(g) The Jessen Procedures 35 and 36 are simple for selection but cannot be rotated easily and score poorly on most other counts.

(h) Sinha's Procedures 42 and 43 look particularly promising for moderate values of n . Since the π_{IJ} are arbitrary they can be chosen to minimize (or using expression (3.7.3) to come close to minimizing) the variance of the Sen-Yates-Grundy variance estimator. For large values of n the procedures become unmanageable.

(j) Because the Systematic Procedures 2 and 3 are so convenient on all counts other than variance estimation, the approximate variance formula (3.7.4) which does not depend on the π_{IJ} may be used to remedy this deficiency.

CHAPTER 4

SELECTION PROCEDURES USING SPECIAL ESTIMATORS

4.1 INTRODUCTION

In Chapter 3 a comparison was made of those selection procedures for which the Horvitz-Thompson estimator possessed the ratio estimator property. It was mentioned, however, in Section 1.7 that certain special estimators had also been devised for use with particular selection procedures, and that in the context of these procedures they also possessed the ratio estimator property. In this Chapter the performance of these special estimators will be compared in the context of their appropriate selection procedures; that is,

(i) Das's estimator with Procedure 4,

(ii) Raj's and Murthy's estimators with Procedure 4,

(iii) the Rao-Hartley-Cochran (RHC) estimator with Procedure 25,

(iv) unbiased and ratio estimators for Poisson sampling with Procedure 27,

(v) unbiased and ratio estimators for Modified Poisson Sampling with Procedure 38,

(vi) unbiased and ratio estimators for Collocated Sampling with Procedure 39, and

(vii) Lahiri's estimator with Procedures 45-46.

4.2 DESCRIPTION OF SPECIAL ESTIMATORS

4.2.1 Das's Estimator

This estimator was devised by Das (1951) for use with Procedure 4, the draw by draw procedure with probabilities proportional to original size at each draw. He first suggested the following ordered linear combinations, which are unbiased estimators of population total Y ,

$$\left.\begin{aligned} t_1' &= y_1/p_1 \\ t_2' &= (1-p_1)y_2/p_1p_2(N-1) \\ &\quad\cdots \\ t_r' &= \left\{\prod_{i=1}^{r-1}\left(1 - \sum_{j=1}^{k} p_j\right)y_r\right\} \div \left\{\prod_{i=1}^{r} p_i \prod_{i=1}^{r-1}(N-i)\right\} \end{aligned}\right\} . \qquad (4.2.1)$$

Every linear combination $t' = \sum_{r=1}^{n} c_r t_r'$, where $\sum_{r=1}^{n} c_r = 1$, is an unbiased estimator of Y . The choice of c_r is free but for simplicity Das chose $c_r = n^{-1}$. The unbiased variance estimator he provided can assume negative values.

Murthy (1957) showed that estimators such as those in (4.2.1) could be improved by *unordering*; that is, taking the expectation of the estimators derived from any given estimator formula by considering all possible orderings (permutations) of the observed sample. The unordered form of t_1' is identical with Murthy's estimator (Pathak 1961). This estimator is considered in Section 4.2.2. Unorderings of t_r' , $r \neq 1$, yield estimators inferior to Murthy's (Samiuddin, Hanif and Asad 1978). These estimators will not be considered further in this monograph.

4.2.2 The Raj and Murthy Estimators

These estimators were devised by Raj (1956a) and Murthy (1957) for use with Procedure 4. The set of unbiased and mutually uncorrelated estimators of population total Y suggested by Raj is

$$\left.\begin{aligned} t_1 &= y_1/p_1 , \\ t_2 &= y_1 + y_2(1-p_1)/p_2 , \\ &\quad\cdots \\ t_n &= y_1 + y_2 + \ldots + y_{n-1} + \frac{y_n}{p_n}(1 - p_1 - p_2 - \ldots - p_{n-1}) \end{aligned}\right\} . \qquad (4.2.2)$$

The estimator t_{mean} of the population total Y is the arithmetic mean of the above set of estimators, which for $n = 2$ yields

$$t_{\text{mean}} = \tfrac{1}{2}\left[(1+p_1)\frac{y_1}{p_1} + (1-p_1)\frac{y_2}{p_2}\right] , \qquad (4.2.3)$$

with variance

$$V(t_{\text{mean}}) = \frac{1}{8}\sum\sum_{\substack{I,J=1 \\ J\neq I}}^{N} P_IP_J(2-P_I-P_J)\left(\frac{Y_I}{P_I} - \frac{Y_J}{P_J}\right)^2 . \qquad (4.2.4)$$

An unbiased estimator of (4.2.4) given by Raj is

$$v(t_{\text{mean}}) = \frac{(1-p_1)^2}{4}\left(\frac{y_1}{p_1} - \frac{y_2}{p_2}\right)^2 . \qquad (4.2.5)$$

Pathak (1967a) derived a formula for the variance of t_{mean} for any n . This variance formula is

$$V(t_{\text{mean}}) = \frac{1}{2n^2}\sum\sum_{\substack{I,J=1 \\ J\neq I}}^{N} P_IP_J\left[1 + \sum_{r=2}^{n} Q_{IJ}(r-1)\right]\left[\frac{Y_I}{P_I} - \frac{Y_J}{P_J}\right]^2 , \qquad (4.2.6)$$

where $Q_{IJ}(r-1)$ denotes the probability of non-inclusion of one or both of the units I and J in the first $(r-1)$ sample units.

An unbiased estimator of variance suggested by Raj (1956a) for any n is

$$v(t_{\text{mean}}) = \frac{1}{n(n-1)}\sum_{k=1}^{n}(t_k-\bar{t})^2 , \qquad (4.2.7)$$

which is non negative for all $n \geq 2$. Here $\bar{t} = \frac{1}{n}\sum_{k=1}^{n} t_k$.

Murthy (1957) suggested that the estimator t_{mean} could be improved by the process of unordering. For $n = 2$ the unordered form of t_{mean} , denoted by t_{symm} , may be written as follows:

$$t_{\text{symm}} = \frac{1}{2-p_1-p_2}\left[\frac{y_1}{p_1}(1-p_2) + \frac{y_2}{p_2}(1-p_1)\right] . \qquad (4.2.8)$$

The variance of t_{symm} for $n = 2$ is

$$V(t_{\text{symm}}) = \tfrac{1}{2}\sum\sum_{\substack{I,J=1 \\ J\neq I}}^{N} P_IP_J\frac{1-P_I-P_J}{2-P_I-P_J}\left[\frac{Y_I}{P_I} - \frac{Y_J}{P_J}\right]^2 . \qquad (4.2.9)$$

An unbiased variance estimator of (4.2.9) is

$$v(t_{\text{symm}}) = \frac{(1-p_1)(1-p_2)(1-p_1-p_2)}{(2-p_1-p_2)^2}\left[\frac{y_1}{p_1} - \frac{y_2}{p_2}\right]^2 . \qquad (4.2.10)$$

Murthy (1957) further showed that an unordered and therefore more efficient iased variance estimator for t_{mean} for $n = 2$ is

$$v_M(t_{\text{mean}}) = \tfrac{1}{4}(1-p_1)(1-p_2)\left[\frac{y_1}{p_1} - \frac{y_2}{p_2}\right]^2 . \qquad (4.2.11)$$

Pathak (1967a) derived the following variance formula for t_{symm} for any 2 :

$$V(t_{\text{symm}}) = \tfrac{1}{2}\sum_{\substack{I,J=1\\J\neq I}}^{N}\sum P_I P_J\left\{1 - \sum_{s\ni IJ}^{*}\frac{p(s|I)p(s|J)}{p(s)}\right\}\left(\frac{Y_I}{P_I} - \frac{Y_J}{P_J}\right)^2 , \qquad (4.2.12)$$

ere $p(s)$ denotes the probability of obtaining the sample s of n units, $s|I)$ denotes the probability of obtaining the sample s given that unit I was awn first, and $\sum\limits_{s\ni IJ}^{*}$ denotes the sum over all samples s containing units I and .

Pathak (1967b) also derived the following unbiased variance estimator for any :

$$v(t_{\text{symm}}) = \tfrac{1}{2}\sum_{\substack{i,j=1\\j\neq i}}^{n}\sum p_i p_j[p(s)p(s|ij)-p(s|i)p(s|j)]p(s)^{-2}\left[\frac{y_i}{p_i} - \frac{y_j}{p_j}\right]^2 , \qquad (4.2.13)$$

ere $p(s|ij)$ denotes the conditional probability of selecting the observed sample , given that units i and j were selected in that order at the first two draws.

