Contributions to the theory of unequal probability sampling without replacement

Chaturvedula Asok
Iowa State University

Follow this and additional works at: https://lib.dr.iastate.edu/rtd
Part of the Statistics and Probability Commons

Recommended Citation
https://lib.dr.iastate.edu/rtd/6322

This Dissertation is brought to you for free and open access by the Iowa State University Capstones, Theses and Dissertations at Iowa State University Digital Repository. It has been accepted for inclusion in Retrospective Theses and Dissertations by an authorized administrator of Iowa State University Digital Repository. For more information, please contact digirep@iastate.edu.
INFORMATION TO USERS

This material was produced from a microfilm copy of the original document. While the most advanced technological means to photograph and reproduce this document have been used, the quality is heavily dependent upon the quality of the original submitted.

The following explanation of techniques is provided to help you understand markings or patterns which may appear on this reproduction.

1. The sign or “target” for pages apparently lacking from the document photographed is “Missing Page(s)”. If it was possible to obtain the missing page(s) or section, they are spliced into the film along with adjacent pages. This may have necessitated cutting thru an image and duplicating adjacent pages to insure you complete continuity.

2. When an image on the film is obliterated with a large round black mark, it is an indication that the photographer suspected that the copy may have moved during exposure and thus cause a blurred image. You will find a good image of the page in the adjacent frame.

3. When a map, drawing or chart, etc., was part of the material being photographed the photographer followed a definite method in “sectioning” the material. It is customary to begin photoing at the upper left hand corner of a large sheet and to continue photoing from left to right in equal sections with a small overlap. If necessary, sectioning is continued again — beginning below the first row and continuing on until complete.

4. The majority of users indicate that the textual content is of greatest value, however, a somewhat higher quality reproduction could be made from “photographs” if essential to the understanding of the dissertation. Silver prints of “photographs” may be ordered at additional charge by writing the Order Department, giving the catalog number, title, author and specific pages you wish reproduced.

5. PLEASE NOTE: Some pages may have indistinct print. Filmed as received.

Xerox University Microfilms
300 North Zeeb Road
Ann Arbor, Michigan 48106
ASOK, Chaturvedula, 1943-
CONTRIBUTIONS TO THE THEORY OF UNEQUAL PROBABILITY SAMPLING WITHOUT REPLACEMENT.

Iowa State University, Ph.D., 1974
Statistics

Xerox University Microfilms, Ann Arbor, Michigan 48106
Contributions to the theory of unequal probability sampling without replacement

by

Chaturvedula Asok

A Dissertation Submitted to the Graduate Faculty in Partial Fulfillment of The Requirements for the Degree of DOCTOR OF PHILOSOPHY

Department: Statistics
Major: Statistics (Survey Sampling)

Approved:

Signature was redacted for privacy.

In Charge of Major Work

Signature was redacted for privacy.

For the Major Department

Signature was redacted for privacy.

For the Graduate College

Iowa State University
Ames, Iowa

1974
TABLE OF CONTENTS

1. INTRODUCTION 1

2. COMPARATIVE STUDIES OF SOME I.P.P.S. SCHEMES 9
   2.1. Schemes for Samples of Size 2 9
   2.2. Some Sampling Schemes for Samples of Size $n^2$ 17
   2.3. Evaluation of the Approximate Expression for $P_{ij}$ of the Sampford's Procedure 27
   2.4. Hanurav's Procedure 62

3. RAO, HARTLEY AND COCHRAN'S PROCEDURE 66
   3.1. Introduction 66
   3.2. Rao, Hartley and Cochran's Procedure 67
   3.3. Inadmissibility of the R.H.C. Estimator 69
   3.4. Horvitz-Thompson Estimator under Rao, Hartley and Cochran Scheme 74
   3.5. Estimation of the Variance 108
   3.6. Comparison with Other Estimators 109
   3.7. Horvitz-Thompson Type Estimator under R.H.C. Scheme 115
   3.8. Numerical Illustration 119
   3.9. Rao, Hartley, Cochran Scheme with Revised Probabilities 122
   3.10. The Improved R.H.C. Estimator 130
   3.11. Improved R.H.C. Estimator for the Case $n=2$ and $N$ Large 136
4. SOME RANDOMIZED VARYING PROBABILITY SCHEMES

4.1. An Exact Sampling Scheme for Sample Size 2

4.2. An Alternative Exact Sampling Scheme Utilizing the Durbin's Procedure

4.3. Role of Random Stratification in Getting Improved Estimates

4.4. Randomized Three Stage Procedure with Durbin's Scheme

4.5. Randomized m-stage Procedure with Durbin's Scheme

5. MISCELLANEOUS TOPICS IN UNEQUAL PROBABILITY SAMPLING

5.1. Model Comparisons of Some Existing Schemes

5.2. Alternative Use of Ancillary Information

6. BIBLIOGRAPHY

7. ACKNOWLEDGMENTS
1. INTRODUCTION

In modern civilization sample survey has come to be considered as an organized fact-finding instrument. Its importance lies in the fact that it can be used to summarize, for the guidance of administration, facts which would be otherwise inaccessible owing to the remoteness and obscurity of the units concerned, or their numerousness. In a scientifically designed sample survey, it is possible to draw valid conclusions from the sample with the help of the available probability theory and statistical inference. Thus it is an interesting fact that the results from a well planned sample survey are expected to be more accurate than those from a complete census, if one such is at all possible to be taken. The technical problems that should receive most careful consideration in planning a sample survey are the manner of selecting the sample and the estimation of population characteristics along with their margin of uncertainty. Since with every sampling and estimation procedure is associated the cost of the survey and the precision of the estimate made, the survey statistician dealing with the problems in the real world must take a very practical attitude in the selection of the procedure and choose a procedure which gives highest precision for a given cost of the survey or the minimum cost for a specified level of precision. As
such it may not be worthwhile and practicable to use some of the theoretically refined results. In large scale surveys, sums and sums of squares may be the only quantities that could possibly be calculated, and thus an estimator with a larger variance but which is cheaper to handle may be preferred to another which requires complicated computations but has a slightly smaller variance. Thus the survey statistician must strike a balance by taking all such facts into consideration and make his own decision regarding the selection of the sampling design in a given situation.

Let a finite population consist of \( N \) distinct units \( U_1, U_2, \ldots, U_N \) with associated values \( Y_t, t = 1, 2, \ldots, N \), of the characteristic \( y \) under study, and consider the problem of estimating the population mean \( \bar{Y} = \frac{1}{N} \sum_{i=1}^{N} Y_i \) or the population total \( Y = \sum_{i=1}^{N} Y_i \) based on a sample of size \( n \) drawn from this population. When data on an ancillary characteristic, say \( x \), which is highly correlated with \( y \), are available for all the units of the population, and \( X_t > 0 \) for \( t = 1, 2, \ldots, N \), it is customary to use this knowledge to provide a more efficient estimate of \( Y \), either by sampling with unequal probabilities or by using a ratio or regression method of estimation after equal probability sampling.

To use the ancillary data in selecting the sample, one simple and straightforward way is to calculate \( p_t = \frac{X_t}{X} \) for
t = 1, 2, ... N where \( X = \sum_{i=1}^{N} X_i \) and then select the units with replacement, the probability of selecting the t-th unit being \( p_t \) at each draw. This method of sampling is called the probability proportional to size (p.p.s.) sampling with replacement. The customary unbiased estimator of the population total under this sampling procedure is

\[
\hat{Y}_{pps} = \frac{1}{n} \sum_{i=1}^{n} \frac{y_i}{p_i} \quad (1.1)
\]

with variance

\[
V(\hat{Y}_{pps}) = \frac{1}{n} \sum_{i=1}^{N} p_i \left( \frac{y_i}{p_i} - \bar{y} \right)^2 \quad (1.2)
\]

Since in the case of simple random sampling, a sample selected with replacement yields a less precise estimate than a sample selected without replacement, it is quite natural to expect similar gains in unequal probability sampling also by shifting to without replacement schemes. Even though this approach under certain conditions gives easily calculated and unbiased estimators of \( Y \), it has the disadvantage that sampling itself may be difficult to carry out and the variances difficult to estimate.

For any given sampling design, Horvitz and Thompson (1952) proposed an unbiased estimator of the population total \( y \), viz.,

\[
\hat{Y}_{HT} = \sum_{i=1}^{n} \frac{y_i}{p_i} \quad (1.3)
\]
where $P_i$ is the probability for the $i$-th unit to be in the sample. In this dissertation, we will be mainly concerned with this estimator in view of its optimal properties established in the literature. Among the several articles in this line mention may be made of Godambe (1955, 1960), Godambe and Joshi (1965), Hájek (1959), and Hanurav (1968).

The variance of the Horvitz-Thompson (H.T.) estimator $\hat{y}_{HT}$ is given by

$$
V(\hat{y}_{HT}) = \sum_i \frac{y_i^2}{P_i} + \sum_{i<j} \frac{P_{ij}}{P_i P_j} y_i y_j - Y^2
$$

(1.4)

where $P_{ij}$ denotes the probability for the $i$-th and $j$-th units to be both in the sample.

From (1.4) one can observe that $V(\hat{y}_{HT})$ reduces to zero when $P_i$ is exactly proportional to $Y_i$, which suggests that by making $P_i$ proportional to $X_i$, considerable reduction in the variance can be achieved if $X_i$ are approximately proportional to $Y_i$. A host of authors have proposed schemes wherein the inclusion probability $P_i$ in a sample is $n p_i$ which imposes the condition $n p_i \leq 1$ on the probabilities $p_i$ which is not a severe one. Such schemes are termed in the literature as inclusion probability proportional to size (I.P.P.S.) schemes or IIPS schemes or exact sampling schemes. The different procedures can be put in four different categories depending upon the manner in which the requirement $P_i = n p_i$ is achieved.
In the first category are the schemes suggested by Durbin (1967) and Sampford (1967) where the first unit is selected with probability \( p_i \) while the subsequent units are selected with unequal probabilities so as to make \( P_i \) equal \( np_i \) for all \( i \). In the second category are the schemes suggested by Midzuno (1952), Lahiri (1951), Narain (1951), Yates and Grundy (1953), Brewer and Undy (1962), and Fellegi (1963). These schemes are based on unit by unit selection with revised probabilities of selection \( p_i' \), \( i = 1,2,\ldots, N \), so calculated that \( \sum_{i=1}^{N} p_i' = 1 \) and the inclusion probability \( P_i \) equals \( np_i \) for all \( i \). In the third category are the schemes suggested by Durbin (1953), Hájek (1964), Sampford (1967), and Hanurav (1967) based on rejective sampling. The units are selected with certain probabilities and with replacement, and the sample is rejected if all the units in the sample are not distinct, otherwise it is accepted. In the fourth category are the schemes suggested by Madow (1949) and Goodman and Kish (1950) where the units are selected in a systematic manner.

Another group of procedures is the pps without replacement sampling procedures. Those suggested by Midzuno (1952), Lahiri (1951) and Horvitz and Thompson (1952) belong to this group. In these procedures the first unit is selected with probability \( p_i \) while the subsequent units are selected with probabilities proportional to \( p_i \) or with equal probabilities.
In spite of so many sampling without replacement procedures being available, none of them has received general acceptance from the point of view of adoption in surveys. The reasons are not far to seek. Most of the authors presented schemes for samples of size two only and have nothing to offer for samples of size greater than two. The methods often lack simplicity, and the algebraic expressions for estimated variance and sometimes even for the estimator itself are complicated and unmanageable for sample size greater than two. Some of the procedures are less efficient than even sampling with replacement. At times, they may involve calculation of revised probabilities of selection which impose restrictive conditions on the initial set of probabilities, or the revised probabilities of selection cannot be obtained easily in practice. These difficulties will get multiplied with increasing sample size. Further, even among the existing schemes, practically nothing is known regarding the relative performance of different schemes as measured by the variances of the estimators proposed.

The I.P.P.S. schemes that are applicable for sample size $n>2$ are those of Midzuno (1952), Goodman and Kish (1950), Sampford (1967) and Hanurav (1967). In Chapter 2 we have established that the H.T. estimator corresponding to the Midzuno scheme has uniformly smaller variance than the customary with replacement estimator for arbitrary sample size, thus
generalizing the result due to Rao (1963a). Also we have compared the variances corresponding to the procedures of Goodman and Kish, Sampford, and Hanurav using the asymptotic approach of Hartley and Rao (1962).

In order to avoid the mathematical complications and the computational difficulties involved in these procedures, Rao, Hartley and Cochran (1962) suggested an ingenious device of selecting a sample of size $n$ with unequal probabilities and without replacement. However, the simplicity of their approach is invalidated by the fact that the estimator they propose is inefficient compared to the H.T. estimator corresponding to most of the I.P.P.S. schemes. In Chapter 3 we have discussed the inadmissibility of the Rao, Hartley and Cochran estimator and brought out the optimal properties of their scheme by suggesting alternate more efficient estimators.

None of the procedures proposed in the literature, owing to the complications involved, are acceptable for use in large scale surveys. In this connection it is worthwhile to quote Durbin (1953, p. 267). He says:

The strict application of the usual methods of unequal probability sampling without replacement, including the calculation of unbiased estimates of sampling error, is out of the question in certain kinds of large-scale survey work on grounds of practicability. There is therefore a need for methods which retain the advantages of unequal probability sampling without replacement but are rather easier to apply in practice and only involve a slight loss of exactness.
In Chapter 4 we have proposed an I.P.P.S. sampling procedure for sample sizes greater than two, that is particularly useful in large scale surveys, and which makes use of the Durbin's procedure (1967) for sample size 2 and established its efficiency in relation to the other existing schemes. We believe the same technique can be used with gain by using any other I.P.P.S. procedure for sample size 2 in place of the Durbin's procedure.

For comparing the efficiencies of various estimators in unequal probability sampling, a super population model is made use of by several authors. However, the average variance is the same for all the I.P.P.S. schemes under this model. In Chapter 5 we have considered a slightly different model and compared the efficiencies of various I.P.P.S. schemes under various a priori distributions of the auxiliary variable. Also we have proposed a new technique of using the ancillary information at the designing stage which is particularly useful in the case of area sampling and cluster sampling and have demonstrated that the estimator proposed under this scheme is always more efficient than the Rao, Hartley and Cochran's estimator.
2. COMPARATIVE STUDIES OF SOME I.P.P.S. SCHEMES

2.1. Schemes for Samples of Size 2

Several authors have proposed schemes for selecting two units from a population of size $N$, with unequal probabilities and without replacement, such that the overall probability of including the $i$-th unit in the sample is proportional to the known size $X_i$ of the $i$-th unit, i.e., $P_i = 2p_i \leq 1$, where $p_i = X_i/X$, $X$ being the total of all the $x$ values in the population. In this section we will discuss the desirable features of some of the schemes that are existent in the literature.

**Theorem 2.1:**

For the scheme of selecting a sample of size two wherein the first unit is selected with probability proportional to the revised sizes $X_j'$ and the second unit with probabilities proportional to the remaining original sizes $X_j$ where the revised sizes $X_j'$ are given by

$$p_j' = x_j'/X = \frac{2p_j(1-p_j)}{(1-2p_j)} \cdot \frac{1}{1 + \sum_1 N \frac{p_t}{(1-2p_t)}}, \quad (2.1.1)$$

the inclusion probabilities $P_i$ and $P_{ij}$ are given by

$$P_i = 2p_i \quad (2.1.2)$$

and
Proof:

The probability $P_i$ of including the $i$-th unit in the sample is given by

$$P_i = \text{prob (i-th unit gets selected at the first draw)} + \text{prob (i-th unit gets selected at the second draw)}$$

$$= \pi_i' + \sum_{j \neq i} \pi_j \frac{\pi_i}{1 - \pi_j}$$

$$= \pi_i' + \pi_i \left[ \sum_{l=1}^{N} \frac{p_t}{(1 - p_t)} - \frac{\pi_i}{(1 - \pi_i)} \right]$$

$$= \frac{2\pi_i(1 - \pi_i)}{(1 - 2\pi_i)} + \pi_i \left[ \sum_{l=1}^{N} \frac{2p_t}{1 - 2p_t} - \frac{2\pi_i}{1 - 2\pi_i} \right]$$

$$= \frac{1}{1 + \sum_{l=1}^{N} p_t/(1 - 2p_t)}$$

$$= 2\pi_i$$

Probability $P_{ij}$ of including the pair $(i, j)$ of units in the sample is given by

$$P_{ij} = \text{Prob (i-th unit gets selected at the first draw and j-th unit gets selected at the second draw)} + \text{Prob (j-th unit gets selected at the first draw and i-th unit gets selected at the second draw)}$$
\[
= \frac{p_i' \cdot p_j}{(1-p_i)} + \frac{p_j' \cdot p_i}{(1-p_j)} \\
= \frac{2p_i' \cdot p_j}{(1-2p_i)} + \frac{2p_j \cdot p_i'}{(1-2p_j)} \cdot \frac{1}{1 + \sum_{l=1}^{N} \frac{p_l}{(1-2p_l)}} \\
= \frac{2p_i'p_j}{N} \cdot \frac{1}{1 + \sum_{l=1}^{N} \frac{p_l}{(1-2p_l)}}
\]

Q.E.D.

We will call the sampling scheme described in Theorem 2.1 as Scheme A. This scheme is due to Brewer (1963), and the expression (2.1.3) is derived by Rao (1965).

**Theorem 2.2:**

Consider the sampling scheme described as follows: two units are selected with replacement, one with probabilities proportional to the revised sizes \(x_j^*\) and the other unit with probabilities proportional to the original sizes \(x_j\). If the two units selected are identical, reject the selections and repeat the process until two different units are selected in the sample. The revised sizes \(x_j^*\) are given by

\[
p_j^* = \frac{x_j^*}{x} = \frac{p_j/(1-2p_j)}{\sum_{l=1}^{N} \frac{p_l}{(1-2p_l)}} \tag{2.1.4}
\]

For this scheme also the inclusion probabilities \(P_i\) and \(P_{ij}\) are given by (2.1.2) and (2.1.3) respectively.
Proof:

It is easy to see that the probability $P_i$ of including the $i$-th unit in the sample is given by

$$P_i = \frac{p_i^* \cdot \sum_{j \neq i} p_j + p_i \cdot \sum_{j \neq i} p_j^*}{N \left(1 - \sum_{l=1}^{N} p_l^* \cdot p_l \right)}, \quad (2.1.5)$$

while the probability $P_{ij}$ of including the pair $(i,j)$ of units in the sample is given by

$$P_{ij} = \frac{p_i^* p_j + p_j^* p_i}{N \left(1 - \sum_{l=1}^{N} p_l^* p_l \right)}, \quad (2.1.6)$$

Substituting the values of $p_i^*$ and $p_j^*$ in (2.1.5) and (2.1.6) we obtain (2.1.2) and (2.1.3). Q.E.D.

We will call the sampling scheme described in Theorem 2.2 as Scheme B. This scheme is due to J.N.K. Rao (1965).

Theorem 2.3:

For the scheme of sampling where the first unit is drawn with probabilities $p_i$ and the second unit from the rest of the population units with probabilities

$$P_{j,i} = \frac{p_j \left(\frac{1}{1-2p_i} + \frac{1}{1-2p_j}\right)}{N \left(1 + \sum_{l=1}^{N} p_l/(1-2p_l)\right)}, \quad (2.1.7)$$

the inclusion probabilities $P_i$ and $P_{ij}$ are given by Equations
(2.1.2) and (2.1.3) respectively.

Proof:

Probability $P_i$ of including the $i$-th unit in the sample is given by

$$P_i = p_i + \sum_{j \neq i}^{N} p_j \cdot p_{i,j} \tag{2.1.8}$$

Substituting from (2.1.7), it can be seen that

$$P_i = 2p_i$$

The inclusion probability $P_{ij}$ is given by

$$P_{ij} = p_i \cdot p_{j,i} + p_j \cdot p_{i,j}$$

$$= \frac{2p_ip_j[\frac{1}{1-2p_i} + \frac{1}{1-2p_j}]}{1 + \frac{N}{1} \cdot p_t/(1-2p_t)}$$

Q.E.D.

We will call the sampling scheme described in Theorem 2.3 as Scheme C. The scheme and the above results are due to Durbin (1967).

Theorem 2.4:

The Horvitz-Thompson estimators, of the population total, $\hat{Y}_A$, $\hat{Y}_B$ and $\hat{Y}_C$ corresponding to the Schemes A, B and C respectively are equally efficient.
Proof:

For any sampling design, the variance of the corresponding Horvitz-Thompson estimator is given by

\[ V(\hat{y}_{\text{H.T.}}) = \frac{N}{\sum_{i=1}^{N} P_i^2} + \sum_{i=1}^{N} \sum_{j \neq i} \frac{P_{ij}}{P_i P_j} \cdot Y_i Y_j - Y^2 \]  

(2.1.9)

Since the expressions for \( P_i \) and \( P_{ij} \) of each of the Schemes A, B and C are given by (2.1.2) and (2.1.3), it follows that the corresponding variances are equal, and thus the estimators are equally efficient.  

Q.E.D.

Theorem 2.5:

The Horvitz-Thompson estimator corresponding to any of the Schemes A, B and C is always more efficient than the customary estimator in the case of probability proportional to size with replacement, and hence the Yates and Grundy estimate of variance for the Schemes A, B, and C is always non-negative.

Proof:

Variance of the customary probability proportional to size with replacement estimator \( \hat{y}_{\text{p.p.s.}} = \frac{1}{2} \frac{\sum Y_i}{P_i} \) is given by

\[ V(\hat{y}_{\text{p.p.s.}}) = \frac{1}{2} \left( \sum \frac{Y_i^2}{P_i} - Y^2 \right) \]  

(2.1.10)

Substituting the values of \( P_i \) and \( P_{ij} \) from (2.1.2) and
(2.1.3) in \( \sum_{ij} \frac{P_{ij}}{P_{i}P_{j}} Y_{i}Y_{j} \) we get

\[
\sum_{ij} \frac{P_{ij}}{P_{i}P_{j}} \cdot Y_{i}Y_{j} = \frac{1}{N} \cdot \frac{1}{2(1+\Sigma p_{t}/(1-2p_{t}))}
\]

Noting that \( 1 + \sum \frac{N}{1-2p_{t}} \cdot \frac{p_{t}}{1-2p_{t}} = 2 \sum \frac{N}{1-2p_{t}} \cdot \frac{p_{t}}{1-2p_{t}} \) and using (2.1.11), (2.1.9) becomes

\[
V(\hat{\theta}_{H.T.}) = \frac{1}{2} \sum \frac{N}{1-2p_{t}} \cdot \frac{Y_{t}^{2}}{p_{t}} + \frac{1}{1+\sum \frac{p_{t}}{p_{t}/(1-2p_{t})}} \cdot \left[ Y_{t} \cdot \sum \frac{N}{1-2p_{t}} \cdot \frac{Y_{t}}{1} - \sum \frac{N}{1-2p_{t}} \cdot \frac{Y_{t}^{2}}{1} - Y^{2} \right]
\]

\[
= \frac{1}{2} \sum \frac{N}{1-2p_{t}} \cdot \frac{Y_{t}^{2}}{p_{t}} - Y^{2} \right] + \frac{1}{1+\sum \frac{p_{t}}{p_{t}/(1-2p_{t})}} \cdot \left[ Y_{t} \cdot \sum \frac{N}{1-2p_{t}} \cdot \frac{Y_{t}}{1} - \sum \frac{N}{1-2p_{t}} \cdot \frac{Y_{t}^{2}}{1} - Y^{2} \cdot \sum \frac{N}{1-2p_{t}} \cdot \frac{p_{t}(1-p_{t})}{1} \right]
\]

Thus we have from (2.1.10) and (2.1.12),
Now it can be easily seen that,

\[ \frac{N p_t (1-p_t)}{1-2p_t} = 1 + \frac{N p_t^2}{1-2p_t} \]

and

\[ \sum_{t=1}^{N} \frac{Y_t}{1-2p_t} = Y + 2 \sum_{t=1}^{N} \frac{p_t}{1-2p_t} \cdot Y_t \]

substituting these values in (2.1.13) we get

\[
V(\hat{\Theta}_{pps}) - V(\hat{\Theta}_{H.T.}) = \frac{1}{N} \cdot \left[ Y^2 \cdot \frac{N p_t (1-p_t)}{1-2p_t} \right]
\]

\[ - Y \cdot \left\{ Y + 2 \sum_{t=1}^{N} \frac{p_t}{1-2p_t} \cdot Y_t \right\} 
+ \sum_{t=1}^{N} \frac{Y_t^2}{1-2p_t} \]

\[ = \frac{1}{N} \cdot \left[ \frac{N p_t^2}{1-2p_t} \cdot \left( \frac{Y_t}{p_t} - Y \right)^2 \right] \]

\[ \geq 0 \]

Thus the H.T. estimator corresponding to either of the Schemes A, B and C is always more efficient than the customary probability proportional to size with replacement estimator.
Since a necessary condition for the without replacement H.T. estimator \( \hat{Y}_{\text{H.T.}} \) for sample size 2 with \( P_i = 2p_i \) to be better than the customary with replacement estimator

\[ \hat{Y}_{\text{pps}} = \frac{1}{2} \frac{\sum Y_t}{P_t} \]

independently of the \( Y_t \)'s is

\[ P_{ij} = P_i P_j, \]

it follows from the above that the Yates and Grundy variance estimator of \( \hat{Y}_{\text{H.T.}} \) for either of the Schemes A, B and C, viz.,

\[ V(\hat{Y}_{\text{H.T.}}) = \frac{P_i P_j - P_{ij}}{P_{ij}} \cdot \left( \frac{Y_i}{P_i} - \frac{Y_j}{P_j} \right)^2 \]  

is always nonnegative. Q.E.D.

2.2. Some Sampling Schemes for Samples of Size \( n > 2 \)

Even though several authors have proposed schemes for sample size two that satisfy the condition \( P_i = 2p_i \leq 1 \), not many of these are useful for generalizing to samples of size \( n > 2 \). The reasons are not far to seek. Often the methods lack simplicity and the algebraic expressions for estimated variance and sometimes even for the estimator itself are complicated and unmanageable. Some of the procedures are less efficient than even sampling with replacement. At times, they may involve calculation of revised probabilities of selection which impose restrictive conditions on the initial set of probabilities or the revised probabilities of selection.
cannot be obtained easily in practice. In this section we will discuss the properties of some of the schemes that are existent in the literature.

2.2.1. **Midzuno scheme with revised probabilities**

Midzuno (1952) has proposed the following scheme for samples of size \( n \geq 2 \).

The first unit is selected with probability proportional to size \( X_i \) and the remaining \((n-1)\) units are selected with equal probabilities and without replacement.

For this scheme of sampling the expressions for \( P_i \) and \( P_{ij} \) are given by

\[
P_i = P_i + (1-P_i) \cdot \frac{n-1}{N-1} \tag{2.2.1}
\]

and

\[
P_{ij} = \frac{(n-1)}{(N-1)} \cdot \left[ \frac{(N-n)}{(N-2)} \cdot (p_i + p_j) + \frac{(n-2)}{(N-2)} \right] \tag{2.2.2}
\]

Horvitz and Thompson (1952) suggested using revised probabilities \( p_i^* \) to make \( P_i \) to be exactly equal to \( np_i \). The revised probabilities \( p_i^* \) are given by the equation

\[
np_i = P_i = p_i^* + (1-p_i^*) \cdot \frac{n-1}{N-1} \tag{2.2.3}
\]

or

\[
p_i^* = \frac{N-1}{N-n} \cdot np_i - \frac{n-1}{N-n} \tag{2.2.4}
\]

This imposes a severe restriction on \( p_i \), viz., \( p_i \geq \frac{n-1}{n} \cdot \frac{1}{N-1} \) since for \( p_i < \frac{n-1}{n(N-1)} \), \( p_i^* \) becomes negative.
Thus a necessary condition for the Midzuno scheme with revised probabilities to be applicable is

\[ p_i > \frac{(n-1)}{n} \cdot \frac{1}{N-1} \]  

(2.2.5)

For samples of size two, J.N.K. Rao (1963a) has shown that the H.T. estimator under the Midzuno scheme with revised probabilities is always more efficient than the customary p.p.s with replacement estimator. Here we will present a proof of the same for arbitrary sample size \( n \geq 2 \).

**Theorem 2.6:**

For the Midzuno scheme with revised probabilities, for arbitrary sample size \( n \), the corresponding H.T. estimator is always more efficient than the customary p.p.s. with replacement estimator.

**Proof:**

Variance of the customary p.p.s. with replacement estimator is

\[ V(\hat{y}_{\text{pps}}) = \frac{1}{n} \left( \sum_{i=1}^{N} \frac{y_i^2}{p_i} - \bar{y}^2 \right) \]  

(2.2.6)

For the Midzuno scheme with revised probabilities,

\[ p_i = np_i \]  

(2.2.7)

and

\[ p_{ij} = \frac{n-1}{N-1} \cdot \left[ \frac{N-n}{N-2} (p_i^+ + p_j^+) + \frac{n-2}{N-2} \right] \]  

(2.2.8)
where \( p_i^* \) is given by (2.2.4). Using these, we have

\[
\frac{p_{ij}}{p_{i}p_{j}} = \frac{(n-1)}{n(N-2)} \cdot \left( \frac{1}{p_{i}} + \frac{1}{p_{j}} - \frac{1}{(N-1)p_{i}p_{j}} \right)
\]

Thus we have

\[
\sum_i \sum_{j \neq i} \frac{p_{ij}}{p_{i}p_{j}} \cdot y_i y_j = \frac{(n-1)}{n \cdot (N-2)} \left[ 2y \cdot \sum \frac{N}{1p_{t}} \cdot \sum \frac{N}{1p_{t}^2} - \sum \frac{N}{1p_{t}} \right]
\]

Using (2.2.9) we have from (2.2.6) and (2.1.9)

\[
V(\hat{\gamma}_{pps}) - V(\hat{\gamma}_{H.T.})_{M} = \frac{1}{n} \cdot \sum \frac{N}{1p_{i}} \cdot \frac{y_i}{p_{i}} - y^2 - \sum \frac{N}{1np_{i}} + y^2
\]

\[
- \frac{(n-1)}{n(N-2)} \cdot \left[ 2y \cdot \sum \frac{N}{1p_{i}} - 2 \sum \frac{N}{1p_{i}} \right]
\]

\[
- \frac{1}{N-1} \cdot \left( \sum \frac{N}{1p_{i}} \right)^2 + \frac{1}{N-1} \cdot \sum \frac{N}{1p_{i}}
\]

(2.2.10)

\[
= \frac{(n-1)}{n(N-1)(N-2)} \cdot \left[ (N-1) (N-2) y^2 - 2(N-1) \cdot \sum \frac{N}{1p_{i}} \right]
\]

\[
+ 2(N-1) \cdot \sum \frac{N}{1p_{i}} \cdot \left( \sum \frac{N}{1p_{i}} \right)^2 - \sum \frac{N}{1p_{i}}
\]
\[ \frac{(n-1)}{n(N-1)(N-2)} \cdot \left[ \sum \left( \frac{Y_i}{P_i} - Y \right)^2 \right] + 2(N-1) \cdot \left( \sum \frac{Y_i^2}{P_i} - Y^2 \right) \]

\[ - \left( \sum \frac{Y_i^2}{P_i} - 2Y \cdot \sum \frac{Y_i}{P_i} + N Y^2 \right) \]

\[ = \frac{(n-1)}{n(N-1)(N-2)} \cdot \left[ \sum \left( \frac{z_i}{P_i} \right)^2 + 2(N-1) \cdot \sum \frac{P_i z_i^2}{P_i} - \sum \frac{z_i^2}{P_i} \right] \]

where

\[ z_i = \frac{Y_i}{P_i} - Y \]

Now we have from (2.2.5)

\[ P_i \geq \frac{n-1}{n} \cdot \frac{1}{N-1} \geq \frac{1}{2(N-1)} \], for \( n \geq 2 \).

Thus we have \( 2(N-1)p_i \geq 1 \).

Using this condition it can be seen from (2.2.11) that

\[ \text{V}(\hat{\gamma}_{\text{pps}}) - \text{V}(\hat{\gamma}_{\text{H.T.}}) M \geq 0 \]

Q.E.D.

Theorem 2.7:

The Yates and Grundy estimate of variance for the Midzuno scheme with revised probabilities is always nonnegative.

Proof:

The proof is exactly the same as given by Sen (1953) and Desraj (1956a) for the Midzuno scheme except for replacing \( p_i \) by \( p_i^* \). Q.E.D.
For the Midzuno scheme with revised probabilities, even if it is guaranteed that the H.T. estimator is always more efficient than the p.p.s. with replacement estimator and that the Yates and Grundy estimate of variance is always nonnegative, it suffers from a severe restriction that the method is applicable only when \( p_i^* > \frac{n-1}{n} \cdot \frac{1}{N-1} \). All the more, since only the first unit is selected with probability proportional to size, the rest being selected with equal probabilities this method is not likely to be as efficient as a method wherein all the \( n \) units are selected with unequal probabilities and without replacement.

2.2.2. Goodman and Kish procedure

The procedure mentioned by Goodman and Kish (1950) is as follows:

Arrange the \( N \) units in a random order and let \( T_j = \sum_{i=1}^{j} np_i \), \( T_0 = 0 \), be the cumulative totals of \( np_i \) in that order. Select a random start by selecting a uniform variate \( d \) with \( 0 < d < 1 \). Then select the \( n \) units whose indices \( j \) satisfy \( T_{j-1} < d + k < T_j \) for some \( k \) between 0 and \( n-1 \). For this procedure of sampling it can be easily verified that

\[
P_i = np_i
\]

The mathematical difficulties involved in evaluating the probabilities \( p_{ij} \) are resolved by Hartley and Rao (1962) by
using an asymptotic theory, and the compact expressions for the variance and the estimate of the variance of the H.T. estimator in terms of $p_t$'s and $y_t$'s have been provided. By assuming that $p_i$ is of $O(N^{-1})$ and $n$ is small relative to $N$, Hartley and Rao derived the approximate expression for $P_{ij}$ to $O(N^{-3})$ and hence for $V(\hat{Y}_{H,T.})_{GK}$ to $O(N^1)$. For the use of moderately large populations they also evaluated $V(\hat{Y}_{H,T.})_{GK}$ to $O(N^0)$ by evaluating $P_{ij}$ to $O(N^{-4})$.

The expression for $P_{ij}$ of the Goodman and Kish procedure obtained by Hartley and Rao correct to $O(N^{-4})$ is

$$P_{ij} = n(n-1)p_ip_j[1+(p_i+p_j)-\Sigma p_t^2]+[2(p_i^2+p_j^2)^2]$$

$$- 2\Sigma p_t^3+2p_ip_j-3(p_i+p_j)\cdot \Sigma p_t^2+3(\Sigma p_t^2)^2]$$

(2.2.14)

and the variance correct to $O(N^0)$ is

$$V(\hat{Y}_{H,T.})_{G.K} = \frac{1}{n} \cdot \left[ \Sigma p_i z_i^2-(n-1)\Sigma p_i z_i^2 \right] \frac{(n-1)}{n} \cdot \left[ 2\Sigma p_i^3 z_i^2 \right.$$ 

$$- \Sigma p_i^2 \cdot \Sigma p_i^2 z_i^2-2 \cdot (\Sigma p_i^2 z_i^2)^2 \right]$$

(2.2.15)

where $z_i$ is given by (2.2.12).

From (2.2.15) we have that $V(\hat{Y}_{H,T.})_{GK}$ correct to $O(N^1)$ in the more familiar form is given by

$$V(\hat{Y}_{H,T.})_{G.K} = \frac{1}{n} \Sigma p_i [1-(n-1)p_i](y_i/p_i-y)^2,$$

(2.2.16)
which clearly shows the principal reduction in the variance by adopting the without replacement scheme instead of the with replacement scheme.

2.2.3. Sampford's procedure

In Section 2.1 we have presented three equivalent schemes for samples of size two that are proposed by Brewer, Rao and Durbin respectively, and we have discussed some of the desirable properties that the H.T. estimator under these schemes possess. So it would be a worthwhile attempt if some or all of these procedures could be generalized for samples of size \( n > 2 \) in view of the simplicity and straightforwardness of these methods. Brewer has described the difficulties involved in generalizing his scheme for \( n > 2 \). Rao tried to generalize his scheme for the case \( n = 3 \), and having faced with the possibility of getting negative values for the revised probabilities he ruled out the possibility for generalizing the scheme for \( n > 2 \). However, Sampford (1967) has generalized the Durbin's scheme and presented a scheme that is applicable for all sample sizes which is described below.

Since the condition \( n p_i = 1 \) ensures the automatic inclusion of the unit in the sample, which reduces the problem to select \( (n-1) \) units only, we may assume without loss of generality that \( n p_i < 1 \) for all \( i \).

Let

\[
\lambda_i = \frac{p_i}{(1-np_i)}
\]  

(2.2.17)
Further, let $S(m)$ denote a set of $m$ different units $i_1, i_2, \ldots, i_m$, and let $L_m$ be defined by

$$L_0 = 1$$

and

$$L_m = \sum_{S(m)} \lambda_{i_1} \lambda_{i_2} \cdots \lambda_{i_m}, \quad (1 \leq m \leq N)$$

where the summation is taken over all possible sets of $m$ units drawn from the population. The procedure consists of selecting the particular sample $S(n)$, consisting of units $i_1, i_2, \ldots, i_n$ with probability

$$P(S(n)) = \frac{n}{K_n} \cdot \lambda_{i_1} \lambda_{i_2} \cdots \lambda_{i_n} \quad (1 - \sum_{u=1}^{n} p_{iu})$$

where

$$K_n = (\sum_{t=1}^{n} tL_n - t/n^t)^{-1}$$

The probabilities (2.2.19) can be achieved in practice in three different ways:

(i) The straightforward way is to evaluate the respective probabilities for the set of all possible samples and to draw one sample from this set with the required probability. However, this is not practicable to adopt for moderately large population sizes.

(ii) Units may be selected without replacement, with the probabilities evaluated at each drawing according to the rule described and illustrated by Sampford (1967).
(iii) The third method is by selecting $n$ units with replacement, the first drawing being made with probabilities $p_i$ and all subsequent ones with probabilities proportional to $p_i/(1-np_i)$ and rejecting completely any sample that does not contain $n$ different units and to start afresh.

In practice, method (iii) could be more convenient because a sample can be discarded as soon as a duplicate unit is drawn. However, for small samples one may take as a guide line in the relative preference of methods (ii) and (iii), the value of the expected number of samples that must be drawn to obtain an acceptable sample which is given by $N \frac{N}{(\sum \lambda_i)^{n-1}}/(n-1)!$. Smaller the value of this expected number, more would be the chance of getting less number of rejections.

For this scheme of sampling Sampford has shown that the expressions for $P_i$ and $P_{ij}$ are given by

$$P_i = np_i$$  \hspace{1cm} (2.2.21)

and

$$P_{ij} = K_n \cdot \lambda_i \lambda_j \phi_{ij}$$  \hspace{1cm} (2.2.22)

where $K_n$ is given by (2.2.20), $\lambda_i$ is given by (2.2.17) and $\phi_{ij}$ is given by

$$\phi_{ij} = \frac{n}{S(n-2)} \cdot \prod_{l=1}^{n-2} \lambda_{l} \cdot \frac{1-(p_i+p_j)\cdot \sum_{u=1}^{n-2} F_u}{(1-(p_i+p_j)\cdot \sum_{u=1}^{n-2} F_u)}$$  \hspace{1cm} (2.2.23)
Sampford has also shown that for this scheme the condition \( P_i P_j - P_{ij} > 0 \), is satisfied which ensures the nonnegativity of the Yates and Grundy variance estimator.

The other sampling schemes that are existent in the literature are the one suggested by K. Vijayan (1968) which is a generalization of one of the procedures suggested by Hanurav (1967) for sample size two and the rejective sampling schemes of Hájek (1964) and Hanurav (1967). The mathematical complications involved in these procedures would make their usefulness much doubtful in practice because the survey practitioner cares much for the simplicity involved in adopting a particular procedure in addition to other requirements like good efficiency compared to other methods.

2.3. Evaluation of the Approximate Expression for \( P_{ij} \) of the Sampford's Procedure

Even though Sampford has given the exact expression (2.2.22) for \( P_{ij} \) and also the computational methods to evaluate these probabilities, the computations become quite cumbersome particularly for \( N \) and/or \( n \) large. It may not be too difficult to carry out the computations on an electronic computer. However, the access to the electronic computers in some developing and underdeveloped countries is restricted and only use of the desk calculators could be made. Since the need of conducting sample surveys in the developing
countries is great, simplicity of computations is one of the important factors in choosing a sampling procedure. Thus the Sampford's scheme suffers from this drawback. In such cases and in cases where quick results are needed one may prefer to use the approximate expressions that would be quite satisfactory and easy for numerical evaluation. Also one would like to know the relative efficiencies of two given schemes to use them as a guideline for their relative preferences. Since the procedure of Goodman and Kish described in Subsection 2.2.2 and the procedure of Sampford described in Subsection 2.2.3 are two competitive schemes, it is worthwhile to compare the efficiencies of the two schemes. Since Hartley and Rao derived the approximate expressions for $P_{ij}$ and the variance of the H.T. estimator for the procedure of Goodman and Kish using an asymptotic theory under some specific assumptions, it would be realistic for comparison purposes to derive the approximate expressions for $P_{ij}$ and the variance for the Sampford's procedure using the same asymptotic approach under the same assumptions. In this section we derive the approximate expressions for $P_{ij}$ for the Sampford's procedure.

