(\nu) = N\left[1 - \left(\frac{N-1}{N}\right)^n\right] ;$$

$$E(\nu^2) = N\left[1 - \left(\frac{N-1}{N}\right)^n\right] + N(N-1)\left[1 - 2\left(\frac{N-1}{N}\right)^n + \left(\frac{N-2}{N}\right)^n\right] ;$$

and $\quad V(\nu) = N\left(\frac{N-1}{N}\right)^n - N^2\left(\frac{N-1}{N}\right)^{2n} + N(N-1)\left(\frac{N-2}{N}\right)^n .$

$$\therefore\ V(\bar{y}^{*}_{\nu(2)}) = \frac{S^2}{N^2\left[1 - \left(\frac{N-1}{N}\right)^n\right]^2}\left\{N\left[1-\left(\frac{N-1}{N}\right)^n\right] - \left[1-\left(\frac{N-1}{N}\right)^n\right] - (N-1)\left[1 - 2\left(\frac{N-1}{N}\right)^n + \left(\frac{N-2}{N}\right)^n\right]\right\}$$

$$+ \frac{\bar{Y}^2}{N^2\left[1-\left(\frac{N-1}{N}\right)^n\right]^2}\left\{N\left(\frac{N-1}{N}\right)^n - N^2\left(\frac{N-1}{N}\right)^{2n} + N(N-1)\left(\frac{N-2}{N}\right)^n\right\} . \quad \ldots (3.5.2)$$

Now,

$$V(\bar{y}_{>}) - V(\bar{y}^{*}_{>(2)}) = S^2\Big[\frac{1^{n-1} + \ldots + (N-1)^{n-1}}{N^n} - \frac{(N-1)}{N^2}\frac{[(1-\frac{1}{N})^n-(1-\frac{2}{N})^n]}{[1-(1-\frac{1}{N})^n]^2}\Big]$$

$$- \frac{\bar{Y}^2}{N}\frac{[(1-\frac{1}{N})^n - N(1-\frac{1}{N})^{2n} + (N-1)(1-\frac{2}{N})^n]}{[1-(1-\frac{1}{N})^n]^2}$$

$$= C_1 S^2 - C_2 \bar{Y}^2 . \quad \text{(say)} \quad \ldots \quad \ldots \qquad (3.5.3)$$

Thus $\bar{y}_{>}$ is better than $\bar{y}^{*}_{>(2)}$ if

$$\frac{S^2}{\bar{Y}^2} < \frac{C_2}{C_1},$$

and worse if

$$\frac{S^2}{\bar{Y}^2} > \frac{C_2}{C_1} .$$

Approximate values of C_1 and C_2 for large populations correct upto terms of order N^{-2}, are given by

$$C_1 = \frac{1}{2nN} + \frac{5(n-1)}{12n N^2} ;$$

$$C_2 = \frac{n-1}{2nN} - \frac{(n-1)(n-2)}{3nN^2} . \qquad \ldots \qquad (3.5.4)$$

The above comparison shows that if the square of population coefficient of variation exceeds (n-1), then $\bar{y}^{*}_{\nu(2)}$ has smaller variance than $\bar{y}_{\nu}$. Therefore, in practical surveys a proper consideration of the above fact must be made before employing these estimators. Moreover, if we have some a priori knowledge of $\bar{Y}$, it would be more pertinent to compare $\bar{y}_{\nu}$ and $\bar{y}^{*}_{\nu(1)}$. It can be seen on similar lines that $\bar{y}_{\nu}$ is better than $\bar{y}^{*}_{\nu(1)}$ if

$$\frac{S^2}{(\bar{Y} - \bar{X})^2} < \frac{C_2}{C_1} ,$$

and worse otherwise. This result shows that if $\bar{X}$ provides a close approximation to $\bar{Y}$, it is always better to use $\bar{y}^{*}_{\nu(1)}$ instead of $\bar{y}_{\nu}$.

We, now, conclude this section with the following admissibility property of $\bar{y}_{\nu}$.

Theorem:

If squared error be the loss function, $\bar{y}_{\nu}$ is admissible among all functions of $\bar{y}_{\nu}$ and ν .

Proof:

Let

$$t = \bar{y}_{\nu} + f(\bar{y}_{\nu} , \nu)$$

be a function of $\bar{y}_{\nu}$ and ν . Suppose that t is uniformly better than $\bar{y}_{\nu}$.

Now, by assumption

$$R(t) = E(\bar{y}_{\nu} - \bar{Y})^2 + E[f(\bar{y}_{\nu}, \nu)]^2 + 2E[(\bar{y}_{\nu} - \bar{Y}) f(\bar{y}_{\nu}, \nu)]$$

$$\leq E[(\bar{y}_{\nu} - \bar{Y})^2] \quad \ldots \qquad (3.5.5)$$

holds for all $Y_1, Y_2, \ldots, Y_N$.

Take in particular $Y_1 = Y_2 = \ldots = Y_N = C$ (say).

Then, the above relation implies that

$$f(C, \nu) = 0 \qquad \ldots \qquad (3.5.6)$$

Since the choice of C is arbitrary, it follows that $f(\bar{y}_{\nu}, \nu)$ is identically zero.

Which proves the above theorem.

3.6. Estimation of variance.

Let us now turn to the problem of estimating the population variance from a simple random sample (with replacement). The usual estimator of the population variance

$$\sigma^2 = \frac{1}{N} \Sigma (Y_i - \bar{Y})^2$$

is given by the sample variance

$$s^2 = \frac{1}{(n-1)} \Sigma (y_i - \bar{y})^2 = \frac{1}{2n(n-1)} \sum_{i \neq i'} (y_i - y_{i'})^2 \quad \ldots \quad (3.6.1)$$

In this section, we derive a uniformly better estimator than s^2.

Theorem:

For any convex loss function[1], a uniformly better estimator than s^2 is given by

$$s^2_\nu = E[s^2 \mid T] = \left[\frac{C_\nu(n) - C_\nu(n-1)}{C_\nu(n)}\right] s^2_d, \quad \ldots \quad (3.6.2)$$

where

$$s^2_d = \begin{cases} \frac{1}{(\nu-1)} \Sigma (y_{(i)} - \bar{y}_\nu)^2 & \text{if } \nu > 1; \\ 0 & \text{otherwise.} \end{cases}$$

Proof:

Since T is a sufficient statistic, by Rao-Blackwell theorem a uniformly better estimator of σ^2 is given by

$$E(s^2 \mid T) = E\left[\frac{1}{2n(n-1)} \sum_{i \neq i'} (y_i - y_{i'})^2 \mid T\right]$$

$$= E\left[\frac{1}{2}(y_1 - y_2)^2 \mid T\right] . \quad \ldots \quad (3.6.3)$$

When $\nu = 1$, (3.6.3) is obviously zero. To derive (3.6.3) when $\nu > 1$, we observe {in terms of the notations of Section 3.2.}

---

1. By convex function, we shall always mean convex (downwards) function .

$$P[x_1 = x_{(i)},\ x_2 = x_{(i)} \mid T] = \frac{\frac{1}{N^2}\,\Sigma'' \frac{(n-2)!}{\alpha_{(1)}! \cdots \alpha_{(\nu)}!} \left(\frac{1}{N}\right)^{\alpha_{(1)}} \cdots \left(\frac{1}{N}\right)^{\alpha_{(\nu)}}}{\Sigma' \frac{n!}{\alpha_{(1)}! \cdots \alpha_{(\nu)}!} \left(\frac{1}{N}\right)^{\alpha_{(1)}} \cdots \left(\frac{1}{N}\right)^{\alpha_{(\nu)}}},$$

... ... (3.6.4)

where Σ' means summation over all integral α's such that

$$\alpha_{(1)} + \alpha_{(2)} + \cdots + \alpha_{(\nu)} = n \quad \text{and} \quad \alpha_{(i)} > 0 \quad \text{for} \quad i = 1, 2, \ldots, \nu\ ;$$

and Σ'' means summation over all integral α's such that

$$\alpha_{(1)} + \alpha_{(2)} + \cdots + \alpha_{(\nu)} = n-2\ ,\quad \alpha_{(i)} \geq 0,\quad \alpha_{(i')} \geq 0,\quad \text{and}$$

$$\alpha_{(k)} > 0 \qquad \text{for } k \neq i \neq i' = 1,2, \ldots, \nu\ .$$

Now, from (2.1.1) and (2.1.3), we have after some simplification

$$\Sigma' \frac{n!}{\alpha_{(1)}! \cdots \alpha_{(\nu)}!} = C_{\nu}(n)\ ;$$

$$\Sigma'' \frac{(n-2)!}{\alpha_{(1)}! \cdots \alpha_{(\nu)}!} = C_{\nu}(n-2) + 2C_{\nu-1}(n-2) + C_{\nu-2}(n-2)$$

$$= \frac{C_{\nu}(n) - C_{\nu}(n-1)}{\nu\,(\nu-1)}\ .$$

... (3.6.5)

$$\therefore \quad P[x_1 = x_{(i)},\ x_2 = x_{(i')} \mid T] = \frac{C_\nu(n) - C_\nu(n-1)}{C_\nu(n)\,\nu(\nu-1)}, \quad \ldots \quad (3.6.6)$$

for $i \neq i' = 1, 2, \ldots, \nu$.

Thus, if $\nu > 1$,

$$E\left[\frac{(y_1 - y_2)^2}{2} \mid T\right] = \sum \frac{(y_{(i)} - y_{(i')})^2}{2} P[x_1 = x_{(i)},\ x_2 = x_{(i')} \mid T]$$

$$= \frac{C_\nu(n) - C_\nu(n-1)}{C_\nu(n)} \left[\frac{1}{2\nu(\nu-1)} \sum (y_{(i)} - y_{(i')})^2\right]$$

$$= \frac{C_\nu(n) - C_\nu(n-1)}{C_\nu(n)} \left[\frac{1}{(\nu-1)} \sum (y_{(i)} - \bar{y}_\nu)^2\right].$$

$$\ldots \qquad \ldots \qquad (3.6.7)$$

Therefore, for any ν,

$$E(s^2 \mid T) = E\left[\frac{(y_1 - y_2)^2}{2} \mid T\right] = \frac{C_\nu(n) - C_\nu(n-1)}{C_\nu(n)}\, s_d^2, \quad (3.6.8)$$

where s_d^2 has been defined earlier.

Hence the theorem is proved.

In practice, the estimator s_ν^2 requires the knowledge of the ratio, $[C_\nu(n-1)/C_\nu(n)]$. Table 3.3 gives values of $[C_\nu(n-1)/C_\nu(n)]$ correct to seven places of decimals for all ν and $n = 1$ to 50. These ratios were computed from values of $[C_\nu(n)/n!]$ tabulated by Gupta [20].

3.7. Some estimators of $V(\bar{y}_\nu)$.

We have seen that

$$V(\bar{y}_\nu) = \frac{1^{n-1} + 2^{n-1} + \ldots + (N-1)^{n-1}}{N^n} S^2 .$$

Various unbiased estimators of $V(\bar{y}_\nu)$ are given by

$$\text{(I)} \quad v_1(\bar{y}_\nu) = \left[\frac{1^{n-1} + 2^{n-1} + \ldots + (N-1)^{n-1}}{N^n}\right] \frac{N}{(N-1)} s^2 ;$$

$$\text{(II)} \quad v_2(\bar{y}_\nu) = \left[\frac{1^{n-1} + 2^{n-1} + \ldots + (N-1)^{n-1}}{N^n}\right]\left[\frac{N}{(N-1)}\right]$$

$$\left[\frac{C_\nu(n) - C_\nu(n-1)}{C_\nu(n)}\right] s_d^2 ;$$

$$\text{(III)} \quad v_3(\bar{y}_\nu) = \frac{C_{\nu-1}(n-1)}{C_\nu(n)} s_d^2 ;$$

$$\text{(IV)} \quad v_4(\bar{y}_\nu) = \left[\left(\frac{1}{\nu} - \frac{1}{N}\right) + \frac{N-1}{N^n - N}\right] s_d^2 ;$$

$$\text{(V)} \quad v_5(\bar{y}_\nu) = \left[\left(\frac{1}{\nu} - \frac{1}{N}\right) + N^{1-n}\left(1 - \frac{1}{\nu}\right)\right] s_d^2 . \text{ (To be used for } \nu > 1.)$$

The estimator (II) is known to be uniformly better than (I). It appears difficult to give direct proofs of relative efficiencies of these estimators. The estimators (IV) and (V) were given by Des Raj and Khamis [14]. The estimator (V) is conditionally unbiased for $\nu > 1$. Des Raj and Khamis have suggested the use of (V) for $\nu > 1$.

It is easy to see that

$$v_4 = v_5 \frac{N^n}{N^n - N} . \qquad \ldots \qquad (3.7.1)$$

A little comparison will, now, show that the conditional variance of (V) is less than the variance of (IV). The amount of decrease in the variance is given by

$$V(v_4) - V(v_5 \mid \nu>1) = \frac{1}{N^{n-1}} E(v_4^2) . \quad \ldots \quad (3.7.2)$$

In general, this leads to the conclusion that any estimator σ_1^2 of σ^2, which is unbiased for σ^2 and is equal to zero for $\nu = 1$, can be reduced to give a conditionally unbiased estimate of σ^2 for $\nu > 1$, which has smaller conditional variance than the variance of σ_1^2. This conditionally improved estimator is related with σ_1^2 by the following equation :

$$\sigma_{1m}^2 = \sigma_1^2 \left[\frac{N^n}{N^n - N}\right] , \qquad \ldots \qquad (3.7.3)$$

where σ_{1m}^2 stands for the conditionally improved estimator of σ^2.

Numerical example:

To study the relative performance of these estimators of $V(\bar{y}_\nu)$, we shall consider the following three populations given by Yates and Grundy [46].

Table 3.1. Three populations of Yates and Grundy.

Population / Unit	A Y_i	B Y_i	C Y_i
1	.5	.8	.2
2	1.2	1.4	.6
3	2.1	1.8	.9
4	3.2	2.0	.8
$\sum_{i=1}^{4} Y_i$	7.0	6.0	2.5

These populations were deliberately chosen by them as being more extreme than will be normally encountered in practice.

The table below gives the variances of unbiased estimators of $V(\bar{y}_{\cdot})$, and $V(\bar{y}_{\cdot})$, when $n = 3$. $V(v_1)$ is not given, since $V(v_2) < V(v_1)$.

Table 3.2. Variances of unbiased estimators of $V(\bar{y}_{\cdot})$.

Population	$V(\bar{y}_{\cdot})$	$V(v_2)$	$V(v_3)$	$V(v_4)$	$V(v_5 \mid \cdot > 1)$
A	.29823	.04940	.05222	.09017	.07897
B	.06125	.00220	.00232	.00396	.00348
C	.020964	.000279	.000293	.000490	.000432

These results show that for the above three populations

$$V(v_2) < V(v_3) < V(v_5 \mid \rho > 1) < V(v_4) \qquad (3.7.4)$$

Thus, v_2 appears to be most efficient estimator of $V(\bar{y}_\rho)$. For $n = 2$, v_2 and v_3 are identical. The above comparison thus strongly suggests the use of v_2 for estimating $V(\bar{y}_\rho)$.

For getting the estimate of $V(t)$, where t is any unbiased estimator of $\bar{Y}$, the following estimator may be used:

$$v(t) = t^2 - \text{est}\ (\bar{Y}^2), \qquad \ldots \qquad (3.7.5)$$

where est $(\bar{Y}^2)$ stands for the unbiased estimator of $\bar{Y}^2$ and can be obtained from any of the relations

$$\text{est}(\bar{Y}^2) = v_i(\bar{y}_\rho) - \bar{y}_\rho^{\,2}. \qquad (i=1,2,3,4,5) \qquad (3.7.6)$$

From the example considered, it is expected that

$$\text{est}\ (\bar{Y}^2) = v_i(\bar{y}_\rho) - \bar{y}_\rho^2 \qquad (i=2,3) \qquad (3.7.7)$$

would fare better than the remaining estimators of $\bar{Y}^2$.

3.8. Comparison between two simple random sampling schemes.

Let us, now, compare the two simple random sampling schemes (with and without replacement) for the purpose of estimation of $\bar{Y}$. If we draw a with replacement simple random sample of size n, then the variance of the sample mean is $n^{-1}\sigma^2$. Further, in a without replacement simple random sample of size n, the variance of the sample mean is $n^{-1}\sigma^2(\frac{N-n}{N-1})$. Since

$$(\frac{N-n}{N-1})\frac{\sigma^2}{n} < \frac{\sigma^2}{n},$$

it is usually claimed that sampling without replacement is better than sampling with replacement. Basu [4] has pointed out that this comparison is not fair, because the cost of selecting a sample of size n in without replacement sampling is greater than the cost of selecting a sample of the same size in with replacement sampling. For comparing the two sampling schemes, it would be appropriate to take into account the cost involved in the selection of two different samples. The comparison, thus, mainly depends on the choice of the cost function and no sampling scheme can be said to be superior to the other unless the cost function is known in advance. Let us, for illustration, consider the case where the cost of sampling is proportional to the number of distinct units drawn. Thus, the expected cost of selecting a with replacement sample of size n is equivalent to the cost of selecting a without replacement sample of size $E(\nu) = N[1 - (\frac{N-1}{N})^n]$. Basu has showed that in this situation

the sample mean of the with replacement sample is worse than the sample mean of the equivalent without replacement sample. We, now, compare the sample mean $\bar{y}$ of the equivalent without replacement sample with the following estimator of with replacement sample:

$$\bar{y}_{\nu(2)} = \frac{\frac{N\nu}{(N-\nu)}}{E\left[\frac{N\nu}{(N-\nu)}\right]} \bar{y}_{\nu} ,$$

it has been shown that

$$V(\bar{y}_{\nu(2)}) = \frac{S^2}{E\left[\frac{N\nu}{(N-\nu)}\right]} + \bar{Y}^2 \; V\left[\frac{\frac{N\nu}{(N-\nu)}}{E\left[\frac{N\nu}{(N-\nu)}\right]}\right], \quad \ldots \quad (3.8.1)$$

and further,

$$V(\bar{y}) = \left[\frac{1}{E(\nu)} - \frac{1}{N}\right] S^2 . \quad \ldots \quad (3.8.2)$$

Since, $\frac{N\nu}{(N-\nu)}$ is a convex function of ν $(1 \leq \nu \leq n < N)$, we have

$$E\left[\frac{N\nu}{(N-\nu)}\right] \geq \frac{N E(\nu)}{[N - E(\nu)]} = \left[\frac{1}{E(\nu)} - \frac{1}{N}\right]^{-1} . \quad (3.8.3)$$

From (3.8.3), it is evident that the first component of $V(\bar{y}_{\nu(2)})$ is smaller than $V(\bar{y})$. Thus, for a population whose

coefficient of variation is sufficiently large, $V(\bar{y}_{\nu(2)})$ would be smaller than $V(\bar{y})$. This comparison shows that the sample mean of without replacement sample cannot be uniformly better than all estimators of with replacement sampling. The comparison made above is not very satisfactory. First , because of the linearity of the cost function and secondly, because $E(\nu)$ is not necessarily an integer. We hope that for some other cost functions also similar situations may be found out where with replacement sampling would fare better than without replacement sampling.