(4.2.13) is non negative but the computation becomes cumbersome as n increases. ayless (1968) developed a computer programme to calculate $p(s|ij)$, $p(s|i)$ and $p(s)$ or *upswor*.

Pathak (1961) showed that Murthy's estimator (4.2.8) could be obtained by nordering any linear combination of the individual Raj estimators (4.2.2).

Note. Basu (1970) suggested that it was natural to estimate the ratio

$$\left(\sum_{I=1}^{N} Y_I - \sum_{i=1}^{n} y_i\right) \div \left(\sum_{I=1}^{N} P_I - \sum_{i=1}^{n} p_i\right)$$

y some sort of an average of the observed ratios. Two particular averages which he uggested were $\sum\limits_{i=1}^{n} y_i \div \sum\limits_{i=1}^{n} p_i$, which led to the conventional ratio estimator, and $n^{-1}\sum\limits_{i=1}^{n} y_i/p_i$ which led to the estimator

$$\hat{y}_B = \sum_{i=1}^{n} y_i + \frac{1}{n}\sum_{i=1}^{n}\frac{y_i}{p_i}\left[1 - \sum_{j=1}^{n} p_j\right] . \qquad (4.2.14)$$

He claimed that these two estimators had 'as much face validity' as unordered forms of the individual Raj estimator (4.2.2), and that although they were not unbiased, they were far simpler to calculate. His argument for 'face validity' appears to be based on their being symmetric functions of the sample values and possessing the ratio estimator property. The authors are not aware of any investigation that has been made as to the performance of Basu's estimators with Procedure 4, but they are not design unbiased, even asymptotically.

4.2.3 The Rao-Hartley-Cochran Estimator

The RHC sampling scheme (Procedure 25) has already been described in Chapter 2. The population units are divided randomly into groups containing N_J units, $J = 1, 2, 3, \ldots, n$, where the N_J are predetermined. One unit is selected from each group, the probabilities of selection being the normed measures of size within the group. The RHC estimation procedure is to form the Horvitz-Thompson estimator for each group separately, and add over the groups.

The unbiased estimator of population total Y is, therefore,

$$y'_{\text{RHC}} = \sum_{i=1}^{n}\frac{y_{it}\pi_i}{p_{it}} , \qquad (4.2.15)$$

where p_{it} is the sample value of the normal measure of size P_{iT} , $\pi_i = \sum\limits_{T=1}^{N_i} P_{iT}$, and $\sum\limits_{i=1}^{n}\pi_i = 1$.

The variance of (4.2.15) is

$$V(y'_{\text{RHC}}) = \left\{n\left(\sum_{i=1}^{n} N_i^2 - N\right) / N(N-1)\right\}\left(\sum_{i=1}^{n}\sum_{T=1}^{N_i}\frac{Y_{iT}^2}{nP_{iT}} - \frac{Y^2}{n}\right) . \qquad (4.2.16)$$

Rao, Hartley and Cochran minimized (4.2.16) by noting that since $N = nR + k$, where $0 < k < n$ and R is a positive integer, it was possible to put $N_1 = N_2 = \ldots = N_k = R+1$ and $N_{k+1} = N_{k+2} = \ldots = N_n = R$, in which case (4.2.16) reduces to

$$V(y'_{\text{RHC}}) = \left\{1 - \frac{n-1}{N-1} + \frac{k(n-k)}{N(N-1)}\right\}\left\{\sum_{i=1}^{n}\sum_{T=1}^{N_i}\frac{Y_{iT}^2}{nP_{iT}} - \frac{Y^2}{n}\right\} . \qquad (4.2.17)$$

f N is a multiple of n , $k = 0$, and the variance further reduces to

$$V(y'_{\mathrm{RHC}}) = \left(1 - \frac{n-1}{N-1}\right)\left(\sum_{i=1}^{n}\sum_{T=1}^{N_i}\frac{Y_{iT}^2}{nP_{iT}} - \frac{Y^2}{n}\right) . \qquad (4.2.18)$$

An unbiased variance estimator of (4.2.16) is

$$v(y'_{\mathrm{RHC}}) = \left[\left(\sum_{i=1}^{n} N_i^2 - N\right) / \left(N^2 - \sum_{i=1}^{n} N_i^2\right)\right] \sum_{i=1}^{n} \pi_i\left(\frac{y_{it}}{p_{it}} - y'_{\mathrm{RHC}}\right)^2 . \qquad (4.2.19)$$

or the simpler forms (4.2.17) and (4.2.18), (4.2.19) reduces to

$$v(y'_{\mathrm{RHC}}) = \frac{N^2 + k(n-k) - Nn}{N^2(n-1) - k(n-k)} \sum_{i=1}^{n} \pi_i\left(\frac{y_{it}}{p_{it}} - y'_{\mathrm{RHC}}\right)^2 \qquad (4.2.20)$$

nd

$$v(y'_{\mathrm{RHC}}) = \frac{1}{n-1}\left(1 - \frac{n}{N}\right) \sum_{i=1}^{n} \pi_i\left(\frac{y_{it}}{p_{it}} - y'_{\mathrm{RHC}}\right)^2 . \qquad (4.2.21)$$

.2.4 Poisson Sampling

Poisson sampling as defined by Hajek (1964) gives each unit in the population a ertain probability of inclusion in the sample which will be denoted by π_I for the th unit, $I = 1, 2, \ldots, N$. To select, a set of N binomial trials is carried out o determine whether each unit in turn is to be included in the sample s or not.

The unbiased Horvitz-Thompson estimator of the population total is

$$y'_{\mathrm{PS}} = \sum_{i \in s} \frac{y_i}{\pi_i} . \qquad (4.2.22)$$

ince the joint probability of inclusion π_{IJ} takes the simple form $\pi_{IJ} = \pi_I\pi_J$, the ariance of (4.2.22) is

$$V(y'_{\mathrm{PS}}) = \sum_{I=1}^{N} (1-\pi_I) \frac{Y_I^2}{\pi_I} , \qquad (4.2.23)$$

nd an unbiased estimator of (4.2.23) is

$$v(y'_{\mathrm{PS}}) = \sum_{i \in s} (1-\pi_i) \frac{y_i^2}{\pi_i^2} . \qquad (4.2.24)$$

ecause the sample size varies in this sampling procedure, the ratio estimator

$$y''_{\mathrm{PS}} = \begin{cases} \dfrac{y'_{\mathrm{PS}}}{m} \cdot n & \text{if } m > 0 , \\ 0 & \text{otherwise,} \end{cases} \qquad (4.2.25)$$

is more efficient than y'_{PS} .

The mean square error of y''_{PS} is given approximately by

$$V(y''_{\mathrm{PS}}) \doteqdot \sum_{I=1}^{N} \pi_I(1-\pi_I)\left(\frac{Y_I}{\pi_I} - \frac{Y}{n}\right)^2 + P_0Y^2 , \qquad (4.2.26)$$

where $P_0 = Pr(m = 0)$ and $n = E(m) = \sum_{I=1}^{N} \pi_I$. (A proof of (4.2.26) is given in Appendix B.)

The conventional estimator of the approximation (4.2.25) is

$$v(y''_{\mathrm{PS}}) = \sum_{i \in s} (1-\pi_i)\left(\frac{y_i}{\pi_i} - \frac{y''_{\mathrm{PS}}}{n}\right)^2 + P_0 y''^2_{\mathrm{PS}} ; \qquad (4.2.27)$$

but a more stable estimator is obtained by multiplying the first expression on the right hand side by n/m .

4.2.5 Modified Poisson Sampling

Modified Poisson sampling is a procedure which ensures that an empty sample is never selected. It was first suggested by Ogus and Clark (1971). An ordinary Poisson sample is drawn first, but if there are no units in that sample, a second Poisson sample is drawn, and so on repeatedly until a non-empty sample is achieved.

