1961

Sampling procedures involving unequal probability selection

Jonnagadda Nalini Kanth Rao
Iowa State University

RAO, Jonnagadda Nalini Kanth. SAMPLING PROCEDURES INVOLVING UNEQUAL PROBABILITY SELECTION.

Iowa State University of Science and Technology
Ph.D., 1961
Mathematics

University Microfilms, Inc., Ann Arbor, Michigan

SAMPLING PROCEDURES INVOLVING
UNEQUAL PROBABILITY SELECTION

by

Jonnagadda Nalini Kanth Rao

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

Major Subject: Statistics

Approved:

Signature was redacted for privacy.

In Charge of Major Work

Signature was redacted for privacy.

Head of Major Department

Signature was redacted for privacy.

Dean of Graduate College

Iowa State University
Of Science and Technology
Ames, Iowa

1961

TABLE OF CONTENTS

I. INTRODUCTION

The use of sample surveys for the estimation of population characteristics is an important tool in modern social and economic planning. Since the idea of using this device is to save the expenditures involved in complete enumeration or censuses of populations the question of the cost of such surveys and the precision of estimates computed from them is of great importance. It has therefore been of major concern to the theory and design of statistical sample surveys to develop methods which yield estimates of high precision at comparatively moderate cost.

The devices which are available for this purpose essentially fall into two groups: (a) Methods in which the mode of computing estimates (of say population mean or total) are developed which have higher precision, or in other words, the development of estimators with smaller variances. The so called "ratio and regression estimators" are examples of these. The theory of ratio and regression methods of estimation has been extensively developed in recent years and unbiased ratio and regression type estimators are now available which correct for bias in the classical ratio and regression estimators. (b) Methods of improving the "design of the sample survey", *i.e.* the mode in which the sample data are collected. In this category fall such devices as choice of sampling unit, stratification, multistage and multiphase sampling and unequal probability sampling. The first two items do not present any difficulties as far as theoretical aspects of estimation etc. are concerned. Multistage and multiphase sampling have been extensively dealt with in the literature. In this dissertation, we will be mainly concerned with the theory of sampling with unequal probabilities. Often, one uses some or all the devices mentioned in groups (a) or (c) simultaneously in order to improve the precision of estimators. For example, a stratified two stage design with the primaries selected with probabilities proportional to sizes is a familiar design in large scale sample surveys.

Unequal probability sampling involves selection of sampling units with probabilities proportional to size of the supplementary variable which is correlated with the characteristic for which the population total or mean is to be estimated. For example, total corn production on a farm is very likely correlated with the supplementary variable, total acreage of the farm. The theory of unequal probability sampling can be directly derived from the properties of the multinomial distribution and presents no inherent difficulties provided the sampling units are drawn with replacement. But, it is well known from the theory of equal probability sampling that sampling with replacement is less precise than sampling without replacement, the proportional reduction in variance being equal to fraction of the population sampled. Therefore, one naturally expects that similar gains in precision can be made by using unequal probability sampling without replacement instead of with replacement.

However, since the probability of drawing a sampling unit does not remain constant with each draw when sampling without replacement, evaluation of selection probabilties and variance formulas involves certain mathematical and computational difficulties and therefore this theory has not yet become popular with survey practitioners. Certain shortcomings of existing published literature on this theory can be listed as follows: 1) Most of the writers deal almost exclusively with sample size of two only, and have very little to offer when sample size is greater than two, since the expressions for selection probabilities become unwieldy and extremely difficult to compute. 2) Some of the procedures proposed have the undesirable property that estimates of the variance can take negative values. 3) Sampling without replacement is sometimes less efficient than sampling with replacement particularly when the sample size is greater than two.
4) These methods do not have the desirable property that the probability of selecting a unit in the sample is proportional to size of the supplementary variable which is universally recognized as a technique yielding considerable reduction in the variance of the estimators. To overcome this contingency, methods such as "revised size measures" of the supplementary

variable are suggested which ensure that this condition is satisfied approximately. However, these methods become cumbersome when the sample size is greater than two and the population size is large, due to the computational difficulties involved in finding "revised size measures". These are some of the main reasons why survey practitioners usually do not favor unequal probability sampling without replacement over sampling with replacement and hence unequal probability sampling with replacement is extensively used in large scale sample surveys.

In this dissertation, we propose to develop an asymptotic theory applicable for any sample size and for large or medium sized populations which takes care of at least all the contingencies mentioned above. We adopt a simple sampling procedure of selecting units with unequal probabilities and without replacement well known to survey practitioners which has been abandoned due to mathematical difficulties in developing the theory. This procedure ensures that the probability of selecting a sampling unit in the sample is exactly proportional to size of the supplementary variable. Compact expressions for the variance and for the estimate of the variance applicable to large and medium sized populations are obtained which are simple to compute and show that this procedure is always more precise than unequal probability sampling with replacement, and that estimates of the variance are always positive. An important merit of this procedure is that it permits ready evaluation of selection probabilities and variance formulas for sample size greater than two, unlike the procedures available in the literature. We hope that these results may stimulate the interest of survey practitioners in unequal probability sampling without replacement, and help in designing efficient sample surveys.

II. REVIEW OF THE LITERATURE

Since ratio and regression methods of estimation are alternative ways of utilizing supplementary information, we shall begin with a brief review of the theory of ratio and regression estimation. Ratio and regression type estimates have been extensively used in the literature for utilizing supplementary information. The well known ratio estimator of the population total Y is

$$\hat{Y}_R = \frac{\bar{Y}}{\bar{x}} \cdot X \tag{2.1}$$

where $\bar{y}$, $\bar{x}$ are the sample means and X is the population total for the supplementary variable x. Bias in this estimator is $\text{cov}(\frac{\bar{Y}}{\bar{x}},\bar{x})$ which is of the order 1/n where n is the sample size so that the bias is negligible for large samples. Hartley and Ross (1954) have developed an unbiased ratio type estimator which seems to compare favorably with $\hat{Y}_R$ regarding efficiency, though the computations involved in using this unbiased estimator are more cumbersome compared with those in using the estimator $\hat{Y}_R$.

The classical regression estimator is based on a linear model

$$y_i = A + Bx_i + e_i \tag{2.2}$$

where x_i's are unspecified and observed without error and e_i and x_i are assumed to be independent and

$$E(e_i | x) = 0 \quad , \quad E(e_i^2 | x) = \sigma^2 \; . \tag{2.3}$$

Under these assumptions the minimum variance unbiased linear estimator of Y is

$$\hat{Y}_B = N\,\bar{y} + b(\bar{X} - \bar{x}) \qquad (2.4)$$

where b is the sample regression coefficient.

However, it is not very realistic to assume such a model in practice so that this estimator is generally biased. Mickey (1954, 1959) has discovered an ingenious and simple procedure of constructing a large variety of unbiased ratio and regression type estimators and this procedure has been further exploited by Williams (1958) to develop and investigate the properties of unbiased regression type estimators.

The possibility of using unequal probabilities for selecting the sampling units to increase the precision of estimates is first considered by Hansen and Hurwitz (1943). Using a two stage stratified sampling design they select one first stage unit from each stratum with probability proportional to number of second stage units in a first stage unit. It is demonstrated that marked reduction in variance over sampling with equal probabilities can be obtained by switching to unequal probability sampling. However, since only one first stage unit is selected from each stratum, no valid estimate of the variance can be obtained and so approximate methods using collapsed strata are suggested for estimating the variance. To avoid this, it has been a common practice in sample surveys to select two or more first stage units with replacement and with p.p.s. (probabilities proportional to size) of the x variable, since the existing theory of sampling with p.p.s. and without replacement presents certain difficulties as will be evident later in the review. An important advantage of sampling with replacement is that an unbiased estimate of the variance for each stratum is simply given by the mean square of estimated totals of the selected first stage units in the stratum and does not depend on the method of selection of second stage units provided separate samples of second stage units are drawn when a first stage unit is selected twice or more. A full account of this theory is available in many of the standard text books on sampling, _e_._g_. Sukhatme (1954), and can be summarized as follows for single stage sampling: Let p_i denote the probability of selecting i^{th} unit in the first draw. Then, an estimate of the total Y is

$$\hat{Y}' = n^{-1} \sum^{n} \frac{y_i}{p_i} , \tag{2.5}$$

the variance of the estimate is

$$V(\hat{Y}') = \sum^{N} np_i \left(\frac{y_i}{np_i} - \frac{Y}{n}\right)^2 \tag{2.6}$$

and an unbiased estimate of the variance is

$$v(\hat{Y}') = \frac{n}{n-1} \sum^{n} \left(\frac{y_i}{np_i} - \frac{\hat{Y}'}{n}\right)^2 . \tag{2.7}$$

Midzuno (1950) has extended Hansen and Hurwitz's theory to sampling a combination of n units with probability proportional to some measure of size of the combination. It is interesting to note that this probability is equal to the total probability of selecting the first unit with p.p.s. and the remaining (n - 1) units with equal probabilities and without replacement. Lahiri (1951) and Des Raj (1954) use Midzuno's procedure in constructing an unbiased ratio estimator by selecting the n units with probabilities proportional to total measure of size of x for the n units. It should be noted that in Hartley and Ross' method, the sampling procedure is not modified as is done by Lahiri and Des Raj, but the usual ratio estimators are modified so that a ratio type estimator is obtained that is unbiased for the usual simple random sampling procedure. Madow (1949) has considered systematic sampling of clusters with probabilities proportional to size, but no valid estimate of the variance can be obtained.

When sampling a finite population without replacement, the class of all unbiased linear estimators can be separated into a number of subclasses of estimators by the nature of coefficients, or weights attached to the observations in the sample. Horvitz and Thompson (1952) have distinguished three subclasses of estimators and Koop (1957) has formulated a more general discussion of the possible subclasses and has investigated some properties of the estimators in each subclass. We shall give a brief review of Koop's formulation below. There are seven different subclasses of unbiased linear estimators. Let T_i denote an estimator in class i. Then, T_1 has weights based on the order of appearance of the units in the sample, T_2 on the presence or absence of a given unit in the sample, T_3 on the set of units composing the sample, T_4 on the appearance of a given unit at a given draw, T_5 on the given unit and the particular sample in which it appears, T_6 on the set of units appearing in a specific order, and T_7 on the unit, the order of its draw and the particular sample in which it appears. Minimum variance unbiased linear estimators are obtained in each subclass using Lagrange's multipliers. However, the weights so obtained depend on the unknown y's. To avoid this, Koop obtains simulated minimum variance unbiased linear estimators by using the relation y = cx where c is a constant.

We feel that this simulation based on the exact relation y = cx is not too realistic in practice and may give a completely false picture if this relationship does not hold. Also, certain systems of linear simultaneous equations have to be solved in order to obtain these weights which become very cumbersome when k is fairly large. Koop states that with the help of electronic computers these calculations can be performed easily. However, in underdeveloped countries access to electronic computers is restricted, and most of the data have to be analyzed on desk calculators. The need for sample surveys in planning economic development is considerable in underdeveloped countries, so that these results have limited use and in any case simplicity of computations is considered as one of the important factors in choosing a sampling procedure.

Godambe (1955) has shown that it is not possible to construct a sampling procedure and associated unbiased linear estimator which is uniformly best for all populations. The efficiency comparisons between the seven subclasses depend on the kind of probability system used except that the variance of T_6 is greater than the variance of T_3. Estimators belonging to the first three subclasses are considered in detail in the literature, though Koop has investigated some properties of estimators in the remaining four subclasses and not many useful results have been obtained regarding their applicability. Lahiri's (1951) unbiased ratio estimator belongs to subclass 3, and estimate of its variance can assume negative values.

Horvitz and Thompson (1952) deal with linear estimators belonging to subclass 2. Their estimator of the total Y is

$$\hat{Y} = \sum_{i=1}^{n} \frac{y_i}{P_i} \qquad (2.8)$$

where P_i is the probability for i^{th} unit to be in the sample. This is the only unbiased estimator possible in subclass 2 and hence the best estimator provided the weights in the linear estimators are assumed to be independent of y's. Koop's (1957) minimum variance unbiased linear estimator in this subclass has weights which depend on y's. In this dissertation, we will be mainly concerned with the estimator $\hat{Y}$ since the sampling procedure adopted is appropriate to this estimator. The variance of $\hat{Y}$ is given by

$$V(\hat{Y}) = \sum^{N} \frac{y_i^2}{P_j} + 2 \sum_{i<i'}^{N} \frac{P_{ii'}}{P_i P_{i'}} y_i y_{i'} - Y^2 \qquad (2.9)$$

where $P_{ii'}$ denotes the probability for the i^{th} and the i'^{th} unit to be both in the sample.

Now, when the P_j are exactly proportional to the y_j, the variance of $\hat{Y}$ is zero which suggests that by making the P_j proportional to the x_j, considerable reduction in the variance of $\hat{Y}$ will result if the x_j are approximately proportional to the y_j. So, the main problem is the evaluation of $P_{ii'}$ and hence $V(\hat{Y})$ when considering sampling procedures which satisfy this "desired optimality" condition, namely,

$$P_i = (n - 1)^{-1} \sum_{i' \neq i}^{N} P_{ii'} = np_i \qquad (2.10)$$

where $p_i = x_i/X$. Since we are mainly concerned with this problem in this dissertation, we shall discuss in detail the available methods and their limitations to deal with this problem after reviewing some more literature on estimators in unequal probability sampling without replacement.

Horvitz and Thompson's (1952) unbiased estimate of the variance of $\hat{Y}$ is

$$v_{HT}(\hat{Y}) = \sum^{n} \frac{1 - P_j}{P_j} y_j^2 + \sum_{i \neq i'}^{n} \frac{P_{ii'} - P_i P_{i'}}{P_i P_{i'} P_{ii'}} y_i y_{i'} \ . \quad (2.11)$$

This estimate of the variance can assume negative values. So, Yates and Grundy (1953) have proposed an alternative estimate of the variance which is believed to be less often negative. Their estimate of the variance is

$$v_{YG}(\hat{Y}) = \sum_{i'>i}^{n} \frac{P_i P_{i'} - P_{ii'}}{P_{ii'}} \left(\frac{y_i}{P_i} - \frac{y_{i'}}{P_{i'}}\right)^2 \ . \quad (2.12)$$

Since this is a weighted sum of squares unlike (2.11), it has some desirable features though it is possible to construct examples to show that (2.12) can be negative (_e.g._ Des Raj, 1956a). It is shown by Sen (1953) and Des Raj (1956a) that (2.12) is always positive at least for the following two important sampling systems: (a) The first unit is selected with p.p.s. and the remaining (n - 1) units are selected with equal probabilities and without replacement. This is due to Midzuno (1950). (b) The first unit is selected with p.p.s. and the second unit with p.p.s. of the remaining units, the sample size being two. This is due to Horvitz and Thompson (1952).

We shall later in Chapter VI, section A, identify a new sampling system with sample size greater than two, for which the Yates and Grundy estimate of the variance is always positive. The expressions for probabilities P_i and $P_{ii'}$ are quite simple for systems (a) and (b) and for the new system so that these systems may be useful. It will be of interest to identify more useful sampling systems for which the Yates and Grundy estimate of the variance is always positive. Another important property of the Yates and Grundy estimate of the variance will be demonstrated in Chapter IV, section C. It will be shown for the case of sample size two that, if there exists a sampling procedure without replacement satisfying the conditions (2.10) and is such that the variance of $\hat{Y}$ given by (2.9) is smaller than the variance of $\hat{Y}'$ when sampling with replacement, namely (2.6), then the Yates and Grundy estimate of the variance is always positive. This is a useful result since we are interested in only those sampling systems for which sampling without replacement is more precise than sampling with replacement. In this connection, one may note Durbin's (1953) comment that the variance of $\hat{Y}$ need not always be smaller than the variance of $\hat{Y}'$ and it is easy to find cases in which the contrary is true.

Since the Yates and Grundy estimate of variance can take negative values, Des Raj (1956a) has considered a set of estimators belonging to subclass 1 with weights based on the order of appearance of the units, while the estimates of the variance of these estimators are always positive. Murthy (1957) has shown that to any ordered estimator there exists an unordered estimator which has smaller variance than the former, and so by unordering Des Raj estimators, unordered estimators with smaller variance than the former are obtained. However, for the case of sample size two only, it is shown that the estimate of variance of the "unordered estimator" is always positive. Mickey (1954, 1959) independently while dealing mainly with unbiased ratio and regression type estimators has developed exactly the same estimators considered by Des Raj and Murthy. Mickey's efficiency comparisons between these estimators and Horvitz and Thompson's estimator $\hat{Y}$ of subclass 2 indicate approximate equality of efficiency.

Returning now to the discussion of methods that ensure the conditions (2.10), namely, the probabilities P_i proportional to the x_i, and the evaluation of $P_{ii'}$ and $V(\hat{Y})$ therefrom, Horvitz and Thompson (1952) suggest two methods that satisfy (2.10) approximately. The first method uses Midzuno's procedure for which

$$P_i = \frac{N - n}{N - 1} p_i^* + \frac{n - 1}{N - 1} \qquad (2.13)$$

and

$$P_{ii'} = \frac{n - 1}{N - 1} \left[\frac{N - n}{N - 2} (p_i^* + p_{i'}^*) + \frac{n - 2}{N - 2}\right] \qquad (2.14)$$

where p_i^* are the revised probabilities such that $P_i = np_i$. Solving (2.13) for p_i^*,

$$p_i^* = \frac{N - 1}{N - n} (np_i) - \frac{n - 1}{N - n} . \qquad (2.15)$$

However, this method is severely restricted since for $p_i < \frac{(n - 1)}{(N - 1)n}$, p_i^* becomes negative. Also, since only one unit is drawn with p.p.s. and the remaining (n - 1) units are drawn with equal probabilities, this method may not be as efficient as a procedure where all the n units are selected with unequal probabilities. The second method for sample size two is based on the assumption that sampling without replacement is not much different from sampling with replacement. Then the p_i^* are obtained by solving the quadratic

$$p_i^{*2} - p_i^* + p_i = 0 . \qquad (2.16)$$

Moreover, this method breaks down if p_i is greater than 0.25 since roots of (2.16) become imaginary.

Yates and Grundy (1953) have suggested a more satisfactory procedure of obtaining revised probabilities, based on iteration using Horvitz and Thompson's procedure of selecting the first unit with p.p.s., the second unit with p.p.s. of the remaining units and so on. Though the iteration process is applicable for any sample size, it becomes extremely cumbersome when sample size is greater than two. For sample size two,

$$P_i = p_i^* + p_i^* \sum_{j=i}^{N} \frac{p_j^*}{1 - p_j^*} \qquad (2.16)$$

and

$$P_{ii'} = p_i^* p_{i'}^* \left(\frac{1}{1 - p_i^*} + \frac{1}{1 - p_{i'}^*}\right) \qquad (2.17)$$

where the p_i^* are such that $P_i = 2p_i$. The p_i^* are obtained from (2.16) by iteration, and the authors think that one iteration should be adequate in most of the populations normally encountered. However, this procedure becomes cumbersome when N is fairly large. Narain (1951) suggests a graphical numerical method for solving (2.16) which is also rather complicated.

Des Raj (1956b) argues that though the above procedures satisfy the conditions (2.10) approximately, the $P_{ii'}$ so obtained may not be optimum. He therefore employs conditions (2.10) as a set of N equations for the $\frac{1}{2}N(N - 1)$ probabilities $P_{ii'}$ and determines the optimum $P_{ii'}$ by minimizing the variance of $\hat{Y}$ given by (2.9) subject to (2.10). This leads to a "linear programming problem" for the $\frac{1}{2}N(N - 1)$ positive $P_{ii'}$ satisfying (2.10). Since the "objective function" (the variance) involves the unknown y_i, these are replaced by the known x_i assuming that

$$y_i = A + Bx_i \qquad (2.18)$$

exactly. There are several limitations of this method. Computations become extremely cumbersome when n is greater than two and/or for large N. Also, as illustrated by Des Raj himself, the method is quite sensitive to the assumption of linear model, and if the model is not satisfied considerable loss in efficiency can result by using these optimum probabilities. Moreover, if it is assumed that the y_i of the population satisfy the linear model (2.18) exactly with unknown A and B, then, clearly the regression estimator has zero variance and even if an error term is introduced into this linear model the regression type estimator would still be the "best" estimator so that it is of little interest to consider other estimators under such assumptions. It may be noted that Des Raj's procedure remains unchanged even if an error term e_i with

$$E(e_i|x) = 0 \qquad \text{and} \qquad Cov(e_i e_j|x) = 0 \qquad i \neq j \tag{2.19}$$

is introduced in the model (2.18), provided the "objective function" is not the variance of $\hat{Y}$ but is the expectation of the variance of $\hat{Y}$ under the assumptions (2.19).

Instead of finding the revised probabilities p_i^* which ensure that conditions (2.10) are satisfied, one would like to have a sampling procedure with the original probabilities p_i for which conditions (2.10) are satisfied. There is a simple sampling procedure well known to survey practitioners having this property, and is mentioned for example in Goodman and Kish (1950). In this procedure, the N units in the population are listed in a random order and their measures of size are cumulated and a systematic selection of n elements from a random atart is then made on the cumulation so that conditions (2.10) are satisfied exactly. Horvitz and Thompson (1952) mention this procedure but say "This selection is easily performed, but there does not appear to be any simple way to determine the probabilities $P_{ii'}$."

In this dissertation, we propose to determine the probabilities $P_{ii'}$ for this sampling procedure explicitly in terms of the P_i. In Chapter III, expressions for $P_{ii'}$ will be given for the cases $n = 2$ and $N = 3$, 4 and 5. As N becomes large, the exact evaluation of $P_{ii'}$ becomes cumbersome, so we shall develop an asymptotic theory in Chapter IV for the case $n = 2$ and in Chapter V for the case of general sample size n. Compact expressions for the probabilities $P_{ii'}$ and the variance of $\hat{Y}$ will be obtained applicable to large and medium sized populations. An important feature of this sampling procedure is that it lends itself to the case of general sample size n unlike the procedures previously mentioned. For example, expressions for P_i and $P_{ii'}$ for Horvitz and Thompson procedure of drawing first unit with p.p.s., second unit with p.p.s. of the remaining units and so on, become unwieldy and not manageable. The only procedure which seems to give simple expressions is Midzuno's procedure of drawing the first unit with p.p.s. and the remaining (n - 1) units with equal probabilities and without replacement. Sen (1955) has proposed a method to deal with the case $n > 2$. Assuming n is a multiple of 2, he suggests to draw the first two units by Horvitz and Thompson procedure, replace the two units, and then draw the next two units by the same procedure and so on. This procedure gives simple expressions for P_i and $P_{ii'}$. However, since each pair of units is replaced before the next pair is drawn, there will be an overlap of units and so this procedure is not as precise as selecting all the n units without replacement. In Chapter V, section D, we prove an interesting result showing that the $P_{ii'}$ values attained through Yates and Grundy iteration procedure and through the sampling procedure mentioned by Goodman and Kish as described before, are exactly the same to order $O(N^{-3})$ so that $V(\hat{Y})$ is the same for both the procedures to order $O(N^{1})$, assuming that P_i is order $O(N^{-1})$ which indicates that both procedures have practically the same efficiency for large N.

Since the strict application of available methods of unequal probability sampling without replacement involves considerable computations, some authors on grounds of practicability have suggested certain methods which retain the advantage of unequal probability sampling without replacement but easier to apply in practice and involve a slight loss of exactness. Yates (1949) suggests using the variance in unequal probability sampling with replacement with the usual finite population correction factor for simple random sampling attached to it, as an approximation for the variance in unequal probability sampling without replacement. Yates and

Grundy (1953) assuming that variation in the quantities y_i/P_i is of random nature unassociated with the P_i, obtain the following simple expression for the variance of $\hat{Y}$ from (2.9) using the relation

$$\sum_{i \neq i'}^{N} P_{ii'} = n(n-1):$$

$$V_{Appr.}(\hat{Y}) = n(1 - n^{-1} \sum^{N} P_i^2) \; V(\frac{y_i}{p_i}) \qquad (2.20)$$

where $V(y_i/p_i)$ is the variance of the quantities y_i/p_i. Durbin (1953) has suggested two approximate methods to obtain simple expressions for the estimate of the variance of $\hat{Y}$.