Table 5.3 Values of $\frac{\theta_m(n-1)}{C_m(n)}$.

n → / m	2	3	4	5	6	7	8	9	10
1				1.0000000 for all n					
2	.0000000	.3333333	.4285714	.4666667	.4838710	.4920635	.4960630	.4980392	.4990215
3	0	0	.1666667	.2400000	.2777778	.2990033	.3115942	.3193388	.3242229
4	↓	0	0	.1000000	.1538462	.1857143	.2057613	.2189189	.2278258
5		↓	↓	0	.0666667	.1071429	.1333333	.1510574	.1634568
				↓					
6					0	.0476190	.0789474	.1005291	.1159154
7					↓	0	.0357143	.0606061	.0785714
8						↓	0	.0277778	.0480000
9							↓	0	.0222222
10								↓	0

	2	3	4	5	6	7	8	9	10

Table 3.3. Values of $\frac{C_m(n-1)}{C_m(n)}$ (contd.)

n → / m	11	12	13	14	15	16	17	18	19	20
1					1.0000000 for all n					
2	.4995112	.4997557	.4998779	.4999390	.4999695	.4999847	.4999924	.4999962	.4999981	.4999990
3	.3273569	.3293923	.3307253	.3316032	.3321838	.3325687	.3328243	.3329943	.3331075	.3331828
4	.2339966	.2383479	.2414585	.2437059	.2453432	.2465438	.2474286	.2480833	.2485693	.2489308
5	.1723544	.1788676	.1837118	.1873611	.1901391	.1922722	.1939216	.1952045	.1962073	.1969942
6	.1271791	.1355998	.1420028	.1469395	.1507901	.1538225	.1562302	.1581551	.1597031	.1609542
7	.0918937	.1019882	.1097724	.1158627	.1206858	.1245448	.1276595	.1301924	.1322655	.1339720
8	.0631313	.0747043	.0837155	.0908370	.0965357	.1011446	.1049062	.1080005	.1105635	.1126991
9	.0389610	.0518519	.0619607	.0700084	.0764970	.0817858	.0861370	.0897460	.0927608	.0952950
10	.0181818	.0322581	.0433566	.0522416	.0594466	.0653539	.0702435	.0743243	.0777549	.0806574
11	0	.0151515	.0271493	.0367965	.0446549	.0511277	.0565107	.0610250	.0648389	.0680820
12	↓	0	.0128205	.0231660	.0316239	.0386164	.0444536	.0493679	.0535361	.0570949
13		↓	0	.0109890	.0200000	.0274725	.0337299	.0390147	.0435116	.0473637
14			↓	0	.0095238	.0174419	.0240896	.0297189	.0345220	.0386476
15				↓	0	.0083333	.0153453	.0212963	.0263854	.0307669
16					↓	0	.0073529	.0136054	.0189628	.0235845
17						↓	0	.0065359	.0121457	.0169935
18							↓	0	.0058480	.0109091
19								↓	0	.0052632
20									↓	0

Table 3.3. Values of $\frac{C_m(n-1)}{C_m(n)}$ (contd.)

n → m	21	22	23	24	25
1		1.0000000 for all n			
2	.4999995	.4999998	.4999999	.4999999	.5000000
3	.3332330	.3332665	.3332888	.3333036	.3333135
4	.2492003	.2494015	.2495518	.2496643	.2497484
5	.1976139	.1981031	.1984904	.1987974	.1990412
6	.1619699	.1627974	.1634737	.1640281	.1644835
7	.1353836	.1365562	.1375340	.1383520	.1390384
8	.1144881	.1159936	.1172659	.1183450	.1192632
9	.0974369	.0992561	.1008081	.1021373	.1032797
10	.0831272	.0852394	.0870540	.0886193	.0899745
11	.0708560	.0732409	.0753009	.0770878	.0786436
12	.0601513	.0627902	.0650794	.0670738	.0688180
13	.0506832	.0535590	.0560624	.0582511	.0601724
14	.0422127	.0453100	.0480141	.0503852	.0524727
15	.0345620	.0378671	.0407595	.0433020	.0455460
16	.0275957	.0310960	.0341657	.0368697	.0392613
17	.0212082	.0248926	.0281295	.0309861	.0335172
18	.0153161	.0191746	.0225696	.0255705	.0282338
19	.0098522	.0138756	.0174206	.0205584	.0233471
20	.0047619	.0089419	.0126294	.0158973	.0188053
21	0	.0043290	.0081522	.0115440	.0145657
22	↓	0	.0039526	.0074627	.0105929
23		↓	0	.0036232	.0068571
24			↓	0	.0033333
25				↓	0

Table 3.3. Values of $\frac{C_m(n-1)}{C_m(n)}$ (contd.)

n → / m	26	27	28	29	30
1		1.0000000 for all n			
2		2.5000000 for n > 24			
3	.3333201	.3333245	.3333275	.3333294	.3333307
4	.2498115	.2498587	.2498940	.2499206	.2499404
5	.1992352	.1993895	.1995125	.1996106	.1996889
6	.1648585	.1651676	.1654230	.1656342	.1658090
7	.1396157	.1401025	.1405137	.1408616	.1411565
8	.1200468	.1207173	.1212923	.1217864	.1222119
9	.1042646	.1051161	.1058542	.1064955	.1070539
10	.0911518	.0921776	.0930737	.0938586	.0945476
11	.0800030	.0811943	.0822415	.0831643	.0839795
12	.0703490	.0716970	.0728875	.0739417	.0748775
13	.0618648	.0633607	.0646867	.0658656	.0669162
14	.0543172	.0559525	.0574068	.0587037	.0598633
15	.0475339	.0493009	.0508763	.0522851	.0535481
16	.0413646	.0432760	.0449662	.0464810	.0478422
17	.0357685	.0377778	.0395768	.0411922	.0426467
18	.0306065	.0327276	.0346298	.0363408	.0378840
19	.0258351	.0280625	.0300629	.0318649	.0334925
20	.0214030	.0237315	.0258255	.0277142	.0294224
21	.0172679	.0196929	.0218762	.0238477	.0256329
22	.0133950	.0159121	.0181807	.0202314	.0220902
23	.0097547	.0123601	.0147103	.0168368	.0187662
24	.0063224	.0090123	.0114409	.0136400	.0156371
25	.0030769	.0058480	.0083516	.0106206	.0126827
26	0	.0028490	.0054250	.0077611	.0098857
27	↓	0	.0026455	.0050463	.0072310
28		↓	0	.0024631	.0047059
29			↓	0	.0022989
30				↓	0

Table 3.3. Values of $\frac{C_m(n-1)}{C_m(n)}$ (contd.)

n → / m	31	32	33	34	35
1			1.0000000	for all n	
2			.5000000	for n > 24	
3	.3333316	.3333322	.3333326	.3333328	.3333330
4	.2499553	.2499665	.2499749	.2499812	.2499859
5	.1997514	.1998012	.1998411	.1998730	.1998984
6	.1659539	.1660741	.1661738	.1662566	.1663255
7	.1414068	.1416195	.1418004	.1419544	.1420856
8	.1225788	.1228958	.1231699	.1234073	.1236131
9	.1075409	.1079665	.1083390	.1086654	.1089519
10	.0951537	.0956879	.0961594	.0965764	.0969457
11	.0847012	.0853413	.0859103	.0864167	.0868683
12	.0757102	.0764525	.0771157	.0777093	.0782413
13	.0678548	.0686950	.0694488	.0701262	.0707361
14	.0609026	.0618362	.0626766	.0634345	.0641193
15	.0546833	.0557058	.0566289	.0574637	.0582202
16	.0490684	.0501756	.0511775	.0520858	.0529109
17	.0439595	.0451473	.0462243	.0472028	.0480935
18	.0392793	.0405439	.0416926	.0427382	.0436916
19	.0349665	.0363044	.0375215	.0386311	.0396445
20	.0309714	.0323793	.0336619	.0348327	.0359036
21	.0272536	.0287284	.0300736	.0313031	.0324290
22	.0237795	.0253185	.0267237	.0280093	.0291680
23	.0205213	.0221217	.0235844	.0249241	.0261535
24	.0174553	.0191148	.0206327	.0220242	.0233023
25	.0145616	.0162777	.0178488	.0192902	.0206152
26	.0118229	.0135936	.0152159	.0167052	.0180754
27	.0092245	.0110478	.0127193	.0142550	.0156687
28	.0067535	.0086276	.0103467	.0119270	.0133827
29	.0043988	.0063218	.0080868	.0097103	.0112066
30	.0021505	.0041209	.0059303	.0075954	.0091310
31	0	.0020161	.0038685	.0055740	.0071475
32	↓	0	.0018939	.0036386	.0052489
33		↓	0	.0017825	.0034286
34			↓	0	.0016807
35				↓	0

Table 3.3. Values of $\frac{C_m(n-1)}{C_m(n)}$ (contd.)

n -> / m	36	37	38	39	40
1			1.0000000 for all n		
2			.5000000 for n > 24		
3	.3333331	.3333332	.3333332	.3333333	.3333333
4	.2499894	.2499921	.2499940	.2499955	.2499966
5	.1999188	.1999350	.1999480	.1999584	.1999667
6	.1663827	.1664302	.1664698	.1665027	.1665301
7	.1421976	.1422931	.1423746	.1424442	.1425037
8	.1237916	.1239467	.1240815	.1241988	.1243008
9	.1092036	.1094250	.1096199	.1097916	.1099431
10	.0972731	.0975637	.0978221	.0980519	.0982566
11	.0872715	.0876321	.0879548	.0882441	.0885036
12	.0787190	.0791485	.0795352	.0798837	.0801982
13	.0712861	.0717828	.0722320	.0726388	.0730075
14	.0647390	.0653007	.0658105	.0662739	.0666957
15	.0589068	.0595311	.0600994	.0606176	.0610907
16	.0536617	.0543460	.0549707	.0555417	.0560643
17	.0489058	.0496476	.0503263	.0509481	.0515186
18	.0445626	.0453597	.0460903	.0467608	.0473773
19	.0405718	.0414218	.0422021	.0429196	.0435802
20	.0368848	.0377855	.0386136	.0393761	.0400792

Table 3.3. Values of $\frac{C_m(n-1)}{C_m(n)}$ (contd.)

n -> m	36	37	38	39	40
21	.0334620	.0344113	.0352853	.0360910	.0368350
22	.0302707	.0312668	.0321848	.0330321	.0338154
23	.0272838	.0283248	.0292852	.0301725	.0309937
24	.0244785	.0255627	.0265639	.0274898	.0283474
25	.0218354	.0229613	.0240017	.0249648	.0258575
26	.0193382	.0205041	.0215825	.0225813	.0235080
27	.0169725	.0181771	.0192920	.0203255	.0212849
28	.0147261	.0159681	.0171183	.0181825	.0191763
29	.0125883	.0138664	.0150508	.0161500	.0171717
30	.0105497	.0118628	.0130802	.0142107	.0152620
31	.0086021	.0099490	.0111984	.0123592	.0134393
32	.0067382	.0081179	.0093983	.0105885	.0116964
33	.0049515	.0063630	.0076735	.0088922	.0100272
34	.0032362	.0046786	.0060183	.0072647	.0084259
35	.0015873	.0030597	.0044277	.0057010	.0068877
36	0	.0015015	.0028972	.0041965	.0054081
37	↓	0	.0014225	.0027473	.0039829
38		↓	0	.0013495	.0026087
39			↓	0	.0012821
40				↓	0

Table 3.3. Values of $\frac{C_m(n-1)}{C_m(n)}$ (contd.)

n -> m	41	42	43	44	45
1			1.0000000 for all n		
2			.5000000 for n > 24		
3			.3333333 for n > 38		
4	.2499975	.2499981	.2499986	.2499989	.2499992
5	.1999734	.1999787	.1999830	.1999864	.1999891
6	.1665530	.1665720	.1665878	.1666009	.1666119
7	.1425546	.1425981	.1426354	.1426672	.1426945
8	.1243897	.1244671	.1245346	.1245935	.1246449
9	.1100767	.1101948	.1102991	.1103913	.1104729
10	.0984390	.0986017	.0987470	.0988767	.0989927
11	.0887367	.0889462	.0891346	.0893043	.0894572
12	.0804824	.0807393	.0809719	.0811826	.0813736
13	.0733423	.0736464	.0739230	.0741749	.0744043
14	.0670799	.0674305	.0677505	.0680431	.0683107
15	.0615232	.0619189	.0622814	.0626139	.0629191
16	.0565434	.0569830	.0573868	.0577581	.0581000
17	.0520426	.0525245	.0529683	.0533774	.0537549
18	.0479446	.0484674	.0489498	.0493954	.0498074
19	.0441893	.0447515	.0452712	.0457520	.0461974
20	.0407284	.0413286	.0418842	.0423991	.0428768
21	.0375228	.0381596	.0387498	.0392976	.0398064
22	.0345404	.0352124	.0358361	.0364155	.0369545
23	.0317545	.0324605	.0331164	.0337264	.0342944
24	.0291428	.0298815	.0305684	.0312080	.0318041
25	.0266863	.0274566	.0281735	.0288416	.0294649

Table 3.3. Values of $\frac{C_m\ (n-1)}{C_m\ (m)}$ (contd.)

n -> m	41	42	43	44	45
26	.0243689	.0251697	.0259157	.0266113	.0272608
27	.0221769	.0230072	.0237812	.0245035	.0251784
28	.0200983	.0209572	.0217583	.0225064	.0232059
29	.0181228	.0190092	.0198366	.0206097	.0213330
30	.0162412	.0171545	.0180073	.0188046	.0195510
31	.0144458	.0153849	.0162624	.0170832	.0178520
32	.0127294	.0136937	.0145950	.0154386	.0162291
33	.0110858	.0120745	.0129991	.0138648	.0146764
34	.0095094	.0105218	.0114690	.0123563	.0131884
35	.0079955	.0090309	.0100000	.0109082	.0117602
36	.0065394	.0075972	.0085877	.0095162	.0103876
37	.0051372	.0062168	.0072281	.0081764	.0090667
38	.0037853	.0048862	.0059176	.0068852	.0077939
39	.0024804	.0036020	.0046531	.0056395	.0065662
40	.0012195	.0023613	.0034317	.0044364	.0053806
41	0	.0011614	.0022506	.0032731	.0042344
42	↓	0	.0011074	.0021475	.0031254
43		↓	0	.0010571	.0020513
44			↓	0	.0010101
45				↓	0

Table 3.3. Values of $\frac{C_m(n-1)}{C_m(n)}$ (contd.)

n -> m	46	47	48	49	50
1			1.0000000 for all n		
2			.5000000 for n > 24		
3			.3333333 for n > 36		
4	.2499994	.2499996	.2499997	.2499997	.2499998
5	.1999913	.1999930	.1999944	.1999955	.1999964
6	.1666210	.1666287	.1666350	.1666403	.1666447
7	.1427178	.1427378	.1427549	.1427695	.1427821
8	.1246897	.1247288	.1247629	.1247928	.1248188
9	.1105451	.1106090	.1106656	.1107158	.1107602
10	.0990965	.0991893	.0992724	.0993468	.0994135
11	.0895950	.0897194	.0898316	.0899330	.0900246
12	.0815469	.0817042	.0818471	.0819771	.0820953
13	.0746136	.0748045	.0749789	.0751383	.0752840
14	.0685558	.0687804	.0689864	.0691754	.0693490
15	.0634995	.0634574	.0636947	.0639132	.0641147
16	.0584150	.0587054	.0589736	.0592212	.0594502
17	.0541035	.0544259	.0547241	.0550003	.0552563
18	.0501888	.0505420	.0508696	.0511736	.0514560
19	.0466104	.0469937	.0473498	.0476809	.0479890
20	.0433204	.0437328	.0441165	.0444738	.0448069
21	.0402796	.0407202	.0411307	.0415135	.0418708
22	.0374563	.0379240	.0383604	.0387679	.0391488
23	.0348239	.0353180	.0357794	.0362108	.0366145
24	.0323603	.0328798	.0333656	.0338201	.0342459
25	.0300469	.0305911	.0311003	.0315773	.0320245

Table 3.3. Values of $\frac{C_m(n-1)}{C_m(n)}$ (contd.)

n -> m	46	47	48	49	50
26	.0278679	.0284359	.0289679	.0294666	.0299345
27	.0258097	.0264008	.0269548	.0274745	.0279626
28	.0238606	.0244740	.0250494	.0255895	.0260971
29	.0220104	.0226455	.0232416	.0238016	.0243281
30	.0202504	.0209065	.0215227	.0221019	.0226468
31	.0185728	.0192493	.0198850	.0204828	.0210456
32	.0169707	.0176670	.0183216	.0189376	.0195177
33	.0154381	.0161537	.0168267	.0174603	.0180573
34	.0139696	.0147040	.0153949	.0160456	.0166590
35	.0125605	.0133131	.0140214	.0146887	.0153181
36	.0112065	.0119767	.0127020	.0133856	.0140305
37	.0099036	.0106911	.0114329	.0121323	.0127924
38	.0086484	.0094527	.0102106	.0109254	.0116003
39	.0074378	.0082585	.0090321	.0097619	.0104512
40	.0062689	.0071056	.0078944	.0086390	.0093423
41	.0051391	.0059913	.0067951	.0075540	.0082710
42	.0040460	.0049135	.0057319	.0065047	.0072351
43	.0029874	.0038698	.0047024	.0054889	.0062324
44	.0019614	.0028584	.0037049	.0045047	.0052611
45	.0009662	.0018773	.0027375	.0035504	.0043192
46	0	.0009251	.0017986	.0026242	.0034053
47	↓	0	.0008865	.0017246	.0025177
48		↓	0	.0008503	.0016552
49			↓	0	.0008163
50				↓	0

Remark :

The values of the ratio $\frac{C_{m-1}(n-1)}{C_m(n)}$ can also be obtained from the values of $\frac{C_m(n-1)}{C_m(n)}$ with the help of the following relation:

$$\frac{1}{m} = \frac{C_m(n-1)}{C_m(n)} + \frac{C_{m-1}(n-1)}{C_m(n)} .$$

CHAPTER IV

ON SAMPLING WITH UNEQUAL PROBABILITIES

Summary.

This chapter deals with the problem of improving estimators in sampling schemes with unequal probabilities of selection. Here, we have derived the improved estimator of the population total, Y, which has been referred to by Basu. In addition, two sets of estimators of Y and Y^2 are given. The first set of estimators is cumbersome to compute, while the second set is simple for computation. The second set of estimators, though less efficient than the first, is more efficient than the usually employed estimators.

4.1. Introduction.

Let us consider an unequal probability selection method; let P_j be the probability of selection of the j-th population unit[1] ($\Sigma P_j = 1$). Suppose a sample of size n is drawn with replacement according to the above probabilities. If, for the i-th sample unit, we record its Y-characteristic y_i, its probability of selection p_i, and its unit-index u_i, then the sample is

---

1. In this and subsequent chapters, we shall be following continuously the notations introduced in Chapter III unless otherwise stated.

$$S = (x_1, x_2, \ldots, x_n),$$

where $x_i = (y_i, p_i, u_i)$.

Let the 'order-statistic' be given by

$$T = (x_{(1)}, \ldots, x_{(\nu)}),$$

where $x_{(1)}, x_{(2)}, \ldots, x_{(\nu)}$ are the ν distinct units arranged in ascending order of their unit-indices.

It can be easily shown that T is a sufficient statistic. Therefore, if $g(S)$ is some estimator depending on S, for any convex loss function, a uniformly better estimator than $g(S)$ is given by $E[g(S) \mid T]$.

4.2. Estimation of the population total.

The usual estimator of the population total

$$Y = \Sigma Y_j$$

is given by

$$\bar{z} = \frac{1}{n} \Sigma z_i , \qquad \ldots \qquad \ldots \qquad (4.2.1)$$

where $z_i = \frac{y_i}{p_i}$.

Theorem 1:

For any convex loss function, a uniformly better estimator than $\bar{z}$ is given by

$$\bar{z}_{\nu} = E(\bar{z} \mid T) = \sum c_{(i)} \frac{y_{(i)}}{p_{(i)}}, \quad \ldots \quad (4.2.2)$$

where

$$c_{(i)} = \frac{p_{(i)}\left[(p_{(1)} + \cdots + p_{(\nu)})^{n-1} - \sum_1^i (p_{(1)} + \cdots + p_{(\nu-1)})^{n-1} + \cdots (-)^{\nu-1} p_{(i)}^{n-1}\right]}{\left[(p_{(1)} + \cdots + p_{(\nu)})^{n} - \sum_1 (p_{(1)} + \cdots + p_{(\nu-1)})^{n} + \cdots + (-)^{\nu-1} \sum_1 p_{(1)}^{n}\right]};$$

$$\ldots \qquad \ldots \qquad (4.2.3)$$

the summations $\sum_1$ and $\sum_1^i$ stand for all combinations of p's and all combinations of p's containing $p_{(i)}$ (chosen out of $p_{(1)}, p_{(2)}, \ldots, p_{(\nu)}$) respectively.