Assuming that the probability of including the Ith population unit in sample is to be held constant at π_I , the probability of selecting this unit in each ordinary Poisson sample drawn must be $\pi_I(1-P_0^*)$, where P_0^* is the probability of selecting an empty sample at each such draw. Then

$$P_0^* = \prod_{I=1}^{N} \{1 - \pi_I(1-P_0^*)\} ,$$

and its value may be obtained iteratively using the initial value zero. For modified Poisson sampling

$$\pi_{IJ} = \pi_I\pi_J(1-P_0^*) \text{ , for } I \neq J .$$

The variance of the Horvitz-Thompson estimator, y'_{MPS} , formed analogously to (4.2.22) but using modified Poisson sampling is

$$V\left(y'_{\mathrm{MPS}}\right) = \sum_{I=1}^{N} \left(1-\pi_I\right) \frac{Y_I^2}{\pi_I} - P_0^*\left[Y^2 - \sum_{I=1}^{N} Y_I^2\right] , \tag{4.2.28}$$

and an unbiased estimator of this variance is

$$v\left(y'_{\mathrm{MPS}}\right) = \sum_{i\in s} \left(1-\pi_i\right) \frac{y_i^2}{\pi_i^2} - \frac{P_0^*}{1-P_0^*} \left[y'^2_{\mathrm{MPS}} - \sum_{i\in s} \frac{y_i^2}{\pi_i^2}\right] . \tag{4.2.29}$$

The mean square error of the ratio estimator y''_{MPS} , formed analogously to (4.2.25) is approximately given by

$$V\left(y''_{\mathrm{MPS}}\right) \doteq \sum_{I=1}^{N} \pi_I\left\{1-\left(1-P_0^*\right)\pi_I\right\}\left[\frac{Y_I}{\pi_I} - \frac{Y}{n}\right]^2 . \tag{4.2.30}$$

The conventional estimator of this approximate mean square error is

$$v\left(y''_{\mathrm{MPS}}\right) = \sum_{i\in s} \left\{1-\left(1-P_0^*\right)\pi_i\right\}\left[\frac{y_i}{\pi_i} - \frac{y''_{\mathrm{MPS}}}{n}\right]^2 , \tag{4.2.31}$$

but a more stable estimator can be obtained by multiplying this expression by n/m . We notice that $V\left(y'_{\mathrm{MPS}}\right) < V\left(y'_{\mathrm{PS}}\right)$ and that provided

$$P_0^* \sum_{I=1}^{N} \pi_I\left[\frac{Y_I}{\pi_I} - \frac{Y}{n}\right]^2 < P_0Y^2 \tag{4.2.32}$$

- a condition which is easily satisfied - it is also true that $V\left(y''_{\mathrm{MPS}}\right) < V\left(y''_{\mathrm{PS}}\right)$. Despite this, the only advantage of modified Poisson sampling over ordinary Poisson sampling is that it ensures a non-empty sample. If the sample selected is much smaller (or much larger) than the target size, modified Poisson sampling provides no remedy and will therefore receive no further consideration in this monograph. A procedure which ensures a more stable sample size is described in the following Subsection.

4.2.6 Collocated Sampling

Collocated sampling is similar to Poisson sampling, but reduces the variation in sample size by requiring the random variable r_I to be uniformly *spaced* instead of uniformly *distributed* over the interval $[0, 1)$. A random ordering L $\left(L_I = 1, 2, \ldots, N\right)$ is chosen with equal probabilities, and a random variable θ is also selected from a uniform distribution over the interval $[0, 1)$. For each I we then define

$$r_I = \frac{L_I+\theta-1}{N} . \tag{4.2.33}$$

The Horvitz-Thompson estimator is still used, but now no simplification of its variance formula is possible. The variance of y'_{CS} , formed analogously to (4.2.22) is therefore

$$V\left(y'_{\mathrm{CS}}\right) = \sum_{I=1}^{N} \left(1-\pi_I\right) \frac{Y_I^2}{\pi_I} + 2 \sum_{\substack{I,J=1\\J>I}}^{N}\sum \left(\pi_{IJ}-\pi_I\pi_J\right) \frac{Y_IY_J}{\pi_I\pi_J} . \tag{4.2.34}$$

An unbiased variance estimator is well known to be

$$v\left(y'_{\mathrm{CS}}\right) = \sum_{i\in s} \left(1-\pi_i\right) \frac{y_i^2}{\pi_i^2} + 2 \sum_{\substack{i,j\in s\\j>i}}\sum \frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}} \frac{y_iy_j}{\pi_i\pi_j} . \tag{4.2.35}$$

The ratio estimator y''_{CS} , formed analogously to (4.2.25),has approximate mean square error (see Appendix B) given by

$$\begin{aligned} V\left(y''_{\mathrm{CS}}\right) \doteq & \sum_{I=1}^{N} \pi_I\left(1-\pi_I\right)\left[\frac{Y_I}{\pi_I} - \frac{Y}{n}\right]^2 \\ & + 2 \sum_{\substack{I,J=1\\J>I}}^{N}\sum \left(\pi_{IJ}-\pi_I\pi_J\right)\left[\frac{Y_I}{\pi_I} - \frac{Y}{n}\right]\left[\frac{Y_J}{\pi_J} - \frac{Y}{n}\right] + P_{0C}Y^2 , \end{aligned} \tag{4.2.36}$$

where P_{0C} is the probability of selecting an empty sample.

The conventional estimator of this approximate mean square error is

$$\begin{aligned} v\left(y''_{\mathrm{CS}}\right) = & \sum_{i\in s} \left(1-\pi_i\right)\left[\frac{y_i}{\pi_i} - \frac{y''}{n}\right]^2 \\ & + 2 \sum_{\substack{i,j\in s\\j>i}}\sum \frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}} \left[\frac{y_i}{\pi_i} - \frac{y''}{n}\right]\left[\frac{y_j}{\pi_j} - \frac{y''}{n}\right] + P_{0C}y''^2 . \end{aligned} \tag{4.2.37}$$

A more stable estimator than (4.2.37) may be obtained by multiplying the first term of (4.2.37) by n/m and the second term by $n(n-1)/m(m-1)$.

The expressions for π_{IJ} and for P_{0C} were devised by Brewer, Early and Hanif (1980) (see Appendices D and E). They are most conveniently expressed in terms of a population listed in ascending order of size, so that $\pi_1 \le \pi_2 \le \ldots \le \pi_N$. Writing $N\pi_I = \left[N\pi_I\right] + K_I$ where $[A]$ denotes the integral part of A , they showed that

$$\pi_{IJ} = \frac{\left[N\pi_I\right]\left(N\pi_J-1\right)+K_I\left[N\pi_J\right]+\max\left\{\left(K_J-K_I\right),0\right\}}{N(N-1)} . \tag{4.2.38}$$

Clearly also

$$P_{0C} = \int_{\theta=0}^{1} \{P_{0C}|\theta\}d\theta \;, \qquad (4.2.39)$$

where $\{P_{0C}|\theta\}$ is the probability of an empty sample given a particular value of θ .

Approximate expressions for π_{IJ} and P_{0C} can be obtained on the assumption that the π_I are integer multiples of N^{-1} , in which case $k_I = 0$ for all I , *viz.*

$$\pi_{IJ} = \frac{\pi_I\left(N\pi_J-1\right)}{N-1} \;, \qquad (4.2.40)$$

and

$$\pi_{IJ} - \pi_I\pi_J = \frac{\pi_I\left(\pi_J-1\right)}{N-1} \;, \qquad (4.2.41)$$

where $\pi_J > \pi_I$.

Even when the π_I are not all integer multiples of N^{-1} , the use of (4.2.41) for $\left(\pi_{IJ}-\pi_I\pi_J\right)$ in the formulae (4.2.34) and (4.2.36) results in good approximations for the variance of y'_{CS} and the mean square error of y''_{CS} . With this same approximate formula for π_{IJ} , P_{0C} becomes

$$P_{0C} = \begin{cases} \dfrac{1}{N!}\displaystyle\prod_{I=1}^{N}\left(I-N\pi_I\right) & \text{if } \displaystyle\min_I \left(I-N\pi_I\right) > 0 \;, \\ 0 & \text{otherwise.} \end{cases} \qquad (4.2.42)$$

This P_{0C} is much smaller than the corresponding P_0 for Poisson sampling.

4.2.7 Lahiri's Estimator

The use of Procedures 45 and 46, for which the probability of selection of a sample is proportional to its total measure of size, ensure that the conventional ratic estimator

$$y'' = \left\{\left(\sum_{i=1}^{n} y_i\right) \Big/ \left(\sum_{i=1}^{n} z_i\right)\right\}Z \qquad (4.2.43)$$

is unbiased.