In order to evaluate the variance expression of the H.T. estimator correct to $O(N^0)$ for the Sampford's procedure, we have to evaluate $P_{ij}$ correct to $O(N^{-4})$ under the assumptions that $n$ is small relative to $N$ and $p_i$ is of $O(N^{-1})$. 
For the Sampford's procedure the exact expression for $P_{ij}$ is given by (2.2.22) viz.,

$$P_{ij} = K_n \lambda_i \lambda_j \phi_{ij}$$ \hspace{1cm} (2.3.1)

From (2.2.17) we have

$$\lambda_i \lambda_j = \frac{p_i}{(1-np_i)} \cdot \frac{p_j}{(1-np_j)}$$

Since $np_i < 1$, expanding in Taylor series we get

$$\lambda_i \lambda_j = p_i p_j [1 + np_i + n^2 p_i^2 + ...] [1 + np_j + n^2 p_j^2 + ...]$$

Retaining the terms up to $O(N^{-4})$ only we get

$$\lambda_i \lambda_j = p_i p_j [1 + n(p_i + p_j) + n^2 (p_i^2 + p_j^2 + p_i p_j)]$$ \hspace{1cm} (2.3.2)

The leading term of $\lambda_i \lambda_j$ above is of $O(N^{-2})$ and thus it would be sufficient to evaluate $K_n$ and $\phi_{ij}$ each correct to $O(N^{-2})$ only in order to evaluate $P_{ij}$ correct to $O(N^{-4})$.

2.3.1. Evaluation of $K_n$ correct to $O(N^{-2})$

The expression for $K_n$ is given by (2.2.20) as

$$K_n = \left( \sum_{t=1}^{n} \frac{tL}{n-t} \right)^{-1}$$ \hspace{1cm} (2.3.3)

For evaluating $K_n$ we first need to evaluate $L_m$ correct to $O(N^{-2})$. 
Consider

\[ L_m = \sum_{S(m)} \lambda_{l_1} \lambda_{l_2} \cdots \lambda_{l_m} \]

\[ = \binom{N}{m} \cdot E[\lambda_{l_1} \lambda_{l_2} \cdots \lambda_{l_m}] \quad (2.3.4) \]

where \( \binom{N}{m} \) stands for the number of ways of choosing \( m \) out of \( N \) units and \( E \) denotes the expectation taken over the scheme of selecting \( m \) units out of \( N \) units with simple random sampling without replacement. Without loss of generality we can assume that the units \( l_1, l_2, \ldots, l_m \) are selected in that order.

Now, substituting the value of \( \lambda_{l_i} \) in (2.3.4) we get

\[ L_m = \frac{N(m)}{m!} \cdot E[p_{l_1} p_{l_2} \cdots p_{l_m} \{1 + np_{l_1} + n^2 p_{l_1}^2 + \cdots\}] \cdot \]

\[ \{1 + np_{l_2} + n^2 p_{l_2}^2 + \cdots\} \cdots \]

\[ \{1 + np_{l_m} + n^2 p_{l_m}^2 + \cdots\} \quad (2.3.5) \]

where \( N(m) = N(N-1) \cdots (N-m+1) \quad (2.3.6) \)

It can be seen that for any set of positive integers \( \alpha_1, \alpha_2, \ldots, \alpha_m \), the contribution of

\[ \prod_{i=1}^{m} E[p_{l_1}^{\alpha_1} p_{l_2}^{\alpha_2} \cdots p_{l_m}^{\alpha_m}] \]

correct to \( O(N^{-2}) \) would be zero if \( \sum_{i=1}^{m} \alpha_i > (m+2) \). Further from
the basic properties of simple random sampling it is also known that

$$E[p_{\ell_1}^{\alpha_1} p_{\ell_2}^{\alpha_2} \cdots p_{\ell_m}^{\alpha_m}]$$

is the same for all the $m!$ permutations of $(\alpha_1, \alpha_2, \ldots, \alpha_m)$.

Hence from (2.3.5) it follows that the expression for $L_m$ that could contribute to $O(N^{-2})$ is given by

$$L_m = \frac{N(m)}{m!} [E(p_{\ell_1} p_{\ell_2} \cdots p_{\ell_m}) + nm \cdot E(p_{\ell_1}^2 p_{\ell_2} p_{\ell_3} \cdots p_{\ell_m})$$

$$+ n^2 m \cdot E(p_{\ell_1}^3 p_{\ell_2} p_{\ell_3} \cdots p_{\ell_m})$$

$$+ \frac{n^2 m(m-1)}{2} \cdot E(p_{\ell_1}^2 p_{\ell_2}^2 p_{\ell_3} p_{\ell_4} \cdots p_{\ell_m})] \quad (2.3.7)$$

Now we will prove here a lemma which will be used in the evaluation of $L_m$.

Lemma 2.1:

Let $\ell_1 \ell_2 \cdots \ell_m$ be the units drawn in that order when a simple random sample of size $m$ is drawn from a population of $N$ units. Under this scheme of sampling, for $m \geq 3$ where $m$ is small relative to $N$ and $p_i$ is of $O(N^{-1})$ the following relations are true correct to $O(N^{-2})$. 

\[ N(m) \cdot E(p_{l_1} p_{l_2} \cdots p_{l_m}) = (\Sigma p_t)^m - \binom{m}{2} (\Sigma p_t)^{m-2} \cdot p_t^2 \]

\[ + 2 \cdot \binom{m}{3} \cdot (\Sigma p_t)^{m-3} \cdot p_t^3 \]

\[ + 3 \cdot \binom{m}{4} \cdot (\Sigma p_t)^{m-4} \cdot (\Sigma p_t^2)^2 \]

(2.3.8)

\[ N(m) \cdot E(p_{l_1}^2 p_{l_2} p_{l_3} \cdots p_{l_m}) = (\Sigma p_t)^{m-1} \cdot p_t^2 \]

\[ - (m-1)(\Sigma p_t)^{m-2} \cdot p_t^3 \]

\[ - \binom{m-1}{2} \cdot \Sigma (p_t)^{m-3} \cdot (\Sigma p_t^2)^2 \]

(2.3.9)

\[ N(m) \cdot E(p_{l_1}^3 p_{l_2} p_{l_3} \cdots p_{l_m}) = (\Sigma p_t)^{m-1} \cdot p_t^3 \]

(2.3.10)

and

\[ N(m) \cdot E(p_{l_1}^2 p_{l_2}^2 p_{l_3} \cdots p_{l_m}) = (\Sigma p_t)^{m-2} \cdot (\Sigma p_t^2)^2 \]

(2.3.11)

wherein

\[ \binom{\mu}{\nu} \]

is to be taken as zero if \( \mu < \nu \)

**Proof:**

First we consider

\[ E[p_{l_1} p_{l_2} p_{l_3}] = E[p_{l_1} p_{l_2} \cdot E(p_{l_3} / l_1, l_2)] \]

(2.3.12)

where \( E(p_{l_3} / l_1, l_2) \) denotes the conditional expectation of \( p_{l_3} \) given that \( l_1 \) and \( l_2 \) are the units selected in the
first two draws.

Thus we have from (2.3.12)

\[ E[p_{1} p_{2} p_{3}] = E[p_{1} p_{2} \cdot \frac{\Sigma p_{t} - p_{1} - p_{2}}{N-2}] \]

\[ = \frac{1}{N-2} \cdot [\Sigma p_{t} \cdot E(p_{1} p_{2}) - E(p_{1}^{2} p_{2}) - E(p_{1} p_{2}^{2})] \]

Proceeding similarly we get correct to \( O(N^{-2}) \)

\[ N(3) \cdot E[p_{1}^{2} p_{2} p_{3}] = (\Sigma p_{t})^{3} - 3 \Sigma p_{t} \cdot \Sigma p_{t}^{2} + 2 \Sigma p_{t}^{3} \quad (2.3.13) \]

which shows that (2.3.8) is true for the value \( m=3 \). Now assuming that

\[ N(m-1) \cdot E[p_{1} p_{2} \ldots p_{m}] = (\Sigma p_{t})^{m-1} - \binom{m-1}{2} \cdot (\Sigma p_{t})^{m-3} \cdot \Sigma p_{t}^{2} \]

\[ + 2 \cdot \binom{m-1}{3} \cdot (\Sigma p_{t})^{m-4} \cdot \Sigma p_{t}^{3} \]

\[ + 3 \cdot \binom{m-1}{4} \cdot (\Sigma p_{t})^{m-5} \cdot (\Sigma p_{t})^{2} \quad (2.3.14) \]

we get

\[ N(m) \cdot E[p_{1} p_{2} \ldots p_{m}] \]

\[ = N \cdot (N-1)(m-1) \cdot E[p_{1} \cdot E(p_{2} p_{3} \ldots p_{m} / \nu_{1})] \]

\[ = N \cdot E[p_{1} \cdot (N-1)(m-1) \cdot E(p_{2} p_{3} \ldots p_{m} / \nu_{1})] \]
\[= n \cdot E[p_{l_1} \cdot (\sum_{l=2}^m p_{l_k}^{-1})^{m-1} \cdot (m-1) \cdot (\sum_{l=2}^m p_{l_1}^{-1})^{m-3} \cdot (\sum_{l=2}^m p_{l_k}^{-2} \cdot p_{l_1}^{-2})
\]
\[+ 2 \cdot (\sum_{l=2}^m p_{l_1}^{-1})^{m-4} \cdot (\sum_{l=2}^m p_{l_1}^{-3})
\]
\[+ 3 \cdot (\sum_{l=2}^m p_{l_1}^{-1})^{m-5} \cdot (\sum_{l=2}^m p_{l_k}^{-2} \cdot p_{l_1}^{-2})
\]
\[= (\sum_{l=2}^m p_{l_1}^{-1})^{m-(m-1)+\binom{m-1}{2}} \cdot (\sum_{l=2}^m p_{l_1}^{-1})^{m-2} \cdot (\sum_{l=2}^m p_{l_1}^{-2})^{m-2} \cdot (\sum_{l=2}^m p_{l_1}^{-3})^{m-3} \cdot (\sum_{l=2}^m p_{l_1}^{-3})
\]
\[+ 3 \cdot (\sum_{l=2}^m p_{l_1}^{-1})^{m-4} \cdot (\sum_{l=2}^m p_{l_k}^{-2} \cdot p_{l_1}^{-2})
\]

Correct to \(O(N^{-2})\). Thus from Equations (2.3.13)-(2.3.15) it follows by induction that (2.3.8) of the Lemma is true for all \(m>3\).

Now considering

\[N(m) \cdot E[p_{l_1}^2 \cdot p_{l_2}^2 \cdot p_{l_3}^2 \cdot \ldots \cdot p_{l_m}^2]
\]
\[= n \cdot E[p_{l_1}^2 \cdot (N-1)(m-1) \cdot E[p_{l_2}^2 \cdot p_{l_3}^2 \cdot \ldots \cdot p_{l_m}^2 / l_1)]
\]

From (2.3.8) we have
\[
(N-1) (m-1) \cdot E[p_{\ell_2} p_{\ell_3} \cdots p_{\ell_m} / \ell_1]
\]
\[
= (\Sigma p_t - p_{\ell_1})^{m-1} \cdot \left( \Sigma p_t - p_{\ell_1} \right)^{m-3} \cdot (\Sigma p_t^2 - p_{\ell_1}^2)^2
\]
\[
+ 2 \cdot (m-1) \cdot (\Sigma p_t - p_{\ell_1})^{m-4} \cdot (\Sigma p_t^3 - p_{\ell_1}^3)
\]
\[
+ 3 \cdot (m-1) \cdot (\Sigma p_t - p_{\ell_1})^{m-5} \cdot (\Sigma p_t^2 - p_{\ell_1}^2)^2
\]
(2.3.17)

substituting this in (2.3.16) we get after simplifying and retaining terms to \(O(N^{-2})\),

\[
N(m) \cdot E(p_{\ell_1}^2 p_{\ell_2} p_{\ell_3} \cdots p_{\ell_m}) = (\Sigma p_t)^{m-1} \cdot \Sigma p_t^2
\]

\[
-(m-1)(\Sigma p_t)^{m-2} \cdot \Sigma p_t^3 - (m-1) \cdot (\Sigma p_t)^{m-3} \cdot (\Sigma p_t^2)^2
\]

which shows that (2.3.9) is true for all \(m \geq 3\).

Considering

\[
N(m) \cdot E(p_{\ell_1}^3 p_{\ell_2} p_{\ell_3} \cdots p_{\ell_m})
\]

\[
= N \cdot E[p_{\ell_1}^3 \cdot (N-1)(m-1) \cdot E(p_{\ell_2} p_{\ell_3} \cdots p_{\ell_m} / \ell_1)]
\]

we get after using (2.3.17) and simplifying

\[
N(m) \cdot E(p_{\ell_1}^3 p_{\ell_2} p_{\ell_3} \cdots p_{\ell_m})
\]

\[
= (\Sigma p_t)^{m-1} \cdot \Sigma p_t^3
\]
correct to $o(N^{-2})$, which shows that (2.3.10) is true for all $m \geq 3$.

Now we consider

$$N(m) \cdot E(p_{k_1}^2 p_{k_2}^2 p_{k_3}^2 p_{k_4} \cdots p_{k_{m}})$$

$$= N \cdot E[p_{k_1}^2 (N-1)(m-1) \cdot E(p_{k_2}^2 p_{k_3}^2 p_{k_4} \cdots p_{k_{m}}/k_1)] \quad (2.3.18)$$

From (2.3.9) we have

$$(N-1)(m-1) \cdot E(p_{k_2}^2 p_{k_3}^2 p_{k_4} \cdots p_{k_{m}}/k_1)$$

$$= (\Sigma p_{k_1} - p_{k_1})^{m-2} \cdot (\Sigma p_{k_1}^2 - p_{k_1}^2)$$

$$- (m-2) \cdot (\Sigma p_{k_1} - p_{k_1})^{m-3} \cdot (\Sigma p_{k_1}^3 - p_{k_1}^3)$$

$$- (m-2) \cdot (\Sigma p_{k_1} - p_{k_1})^{m-4} \cdot (\Sigma p_{k_1}^2 - p_{k_1}^2)^2$$

Substituting this value in (2.3.9) we get after simplifying and retaining terms to $o(N^{-2})$,

$$N(m) \cdot E(p_{k_1}^2 p_{k_2}^2 p_{k_3}^2 p_{k_4} \cdots p_{k_{m}}) = (\Sigma p_{k_1})^{m-2} \cdot (\Sigma p_{k_1}^2)^2$$

which shows that (2.3.11) is true for all $m \geq 3$.

**Remark:**

Even though the proof of the Lemma assumes that $m \geq 3$, the statement of the Lemma is true for the values $m = \tilde{u}, 1$ and $2$ also which can be easily verified.
Now using the results of Lemma 2.1, we get from (2.3.7),

\[ N \]

after observing that \( \sum p_t = 1 \), for \( m \geq 3 \),

\[
L_m = \frac{1}{m!} \left\{ \left( 1 - \left( \frac{m}{2} \right) \sum p_t^2 + 2 \cdot \left( \frac{m}{3} \right) \sum p_t^3 + 3 \cdot \left( \frac{m}{4} \right) \sum p_t^2 \right) \right\} \\
+ nm \cdot \left( \sum p_t^2 - \sum p_t^3 \right) \left( \sum p_t^2 \right) \\
+ n^2 m \cdot \sum p_t^3 + \frac{n^2 m (m-1)}{2} \cdot \left( \sum p_t^2 \right) \\
= \frac{1}{m!} \left\{ 1 + \left( \frac{m}{1} \right) n - \left( \frac{m}{2} \right) \right\} \sum p_t^2 + \left( \frac{m}{1} \right) n^2 \sum p_t^3 + \left( \frac{m}{2} \right) n^2 + \left( \frac{m}{3} \right) \sum p_t^3 \\
+ \left( \frac{m}{2} \right) n^2 - 3 \cdot \left( \frac{m}{3} \right) n + 3 \cdot \left( \frac{m}{4} \right) \sum p_t^2 \right\} \left( \sum p_t^2 \right)^2 \]  

(2.3.19)

By definition, \( L_0 = 1 \)

(2.3.20)

Also it can be easily verified that correct to \( O(N^{-2}) \),

\[
L_1 = 1 + n \sum p_t^2 + n^2 \sum p_t^3 \]  

(2.3.21)

and

\[
L_2 = \frac{1}{2} \cdot \left\{ 1 + (2n-1) \sum p_t^2 + 2n(n-1) \sum p_t^3 + n^2 \left( \sum p_t^2 \right)^2 \right\} \]  

(2.3.22)

From (2.3.3) we have

\[
\frac{1}{K_n} = \sum_{t=1}^{n} \frac{tL_{n-t}}{n} \]  

(2.3.23)

Thus by using (2.3.20) and (2.3.21) we get for \( n = 2 \),

\[
\frac{1}{K_2} = \frac{(L_0 + L_1)}{2} = 1 + \sum p_t^2 + 2 \sum p_t^3 \]  

(2.3.24)
Similarly by substituting the values from (2.3.19) to (2.3.22) for the respective terms in (2.3.23) we get after simplifying

\[
\frac{1}{K_3} = \frac{1}{2} \cdot [1+3\Sigma p_t^2 + 8\Sigma p_t^3 + 3(\Sigma p_t^2)^2] \tag{2.3.25}
\]

and

\[
\frac{1}{K_4} = \frac{1}{6} \cdot [1+6\Sigma p_t^2 + 20\Sigma p_t^3 + 15(\Sigma p_t^2)^2] \tag{2.3.26}
\]

**Theorem 2.7:**

For \(n \geq 5\), the expression for \(1/K_n\) correct to \(O(N^{-2})\) is given by

\[
\frac{1}{K_n} = \frac{1}{(n-1)!} + \frac{n}{2(n-2)!} \Sigma p_t^2 + \frac{n(n+1)}{3(n-2)!} \Sigma p_t^3 + \frac{n(n+1)}{8(n-3)!} (\Sigma p_t^2)^2 \tag{2.3.27}
\]

**Proof:**

Using the transformation \(s = n-t\) in (2.3.23) we get

\[
\frac{1}{K_n} = \sum_{s=0}^{n-1} \frac{(n-s) \cdot L_s}{n} \cdot \frac{n-s}{n} \tag{2.3.28}
\]

\[
= \frac{L_0}{n^{n-1}} + (n-1) \cdot \frac{L_1}{n^{n-1}} + (n-2) \cdot \frac{L_2}{n^{n-2}} + G,
\]

where

\[
G = \sum_{s=3}^{n-1} \frac{(n-s) \cdot L_s}{n} \cdot \frac{n-s}{n} \tag{2.3.28}
\]
Substituting the value of $L_m$ from (2.3.19) the above expression for $G$ can be written as

$$G = \frac{1}{n^n} \cdot \left[ \sum_{s=3}^{n-1} (n-s) \cdot T_s + n^2 \cdot \left\{ \sum_{s=3}^{n-1} (n-s) \cdot T_{s-1} \right\} - \frac{1}{2} \cdot \sum_{s=3}^{n-1} (n-s) \cdot T_{s-2} \cdot \Sigma p_t^2 \right]$$

$$+ n^3 \cdot \left\{ \sum_{s=3}^{n-1} (n-s) \cdot T_{s-1} - \sum_{s=3}^{n-1} (n-s) \cdot T_{s-2} \right\}$$

$$+ \frac{1}{3} \cdot \sum_{s=3}^{n-1} (n-s) \cdot T_{s-3} \cdot \Sigma p_t^3$$

$$+ \frac{4}{2} \cdot \left\{ \sum_{s=3}^{n-1} (n-s) \cdot T_{s-2} - \sum_{s=3}^{n-1} (n-s) \cdot T_{s-3} \right\}$$

$$+ \frac{1}{4} \cdot \sum_{s=4}^{n-1} (n-s) \cdot T_{s-4} \cdot [\Sigma p_t^2]^2$$

(2.3.29)

where

$$T_s = \frac{n^s}{s!}$$

(2.3.30)

For any nonnegative integers $0 \leq \ell \leq m$, let

$$I(\ell, m) = \sum_{s=\ell}^{m} T_s$$

and

$$J(\ell, m) = \sum_{s=\ell}^{m} s \cdot T_s$$

Then the following results can easily be established:
For any $0 < l < m$,

\[
\sum_{s=l}^{m} (n-s) \cdot T_{s-a} = (n-l) \cdot I(l-a, m-a) - J(l-a, m-a) \tag{2.3.31}
\]

\[
J(0, m) = J(1, m) \tag{2.3.32}
\]

For any $1 < l < m$, \( J(l, m) = n \cdot I(l-1, m-1) \tag{2.3.33} \)

and for any $0 < l < m$,

\[
I(l+1, m+1) - I(l, m) = T_{m+1} - T_l \tag{2.3.34}
\]

In view of (2.3.31), expression (2.3.29) for $G$ reduces to

\[
G = \frac{1}{n^n} \cdot \left\{ \frac{n \cdot I(3, n-1) - J(3, n-1)}{n^2} + \frac{(n-1)I(2, n-2)}{n} \right\} + \frac{2}{n} \cdot J(2, n-2) - \frac{1}{2} \cdot J(1, n-3) + \frac{1}{3} \cdot J(0, n-4) + \frac{4}{n} \cdot \frac{(n-2)I(1, n-3) - J(1, n-3)}{n^4} - \frac{1}{3} \cdot J(0, n-4) + \frac{1}{4} \cdot J(0, n-5) \right\} (\exp t)^2
\]

\[
+ \frac{1}{n^3} \cdot (n-1)I(2, n-2) + \left\{ \frac{n-2}{2} \cdot I(1, n-3) + \frac{1}{2} \cdot J(1, n-3) \right\} (\exp t)^2
\]

\[
+ \frac{n^3}{3} \cdot (n-1)I(2, n-2) + \frac{1}{3} \cdot J(1, n-3) \right\} (\exp t)^3
\]

\[
+ \frac{n^4}{4} \cdot \frac{(n-2)I(1, n-3) - J(1, n-3)}{2} + \frac{1}{4} \cdot J(0, n-5) \right\} (\exp t)^2
\]
Using relations (2.3.32)-(2.3.34) the above expression for $G$ can be written, after suitable rearrangement of the terms, as

$$G = \frac{1}{n} \cdot [n(T_{n-1} - T_2) + n^2 \cdot (n-1)(T_{n-2} - T_1)$$

$$- \frac{n}{2}(T_{n-3} - T_0) \cdot \Sigma P_t^2$$

$$+ n^3 \cdot (n-1)(T_{n-2} - T_1) - (n-1)(T_{n-3} - T_0)$$

$$+ \frac{n}{3} \cdot T_{n-4} \Sigma P_t^3$$

$$+ n^4 \cdot \left\{ \frac{n-2}{2} \cdot (T_{n-3} - T_0) - \frac{(n-1)}{2} \cdot T_{n-4} \right\}$$

$$+ \frac{n}{8} \cdot T_{n-5} \cdot (\Sigma P_t^2)^2 \right\}$$

(2.3.35)

Substituting the values of $L_0$, $L_1$ and $L_2$ from (2.3.20)-(2.3.22), and the value of $G$ from (2.3.35) in Equation (2.3.28) we get for $n \geq 5$,

$$\frac{1}{K_n} = \frac{1}{n^{n-1}} + \frac{(n-1) \cdot (1+n\Sigma P_t^2 + n^2 \Sigma P_t^3)}{n^{n-1}}$$

$$+ \frac{(n-2)}{2n^{n-2}} \cdot [1+(2n-1)\Sigma P_t^2 + 2n(n-1)\Sigma P_t^3 + n^2(\Sigma P_t^2)^2]$$

$$+ \frac{1}{n} \cdot [n(T_{n-1} - T_2) + n^2 \cdot (n-1)(T_{n-2} - T_1)$$

$$- \frac{n}{2}(T_{n-3} - T_0) \Sigma P_t^2$$
Now, let
\[ \frac{1}{K_n} = C_0 + C_1 \Sigma p_t^2 + C_2 \Sigma p_t^3 + C_3 (\Sigma p_t^2)^2 \] (2.3.37)

Equating the coefficients of the like terms in (2.3.36) and (2.3.37) we get after substituting the value of \( T_s \) from (2.3.30),

\[ C_0 = \frac{1}{n^{n-1}} + \frac{(n-1)}{n^{n-1}} + \frac{(n-2)}{2n^{n-2}} + \frac{1}{n^{n-1}} \left[ \frac{n^{n-1}}{(n-1)!} - \frac{n^2}{2} \right] \]

\[ = \frac{1}{(n-1)!}, \quad (2.3.38) \]

\[ C_1 = \frac{(n-1)}{n^{n-2}} + \frac{(n-2)(2n-1)}{2n^{n-2}} + \frac{1}{n^{2n-2}} \left( \frac{n^{n-2} \cdot (n-1)}{(n-2)!} - n(n-1) \right) \]

\[ - \frac{n^{n-2}}{2(n-3)!} + \frac{n}{2} \]

\[ = \frac{n}{2(n-2)!}, \quad (2.3.39) \]
\[ C_2 = \frac{(n-1)}{n^{n-3}} + \frac{(n-1)(n-2)}{n^{n-3}} + \frac{1}{n^{n-3}} \left[ (n-1) \cdot \frac{n^{n-2}}{(n-2)!} - n \right] \]

\[ = (n-1) \cdot \frac{n^{n-3}}{(n-3)!} - 1 + \frac{n^{n-3}}{(n-4)!} \]

\[ = \frac{n(n+1)}{3(n-2)!} \quad (2.3.40) \]

and

\[ C_3 = \frac{(n-2)}{2n^{n-4}} + \frac{1}{n^{n-4}} \cdot \left[ \frac{(n-2)}{2} \cdot \frac{n^{n-3}}{(n-3)!} - 1 \right] - \frac{(n-1) \cdot n^{n-4}}{2(n-4)!} + \frac{n^{n-4}}{8(n-5)!} \]

\[ = \frac{n(n+1)}{8(n-3)!} \quad (2.3.41) \]

Hence from Equations (2.3.37)-(2.3.41) it follows that (2.3.27) holds. Q.E.D.

Remark:

Even though Equation (2.3.27) of Theorem 2.7 is derived for \( n \geq 5 \), observation of Equations (2.3.25) and (2.3.26) tells us that (2.3.27) is true for the values \( n = 3 \) and 4 also.

Thus we have for \( n \geq 3 \),

\[ \frac{1}{K_n} = \frac{1}{(n-1)!} + \frac{n}{2(n-2)!} \cdot \Sigma P_t^2 + \frac{n(n+1)}{3(n-2)!} \Sigma P_t^3 + \frac{n(n+1)}{8(n-3)!} (\Sigma P_t^2)^2 \quad (2.3.42) \]
Now from (2.3.24) we have

\[ K_2 = \frac{1}{1 + \Sigma p_t^2 + 2 \Sigma p_t^3}, \]

which after expanding the denominator and retaining terms to \( O(N^{-2}) \) gives

\[ K_2 = 1 - \Sigma p_t^2 - 2 \Sigma p_t^3 + (\Sigma p_t^2)^2 \] (2.3.43)

After a similar operation we get from (2.3.42) for \( n \geq 3 \),

\[ K_n = \frac{1}{c_0} \cdot [1 - \frac{c_1}{c_0} \cdot \Sigma p_t^2 - \frac{c_2}{c_0} \cdot \Sigma p_t^3 + (\frac{c_1}{c_0} - \frac{c_3}{c_0}) (\Sigma p_t^2)^2] \] (2.3.44)

2.3.2. Evaluation of \( \phi_{ij} \) correct to \( O(N^{-2}) \)

The expression for \( \phi_{ij} \) from (2.2.23) is given by

\[ \phi_{ij} = n \cdot \Sigma_{S(n-2)} \lambda_1 \lambda_2 \cdots \lambda_{n-2} \{1 - (p_i + p_j) - \Sigma_{u=1}^{n-2} p_u \} \] (2.3.45)

Since the right hand side of (2.3.45) is not meaningful to consider for \( n=2 \), we derive here the approximate expression for \( \phi_{ij} \) assuming that \( n \geq 3 \).

(2.3.45) can alternatively be written as

\[ \phi_{ij} = n \cdot \binom{n-2}{n-2} \cdot E' \left[ \lambda_1 \lambda_2 \cdots \lambda_{n-2} \{1 - (p_i + p_j) \right. \\
- \Sigma_{u=1}^{n-2} p_u \left. \right] \] (2.3.46)
where $E'$ denotes the expectation taken over the scheme of selecting $(n-2)$ units from among the population excluding the $i$th and $j$th units with simple random sampling without replacement. Without loss of generality we can assume that $\ell_1, \ell_2, \ldots, \ell_{n-2}$ are the units selected in that order.

Thus we have from (2.3.46),

$$
\phi_{ij} = \frac{n}{(n-2)!} \cdot (N-2) (n-2) \cdot [E'(\lambda_{\ell_1} \lambda_{\ell_2} \ldots \lambda_{\ell_{n-2}})]
- (p_i + p_j) \cdot E'(\lambda_{\ell_1} \lambda_{\ell_2} \ldots \lambda_{\ell_{n-2}})
- E'(\lambda_{\ell_1} \lambda_{\ell_2} \ldots \lambda_{\ell_{n-2}} \cdot \sum_{u=1}^{n-2} p_{\ell_u})]
\quad (2.3.47)
$$

First we consider

$$(N-2) (n-2) \cdot E'(\lambda_{\ell_1} \lambda_{\ell_2} \ldots \lambda_{\ell_{n-2}})$$

$$\quad = (N-2) (n-2) \cdot E'[\prod_{i=1}^{n-2} p_{\ell_i} (1 + np_{\ell_i} + np_{\ell_i}^2 + \ldots)]$$

By using the fact that $E'[p_{\ell_1}^{\alpha_1} p_{\ell_2}^{\alpha_2} \ldots p_{\ell_{n-2}}^{\alpha_{n-2}}]$ is invariant over all the permutations of $(\alpha_1 \alpha_2 \ldots \alpha_{n-2})$ for any positive integers $\alpha_1, \alpha_2 \ldots \alpha_{n-2}$, we get by retaining terms that contribute to $O(N^{-2})$ only,

$$(N-2) (n-2) \cdot E'[\lambda_{\ell_1} \lambda_{\ell_2} \ldots \lambda_{\ell_{n-2}}]$$

$$\quad = (N-2) (n-2) \cdot E'[p_{\ell_1}^{\ell_1} p_{\ell_2}^{\ell_2} \ldots p_{\ell_{n-2}}^{\ell_{n-2}}]$$
Now, using the Equations (2.3.8)-(2.3.11) of Lemma 2.1, for the population of (N-2) units excluding the ith and jth units and with \( m = n-2 \), we get

\[
(n-2) \cdot \mathbb{E}'[\lambda_1 \lambda_2 \cdots \lambda_{n-2}]
\]

\[
= [(\Sigma \pi_t - \pi_i - \pi_j)^{n-2} - (\Sigma \pi_t - \pi_i - \pi_j)^{n-4} \cdot (\pi_t^2 - \pi_i^2 - \pi_j^2) + 2 \cdot \binom{n-2}{3} \cdot (\Sigma \pi_t - \pi_i - \pi_j)^{n-5} \cdot (\pi_t^3 - \pi_i^3 - \pi_j^3) + 3 \cdot \binom{n-2}{4} \cdot (\Sigma \pi_t - \pi_i - \pi_j)^{n-6} \cdot (\pi_t^2 - \pi_i^2 - \pi_j^2)^2] + n(n-2) \cdot [(\Sigma \pi_t - \pi_i - \pi_j)^{n-3} \cdot (\pi_t^2 - \pi_i^2 - \pi_j^2)
\]

\[
- (n-3) \cdot (\Sigma \pi_t - \pi_i - \pi_j)^{n-4} \cdot (\pi_t^3 - \pi_i^3 - \pi_j^3)
\]

\[
- \binom{n-3}{2} \cdot (\Sigma \pi_t - \pi_i - \pi_j)^{n-5} \cdot (\pi_t^2 - \pi_i^2 - \pi_j^2)^2] + n^2(n-2) \cdot (\Sigma \pi_t - \pi_i - \pi_j)^{n-3} \cdot (\pi_t^3 - \pi_i^3 - \pi_j^3) + n^2 \cdot \binom{n-2}{2} \cdot (\Sigma \pi_t - \pi_i - \pi_j)^{n-4} \cdot (\pi_t^2 - \pi_i^2 - \pi_j^2)^2
\]
which, by noting that \( \sum_{l=1}^{N} p_l = 1 \) and retaining terms to \( O(N^{-2}) \)
only reduces to

\[
(N-2) (n-2) \cdot E' \left[ \lambda_{l_1} \lambda_{l_2} \cdots \lambda_{l_{n-2}} \right] = 1 + \left\{ \left( \frac{(n-2)(n+3)}{2} \right) \cdot \sum p_t^2 - (n-2) (p_i + p_j) \right\} + \left\{ (n-2)(n-3) p_i p_j - 3(n-2)(p_i^2 + p_j^2) \right. \\
\left. - \frac{(n-2)(n-3)(n+4)}{2} \cdot (p_i + p_j) \cdot \sum p_t^2 \right\} + \left\{ \frac{(n-2)(n^2+2n+12)}{3} \cdot \sum p_t^3 \right. \\
\left. + \frac{(n-2)(n-3)(n^2+7n+20)}{8} \cdot \left( \sum p_t^2 \right)^2 \right\} 
\]

Using this expression we get correct to \( O(N^{-2}) \),

\[
(p_i + p_j) \cdot (N-2) (n-2) \cdot E' \left[ \lambda_{l_1} \lambda_{l_2} \cdots \lambda_{l_{n-2}} \right] = (p_i + p_j) \cdot \left[ 1 + \left( \frac{(n-2)(n+3)}{2} \cdot \sum p_t^2 - (n-2) (p_i + p_j) \right) \right] 
\]

We now consider

\[
(N-2) (n-2) \cdot E' \left[ \lambda_{l_1} \lambda_{l_2} \cdots \lambda_{l_{n-2}} \cdot \sum p_{l_u} \right] = (N-2) (n-2) \cdot E' \left[ \left\{ \prod_{u=1}^{n-2} p_{l_u} \cdot (1 + np_{l_u} + n^2 p_{l_u}^2 + \ldots ) \right\} \cdot \sum p_{l_u} \right] 
\]
By the symmetry of $E'(p_{\ell_1}^{\alpha_1} p_{\ell_2}^{\alpha_2} \ldots p_{\ell_{n-2}}^{\alpha_{n-2}})$ in $\alpha_1, \alpha_2, \ldots, \alpha_{n-2}$ we get after retaining only the terms that contribute to $O(N^{-2})$,

$$(N-2)^{(n-2)} \cdot E'[\lambda_{\ell_1}^{\alpha_1} \lambda_{\ell_2}^{\alpha_2} \ldots \lambda_{\ell_{n-2}}^{\alpha_{n-2}} \cdot \sum_{u=1}^{n-2} p_{\ell_u}]$$

$$= (n-2) \cdot (N-2)^{(n-2)} \cdot E'[p_{\ell_1}^{2} p_{\ell_2}^{2} p_{\ell_3}^{2} \ldots p_{\ell_{n-2}}^{2}]$$

$$+ n(n-2)(n-3) \cdot (N-2)^{(n-2)} \cdot E'[p_{\ell_1}^{3} p_{\ell_2}^{3} p_{\ell_3}^{3} \ldots p_{\ell_{n-2}}^{3}]$$

Again by using the results of Lemma 2.1 with suitable changes we get after noting that $\sum_{N}^{N} p_{t} = 1$,

$$(N-2)^{(n-2)} \cdot E'[\lambda_{\ell_1}^{\alpha_1} \lambda_{\ell_2}^{\alpha_2} \ldots \lambda_{\ell_{n-2}}^{\alpha_{n-2}} \cdot \sum_{u=1}^{n-2} p_{\ell_u}]$$

$$= (n-2) \cdot \left\{ (\sum_{1}^{n-2} p_{t} - p_{i} - p_{j})^{n-3} \cdot (\sum_{1}^{n-2} p_{t}^{2} - p_{i}^{2} - p_{j}^{2}) \right\}$$

$$- (n-3) \cdot (\sum_{1}^{n-2} p_{t} - p_{i} - p_{j})^{n-4} \cdot (\sum_{1}^{n-2} p_{t}^{3} - p_{i}^{3} - p_{j}^{3})$$

$$- (n-3) \cdot (\sum_{1}^{n-2} p_{t} - p_{i} - p_{j})^{n-5} \cdot (\sum_{1}^{n-2} p_{t}^{2} - p_{i}^{2} - p_{j}^{2})^{2}$$

$$+ n(n-2)(n-3) \cdot (\sum_{1}^{n-2} p_{t} - p_{i} - p_{j})^{n-4} \cdot (\sum_{1}^{n-2} p_{t}^{2} - p_{i}^{2} - p_{j}^{2})^{2}$$

$$+ n(n-2) \cdot (\sum_{1}^{n-2} p_{t} - p_{i} - p_{j})^{n-3} \cdot (\sum_{1}^{n-2} p_{t}^{3} - p_{i}^{3} - p_{j}^{3})$$
\[= (n-2)[\Sigma p_t^2-(n-3) \cdot (p_i+p_j) \cdot \Sigma p_t^2-(p_i^2+p_j^2)]
\]
\[+ 3\Sigma p_t^3 + \frac{(n-3)(n+4)}{2} \cdot (\Sigma p_t^2)^2 \]  \hspace{1cm} (2.3.50)
correct to \(O(N^{-2})\).

Substituting the values from (2.3.48), (2.3.49) and (2.3.50) in (2.3.47) we get, for \(n \geq 3\), after some simplification,

\[
\phi_{ij} = \frac{n}{(n-2)!} \cdot [1+\{\frac{(n-2)(n+1)}{2} \cdot \Sigma p_t^2-(n-1)(p_i+p_j)\}]
\]
\[+ \{(n-1)(n-2) \cdot p_i p_j-(n-2)(p_i^2+p_j^2)\]
\[\frac{(n-2)(n^2-3)}{2} \cdot (p_i+p_j) \Sigma p_t^2
\]
\[+ \frac{(n-2)(n+2n+3)}{3} \Sigma p_t^3
\]
\[+ \frac{(n-2)(n-3)(n^2+3n+4)}{8} \cdot (\Sigma p_t^2)^2]\]  \hspace{1cm} (2.3.51)

2.3.3. Evaluation of \(P_{ij}\) correct to \(O(N^{-4})\)

Case (i): \(n=2\):

As we mentioned earlier Sampford's procedure for sample size two is the same as Durbin's scheme (1967). The expression for \(P_{ij}\) of the Durbin's scheme is

\[P_{ij} = K_2 p_i p_j \left(\frac{1}{1-2p_i} + \frac{1}{1-2p_j}\right)\]

substituting the value of \(K_2\) from (2.3.50) and after ex-
panding \(1/(1-2p_i)\) and \(1/(1-2p_j)\) in the above expression we get after retaining terms to \(O(N^{-4})\) only,

\[
P_{ij} = 2p_i p_j \left[ 1 + \left( \sum p_i^2 + \sum p_j^2 \right) \right] + \left( \sum p_i^2 + \sum p_j^2 \right) - 2 \sum p_t^2 \]

\[
- \left( p_i + p_j \right) \cdot \sum p_t^2 + \left( \sum p_t^2 \right)^2 \] (2.3.52)

Case (ii): \(n \geq 3\):

From (2.3.2) and (2.3.44) we get, after multiplying and retaining terms to \(O(N^{-4})\)

\[
K_{ij} = \left[ 1 + \left( n(p_i + p_j) - \sum p_t^2 \right) \right] + \left( \sum p_i^2 + \sum p_j^2 + p_i p_j \right)

- \left( \frac{C_2}{C_0} \sum p_t^3 + \left( \frac{C_1}{C_0} \right)^2 + \left( \frac{C_3}{C_0} \right)^2 \right) n \cdot \left( p_i + p_j \right) \sum p_t^2 \] (2.3.53)

Now, it can be seen that

\[
\frac{1}{C_0} = (n-1)!

\[
\frac{C_1}{C_0} = \frac{n(n-1)}{2}

\[
\frac{C_2}{C_0} = \frac{n(n-1)(n+1)}{3}

\[
\frac{C_3}{C_0} = \frac{n(n-1)(n-2)(n+1)}{8}

and
\[
\frac{C_1^2}{C_0^2} - \frac{C_3}{C_0} = \frac{n(n-1)(n^2-n+2)}{8}
\]

Substituting these values in (2.3.58) we get

\[
K_n \cdot \lambda_i \lambda_j = (n-1)!p_i p_j [1 + \{n(p_i + p_j) - \frac{n(n-1)}{2} \cdot \Sigma p_t^2\} + \frac{n(n-1)(n^2-n+2)}{8} \cdot \Sigma p_t^2 + n^2(p_i^2 + p_j^2) - \frac{n(n-1)(n+1)}{3} \cdot \Sigma p_t^3 + n^2p_i p_j - \frac{n^2(n-1)}{2} \cdot (p_i + p_j) \cdot \Sigma p_t^2]}
\]

using this expression and the expression for \(\phi_{ij}\) from (2.3.51) we get after simplifying, when terms to \(O(N^{-4})\) are retained,

\[
P_{ij} = n(n-1)p_i p_j [1 + \{p_i + p_j - \Sigma p_t^2\} + \{2(p_i^2 + p_j^2) - 2\Sigma p_t^3\} - (n-2)p_i p_j + (n-3) \cdot (p_i + p_j) \Sigma p_t^2 - (n-3)(\Sigma p_t^2)^2)]
\]

(2.3.54)

Even though this expression is derived for the case \(n \geq 3\), Equation (2.3.52) shows that (2.3.54) is valid for the value \(n=2\), also. Thus the expression for \(P_{ij}\) correct to \(O(N^{-4})\) under Sampford's scheme of selection for samples of size \(n\) is given by (2.3.54) which is true for all \(n \geq 2\).
A check on (2.3.54) is provided by verifying that

\[ P_{ij} = \frac{n(n-1)}{N^2}[1 + \frac{1}{N} + \frac{1}{N^2}] \]

when all \( p_i \) are equal to \( \frac{1}{N} \), which is the expression for \( P_{ij} \) correct to \( O(N^{-4}) \) in the case of simple random sampling without replacement. A more thorough check on (2.3.54) is provided by verifying that

\[ \sum_{j \neq i} P_{ij} = (n-1)P_i \]

satisfied to \( O(N^{-3}) \) which confirms that (2.3.54) is correct to \( O(N^{-4}) \).