Proof:

Obviously, by Rao-Blackwell theorem, a uniformly better estimator than $\bar{z}$ is given by

$$E(\bar{z} \mid T) = E\left(\frac{y_1}{p_1} \mid T\right)$$

$$= \sum \frac{y_{(i)}}{p_{(i)}} P\left[x_1 = x_{(i)} \mid T\right] \quad \ldots \qquad (4.2.4)$$

But

$$P[x_1 = x_{(i)} \mid T] = \frac{p_{(i)} \sum'' \frac{(n-1)!}{\alpha_{(1)}! \cdots \alpha_{(\nu)}!} p_{(1)}^{\alpha_{(1)}} \cdots p_{(\nu)}^{\alpha_{(\nu)}}}{\sum' \frac{n!}{\alpha_{(1)}! \cdots \alpha_{(\nu)}!} p_{(1)}^{\alpha_{(1)}} \cdots p_{(\nu)}^{\alpha_{(\nu)}}}, \qquad (4.2.5)$$

where $\sum''$ and $\sum'$ have meanings similar to those defined in (3.2.4).

It has been shown in Appendix I that

$$\sum'' \frac{(n-1)!}{\alpha_{(1)}! \cdots \alpha_{(\nu)}!} p_{(1)}^{\alpha_{(1)}} \cdots p_{(\nu)}^{\alpha_{(\nu)}} = [(p_{(1)} + \cdots + p_{(\nu)})^{n-1} -$$

$$\sum_1^i (p_{(1)} + \cdots + p_{(\nu-1)})^{n-1} + \cdots (-)^{\nu-1} p_{(i)}^{n-1}];$$

and

$$\sum' \frac{n!}{\alpha_{(1)}! \cdots \alpha_{(\nu)}!} p_{(1)}^{\alpha_{(1)}} \cdots p_{(\nu)}^{\alpha_{(\nu)}} = [(p_{(1)} + \cdots + p_{(\nu)})^{n} -$$

$$\sum_1 (p_{(1)} + \cdots + p_{(\nu-1)})^{n} + \cdots (-)^{\nu-1} \sum_1 p_{(i)}^{n}].$$

$$\cdots \qquad \cdots \qquad (4.2.6)$$

Using (4.2.4), (4.2.5) and (4.2.6), we get

$$\bar{z}_{\nu} = E(\bar{z} \mid T) = \sum c_{(i)} \frac{y_{(i)}}{p_{(i)}} .$$

Hence the theorem is proved.

The above estimator, though better, is not very useful in large scale sample surveys because of the cumbersome computations of $c_{(i)}$'s. In Section 4.4, we shall derive a simpler estimator of Y, better than $\bar{z}$. Table 4.1, given below, gives the exact expressions of this estimator, $\bar{z}_{\nu}$, for $n = 3$, 4 and 5. For $n = 1$ and $n = 2$, $\bar{z}_{\nu}$ is same as $\bar{z}$.

Table 4.1.

$n \rightarrow$ / ν	3	4
1	$\frac{y_{(1)}}{p_{(1)}}$	$\frac{y_{(1)}}{p_{(1)}}$
2	$\frac{\sum [2p_{(1)} + p_{(2)}] \frac{y_{(1)}}{p_{(1)}}}{3[p_{(1)} + p_{(2)}]}$	$\frac{\sum [(p_{(1)}+p_{(2)})^3 - p_{(1)}^3] y_{(1)}}{[(p_{(1)}+p_{(2)})^4 - p_{(1)}^4 - p_{(2)}^4]}$
3	$\sum \frac{y_{(1)}}{p_{(1)}}$	$\frac{\sum [2p_{(1)} + p_{(2)} + p_{(3)}] \frac{y_{(1)}}{p_{(1)}}}{4[p_{(1)} + p_{(2)} + p_{(3)}]}$
4	-	$\sum \frac{y_{(1)}}{p_{(1)}}$

Table 4.2 (contd.)

n → / ↓	5
2	$\frac{\Sigma\, r_{(1)}}{p_{(1)}}$
3	$\frac{\Sigma\,[(p_{(1)}+p_{(2)})^4 - p_{(1)}^4]\, r_{(1)}}{[(p_{(1)}+p_{(2)})^3 - p_{(1)}^3 - p_{(2)}^3]}$
4	$\frac{\Sigma\,[3p_{(1)}(p_{(1)}+p_{(2)}+p_{(3)})^2 + 4(p_{(2)}^3+p_{(3)}^3) + 6p_{(2)}p_{(3)}]\frac{r_{(1)}}{p_{(1)}}}{5[4(p_{(1)}^3+p_{(2)}^3+p_{(3)}^3)+6(p_{(1)}^2p_{(2)}+p_{(2)}^2p_{(3)}+p_{(3)}^2p_{(1)})]}$
5	$\frac{\Sigma\,[2p_{(1)}+p_{(2)}+p_{(3)}+p_{(4)}]\frac{r_{(1)}}{p_{(1)}}}{5\,[p_{(1)}+p_{(2)}+p_{(3)}+p_{(4)}]}$
6	$\Sigma\,\frac{r_{(1)}}{p_{(1)}}$

4.3. Estimation of Y^2.

The problem of finding an unbiased estimator of Y^2 arises in most problems of variance estimation of estimators of Y. The usual estimator of Y^2 is

$$z_p = \frac{1}{n(n-1)} \sum_{i \neq i'} z_i z_{i'} , \quad \ldots \quad (4.3.1)$$

where $z_i = \frac{y_i}{p_i}$.

Theorem 2:

For any convex loss function, a uniformly better estimator than z_p is given by

$$E(z_p \mid T) = \sum c_{(i,i)} z^2_{(i)} + \sum_{i \neq i'} c_{(i,i')} z_{(i)} z_{(i')} , \quad \ldots \quad (4.3.2)$$

where

$$c_{(i,i)} = \frac{p^2_{(i)} \left[(p_{(1)} + p_{(2)} + \ldots + p_{(\nu)})^{n-2} - \Sigma^{i}_{1} (p_{(1)} + \ldots + p_{(\nu-1)})^{n-2} + \ldots (-)^{\nu-1} p^{n-2}_{(i)} \right]}{\left[(p_{(1)} + \ldots + p_{(\nu)})^{n} - \Sigma_1 (p_{(1)} + \ldots + p_{(\nu-1)})^{n} + \ldots (-)^{\nu-1} \Sigma_1 p^{n}_{(1)} \right]} ,$$

and

$$c_{(i,i')} = \frac{p_{(i)} p_{(i')} \left[(p_{(1)} + \ldots + p_{(\nu)})^{n-2} - \Sigma^{ii'}_{1} (p_{(1)} + \ldots + p_{(\nu-1)})^{n-2} + \ldots (-)^{\nu-2} (p_{(i)} + p_{(i')})^{n-2} \right]}{\left[(p_{(1)} + \ldots + p_{(\nu)})^{n} - \Sigma_1 (p_{(1)} + \ldots + p_{(\nu-1)})^{n} + \ldots (-)^{\nu-1} \Sigma_1 p^{n}_{(1)} \right]} ;$$

$$\ldots \quad (4.3.3)$$

the summations Σ_1 and Σ_1^i have been defined in (4.2.3). The summation $\Sigma^{ii'}$ stands for all combinations of p's containing $p_{(i)}$ and $p_{(i')}$.

Proof:

Obviously, a uniformly better estimator than z_p is given by

$$E(z_p \mid T) = E\left(\frac{1}{n(n-1)} \sum_{i \neq i'} z_i z_{i'} \mid T\right) = E(z_1 z_2 \mid T)$$

$$= \Sigma\, z_{(i)}\, z_{(i')}\, P[x_1 = x_{(i)},\ x_2 = x_{(i')} \mid T]. \qquad \ldots \quad (4.3.4)$$

It is easy to see that

$$P[z_1 = z_{(i)}, z_2 = z_{(i)} \mid T] = \frac{p_{(i)}^2 \, \Sigma'' \frac{(n-2)!}{\alpha_{(1)}! \cdots \alpha_{(\nu)}!} p_{(1)}^{\alpha_{(1)}} \cdots p_{(\nu)}^{\alpha_{(\nu)}}}{\Sigma' \frac{n!}{\alpha_{(1)}! \cdots \alpha_{(\nu)}!} p_{(1)}^{\alpha_{(1)}} \cdots p_{(\nu)}^{\alpha_{(\nu)}}},$$

and

$$P[z_1 = z_{(i)}, z_2 = z_{(i')} \mid T] = \frac{p_{(i)} p_{(i')} \, \Sigma''' \frac{(n-2)!}{\alpha_{(1)} \cdots \alpha_{(\nu)}!} p_{(1)}^{\alpha_{(1)}} \cdots p_{(\nu)}^{\alpha_{(\nu)}}}{\Sigma' \frac{n!}{\alpha_{(1)}! \cdots \alpha_{(\nu)}!} p_{(1)}^{\alpha_{(1)}} \cdots p_{(\nu)}^{\alpha_{(\nu)}}}$$

$$(i \neq i'); \qquad \ldots \quad (4.3.5)$$

where Σ', Σ'' and Σ''' have meanings similar to those defined in (3.2.4) and (3.6.4). Therefore, we have from appendix I,

$$P[\, z_{(1)} = z_{(i)},\ z_{(2)} = z_{(i)} \mid T] = c_{(i,i)} ,$$

and

$$P[\, z_{(1)} = z_{(i)},\ z_{(2)} = z_{(i')} \mid T] = c_{(i,i')} . \qquad (4.3.6)$$

Using (4.3.4) and (4.3.6), we get

$$E(z_p \mid T) = \Sigma\, c_{(i,i)}\, z^2_{(i)} + \sum_{i \neq i'} c_{(i,i')}\, z_{(i)}\, z_{(i')} ,$$

$$\ldots \qquad (4.3.7)$$

which was to be proved.

Corollary 1. Improved estimator of σ_z^2 .

The usual estimator of $\sigma_z^2 = \Sigma P_j \left(\frac{Y_j}{P_j} - Y\right)^2$ is given by

$$s_z^2 = \frac{1}{(n-1)} \Sigma (z_i - \bar{z})^2 = \frac{1}{2n(n-1)} \sum_{i \neq i'} (z_i - z_{i'})^2 . \ldots (4.3.8)$$

Thus a uniformly better estimator than s_z^2 is given by

$$E(s_z^2 \mid T) = E\left[\frac{(z_1 - z_2)^2}{2} \mid T \right]$$

$$= \sum_{i \neq i'} c_{(i,i')} \frac{[z_{(i)} - z_{(i')}]^2}{2} . \qquad (4.3.9)$$

Corollary 2.

We can express $E(z_p \mid T)$ in a different form as follows:

Since

$$s_z^2 = \frac{1}{n} \sum z_i^2 - z_p ,$$

$$\therefore \quad E(z_p \mid T) = E\left[\frac{1}{n} \sum z_i^2 \mid T\right] - E\left[s_z^2 \mid T\right]$$

$$= \sum c_{(i)} z_{(i)}^2 - \sum_{i \neq i'} c_{(i,i')} \frac{[z_{(i)} - z_{(i')}]^2}{2} . \qquad \ldots \quad (4.3.10)$$

Corollary 3. Estimator of $V(\bar{z}_\nu)$.

An unbiased estimator of $V(\bar{z}_\nu)$ is given by

$$v(\bar{z}_\nu) = \bar{z}_\nu^2 - \sum c_{(i,i)} z_{(i)}^2 - \sum_{i \neq i'} c_{(i,i')} z_{(i)} z_{(i')} . \qquad (4.3.11)$$

Since this estimator is quite complicated for use in large samples, Basu has suggested the use of

$$\frac{1}{n(n-1)} \sum (z_i - \bar{z})^2 \qquad \ldots \quad (4.3.12)$$

as an estimator of $V(\bar{z}_\nu)$. As it over-estimates $V(\bar{z}_\nu)$, we are always on the safe side to use (4.3.12) as our estimator.

The estimators derived in this and preceding sections, though superior to the usually employed estimators, are not of much use for large scale sample surveys. The main advantage of simplicity of these sampling schemes would be lost if we use these estimators. In the next section, we give simpler estimators of Y and Y^2. These estimators though less efficient than the above derived estimators, are

superior to the usually employed estimators.

4.4. Simple improved estimators of Y and Y^2.

Let us suppose that the observed samples are segregated into groups of equal p_i's. For instance, consider the problem of estimating the total yield of a crop from a sample of farms. Every sample-farm is selected with probability proportional to its area. Here, if some crude approximation (say correct to an acre) is used to measure their areas, we expect to get a number of farms with same p_i in the sample . In the sequel, by the p-value of a unit, we mean the probability of selection associated with that unit. Let $p_{(1)}, p_{(2)}, \ldots, p_{(k)}$ be the distinct p-values of the sample units arranged in an increasing order of their magnitude. Let $n_{(i)}$ be the number of sample units having $p_{(i)}$ as their p-value. However, not all these $n_{(i)}$ units will be distinct, let $\nu_{(i)}$ be the number of distinct units among them. Now, if we arrange these $\nu_{(i)}$ distinct units in an ascending order of their unit-indices and call them $x_{(i1)}, x_{(i2)}, \ldots, x_{(i\,\nu_{(i)})}$, then it is not difficult to see that the statistic

$$T^* = \left[\left\{x_{(11)}, \ldots, x_{(1\,\nu_{(1)})}; n_{(1)}\right\}, \ldots, \left\{x_{(k1)}, \ldots, x_{(k\,\nu_{(k)})}; n_{(k)}\right\}\right] \quad \ldots \quad (4.4.1)$$

is sufficient.

It should be noted that if we take away the ancillary statistics $n_{(1)}, n_{(2)}, \ldots, n_{(k)}$ from the sufficient statistic T^*, then it reduces to the 'order-statistic', T, defined in the earlier section. The 'unnecessarily wide' sufficient statistic T^* is used here for the purpose of deriving estimators of Y and Y^2 that are much simpler (though somewhat less efficient) than those considered in the previous sections.

4.4A. Estimation of Y.

Theorem 3:

For any convex loss function, an estimator uniformly better than $\bar{z}$ is given by

$$\bar{z}^*_{\nu} = \frac{1}{n} \sum_{i=1}^{k} \frac{n_{(i)}}{p_{(i)}} \bar{y}_{\nu_{(i)}} , \ldots \qquad (4.4.2)$$

where

$$\bar{y}_{\nu_{(i)}} = \frac{1}{\nu_{(i)}} \sum_{r=1}^{\nu_{(i)}} y_{(ir)} .$$

Proof:

Evidently, by Rao-Blackwell theorem, an estimator uniformly better than $\bar{z}$ is given by

$$E(\bar{z} \mid T^*) = E\left[\frac{y_1}{p_1} \mid T^*\right] . \qquad \ldots \qquad (4.4.3)$$

Further, the probability of getting a sample with a given T^* is

$$P(T^*) = \frac{n!}{n_{(1)}! \cdots n_{(k)}!} \; p_{(1)}^{n_{(1)}} \cdots p_{(k)}^{n_{(k)}} \cdot c_{\nu_{(1)}}(n_{(1)}) \cdot \cdots c_{\nu_{(k)}}(n_{(k)}),$$

$$\cdots \qquad (4.4.4)$$

and

$$P[x_1 = x_{(ir)} \mid T^*] = \frac{p_{(i)} \dfrac{(n-1)!}{n_{(1)}! \cdots (n_{(i)}-1)! \cdots n_{(k)}!} p_{(1)}^{n_{(1)}} \cdots p_{(i)}^{n_{(i)}-1} \cdots p_{(k)}^{n_{(k)}}}{\dfrac{n!}{n_{(1)}! \cdots n_{(i)}! \cdots n_{(k)}!} p_{(1)}^{n_{(1)}} \cdots p_{(i)}^{n_{(i)}} \cdots p_{(k)}^{n_{(k)}}} \times$$

$$\frac{c_{\nu_{(1)}}(n_{(1)}) \cdots \dfrac{c_{\nu_{(i)}}(n_{(i)})}{\nu_{(i)}} \cdots c_{\nu_{(k)}}(n_{(k)})}{c_{\nu_{(1)}}(n_{(1)}) \cdots c_{\nu_{(i)}}(n_{(i)}) \cdots c_{\nu_{(k)}}(n_{(k)})}$$

$$= \frac{n_{(i)}}{n} \cdot \frac{1}{\nu_{(i)}} \cdot \qquad \cdots \qquad (4.4.5)$$

From (4.4.3) and (4.4.5), it follows that

$$E(\bar{z} \mid T^*) = \frac{1}{n} \sum_{i=1}^{k} \frac{n_{(i)}}{p_{(i)}} \bar{y}_{\nu_{(i)}},$$

which completes the proof of the theorem.

A simple comparison of $\bar{z}^*_{\nu}$ and $\bar{z}$ will show that $\bar{z}^*_{\nu}$ will be superior to $\bar{z}$ if and only if the sample size is greater than two and at least three units in the population have the same p-value.

otherwise $\bar{\bar{z}}^*_{>}$ and $\bar{z}$ will be identical. It is not difficult to give a direct proof of the fact that $V(\bar{\bar{z}}^*_{>}) \leq V(\bar{z})$. The strict sign of inequality holds only when the above condition is satisfied.

Variance of $\bar{\bar{z}}^*_{>}$.

We have

$$V(\bar{\bar{z}}^*_{>}) = E[V(\bar{\bar{z}}^*_{>} \mid n_{(1)}, \ldots, n_{(k)})] + V[E(\bar{\bar{z}}^*_{>} \mid n_{(1)}, \ldots, n_{(k)})]$$

$$= E\left[\sum_{i=1}^{k} \frac{n_{(i)}^2}{n^2 p_{(i)}^2} V(\bar{y}_{>(i)} \mid n_{(i)})\right] + V\left[\sum_{i=1}^{k} \frac{n_{(i)}}{n} \cdot \frac{\bar{Y}_{(i)}}{p_{(i)}}\right], \ldots (4.4.6)$$

where $\bar{Y}_{(i)}$ is the average of the population units having the p-value $p_{(i)}$.

Assuming that $P_1, P_2, \ldots, P_j, \ldots, P_K$ are the distinct p-values in the population, we get after simplifying (4.4.6)

$$V(\bar{\bar{z}}^*_{>}) = \frac{1}{n} \sum_{j=1}^{K} \frac{1}{P_j} \sum_{h=1}^{N_j-1} \left\{1 - (N_j-h)P_j\right\}^{n-2} \left\{1-(N_j-nh)\right\} S_j^2 + \frac{\sigma_{bz}^2}{n}, \quad \ldots \quad (4.4.7)$$

where

$$S_j^2 = \frac{1}{(N_j-1)} \Sigma (Y_j - \bar{Y}_j)^2, \quad \bar{Y}_j = \frac{1}{N_j} \Sigma Y_j$$

and

$$\sigma^2_{bz} = \sum_{j=1}^{K} N_j P_j \left(\frac{\bar{Y}_j}{P_j} - Y \right)^2 ,$$

the summation Σ runs over all N_j population units with the p-value P_j.

We, thus, see that $V(\bar{z}^*_{\nu})$ is made up of two components. The second component is unaltered if instead of using $\bar{y}_{\nu(i)}$ we use some other unbiased estimator of $\bar{Y}_{(i)}$. Consequently, in order to minimise the first component, various other estimators of $\bar{Y}_{(i)}$ of the form

$$f_1(\nu_{(i)})\, \bar{y}_{\nu(i)} + f_2(\nu_{(i)})$$

(where $E[f_1(\nu_{(i)}) \mid n_{(i)}] = 1$ and $E[f_2(\nu_{(i)}) \mid n_{(i)}] = 0$) can be used. To choose a reasonable estimator in this class of estimators of $\bar{Y}_{(i)}$, one may use the same criteria as discussed in Chapter III.