When n is large, the probabilities of selecting all particular samples tend to equality, and the variance of y'' may be evaluated and estimated using the Taylor expansion expressions familiar from equal probability sampling.

When, however, n is small and the inequalities in selection probability cannot be ignored, difficulties arise in the estimation of variance. For $n = 2$ the variance is

$$V(y'') = \frac{1}{N-1}\sum_{\substack{I,J=1\\J>I}}^{N}\sum \frac{\left(Y_I+Y_J\right)^2}{P_I+P_J} - Y^2 \;. \qquad (4.2.44)$$

Rao and Bayless (1969) used model (1.8.5) and obtained the following expression for the expected variance of (4.2.43) for $n = 2$:

$$E^*V(y'') = \sigma^2Z^2(N-1)^{-1}\sum_{J>I}^{N}\sum \frac{P_I^{2\gamma}+P_J^{2\gamma}}{P_I+P_J} - \sum_{I=1}^{N} P_I^{2\gamma} \;. \qquad (4.2.45)$$

They also found that for $n = 2$ the Lahiri estimator was more efficient than the Horvitz-Thompson, Raj, Murthy and RHC estimators when either

(a) few units in the population had large sizes relative to the sizes of remaining units in the population, and samples containing those units gave good estimates of Y , or

(b) the coefficient of variation of the benchmark variable was small.

For other populations it had poor efficiency. Raj (1954) and Sen (1955) provided an unbiased variance estimator for $n = 2$, namely

$$v_R(y'') = y''^2 - \frac{1}{p_1+p_2}\left[\left(y_1-y_2\right)^2+2Ny_1y_2\right] \;. \qquad (4.2.46)$$

This can take negative values, and was found by Rao and Bayless (1969) to have very poor stability. The modification suggested by Sen (1955), replacing negative values of $v_R(y'')$ by zero, did not lead to any substantial improvement.

Bayless and Rao (1970) in extending their earlier investigations to the cases $n = 3$ and $n = 4$, arrived at the same conclusions as for $n = 2$, both with respect to the efficiency of the estimator of total and the poor performance of the variance estimators. The reader is referred to that paper for variance formulae and estimators.

More recently Rao and Vijayan (1977) have proposed two new unbiased variance estimators which for some populations are nonnegative. For the case $n = 2$ these estimators coincide and take the form

$$v_a(y'') = -a_{12}(s)z_1z_2\left(\frac{y_1}{z_1} - \frac{y_2}{z_2}\right)^2 \;, \qquad (4.2.47)$$

where

$$a_{12}(s) = \frac{Z}{z_1+z_2}\left[\frac{Z}{z_1+z_2} - (N-1)\right] \;. \qquad (4.2.48)$$

For $n > 2$ the first estimator suggested is

$$v_{a1}(y'') = -\sum_{\substack{i,j\in s\\ j>i}}\sum \frac{a_{ij}}{\pi_{ij}} z_i z_j \left(\frac{y_i}{z_i} - \frac{y_j}{z_j}\right)^2 \qquad (4.2.49)$$

where

$$\pi_{ij} = \frac{(n-1)(N-n)}{(N-1)(N-2)} \frac{z_i+z_j}{Z} + \frac{(n-1)(n-2)}{(N-1)(N-2)} \qquad (4.2.50)$$

and

$$a_{ij} = Z\sqrt{\binom{N-1}{n-1}} \sum_{s\ni i,j} \sqrt{\left(\sum_{k\in s} z_k\right)} - 1 \ . \qquad (4.2.51)$$

The second estimator suggested is

$$v_{a2}(y'') = Z\sqrt{\left(\sum_{k\in s} z_k\right)}\left[\frac{N-1}{n-1} - Z\sqrt{\left(\sum_{k\in s} z_k\right)}\right] \sum_{\substack{i,j\in s\\ j>i}} z_i z_j \left(\frac{y_i}{z_i} - \frac{y_j}{z_j}\right)^2 \ . \qquad (4.2.52)$$

This second estimator is computationally simpler than the first, but is consistently less efficient, and usually has a greater probability of producing a negative estimate. Both estimators are typically (though not invariably) much more efficient than $v_R(y'')$.

4.3. COMPARISON OF SAMPLING SCHEMES USING SPECIAL ESTIMATORS

The criteria for comparison will be as in Chapter 3; limitation to the case $n > 2$, simplicity in selection, simplicity in variance estimation, the efficiency of the estimator of total, the unbiasedness and stability of the variance estimator and rotatability. All these concepts have been described in Chapter 3.

4.4. LIMITATION TO SAMPLE SIZE $n = 2$

When the Horvitz-Thompson estimator was used, the limit to the number of units which could be selected in sample was $P_{\max}^{-1} = Z/Z_{\max}$. This was because the probability of its inclusion in sample, $nP_{\max}$, was not allowed to exceed unity. For four of the seven sampling procedures considered in this Chapter (Procedures 4, 25, 45 and 46) this limit is not relevant and the number of units in sample can be set at any value up to N itself. For the Poisson sampling group (Procedures 27, 38 and 39) the upper limit to the expected number of sample units remains $P_{\max}^{-1}$.

4.5. SIMPLICITY IN SELECTION PROCEDURE

It was mentioned in Chapter 3 that systematic procedures have an obvious advantage over all other procedures in simplicity of selection procedure. The Raj and Murthy sample schemes use a selection procedure which is hardly more complicated than systematic selection. The RHC Procedure 25 involves the formation of n random groups. It is therefore slightly more tedious than that of the Raj and Murthy schemes but perhaps slightly easier to apply than the rejective Rao-Sampford Procedure 11.

Poisson Sampling uses a series of N binomial trials to determine whether each population unit is to be included in sample or not. Although this is more tedious than the procedures mentioned above, it is appreciably simpler to use than those which involve iteration. Collocated sampling is not feasible without a computer, at least for the comparatively large populations for which it was devised.

The Ikeda-Midzuno Procedure 46 appears to be somewhat less cumbersome than Lahiri's Procedure 45 if a sample selected with probability proportional to aggregate size is desired.

4.6. SIMPLICITY IN VARIANCE ESTIMATION

In Chapter 3 it was pointed out that for procedures using the Horvitz-Thompson estimator, the problem of estimating variance was virtually identical with the problem of determining the π_{IJ} . In consequence it was the case for most, though not all, procedures that the simplicity of the variance estimation procedure was directly related to the simplicity of the selection procedure.

For the estimators discussed in this Chapter the variance estimation formulae for these procedures have already been set out in equations (4.2.5), (4.2.10), (4.2.19), (4.2.24), (4.2.27), (4.2.35) and (4.2.37). It will be seen that Raj's estimator, the RHC estimator and the estimators for Poisson Sampling all have quite simple variance estimators for any n . The same is true for collocated sampling provided the approximate formula (4.2.40) is used. The Murthy variance estimator is simple for $n = 2$ but becomes rapidly more complicated as n increases.

It has already been mentioned that $v_{a2}(y'')$ is simpler than $v_{a1}(y'')$ for estimating the variance of Lahiri's estimator.

4.7. EFFICIENCY OF ESTIMATOR OF TOTAL

In this Section the efficiency of the various procedures will be considered empirically and semi-empirically using the model (1.8.5).

[illegible] facilitate comparison with the formulae relevant to the use of the Horvitz-Thompson estimator with exact selection procedures, the symbols π_I, π_J, and so on, will be used to denote nP_I, nP_J and so on. Note that there are *not* the probabilities of inclusion in the sample for either the Raj-Murthy or for the Rao-Hartley-Cochran Procedures although they do have this meaning in the Poisson sampling group.