2.3.4. Evaluation of the variance expression to \( O(N^0) \)

We will first prove a theorem which is applicable for various without replacement schemes as we will be seeing later in this chapter as well as in the subsequent chapters.

**Theorem 2.8:**

Given any varying probability sampling scheme for selecting a sample of size \( n \) whose \( P_i \) is given by

\[ P_i = np_i \quad (2.3.55) \]

and whose \( P_{ij} \) correct to \( O(N^{-4}) \) is given by

\[ P_{ij} = n(n-1)p_ip_j[1 + ((p_i + p_j) - \Sigma p_t^2) + 2(p_i^2 + p_j^2) - 2\Sigma p_t^3 \\
+ a_np_ip_j - (a_n + 1)(p_i + p_j) \cdot \Sigma p_t^2 + (a_n + 1)(\Sigma p_t^2)^2] \quad (2.3.56) \]

for some constant \( a_n \) that does not depend on \( p_t \)'s but may
depend on \( n \), the variance expression correct to \( O(N^0) \) of the corresponding H.T. estimator is given by

\[
V(\hat{\gamma}_{HT}) = \frac{1}{n} \left[ \sum p_i z_i^2 - (n-1) \sum p_i^2 z_i^2 \right] - \frac{(n-1)}{n} \left[ 2 \sum p_i^3 z_i^2 - \sum p_i^2 \sum p_i^2 z_i^2 - \frac{a_n}{n} \cdot (\sum p_i^2 z_i^2) \right]
\]

(2.3.57)

where \( z_i \) is given in (2.2.12).

**Proof:**

Variance of the H.T. estimator is given by

\[
V(\hat{\gamma}_{HT}) = \sum y_i^2/p_i + \sum \sum p_{ij}/p_i p_j \cdot y_i y_j - y^2
\]

(2.3.58)

From (2.3.55) we have

\[
\sum y_i^2/p_i = \sum y_i^2/np_i
\]

(2.3.59)

Also by using (2.3.55) and (2.3.56) we get

\[
\begin{align*}
\sum \sum \frac{p_{ij}}{p_i p_j} \cdot y_i y_j \\
= \sum \sum \frac{(n-1)}{n} \cdot \left[ 1+\{(p_i+p_j)-\sum p_t^2\}+\{2(p_i^2+p_j^2)\} \right. \\
- 2\sum p_t^2 + a_n p_i p_j - (a_n+1)(p_i+p_j) \sum p_t^2 \\
+ \left. (a_n+1) \cdot (\sum p_t^2)^2 \right] \cdot y_i y_j
\end{align*}
\]
\[
\frac{(n-1)}{n} \left[ (1+p_i - \Sigma \rho_t^2)^2 - 2p_i^2 - 2\Sigma \rho_t^3 - (a_n + 1)p_i \Sigma \rho_t^2 \right. \\
+ (a_n + 1) (\Sigma \rho_t^2)^2 \cdot y_i (y - y_i) \\
+ \{ y_i (\Sigma \rho_t^2 - p_i y_i) + 2y_i (\Sigma \rho_t^2 - p_i^2 y_i) \\
+ a_n p_i y_i (\Sigma \rho_t^2 - p_i y_i) - (a_n + 1) \cdot (\Sigma \rho_t^2 - p_i y_i) \cdot y_i \cdot \Sigma \rho_t^2 \} \right]
\]

\[
= \frac{(n-1)}{n} \cdot \left[ (1 - \Sigma \rho_t^2 - 2\Sigma \rho_t^3 + (a_n + 1) (\Sigma \rho_t^2)^2) \cdot (y^2 - \Sigma y_i^2) \\
+ y \cdot \{ \Sigma \rho_t y_t + 2\Sigma \rho_t^2 y_t - (a_n + 1) \Sigma \rho_t y_t \} \\
- \Sigma \rho_t y_t^2 - 2\Sigma \rho_t^2 y_t^2 + (a_n + 1) \Sigma \rho_t^2 \cdot \Sigma \rho_t y_t^2 \\
+ y \cdot \Sigma \rho_t y_t - \Sigma \rho_t y_t^2 + 2y \cdot \Sigma \rho_t^2 y_t - 2\Sigma \rho_t^2 y_t^2 \\
+ a_n (\Sigma \rho_t y_t)^2 - a_n \Sigma \rho_t^2 y_t^2 - (a_n + 1) \cdot y \cdot \Sigma \rho_t^2 \cdot \Sigma \rho_t y_t \\
+ (a_n + 1) \Sigma \rho_t^2 \cdot \Sigma \rho_t y_t^2 \right]
\]

Retaining only terms to \(O(N^0)\), we get

\[
\sum_{i=1}^{p} \sum_{j=1}^{p} \frac{\rho_{ij}}{\rho_{ij}} \cdot y_i y_j = \frac{(n-1)}{n} \cdot \left[ y^2 - (y^2 \Sigma \rho_t^2 - 2y \cdot \Sigma \rho_t y_t + \Sigma y_t^2) \\
+ \{ (a_n + 1) y^2 (\Sigma \rho_t^2)^2 - 2y^2 \Sigma \rho_t^3 \\
+ 4y \cdot \Sigma \rho_t^2 y_t + \Sigma \rho_t^2 \cdot \Sigma y_t^2 - 2(a_n + 1) \cdot y \cdot \Sigma \rho_t^2 \cdot \Sigma \rho_t y_t \\
- 2\Sigma \rho_t y_t^2 + a_n (\Sigma \rho_t y_t)^2 \} \right]
\]

\(\text{(2.3.60)}\)
Substituting in (2.3.58) the values in (2.3.59) and (2.3.60) we get, after simplifying and putting in suitable form, the variance correct to $O(N^0)$ as

$$V(\hat{Y}_{HT}) = \frac{1}{n} \left[ \sum P_i z_i^2 - (n-1) \sum P_i^2 z_i^2 \right]$$

$$- \frac{(n-1)}{n} \cdot \left[ 2 \sum P_i z_i^2 - \sum P_i^2 \cdot \sum P_i^2 z_i^2 - a_n \cdot (\sum P_i^2 z_i^2)^2 \right]$$

(2.3.61)

Q.E.D.

From (2.3.61), $V(\hat{Y}_{HT})$ correct to $O(N^2)$ is given by

$$V(\hat{Y}_{HT}) = \frac{1}{n} \sum P_i z_i^2$$

which is equal to the variance of the customary estimator in probability proportional to size with replacement sampling.

$$V(\hat{Y}_{HT})$$ correct to $O(N^1)$ is given by

$$V(\hat{Y}_{HT}) = \frac{1}{n} \left[ \sum P_i z_i^2 - (n-1) \sum P_i^2 z_i^2 \right]$$

$$= \frac{1}{n} \left[ \sum P_i \left[ 1 - (n-1)P_i \right] z_i^2 \right]$$

(2.3.62)

which clearly shows the reduction in variance compared to the with replacement procedure. Thus $\frac{(n-1)}{n} \sum P_i^2 z_i^2$ represents the principal gain of the without replacement procedure relative to the with replacement scheme.

In Subsection 2.2.2 we have mentioned that Hartley and Rao derived the expression for $P_{ij}$ of the Goodman and Kish procedure correct to $O(N^{-4})$ which is given by (2.2.14).
We can observe that (2.2.14) is the same as (2.3.56) with $a_n=2$ and thus Theorem 2.8 is applicable to the Goodman and Kish procedure.

Also we observe that (2.3.54) is the same as (2.3.56) with $a_n=-(n-2)$.

Thus Theorem 2.8 is also applicable to Sampford's procedure. Thus using Theorem 2.8 we can compare the efficiencies of the H.T. estimator under the procedures of (i) Goodman and Kish and (ii) Sampford.

**Theorem 2.9:**

When the variance is considered to $O(N^1)$, the H.T. estimators corresponding to Sampford's procedure and the Goodman and Kish procedure are equally efficient, and when the variance is considered to $O(N^0)$, the H.T. estimator corresponding to the Sampford's procedure is always more efficient than the H.T. estimator corresponding to the Goodman and Kish procedure and the relative gain in precision will be larger for larger sample sizes.

**Proof:**

Since Theorem 2.8 is applicable to both the schemes it follows from (2.3.62) that to $O(N^1)$ the H.T. estimators corresponding to both the schemes are equally efficient.

From (2.3.61) we have correct to $O(N^0)$,
Thus we get by considering the difference

\[ V(\hat{\gamma}_{H.T.})_{G.K} - V(\hat{\gamma}_{H.T.})_{samp} = (n-1) \cdot (\Sigma p_i \Sigma z_i^2) \]

\[ > 0 \]

Thus the estimator corresponding to the Sampford's scheme is always more efficient than the estimator corresponding to the Goodman and Kish procedure.

Also the percentage gain in efficiency is given by

\[ E = \left( \frac{(n-1) \cdot (\Sigma p_i \Sigma z_i^2)}{V(\hat{\gamma}_{H.T.})_{G.K}} \right) \times 100 \]

Thus \( E \) would be an increasing function of the sample size since the numerator increases and the denominator decreases as the sample size increases.

O.E.D.
Thus as a conclusion it is mentioned that one would gain by preferring Sampford's procedure over the Goodman and Kish procedure especially for larger sample sizes.

2.3.5. Numerical illustration

The data we consider here is that of 35 Scottish farms, appearing as Table 5.1 in Sampford (1962) which is reproduced on the following page.

In order to have an idea as to how good the approximate expressions (2.3.56) for \( P_{ij} \) are in a real situation, the \( P_{ij} \) are calculated for the population in Table 1 by using both the exact (2.2.22) expressions and the approximate expressions (2.3.56) for samples of size 3. The variance also is evaluated using both the sets of \( P_{ij} \). The set of probabilities \( p_{ij} \) \((j = 2,3,...35)\) are tabulated in Table 2.2 along with the corresponding approximate \( P_{ij} \) \((j = 2,...35)\).

The variance calculated using the exact \( P_{ij} \) is found to be

\[ V(\hat{y}) = 68318.56 \]

whereas the variance computed using the approximate \( P_{ij} \) is found to be

\[ V(\hat{y}) = 68341.43 \]
Table 2.1. Recorded acreage of crops and grass for 1947 and acreage under oats in 1957, for 35 farms in Orkney

<table>
<thead>
<tr>
<th>Farm Number</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recorded crops + grass $x_i$</td>
<td>50</td>
<td>50</td>
<td>52</td>
<td>58</td>
<td>60</td>
<td>60</td>
<td>62</td>
<td>65</td>
<td>65</td>
<td>68</td>
<td>71</td>
<td>74</td>
<td>78</td>
<td>90</td>
</tr>
<tr>
<td>Oats 1957 $y_i$</td>
<td>17</td>
<td>17</td>
<td>10</td>
<td>16</td>
<td>6</td>
<td>15</td>
<td>20</td>
<td>18</td>
<td>14</td>
<td>20</td>
<td>24</td>
<td>18</td>
<td>23</td>
<td>0</td>
</tr>
<tr>
<td>Farm number $x_i$</td>
<td>91</td>
<td>92</td>
<td>96</td>
<td>110</td>
<td>140</td>
<td>140</td>
<td>156</td>
<td>156</td>
<td>190</td>
<td>198</td>
<td>209</td>
<td>240</td>
<td>274</td>
<td>300</td>
</tr>
<tr>
<td>$y_i$</td>
<td>27</td>
<td>34</td>
<td>25</td>
<td>24</td>
<td>43</td>
<td>48</td>
<td>44</td>
<td>45</td>
<td>60</td>
<td>63</td>
<td>70</td>
<td>28</td>
<td>62</td>
<td>59</td>
</tr>
<tr>
<td>Farm number $x_i$</td>
<td>29</td>
<td>30</td>
<td>31</td>
<td>32</td>
<td>33</td>
<td>34</td>
<td>35</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$y_i$</td>
<td>303</td>
<td>311</td>
<td>324</td>
<td>330</td>
<td>356</td>
<td>410</td>
<td>430</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$y_i$</td>
<td>66</td>
<td>58</td>
<td>128</td>
<td>38</td>
<td>69</td>
<td>72</td>
<td>103</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 2.2. Exact $P_{lj}$'s and the approximate $P_{lj}$'s of the Sampford's procedure with $n=3$ for the data in Table 2.1

<table>
<thead>
<tr>
<th>Unit no. $j$</th>
<th>Exact $P_{lj}$</th>
<th>Approximate $P_{lj}$</th>
<th>Unit no. $j$</th>
<th>Exact $P_{lj}$</th>
<th>Approximate $P_{lj}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>.000439</td>
<td>.000439</td>
<td>20</td>
<td>.001249</td>
<td>.001250</td>
</tr>
<tr>
<td>3</td>
<td>.000546</td>
<td>.000510</td>
<td>21</td>
<td>.001396</td>
<td>.001397</td>
</tr>
<tr>
<td>4</td>
<td>.000527</td>
<td>.000528</td>
<td>22</td>
<td>.001396</td>
<td>.001397</td>
</tr>
<tr>
<td>5</td>
<td>.000527</td>
<td>.000528</td>
<td>23</td>
<td>.001712</td>
<td>.000713</td>
</tr>
<tr>
<td>6</td>
<td>.000527</td>
<td>.000528</td>
<td>24</td>
<td>.001787</td>
<td>.001788</td>
</tr>
<tr>
<td>7</td>
<td>.000545</td>
<td>.000546</td>
<td>25</td>
<td>.001891</td>
<td>.001891</td>
</tr>
<tr>
<td>8</td>
<td>.000572</td>
<td>.000572</td>
<td>26</td>
<td>.002185</td>
<td>.002185</td>
</tr>
<tr>
<td>9</td>
<td>.000572</td>
<td>.000572</td>
<td>27</td>
<td>.002512</td>
<td>.002512</td>
</tr>
<tr>
<td>10</td>
<td>.000599</td>
<td>.000599</td>
<td>28</td>
<td>.002765</td>
<td>.002765</td>
</tr>
<tr>
<td>11</td>
<td>.000625</td>
<td>.000626</td>
<td>29</td>
<td>.002794</td>
<td>.002794</td>
</tr>
<tr>
<td>12</td>
<td>.000652</td>
<td>.000653</td>
<td>30</td>
<td>.002873</td>
<td>.002873</td>
</tr>
<tr>
<td>13</td>
<td>.000688</td>
<td>.000688</td>
<td>31</td>
<td>.003001</td>
<td>.003001</td>
</tr>
<tr>
<td>14</td>
<td>.000796</td>
<td>.000796</td>
<td>32</td>
<td>.003061</td>
<td>.003060</td>
</tr>
<tr>
<td>15</td>
<td>.000804</td>
<td>.000805</td>
<td>33</td>
<td>.003321</td>
<td>.003319</td>
</tr>
<tr>
<td>16</td>
<td>.000813</td>
<td>.000814</td>
<td>34</td>
<td>.003870</td>
<td>.003866</td>
</tr>
<tr>
<td>17</td>
<td>.000850</td>
<td>.000850</td>
<td>35</td>
<td>.004077</td>
<td>.004072</td>
</tr>
<tr>
<td>18</td>
<td>.000976</td>
<td>.000977</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>.001249</td>
<td>.001250</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
which suggests that in many practical situations the approximations will serve the purpose quite adequately.

In Table 2.3 are presented the variances computed to various orders for both the schemes of (i) Sampford and (ii) Goodman and Kish when samples of size 4 are considered.

Table 2.3. Approximations to $V(\hat{Y}_{H.T.})$

<table>
<thead>
<tr>
<th>Order of Approximation</th>
<th>Sampfords Procedure</th>
<th>Goodman and Kish Procedure</th>
</tr>
</thead>
<tbody>
<tr>
<td>$O(N^2)$</td>
<td>55852.450</td>
<td>55852.450</td>
</tr>
<tr>
<td>$O(N^1)$</td>
<td>49320.940</td>
<td>49320.940</td>
</tr>
<tr>
<td>$O(N^0)$</td>
<td>48952.190</td>
<td>48979.130</td>
</tr>
</tbody>
</table>

The value computed to $O(N^2)$ represents the true variance of the customary estimator in the varying probability with replacement scheme. Values of the successive approximations suggest that the convergence is quite satisfactory even though the population size, $N=35$, is much smaller than the sizes we actually come across in practice. For larger population sizes, the relative difference is however expected to be higher than it is in this example.
2.4. Hanurav's Procedure

Hanurav (1967) presented an unequal probability sampling scheme for sample size 2 which satisfies the condition \( P_i = 2p_i \). Vijayan (1968) has extended this procedure for sample size \( n \geq 2 \). The procedure for general sample size is much too complicated to adopt in practice and thus we consider the simple case of sample size two only. The scheme is described as follows:

Two units are selected with probabilities \( \{p_i\} \), \( i = 1, 2, \ldots, N \), with replacement and if the two units are distinct, the sample is retained, otherwise this sample is rejected and another sample of two units is selected with probabilities proportional to \( \{p_i^2\} \) and with replacement. If the two units selected are distinct, the sample is retained; otherwise a further sample of two units is selected with probabilities proportional to \( \{p_i^4\} \) and with replacement, and so on.

Hanurav has shown that under some restrictions this procedure terminates with probability one and that the expressions for \( P_i \) and \( P_{ij} \) are given by

\[
P_i = 2p_i \quad \quad (2.4.1)
\]

and

\[
P_{ij} = 2p_ip_j[1 + \sum_{k=1}^{\infty} w_k], \quad (2.4.2)
\]
where
\[ w_k = \frac{(p_i p_j)^{2k-1}}{S(1)S(2)\cdots S(k)}, \quad (2.4.3) \]
and
\[ S(r) = \sum_{t=1}^{N} p_t^r, \quad (2.4.4) \]

Since the expression for \( P_{ij} \) involves an infinite series it is not possible in practice to get the exact variance of the corresponding H.T. estimator. However for large values of \( N \), we can derive the approximate expressions for \( P_{ij} \) by assuming that \( p_i \) is of \( O(N^{-1}) \). From (2.4.3) and (2.4.4) it can be easily seen that each \( S(r) \) is of \( O(N^{-2r+1}) \) and consequently each \( w_k \) will be of \( O(N^{-k}) \).

Thus the expression for \( P_{ij} \) correct to \( O(N^{-4}) \) is

\[
P_{ij} = 2p_i p_j [1 + w_1 + w_2]
= 2p_i p_j [1 + \frac{p_i p_j}{\sum p_t} + \frac{p_i^3 p_j^3}{\sum p_t \cdot \sum p_t^2}]. \quad (2.4.5)
\]

Equations (2.4.1) and (2.4.5) do not satisfy the conditions of Theorem 2.8 because \( P_{ij} \) given in (2.4.5) is not of the form given in (2.3.56) and hence the theorem is not applicable here. However, substituting from (2.4.1) and (2.4.5) in the expression

\[
\bar{\nu}^2_{(\text{H.T.})} = \frac{\bar{y}_i^2}{\sum \frac{P_i}{p_i}} + \sum \frac{P_{ij}}{p_i p_j} \bar{y}_i \bar{y}_j - \bar{x}^2,
\]
we get after much simplification the variance to $O(N^0)$ of the H.T. estimator corresponding to the Hanurav's procedure as

$$V(\hat{\Phi}_{\text{H.T.}})_H = \frac{1}{2} \Sigma_{i=1}^{p} z_i^2 - \frac{1}{2}[\Sigma_{i=1}^{p} z_i^2 - \frac{\left(\Sigma_{i=1}^{p} z_i\right)^2}{\Sigma_{i=1}^{p}}]$$

$$- \frac{1}{2} \left[\Sigma_{i=1}^{p} z_i^2 - \frac{\left(\Sigma_{i=1}^{p} z_i\right)^2}{\Sigma_{i=1}^{p}}\right]$$

(2.4.6)

where

$$z_i = \frac{y_i}{p_i} - \bar{y}$$

variance correct to $O(N^1)$ is

$$V(\hat{\Phi}_{\text{H.T.}})_H = \frac{1}{2} \Sigma_{i=1}^{p} z_i^2 - \frac{1}{2}[\Sigma_{i=1}^{p} z_i^2 - \frac{\left(\Sigma_{i=1}^{p} z_i\right)^2}{\Sigma_{i=1}^{p}}]$$

(2.4.7)

Variance correct to $O(N^1)$ of the H.T. estimator corresponding to the Goodman and Kish procedure as given in (2.2.16) is

$$V(\hat{\Phi}_{\text{H.T.}})_{G.K} = \frac{1}{2} \Sigma_{i=1}^{p} z_i^2 - \frac{1}{2} \Sigma_{i=1}^{p} z_i^2$$

(2.4.8)

From (2.4.7) and (2.4.8) we get

$$V(\hat{\Phi}_{\text{H.T.}})_H - V(\hat{\Phi}_{\text{H.T.}})_{G.K} = \frac{1}{2} \cdot \frac{\left(\Sigma_{i=1}^{p} z_i\right)^2}{\Sigma_{i=1}^{p}} \geq 0$$

(2.4.9)

showing that for large $N$, the H.T. estimator corresponding to G.K. procedure is always more efficient than the one
corresponding to Hanurav's procedure. However, for moderately large populations one has to consider the variance to $O(N^0)$ and no valid conclusion can be drawn from the comparison of $V(\hat{Y}_{H.T.})_H$ to $O(N^0)$ given in (2.4.6) with $V(\hat{Y}_{H.T.})_G.K$ given in (2.2.16) for sample size two.
3. RAO, HARTLEY AND COCHRAN'S PROCEDURE

3.1. Introduction

As there are not many schemes of unequal probability sampling without replacement that are simple to adopt in practice and are applicable for sample sizes \( n^2 \), Rao, Hartley and Cochran suggested a simple procedure which is applicable for sample size \( n^2 \) and provided an unbiased estimator for the population total. However, it is often commented by several authors that the estimator most often turns out to be inefficient relative to some estimators under other unequal probability sampling procedures that are existent in the literature. In this chapter a mathematical proof has been given showing the inadmissibility of the Rao, Hartley, Cochran (R.H.C.) estimator and several other alternative estimators under the Rao, Hartley, Cochran (R.H.C.) scheme are suggested which are almost always more efficient than the other existing estimators under unequal probability sampling procedures. The efficiency of these proposed estimators in relation to the other existing estimators is illustrated numerically by considering several populations that are considered in the literature as the most suitable data for the unequal probability sampling procedures.
3.2. Rao, Hartley and Cochran's Procedure

The procedure of unequal probability sampling without replacement proposed by Rao, Hartley and Cochran for selecting a sample of size n is described as follows:

(i) split the population at random into n groups of sizes \( N_1, N_2, \ldots, N_n \) where \( N_1 + N_2 + \ldots + N_n = N \) and,

(ii) select one unit with probability proportional to \( p_i \) from each of these n groups independently.

The primary advantage of this scheme compared to the other without replacement unequal probability sampling procedures is that it does not require heavy computations for drawing the sample even for sample size \( n > 2 \) and thus is very simple to operate.

Let \( \sum_{i=1}^{n} p_i = S_i \), say

For the above scheme of sampling Rao, Hartley and Cochran proposed

\[
T_1 = \sum_{i=1}^{n} \frac{y_i}{p_i} \cdot S_i \tag{3.2.1}
\]

as an estimator for the population total \( Y \), where \( y_i \) is the value of the unit selected in the \( i \)-th group and \( p_i \) is the corresponding probability.
Theorem 3.1:
Under the R.H.C. scheme of sampling the estimator $T_1$ is unbiased for estimating the population total $Y$ and the variance of $T_1$ is given by

$$V(T_1) = \frac{1}{\binom{N}{N-1}} \left( \sum_{i=1}^{N} N_i^2 - N \right) \cdot \left[ \sum_{t=1}^{N} \frac{Y_t^2}{p_t} - Y^2 \right]$$

which attains its minimum when $N_1 = N_2 = \ldots = N_n = \frac{N}{n}$. 

Proof:
The details of the proof are given in the paper by Rao, Hartley and Cochran, and thus are not furnished here. Q.E.D.

Hereafter in this chapter we will assume for the sake of mathematical tractability and for the comparison purposes that $N$ is a multiple of $n$. Also we will make the choice

$$N_1 = N_2 = \ldots = N_n = \frac{N}{n} = M, \text{ say}$$

Under these assumptions (3.2.2) reduces to

$$V(T_1) = \left(1 - \frac{n-1}{N-1}\right) \cdot \frac{1}{n} \left(\sum_{t=1}^{N} \frac{Y_t^2}{p_t} - Y^2\right),$$

which clearly shows the reduction in the variance as compared to sampling with replacement estimator.
3.3. Inadmissibility of the R.H.C. Estimator

For simplicity of notation let the elements of the population \( U \) be represented by integers \( 1, 2, \ldots, N \); \( U = \{1, 2, \ldots, N\} \). Let \( s \) denote a typical subset of \( U \). Now, depending on the specific sampling scheme used there could possibly be various ways of expressing the outcome of the sampling experiment. Sometimes the outcome of the experiment, denoted by \( \omega \), can be described as \( \omega_0 = (s, y) \) where \( s \) is the subset of \( U \) that has been selected and \( y \) is the vector of corresponding \( y \)-values written in the same order as the elements of \( s \). For example, in the case of ordinary systematic sampling \( \omega \) is described in this way. In some situations \( \omega \) could possibly be described in a more detailed way. For example \( \omega \) can be described as \( \omega = (s', y') \) where \( s' \) is the ordered subset of \( U \) that is selected, the order being the order in which the units are selected and \( y' \) is the vector of corresponding \( y \)-values that appear in \( s' \). Thus \( s' \) fully describes the unit by unit sampling without replacement. In some situations \( \omega \) can also be described as \( \omega = (s'', y'') \) where \( s'' \) is the sequence that is selected all members of which belong to \( U \) and \( y'' \) is the vector of corresponding \( y \)-values. In \( s'' \) if we remove all the members that appear in some of the preceding places we get the ordered set \( s' \).
That is if $s'' = (3,5,2,3,2)$ then the corresponding ordered set is given by $s' = (3,5,2)$. Similarly if we ignore the ordering in $s'$ we get a set $s$. For example if $s' = (3,5,2)$ then the corresponding $s = (2,3,5)$. Thus $s'$ is an abstract function of $s''$ and $s$ is an abstract function of $s'$ and consequently also an abstract function of $s''$. Symbolically we can write $s' = f_1(s''), s = f_2(s') = f_2(f_1(s'')) = f_3(s'')$.

Thus we see that depending on the sampling scheme adopted there could possibly be different ways of describing the outcome $\omega$ of the sampling experiment. Having defined the outcome $\omega$, we can define an estimate $t$ of a population parameter as some function of the outcome $\omega$. That is $t = t(\omega)$, or equivalently $t = t''(s'', \underline{y}''), t = t'(s', \underline{y}')$, $t = t(s, \underline{y})$ as the case may be. Now, when the outcome $\omega$ is described as $(s'', \underline{y}'')$, the knowledge of $(s'', \underline{y}'')$ is enough to know any $(z'', \underline{y}'')$ such that

$$f_3(z'') = f_3(s'') = s$$

(3.3.1)

To be specific, if $(s'', \underline{y}'') = (2,5,3,6,3; 10,14,16,7,16)$ then for $z'' = (3,6,5,6,2)$ we have $(z'', \underline{y}'') = (3,6,5,6,2;16,7,14,7,10)$. This is because we know from $(s'', \underline{y}'')$ that $y_2 = 10$, $y_5 = 14$, $y_3 = 16$, and $y_6 = 7$. Thus if we have an estimate $t'' = t''(s'', \underline{y}'')$ we can evaluate the estimate $t = t(s, \underline{y})$ given by
\[ t = t(s,y) = \frac{\sum t'(z'',y'') \cdot p(z'')}{\sum p(z'')} \quad (3.3.2) \]

where the summation is over all \( z'' \) such that (3.3.1) holds and \( p(z'') \) is the probability of observing \( (z'',y'') \). Careful observation of (3.3.2) tells us that \( t \) is the conditional mean value of \( t'' \) with respect to \( (s,y) \). Thus we have from (3.3.2) that, \( E(t) = E(t'') \) and if \( t'' \) is unbiased,

\[ \text{Var}(t) = \text{var}(t'') - E(t-t'')^2. \]

So, as an estimate \( t \) is at least as good as \( t'' \). Thus any estimate that is a function of \( (s'',y'') \) can always be improved upon by using the above technique. Similarly estimators that are functions of \( (s',y') \) also can be improved upon by using the same technique. Thus any good estimate is a function of \( (s,y) \). In the case of simple random sampling with replacement and varying probability sampling with replacement the customary estimators are functions of \( (s'',y'') \) ignoring the order in which the units are drawn. In the case of unequal probability sampling without replacement the estimator proposed by Desraj (1956a) is a function of \( (s',y') \). In all these three cases Basu (1958) has shown that the 'order statistic' forms a sufficient statistic, and therefore any estimator which is not a function of the order statistic, can be improved by the use of Rao-Blackwell theorem. In fact the technique used in
(3.3.2) to get the estimate $t(s,y)$ is nothing but Rao-Blackwellisation of $t''(s'',y'')$. Thus the estimators based on distinct units obtained by using Rao-Blackwell theorem in the case of with replacement scheme, and the Basu-Murthy unordered estimator obtained by using Rao-Blackwell theorem to improve Desraj's estimator are uniformly better than the respective customary estimators. The technique of obtaining these estimators is discussed in detail by Murthy (1957), Basu (1958) and Pathak (1961).

In general if a sampling scheme defines probability distributions $P_1(\omega)$ on the outcomes $\omega \in \Omega$, we can define the projection of $P_1(\omega)$ into the space of probability distributions defined on the samples $s$ by

$$P(s) = P_1(\omega | \Omega_s),$$

where $\Omega_s \subseteq \Omega$ consists of all $\omega$'s which result in $(s,y)$.

So we have

$$P_1(\omega) = P(s) \cdot P_2(\omega | \Omega_s), \quad s \subseteq U.$$

Now, for the purpose of estimating the population total, the conditional distributions $P_2(\omega | \Omega_s)$ are not useful which shows that $(s,y)$ is sufficient for the estimation purposes. This is the reason why it is genuine to define a sampling design as a probability distribution $P(s)$ defined on the space of samples $s$ as done by Godambe, Godambe and Joshi,
Hájek and others.

In case of the Rao, Hartley and Cochran's procedure, the authors have implicitly defined the outcome $u^*$ of their procedure as

$$ u^* = (i_1, y_i_1, G_i_1; i_2, y_i_2, G_i_2; \ldots; i_n, y_i_n, G_i_n) \quad (3.3.3) $$

where $s = (i_1, i_2, \ldots, i_n)$ is the subset $s$ of $U$ that is selected, $(y_i_1, y_i_2, \ldots, y_i_n)$ are the corresponding y-values and $G_i_1, G_i_2, \ldots, G_i_n$ are the random groups that contain the units $i_1, i_2, \ldots, i_n$ respectively. The estimator $T_1$ in (3.2.1) proposed by Rao, Hartley and Cochran is a function of $u^*$ and not just a function of $w_0 = (s, y)$, $s$ being the subset of $U$ that is selected. Thus the Rao, Hartley and Cochran's estimator can be improved by using the Rao-Blackwell Theorem. This establishes the inadmissibility of the Rao, Hartley and Cochran's estimator. This has been observed first by Hájek (1964) who also mentioned as a passing remark that further study is necessary in this direction. It seems that this point has been overlooked by other researchers including Pathak (1961, 1964) who dealt in detail with the concept of sufficiency in sampling theory and considered several specific situations where it can be used. The reason for this, perhaps, could be that the estimator $T_1$ outwardly looks to be a function of the subset $s$ of $U$ that has been selected, unlike the
customary estimators in the case of sampling with replacement and the Desraj estimator in the case of sampling with unequal probabilities and without replacement. We will deal in detail with this aspect of improving the Rao, Hartley and Cochran's estimator in a later section.

Since \((s,y)\) is a sufficient statistic any good estimate belongs to the class of estimates that are functions of \((s,y)\) and this class is complete in the sense that for any estimator not belonging to this class there exists a corresponding estimate belonging to this class which is uniformly better. As is well known, H.T. estimator is a member of this class. Also in view of the admissible property of the H.T. estimator in the class of all unbiased estimators proved by Godambe and Joshi (1965) it is interesting to investigate the properties of the H.T. estimator under R.H.C. scheme.

3.4. Horvitz-Thompson Estimator under Rao, Hartley and Cochran Scheme

3.4.1. Definitions, notations and basic results pertaining to randomization

In order to study the properties of the H.T. estimator under the R.H.C. scheme one must first solve the two problems: (i) to find the relation between the inclusion probabilities \(P_1, P_2, \ldots, P_N\) and the initial probabilities, \(P_1, P_2, \ldots, P_N\) and (ii) to find the relation between the probabilities \(P_{ij} (1 \leq i \neq j)\) and the initial probabilities.
In this section we will consider these two problems of evaluating $P_i$ $(i = 1, 2, \ldots N)$ and $P_{ij}$ $(1 < i \neq j < N)$ in terms of $P_1, P_2, \ldots P_N$. Let $\mathcal{U} = \{U_1, U_2, \ldots U_N\}$ denote the set of all the population units and let $G$ denote a typical group of $M$ units out of $N$ units. There are in fact $\binom{N}{M}$ such groups. Let $\mathcal{Q} = \{G_1, G_2, \ldots G_{\binom{N}{M}}\}$ be the set of all such groups.

**Definition 3.1:**

An ordered $n$-tuple $\varphi = (G_{i_1}, G_{i_2}, \ldots G_{i_n})$ is said to be a partition of the population $\mathcal{U}$ if

- $G_{i_j} \in \mathcal{Q}, j = 1, 2, \ldots n$
- $G_{i_j} \cap G_{i_{j'}} = \emptyset, j \neq j'$
- $\mathcal{U} = \bigcup_{j=1}^{n} G_{i_j}$

**Definition 3.2:**

Two partitions $\varphi_i = (G_{i_1}, G_{i_2}, \ldots G_{i_n})$ and $\varphi'_i = (G_{i_1}', G_{i_2}', \ldots G_{i_n}')$ of the population $\mathcal{U}$ are said to be equivalent if one is just a rearrangement of the other, i.e., if each $G_{i_j}$ is some $G_{i_j}'$, and vice versa.
Definition 3.3:

Two partitions \( \varnothing_1 \) and \( \varnothing_2 \) are said to be distinct if they are not equivalent.

Theorem 3.2:

The total number, \( A \), of distinct partitions of the population \( \mathcal{U} \) with groups of size \( M \) each is \( \frac{N!}{n! (M!)^n} \).

Proof:

Total number of ways of selecting the first group = \( \binom{N}{M} \).

Total number of ways of selecting the second group having selected the first group = \( \binom{N-M}{M} \).

In general, total number of selecting the jth group having selected the first \((j-1)\) groups = \( \binom{N-J+1}{M} \), \( j = 2, 3, \ldots, n \). Therefore total number of possible partitions

\[
= \binom{N}{M} \cdot \binom{N-M}{M} \cdot \binom{N-2M}{M} \cdots \binom{2M}{M} \\
= \frac{N!}{(M!)^n}
\]

Therefore the total number of distinct partitions, \( A \) is given by

\[
A = \frac{N!}{n! (M!)^n} \quad \text{(3.4.1)}
\]

Q.E.D.
Let \( G = \{ \theta_1, \theta_2, \ldots, \theta_A \} \) denote the set of all distinct partitions.

**Theorem 3.3:**

The total number, \( A_1 \), of distinct partitions of the population \( U \) with groups of size \( M \) each such that a particular pair of units \((U_i, U_j)\) falls in the same group is given by

\[
A_1 = \frac{(N-2)!}{(n-1)!(M-2)!(M!)^{n-1}}.
\]

**Proof:**

The group that contains the pair of units \((U_i, U_j)\) can be formed in a total number of \( \binom{N-2}{M-2} \) ways.

Given this group, the total number of distinct partitions that can be made of the rest of the units into groups of size \( M \) each is given by \( \frac{(N-M)!}{(n-1)!(M!)^{n-1}} \) which follows from Theorem 3.2. Therefore the total number of possible distinct partitions such that the pair \((U_i, U_j)\) falls in the same group is given by

\[
A_1 = \frac{(N-2)!}{(n-1)!(M-2)!(M!)^{n-1}}.
\]

Since this number does not depend on the particular pair \((U_i, U_j)\) we are justified in denoting this number by \( A_1 \).

Q.E.D.

Let \( G_1(i,j) = \{ \theta_1, \theta_2, \ldots, \theta_{A_1} \} \) denote the set of all distinct partitions such that \((U_i, U_j)\) is in the same group.
Theorem 3.4:

The total number, $A_2$, of distinct partitions of the population $\mathcal{U}$ with groups of size $M$ each such that a particular pair of units $(U_i, U_j)$ falls in different groups is given by

$$
\frac{(N-2)!}{(n-2)!(M-1)!^2(M^n)!^{n-2}}.
$$

Proof:

The two groups that contain the $i$th and $j$th units can be formed in $\binom{N-2}{M-1} \cdot \binom{N-M-1}{M-1}$ ways. Given these two groups, the total number of distinct partitions that can be made of the rest of the units into groups of size $M$ each is given by

$$
\frac{(N-2M)!}{(n-2)!(M^n)!^{n-2}}
$$

which follows from Theorem 3.2.

Therefore the total number of distinct partitions that can be made such that the pair of units $(U_i, U_j)$ fall in different groups is given by,

$$
A_2 = \frac{(N-2)(N-M-1) \cdot (N-2M)!}{(n-2)!(M^n)!^{n-2}}
$$

(3.4.3)

Since this number does not depend on the particular pair $(U_i, U_j)$ we are justified in denoting this number by $A_2$.

Q.E.D.

Let $\mathcal{Q}_2(i,j) = \{\varnothing_1, \varnothing_2, \ldots \varnothing_{A_2}\}$ denote the set of all
distinct partitions such that \((U_i, U_j)\) are in different groups.

From the way they have been defined, the following relations among \(G_1(i,j), G_2(i,j)\) and \(G\) are immediate

\[ G_1(i,j) \cup G_2(i,j) = G \]

and

\[ G_1(i,j) \cap G_2(i,j) = \emptyset \]

An obvious check on formulas (3.4.1)-(3.4.3) is provided by the relation

\[ A_1 + A_2 = A \quad (3.4.4) \]

Considering the R.H.C. scheme, the procedure of randomly dividing the population into \(n\) groups amounts to choosing at random a partition \(\Theta\) from \(G\), the set of all possible distinct partitions.

3.4.2. Exact expressions for \(P_i\) and \(P_{ij}\) under R.H.C. scheme

**Theorem 3.5:**

The probability \(P_i\) of including the \(i\)th population unit in the sample under R.H.C. scheme is given by

\[ P_i = \frac{1}{A} \sum_{\Theta \in G} p_i/S(\Theta, i), \] where \(S(\Theta, i)\)

denotes the sum of the \(p_i\)'s of all the units in the group, that contains the \(i\)th unit, of the partition \(\Theta\), and the summation is over all the partitions \(\Theta\) belonging to \(G\).
Proof:

From the elementary definition of probability, we have, 
probability of including the ith unit in the sample

\[ P_i = \sum_{\Theta \in \mathcal{G}} \frac{\text{Prob. of selecting the partition } \Theta}{\text{Prob. of selecting the ith unit/ the partition } \Theta} \]

\[ = \sum_{\Theta \in \mathcal{G}} \frac{1}{A} \cdot \frac{P_i}{S(\Theta, i)} \]

\[ = \frac{1}{A} \cdot \sum_{\Theta \in \mathcal{G}} \frac{P_i}{S(\Theta, i)} \]

(3.4.5)

Q.E.D.

Theorem 3.6:

The probability \( P_{ij} \) of including the pair of units \((U_i, U_j)\) in the sample under R.H.C. scheme is

\[ P_{ij} = \frac{1}{A} \cdot \sum_{\Theta \in \mathcal{G}_2(i, j)} \frac{P_i P_j}{S(\Theta, i) S(\Theta, j)} \]

where the summation runs over all the partitions belonging to \( \mathcal{G}_2(i, j) \).

Proof:

For the inclusion probability \( P_{ij} \) we have
\( P_{ij} = \text{probability of including the pair of units} \)

\((U_i, U_j)\)

\[ = \sum_{\Theta \in \mathcal{G}} \left[ \text{Prob. of selecting the partition } \Theta \right] \times \]

\[ \left[ \text{Prob. of selecting the pair } (U_i, U_j)/ \right. \]

\[ \text{the partition } \Theta \left. \right] \]

\[ = \sum_{\Theta \in \mathcal{G}_1(i, j)} \left[ \text{Prob. of selecting the partition } \Theta \right] \times \]

\[ \left[ \text{Prob. of selecting the pair } (U_i, U_j)/ \right. \]

\[ \text{the partition } \Theta \left. \right] \]

\[ + \sum_{\Theta \in \mathcal{G}_2(i, j)} \left[ \text{Prob. of selecting the partition } \Theta \right] \times \]

\[ \left[ \text{Prob. of selecting the pair } \right. \]

\[ (U_i, U_j)/ \text{the partition } \Theta \left. \right] \]

\[ = \frac{1}{A} \cdot \sum_{\Theta \in \mathcal{G}_2(i, j)} \frac{P_i P_j}{S(\Theta, i) \cdot S(\Theta, j)} \]

\[(3.4.6)\]

Since the first term is obviously zero.