4.4B. Estimation of Y^2.

Theorem 4:

For any convex loss function, an estimator uniformly better than z_p is given by

$$s_p^* = \frac{1}{n(n-1)}\left[\left\{\left(\sum_{i=1}^{k} n_{(i)}\ \frac{\bar{y}_{\nu(i)}}{p_{(i)}}\right)^2 - \sum_{i=1}^{k} n_{(i)}\ \frac{\bar{y}^2_{\nu(i)}}{p^2_{(i)}}\right\}\right.$$

$$\left. - \sum_{i=1}^{k} n_{(i)}\,(n_{(i)}-1)\ \frac{c_{\nu(i)-1}\,(n_{(i)}-1)}{c_{\nu(i)}\,(n_{(i)})}\ \frac{s^2_{\nu(i)}}{p^2_{(i)}}\right], \quad \ldots \quad (4.4.8)$$

where

$$s^2_{\nu(i)} = \frac{1}{(\nu(i)-1)} \sum_{r=1}^{\nu(i)} \left(y_{(ir)} - \bar{y}_{\nu(i)} \right)^2 .$$

Proof:

Obviously, an estimator uniformly better than s_p is given by

$$E(s_p \mid T^*) = E\left[\frac{1}{n(n-1)} \sum_{i \neq i'} \left(\frac{y_i}{p_i}\right)\left(\frac{y_{i'}}{p_{i'}}\right) \mid T^*\right]$$

$$= E\left[\left(\frac{y_1}{p_1}\right)\left(\frac{y_2}{p_2}\right) \mid T^*\right] . \qquad \ldots \qquad (4.4.9)$$

Further, it is easy to verify that

$$P[x_1 = x_{(ir)},\ x_2 = x_{(ir)} \mid T^*] = \frac{n_{(i)}(n_{(i)}-1)}{n(n-1)} \cdot \frac{c_{\nu(i)}\,(n_{(i)}-1)}{\nu(i)\, c_{\nu(i)}\,(n_{(i)})} ;$$

$$P[x_1 = x_{(ir)},\ x_2 = x_{(ir')} \mid T^*] = \frac{n_{(i)}(n_{(i)}-1)}{n\,(n-1)}\left[\frac{c_{\nu(i)}(n_{(i)}) - c_{\nu(i)}(n_{(i)}-1)}{\nu(i)\,(\nu(i)-1)\,c_{\nu(i)}(n_{(i)})}\right];$$

$$(r \neq r')$$

and

$$P[x_1 = x_{(ir)},\ x_2 = x_{(i'r')} \mid T^*] = \frac{n_{(i)}n_{(i')}}{n(n-1)} \cdot \frac{1}{\nu_{(i)}} \cdot \frac{1}{\nu_{(i')}}$$

$$(i \neq i') . \qquad \ldots \qquad (4.4.10)$$

Therefore,

$$E(z_p \mid T^*) = \sum_{i=1}^{k} \frac{n_{(i)}(n_{(i)}-1)}{n(n-1)} \sum_{r=1}^{\nu_{(i)}} \frac{y_{(ir)}^2}{p_{(i)}^2} \cdot \frac{e^{\nu_{(i)}(n_{(i)}-1)}}{\nu_{(i)}\, e^{\nu_{(i)}(n_{(i)})}} +$$

$$\sum_{i=1}^{k} \frac{n_{(i)}(n_{(i)}-1)}{n(n-1)} \sum_{r\neq r'=1}^{\nu_{(i)}} \frac{y_{(ir)}y_{(ir')}}{p_{(i)}^2} \cdot \frac{[e^{\nu_{(i)}(n_{(i)})} - e^{\nu_{(i)}(n_{(i)}-1)}]}{\nu_{(i)}(\nu_{(i)}-1)e^{\nu_{(i)}(n_{(i)})}} +$$

$$\sum_{i\neq i'=1}^{k} \frac{n_{(i)}n_{(i')}}{n(n-1)} \sum_{r=1}^{\nu_{(i)}} \sum_{r'=1}^{\nu_{(i')}} \frac{y_{(ir)}}{p_{(i)}} \cdot \frac{y_{(i'r')}}{p_{(i')}} \cdot \frac{1}{\nu_{(i)}} \cdot \frac{1}{\nu_{(i')}} .$$

$$\ldots \qquad (4.4.11)$$

Using the equality

$$\frac{e^{\nu_{(i)}-1}(n_{(i)}-1)}{e^{\nu_{(i)}(n_{(i)})}} s^2_{\nu_{(i)}} = \bar{y}^2_{\nu_{(i)}} - \frac{e^{\nu_{(i)}(n_{(i)}-1)}}{\nu_{(i)}\, e^{\nu_{(i)}(n_{(i)})}} \sum_{r=1}^{\nu_{(i)}} y^2_{(ir)} -$$

$$\frac{[e^{\nu_{(i)}(n_{(i)})} - e^{\nu_{(i)}(n_{(i)}-1)}]}{\nu_{(i)}\ (\nu_{(i)}-1)\ e^{\nu_{(i)}(n_{(i)})}} \sum_{r\neq r'=1}^{\nu_{(i)}} y_{(ir)}y_{(ir')},$$

and simplifying (4.4.11), we get

$$s_p^* = E(s_p \mid T^*) = \frac{1}{n(n-1)} \left[\left\{ \left(\sum_{i=1}^{k} n_{(i)} \frac{\bar{y}_{\nu(i)}}{p_{(i)}} \right)^2 - \sum_{i=1}^{k} n_{(i)} \frac{\bar{y}^2_{\nu(i)}}{p^2_{(i)}} \right\} \right.$$

$$\left. - \sum_{i=1}^{k} n_{(i)}(n_{(i)}-1) \frac{e_{\nu(i)-1}(n_{(i)}-1)}{e_{\nu(i)}(n_{(i)})} \frac{s^2_{\nu(i)}}{p^2_{(i)}} \right] .$$

This completes the proof.

Corollary 1:

It is easy to see that

$$E(s_z^2 \mid T^*) = E\left(\frac{y_1^2}{p_1^2} \mid T^* \right) - E\left(\frac{y_1}{p_1} \cdot \frac{y_2}{p_2} \mid T^* \right)$$

$$= \sum_{i=1}^{k} \frac{n_{(i)}}{n} \frac{1}{\nu_{(i)}} \sum_{r=1}^{\nu(i)} \frac{y^2_{(ir)}}{p^2_{(i)}} + \sum_{i=1}^{k} \frac{n_{(i)}(n_{(i)}-1)}{n(n-1)} \frac{e_{\nu(i)-1}(n_{(i)}-1)}{e_{\nu(i)}(n_{(i)})} \frac{s^2_{\nu(i)}}{p^2_{(i)}}$$

$$- \frac{1}{n(n-1)} \left[\left(\sum_{i=1}^{k} n_{(i)} \frac{\bar{y}_{(i)}}{p_{(i)}} \right)^2 - \sum_{i=1}^{k} n_{(i)} \frac{\bar{y}^2_{\nu(i)}}{p^2_{(i)}} \right] \quad \ldots (4.4.12)$$

is a simple-improved estimator of σ_z^2 .

Corollary 2:

An unbiased estimator of $V(\bar{z}^*_{\nu})$ is given by

$$v(\bar{z}^*_{\nu}) = \bar{z}^{*2}_{\nu} + \sum_{i=1}^{k} \frac{n_{(i)}(n_{(i)}-1)}{n(n-1)} \frac{c_{\nu(i)-1}\,(n_{(i)}-1)}{c_{\nu(i)}\,(n_{(i)})} \frac{s^2_{\nu(i)}}{p^2_{(i)}}$$

$$- \frac{1}{n(n-1)} \left[\left(\sum_{i=1}^{k} n_{(i)} \frac{\bar{y}_{\nu(i)}}{p_{(i)}} \right)^2 - \sum_{i=1}^{k} n_{(i)} \frac{\bar{y}^2_{\nu(i)}}{p^2_{(i)}} \right] . \quad \ldots(4.4.13)$$

However, in practice it seems reasonable to use

$$\frac{1}{n(n-1)} \sum_{i=1}^{n} \left(\frac{y_i}{p_i} - \bar{z} \right)^2$$

as an estimator of $V(\bar{z}^*_{\nu})$. First, because it is simple to compute; secondly, because it is always non-negative. Besides this, we are on the safe side as it always over-estimates the variance of $\bar{z}^*_{\nu}$.

CHAPTER V

USE OF 'ORDER-STATISTIC' IN WITHOUT REPLACEMENT SAMPLING

Summary.

In sampling without replacement from a finite population, the order in which the units are selected, is immaterial for the purpose of estimation. This point was noted by Basu [4] and Murthy [31]. Basu showed that the 'order-statistic' (sample units arranged in ascending order of their unit-indices) forms a sufficient statistic, and therefore, any estimator which is not a function of the 'order-statistic', can be uniformly improved by the use of Rao-Blackwell theorem. In this chapter certain results obtained by Murthy [31] are shown to be immediate consequences of the above observation.

It is shown that sampling with different probabilities with replacement until we get a specified number of distinct units, is equivalent in some sense to sampling with different probabilities without replacement. Some other related problems are also considered here.

5.1. Sampling without replacement.

In sampling without replacement from a population containing N units, a particular sample may be recorded as

$$s = (x_1, x_2, \ldots, x_n),$$

where $x_i = (y_i, p_i, u_i)$ $(i = 1, 2, \ldots, n)$ and n is the sample size; all these symbols have already been defined in Chapter III.

The probability of drawing such a particular sample is given by

$$P(S) = p_1 \cdot \frac{p_2}{(1-p_1)} \cdot \frac{p_3}{(1-p_1-p_2)} \cdots \cdots \cdot \frac{p_n}{(1-p_1-p_2-\cdots-p_{n-1})} \cdot$$

$$\cdots \qquad (5.1.1)$$

If we record the 'order-statistic' by

$$T = [\, x_{(1)}, x_{(2)}, \ldots, x_{(n)}] \,,$$

where $x_{(i)} = [\, y_{(i)}, p_{(i)}, u_{(i)}]$ is the i-th order-statistic $(i=1,\ldots,n)$. We have

$$P(T) = \sum_{S \supset T} \frac{p_1 p_2 \cdots p_n}{(1-p_1)(1-p_1-p_2) \cdots (1-p_1-p_2-\cdots-p_{n-1})} \,, \quad \cdots \quad (5.1.2)$$

where the summation is taken over all possible samples giving rise to the 'order-statistic' T.

It has been shown by Basu [4] that T is a sufficient statistic. Thus, if $g(S)$ is some estimator depending on S, by Rao-Blackwell theorem, a uniformly better estimator than $g(S)$ is given by $E[g(S) \mid T]$. For any convex loss function, the risk associated with $E[g(S) \mid T]$ is smaller than the risk associated with $g(S)$.

5.2. Sampling with replacement. [when the number of distinct units is fixed in advance]

In this case, units are drawn with unequal probabilities and with replacement until we get a specified number 'n' of distinct units. If r denotes the number of draws in a particular case, the sample S may be recorded as

$$S = (x_1, x_2, \ldots, x_r).$$

If we denote the 'order-statistic' T by

$$T = [\, x_{(1)}, x_{(2)}, \ldots, x_{(n)}],$$

where $x_{(i)}$ is the i-th order-statistic $(i = 1, \ldots, n)$.
It is not difficult to show that

$$P(T) = \sum_{r=2}^{\infty} \Big[\sum_{i=1}^{n} p_{(i)} \Big\{ \Sigma^{(i)} (p_{(1)} + \ldots + p_{(n)})^{r-1} -$$

$$\Sigma^{(i)} (p_{(1)} + \ldots + p_{(n-1)})^{r-1} + \ldots (-)^{n-2} \Sigma^{(i)} p_{(1)}^{r-1} \Big\} \Big], \ldots \quad (5.2.1)$$

where $\Sigma^{(i)}$ denotes the summation over all possible combinations out of $p_{(1)}, \ldots, p_{(i-1)}, p_{(i+1)}, \ldots, p_{(n)}$; and the term inside the square brackets denotes the probability of getting T in r draws. Assuming

without any loss of generality that $p_{(1)} + \cdots + p_{(n)} < 1$, we get on summing (5.2.1) over r

$$P(T) = \sum_{i=1}^{n} p_{(i)} \left[\Sigma^{(i)} \frac{p_{(1)} + \cdots + p_{(n)}}{1-p_{(1)} - \cdots - p_{(n)}} - \Sigma^{(i)} \frac{p_{(1)} + \cdots + p_{(n-1)}}{1-p_{(1)} - \cdots - p_{(n-1)}} + \right.$$

$$\cdots \quad \cdots \quad \left. + (-)^{n-2} \Sigma^{(i)} \frac{p_{(1)}}{1-p_{(1)}} \right]. \quad \cdots \quad (5.2.2)$$

It can be proved by induction over n that (5.2.2) and (5.1.2) are equal.

Thus, if we rely only on the 'order-statistic' T, the two methods of sampling are essentially the same.

5.3. Improving Des Raj's estimators.

Des Raj [13], in sampling without replacement, gave the following set of uncorrelated estimators of $Y = \Sigma Y_j$.

$$t_1(s) = \frac{y_1}{p_1} ;$$

$$t_2(s) = y_1 + \frac{y_2}{p_2}(1 - p_1);$$

$$\cdots \quad \cdots \quad \cdots$$

$$t_i(s) = y_1 + y_2 + \cdots + y_{i-1} + \frac{y_i}{p_i}(1 - p_1 - p_2 - \cdots - p_{i-1});$$

$$\cdots \quad \cdots \quad \cdots$$

$$t_n(s) = y_1 + y_2 + \cdots + y_{n-1} + \frac{y_n}{p_n}(1 - p_1 - p_2 - \cdots - p_{n-1}). \quad (5.3.1)$$

Theorem 1:

For any convex loss function, a uniformly better estimator than

$$t(S) = \sum_{i=1}^{n} c_i \, t_i(S) \qquad \left(\sum_{i=1}^{n} c_i = 1 \right)$$

is given by

$$\sum_{i=1}^{n} y_{(i)} \quad \frac{P[T|(i)]}{P(T)} \quad , \qquad \cdots \qquad \cdots \qquad (5.3.2)$$

where $P(T \mid (i))$ is the conditional probability of getting the 'order-statistic' T given that i-th order unit was drawn first.

Proof:

Since

$$\frac{P[\, x_i = x_{(h)},\, x_{i+1} = x_{(k)} \mid x_1, x_2, \cdots, x_{i-1}]}{P[\, x_i = x_{(k)},\, x_{i+1} = x_{(h)} \mid x_1, x_2, \cdots, x_{i-1}]}$$

$$= \left[\frac{1-p_1-p_2- \cdots - p_{i-1} - p_{(k)}}{1-p_1-p_2- \cdots - p_{i-1} - p_{(h)}} \right] \cdot \cdot \qquad \cdots \qquad \cdots \qquad (5.3.3)$$

for all $h \neq k = 1, 2, \ldots, n$, it follows that

$$E\left[t_{i+1}(S) - t_i(S) \mid T \right] = E\left[\frac{y_{i+1}}{p_{i+1}} (1-p_1-\cdots- p_i) - \frac{y_i}{p_i} (1-p_1-p_2-\cdots-p_i) \mid T \right]$$

$$= 0 \, .$$

Therefore,

$$E\Big[\sum_{i=1}^{n} c_i t_i(S) \mid T\Big] = E\big[t_1(S) \mid T\big] = \sum_{i=1}^{n} y_{(i)} \frac{P[T \mid (i)]}{P[T]}.$$

Corollary 1:

When $n = 2$, we have

$$E\Big[\sum_{i=1}^{2} c_i t_i(S) \mid T\Big] = \frac{1}{(2-p_{(1)}-p_{(2)})}\Big[(1-p_{(2)})\frac{y_1}{p_{(1)}} + (1-p_{(1)})\frac{y_{(2)}}{p_{(2)}}\Big].$$

Corollary 2:

In simple random sampling(without replacement)

$$t_i(S) = y_1 + \ldots + y_{i-1} + (N-i+1)\, y_i, \quad i=1, \ldots, n;$$

and

$$E\big[t_i(S) \mid T\big] = E\big[t_1(S) \mid T\big] = \frac{N}{n} \sum_{i=1}^{n} y_{(i)}. \qquad \ldots (5.5.4)$$

Theorem 2:

A uniformly better estimator than

$$g(S) = \sum_{i \neq j=1}^{n} c_{ij}\, t_i(S) t_j(S) \qquad \Big(\sum_{i \neq j=1}^{n} c_{ij} = 1\Big)$$

of Y^2, is given by

$$g(T) = E[\, g(S) \mid T] = \frac{\sum_{i=1}^{n} y_{(i)}^2 \, P[\, T \mid (i)] + \sum_{i \neq i'=1}^{n} y_{(i)} y_{(i')} P[T \mid (i),(i')]}{P(T)}, \qquad \ldots \quad (5.3.5)$$

where $P[\, T \mid (i),(i')]$ is the conditional probability of getting T given that $x_1 = x_{(i)}$ and $x_2 = x_{(i')}$.

Proof:

Using (5.3.5), it can be seen that

$$E[\, t_1(S)\, t_{j+1}(S) - t_1(S)\, t_j(S) \mid T] = 0 \quad (j = 2, \ldots, n)$$

and

$$E[\, t_{i+1}(S)\, t_j(S) - t_i(S)\, t_j(S) \mid T\,] = 0, \quad (i = 1, 2, \ldots, j-2)$$

and hence,

$$E[\, g(S) \mid T] = E[\, t_1(S)\, t_2(S) \mid T]$$

$$= \frac{\sum_{i=1}^{n} y_{(i)}^2 \, P[T \mid (i)] + \sum_{i \neq i'=1}^{n} y_{(i)}\, y_{(i')}\, P[\, T \mid (i), (i')]}{P(T)},$$

which was to be proved.

Remark: The estimators (5.3.2) and (5.3.5) can also be got by improving the usual estimators of Y and Y^2 under the sampling scheme discussed in Section 5.2.

Corollary 1:

When $n = 2$, we have

$$E[\, t_1(s)\, t_2(s) \mid T] = \frac{1}{(2-p_1-p_2)} \left[(1-p_2)\frac{y_1^2}{p_1} + (1-p_1)\frac{y_2^2}{p_2} + 2(1-p_1)(1-p_2)\frac{y_1}{p_1}\frac{y_2}{p_2}\right] \quad \ldots (5.3.5a)$$

Corollary 2:

In simple random sampling (without replacement)

$$E[\, t_i(s)\, t_j(s) \mid T] = \frac{N}{n} \sum_{i=1}^{n} y_{(i)}^2 + \frac{N(N-1)}{n(n-1)} \sum_{i \neq j=1}^{n} y_{(i)} y_{(j)}. \qquad (5.3.5b)$$

The estimator (5.3.5) is used to derive unbiased variance estimator of $\sum_{i=1}^{n} y_{(i)} \frac{P(T \mid (i))}{P(T)}$.

5.4. Improving Das' estimators.

The set of estimators of Y given by Das [8] is as follows:

$$u_1(s) = \frac{y_1}{p_1};$$

$$u_2(s) = \frac{y_2}{p_2} \cdot \frac{(1-p_1)}{p_1} \cdot \frac{1}{(N-1)};$$

$$\ldots \qquad \ldots \qquad \ldots$$

$$u_r(s) = \frac{y_r}{p_r} \cdot \frac{(1-p_1-p_2-\ldots-p_{r-1})}{p_{r-1}} \cdots \frac{(1-p_1-p_2)}{p_2} \cdot \frac{(1-p_1)}{p_1} \cdot \frac{1}{(r-1)!\binom{N-1}{r-1}};$$

$$\ldots \qquad \ldots \qquad \ldots \qquad \ldots \qquad \ldots$$

$$u_n(s) = \frac{y_n}{p_n} \cdot \frac{(1-p_1-\ldots-p_{n-1})}{p_{n-1}} \cdot \ldots \cdot \frac{(1-p_1)}{p_1} \cdot \frac{1}{(n-1)!\binom{N-1}{r-1}} \cdot \quad \ldots (5.4.1)$$

A uniformly better estimator than $u_r(S)$ is given by

$$u_r(T) = E\left[u_r(S) \mid T\right]$$

$$= \sum_{i=1}^{n} y_{(i)} \frac{\sum \frac{1}{(N-1)(N-2)\ldots(N-r+1)} P[T \mid x_1, \ldots, x_{r-1}, x_r = x_{(i)}]}{P[T]}, \quad \ldots \quad \ldots \quad (5.4.2)$$

where the summation $\sum$ is taken over all possible $x_1, \ldots, x_{r-1}$.