4.7.1 The Raj and Murthy Estimators

Writing π_I for $2P_I$ in (4.2.3), (4.2.4), (4.2.8) and (4.2.9), we have for $n = 2$:

$$t_{\text{mean}} = \tfrac{1}{2}\left[\left(2+\pi_1\right) \frac{y_1}{\pi_1} + \left(2-\pi_1\right) \frac{y_2}{\pi_2}\right] , \tag{4.7.1}$$

$$V\left(t_{\text{mean}}\right) = \tfrac{1}{8} \sum\sum_{\substack{I,J=1\\J\neq I}}^{N} \pi_I\pi_J\left(n-\pi_I-\pi_J\right)\left(\frac{Y_I}{\pi_I} - \frac{Y_J}{\pi_J}\right)^2 , \tag{4.7.2}$$

$$t_{\text{symm}} = \frac{2}{4-\pi_1-\pi_2} \left[\left(2-\pi_2\right) \frac{y_1}{\pi_1} + \left(2-\pi_1\right) \frac{y_2}{\pi_2}\right] , \tag{4.7.3}$$

and

$$V\left(t_{\text{symm}}\right) = \tfrac{1}{2} \sum\sum_{\substack{I,J=1\\J\neq I}}^{N} \pi_I\pi_J \frac{2-\pi_I-\pi_J}{4-\pi_I-\pi_J} \left(\frac{Y_I}{\pi_I} - \frac{Y_J}{\pi_J}\right)^2 . \tag{4.7.4}$$

Rao and Bayless (1969) used the model (1.8.5) and obtained the expected variances of the estimators t_{mean} and t_{symm}, from (4.7.2) and (4.7.4), *viz.*

$$E^*\left[V\left(t_{\text{mean}}\right)\right] = \frac{\sigma^2}{4} (Z/2)^{2\gamma} \sum\sum_{\substack{I,J=1\\J\neq I}}^{N} \pi_I^{2\gamma-1}\pi_J\left(4-\pi_I-\pi_J\right) \tag{4.7.5}$$

and

$$E^*\left[V\left(t_{\text{symm}}\right)\right] = \sigma^2(Z/2)^{2\gamma} \sum\sum_{\substack{I,J=1\\J\neq I}}^{N} \pi_I^{2\gamma-1}\pi_J \frac{2-\pi_I-\pi_J}{4-\pi_I-\pi_J} . \tag{4.7.6}$$

Hanurav (1966b) and Vijayan (1966) compared the relative efficiencies under the model (1.8.5) of the Horvitz-Thompson estimator, y'_{HT} and t_{symm}. They proved that the Horvitz-Thompson estimator was more efficient than t_{symm} for $\gamma = 1$. Rao (1966b) further proved that y'_{HT} was more efficient than y'_{HH} for all values of γ. Rao (1966b) and Vijayan (1966) also proved that the t_{symm} was better than y'_{HT} for $\gamma = 0.5$.

Extensive empirical and semi-empirical studies were carried out by Rao and Bayless (1969) for the case $n = 2$, and by Bayless and Rao (1970) for the cases $n = 3$ and $n = 4$. In their empirical studies they found that Murthy's estimator was nearly always more efficient than the Horvitz-Thompson estimator, except in certain artificial populations. In their semi-empirical studies of the case $n = 2$, the values of γ which they used were 0.5, 0.75, 0.875, and 1.0. For all these values of γ, Murthy's procedure was consistently more efficient than Raj's procedure. Raj's estimator was usually more efficient than the Horvitz-Thompson estimator for $\gamma = 0.5$ and usually less efficient for $\gamma > 0.5$. Murthy's estimator was more efficient than the Horvitz-Thompson estimator for $\gamma \leq 0.875$ and less efficient for $\gamma = 1.0$. For $\gamma = 0.875$ Murthy's estimator was nearly always the more efficient but the difference was very small.

For $n = 3$ and 4 Bayless and Rao investigated the cases $\gamma = 0.75$, 0.875 and 1.0 only. Raj's estimator was less efficient than the Horvitz-Thompson estimators in almost every case. Murthy's estimator was again more efficient than the Horvitz-Thompson estimator for $\gamma \leq 0.875$.

Few of the differences in efficiency between Murthy's estimator and the Horvitz-Thompson estimator for natural populations exceeded 10%. The same was true for the comparison of the Raj and Horvitz-Thompson estimators.

The close agreements between the empirical and the semi-empirical results of Rao and Bayless tend to suggest that the form of the linear stochastic model assumed by them is reasonably appropriate. However, Samiuddin *et al* (1978) studied the behaviour of t_{symm}, y'_{HT} and several other estimators with six semi-empirical and six artificial populations. The Horvitz-Thompson estimator was found to be reasonably efficient in all cases. Murthy's estimator was reasonably efficient for the semi-empirical populations but somewhat less satisfactory for the artificial ones.

4.7.2 The Rao-Hartley-Cochran Estimator

When N is a multiple of n, the Rao-Hartley-Cochran variance estimator attains the minimum value (4.2.18). The expected variance of the RHC estimator is (Rao and Bayless, 1969)

$$E^*V\left(y'_{\text{RHC}}\right) = \sigma^2 c_0 c_1 (Z/2)^{2\gamma} \sum_{I=1}^{N} \left(2-\pi_I\right)\pi_I^{2\gamma-1} , \tag{4.7.7}$$

where

$$c_0 = \frac{N_1^2+N_2^2-N}{N^2-N_1^2-N_2^2} \quad \text{and} \quad c_1 = \frac{N^2-N_1^2-N_2^2}{N(N-1)} .$$

A corresponding formula for $n > 2$ is given by Bayless and Rao (1970).

In [illegible] comparison of the relative efficiencies of various estimators under the [illegible] showed that the RHC estimator was less efficient than both Murthy's estimator and the Horvitz-Thompson estimator for $\gamma = 1$. The Horvitz-Thompson estimator was more, equally or less efficient than the RHC estimator according as γ was greater than, equal to, or less than 0.5 respectively. Further comparisons of the efficiency of the RHC estimator with that of the Horvitz-Thompson estimator are given by Pedgaonkar and Prabhu Ajgaonkar (1978). Pathak (1966) also proved that for large N the RHC estimator is less efficient than Murthy's estimator for $\gamma \geq 0.5$. Singh and Kishore (1975) showed that after taking expected cost into account the Hansen-Hurwitz estimator based on multinomial sampling was sometimes superior to the RHC estimator.

Rao and Bayless (1969) and Bayless and Rao (1970) in their empirical studies for $n = 2, 3$ and 4 concluded that the RHC estimator was consistently less efficient than Murthy's estimator, and that it was sometimes slightly more and sometimes slightly less efficient than the Horvitz-Thompson estimators.

In the semi-empirical studies carried out by the same authors, the RHC estimator was found to be consistently less efficient than both the Murthy and the Horvitz-Thompson estimators. Its efficiency vis-à-vis the Murthy estimator was not greatly affected by the value of γ , but vis-à-vis the Horvitz-Thompson estimators it was least efficient for $\gamma = 1$. As with Murthy's and Raj's estimators, most of the differences were only of the order of a few percent, except for $n = 4$ where differences of 20% and 30% were not uncommon.

4.7.3. Poisson and collocated sampling

Two empirical populations were used by Brewer, Early and Hanif (1980) to compare Poisson and collocated sampling with other unequal probability sampling strategies. The first of these was the population of 49 cities listed in Cochran (1963), p. 156, and the second that of 270 blocks listed in Kish (1965), p. 624. The Cochran population contained one exceptional unit with very low π_I and high ratio Y_I/π_I . The Kish population contained no such maverick.

The strategies compared were as follows:

(i) Sampling with replacement (that is multinomial sampling) with the Hansen-Hurwitz (1943) estimator.

(ii) Sampling without replacement (m fixed) with the Horvitz-Thompson (1952) estimator. For this strategy the asymptotic variance formula

$$V(y'_{\text{HT}}) \doteqdot \sum_{I=1}^{N} \pi_I\left(1 - \frac{n-1}{n}\pi_I\right)\left(\frac{Y_I}{\pi_I} - \frac{Y}{n}\right)^2 \qquad (1.8.4)$$

was used.

(iii) Poisson sampling with the unbiased estimator y'_{PS} .

(iv) Poisson sampling with the ratio estimator y''_{PS} .

(v) Collocated sampling with the unbiased estimator y'_{CS} .

(vi) Collocated sampling with the ratio estimator y''_{CS} .

For Poisson and collocated sampling, variances were calculated both excluding and including the terms P_0Y^2, $P_{0C}Y^2$, so as to indicate the importance of the non-zero probability of an empty sample.

For collocated sampling the mean square errors were calculated using

(a) the exact π_{IJ} values given in (4.2.38),

(b) the approximate π_{IJ} values given by (4.2.40),

(c) the approximate π_{IJ} values given by (4.2.40) wherever these exceeded zero, but otherwise replaced by zero.

