1982

A study of the effects of measurement error in survey sampling

Promod Kumar Chandhok
Iowa State University

Follow this and additional works at: https://lib.dr.iastate.edu/rtd

Part of the Statistics and Probability Commons

Recommended Citation
Chandhok, Promod Kumar, "A study of the effects of measurement error in survey sampling " (1982). Retrospective Theses and Dissertations. 7030.
https://lib.dr.iastate.edu/rtd/7030

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

This reproduction was made from a copy of a document sent to us for microfilming. While the most advanced technology has been used to photograph and reproduce this document, the quality of the reproduction is heavily dependent upon the quality of the material submitted.

The following explanation of techniques is provided to help clarify markings or notations which may appear on this reproduction.

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

2. When an image on the film is obliterated with a round black mark, it is an indication of either blurred copy because of movement during exposure, duplicate copy, or copyrighted materials that should not have been filmed. For blurred pages, a good image of the page can be found in the adjacent frame. If copyrighted materials were deleted, a target note will appear listing the pages in the adjacent frame.

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

4. For illustrations that cannot be satisfactorily reproduced by xerographic means, photographic prints can be purchased at additional cost and inserted into your xerographic copy. These prints are available upon request from the Dissertations Customer Services Department.

5. Some pages in any document may have indistinct print. In all cases the best available copy has been filmed.

University Microfilms International
300 N. Zeeb Road
Ann Arbor, MI 48106
Chandhok, Promod Kumar

A STUDY OF THE EFFECTS OF MEASUREMENT ERROR IN SURVEY SAMPLING

Iowa State University

University Microfilms International 300 N. Zeeb Road, Ann Arbor, MI 48106
A study of the effects of measurement error in survey sampling

by

Promod Kumar Chandhok

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

Major: Statistics

Approved:

[Signatures redacted for privacy]

In Charge of Major Work

For the Major Department

For the Graduate College

Iowa State University
Ames, Iowa

1982
TABLE OF CONTENTS

1. INTRODUCTION
   1.1. The Modern Theory of Survey Sampling 1
   1.2. Response Errors in Surveys 2
   1.3. Mathematical-Statistical Models for Response Errors 6
   1.4. Scope of the Thesis 9

2. REVIEW OF LITERATURE 11

3. EFFECTS OF MEASUREMENT ERROR IN EQUAL AND UNEQUAL PROBABILITY SAMPLING
   3.1. Introduction 23
   3.2. Unequal Probability Sampling with Replacement Under the Simple Correlation Model 26
   3.3. Unequal Probability Sampling Without Replacement Under the Simple Correlation Model 32
   3.4. Simple Random Sampling Under the Simple Correlation Model 38

4. RATIO ESTIMATION IN THE PRESENCE OF MEASUREMENT ERRORS 47
   4.1. Introduction 47
   4.2. Ratio Estimation Under Equal Probability Sampling 50
   4.3. Ratio Estimation Under Midzuno's Scheme of Sampling 62
   4.4. Comparison of the Two Strategies 72

5. EXTENSIONS TO TWO-STAGE AND STRATIFIED SAMPLING 76
   5.1. Introduction 76
   5.2. Unequal Probability Sampling with Replacement 78
   5.3. Unequal Probability Sampling without Replacement 81
   5.4. Stratified Sampling 92
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.5. Systematic Sampling</td>
<td>96</td>
</tr>
<tr>
<td>6. EMPIRICAL STUDY</td>
<td>98</td>
</tr>
<tr>
<td>6.1. Comparison of ppswr Sampling With Simple Random Sampling</td>
<td>98</td>
</tr>
<tr>
<td>6.2. Ratio Estimation</td>
<td>106</td>
</tr>
<tr>
<td>7. SUMMARY</td>
<td>113</td>
</tr>
<tr>
<td>8. BIBLIOGRAPHY</td>
<td>115</td>
</tr>
<tr>
<td>9. ACKNOWLEDGEMENTS</td>
<td>119</td>
</tr>
<tr>
<td>10. APPENDIX</td>
<td>120</td>
</tr>
</tbody>
</table>
1. INTRODUCTION

1.1. The Modern Theory of Survey Sampling

In survey sampling, inference on population characteristics is based on the information contained in the sample. The method of survey sampling is recognized as an organized system of fact-finding, and is being increasingly used in many fields of study. The most extensive users of this method are the governments of both developed and developing countries. As an example, the annual number of births, deaths and marriages, and the number migrating from rural to urban areas and vice versa, are some items of interest to a government interested in keeping track of the population. In the field of fisheries, information is needed on the quantity of fish harvested, its distribution by species, and the fishing effort involved. In market research, information is collected on the size of radio and television audiences, readership of newspapers and magazines, the reaction of consumers to new products being manufactured, and the reasons for preference of one product over another. In the social sciences, a large amount of data is collected on variables such as people's behavior, feelings and motivations.