It is easy to see that the estimators $u_r(T)$ $(r=1,\ldots, n)$ are identical if and only if the sample is drawn by simple random sampling (without replacement). In this case (5.4.2) is same as (5.3.4).

This shows that in simple random sampling (without replacement) the estimator based on the sample mean is more efficient than Das' as well as Des Raj's estimators.

An unbiased estimator of Y^2 based on $u_r(S)$ and y_q $(q < r)$, is given by

$$v_{qr}(S) = u_r(S)y_r + (N-1)u_r(S)y_q . \quad \ldots \quad \ldots \quad (5.4.3)$$

$$(q < r = 1, 2, \ldots, n)$$

A uniformly better estimator than this is given by

$$E[v_{qr}(S)|T] = \sum_{i=1}^{n} y_{(i)}^2 \frac{\left[\sum \frac{1}{(N-1)}\cdots\frac{1}{(N-r+1)} P(T|x_1,\ldots,x_{r-1},x_r=x_{(i)})\right]}{P(T)} +$$

$$(N-1) \sum_{i\neq i'=1}^{n} y_{(i)}y_{(i')} \frac{\left[\sum \frac{1}{(N-1)}\cdots\frac{1}{(N-r+1)} P(T|x_1,\ldots,x_q=x_{(i)},\ldots x_r=x_{(i')})\right]}{P(T)}.$$

... ... (5.4.4)

This expression will also be identical for all r and k if and only if the sample is drawn by simple random sampling (without replacement).

Further, it may be seen on similar lines that in a more general (without replacement) sampling scheme which has been considered by Des Raj (13), the estimators of Y (or of Y^2) obtained by improving Das' estimators will be identical if and only if:

The first unit in the sample is selected with pre-assigned probabilities and the remaining units are selected by simple random sampling (without replacement). For further reference about this, one may refer to Des Raj (13) and Murthy (31).

CHAPTER VI

ESTIMATION PROBLEM IN SOME GENERAL SAMPLING SCHEMES

Summary.

In this chapter, we have extended the technique of improving estimators to two-stage ~~xxx~~ sampling schemes. Improved estimators of the population total, Y, and their variance estimators are derived. Similar to Chapter IV, two sets of estimators of Y are given here. The first estimator is easy to compute in practice, whereas the second though tedious to compute, is more efficient than the first.

6.1. Preliminaries for two-stage sampling schemes.

Let $X_1, X_2, \ldots, X_j, \ldots, X_N$ be the N first-stage units of a population. Suppose that X_j consists of M_j second-stage units. Let Y_{jh} be some real valued characteristic of the h-th second-stage unit of X_j $(h = 1, \ldots, M_j)$[1] in which we are interested. In conformity with the notations used in previous chapters, capital letters refer to the population and small letters refer to the sample. For example, $x_1, x_2, \ldots, x_n$ stand for the n first-stage (in order of draw) sample units[2]; by x_{ir}, we mean the r-th (in order of draw)

---

1. Throughout this chapter, j runs from 1 to N, h from 1 to M_j, i from 1 to n, r from 1 to s_{ui}, (i) from (1) to (ν), and (ir) from (i1) to $(i\,\nu_{(i)})$, unless otherwise stated.

2. All relevant information about the units, such as, their unit-indices, probabilities of selection etc., are incorporated in the symbols x_i and x_{ir} .

second-stage unit in the i-th (in order of draw) first-stage sample unit. We assume that first-stage units are selected with unequal probabilities (with replacement), and if the j-th first-stage unit is included λ_j times in the sample, λ_j sub-samples of m_j units each will be drawn therefrom independently of each other, every sub-sample being drawn according to a given sampling method.

6.2. Application to two-stage sampling [unequal probabilities for first-stage and equal probabilities with replacement for second-stage].

In this sampling scheme, the first-stage units are selected with unequal probabilities (with replacement), and the second-stage units are selected with equal probabilities (with replacement). Let P_j be the probability of selection of the j-th first-stage unit ($\Sigma P_j = 1$), and let us call

$$z_{jh} = M_j \frac{Y_{jh}}{P_j}, \qquad \ldots \qquad (6.2.1)$$

the z-value of the h-th second-stage unit of X_j.

The usual estimator of the population total, $Y = \sum_j \sum_h Y_{jh}$, is given by

$$\bar{z}_n = \frac{1}{n} \Sigma \bar{z}_i, \qquad \ldots \qquad (6.2.2)$$

where $\bar{z}_i = \frac{1}{m} \sum_{r=1}^{m_{u_i}} z_{ir}$, z_{ir} is the z-value of x_{ir} and u_i is the unit-index of the i-th (in order of draw) first-stage sample unit[1].

1. The symbol m_{u_i} is commonly denoted by m_i; we have used this symbol to avoid confusion with the previous use of m_j for the sub-sample size of X_j.

Let $x_{(1)}, x_{(2)}, \ldots, x_{(\nu)}$ be the distinct first-stage units in the sample arranged in an increasing order of their unit-indices. Let $\lambda_{(i)}$ be the number of times $x_{(i)}$ is selected in the sample. Finally, let $x_{(i1)}, \ldots, x_{(i\nu_{(i)})}$ be the distinct second-stage units of $x_{(i)}$ arranged in an increasing order of their unit-indices.

Now it is not difficult to show that the statistic

$$T^* = \left[\left\{ x_{(i)}, \lambda_{(i)}; x_{(i1)}, \ldots, x_{(i\nu_{(i)})} \right\} i = 1, \ldots, \nu \right] \quad \ldots \quad (6.2.3)$$

is sufficient; and the probability of getting a sample with a given T^* is

$$P(T^*) = \frac{n!}{\lambda_{(1)}! \cdots \lambda_{(\nu)}!} p_{(1)}^{\lambda_{(1)}} \cdots p_{(\nu)}^{\lambda_{(\nu)}} \prod_{(i)} \frac{e^{\nu_{(i)}} (m_{(i)}^{\lambda_{(i)}})}{M_{(i)}^{m_{(i)}\lambda_{(i)}}} . \quad \ldots \quad (6.2.4)$$

We, therefore, have the following:

Theorem 1:

For any convex loss function, an estimator uniformly better than $\bar{z}_n$ is given by

$$\bar{z}^*_\nu = \frac{1}{n} \sum \lambda_{(i)} \bar{z}_{\nu_{(i)}} , \quad \ldots \quad (6.2.5)$$

where $\bar{z}_{\nu_{(i)}} = \frac{1}{\nu_{(i)}} \sum_{(ir)} z_{(ir)}$, $z_{(ir)}$ being the z-value of $x_{(ir)}$.

Proof:

Clearly, an estimator uniformly better than $\bar{\bar{z}}_n$ is given by

$$E\left(\bar{\bar{z}}_n \mid T^*\right) = E\left(z_{11} \mid T^*\right) \quad \ldots \quad (6.2.6)$$

Since

$$P[x_{11} = x_{(ir)} \mid T^*] = \frac{p_{(i)} \dfrac{(n-1)!}{\lambda_{(1)}! \cdots (\lambda_{(i)}-1)! \cdots \lambda_{(\nu)}!} p_{(1)}^{\lambda_{(1)}} \cdots p_{(i)}^{\lambda_{(i)}-1} \cdots p_{(\nu)}^{\lambda_{(\nu)}}}{\dfrac{n!}{\lambda_{(1)}! \cdots \lambda_{(i)}! \cdots \lambda_{(\nu)}!} p_{(1)}^{\lambda_{(1)}} \cdots p_{(i)}^{\lambda_{(i)}} \cdots p_{(\nu)}^{\lambda_{(\nu)}}} \times$$

$$\frac{\dfrac{1}{M_{(i)}} \left[\prod_{i' \neq i=1}^{\nu} \dfrac{c_{\nu(i')}\left(m_{(i')}\, \lambda_{(i')}\right)}{m_{(i')}^{\lambda_{(i')}}} \right] \dfrac{c_{\nu(i)}\left(m_{(i)}\, \lambda_{(i)}\right)}{\nu_{(i)}\, m_{(i)}^{\lambda_{(i)}-1}}}{\prod_{i'=1}^{\nu} \dfrac{c_{\nu(i')}\left(m_{(i')}\, \lambda_{(i')}\right)}{m_{(i')}^{\lambda_{(i')}}}}$$

$$= \frac{\lambda_{(i)}}{n} \cdot \frac{1}{\nu_{(i)}}, \quad \ldots \quad (6.2.7)$$

we thus have from (6.2.6)

$$E\left(\bar{\bar{z}}_n \mid T^*\right) = \frac{1}{n} \sum_{(i)} \frac{\lambda_{(i)}}{\nu_{(i)}} \sum_{(i\,r)} z_{(ir)} = \frac{1}{n} \sum \lambda_{(i)}\, \bar{\bar{z}}_{\nu(i)} .$$

Hence the theorem is proved.

Variance of $\bar{z}^*_{\nu}$.

Now

$$V(\bar{z}^*_{\nu}) = E[V(\frac{1}{n}\Sigma \lambda_{(i)}\, \bar{z}_{\nu(i)} \mid \lambda_{(1)}, \cdots, \lambda_{(\nu)})] +$$

$$V[E(\frac{1}{n}\Sigma \lambda_{(i)}\, \bar{z}_{\nu(i)} \mid \lambda_{(1)}, \cdots, \lambda_{(\nu)})]$$

$$= \frac{1}{n^2} E[\Sigma \lambda^2_{(i)}\, V(\bar{z}_{\nu(i)} \mid \lambda_{(i)})] + V[\frac{1}{n}\Sigma \lambda_{(i)}\, \bar{Z}_{(i)}], \qquad (6.2.8)$$

where $\bar{Z}_{(i)} = \frac{1}{M_{(i)}} \sum_{h=1}^{M_{(i)}} Z_{(ih)}$.

The above equation after some simplification reduces to

$$V(\bar{z}^*_{\nu}) = \frac{1}{n}\Sigma P_j S^2_j(z) \sum_{\ell=1}^{M_j-1} \frac{1}{\ell} \left(\frac{\ell}{M_j}\right)^{m_j} \left\{1-P_j+n\left(\frac{\ell}{M_j}\right)^{m_j} P_j\right\} \cdot$$

$$\left\{1-P_j+\left(\frac{\ell}{M_j}\right)^{m_j} P_j\right\}^{n-2} + \frac{\sigma^2_{bz}}{n}, \quad \cdots \qquad (6.2.9)$$

where

$$S^2_j(z) = \frac{1}{(M_j-1)} \sum_h (Z_{jh} - \bar{Z}_j)^2$$

and

$$\sigma^2_{bz} = \Sigma P_j (\bar{Z}_j - Y)^2 .$$

An unbiased estimator of $V(\bar{\bar{z}}^*_{\nu})$ is given by

$$v_1(\bar{\bar{z}}^*_{\nu}) = \bar{\bar{z}}^{*2}_{\nu} - \frac{\sum_{i \neq i'=1}^{n} \bar{z}_i \bar{z}_{i'}}{n(n-1)}, \quad \ldots \quad (6.2.10)$$

where $\bar{z}_i$ has been defined in (6.2.2).

An estimator uniformly better than $v_1(\bar{\bar{z}}^*_{\nu})$ is given by

$$v_2(\bar{\bar{z}}^*_{\nu}) = \bar{\bar{z}}^{*2}_{\nu} + \sum \frac{\lambda_{(i)}(\lambda_{(i)}-1)}{n(n-1)} \cdot \frac{c_{\nu_{(i)}-1}(m_{(i)}\lambda_{(i)}-1)}{c_{\nu_{(i)}}(m_{(i)}\lambda_{(i)})} s^2_{\nu_{(i)}}(z)$$

$$- \frac{1}{n(n-1)}\left[\left(\sum \lambda_{(i)} \bar{z}_{\nu_{(i)}} \right)^2 - \sum \lambda_{(i)} \bar{z}^2_{\nu_{(i)}} \right], \quad \ldots \quad (6.2.11)$$

where

$$s^2_{\nu_{(i)}}(z) = \begin{cases} \frac{1}{(\nu_{(i)}-1)} \sum_{(ir)} (z_{(ir)} - \bar{z}_{\nu_{(i)}})^2 & \text{if } \nu_{(i)} > 1, \\ 0 & \text{otherwise.} \end{cases}$$

Note the similarity between the expressions (6.2.11) and (4.4.13). The proof for the above formula runs on lines parallel to that for (4.4.13).

An estimator better than $\bar{x}^*_{\nu}$.

If from the statistic T^*, we take out ancillary statistics $\lambda_{(1)}, \ldots, \lambda_{(\nu)}$, we get another sufficient statistic

$$T = \left[\left\{ x_{(i)}; x_{(i1)}, \ldots, x_{(i\,\nu_{(i)})} \right\} \quad i = 1,2, \ldots, \nu \right]. \qquad \ldots \qquad (6.2.12)$$

The statistic T is smaller than T^*. Therefore, any estimator, which depends on T^*, can again be uniformly improved by the use of the well-known Rao-Blackwell theorem.

Clearly, the probability of getting a sample with a given T is

$$P(T) = \Sigma' \frac{n!}{\lambda_{(1)}! \cdots \lambda_{(\nu)}!} P_{(1)}^{\lambda_{(1)}} \cdots P_{(\nu)}^{\lambda_{(\nu)}} \prod_{(i)} \frac{c_{\nu_{(i)}}(m_{(i)}\lambda_{(i)})}{m_{(i)}^{\lambda_{(i)}}}, \qquad \ldots \qquad (6.2.13)$$

where Σ' stands for the summation over all positive integral $\lambda_{(i)}$'s such that $\Sigma\, \lambda_{(i)} = n$.

From the results of Appendix II, it follows that

$$P(T) = \sum_{r=0}^{[\sum \nu_{(i)}-1]} (-)^r \sum_1 \binom{\nu_{(1)}}{\alpha_1} \cdots \binom{\nu_{(2)}}{\alpha_\nu} [p_{(1)} \left(\frac{\nu_{(1)} - \alpha_1}{M_{(1)}}\right)^{m(1)} + \cdots +$$

$$p_{(\nu)} \left(\frac{\nu_{(\nu)} - \alpha_\nu}{M_{(\nu)}}\right)^{m(\nu)}]^n , \quad \cdots \qquad (6.2.14)$$

where $\sum_1$ stands for the summation over all non-negative integral α_i's such that $\alpha_1 + \alpha_2 + \cdots + \alpha_\nu = r$.

Theorem 2:

For any convex loss function, an estimator uniformly better than $\bar{z}^*_\nu$ is given by

$$\bar{z}_\nu = \sum c_{(i)} \bar{z}_{\nu(i)} , \quad \cdots \qquad (6.2.15)$$

where

$$c_{(i)} = p_{(i)} \Bigg\{ \sum_{r=0}^{[\sum \nu_{(i)}-1]} (-)^r \sum_1 \binom{\nu_{(1)}}{\alpha_1} \cdots \binom{\nu_{(2)}}{\alpha_\nu} \left(\frac{\nu_{(i)} - \alpha_i}{M_{(i)}}\right)^{m(i)} [p_{(1)} \left(\frac{\nu_{(1)} - \alpha_1}{M_{(1)}}\right)^{m(1)} + \cdots$$

$$\cdots + p_{(\nu)} \left(\frac{\nu_{(\nu)} - \alpha_\nu}{M_{(\nu)}}\right)^{m(\nu)}]^{n-1} \Bigg/ P(T)$$

Proof:

It is obvious that an estimator uniformly better than $\bar{z}^*_\nu$ is given by

$$E(\bar{z}^{*}_{\nu} \mid T) = \sum E\left(\frac{\lambda_{(i)}}{n} \mid T\right) \bar{z}_{\nu} \quad . \quad \ldots \qquad (6.2.16)$$

Moreover,

$$E\left[\frac{\lambda_{(i)}}{n} \mid T\right] = \frac{\sum' \frac{\lambda_{(i)}}{n} \cdot \frac{n!}{\lambda_{(1)}! \ldots \lambda_{(\nu)}!} p_{(1)}^{\lambda_{(1)}} \ldots p_{(\nu)}^{\lambda_{(\nu)}} \prod_{(i)} \frac{\theta_{\nu(i)}(m_{(i)}\lambda_{(i)})}{m_{(i)}\, \theta_{(i)}}}{P(T)},$$

where Σ' has been defined in (6.2.15).

On the lines similar to those given in Appendix II, it can be shown that

$$E\left(\frac{\lambda_{(i)}}{n} \mid T\right) = \theta_{(i)}, \qquad \ldots \qquad (6.2.17)$$

The theorem follows by combining (6.2.16) and (6.2.17).

Corollary 1:

When $P_j = \frac{M_j}{\sum M_j}$ and $m_j = 1$, the above estimator takes the simple form

$$\bar{z}_{\nu} = \frac{1}{\sum \nu_{(i)}} \left[\sum \nu_{(i)}\, \bar{z}_{\nu(i)} \right].$$

Though the estimator $\bar{z}_{\nu}$ is superior to $\bar{z}^{*}_{\nu}$, it cannot be of much use in practice, unless $P_j = \frac{M_j}{\sum M_j}$ and $m_j = 1$. It is better to rely on the estimator $\bar{z}^{*}_{\nu}$, which though less efficient than $\bar{z}_{\nu}$, has the merit of simplicity.

6.3. Application to two-stage sampling [unequal probabilities for first-stage and equal probabilities without replacement for second-stage].

Let us now consider the commonly adopted procedure of two-stage sampling. Here, the first-stage units are selected as usual with unequal probabilities, but whenever a specified first-stage unit, say the j-th, is included in the sample, a sub-sample of m_j second-stage units is drawn therefrom by simple random sampling (without replacement). If the j-th first-stage unit is included λ_j times, λ_j such sub-samples are drawn independently of each other.

Following the notations defined in Section 6.1 and Section 6.2, we record the following sufficient statistic:

$$T^* = \left[\left\{ x_{(i)}, \lambda_{(i)}, x_{(i1)}, \ldots, x_{(i\,\nu_{(i)})} \right\} i = 1, \ldots, \nu \right] . \qquad \ldots \quad (6.3.1)$$

A little investigation will now show that the probability of getting a sample with a given T^* is (Feller, 16)

$$P(T^*) = \frac{n!}{\lambda_{(1)}! \cdots \lambda_{(\nu)}!} P_{(1)}^{\lambda_{(1)}} \cdots P_{(\nu)}^{\lambda_{(\nu)}} \times \prod_{(i)} \frac{1}{\left[\binom{M_{(i)}}{m_{(i)}}\right]^{\lambda_{(i)}}} \left\{ \left[\binom{\nu_{(i)}}{m_{(i)}}\right]^{\lambda_{(i)}} - \binom{\nu_{(i)}}{1}\left[\binom{\nu_{(i)}-1}{m_{(i)}}\right]^{\lambda_{(i)}} + \cdots \right.$$

$$\left. \cdots (-)^{\nu_{(i)}-m_{(i)}} \binom{\nu_{(i)}}{\nu_{(i)}-m_{(i)}} \left[\binom{m_{(i)}}{m_{(i)}}\right]^{\lambda_{(i)}} \right\} \cdots \quad (6.3.2)$$

We thus have:

Theorem 3:

For any convex loss function, an estimator uniformly better than $\bar{z}_n$ is

$$\bar{z}_3^* = \frac{1}{n} \sum \lambda_{(i)} \bar{z}_{\nu(i)} \quad . \qquad \ldots \qquad (6.3.3)$$

Proof:

Obviously, an estimator uniformly better than $\bar{z}_n$ is

$$E(\bar{z}_n \mid T^*) = E(\bar{z}_1 \mid T^*) \quad . \qquad \ldots \qquad (6.3.4)$$

It may be noted that $\bar{z}_1$ is the average of the z-values of $x_{11}, x_{12}, \ldots, x_{1m_{u_1}}$.