In every case the probabilities of inclusion in sample were taken to be proportional to the Z-values supplied. The use of the approximate formula (4.2.40) for the π_{IJ} resulted in reasonable approximations for the variance and mean square error formulae for collocated sampling. The better of the two approximations was achieved when the negative values obtained from (4.2.40) are set equal to zero, but the advantage held only when n is small.

The results based on exact π_{IJ} values are given in Tables 4.1 and 4.2. Some highlights of these are as follows:

1. When the probability of an empty sample is small or zero, the mean square error of the ratio estimator for Poisson or collocated sampling is comparable with the variance of the Horvitz-Thompson estimator when m is fixed. (The calculations actually show the ratio estimator mean square error to be smaller, but this is due to the Taylor series approximation.)

2. When the probability of an empty sample is of the order of 0.003 or greater, the contribution to the variance from the empty sample term is too large to be ignored.

3. The probability of an empty sample is at least an order of magnitude smaller in collocated sampling than in Poisson sampling, and becomes exactly zero for large samples.

TABLE 4.1

Comparisons of Efficiencies for Different Strategies with Hansen-Hurwitz *ppswr* as Standard

Strategy	Cochran's Population N = 49			Kish's Population N = 270			
	n = 2	5	9	n = 2	10	20	30
Hansen-Hurwitz *ppswr*	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
Horvitz-Thompson *πpswor* (fixed sample size)	1.0037	1.0151	1.0307	1.0054	1.0511	1.1144	1.1859
Poisson							
- Unbiased	0.1812	0.2022	0.2391	0.0817	0.0877	0.0967	0.1078
- Ratio							
- Ignoring P_0Y^2	1.0075	1.0109	1.0346	1.0109	1.0571	1.1212	1.1935
- Including P_0Y^2	0.4534	0.9380	1.0344	0.2472	1.0532	1.1212	1.1935
Collocated							
- Unbiased	0.3127	0.3627	0.4300	0.1303	0.1440	0.1592	0.1776
- Ratio							
- Ignoring $P_{0C}Y^2$	1.0169	1.0295	1.0472	1.0744	1.1519	1.2375	1.3287
- Including $P_{0C}Y^2$	0.8745	1.0292	1.0472	0.5771	1.1519	1.2375	1.3287

TABLE 4.2

Comparisons of Probabilities of Empty Samples for Poisson and Collocated Sampling

Strategy	Cochran's Population, N = 49			Kish's Population, N = 270		
	n = 2	5	9	n = 2	10	20
Poisson, P_0	0.1237	0.3459×10^{-2}	0.4139×10^{-5}	0.1333	0.3055×10^{-4}	0.3666×10^{-9}
Collocated, P_{0C}	0.1633×10^{-1}	0.1155×10^{-4}	zero	0.3499×10^{-1}	0.2787×10^{-7}	0.9240×10^{-16}

NOTE: The probabilities of empty samples for the above tables have been calculated using the approximation (E.2) from Appendix E.

4.7.4. [illegible]ator

Exc[illegible] is very small, sampling with probability proportional to aggregate size approximates sampling with equal probabilities, Cochran (1953) showed that the conventional ratio estimator with equal probability sampling was more efficient than the Horvitz-Thompson estimator with *πpswor* for low values of γ with a break-even point close to $\gamma = \frac{1}{2}$. While much the same kind of conclusion was reached for the RHC estimator, the contrast here is much more severe. The RHC estimator closely resembles the Horvitz-Thompson estimator, while the conventional ratio estimator is entirely different. Similarly the RHC Procedure 25 is nearly an exact *πpswor* scheme, while Procedures 45 and 46 approximate equal probability sampling.

4.8. UNBIASEDNESS AND STABILITY OF VARIANCE ESTIMATORS

4.8.1. The Raj and Murthy Estimators

Rao and Bayless (1969) used the model (1.8.5) to find the stability of variance estimators (4.2.5) and (4.2.10). They had shown that the leading terms in the expected variances of (4.2.5) and (4.2.10) for $n = 2$ were

$$E^*\left[Ev^2\left(t_{\text{mean}}\right)\right] = \frac{3}{64}\,\sigma^4 (Z/2)^{4\gamma} \sum_{\substack{I,J=1\\ J\neq I}}^{N}\sum \pi_I \pi_J \left(2-\pi_I\right)^3 \left(\pi_I^{2\gamma-2}+\pi_I^{2\gamma-2}\right)^2 , \qquad (4.8.1)$$

and

$$E^*\left[Ev^2\left(t_{\text{symm}}\right)\right] = 3\sigma^4 (Z/2)^{4\gamma} \sum_{\substack{I,J=1\\ J\neq I}}^{N}\sum \frac{\pi_I\pi_J\left(2-\pi_I\right)\left(2-\pi_J\right)\left(2-\pi_I-\pi_J\right)}{\left(4-\pi_I-\pi_J\right)^3} \left(\pi_I^{2\gamma-2}+\pi_J^{2\gamma-2}\right)^2 . \qquad (4.8.2)$$

The leading term in the expected variance of (4.2.11) is also presented here:

$$E^*\left[Ev_M^2\left(t_{\text{symm}}\right)\right] = \frac{3}{128}\,\sigma^4 (Z/2)^{4\gamma} \sum_{\substack{I,J=1\\ J\ \ I}}^{N}\sum \pi_I\pi_J\left(4-\pi_I-\pi_J\right)\left(2-\pi_I\right)\left(2-\pi_J\right)\left(\pi_I^{2\gamma-2}+\pi_J^{2\gamma-2}\right)^2 . \qquad (4.8.3)$$

Rao and Bayless (1969) and Bayless and Rao (1970) made semi-empirical and empirical studies of the stabilities of variance estimators for $n = 2, 3$, and 4 . They concluded from their semi-empirical studies that Murthy's variance estimator was consistently more stable than the Sen-Yates-Grundy variance estimator. This was particularly the case for the smaller values of γ . Murthy's variance estimator also tended to be more stable than Raj's variance estimator, especially for the larger values of γ and of n .

In their empirical studies Rao and Bayless concluded that Raj's and Murthy's variance estimators were essentially equivalent in stability for $n = 2$, but that Murthy's was usually slightly more stable for $n = 4$. Both these variance estimato were almost always more stable than the Sen-Yates-Grundy variance estimator, and the gains were often appreciable.

4.8.2. The Rao-Hartley-Cochran Estimator

Rao and Bayless (1969) and Bayless and Rao (1970) used the linear stochastic model (1.8.5) to derive the expected variance of the RHC variance estimator. The formulae, which are extremely complicated, are given in their 1969 paper for $n = 2$ and in Appendix B of their 1970 paper for any n .

In their semi-empirical studies they also concluded that for $n = 2$, the RHC variance estimator was consistently more stable than the Raj, Murthy, and Sen-Yates-Grundy estimators for all values of γ ; however the gains over Murthy's variance estimator were not large. For $n = 3$ and 4 , the RHC variance estimator was still almost always more stable than the Murthy variance estimator for $\gamma = 0.875$, but for $\gamma = 1$ the reverse was the case. It was consistently more stable than the Sen-Yates-Grundy variance estimator for all values of γ . In their empirical studies they found that for all values of n considered, the RHC variance estimator was more stable than the Raj, Murthy, and Sen-Yates-Grundy variance estimators.

These special variance estimators are much more stable than the Sen-Yates-Grundy variance estimator, even when the joint probabilities of selection are chosen specifically to stabilise the latter. This result is consonant with Raj's own findings (1956a) and is also heuristically plausible in that the coefficients of $\left(\left(y_1/\pi_1\right)-\left(y_2/\pi_2\right)\right)^2$ for all these three variance estimators are usually close to and always less than unity, whereas for the Sen-Yates-Grundy variance estimator the coefficients are $\left\{\pi_1\pi_2\pi_{12}^{-1}-1\right\}$, which tend to be rather variable (see for instance Table 3.1).

4.8.3. Lahiri's Estimator

It has already been mentioned that $v_{a1}(y'')$ is appreciably more stable than $v_{a2}(y'')$ for estimating the variance of Lahiri's estimator.

4.9. ROTATABILITY

Raj's and Murthy's sample schemes are not appropriate for rotation except using the Alternative III mentioned in Section 3.8. Samples selected by the Rao, Hartley and Cochran Procedure 25 may be rotated using a slightly modified version of Alternative I. Since selection within each of the n groups occurs independently, each selection may be rotated around the population units allocated to that group, starting [illegible] point within the first unit selected (for the reason indicated in Section [illegible]

A method of Poisson sampling which allows for rotation and updating in a simple way was presented by Brewer, Early and Joyce (1972).