Q.E.D.

Having obtained the expressions for \( P_i \) and \( P_{ij} \) we can talk of the H.T. estimator under R.H.C. scheme, its variance and the Yates-Grundy variance estimator. The expressions for \( P_i \) and \( P_{ij} \) are quite easy to evaluate for moderately large values of \( n \) and \( N \) using the now prevailing computer facilities. However, in order to be useful
in studying the estimator's relative performance in relation
to some other existing strategies where by a strategy we mean
a sampling scheme together with an estimator, we have
derived in Subsections 3.4.3 and 3.4.4 the approximate
expressions for $P_i$ and $P_{ij}$ under some regularity assumptions.

Hartley and Rao (1962) have derived approximate ex­
pressions for $P_{ij}$ and hence to the variance expression of
the H.T. estimator for the randomized systematic sampling
proposed by Goodman and Kish, using an asymptotic theory
which is applicable for large and moderate $N$. Using the same
asymptotic theory, by assuming that $p_i$ is of $O(N^{-1})$ and $N$ is
much larger than $n$, Rao (1963) has derived the variance
expressions for the schemes of Durbin (1953) and Yates and
Grundy (1953) to $O(N^0)$ for sample size 2 and to $O(N^1)$ for
sample size $n>2$. We use here the same technique by assuming
that $p_i$ is of $O(N^{-1})$, $N$ is large and $n$ is small relative to
$N$ to derive approximate expressions for $P_i$ and $P_{ij}$ and
hence the variance of the H.T. estimator under the R.H.C.
scheme.

3.4.3. Approximate expression for $P_i$

The exact expression for $P_i$ under R.H.C. scheme from
(3.4.5) is

$$P_i = \frac{1}{A} \cdot \sum_{\phi \in \Omega} \frac{P_i}{S(\phi,1)}$$

(3.4.7)
Now, in a given partition there are \((M-1)\) units occurring in a group along with the \(i\)th unit. This particular set of \((M-1)\) units is one among the set of all possible combinations \(\binom{N-1}{M-1}\) that is chosen from the population of the rest of \((N-1)\) units excluding the \(i\)th unit. \(p_i/S(\theta,i)\) is the same for all those partitions in which the \(i\)th unit occurs in a group with a particular set of \((M-1)\) units.

Further, among the \(A\) distinct partitions, each of the \(\binom{N-1}{M-1}\) sets occur equally frequently along with the \(i\)th unit, say \(\alpha\) times each where \(\alpha\) is given by

\[
\alpha = \frac{(N-M)!}{(n-1)! (M!)^{n-1}}
\]

or

\[
\alpha = \frac{(N-M)!}{(n-1)! (M!)^{n-1}}
\]

Now, let \(C(\tau) = \{C_1(\tau), C_2(\tau)\ldots C(\tau)\}\) be the set of all \(\binom{N-1}{M-1}\) possible combinations of \((M-1)\) units selected from the \((N-1)\) units of the population excluding \(U_i\). Thus we have from (3.4.7),

\[
P_i = \frac{1}{A} \sum_{\theta \in \mathcal{G}} \frac{P_i}{S(\theta,i)} = \frac{\alpha}{A} \sum_{g=1}^{(N-1)} \frac{P_i}{p_i + S'_g}
\]

where \(S'_g\) is the sum of the \(p_i\)'s of the units belonging to the set \(C_g(\tau)\) of \(C(\tau)\). Note that for notational convenience \(S(\theta,i)\) is written in a different way as \(p_i + S'_g\) which is equal
to \( S_g \) where \( g \) denotes the group of units that contain the
ith unit in a given partition \( \phi \). Substituting the value
of \( \frac{q}{A} \) from (3.4.8) we get

\[
P_i = \frac{1}{N-1} \left( \frac{P_i}{M-1} \right) \sum_{g=1}^{N-1} \frac{P_i}{P_i + S_g'},
\]

which can alternatively be written as

\[
P_i = E_{g} \left[ \frac{P_i}{S_g'} \right] = p_i \ E_{g} \left[ \frac{1}{S_g'} \right] \quad (3.4.9)
\]

where \( E \) denotes the expectation taken over the scheme of
randomly selecting \((M-1)\) units from out of \((N-1)\) population
units excluding the ith unit with simple random sampling
without replacement and \( S_g' = S_g - p_i \) is the sum of the \( p_i \)'s
of all the \((M-1)\) units thus selected.

Now, we can write (3.4.9) as

\[
P_i = \frac{P_i}{p_i + (M-1) \cdot E(S_g')} \cdot E \left[ \frac{1}{(M-1) \bar{A}} \right] \quad (3.4.10)
\]

\[
= np_i \theta_i \cdot E \left[ \frac{1}{1 + N \gamma_i \bar{A}_g} \right]
\]

\[
= np_i \theta_i \cdot E \left[ 1 + N \gamma_i \bar{A}_g \right]^{-1},
\]

where

\[
\bar{S}_g' = S_g'/(M-1) \quad (3.4.11)
\]
\[
\overline{A}_g = \overline{S'}_g - E(\overline{S'}_g)
\]

and

\[
\gamma_i = (1 - \frac{n}{N}) \theta_i = \frac{(1 - \frac{1}{N})(1 - \frac{n}{N})}{n - (n-1)Np_i}
\]

In order to evaluate the expectation of the expression on the right side of (3.4.10), by assuming that \(|N\gamma_i \overline{A}_g| < 1\), we can expand \([1 + N\gamma_i \overline{A}_g]^{-1}\) as a power series in powers of \(N\gamma_i \overline{A}_g\). However in view of the inequality

\[
\frac{(M-1) \cdot E(\overline{S'}_g)}{p_i + (M-1) \cdot E(\overline{S'}_g)} < 1,
\]

it is clear that the usual assumption made in the theory of ratio estimation viz., \(|\frac{\overline{A}_g}{E(\overline{S'}_g)}| < 1\), is sufficient to ensure that \(|N\gamma_i \overline{A}_g| < 1\). Since \((M-1)\) is sufficiently large it is quite likely that the assumption \(|\frac{\overline{A}_g}{E(\overline{S'}_g)}| < 1\) is valid.

So by expanding \([1 + N\gamma_i \overline{A}_g]^{-1}\) as a power series we get from (3.4.10) that

\[
P_i = np_i \theta_i \cdot E[1 - N\gamma_i \overline{A}_g + N^2 \gamma_i \overline{A}_g^2 - N^3 \gamma_i \overline{A}_g^3 + ...]
\]

In order to derive the variance expression of the H.T. estimator correct to \(O(N^0)\) we first have to get the expression for \(P_i\) correct to \(O(N^{-3})\). Since \(p_i\) is assumed
to be of $O(N^{-1})$, in order to get $P_i$ correct to $O(N^{-3})$ we have to evaluate $\theta_i$ and the expectation of the infinite series on the right hand side of (3.4.14), correct to $O(N^{-2})$.

From (3.4.13) we have

$$\theta_i = \frac{(1-\frac{1}{N})}{n-(n-1)NP_i} \{1-\frac{N}{1-\frac{1}{N}}\}$$

Expanding the denominator as a power series we get after retaining terms to $O(N^{-2})$,

$$\theta_i = 1 + \frac{(n-1)(1-NP_i)}{N} + \frac{(n-1)(1-NP_i) \cdot \{n-(n-1)NP_i\}}{N^2}$$

(3.4.15)

In evaluating $E[1-N\gamma_i \Delta_g + N^2 \gamma_i^2 \Delta_g^2 - N^3 \gamma_i^3 \Delta_g^3 + \ldots]$ we will be using a result due to David (1971) which is stated below without proof.

Lemma 3.1:

Let $u'_n$ be a variate defined over a population of size $N$, the mean value of which is assumed without loss of generality to be zero. If $\bar{u}'_n$ denotes the sample mean of the variate $u'_n$ for a simple random sample without replacement of size $n$, then we have for any positive integer $r$, 

\[ E(\bar{u}_n^r) = O(n^{\frac{-r}{2}}), \text{ if } r \text{ is even} \]

\[ = O(n^{\frac{-(r+1)}{2}}), \text{ if } r \text{ is odd}. \]

Analogous to the set up in this lemma, the population size in our case is \( N-1 \), sample size is \( (M-1) = \frac{N}{n} - 1 \); and the variate under consideration is \( p_t \) which is assumed to be of \( O(N^{-1}) \). So, in order to make use of the lemma we will consider the variate \( v_t = Np_t - \frac{N(1-p_t)}{N-1} \), for \( t \neq i \), which is of \( O(N^0) \). Then we have \( \bar{v}_{(N-1)} = 0 \) and since the sample size \( (M-1) \) is of \( O(N^1) \), it follows from the above lemma that

\[ E(\bar{v}_{(M-1)}^r) = O(N^{\frac{-r}{2}}), \text{ if } r \text{ is even} \]

\[ = O(N^{\frac{-(r+1)}{2}}), \text{ if } r \text{ is odd}. \]

Now from (3.4.12) we have

\[ N^r \Delta g = N^r [S_g' - E(S_g')]^r \]

\[ = \bar{v}_{(M-1)}^r \tag{3.4.17} \]

Since the leading term in \( \gamma_1 \) is 1 it follows from (3.4.16) and (3.4.17) that
\[ E[N^r \gamma_i r \Delta g^r] = \gamma_i r E[N^r \Delta g^r] = O(N^{\frac{r}{2}}), \text{ if } r \text{ is even} \]

\[ = O(N^{\frac{-(r+1)}{2}}), \text{ if } r \text{ is odd} \]

(3.4.18)

Hence it follows that \( E[N^r \gamma_i r \Delta g^r] \) for \( r \geq 5 \) would not contribute to \( P_i \) when considered correct to \( O(N^{-3}) \). Thus we have from (3.4.14) that

\[ p_i = np_i \theta_i \cdot E[1 - N \gamma_i \Delta g^2 + N^2 \gamma_i \Delta g^2 - N^3 \gamma_i \Delta g^2 + N^4 \gamma_i \Delta g^4], \]

(3.4.19)

correct to \( O(N^{-3}) \).

Obviously we have \( E[N \gamma_i \Delta g] = N \gamma_i E[\Delta g] = 0 \) (3.4.20)

In evaluating \( E[N^r \gamma_i r \Delta g^r] \) for \( r = 2, 3 \) and \( 4 \) correct to \( O(N^{-2}) \), the formulae presented by Sukhatme (1944) have been used here. Thus using formulae 2, 5 and 10 of his article we have,

\[ E(\Delta g^2) = (N-1) \mu_2 \frac{(e_1-e_2)}{(N^n -1)^2} \]

\[ E(\Delta g^3) = (N-1) \mu_3 \frac{(e_1-3e_2+2e_3)}{(N^n -1)^3} \]

and
\[ E[\overline{\Lambda}_g^4] = (N-1) \mu_4 \left[ \frac{(e_1 - 7e_2 + 12e_3 - 6e_4)}{(N/n - 1)^4} \right] \]

\[ + (N-1)^2 \mu_2^2 \cdot \frac{3(e_2 - 2e_3 + e_4)}{(N/n - 1)^4}, \]

where

\[ (N-1) \mu_r = \sum_{t \neq i}^{N} \left( p_t \cdot \frac{1-p_i}{N-1} \right)^r, \]

and

\[ e_r = \frac{(N/n - 1)(r)}{(N-1)(r)} \]

Using (3.4.13) together with these equations it can be seen after some simplification that correct to \( O(N^{-2}) \), we have

\[ E[N^2 \gamma_1 \overline{\Lambda}_g^2] = (n-1)[(\Sigma p_t^2 - \frac{1}{N}) + \frac{n+1}{N} \cdot (\Sigma p_t^2 - \frac{1}{N}) \]

\[ - 2(n-1)(\Sigma p_t^2 - \frac{1}{N}) \cdot p_i - (p_i - \frac{1}{N})^2] \] (3.4.21)

\[ E[N^3 \gamma_1 \overline{\Lambda}_g^3] = (n-1)(n-2) \left[ \Sigma p_t^3 - \frac{3\Sigma p_t^2}{N} + \frac{2}{N^2} \right] \] (3.4.22)

and

\[ E[N^4 \gamma_1 \overline{\Lambda}_g^4] = 3(n-1)^2 \cdot (\Sigma p_t^2 - \frac{1}{N})^2 \] (3.4.23)

Using Equations (3.4.15) and (3.4.20)-(3.4.23) we get from (3.4.19) after some simplification,
\[ P_i = np_i\left[1 + (n-1)(\Sigma p_t^2 - p_i) + \frac{N}{n}(p_i - \Sigma p_t^2)\right] + (n-2)(p_i^2 - \Sigma p_t^3) - 3(n-1)(p_i^2 - \Sigma p_t^2)\cdot\Sigma p_t^2 \]  
(3.4.24)

correct to \(O(N^{-3})\).

In the situation when all the \(p_i\)'s are equal, i.e., when \(p_i = \frac{1}{N}\), R.H.C. scheme would reduce to simple random sampling without replacement in which case \(P_i\) is known to be equal to \(n/N\). A check on (3.4.24) is provided by verifying that the right hand side in fact reduces to \(n/N\) when \(p_i\) is replaced by \(1/N\). A more rigorous check on (3.4.24) is provided by verifying that \(\sum_{i=1}^{N} p_i = n\) when the value of \(p_i\) is substituted from (3.4.24).

3.4.4. Approximate expression for \(P_{ij}\) correct to \(O(N^{-4})\)

The exact expression for \(P_{ij}\) under the R.H.C. scheme from (3.4.6) is

\[ P_{ij} = \frac{1}{A} \cdot \sum_{\phi \in G_2(i,j)} \frac{P_i P_j}{S(\phi, i) S(\phi, j)} \]  
(3.4.25)

In a given partition \(\phi\) belonging to \(G_2(i,j)\) there are \((M-1)\) units occurring in a group along with the \(i\)th unit and there are another set of \((M-1)\) units occurring in another group along with the \(j\)th unit. This particular ordered pair of groups is one among the possible number
of ordered pairs of groups \((\binom{N-2}{M-1}, \binom{N-M-1}{M-1})\), and the product 
\[ P_i P_j/ (\Theta, i) \cdot S(\Theta, j) \]
remains the same for all those partitions \(\Theta\) of \(G_2(i,j)\) where in the ordered pair \((U_i, U_j)\) is associated with a particular member of the set of 
\((\binom{N-2}{M-1}, \binom{N-M-1}{M-1})\) ordered pairs of groups.

Let \(\mathcal{D}(i, j) = \{D_1(i, j), D_2(i, j), \ldots D(i, j)\}\) denote the collection of all possible ordered pairs of groups of sizes \((M-1)\) from the population excluding \(U_i\) and \(U_j\). Among the \(A_2\) partitions of \(G_2(i, j)\) each of the \((\binom{N-2}{M-1}, \binom{N-M-1}{M-1})\) possible ordered pairs of groups occur equally frequently along with the ordered pair \((U_i, U_j)\), say, \(v\) times where \(v\) is given by

\[
v \cdot \binom{N-2}{M-1} \cdot \binom{N-M-1}{M-1} = A_2 = \frac{(N-2)!}{(n-2)! \{\binom{(M-1)!}{2}(M!\}^{n-2}}
\]

(3.4.26)

or

\[
v = \frac{(N-2M)!}{(n-2)! \{\binom{(M-1)!}{2}(M!)\}^{n-2}}
\]

Therefore, from (3.4.25) we have,

\[
P_{ij} = \frac{1}{A} \cdot \sum_{\Theta \in G_2(i, j)} \frac{P_i P_j}{S(\Theta, i) \cdot S(\Theta, j)}
\]
where the summation runs over all the members of $\mathcal{S}(I,J)$, and $S'_r$ and $S'_s$ denote the sum of the $p_t$'s of the set of units that correspond to the first and second groups respectively of a given pair of groups of $\mathcal{S}(I,J)$.

(3.4.27) can be written as

$$P_{ij} = p_ip_j \cdot \frac{A_2}{A} \cdot \frac{1}{(p_i + S'_r)(p_j + S'_s)}$$

which after substituting the value of $v/A_2$ from (3.4.26) gives

$$P_{ij} = p_ip_j \cdot \frac{A_2}{A} \cdot \frac{1}{N-2} \cdot \frac{1}{(M-1)^2} \cdot \frac{1}{(p_i + S'_r)(p_j + S'_s)}

(3.4.28)$$

In order to evaluate the variance of the H.T. estimator correct to $O(N^0)$, we have to get the value of $P_{ij}$ correct to $O(N^{-4})$.

From (3.4.1) and (3.4.3) we get,

$$\frac{A_2}{A} = \frac{N(n-1)}{n(N-1)}

(3.4.29)$$

which when expanded in powers of $1/N$ gives

$$\frac{A_2}{A} = \frac{(n-1)}{n} \left[ 1 + \frac{1}{N} + \frac{1}{N^2} \right]

(3.4.30)$$
correct to $O(N^{-2})$.

Since $p_i$ is assumed to be of $O(N^{-1})$ it follows that

$$\frac{1}{(N-2)} \sum_{i=1}^{N-M} \frac{1}{(p_i + S_i') (p_j + S_j')}$$

to $O(N^{-2})$. Before evaluating this, we will prove a lemma which will be used also at different places in the subsequent portions of this dissertation.

**Lemma 3.2:**

Let $p_t$ be a variate defined over a population of size $N$ where in $p_t$ is assumed to be of $O(N^{-1})$. Let $N$ be a multiple of $K$ where $K$ is small relative to $N$. Consider the scheme of selecting two without replacement simple random samples of size $(\frac{N}{K}-1)$ each from the population of $(N-2)$ units excluding $U_i$ and $U_j$. Let $S_r'$ and $S_s'$ be the sum of the $p_t$'s of the units belonging to these two samples respectively. Let $S_r = p_1 + S_r'$ and $S_s = p_j + S_s'$, then we have correct to $O(N^{-2})$,

$$E[\frac{S_r}{S_s}] = K[1 + \left((p_j - \frac{1}{N}) + (K-1)(\Sigma p_t^2 - p_i)\right]$$

$$+ \left((K-1)(K-2)p_i^2 - (K-2)p_j^2 - 2(K-1)p_ip_j\right)$$

$$+ \left(K^2 + K - 1)p_i/N - p_j/N - (K^2 + K - 1) \cdot \Sigma p_t^2 / N\right)$$

$$- 3(K-1)^2 \Sigma p_t^2 p_i + 3(K-1) \Sigma p_t^2 p_j - (K-1)(K-2) \Sigma p_t^3$$

$$+ 3(K-1)^2 (\Sigma p_t^2)^2 \right]$$

(3.4.31)
\[
E\left[ \frac{1}{S_s} \right] = K\left[ 1 + \left\{ (p_i - \frac{1}{N}) + (K-1) (\Sigma p_t^2 - p_j) \right\} + \left\{ (K-1)(K-2)p_j^2 - (K-2)p_i^2 - 2(K-1)p_ip_j \right\} + \left( K^2 + K - 1 \right) \frac{p_j}{N-p_i/N} - (K^2 + K - 1) \cdot \frac{\Sigma p_f^2}{N} - 3(K-1)^2 \Sigma p_t^2 \cdot p_j + 3(K-1)^2 (\Sigma p_t^2)^2 \right] \tag{3.4.32}
\]

and

\[
E\left[ \frac{1}{S_r S_s} \right] = K^2 \left[ 1 + \left\{ (2K-3) \Sigma p_t^2 - (K-2) (p_i + p_j) - 1/N \right\} + \left\{ (K^2 - 2) (p_i + p_j)/N + (K-2)(K-3) (p_i^2 + p_j^2) \right\} - 2(K-2)(2K-3) (p_i + p_j) \cdot \Sigma p_t^2 - (2K^2 - 3) \cdot \Sigma p_t^2 / N \right] - 2(K-2)^2 \Sigma p_t^3 + (7K^2 - 20K + 15) (\Sigma p_t^2)^2 + \left(K^2 - 6K + 6\right)p_ip_j \right] \tag{3.4.33}
\]

Proof:

Analogous to (3.4.10) we can write

\[
\frac{1}{S_r} = \frac{1}{p_i + (\frac{N}{K} - 1) \cdot E(S'_r)} \cdot \frac{1}{[1 + \frac{(N-1)\Delta_r}{p_i + (\frac{N}{K} - 1) E(S'_r)}]}
\]

\[
= K \cdot \theta_{ij} \cdot \frac{1}{1 + \hat{\theta}_{ij} \Delta_r} \tag{3.4.34}
\]
and

\[
\frac{1}{S_s} = \frac{1}{p_j^+ \left( \frac{N}{K} - 1 \right) \cdot E(S'_s)} \cdot \frac{1}{\left[ 1 + \frac{\left( \frac{N}{K} - 1 \right) \bar{\Lambda}}{p_j^+ \left( \frac{N}{K} - 1 \right) E(S'_s)} \right]}
\]

\[
= K \cdot \theta_{ji}^+ \cdot \frac{1}{1 + N \gamma_{ji}^+ \bar{\Lambda}}
\]

(3.4.35)

where

\[
\gamma_{ij}^+ = (1 - \frac{K}{N}) \theta_{ij} = \frac{(1 - K/N)(1 - 2/N)}{\left[ 1 - \{p_j^+ - (K - 1)p_j^+ / N \cdot (1 + p_j^+ - p_j) \} \right]}
\]

(3.4.36)

\[
\gamma_{ji}^+ = (1 - K/N) \theta_{ji} = \frac{(1 - K/N)(1 - 2/N)}{\left[ 1 - \{p_i^+ - (K - 1)p_j^+ / N \cdot (1 + p_j^+ - p_j) \} \right]}
\]

(3.4.37)

\[
\bar{\Lambda}_r = \bar{S}_r' - E(S'_r)
\]

(3.4.38)

and

\[
\bar{\Lambda}_s = \bar{S}_s' - E(S'_s)
\]

(3.4.39)

Thus we have from (3.4.34),

\[
E[\frac{1}{S_r}] = K \cdot \theta_{ij}^+ \cdot E[\frac{1}{1 + N \gamma_{ij}^+ \bar{\Lambda}_r}]
\]

\[
= K \cdot \theta_{ij}^+ \cdot E[1 + N \gamma_{ij}^+ \bar{\Lambda}_r]^{-1}
\]

(3.4.40)
By assuming that \( |N\Gamma_{ij\sigma}^r| < 1 \), we can expand \([1+N\Gamma_{ij\sigma}^r]^{-1}\) as a power series in powers of \(N\Gamma_{ij\sigma}^r\). Here again, the usual assumption made in the theory of ratio estimation, viz., \( \frac{\Delta_r}{E(S'_r)} \) \( |N\Gamma_{ij\sigma}^r| < 1 \) is sufficient to ensure that \( |N\Gamma_{ij\sigma}^r| < 1 \). Since \( \left(\frac{N}{K} - 1\right) \) is sufficiently large it is quite likely that the assumption \( \frac{\Delta_r}{E(S'_r)} \) \( |N\Gamma_{ij\sigma}^r| < 1 \) is valid. So by expanding \([1+N\Gamma_{ij\sigma}^r]^{-1}\) as a power series we get

\[
E\left[ \frac{1}{S_r} \right] = K\theta_{ij\sigma} E\left[ 1 - N\Gamma_{ij\sigma}^r + N^2 \gamma_{ij\sigma}^2 \Delta_r^2 \right] - N^3 \gamma_{ij\sigma}^3 \Delta_r^3 + \ldots 
\]

(3.4.41)

From (3.4.34) we have,

\[
\theta_{ij\sigma} = \frac{(1-2/N)}{[1-{p_j-(K-1)p_i+N(1+p_i-p_j)}]} 
\]

Expanding the denominator as a power series we get after retaining terms to \( O(N^{-2}) \),

\[
\theta_{ij\sigma} = [1+{p_j-(K-1)p_i+(K-2)/N}+{p_j^2+(K-1)^2p_i^2}
\]

\[-2(K-1)pi p_j + \frac{K(K-2)}{N^2} - (K-2)(2K-1)p_i/N
\]

\[+ (K-2)p_j/N] \]

(3.4.42)
Since $p_i$ is of $O(N^{-1})$ and the leading term in $\theta_{ij}$ is 1,

\[ E\left[ \frac{1}{S_r} \right] \text{ in (3.4.41), correct to } O(N^{-2}) \text{ is given by,} \]

\[ E\left[ \frac{1}{S_r} \right] = K \theta_{i,j} \cdot E[1 - N \gamma_{ij} \hat{A}_r + N^2 \gamma_{ij} \hat{A}_r^2 - N^3 \gamma_{ij} \hat{A}_r^3 - N^4 \gamma_{ij} \hat{A}_r^4] \]

\[ + N^4 \gamma_{ij} \hat{A}_r^4], \tag{3.4.43} \]

which follows by the application of Lemma 3.1. Obviously we have $E[N \gamma_{ij} \hat{A}_r] = N \gamma_{ij} E[\hat{A}_r] = 0$ \tag{3.4.44}

Using formulae 2, 5, and 10 of Sukhatme (1944) we get,

\[ E[\hat{A}_r^2] = (N-2) \mu_2 \left[ \frac{(e_1 - e_2)}{(N-1)^2} \right] \]

\[ E[\hat{A}_r^3] = (N-2) \mu_3 \left[ \frac{(e_1 - 3e_2 + 2e_3)}{(N-1)^3} \right] \]

and

\[ E[\hat{A}_r^4] = (N-2) \mu_4 \left[ \frac{(e_1 - 7e_2 + 12e_3 - 6e_4)}{(N-1)^4} \right] \]

\[ + (N-2)^2 \mu_2 \left[ \frac{3(e_2 - 2e_3 + e_4)}{(N-1)^4} \right] \]

where

\[ (N-2) \mu_r = \sum_{t(\neq i,j)} N \left( p_t - \frac{1-p_i-p_j}{N-2} \right) r \]

and
\[ e_r = \frac{\left(\frac{N}{K} - 1\right)(r)}{(N-2)(r)} \]

Thus, using (3.4.36) it can be seen after some simplification that correct to \(O(N^{-2})\)

\[
E[N^2 \gamma_{i,j}^2 \alpha_r^2] = (K-1) (\Sigma p_t^2 - 1/N) + 2K(K-1) \cdot p_i/N
\]

\[
- (K-1) (p_i^2 + p_j^2) - 2(K-1)^2 \Sigma p_t^2 \cdot p_i
\]

\[
+ 2(K-1) \Sigma p_t^2 \cdot p_j + (K^2 - K-1) \cdot \frac{\Sigma p_t^2}{N} - (K^2 + K-3) \cdot \frac{1}{N^2}
\]

(3.4.45)

\[
E[N^3 \gamma_{i,j}^3 \alpha_r^3] = (K-1) (K-2) \cdot [\Sigma p_t^3 - \frac{3\Sigma p_t^2}{N} + \frac{2}{N^2}]
\]

(3.4.46)

and

\[
E[N^4 \gamma_{i,j}^4 \alpha_r^4] = 3(K-1)^2 \cdot (\Sigma p_t^2 - 1/N)^2
\]

(3.4.47)

Using (3.4.42), and (3.4.44)-(3.4.47) we get from (3.4.43) that, correct to \(O(N^{-2})\),

\[
E[\frac{1}{S_r}] = K[1 + \{(p_j - 1/N) + (K-1) (\Sigma p_t^2 - p_i)\}]
\]

\[
+ \{(K-1) (K-2) p_i^2 - (K-2) p_j^2 - 2(K-1) p_i p_j
\]

\[
+ (K^2 + K-1) \cdot p_i/N - p_j/N - (K^2 + K-1) \cdot \Sigma p_t^2/N
\]

\[
-3(K-1)^2 \Sigma p_t^2 \cdot p_i + 3(K-1) \Sigma p_t^2 \cdot p_j - (K-1) (K-2) \Sigma p_t^3
\]

\[
+3(K-1)^2 (\Sigma p_t^2)^2]
\]

(3.4.48)
By symmetry we get $E[\frac{1}{S^2}]$ correct to $O(N^{-2})$ by interchanging $p_i$ and $p_j$ in (3.4.48) which will yield (3.4.32).

Now, from (3.4.34) and (3.4.35) we get,

$$E[\frac{1}{S^2}] = K^2 \theta_{ij} \theta_{ji} \cdot E[\frac{1}{(1 + N\gamma_{ij} \Delta_r)} \cdot \frac{1}{(1 + N\gamma_{ji} \Delta_s)}]$$

$$= K^2 \theta_{ij} \theta_{ji} \cdot E[1 - \{N\gamma_{ij} \Delta_r + N\gamma_{ji} \Delta_s\}]$$

$$+ \{N^2 \gamma_{ij} \Delta_r^2 + N^2 \gamma_{ji} \Delta_s^2 + N^2 \gamma_{ij} \gamma_{ji} \Delta_r \Delta_s\}$$

$$- \{N^3 \gamma_{ij} \Delta_r^3 + N^3 \gamma_{ji} \Delta_s^3 + N^3 \gamma_{ij} \gamma_{ji} \Delta_r \Delta_s^2\}$$

$$+ N^3 \gamma_{ij} \gamma_{ji} \Delta_r \Delta_s^2\}$$

$$+ \{N^4 \gamma_{ij} \Delta_r^4 + N^4 \gamma_{ji} \Delta_s^4 + N^4 \gamma_{ij} \gamma_{ji} \Delta_r \Delta_s^3\}$$

$$+ N^4 \gamma_{ij} \gamma_{ji} \Delta_r \Delta_s^3 + N^4 \gamma_{ij} \gamma_{ji} \Delta_r \Delta_s^2\}]}$$

(3.4.49)

correct to $O(N^{-2})$.

Obviously $E[N\gamma_{ij} \Delta_r + N\gamma_{ji} \Delta_s] = 0$ (3.4.50)
From (3.4.45)-(3.4.47) by symmetry it follows that, correct to $O(N^{-2})$,

$$E[N^2 \gamma_{jj} - \bar{\Delta}_s^2] = (K-1)(\Sigma \sigma_i^2 - 1/N)$$

$$+ 2K(K-1) \cdot p_j / (K-1) (p_i^2 + p_j^2) - 2(K-1)^2 \cdot \sigma_i^2 \cdot p_j$$

$$+ 2(K-1) \Sigma \sigma_t^2 \cdot p_i + (K^2 - K - 1) \cdot \sigma_t^2 / N - (K^2 + K - 3) \cdot 1/N^2 \quad (3.4.51)$$

$$E[N^3 \gamma_{jj} - \bar{\Delta}_s^3] = (K-1)(K-2) \cdot [\Sigma \sigma_t^3 - 3\Sigma \sigma_t^2 / N + 2/N^2] \quad (3.4.52)$$

$$E[N^4 \gamma_{jj} - \bar{\Delta}_s^4] = 3(K-1)^2 \cdot (\Sigma \sigma_t^2 - 1/N)^2 \quad (3.4.53)$$

From the basic properties of simple random sampling we have

$$E[\bar{\Delta}_r \cdot \bar{\Delta}_s] = - \frac{1}{(N-2)} \cdot \frac{1}{(N-3)} \cdot [\Sigma \sigma_t^2 \cdot p_i^2 \cdot p_j^2 \cdot (1-p_i-p_j)^2 / N-2] \quad (3.4.54)$$

Using this we will get after simplifying and retaining terms to $O(N^{-2})$,

$$E[N^2 \gamma_{ij} - \bar{\Delta}_r \cdot \bar{\Delta}_s] = \left[-(\Sigma \sigma_t^2 - 1/N) + \frac{(p_i + p_j)}{N} - \frac{3}{N^2} \right]$$

$$- (p_i^2 + p_j^2) + \Sigma \sigma_t^2 / N - (K-2)(p_i + p_j) \cdot \sigma_t^2 \right) \quad (3.4.55)$$
From (3.4.45), (3.4.51) and (3.4.55) we get

\[
E[N^2 \gamma_{ij}^2 \bar{A}_x^2 + N^2 \gamma_{ji}^2 \bar{A}_y^2 + N^2 \gamma_{ij}^2 \gamma_{ji} \bar{A}_x \bar{A}_y]
\]

\[
= (2K-3)(\Sigma_p t^2 - 1/N) + K(2K-3) \frac{(p_i + p_j)}{N} - (2K-3)(p_i^2 + p_j^2)
\]

\[
- (K-2)(2K-3)(p_i + p_j) \Sigma_p t^2 + (2K^2 - 2K - 3) \cdot \Sigma_p t^2 / N
\]

\[
- \frac{(2K^2 + 2K - 9)}{N^2}
\]

(3.4.56)

correct to \(O(N^{-2})\).

Using the conditional expectation approach, it can be seen by symmetry that,

\[
E[N^3 \gamma_{ij}^2 \gamma_{ji} \bar{A}_x \bar{A}_y]
\]

\[
= -N^3 \gamma_{ij}^2 \gamma_{ji} \cdot \frac{(N/K - 1)}{(N\cdot N/K - 1)} \cdot E[\bar{A}_x^3]
\]

\[
= -(K-2) \cdot [\Sigma_p t^3 - \frac{3\Sigma_p t^2}{N} + \frac{2}{N^2}],
\]

(3.4.57)

correct to \(O(N^{-2})\).

Thus from (3.4.46), (3.4.52) and (3.4.57) we get

\[
E[N^3 \gamma_{ij}^3 \bar{A}_x^3 + N^3 \gamma_{ji}^3 \bar{A}_y^3 + N^3 \gamma_{ij}^3 \gamma_{ji} \bar{A}_x \bar{A}_y + N^3 \gamma_{ij}^3 \gamma_{ji} \bar{A}_x \bar{A}_y]
\]

\[
= 2(K-2)^2 \cdot [\Sigma_p t^3 - \frac{3\Sigma_p t^2}{N} + \frac{2}{N^2}],
\]

(3.4.58)
correct to $O(N^{-2})$.

Using the same conditional approach it can be seen that,

$$E[N^4 Y_{i}`Y_{j}`\gamma_{i}`\gamma_{j}`\bar{\Delta}_{r}`\bar{\Delta}_{s}] = E[N^4 Y_{i}`Y_{j}`\gamma_{i}`\gamma_{j}`\bar{\Delta}_{r}`\bar{\Delta}_{s}]$$

$$= -3(K-1) \cdot (\Sigma p_{t} - 1/N)^2,$$  \hspace{1cm} (3.4.59)

and

$$E[N^4 Y_{i}`Y_{j}`\gamma_{i}`\gamma_{j}`\bar{\Delta}_{r}`\bar{\Delta}_{s}] = (K^2 - 2K + 3) \cdot (\Sigma p_{t} - 1/N)^2,$$  \hspace{1cm} (3.4.60)

correct to $O(N^{-2})$.

Thus, from (3.4.47), (3.4.53), (3.4.59) and (3.4.60) we get,

$$E[N^4 Y_{i}`Y_{j}`\gamma_{i}`\gamma_{j}`\bar{\Delta}_{r}`\bar{\Delta}_{s}] = (7K^2 - 20K + 15)(\Sigma p_{t} - 1/N)^2,$$  \hspace{1cm} (3.4.61)

correct to $O(N^{-2})$.

From (3.4.42) we get by symmetry correct to $O(N^{-2})$,

$$\theta_{j}` = [1+(p_{i} - (K-1)p_{j} + \frac{(K-2)}{N}) + (p_{i}^2 + (K-1)^2p_{j}^2 - 2(K-1)p_{i}p_{j} + \frac{K(K-2)}{N^2} - (K-2)(2K-1)\cdot \frac{p_{i}}{N} + (K-2)\cdot \frac{p_{i}}{N}]$$  \hspace{1cm} (3.4.62)
Substituting from (3.4.42), (3.4.50), (3.4.56), (3.4.58) and (3.4.61) into (3.4.49), we get

\[ E[ \frac{1}{S_r S_s} ] = K^2 \left[ 1 + \{(2K-3) \sum p_t^2 - (K-2)(p_i + p_j) - \frac{1}{N}\right) \]

\[ + \{ (K^2-2) \frac{(p_i + p_j)}{N} + (K-2)(K-3)(p_i^2 + p_j^2) \}

\[ - 2(K-2)(2K-3)(p_i + p_j) \cdot \sum p_t^2 - (2K^2-3) \cdot \frac{\sum p_t^2}{N} \]

\[ - 2(K-2)^2 \sum p_t^3 + (7K^2-20K+15)(\sum p_t^2)^2 \]

\[ + (K^2-6K+6)p_i p_j \} \] (3.4.63)
correct to \( O(N^{-2}) \).

Thus the proof of Lemma 3.2 is completed. Q.E.D.

Now, going back to the problem of evaluating \( p_{ij} \) in (3.4.28) correct to \( O(N^{-4}) \), we have observed that

\[ \frac{1}{(N-2)(N-M-1)} \sum' \frac{1}{(p_i + s') (p_j + s')} \]

is to be evaluated correct to \( O(N^{-2}) \). It can be observed that this expression can be considered as \( E[ \frac{1}{S_r S_s} ] \) where \( E \) denotes the expectation taken over the scheme described in Lemma 3.2 with \( K \) replaced by \( n \).

Thus, from (3.4.33) we have,
\[
\frac{1}{M-1} \frac{1}{N-M-1} \sum' \left( \frac{1}{(p_i + s_i')(p_j + s_j')} \right) = n^2 \left[ 1 + \left( 2n-3 \right) \sum p_t^2 \right.
\]

\[
- \left( n-2 \right) \left( p_i + p_j \right) \frac{1}{N} + \left( n^2 - 2 \right) \frac{p_i + p_j}{N}
\]

\[
+ \left( n-2 \right) \left( n-3 \right) \left( p_i^2 + p_j^2 \right)
\]

\[
- 2 \left( n-2 \right) \left( 2n-3 \right) \left( p_i + p_j \right) \sum p_t^2
\]

\[
- \left( 2n^2 - 3 \right) \frac{\sum p_t^2}{N} - 2 \left( n-2 \right)^2 \sum p_t^3
\]

\[
+ \left( 7n^2 - 20n + 15 \right) \left( \sum p_t^2 \right)^2
\]

\[
+ \left( n^2 - 6n + 6 \right) p_ip_j \}
\]

(3.4.64)

correct to \( O(N^{-2}) \).

Substituting from (3.4.30) and (3.4.64) in (3.4.28) we get after simplifying and retaining terms to \( O(N^{-4}) \),

\[
p_{ij} = n(n-1)p_ip_j \left[ 1 + (2n-3) \sum p_t^2 - (n-2) \left( p_i + p_j \right) \right]
\]

\[
+ \left( n^2 - 5n + 6 \right) \left( p_i^2 + p_j^2 \right) - 2 \left( n-2 \right)^2 \sum p_t^3
\]

\[
+ \left( n^2 - 6n + 6 \right) p_ip_j - 2 \left( n-2 \right) \left( 2n-3 \right) \left( p_i + p_j \right) \sum p_t^2
\]

\[
+ \left( 7n^2 - 20n + 15 \right) \left( \sum p_t^2 \right)^2 + n(n-1) \frac{p_i + p_j}{N}
\]

\[
- 2n(n-1) \left( \frac{\sum p_t^2}{N} \right) \}
\]

(3.4.65)

correct to \( O(N^{-4}) \).
In case when all the $p_i$'s are equal R.H.C. scheme reduces to simple random sampling without replacement. When we substitute the value $p_i = \frac{1}{N}$ for $i = 1, 2, \ldots N$; (3.4.65) reduces to $\frac{n(n-1)}{N^2} \left[ 1 + \frac{1}{N} + \frac{1}{N^2} \right]$ which is the value for $P_{ij}$ to $O(N^{-4})$ in the case of simple random sampling without replacement thus providing a check on (3.4.65). A more thorough check on (3.4.65) is provided by verifying that

$$\sum_{j(\neq i)}^{N} P_{ij} = (n-1)P_i,$$

when the terms to $O(N^{-3})$ are retained in the above summation, wherein the value for $p_i$ is as in (3.4.24).

3.4.5. Approximate expression for the variance of the H.T. estimator correct to $O(N^0)$ and to $O(N^1)$

The variance expression for the H.T. estimator denoted by $T_2$, is given by

$$V(T_2) = \sum_{i=1}^{N} \frac{Y_i^2}{P_i} + \sum_{i=1}^{N} \sum_{j(\neq i)}^{N} \frac{P_{ij}}{P_i P_j} \cdot Y_i Y_j - Y^2 \quad (3.4.66)$$

Substituting the value of $P_i$ from (3.4.24) we get,

$$\sum_{i=1}^{N} \frac{Y_i^2}{P_i} = \sum_{i=1}^{N} \frac{Y_i^2}{n P_i} \cdot \left[ 1 + (n-1)(\Sigma P_t^2 - p_i) + (n-1)(\frac{n}{N} (p_i - \Sigma P_t^2)) \\
+ (n-2) (p_i^2 - \Sigma P_t^3) - 3(n-1)(p_i - \Sigma P_t^2)(\Sigma P_t^2) \right]^{-1}$$
Expanding the expression binomially we get correct to $O(N^0)$,

$$
\sum_{i=1}^{N} \frac{Y_i^2}{P_i} = \frac{Y_t^2}{nP_t} - \frac{n-1}{n}. \left[ \sum_{i} P_i \cdot \frac{Y_t^2}{P_t} - \sum_{t} Y_t^2 \right]
$$

$$
- \frac{n-1}{n} \cdot \left[ \frac{n}{N} \sum_{t} Y_t^2 - \frac{n}{N} \sum_{t} P_t \cdot \frac{Y_t^2}{P_t} \right]
$$

$$
- (n-1) \sum_{t} P_t^2 \cdot Y_t^2 + 2(n-1)(\sum_{t} P_t^2)^2 \cdot \sum_{t} Y_t^2 / P_t \right] \tag{3.4.67}
$$

Using (3.4.24) and (3.4.65) we get,

$$
\frac{P_{ij}}{P_i P_j} = \frac{n-1}{n} \cdot \left[ 1 + \{(p_i + p_j) - \sum_{t} P_t^2\} + \{(3n-5)(p_i + p_j) \sum_{t} P_t^2 \right. \\
- \left. (n-3)(p_i^2 + p_j^2) + 2(n-2) \sum_{t} P_t^3 \right. \\
- \left. 2(2n-3)(\sum_{t} P_t^2)^2 - (2n-3)p_i p_j \} \right]
$$

correct to $O(N^{-2})$.