Next, it can be seen that[1]:

$$P[x_1 = x_{(i)}; x_{11} = x_{(i1)}, \ldots, x_{1m_{u_1}} = x_{(im_{(i)})} \mid T^*] = \frac{\lambda_{(i)}}{n} \frac{1}{\binom{\nu(i)}{m_{(i)}}}, \qquad \ldots \qquad (6.3.5)^2$$

where $x_{(i1)}, \ldots, x_{(im_{(i)})}$ are the $m_{(i)}$ distinct second-stage units of $x_{(i)}$ taken from $x_{(i1)}, x_{(i2)}, \ldots, x_{(i\nu(i))}$. As the choice of these $m_{(i)}$ second-stage units is arbitrary, it follows from (6.3.4)

---

1. Note that the assumption $x_1 = x_{(i)}$ implies that $m_{u_1} = m_{(i)}$.

2. The equations(6.3.2) and (6.3.5) will be obvious from the similar equations of this nature derived in Chapter VIII.

and (6.3.5) that

$$E(\bar{\bar{z}}_n \mid T^*) = \frac{1}{n} \sum_{(i)} \lambda_{(i)} \frac{1}{\binom{\nu_{(i)}}{m_{(i)}}} \sum \bar{z}_{(i)}(m_{(i)}),$$

where the summation $\sum$ runs over all possible combinations of $m_{(i)}$ distinct $x_{(ir)}$'s chosen out of $x_{(i1)}, \cdots, x_{(i\,\nu_{(i)})}$, and $\bar{z}_{(i)}(m_{(i)})$'s denote the averages of the z-values of these combinations of $x_{(ir)}$'s. Therefore,

$$E(\bar{\bar{z}}_n \mid T^*) = \frac{1}{n} \sum_{(i)} \lambda_{(i)} \bar{z}_{\nu_{(i)}} ,$$

which completes the proof of the theorem.

Variance of $\bar{\bar{z}}^*_\nu$:

Similar to (6.2.9), it can be shown that

$$V(\bar{\bar{z}}^*_\nu) = \frac{1}{n} \sum_{j=1}^{N} P_j S_j^2(z) \sum_{\ell=1}^{M_j - m_j} \frac{1}{(M_j - \ell)} \phi_j(\ell) \Big\{ 1 - P_j + n \phi_j(\ell) P_j \Big\} \Big\{ 1 - P_j + \phi_j(\ell) P_j \Big\}^{n-2} + \frac{\sigma^2_{bs}}{n} , \quad (6.3.6)$$

where

$$\phi_j(\ell) = \frac{\binom{M_j - \ell}{m_j}}{\binom{M_j}{m_j}} .$$

An unbiased estimator of $V(\bar{z}^*_{>})$ is given by

$$v_1(\bar{z}^*_{>}) = \bar{z}^{*2}_{>} - \frac{\sum_{i \neq i'=1}^{n} \bar{z}_i \bar{z}_{i'}}{n(n-1)} . \quad \ldots \quad (6.3.7)$$

It can be verified on similar lines that $\bar{z}^*_{>}$ is also uniformly better than $\bar{z}_n$ in a different two-stage sampling scheme, where the first-stage units are drawn with unequal probabilities and the second-stage units are drawn by circular-systematic sampling.
This sampling scheme is in current use in the National Sample Survey of India.

An estimator better than $\bar{z}^*_{>}$.

The procedure leading to an estimator better than $\bar{z}^*_{>}$ is same as that given in the previous section. We state below only the final result.

Theorem 4:

An estimator uniformly better than $\bar{z}^*_{>}$ is given by

$$\bar{z}_{>} = \Sigma c_{(i)} \bar{z}_{>(i)} , \quad \ldots \quad (6.3.8)$$

where

$$c_{(i)} = p_{(i)} \left\{ \sum_{r=0}^{[\sum \nu_{(i)}-1]} (-)^r \Sigma_1 \binom{\nu_{(1)}}{\alpha_1} \cdots \binom{\nu_{(\nu)}}{\alpha_\nu} \frac{\binom{\nu_{(i)}-\alpha_i}{m_{(i)}}}{\binom{M_{(i)}}{m_{(i)}}} \left[p_{(1)} \frac{\binom{\nu_{(1)}-\alpha_1}{m_{(1)}}}{\binom{M_{(1)}}{m_{(1)}}} + \cdots \right.\right.$$

$$\left.\left. \cdots + p_{(\nu)} \frac{\binom{\nu_{(\nu)}-\alpha_\nu}{m_{(\nu)}}}{\binom{M_{(\nu)}}{m_{(\nu)}}} \right]^{n-1} \right\} \Big/ P(T),$$

and

$$P(T) = \sum_{r=0}^{[\sum \nu_{(i)}-1]} (-)^r \Sigma_1 \binom{\nu_{(1)}}{\alpha_1} \cdots \binom{\nu_{(\nu)}}{\alpha_\nu} \left[p_{(1)} \frac{\binom{\nu_{(1)}-\alpha_1}{m_{(1)}}}{\binom{M_{(1)}}{m_{(1)}}} + \cdots \right.$$

$$\left. \cdots + p_{(\nu)} \frac{\binom{\nu_{(\nu)}-\alpha_\nu}{m_{(\nu)}}}{\binom{M_{(\nu)}}{m_{(\nu)}}} \right]^n ;$$

the summation Σ_1 stands for all non-negative integral α_1's such that $\alpha_1 + \alpha_2 + \cdots + \alpha_\nu = r$.

However, in practical situations, where simplicity of the estimator is the main criterion, the estimator $\bar{z}_\nu$ is not useful. In such cases, we recommend the use of $\bar{z}^*_\nu$ as an estimator of Y, and

$$\frac{1}{n(n-1)} \Sigma (\bar{z}_i - \bar{z}_n)^2$$

as an estimator of $V(\bar{z}^*_\nu)$.

The extension of the method of improving estimators in multi-stage sampling schemes can be given on similar lines. From the point of view of brevity, we think it unnecessary to consider these sampling schemes. It may be remarked that useful improved estimators of Y in these sampling schemes will be essentially of the form of the estimator $\bar{z}^*$. [By useful improved estimators, we mean improved estimators that are easy to compute in practice.]

CHAPTER VII

SAMPLING SCHEMES PROVIDING UNBIASED RATIO ESTIMATORS

Summary.

The problem of finding unbiased ratio estimator of the population total of some character with the help of an auxiliary character, has drawn much attention in recent years. Some references to this are given in the bibliography. Manjunma, Murthy and Sethi (33) have given unbiased ratio estimators under different sampling schemes. These schemes have been obtained by simple modifications of the commonly adopted sampling schemes. For some such sampling schemes, we have derived ratio estimators which are more efficient than those given by Manjunma, Murthy and Sethi. The method of improving the ratio estimators for other sampling schemes of the above type, where samples are drawn with replacement at some stage of sampling, is analogous to that given in this chapter, and is essentially based on the Rao-Blackwell theorem.

7.1. Introduction.

For the sake of simplicity of exposition, we shall follow the notations already introduced in preceding chapters. Further, the letter W will stand for some real valued auxiliary characteristic related to Y characteristic of a population. We assume that the value of the

W characteristic of every population unit is known in advance and is greater than zero.

Instead of giving improved ratio estimators of the population total of Y characteristic, we give unbiased ratio estimators of the ratio of the population totals of Y and W characteristics. Improved unbiased ratio estimators of the population total of Y- characteristic can be obtained by multiplying them by the population total of W-characteristic.

7.2. Sampling with unequal probabilities.

The modification of sampling with unequal probabilities which provides unbiased ratio estimators, is as follows:

1) Draw one unit with ppw and replace it.[1]

2) Draw the remaining (n-1) sample units from the whole population in the usual manner, i.e., with unequal probabilities (with replacement), P_j being the probability of selection associated with the j-th population unit (j = 1,2,...,N).

Let us now record the observed sample as

$$S = [\ (x_{(1)},\ \lambda_{(1)}),\ \ldots,\ (x_{(\nu)},\ \lambda_{(\nu)})],$$

where $x_{(i)} = [\ y_{(i)},\ P_{(i)},\ u_{(i)},\ w_{(i)}]$ is the i-th order statistic and $\lambda_{(i)}$ is the number of times $x_{(i)}$ is included in the sample.

---

1. The symbol ppw is an abbreviation for 'probabilities proportional to W'.

The probability of getting a particular sample S is given by

$$P(S) = \frac{n! \prod_{i=1}^{\nu} p_{(i)}^{\lambda_{(i)}}}{\prod_{i=1}^{\nu} \lambda_{(i)}!} \frac{1}{W} \left(\frac{1}{n} \sum \lambda_{(i)} \frac{w_{(i)}}{p_{(i)}} \right), \qquad (7.2.1)$$

where $W = \sum W_j$.

In this sampling scheme, an unbiased estimator of the ratio

$$R = \frac{\sum Y_j}{\sum W_j} = \frac{Y}{W}$$

is given by

$$\hat{R} = \frac{\frac{1}{n} \sum \lambda_{(i)} \frac{y_{(i)}}{p_{(i)}}}{\frac{1}{n} \sum \lambda_{(i)} \frac{w_{(i)}}{p_{(i)}}} \quad \ldots \qquad (7.2.2)$$

To get an estimator uniformly better than $\hat{R}$, let us record the 'order-statistic'

$$T = [x_{(1)}, x_{(2)}, \ldots, x_{(\nu)}]. \qquad (7.2.3)$$

Now, as T is a sufficient statistic, we have:

Theorem 1:

For any convex loss function, an estimator uniformly better than

$\hat{R}$ is given by

$$\hat{R}_{\nu} = \frac{\Sigma\, e_{(i)}\, \frac{y_{(i)}}{p_{(i)}}}{\Sigma\, e_{(i)}\, \frac{w_{(i)}}{p_{(i)}}}, \qquad \ldots \qquad \ldots \qquad (7.2.4)$$

where $e_{(i)}$ is given by (4.2.3).

Proof:

Obviously, an estimator uniformly better than $\hat{R}$ is given by

$$E(\hat{R} \mid T) = \frac{\Sigma^{*}\, \hat{R}\, P(S)}{\Sigma^{*}\, P(S)}, \qquad \ldots \qquad \ldots \qquad (7.2.5)$$

where the summation Σ^{*} is taken over all samples giving rise to the 'order-statistic' T.

On putting the value of P(S) and simplifying, we can write (7.2.5) as

$$E(\hat{R} \mid T) = \frac{\Sigma^{*} \frac{n!}{\prod_{i=1}^{\nu} \lambda_{(i)}!} \prod_{i=1}^{\nu} p_{(i)}^{\lambda_{(i)}} \left(\frac{1}{n} \Sigma\, \lambda_{(i)} \frac{y_{(i)}}{p_{(i)}} \right) \Big/ \Sigma^{*} \frac{n!}{\prod_{i=1}^{\nu} \lambda_{(i)}!} \prod_{i=1}^{\nu} p_{(i)}^{\lambda_{(i)}}}{\Sigma^{*} \frac{n!}{\prod_{i=1}^{\nu} \lambda_{(i)}!} \prod_{i=1}^{\nu} p_{(i)}^{\lambda_{(i)}} \left(\frac{1}{n} \Sigma\, \lambda_{(i)} \frac{w_{(i)}}{p_{(i)}} \right) \Big/ \Sigma^{*} \frac{n!}{\prod_{i=1}^{\nu} \lambda_{(i)}!} \prod_{i=1}^{\nu} p_{(i)}^{\lambda_{(i)}}}$$

$$\ldots \qquad \ldots \qquad (7.2.6)$$

It at once follows from Theorem 1 of Chapter IV that the numerator of (7.2.6) is given by $\Sigma c_{(i)} \frac{y_{(i)}}{p_{(i)}}$ and the denominator, by $\Sigma c_{(i)} \frac{w_{(i)}}{p_{(i)}}$.

$$\therefore E[\hat{R} \mid T] = \frac{\Sigma c_{(i)} \frac{y_{(i)}}{p_{(i)}}}{\Sigma c_{(i)} \frac{w_{(i)}}{p_{(i)}}}.$$

Hence the theorem is proved.

Estimation of R^2.

For estimating $V(\hat{R}_{\cdot})$, we require unbiased estimators of R^2. Manjamma, Murthy and Sethi gave the following estimator of R^2:

$$\hat{R}^2 = \frac{\sum_{i=1}^{\nu} \lambda_{(i)} (\lambda_{(i)} - 1) \frac{y_{(i)}^2}{p_{(i)}^2} + \sum_{i \neq i'=1}^{\nu} \lambda_{(i)} \lambda_{(i')} \frac{y_{(i)}}{p_{(i)}} \frac{y_{(i')}}{p_{(i')}}}{n(n-1) \; w \left(\frac{1}{n} \Sigma \lambda_{(i)} \frac{w_{(i)}}{p_{(i)}} \right)}. \quad (7.2.7)$$

With the help of Theorem 2 of Chapter IV, it can be proved similarly that

Theorem 2:

For any convex loss function, an estimator uniformly better than $\hat{R}^2$ is given by

$$\hat{R}^2_{\nu} = \frac{\sum_{i=1}^{\nu} e_{(i,i)} \frac{y^2_{(i)}}{p^2_{(i)}} + \sum_{i \neq i'=1}^{\nu} e_{(i,i')} \frac{y_{(i)}}{p_{(i)}} \frac{y_{(i')}}{p_{(i')}}}{\left(\sum_{i=1}^{\nu} e_{(i)} \frac{y_{(i)}}{p_{(i)}} \right)^2}, \qquad (7.2.8)$$

where $e_{(i,i)}$ and $e_{(i,i')}$ have been defined in (4.3.3).

Corollary 1:

An unbiased estimator of $V(\hat{R}_{\nu})$, uniformly better than

$$v_1(\hat{R}_{\nu}) = \hat{R}^2_{\nu} - \hat{R}^2,$$

is given by

$$v_2(\hat{R}_{\nu}) = \hat{R}^2_{\nu} - \hat{R}^2_{\nu} . \qquad \ldots \qquad (7.2.9)$$

Simple improved estimators of R and R^2.

As indicated in Chapter IV, the type of estimators derived above are tedious to compute in practice. In terms of the notations of Section 4.4, we give below estimators of R and R^2 which though less efficient than the above derived estimators, are much simpler to compute in practice.

Theorem 3:

For any convex loss function, another estimator uniformly better than $\hat{R}$ is given by

$$\hat{R}^*_{\nu} = \frac{\sum_{i=1}^{k} \frac{n_{(i)}}{p_{(i)}} \bar{y}_{\nu(i)}}{\sum_{i=1}^{k} \frac{n_{(i)}}{p_{(i)}} \bar{w}_{\nu(i)}} . \qquad (7.2.10)$$

Proof:

The proof is an easy consequence of Theorem 3 of Chapter IV and of the fact that

$$E(\hat{R} \mid T^*) = \frac{\sum_{S \supset T^*} \hat{R}\, P(S)}{\sum_{S \supset T^*} P(S)}, \qquad \ldots \qquad (7.2.11)$$

where T^* is the similar statistic as defined by (4.4.1), and the summation runs over all samples giving rise to T^*.

After substituting the values of $\hat{R}$ and $P(S)$ in (7.2.11) and simplifying, the theorem follows at once with the help of Theorem 3 of Chapter IV.

Using Theorem 4 of Chapter IV, we can prove similarly the following:

Theorem 4:

For any convex loss function, another estimator uniformly better than $\hat{R}^2$ is given by

$$\hat{R}^{2*}_{\nu} = \frac{1}{n(n-1)\, w\left(\sum_{i=1}^{k} \frac{n_{(i)}}{p_{(i)}}\, \bar{w}_{\nu(i)}\right)} \left[\left\{ \left(\sum_{i=1}^{k} n_{(i)}\, \frac{\bar{y}_{\nu(i)}}{p_{(i)}} \right)^2 - \sum_{i=1}^{k} n_{(i)}\, \frac{\bar{y}^2_{\nu(i)}}{p^2_{(i)}} \right\} - \sum_{i=1}^{k} \frac{n_{(i)}(n_{(i)}-1)\, e_{\nu(i)-1}\, (n_{(i)}-1)\, s^2_{\nu(i)}}{e_{\nu(i)}\, (n_{(i)})\, p^2_{(i)}} \right].$$

$$\ldots \qquad (7.2.12)$$

This theorem will be found useful for deriving unbiased estimators of $V(\hat{R}^*)$ and $V(\hat{R})$.

7.3. Two-stage sampling.

Let us now turn to the problem of deriving improved ratio estimators in case of two-stage sampling. We consider only the modification of the two-stage sampling scheme discussed in Section 6.3, where the first-stage units are drawn with unequal probabilities (with replacement) and the second-stage units, with equal probabilities (without replacement). Similar procedure can be followed for other two-stage sampling schemes. The modification of the above scheme is as follows:

1) Draw one second-stage unit from the whole population of second-stage units with ppw[1], say X_{jh}, and then select (m_j-1) second-stage units from the remaining (M_j-1) second-stage units of the j-th first-stage unit by simple random sampling without replacement.

2) Draw the remaining (n-1) first-stage sample units and their sub-samples in the usual manner (i.e, unequal probabilities for first-stage and equal probabilities (without replacement) for second-stage).

---

1. This can also be achieved in a different manner by first selecting a first-stage unit with probabilities proportional to the total w-characteristics in the first-stage units and then selecting a second-stage unit with ppw from the selected first-stage unit.

Using the notations of Chapter VI, we record the observed sample:

$$S = [\{x_{11}, x_{12}, \ldots, x_{1m_{u_1}}\}, \ldots, \{x_{n1}, x_{n2}, \ldots, x_{nm_{u_n}}\}].$$

The probability of getting a particular sample S is given by

$$P(S) = \frac{v_{11}}{W} \cdot P'(S) \quad , \quad \ldots \qquad (7.3.1)$$

where $W = \sum W_{jh}$ and v_{11} is the v-value of x_{11} [1]. And $P'(S)$ is the probability of getting the above sample under the usual two-stage sampling scheme considered in Section 6.3.

In this modified sampling scheme, an unbiased estimator of the ratio

$$R = \frac{\sum Y_{jh}}{\sum W_{jh}} = \frac{Y}{W}$$

is given by

$$\hat{R} = \frac{\frac{1}{n}\sum_{i=1}^{n} \bar{z}_i}{\frac{1}{n}\sum_{i=1}^{n} \bar{v}_i} \quad , \quad \ldots \qquad (7.3.2)$$

---

1. We define by

$$v_{jh} = M_j \frac{W_{jh}}{P_j},$$

the v-value of the h-th second-stage unit of X_j analogous to the z-value as defined by (6.2.1).

where $\bar{z}_i$ and $\bar{v}_i$ are averages of the z-values and the v-values of $x_{i1}, x_{i2}, \ldots, x_{im_{ui}}$ respectively.

Now, if by

$$T^* = \left[\left\{ x_{(i)}, \lambda_{(i)}, x_{(i1)}, \ldots, x_{(i\nu_{(i)})} \right\} i = 1,2,\ldots, \nu \right],$$

we denote the similar sufficient statistic as defined in (6.2.5), we have:

Theorem 5:

For any convex loss function, an estimator uniformly better than $\hat{R}$ is given by

$$\hat{R}^*_{\nu} = \frac{\sum \lambda_{(i)} \bar{z}_{\nu(i)}}{\sum \lambda_{(i)} \bar{v}_{\nu(i)}} \quad , \quad \ldots \qquad (7.3.3)$$

where $\bar{z}_{\nu(i)} = \frac{1}{\nu_{(i)}} \sum_{(i\,r)} z_{(ir)}$ and $\bar{v}_{\nu(i)} = \frac{1}{\nu_{(i)}} \sum_{(ir)} v_{(ir)}$,

$z_{(ir)}$ and $v_{(ir)}$ are the z-value and the v-value of $x_{(ir)}$ respectively.