Stevens (1958) has a method of sampling without replacement if the values of x are or can be grouped into groups of units having the same measure of size, x. Then, the procedure is to select n groups with replacement and with probabilities proportional to total size of the groups, e.g. if in the ith group there are N_i units each of size x_i, then the total size of the group is $N_i x_i$. If the group i is chosen t_i times, select without replacement t_i units with equal probability and without replacement from this group. Stevens derives formulas for the variance etc. at length using this procedure. It is of interest to note that these formulas can be obtained as particular cases from a well known two stage sampling procedure (Sukhatme, 1954) in which the first stage units are selected with p.p.s. and with replacement and if the ith first stage unit is selected t_i times, $m_i t_i$ secondaries are selected with equal probability and without replacement from it. To obtain Steven's results, one simply has to identify the groups as first stage units, the units in a group or second stage units and put $m_i = 1$ in Sukhatme's formulas.

There are several other interesting problems in unequal probability sampling without replacement. It is of interest to estimate the variance in simple random sampling from a sample drawn with unequal probabilities in order to estimate the gain in efficiency of unequal probability sampling over simple random sampling. In most of the sample surveys we are usually interested in estimating the means or totals of several characteristics. If the sample is selected with p.p.s. of x, it may often happen that x is not highly correlated with all the characteristics of interest. For some of the characteristics y the correlation between y and x may be quite small so that using the usual estimators in unequal probability sampling may give large variances for the estimates of these characteristics. In such circumstances, one would like to save the situation with the help of alternative estimators that have smaller variances. Another important problem is the estimation of the gain in efficiency due to stratification for unequal probability sampling without replacement. Efficiency of stratification has been considered by Cochran (1953) for simple random sampling and by Sukhatme

(1954) for unequal probability sampling with replacement. In Chapter VI, sections B and C, we consider these problems.

It is of importance to make efficiency comparisons between unequal probability sampling and other methods of utilizing supplementary information, e.g. ratio and regression methods of estimation, stratification. Since in practice, no functional form of the distribution followed by the data is assumed, it is difficult to make meaningful comparisons. So, Cochran (1953) assuming the model

$$y_i = Yp_i + e_i \tag{2.21}$$

with

$$E(e_i | x) = 0 \quad \text{and} \quad E(e_i^2 | x) = ap_i^g, \quad g > 0, \ a > 0 \tag{2.22}$$

has shown that the variance in p.p.s. sampling with replacement is smaller than the variance of the ratio estimate $\hat{Y}_R$ (for large samples) without the finite population correction factor, if $g > 1$. It is also remarked that in practice g usually lies between 1 and 2 so that the p.p.s. estimate is generally more precise. Also, it is noted that if it costs more to obtain data from a larger unit than from a smaller one, the comparison is biased in favor of p.p.s. sampling, which tends to concentrate on the larger units. Said (1955) has made extensive investigations on efficiency comparisons between unequal probability sampling, ratio and regression methods of estimation and stratification, under certain specific relationships between y and x and assuming x has a

Pearson's Type III distribution. It is hard to know how good these assumptions are in practice, and so we think that these results have limited use. Des Raj (1958) has suggested using Cochran's idea of regarding the finite population as drawn at random from an infinite super-population with certain properties, so that the results obtained apply to the average of all finite populations that can be drawn from the infinite population. He makes certain efficiency comparison using this concept. Zarkovic (1960) expands the variance in p.p.s. sampling with replacement by Taylor's expansion neglecting terms with powers higher than second, and compares it with the variance of ratio and regression estimates. Since we obtain compact expressions for the variance of $\hat{Y}$ in unequal probability sampling without replacement, we shall make comparisons in Chapter V, section E, with the variance of the ratio estimate with the finite population correction factor included.

Finally, mention should be made of the criticism on the logic of unequal probability sampling. It is worth quoting Weibull (1960, p. 84) in this connection. He says:

> The method of sampling with varying probabilities in sample survey theory is based on a criterion of minimizing the expected variance, a criterion which is not appropriate when only a single sample is drawn. The supposed reduction of the variance in the estimates is illusory and has no real significance. Intutively this is fairly clear. If it is known that some units contain more information - or from other points of view are more desirable to sample - than some other units, there is no reason

> to let the actual selection depend on a random procedure.

If this implies that units with high weight should be sampled and units with low weight ignored, then obviously no valid estimate of the variance can be found. However, these sentiments can be incorporated in a probability design with stratification and sampling with unequal probabilities within each or some of the strata. Such a design is described in Chapter VI, section D.

III. A SIMPLE PROCEDURE OF UNEQUAL PROBABILITY SAMPLING WITHOUT REPLACEMENT

The problem is to draw a sample of n units without replacement from a finite population of N units such that the probability P_i for the i^{th} unit to be in the sample is proportional to $p_i = x_i/X$ and $\sum^N p_i = 1$, i.e.

$$P_i = \text{Pr. } (i^{th} \text{ unit in the sample}) = cp_i \qquad (3.1)$$

where c is a constant. We now prove the following theorem:

Theorem 3.1. If there is a sampling procedure which satisfies equation (3.1), then $c = n$ and $np_i \leq 1$.

Proof. Let a_i denote the "indicator variable" such that

$$a_i = \begin{cases} 1 & \text{if } i^{th} \text{ unit is in the sample} \\ 0 & \text{if } i^{th} \text{ unit is not in the sample.} \end{cases} \qquad (3.2)$$

Then

$$E(a_i) = 1 \cdot \text{Pr.}(a_i = 1) = P_i = cp_i \ . \qquad (3.3)$$

Since the n units in the sample are drawn without replacement, exactly n of the a_i take the value 1 and the remaining $(N - n)$ of the a_i take the value 0 so that

$$\sum^N a_i = n \ . \qquad (3.4)$$

Taking expectations of (3.4) and using (3.3) we find

$$n = \sum^N E(a_i) = c \sum^N p_i = c \qquad (3.5)$$

so that $c = n$ and since the probabilities P_i cannot be greater than 1, it immediately follows that

$$P_i = np_i \leq 1 \ . \qquad (3.6)$$

We shall now describe the sampling procedure adopted in this dissertation which has been mentioned by Goodman and Kish (1950), and which satisfies (3.6). So, to apply this sampling procedure we have to confine to those p_i for which $np_i \leq 1$. If the p_i for some of the units in the population do not satisfy this condition, one can include these units automatically in the sample or subdivide each of these units into two or more subunits such that the p_i corresponding to the subunits satisfy this conditions.

A. Description and Illustration of the Sampling Procedure

The sampling procedure can be described in two steps as follows:

Step 1. Arrange the N units in a random order and denote (without loss of generality) by $j = 1, 2, \ldots, N$ this random order, and by

$$\pi_j = \sum_{i=1}^{j} (np_i) \ , \quad \pi_0 = 0 \qquad (3.7)$$

the cumulative totals of the np_i in that order.

Step 2. Select a "random start", i.e. select a "uniform variate" d with $0 \leq d < 1$. Then the n selected units are those whose index j satisfies

$$\pi_{j-1} \leq d + k < \pi_j \qquad (3.8)$$

for some integer k between 0 and $(n - 1)$. Since each $np_i \leq 1$, every one of the n integers $k = 0, 1, \ldots, (n-1)$ will select a different unit j.

Though it is known that (3.6) is satisfied by this sampling procedure, no formal proof seems to have been given in the literature. Theorem 3.2 below gives a proof to this effect.

Theorem 3.2. For the above sampling procedure the probability of selecting the j^{th} unit in the sample, P_j, is equal to np_j.

Proof. Consider a particular arrangement of the N units in an ordered sequence and single out a particular unit j in that sequence. Let I denote the largest integer with $I \leq \pi_{j-1}$. Now if $\pi_j - I \leq 1$, from (3.8) it immediately follows that unit j is selected if $\pi_{j-1} - I \leq d < \pi_j - I$ for $k = I$. If, on the other hand, $\pi_j - I > 1$, the unit j is selected if $\pi_{j-1} - I \leq d < 1$ for $k = I$ or if $0 \leq d < \pi_j - I - 1$ for $k = I + 1$. Since d is a uniform variate we see that in

Case 1: $\pi_j - I \leq 1$.

$$P_j = Pr.(\pi_{j-1} - I \leq d < \pi_j - I) = \pi_j - \pi_{j-1} = np_j \quad (3.9)$$

and in

Case 2: $\pi_j - I > 1$

$$\begin{aligned} P_j &= Pr.(\pi_{j-1} - I \leq d < 1) + Pr.(0 \leq d < \pi_j - I - 1) \\ &= (1 - \pi_{j-1} + I) + (\pi_j - I - 1) = np_j . \end{aligned} \quad (3.10)$$

Therefore in either case we have $P_j = np_j$. It may be noted that the randomization of the N units in step 1 is not necessary to prove Theorem 3.2. However, this is required to obtain compact variance formulas for the estimate $\hat{Y}$ using an asymptotic theory as will be evident in Chapters IV and V.

1. A cyclical analogue to the sampling procedure

We consider a cyclical analogue to the sampling procedure which is more convenient to use from the point of view of mathematical treatment and is stochastically equivalent to the original sampling procedure. Steps 1 and 2 are modified as follows:

Step 1'. Arrange the N units in a random order, denote by $j = 1, 2, \ldots, N$ this random order and form (as before) the cumulative totals π_j given by (3.7). Since $\sum^{N} (np_j) = n$, consider a circle with circumference of n or of radius $n/2\pi$ and then mark off on the perimeter of the circle arcs of lengths P_j in clockwise direction starting at the top.

Step 2'. Select a uniform arc s with $0 \leqslant s < n$. Then the n selected units are those whose indices j satisfy

$$\pi_{j-1} \leqslant s + k < \pi_j \qquad (3.11)$$

for some integer k between $-(n - 1)$ and $(n - 1)$. Only n of the $(2n - 1)$ integers k will actually select the n different units. Theorem 3.2 holds here because we know with certainty that one of the arcs π_j will fall within the range 0 to 1 and this may be identified with the variate d in step 2.

2. Illustration of the sampling procedure

To demonstrate the actual method of selecting the units by the present sampling procedure, we take the population of 20 blocks in Ames, Iowa, considered by Horvitz and Thompson (1952). We have chosen this example here because we will be making efficiency comparisons later in Chapter IV, section D, using the same data. The variate y denotes the number of households on a block and the variate x denotes the eye estimated number of households on a block. The data are given below in Table 1 and the population totals are Y = 434 and X = 394. It is not necessary to compute the quantities $p_i = x_i/X$ and $P_i = np_i$ in order to use the sampling procedure, since by scaling all computations up by the factor X/n we have to compute only the cumulative totals of x_i instead of the cumulative totals of P_i. Then select a random integer (start) between 1 and X/n say D and use (3.3) as

$$\sum_{i=1}^{j-1} x_i \leq D + \frac{X}{n} \cdot k < \sum_{i=1}^{j} x_i \qquad (3.12)$$

to select the n units.

Suppose a sample of size n = 3 units is to be drawn and suppose the random number D between 1 and X/n = 394/3 = 131 (approx.) is 45. Then, we must find the lines (j) where the column $\sum_{i=1}^{j} x_i$ passes through the levels D = 45 (for k = 0),

Table 1. Selection of n = 3 units from a population of N = 20 units

Block	No. of households	Eye estimated no. of households	Cumulative sum	Start = 45
				Step = X/n = 131

j	y_j	x_j	$\sum_{i=1}^{j} x_i$	
1	19	18	18	
2	9	9	27	
3	17	14	41	
4	14	12	53	k = 0, D = 45
5	21	24	77	
6	22	25	102	
7	27	23	125	
8	35	24	149	
9	20	17	166	
10	15	14	180	k=1, D+131=176
11	18	18	198	
12	37	40	238	
13	12	12	250	
14	47	30	280	
15	27	27	307	k=2, D+262=307
16	25	26	333	
17	25	21	354	
18	13	9	363	
19	19	19	382	
20	12	12	394	
Total	434	394		

D + 131 = 176 (for k = 1) and D + 262 = 307 (for k = 2). From Table 1, it is seen that the units j = 4, 10 and 16 are selected in the sample.

B. Variance Formulas for the Cases $n = 2$, $N = 3$, 4 and 5

To find the variance of $\hat{Y}$ in terms of P_j and y_j, one has to evaluate $P_{ii'}$ explicitly in terms of P_j and then substitute in (2.9), namely,

$$V(\hat{Y}) = \sum^{N} \frac{y_j^2}{P_j} + 2 \sum_{i<i'}^{N} \frac{P_{ii'}}{P_i P_{i'}} \cdot y_i y_{i'} - Y^2 . \qquad (3.13)$$

To find an estimate of the variance of $\hat{Y}$, we substitute the value of $P_{ii'}$ in the Yates and Grundy estimate of the variance, namely,

$$v_{YG}(\hat{Y}) = \sum_{i<i'}^{n} \frac{P_i P_{i'} - P_{ii'}}{P_{ii'}} \left(\frac{y_i}{P_i} - \frac{y_{i'}}{P_{i'}}\right)^2 . \qquad (3.14)$$

1. <u>The case $n = 2$, $N = 3$</u>

Since there are only three units in the population,

$$P_{ii'} = 1 - \text{Pr. (i" in the sample)} \qquad (3.15)$$

where i" is the remaining unit in the population. Thus,

$$P_{ii'} = 1 - P_{i''} = P_i + P_{i'} - 1 \qquad (3.16)$$

since

$$P_i + P_{i'} + P_{i''} = 2 . \qquad (3.17)$$

From (3.16) it follows that $P_{ii'} > 0$ except in the obvious case $P_{i''} = 1$. Substituting $P_{ii'}$ from (3.16) in (3.13), we find

$$V(\hat{Y}) = \sum_{i<i'}^{3} (1 - P_i)(1 - P_{i'})\left(\frac{y_i}{P_i} - \frac{y_{i'}}{P_{i'}}\right)^2 . \qquad (3.18)$$

Similarly from (3.16) and (3.14) we obtain

$$v_{YG}(\hat{Y}) = \frac{(1 - P_i)(1 - P_{i'})}{P_i + P_{i'} - 1}\left(\frac{y_i}{P_i} - \frac{y_{i'}}{P_{i'}}\right)^2 \qquad (3.19)$$

which is nonnegative since $P_i + P_{i'} \geqslant 1$ and $P_j \leqslant 1$. It is interesting to note that (3.16) is true for the more general case $n = N - 1$, $N = N$, since

$$P_{ii'} = \sum_{j \neq i,i'}^{N} \left[1 - \text{Pr.}(j \text{ in the sample})\right]$$

$$= (N - 2) - \sum_{j \neq i,i'}^{N} P_j$$

$$= (N - 2) - \ (N - 1) - P_i - P_{i'} \ = P_i + P_{i'} - 1 .$$

In fact, in this special case, it is easy to evaluate $P_{ij\ldots m}$, the probability of including r units i, j, ..., m, since

$$P_{ij\ldots m} = \sum_{s \neq (i,j,\ldots,m)}^{N} \left[1 - \text{Pr.}(s \text{ in the sample})\right]$$

$$= (N - r) - \left[(N - 1) - P_i - P_j \ \ldots - P_m\right]$$

$$= P_i + P_j + \ldots + P_m - r + 1 . \qquad (3.20)$$

However, this case may not be of much practical importance.

2. The case n = 2, N = 4

Without loss of generality, let us assume that

$$P_i \geqslant P_{i'} \quad \text{and} \quad P_{i'''} \geqslant P_{i''} \tag{3.21}$$

where i" and i"' denote the remaining two units in the population. In order to evaluate $P_{ii'}$ we have to distinguish the following two cases of the randomization results:

Case 1. The units i and i' are adjacent.

Case 2. The units i and i' are separated by one unit.

Now, for case 1 there are 16 possible configurations of the P_j on the circle and 8 possible configurations for case 2. The probability $P'_{ii'}$ that the units i and i' are the sampled units in case 1 for a typical configuration, say, first two arcs from the top correspond to P_i and $P_i + P_{i'}$ respectively, is

$$P'_{ii'} = \Pr.(0 \leqslant d < P_i;\ P_i \leqslant d + 1 < P_i + P_{i'})$$

$$= \begin{cases} P_i + P_{i'} - 1 & \text{if} \quad P_i + P_{i'} \geqslant 1 \\ \quad 0 & \text{if} \quad P_i + P_{i'} < 1 \end{cases} \tag{3.22}$$

where d is the uniform variate with $0 \leqslant d < 1$. All the remaining configurations have the same $P'_{ii'}$. The probability $P''_{ii'}$ that the units i and i' are the sampled units in case 2 for a typical configuration, say, first three arcs from the top correspond to P_i, $P_i + P_{i''}$ and $P_i + P_{i''} + P_{i'}$ respectively, is

$$P''_{ii'} = \Pr.(0 \leq d < P_i;\ P_i + P_{i''} \leq d + 1 < P_i + P_{i''} + P_{i'})$$

$$= \begin{cases} P_i + P_{i'} + P_{i''} - 1 & \text{if} \quad P_i + P_{i''} \leq 1 \\ P_{i'} & \text{if} \quad P_i + P_{i''} > 1 \end{cases} \tag{3.23}$$

using conditions (3.21). All the remaining configurations have the same $P''_{ii'}$. Therefore the overall probability $P_{ii'}$ is given by

$$P_{ii'} = \frac{16}{24} P'_{ii'} + \frac{8}{24} P''_{ii'}$$

$$= \frac{2}{3} P'_{ii'} + \frac{1}{3} P''_{ii'} \tag{3.24}$$

where $P'_{ii'}$ and $P''_{ii'}$ are given by (3.22) and (3.23) respectively.

The substitution of $P_{ii'}$ from (3.24) in (3.13) yields the variance of $\hat{Y}$. It may be noted that $P_{ii'} > 0$ except in the obvious case $P_{i'''} = 1$. However, if the P_j are arranged systematically, $P_{ii'}$ can be zero even if $P_{i'''} < 1$.

3. The case n = 2, N = 5

Let the numbering of the units before randomization be 1, 2, 3, 4 and 5 and let $i = 1$ and $i' = 2$ and $P_1 \geq P_2$ without loss of generality. Again we distinguish the two cases:

Case 1. The units 1 and 2 are adjacent.

Case 2. The units 1 and 2 are separated by one unit.

There are 60 possible configurations for case 1 and 60 for case 2. The probability P'_{12} that the units 1 and 2 are the sampled units in case 1 for a typical configuration is

$$P'_{12} = \begin{cases} P_1 + P_2 - 1 & \text{if} \quad P_1 + P_2 \geqslant 1 \\ 0 & \text{if} \quad P_1 + P_2 < 1 . \end{cases} \tag{3.25}$$

All the remaining configurations have the same P'_{12}. Now in case 2, we have to distinguish the following three sub-cases each with 20 possible configurations, in order to evaluate the probability that the units 1 and 2 are the sampled units:

<u>Case (2a)</u>. P_4 and P_5 are adjacent and separated from P_3 by P_1 and P_2.

<u>Case (2b)</u>. P_3 and P_5 are adjacent and separated from P_4 by P_1 and P_2.

<u>Case (2c)</u>. P_3 and P_4 are adjacent and separated from P_5 by P_1 and P_2.

In case (2a) if $P_3 \leqslant P_4 + P_5$, the probability $P''_{12}(a)$ that the units 1 and 2 are the sampled units for a typical configuration is

$$P''_{12}(a) = \begin{cases} 0 & \text{if} \quad P_1 + P_2 + P_3 < 1 \\ P_1 + P_2 + P_3 - 1 & \text{if} \quad P_1 + P_3 \leqslant 1 \quad \text{and} \\ & \phantom{\text{if}} \quad P_1 + P_2 + P_3 \geqslant 1 \\ P_2 & \text{if} \quad P_1 + P_3 > 1 . \end{cases} \tag{3.26}$$

However, if $P_3 \geqslant P_4 + P_5$ then

$$P''_{12}(a) = \begin{cases} P_1 + P_2 + P_4 + P_5 - 1 & \text{if} \quad P_1 + P_4 + P_5 \leqslant 1 \\ P_2 & \text{if} \quad P_1 + P_4 + P_5 > 1 . \end{cases} \tag{3.27}$$

All the remaining configurations in case (2a) have the same $P''_{12}(a)$. Expressions analogous to (3.26) and (3.27) hold for

$P''_{12}(b)$ and $P''_{12}(c)$. Therefore the overall probability P_{12} is

$$P_{12} = \frac{60}{120} P'_{12} + \frac{20}{120} P''_{12}(a) + \frac{20}{120} P''_{12}(b) + \frac{20}{120} P''_{12}(c)$$

$$= \frac{1}{2} P'_{12} + \frac{1}{6} P''_{12}(a) + \frac{1}{6} P''_{12}(b) + \frac{1}{6} P''_{12}(c) \ . \qquad (3.28)$$

Again, it is obvious that $P_{12} = 0$ if $P_3 = 1$ or $P_4 = 1$ or $P_5 = 1$, but in this case it is interesting to note that P_{12} can also be zero if with all $P_j < 1$ the following conditions are satisfied:

$$P_1 + P_2 + P_t < 1 \ (t = 3, 4 \text{ and } 5) \ . \qquad (3.29)$$

This contradicts a statement made by Thompson (1952), p. 58, to the effect that $P_{12} > 0$ if all $P_i < 1$ and randomization is used. The following example illustrates the computations and shows that $P_{12} = 0$.

It is now evident that the exact evaluation of $P_{ii'}$ becomes cumbersome as N increases, and in any case the resulting formulas are too complicated to yield a compact formula for $V(\hat{Y})$. Therefore, an asymptotic theory for the present sampling procedure is developed in Chapters IV and V which yields compact formulas for $V(\hat{Y})$ applicable to moderately large populations.

4. Example

Let $P_1 = 0.20$, $P_2 = 0.20$, $P_3 = 0.55$, $P_4 = 0.55$ and $P_5 = 0.50$ so that $\sum^{5} P_j = 2$ and (3.29) are satisfied.

Therefore $P'_{12} = 0$ and $P''_{12}(a) = P''_{12}(b) = P''_{12}(c) = 0$ and $P_{12} = 0$. Let us illustrate the computation of P_{13} where $P_3 = 0.55 > P_1 = 0.2$. Now

$$P'_{13} = 0 \quad \text{since} \quad P_1 + P_3 = 0.75 < 1$$

$$P''_{13}(a) = 0 \quad \text{since} \quad P_4 + P_5 = 0.95 > P_2 = 0.20$$
$$\text{and} \quad P_1 + P_3 + P_2 = 0.95 < 1$$

$$P''_{13}(b) = P_1 = 0.20 \quad \text{since} \quad P_2 + P_5 = 0.70 > P_4 = 0.55$$
$$\text{and} \quad P_3 + P_4 = 1.10 > 1$$

$$P''_{13}(c) = P_1 = 0.20 \quad \text{since} \quad P_2 + P_4 = 0.75 > P_5 = 0.50$$
$$\text{and} \quad P_3 + P_5 = 1.05 > 1 .$$

Therefore

$$P_{13} = \frac{1}{2}(0) + \frac{1}{6}(0) + \frac{1}{6}(0.20) + \frac{1}{6}(0.20) = \frac{0.40}{6} .$$

Similar calculations lead to the following table of $P_{ii'}$ values. A check is provided on the calculations by noting

Table 2. $P_{ii'}$ values for the above example

i i'	1	2	3	4	5	Total = P_i
1	--	0	$\frac{0.40}{6}$	$\frac{0.40}{6}$	$\frac{0.40}{6}$	0.20
2	0	--	$\frac{0.40}{6}$	$\frac{0.40}{6}$	$\frac{0.40}{6}$	0.20
3	$\frac{0.40}{6}$	$\frac{0.40}{6}$	--	$\frac{1.40}{6}$	$\frac{1.10}{6}$	0.55
4	$\frac{0.40}{6}$	$\frac{0.40}{6}$	$\frac{1.40}{6}$	--	$\frac{1.10}{6}$	0.55
5	$\frac{0.40}{6}$	$\frac{0.40}{6}$	$\frac{1.10}{6}$	$\frac{1.10}{6}$	--	0.50

that the marginal totals P_i in Table 2 agree with the given values of P_i.