Substitution of this yields,

$$
\sum_{i=1}^{N} \sum_{j(i \neq j)}^{N} P_{ij} Y_i Y_j = \frac{(n-1)}{n} \cdot \left[ Y^2 - \{(2n+1) \sum_{t} P_t^2 \right. \\
- \left. 2(2n-3)Y \sum_{t} P_t \sum_{t} Y_t + \sum_{t} Y_t^2 \right. \\
+ \left. (\sum_{t} P_t)^2 \cdot \sum_{t} Y_t^2 - 2 \sum_{t} P_t Y_t Y_t - 2(2n-3) Y \sum_{t} P_t \sum_{t} Y_t \right. \\
+ \left. 2(3n-5)Y \cdot \sum_{t} P_t^2 \cdot \sum_{t} Y_t + 2(n-2) Y^2 \cdot \sum_{t} P_t^3 \right. \\
- \left. 2(2n-3) Y^2 \cdot (\sum_{t} P_t^2)^2 - (2n-3) \cdot (\sum_{t} P_t Y_t)^2 \right. \right]
$$

correct to $O(N^0)$.

(3.4.68)
Substitution from (3.4.67) and (3.4.68) in (3.4.66) yields after retaining terms to $O(N^0)$ only,

$$V(T_2) = \frac{1}{n} \left[ \frac{\bar{y}_t^2}{\bar{p}_t} - y^2 \right] - \frac{n-1}{n} \left[ \sum \frac{\bar{y}_t^2}{\bar{p}_t} \right] - 2\bar{y} \cdot \sum \bar{p}_t \bar{y}_t$$

$$+ \bar{y}^2 \sum \bar{p}_t^2 - \frac{n-1}{n} \left[ \frac{\sum \bar{y}_t^2}{N} - \frac{\sum \bar{y}_t^2}{N} \right]$$

$$- 2(n-1) \left[ \sum \bar{p}_t^2 \right] \cdot \frac{\bar{y}_t^2}{\bar{p}_t}$$

$$+ \sum \bar{p}_t \bar{y}_t^2 - n \sum \bar{p}_t^2 \cdot \sum \bar{y}_t^2 + 2(n-3) \cdot \bar{y} \cdot \sum \bar{p}_t^2 \bar{y}_t$$

$$- 2(3n-5) \bar{y} \cdot \sum \bar{p}_t^2 \cdot \sum \bar{y}_t^2 - 2(n-2) \bar{y}^2 \sum \bar{p}_t^3$$

$$+ 2(2n-3) \bar{y}^2 \sum \bar{p}_t^2 \cdot \sum \bar{p}_t^2 \bar{y}_t^2 + (2n-3) \left[ \sum \bar{p}_t \bar{y}_t^2 \right], \quad (3.4.69)$$

On the other hand, if terms only to $O(N^1)$ are retained, from (3.4.69) we find to $O(N^1)$, the simplified expression,

$$V(T_2) = \frac{1}{n} \left[ \frac{\bar{y}_t^2}{\bar{p}_t} - y^2 \right] - \frac{n-1}{n} \left[ \sum \frac{\bar{y}_t^2}{\bar{p}_t} \right] - 2\bar{y} \cdot \sum \bar{p}_t \bar{y}_t + \bar{y}^2 \sum \bar{p}_t^2$$

$$+ \sum \bar{p}_t \bar{y}_t^2 - n \sum \bar{p}_t^2 \cdot \sum \bar{y}_t^2 + 2(n-3) \cdot \bar{y} \cdot \sum \bar{p}_t^2 \bar{y}_t$$

$$- 2(3n-5) \bar{y} \cdot \sum \bar{p}_t^2 \cdot \sum \bar{y}_t^2 - 2(n-2) \bar{y}^2 \sum \bar{p}_t^3$$

$$+ 2(2n-3) \bar{y}^2 \sum \bar{p}_t^2 \cdot \sum \bar{p}_t^2 \bar{y}_t^2 + (2n-3) \left[ \sum \bar{p}_t \bar{y}_t^2 \right], \quad (3.4.70)$$

For the special case of equal probabilities $p_i = \frac{1}{N}$, (3.4.69) reduces to the familiar variance formula for the estimator in simple random sampling without replacement. This provides a check for the variance expression (3.4.69) correct to $O(N^0)$. 
3.5. Estimation of the Variance

Yates-Grundy estimate of variance for the H.T. estimator is

\[ \nu_{Y-G}(T_2) = \sum_{i>j} \frac{n}{P_i P_j} (\frac{Y_i}{P_i} - \frac{Y_j}{P_j})^2 \]  

(3.5.1)

From (3.4.24) and (3.4.65) we get,

\[ P_i = n p_i [1+(n-1)(\Sigma P_i^2 - p_i)] \]

(3.5.2)

to O(N^{-2}) and

\[ P_{ij} = n(n-1)p_i p_j [1+(2n-3)(\Sigma P_i^2 - (n-2)(p_i+p_j))] \]

(3.5.3)

to O(N^{-3}).

Substituting (3.5.2) and (3.5.3) in (3.5.1), we get after simplifying and retaining terms to O(N^1),

\[ \nu_{Y-G}(T_2) = \frac{1}{n^2(n-1)} \sum_{i>j} \{[1-n(p_i+p_j) - (n-2)\Sigma P_i^2] (\frac{Y_i}{P_i} - \frac{Y_j}{P_j})^2 + 2(n-1) \cdot (Y_i - Y_j) (\frac{Y_i}{P_i} - \frac{Y_j}{P_j}) \} \]

(3.5.4)

to O(N^1).

For the special case of equal probabilities \( p_i = \frac{1}{N} \),

(3.5.4) agrees with the formula for the estimate of the variance in equal probability sampling without replacement, noting that
\[ \sum_{i>j}^{n} (y_i - y_j)^2 = n \cdot \sum_{i<j}^{n} (y_i - \bar{y})^2 \quad (3.5.5) \]

By substituting the value of \( P_i \) to \( O(N^{-3}) \) and \( P_{ij} \) to \( O(N^{-4}) \) one can get the Y-G estimate of the variance to \( O(N^0) \) which could be used for smaller size populations.

### 3.6. Comparison with Other Estimators

Hartley and Rao (1962) derived the approximate variance formula of the H.T. estimator for the randomized systematic scheme proposed by Goodman and Kish (1950) by using the same asymptotic theory. They have shown that to \( O(N^1) \),

\[ V(\hat{y}_{H.T.})_{G.K} = \frac{1}{n} \left( \frac{y_t^2}{p_t} - y^2 \right) - \frac{(n-1)}{n} \cdot \sum_t p_t \left( \frac{y_t}{p_t} - y \right)^2 \quad (3.6.1) \]

From (3.2.4) we have for the R.H.C. scheme,

\[ V(T_1) = \frac{1}{n} \left( \frac{y_t^2}{p_t} - y^2 \right) - \frac{(n-1)}{n} \cdot \frac{1}{N} \sum_t p_t \left( \frac{y_t}{p_t} - y \right)^2 \quad (3.6.2) \]

correct to \( O(N^1) \), by considering the leading term and the next lower order term in \( N^{-1} \). Since sampling with unequal probabilities is used mostly in situations wherein \( Y_t \) is approximately proportional to \( p_t \), a simple model that is relevant and has been used by many research workers in survey sampling is

\[ Y_t = Yp_t + e_t, \quad (3.6.3) \]
where

\[ E(e_t \mid p_t) = 0, \quad E(e_t^2 \mid p_t) = \alpha p_t^g; \quad \alpha > 0, \quad g > 0 \]  \hspace{1cm} (3.6.4)

Using this model Cochran (1963) has compared the variance of the customary estimator in unequal probability sampling with replacement and the variance of the ratio estimate without the usual finite population correction factor. Cochran has shown that the estimate in unequal probability sampling with replacement is more precise than the ratio estimate if \( g > 1 \) and less precise if \( g < 1 \). Also it is stated that, because of the positive correlation that usually exists between elements in the same cluster unit, \( g \) is likely to lie between 1 and 2.

Hartley and Rao assuming the same model have shown that \( V(\hat{Y}_{H.T.})_{G.K.} \) is smaller or greater than the variance of the ratio estimate with the correction factor according as \( g \) is greater or smaller than 1. Also Rao, Hartley and Cochran assuming the same model have shown that \( V(\hat{Y}_{H.T.})_{G.K.} \) in (3.6.1) is smaller or greater than \( V(T_1) \) in (3.6.2) according as \( g > 1 \) or \( g < 1 \).

Here we will compare the variance expression for \( T_2 \) derived to \( O(N^{-1}) \) with (3.6.1) and (3.6.2) assuming the same model.

It can be easily seen that under the model assumptions (3.6.3) we have
\[ \epsilon V(\hat{g}_{H.T.})_{G.K.} = \frac{a}{n}[\sum_p g^{-1} - (n-1)\sum_p g], \quad (3.6.4) \]

\[ \epsilon V(T_1) = \frac{a}{N}[\sum_p g^{-1} - (n-1)\cdot \frac{1}{N} \sum_p g^{-1}], \quad (3.6.5) \]

and

\[ \epsilon V(T_2) = \frac{a}{n}[\sum_p g^{-1} - (n-1)\cdot \sum_p \sum_p g^{-1}] \quad (3.6.6) \]

where \( \epsilon V(\cdot) \) denotes the average variance under model (3.6.3).

**Theorem 3.7:**

Under the model (3.6.3), variance of the H.T. estimator under the R.H.C. scheme is smaller than that of the R.H.C. estimator for all \( g, 0 \leq g \leq 2 \).

**Proof:**

From (3.6.5) and (3.6.6) we have

\[ \epsilon V(T_1) - \epsilon V(T_2) = \frac{n-1}{n} \cdot a \cdot (\sum_p \sum_p - \frac{1}{N}) \cdot \sum_p g^{-1} \]

\[ \geq 0, \text{ for all } g, \ 0 \leq g \leq 2 \]

because of the inequality \( \sum_p 2 \geq \frac{1}{N} \)

Q.E.D.

**Lemma 3.3:**

For any given set of \( p_t \) 's such that \( 0 \leq p_t \leq 1 \) and \( \sum p_t = 1 \), the expression \( \sum_p \sum_p g^{-1} - \sum_p g \) as a function of \( g \) is monotonically decreasing to the value zero in the domain \([0,2]\).
Proof:

Let
\[ f(g) = \Sigma p_t^2 \cdot \Sigma p_t g^{-1} \Sigma p_t^g = \Sigma p_t^2 \cdot \Sigma p_t g^{-1} \Sigma p_t^g \cdot \Sigma p_t \]

\[ = \Sigma p_t^{g+1} + \sum_{t=1}^{N} p_t^2 \left( \Sigma p_t^{g-1} \right) - \Sigma p_t^{g+1} \]

\[ - \sum_{t=1}^{N} p_t \left( \sum_{t' (\neq t)} p_t^g \right) \]

\[ = \sum_{t=1}^{N} \sum_{t' (\neq t)} p_t^{g-1} (p_t - p_{t'}) \]

\[ = \sum_{t < t'} (p_t - p_{t'}) (p_t p_{t'}, g^{-1} - p_{t'} p_t^{g-1}) \]

\[ = \sum_{t < t'} p_t p_{t'} (p_t - p_{t'}) \cdot \left( \frac{1}{p_t} \right)^{2-g} \cdot \left( \frac{1}{p_{t'}} \right)^{2-g} \]

so for any \(0 \leq g_1 < g_2 < 2\) we have

\[ f(g_1) - f(g_2) = \sum_{t < t'} p_t p_{t'} (p_t - p_{t'}) \cdot \left( \frac{1}{p_t} \right)^{2-g_1} \cdot \left( \frac{1}{p_{t'}} \right)^{2-g_1} \]

\[ - \left( \frac{1}{p_t} \right)^{2-g_2} + \left( \frac{1}{p_{t'}} \right)^{2-g_2} \]
Now, 

\[ p_t < p_t' \iff \left( \frac{1}{p_t} \right)^{2-g_1} < \left( \frac{1}{p_t} \right)^{2-g_1}, \text{ since } g_1 < 2 \]

Also

\[ p_t < p_t' \iff p_t^{g_2 - g_1} < p_t^{g_2 - g_1}, \text{ since } g_2 > g_1 \]

\[ \iff 1 - p_t^{g_2 - g_1} < 1 - p_t^{g_2 - g_1} \]

Therefore

\[ p_t < p_t' \iff \left( \frac{1}{p_t} \right)^{2-g_1} \cdot (1-p_t^{g_2 - g_1}) < \left( \frac{1}{p_t} \right)^{2-g_1} (1-p_t^{g_2 - g_1}) \]

Thus we have,

\[ (p_t - p_t') \cdot \left\{ \left( \frac{1}{p_t} \right)^{2-g_1} (1-p_t^{g_2 - g_1}) - \left( \frac{1}{p_t} \right)^{2-g_1} (1-p_t^{g_2 - g_1}) \right\} \]

\[ > 0, \text{ for } t \neq t' \]

Hence it follows from (3.6.7) that
Therefore we have

\[ f(g_1) - f(g_2) \geq 0 \]

for all \( 0 < g_1 < g_2 \leq 2 \)

In particular we have, for \( 0 < g < 2 \)

\[ f(g) = f(2) = \sum_p t^2 - \sum_p g^2 = 0 \]

Hence \( f \) is monotone decreasing to the value zero in the domain \( [0, 2] \).

Q.E.D.

**Theorem 3.8:**

Under the model (3.6.3) variance of the H.T. estimator under the R.H.C. scheme is smaller than the variance of the H.T. estimator under the Goodman and Kish procedure for all \( g, 0 < g < 2 \).

**Proof:**

From (3.6.4) and (3.6.6) we have

\[ \epsilon V(G.H.T.) - \epsilon V(T_2) = \frac{n-1}{n} \cdot a \cdot \left[ \sum_p t^2 \cdot \sum_p g - \sum_p t^2 \right] \]

\[ > 0 \text{ for all } g, 0 < g \leq 2 \]  \hfill (3.6.8)

in view of Lemma (3.3). Q.E.D.

The difference in the two variances (3.6.4) and (3.6.6) as given by (3.6.8) will be smaller for larger values of \( g \). When \( g = 2 \) both the estimates are equally efficient as it should be expected. However when \( g = 2 \), one should prefer the
Horvitz-Thompson estimator under any scheme with \( P_i = np_i \) in view of its optimum properties under the above model when \( g=2 \) as proved by Godambe (1955) who has established that H.T. estimator is the Bayes estimator under the a priori distribution given by (3.6.3) with \( g=2 \), when any scheme with \( P_i = np_i \) is adopted.

However, it is seldom known in practice whether \( g \) is exactly 2 or not. Thus the H.T. estimator under R.H.C. scheme seems to be more precise in many practical situations than the R.H.C. estimator as well as the H.T. estimator under the Goodman and Kish procedure.

### 3.7. Horvitz-Thompson Type Estimator under R.H.C. Scheme

Under the R.H.C. scheme even if it is quite feasible to compute the exact values of \( P_i \) for the selected units from (3.4.5) to get the unbiased estimate using the computer facilities it may not always be worthwhile due to cost considerations to get the exact values as the approximate expression (3.4.24) may be quite adequate.

As given by (3.4.24), the approximate expression for \( P_i \) correct to \( O(N^{-3}) \), say \( a_i \), is

\[
a_i = np_i [1 + (n-1)(\Sigma p_t^2 - p_i) + (n-1)\frac{n}{N}(p_i - \Sigma p_t^2)] + (n-2)(p_i^2 - \Sigma p_t^3) - 3(n-1) \cdot (p_i - \Sigma p_t^2) \Sigma p_t^2)
\]

(3.7.1)
Let $P_i = a_i + R_i$ \hfill (3.7.2)

where the leading term in $R_i$ is of $O(N^{-4})$. Then the Horvitz-Thompson type estimate proposed is

$$T_2' = \frac{1}{n} \sum_{i=1}^{n} \frac{y_i}{a_i}$$ \hfill (3.7.3)

**Theorem 3.9:**

Bias of the H.T. type estimator $T_2'$, as an estimate of $Y$, is of $O(N^{-2})$ and the mean square error of $T_2'$ correct to $O(N^0)$ is the same as the variance of the H.T. estimator $T_2$ correct to $O(N^0)$.

**Proof:**

The bias of $T_2'$ is

$$B(T_2') = E\left[\frac{y_i}{a_i}\right] - Y$$

Let

$$W_i = \frac{Y_iP_i}{a_i}$$

Thus we have,

$$B(T_2') = E\left[\frac{\sum W_i}{P_i}\right] - Y$$

$$= \frac{\sum W_i - Y}{P_i}$$

$$= \frac{\sum \frac{y_i}{a_i} - \sum P_i}{a_i}$$

$$= \frac{\sum y_i R_i}{\sum a_i}$$ \hfill (3.7.4)
Thus the leading term in the bias is of $O(N^{-2})$ and hence the bias is of $O(N^{-2})$.

Now, we have for the variance of $T_2'$,

$$V(T_2') = V\left[\sum_{i=1}^{n} \frac{y_i}{a_i}\right] = V\left[\sum_{i=1}^{n} \frac{w_i}{p_i}\right] = \sum_{i=1}^{n} \frac{w_i^2}{p_i} + \sum_{i,j\neq i}^{\Sigma} \frac{p_{ij}}{p_ip_j} w_i w_j - (\sum_{i=1}^{n} w_i)^2$$

(3.7.5)

$$\Sigma \frac{w_i^2}{p_i} = \Sigma \frac{y_i^2}{a_i} + \Sigma \frac{y_i^2 \cdot R_i}{a_i}$$

Since the leading term in $\Sigma \frac{y_i^2 \cdot R_i}{a_i}$ is of $O(N^{-1})$, we have

$$\Sigma \frac{w_i^2}{p_i} = \Sigma \frac{y_i^2}{a_i} + O(N^{-1})$$

(3.7.6)
In a similar way it can be seen that

\[ \sum_{i \neq j} \frac{p_{ij}}{p_i p_j} w_i w_j = \sum_{i \neq j} \frac{p_{ij}}{a_i a_j} y_i y_j + O(N^{-1}) \]

\[ = \sum_{i \neq j} \frac{p_{ij}}{p_i p_j} y_i y_j \quad (3.7.7) \]

\[ \frac{N}{(\sum w_i)^2} = Y^2 + \left( \sum \frac{Y_i R_i}{a_i} \right)^2 + 2Y \cdot \sum \frac{Y_i R_i}{a_i} \]

\[ = Y^2 + O(N^{-1}) \quad (3.7.8) \]

Substituting (3.7.6)-(3.7.8) in (3.7.5) we have

\[ V(T'_2) = \frac{N}{\sum p_i} + \sum_{i \neq j} \frac{p_{ij}}{p_i p_j} y_i y_j - Y^2 + O(N^{-1}) \]

\[ = V(T_2), \]

correct to \( O(N^0) \).

Thus we have

\[ \text{MSE}(T'_2) = V(T'_2) + b^2(T'_2) \]

\[ = V(T'_2) + \left( \sum \frac{Y_i R_i}{a_i} \right)^2 \]

\[ = V(T'_2) + O(N^{-4}) \]

\[ = V(T_2), \]

correct to \( O(N^0) \). \( \text{Q.E.D.} \)
3.8. Numerical Illustration

We use the data given in Table 3.1 which is taken from Sukhatme (1953) for comparing the efficiencies of \((\hat{Y}_{H.T.})_{G.K.}, T_1\) and \(T_2\) for estimating the population total. The data gives the number of villages \((X_t)\) and the area under wheat \((Y_t)\) in each of the first 20 administrative areas (out of the total 89) in the Hapur Subdivision of Meerut District (India).

It is required to estimate the total area under wheat in the subdivision using an administrative area (circle) as the unit of sampling.

In Table 3.2 we have presented the numerical values of the variance expressions of each of the estimators \((\hat{Y}_{H.T.})_{G.K.}, T_1\) and \(T_2\) correct to \(O(N^2)\), \(O(N^1)\) and \(O(N^0)\).

The convergence in this example appears to be quite satisfactory although the population size \((N=20)\) is much smaller than those usually encountered in survey work. This indicates that in most of the practical situations the variance formulae to \(O(N^1)\) which is relatively simple should be quite satisfactory. The approximation to \(O(N^2)\) in each of the three cases represents the true variance of the customary estimator in the case of probability proportional to size with replacement estimator, the numerical value of which in this example is given in column 2 of Table 3.2.
Table 3.1. Number of villages and the area under wheat in the administrative circles of Hapur

<table>
<thead>
<tr>
<th>Circle No. (i)</th>
<th>Number of villages $X_i$</th>
<th>Area under wheat $Y_i$</th>
<th>Circle No. (i)</th>
<th>Number of villages $X_i$</th>
<th>Area under wheat $Y_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6</td>
<td>1562</td>
<td>11</td>
<td>3</td>
<td>1027</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>1003</td>
<td>12</td>
<td>4</td>
<td>1393</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>1691</td>
<td>13</td>
<td>3</td>
<td>692</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>271</td>
<td>14</td>
<td>1</td>
<td>524</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>458</td>
<td>15</td>
<td>1</td>
<td>602</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
<td>736</td>
<td>16</td>
<td>3</td>
<td>1522</td>
</tr>
<tr>
<td>7</td>
<td>4</td>
<td>1224</td>
<td>17</td>
<td>4</td>
<td>2087</td>
</tr>
<tr>
<td>8</td>
<td>2</td>
<td>996</td>
<td>18</td>
<td>8</td>
<td>2474</td>
</tr>
<tr>
<td>9</td>
<td>5</td>
<td>475</td>
<td>19</td>
<td>2</td>
<td>461</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>34</td>
<td>20</td>
<td>4</td>
<td>846</td>
</tr>
</tbody>
</table>
Table 3.2. Approximations to the variances to $O(N^2)$, $O(N^1)$, and $O(N^0)$

<table>
<thead>
<tr>
<th>Estimator</th>
<th>$O(N^2)$</th>
<th>$O(N^1)$</th>
<th>$O(N^0)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>H.T. estimator for the Goodman and Kish procedure</td>
<td>51272860</td>
<td>48664940</td>
<td>48523770</td>
</tr>
<tr>
<td>$T_1$ under R.H.C. procedure</td>
<td>51272860</td>
<td>48709210</td>
<td>48581020</td>
</tr>
<tr>
<td>$T_2$ under R.H.C. procedure</td>
<td>51272860</td>
<td>46805760</td>
<td>46726330</td>
</tr>
</tbody>
</table>

Comparing the figures in columns 2, 3, and 4 it is clear that $(\hat{\theta}_{H.T.})_{GK}$, $T_1$ and $T_2$ all fared better relative to the with replacement estimator, $(\hat{\theta}_{H.T.})_{GK}$ fared better than $T_1$, and $T_2$ fared better than both $T_1$ and $(\hat{\theta}_{H.T.})_{GK}$. Concentrating on the column corresponding to $O(N^1)$, it seems that the model (3.6.3) holds good for this population and in particular the value of $g$ lies in between 1 and 2. This particular fact that $g$ most often lies in between 1 and 2 was stated by several authors, as evidenced by the empirical studies conducted.

In order to investigate the validity of the model and the relative performance of the estimators $(\hat{\theta}_{H.T.})_{GK}$, $T_1$ and $T_2$ we have calculated the numerical values of the variance expressions to $O(N^1)$ for several populations and
presented the results in Table 3.3. The several populations that are considered here are taken from the literature. All these populations are the data from actual surveys. Rao and Bayless (1969) have also considered these populations for empirical studies in a different context.

For populations 1 to 6 the relation $V(T_2) < V(\hat{\phi}_{H,T})_{GK} < V(T_1)$ holds which suggests that model (3.6.3) holds good with $1 \leq g \leq 2$. Populations 7, 8 and 9 have been chosen by Cochran as most suitable for the ratio estimate. This fact is stated by Cochran (1963) on page 156. This statement suggests that the model (3.6.3) holds with $g < 1$ because the ratio estimate has lesser variance than the varying probability estimate when $g < 1$. In view of Theorems 3.7 and 3.8, our results for these three populations viz., $V(T_2) < V(T_1) < V(\hat{\phi}_{H,T})_{GK}$ also show the evidence in the same direction. In all the 9 cases the H.T. estimator under R.H.C. scheme is superior to both $(\hat{\phi}_{H,T})_{GK}$ and $T_1$.

3.9. Rao, Hartley, Cochran Scheme with Revised Probabilities

It has already been mentioned that the variance of the H.T. estimator will be zero for any sampling design with $P_i \propto Y_i$. Since sampling with unequal probabilities is resorted to in the situations where $Y_i \propto P_i$ one would expect to
Table 3.3. Table of variances to $O(N^1)$ of the estimators $(\hat{Y}_{H,T})_{GK}$, $T_1$ and $T_2$

<table>
<thead>
<tr>
<th>No.</th>
<th>Source</th>
<th>$Y_i$</th>
<th>$X_i$</th>
<th>$N$</th>
<th>$V(\hat{Y}<em>{H,T})</em>{GK}$</th>
<th>$V(T_1)$</th>
<th>$V(T_2)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Horvitz-Thompson (1952) pp. 663-85</td>
<td>No. of households</td>
<td>Eye-estimated no. of households</td>
<td>20</td>
<td>3031</td>
<td>3084</td>
<td>2988</td>
</tr>
<tr>
<td>2</td>
<td>Sukhatme (1953) circles 1-20 pp. 279-80</td>
<td>Wheat acreage</td>
<td>No. of villages</td>
<td>20</td>
<td>48664940</td>
<td>48709210</td>
<td>46805760</td>
</tr>
<tr>
<td>3</td>
<td>Sukhatme (1953) circles 21-40 pp. 279-80</td>
<td>Wheat acreage</td>
<td>No. of villages</td>
<td>20</td>
<td>26177180</td>
<td>26213980</td>
<td>25956700</td>
</tr>
<tr>
<td>4</td>
<td>Sampford (1962) p. 61</td>
<td>Oats acreage in 1957</td>
<td>Total acreage in 1947</td>
<td>34</td>
<td>99008</td>
<td>100324</td>
<td>94773</td>
</tr>
<tr>
<td>5</td>
<td>Sukhatme (1953) villages 1-34 p. 183</td>
<td>No. of wheat acres in 1936</td>
<td>No. of wheat acres in 1936</td>
<td>34</td>
<td>831885</td>
<td>860902</td>
<td>778324</td>
</tr>
<tr>
<td>6</td>
<td>Desraji (1965) Modification of Horvitz and Thompson's population</td>
<td>No. of households</td>
<td>Eye estimated no. of households</td>
<td>20</td>
<td>8884</td>
<td>8963</td>
<td>8432</td>
</tr>
<tr>
<td>No.</td>
<td>Source</td>
<td>Yi</td>
<td>Xi</td>
<td>N</td>
<td>V(\hat{Y}<em>{H.T.})</em>{GK}</td>
<td>V(T_1)</td>
<td>V(T_2)</td>
</tr>
<tr>
<td>-----</td>
<td>--------</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td>------------------------</td>
<td>----------</td>
<td>--------</td>
</tr>
<tr>
<td>7</td>
<td>Cochran (1963) cities 1-16 p. 156</td>
<td>No. of people in 1930</td>
<td>No. of people in 1920</td>
<td>16</td>
<td>55323</td>
<td>53557</td>
<td>48972</td>
</tr>
<tr>
<td>8</td>
<td>Cochran (1963) cities 17-32 p. 156</td>
<td>No. of people in 1930</td>
<td>No. of people in 1920</td>
<td>16</td>
<td>988912</td>
<td>932978</td>
<td>827495</td>
</tr>
<tr>
<td>9</td>
<td>Cochran (1963) cities 33-48 p. 156</td>
<td>No. of people in 1930</td>
<td>No. of people in 1920</td>
<td>16</td>
<td>188591</td>
<td>182548</td>
<td>145734</td>
</tr>
</tbody>
</table>
gain considerably by choosing a scheme in which $p_i = p_i^*$. This aspect led to the technique of revising the initial probabilities, for a given scheme and selecting the units with these revised probabilities $p_i^*$ where in the $p_i^*$ are chosen so that the condition $P_i = np_i$ is satisfied where $n$ is the sample size. The Midzuno scheme with revised probabilities and the Sampford's scheme that have been dealt with in the previous chapter belong to this category. There are several other schemes in the literature that belong to this group. Each of these schemes has its own limitations and none of these is satisfactory in the survey practitioner's point of view.

In this section we have considered the problem of revising the probabilities $p_i$ and adopting the Rao, Hartley, Cochran scheme with the revised probabilities $p_i^*$ wherein the probabilities $p_i^*$ are chosen so that the condition $P_i = np_i$ is satisfied.

Under the Rao, Hartley, Cochran scheme with revised probabilities $p_i^*$, the expression for $P_i$ correct to $O(N^{-3})$ from (3.4.24) is

$$P_i = np_i^*[1+(n-1)(\Sigma p_t^* - p_i^*)+(n-1)\{\frac{n}{N}(p_i^* - \Sigma p_t^*)^2

+ (n-2)(p_i^* - \Sigma p_t^*)^3 - 3(n-1)(p_i^* - \Sigma p_t^*)^2,\Sigma p_t^* = 2\}]

(3.9.1)$$
where the $p_i^*$ (like $p_i$) is assumed to be of $O(N^{-1})$ and is chosen so that

$$p_i = np_i$$  \hspace{1cm} (3.9.2)

From (3.9.1) and (3.9.2) we get to $O(N^{-2})$,

$$p_i = np_i^* [1 + (n-1)(\Sigma p_t^* - p_i^*)] = np_i$$

Therefore

$$p_i^* = p_i [1 + (n-1)(\Sigma p_t^* - p_i^*)]^{-1}$$
$$= p_i [1 - (n-1)(\Sigma p_t^* - p_i^*)],$$ \hspace{1cm} (3.9.3)

to $O(N^{-2})$.

Therefore $p_i^* = p_i^2 [1 - 2(n-1)(\Sigma p_t^* - p_i^*)]$

$$+ (n-1)^2 (\Sigma p_t^* - p_i^*)^2$$

Summing over all $i$, we get

$$\Sigma p_t^* = \Sigma p_t^2, \hspace{0.5cm} to \hspace{0.5cm} O(N^{-1})$$

So from (3.9.3) we have

$$p_i^* = p_i [1 - (n-1)(\Sigma p_t^2 - p_i)],$$ \hspace{1cm} (3.9.4)

to $O(N^{-2})$.

From (3.9.1) and (3.9.2) we get
\[ p_i^* = p_i \left[ 1 - (n-1)(\sum p_t^* - p_i^*) \right] - (n-1)\left( \frac{n}{N} (p_i^* - \sum p_t^*)^2 \right) \]

\[ + (n-2) (p_i^* - p_i^2) - 3(n-1) (p_i^* - \sum p_t^* - 3(n-1) (p_i^* - \sum p_t^* - 3(n-1) (p_i^* - \sum p_t^*) - 3(n-1) (p_i^* - \sum p_t^*)^2) \right] \]

(3.9.5)

to \(O(N^{-3})\).

Substituting the value of \(p_i^*\) to \(O(N^{-2})\) from (3.9.4) in the right hand side of (3.9.5) we get after simplifying and retaining terms to \(O(N^{-3})\),

\[ p_i^* = p_i \left[ 1 - (n-1)(\sum p_t^* - p_i^*) + \frac{n(n-1)}{N} (\sum p_t^2 - p_i^2) \right] - n(n-1)(\sum p_t^3 - p_i^2) \]

(3.9.6)

to \(O(N^{-3})\).

As a check it can be verified that the \(p_i^*\) given by (3.9.6) satisfies the equation \(\sum p_i^* = 1\). As a more thorough check it can be verified by substituting the value of \(p_i^*\) from (3.9.6) in (3.9.1) and retaining terms to \(O(N^{-3})\), that

\[ P_i = np_i \]  

(3.9.7)

Thus the R.H.C. scheme adopted with probabilities \(p_i^*\) given by (3.9.6) would ensure that \(P_i = np_i\) to \(O(N^{-3})\). The pairwise inclusion probability \(P_{ij}\) for the R.H.C. scheme with probabilities \(p_i^*\) is given by (3.4.64) with \(p_i\) replaced by \(p_i^*\).
Thus

\[ P_{ij} = n(n-1)p_i^*p_j^*[1+\{(2n-3)\Sigma p_t^*2-(n-2)(p_i^*+p_j^*)}\]

\[ + \{(n^2-5n+6)(p_i^*2+p_j^*2)-2(n-2)^2\Sigma p_t^*3\}

\[ + (n^2-6n+6)p_i^*p_j^*-2(n-2)(2n-3)(p_i^*+p_j^*)\Sigma p_t^*2\]

\[ + (7n^2-20n+15)(\Sigma p_t^*2)^2+n(n-1)\frac{(p_i^*+p_j^*)}{N}\]

\[ - 2n(n-1)\frac{\Sigma p_t^*2}{N}\}\]  \hspace{1cm} (3.9.8)

Substituting the value of \( p_i^* \) from (3.9.6) we get after retaining terms to \( O(N^{-4}) \) only,

\[ P_{ij} = n(n-1)p_i^*p_j^*[1+\{p_i^*+p_j^*\}-\Sigma p_t^2]+\{2(p_i^*2+p_j^*2)-2\Sigma p_t^3\}

\[ -(2n-3)p_i^*p_j^*+2(n-2)(p_i^*+p_j^*)\Sigma p_t^2\]

\[ -2(n-2)(\Sigma p_t^2)^2\}], \hspace{1cm} (3.9.9)\]

correct to \( O(N^{-4}) \).

(3.9.9) is the same as (2.3.56) of Theorem 2.8 in Chapter 2, with \( a_n = -(2n-3) \).

Thus (3.9.7) and (3.9.9) show that the R.H.C. scheme with revised probabilities satisfies the conditions of Theorem 2.8. Hence it follows from Theorem 2.8 that the variance of the H.T. estimator denoted by \( T_j \), under the R.H.C. scheme with revised probabilities is given by,
\[ V(T_3) = \frac{1}{n} [\sum p_i z_i^2 - (n-1) \sum p_i^2 z_i ] \]

\[ - \frac{(n-1)}{n} \cdot [2 \sum p_i^3 z_i^2 - \sum p_t^2 \cdot \sum p_i^2 z_i^2 + (2n-3)(\sum p_i^2 z_i^2)] \]

(3.9.10)

correct to \( O(N^0) \), where \( z_i = \left( \frac{y_i}{p_i} - y \right) \).

As an alternative, we can substitute the value of \( p_i^* \) in the variance expression (3.4.69) with \( p_i \) replaced by \( p_i^* \), simplify and retain the terms to \( O(N^0) \). Then also we will arrive at the same expression (3.9.10) which provides a check.

In Chapter II, we have derived the variance of the H.T. estimator for the Sampfords procedure, correct to \( O(N^0) \) which is given by

\[ V(\hat{y}_{H.T.})_{Samp} = \frac{1}{n} [\sum p_i z_i^2 - (n-1) \sum p_i^2 z_i ] \]

\[ - \frac{(n-1)}{n} [2 \sum p_i^3 z_i^2 - \sum p_t^2 \cdot \sum p_i^2 z_i^2 \]

\[ + (n-2)(\sum p_i^2 z_i^2)] \]

(3.9.11)

correct to \( O(N^0) \).

Also we have shown that

\[ V(\hat{y}_{H.T.})_{Samp} \leq V(\hat{y}_{H.T.})_{GK}, \text{ for all } n. \]  

(3.9.12)

Now, from (3.9.10) and (3.9.11) we have

\[ V(\hat{y}_{H.T.})_{Samp} - V(T_3) = \frac{(n-1)}{n} \cdot (\sum p_i^2 z_i^2)^2 \geq 0 \]  

(3.9.13)
Thus from (3.9.12) and (3.9.13) we have

\[ V(T_3) \leq V(\hat{\gamma}_{H.T.})_{\text{samp}} \leq V(\hat{\gamma}_{H.T.})_{\text{GK}}, \]

for all sample sizes.

In fact (3.9.13) together with the equation

\[ V(\hat{\gamma}_{H.T.})_{\text{GK}} - V(\hat{\gamma}_{H.T.})_{\text{samp}} = (n-1) \cdot (\sum \pi_i \delta_i^2 \gamma_i) \]

as given in (2.3.65), imply that the Sampford's scheme is almost in the midway between the R.H.C. scheme with revised probabilities and the Goodman and Kish procedure, in regard to its performance as measured by the variance. Thus it seems that for sample sizes that are large in absolute terms but small relative to N, it will be advantageous to adopt the R.H.C. scheme with revised probabilities. Also the procedure is considerably simpler to adopt relative to the procedures of Goodman and Kish, and Sampford. Of course when one is confident that model (3.6.3) holds, it would be more advantageous to use the R.H.C. scheme with the original probabilities, the estimator to be used being the H.T. estimator.

3.10. The Improved R.H.C. Estimator

In section 3.3. we have given a heuristic argument and concluded that in the case of R.H.C. scheme the subset s of the population U that has been selected together with the
respective \( y \) values forms a sufficient statistic. Here we will give a more rigorous proof for the same.

Suppose the outcome \( \omega \) of the sampling experiment is given by \( \omega = (x_{i_1}, x_{i_2}, \ldots x_{i_n}) \) where \( x_{i_j} = (i_j, G_{i_j}), \ j = 1, 2, \ldots n; \ (i_1, i_2, \ldots i_n) \) is the subset \( s \) of \( U \) that has been selected, \( G_{i_j} \) is the random group of units to which \( i_j \) belongs, \( j = 1, 2 \ldots n. \)

Now consider the subset \( s = (i_1, i_2, \ldots i_n) \) of \( U \) that has been selected by the sampling experiment. From (3.4.1), the total number of distinct partitions is given by \( A = \frac{N!}{n! (M!)^n}. \) Among these there are only

\[
\nu_{i_1, i_2 \ldots i_n} = \frac{n-2}{r=0} (n-r)(M-1) ^{M-1}
\]

number of partitions that could possibly give rise to the sample \( (i_1, i_2, \ldots i_n) \) and each of these partitions has a probability of \( \frac{1}{A} \) to materialize.

Conditional probability of the units \( (i_1, i_2, \ldots i_n) \) to get selected for a given partition is given by

\[
P(i_1, i_2 \ldots i_n | \text{partition}) = \frac{p_{i_1}}{S_{i_1}} \cdot \frac{p_{i_2}}{S_{i_2}} \ldots \frac{p_{i_n}}{S_{i_n}},
\]

where \( S_{i_j} \) is the sum of the \( p_t \)'s of the units in \( G_{i_j} \) (\( j = 1, 2, \ldots n \)).

Therefore probability of selecting the sample \( (i_1, i_2 \ldots i_n) \) is
Thus from the standard relation,

\[ P(x_{i_1}, x_{i_2}, \ldots, x_{i_n} | i_1, i_2, \ldots, i_n) = \frac{P(x_{i_1}, x_{i_2}, \ldots, x_{i_n}, i_1, i_2, \ldots, i_n)}{P(i_1, i_2, \ldots, i_n)} \]  

(3.10.2)

we get

\[ P(x_{i_1}, x_{i_2}, \ldots, x_{i_n} | i_1, i_2, \ldots, i_n) = \frac{P(x_{i_1}, x_{i_2}, \ldots, x_{i_n})}{P(i_1, i_2, \ldots, i_n)} \]  

(3.10.3)

because

\[ P(x_{i_1}, x_{i_2}, \ldots, x_{i_n}; i_1, i_2, \ldots, i_n) = P(x_{i_1}, x_{i_2}, \ldots, x_{i_n}) \]

But

\[ P(x_{i_1}, x_{i_2}, \ldots, x_{i_n}) = \frac{1}{A} \cdot \frac{P_{i_1}}{S_{i_1}} \cdot \frac{P_{i_2}}{S_{i_2}} \cdot \cdots \cdot \frac{P_{i_n}}{S_{i_n}} \]  

(3.10.4)

Thus substituting (3.10.1) and (3.10.4) in (3.10.2) we get

\[ P(x_{i_1}, x_{i_2}, \ldots, x_{i_n} | i_1, i_2, \ldots, i_n) = \frac{1}{S_{i_1} S_{i_2} \cdots S_{i_n}} \sum_{v_{i_1 i_2 \ldots i_n}} \frac{1}{S_{i_1} S_{i_2} \cdots S_{i_n}} \]  

(3.10.5)
Using (3.10.5) we get from (3.10.3),

\[
P(X_1, X_2, \ldots, X_n) = P(i_1, i_2, \ldots, i_n) \cdot \prod_{i=1}^{n} \frac{1}{S_i} \frac{S_{i_1} \cdots S_{i_n}}{S_{i_1} \cdots S_{i_n}}
\]

(3.10.6)

The expression in the square brackets on the right hand side of (3.10.6) can be calculated from the information of \( s = (i_1, i_2, \ldots, i_n) \) alone. Thus from the Neyman's factorization criterion it follows that \( s = (i_1, i_2, \ldots, i_n) \) together with its \( y \)-values forms a sufficient statistic. Hence any estimator that is not a function of \( s \) alone can be uniformly improved upon by using the Rao-Blackwell theorem.