Proof:

An estimator uniformly better than $\hat{R}$ is given by

$$E\left(\hat{R} \mid T^*\right) = \frac{\sum_{S \supset T^*} \hat{R}\, P(S)}{\sum_{S \supset T^*} P(S)}, \quad \ldots \quad (7.3.4)$$

where the summation runs over all samples giving rise to T^*.

Some consideration on the lines of Theorem 3 of Chapter VI will show that

$$\sum_{S \supset T^*} P(S) = \left[\frac{\sum \lambda_{(i)}\, \bar{v}_{\gamma(i)}}{n\,\bar{W}}\right]\left[\sum_{S \supset T^*} P'(S)\right],$$

and

$$\sum_{S \supset T^*} \hat{R}\, P(S) = \left[\frac{\sum \lambda_{(i)}\, \bar{s}_{\gamma(i)}}{n\,\bar{W}}\right]\left[\sum_{S \supset T^*} P'(S)\right]. \quad \ldots$$

$$\ldots \quad (7.3.5)$$

The theorem follows by combining (7.3.4) and (7.3.5).

Further, by proceeding on the lines of Theorem 4 of Chapter VI, it can be proved that an estimator still better than $\hat{R}^*_{\gamma}$ is given by

$$\hat{R}_{\gamma} = \frac{\sum c_{(i)}\, \bar{s}_{\gamma(i)}}{\sum c_{(i)}\, \bar{v}_{\gamma(i)}}, \quad \ldots \quad (7.3.6)$$

where $c_{(i)}$ has the same meaning as in (6.3.8).

The extension of this technique to any general multi-stage sampling scheme can be given on similar lines. We conclude this chapter with the following obvious extension to stratified sampling.

7.4. Stratified sampling.

Here, we shall consider the modification of stratified sampling with unequal probabilities with replacement.

Let k be the number of strata, N_ℓ and n_ℓ be the number of units in the population and the sample respectively for the ℓ-th stratum ($\ell = 1, \ldots, k$). Let $P_{\ell j}$ be the probability of selection associated with the j-th unit of the ℓ-th stratum ($\sum_{j=1}^{N_\ell} P_{\ell j} = 1$, $\ell = 1, \ldots, k$), the corresponding values of W and Y characteristics being $W_{\ell j}$ and $Y_{\ell j}$ respectively.

In stratified sampling with unequal probabilities, n_ℓ units are drawn independently from the ℓ-th stratum with unequal probabilities ($\ell = 1, \ldots, k$). The modified sampling scheme which provides an unbiased ratio estimator is as follows:

1) Draw one unit, say the j-th unit in the ℓ-th stratum, from the whole population with ppw and replace it.
2) Draw the remaining $(n_\ell - 1)$ units from the ℓ-th stratum, and $n_{\ell'}$ units from ℓ'-th stratum ($\ell' \neq \ell = 1,2,\ldots,k$) in the usual way, i.e., with stratified sampling with unequal probabilities.

In this case, a particular sample may be recorded as

$$S = [\{(x_{(\ell 1)}, \lambda_{(\ell 1)}), \ldots, (x_{(\ell \nu_\ell)}, \lambda_{(\ell \nu_\ell)})\} \quad \ell = 1, \ldots, k],$$

where $x_{(\ell i)}$ is the i-th order-statistic of the sample selected from the ℓ-th stratum, and $\lambda_{(\ell i)}$ is the number of times it is included in the sample ($\ell = 1, \ldots, k$; $i = 1, \ldots, \nu_\ell$).

The probability of getting such a sample is given by

$$P(S) = \frac{\sum_{\ell=1}^{k} \frac{1}{n_\ell} \sum_{i=1}^{\nu_\ell} \lambda_{(\ell i)} \frac{w_{(\ell i)}}{p_{(\ell i)}}}{W} P'(S), \quad \ldots \quad (7.4.1)$$

where $W = \sum_{\ell=1}^{k} \sum_{j=1}^{N_\ell} w_{\ell j}$ is the population total of the auxiliary characteristic, and $P'(S)$ is the probability of getting the above sample under stratified sampling with unequal probabilities (without any modification).

An unbiased estimator of the ratio

$$R = \frac{\sum_{\ell=1}^{k} \sum_{j=1}^{N_\ell} Y_{\ell j}}{\sum_{\ell=1}^{k} \sum_{j=1}^{N_\ell} w_{\ell j}} = \frac{Y}{W}$$

is given by

$$\hat{R} = \frac{\sum_{\ell=1}^{k} \frac{1}{n_\ell} \sum_{i=1}^{\nu_\ell} \lambda_{(\ell i)} \frac{y_{(\ell i)}}{p_{(\ell i)}}}{\sum_{\ell=1}^{k} \frac{1}{n_\ell} \sum_{i=1}^{\nu_\ell} \lambda_{(\ell i)} \frac{w_{(\ell i)}}{p_{(\ell i)}}} . \quad \ldots \quad (7.4.2)$$

It may be noted here that the estimators $\frac{1}{n_\ell} \sum_{i=1}^{\nu_\ell} \lambda_{(\ell i)} \frac{y_{(\ell i)}}{p_{(\ell i)}}$ and $\frac{1}{n_\ell} \sum_{i=1}^{\nu_\ell} \lambda_{(\ell i)} \frac{w_{(\ell i)}}{p_{(\ell i)}}$ are unbiased estimators of the population totals of Y and W characteristics of the ℓ-th stratum respectively under stratified sampling with unequal probabilities, and are of the same form as given by (4.2.1).

We state without proof that better estimators than $\hat{R}$ are obtained by replacing $\frac{1}{n_\ell} \sum_{i=1}^{\nu_\ell} \lambda_{(\ell i)} \frac{y_{(\ell i)}}{p_{(\ell i)}}$ and $\frac{1}{n_\ell} \sum_{i=1}^{\nu_\ell} \lambda_{(\ell i)} \frac{w_{(\ell i)}}{p_{(\ell i)}}$ by estimators of the forms (4.2.2) and (4.4.2) as done in Chapter IV.

CHAPTER VIII

A GENERAL SAMPLING SCHEME AND ITS APPLICATIONS

8.1. Introduction and summary.

The problem of estimating the total size of a population is known to be of great importance in biological and other related problems, e.g., one may be interested to find out the total number of fish in a lake, or to find out the total number of cycles operating in a city etc.. Several authors have devised methods of sampling such populations references to which are given in the bibliography. In this chapter, we consider simple random sampling (without replacement) at several stages for this purpose. As it has been mentioned by Bailey (1) that in certain ecological problems we may be more concerned to use the reciprocal of the population size rather than the use of the population size itself, the problem of estimating the reciprocal of the population size is also considered here. In addition, the problem of estimating the population mean of some characteristic (say fish weight) and of the ratio of the population means of two characteristics are also considered.

Consider a population of N units. Let Y_j and W_j be the values of Y - and W - characteristics associated with the j-th population unit. (W is an auxiliary characteristic related to Y - characteristic, and j varies from 1 to N.) We shall begin with the problem of finding unbiased estimators of

$$\bar{Y} = \frac{1}{N}\sum_{j=1}^{N} Y_j \qquad \text{and} \qquad \bar{\bar{Y}} = \frac{\sum Y_j}{\sum W_j} . \quad \ldots \quad \ldots \quad (8.1.1)$$

The sampling scheme to be considered is as follows:

k simple random (without replacement) samples of sizes $n_1, n_2, \ldots, n_k$, are drawn independently of each other, i.e., each sample is drawn by simple random sampling (without replacement) and is replaced to the population for subsequent selection of samples.

Considering first the problem of estimating $\bar{Y}$, we see that the usual estimator of $\bar{Y}$ based on the i-th sample is given by the sample mean

$$\bar{y}_i = \frac{1}{n_i}\sum y \quad , \qquad \ldots \qquad (8.1.2)$$

where the summation is taken over all units of the ith sample. Obviously, any linear function $\sum_{i=1}^{k} c_i \bar{y}_i$ is also an unbiased estimator of $\bar{Y}$ provided $\sum_{i=1}^{k} c_i = 1$.

Suppose that along with recording the values of Y - and W - characteristics, we also record the unit-indices of the sample units. Let $x_{(1)}, x_{(2)}, \ldots, x_{(m)}$ be the m distinct units observed in the sample, then it is easy to see that the statistic

$$T = [\, x_{(1)}, x_{(2)}, \ldots, x_{(m)} \,] \qquad \ldots \qquad (8.1.3)$$

is sufficient.

Further, it can be verified that the probability of getting any preassigned m distinct units, in this sampling scheme, is given by

$$P_1 = \frac{\prod_{i=1}^{k}\binom{m}{n_i} - \binom{m}{1}\prod_{i=1}^{k}\binom{m-1}{n_i} + \ldots(-)^{m-\max n_i}\binom{m}{m-\max n_i}\prod_{i=1}^{k}\binom{\max n}{n_i}}{\prod_{i=1}^{k}\binom{N}{n_i}} \qquad \ldots \quad (8.1.4)$$

Also the probability that at least $(m - n_1)$ specified units out of m given units will be drawn in the last $(k - 1)$ samples, is given by [1]

$$P_2 = \frac{\prod_{i=2}^{k}\binom{m}{n_i} - \binom{m-n_1}{1}\prod_{i=2}^{k}\binom{m-1}{n_i} + \ldots\binom{m-n_1}{2}\prod_{i=2}^{k}\binom{m-2}{n_i} - \ldots}{\prod_{i=2}^{k}\binom{N}{n_i}} \qquad \ldots \quad (8.1.5)$$

Clearly, for any convex loss function, an estimator uniformly better than $\bar{y}_1$ is given by $E(\bar{y}_1 \mid T)$. Using (8.1.4), (8.1.5) and proceeding on the lines of Theorem 3 of Chapter VI, we can prove that

$$E(\bar{y}_1 \mid T) = \bar{y}_m, \qquad \ldots \qquad (8.1.6)$$

where $\bar{y}_m$ denotes the average of the m distinct units observed.

---

1. $\binom{m}{r}$ is to regarded as zero if $r > m$, or if m is negative.

In a similar manner, we can prove that

$$E(\bar{y}_i \mid T) = \bar{y}_m .$$

Thus $\bar{y}_m$ is uniformly better than any linear function $\sum c_i \bar{y}_i$ $(\sum c_i = 1)$ unbiased for $\bar{Y}$.

Variance of $\bar{y}_m$.

Obviously

$$V(\bar{y}_m) = E[V(\bar{y}_m \mid m)] + V[E(\bar{y}_m \mid m)]$$

$$= E[V(\bar{y}_m \mid m)] = E\left[\left(\frac{1}{m} - \frac{1}{N}\right) S^2\right], \qquad .. \quad (8.1.7)$$

where $S^2 = \frac{1}{(N-1)} \sum (Y_j - \bar{Y})^2$, and it follows from (2.2.8) that

$$E\left(\frac{1}{m}\right) = \frac{1}{\prod_{i=1}^{k} \binom{N}{n_i}} \left[\frac{\prod_{i=1}^{k} \binom{N}{n_i}}{N} + \frac{\prod_{i=1}^{k} \binom{N-1}{n_i}}{(N-1)} + \ldots + \frac{\prod_{i=1}^{k} \binom{\max n_i}{n_i}}{\max n_i} \right].$$

Estimation of $\bar{Y}^2$.

For getting an unbiased estimator of $V(\bar{y}_m)$, the following unbiased estimator of $\bar{Y}^2$:

$$\hat{\bar{Y}}_1^2 = \frac{\sum_{i\neq i'=1}^{k} n_i\, n_{i'}\, \bar{y}_i\, \bar{y}_{i'}}{\sum_{i\neq i'=1}^{k} n_i\, n_{i'}} \quad \ldots \quad \ldots \quad (8.1.8)$$

may be used.

It can be shown that an estimator uniformly better than $\hat{\bar{Y}}_1^2$ is given by

$$\hat{\bar{Y}}_2^2 = \sum_{i=1}^{m} y_{(i)}^2 \frac{\left\{\frac{1}{m^2}\prod_{i=1}^{k}\binom{m}{n_i} - \frac{\binom{m-1}{1}}{(m-1)^2}\prod_{i=1}^{k}\binom{m-1}{n_i} + \ldots\right\}}{\left\{\prod_{i=1}^{k}\binom{m}{n_i} - \binom{m}{1}\prod_{i=1}^{k}\binom{m-1}{n_i} + \ldots\right\}}$$

$$+ \sum_{i\neq i'=1} y_{(i)} y_{(i')} \frac{\left\{\frac{1}{m^2}\prod_{i=1}^{k}\binom{m}{n_i} - \frac{\binom{m-2}{1}}{(m-1)^2}\prod_{i=1}^{k}\binom{m-1}{n_i} + \ldots\right\}}{\left\{\prod_{i=1}^{k}\binom{m}{n_i} - \binom{m}{1}\prod_{i=1}^{k}\binom{m-1}{n_i} + \ldots\right\}}$$

$$\ldots \quad (8.1.9)$$

As the expression (8.1.9) is unwieldy for the purpose of computation, the expression (8.1.8) seems preferable to (8.1.9) in actual practice.

Application to the estimation of average fish weight.

In order to estimate the average fish weight of fish in a lake, the following procedure may be adopted:

We catch fish k times, each time after catch every fish is weighed, its weight noted, marked with a black spot and thrown back in the lake. In this way, the total number of distinct fish observed is equal to the number of unspotted fish caught; these distinct fish form a sufficient statistic. The average weight of the unspotted fish will be the required estimator of the average fish weight.

8.2. Estimation of the number of fish in a lake.

In this section, we shall take up the problem of estimating the total number and the reciprocal of the total number of fish in a lake. To render our exposition clear, we shall consider the particular case of the general sampling scheme given in Section 8.1 when $n_i = 1$ (i = 1, ..., k). This scheme is known as direct sampling.

8.2A. Direct sampling.

In this sampling scheme, if N is the unknown number of fish in a lake, the probability of getting any m distinct fish is given by

$$P_{(m)} = \frac{\binom{N}{m} c_k(m)}{N^k} \qquad \ldots \qquad \ldots \qquad (8.2.1)$$

It is not difficult to show that complicated but unique estimator of $\frac{1}{N}$ is given by

$$t_{-1}(m) = \frac{c_m(k-1)}{c_m(k)}, \quad \ldots \quad \ldots \quad (8.2.2)$$

However, the search for an estimator of N leads to the following theorem:

Theorem 1:

A unique unbiased estimator of N exists if and only if the number of fish caught is not less than the number of fish in the lake (i.e., $k \geq N$), in that case, the required estimator is given by

$$t_1(m) = \frac{c_m(k+1)}{c_m(k)}, \quad \ldots \quad (8.2.3)$$

otherwise, no unbiased estimator exists.

Proof:

Suppose that there exists an unbiased estimator of N. Let it be $t_1(m)$, then we have from the condition of unbiasedness

$$\sum_{m=1}^{\min(k,N)} t_1(m) \frac{c_m(k)\binom{N}{m}}{N^k} = N \text{ for all } N \geq 1. \quad (8.2.4)$$

Putting successively $N = 1, 2, \ldots$, we get the only possible estimator:

$$t_1(m) = \frac{c_m(k+1)}{c_m(k)}, \quad \ldots \quad \ldots \quad (8.2.5)$$

this is unbiased for N if $k \geq N$, otherwise

$$E(t_1(m)) = \left[N - \frac{(k+1)\, !\, \binom{N}{k+1}}{N^k} \right]. \qquad \ldots \qquad (8.2.6)$$

The bias of $t_1(m)$ decreases as k increases and would be negligible if k is large enough. Moreover, if some crude approximation for N is available in advance, a correction for the bias can be made.

An estimator of $V(t_1(m))$ (unbiased if $k \geq N$) is given by

$$v_1 \left[t_1(m) \right] = t_1^2(m) - \frac{c_m (k+2)}{c_m (k)}. \qquad \ldots \qquad (8.2.7)$$

8.2B. General sampling scheme.

Proceeding on the above lines, it can be shown that in the sampling scheme given in Section 8.1, an unbiased estimator of $\frac{1}{N}$ is given by

$$t_{-1}(m) = \frac{\left[\frac{1}{m} \prod_{i=1}^{k} \binom{m}{n_i} - \frac{\binom{m}{1}}{(m-1)} \prod_{i=1}^{k} \binom{m-1}{n_i} + \frac{\binom{m}{2}}{(m-2)} \prod_{i=1}^{k} \binom{m-2}{n_i} - \ldots \right]}{\left[\prod_{i=1}^{k} \binom{m}{n_i} - \binom{m}{1} \prod_{i=1}^{k} \binom{m-1}{n_i} + \binom{m}{2} \prod_{i=1}^{k} \binom{m-2}{n_i} - \ldots \right]}$$

$$\ldots \qquad (8.2.8)$$

and an estimator of N (unbiased if $\sum_{i=1}^{k} n_i \geq N$) is given by

$$t_1(m) = \frac{\left[m \prod_{i=1}^{k} \binom{m}{n_i} - (m-1)\binom{m}{1} \prod_{i=1}^{k} \binom{m-1}{n_i} + \cdots \right]}{\left[\prod_{i=1}^{k} \binom{m}{n_i} - \binom{m}{1} \prod_{i=1}^{k} \binom{m-1}{n_i} + \cdots \right]} \quad \ldots (8.2.9)$$

An estimator of $V(t_1(m))$ (unbiased if $\sum_{i=1}^{k} n_i \geq N$) is given by

$$v(t_1(m)) = t_1^2(m) - \frac{\left[m^2 \prod_{i=1}^{k} \binom{m}{n_i} - (m-1)^2 \binom{m}{1} \prod_{i=1}^{k} \binom{m-1}{n_i} + \cdots \right]}{\left[\prod_{i=1}^{k} \binom{m}{n_i} - \binom{m}{1} \prod_{i=1}^{k} \binom{m-1}{n_i} + \cdots \right]}$$

$\ldots \quad \ldots \qquad (8.2.10)$

If the number of fish caught on successive occassions is small as compared to the total number of fish, these expressions can be approximated by

$$t_{-1}(m) \doteq \frac{e_m(\Sigma n_i - 1)}{e_m(\Sigma n_i)},$$

$$t_1(m) \doteq \frac{e_m(\Sigma n_i + 1)}{e_m(\Sigma n_i)}. \qquad \ldots \qquad (8.2.11)$$

The exact values of these terms can be got from Table 3.3 upto $\Sigma n_i = 50$.

When n and m are large and are such that $m e^{-\frac{n}{m}} = \lambda$ is of moderate value (say ≤ 5) (Feller page 93), the approximate values of the above expressions can be computed by the relation

$$e_m(n) = m^n e^{-\lambda} . \qquad \ldots \qquad (8.2.12)$$

The estimators derived above are expected to fare better than the usually adopted estimators as they are functions of the sufficient statistic 'm'. However, their practical application is restricted only to cases where the sample size is large enough so that their bias may be negligible. Another difficulty about their use is the difficulty of computing $e_m(n)$; this difficulty can be overcome if the Table 3.3 is extended for sufficiently large values of n. In the next section, where we consider the problem of estimating $\bar{Y}/\bar{W}$, it will be shown that a simple modification of the above sampling scheme will provide simpler estimator of N, though not exactly unbiased. The bias of this estimator, in the particular case of direct sampling, is equal to the bias of Bailey's estimator. But for any convex loss function, our estimator is uniformly better than Bailey's estimator.

8.3. Estimation of $\bar{Y}/\bar{W}$.

The importance of this problem arises, as for example, in estimating the proportions of persons in a city using a particular brand of cycle, or in estimating the ratio of the average fish weight to the average fish length of fish in a lake etc.. The method to

derive an unbiased estimator of $\bar{Y}/\bar{W}$, based on the lines of Nanjamma, Murthy and Sethi (33), is as follows:

1) Draw one unit from the whole population with ppw. ($W_j > 0$ for all j.)