C. An Example for Efficiency Comparisons

To compare the efficiency of the present sampling procedure with both the procedures of Yates and Grundy of finding the revised probabilities and that of Des Raj (1956b) which consists of finding the optimum $P_{ii'}$ under the assumption of a linear model, we consider the case n = 2, N = 4 and use the three populations examined by these authors. Yates and Grundy who introduce these data for purposes of illustration state that these populations have been deliberately chosen to represent situations more extreme than those normally encountered in practice. The three populations (all of size N = 4) have the same set of four p_j values with different sets of y_j values attached to them and are given in Table 3 below.

Table 3. Three populations of size N = 4

Unit number	p_j	Population A y_j	Population B y_j	Population C y_j
1	0.1	0.5	0.8	0.2
2	0.2	1.2	1.4	0.6
3	0.3	2.1	1.8	0.9
4	0.4	3.2	2.0	0.8

Table 4 below gives the values of $P_{ii'}$ for the above three sampling procedures. Tables 4.1 and 4.2 are taken from Des Raj and Table 4.3 is computed using (3.24).

The variance of $\hat{Y}$ for the three sampling procedures and the three populations are given in Table 5 below using the $P_{ii'}$ values of Table 4 and equation (3.13), the formula for

Table 4. Values of $P_{ii'}$ for populations in Table 3

i	i'	1	2	3	4
		4.1. Yates and Grundy procedure			
1		--	0.032	0.059	0.113
2		0.032	--	0.122	0.246
3		0.059	0.122	--	0.428
4		0.113	0.246	0.428	--
		4.2 Des Raj optimum procedure			
1		--	0.0	0.0	0.2
2		0.0	--	0.2	0.2
3		0.0	0.2	--	0.4
4		0.0	0.2	0.4	--
		4.3 Present procedure			
1		--	0.067	0.067	0.067
2		0.067	--	0.067	0.267
3		0.067	0.067	--	0.467
4		0.067	0.267	0.467	--

Table 5. Comparative efficiency of four sampling procedures

Procedure	Population A		Population B		Population C	
	Var.	Eff.%	Var.	Eff.%	Var.	Eff.%
1. Des Raj	0.200	100.0	0.200	100.0	0.100	100.0
2. Yates and Grundy	0.329	61.9	0.269	74.3	0.057	175.4
3. Present procedure	0.367	54.5	0.367	54.5	0.033	333.3
4. With replacement	0.500	40.0	0.500	40.0	0.125	80.0

$V(\hat{Y})$. Moreover the values of the variance of $\hat{Y}'$ for sampling with replacement using equation (2.6) are shown in Table 5 for comparison.

For populations A and B, the linear model assumption seems to be fairly well satisfied since from Table 5 it is seen that Des Raj optimum procedure yields the smallest variance. For population C, the model does not seem to be appropriate since it is seen that considerable loss in efficiency results for Des Raj procedure. Also it is seen from Table 5 that the variances of Yates and Grundy procedure and the present procedure are approximately of the same size. In fact, in Chapter IV, section E, it is proved that Yates and Grundy procedure and the present procedure have the same asymptotic efficiency, i.e. the formulas for $V(\hat{Y})$ agree to order N^{1}. For the present (artificial) populations these results for "large N" do not, of course, apply. However, these asymptotic results are illustrated in a later example of a population of size N = 20, in Chapter IV, section D.

IV. THE CASE $n = 2$ AND N LARGE

The difference between sampling with and without replacement gradually disappears as N tends to infinity, so that the expected gain in precision through sampling without replacement will become negligible. Now, for sampling with replacement with probabilities p_i, we have from the properties of the multinomial distribution

$$P_{ii'} = n(n - 1)p_i p_{i'} = \frac{(n - 1)}{n} P_i P_{i'} \qquad (4.1)$$

with $P_i = np_i$, so that if P_i is assumed to be of order $O(N^{-1})$, $P_{ii'}$ will be of order $O(N^{-2})$. In sampling without replacement this will be the leading term, and hence in order to supply formulas for moderately large populations N, we have to evaluate the next lower order terms, namely terms of $O(N^{-3})$. These terms will represent the gain in precision due to the so called finite population correction. The variance of the estimate $\hat{Y}$ for sampling with replacement is of $O(N^2)$, and so in sampling without replacement, the next lower order terms $O(N^1)$ which represent the reduction in variance accomplished by sampling without replacement, have to be evaluated. This is equivalent to evaluating $P_{ii'}$ to $O(N^{-3})$ and substituting it in the variance formula for $\hat{Y}$. So, we evaluate here for our sampling procedure, $P_{ii'}$ to $O(N^{-3})$ and hence $V(\hat{Y})$ to $O(N^1)$, assuming P_i is $O(N^{-1})$. Also, for the benefit of smaller size populations, we evaluate here, $P_{ii'}$ to $O(N^{-4})$

and hence $V(\hat{Y})$ to $O(N^0)$.

As pointed out earlier, the present sampling procedure lends itself for the sample size $n > 2$ unlike the procedures previously published. We discuss the case $n = 2$ in this chapter in detail, and consider the case $n > 2$ in the next chapter. The methods of attack for the case $n > 2$ are similar to those for the case $n = 2$. However, the case $n > 2$ presents certain new features other than those encountered for the case $n = 2$.

A. Derivation of the Probabilities $P_{ii'}$ to Orders $O(N^{-3})$ and $O(N^{-4})$

The total number of arrangements of the N units on the circle, namely N!, can be divided into (N - 1) groups according as to whether there are v = 0, 1, ..., (N - 2) units "between" P_i and $P_{i'}$, where "between" means that there are v units when proceeding from P_i to $P_{i'}$ in clockwise direction. There are N x (N - 2)! arrangements in each of these (N - 1) groups so that the probability for each of these arrangements is the same and is equal to N x (N - 2)!/N! = 1/(N - 1). Let us consider now the contribution to $P_{ii'}$ from a <u>particular</u> group with v units between P_i and $P_{i'}$. For the unit i to be in the sample, we know from our sampling procedure, the inequalities

$$\pi_{i-1} \leq s + \underline{x} < \pi_i \qquad (4.2)$$

must be satisfied where k may be any integer between -1 and 1 and s is a uniform arc with $0 \leq s < 2$. This means that s must lie within one of the following ranges each of length P_i. The first range is $\pi_{i-1} \leq s < \pi_i$ and the other range is displaced from the above range by a unit arc, i.e., $\pi_{i-1} - 1 \leq s < \pi_i - 1$ if $\pi_{i-1} \geqslant 1$ and $\pi_{i-1} + 1 \leq s < \pi_i + 1$ if $\pi_i \leq 1$. So, to evaluate $P_{ii'}$ we have to add the contributions to $P_{ii'}$ from the first range, say $P'_{ii'}$, and from the second range, say $P''_{ii'}$. These two ranges give identical contributions to $P_{ii'}$ since in both cases the length of the range for s is equal to P_i.

Let us consider now the evaluation of $P'_{ii'}$. Since the uniform variate s lies inside the range

$$\pi_{i-1} \leq s < \pi_i \qquad (4.3)$$

a positive contribution to $P'_{ii'}$ can be made only if the variate s + 1 also lies on the arc covered by $P_{i'}$. This means that if we denote by T_v to total length of the v arcs P_j which lie "between" the arcs P_i and $P_{i'}$, the inequalities

$$\pi_i + T_v \leq s + 1 < \pi_i + T_v + P_{i'} \qquad (4.4)$$

or

$$1 + t - P_i - P_{i'} < T_v \leq 1 + t - P_i \qquad (4.5)$$

where

$$t = s - \pi_{i-1} = s + P_i - \pi_i \qquad (4.6)$$

must be satisfied. Since the uniform variate t lies inside the range

$$0 \leq t < P_i \tag{4.7}$$

and has an ordinate density of 1/2 like the variate s, the integrated contribution to $P'_{ii'}$ is given by

$$\int_0^{P_i} \frac{1}{2} \Pr.(1 + t - P_i - P_{i'} < T_v \leq 1 + t - P_i)dt$$

$$= \frac{1}{2} \int_0^{P_i} \left[F_v(1 + t - P_i) - F_v(1 + t - P_i - P_{i'})\right] dt \tag{4.8}$$

where $F_v(T)$ denotes the cumulative distribution function of the total (T_v) of v values of the P_j. Since the units are randomized prior to drawing the sample, T_v represents the total of v values of the P_j sampled without replacement and equal probability from the finite population of $(N - 2)$ arcs P_j excluding the specific pair P_i and $P_{i'}$. Therefore, noting that $\sum_1^N P_j = 2$ we find that

$$E(T_v) = v(2 - P_i - P_{i'})/(N - 2)$$

$$\text{Var} \cdot (T_v) = v(1 - \frac{v}{N - 2})S^2_{ii'} \tag{4.9}$$

where

$$S^2_{ii'} = (N - 3)^{-1} \sum_{j \neq (i,i')}^{N} \left[P_j - \frac{(2 - P_i - P_{i'})}{(N - 2)}\right]^2$$

$$= (N - 3)^{-1} \left[\sum_1^N P_j^2 - P_i^2 - P_{i'}^2 - \frac{(2 - P_i - P_{i'})^2}{(N - 2)}\right]. \tag{4.10}$$

This important aspect of the randomization of the units prior to drawing the sample will now be used to develop an asymptotic theory for the evaluation of $P_{ii'}$.

Adding now the two (identical) contributions to $P'_{ii'}$ and $P''_{ii'}$ from (4.8) and summing over v we obtain

$$P_{ii'} = (N - 1)^{-1} \sum_{v=0}^{N-2} \int_0^{P_i} \Big[F_v(1 + t - P_i) - F_v(1 + t - P_i - P_{i'})\Big] dt \qquad (4.11)$$

where the factor $(N - 1)^{-1}$ represents the (constant) probability of a random arrangement of the N arcs P_j in which exactly v units lie "between" P_i and $P_{i'}$. It may be noted that the value of $P_{ii'}$ given by (4.11) is exact. We now find an approximation to (4.11) by expanding F_v in an Edgeworth series of which the cumulative normal integral is the leading term, in order to obtain usable results. In the literature, this problem of expressing a cumulative distribution function by an Edgeworth series is considered only for sampling without replacement from an infinite population (or for sampling with replacement from a finite population). However, the present problem involves sampling without replacement from a finite population. To deal with this, we make use of results in the literature on the moments of a sample total or mean in sampling without replacement and equal probability from a finite population.

Let i = 1 and i' = 2 without loss of generality. From the inversion theorem for the characteristic function of the cumulative distribution function F(x) of a statistical variate x we have (_e.g._ Kendall and Stuart (1958, p. 158)

$$F(x) = \exp.\left\{ \sum_{i=3}^{\infty} D^i \frac{k_i}{i!} (-1)^i \right\} P(x) \qquad (4.12)$$

where P(x) denotes the normal cumulative distribution

$$P(x) = (2\pi)^{-\frac{1}{2}} \int_{-\infty}^{x} \exp.(-\tfrac{1}{2} y^2)dy \ . \qquad (4.13)$$

D^i is the i^{th} order derivative w.r.k. x. and k_i are the standardized cumulants. In our case the formula (4.12) is applied to the standardized variate

$$z_v = \frac{T_v - v(2 - P_1 - P_2)(N - 2)^{-1}}{S_{12}\left[v(1 - \frac{v}{N - 2})\right]^{\frac{1}{2}}} \qquad (4.14)$$

in place of x so that F(x) is the finite proportion $F_v(z)$ say, of values z_v with $z_v \leq z$. This function is therefore a step function with a _finite_ number of discontinuities which do not affect the evaluation of (4.11). The r.h.s. of (4.12) is equal to $F_v(z)$ for almost all values of z_v whereas at the points of discontinuity the r.h.s. of (4.12) is equal to $Pr.(z_v < z) + \frac{1}{2} Pr.(z_v = z)$, _e.g._ Kendall and Stuart(1958), p. 97. We therefore have from (4.12),

$$F_v(z) = P(z) - \frac{k_3}{6} D^3 P(z) + R(v) \qquad (4.15)$$

where

$$R(v) = \exp\cdot\left\{\sum_{i=3}^{\infty} D^i \frac{k_i}{i!} (-1)^i\right\}P(z) - \left\{1 - \frac{k_3}{6} D^3\right\}P(z) \qquad (4.16)$$

and k_i are the cumulants of z_v. The remainder term $R(v)$ is a double infinite series each term involving a power product of the cumulants k_i and an associated high order derivative $D^rP(z)$, the term with the least order differential being $\frac{k_4}{4!} D^4P(z)$. Using Wishart's (1952) results, the cumulant k_3 of z_v in terms of the standardized cumulant K_3 of the finite population of P_j, is given by

$$k_3 = \left[v^{-\frac{1}{2}}\left(1 - \frac{v}{N - 2}\right)^{\frac{1}{2}} - \frac{v^{\frac{1}{2}}}{N - 2} \cdot \left(1 - \frac{v}{N - 2}\right)^{-\frac{1}{2}}\right]K_3 \; . \qquad (4.17)$$

Substituting now (4.15) in (4.11) we obtain

$$P_{12} = (N - 1)^{-1} \sum_{v=0}^{N-2} \int_0^{P_1} \Big\{P(z_1) - P(z_2) - \frac{1}{6} k_3\left[P^{(3)}(z_1) - P^{(3)}(z_2)\right]\Big\}\, dt + \rho \qquad (4.18)$$

where

$$z_1 = \frac{t + 1 - P_1 - v(2 - P_1 - P_2)(N - 2)^{-1}}{S_{12}v^{\frac{1}{2}}\left(1 - \frac{v}{N - 2}\right)^{\frac{1}{2}}}$$

$$(4.19)$$

$$z_2 = \frac{t + 1 - P_1 - P_2 - v(2 - P_1 - P_2)(N - 2)^{-1}}{S_{12}v^{\frac{1}{2}}\left(1 - \frac{v}{N - 2}\right)^{\frac{1}{2}}}$$

$$\rho = (N - 1)^{-1} \sum_{v=0}^{N-2} \int_0^{P_1} \left[R(z_1) - R(z_2)\right] dt \qquad (4.20)$$

and k_3 is given by (4.17) and $P^{(r)}(z)$ denotes the r^{th} order derivative of $P(z)$.

We now apply the Euler-Maclaurin formula

$$\int_a^b g^{(1)}(t)dt = g(b) - g(a) = (b - a)g^{(1)}\left(\frac{a + b}{2}\right) + \frac{(b - a)^3}{24} g^{(3)}\left(\frac{a + b}{2}\right) + \frac{(b - a)^5}{1920} g^{(5)}(\bar{t}) \qquad (4.21)$$

here given for a general function $g(x)$ satisfying the required continuity conditions and $\bar{t}$ is such that $a \leq \bar{t} \leq b$. Applying this formula first to the differences $P(z_1) - P(z_2)$ and $P^{(3)}(z_1) - P^{(3)}(z_2)$ in (4.18), we find

$$P_{12} = (N - 1)^{-1} \sum_{v=0}^{N-2} \int_0^{\bar{P}_1} \Big[\frac{P_2}{S_{12}} v_1^{-\frac{1}{2}} P^{(1)}\left(\frac{z_1 + z_2}{2}\right) + \frac{P_2^3}{24S_{12}^3} v_1^{-\frac{3}{2}} P^{(3)}\left(\frac{z_1 + z_2}{2}\right) - \frac{k_3}{6} \frac{P_2}{S_{12}} v_1^{-\frac{1}{2}} P^{(4)}\left(\frac{z_1 + z_2}{2}\right) + \omega(t)\Big] dt + \rho \qquad (4.22)$$

where

$$v_1 = v\left(1 - \frac{v}{N - 2}\right) \qquad (4.23)$$

and $\omega(t)$ represents the aggregate of the remainder terms in the application of (4.21). Now integrating (4.22) over t again using (4.21), we obtain retaining only the relevant terms,

$$P_{12} = (N-1)^{-1} \int_0^{N-2} \Big[\frac{P_1P_2}{S_{12}} v_1^{-\frac{1}{2}} p^{(1)}(v_2) + \frac{P_1P_2^3}{24S_{12}^3} v_1^{-\frac{3}{2}} p^{(3)}(v_2) + \frac{P_1^3P_2}{24S_{12}} v_1^{-\frac{1}{2}} p^{(3)}(v_2) - \frac{k_3}{6} \frac{P_1P_2}{S_{12}} v_1^{-\frac{1}{2}} p^{(4)}(v_2)\Big] dv + \rho + \omega + \rho' \qquad (4.24)$$

where

$$v_2 = \frac{1 - \frac{1}{2}(P_1 + P_2) - v \cdot \dfrac{2 - P_1 - P_2}{N - 2}}{S_{12} v_1^{\frac{1}{2}}} \qquad (4.25)$$

ρ is given by (4.20), ω denotes the aggregated remainder terms in the application of (4.21) on (4.22) and ρ' is the remainder term arising from the approximation of $\sum_v$ by $\int dv$. Since we are interested in finding P_{12} to $O(N^{-4})$, only those terms in the evaluation of (4.24) that contribute to $O(N^{-4})$ or to larger orders i.e. $O(N^{-3})$ and $O(N^{-2})$, are to be retained. We now evaluate the terms in (4.24) one by one. The first term is

$$A = (N - 1)^{-1} \cdot \frac{P_1P_2}{S_{12}} \int_0^{N-2} v_1^{\frac{1}{2}} \, p^{(1)}(v_2)dv \qquad (4.26)$$

where

$$p^{(1)}(v_2) = (2\pi)^{-\frac{1}{2}} e^{-\frac{1}{2}v_2^2} . \qquad (4.27)$$

Making the transformation

$$u = v - \tfrac{1}{2}(N - 2) \qquad (4.28)$$

and expanding the exponential in (4.27) as well as $v_1^{-\frac{1}{2}}$ where v_1 is given by (4.23), we find

$$A = \frac{(N - 2)}{(N - 1)} \frac{P_1P_2}{(2 - P_1 - P_2)} (2\pi)^{-\frac{1}{2}} \int_{-h}^{h} e^{-\frac{1}{2}p^2}$$

$$\cdot \exp\left\{- \frac{1}{2} h^{-2}p^4 - \frac{1}{2} h^{-4}p^6 + \text{higher terms}\right\}$$

$$\times \left(1 + \frac{1}{2} h^{-2}p^2 + \frac{3}{8} h^{-4}p^4 + \text{higher terms}\right)dp \qquad (4.29)$$

where

$$h = (2 - P_1 - P_2)(N - 2)^{-\frac{1}{2}} S_{12}^{-1} \qquad (4.30)$$

and the variable of integration is changed to

$$p = 2\, uh(N - 2)^{-1} . \qquad (4.31)$$

Now from (4.29), expanding the exponential $\{\ \}$ and multiplying by the series in ( ) and simplifying, we obtain

$$A = \frac{(N-2)}{(N-1)} \frac{P_1 P_2}{(2 - P_1 - P_2)} \cdot (2\pi)^{-\frac{1}{2}} \int_{-h}^{h} e^{-\frac{1}{2}p^2}$$

$$\cdot \left[1 + \frac{1}{2} h^{-2}(p^2 - p^4) + \frac{1}{8} h^{-4}(3p^4 - 6p^6 + p^8) \right.$$

$$\left. + \text{higher terms } dp \right] . \qquad (4.32)$$

Since P_1 is $O(N^{-1})$, S_{12} is $O(N^{-1})$ so that from (4.30), h is $O(N^{\frac{1}{2}})$. Therefore, we can replace the integration limits in (4.32) by $-\infty$ and $+\infty$ apart from errors which are $O(e^{-N}N^{\varepsilon})$. Using now the standardized normal moments

$$\mu_2 = 1, \ \mu_4 = 3, \ \mu_6 = 15 \text{ and } \mu_8 = 105 \qquad (4.33)$$

we find from (4.32) to $O(N^{-4})$

$$A = \frac{(N-2)}{(N-1)} \frac{P_1 P_2}{(2 - P_1 - P_2)} (1 - h^{-2} + 3h^{-4}) \ . \qquad (4.34)$$

The second term is

$$B = (N-1)^{-1} \cdot \frac{P_1 P_2^3}{24 S_{12}^3} \int_0^{N-2} v_1^{-\frac{3}{2}} P^{(3)}(v_2) dv \qquad (4.35)$$

where

$$P^{(3)}(v_2) = (2\pi)^{-\frac{1}{2}} e^{-\frac{1}{2}v_2^2} (v_2^2 - 1) \ . \qquad (4.36)$$

By a similar argument, using the transformations u and p given by (4.28) and (4.31), and expanding the exponential in (4.36) as well as $v_1^{-\frac{3}{2}}$ and $(v_2^2 - 1)$ in terms of p and multiplying out the resulting series, we find after simplification

$$B = (N - 1)^{-1} \cdot \frac{P_1 P_2^3 S_{12}^{-2}}{6(2 - P_1 - P_2)} \cdot (2\pi)^{-\frac{1}{2}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}p^2}$$

$$\cdot \Big[(p^2 - 1) - \tfrac{1}{2}(p^6 - 6p^4 + 3p^2)h^{-2}$$

$$+ \text{higher terms}\Big] dp. \qquad (4.37)$$

Using the standardized normal moments (4.33), it is seen from (4.37) that B is zero to $O(N^{-4})$ and hence B does not contribute to P_{12} to $O(N^{-4})$. Similarly, we find that the next term

$$C = (N - 1)^{-1} \cdot \frac{P_1^3 P_2}{24 S_{12}} \int_0^{N-2} v_1^{-\frac{1}{2}} p^{(3)}(v_2) dv \qquad (4.38)$$

is reduced to

$$C = \frac{(N - 2)}{(N - 1)} \cdot \frac{P_1^3 P_2}{24(2 - P_1 - P_2)} \cdot (2\pi)^{-\frac{1}{2}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}p^2}$$

$$\cdot \Big[(p^2 - 1) - \tfrac{1}{2}(p^6 - 4p^4 + p^2)h^{-2}$$

$$+ \text{higher terms}\Big] dp \ . \qquad (4.39)$$

The evaluation of the terms retained in (4.39) yields

$$C = -\frac{(N - 2)}{(N - 1)} \cdot \frac{P_1^3 P_2}{12(2 - P_1 - P_2)} \cdot h^{-2} \qquad (4.40)$$

which is $O(N^{-5})$, so that C does not contribute to P_{12} to $O(N^{-4})$. The next term is

$$D = -(N - 1)^{-1} \cdot \frac{P_1 P_2}{6S_{12}} \int_0^{N-2} k_3\, v_1^{-\frac{1}{2}}\, p^{(4)}(v_2)\, dv \qquad (4.41)$$

where

$$p^{(4)}(v_2) = (2\pi)^{-\frac{1}{2}} e^{-\frac{1}{2}v_2^2} (3v_2 - v_2^3) \qquad (4.42)$$

and k_3 is a function of v given by (4.17). Now using the same transformations u and p, expanding the quantities $v_1^{-\frac{1}{2}}$, $(v_2^3 - 3v_2)$ and k_3 and the exponential in (4.42) in terms of p and multiplying out the resulting series, we find after considerable simplification

$$D = \frac{2(N - 2)^{\frac{1}{2}}}{3(N - 1)} \cdot \frac{P_1 P_2 K_3}{(2 - P_1 - P_2)} \cdot (2\pi)^{-\frac{1}{2}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}p^2} \cdot \left[\tfrac{1}{2}h^{-1}(p^4 - 3p^2) + \tfrac{1}{4}h^{-3}(6p^6 - 9p^4 - p^8) + \text{higher terms}\right] dp \,. \qquad (4.43)$$

Using the standardized normal moments (4.33), the evaluation of the terms retained in (4.43) yields

$$D = -\frac{2(N - 2)}{(N - 1)} \cdot \frac{P_1 P_2 K_3}{(2 - P_1 - P_2)} \cdot h^{-3}(N - 2)^{-\frac{1}{2}} \qquad (4.44)$$

$$= - \frac{2(N - 2)^2}{(N - 1)} \cdot \frac{P_1 P_2 K_3 S_{12}^3}{(2 - P_1 - P_2)^4} \qquad (4.45)$$

which is $O(N^{-4})$ since

$$K_3 S_{12}^3 = (N - 2)^{-1} \sum_{3}^{N} (P_j - \frac{2 - P_1 - P_2}{N - 2})^3 \qquad (4.46)$$

is $O(N^{-3})$. We shall presently show that the remainder terms ρ, ω and ρ' do not contribute to P_{12} to $O(N^{-4})$, so that adding the expressions A and D (since B and C are zero to $O(N^{-4})$)given by (4.34) and (4.44), we obtain for the probability P_{12}, an approximation to $O(N^{-4})$ given by

$$P_{12} = \frac{(N - 2)}{(N - 1)} \cdot \frac{P_1 P_2}{(2 - P_1 - P_2)} \left[1 - h^{-2} + 3h^{-4} - 2K_3 h^{-3} (N - 2)^{-\frac{1}{2}} \right]$$

(4.47)

where h is given by (4.30) and K_3 by (4.46). Since the last two terms in (4.47) are $O(N^{-4})$, we obtain to $O(N^{-3})$ the simplified expression

$$P_{12} = \frac{(N - 2)}{(N - 1)} \frac{P_1 P_2}{(2 - P_1 - P_2)} (1 - h^{-2}) \ . \qquad (4.48)$$

Let us now consider the remainder terms ρ, ω and ρ'. The remainder term ω represents the aggregated remainder terms in applying the Euler-Maclaurin formula (4.21) to the differences $P(z_1) - P(z_2)$ and $P^{(3)}(z_1) - P^{(3)}(z_2)$ in (4.18). The remainder term in the application of (4.21) to the difference $P(z_1) - P(z_2)$ is $\frac{(z_1 - z_2)^5}{1920} P^5[z_1 + \theta(z_2 - z_1)]$ with $0 \leq \theta \leq 1$. Therefore, the contribution to P_{12} from this remainder term, say ω_1, is

$$\omega_1 = \frac{(N-1)^{-1}}{1920} \sum_{v=0}^{N-2} \int_0^{P_1} (z_1 - z_2)^5 \, P^{(5)}[z_1 + \theta(z_2 - z_1)]dt \tag{4.49}$$

where

$$z_1 - z_2 = P_2 S_{12}^{-1} v^{-\frac{1}{2}} (1 - \frac{v}{N-2})^{-\frac{1}{2}} . \tag{4.50}$$

Now consider the first term in the application of (4.21) to integrate (4.49) over t, say ω_2, i.e.,

$$\omega_2 = \frac{(N-1)^{-1}}{1920} \sum_{v=0}^{N-2} P_1 \cdot \left[P_2 S_{12}^{-1} v^{-\frac{1}{2}} (1 - \frac{v}{N-2})^{-\frac{1}{2}}\right]^5$$

$$\cdot P^{(5)}\left[v_2 + \theta' P_2 S_{12}^{-1} v^{-\frac{1}{2}} (1 - \frac{v}{N-2})^{-\frac{1}{2}}\right] \tag{4.51}$$

with $-\frac{1}{2} \leq \theta' \leq \frac{1}{2}$. Making the transformations u and p given by (4.26) and (4.31) and proceeding as before, we find after simplification

$$\omega_2 = c \, P_1\left[P_2(N-2)^{-\frac{1}{2}} S_{12}^{-1}\right]^5 \cdot \frac{(N-2)^{\frac{3}{2}}}{(N-1)} S_{12} \int_{-\infty}^{\infty}$$

$$\cdot \left[1 + O(p^2 n^{-1})\right] x \left[P^{(5)}(p) + O(N^{-\frac{1}{2}})\right] dp \tag{4.52}$$

where c is a constant.