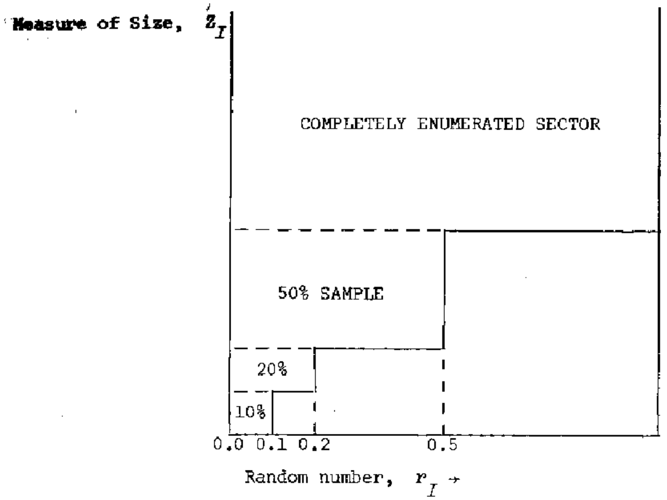

FIGURE 4.1. Stratified random Poisson sample.

Figure 4.1 shows how Poisson sampling works for a stratified random sample with three sampled strata and a completely enumerated sector. The units of the population correspond to points on the chart specified by r and Z . The sample consists of all points in the chart to the left of the thick line. Rotation can be effected by shifting the sample area to the right. If the shift is 0.02 in r , all units for which r is less than 0.02 are rotated out of sample and replaced by units for which r lies between $\pi(Z_I)$ and $\pi(Z_I) + 0.02$ where $\pi(Z_I)$ is the probability of inclusion in sample of a unit with size Z_I . This would give a 20% rotation in the lowest sampled stratum, 10% in the next, and 4% in the highest. The chart should be thought of as cylindrical, so that for the completely enumerated sector where $\pi(Z_I) = 1.00$, the new limit of $\pi(Z_I) + 0.02$ or 1.02 brings in again those units which would otherwise be rotated out, giving nil rotation in the completely enumerated sector.

Figure 4.1 can obviously be used to select other samples of various sizes with minimum or maximum overlap, and shows at a glance what is feasible and what is not feasible about, say, different rates of rotation for samples with minimum overlap.

Figure 4.2 illustrates two different ways of rotating a sample drawn with probabilities proportional to size (*πps*). The original sample is that of all points contained in the area *AOB* . A fixed proportion rotation, such as 20% , gives the new sample area as *A'OB'* . A usually preferable alternative is represented by *A"O"B"* . This gives fast rotation for small units and slow for large. Similar procedures may be used if the probability of selection is any function of size.

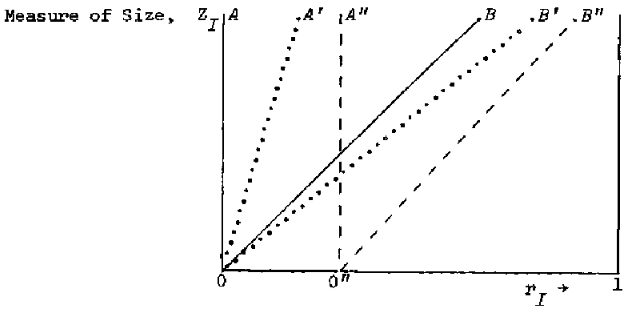

FIGURE 4.2. Two ways of rotating a *πps* Poisson sample.

A formal description of this method was presented by Brewer, Early and Hanif (1980). Choose an arbitrary fixed number c and (for all I) a uniformly distributed random number r_I in the interval $[0, 1)$. Then the Ith unit is in sample if $r_I < \max\{0, \pi_I-1+c\}$ or $c \le r_I < \min\{\pi_I+c, 1\}$; that is, the Ith unit is selected if $\{\pi_I, r_I\}$ lies in the shaded area in Figure 4.3. Since r_I is uniformly distributed over $[0, 1)$, the probability that the Ith unit is selected is $\max\{0, \pi_I-1+c\} + \min\{\pi_I+c, 1\} - c = \pi_I$ for all I , as required.

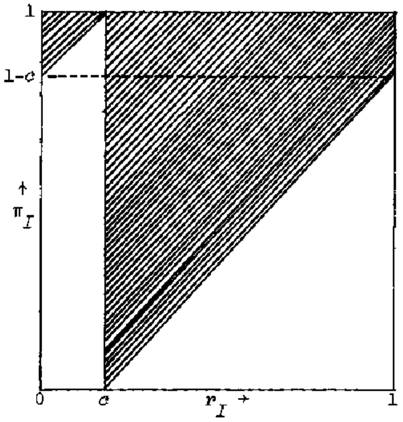

FIGURE 4.3. Diagram showing selection region.

[illegible] sample with (updated) values π_I' for each I , we can contr[illegible] overlap between the samples by varying the value of c . Suppose the va[illegible] altered for the revised sample to $c + d$ where $0 \le c+d < 1$ and $d > 0$ [illegible] Appendix C) the probability of inclusion of the Ith unit in both samples, $B_I(d)$, takes the values set out in the following table.

	$d \ge \pi_I$	$d \le \pi_I$
$d \le 1-\pi_I'$	0	$\min\{\pi_I', \pi_I-d\}$
$d \ge 1-\pi_I'$	$\min\{\pi_I, \pi_I'-1+d\}$	$\pi_I'+\pi_I-1$

(4.9.1)

Noticing that when $d \le \pi_I$ and $d \ge 1-\pi_I'$,

$$\pi_I' + \pi_I - 1 = \min(\pi_I, \pi_I'-1+d) + \min(\pi_I', \pi_I-d) ,$$

we have that

$$\sum_{I=1}^{N} B_I(d) = \sum_{\pi_I \ge d} \min(\pi_I', \pi_I-d) + \sum_{\pi_I' \ge 1-d} \min(\pi_I, \pi_I'-1+d) \qquad (4.9.2)$$

is the expected number of population units in the overlap of the samples, so that $n^{-1} \sum_I B_I(d)$ will be a measure of the expected proportion of original sample units retained. The maximum expected overlap is $\sum_I \min(\pi_I, \pi_I')$ which occurs if $d = 0, 1$. The minimum expected overlap will be approximated closely in practice by taking $d = \frac{1}{2}$, and its value will be close to $\sum_{\pi_I+\pi_I'>1} (\pi_I+\pi_I'-1)$ (see Appendix C).

These maximum and minimum overlap situations may be represented diagramatically as follows:

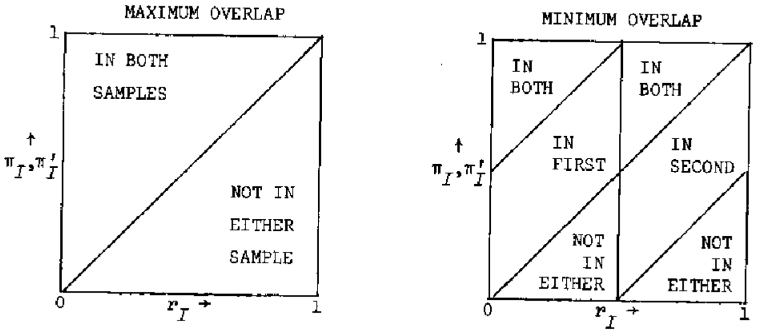

FIGURE 4.4. Diagrams showing maximum and minimum overlap.

An alternative method of achieving minimum overlap (when the π_I and π_I' are comparable) is illustrated in Figure 4.5.

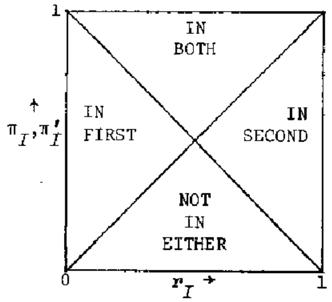

FIGURE 4.5. An alternative method of achieving minimum overlap.