2) Draw $(n_1 - 1)$ units with equal probabilities (without replacement) from the remaining $(N-1)$ units.

3) Draw the remaining $(k - 1)$ samples in the usual way, i.e., by independent simple random sampling (without replacement).

In this sampling scheme, the probability of getting a particular set of k samples, on deleting the information which unit was selected first, is given by

$$P(S) = \frac{\bar{w}_1}{\bar{W}} \prod_{i=1}^{k} \frac{1}{\binom{N}{n_i}} . \qquad \ldots \qquad (8.3.1)$$

Thus an unbiased estimator of $\frac{\bar{Y}}{\bar{W}}$ is given by

$$\hat{R} = \frac{\sum_{i=1}^{k} c_i \bar{y}_i}{\bar{w}_1}, \qquad \ldots \quad (8.3.2)$$

where $c_1, c_2, \ldots, c_k$ are such that $\sum_{i=1}^{k} c_i = 1$.

Similar to (8.1.3), the statistic

$$T = [\, x_{(1)}, \ldots, x_{(m)}] \qquad \ldots \qquad (8.3.3)$$

is sufficient, and the probability of getting samples with a given T is

$$P(T) = \sum_{S \supset T} P(S) = \frac{\bar{w}_m}{\bar{W}} P_1, \ \ldots \qquad (8.3.4)$$

where P_1 is given by (8.1.4) and $\bar{w}_m = \frac{1}{m} \sum_{i=1}^{m} w_{(i)}$.

Proceeding similarly, it can be proved that an estimator uniformly better than $\hat{R}$ is

$$E(\hat{R} \mid T) = \frac{\sum_{S \supset T} \hat{R} P(S)}{\sum_{S \supset T} P(S)} = \frac{\bar{y}_m}{\bar{w}_m}, \qquad \ldots \qquad (8.3.5)$$

since $\sum_{S \supset T} \hat{R} P(S) = \sum_{S \supset T} \frac{(\sum_{i=1}^{k} c_i \bar{y}_i)}{\bar{W}} \prod_{i=1}^{k} \frac{1}{\binom{N}{n_i}} = \frac{\bar{y}_m}{\bar{W}} P_1$.

If we know $\bar{W}$, an unbiased estimator of $V\left(\frac{\bar{y}_m}{\bar{w}_m}\right)$ is given by

$$v\left(\frac{\bar{y}_m}{\bar{w}_m}\right) = \frac{\bar{y}_m^2}{\bar{w}_m^2} - \frac{\hat{\bar{Y}}_2^2}{\bar{w}_m \bar{W}}, \qquad \ldots \qquad \ldots \qquad (8.3.6)$$

where $\hat{\bar{Y}}_2^2$ is given by (8.1.9).

Application to the estimation of fish population.

We now apply this technique to derive an estimator for the total number of fish in a lake. The procedure is as follows:

1) First, a large number of fish are caught and are marked with a red spot, one red fish is then taken out and the remaining are thrown back.

2) We recapture and observe every fish including the previous detained fish into this sample, and mark them with a black spot.

3) All the previous captured and recaptured fish are thrown back into the lake, and then (k-1) independent recaptures are made. Every time all the fish caught are observed and marked with a black spot.

We now associate with every fish in the lake variate values

$$Y_j = 1 \qquad j = 1, \ldots, N.$$

$$W_j = \begin{cases} 1 & \text{if the j-th fish has a red spot;} \\ 0 & \text{otherwise.} \end{cases}$$

In this formulation, the total number of distinct fish is equal to the number of fish observed without any black spot. Then, similarly the sufficient statistic is given by

$$T = [\, x_{(1)}, \ldots, x_{(m)}; \; n_{(1)}, \ldots, n_{(k)} \,]; \quad \ldots \qquad (8.3.7)$$

where $n_{(1)}, \ldots, n_{(k)}$ are recapture sizes arranged in an increasing order. If by N_1, we denote the number of red fish in the lake, the probability of getting samples with a particular T is given by[1]

$$P(T) = \frac{N}{N_1} \frac{m_1}{m} \cdot P_1 , \qquad \ldots \qquad .. \qquad (8.3.8)$$

where m is the total number of distinct fish observed, m_1 is the total number of distinct red fish observed, and P_1 is given by (8.1.4).

From (8.3.5), it follows that an estimator of N is given by

$$\hat{N} = \frac{m}{m_1} \cdot N_1 \qquad . \qquad \ldots \qquad (8.3.9)$$

This estimator is not exactly unbiased as

$$E\left(\frac{m}{m_1} \cdot N_1\right) = N \sum_{m_1 > 0} P_1 = N \left[1 - \prod_{i=1}^{k} \frac{\binom{N-N_1}{n_i}}{\binom{N}{n_i}} \right] .$$

$$\ldots \qquad (8.3.10)$$

The reason that (8.3.5) is unbiased and (8.3.9) biased, is that in this case the condition $W_j > 0$ for all N units, is not satisfied, but this bias would be negligible if N_1 is large and the number of

---

1. This can be obtained from (8.3.4) by putting

$$W_j = \begin{cases} 1 & \text{if the fish has a red spot;} \\ 0 & \text{otherwise.} \end{cases}$$

recaptures is also large.

In the particular case of this scheme when $n_i = 1$ for all i=1,..., k, Bailey suggested the following estimator of N

$$\text{est}(N) = \frac{k}{r}, \qquad \ldots \qquad \ldots \qquad (8.3.11)$$

where r is the total number of red fish observed in k recaptures.

It is well-known that this is a biased estimator with bias negligible for large samples. For any convex loss function, it can be shown that a better estimator than (8.3.11) is given by (8.3.9). Though the former estimator has the same bias as Bailey's estimator, it has smaller risk function than Bailey's estimator for any convex loss function.

It is remarked that the above estimators are derived on the assumption that the recaptures form simple random samples from the fish population, and that the deaths and births of fish population can be neglected during the process of sampling. Consequently, these estimators are applicable only to those populations, where these assumptions are satisfied.

Appendix I

As indicated in Chapter II and Chapter IV, we shall present here a theorem and its corollaries which have been used extensively in the thesis.

Theorem:

$$\Sigma' \frac{n!}{\alpha_1!\,\alpha_2!\,\dots\,\alpha_\nu!}\, x_1^{\alpha_1} x_2^{\alpha_2} \dots x_\nu^{\alpha_\nu}$$

$$= (x_1 + x_2 + \dots + x_\nu)^n - \Sigma_1 (x_1 + x_2 + \dots + x_{\nu-1})^n + \dots (-)^{\nu-1} \Sigma_1 x_1^n ,$$

where the summation Σ' is taken over all positive integers $\alpha_1, \alpha_2, \dots, \alpha_\nu$ such that $\alpha_1 + \alpha_2 + \dots + \alpha_\nu = n$, and the summation Σ_1 extends over all possible combinations of x's chosen out of $x_1, x_2, \dots, x_\nu$.

Proof:

We shall prove this theorem by induction over n. The theorem is evidently true for any n when $\nu = 1$ or 2.

Let us suppose that the theorem is true for $n = 1, 2, \dots, m-1$. Then, by supposition

$$\Sigma' \frac{m!}{\alpha_1! \cdots \alpha_\nu!} x_1^{\alpha_1} \cdots x_\nu^{\alpha_\nu} = \sum_{\alpha_1 > 0} \frac{m!\, x_1^{\alpha_1}}{\alpha_1!\,(m-\alpha_1)!} \left[\Sigma \frac{(m-\alpha_1)!}{\alpha_2! \cdots \alpha_\nu!} x_2^{\alpha_2} \cdots x_\nu^{\alpha_\nu} \right],$$

$$= \sum_{\alpha_1 > 0} \frac{m!\, x_1^{\alpha_1}}{\alpha_1!\,(m-\alpha_1)!} \left[(x_2+\ldots+x_\nu)^{m-\alpha_1} - \Sigma_1 (x_2+\ldots+x_{\nu-1})^{m-\alpha_1} + \ldots (-)^{\nu-2} \Sigma_1 x_2^{m-\alpha_1} \right],$$

where the summations inside the square brackets are carried over $x_2, x_3, \ldots, x_\nu$ only.

The above expression, by summing over α_1, reduces to

$$\Sigma' \frac{n}{\alpha_1! \ldots \alpha_\nu!} x_1^{\alpha_1} \cdots x_\nu^{\alpha_\nu} = (x_1+\ldots+x_\nu)^n - \Sigma_1 (x_1+\ldots+x_{\nu-1})^n + \ldots (-)^{\nu-1} \Sigma_1 x_1^n .$$

This proves the truth of the theorem for $n = m$. Since it is true for $n = 1$, it is true in general.

Corollary 1:

$$\Sigma'' \frac{n!}{\alpha_1! \cdots \alpha_\nu!} x_1^{\alpha_1} \cdots x_\nu^{\alpha_\nu}$$

$$= \left[(x_1+\ldots+x_\nu)^n - \Sigma_1^i (x_1+\ldots+x_{\nu-1})^n + \ldots (-)^{\nu-1} x_1^n \right],$$

where Σ'' means summation over $\alpha_1, \ldots \alpha_\nu$ such that

$$\alpha_i \geq 0, \quad \alpha_{i'} > 0 \quad i' \neq i = 1, \ldots, \nu \text{ and } \sum_{i=1}^{\nu} \alpha_i = n,$$

and Σ_1^i means summation over all combinations of x's containing x_1.

Proof:

The proof is obtained directly by using the equality:

$$\Sigma'' \frac{n!}{\alpha_1! \cdots \alpha_\nu!} x_1^{\alpha_1} \cdots x_\nu^{\alpha_\nu} = \Sigma' \frac{n!}{\alpha_1! \cdots \alpha_i! \cdots \alpha_\nu!} x_1^{\alpha_1} \cdots x_i^{\alpha_i} \cdots x_\nu^{\alpha_\nu} +$$

$$\Sigma' \frac{n!}{\alpha_1! \cdots \alpha_{i-1}! \, \alpha_{i+1}! \cdots \alpha_\nu!} x_1^{\alpha_1} \cdots x_{i-1}^{\alpha_{i-1}} \, x_{i+1}^{\alpha_{i+1}} \cdots x_\nu^{\alpha_\nu} .$$

Corollary 2:

Putting $x_1 = x_2 = \cdots = x_\nu = 1$, we get

$$\Sigma' \frac{n!}{\alpha_1! \cdots \alpha_\nu!} = \nu^n - \binom{\nu}{1}(\nu - 1)^n + \cdots (-)^{\nu-1} \binom{\nu}{1} 1^n ,$$

and

$$\Sigma'' \frac{n!}{\alpha_1! \cdots \alpha_\nu!} = \nu^n - \binom{\nu-1}{1}(\nu - 1)^n + \cdots (-)^{\nu-1} \binom{\nu-1}{\nu-1} 1^n .$$

Alternative versions of these equalities are given in Chapter II in slightly different forms.

Appendix II

In this appendix, we give the following theorem which has been referred to in Chapter VI.

Theorem:

$$\Sigma' \frac{n!}{\lambda_1! \cdots \lambda_\nu!} \prod_{i=1}^{\nu} \left[\{ c_{\nu_i}(m_i \lambda_i) \} x_i^{\lambda_i} \right]$$

$$= \sum_{r=0}^{[\sum \nu_i - 1]} (-)^r \Sigma_1 \binom{\nu_1}{\alpha_1} \cdots \binom{\nu_\nu}{\alpha_\nu} [x_1 (\nu_1 - \alpha_1)^{m_1} + \cdots + x_\nu (\nu_\nu - \alpha_\nu)^{m_\nu}]^n ,$$

where Σ' stands for the summation over all possitive integral λ_i's such that $\sum_{i=1}^{\nu} \lambda_i = n$, and Σ_1 stands for the summation over non-negative α_i's such that $\alpha_1 + \alpha_2 + \cdots + \alpha_\nu = r$

Proof:

Since $c_{\nu_i}(m_i \lambda_i) = 0$ whenever $\lambda_i = 0$, the summation Σ' is equivalent to taking the summation over all non-negative λ_i's such that $\sum_{i=1}^{\nu} \lambda_i = n$. Further, as it is not difficult to show (by direct expansion of $c_{\nu_i}(m_i \lambda_i)$) that

$$\prod_{i=1}^{\nu}\left[\left\{c_{\nu_i}(m_i\ \lambda_i)\right\}\ x_i^{\lambda_i}\right] = \sum_{r=0}^{[\Sigma\nu_i - 1]} (-)^r \Sigma_1 \binom{\nu_1}{\alpha_1}\cdots\binom{\nu_\nu}{\alpha_\nu}\left[\left\{x_1(\nu_1 - \alpha_1)^{m_1}\right\}^{\lambda_1}\right] \cdot$$

$$\cdots\left\{x_\nu(\nu_\nu - \alpha_\nu)^{m_\nu}\right\}^{\lambda_\nu},$$

we get the result on putting this expression in the above equation and summing by multinomial theorem.

BIBLIOGRAPHY

1. Bailey, N. T. J., 'On estimating the size of mobile population from recapture data', Biometrika, Vol. 38 (1951), pp. 293-306.

2. Basu, D., 'On symmetric estimators in point estimation with convex weight functions', Sankhya, Vol. 12 (1953), pp. 45-52.

3. Basu, D., 'On the optimum character of some estimators used in multistage sampling problems', Sankhya, Vol. 13 (1954), pp. 365-368.

4. Basu, D., 'On sampling with and without replacement', Sankhya, Vol. 20 (1958), pp. 287-294.

5. Chapman, D.G., 'Inverse, multiple and sequential simple censuses', Biometrics, Vol. 8 (1952), pp. 286-306.

6. Cochran, W.G., Sampling Techniques, New York, John Wiley and Sons, 1953.

7. Craig , C. C., 'On the utilization of marked specimens in estimating populations of flying insects', Biometrika, Vol. 40 (1953), pp. 170-176.

8. Das, A. C., 'On two phase sampling and sampling with varying probabilities', Bull. Int. Stat. Inst., Vol. 33 (1951), pp. 105-112.

9. Davis, H. T. Tables of the Higher Mathematical Functions, Vol. II, Bloomington, Indiana, The Principia Press, Inc., 1935.

10. De Lury, D. B., 'On estimation of biological populations', Biometrics, Vol. 3 (1947), pp. 145-167.

11. Deming, W. E., Some Theory of Sampling, New York, John Wiley and Sons, 1950.

12. Des Raj, 'Ratio estimation in sampling with equal and unequal probabilities' J. Ind. Soc. Agr. Stat., Vol. 6 (1954), pp. 127-138.

13. Des Raj, 'Some estimators in sampling with varying probabilities without replacement', J. Amer. Stat. Assn., Vol. 51 (1956), pp. 269-284.

14. Des Raj and Khamis, S. H., 'Some remarks on sampling with replacement', Ann. Math. Stat., Vol. 39 (1958), pp. 550-557.

15. Durbin, J., 'Some results in sampling theory when the units are selected with unequal probabilities', Jour. Roy. Stat. Soc., Series B, Vol. 15 (1953), pp. 262-269.

16. Feller, W., An Introduction to Probability Theory and Its Applications, Vol. I, Bombay, Asia Publishing House, 1960.

17. Fraser, D.A.S., Nonparametric Methods in Statistics, New York, John Wiley and Sons, 1957.

18. Godambe, V.P., 'A unified theory of sampling from finite populations', J. Roy. Stat. Soc., Series B, Vol.7 (1955), pp. 269-277.

19. Godambe, V.P., 'An admissible estimate for any sampling design', Sankhya, Vol. 22 (1960), pp. 285-288.

20. Gupta, H., 'Tables of distributions', Research Bulletin of the East Punjab University, No.2,3 (1950), pp. 13-44.

21. Haldane, J.B.S., 'On method of estimating frequencies', Biometrika, Vol. 33 (1945), pp. 222-225.

22. Hansen, M.H., Hurwitz, W.N. and Madow, W.G., Sample Survey Method and Theory, Vol. I and Vol. II, New York, John Wiley and Sons, 1953.

23. Horvitz, D.G. and Thompson, D.J., 'A generalization of sampling without replacement from a finite universe', J. Amer. Stat. Assn., Vol. 47 (1952), pp.663-685.

24. Lahiri, D.B., 'A method of sample selection providing unbiased ratio estimators', Bull. Int. Stat. Inst., Vol. 23 (1951), pp. 133-140.

25. Lehman, E.L. and Scheffe, H., 'Completeness, similar regions and unbiased estimation', Part I, Sankhya, Vol. 10 (1950), pp. 305-340.

26. Leslie, P.H., 'The estimation of population parameters from data obtained by means of the capture - recapture method', Part II, Biometrika, Vol. 39 (1953), pp. 363-388.

27. Leslie, P.H., Chitty, D. and Chitty, H., 'The estimation of population parameters by capture - recapture method', Biometrika, Vol. 40 (1953), pp. 137-169.

28. Mickey, M.R., 'Some finite population unbiased ratio and regression estimators', J. Amer. Stat. Assn., Vol. 54 (1959), pp. 596-612.

29. Midzuno, H., 'On sampling system with probability proportional to sum of sizes', Ann. Inst. Stat. Math., Vol. 3 (1952), pp. 99-107.

30. Moran, P.A.P., 'A mathematical theory of animal trapping', Biometrika, Vol. 38 (1951), pp. 307-311.

31. Murthy, M.N., 'Ordered and unordered estimators in sampling without replacement', Sankhya, Vol. 18 (1957), pp. 379-390.

32. Murthy, M.N. and Nanjamma, N.S., 'Almost unbiased ratio estimators based on interpenetrating sub-sample estimates', Sankhya, Vol. 21 (1959), pp. 381-392.

33. Nanjamma, N.S., Murthy, M.N. and Sethi, V.K., 'Some sampling systems providing unbiased ratio estimators', Sankhya, Vol. 21 (1959), pp. 299-314.

34. Narain, R. D., 'On sampling without replacement with varying probabilities', J. Ind. Soc. Agr. Stat., Vol. 3 (1951), pp. 169-174.

35. Pathak, P.K., 'On the evaluation of moments of distinct units in a sample', accepted for publication in Sankhya, 1961.

36. Pathak, P.K., "Use of 'order-statistic' in without replacement sampling", accepted for publication in Sankhya, 1961.

37. Roy Chowdhury, D.K., 'Sampling with varying probabilities', part of the thesis submitted for the Associateship of the Indian Statistical Institute, 1956.

38. Roy, J. and Chakravarty, I.M., 'Estimating the mean of a finite population', Ann. Math. Stat., Vol. 31 (1960), pp. 392-398.

39. Schnabel, Z.E., 'The estimation of total fish in a lake', Amer. Math. Mon., Vol. 45 (1938), pp. 348-350.

40. Sen, A.R., 'Present states of probability sampling and its use in estimation of characteristics' (an abstract), Econometrika, Vol. 20 (1952), pp. 103.

41. Sen, A.R., 'On the selection of n primary sampling units from a stratum structure $(n \geq 2)$', Ann. Math. Stat., Vol. 26 (1955), pp. 744-751.

42. Stephan, F.F., 'The expected value and variance of the reciprocal and other negative powers of a positive Bernoullian variate', Ann. Math. Stat., Vol. 16 (1945), pp. 50-61.

43. Stevens, W.L., 'Sampling without replacement proportional to size', J. Roy. Stat. Soc., Series B, Vol. 20 (1958), pp. 393-397.

44. Sukhatme, P.V., Sampling Theory of Surveys with Applications, Indian Society of Agricultural Statistics, 1953.

45. Yates, F., Sampling Method for Census and Surveys, London, Griffin, 1949.

46. Yates, F. and Grundy, P.M., 'Selection without replacement from within strata with probability proportional to size', J. Roy. Stat. Soc., Series B, Vol. 15 (1953), pp. 253-261.