Since $\int_{-\infty}^{\infty} p^{(5)}(p)dp$ is zero, we find from (4.52) that ω_2 is at least of order $O(N^{-4\frac{1}{2}})$, so that it does not contribute to P_{12} to $O(N^{-4})$. Similar arguments apply to the differences $P^{(3)}(z_1) - P^{(3)}(z_2)$ as well as the remainder terms arising from applying (4.21) in integrating (4.18) over t so that the aggregated remainder term ω does not contribute to P_{12} to $O(N^{-4})$.

Consider now the remainder term ρ' arising from the approximation of $\sum\limits_{v}$ by $\int dv$. From the following version of Euler-Maclaurin formula:

$$\sum_{v=0}^{N-2} f(v) - \int_{0}^{N-2} f(v)dv = \frac{1}{2} f(0) + \frac{1}{2} f(N - 2)$$

$$+ \sum_{s=1}^{m-1} \frac{B_{2s}}{(2s)!} \left\{ f^{(2s-1)}(N - 2) - f^{(2s-1)}(0) \right\}$$

$$+ \frac{(N - 2)}{(2m)!} B_{2m} \; f^{(2m)}(\overline{N - 2}\,\theta_N) \qquad (4.53)$$

where B_{2s} are the Bernoulli numbers and $f^{(r)}$ is the r^{th} derivative w.r.t. v of any of the integrand functions involved in (4.24) and $0 \leq \theta_N \leq 1$ while 2m, the order of the remainder term in (4.53) is at our disposal, it is seen that ρ' involves the terminal differentials of the integrands at the end points of integration $v = 0$ and $v = N - 2$ which are zero since v_2 becomes infinite and the integrands involve the term $e^{-\frac{1}{2}v_2^2}$. Now consider the remainder term in (4.53). At $v = (N - 2)\theta_N$, from (4.25),

$$v_2 = (1 - \frac{P_1 + P_2}{2})(1 - 2\theta_N)\left[\theta_N(1 - \theta_N)\right]^{-\frac{1}{2}} S_{12}^{-\frac{1}{2}}(N - 2)^{-\frac{1}{2}}. \qquad (4.54)$$

We now separate the values of θ_N between 0 and 1 into two groups. In the first group, θ_N is equal to 1/2 or the leading term of the difference between θ_N and 1/2 is proportional to N^{-r_N} with $r_N > 0$. The remaining values of θ_N fall in the second group. It is easily seen from (4.54) that for the values of θ_N in the second group v_2 is $O(N^s)$ with $s > 0$ since S_{12}^{-1} is $O(N^1)$, and the argument to be used for the remainder term in case (b) below also applies to the values of θ_N in this group. Now from (4.54), for values of θ_N in the first group either v_2 is zero or is $O(N^{\frac{1}{2}-r_N})$. So, we now distinguish the two cases (a) $r_N \geq 1/2$ and (b) $r_N < 1/2$. Consider first the case (a). In terms of the variable u where u is given by (4.28),

$$v_2 = \text{const.}(N - 2)^{-\frac{3}{2}} S_{12}^{-1} u\left[1 - \left[\frac{2u}{N - 2}\right]^2\right]^{-\frac{1}{2}}$$

$$= \text{const.}(N - 2)^{-\frac{1}{2}} S_{12}^{-1} \sum_{i=0}^{\infty} \binom{-\frac{1}{2}}{i} (-1)^i \left(\frac{2u}{N - 2}\right)^{2i+1}. \qquad (4.55)$$

Therefore, by repeated differentiation of (4.55), we have for the largest value of $|u|$,

$$\frac{d^t v_2}{dv^t} = \frac{d^t v_2}{du^t} = O(N^{\frac{1}{2}-t}) \ . \qquad (4.56)$$

The repeated differentiation of the function involving v_2 only in the integrand of (4.24), say $g(v_2)$, is now seen to have a leading term of the form $\frac{d^t g}{dv_2^t} \cdot (\frac{dv_2}{du})^t$ which is of order $O(N^{-\frac{1}{2}t})$. Therefore, from the Leibnitz formula of differentiation of a product it is evident that every integrand function in (4.24) which is seen to be of the type $v_1^{-b}\ g(v_2)$, $b > 0$, is $O(N^{-k})$ with $k > 4$ provided $2m$ is taken sufficiently large.

In case (c), the remainder term goes down as $O(e^{-bN^s} \cdot N^a)$ where $s = 1 - 2r_N > 0$ and hence smaller than $O(N^{-4})$. So, the remainder term ρ' does not contribute to P_{12} to $O(N^{-4})$. Finally consider the remainder term ρ given by (4.20). From (4.17) it is seen that the sum of the exponents of the power products in v and N in the formula for k_3 is equal to $- 1/2$. Now in the p-scale, $v = \frac{N - 2}{2}(1 + h^{-1}p) = qN$ with $q = \frac{(N - 2)}{2N}(1 + h^{-1}p)$. So, k_3 is order $O(N^{-\frac{1}{2}})$ in the p-scale since $q = \frac{1}{2} + O(N^{-\frac{1}{2}})$ in the p-scale because h^{-1} is $O(N^{-\frac{1}{2}})$. The Appendix in Chapter IX gives a heuristic argument to show that the sum of the exponents of the power products in v and N in the formula for k_r is equal to $(-\frac{r}{2}+1)$, i.e. k_r is $O(N^{-\frac{r}{2}+1})$ with $v = qN$ and this is actually verified up to $r = 8$. Now, in the remainder term ρ, k_4 and k_3^2 are the largest order terms, i.e. $O(N^{-1})$ with $v = qN$. An analysis similar to that of the k_3 terms shows that the terms with k_4 and k_3^2 are of smaller order than $O(N^{-4})$ and so do not contribute to P_{12} to $O(N^{-4})$. Note from (4.44) that the term with k_3 contributes to P_{12} only terms of order $O(N^{-4})$ and smaller. Since all the remaining terms in ρ involve the higher order cumulants and their powers which are of smaller order than $O(N^{-1})$ with $v = qN$, it follows the the terms in ρ do not contribute to P_{12} to $O(N^{-4})$. We shall not discuss here the inversion of the double summation in (4.16) and its convergence.

Independently of the above argument that the remainder terms ρ, ω and ρ' do not contribute to P_{12} to $O(N^{-4})$, the following two checks provide additional evidence that all the terms of $O(N^{-4})$ and larger are included in (4.47). The first check is the special case when all probabilities P_i are equal to $2/N$ so that $S_{12} = 0$ and $h^{-1} = 0$. This check tests only the leading term of (4.47) since $h^{-1} = 0$ so that the coefficients of the remaining terms in (4.47) are not affected by this check. In this case, P_{12} given by (4.47) reduces to $2/N(N-1)$ which is the correct probability for units 1 and 2 to be in a sample of size 2. A more searching check which takes account of all the terms in (4.47) is provided by testing the order to which the equation

$$\sum_{i' \neq i}^{N} P_{ii'} = (n - 1)P_i \tag{4.57}$$

which in our case $n = 2$ reduces to

$$\sum_{i' \neq i}^{N} P_{ii'} = P_i \tag{4.58}$$

is satisfied. We now show that (4.58) is in fact satisfied to an order $(N - 1)\ O(N^{-4}) = O(N^{-3})$ if (4.47) is substituted in (4.58) which confirms that (4.47) is correct to $O(N^{-4})$. Using the formula (4.30) for h and (4.46) for K_3, (4.47) can be written in the form

$$\begin{aligned} P_{ii'} = \frac{P_iP_{i'}}{(2 - P_i - P_{i'})} \cdot \frac{(N - 2)}{(N - 1)} \Bigg[1 &- \frac{\sum^{N} P_t^2 - P_i^2 - P_{i'}^2}{(2 - P_i - P_{i'})^2} \\ &\cdot \left(1 + \frac{1}{N - 3} + \frac{6}{N - 3}\right) + \frac{3(\sum P_t^2 - P_i^2 - P_{i'}^2)^2}{(2 - P_i - P_{i'})^4} \\ &+ \frac{1}{N - 3} + \frac{3}{(N - 3)^2} - \frac{4}{(N - 2)^2} - \frac{2\sum P_t^3}{(2 - P_i - P_{i'})^3} \\ &+ \frac{6\sum P_t^2}{(N - 2)(2 - P_i - P_{i'})^2} \Bigg] \end{aligned} \tag{4.59}$$

which to $O(N^{-4})$, reduces to

$$P_{ii'} = \frac{P_iP_{i'}}{(2 - P_i - P_{i'})} - \frac{P_iP_{i'}(\sum P_t^2 - P_i^2 - P_{i'}^2)}{(2 - P_i - P_{i'})^3}$$

$$+ \frac{3P_iP_{i'}(\sum P_t^2)^2}{(2 - P_i - P_{i'})^5} - \frac{2P_iP_{i'}\sum P_t^3}{(2 - P_i - P_{i'})^4} \qquad (4.60)$$

where the subscripts 1 and 2 are replaced by i and i' respectively. Expanding all denominators in (4.60) binomially, retaining all terms to $O(N^{-4})$, we find after simplication

$$P_{ii'} = [\tfrac{1}{2} P_iP_{i'} + \tfrac{1}{4}(P_i^2P_{i'} + P_iP_{i'}^2) - \tfrac{1}{8} P_iP_{i'} \sum P_t^2]$$

$$+ \tfrac{1}{8}(2P_i^3P_{i'} + 2P_iP_{i'}^3 + 2P_i^2P_{i'}^2)$$

$$- \tfrac{3}{16}(P_i^2P_{i'} + P_iP_{i'}^2) \sum P_t^2 + \tfrac{3}{32}(\sum P_t^2)^2 P_iP_{i'}$$

$$- \tfrac{1}{8} P_iP_{i'} \sum P_t^3 \, . \qquad (4.61)$$

Summing (4.61) now over i' from 1 to N excepting i' = i and noting that $\sum P_t = 2$, we obtain to $O(N^{-3})$,

$$\sum_{i' \neq i}^{N} P_{ii'} = \tfrac{1}{2} P_i(2 - P_i) + \tfrac{1}{4} P_i^2(2 - P_i) + \tfrac{1}{4} P_i(\sum P_t^2 - P_i^2)$$

$$+ \tfrac{1}{2} P_i^3 - \tfrac{1}{8} P_i^2 \sum P_t^2 - \tfrac{1}{8} P_i(2 - P_i) \sum P_t^2$$

$$(4.62)$$

which reduces to P_i thereby providing the desired check.

B. Variance Formulas to Orders $O(N^1)$ and $O(N^0)$

Substituting for $P_{ii'}$ from (4.61) in the variance formula for $\hat{Y}$, namely,

$$V(\hat{Y}) = \sum^{N} \frac{y_j^2}{P_j} + \sum^{N}_{i \neq i'} \frac{P_{ii'}}{P_i P_{i'}} y_i y_{i'} - Y^2 \tag{4.63}$$

we find

$$\begin{aligned} V(\hat{Y}) = \; & \sum \frac{y_j^2}{P_j} + \frac{1}{2} \sum_{i \neq i'} y_i y_{i'} + \frac{1}{4} \sum_{i \neq i'} (P_i + P_{i'}) y_i y_{i'} \\ & - \frac{1}{8}(\sum P_t^2)(\sum_{i \neq i'} y_i y_{i'}) \\ & - \frac{3}{16}(\sum P_t^2)\Big[\sum_{i \neq i'} (P_i + P_{i'}) y_i y_{i'}\Big] \\ & + \frac{1}{4} \sum_{i \neq i'} (P_i^2 + P_{i'}^2) y_i y_{i'} + \frac{1}{4} \sum_{i \neq i'} (P_i y_i)(P_{i'} y_{i'}) \\ & + \frac{3}{32}(\sum P_t^2)^2 (\sum_{i \neq i'} y_i y_{i'}) \\ & - \frac{1}{8}(\sum P_t^3)(\sum_{i \neq i'} y_i y_{i'}) - Y^2 \, . \end{aligned} \tag{4.64}$$

Retaining terms to $O(N^0)$, (4.64) reduces to

$$\begin{aligned} V(\hat{Y}) = \; & \sum \frac{y_j^2}{P_j} - \frac{1}{2} Y^2 - \frac{1}{2} \sum y_j^2 + \frac{1}{2} Y \sum P_j y_j - \frac{1}{8}(\sum P_t^2) Y^2 \\ & - \frac{1}{2} \sum P_j y_j^2 + \frac{1}{8}(\sum P_t^2)(\sum y_j^2) - \frac{3}{8}(\sum P_t^2)(\sum P_j y_j) Y \\ & + \frac{1}{2} Y(\sum P_j^2 y_j) + \frac{3}{32}(\sum P_t^2)^2 Y^2 - \frac{1}{8} Y^2 (\sum P_t^3) \\ & + \frac{1}{4}(\sum P_j y_j)^2 \end{aligned}$$

$$= \sum^{N} P_j(1 - \tfrac{1}{2} P_j)(\tfrac{y_j}{P_j} - \tfrac{Y}{2})^2 - \tfrac{1}{2} \sum^{N} (P_j^3 - \tfrac{1}{4} P_j^2 \sum P_t^2)$$

$$\cdot (\tfrac{y_j}{P_j} - \tfrac{Y}{2})^2 + \tfrac{1}{4}(\sum P_j y_j - \tfrac{1}{2} Y \sum P_t^2)^2 . \qquad (4.66)$$

On the other hand, if terms only to $O(N^1)$ are retained,

$$V(\hat{Y}) = \sum \frac{y_j^2}{P_j} - \tfrac{1}{2} Y^2 - \tfrac{1}{2} \sum y_j^2 + \tfrac{1}{2} Y \sum P_j y_j - \tfrac{1}{8}(\sum P_t^2)Y^2$$

$$= \sum^{N} P_j(1 - \tfrac{1}{2} P_j)(\tfrac{y_j}{P_j} - \tfrac{Y}{2})^2 . \qquad (4.67)$$

The variance of the estimate of the total Y in sampling with replacement is

$$V(\hat{Y}') = \sum^{N} P_j(\frac{y_j}{P_j} - \frac{Y}{2})^2 . \qquad (4.68)$$

Equation (4.67) which is correct to $O(N^1)$ compared with (4.68) shows the characteristic reduction in the variance through the "finite population corrections" $(1 - \frac{1}{2} P_j)$. Hence, the present sampling procedure without replacement yields a smaller variance asymptotically for the estimate of the total than sampling with replacement. For the special case of equal probabilities $P_i = \frac{2}{N}$, (4.63) to $O(N^0)$ reduces to the familiar variance formula for the estimate of the total in sampling with equal probability and without replacement, i.e.,

$$V(\hat{Y}) = \frac{N^2}{2(N - 1)} \cdot (1 - \tfrac{2}{N}) \sum^{N} (y_j - \tfrac{Y}{N})^2 . \qquad (4.69)$$

C. Estimation of the Variance

The method is to substitute for $P_{ii'}$ in the Yates and Grundy estimate of the variance, which for $n = 2$ is

$$v_{YG}(\hat{Y}) = \frac{P_iP_{i'} - P_{ii'}}{P_{ii'}} \left(\frac{y_i}{P_i} - \frac{y_{i'}}{P_{i'}}\right)^2 . \qquad (4.70)$$

From (4.61) to $O(N^{-3})$,

$$P_{ii'} = \frac{1}{2} P_iP_{i'} \left[1 + \frac{1}{2}(P_i + P_{i'}) - \frac{\Sigma P_t^2}{4}\right] . \qquad (4.71)$$

Therefore, substituting for $P_{ii'}$ from (4.71) in (4.70), we find to $O(N^1)$,

$$v_{YG}(\hat{Y}) = \frac{(1 - \frac{P_i + P_{i'}}{2} + \frac{\Sigma P_t^2}{4})}{(1 + \frac{P_i + P_{i'}}{2} - \frac{\Sigma P_t^2}{4})} \cdot \left(\frac{y_i}{P_i} - \frac{y_{i'}}{P_{i'}}\right)^2 . \qquad (4.72)$$

Expanding the denominator binomially and retaining terms to $O(N^1)$,

$$v_{YG}(\hat{Y}) = (1 - P_i - P_{i'} + \frac{\Sigma P_t^2}{2}) \left(\frac{y_i}{P_i} - \frac{y_{i'}}{P_{i'}}\right)^2 . \qquad (4.73)$$

For the special case of equal probabilities $P_i = \frac{2}{N}$, (4.73) to $O(N^1)$ agrees with the familiar formula for the estimate of the variance in equal probability sampling without replacement, i.e.,

$$v(\hat{Y}) = \frac{N^2}{2}(1 - \frac{2}{N}) \sum^{2} (y_j - \bar{y})^2 \qquad (4.74)$$

where $\bar{y}$ is the sample mean of the two units i and i'. To find $v_{YG}(\hat{Y})$ to $O(N^0)$, substituting for $P_{ii'}$ in (4.70) from (4.61) which is correct to $O(N^{-4})$, and expanding the denominator binomially and retaining terms to $O(N^0)$, we obtain after simplification

$$v_{YG}(\hat{Y}) = \Big[1 - (P_i + P_{i'}) + \frac{1}{2} \Sigma P_t^2 - \frac{1}{2}(P_i^2 + P_{i'}^2) - \frac{1}{4}(\Sigma P_t^2)^2 + \frac{1}{4}(P_i + P_{i'}) \Sigma P_t^2 + \frac{1}{2} \Sigma P_t^3\Big](\frac{y_i}{P_i} - \frac{y_{i'}}{P_{i'}})^2 \qquad (4.75)$$

which agrees to $O(N^0)$ with (4.74) when all $P_i = \frac{2}{N}$.

In this connection, it is worthwhile to point out an important aspect of the Yates and Grundy estimate of the variance for the case n = 2. From (4.63) and (4.68), it can be easily shown that a necessary condition for $V(\hat{Y})$ to be smaller than $V(\hat{Y}')$ is

$$P_{ii'} \leq P_i P_{i'} \ . \qquad (4.76)$$

For general sample size n, this condition is

$$P_{ii'} \leq \frac{2(n-1)}{n} P_i P_{i'} \ . \qquad (4.77)$$

This condition is given by Narain (1951). Therefore, it immediately follows from (4.76) and (4.70) that the Yates and Grundy estimate of the variance is always positive if a sampling procedure without replacement for which $P_i = np_i$ is more efficient than sampling with replacement, and $n = 2$. That is, if there is a sampling procedure without replacement for which the variance is smaller than the variance in sampling with replacement independent of the y_i, which is the case we are interested in, then the Yates and Grundy estimate of the variance is always positive. It may be noted that this result is true only for the case $n = 2$, since conditions (4.77) are not sufficient to show that

$$v_{YG}(\hat{Y}) = \sum_{i'>i}^{n} \frac{P_i P_{i'} - P_{ii'}}{P_{ii'}} \left(\frac{y_i}{P_i} - \frac{y_{i'}}{P_{i'}}\right)^2 \qquad (4.78)$$

is always positive. (4.78) is positive if conditions (4.76) for all i and i' ($i \neq i'$) are satisfied. However, conditions (4.77) do not imply (4.76) except when $n = 2$.

For our particular sampling procedure, condition (4.76) is in fact satisfied to $O(N^{-3})$ since from (4.71),

$$P_i P_{i'} - P_{ii'} = \frac{P_i P_{i'}}{2} \left[1 - \frac{P_i + P_{i'}}{2} + \frac{\sum P_t^2}{4}\right] \qquad (4.79)$$

which is greater than zero since $\frac{P_i + P_{i'}}{2} \leq 1$. This fact could of course have been inferred from (4.67) which shows that $V(\hat{Y})$ is smaller than $V(\hat{Y}')$ so that (4.76) would have followed as a necessary condition.

D. An Example for Efficiency Comparisons

We use the data given in Table 1, Chapter III, which are taken from Horvitz and Thompson (1952). The population here consists of $N = 20$ blocks in Ames, Iowa, and y_j and x_j denote respectively the actual number of households and "eye-estimated" number of households in the j^{th} block ($j = 1$ to 20). The probability P_j for the j^{th} unit to be in a sample of size 2 is taken proportional to the "eye-estimated" number of households x_j, i.e. $P_j = 2x_j/\sum_{j=1}^{20} x_j$. In Table 6 below, the evaluations of the variance of the estimated total for the present sampling procedure and for different sampling systems considered in the literature are given. These efficiency comparisons ignore cost.