Thus the improved R.H.C. estimator is given by

\[
T'_1 = E[T_1 | i_1, i_2, \ldots, i_n],
\]

(3.10.7)

where

\[
T_1 = \sum_{j=1}^{n} \frac{Y_i}{P_{ij}} \cdot S_{ij}
\]

(3.10.8)

Therefore we have

\[
T'_1 = \sum_{\{v_1, i_1, i_2, \ldots, i_n\}} \frac{S_{i_1} \cdots S_{i_n}}{S_{i_1} \cdots S_{i_n}} T_1 \cdot P(X_1, X_2, \ldots, X_n | i_1, i_2, \ldots, i_n),
\]

(3.10.9)

which upon using (3.10.5) gives,
Now, we have from (3.10.9),

\[
E(T'_1) = \sum_{1}^{N} T'_1 \cdot P(i_1, i_2, \ldots, i_n),
\]

where the summation is over all possible \((\binom{N}{n})\) subsets of \(U\).

Therefore,

\[
E(T'_1) = \sum_{1}^{N} \left\{ \sum_{\{v_{i_1, i_2, \ldots, i_n}\}} T'_1 \cdot P[x_{i_1}, x_{i_2}, \ldots, x_{i_n} | i_1, i_2, \ldots, i_n] \right\} \cdot P(i_1, i_2, \ldots, i_n)
\]

\[
= \sum_{1}^{N} \left\{ \sum_{\{v_{i_1, i_2, \ldots, i_n}\}} T'_1 \cdot P[x_{i_1}, x_{i_2}, \ldots, x_{i_n}] \right\}
\]

\[
= E(T'_1) = Y
\]

Therefore \(T'_1\) is an unbiased estimate of \(Y\). Also

\[
V(T'_1) = E(T'_1^2) - E^2(T'_1)
\]

\[
= \sum_{1}^{N} T'_1^2 \cdot P(i_1, i_2, \ldots, i_n) - Y^2
\]
and

\[ V(T_1) = E(T_1^2) - E^2(T_1) \]

\[
\begin{align*}
N \\
(n) \\
= \sum_l \sum \{v_{i_1i_2\ldots i_n} \} T_{1,l}^2 \cdot P(x_{i_1}x_{i_2}\ldots x_{i_n}) - \gamma^2
\end{align*}
\]

Therefore we have,

\[
\begin{align*}
V(T_1) - V(T_1') &= \sum_l \sum \{v_{i_1i_2\ldots i_n} \} T_{1,l}^2 \cdot P(x_{i_1}x_{i_2}\ldots x_{i_n}) \\
&- \sum_l T_{1,l}^2 \cdot P(i_1,i_2,\ldots i_n)
\end{align*}
\]

\[
\begin{align*}
&= \sum_l \sum \{v_{i_1i_2\ldots i_n} \} T_{1,l}^2 \cdot P(x_{i_1}x_{i_2}\ldots x_{i_n}) \\
&- \sum_l \sum \{v_{i_1i_2\ldots i_n} \} [T_{1,l} - \sum \{v_{i_1i_2\ldots i_n} \} T_{1,l} \cdot \frac{P(x_{i_1}x_{i_2}\ldots x_{i_n})}{P(i_1,i_2,\ldots i_n)}]^2 \cdot P(i_1,i_2,\ldots i_n)
\end{align*}
\]
This represents the improvement of the estimator $T'_1$ over the estimator $T_1$. For general sample size, however, it is rather troublesome to compute the conditional probabilities (3.10.5), which is needed to compute the estimator $T'_1$.

It is relatively simple for sample size 2 to investigate the properties of $T'_1$. So we will deal only with the case $n=2$.

3.11. Improved R.H.C. Estimator for the Case $n=2$ and $N$ Large

For sample size 2, the probability of getting the subset $(i_1, i_2)$, $P(i_1, i_2)$ is the same as the inclusion probability $P_{i_1i_2}$ considered in Section 3.4.

Therefore from (3.10.10) we have

$$T'_1 = \sum_{G_2(i_1, i_2)} \left\{ \frac{Y_{i_1}}{P_{i_1}} \cdot S_{i_1} + \frac{Y_{i_2}}{P_{i_2}} \cdot S_{i_2} \right\}$$

$$= \sum_{G_2(i_1, i_2)} \frac{1}{A} \cdot \frac{P_{i_1}P_{i_2}}{S_{i_1}S_{i_2}} \cdot \left\{ \frac{1}{P(i_1, i_2)} \cdot \frac{Y_{i_1}}{P_{i_1}} + \frac{Y_{i_2}}{P_{i_2}} \cdot S_{i_2} \right\}$$
\[ A_{P(i_1', i_2')} = \frac{A^2}{A \cdot P(i_1, i_2)} \cdot [Y_{i_1} P_{i_2} \cdot \frac{1}{A^2} \sum_{i_1, i_2} \frac{1}{G_2(i_1, i_2) S_{i_2}} ] + Y_{i_2} P_{i_1} \cdot \frac{1}{A^2} \sum_{i_1, i_2} \frac{1}{G_2(i_1, i_2) S_{i_1}} \] (3.11.1)

From (3.4.28) we have
\[ P(i_1, i_2) = P_{i_1} P_{i_2} \cdot \frac{A^2}{A^2} \cdot \frac{1}{E[S_{i_1} S_{i_2}]} \]

where \( E \) denotes the expectation taken over the scheme considered in Lemma 3.2 with \( K=2 \).

Thus, we have
\[ \frac{A^2}{A \cdot P(i_1, i_2)} = \frac{1}{P_{i_1} P_{i_2}} \cdot \frac{1}{E[S_{i_1} S_{i_2}]} \]

Substituting the value of \( E[S_{i_1} S_{i_2}] \) from Lemma 3.2 with \( K=2 \), we get after simplifying and retaining terms to \( O(N^0) \),
\[ \frac{A^2}{A \cdot P(i_1, i_2)} = \frac{1}{4P_{i_1} P_{i_2}} \cdot \frac{2(p_{i_1} + p_{i_2})}{[1 - (\Sigma p_t^2 - 1/N) - \frac{2(p_{i_1} + p_{i_2})}{N} + \frac{3\Sigma p_t^2}{N} + 2(\Sigma p_t^2)^2 - \frac{1}{N^2} - 2p_{i_1} p_{i_2}]} \] (3.11.2)

Also,
\[
\frac{1}{A_2} \sum_{G_2(i_1, i_2)} \frac{1}{S_{i_1}} = E\left[\frac{1}{S_{i_1}}\right],
\]

and

\[
\frac{1}{A_2} \sum_{G_2(i_1, i_2)} \frac{1}{S_{i_2}} = E\left[\frac{1}{S_{i_2}}\right]
\]

where \( E \) denotes the same as above.

Thus substituting the values of \( E\left[\frac{1}{S_{i_2}}\right] \) from Lemma 3.2
we have

\[
\frac{1}{A_2} \sum_{G_2(i_1, i_2)} \frac{1}{S_{i_1}} = 2\left[1 + \left(\sum_{\Sigma_{p_i}}^2 - \frac{1}{N}p_{11} - p_{i2}\right) \right. \\
+ \left\{3\left(\sum_{\Sigma_{p_i}}^2 \right)^2 - \frac{5}{5} \sum_{\Sigma_{p_i}}^2 \left. \left[\frac{p_{11}}{N} + \frac{5p_{i2}}{N} - 2p_{i1}p_{i2}\right]\right\},
\]

correct to \( O(N^{-2}) \).

Thus, from (3.11.2) and (3.11.3) we get

\[
\frac{A_2}{A \cdot P(i_1, i_2) \cdot Y_{i_1}p_{i_2}} \cdot \sum_{G_2(i_1, i_2)} \frac{1}{S_{i_2}} \frac{1}{A_2} \sum_{G_2(i_1, i_2)} \frac{1}{S_{i_2}}
\]

\[
= \frac{Y_{i_1}}{2p_{i_1}} \left[1 + (p_{i_1} - p_{i_2}) + 2(p_{i_1} - p_{i_2}) \left(\sum_{\Sigma_{p_i}}^2 - \frac{1}{N}\right)\right]
\]

correct to \( O(N^{-1}) \).
By interchanging $i_1$ and $i_2$ in the above we get

$$\frac{A_2}{A \cdot P(i_1', i_2')} \cdot Y_{i_2} P_{i_1} \cdot \frac{1}{A_2} G_2(i_1', i_2') \frac{1}{S_{i_1}}$$

$$= \frac{Y_{i_2}}{2P_{i_2}} [1+(P_{i_2}-P_{i_1})+2(P_{i_2}-P_{i_1})(\Sigma P_t^2 - \frac{1}{N})]$$

(3.11.5)

correct to $O(N^{-1})$.

Substituting (3.11.4) and (3.11.5) in (3.11.1) we get,

$$T_1' = \frac{1}{2} \left[ \frac{Y_{i_1}}{P_{i_1}} + \frac{Y_{i_2}}{P_{i_2}} \right] + \frac{1+2(\Sigma P_t^2 - \frac{1}{N})}{(P_{i_1}-P_{i_2})(\Sigma P_t^2 - \frac{1}{N})} (P_{i_1}-P_{i_2})$$

(3.11.6)

correct to $O(N^{-1})$.

Variance of the estimator $T_1'$ is

$$V(T_1') = \frac{1}{2} \cdot \sum_{i_1=1}^{N} \sum_{i_2 \neq i_1}^{N} T_1'^2 \cdot P(i_1, i_2) - Y^2$$

(3.11.7)

Using the value of $P(i_1, i_2)$ from (3.4.64) with $n=2$ and the expression for $T_1'$ from (3.11.6) we get

$$T_1'^2 \cdot P(i_1, i_2) = \frac{(1+\Sigma P_t^2)}{2} \left[ Y_{i_1}^2 \left( \frac{P_{i_2}}{P_{i_1}} + 2P_{i_2} - \frac{2P_{i_2}^2}{P_{i_1}} \right) + Y_{i_2}^2 \left( \frac{P_{i_1}}{P_{i_2}} + 2P_{i_1} - \frac{2P_{i_1}^2}{P_{i_2}} \right) + 2Y_{i_1}Y_{i_2} \right]$$
Summing this over all the combinations \((i_1, i_2)\) we get

\[
\sum_{i_1}^N \sum_{i_2}^2 \Sigma T_{1}^{2} P(i_1, i_2) = \frac{1}{2} \left[ \Sigma \frac{y_t^2}{p_t} + \Sigma \frac{y_t^2}{p_t} (\Sigma \frac{y_t^2}{p_t} - y^2) \right],
\]

correct to \(O(N^1)\).

Substituting this in (3.11.7) we get

\[
V(T_1') = \frac{1}{2} (1 - \Sigma \frac{2}{p_t}) \left[ \Sigma \frac{y_t^2}{p_t} - y^2 \right]
\]

correct to \(O(N^1)\).

To the same order, from (3.6.2) we have

\[
V(T_1) = \frac{1}{2} (1 - \frac{1}{N}) \cdot \left[ \Sigma \frac{y_t^2}{p_t} - y^2 \right]
\]

From (3.11.8) and (3.11.9) we have

\[
V(T_1) - V(T_1') = \frac{1}{2} (\Sigma p_t^2 - \frac{1}{N}) \left[ \Sigma \frac{y_t^2}{p_t} - y^2 \right]
\]

which is always nonnegative because of the inequality

\[
\Sigma p_t^2 \geq \frac{1}{N}.
\]

Under the model (3.6.3) we get

\[
\epsilon V(T_1') = \frac{a}{2} [\Sigma p_t g - 1 - \Sigma p_t^2 \Sigma p_t g - 1]
\]

which is the same as \(\epsilon V(T_2)\) for \(n=2\) as given in (3.6.6).

This implies that the improved R.H.C. estimator is more efficient than the R.H.C. estimator as well as the H.T. estimator under the Goodman and Kish procedure under
model (3.6.3) for all values of g.

Also under the model (3.6.3), the improved R.H.C. estimator and the H.T. estimator for the R.H.C. scheme are equally efficient for all values of g. However in view of the other optimal properties of the H.T. estimator one should prefer this estimator.
4. SOME RANDOMIZED VARYING PROBABILITY SCHEMES

We have mentioned already the optimal properties that sampling schemes satisfying the condition $P_i = np_i$ possess when the $y_i$ is approximately proportional to $p_i$. We have also mentioned the scarcity of such schemes in the literature which are practically useful for samples of arbitrary size. Moreover the strict applicability of the existing methods of unequal probability sampling without replacement including the calculation of unbiased estimates of sampling error is out of question in certain kinds of large scale survey work on grounds of practicability. Thus there is a need for evolving methods which retain the advantages of unequal probability sampling without replacement but are rather easier to apply in practice and only involve a slight loss of exactness. In this chapter we will investigate the role of randomization in getting schemes, that are practically useful and are applicable in large scale surveys, by making use of the schemes that are useful for smaller sample sizes.

4.1. An Exact Sampling Scheme for Sample Size 2

In this section we will present a scheme for sample size 2 such that the overall probability $P_i$ of selecting the $i$th unit in the sample is proportional to $p_i$. The
scheme is described as follows:

(i) Split the population at random into three groups of equal sizes, and select two groups from among these three such that the inclusion probability of any group is proportional to the sum of the \( p_t \)'s of the units belonging to that group.

(ii) Select one unit each from the two selected groups independently with probability proportional to \( p_t \)'s.

We will denote this as scheme 4.1.

Adopting the same notations as in Chapter 3, we get that for scheme 4.1, the probability of including the \( i \)th unit in the sample is given by

\[
P_i = \frac{1}{A} \cdot \sum_{G} P(g) \cdot \frac{P_i}{S_g},
\]

where \( P(g) \) is the probability of including the \( g \)th primary stage unit (p.s.u.) which contains \( U_i \), in a given partition and is given by \( P(g) = 2S_g \), where \( S_g \) is the sum of the \( p_t \)'s of the units belonging to the \( g \)th p.s.u.

Thus from (4.1.1) we have,

\[
P_i = \frac{1}{A} \cdot \sum_{G} 2S_g \cdot \frac{P_i}{S_g}
= 2p_i
\]

The probability \( P_{ij} \) of including the pair \( (U_i, U_j) \) in the sample is
where $P(r,s)$ is the probability of selecting the $r$th and $s$th p.s.u.'s together which contain respectively the $i$th and $j$th population units in a given partition of $G_2(i,j)$.

The expression for $P(r,s)$ is known to be given by

$$p(r,s) = p(r) + p(s) - 1,$$  \hspace{1cm} (4.1.4)

where $p(r)$ is the probability of including the $r$th p.s.u. and $p(s)$ is the probability of including the $s$th p.s.u. and are given by

$$p(r) = 2S_r$$  \hspace{1cm} (4.1.5)

and

$$p(s) = 2S_s$$  \hspace{1cm} (4.1.6)

substituting the values from (4.1.4)-(4.1.6) in (4.1.3) we get,

$$p_{ij} = \frac{1}{A} \cdot \sum_{G_2(i,j)} p(r,s) \cdot \frac{p_i}{S_r} \cdot \frac{p_j}{S_s},$$  \hspace{1cm} (4.1.3)

where $E$ denotes the operation of taking the expectation over the scheme of selecting two without replacement simple random samples of size $(N-2)/3$ each from the population of $(N-2)$ units excluding $U_i$ and $U_j$. 

$$= 2p_ip_j \cdot \frac{A^2}{A} \cdot E\left[\frac{1}{S_r} + \frac{1}{S_s} - \frac{1}{2S_rS_s}\right],$$  \hspace{1cm} (4.1.7)
From (3.4.30), since the population is divided into three
groups, we have correct to \( O(N^{-2}) \),

\[
\frac{A_2}{A} = \frac{2}{3} \left[ 1 + \frac{1}{N} + \frac{1}{N^2} \right] \tag{4.1.8}
\]

Further, by using Equations (3.4.31)-(3.4.33) of Lemma 3.2
with \( K=3 \), we get

\[
E \left[ \frac{1}{S_r} + \frac{1}{S_s} - \frac{1}{2S_r S_s} \right]
\]

\[
= \frac{3}{2} \left[ 1 + \{(p_i + p_j) - \Sigma p_t^2 - \frac{1}{N}\} + \{2(p_i^2 + p_j^2) \right.

\]

\[
- 7p_i p_j - \frac{(p_i + p_j)}{N} + 6(p_i + p_j) \Sigma p_t^2
\]

\[
+ \frac{\Sigma p_t^2}{N} - 2 \Sigma p_t^3 - 6(\Sigma p_t^2)^2 \right] \right], \tag{4.1.9}
\]

correct to \( O(N^{-2}) \).

Using (4.1.8) and (4.1.9) we get from (4.1.7),

\[
P_{ij} = 2p_i p_j \left[ 1 + \{(p_i + p_j) - \Sigma p_t^2\} + \{2(p_i^2 + p_j^2) - 2 \Sigma p_t^3 \right.

\]

\[
- 7p_i p_j + 6(p_i + p_j) \Sigma p_t^2 - 6(\Sigma p_t^2)^2 \right] \right], \tag{4.1.10}
\]

correct to \( O(N^{-4}) \).

Thus from (4.1.2) and (4.1.10) it follows that the
scheme under consideration satisfies the conditions of
Theorem 2.8 with \( a_n = -7 \) and so we have by applying the theorem,

\[
V(\hat{\gamma}_{H.T.}) = \frac{1}{2}[\sum z_i^2 - \sum z_i^2] - \frac{1}{2}[2\sum z_i^2 - \sum z_i^2]
+ 7(\sum z_i^2),
\]

(4.1.11)
correct to \( O(N^0) \).

To the same approximation the variance of the corresponding H.T. estimator under the Durbin's (1967) procedure is given by (2.3.63) with \( n=2 \), because Sampford's procedure is a generalization of the Durbin's procedure for sample size 2.

Thus we have,

\[
V(\hat{\gamma}_{H.T.}) = \frac{1}{2}[\sum z_i^2 - \sum z_i^2] - \frac{1}{2}[2\sum z_i^2 - \sum z_i^2]
\]

(4.1.12)
correct to \( O(N^0) \).

From (4.1.11) and (4.1.12) we get,

\[
V(\hat{\gamma}_{H.T.}) - V(\hat{\gamma}_{H.T.}) = \frac{7}{2}(\sum z_i^2) \geq 0,
\]

which implies that the H.T. estimator under scheme 4.1 is always more efficient than the corresponding H.T. estimator under the Durbin's procedure.
4.2. An Alternative Exact Sampling Scheme Utilizing the Durbin's Procedure

In this section we will present a scheme for sample size 2, such that the overall probability $P_i$ of selecting the $i$th unit in the sample is $2p_i$, which utilizes the Durbin's method of sampling. The scheme is as follows:

(i) Split the population at random into three groups of equal sizes and select one group from among the three groups with probability proportional to the sum of the $p_t$'s of the units belonging to that group.

(ii) Select two units utilizing the Durbin's procedure from the group that has been selected in step (i) utilizing the $p_t$'s.

We will denote this as scheme 4.2. For scheme 4.2, the probability $P_i$, of including the $i$th unit in the sample is given by

$$P_i = \frac{1}{A} \cdot \sum_{g} p(g) \cdot \frac{2p_i}{S_g}$$  \hspace{1cm} (4.2.1)

where $p(g)$, the probability of selecting the $g$th p.s.u. is $p(g) = S_g$.

Thus we have from (4.2.1)

$$P_i = 2p_i$$  \hspace{1cm} (4.2.2)

The expression for $P_{ij}$ of this scheme is
\[ P_{ij} = \frac{1}{A} \cdot \sum_{G(i,j)} S_q D_{ij} \]  

(4.2.3)  

where \( S_q \) is the sum of the \( p_t \)'s of the units belonging to the \( q \)th p.s.u., that contain the pair \((U_i, U_j)\), of a given partition of \( G_1(i,j) \); and \( D_{ij} \) is the probability of including the pair \((U_i, U_j)\) together under the Durbin's procedure, given the \( q \)th p.s.u. From (2.1.3) we have

\[ D_{ij} = \frac{2 \frac{p_i}{S_q} \frac{p_j}{S_q}}{\frac{1}{1+\Sigma'p_t/S_q} - \frac{2\Sigma'p_t/S_q}{1-2p_t/S_q}} \]  

(4.2.4)

where \( \Sigma' \) denotes the summation taken over all the units that belong to the \( q \)th p.s.u. From (2.3.52) we have after replacing \( p_t \) by \( p_t/S_q \).

\[ D_{ij} = \frac{2p_i p_j}{S_q^2} \left[ 1 + \left( \frac{p_i + p_j}{S_q} \right) - \Sigma' \frac{p_t}{S_q}^2 - \frac{2(\Sigma'p_t/S_q)^3}{S_q^3} \right] \]  

(4.2.5)

\[ + \frac{2(\Sigma'p_t/S_q)^2}{S_q} - \frac{\Sigma'p_t}{S_q}^2 \]  

correct to \( O(N^{-4}) \).

(4.2.3) can be written as

\[ P_{ij} = \frac{A_1}{A} \cdot \frac{1}{A_1} \cdot \sum_{G_1(i,j)} S_q D_{ij} \]
\[ \frac{A_1}{A} \cdot E[\sum_{q} D_{ij}] \tag{4.2.6} \]

where \( E \) denotes the operation of taking the expectation over the scheme of selecting \( \left( \frac{N}{3} - 2 \right) \) units from among the \( (N-2) \) population units excluding \( U_i \) and \( U_j \). Utilizing (4.2.5) we get after retaining terms that contribute to \( P_{ij} \) up to \( O(N^{-4}) \),

\[ P_{ij} = 2p_ip_j \cdot \frac{A_1}{A} \cdot E[\frac{1}{S_q}] + \frac{(p_i + p_j)}{S_q^2} + \frac{(p_i^2 + p_j^2)}{S_q^3} \]

\[ - \frac{\Sigma^p_t}{S_q^3} - (p_i + p_j) \cdot \frac{\Sigma^p_t}{S_q^4} - \frac{2\Sigma^p_t}{S_q^4} + \frac{(\Sigma^p_t)^2}{S_q^5} \],

\[ \tag{4.2.7} \]

where \( \Sigma^p \) denotes the summation running over all the units belonging to the \( q \)th p.s.u. excepting \( U_i \) and \( U_j \).

For evaluating the expectations of the individual terms in (4.2.7) we will state a lemma the results of which will be used in the later sections also.

**Lemma 4.1:**

Let \( p_t \) be a variate defined over a population of size \( N \) where in \( p_t \) is assumed to be of \( O(N^{-1}) \). Let \( N \) be a multiple of \( K \) where \( K \) is small relative to \( N \). Consider the scheme of selecting a simple random sample of size \( \frac{N}{K} - 2 \) from the population of \( (N-2) \) units excluding \( U_i \) and \( U_j \). Let \( S_q \)' be the sum of the \( p_t \)'s of the units belonging to this sample and
let \( S_q = p_i + p_j + s_q' \). Then we have

\[
E \left[ \frac{1}{S_q} \right] = K \left[ 1 + (K-1) \left\{ \Sigma p_t^2 + \frac{1}{N} - (p_i + p_j) \right\} \right]
\]

\[+ (K-1) \left\{ (K-2) (p_i^2 + p_j^2) - (K-1) \frac{(p_i + p_j)}{N} \right\}
\]

\[+ 2(K-1)p_ip_j - 3(K-1) (p_i + p_j) \Sigma p_t^2 - (K-2) \Sigma p_t^3
\]

\[+ 3(K-1) (\Sigma p_t^2)^2 + (K-1) \frac{\Sigma p_t^2}{N} + \frac{K}{N^2} \} \right], \quad (4.2.8)
\]

correct to \( O(N^{-2}) \),

\[
E \left[ \frac{1}{S_q^2} \right] = K^2 \left[ 1 + (K-1) \left\{ 3 \Sigma p_t^2 + \frac{1}{N} - 2(p_i + p_j) \right\} \right], \quad (4.2.9)
\]

correct to \( O(N^{-1}) \),

\[
E \left[ \frac{1}{S_q^3} \right] = K^3, \quad (4.2.10)
\]

correct to \( O(N^0) \),

\[
E \left[ \frac{\Sigma p_t^2}{S_q} \right] = K^2 \left[ \Sigma p_t^2 - (p_i^2 + p_j^2) - 3(K-1) (p_i + p_j) \Sigma p_t^2
\]

\[- 3(K-1) \Sigma p_t^3 + 6(K-1) (\Sigma p_t^2)^2 + (K-1) \frac{\Sigma p_t^2}{N} \] \right], \quad (4.2.11)
\]

correct to \( O(N^{-2}) \),

\[
E \left[ \frac{\Sigma p_t^2}{S_q^2} \right] = K^3 \Sigma p_t^2, \quad (4.2.12)
\]

Correct to \( O(N^{-1}) \),
\[
E\left[\frac{\Sigma P_t^3}{\Sigma q}\right] = k^3 \cdot \Sigma P_t^3, \quad (4.2.13)
\]
correct to \(O(N^{-2})\),

and

\[
E\left[\frac{(\Sigma P_t^2)^2}{\Sigma q}\right] = k^3 \cdot (\Sigma P_t^2)^2, \quad (4.2.14)
\]

Proof of the above lemma is in the same lines as of Lemmas 3.1 and 3.2 and hence is omitted.

From (4.1.8) we have

\[
\frac{A_1}{A} = \frac{1}{3} \left[ 1 - \frac{2}{N} - \frac{2}{N^2} \right] \quad (4.2.15)
\]
correct to \(O(N^{-2})\).

Using the results of Lemma 4.1 and Equation (4.2.15) we get from (4.2.7),

\[
P_{ij} = 2p_i p_j \left[ 1 + \{p_i + p_j \} - \Sigma P_t^2 \right] + \left\{ 2(p_i^2 + p_j^2) - 2\Sigma P_t^3 \right\}
- 16p_i p_j + 15(p_i + p_j)^2 P_t^2 - 15(\Sigma P_t^2)^2 \quad (4.2.16)
\]
correct to \(O(N^{-4})\).

Thus from (4.2.2) and (4.2.16) it follows that the scheme under consideration satisfies the conditions of Theorem 2.8 with \(a_n = -16\) and so we have by applying the theorem.
\[
V(\hat{y}_{H.T.}) = \frac{1}{2} \left[ \sum p_i z_i^2 - \sum p_i z_i^2 \right] - \frac{1}{2} \left[ 2 \sum p_i \sum z_i^2 - \sum p_i \sum z_i^2 \right] + 16 \left( \sum p_i z_i \right)^2
\]  

(4.2.17)
correct to \(O(N^0)\).

A direct comparison of expression (4.2.17) with the expression (4.1.11) shows that \(V(\hat{y}_{H.T.})\) for the scheme 4.2 is uniformly smaller than \(V(\hat{y}_{H.T.})\) for the scheme 4.1 and hence than \(V(\hat{y}_{H.T.})\) corresponding to the Durbin's procedure.

4.3. Role of Random Stratification in Getting Improved Estimates

In Sections 4.1 and 4.2 we have presented two unequal probability schemes for sample size 2 which give better estimates than most of the existing schemes. The idea of random stratification has been utilized in both the schemes as in the Rao, Hartley and Cochran's procedure. In this section we will discuss the role of random stratification in getting an improved estimate for any given scheme that satisfy the conditions (2.3.55) and (2.3.56) of Theorem 2.8.

We will adopt the following procedure for selecting a sample of size 2n:

(i) Split the population at random into three groups of equal sizes, and select two groups from among these
three such that the inclusion probability of any group is proportional to the sum of the p_t's of the units belonging to that group.

(ii) Select n units each from the two selected groups independently by adopting any I.P.P.S. (inclusion probability proportional to size) scheme that satisfy the conditions (2.3.55) and (2.3.56) of Theorem 2.8. We will call this procedure as scheme 4.3. With the same notations used in sections 4.1 and 4.2 we have for the scheme 4.3, the inclusion probability \( P_i \) given by

\[
P_i = \frac{1}{A} \sum_{G} 2S_g \cdot \frac{np_i}{S_g} = 2np_i
\]

and

\[
P_{ij} = \frac{1}{A} \sum_{G_1(i,j)} 2S_q \cdot p_{ij}^{(q)} + \frac{1}{A} \sum_{G_2(i,j)} p(r,s) \cdot \frac{np_i}{S_r} \cdot \frac{np_j}{S_s}
\]

where \( p_{ij}^{(q)} \) is the probability of including the pair of units \((U_i,U_j)\) when step (ii) is adopted in the qth group that contains \(U_i\) and \(U_j\) in a given partition of \(G_1(i,j)\); and \( p(r,s) \) is the probability of including the rth and sth groups together when step (i) is adopted where the rth group contains \(U_i\) and sth group contains \(U_j\) in a given partition of \(G_2(i,j)\).
From (2.3.55) and (2.3.56) we have

\[ p_{ij}^{(q)} = \frac{n(n-1)p_ip_j}{S_q^2} \left[ 1 + \left( \frac{p_i+p_j}{S_q} - \frac{\Sigma'p_t^2}{S_q^2} \right) \right] \]

\[ + \left\{ \frac{2(p_i^2+p_j^2)}{S_q^2} - \frac{2\Sigma'p_t^3}{S_q^3} + \frac{a_n p_ip_j}{S_q^2} \right\} \]

\[ - (a_n+1) \cdot \frac{(p_i+p_j)\Sigma'p_t^2}{S_q^3} + (a_n+1) \cdot \left( \frac{\Sigma'p_t^2}{S_q^2} \right)^2 \] (4.3.3)

correct to \( O(N^{-4}) \), where \( \Sigma' \) denotes the summation over all the units belonging to the \( q \)th group, and \( a_n \) is a constant that may depend on \( n \).

\[ \frac{1}{A} \cdot \epsilon_{ij}^{(q)} = \frac{A_1}{A} \cdot E[2S_q \cdot p_{ij}^{(q)}] = \frac{A_1}{A} \cdot E[2S_q \cdot p_{ij}] \] (4.3.4)

where \( E \) denotes the operation of taking the expectation with respect to the scheme of selecting \( \left( \frac{N}{3} - 2 \right) \) units from among the \( (N-2) \) population units excluding \( U_i \) and \( U_j \).

Now, using (4.3.3) we get by retaining only the terms that contribute to \( O(N^{-4}) \),

\[ E[2S_q \cdot p_{ij}^{(q)}] = 2n(n-1)p_ip_j \cdot E\left[ \frac{1}{S_q} \right] + \frac{p_i+p_j}{S_q} \]

\[ + \left( \frac{p_i^2+p_j^2+a_n p_ip_j}{S_q^3} - \frac{\Sigma'p_t^2}{S_q^3} \right) \]
\[
\begin{align*}
- (a_n+1)(p_i+p_j) & \sum_{p_t}^2 \frac{s_q^4}{s_q^4} - \frac{2\sum_{p_t}^3}{s_q^4} \\
+ (a_n+1) \frac{\left(\sum_{p_t}^2\right)^2}{s_q^5},
\end{align*}
\]

where \( \sum \) denotes the summation over all the units belonging to the \( q \)th group excluding \( U_i \) and \( U_j \). Using the results of Lemma 4.1 with \( K=3 \), we get from (4.3.5) after some simplification,

\[
E[2S_{q \cdot p_i^j}^{(q)}] = 6n(n-1)p_ip_j \cdot [1+ \{(p_i+p_j)-\Sigma p_t^2 - \frac{2}{N}\} \\
+ \left\{ \frac{6}{N^2} + \frac{2(p_i+p_j)}{N} - \frac{2\Sigma p_t^2}{N} + 2(p_i^2+p_j^2) \\
+ (9a_n-16)p_ip_j+(9a_n-15)(\Sigma p_t^2)^2 \\
- 2\Sigma p_t^3-(9a_n-15)(p_i+p_j)\Sigma p_t^2 \} \right\] (4.3.6)
\]

Using the fact that \( \frac{A_1}{A} = \frac{1}{3}[1- \frac{2}{N} - \frac{2}{N^2}] \) correct to \( O(N^{-2}) \), and (4.3.6) we get from (4.3.4),

\[
\frac{1}{A} \sum_{q \cdot p_i^j}^{A} = 2n(n-1)p_ip_j \cdot [1+ \{(p_i+p_j)-\Sigma p_t^2\} \\
+ \{(p_i^2+p_j^2)- 2\Sigma p_t^3 \} \\
+ (9a_n-15)p_ip_j+(9a_n-15)(\Sigma p_t^2)^2 \} + (9a_n-15)(\Sigma p_t^2)^2 \} ] \] (4.3.7)
correct to $O(N^{-4})$.

Now, the second component in (4.3.2) is

$$\frac{1}{A} \cdot \sum_{G_2(i,j)} p(r,s) \cdot \frac{np_i}{S_r} \cdot \frac{np_j}{S_s} = n^2 \frac{1}{A} \cdot \sum_{G_2(i,j)} p(r,s) \frac{P_i P_j}{S_r S_s}$$

(4.3.8)

The factor $\frac{1}{A} \cdot \sum_{G_2(i,j)} p(r,s) \cdot \frac{P_i}{S_r} \cdot \frac{P_j}{S_s}$ is exactly the same as the right hand side expression of the Equation (4.1.3) whose value correct to $O(N^{-4})$ is given by (4.1.10).

Thus we have by substituting the value from (4.1.10), in (4.3.8),

$$\frac{1}{A} \cdot \sum_{G_2(i,j)} p(r,s) \cdot \frac{np_i}{S_r} \cdot \frac{np_j}{S_s} = 2n^2 P_i P_j \left[1 + \{(P_i + P_j - \Sigma P_t^2\} + \{2(P_i^2 + P_j^2) - 2\Sigma P_t^3 - 7P_i P_j + 6(P_i + P_j) \Sigma P_t^2 \right. \left. - 6(\Sigma P_t^2)^2\} \right]$$

(4.3.9)

correct to $O(N^{-4})$.

Substituting the values from (4.3.7) and (4.3.9) into (4.3.2), we get after some simplification,

$$P_{ij} = 2n(2n-1) P_i P_j \left[1 + \{(P_i + P_j - \Sigma P_t^2\} + \{2(P_i^2 + P_j^2) \right. \left. - 2\Sigma P_t^3 + b_n \cdot P_i P_j - (b_n + 1)(P_i + P_j) \Sigma P_t^2 \right. \left. + (b_n + 1)(\Sigma P_t^2)^2\} \right]$$

(4.3.10)
correct to $O(N^{-4})$, where

$$b_n = \frac{(n-1)(9a_n - 16) - 7n}{(2n-1)} \quad (4.3.11)$$

Equations (4.3.1) and (4.3.11) show that this scheme again satisfies the conditions of Theorem 2.8. Hence it follows from Theorem 2.8 that instead of using any given I.P.P.S. scheme for sample size $2n$, we will get a better estimate by adopting the procedure described in scheme 4.3, if the condition

$$a_{2n} - b_n > 0 \quad (4.3.12)$$

is satisfied.

Illustrations:

(i) Goodman and Kish procedure:

For the Goodman and Kish procedure we have

$$a_n = 2, \text{ for all } n.$$  

Substituting this value in (4.3.11) we get

$$b_n = \frac{2(n-1) - 7n}{(2n-1)} = -\frac{(5n+2)}{(2n-1)} < 2 = a_{2n}$$

(ii) Sampford's procedure:

For the Sampford's procedure

$$a_n = -(n-2)$$

substituting this in (4.3.11), we get
\[ b_n = -\frac{\{9(n-1)(n-2)+16(n-1)+7n\}}{(2n-1)} \]

from which we get

\[ a_{2n} - b_n = \frac{n(5n+2)}{(2n-1)} > 0 \]

(iii) Rao, Hartley, Cochran scheme with revised probabilities:

For the Rao, Hartley, Cochran scheme with revised probabilities

\[ a_n = -(2n-3) \]

Substituting this in (4.3.11), we get

\[ b_n = -\frac{(n-1)(9(2n-3)+16)+7n}{(2n-1)} \]

from which it follows that,

\[ a_{2n} - b_n = \frac{2}{(2n-1)} \cdot [(n-1)(5n-1)+3] \]

\[ > 0 \text{ for } n \geq 2 \]

The unequal probability schemes that are easily applicable for general sample sizes are rather scarce in the literature owing to the complications involved. Thus the above mentioned procedure would be advantageous to adopt for getting a sample of four units by applying it to any given simple procedure presented for sample size 2.

From (4.3.11) we have
Thus for all those schemes useful for sample size 2, we can adopt the above mentioned procedure advantageously if the condition (4.3.13) is satisfied. For example, for the procedure of Yates and Grundy (1953) and the procedure of Durbin (1953), the condition (4.3.13) is satisfied.

4.4. Randomized Three Stage Procedure with Durbin's Scheme

When the Durbin's scheme (1967) is adopted in step (ii) of the procedure described in section 4.3 for getting a sample of size 4, it follows from (4.3.1) that

\[ P_i = 4p_i \] (4.4.1)

Also putting \( n=2 \) in (4.3.11) we get, \( b_2 = -10 \), since \( a_2 = 0 \) for the Durbin's scheme.

Thus the \( P_{ij} \) of the randomized Durbin's scheme for sample size 4 is got by substituting the value \( b_2 = -10 \) and \( n = 2 \) in (4.3.10). Thus we have,

\[
P_{ij} = 12p_i p_j [1 + (p_i + p_j) - \Sigma p_t^2] + 2(p_i^2 + p_j^2) - 2\Sigma p_t^3 - 10p_i p_j + 9(p_i + p_j) \Sigma p_t^2 - 9(\Sigma p_t^2)^2)
\] (4.4.2)

correct to \( O(N^{-2}) \).

In section 4.3 we have illustrated that this procedure
gives a more efficient estimator for sample size 4 which could be considered as a generalization of the Durbin's scheme. As the procedure described in section 4.3 seems to be advantageous to apply in practice it can easily be seen through a conditional argument that the same procedure can be adopted in successive stages, for any sample size of the form \( n = 2^m \), where \( m \) is any positive integer, which provides more efficient estimators. In this section we will describe the scheme utilizing the Durbin's procedure by considering a randomized three stage design. This scheme is for getting a sample of size 8 which is an extension of the procedure described in Section 4.3.

The procedure is described as follows:

(i) Split the population of \( N \) units at random into 3 equal groups and select 2 groups from among the three groups such that the inclusion probability of any group is proportional to the sum of the \( p_t \)'s of the units belonging to that group.

(ii) Perform the following procedure independently within each of the two groups that are selected in step (i).

(ii)(a) Split the \( \frac{N}{3} \) units at random into three equal groups and select 2 groups from among the three groups such that the inclusion probability of any group is proportional to the sum of the \( p_t \)'s of the units belonging to that group.
(ii)(b) Select two units by using Durbin's procedure independently within each of the two groups selected in step (ii)(a).

Thus we get a sample of 8 units by adopting the above scheme. Since the selection in this procedure is made in three stages we will call this scheme as randomized three stage procedure with Durbin's scheme.

Analogous to this, the scheme described in section 4.3 for getting a sample of size 4 by using the Durbin's scheme may be called as the randomized two stage procedure with Durbin's scheme. The procedure of randomized three stage Durbin's scheme can alternatively be adopted as follows:

Stratify the population at random into 9 equal groups. Without loss of generality let the first three groups constitute the first primary stage unit (p.s.u.), the second three groups constitute the second primary stage unit and the last three groups constitute the third primary stage unit. The individual groups may be viewed as the secondary stage units (s.s.u.) and the units within the secondary stage units may be viewed as the third stage units (t.s.u.). Except at the ultimate stage, selection at each of the first two stages is very simple to adopt in practice because of the fact that for any scheme of selecting two units from among the three units such that the probability of including the
unit $i$ is $p_i$, the pairwise inclusion probability $p_{ij}$ which is the same as the probability of selecting the sample $(i,j)$, is given by the formula

$$p_{ij} = p_i + p_j - 1 \quad (4.4.3)$$

In the ultimate stage however, we adopt the Durbin's scheme of selecting two units within each of the selected penultimate stage units, which is again simple to operate.

We denote the total number of distinct arrangements that can be made of the population of $N$ units into three equal groups by $R_N(2)$, the total number of distinct arrangements that can be made of the population of $N$ units into three equal groups such that a given pair of units $(u_i, u_j)$ belong to two different groups by $R_N(2,1)$ and the total number of distinct arrangements that can be made of the population of $N$ units into three equal groups such that a given pair of units $(u_i, u_j)$ belong to the same group by $R_N(2,2)$.

It follows from Theorems 3.2, 3.3 and 3.4 that

$$R_N(2) = \frac{N!}{6(\frac{N}{3})^3} \quad (4.4.4)$$

$$R_N(2,1) = \frac{(N-2)!}{\{(\frac{N}{3} - 1)\}^2 \cdot (\frac{N}{3})!} \quad (4.4.5)$$

and
\[
R_N(2,2) = \frac{(N-2)!}{2\left(\binom{N}{3}-2\right)!\left(\binom{N}{3}\right)^2}
\]  

(4.4.6)

As a check the equation,

\[
R_N(2,1) + R_N(2,2) = R_N(2)
\]

(4.4.7)

can be easily verified.

Further we have from (4.4.4)-(4.4.6)

\[
\frac{R_N(2,1)}{R_N(2)} = \frac{2N}{3(N-1)}
\]

(4.4.8)

and

\[
\frac{R_N(2,2)}{R_N(2)} = \frac{(N-3)}{3(N-1)}
\]

(4.4.9)

Now, considering the randomized three stage Durbin's scheme, let \(R_N(3)\) denote the total number of arrangements such that within each stage the arrangements are distinct.