This second method of achieving minimum overlap is more powerful than the first in situations where the π_I and π_I' are roughly proportional but differ substantially in magnitude. To simplify the discussion suppose that $\pi_I' = \frac{1}{2}\pi_I$ for all I for which π_I is definable, and that π_I' goes on increasing to a maximum of unity for large units which in the first survey lie in the completely enumerated (CE) sector. In this situation maximum and minimum overlaps are achieved by using the diagrams set out in Figure 4.6. The annotations in these diagrams indicate whether the regions are CE, 'IN' sample or 'OUT' of sample on each occasion.

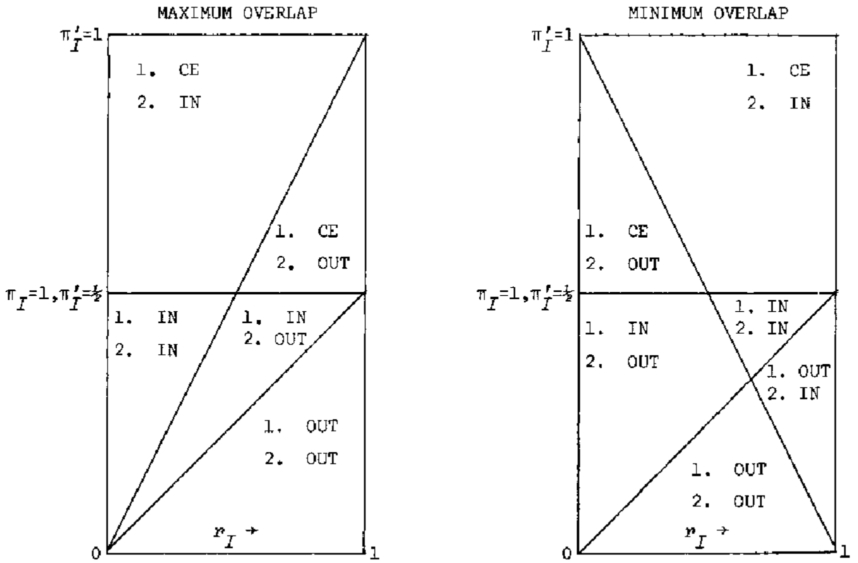

FIGURE 4.6. Maximum and minimum overlap where samples are of markedly different size.

Th[illegible] achieved in the control of rotation and respondent bur[illegible] obvious. They are applicable only to Poisson and coll[illegible] be suggested in Chapter 6 that these procedures are for this [illegible] appropriate for use in large scale surveys of businesses, farm[illegible], where units differ greatly in size and it is not unusual to [illegible] samples in use simultaneously which have been selected from the same population.

Another approach to achieving a desired overlap between successive Poisson samples is to stipulate a new inclusion probability for each population unit, conditional on whether or not that unit was included in the first sample. This Keyfitz (1951) inspired method was used by the US Bureau of the Census (Ogus and Clark, 1971) and yields the same overlap as the method first described, but does not share its simple control properties.

As already mentioned, the method of rotation and updating for controlling overlap between successive samples applies equally to collocated sampling. The only complication arises when units enter and leave the population between the selection of the original and revised samples. In these cases we must allow the values of r_I to change between samples. We can minimize this problem (improving on the procedure suggested by Brewer, Early, and Joyce, 1972) by defining a new ordering H to replace L as follows.

Assume that l new units have been added to the population and k original units deleted. There are two cases to consider.

Case 1: $l > k$

The population labels of the k deleted units are attached at random to k of the l new units. The remaining $l - k$ new units are assigned population labels $L_I = N+1, \ldots, N+l-k$. Choose, using simple random sampling without replacement, $l - k$ distinct integers $H_{N+1}, \ldots, H_{N+l-k}$ from the set $\{1, \ldots, N+l-k\}$. Then let the Jth largest of the remaining $N - k$ integers define the new label value H_I where I is the original population label of the surviving units with the Jth largest value of L_I .

Case 2: $l \leq k$

The population labels L_I of a random sample of l of the k deleted units are attached at random to the l new units. The remaining unattached $k - l$ population labels I are destroyed. Then let the integer J define the new label value H_I where I is the original population label of a new or surviving unit with the Jth largest value of L_I .

We have thus defined H_I for each unit in the new population, and the values of H are uniformly distributed over the set $\{1, \ldots, N+l-k\}$. For each I as above, set

$$t_I = \frac{\theta+H_I-1}{N+l-k} . \qquad (4.9.3)$$

Provided k and l are small in comparison to N we may assume $t_I \doteq r_I$ for any surviving unit, and thus use the result of Poisson sampling to control overlap.

Collocated sampling thus retains the most desirable properties of Poisson sampling (simplicity of selection and estimation of variance, and control over the sample). The π_{IJ} can be evaluated quite straightforwardly and are given in (4.2.40).

4.10. SUMMARY

In Tables 4.3 and 4.4, summaries of the properties of the procedures using special estimators are given for $n = 2$ and $n > 2$ respectively.

TABLE 4.3

Summary of Properties of the Procedures using Special Estimators, $n = 2$
(with Rao-Sampford Procedure for Comparison)

Procedure	Raj	Murthy	RHC	Lahiri	Rao-Sampford
Is number in sample fixed?	yes	yes	yes	yes	yes
Is estimator unbiased?	yes	yes	yes	yes	yes
Efficiency* $\gamma < \frac{1}{2}$	close to standard	better than standard	better than standard	much better than standard	standard
$\frac{1}{2} < \gamma < 1$	below standard	about standard	below standard	much below standard	standard
$\gamma = 1$	much below standard	below standard	much below standard	much below standard	standard
Is variance estimator unbiased?	yes	yes	yes	yes	yes
Stability of variance estimator*	good	very good	excellent	unknown	standard
Simplicity in selection	excellent	excellent	very good	excellent	good
Simplicity in variance estimation	excellent	excellent	excellent	excellent	good
Which alternatives may be used for rotation?	III	III	II[1], III	III	II[2], III

* The efficiency of the Horvitz-Thompson estimator and of the Sen-Yates-Grundy variance estimator have been taken as the standard in these parts of the table.

[1] If Alternative II is used, the procedure is to replace the selected unit within each group in turn by a new unit selected PPS within the same group.

[2] If Alternative II is used, oversampling is needed.

TABLE 4.4

Summary of Properties of the Procedures using Special Estimators, $n > 2$
(with Rao-Sampford Procedure for Comparison)

Procedure	Raj	Murthy	RHC	Poisson (Unbiased Estimator)	Poisson (Ratio Estimator)	Collocated (Unbiased)	Collocated (Ratio)	Lahiri	Rao-Sampford
Is number in sample fixed?	yes	yes	yes	no	no	no	no	yes	yes
Limit on (expected) number in sample	N	N	N	Z/Z_{max}	Z/Z_{max}	Z/Z_{max}	Z/Z_{max}	N	Z/Z_{max}
Is estimator unbiased?	yes	yes	yes	yes	nearly	yes	nearly	yes	yes
Efficiency* $\gamma < \frac{1}{2}$	close to standard	better than standard	better than standard	much below standard	close to standard	much below standard	close to standard	much better than standard	standard
$\frac{1}{2} < \gamma < 1$	below standard	about standard	below standard	much below standard	close to standard	much below standard	close to standard	much below standard	standard
$\gamma = 1$	much below standard	below standard	much below standard	much below standard	close to standard	much below standard	close to standard	much below standard	standard
Is variance estimator unbiased?	yes	yes	yes	yes	nearly	yes	nearly	yes	yes
Stability of variance estimator*	good	very good	excellent	unknown	unknown	unknown	unknown	unknown	standard
Simplicity in selection	excellent	excellent	very good	fair	fair	needs computer	needs computer	excellent	good
Simplicity in variance estimation	excellent	poor	excellent	excellent	good	good[4]	good[4]	excellent	good
Which alternatives may be used for rotation?	III	III	II[1], III	I[2]	I[2]	I[2]	I[2]	III	II[3]

* The efficiency of the Horvitz-Thompson Estimator and of the Sen-Yates-Grundy variance estimator have been taken as the standard "average" values in these parts of the table.

[1] If Alternative II is used, the procedure is to replace the selected unit with each group in turn by a new unit selected PPS within the same group with replacement.

[2] If Alternative I us used, rotation is best specified using diagrams such as Figures 4.1 and 4.2.

[3] If Alternative II is used, over sampling is needed.

[4] Using approximate expressions for joint inclusion probabilities.