Sampling systems 2 to 10 correspond to different methods of utilizing supplementary information x_j, and sampling system 1 is equal probability sampling without utilizing supplementary information. It is evident from Table 6, that all these methods of utilizing x_j are vastly superior to system 1. The estimator 2 is the well known ratio estimator in equal probability sampling and here the bias of this estimator which equals 1.17 is neglected. In system 3, the 20 blocks are divided into two strata according to the measure of size x_j, the ten largest belong to stratum 1 and the remaining ten belong to stratum 2, and X_t denotes the stratum total of x_j. Since only one unit is drawn with p.p.s. from each stratum, no

Table 6. Variances of various estimators of the total of the y_i for the population given in Table 1

Sampling system	Method of selection	Form of the estimator	Variance of the estimator	% relative efficiency
1.	Equal probability without replacement	$N\bar{y}$	16,219	100
2.	"	$\frac{\sum^{2} y_j}{\sum^{2} x_j} \cdot X$	3,280	497
3.	Stratified; one unit with p.p.s. from each of 2 strata	$\sum^{2} \frac{y_t}{x_t} \cdot X_t$	3,934	412
4.	Lahiri: Unbiased ratio estimator	$\frac{(\sum y_j)_2}{(\sum x_j)_2} \cdot X$	3,579	453
5.	Horvitz and Thompson (Method 1)	$\sum^{2} \frac{y_j}{P_j}$	3,095	524
6.	Horvitz and Thompson (Method 2)	"	3,075	527
7.	Mickey, ordered estimator	$\bar{u}$	3,055	531
8.	Mickey, unordered estimator	u^*	$3,026 < V(u^*) < 3,038$	$534 < E < 536$
9.	P.p.s. with replacement	$\sum^{2} \frac{y_j}{P_j}$	3,241	500

Table 6. (Continued)

Sampling system	Method of selection	Form of the estimator	Variance of the estimator	% relative efficiency
10.	Present procedure			
(a)	$O(N^1)$	$\sum^{2} \frac{y_j}{P_j}$	3,025	536
(b)	$O(N^0)$	"	3,007	539

valid estimate of the variance can be found for system 3. In system 4, the two units are selected with probability proportional to the sum of the measures for the two units, i.e. $(\sum x_j)_2/X$ where $(\ )_2$ denotes a set of 2 units. The estimator 4 belongs to class 3 according to the classification of the estimators in Chapter II. Sampling systems 5 and 6 and their limitations have been described in Chapter II. The estimator 7 belongs to class 1. From Mickey (1959),

$$\bar{u} = \frac{1}{2}\left(\frac{y_1}{p_1} + \frac{y_2}{p_2}\right) + \frac{p_1}{2}\left(\frac{y_1}{p_1} - \frac{y_2}{p_2}\right) . \qquad (4.80)$$

The estimator 8, u*, obtained by unordering $\bar{u}$ is

$$u^* = \frac{1}{2}\left(\frac{y_1}{p_1} + \frac{y_2}{p_2}\right) + \frac{p_1 - p_2}{2(2 - p_1 - p_2)}\left(\frac{y_1}{p_1} - \frac{y_2}{p_2}\right) . \qquad (4.81)$$

The variance for the first six systems are taken from Horvitz and Thompson (1952) and the variance for the estimators 7 and 8 are taken from Mickey (1959). For the estimate 8, only bounds on the variance are available. The variance for systems 9, 10a and 10b is computed from the formulas (4.68), (4.67) and (4.66) respectively. Systems 5 to 10 have approximately the same variance in magnitude where systems 5, 6 and 10 belong to class 2, and systems 7 and 8 belong to class 1. This may indicate the approximate equality of efficiency of estimators in classes 1 and 2 (a discussion on this aspect is given in Mickey, 1959). Incidentally, our sampling procedure 10 has the smallest variance compared to the other systems 1 to 9, though the gain in efficiency is comparatively small. Also, there is a gain in efficiency of about 7% (234/3241) through sampling without replacement as compared to sampling with replacement (10b vs. 9). Finally, it is of interest to exhibit the nature of convergence of approximations $O(N^1)$ and $O(N^0)$ to $V(\hat{Y})$, by regarding the variance formula (4.68) for sampling with replacement as an approximation to $O(N^2)$ as set out in Table 7 below.

Table 7. Approximations to the variance of $\hat{Y}$

Order of approximation	Formula used	$V(\hat{Y})$	Difference
$O(N^2)$	Eq. (4.68)	3,241	216
$O(N^1)$	Eq. (4.67)	3,025	18
$O(N^0)$	Eq. (4.66)	3,007	

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 formula (4.67) to $O(N^1)$ which is fairly simple to compute, should be satisfactory.

E. Comparison with the Method of Revised Probabilities of Yates and Grundy

The iteration procedure of Yates and Grundy (1953) to obtain revised probabilities which ensure that $P_j = np_j$, has been described in Chapter II. It is proved here that, for the case $n = 2$, the $P_{ii'}$ values attained through the Yates and Grundy procedure and through the present sampling procedure are exactly the same to $O(N^{-3})$, but not to $O(N^{-4})$ so that $V(\hat{Y})$ is the same for both the procedures to $O(N^1)$ but not to $O(N^0)$. Since the terms of $O(N^1)$ are the important terms contributing to the gain in precision of sampling without replacement over sampling with replacement for moderately large N, this result shows that both the procedures have practically the same efficiency. However, with our procedure there is no need to compute the revised probabilities which involves heavy computation as N increases.

Now from (4.71), the probability of selecting units i and i' for our procedure to $O(N^{-3})$ is

$$P_{ii'} = 2p_ip_{i'} + 2(p_i^2p_{i'} + p_ip_{i'}^2) - 2p_ip_{i'}\sum p_t^2 \qquad (4.82)$$

since $P_i = 2p_i$. For the Yates and Grundy procedure, the probability of selecting units i and i', say $P_{ii'}^{(a)}$, is given by

$$P_{ii'}^{(a)} = \frac{p_i^*p_{i'}^*}{1 - p_i^*} + \frac{p_i^*p_{i'}^*}{1 - p_{i'}^*} \qquad (4.83)$$

and

$$P_i = p_i^* + p_i^* \sum_{j\neq i}^{N} \frac{p_j^*}{1 - p_j^*} = 2p_i \qquad (4.84)$$

where p_i^* are the revised probabilities which ensure that $P_i = 2p_i$. Now, expanding (4.84) binomially, we obtain to $O(N^{-2})$,

$$P_i = p_i^* \left[2 + (\sum p_t^{*2} - p_i^*)\right] = 2p_i \qquad (4.85)$$

or

$$p_i^* = p_i \left[1 + \frac{(\sum p_t^{*2} - p_i^*)}{2}\right]^{-1}$$

$$= p_i \left[1 - \frac{(\sum p_t^{*2} - p_i^*)}{2}\right] \text{to } O(N^{-2})$$

$$= p_i \left[1 - \frac{(\sum p_t^2 - p_i)}{2}\right] \text{to } O(N^{-2}) \qquad (4.86)$$

since

$$p_i^* = p_i \left[1 + \text{terms of } O(N^{-1})\right] . \qquad (4.87)$$

Further, expanding (4.83) binomially, we obtain to $O(N^{-3})$,

$$P_{ii'}^{(a)} = p_i^*p_{i'}^*(1 + p_i^*) + p_i^*p_{i'}^*(1 + p_{i'}^*) . \qquad (4.88)$$

Substituting for p_i^* from (4.86) in (4.88), we find to $O(N^{-3})$,

$$P^{(a)}_{ii'} = 2p_ip_{i'} + 2(p_i^2p_{i'} + p_ip_{i'}^2) - 2p_ip_{i'}\sum p_t^2 \tag{4.89}$$

which is exactly the same as (4.82). Now let us examine the comparison of these formulas to $O(N^{-4})$. From (4.61), we obtain to $O(N^{-4})$,

$$\begin{aligned} P_{ii'} = {} & 2p_ip_{i'} + 2(p_i^2p_{i'} + p_ip_{i'}^2) - 2p_ip_{i'}\sum p_t^2 \\ & + 4(p_i^3p_{i'} + p_ip_{i'}^3) + 4p_i^2p_{i'}^2 \\ & - 6(p_i^2p_{i'} + p_ip_{i'}^2)\sum p_t^2 + 6p_ip_{i'}(\sum p_t^2)^2 \\ & - 4p_ip_{i'}\sum p_t^3 \end{aligned} \tag{4.90}$$

since $P_i = 2p_i$. On the other hand, for the Yates and Grundy procedure, we may write to $O(N^{-3})$,

$$\begin{aligned} P_i &= p_i^* + p_i^* \sum_{j\neq i}^{N} p_j^*(1 + p_j^* + p_j^{*2}) \\ &= p_i^*\left[2 + \sum p_t^{*2} + \sum p_t^{*3} - p_i^* - p_i^{*2}\right] = 2p_i \end{aligned} \tag{4.91}$$

so that

$$\begin{aligned} p_i^* &= p_i\left[1 + \frac{(\sum p_t^{*2} + \sum p_t^{*3} - p_i^* - p_i^{*2})}{2}\right]^{-1} \\ &= p_i\left[1 - \frac{(-\sum p_t^{*2} + \sum p_t^{*3} - p_i^* - p_i^{*2})}{2}\right. \\ &\qquad \left. + \frac{(\sum p_t^{*2})^2 + p_i^{*2} - 2p_i^*\sum p_t^{*2}}{4}\right] \end{aligned} \tag{4.92}$$

to $O(N^{-3})$. Now substituting the value of p_i^* to $O(N^{-2})$ from (4.86) in r.h.s. of (4.92), we find after simplification, to

$O(N^{-3})$ that

$$p_i^* = p_i\left[1 + \frac{3(\sum p_t^2)^2 - 2\sum p_t^2 + 2p_i + 4p_i^2 - 3p_i\sum p_t^2 - 4\sum p_t^3}{4}\right]. \tag{4.93}$$

Moreover, we obtain from (4.83) to $O(N^{-4})$,

$$p_{ii'}^{(a)} = p_i^* p_{i'}^*(1 + p_i^* + p_i^{*2}) + p_i^* p_{i'}^*(1 + p_{i'}^* + p_{i'}^{*2}). \tag{4.94}$$

Finally, substituting the p_i^* given by (4.93) in (4.94), we obtain after simplification, and to $O(N^{-4})$ that

$$\begin{aligned} p_{ii'}^{(a)} = {} & 2p_ip_{i'} + 2(p_i^2p_{i'} + p_ip_{i'}^2) - 2p_ip_{i'}\sum p_t^2 \\ & + 4(p_i^3p_{i'} + p_ip_{i'}^3) + \frac{3}{2}\, p_i^2p_{i'}^2 - \frac{7}{2}(p_i^2p_{i'} + p_ip_{i'}^2)\sum p_t^2 \\ & + \frac{7}{2}\, p_ip_{i'}(\sum p_t^2)^2 - 4\, p_ip_{i'}\sum p_t^3 \,. \end{aligned} \tag{4.95}$$

Comparing now (4.90) with (4.95) it is seen that $P_{ii'}$ and $p_{ii'}^{(a)}$ differ in three terms which are $O(N^{-4})$. For the special case of equal probabilities $P_i = \frac{2}{N}$ or $p_i = \frac{1}{N}$, the probability $p_{ii'}^{(a)}$ like $P_{ii'}$ reduces to $2/N(N - 1)$ which is the probability for selecting units i and i' in the equal probability case, the sample size being two. The check (4.58) which was used for $P_{ii'}$ can also be applied to check the order of $p_{ii'}^{(a)}$. It has been verified that

$$\sum_{i' \neq i}^{N} p_{ii'}^{(a)} = P_i = 2p_i \tag{4.96}$$

to $O(N^{-3})$, by substituting for $p_{ii'}^{(a)}$ from (4.95) in (4.96).

Now, using the values of $P_{ii'}^{(a)}$ in (4.96) and proceeding exactly as in section B, it is found that the variance of $\hat{Y}$ to $O(N^0)$ is

$$V_1(\hat{Y}) = \sum^{N} P_j(1 - \frac{P_j}{2})(\frac{y_j}{P_j} - \frac{Y}{2})^2 - \frac{1}{2} \sum^{N} (P_j^3 - \frac{P_j^2 \sum P_t^2}{4})$$

$$\cdot (\frac{y_j}{P_j} - \frac{Y}{2})^2 + \frac{3}{32}(\sum P_t y_t)^2 + \frac{3}{128}(\sum P_t^2)^2 Y^2$$

$$+ \frac{1}{64}(\sum P_t^2)(\sum P_t y_t)Y \ . \qquad (4.97).$$

On the other hand, for our sampling procedure, from (4.66)

$$V(\hat{Y}) = \sum^{N} P_j(1 - \frac{P_j}{2})(\frac{y_j}{P_j} - \frac{Y}{2})^2 - \frac{1}{2} \sum^{N} (P_j^3 - \frac{P_j^2 \sum P_t^2}{4})$$

$$\cdot (\frac{y_j}{P_j} - \frac{Y}{2})^2 + \frac{1}{4}(\sum P_t y_t)^2 + \frac{1}{16}(\sum P_t^2)^2 Y^2$$

$$- \frac{1}{4}(\sum P_t^2)(\sum P_t y_t)Y \qquad (4.98)$$

to $O(N^0)$. Equations (4.97) and (4.98) differ in their last three terms which are $O(N^0)$, and it is not quite clear which variance is smaller and this may depend on the structure of the P_j and y_j values.

V. THE GENERAL CASE $n \geq 2$ AND N LARGE

Since the methods to be employed for $n > 2$ are similar to those used for $n = 2$, we shall briefly describe these methods but concentrate on the new features that are not encountered in the case $n = 2$.

A. Derivation of the Probabilities $P_{ii'}$ to Orders $O(N^{-3})$ and $O(N^{-4})$

As before, the total number of arrangements N! can be divided into (N - 1) groups according as to whether there are $v = 0, 1, \ldots, (N - 2)$ units "between" P_i and $P_{i'}$. There are $N \times (N - 2)!$ arrangements in each of these (N - 1) groups so that all of these arrangements are represented with equal probability $\frac{1}{N - 1}$. Consider now the contribution to $P_{ii'}$ from a particular group with v units between P_i and $P_{i'}$. For the i^{th} unit to be in the sample, we know from our sampling procedure, the inequalities

$$\pi_{i-1} \leq s + k < \pi_i \tag{5.1}$$

must be satisfied where k may be any integer between $-(n - 1)$ and $(n - 1)$ and s is a uniform variate with $0 \leq s < n$. This means that s must be within one of the following ranges each of length P_i. The first of these is $\pi_{i-1} \leq s < \pi_i$ and the other ranges are displaced from the above range in the anti-clockwise direction by 1 or 2 ... or $(n - 1)$ according as $\pi_{i-1} \geq 1$ or $\pi_{i-1} \geq 2$ ... or $\pi_{i-1} \geq (n - 1)$ or in the clockwise direction according as $\pi_i \leq 1$ or $\pi_i \leq 2$ ... or $\pi_i \leq (n - 1)$ respectively. All these ranges make contributions to $P_{ii'}$ identical to that from the range $\pi_{i-1} \leq s \leq \pi_i$ since the length of the range of s is equal to P_i in all the cases. Therefore, we have to evaluate only the contribution to $P_{ii'}$ from the first range $\pi_{i-1} \leq s < \pi_i$, say $P'_{ii'}$, those from the remaining (n - 1) ranges being identical.

A positive contribution to $P'_{ii'}$ can only be made if both $\pi_{i-1} \leq s < \pi_i$ and one of the following (n - 1) inequalities is satisfied at the same time:

Inequality (1). $\pi_i + T_v \leq s + 1 < \pi_i + T_v + P_{i'}$

Inequality (2). $\pi_i + T_v \leq s + 2 < \pi_i + T_v + P_{i'}$

$\vdots$

Inequality (j). $\pi_i + T_v \leq s + j < \pi_i + T_v + P_{i'}$

$\vdots$

Inequality (n - 1). $\pi_i + T_v \leq s + (n - 1) < \pi_i + T_v + P_{i'}$

(5.2)

where T_v is the total length of the v arcs which lie "between" P_i and $P_{i'}$ in clockwise direction. This means that we consider the probability that the given i^{th} unit is drawn for k = 0 and i'^{th} unit is drawn for either k = 1 or k = 2 ... or k = (n - 1). Making the transformation

$$t = s - \pi_{i-1} = s + P_i - \pi_i \qquad (5.3)$$

so that the first range is

$$0 \leq t < P_i \qquad (5.4)$$

where t is a uniform variate with ordinate density 1/n like s, equations (5.2) can be written as

$$\text{Inequality (1).}\quad 1 + t - P_i - P_{i'} < T_v \leq 1 + t - P_i$$

$$\text{Inequality (2).}\quad 2 + t - P_i - P_{i'} < T_v \leq 2 + t - P_i$$

$$\vdots$$

$$\text{Inequality (j).}\quad j + t - P_i - P_{i'} < T_v \leq j + t - P_i$$

$$\vdots$$

$$\text{Inequality (n - 1).}\quad (n - 1) + t - P_i - P_{i'} < T_v \leq (n - 1) + t - P_i \; . \tag{5.5}$$

Therefore, the integrated contribution to P'_{ii} from inequality (j) is

$$\frac{1}{n}\int_0^{P_i} Pr.(j + t - P_i - P_{i'} < T_v \leq j + t - P_i)dt \; . \tag{5.6}$$

If the i'^{th} unit is drawn for $k = j$, then from inequality (j) of (5.2), it is seen that v ranges from (j - 1) to (N - n + j - 1) since i^{th} unit is drawn for k = 0 and each $P_r \leq 1$. Therefore, summing over the appropriate ranges of v for these (n - 1) different cases, and multiplying by the constant probability 1/(N - 1), the total integrated contribution to P'_{ii} is seen to be

$$P'_{ii'} = \frac{1}{n(N - 1)}\left\{\sum_{v=0}^{N-n}\int_0^{P_i} Pr.\left[1 + t - P_i - P_{i'} < T_v \leq 1 + t - P_i\right]dt + \ldots\right.$$

$$+ \sum_{v=m}^{N-n+m} \int_0^{P_i} \Pr.\Big[m + 1 + t - P_i - P_{i'} < T_v \leq m + 1$$

$$+ t - P_i\Big] dt + \ldots$$

$$+ \sum_{v=n-k}^{N-k} \int_0^{P_i} \Pr.\Big[(n - 1) + t - P_i - P_{i'} < T_v \leq$$

$$(n - 1) + t - P_i\Big] dt \Big\} . \qquad (5.7)$$

Adding now the contributions to $P_{ii'}$ from all the remaining $(n - 1)$ ranges which are identical with (5.7), we find the total contribution to $P_{ii'}$ as

$$P_{ii'} = (N - 1)^{-1} \Big\{ \sum_{v=0}^{N-n} \int_0^{P_i} \Big[F_v(1 + t - P_i)$$

$$- F_v(1 + t - P_i - P_{i'})\Big] dt + \ldots$$

$$+ \sum_{v=m}^{N-n+m} \int_0^{P_i} \Big[F_v(m + 1 + t - P_i)$$

$$- F_v(m + 1 + t - P_i - P_{i'})\Big] dt + \ldots$$

$$+ \sum_{v=n-k}^{N-k} \int_0^{P_i} \Big[F_v(n - 1 + t - P_i)$$

$$- F_v(n - 1 + t - P_i - P_{i'})\Big] dt \Big\} \cdot \qquad (5.8)$$

where $F_v(T)$ denotes the cumulative distribution function of the total (T_v) of the v values P_r. As before

$$E(T_v) = \frac{v(2 - P_i - P_{i'})}{N - 2}$$

$$V(T_v) = v(1 - \frac{v}{N - 2})S^2_{ii'} \qquad (5.9)$$

where $S^2_{ii'}$ is given by (4.10). It may be noted that $P_{ii'}$ given by (5.8) reduces to $P_{ii'}$ given by (4.11) in the special case $n = 2$. It will be shown below that each of the $(n - 1)$ integrals summed over v in (5.8) contribute identically to $P_{ii'}$ to $O(N^{-4})$ assuming that $P_i = np_i$ is $O(N^{-1})$.

Let us consider the m^th term $(m = 0, 1, \ldots, n - 2)$ in (5.8), say $P^{(m)}_{ii'}$, given by

$$P^{(m)}_{ii'} = (N - 1)^{-1} \sum_{v=m}^{N-n+m} \int_0^{P_i} \Big[F_v(m + 1 + t - P_i)$$

$$- F_v(m + 1 + t - P_i - P_{i'})\Big]dt \qquad (5.10)$$

and let $i = 1$ and $i' = 2$ without loss of generality. Proceeding now exactly as in the case of $n = 2$, by expanding $F_v(T)$ in an Edgeworth series and applying Euler-Maclaurin formula (4.21) twice, and approximating $\sum\limits_v$ by $\int dv$, we find

$$P^{(m)}_{12} = (N - 1)^{-1} \int_m^{N-n+m} \Big\{ \frac{P_1P_2}{S_{12}} \cdot v_1^{-\frac{1}{2}} p^{(1)}(v_2)$$

$$+ \frac{P_1P_2^3}{24S^3_{12}} \cdot v_1^{-\frac{3}{2}} p^{(3)}(v_2) + \frac{P_1^3P_2}{24S_{12}} \cdot v_1^{-\frac{1}{2}} p^{(3)}(v_2)$$

$$- \frac{k_3P_1P_2}{3S_{12}} \cdot v_1^{-\frac{1}{2}} p^{(4)}(v_2)\Big\}dv + \rho_m + \omega_m + \rho'_m \qquad (5.11)$$

where

$$v_1 = v(1 - \frac{v}{N - 2}) \tag{5.12}$$

$$v_2 = \frac{m + 1 - \frac{P_1 + P_2}{2} - v \cdot \frac{n - P_1 - P_2}{N - 2}}{v_1 S_{12}} \tag{5.13}$$

ρ_m, ω_m and ρ_m' are the remainder terms defined exactly as in the case n = 2, $P^{(r)}(x)$ denotes the r^{th} order derivative of the normal cumulative distribution $P(x)$, and k_3 is the standardized cumulant of the total T_v given by (4.17). Note that v_2 depends on m.

Let us now evaluate the terms in (5.11) one by one. The first term is

$$A_m = (N - 1)^{-1} \frac{P_1 P_2}{S_{12}} \int_m^{N-n+m} v_1^{-\frac{1}{2}} P^{(1)}(v_2) dv \tag{5.14}$$

where

$$P^{(1)}(v_2) = (2\pi)^{-\frac{1}{2}} e^{-\frac{1}{2}v_2^2} . \tag{5.15}$$

Now make the transformation

$$v - c = u \tag{5.16}$$

where

$$c = (N - 2) \cdot \frac{2(m + 1) - P_1 - P_2}{2(n - P_1 - P_2)} . \tag{5.17}$$

Then

$$v_1 = \left(c - \frac{c^2}{N - 2}\right)\left[1 + \frac{u\left(1 - \frac{2c}{N - 2}\right)}{c - \frac{c^2}{N - 2}} - \frac{u^2}{(N - 2)\left(c - \frac{c^2}{N - 2}\right)}\right]. \tag{5.18}$$

For the case $n = 2$, $m = 0$ so that

$$c = \frac{N - 2}{2} \text{ and } v_1 = \frac{(N - 2)}{4}\left(1 - \frac{4u^2}{(N - 2)^2}\right). \tag{5.19}$$

Now, in order to expand $v_1^{-\frac{1}{2}}$ binomially, it is necessary to show that

$$F = \frac{u\left(1 - \frac{2c}{N - 2}\right)}{c - \frac{c^2}{N - 2}} - \frac{u^2}{(N - 2)\left(c - \frac{c^2}{N - 2}\right)} \tag{5.20}$$

is less than one in absolute value for all u ranging from $m - c$ to $N - n + m - c$. This is immediately seen to be true for $n = 2$ since u ranges from $-(N - 2)/2$ to $(N - 2)/2$ and $F = -4u^2/(N - 2)^2$. Now at $u = m - c$, (5.20) reduces to

$$F = -1 + \frac{(n - P_1 - P_2)^2 m(N - 2 - m)}{(N - 2)^2\left(m + 1 - \frac{P_1 + P_2}{2}\right)\left(n - m - 1 - \frac{P_1 + P_2}{2}\right)} \tag{5.21}$$

which is less than 1 in absolute value. Also for any value between $m - c$ and 0, say $m - c + e$ with $e > 0$,

$$F = -1 + \frac{(n - P_1 - P_2)^2(m + e)(N - 2 - m - e)}{(N - 2)^2\left(m + 1 - \frac{P_1 + P_2}{2}\right)\left(n - m - 1 - \frac{P_1 + P_2}{2}\right)} \tag{5.22}$$

which is less than 1 in absolute value. Similarly, at u = N - n + m - c,

$$F = -1 + \frac{(n - P_1 - P_2)^2 (n - m - 2)(N - n + m)}{(N - 2)^2 (m + 1 - \frac{P_1 + P_2}{2})(n - m - 1 - \frac{P_1 + P_2}{2})} \tag{5.23}$$

which is less than 1 in absolute value, and for any value between 0 and N - n + m - c, say N - n + m - c - e,

$$F = -1 + \frac{(n - P_1 - P_2)^2 (n - m - 2 + e)(N - n + m - e)}{(N - 2)^2 (m + 1 - \frac{P_1 + P_2}{2})(n - m - 1 - \frac{P_1 + P_2}{2})} \tag{5.24}$$

which is less than 1 in absolute value. Hence, F is less than 1 in absolute value for all values of u ranging from m - c to N - n + m - c.