Then by an extension of (4.4.4) we have

\[
R_N(3) = R_N(2).\left\{\frac{R_N(2)}{3}\right\}^3.
\]

(4.4.10)

With respect to any pair of units \((U_i,U_j)\) of the population the \(R_N(3)\) arrangements can be divided into three categories, viz., (i) arrangements in which the \(i\)th and \(j\)th units come in different primary stage units, (ii) arrangements in which \(i\)th and \(j\)th units come in the same primary stage unit but in different second stage units, and (iii) arrangements in which the \(i\)th and \(j\)th units come in the
same primary stage unit and in the same secondary stage unit. We denote the number of arrangements in the categories (i), (ii) and (iii) above by \( R_N(3,1) \), \( R_N(3,2) \) and \( R_N(3,3) \) respectively. A direct extension of formulae (4.4.4)-(4.4.6) yields the relations

\[
R_N(3,1) = R_N(2,1) \cdot \left\{ R_N(2) \right\}^3 \tag{4.4.11}
\]

\[
R_N(3,2) = R_N(2,2) \cdot R_N(2,1) \cdot \left\{ R_N(2) \right\}^2 \tag{4.4.12}
\]

and

\[
R_N(3,3) = R_N(2,2) \cdot R_N(2,2) \cdot \left\{ R_N(2) \right\}^2 \tag{4.4.13}
\]

As a check the relation

\[
R_N(3,1) + R_N(3,2) + R_N(3,3) = R_N(3) \tag{4.4.14}
\]

can be easily verified with the help of (4.4.7). Now for the randomized three stage Durbin's procedure the inclusion probability \( P_i \) is given by

\[
P_i = \frac{1}{R_N(3)} \sum_{i=1}^{R_N(3)} \left[ \frac{2P_i}{S_{g_1g_2}} + \frac{2S_{g_1g_2}}{S_{g_1}} + 2S_{g_1} \right] \tag{4.4.15}
\]

where the summation runs over all the arrangements belonging to \( R_N(3) \), the collection of all distinct arrangements, \( S_{g_1g_2} \) is the sum of the \( p_t \)'s of the units belonging to the \( g_2 \)th second stage unit of the \( g_1 \)th first stage unit and \( S_{g_1} \) is the sum of the \( p_t \)'s of the units belonging to the \( g_1 \)th first stage unit. Here the \( i \)th unit is assumed to belong to the \( g_2 \)th second stage unit of the \( g_1 \)th first stage unit for a given
arrangement. From (4.4.15) we have

$$P_i = 8p_i$$  \hspace{1cm} (4.4.16)

showing that the scheme is an 'Inclusion probability proportional to size scheme'.

The pairwise inclusion probability $P_{ij}$ is given by

$$P_{ij} = \frac{1}{R_N(3)} \cdot \left[ \sum_{3,1} \frac{C_1}{R_N(3,1)} + \sum_{3,2} \frac{C_2}{R_N(3,2)} + \sum_{3,3} \frac{C_3}{R_N(3,3)} \right],$$  \hspace{1cm} (4.4.17)

where $C_1$, $C_2$, and $C_3$ are the conditional probabilities of selecting the pair $(U_i, U_j)$ given the arrangement corresponding to categories (i), (ii) and (iii) respectively.

In a given arrangement corresponding to category (i) let the $i$th unit belong to the $r_2$th second stage unit of the $r_1$th first stage unit and the $j$th unit belong to the $s_2$th second stage unit of the $s_1$th first stage unit.

Then we have,

$$C_1 = \frac{2p_i}{r_1 r_2} \cdot \frac{2p_j}{s_1 s_2} \cdot \frac{2S_{r_1 r_2}}{S_{r_1}} \cdot \frac{2S_{s_1 s_2}}{S_{s_1}} \cdot \left[ \frac{R_N(3,1) + \frac{1}{2S_{r_1 s_1}} - \frac{1}{2S_{r_1 s_1}}}{2S_{r_1 s_1}} \right],$$

$$= 32p_i p_j \left[ \frac{1}{r_1} + \frac{1}{s_1} - \frac{1}{2S_{r_1 s_1}} \right],$$  \hspace{1cm} (4.4.18)

from which we get

$$\frac{1}{R_N(3)} \cdot \frac{\sum_{3,1} R_N(3,1)}{R_N(3,1)} \cdot \sum_{3,1} \left[ \frac{1}{r_1} + \frac{1}{s_1} - \frac{1}{2S_{r_1 s_1}} \right].$$
where \( E \) denotes the expectation over the scheme of selecting two without replacement groups of size \( \frac{N}{3} - 1 \) units each from the population excluding the \( i \)th and \( j \)th units and attaching the \( i \)th unit to one group and the \( j \)th unit to the other. Observing that \( E\left[ \frac{1}{S} + \frac{1}{S} - \frac{1}{2Sr_1S_1} \right] \) correct to \( O(N^{-2}) \) is the same as the expression considered in Equation (4.1.9) we get, from (4.1.9)

\[
E\left[ \frac{1}{S} + \frac{1}{S} - \frac{1}{2Sr_1S_1} \right] = \frac{3}{2}\left[ 1 + \left\{ (p_i + p_j) - \Sigma p_t^2 - \frac{1}{N} \right\} + 2(p_i^2 + p_j^2 - 7p_i p_j - \frac{(p_i + p_j)}{N}) + 6(p_i + p_j) \Sigma p_t^2 + \frac{\Sigma p_t^2}{N} - 2\Sigma p_t^3 - 6(\Sigma p_t^2)^2 \right],
\]

(4.4.20) correct to \( O(N^{-2}) \).

Further, we have from (4.4.10) and (4.4.11),

\[
\frac{R_N(3,1)}{R_N(3)} = \frac{R_N(2,1)}{R_N(2)} = \frac{2N}{3(N-1)}
\]
167

\[ = \frac{2}{3} \left[ 1 + \frac{1}{N} + \frac{1}{N^2} \right], \quad (4.4.21) \]

correct to \( O(N^{-2}) \), which follows from (4.4.8). Substituting from (4.4.20) and (4.4.21) in (4.4.19) we get after simplifying,

\[
\frac{1}{R_n(3)} \sum_{R_n(3,1)} C_1 = 32p_ip_j \left[ 1 + \left\{(p_i + p_j) - \Sigma p_t^2\right\} + 2(p_i^2 + p_j^2) \right.

- 2\Sigma p_t^3 - 7p_ip_j + 6(p_i + p_j)\Sigma p_t^2

- 6(\Sigma p_t^2)^2 \right], \quad (4.4.22)
\]
correct to \( O(N^{-4}) \).

In a given arrangement corresponding to category (ii), let the \( i \)th unit belong to the \( r_2 \)th second stage unit of the \( q_1 \)th first stage unit and the \( j \)th unit belong to the \( s_2 \)th second stage unit of the \( q_1 \)th first stage unit. Thus we have

\[
C_2 = \frac{2p_i}{S_{q_1 r_2}} \cdot \frac{2p_j}{S_{q_1 s_2}} \cdot \left[ \frac{2S_{q_1 r_2}}{S_{q_1}} + \frac{2S_{q_1 s_2}}{S_{q_1}} - 1 \right] . 2S_{q_1}

= 16p_ip_j \left[ \frac{1}{S_{q_1 r_2}} + \frac{1}{S_{q_1 s_2}} - \frac{S_{q_1}}{2S_{q_1 r_2} S_{q_1 s_2}} \right] \quad (4.4.23)
\]

from which we get,
where \( E \) denotes the expectation taken over the mechanism of randomly splitting the population units such that an arrangement belonging to the category (ii) would emerge.

In the following lemma we present a result which we will be using in the next section also.

**Lemma 4.2:**

Consider the sampling mechanism described below:

(i) Select a simple random sample of size \( \frac{N}{3K} - 2 \) units, where \( N \) is assumed to be a multiple of 3K, from the population excluding the \( i \)th and \( j \)th units. Let \( S_{q_{1}} \) denote the sum of the \( p_{t} \)'s of these \( \frac{N}{3K} - 2 \) units and also \( p_{i} \) and \( p_{j} \).

(ii) Select a simple random sample of size \( \frac{N}{3K} - 1 \) from the \( \frac{N}{3K} - 2 \) units that are selected in step (i). Let \( S_{q_{1}r_{2}} \) denote the sum of the \( p_{t} \)'s of these \( \frac{N}{3K} - 1 \) units and also \( p_{i} \).

(iii) Select a simple random sample of size \( \frac{N}{3K} - 1 \) units from the remaining \( \frac{2N}{3K} - 1 \) units. Let \( S_{q_{1}s_{2}} \) denote the sum of the \( p_{t} \)'s of these \( \frac{N}{3K} - 1 \) units and also \( p_{j} \).
Assume further that \( p_t \) is of \( O(N^{-1}) \), \( K \) is small relative to \( N \), and \( N \) is moderately large. Under these assumptions, for the sampling scheme described above we have,

\[
E\left[ \frac{1}{S q_1 r_2} + \frac{1}{S q_1 s_2} - \frac{S q_1}{2S q_1 r_2 S q_1 s_2} \right] = \frac{3K}{2} \left[ 1 + ((p_i + p_j) - \Sigma p_t^2) \right] - \frac{1}{N} \\
+ (2(p_i^2 + p_j^2) - (9K^2 - 2)p_ip_j \\
- \frac{(p_i + p_j)}{N} + (9K^2 - 3)(p_i + p_j) \Sigma p_t^2 \\
+ \frac{\Sigma p_t^2}{N} - 2\Sigma p_t^3 - (9K^2 - 3)(\Sigma p_t^2)^2], \tag{4.4.25}
\]
correct to \( O(N^{-2}) \).

Considering the Equation (4.4.24) we have from (4.4.10) and (4.4.12),

\[
\frac{R_N(3,2)}{R_N(3)} = \frac{R_N(2,2)}{R_N(2)} \cdot \frac{R_{N/3}(2,1)}{R_{N/3}(2)} \tag{4.4.26}
\]
substituting the values from (4.4.8) and (4.4.9) we get,

\[
\frac{R_N(3,2)}{R_N(3)} = \frac{2N}{9(N-1)} = \frac{2}{9}[1 + 1/N + 1/N^2] \tag{4.4.27}
\]
correct to \( O(N^{-2}) \).

It can easily be seen that \( E\left[ \frac{1}{S q_1 r_2} + \frac{1}{S q_1 s_2} - \frac{S q_1}{2S q_1 r_2 S q_1 s_2} \right] \)

of (4.4.24) can be obtained from Lemma 4.2 when \( K=3 \). Thus, we get
\[
E\left[\frac{1}{S_{q_1 r_2}} + \frac{1}{S_{q_1 s_2}} - \frac{S_{q_1}}{2S_{q_1 r_2} S_{q_1 s_2}}\right] = \frac{9}{2}\left[1 + \{(p_i + p_j)^2 - \Sigma p_t^2\} - \frac{1}{N}\right]
\]

\[
+ \left\{2(p_i^2 + p_j^2) - 79p_i p_j - \frac{(p_i + p_j)}{N}\right\} + 78(p_i + p_j) \Sigma p_t^2 + \frac{\Sigma p_t^2}{N}
\]

\[- 2\Sigma p_t^3 - 78(\Sigma p_t^2)^2\}\] ,

(4.4.28)
correct to \(O(N^{-2})\).

Now, substituting from (4.4.27) and (4.4.28) in (4.4.24) we get after simplifying and retaining terms to \(O(N^{-4})\),

\[
\frac{1}{R_N(3)} \cdot \Sigma C_2 = 16p_i p_j \left\{1 + \{(p_i + p_j)^2 - \Sigma p_t^2\} + \{2(p_i^2 + p_j^2)ight.
\]

\[- 2\Sigma p_t^3 - 79p_i p_j + 78(p_i + p_j) \Sigma p_t^2 - 78(\Sigma p_t^2)^2\}\]

(4.4.29)

In a given arrangement corresponding to the category (iii) where in the \(i\)th and \(j\)th units come in the same first stage unit and in the same second stage unit, let the pair \((U_i, U_j)\) belong to the \(q_2\)th second stage unit of the \(q_1\)th first stage unit. Thus we have the conditional probability \(C_3\) of selecting the pair of units \(U_i\) and \(U_j\) in the sample is given by,
\[
C_3 = \left[ \frac{\frac{2p_i}{S_{q_1q_2}} \cdot \frac{p_j}{S_{q_1q_2}} \left\{ \frac{1}{1-2p_i/S_{q_1q_2}} + \frac{1}{1-2p_j/S_{q_1q_2}} \right\}}{1 + \sum' \frac{p_t/S_{q_1q_2}}{1-2p_t/S_{q_1q_2}}} \right] \times \frac{2S}{S_{q_1q_2}} \times 2S_{q_1} \quad (4.4.30)
\]

where \( \sum' \) denotes the summation over all the units belonging to the \( q_2 \)th second stage unit of the \( q_1 \)th first stage unit.

From (2.3.52) we get after replacing \( p_t \) by \( p_t/S_{q_1q_2} \),

\[
C_3 = \frac{8p_i p_j}{S_{q_1q_2}} \cdot \left[ 1 + \left\{ \frac{(p_i+p_j)}{S_{q_1q_2}} - \frac{\sum' p_t^2}{S_{q_1q_2}} \right\} + \left( \frac{2(p_i^2+p_j^2)}{S_{q_1q_2}} \right) \right] \quad (4.4.31)
\]
correct to \( O(N^{-4}) \).

\[
\frac{1}{R_N(3)} \Sigma_{\Phi_N(3,3)} C_3 = \frac{R_N(3,3)}{R_N(3)} \cdot \text{E[C}_3] \quad (4.4.32)
\]

where \( \text{E} \) denotes the expectation taken over the mechanism of randomly splitting the population units such that an arrangement belonging to the category (iii) would emerge.
From (4.4.10) and (4.4.13) we get,

\[ \frac{R_N(3,3)}{R_N(3)} = \frac{R_N(2,2)}{R_N(2)} \cdot \frac{R_{N/3}(2,2)}{R_{N/3}(2)}, \]

which upon using (4.4.9) yields,

\[ \frac{R_N(3,3)}{R_N(3)} = \frac{(N-9)}{9(N-1)} = \frac{1}{9} [1 - \frac{8}{N} - \frac{8}{N^2}], \tag{4.4.33} \]

correct to \( O(N^{-2}) \).

From (4.4.31) we get after retaining terms that contribute to \( O(N^{-4}) \),

\[
E[C_3] = 8p_ip_j \cdot E\left[ \frac{1}{s_1q_2^2} + \frac{p_i^2 + p_j^2}{s_2q_1q_2} + \frac{p_i^2 + p_j^2}{s_3q_1q_2} - \frac{\Sigma''p_t^2}{s_3q_1q_2} \right]^{(4.4.34)}
\]

where \( \Sigma'' \) denotes the summation over all the units excepting the \( i \)th and \( j \)th units, that belong to the \( q_2 \)th second stage unit of the \( q_1 \)th first stage unit.

Using the results of Lemma 4.1 with \( K=9 \), we get from (4.4.34), after some simplification,

\[
E[C_3] = 72p_ip_j [1 + \{(p_i^2 + p_j)^{-2}\Sigma p_t^2} + \frac{8}{N}] + \{2(p_i^2 + p_j^2) + \frac{8(p_i^2 + p_j)}{N} - 160p_ip_j + 159(p_i^2 + p_j)\Sigma p_t^2 \]
- \frac{2\pi}{3} \left( \frac{\pi^2}{N} - \frac{\pi^2}{N^2} \right) - 159 \left( \frac{\pi^2}{N} \right)^2 \right)_{(4.4.35)}

\text{correct to } O(N^{-2}).

Substituting from (4.4.33) and (4.4.35) in (4.4.32) we get after simplifying

\frac{1}{R(3)} \sum_{N(3,3)} C_3 = 8\pi p_j [1+\{(p_i+p_j)-\pi p_t^2\}+(2(p_i^2+p_j^2) - 159(p_i+p_j)^2)\pi p_t^2]

\text{correct to } O(N^{-4}).

Using (4.4.22), (4.4.29) and (4.4.36), we get from (4.4.17),

\frac{\pi}{(pi^2+p_j^2)} = 56\pi p_j [1+\{(p_i+p_j)-\pi p_t^2\}+(2(p_i^2+p_j^2)-2\pi p_t^3]

\text{correct to } O(N^{-4}).

Thus for the randomized three stage Durbin's procedure the expression for \(p_i\) is given by (4.4.16) and the expression for \(p_{ij}\) correct to \(O(N^{-4})\) is given by (4.4.37).

Hence for this scheme when the H.T. estimator is considered to estimate the population total the variance ex-
pression can be obtained by applying Theorem 2.8 because the assumptions of the theorem are satisfied with the value for \( a_n \) being \(-\frac{346}{7}\).

Hence the variance of the H.T. estimator correct to \( O(N^0) \) is given by

\[
V(\hat{\phi}_{\text{H.T.}})_{\text{RD}} = \frac{1}{8}[\Sigma p_i z_i^2 - 7\Sigma p_i^2 z_i^2] - \frac{7}{8}[2\Sigma p_i^3 z_i^2 - \Sigma p_i 2\Sigma p_i^2 z_i^2 + \frac{346}{7}(\Sigma p_i^2 z_i^2)] \tag{4.4.38}
\]

where \( z_i = \frac{y_i}{p_i} - y \).

To the same approximation, variance of the H.T. estimator for the Sampford's procedure is given by

\[
V(\hat{\phi}_{\text{H.T.}})_{\text{Samp}} = \frac{1}{8}[\Sigma p_i z_i^2 - 7\Sigma p_i^2 z_i^2] - \frac{7}{8}[2\Sigma p_i^3 z_i^2 - \Sigma p_i 2\Sigma p_i^2 z_i^2 + 6(\Sigma p_i^2 z_i^2)] \tag{4.4.39}
\]

Thus we have

\[
V(\hat{\phi}_{\text{H.T.}})_{\text{Samp}} - V(\hat{\phi}_{\text{H.T.}})_{\text{RD}} = \frac{151}{4}(\Sigma p_i^2 z_i^2)^2 \geq 0
\]

which shows that the H.T. estimator under the randomized three stage procedure with the Durbin's scheme is uniformly more efficient than the H.T. estimator under the Sampford's procedure for selecting samples of size 8.
4.5. Randomized m-stage Procedure with Durbin's Scheme

This is a procedure for selecting samples of size $2^m$, where $m$ is any positive integer, and is a generalization of the procedure described in the previous section for selecting a sample of size 8.

The procedure is as follows:

(i) Split the population of $N$ units at random into 3 equal groups and select 2 groups from among the three groups such that the inclusion probability of any group is proportional to the sum of the $p_t$'s of the units belonging to that group.

Within each of the above selected groups, which could be denoted as primary stage units, perform the following procedure independently.

(ii) Split the units belonging to this group at random into 3 equal groups and select two groups from among the three such that the inclusion probability of any group is proportional to the sum of the $p_t$'s of the units belonging to that group.

Repeat the procedure described in step (ii) independently within each of the selected units at each stage until we select $2^{m-1}$ units of the $(m-1)$th stage.

(iii) Within each of the $(m-1)$th stage units that are selected in step (ii), apply the Durbin's procedure inde-
pendently for selecting a sample of size 2.

The above procedure would yield a sample of size $2^m$. We will call this procedure as the "Randomized m-Stage Procedure with Durbin's Scheme". In what follows, we will assume for mathematical convenience that $N$ is a multiple of $3^{m-1}$.

The notations we use in this section are similar to those adopted in the previous section.

$R_N(m)$ denotes the collection of all arrangements such that within each stage the arrangements are distinct, and $R_N(m)$ is the cardinality of the set $R_N(m)$.

By the inductive argument we get

$$R_N(m) = R_N(2) \cdot \{R_{N/3}(m-1)\}^3$$  \hspace{1cm} (4.5.1)

where $R_N(2)$ from (4.4.4) is given by

$$R_N(2) = \frac{N!}{6 \cdot (\frac{N}{3})!^3}$$  \hspace{1cm} (4.5.2)

By the recursive relationship, we get from (4.5.1),

$$R_N(m) = R_N(2) \cdot [R_{N/3}(2)]^3 \cdot [R_{N/3^2}(2)]^2 \cdot [R_{N/3^m}(2)]$$  \hspace{1cm} (4.5.3)

With respect to any particular pair $(U_i, U_j)$ of the population units, the collection, $R_N(m)$, of all arrangements is the union of mutually disjoint sets $R_N(m, t) \ (t=1, 2, \ldots, m)$ where $R_N(m, 1)$ denotes the collection of all arrangements where in the
pair \((U_i, U_j)\) belong to different primary stage units, 
\(\mathcal{R}_N(m, 2)\) denotes the collection of all arrangements where in the pair \((U_i, U_j)\) belong to the same primary stage unit but different second stage units, \(\ldots \mathcal{R}_N(m, t), 1 \leq t \leq m-1,\) denotes the collection of all arrangements wherein the pair \((U_i, U_j)\) belong to the same primary stage unit, same secondary stage unit...same \((t-1)\)th stage unit, but different \(t\)th stage units; \(\mathcal{R}_N(m, m)\) denotes the collection of all arrangements wherein the pair \((U_i, U_j)\) belong to the same \((m-1)\)th stage unit.

Let \(R_N(m, t)\) be the cardinality of the set \(\mathcal{R}_N(m, t), 1 \leq t \leq m\). As a direct extension of relations (4.4.11)-(4.4.13), we get,

\[
R_N(m, 1) = R_N(2, 1) \cdot \{R_N/3(m-1)\}^3
\]

(4.5.4)

for \(2 \leq t \leq m-1,\)

\[
R_N(m, t) = \prod_{\ell=2}^{t} \left[ R_{N/3^{\ell-2}}(2, 2) \cdot \{R_{N/3^{\ell-1}}(m-\ell+1)\}^2 \right]
\]

\[\cdot R_{N/3^{t-1}}(m-t+1, 1)\]

(4.5.5)

\[
R_N(m, m) = \prod_{\ell=2}^{m-1} \left[ R_{N/3^{\ell-2}}(2, 2) \cdot \{R_{N/3^{\ell-1}}(m-\ell+1)\}^2 \right] \cdot R_{N/3^{m-2}}(2, 2)
\]

(4.5.6)

Using these formulae it can easily be verified that
For the $R_N(m,t)$, $1 \leq t \leq m$, the following relations hold.

\[ R_N(m,t) = \sum_{t=1}^{m} R_N(m,t) = R_N(m) \]  \hspace{1cm} (4.5.7)

**Theorem 4.1:**

For the $R_N(m,t)$, $1 \leq t \leq m$, the following relations hold.

\[ \frac{R_N(m,t)}{R_N(m,t-1)} = \frac{1}{3}, \text{ for } 2 \leq t \leq m-1 \]  \hspace{1cm} (4.5.8)

and

\[ \frac{R_N(m,m)}{R_N(m,m-1)} = \frac{(N-3m-1)}{2N} \]  \hspace{1cm} (4.5.9)

**Proof:**

From (4.5.4) and (4.5.5) we get

\[ \frac{R_N(m,2)}{R_N(m,1)} = \frac{R_N(2,2)}{R_N(2,1)} \cdot \frac{R_{N/3}(m-1,1)}{R_{N/3}(m-1)} \]  \hspace{1cm} (4.5.10)

From (4.5.1) and (4.5.4) we get,

\[ \frac{R_{N/3}(m-1,1)}{R_{N/3}(m-1)} = \frac{R_{N/3}(2,1)}{R_{N/3}(2)} \]  \hspace{1cm} (4.5.11)

Using (4.4.8), (4.4.9) and (4.5.11), we get from (4.5.10),

\[ \frac{R_N(m,2)}{R_N(m,1)} = \frac{1}{3} \]  \hspace{1cm} (4.5.12)

For $2 \leq t \leq m-1$, we have from (4.5.5),
\[
\frac{R_N(m, t)}{R_N(m, t-1)} = \frac{R_{N/3t-2}(2, 2)}{R_{N/3t-2}(2, 1)} \cdot \frac{R_{N/3t-1}(m-t+1, 1)}{R_{N/3t-1}(m-t+2, 1)}
\]

which with the help of (4.5.4) yields

\[
\frac{R_N(m, t)}{R_N(m, t-1)} = \frac{R_{N/3t-2}(2, 2)}{R_{N/3t-2}(2, 1)} \cdot \frac{R_{N/3t-1}(m-t+1, 1)}{R_{N/3t-1}(m-t+2, 1)} (4.5.13)
\]

Using (4.4.8), (4.4.9) and (4.5.11), Equation (4.5.13) gives

\[
\frac{R_N(m, t)}{R_N(m, t-1)} = \frac{1}{3} (4.5.14)
\]

From (4.5.5) and (4.5.6) we get

\[
\frac{R_N(m, m)}{R_N(m, m-1)} = \frac{R_{N/3m-2}(2, 2)}{R_{N/3m-2}(2, 1)}
\]

which on using (4.4.8) and (4.4.9) gives

\[
\frac{R_N(m, m)}{R_N(m, m-1)} = \frac{(N-3m-1)}{2N} (4.5.15)
\]

Q.E.D.

**Theorem 4.2:**

For the \(R_N(m, t)\) and \(R_N(m)\), \(1 \leq t \leq m\), the following relations hold

\[
\frac{R_N(m, t)}{R_N(m)} = \frac{2N}{3^t(N-1)}, 1 \leq t \leq m-1 (4.5.16)
\]
and
\[
\frac{R_N(m,m)}{R_N(m)} = \frac{(N-3^{m-1})}{3^{m-1}(N-1)}
\] (4.5.17)

**Proof:**

Equations (4.4.8), (4.5.1) and (4.5.4) yield,

\[
\frac{R_N(m,1)}{R_N(m)} = \frac{R_N(2,2)}{R_N(2)} = \frac{2N}{3(N-1)}
\] (4.5.18)

Assume that for \(1 \leq t-1 < m-1\),

\[
\frac{R_N(m,t-1)}{R_N(m)} = \frac{2N}{3^{t-1}(N-1)}
\] (4.5.19)

Then for \(1 \leq t-1 < t < m-1\), we have by using (4.5.14),

\[
\frac{R_N(m,t)}{R_N(m)} = \frac{2N}{3^t(N-1)}
\] (4.5.20)

Hence by induction it follows that for \(1 \leq t < m-1\)

\[
\frac{R_N(m,t)}{R_N(m)} = \frac{2N}{3^t(N-1)}
\] (4.5.21)

From (4.5.15) and (4.5.21) we get

\[
\frac{R_N(m,m)}{R_N(m)} = \frac{R_N(m,m)}{R_N(m,m-1)} \cdot \frac{R_N(m,m-1)}{R_N(m)} = \frac{(N-3^{m-1})}{3^{m-1}(N-1)}
\] (4.5.22)

Q.E.D.
Now for the randomized m-stage procedure with the Drubin's scheme the inclusion probability for the ith population unit is,

\[
P_i = \frac{1}{R_N(m)} \cdot \frac{2p_i}{S_{g_1g_2\ldots g_{m-1}}} \cdot \frac{2S_{g_1g_2\ldots g_{m-2}}}{S_{g_1g_2\ldots g_{m-3}}} \cdot \frac{2S_{g_1g_2\ldots g_{m-2}}}{S_{g_1g_2\ldots g_{m-3}}} \cdot \frac{2S_{g_1g_2}}{S_{g_1}} \cdot \frac{2S_{g_1}}{g_1} \]

(4.5.23)

where \( S_{g_1g_2\ldots g_{g}} \) denotes the sum of the \( p_i \)'s of the units belonging to the \( g_1 \)th stage unit of the \( g_2 \)th stage unit of the \( \ldots \) \( g_{g-1} \)th stage unit of the \( g \)th primary stage unit.

(4.5.23) reduces to

\[
P_i = \frac{1}{R_N(m)} \cdot \frac{2^m \cdot p_i}{S_{g_1}}
\]

\[
= 2^m \cdot p_i
\]

\[
= np_i
\]

(4.5.24)

Probability of including the pair of units \((U_i, U_j)\) together in the sample is given by,

\[
P_{ij} = \frac{1}{R_N(m)} \cdot \left[ \sum_{\alpha_N(m, 1)} C_1 + \sum_{\alpha_N(m, 2)} C_2 + \ldots + \sum_{\alpha_N(m, t)} C_t \right] \]

\[
+ \ldots \sum_{\alpha_N(m, m)} C_m \]

(4.5.25)
where \( C_t \) \((1 \leq t \leq m)\) is the conditional probability of selecting the pair \((U_i, U_j)\) given the arrangement belonging to the \(t\)th category.

Evaluation of \( \frac{1}{R_{N(m)}} \sum R_{N(m,1)} C_1 \):

In a given arrangement of the first category let \( U_i \) belong to the \( r_{m-1} \)th \((m-1)\)th stage unit of the \( r_{m-2} \)th \((m-2)\)th stage unit of the \( \ldots \) \( r_2 \)th second stage unit of the \( r_1 \)th primary stage unit and let \( U_j \) belong to the \( s_{m-1} \)th \((m-1)\)th stage unit of the \( s_{m-2} \)th \((m-2)\)th stage unit of the \( \ldots \) \( s_2 \)th second stage unit of the \( s_1 \)th primary stage unit.

The conditional probability \( C_1 \) is given by

\[
C_1 = \sum \left\{ \frac{2p_i}{S_{r_1 r_2 \ldots r_{m-1}}} \cdot \frac{2S_{r_1 r_2 \ldots r_{m-1}}}{S_{r_1 r_2 \ldots r_{m-2}}} \cdot \frac{2S_{r_1 r_2 \ldots r_{m-3}}}{S_{r_1 r_2 \ldots r_{m-2}}} \cdots \right\} \times (2S_{s_1 s_2 \ldots s_{m-1}} + 2S_{s_1 s_2 \ldots s_{m-2}} - 1) \times 2^{m-1} \left[p_i p_j \left( \frac{1}{S_{r_1}} + \frac{1}{S_{s_1}} - \frac{1}{2S_{r_1 s_1}} \right) \right]
\]

(4.5.26)

Thus we have
\[
\frac{1}{R_N(m)} \sum_{R_N(m, l)} \frac{R_N(m, l)}{R_N(m)} \cdot E\left[ \frac{1}{r_1} + \frac{1}{s_1} - \frac{1}{2s_1} \right]
\]

where \( E \) denotes the average taken over all the arrangements belonging to the first category. From (4.5.16) we get,

\[
\frac{R_N(m, l)}{R_N(m)} = \frac{2}{3}[1 + \frac{1}{N} + \frac{1}{N^2}]
\]

(4.5.28)
correct to \( O(N^{-2}) \).

It can be easily seen that \( E\left[ \frac{1}{s_1} + \frac{1}{s_1} - \frac{1}{2s_1} \right] \) is given by the right hand side expression of Equation (4.1.9) and thus we have,

\[
E\left[ \frac{1}{s_1} + \frac{1}{s_1} - \frac{1}{2s_1} \right]
\]

\[
= \frac{3}{2}(1+((p_i+p_j)^2-\Sigma p_t^2) - \frac{1}{N}) + \{2(p_i^2+p_j^2) - 7p_ip_j - \frac{\Sigma p_t^2}{N} \}
\]

\[
- \frac{(p_i+p_j)}{N} + 6(p_i+p_j)\Sigma p_t^2 + \frac{\Sigma p_t^2}{N}
\]

\[
- 2\Sigma p_t^3 - 6(\Sigma p_t^2)^2)
\]

(4.5.29)
correct to \( O(N^{-2}) \).
Hence, using (4.5.28) and (4.5.29) we get from (4.5.27) after simplifying and retaining terms to \( O(N^{-4}) \) only,

\[
\frac{1}{\mathcal{R}_N(m)} \sum_{\mathcal{R}_N(m,1)} C_1 = 2^{2m-1} \prod_{i,j} \left[ 1 + (p_i + p_j - 2\Sigma P_t)^2 \right] 
\]

\[
+ \{2(p_i^2 + p_j^2) - 2\Sigma P_t^3 + 7p_i p_j 
\]

\[
+ 6(p_i + p_j)^2 \Sigma P_t^2 - 6(\Sigma P_t)^2 \} \tag{4.5.30}
\]

Evaluation of \( \frac{1}{\mathcal{R}_N(m)} \sum_{\mathcal{R}_N(m,t)} C_t \) for \( 2 \leq t \leq m-1 \):

In a given typical arrangement of the \( t \)th category let \( U_i \) belong to the \( r_{m-1} \)th \((m-1)\)th stage unit of the \( r_{m-2} \)th \((m-2)\)th stage unit of the \( \ldots \) \( q_{t-1} \)th \((t-1)\)th stage unit of the \( \ldots \) \( q_2 \)th 2nd stage unit of the \( q_1 \)th primary stage unit and let \( U_j \) belong to the \( s_{m-1} \)th \((m-1)\)th stage unit of the \( s_{m-2} \)th \((m-2)\)th stage unit of the \( \ldots \) \( q_{t-1} \)th \((t-1)\)th stage unit of the \( \ldots \) \( q_2 \)th 2nd stage unit of the \( q_1 \)th primary stage unit.

The conditional probability \( C_t \) is given by

\[
C_t = \left\{ \frac{2S}{q_1} \cdot \frac{2S}{q_2} \cdot \frac{2S}{q_2 q_3} \cdots \frac{2S}{q_1 q_2 \cdots q_{t-1}} \right\} 
\]

\[
\times \left\{ \frac{2S}{q_1 q_2 \cdots q_{t-1}} \frac{s_t}{q_1 q_2 \cdots q_{t-1} s_t} \right\} 
\]

\[
+ \frac{2S}{q_1 q_2 \cdots q_{t-1}} \frac{s_t}{q_1 q_2 \cdots q_{t-1} s_t} - 1 \right\}
\]
\[
\times \left\{ \frac{2p_i}{S q_1 q_2 \cdots q_t r_t r_{t+1} \cdots r_{m-1}} \right\} \cdot \frac{2S q_1 q_2 \cdots q_t r_t \cdots r_{m-1}}{S q_1 q_2 \cdots q_t r_t \cdots r_{m-2}}
\]

\[
\times \frac{2S q_1 q_2 \cdots q_t r_t r_{t+1}}{S q_1 q_2 \cdots q_t r_t}
\]

\[
\times \left\{ \frac{2p_j}{S q_1 q_2 \cdots q_t s_t s_{t+1} \cdots s_{m-1}} \right\} \cdot \frac{2S q_1 q_2 \cdots q_t s_t \cdots s_{m-1}}{S q_1 q_2 \cdots q_t s_t \cdots s_{m-2}}
\]

\[
\times \frac{2S q_1 q_2 \cdots q_t s_t s_{t+1}}{S q_1 q_2 \cdots q_t s_t}
\]

\[
= 2^{t-1} s q_1 q_2 \cdots q_t \left\{ \frac{2S q_1 q_2 \cdots q_t r_t}{S q_1 q_2 \cdots q_t} \right\} - 1
\]

\[
= 2^{m-t} p_i \times \frac{s q_1 q_2 \cdots q_t r_t}{S q_1 q_2 \cdots q_t} + 2^{m-t} p_j \times \frac{s q_1 q_2 \cdots q_t s_t}{S q_1 q_2 \cdots q_t}
\]

\[
= 2^{2m-t} p_i p_j \left[ \frac{s q_1 q_2 \cdots q_t r_t}{S q_1 q_2 \cdots q_t} \right] + \frac{s q_1 q_2 \cdots q_t s_t}{S q_1 q_2 \cdots q_t} - \frac{s q_1 q_2 \cdots q_t r_t}{2S q_1 q_2 \cdots q_t r_t q_1 q_2 \cdots q_t s_t}
\]

\[\text{(4.5.31)}\]
Thus we have

\[
\frac{1}{R_N(m)} \sum_{R_N(m,t)} C_t = 2^{2m-t} \prod_{i \neq j} R_N(m) \frac{R_N(m,t)}{R_N(m)}
\]

\[
E \left[ \frac{1}{q_1 q_2 \cdots q_{t-1} r_t} + \frac{1}{s q_1 q_2 \cdots q_{t-1} s_t} \right] = \frac{S q_1 q_2 \cdots q_{t-1} s_t}{2 s q_1 q_2 \cdots q_{t-1} r_t q_1 q_2 \cdots q_{t-1} s_t} \] (4.5.32)

where \( E \) denotes the average taken over all the arrangements belonging to the \( t \)th category.

From (4.5.16) we get,

\[
R_N(m,t) = \frac{2}{3} \left[ 1 + \frac{1}{N} + \frac{1}{N^2} \right], \] (4.5.33)

correct to \( O(N^{-2}) \).

In view of the well known fact that a simple random sample taken from a simple random sample of a population would itself be a simple random sample, it can be easily observed that

\[
E \left[ \frac{1}{q_1 q_2 \cdots q_{t-1} r_t} + \frac{1}{s q_1 q_2 \cdots q_{t-1} s_t} \right]
\]
can be evaluated by using Lemma 4.2 with the value of \( K \) being \( 3^{t-1} \).

Thus using (4.4.25) with \( K=3^{t-1} \) and (4.5.33) we get, from (4.5.32), after simplifying and retaining terms to \( O(N^{-4}) \) only,

\[
\frac{1}{R_N(m)} \sum_{R_N(m,t)} C_t = 2^{2m-t} p_i p_j \{ \left[ 1 + \left( p_i + p_j \right) - \Sigma p_t^2 \right] \\
+ \{ 2(p_i^2 + p_j^2) - 2 \Sigma p_t^3 - (3^t - 2) p_i p_j + (3^t - 3) (p_i + p_j) \Sigma p_t^2 \\
- (3^t - 3) \left( \Sigma p_t^2 \right)^2 \} \}
\] (4.5.34)

Observing (4.5.30) it can be seen that (4.5.34) is valid for the case \( t=1 \) also.

Thus we have for \( 1 \leq t \leq m-1 \),

\[
\frac{1}{R_N(m)} \sum_{R_N(m,t)} C_t = 2^{2m-t} p_i p_j \{ \left[ 1 + \left( p_i + p_j \right) - \Sigma p_t^2 \right] \\
+ \{ 2(p_i^2 + p_j^2) - 2 \Sigma p_t^3 - (3^t - 2) p_i p_j \\
+ (3^t - 3) (p_i + p_j) \Sigma p_t^2 - (3^t - 3) \left( \Sigma p_t^2 \right)^2 \} \}
\] (4.5.35)
correct to $O(N^{-4})$.

Evaluation of $\frac{1}{R_N(m)} \sum \rho_{N,m} C_m$:

In a given arrangement of the $m$th category let the pair of units $U_i$ and $U_j$ belong to the $q_{m-1}$th $(m-1)$th stage unit of the $q_{m-2}$th $(m-2)$th stage unit of the $\ldots q_2$th 2nd stage unit of the $q_1$th primary stage unit.

The conditional probability $C_m$ is given by

$$C_m = \mathcal{P}_{ij} \frac{2S_{q_1q_2\ldots q_{m-1}}}{S_{q_1q_2\ldots q_{m-2}}} \frac{2S_{q_1q_2}}{S_{q_1}} \cdot \frac{2S_{q_1}}{2S_{q_1}} \cdot \frac{2S_{q_1}}{2S_{q_1}}$$

where $\mathcal{P}_{ij}$ is the conditional probability of selecting $U_i$ and $U_j$ together by the Durbin's procedure given the $(m-1)$th stage unit containing $U_i$ and $U_j$.

Equation (4.5.36) reduces to

$$C_m = 2^{m-1} \cdot S_{q_1q_2\ldots q_{m-1}} \cdot \mathcal{P}_{ij}$$

(4.5.37)

Now

$$\frac{1}{R_N(m)} \sum \rho_{N,m} C_m = \frac{R_N(m,m)}{R_N(m,m)} \cdot E[C_m],$$

(4.5.38)

where $E$ denotes the average taken over all the arrangements belonging to $\rho_{N,m}$. From (4.5.17) we get,
The conditional probability $P_{ij}$ under the Durbin's procedure is,

\[
P_{ij} = (q_1q_2 \cdots q_{m-1})
\]

where $\Sigma'$ denotes the summation taken over all the units belonging to the $(m-1)$th stage unit containing the pair $(U_i, U_j)$.

Using Equations (4.5.37) and (4.5.40) we get,

\[
E[C_m] = 2^m p_i p_j E[S^{-1}] + \frac{(p_i + p_j)}{S^2} + \frac{(p_i^2 + p_j^2)}{S^3}
\]

\[
- \frac{\Sigma' p_t^2}{S^2 q_1q_2 \cdots q_{m-1}}
\]

where $\Sigma'$ denotes the summation taken over all the units belonging to the $(m-1)$th stage unit containing the pair $(U_i, U_j)$. 

The conditional probability $P_{ij}$ under the Durbin's procedure is,

\[
P_{ij} = \frac{p_i p_j}{S q_1q_2 \cdots q_{m-1}} \left[ \frac{1}{2p_i} + \frac{1}{2p_j} \right]
\]

\[
1 + \Sigma'
\]

\[
= \frac{p_t}{S q_1q_2 \cdots q_{m-1}} \frac{1}{1 - \frac{2p_t}{S q_1q_2 \cdots q_{m-1}}}
\]

where $\Sigma'$ denotes the summation taken over all the units belonging to the $(m-1)$th stage unit containing the pair $(U_i, U_j)$. 

Using Equations (4.5.37) and (4.5.40) we get,

\[
E[C_m] = 2^m p_i p_j E[S^{-1}] + \frac{(p_i + p_j)}{S^2} + \frac{(p_i^2 + p_j^2)}{S^3}
\]

\[
- \frac{\Sigma' p_t^2}{S^2 q_1q_2 \cdots q_{m-1}}
\]
where $\Sigma^m$ denotes the summation over all the units belonging to the (m-1)th stage unit, containing $U_i$ and $U_j$, excepting $U_i$ and $U_j$.