Now, as in the case n = 2, expanding the exponential in (5.16) as well as $v_1^{-\frac{1}{2}}$ in terms of u binomially, and changing the variable of integration u to p where

$$p = uh(N - 2)^{-\frac{1}{2}} (c - \frac{c^2}{N - 2})^{-\frac{1}{2}} \tag{5.25}$$

where

$$h = (n - P_1 - P_2)(N - 2)^{-\frac{1}{2}} S_{12}^{-1} \tag{5.26}$$

we find after considerable simplification

$$A_m = \frac{(N - 2)}{(N - 1)} \frac{P_1P_2}{(n - P_1 - P_2)} (2\pi)^{-\frac{1}{2}} \int_{-\infty}^{\infty} e^{-\frac{p^2}{2}}$$

$$\cdot \Big[1 + h^{-2}(p^2 - p^4) + \frac{1}{2} h_1(\frac{3}{4} p^2 - \frac{3}{2} p^4 + \frac{1}{4} p^6)$$

$$+ h^{-4}(\frac{3}{8} p^4 - \frac{3}{4} p^6 + \frac{1}{8} p^8) + h_1^2h^{-2}(\frac{15}{16} p^4 - \frac{45}{16} p^6$$

$$+ \frac{15}{16} p^8 - \frac{1}{16} p^{10}) + h_1^4(\frac{35}{128} p^4 - \frac{35}{32} p^6 + \frac{35}{64} p^8$$

$$- \frac{7}{96} p^{10} + \frac{1}{384} p^{12}) + \text{higher terms}\Big]dp \qquad (5.27)$$

where

$$h_1 = \frac{(N - 2)S_{12}\left(1 - \frac{2c}{N - 2}\right)}{(n - P_1 - P_2)(c - \frac{c^2}{N - 2})^{\frac{1}{2}}} \qquad (5.28)$$

and the limits of integration in (5.27) are respectively $h(m - c)(N - 2)^{-\frac{1}{2}} (c - \frac{c^2}{N - 2})^{-\frac{1}{2}}$ and $h(N - n + m - c)(N - 2)^{-\frac{1}{2}} \cdot (c - \frac{c^2}{N - 2})^{-\frac{1}{2}}$. These integration limits are respectively $-O(N^{\frac{1}{2}})$ and $O(N^{\frac{1}{2}})$ so that these can be replaced by $-\infty$ and $+\infty$ apart from errors which are $O(e^{-N_{K}a})$. The main feature here is the appearance of a noncentrality type parameter h_1 which depends on m and is zero when n = 2. However, it will be shown now that the coefficients corresponding to terms involving h_1 are zero so that all the terms A_m contribute identically to P_{12}. Using the standardized normal moments

$$\mu_4 = 3,\ \mu_6 = 15,\ \mu_8 = 105,\ \mu_{10} = 945,\ \mu_{12} = 10395$$

$$\text{and } \mu_{2r+1} = 0,\ r = 1, 2, 3, 4 \qquad (5.29)$$

we find from (5.27),

$$A_m = \frac{(N - 2)}{(N - 1)} \frac{P_1 P_2}{(n - P_1 - P_2)} (1 - h^{-2} + 3h^{-4}) \qquad (5.30)$$

to $O(N^{-4})$, which shows that A_m is independent of m since h does not depend on m. Similar analysis for the second term

$$B_m = (N - 1)^{-1} \frac{P_1 P_2^3}{24 S_{12}^3} \int_m^{N-n+m} v_1^{-\frac{3}{2}} p^{(3)}(v_2) dv \qquad (5.31)$$

where

$$p^{(3)}(v_2) = (2\pi)^{-\frac{1}{2}} e^{-\frac{1}{2}v_2^2} (v_2^2 - 1) \qquad (5.32)$$

shows that

$$B_m = \frac{(N - 2)}{(N - 1)} \frac{P_1 P_2^3 S_{12}^{-2}}{24(n - P_1 - P_2)} (2\pi)^{-\frac{1}{2}} \left(c - \frac{c^2}{N - 2}\right)^{-1}$$

$$\cdot \int_{-\infty}^{\infty} e^{-\frac{1}{2}p^2} \Big[(p^2 - 1) + \frac{1}{2} h_1 (p^5 - 6p^3 + 3p)$$

$$- \frac{1}{2} h^{-2} (p^6 - 6p^4 + 3p^2) + \frac{1}{8} h_1^2 (p^8 - 15p^6 + 45p^4 - 15p^2)$$

$$+ \text{higher terms}\Big] dp \qquad (5.33)$$

which is seen to be zero to $O(N^{-4})$ using the normal moments (5.29) and hence B_m does not contribute to P_{12} to $O(N^{-4})$. Similarly, we find that the next term

$$C_m = (N - 1)^{-1} \frac{P_1^3 P_2}{24 S_{12}} \int_m^{N-n+m} v_1^{-\frac{1}{2}} p^{(3)}(v_2) dv \qquad (5.34)$$

is reduced to

$$C_m = \frac{(N - 2)}{(N - 1)} \frac{P_1^3 P_2}{24(n - P_1 - P_2)} (2\pi)^{-\frac{1}{2}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}p^2}$$

$$\cdot \Big[(p^2 - 1) + \frac{1}{2} h_1(p^5 - 4p^3 + p) - \frac{1}{2} h^{-2}(p^6 - 4p^4 + p^2)$$

$$+ \frac{1}{8} h_1^2(p^8 - 11p^6 + 21p^4 - 3p^2) + \text{higher terms}\Big] dp \, . \qquad (5.35)$$

Using the normal moments, the evaluation of the terms retained in (5.35) yields

$$C_m = - \frac{(N - 2)}{(N - 1)} \frac{P_1^3 P_2 h^{-2}}{12(n - P_1 - P_2)} \qquad (5.36)$$

which is $O(N^{-5})$ since h^{-2} is $O(N^{-1})$ and hence C_m does not contribute to P_{12} to $O(N^{-4})$. The next term is

$$D_m = - (N - 1)^{-1} \frac{P_1 P_2}{6 S_{12}} \int_m^{N-n+m} k_3 v_1^{-\frac{1}{2}} p^{(4)}(v_2) dv \qquad (5.37)$$

where

$$-p^{(4)}(v_2) = (2\pi)^{-\frac{1}{2}} e^{-\frac{1}{2}v_2^2} (v_2^2 - 3v_2) \qquad (5.38)$$

and

$$k_3 v_1^{-\frac{1}{2}} = \left[v_1^{-1} \left(1 - \frac{v}{N - 2}\right) - (N - 2)^{-1} \left(1 - \frac{v}{N - 2}\right)^{-1} \right] K_3 \; . \tag{5.39}$$

Now, making the transformations u and p, expanding $k_3 v_1^{-\frac{1}{2}}$, $(v_2^3 - 3v_2)$ and the exponential in (5.38) in terms of p, and multiplying out the resulting series, we find after considerable simplification

$$D_m = - \frac{(N - 2)}{(N - 1)} \frac{P_1 P_2 K_3}{6(n - P_1 - P_2)} (2\pi)^{-\frac{1}{2}}$$

$$\cdot \int_{-\infty}^{\infty} e^{-\frac{1}{2}p^2} \left\{ \frac{\left(1 - \frac{2c}{N - 2}\right)}{\left(c - \frac{c^2}{N - 2}\right)^{\frac{1}{2}}} \left[(p^3 - 3p) \right.\right.$$

$$+ \frac{1}{2} h_1 (p^6 - 6p^4 + 3p^2) - \frac{1}{2} h^{-2} (p^7 - 6p^5 + 3p^3)$$

$$+ \frac{1}{8} h_1^2 (p^9 - 13p^7 + 33p^5 - 9p^3) - \frac{1}{4} h_1 h^{-2} (p^{10} - 13p^8$$

$$+ 33p^6 - 9p^4) + \frac{1}{48} h_1^3 (p^{12} - 24p^{10} + 150p^8 - 240p^6$$

$$\left. + 45p^4) \right] - n^{-1} (N - 2)^{\frac{1}{2}} \left(c - \frac{c^2}{N - 2}\right)$$

$$\cdot \left[\frac{1}{c^2} + \frac{1}{(N - 2)^2 \left(1 - \frac{c}{N - 2}\right)^2} \right] \left[(p^4 - 3p^2) \right.$$

$$+ \frac{1}{2} h_1 (p^7 - 6p^5 + 3p^3) - \frac{1}{2} h^{-2} (p^8 - 6p^6 + 3p^4)$$

$$+ \frac{1}{8} h_1^2(p^{10} - 13p^8 + 33p^6 - 9p^4)\Big] + h^{-2}(N - 2)(c - \frac{c^2}{N - 2})^{\frac{3}{2}}$$

$$\cdot\Big[\frac{1}{c^3} - \frac{1}{(N - 2)^3(1 - \frac{c}{N - 2})^3}\Big]\Big[(p^5 - 3p^3)$$

$$+ \frac{1}{2} h_1(p^8 - 6p^6 + 3p^4)\Big] - h^{-3}(N - 2)^{\frac{3}{2}}(c - \frac{c^2}{N - 2})^2$$

$$\cdot\Big[\frac{1}{c^4} + \frac{1}{(N - 2)^4(1 - \frac{c}{N - 2})^4}\Big](p^6 - 3p^4)$$

$$+ \text{higher terms}\Big\}dp \;. \qquad (5.40)$$

Using now the standardized normal moments (5.29), the evaluation of the terms retained in (5.40) yields

$$D_m = - \frac{(N - 2)}{(N - 1)} \frac{P_1P_2K_3}{8(n - P_1 - P_2)} h^{-3}(N - 2)^{\frac{3}{2}}$$

$$\cdot\Big[- \frac{12(1 - \frac{2c}{N - 2})^2}{(N - 2)(c - \frac{c^2}{N - 2})} - 6(1 - \frac{2c}{N - 2})^2$$

$$\cdot\Big\{\frac{1}{c^2} + \frac{1}{(N - 2)^2(1 - \frac{c}{N - 2})^2}\Big\} + 6c^{-2}(1 - \frac{c}{N - 2})^2$$

$$+ \frac{6c^2}{(N - 2)^4(1 - \frac{c}{N - 2})^2}\Big] \;. \qquad (5.41)$$

Further simplification of (5.41) results in

$$D_m = - \frac{2(N - 2)}{(N - 1)} \frac{P_1P_2K_3}{(n - P_1 - P_2)} h^{-3}(N - 2)^{-\frac{1}{2}} \qquad (5.42)$$

which is $O(N^{-4})$ and does not depend on m.

The argument to show that the remainder terms ρ_m, ω_m and ρ'_m do not contribute to P_{12} to $O(N^{-4})$ is similar to that given in the case n = 2, for the remainder terms ρ, ω and ρ'. Therefore, adding the expressions A_m and D_m (since B_m and C_m are zero to $O(N^{-4})$) given by (5.30) and (5.42) respectively, we find to $O(N^{-4})$,

$$P_{12}^{(m)} = \frac{(N-2)}{(N-1)} \frac{P_1 P_2}{(n - P_1 - P_2)} \left[1 - h^{-2} + 3h^{-4} - 2K_3 h^{-3}(N-2)^{-\frac{1}{2}}\right]. \tag{5.43}$$

Since (5.43) does not depend on m, it follows that, to $O(N^{-4})$,

$$P_{12} = \sum_{m=0}^{n-2} P_{12}^{(m)}$$

$$= (n-1) \frac{(N-2)}{(N-1)} \frac{P_1 P_2}{(n - P_1 - P_2)}$$

$$\cdot \left[1 - h^{-2} + 3h^{-4} - 2K_3 h^{-3}(N-2)^{-\frac{1}{2}}\right]. \tag{5.44}$$

For the special case n = 2, (5.44) reduces to (4.47) derived in Chapter IV. Since the last two terms in (5.44) are $O(N^{-4})$, we obtain to $O(N^{-3})$, the simplified expression

$$P_{12} = (n-1) \frac{(N-2)}{(N-1)} \frac{P_1 P_2}{(n - P_1 - P_2)} (1 - h^{-2}) . \tag{5.45}$$

As in the case of n = 2, we can apply the two checks to test the order of (5.44). In the first check, when all the probabilities P_i are equal to n/N, (5.44) for P_{12}

reduces to n(n - 1)/N(N - 1) which is the correct probability for two units to be in a sample of size n drawn with equal probabilities and without replacement. The second check is to test that

$$\sum_{i' \neq i}^{N} P_{ii'} = (n - 1)P_i \qquad (5.46)$$

is satisfied to $O(N^{-3})$ when (5.44) is substituted in (5.46) where the suffixes 1 and 2 are replaced by i and i' respectively. Now, proceeding exactly as in the case n = 2, (5.44) to $O(N^{-4})$ can be simplified as

$$\begin{aligned} P_{ii'} = {} & \frac{(n - 1)}{n} P_i P_{i'} + \frac{(n - 1)}{n^2} (P_i^2 P_{i'} + P_i P_{i'}^2) \\ & - \frac{(n - 1)}{n^3} P_i P_{i'} \sum P_t^2 + \frac{(n - 1)}{n^3} (2P_i^3 P_{i'} + 2P_i P_{i'}^3 \\ & + 2P_i^2 P_{i'}^2) - \frac{3(n - 1)}{n^4} (P_i^2 P_{i'} + P_i P_{i'}^2) \sum P_t^2 \\ & + \frac{3(n - 1)}{n^5} P_i P_{i'} (\sum P_t^2)^2 - \frac{2(n - 1)}{n^4} P_i P_{i'} \sum P_t^3 \end{aligned} \qquad (5.47)$$

where P_1 and P_2 are replaced by P_i and $P_{i'}$ respectively. Summing (5.47) now over i' from 1 to N except i' = i, and noting that $\sum P_t = n$, we obtain to $O(N^{-3})$,

$$\sum_{i'\neq i}^{N} P_{ii'} = \frac{(n-1)}{n} P_i(n - P_i) + \frac{(n-1)}{n^2} P_i^2(n - P_i)$$

$$+ \frac{(n-1)}{n^2} P_i\left(\sum P_t^2 - P_i^2\right) + \frac{2(n-1)}{n^2} P_i^3$$

$$- \frac{(n-1)}{n^3} P_i^2 \sum P_t^2 - \frac{(n-1)}{n^3} P_i(n - P_i) \sum P_t^2$$

$$= (n-1)P_i \qquad (5.48)$$

thereby providing the desired check.

B. Variance Formulas to Orders $O(N^1)$ and $O(N^0)$

Substituting for $P_{ii'}$ from (5.47) in

$$V(\hat{Y}) = \sum^{N} \frac{y_j^2}{P_j} + \sum_{i\neq i'}^{N} \frac{P_{ii'}}{P_i P_{i'}} y_i y_{i'} - Y^2 \qquad (5.49)$$

and retaining terms to $O(N^0)$, we find

$$V(\hat{Y}) = \sum \frac{y_j^2}{P_j} - \frac{Y^2}{n} - \frac{(n-1)}{n} \sum y_j^2 + \frac{2(n-1)}{n^2} \left(\sum P_j y_j\right)Y$$

$$- \frac{(n-1)}{n^3} \left(\sum P_t^2\right)Y^2 - \frac{2(n-1)}{n^2} \sum P_j y_j^2$$

$$+ \frac{(n-1)}{n^3} \left(\sum P_t^2\right)\left(\sum y_j^2\right) - \frac{6(n-1)}{n^4}\left(\sum P_t^2\right)\left(\sum P_j y_j\right)Y$$

$$+ \frac{4(n-1)}{n^3}\left(\sum P_j^2 y_j\right)Y + \frac{3(n-1)}{n^5}\left(\sum P_t^2\right)^2 Y^2$$

$$- \frac{2(n-1)}{n^4} Y^2\left(\sum P_t^3\right) + \frac{2(n-1)}{n^3} \left(\sum P_j y_j^2\right)$$

$$= \sum^{N} P_j \left[1 - \frac{(n-1)}{n} P_j\right]\left(\frac{y_j}{P_j} - \frac{Y}{n}\right)^2$$

$$- \frac{(n-1)}{n^2} \sum^{N} \left(2P_j^3 - \frac{P_j^2}{n} \sum^{N} P_t^2\right)\left(\frac{y_j}{P_j} - \frac{Y}{n}\right)^2$$

$$+ \frac{2(n-1)}{n^3} \left(\sum^{N} P_j y_j - \frac{Y}{n} \sum^{N} P_t^2\right)^2 \qquad (5.51)$$

to $O(N^0)$. On the other hand, if terms only to $O(N^1)$ are retained, from (5.50) we find to $O(N^1)$, the simplified expression

$$V(\hat{Y}) = \sum \frac{y_j^2}{P_j} - \frac{Y^2}{n} - \frac{(n-1)}{n} \sum y_j^2 + \frac{2(n-1)}{n^2} \left(\sum P_j y_j\right) Y$$

$$- \frac{(n-1)}{n^3} \left(\sum P_t^2\right) Y^2 \qquad (5.52)$$

$$= \sum^{N} P_j \left[1 - \frac{(n-1)}{n} P_j\right]\left(\frac{y_j}{P_j} - \frac{Y}{n}\right)^2 . \qquad (5.53)$$

Equation (5.53) shows the characteristic reduction in the variance when compared with the variance in sampling with replacement, through the "finite population corrections" $(1 - \frac{(n-1)}{n} P_j)$. Hence, the present sampling procedure without replacement yields a smaller variance for $\hat{Y}$ asymptotically compared with unequal probability sampling with replacement, for the general sample size n. For the special case of equal probabilities $P_j = n/N$, (5.51) to $O(N^0)$ reduces to the familiar variance formula for sample total in equal probability sampling without replacement.

C. Estimation of the Variance

The method is, as before, to substitute for $P_{ii'}$ in the Yates and Grundy estimate of the variance

$$v_{YG}(\hat{Y}) = \sum_{i'>i}^{n} \frac{P_iP_{i'} - P_{ii'}}{P_{ii'}} \left(\frac{y_i}{P_i} - \frac{y_{i'}}{P_{i'}}\right)^2 . \qquad (5.54)$$

From (5.47) to $O(N^{-3})$, we have

$$P_{ii'} = \frac{(n-1)}{n} P_iP_{i'}\left[1 + \frac{1}{n}(P_i + P_{i'}) - \frac{1}{n^2} \sum P_t^2\right] . \qquad (5.55)$$

Therefore, substituting (5.55) in (5.54), we find

$$v_{YG}(\hat{Y}) = (n-1)^{-1} \sum_{i'>i}^{n} \frac{1 - \frac{(n-1)}{n}(P_i + P_{i'}) + \frac{(n-1)}{n^2} \sum P_t^2}{1 + \frac{1}{n}(P_i + P_{i'}) - \frac{1}{n^2} \sum P_t^2}$$

$$\cdot \left(\frac{y_i}{P_i} - \frac{y_{i'}}{P_{i'}}\right)^2 . \qquad (5.56)$$

Expanding the denominator binomially and retaining terms to $O(N^{1})$, we find

$$v_{YG}(\hat{Y}) = (n-1)^{-1} \sum_{i'>i}^{n} \left(1 - P_i - P_{i'} + \frac{1}{n} \sum^{N} P_t^2\right)\left(\frac{y_i}{P_i} - \frac{y_{i'}}{P_{i'}}\right)^2 \qquad (5.57)$$

to $O(N^{1})$. For the special case of equal probabilities $P_i = n/N$, (5.57) agrees with the familiar formula for the estimate of the variance in equal probability sampling without replacement, noting that

$$\sum_{i'>i}^{n} (y_i - y_{i'})^2 = n \sum^{n} (y_i - \bar{y})^2 . \qquad (5.58)$$

On the other hand, by substituting for $P_{ii'}$ from (5.47) in (5.54) and expanding the denominator binomially and retaining terms to $O(N^0)$, we obtain

$$v_{YG}(\hat{Y}) = (n - 1)^{-1} \sum_{i'>i}^{n} \Big[1 - (P_i + P_{i'}) + \frac{1}{n} \sum^{N} P_t^2$$
$$- \frac{1}{n}(P_i^2 + P_{i'}^2) - \frac{2}{n^3} \Big(\sum^{N} P_t^2\Big)^2$$
$$+ \frac{1}{n^2}(P_i + P_{i'}) \sum^{N} P_t^2 + \frac{2}{n^2} \sum^{N} P_t^3 \Big]\Big(\frac{y_i}{P_i} - \frac{y_{i'}}{P_{i'}}\Big)^2 \qquad (5.59)$$

to $O(N^0)$, which agrees with the estimate of the variance in equal probability sampling without replacement, when all $P_i = n/N$.

D. Comparison with the Method of Revised Probabilities of Yates and Grundy

It is shown here for the case of general sample size n, that the $P_{ii'}$ values attained through the Yates and Grundy procedure of revised probabilities to ensure $P_j = np_j$, and through our procedure are exactly the same to $O(N^{-3})$, so that $V(\hat{Y})$ is the same for both procedures to $O(N^1)$. We shall not evaluate here the $P_{ii'}$ values to $O(N^{-4})$ for the Yates and

Grundy procedure as was done in the case $n = 2$, since the evaluation seems to involve heavy algebra.

Now, from (5.47), the probability of selecting the units i and i' in a sample of size n for our procedure, to $O(N^{-3})$ is

$$\begin{aligned} P_{ii'} &= n(n-1)p_ip_{i'} + n(n-1)(p_i^2p_{i'} + p_ip_{i'}^2) \\ &\quad - n(n-1)p_ip_{i'} \sum^{N} p_t^2 \end{aligned} \tag{5.60}$$

since $P_i = np_i$. For the Yates and Grundy procedure, the probability for selecting the units i and i' is given by

$$\begin{aligned} P_{ii'}^{(a)} &= \left(\frac{p_i^*p_{i'}^*}{1-p_i^*} + \frac{p_i^*p_{i'}^*}{1-p_{i'}^*}\right) + \sum_{k=3}^{n} \Bigg\{ \sum_{\ell=1}^{k-1} \Bigg[\overbrace{\sum_{\substack{j \neq s \neq \cdots \\ \neq i \neq i'}} \sum \cdots \sum}^{(k-2)\text{ sums}} p_j^* \\ &\quad \cdot \frac{p_s^*}{(1-p_j^*)} \cdots \underbrace{\frac{p_i^*}{(1-p_j^*-p_s^*-\cdots)}}_{\ell^{\text{th}}\text{ position}} \\ &\quad \cdots \frac{p_{i'}^*}{(1-p_j^*-p_s^*\cdots-p_i^*\cdots)} \Bigg] \\ &\quad + \sum_{\ell=1}^{k-1} \Bigg[\overbrace{\sum_{\substack{j \neq s \neq \cdots \\ \neq i \neq i'}} \sum \cdots \sum}^{(k-2)\text{ sums}} p_j^* \cdot \frac{p_s^*}{(1-p_j^*)} \cdots \overbrace{\frac{p_{i'}^*}{(1-p_j-p_s-\cdots)}}^{\ell^{\text{th}}\text{ position}} \\ &\quad \cdots \frac{p_i^*}{(1-p_j^*-p_s^*\cdots-p_{i'}^*\cdots)} \Bigg] \Bigg\} \end{aligned} \tag{5.61}$$

and

$$P_i = p_i^* + p_i^* \sum_{j \neq i}^{N} \frac{p_j^*}{(1 - p_j^*)} + \sum_{k=3}^{n} \Bigg\{ \overbrace{\sum \sum \cdots \sum_{j \neq s \neq \cdots \neq i}}^{(k-1)\,\text{sums}} p_j^*$$

$$\cdot \frac{p_s^*}{(1 - p_j^*)} \cdots \frac{p_i^*}{(1 - p_j^* - p_s^* - \cdots)} \Bigg\} \qquad (5.62)$$

where p_i^* are the revised probabilities which ensure that $P_i = np_i$. Now, expanding the denominators in (5.62) binomially, we find after some algebra, to $O(N^{-2})$,

$$P_i = p_i^* \Big\{ 2 + (\sum p_t^{*2} - p_i^*) + \sum_{k=3}^{n} \Big[1 - (k - 1)p_i^*$$

$$+ (k - 1) \sum p_t^{*2} \Big] \Big\}$$

$$= np_i^* \left[1 - \frac{(n - 1)}{2} p_i^* + \frac{(n - 1)}{2} \sum p_t^{*2} \right] = np_i \ . \qquad (5.63)$$

Therefore

$$p_i^* = p_i \left[1 - \frac{(n - 1)}{2} p_i^* + \frac{(n - 1)}{2} \sum p_t^{*2} \right]^{-1}$$

$$= p_i \left[1 + \frac{(n - 1)}{2} p_i^* - \frac{(n - 1)}{2} \sum p_t^{*2} \right] \text{ to } O(N^{-2})$$

$$= p_i \left[1 + \frac{(n - 1)}{2} p_i - \frac{(n - 1)}{2} \sum p_t^{2} \right] \text{ to } O(N^{-2}) \quad (5.64)$$

since

$$p_i^* = p_i \left[1 + \text{terms of } O(N^{-1}) \right] \ . \qquad (5.65)$$

Further, expanding the denominators in (5.61) binomially, we obtain after considerable simplification, to $O(N^{-3})$,

$$\begin{aligned} P^{(a)}_{ii'} = {} & p_i^* p_{i'}^*(1 + p_i^*) + p_i^* p_{i'}^*(1 + p_{i'}^*) + p_i^* p_{i'}^* \sum_{k=3}^{n} \Big[2(k - 1) \\ & - (p_i^* + p_{i'}^*)\Big\{2(k - 2)(k - 1) - \frac{k(k - 1)}{2}\Big\} \\ & - \sum p_t^{*2}\Big\{(k - 1)(k - 2)(k - 3) - k(k - 1)(k - 2)\Big\}\Big] \end{aligned} \tag{5.66}$$

$$\begin{aligned} = {} & n(n - 1)p_i^* p_{i'}^* + (p_i^{*2} p_{i'}^* + p_i^* p_{i'}^{*2})\Big[\frac{(n - 1)n(n + 1)}{6} \\ & - \frac{2(n - 2)(n - 1)n}{3}\Big] + (n - 2)(n - 1)n \cdot p_i^* p_{i'}^* \sum p_t^{*2} \ . \end{aligned} \tag{5.67}$$

Substituting for p_i^* from (5.64) in (5.67), we finally obtain to $O(N^{-3})$,

$$\begin{aligned} P^{(a)}_{ii'} = {} & n(n - 1)p_i p_{i'} + (p_i^2 p_{i'} + p_i p_{i'}^2)\Big[\frac{(n - 1)n(n + 1)}{6} \\ & - \frac{2(n - 2)(n - 1)n}{3} + \frac{(n - 1)^2 n}{2}\Big] + (p_i p_{i'} \sum p_t^2) \\ & \cdot \Big[(n - 2)(n - 1)n - (n - 1)^2 n\Big] \\ = {} & n(n - 1)p_i p_{i'} + n(n - 1)(p_i^2 p_{i'} + p_i p_{i'}^2) \\ & - n(n - 1)p_i p_{i'} \sum p_t^2 \end{aligned} \tag{5.68}$$

which is exactly the same as the $P_{ii'}$ to $O(N^{-3})$ for our procedure, namely, equation (5.60).

E. A Comparison with Ratio Method of Estimation

It is of importance to make efficiency comparisons with alternative methods of utilizing supplementary information such as ratio and regression methods of estimation and stratification. The difficulties involved in such comparisons and the limitations of the available results in the literature have already been mentioned in Chapter II. As mentioned earlier, Cochran (1953) has compared the variance of the estimate in unequal probability sampling with replacement and the variance of the ratio estimate without the usual finite population correction factor. Since we have obtained a compact expression for the variance of the estimate $\hat{Y}$ in unequal probability sampling without replacement, namely (5.53), it will be of interest to compare this with the variance of the ratio estimate not ignoring the finite population correction factor. Now from (5.53),

$$V(\hat{Y}) = \frac{1}{n} \sum^{N} \frac{1}{p_j}(y_j - Yp_j)^2 - \frac{(n-1)}{n} \sum^{N} (y_j - Yp_j)^2 \qquad (5.69)$$

to $O(N^1)$, since $P_j = np_j$. On the other hand, the variance of the ratio estimate $\hat{Y}_R$ for large samples (ignoring its bias) is given by

$$V(\hat{Y}_R) = \frac{N^2}{n(N-1)} \cdot (1 - \frac{n}{N})\sum^{N}(y_j - Yp_j)^2 \qquad (5.70)$$

where $p_j = \frac{x_j}{X}$.

$$= \frac{N}{n}(1 + \frac{1}{N})(1 - \frac{n}{N})\sum^{N}(y_j - Yp_j)^2 \text{ to } O(N^1)$$

$$= \frac{N}{n}\sum^{N}(y_j - Yp_j)^2 - \frac{(n-1)}{n}\sum^{N}(y_j - Yp_j)^2$$

$$\text{to } O(N^1) \; . \qquad (5.71)$$

The first term of (5.69) represents the variance in unequal probability sampling with replacement. It is interesting to note from (5.69) and (5.71) that the finite population correction factors for $\hat{Y}$ and $\hat{Y}_R$ are exactly the same. Therefore, the comparison reduces to the comparison of the variance in unequal probability sampling with replacement and the variance of the ratio estimate without the correction factor, so that Cochran's results apply here. Assuming the model

$$y_j = Yp_j + e_j \tag{5.72}$$

where

$$E(e_j | p_j) = 0 \ ; \ E(e_j^2 | p_j) = ap_j^g \ , \quad a > 0, \ g > 0 \ . \tag{5.73}$$

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 in practice g usually lies between 1 and 2, so that the estimate $\hat{Y}$ is generally more precise than the ratio estimate $\hat{Y}_R$. We do not propose to investigate here further possibilities of efficiency comparisons with other methods of utilizing supplementary information, _e.g._ stratification.

VI. MISCELLANEOUS TOPICS IN UNEQUAL PROBABILITY SAMPLING

In Chapters III to V, we have developed the theory for a particular sampling procedure of unequal probability sampling without replacement, the advantages of which have already been described. We shall now discuss some interesting topics in unequal probability sampling in general.

A. A New Sampling System for which the Yates and Grundy Estimate of the Variance is Always Positive

As mentioned earlier in Chapter II, the Horvitz and Thompson estimate of the variance of $\hat{Y}$ can take negative values. The Yates and Grundy estimate of the variance of $\hat{Y}$ is given by

$$v_{YG}(\hat{Y}) = \sum_{i'>i}^{n} \frac{P_i P_{i'} - P_{ii'}}{P_{ii'}} \left(\frac{y_i}{P_i} - \frac{y_{i'}}{P_{i'}}\right)^2 \qquad (6.1)$$

and it is believed to be "less often negative". Also as mentioned earlier, the estimator (6.1) is always positive in the following two important situations:

(1) The first unit is selected with p.p.s., i.e. with probabilities p_i and the remaining (n - 1) units in the sample are selected with equal probabilities and without replacement.

(2) The first unit is selected with p.p.s. and the second unit is selected with p.p.s. of the remaining units, the sample size being two.

This means that the Yates and Grundy estimate of the variance is always positive whenever two units are drawn by the above plan (2) which is the one originally proposed by Horvitz and Thompson and also employed by Yates and Grundy.

It may be noted that for these two systems P_i is not proportional to p_i unless the revised probabilities p_i^* are introduced. We shall not be concerned here with the problem of making P_i proportional to p_i. It will be of interest to identify more sampling systems which yield simple expressions for P_i and $P_{ii'}$ as in the case of systems 1 and 2, and for which the Yates and Grundy estimate of the variance is always positive. We identify here a new sampling system with $n > 2$ which yields simple expressions for P_i and $P_{ii'}$ and for which the Yates and Grundy estimate of the variance is always positive. The sampling system is as follows:

(3) The first unit is selected with p.p.s., second unit with p.p.s. of the remaining units as in (2) and the remaining $(n - 2)$ units in the sample are selected with equal probabilities and without replacement.

Then, from the above description it follows that

$$P_i = p_i + p_i \sum_{j \neq i}^{N} \frac{p_j}{1 - p_j} + \sum_{j \neq k \neq i}^{N}\sum \frac{p_j p_k}{1 - p_j} \cdot \frac{n - 2}{N - 2} . \qquad (6.2)$$

Noting that $\sum^{N} p_t = 1$, (6.2) can be simplified as

$$P_i = \frac{(N - n)}{(N - 2)} p_i \left[\frac{1}{1 - p_{i'}} + A_{ii'} \right] + \frac{n - 2}{N - 2} \qquad (6.3)$$

where

$$A_{ii'} = \sum_{j \neq (i,i')}^{N} \frac{p_j}{1 - p_j} \ . \qquad (6.4)$$

Also

$$\begin{aligned} P_{ii'} = {} & p_i p_{i'} \left(\frac{1}{1 - p_i} + \frac{1}{1 - p_{i'}}\right) + \left(\sum_{j \neq (i,i')}^{N} p_j \right) \\ & \cdot \left(\frac{p_i}{1 - p_i} + \frac{p_{i'}}{1 - p_{i'}}\right) \frac{n - 2}{N - 2} \\ & + (p_i + p_{i'}) \left[\sum_{j \neq (i,i')}^{N} \frac{p_j}{1 - p_j} \right] \frac{n - 2}{N - 2} \\ & + \frac{(n - 2)(n - 3)}{(N - 2)(N - 3)} \sum_{\substack{j \neq j' \\ \neq (i,i')}}^{N} \sum^{N} \frac{p_j p_{j'}}{1 - p_j} \qquad (6.5) \end{aligned}$$

$$\begin{aligned} = {} & p_i p_{i'} \left(\frac{1}{1 - p_i} + \frac{1}{1 - p_{i'}}\right) \frac{N - n}{N - 2} + \frac{(n - 2)(N - n)}{(N - 2)(N - 3)} (p_i + p_{i'}) \\ & + \frac{(n - 2)(N - n)}{(N - 2)(N - 3)} (p_i + p_{i'}) A_{ii'} + \frac{(n - 2)(n - 3)}{(N - 2)(N - 3)} \ . \end{aligned} \qquad (6.6)$$

For the special case of equal probabilities $p_i = \frac{1}{N}$, (6.3) reduces to n/N and (6.5) to $n(n - 1)/N(N - 1)$ thus providing a check. Now $v_{YG}(\hat{Y})$ is always positive when $P_i P_{i'} - P_{ii'} > 0$ for every pair (i,i'). So it is sufficient if we prove for system 3 that

$$P_i P_{i'} - P_{ii'} > 0 \qquad (i \neq i' = 1, 2, \ldots, N) \ . \tag{6.7}$$

After some simplification, we find from (6.3) and (6.6) that

$$\begin{aligned} P_i P_{i'} - P_{ii'} = \frac{(N - n)}{(N - 2)^2} \Big[\frac{(n - 2)}{(N - 3)} \Big\{ &(1 - p_i - p_{i'}) \\ &- A_{ii'}(p_i + p_{i'}) \Big\} - \frac{p_i p_{i'}(1 - p_i - p_{i'})(N - n)}{(1 - p_i)(1 - p_{i'})} \\ &+ (N - n) A_{ii'} \frac{p_i p_{i'}(2 - p_i - p_{i'})}{(1 - p_i)(1 - p_{i'})} \\ &+ (N - n) p_i p_{i'} A_{ii'}^2 \Big] \ . \end{aligned} \tag{6.8}$$

Consider now the term

$$M = (1 - p_i - p_{i'}) - A_{ii'}(p_i + p_{i'}) \ . \tag{6.9}$$

Since

$$1 - p_j > p_i + p_{i'} \qquad \text{for } j \neq (i, i') \tag{6.10}$$

we have

$$\begin{aligned} A_{ii'}(p_i + p_{i'}) &= \sum_{j \neq (i,i')}^{N} \frac{p_j}{1 - p_j}(p_i + p_{i'}) < \sum_{j \neq (i,i')}^{N} p_j \\ &= 1 - p_i - p_{i'} \end{aligned} \tag{6.11}$$

so that

$$M > (1 - p_i - p_{i'}) - (1 - p_i - p_{i'}) = 0 \ . \tag{6.12}$$

Therefore

$$P_iP_{i'} - P_{ii'} > \frac{(N-n)}{(N-2)^2}\Big[(N-n)p_ip_{i'}A^2_{ii'} + (N-n)A_{ii'}$$

$$\cdot \frac{p_ip_{i'}(2-p_i-p_{i'})}{(1-p_i)(1-p_{i'})} - \frac{p_ip_{i'}(1-p_i-p_{i'})(N-n)}{(1-p_i)(1-p_{i'})}\Big]. \qquad (6.13)$$

To prove that (6.13) is greater than zero, one can use the proof of Sen (1953) and Des Raj (1956a) for system (2), which consists of finding the minimum of $A_{ii'}$ and substituting it in (6.13). However, we give below an elementary and simpler proof to show that (6.13) is greater than zero. This proof, of course, can be used as an alternative and simpler proof to show that the Yates and Grundy estimate of the variance is always positive for system (2). Since

$$A_{ii'} = \sum_{j\neq(i,i')}^{N} \frac{p_j}{1-p_j} > \sum_{j\neq(i,i')}^{N} p_j = 1 - p_i - p_{i'} \qquad (6.14)$$

by substituting for $A_{ii'}$ from (6.14) in (6.13), it follows that

$$P_iP_{i'} - P_{ii'} > \frac{(N-n)}{(N-2)^2}\Big[(N-n)p_ip_{i'}A^2_{ii'}$$

$$+ \frac{(N-n)p_ip_{i'}}{(1-p_i)(1-p_{i'})}(1-p_i-p_{i'})^2\Big] \qquad (6.15)$$

which is greater than zero. Hence, the Yates and Grundy estimate of the variance is always positive for sampling system (3).

B. Two Problems in Unequal Probability Sampling

1. Estimation of the efficiency of unequal probability sampling over equal probability sampling

It is of interest to estimate the gain in efficiency in using unequal probability sampling over equal probability sampling. The variance of the estimate of the total in equal probability sampling without replacement is

$$V(N\bar{y}) = \frac{N^2}{n}\left(1 - \frac{n}{N}\right)\frac{1}{N - 1}\left(\sum^{N} y_t^2 - \frac{Y^2}{N}\right). \qquad (6.16)$$

So, the problem is to estimate (6.16) from a sample drawn with unequal probabilities, specifically P_i is the probability for including the i^{th} unit in a sample of size n. Now

$$E\sum^{n}\frac{y_i^2}{p_i} = \sum^{N} y_t^2 . \qquad (6.17)$$

Also since

$$V(\hat{Y}) = E(\hat{Y}^2) - Y^2 \qquad (6.18)$$

where E denotes the expectation, it follows that

$$\text{Est. } Y^2 = \hat{Y}^2 - \text{Est. } V(\hat{Y}) \ . \qquad (6.19)$$

For the estimate of $V(\hat{Y})$, we use the Yates and Grundy estimate of the variance, $v_{YG}(\hat{Y})$. Therefore, an unbiased estimate of $V(N\bar{y})$ from the sample drawn with unequal probabilities is

$$v'(N\bar{y}) = \frac{N^2}{n}\left(1 - \frac{n}{N}\right)\frac{1}{N - 1}\left[\sum^{n}\frac{y_i^2}{P_i} - \frac{\hat{Y}^2}{N} + \frac{v_{YG}(\hat{Y})}{N}\right]. \qquad (6.20)$$

Comparing this with $v_{YG}(\hat{Y})$, an estimate of the percentage gain in efficiency in using unequal probability sampling over equal probability sampling is

$$\frac{v'(N\bar{y}) - v_{YG}(\hat{Y})}{v_{YG}(\hat{Y})} \times 100 \ . \qquad (6.21)$$

It may be noted that for the special case of equal probabilities $P_i = n/N$, (6.20) reduces to the familiar formula for the estimate of the variance in equal probability sampling without replacement. The above formulas are not, of course, intended to indicate for which populations $V(\hat{Y}) \leq V(N\bar{y})$ and for which populations the inequality is inverted. They are merely intended to provide estimates for the variances computed from data with unequal probability sampling. An example illustrating this is given below.

Example. Let us take the example of Horvitz and Thompson (1952), namely, the 20 blocks of Ames, Iowa, given in Table 1, Chapter III. Using our particular sampling procedure for $n = 2$, the units 5 and 14 are selected with probabilities proportional to size and without replacement, assuming that the ordering of the units given in Table 1 is random. The following values are obtained:

$$\hat{Y} = \frac{y_5}{P_5} + \frac{y_{14}}{P_{14}} = 481.34.$$

Using the formulas (4.73) and (4.75),

$$v_{YG}(\hat{Y}) = 15805 \text{ to } O(N^1)$$
$$v_{YG}(\hat{Y}) = 15777 \text{ to } O(N^0) \ .$$

These two values show that the approximation to $O(N^1)$ is quite satisfactory. Also from (6.20),

$$v'(N\bar{y}) = 69663 .$$

Therefore, an estimate of percentage gain in efficiency is equal to

$$100\,(\frac{69663}{15805} - 1) = 341 .$$

Obviously in this example the variance estimates based on sample size of two units are very unreliable. In practice, such estimates each computed from one of a large number of strata would be pooled.

2. Alternative estimators in unequal probability sampling

In most of the large scale sample surveys, we are usually interested in estimating the population totals or means of several characteristics. If the sample is selected with p.p.s. of the supplementary variable x, it may often happen that x is not highly correlated with all the characteristics of interest. For some of the characteristics y, the correlation between y and x may be quite small so that using the usual estimators in unequal probability sampling may give large variance for the estimates of these characteristics. In such circumstances, one would like to use alternative estimators that have smaller variance. In equal probability sampling when the supplementary variable x is utilized through ratio or regression estimates, there is no difficulty in the above circumstances, since we can ignore the information on x and use the familiar estimate $N\bar{y}$ to estimate the population total. One naturally thinks of using $N\bar{y}$ as an estimate of the total in p.p.s. sampling also for just *those* characteristics y for which the correlation between y and x is quite small. Now, under the p.p.s. system

$$E(N\bar{y}) = \frac{N}{n} \sum^{N} y_iP_i = Y + \left(\frac{N}{n} \sum^{N} y_iP_i - Y \right). \qquad (6.22)$$

Also from the ordinary definition of population covariance,

$$\text{Cov.}(y_i,P_i) = \frac{1}{N}\left[\sum y_iP_i - \frac{Y\sum P_i}{N} \right]$$

$$= \frac{n}{N^2}\left(\frac{N}{n} \sum y_iP_i - Y \right) \qquad (6.23)$$

since $\sum^{N} P_i = n$. Since we are usually interested in the sampling procedures for which $P_i = np_i$ where $p_i = x_i/X$,

$$\text{Cov.}(y_i,P_i) = \frac{n}{X}\,\text{Cov.}(y_i,x_i) \ . \qquad (6.24)$$

Therefore, from (6.23) and (6.24),

$$E(N\bar{y}) = Y + \frac{N^2}{X}\,\text{Cov.}(y_i,x_i) \ . \qquad (6.25)$$

Since we expect to have a very small correlation between y and x for just *those* characteristics y for which we may wish to use the estimate $N\bar{y}$, the bias in (6.25) is small and can be neglected. In fact, if there is no correlation, $N\bar{y}$ is an unbiased estimate of Y. To compare the variance of $N\bar{y}$ and the usual estimator $\hat{Y}$, under the p.p.s. system, let us consider our particular sampling procedure. We have, to $O(N^1)$,

$$V(\hat{Y}) = V\left(\sum^{n} \frac{y_i}{P_i} \right) = \sum^{N} P_i \left[1 - \frac{(n - 1)}{n} P_i\right]\left(\frac{y_i}{P_i} - \frac{Y}{n}\right)^2 . \quad (6.26)$$

Now

$$V(N\bar{y}) = \frac{N^2}{n^2} V\left(\sum^{n} y_i \right) = \frac{N^2}{n^2} V\left(\sum^{n} \frac{y_i P_i}{P_i} \right) . \quad (6.27)$$

Therefore, $V(\sum^{n} y_i)$ to $O(N^1)$ is obtained by replacing y_i by $y_i P_i$ in (6.26). Hence, to $O(N^{-3})$,

$$V(N\bar{y}) = \frac{N^2}{n^2} \sum^{N} P_i \left[1 - \frac{(n - 1)}{n} P_i\right]\left(y_i - \frac{\sum y_i P_i}{n}\right)^2 . \quad (6.28)$$

Since the correlation between y and x is expected to be quite small,

$$\frac{N}{n} \sum y_i P_i \doteq Y . \quad (6.29)$$

Therefore, to $O(N^{-3})$,

$$V(N\bar{y}) \doteq \sum^{N} P_i \left[1 - \frac{(n - 1)}{n} P_i\right]\left(\frac{N}{n} y_i - \frac{Y}{n}\right)^2 . \quad (6.30)$$

Now, if the correlation between y and x is small, we expect that the variation between the variates $\frac{N}{n} y_i$ is smaller than that between the variates $\frac{y_i}{P_i} = \frac{X}{n} \cdot \frac{y_i}{x_i}$. Now noting that the equations (6.30) and (6.26) are weighted sums of squares of deviations of the variates $\frac{N}{n} y_i$ and y_i/P_i from Y/n respectively with the same weights, it follows that under the above circumstances we expect $V(N\bar{y})$ to be smaller than $V(\hat{Y})$.

In unequal probability sampling with replacement, the variance of the usual estimator $\hat{Y}' = \sum^{n} y_i/np_i$ is greater than or equal to the variance of the estimator $N\bar{y}$, if it is assumed that y_i and p_i (or x_i) are approximately independent as shown below. This assumption may not be too unrealistic when the correlation between y_i and x_i is very small and sampling is done with replacement. Now

$$V(\hat{Y}') = n^{-1} \sum \frac{y_i^2}{p_i} - \frac{Y^2}{n} \qquad (6.31)$$

and

$$V(N\bar{y}) = \frac{N^2}{n} \sum y_i^2 p_i - \frac{N^2}{n} (\sum y_i p_i)^2 \; . \qquad (6.32)$$

Since y_i and p_i are assumed to be approximately independent,

$$\sum y_i p_i \doteq \frac{Y \sum p_i}{N} = \frac{Y}{N}$$

$$\sum y_i^2 p_i \doteq \frac{\sum y_i^2 \sum p_i}{N}$$

$$\text{and} \quad \sum \frac{y_i^2}{p_i} \doteq \frac{\sum y_i^2 \sum \frac{1}{p_i}}{N} \; . \qquad (6.33)$$

Therefore, $V(N\bar{y})$ is smaller than or equal to $V(\hat{Y}')$ if

$$\frac{N}{n}(\sum y_i^2)(\sum p_i) - \frac{Y^2}{n} \leq \frac{1}{Nn}(\sum y_i^2)(\sum \frac{1}{p_i}) - \frac{Y^2}{n} \qquad (6.34)$$

or

$$N^{-1} \sum \frac{1}{p_i} \geq N \sum p_i = N \; . \qquad (6.35)$$

Now, the harmonic mean of the p_i's is smaller than or equal to the aritnmetic mean of the p_i's, i.e.

$$\frac{N}{\sum \frac{1}{p_i}} \leq \frac{\sum p_i}{N} = N^{-1}$$

$$\text{or } N^{-1} \sum \frac{1}{p_i} \geq N \tag{6.36}$$

which is the same as (6.35). Hence, the variance of $N\bar{y}$ is smaller than or equal to the variance of $\hat{Y}'$.