Since the $(N-2)$ units that constitute, together with the pair of units $U_i$ and $U_j$, the (m-1)th stage unit can be considered as a simple random sample from the population of (N-2) units excluding $U_i$ and $U_j$, we can use Lemma 4.1, with the value of $K$ being $3^{m-1}$ for evaluating the right hand side expression of (4.5.41).

Thus we have,

$$
E[C_m] = K \cdot 2^m p_ip_j \left[ 1 + \left\{ (p_i+p_j) - \Sigma p_t^2 + \frac{(K-1)}{N} \right\} 
+ 2 (p_i^2 + p_j^2) + (K-1) \frac{(p_i+p_j)}{N} 
- 2 (K^2-1)p_ip_j + (2K^2-3) (p_i+p_j) \Sigma p_t^2 
- 2 \Sigma p_t^3 - (2K^2-3) (\Sigma p_t^2)^2 
- (K-1) \cdot \frac{\Sigma p_t^2}{N} + \frac{K(K-1)}{N^2} \right] 
$$

(4.5.42)
correct to $O(N^{-4})$, where $K = 3^{m-1}$.

Substituting from (4.5.39) and (4.5.42) into (4.5.38) we get after simplifying and retaining terms to $O(N^{-4})$ only,

\[
\frac{1}{R_N(m)} \sum_{m,m} C_m = 2^m p_i p_j [1+\{(p_i+p_j)-\Sigma p_t^2\}
\]

\[
+ \{2(p_i^2+p_j^2)-2\Sigma p_t^3
\]

\[
- (2.3^{2m-2}-2)p_ip_j
\]

\[
+ (2.3^{2m-2}-3)(p_i+p_j)\Sigma p_t^2
\]

\[
- (2.3^{2m-2}-3)(\Sigma p_t^2)^2\}
\]  

Substituting from (4.5.35) and (4.5.43) in (4.5.25), we get

\[
P_{ij} = (\sum_{t=1}^{m} 2^{2m-t}) \cdot p_i p_j [1+\{(p_i+p_j)-\Sigma p_t^2\}
\]

\[
+ \{2(p_i^2+p_j^2)-2\Sigma p_t^3+2p_ip_j
\]

\[
- 3(p_i+p_j)\Sigma p_t^2+3(\Sigma p_t^2)^2\}
\]

\[
+ (\sum_{t=1}^{m-1} 2^{2m-t+1}+2^{m+1}.3^{2m-2}) \cdot p_i p_j [(p_i+p_j)\Sigma p_t^2
\]

\[
- p_i p_j-(\Sigma p_t^2)^2\}
\]  

(4.5.44)

correct to $O(N^{-4})$.

Observing that
\[
\sum_{t=1}^{2^{m-t}} = 2^m(2^m - 1),
\]
and
\[
\sum_{t=1}^{2^{m-t}} 3 \cdot 2^t = \frac{9}{7} \cdot 2^{m+1}(9^{m-1} - 2^{m-1}),
\]
we can write (4.5.44) as
\[
P_{ij} = 2^m(2^m - 1)P_iP_j[1 + \{(p_i + p_j) - \sum p_t^2\} + \{(p_i^2 + p_j^2) - 2p_t^2 + B_t^2\} + (B_m + 1)(\sum p_t^2)^2]\]
(4.5.45)
correct to \(O(N^{-4})\), where
\[
B_m = \frac{1}{7(2^m - 1)}[23 \cdot 2^m - 32 \cdot 9^{m-1} - 14]
\]
Thus for the randomized m-stage procedure with the Durbin's scheme, the expression for \(P_i\) is given by (4.5.24) and the expression for \(P_{ij}\) correct to \(O(N^{-4})\) is given by (4.5.46).

Since the conditions of Theorem 2.8 are satisfied, the variance of the H.T. estimator correct to \(O(N^0)\) for this procedure is
\[
V(\hat{\gamma}_{H.T.})_{RD} = \frac{1}{2^m} \cdot [\sum p_i z_i^2 - (2^m - 1) \sum p_i^2 z_i^2]
\]
\[
- \frac{(2^m - 1)}{2^m} \cdot \left[ z_i^2 \sum p_i^3 - \sum p_i^2 z_i^2 - \sum p_i^2 z_i^2 - B_m \cdot (\sum p_i^2 z_i^2)^2 \right]
\]
(4.5.47)
where \( z_i = \frac{Y_i}{p_i} - Y \) and \( B_m \) is given by (4.5.46).

Randomized m-stage procedure with the Durbin's scheme is an alternative to the Sampford's procedure as a generalization of the Durbin's scheme for samples of size \( n>2 \). Since the simplicity of this randomized procedure to adopt in large scale surveys is evident relative to the procedure of Sampford, it will be interesting to study the relative performance of the two methods.

**Theorem 4.3:**

When the variance of the corresponding Horvitz-Thompson estimator is considered correct to \( O(N^0) \), variance corresponding to the randomized m-stage procedure with the Durbin's scheme for sample size \( 2^m \) is uniformly smaller than the variance corresponding to the Sampford's procedure for sample size \( 2^m \) and the difference between the two variances would be larger for larger values of \( m \).

**Proof:**

Variance of the Horvitz-Thompson estimator correct to \( O(N^0) \) for the Sampford's procedure with sample size \( 2^m \) as given by (2.3.63) is
The corresponding expression for the randomized m-stage procedure with the Durbin's scheme is given by (4.5.47).

From (4.5.47) and (4.5.48) we get,

\[ V(\hat{Y}_{H.T.})_{Samp} - V(\hat{Y}_{H.T.})_{RD} = \frac{1}{7} \cdot D_m \cdot (\Sigma_{i} \beta_i z_i)^2 \]  

(4.5.49)

where

\[ D_m = \frac{1}{2^m} \cdot (32 \cdot 9^{m-1} - 7.4^{m-2} + 1) \]  

(4.5.50)

It follows from (4.5.50) that

\[ D_2 = 42 > 0 \]  

(4.5.51)

and

\[ D_{m+1} = \frac{1}{2^{m+1}} (32 \cdot 9^m - 7.4^{m+1} - 2^m + 2) \]

\[ > \frac{9}{2^{m+1}} \cdot (32 \cdot 9^{m-1} - 7.4^{m-2} + 1) \]

\[ = \frac{9}{2} D_m, \text{ for all } m \]  

(4.5.52)

(4.5.51) and (4.5.52) together imply that \( D_m \) is non-negative and monotone increasing.

Hence it follows from (4.5.49) that \( V(\hat{Y}_{H.T.})_{Samp} - V(\hat{Y}_{H.T.})_{RD} \) is nonnegative and is larger for larger values
Remark:

The fact that \( V(\hat{\gamma}_{H.T.})_{\text{Samp}} - V(\hat{\gamma}_{H.T.})_{\text{RD}} \) is a monotone increasing function of \( m \) implies that the relative efficiency of the randomized \( m \)-stage procedure adopted with the Durbin's scheme compared to the Sampford's procedure increases as the sample size increases.

In Chapter 3 we have proposed the Rao, Hartley and Cochran's procedure with revised probabilities which ensures the condition \( p_i = np_i \). Since the Rao, Hartley and Cochran's procedure is also practically convenient to adopt in large scale surveys for any sample size it would be of interest to compare the relative performance of the randomized \( m \)-stage procedure using the Durbin's scheme with respect to the Rao, Hartley and Cochran's procedure with the revised probabilities.

Theorem 4.4:

Variance of the Horvitz-Thompson estimator correct to \( O(N^0) \) for the randomized \( m \)-stage procedure using the Durbin's scheme is uniformly smaller than the corresponding expression in the case of the Rao, Hartley and Cochran's procedure with the revised probabilities and the difference between the two variances would be larger for larger values of \( m \).
Proof:

Variance of the Horvitz-Thompson estimator correct to \( O(N^0) \) for the Rao, Hartley and Cochran's scheme with the revised probabilities for selecting a sample of size \( 2^m \), as given by (3.9.10) is

\[
V(\hat{\theta}_{H.T.})_{RHC-RP} = \frac{1}{2^m} [\sum p_i z_i^2 - (2^m-1) \cdot \sum p_i^2 z_i^2]
\]

\[
- \frac{(2^m-1)}{2^m} \cdot [2 \sum p_i z_i^2 - \sum p_i^2 \cdot \sum p_i z_i^2]
\]

\[
+ (2^{m+1}-3) \cdot (\sum p_i^2 z_i) \}
\]

(4.5.53)

Thus from (4.5.47) and (4.5.53) we get,

\[
V(\hat{\theta}_{H.T.})_{RHC-RP} - V(\hat{\theta}_{H.T.})_{RD} = \frac{1}{7} \cdot J_m \cdot (\sum p_i^2 z_i) \}
\]

(4.5.54)

where

\[
J_m = \frac{1}{2^m} \cdot (32.9^{m-1} - 56.4^{m-1} + 12.2^{m-7})
\]

(4.5.55)

It follows from (4.5.55) that

\[
J_2 = \frac{105}{4} > 0
\]

(4.5.56)

and

\[
J_{m+1} - J_m = \frac{1}{2^m} (112.9^{m-1} - 56.4^{m-1} + \frac{7}{2})
\]

\[
> \frac{1}{2^m} (56.9^{m-1} + \frac{7}{2})
\]

(4.5.57)
(4.5.56) and (4.5.57) together imply that $J_m$ is nonnegative and monotone increasing. Hence it follows from (4.5.54) that $V(\hat{\gamma}_{H.T.})_{RHC-RP} - V(\hat{\gamma}_{H.T.})_{RD}$ is always nonnegative and is larger for larger values of $m$.

Q.E.D.

Remark:

As with the case of Sampford's procedure here also it follows that the relative efficiency of the randomized $m$-stage procedure adopted with the Durbin's scheme compared to the Rao, Hartley, and Cochran's procedure with revised probabilities increases as the sample size increases.

Theorems 4.3 and 4.4 suggest that the gains would be substantial when we adopt the randomized $m$-stage procedure using the Durbin's scheme in large scale surveys.

Instead of the Durbin's scheme one can use any efficient scheme at the $(m-1)$th stage of the randomized $m$-stage procedure where in the gains are expected to be substantial. The formulae for $P_{ij}$ and hence the variance of the corresponding H.T. estimator, could be derived using exactly the same technique. Applicability of these randomized varying probability schemes in large scale surveys is quite evident compared to the complicated procedures that are existent in the literature whose applicability is doubtful in large scale surveys.
5. MISCELLANEOUS TOPICS IN UNEQUAL PROBABILITY SAMPLING

5.1. Model Comparisons of Some Existing Schemes

In order to study the relative performance of different I.P.P.S. schemes as measured by the variance of the corresponding H.T. estimator, it is convenient to assume some knowledge regarding the relationship between the variate y and the auxiliary characteristic x. Since unequal probability sampling is resorted to in the situations where y is approximately proportional to x it is reasonable to assume the model

\[ y_i = \alpha + \beta x_i + e_i \]  \hspace{1cm} (5.1.1)

where \( \alpha \) and \( \beta \) are unknown constants and \( e_i \) is a random variable such that \( \mathbb{E}(e_i | x_i) = 0 \), \( \mathbb{E}(e_i^2 | x_i) = ax_i^g \), \( a > 0 \), \( g > 0 \); and \( \mathbb{E}(e_i e_j | x_i, x_j) = 0 \).

Theorem 5.1:

Average variance of the corresponding H.T. estimator for any I.P.P.S. scheme under the model (5.1.1) is

\[ \mathbb{V}^*(\hat{\gamma}_{H.T.}) = \alpha^2 \left[ \frac{1}{\bar{p}_i} + \sum_{i} \sum_{j \neq i} \frac{p_{ij}}{p_i p_j} - N^2 \right] \\
+ \alpha x^g \left[ \frac{\sum_{t} p_t^{g-1}}{n} - \sum_{t} p_t^g \right] \]  \hspace{1cm} (5.1.2)
Proof:

Taking the expectation of $V(\hat{Y}_{H.T.})$ under the model (5.1.1) we get

$$V^*(\hat{Y}_{H.T.}) = E[V(\hat{Y}_{H.T.})] = \frac{N aX_i^g + (\alpha + \beta X_i)^2}{p_i}$$

$$+ \sum_{i} \sum_{j(i) \neq i} \frac{P_{ij}}{P_i P_j} \cdot (\alpha + \beta X_i) (\alpha + \beta X_j)$$

$$- (N\alpha + \beta X)^2 - a\Sigma X_i^g$$

$$= \alpha^2 [\sum_{i} \sum_{j(i) \neq i} \frac{P_{ij}}{P_i P_j} - N^2]$$

$$+ \alpha \beta [2\Sigma_{i} \frac{X_i}{P_i} + \sum_{i} \sum_{j(i) \neq i} \frac{P_{ij}}{P_i P_j} (X_i + X_j) - 2NX]$$

$$+ \beta^2 [\sum_{i} \frac{X_i^2}{P_i} + \sum_{i} \sum_{j(i) \neq i} \frac{P_{ij}}{P_i P_j} X_i X_j - X^2]$$

$$+ a [\sum_{i} \frac{X_i^g}{P_i} - \Sigma X_i^g],$$

Which upon using the relations $P_i = np_i$ and $\sum_{j(i) \neq i} P_{ij} = (n-1)p_i$ reduces to (5.1.2).

Q.E.D.

Thus from (5.1.2) it follows that when $\alpha = 0$, the average variance of the corresponding H.T. estimator will be the same for all the I.P.P.S. schemes. However, if $\alpha \neq 0$, it can be observed from (5.1.2) that among all the I.P.P.S.
schemes, the H.T. estimator corresponding to the scheme for which the value of \( \sum \sum \frac{P_{ij}}{\bar{P}_{i}P_{j}} \) is least will have the least average variance. Thus a reasonable investigation will be to rank the various I.P.P.S. schemes according to the value of \( \sum \sum \frac{P_{ij}}{\bar{P}_{i}P_{j}} \quad (= C, \text{ say}) \). For this investigation we will confine to the case \( n=2 \) only.

For the schemes of Durbin (1967), Yates and Grundy (1953), Durbin (1953), Goodman and Kish (1950) and Hanurav (1967) the approximate expressions for \( P_{ij} \) correct to \( O(N^{-4}) \) are respectively given by

\[
P_{ij}^{(1)} = 2p_{i}p_{j}[1+\{(p_{i}+p_{j})-\Sigma p_{t}^{2}\}+2(p_{i}^{2}+p_{j}^{2})-2\Sigma p_{t}^{3} \]
\[
- (p_{i}+p_{j})\Sigma p_{t}^{2} + (\Sigma p_{t}^{2})^{2}] \quad (5.1.3)
\]

\[
P_{ij}^{(2)} = 2p_{i}p_{j}[1+\{(p_{i}+p_{j})-\Sigma p_{t}^{2}\}+2(p_{i}^{2}+p_{j}^{2})-2\Sigma p_{t}^{3} + \frac{3}{4} p_{i}p_{j} \]
\[
- \frac{7}{4}(p_{i}+p_{j})\Sigma p_{t}^{2} + \frac{7}{4}(\Sigma p_{t}^{2})^{2}] \quad (5.1.4)
\]

\[
P_{ij}^{(3)} = 2p_{i}p_{j}[1+\{(p_{i}+p_{j})-\Sigma p_{t}^{2}\}+2(p_{i}^{2}+p_{j}^{2})-2\Sigma p_{t}^{3}+p_{i}p_{j} \]
\[
- 2(p_{i}+p_{j})\Sigma p_{t}^{2} + 2(\Sigma p_{t}^{2})^{2}] \quad (5.1.5)
\]

\[
P_{ij}^{(4)} = 2p_{i}p_{j}[1+\{(p_{i}+p_{j})-\Sigma p_{t}^{2}\}+2(p_{i}^{2}+p_{j}^{2})-2\Sigma p_{t}^{3}+2p_{i}p_{j} \]
\[
- 3(p_{i}+p_{j})\Sigma p_{t}^{2} + 3(\Sigma p_{t}^{2})^{2}] \quad (5.1.6)
\]
Expressions (5.1.3), (5.1.6) and (5.1.7) are from Chapter 2 and expressions (5.1.4) and (5.1.5) are from Rao (1963b).

Using Equations (5.1.3)-(5.1.7) and the relation

\[ P_i = 2p_i \]

the value of \( \sum_{i \neq j} p_i p_j \) correct to \( O(N^0) \) for the above five schemes is respectively given by

\[ C_1 = \frac{1}{2} \left[ N^2 + N(1-N\Sigma p_t^2) + \left\{ 3N\Sigma p_t^2 + N^2(\Sigma p_t^2)^2 - 2N^2\Sigma p_t^3 - 2N^2\Sigma p_t^3 - 2 \right\} \right] \] (5.1.8)

\[ C_2 = \frac{1}{2} \left[ N^2 + N(1-N\Sigma p_t^2) + \left\{ \frac{3}{2}N\Sigma p_t^2 + \frac{7}{4}N^2(\Sigma p_t^2)^2 - 2N^2\Sigma p_t^3 - \frac{5}{4} \right\} \right] \] (5.1.9)

\[ C_3 = \frac{1}{2} \left[ N^2 + N(1-N\Sigma p_t^2) + \left\{ N\Sigma p_t^2 + 2N^2(\Sigma p_t^2)^2 - 2N^2\Sigma p_t^3 - 1 \right\} \right] \] (5.1.10)

\[ C_4 = \frac{1}{2} \left[ N^2 + N(1-N\Sigma p_t^2) - \left\{ N\Sigma p_t^2 - 3N^2(\Sigma p_t^2)^2 + 2N^2\Sigma p_t^3 \right\} \right] \] (5.1.11)

and

\[ C_5 = \frac{1}{2} \left[ N^2 + \frac{1}{\Sigma p_t^2}(1-N\Sigma p_t^2) + \left\{ \frac{(\Sigma p_t^3)^2}{\Sigma p_t^2 \cdot \Sigma p_t^4} - 1 \right\} \right] \] (5.1.12)

It can be easily verified from (5.1.8)-(5.1.11) that

\[ C_1 < C_2 < C_3 < C_4 \] (5.1.13)

This is also a direct consequence of the comparisons made by Rao (1963b, 1965) of the above four schemes without any model assumptions.
From (5.1.11) and (5.1.12) we get

\[ C_5 - C_4 = \frac{1}{2} \left[ \frac{1}{2} \cdot (1 - N \Sigma P_t^2)^2 \right] + \left[ \frac{(\Sigma P_t^3)^2}{2} \right] - 3N^2(\Sigma P_t^2)^2 + N\Sigma P_t^2 + 2N^2\Sigma P_t^3 - 1 \]  

(5.1.14)

Now, assuming \( p_1, p_2, \ldots, p_N \) to be having a specific distribution \( \Delta \), with moments \( \mu_r^1 \) we can replace \( \Sigma P_t^r \) in (5.1.14) by \( N\mu_r^1 \) because we have from Khintchine's law of large numbers

\[ \text{plim} \frac{1}{N} \Sigma P_t^r = \mu_r \]  

(5.1.15)

In view of the relation \( \Sigma P_t=1 \), we however should have

\[ \mu_1^1 = \frac{1}{N}. \]  

(5.1.16)

In the following we will investigate the relative efficiency of the Hanurav's procedure in relation to the other procedures considered above under different distributions of \( p_t \).

Case (i) - \( \chi^2 \) distribution:

When the \( p_t \)'s are distributed as \( \frac{1}{N} \chi^2(v) \) where \( \chi^2(v) \) is the chi-square variate with \( v \) degrees of freedom, from the relation

\[ \Sigma P_t^r = N\mu_r^1 \]  

(5.1.17)
we get

$$\Sigma p_t^2 = \frac{v+2}{vN}$$  \hspace{1cm} (5.1.18)

$$\Sigma p_t^3 = \frac{(v+2)(v+4)}{v^2 N^2}$$  \hspace{1cm} (5.1.19)

and

$$\Sigma p_t^4 = \frac{(v+2)(v+4)(v+6)}{v^3 N^3}$$  \hspace{1cm} (5.1.20)

Substituting these values in (5.1.14) we get

$$C_5-C_4 = \frac{1}{2} \left[ \frac{vN}{(v+2)} \left( 1 - \frac{v+2}{v} \right)^2 + \left( \frac{v+4}{v+6} - \frac{3(v+2)^2}{v^2} + \frac{v+2}{v} ight) \right] + \frac{2(v+2)(v+4)}{v^2} - 1} \right]$$

which after simplification reduces to

$$C_5-C_4 = \frac{2N}{v(v+2)} + \frac{4(2v+3)}{v^2(v+6)} \geq 0$$  \hspace{1cm} (5.1.21)

Case (ii) - $\beta$ distribution:

When the $p_t$'s follow a beta distribution of the first kind with parameters $(a_1-1,a_2)$ where $a_1$ and $a_2$ are related by the equation

$$\nu_1 = \frac{a_1}{a_1+a_2+1} = \frac{1}{N}$$

or

$$a_2 = (N-1)a_1 - 1$$  \hspace{1cm} (5.1.22)
we get after substituting $\text{Nu}_1$ for $\Sigma p_t$, 

$$\Sigma p_t^2 = \frac{\alpha_1 + 1}{Na_1 + 1} \tag{5.1.23}$$

$$\Sigma p_t^3 = \frac{(\alpha_1 + 1)(\alpha_1 + 2)}{(Na_1 + 1)(Na_1 + 2)} \tag{5.1.24}$$

and

$$\Sigma p_t^4 = \frac{(\alpha_1 + 1)(\alpha_1 + 2)(\alpha_1 + 3)}{(Na_1 + 1)(Na_1 + 2)(Na_1 + 3)} \tag{5.1.25}$$

Substituting from (5.1.23)-(5.1.25) in (5.1.14) we get

$$C_5 - C_4 = \frac{1}{2(\alpha_1 + 1)(\alpha_1 + 3)(Na_1 + 1)^2(Na_1 + 2)}$$

$$[Na_1^3(N^2 + 2N^2 - 7N + 4)$$

$$+ \alpha_1^2(3N^4 + 4N^3 - 22N^2 + 14N + 1)$$

$$+ \alpha_1(12N^3 - 33N^2 + 18N + 3) - 6(N - 1)] \tag{5.1.26}$$

Now,

$$N^3 + 2N^2 - 7N + 4 = N(N^2 - 1) + 2N(N - 3) + 4$$

$$> 0 \text{ for } N \geq 3 \tag{5.1.27}$$

$$3N^4 + 4N^3 - 22N^2 + 14N + 1 = N^3(3N - 7) + 11N^2(N - 2) + 14N + 1$$

$$> 0 \text{ for } N \geq 3 \tag{5.1.28}$$

and

$$\alpha_1(12N^3 - 33N^2 + 18N + 3) - 6(N - 1)$$

$$= (\alpha_1 - 1)[3N^2(4N - 11) + 3(6N + 1)] + 3N^2(4N - 11) + 3(4N + 3)$$

$$> 0 \text{ for } N \geq 3, \tag{5.1.29}$$
because
\[ a_1 > 1 \]

(5.1.26)-(5.1.28) imply that

\[ C_5 - C_4 > 0 \quad (5.1.30) \]

Case (iii) - uniform distribution:

When the \( p_t \)'s follow a uniform distribution over the interval \((0, \frac{2}{N})\), we get from \( \Sigma p_t^x = N \mu_t^x \),

\[ \Sigma p_t^2 = \frac{4}{3N} \]
\[ \Sigma p_t^3 = \frac{2}{N^2} \]

and

\[ \Sigma p_t^4 = \frac{16}{5N^3} \]

substitution of these values in (5.1.14) gives,

\[ C_5 - C_4 = \frac{(4N-3)}{96} > 0 \quad (5.1.31) \]

In view of Equations (5.1.13), (5.1.21), (5.1.30) and (5.1.31) it follows that when the variance is considered to \( O(N^0) \), Hanurav's strategy would be inferior to those of Durbin (1967), Yates and Grundy (1953), Durbin (1953), and Goodman and Kish (1950) when the \( p_t \)'s follow chi-square, beta or uniform distributions.
5.2. Alternative Use of Ancillary Information

In this section we consider an alternative way of using the ancillary information in providing a better estimate for the population total than the Rao, Hartley and Cochran's estimator.

Suppose the $i$th unit $U_i$ having the size $X_i$ is considered as made up of $X_i$ sub-units having the same value $Y_i/X_i$, which means that the $j$th sub-unit of the $i$th unit is taken as having the value $Z_{ij} = Y_i/X_i$, $j = 1, 2, \ldots, X_i$. Then the process of selecting one sub-unit with simple random sampling from the population of $X(=\Sigma X_i)$ sub-units and considering the unit to which it belongs as selected is equivalent to selecting a unit with probability proportional to size because the probability of selecting any unit is proportional to the number of sub-units in it. Thus the equal probability estimator based on a selected sub-unit is the same as the probability proportional to size estimator.

If we denote the $j$th sub-unit of the $i$th unit as $U_{ij}$, we can arrange the $X$ sub-units as $U_{11}, U_{12}, \ldots, U_{1X_1}; U_{21}, U_{22}, \ldots, U_{2X_2}; \ldots; U_{N_1}, U_{N_2}, \ldots, U_{NX_N}$.

Redesignating these sub-units preserving the order as $V_1, V_2, \ldots, V_X$ we can rewrite the set of sub-units as

$$V = \{V_1, V_2, \ldots, V_X\} \quad (5.2.1)$$
It can be observed that each of the sub-units \( V_1, V_2 \ldots V_{X_1} \) has the same y-value \( Y_1/X_1 \); each of the sub-units \( V_{X_1+1}', V_{X_1+2}' \ldots V_{X_1+X_2}' \) has the same y-value \( Y_2/X_2 \) and in general each of the sub-units \( V_{j-1}', V_{j-1}' \ldots \sum_{i=1}^{X_j+1} X_i+1, \sum_{i=1}^{X_j+2} X_i \) has the same y-value \( Y_j/X_j \) \((j=1,2 \ldots N)\). Now since each \( U_{ij} \) is some \( V_\alpha \) and each \( V_\alpha \) is some \( U_{ij} \), if we denote the y-value corresponding to \( V_\alpha \) as \( Z_\alpha \), we get

\[
Z_\alpha = Z_{ij} = Y_i/X_i \quad (5.2.2)
\]

Assuming the number of sub-units \( X \) to be a multiple of \( N \), we define a new sampling frame \( U' \), by defining a new set of units \( U'_1', U'_2' \ldots U'_N' \), wherein, the first \( \bar{X}(=X/N) \) sub-units constitute \( U'_1' \), the subsequent \( \bar{X} \) sub-units constitute \( U'_2' \) and in general the sub-units \( V_{(j-1)\bar{X}+1}', V_{(j-1)\bar{X}+2}' \ldots V_{j\bar{X}}' \) constitute \( U'_j' \) \((j=1,2 \ldots N)\). Thus we get the new population frame

\[
U' = \{U'_1', U'_2' \ldots U'_N'\} \quad (5.2.3)
\]

Denoting the y-value corresponding to \( U'_j' \) by \( Y'_j \) we can observe that \( Y'_j \) will be the sum of the y-values corresponding to the sub-units constituting \( U'_j' \) and thus we have
\[ Y_j = \sum_{\alpha=(j-1)X+1}^j Z_{\alpha} \quad (j = 1, 2, \ldots N) \quad (5.2.4) \]

If \( Y' \) and \( Y \) denote the column vectors

\[
Y' = \begin{bmatrix} Y'_1 \\ Y'_2 \\ \vdots \\ Y'_N \end{bmatrix} \quad \text{and} \quad Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_N \end{bmatrix},
\]

the nature of relationship between \( Y' \) and \( Y \) can best be visualized with the help of a simple example.

Let us consider the population of size 4 and with

\( (Y_1', X_1) = (15, 3); \quad (Y_2', X_2) = (22, 4); \quad (Y_3', X_3) = (19, 4) \) and

\( (Y_4', X_4) = (24, 5) \). Here

- \( U_1 \) has 3 sub-units with \( \frac{Z_{1j}}{X_1} = \frac{15}{3} \) \( (j = 1, 2, 3) \);
- \( U_2 \) has 4 sub-units with \( \frac{Z_{2j}}{X_2} = \frac{22}{4} \) \( (j = 1, 2, 3, 4) \);
- \( U_3 \) has 4 sub-units with \( \frac{Z_{3j}}{X_3} = \frac{19}{4} \) \( (j = 1, 2, 3, 4) \); and
- \( U_4 \) has 5 sub-units with \( \frac{Z_{4j}}{X_4} = \frac{24}{5} \) \( (j = 1, 2, \ldots 5) \).

From (5.2.4) we get

\[
Y'_1 = \frac{15}{3} + \frac{15}{3} + \frac{15}{3} + \frac{22}{4} = 1 \cdot Y_1 + \frac{1}{4} \cdot Y_2
\]

\[
Y'_2 = \frac{22}{4} + \frac{22}{4} + \frac{22}{4} + \frac{19}{4} = \frac{3}{4} \cdot Y_2 + \frac{1}{4} \cdot Y_3
\]

\[
Y'_3 = \frac{19}{4} + \frac{19}{4} + \frac{19}{4} + \frac{24}{5} = \frac{3}{4} \cdot Y_3 + \frac{1}{5} \cdot Y_4 \quad (5.2.5)
\]

and

\[
Y'_4 = \frac{24}{5} + \frac{24}{5} + \frac{24}{5} + \frac{24}{5} = \frac{4}{5} \cdot Y_4
\]

These equations can be written in a matrix form as
\[ Y' = A \cdot Y \]

where the transformation matrix \( A \) is given as

\[
A = \begin{bmatrix}
1 & 1/4 & 0 & 0 \\
0 & 3/4 & 1/4 & 0 \\
0 & 0 & 3/4 & 1/5 \\
0 & 0 & 0 & 4/5
\end{bmatrix}
\]

Thus, in general the relationship between \( Y' \) and \( Y \) can be written as

\[ Y' = AY \quad (5.2.6) \]

where

\[
A = \begin{bmatrix}
a_{11} & a_{12} & \ldots & a_{1N} \\
a_{21} & a_{22} & \ldots & a_{2N} \\
\vdots & \vdots & \ddots & \vdots \\
a_{N1} & a_{N2} & \ldots & a_{NN}
\end{bmatrix}
\quad (5.2.7)
\]

The elements \( a_{ij} \) of the matrix are given by

\[ a_{ij} = \frac{a_{ij}}{x_j} \quad (5.2.8) \]

where \( a_{ij} \) satisfy the conditions,

(i) \( a_{ij} > 0 \) \quad (5.2.9)

(ii) \( \sum_{i=1}^{N} a_{ij} = x_j \), for \( j = 1, 2, \ldots N \) \quad (5.2.10)

(iii) \( \sum_{j=1}^{N} a_{ij} = \bar{x} \), for \( i = 1, 2, \ldots N \) \quad (5.2.11)
Now we recall the definition of a stochastic matrix.

**Definition 5.1:**

An \( nxn \) matrix \( P = (p_{ij}) \) is called a stochastic matrix if

(i) \( p_{ij} \geq 0 \)

and

(ii) \( \sum_{j=1}^{N} p_{ij} = 1, \ i = 1, 2 \ldots N \)

**Lemma 5.1:**

Every \( nxn \) stochastic matrix \( P = (p_{ij}) \) satisfies the equation \( P \cdot 1 = 1 \) where \( 1 \) denotes the \( nx1 \) column vector of 1's.

Proof is immediate from the definition.

Now, since each \( x_j \geq 0 \) we get from (5.2.8)-(5.2.10) that

\[
a_{ij} \geq 0
\]

and

\[
\sum_{i=1}^{N} a_{ij} = 1
\]

Using (5.2.12) and (5.2.13), we get from Definition 5.1 and Lemma 5.1 that:

The transpose \( A^T \) of the matrix \( A \) of (5.2.7) is a stochastic matrix and hence satisfies the equation

\( A^T \cdot 1 = 1 \).

Hence from (5.2.6) we get
Thus, in order to estimate the population total $Y$, we can use the sampling frame $U'$. Now, we consider the procedure of selecting a simple random sample without replacement of size $n$ from the population $U'$. We will call this procedure as 'Modified Simple Random Sampling' (M.S.R.S.), since we are adopting the simple random sampling procedure after modifying the sampling frame. The estimator of the population total proposed is

$$\hat{Y}_{MSRS} = \frac{N}{n} \sum_{i=1}^{n} Y'_i$$

(5.2.15)

As is well known,

$$\text{Var}(\hat{Y}_{MSRS}) = \frac{N-n}{n} \cdot \frac{N}{N-1} \left( \frac{N}{\sum_{i=1}^{N} Y_i^2 - N\bar{Y}^2} \right)$$

(5.2.16)

Theorem 5.2:

As an estimator of the population total $Y$, $\hat{Y}_{MSRS}$ has uniformly smaller variance than the Rao, Hartley and Cochran's estimator.

Proof:

Variance of the Rao, Hartley and Cochran's estimator is

$$\text{Var}(\hat{Y}_{RHC}) = \left(1 - \frac{n-1}{N-1}\right) \frac{1}{n} \left(\frac{\sum_{i=1}^{n} Y_i^2}{P_i} - \bar{Y}^2 \right)$$

(5.2.17)
From (5.2.16) and (5.2.17) we get

\[
V(\hat{\phi}_{\text{RHC}}) - V(\hat{\phi}_{\text{MSRS}}) = \frac{N(N-n)}{n(N-1)} \left( \sum_{i=1}^{N} \frac{Y_i^2}{Np_i} - \sum_{i=1}^{N} Y_i^2 \right)
\]

\[
= \frac{N(N-n)}{n(N-1)} (Y^T D Y - Y^T A^T A Y)
\]

\[
= \frac{N(N-n)}{n(N-1)} \cdot Y^T C Y
\]

(5.2.18)

where

\[
D = \begin{bmatrix}
d_1 & 0 & 0 & \ldots & 0 \\
0 & d_2 & 0 & \ldots & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & \ldots & d_N \\
\end{bmatrix}
\]

(5.2.19)

with

\[
d_i = \frac{1}{Np_i} = \frac{\bar{X}}{X_i}
\]

(5.2.20)

and

\[
C = D - A^T A
\]

(5.2.21)

Now, let

\[
B = (b_{ij}) = A^T A
\]

(5.2.22)

Then,

\[
b_{ij} = \sum_{K=1}^{N} a_{K_i} a_{K_j}
\]

(5.2.23)

Thus, we get

\[
Y^T C Y = \sum_{i=1}^{N} (d_i - b_{ii}) Y_i^2 - 2 \sum_{i<j} b_{ij} Y_i Y_j
\]

(5.2.24)
If we define
\[ \beta_{ij} = \sum_{K=1}^{N} \alpha_{K_i} \alpha_{K_j} \geq 0 \]  
(5.2.25)
we get by using Equations (5.2.10) and (5.2.11),
\[ \beta_{ij} = \sum_{j=1}^{N} \beta_{ij} = \sum_{j=1}^{N} \alpha_{K_i} (\sum_{j=1}^{N} \alpha_{K_j}) = \overline{x} \cdot \sum_{K=1}^{N} \alpha_{K_i} \]
\[ = \overline{x} \cdot x_i = x_i^2 \cdot d_i \]  
(5.2.26)
By symmetry of \( \beta_{ij} \) we get
\[ \beta_{ji} = \sum_{i=1}^{N} \beta_{ij} = \sum_{i=1}^{N} \beta_{ji} = \beta_{ji} = \overline{x} \cdot x_j = x_j^2 \cdot d_j \]  
(5.2.27)
Also from (5.2.8) we get
\[ \beta_{ii} = \sum_{K=1}^{N} \alpha_{K_i}^2 = x_i^2 \sum_{K=1}^{N} \alpha_{K_i}^2 = x_i^2 \cdot b_{ii} \]  
(5.2.28)
Equation (5.2.24) can be written as
\[ y^T Cy = \sum_{i=1}^{N} (d_i - b_{ii}) y_i^2 - 2 \sum_{i<j} \beta_{ij} \frac{y_i}{x_i} \cdot \frac{y_j}{x_j} \]
\[ = \sum_{i=1}^{N} (d_i - b_{ii}) y_i^2 + \sum_{i<j} \beta_{ij} \left(\frac{y_i}{x_i} - \frac{y_j}{x_j}\right)^2 \]
\[ - \sum_{i<j} \beta_{ij} \left(\frac{y_i^2}{x_i^2} + \frac{y_j^2}{x_j^2}\right) \]  
(5.2.29)
Now using (5.2.26)-(5.2.28) we get
Thus, from (5.2.29) and (5.2.30) we get

\[ V(Y_{\text{RHC}}) - V(Y_{\text{MSRS}}) = \frac{N(N-n)}{n(N-1)} Y_{\text{TCY}} > 0 \]

Q.E.D.

The technique of cluster sampling is widely used in large scale surveys in view of its operational conveniences and particularly due to its advantages in view of cost considerations. In situations where it is convenient to take certain naturally formed groups of units as clusters, the cluster size would, in general, vary from cluster to cluster.
Households which are groups of persons and villages which are groups of households are most often considered as clusters for purposes of sampling. Since in most of the practical situations the cluster total of the variable under study is likely to be positively correlated with the number of units in the cluster, it would be profitable to select the clusters with probability proportional to the number of units in the cluster. In particular one can adopt the Rao, Hartley and Cochran's procedure in view of its applicability in large scale surveys. Alternatively one can use the 'Modified Simple Random Sampling' procedure described in this section with advantage. Instead of assuming that all the sub-units of a cluster have the same y-values, which we did for theoretical purposes, we actually observe the corresponding y-value for each sub-unit that gets selected in the sample through the method of MSRS procedure. The results, however, are not expected to deviate from the theoretical studies in view of the approximate proportionality that usually exists between the study variable and the auxiliary variable.

**Numerical example:**

The relative performance of the RHC estimator and the MSRS estimator is studied through the help of an example. The data considered here is the 1960 population of the first fifteen counties of Iowa by minor civil divisions.
Considering this as our population under study, our purpose is to estimate the total number of inhabitants in the first fifteen counties. The counties are considered as clusters and the minor civil divisions within the county are the elements within the cluster. Instead of presenting the y-values for each minor civil division, we presented in Table 5.1 only the y-values and the number of minor civil divisions for each county. We also presented the y-values corresponding to the modified frame which are denoted by \( Y_j \).

The true variances of the RHC estimator as well as the MSRS estimator are calculated for samples of 3 clusters and we obtained

\[
V(\hat{\phi}_{RHC}) = 1475965 \times 10^5
\]

and

\[
V(\hat{\phi}_{MSRS}) = 1022261 \times 10^5
\]

Relative efficiency of the MSRS estimate with respect to the RHC estimate is

\[
\frac{1475965}{1022261} \approx 1.44
\]
Table 5.1. Table of the county totals $Y_i$, number of inhabitants $X_i$ and the corresponding $Y'_i$

<table>
<thead>
<tr>
<th>County No.</th>
<th>Number of inhabitants $Y_i$</th>
<th>Number of minor civil divisions $X_i$</th>
<th>$Y'_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>15534</td>
<td>27</td>
<td>15903</td>
</tr>
<tr>
<td>2</td>
<td>10206</td>
<td>17</td>
<td>17943</td>
</tr>
<tr>
<td>3</td>
<td>23538</td>
<td>25</td>
<td>33400</td>
</tr>
<tr>
<td>4</td>
<td>26358</td>
<td>31</td>
<td>21307</td>
</tr>
<tr>
<td>5</td>
<td>15715</td>
<td>18</td>
<td>19876</td>
</tr>
<tr>
<td>6</td>
<td>36499</td>
<td>38</td>
<td>31541</td>
</tr>
<tr>
<td>7</td>
<td>227737</td>
<td>30</td>
<td>162557</td>
</tr>
<tr>
<td>8</td>
<td>45251</td>
<td>33</td>
<td>17034</td>
</tr>
<tr>
<td>9</td>
<td>33479</td>
<td>24</td>
<td>36906</td>
</tr>
<tr>
<td>10</td>
<td>33063</td>
<td>28</td>
<td>32372</td>
</tr>
<tr>
<td>11</td>
<td>34457</td>
<td>29</td>
<td>37156</td>
</tr>
<tr>
<td>12</td>
<td>26383</td>
<td>30</td>
<td>21696</td>
</tr>
<tr>
<td>13</td>
<td>24724</td>
<td>31</td>
<td>30189</td>
</tr>
<tr>
<td>14</td>
<td>37888</td>
<td>34</td>
<td>26718</td>
</tr>
<tr>
<td>15</td>
<td>29031</td>
<td>25</td>
<td>31265</td>
</tr>
</tbody>
</table>

\[ a \text{The total number of inhabitants} = 619863. \]

\[ b \text{The total number of minor civil divisions} = 420, \]
\[ \text{the average number of minor civil divisions} = 28. \]
6. BIBLIOGRAPHY


7. ACKNOWLEDGMENTS

The author sincerely acknowledges Dr. B. V. Sukhatme for his guidance and encouragement during the course of this research and preparation of the manuscript.

Thanks are also due to Dr. T. A. Bancroft and Dr. H. A. David for their encouragement and assistance to the author during his stay at Iowa State University.

Acknowledgment is also made to the Statistical Reporting Service, U.S. Department of Agriculture and to the U.S. Office of Education, Department of Health, Education and Welfare under contract No. OEC-0-73-6640 for their support of the author's graduate work and research.

Last but not the least, the author is prompted by his sacred filial duty to record his deep appreciation of his parents for the noble qualities of the highest order they cherished and for the enormous encouragement they gave to pursue his doctoral studies.