C. Efficiency of Stratification

Stratification is an important device to increase the precision of the estimators. A useful stratified unequal probability sampling design is described in the next section. Here we consider efficiency of stratification for unequal probability sampling without replacement. Cochran (1953) has considered the efficiency of stratification in equal probability sampling without replacement and has estimated the gain in efficiency due to stratification. Sukhatme (1954) has considered the case of unequal probability sampling with replacement. The problem involved is to compare the estimate of the variance of the given stratified sample with the estimate of the variance of an unstratified random sample of same size expressed in terms of the units in the stratified sample. Efficiency of stratification for unequal probability sampling without replacement has not been considered in the literature, the reason probably is due to the difficulties involved in evaluating the probabilities P_i and $P_{ii'}$ involved in the variance formula, when $n > 2$. The only procedure available which gives simple expressions for P_i and $P_{ii'}$ when $n > 2$ seems to be that of Midzuno, which has some restrictive features due to the fact that only one unit is selected with unequal probabilities and the remaining $(n - 1)$ units are selected with equal probabilities.

Since we have developed an asymptotic theory for a particular unequal probability sampling procedure which provides compact expressions for the variance when $n > 2$, it may be useful to spell out here the formulas for evaluating efficiency of stratification in unequal probability sampling without replacement.

Let there be L strata with N_h units in the h^{th} stratum $(h = 1, \ldots, L)$. A sample of size n_h is drawn from the h^{th} stratum with unequal probabilities and without replacement so that

$$\hat{Y}_s = \sum_{h}^{L} \sum_{t}^{n_h} \frac{y_{ht}}{p_{ht}} = \sum_{h}^{L} \hat{Y}_h \qquad (6.37)$$

is an unbiased estimate of the population total Y where p_{ht} is the probability for selecting the t^{th} unit of the h^{th} stratum. Since the samples are drawn independently from each stratum,

$$V(\hat{Y}_s) = \sum_{h}^{L} V(\hat{Y}_h) = \sum_{h}^{L} \Big[\sum_{t}^{N_h} \frac{y_{ht}^2}{P_{ht}}$$

$$+ \sum_{t \neq t'}^{N_h} \frac{P_{htt'}}{P_{ht}P_{ht'}} y_{ht}y_{ht'} - Y_h^2 \Big] \qquad (6.38)$$

where $P_{htt'}$ is the probability for selecting both units t and t' of the h^th stratum. Equation (6.38) is a general formula for any sampling procedure. Now, for our particular sampling procedure, assuming that N_h is large we have to $O(N_h^1)$,

$$V(\hat{Y}_h) = \sum_{t}^{N_h} P_{ht} \Big[1 - \frac{(n_h - 1)}{n_h} P_{ht}\Big]\Big(\frac{y_{ht}}{P_{ht}} - \frac{Y_h}{n_h}\Big)^2 \qquad (6.39)$$

where $P_{ht} = n_h p_{ht}$ and Y_h is the population total for the h^th stratum. If the size measures x_i are good for the population as a whole, we can often expect that the same size measures to be good for each of the strata so that usually we take $p_{ht} = x_{ht}/X_h$ where X_h is the h^th stratum total for the x_i. Using (6.39) it follows that

$$V(\hat{Y}_s) = \sum_{h}^{L} \sum_{t}^{N_h} \frac{p_{ht}}{n_h}\Big(\frac{y_{ht}}{p_{ht}} - Y_h\Big)^2$$

$$- \sum_{h}^{L} \frac{(n_h - 1)}{n_h} \sum_{t}^{N_h} (y_{ht} - Y_h p_{ht})^2 \qquad (6.40)$$

for our sampling procedure. Also, the Yates and Grundy estimate of the variance of $\hat{Y}_s$ for any sampling procedure is

$$v_{YG}(\hat{Y}_s) = \sum_h^L v_{YG}(\hat{Y}_h) = \sum_h^L \sum_{t>t'}^{n_h} \frac{P_{ht}P_{ht'} - P_{htt'}}{P_{htt'}} \cdot \left(\frac{y_{ht}}{P_{ht}} - \frac{y_{ht'}}{P_{ht'}}\right)^2 . \tag{6.41}$$

For our sampling procedure, using the estimate of the variance of $\hat{Y}_h$ to $O(N_h^{1})$, we have

$$v_{YG}(\hat{Y}_s) = \sum_h^L (n_h - 1)^{-1} \sum_{t>t'}^{n_h} \cdot \left(1 - P_{ht} - P_{ht'} + n_h^{-1} \sum_t^{N_h} P_{ht}^2\right)\left(\frac{y_{ht}}{P_{ht}} - \frac{y_{ht'}}{P_{ht'}}\right)^2 . \tag{6.42}$$

Also, for an unstratified sample of size $n = \sum^L n_h$, the variance formula for the estimate $\hat{Y}$ is

$$V(\hat{Y}) = \sum^N \frac{y_j^2}{P_j} + \sum_{i\neq i'}^N \frac{P_{ii'}}{P_iP_{i'}} y_i y_{i'} - Y^2$$

$$= \sum_{i>i'}^N (P_iP_{i'} - P_{ii'})\left(\frac{y_i}{P_i} - \frac{y_{i'}}{P_{i'}}\right)^2 \tag{6.43}$$

where P_i and $P_{ii'}$ are respectively the probability for selecting the ith unit and the probability for selecting both the units i and i' in an unstratified sample of size n. For our sampling procedure, to $O(N^1)$ we have

$$V(\hat{Y}) = \sum^{N} P_i \left[1 - \frac{(n-1)}{n} P_i\right]\left(\frac{y_i}{P_i} - \frac{Y}{n}\right)^2 \qquad (6.44)$$

where $P_i = np_i$ with $p_i = x_i/X$.

Let the i^{th} unit in the population correspond to the t^{th} unit in the h^{th} stratum so that $p_i = p_{ht}p_{h.}$ where $p_{h.} = \sum^{N_h} p_i = \frac{X_h}{X}$. Then (6.44) can be written as

$$V(\hat{Y}) = \sum_h^L \sum_t^{N_h} \frac{p_{ht}}{np_{h.}}\left(\frac{y_{ht}}{p_{ht}} - Y_h\right)^2 + \frac{1}{n}\sum_h^L p_{h.}\left(\frac{Y_h}{p_{h.}} - Y\right)^2$$

$$- \frac{(n-1)}{n}\sum_h^L \sum_t^{N_h} (y_{ht} - p_{ht}p_{h.}Y)^2 . \qquad (6.45)$$

Therefore from (6.45) and (6.40)

$$V(\hat{Y}) - V(\hat{Y}_s) = \sum_h^L \sum_t^{N_h} \left(\frac{1}{np_{h.}} - \frac{1}{n_h}\right)p_{ht}\left(\frac{y_{ht}}{p_{ht}} - Y_h\right)^2$$

$$+ \frac{1}{n}\sum_h p_{h.}\left(\frac{Y_h}{p_{h.}} - Y\right)^2$$

$$- \frac{(n-1)}{n}\sum_h^L \sum_t^{N_h} (y_{ht} - p_{ht}p_{h.}Y)^2$$

$$+ \sum_h^L \frac{(n_h - 1)}{n_h}\sum_t^{N_h} (y_{ht} - p_{ht}Y)^2 . \qquad (6.46)$$

In the r.h.s of (6.46), the first two terms are of larger order than the last two terms. So, if the allocation of the

n_h is such that $n_h = np_{h.}$, we expect $V(\hat{Y}_s)$ to be smaller than $V(\hat{Y})$.

Let us now consider the estimation of the efficiency of stratification. Let P'_{ht} and $P'_{htt'}$ denote P_i and $P_{ii'}$ respectively where i and i' correspond to t and t'th units in the hth stratum. Similarly let $P'_{hh'tt'}$ denote $P_{ii'}$ where i and i' correspond to units t and t' in the hth and h'th strata respectively. Then, (6.43) can be written as

$$V(\hat{Y}) = \sum_{h}^{L} \sum_{t>t'}^{N_h} (P'_{ht}P'_{ht'} - P'_{htt'})\left(\frac{y_{ht}}{P'_{ht}} - \frac{y_{ht'}}{P'_{ht'}}\right)^2$$

$$+ \sum_{h>h'}^{L} \sum_{t}^{N_h} \sum_{t'}^{N_{h'}} (P'_{ht}P'_{h't'} - P'_{hh'tt'})\left(\frac{y_{ht}}{P'_{ht}} - \frac{y_{h't'}}{P'_{h't'}}\right)^2 . \tag{6.47}$$

Therefore, an unbiased estimate of (6.47) from the given stratified sample is

$$v(\hat{Y}) = \sum_{h}^{L} \sum_{t>t'}^{n_h} \frac{(P'_{ht}P'_{ht'} - P'_{htt'})}{P_{htt'}} \left(\frac{y_{ht}}{P'_{ht}} - \frac{y_{ht'}}{P'_{ht'}}\right)^2$$

$$+ \sum_{h>h'}^{L} \sum_{t}^{n_h} \sum_{t'}^{n_{h'}} \frac{(P'_{ht}P'_{h't'} - P'_{hh'tt'})}{P_{ht}P_{h't'}} \left(\frac{y_{ht}}{P'_{ht}} - \frac{y_{h't'}}{P'_{h't'}}\right)^2 . \tag{6.48}$$

In the special case of equal probabilities, we have $P_i = n/N$, $P_{ii'} = n(n - 1)/N(N - 1)$, $P_{ht} = n_h/N_h$ and $P_{htt'} = n_h(n_h - 1)/N_h(N_h - 1)$ and it can be shown after some manipulation that

(6.48) reduces to the expression given by Cochran (1953) which is

$$v(\hat{Y}) = \frac{(N - n)}{n(N - 1)} \Big[N \sum N_h s_h^2 - N \sum \frac{N_h s_h^2}{n_h} + \sum \frac{N_h^2 s_h^2}{n_h}$$

$$- \sum N_h s_h^2 + N \sum N_h \bar{y}_h^2 - (\sum N_h \bar{y}_h)^2 \Big] \qquad (6.49)$$

where $\bar{y}_h$ and s_h^2 are respectively the sample mean and the sample mean square for the h^{th} stratum. Also it may be noted that the situations in which $v(\hat{Y})$ is positive are similar to those in which the Yates and Grundy estimate of the variance is positive.

For our particular sampling procedure, the general formula (6.48) reduces to

$$v(\hat{Y}) = \sum_{h}^{L} \sum_{t>t'}^{n_h} \frac{1}{nn_h(n_h - 1)} \Big[1 - (p_{ht} + p_{ht'})$$

$$\cdot \{ (n - 1)p_{h.} + 1 \} + (n - 1) \sum_{h} p_{h.}^2 . \sum_{t} p_{ht}^2 + \sum_{t} p_{ht}^2 \Big]$$

$$x \, (\frac{y_{ht}}{p_{ht}} - \frac{y_{ht'}}{p_{ht'}})^2$$

$$+ \sum_{h>h'}^{L} \sum_{t}^{n_h} \sum_{t'}^{n_{h'}} \frac{p_{h.} p_{h'.}}{nn_h n_{h'}} \Big[1 - (n - 1)(p_{ht} p_{h.}$$

$$+ p_{h't'} p_{h'.}) + (n - 1) \sum_{h} p_{h.}^2 . \sum_{t} p_{ht}^2 \Big]$$

$$x \, (\frac{y_{ht}}{p_{ht} p_{h.}} - \frac{y_{h't'}}{p_{h't'} p_{h'.}})^2 \qquad (6.50)$$

after sucstituting for $P_{ii'}$ to $O(N^{-3})$ and $P_{htt'}$ to $O(N_h^{-3})$. Finally, the estimated percentage gain in efficiency due to stratification is given by

$$\frac{v(\hat{Y}) - v_{YG}(\hat{Y}_s)}{v_{YG}(\hat{Y}_s)} \times 100 \ . \qquad (6.51)$$

D. A Stratified Unequal Probability Sampling Design

As mentioned earlier in Chapter II, the sentiments expressed by Weibull (1960) regarding the desirability of sampling the units with higher weights than the units with lower weights, can be incorporated in the following stratified design: First, rank the units according to their weights (say weight of a unit is proportional to its size measure). Then form several strata by grouping the units in that order, such that each stratum has approximately the same total size. Draw two units from each stratum with unequal probabilities, usually with p.p.s. (assuming of course that there are at least two units in each stratum). It is not necessary that unequal probability sampling has to be used in each stratum. In fact, in some of the strata we may prefer to use equal probability sampling since the size measures of the units may not vary much in these strata. By stratifying in this manner, the number of units in a stratum with higher size measures is smaller than the number of units in a stratum with lower size measures since the strata are all approximately of equal total size. Therefore, the intensity of sampling in the strata with higher size measures is greater than the intensity of sampling in the strata with lower size measures since the sample size is the same (namely two) in all the strata. For example, if two units have very large size measures, these two units may constitute a stratum so that these two units will be included in the sample with certainty, *i.e.* the sampling intensity in this stratum is hundred percent. The above stratified design provides a valid estimate of the variance of the estimate of the population total or mean unlike the design where only units with higher weights are sampled.

VII. LITERATURE CITED

Abdel-Aty, S. H. 1954. Tables of generalized k-statistics. Biometrika, 41:253-260.

Cochran, W. G. 1953. Sampling techniques. John Wiley and Sons, Incorporated, New York.

Des Raj. 1954. Ratio estimation in sampling with equal and unequal probabilities. Journal of the Indian Society of Agricultural Statistics, 6:127-138.

________. 1956a. Some estimators in sampling with varying probabilities without replacement. Journal of the American Statistical Association, 51:269-284.

________. 1956b. A note on the determination of optimum probabilities in sampling without replacement. Sankhayā, 17:197-200.

________. 1958. On the relative accuracy of some sampling techniques. Journal of the American Statistical Association, 53:98-101.

Durbin, J. 1953. Some results in sampling theory when the units are selected with unequal probabilities. Journal of the Royal Statistical Society, Series B, 15:262-269.

Godambe, V. P. 1955. A unified theory of sampling from finite populations. Journal of the Royal Statistical Society, Series B, 17:269-277.

Goodman, R. and L. Kish. 1950. Control beyond stratification; a technique in probability sampling. Journal of the American Statistical Association, 45:350-372.

Hansen, M. H. and W. N. Hurwitz. 1943. On the theory of sampling from finite populations. Annals of Mathematical Statistics, 14:333-362.

Hartley, H. O. and A. Ross. 1954. Unbiased ratio estimators. Nature, 174:270.

Horvitz, D. G. and D. J. Thompson. 1952. A generalization of sampling without replacement from a finite universe. Journal of the American Statistical Association, 47:663-685.

Kendall, M. G. and A. Stuart. 1958. The advanced theory of statistics. Hafner Publishing Company, Incorporated, New York.

Koop, J. C. 1957. Contributions to the general theory of sampling finite populations without replacement and with unequal probabilities. Unpublished Ph. D. Thesis. Library, North Carolina State College, Raleigh, North Carolina.

Lahiri, D. B. 1951. A method of sample selection providing unbiased ratio estimates. Bulletin of International Statistical Institute, 33:133-140.

Madow, W. G. 1948. On the limiting distributions of estimates based on samples from finite universes. Annals of Mathematical Statistics, 19:535-545.

________. 1949. On the theory of systematic sampling. II. Annals of Mathematical Statistics, 20:333-354.

Mickey, M. R. 1954. Some finite population unbiased ratio and regression estimators. (Mimeo.) Statistical Laboratory, Iowa State University of Science and Technology, Ames, Iowa.

________. 1959. Some finite population unbiased ratio and regression estimators. Journal of the American Statistical Association, 54:594-612.

Midzuno, H. 1950. An outline of the theory of sampling systems. Annals of the Institute of Statistical Mathematics (Japan), 1:149-156.

Murthy, M. N. 1957. Ordered and unordered estimators in sampling without replacement. Sankhayā, 18:379-390.

Narain, R. D. 1951. On sampling without replacement with varying probabilities. Journal of the Indian Society of Agricultural Statistics, 3:169-174.

Said, E. 1955. A comparison between alternative techniques using supplementary information in sample survey design. Unpublished Ph.D. Thesis. Library, North Carolina State College, Raleigh, North Carolina.

Sen, A. R. 1953. On the estimate of the variance in sampling with varying probabilities. Journal of the Indian Society of Agricultural Statistics, 5:119-127.

_________. 1955. A simple design in sampling with varying probabilities. Journal of the Indian Society of Agricultural Statistics, 7:57-69.

Stevens, W. L. 1958. Sampling without replacement with probability proportional to size. Journal of the Royal Statistical Society, Series B, 20:393-397.

Sukhatme, P. V. 1954. Sampling theory of surveys with application. Iowa State College Press, Ames, Iowa.

Thompson, D. J. 1952. A theory of sampling finite universes with arbitrary probabilities. Unpublished Ph. D. Thesis. Library, Iowa State University of Science and Technology, Ames, Iowa.

Weibull, C. 1960. Some aspects of statistical inference with applications to sample survey theory. Statistical Institute, University of Gothenburg, Sweden.

Williams, W. H. 1958. Unbiased regression estimators and their efficiencies. Unpublished Ph. D. Thesis. Library, Iowa State University of Science and Technology, Ames, Iowa.

Wishart, J. 1952. Moment coefficients of the k-statistics in samples from a finite population. Biometrika, 39: 1-13.

Yates, F. 1949. Sampling methods for censuses and surveys. Charles Griffin and Company, Incorporated, London.

_________ and P. M. Grundy. 1953. Selection without replacement from within strata with probability proportional to size. Journal of the Royal Statistical Society, Series B, 15:235-261.

Zarkovic, S. S. 1960. On the efficiency of sampling with varying probabilities and the selection of units with replacement. Metrika, 3:53-60.

VIII. ACKNOWLEDGMENTS

The author is greatly indebted to Professor H. O. Hartley for his valuable guidance during the preparation of this thesis. Sincere thanks are also expressed to Professor T. A. Bancroft for his kind advice from time to time.

Acknowledgment is also made for the support of this work by a basic research grant from the Office of Ordinance Research, United States Army.

IX. APPENDIX

Madow (1948) has shown that the distribution of a standardized total of v units from a population of size N, tends to a normal distribution with mean zero and standard deviation 1, provided that an $e > 0$ exists such that $\frac{v}{N} < 1 - e$ if v and N are sufficiently large. That is, all k_r ($r \geq 3$) of the standardized total are zero in the limit. Moreover it follows from Madow's results that k_r ($r \geq 3$) is at least $O(N^{-\frac{1}{2}})$ with $v = qN$. This result immediately shows that none of the k_r's ($r \geq 3$) contribute to P_{12} to $O(N^{-3})$. However, Madow's result is not sufficient to show that (4.47) for P_{12} is correct to $O(N^{-4})$ since we need to show that k_r ($r \geq 4$) is at least $O(N^{-c})$ with $c > 1/2$. We now give a heuristic argument to show that k_r is in fact $O(N^{-\frac{r}{2}+1})$ with $v = qN$. It may be noted that for infinite populations it is well known that k_r is $O(v^{-\frac{r}{2}+1})$. Using the results of Wishart (1952) and Abdel-Aty (1954), it is verified below that k_r ($r \leq 8$) is $O(N^{-\frac{r}{2}+1})$ with $v = qN$. The difficulty involved in giving a general proof is that no general relations for the standardized polykays $K_{ijt...}$ in terms of the standardized cumulants K_i are available except that Abdel-Aty provides a table giving the relations up to $r = 12$. In general we have

$$K_{ijt...} = K_iK_jK_t \; ... + \text{terms of } O(N^{-1}) \text{ and smaller.}$$

Now from Wishart, the fourth moment μ_4 of the standardized total is

$$\mu_4 = a^{-2}\left[K_4\left\{a^3 - \frac{a}{N}(a - N^{-1})\right\} + 3a^2 K_{22}\right] \qquad (9.1)$$

where $a = 1/v - 1/N$. Using the relations

$$K_{22} = \frac{N - 1}{N + 1} K_2^2 - \frac{(N - 1)}{N(N - 1)} K_4 \qquad (9.2)$$

and

$$k_4 = \mu_4 - 3\mu_2^2 \qquad (9.3)$$

it is seen that

$$k_4 = K_4\left[a - \frac{(a - N^{-1})}{Na} - \frac{3(N - 1)}{N(N + 1)}\right] - \frac{6}{N + 1} K_2^2 \qquad (9.4)$$

which is $O(N^{-1})$ with $v = qN$. Similarly,

$$k_5 = K_5\left[a^{\frac{3}{2}} - \frac{v}{Na^{\frac{3}{2}}}(a^3 + N^{-3})\right]$$

$$+ \left[\frac{6}{N + 5} K_3 + \frac{(N - 1)}{N(N + 5)} K_5\right]\left[\frac{6}{Na^{\frac{1}{2}}} - \frac{4}{a^{\frac{1}{2}}}(a - N^{-1}) - 6a^{\frac{1}{2}}\right] \qquad (9.5)$$

which is $O(N^{-\frac{3}{2}})$ with $v = qN$. Also

$$\mu_6 = a^{-3}\left[av(a^5 + N^{-5})K_6 + 15va^2(a^3 + N^{-3})K_{42}\right.$$

$$\left. + 10a^2(a - N^{-2})^2 K_{33} + 15a^3 K_{222}\right]. \qquad (9.6)$$

Note that in the r.h.s. of (9.6), the first term is $O(N^{-2})$, the next two terms are $O(N^{-1})$ and the last term is $O(N^0)$ with

$v = qN$. Using the relation

$$k_6 = \mu_6 - 15\mu_4\mu_2 - 10\mu_3^2 + 30\mu_2^3 \tag{9.7}$$

and the relations for $K_{ijt\ldots}$ in terms of K_i from Wishart, it is found that all terms of $O(N^{-1})$ and $O(N^0)$ in (9.6) become cumulant corrections which cancel so that k_6 is $O(N^{-2})$. Now from Abdel-Aty

$$\mu_7 = a^{-\frac{7}{2}}\left[K_7A_7 + 21K_{52}A_5A_2 + 35K_{43}A_4A_3 + 105K_{322}A_3A_2^2\right] \tag{9.8}$$

where

$$A_r = a^{r-1} - \frac{a^{r-2}}{N} + \frac{a^{r-3}}{N^2} \cdots (-1)^{r-2}\frac{a}{N^{r-2}}\ . \tag{9.9}$$

Using the relations

$$k_7 = \mu_7 - 21\mu_5\mu_2 - 35\mu_4\mu_3 + 210\mu_3\mu_2^2 \tag{9.10}$$

and the relations for $K_{ijt\ldots}$ in terms of K_i, it is verified that k_7 is $O(N^{-\frac{5}{2}})$. Similarly, it is verified that k_8 is $O(N^{-3})$. In general we have from Abdel-Aty

$$\mu_r = a^{-\frac{r}{2}}\Big[K_rA_r + \sum_{\substack{(i,j\geq 2)\\ i+j=r}} C_{i-1,j-1}(v,N)\ K_{ij} + \cdots$$

$$+ \sum_{\substack{(i,j,t\cdots\geq 2)\\ i+j+t+\cdots=r}} C_{i-1,j-1,t-1,\ldots}\ (v,N)\ K_{ijt\ldots}\Big] \tag{9.11}$$

where

$$C_{i-1,j-1,t-1,\ldots}(v,N) = \frac{r!}{(i!j!k!\cdots)(s!m!\cdots)}\hat{A}_i\hat{A}_j\hat{A}_t\cdots \tag{9.12}$$

where s of the A's are equal, m of the A's are equal and so on. Note that $a^{-\frac{r}{2}} C_{i-1,j-1,t-1,\ldots}(v,N)$ is $O(N^{-[(i-1)+(j-1)+(t-1)+\cdots]+\frac{r}{2}})$. Since we have verified that k_r is $O(N^{-\frac{r}{2}+1})$ up to r = 8, we conjecture that using (9.11) and the relation for k_r in terms of $\mu_i (i \leq r)$ which involves Bernoulli numbers, and also the relations for $K_{ijt\ldots}$ in terms of K_i, all terms of order larger than $O(N^{-\frac{r}{2}+1})$ become cumulant corrections which cancel so that k_r is $O(N^{-\frac{r}{2}+1})$. We should recall here that an independent check on the order of magnitude of P_{12} was given earlier in Chapters IV and V.