The modern theory of survey sampling is highly advanced. Several methods of sample selection are now available — methods such as simple random sampling, systematic sampling and sampling with unequal probabilities with and without replacement. Considerable use is made of the auxiliary information available on the population to be studied. This is done through stratification, ratio estimation and regression estima-
tion. Cluster sampling, multistage sampling and double sampling enable us to make effective use of all relevant information available. An important feature of the sampling method is that a probability sample provides both an estimate of the parameter and a measure of its precision calculable from the sample itself. But all these mathematical advances have been made subject to the provision that there are no measurement errors at the stage of data collection. This means that when an observation is made on a unit for the character \( y \) (say, age of the person), the true value of \( y \) (exact age in this case) is always obtained. The assumption of no measurement error is usually violated since actual survey experience shows that all stages of a sample survey are potential sources of error. The following section will discuss some of the important sources of error in sample surveys.

1.2. Response Errors in Surveys

Some of the important sources of error in surveys are:

(i) In the first place, the questionnaire used in the survey may be defective. The question asked may be incorrect. For example, the question, "What was your income last year?" may refer to family income or the respondent's income. The question, in this case, is not precisely worded.

(ii) The question asked may be correct, but it is asked to the wrong person (respondent). For example, in a consumer expenditure survey it is required to determine the amount of money spent by a member of the household (say, the son...
studying at a college) on foods consumed by him away from home during the previous month. If the son is not available for an interview and information is collected from his mother (who is available at home when the interviewer has called), the response obtained is likely to be in error since the mother may not know the answer and will make an appropriate estimate of the amount involved.

(iii) Although the question is asked to the proper respondent, the respondent may not have the desired information. Many families in African countries do not record a birth in the family, and no birth certificate is taken as the birth is not registered. In this case, the age recorded in the survey for a person selected in the sample could only be an approximation, as the respondent does not possess the information needed. Quite often, a calendar of events is used in this case to place the person in some age group. For a detailed discussion, see Brass, et al. (1968, p. 88).

(iv) The respondent may be able to produce the information needed with some effort on his part, but there is no motivation on the part of the respondent to make this effort. For example, a person may be requested to maintain an itemized diary of all expenses incurred by him during a period of one month. If he has the will to do it, he can produce an accurate diary of all items purchased on a daily basis. But if he is not interested in the matter, he may put down a few items in
a careless manner and this will lead to serious response errors.

(v) Even if the respondent has the will to cooperate in the survey, the response obtained may not be accurate. The fact that the respondent is being asked to maintain a record of his expenditures may bring about a change in his buying habits for the period of the survey. He might stop buying unnecessary articles for a while, or buy much more than he normally does. This happens unconsciously on his part, and he is not aware of it. This is an example of "conditioning" in surveys (Raj, 1968).

(vi) The respondent may deliberately misrepresent facts. For fear of taxation, a person may not give his total income correctly. The person is not sure that the information collected will be kept strictly confidential. As a matter of prestige, a respondent may overstate his household expenditure to show that he belongs to a high stratum of society. This brings about response errors.

(vii) Because of the passage of time, and consequent loss of memory, a person may not be able to place an event properly in time. When asked about the number of children ever born to a woman, she may fail to report the birth of a baby that took place many years back and the baby died soon after birth (Brass, et al., 1968, p. 64). If information is collected on the expenditure incurred during the last week, the respondent
may unduly include in this week some expenditure that actually took place a day or two prior to the beginning of the week. This phenomenon is called "telescoping of events" (Raj, 1968).

(viii) If information is collected through an interviewer, the interviewer may bring in his or her own ideas in asking questions and recording the response. In an employment survey, an interviewer may not ask women whether they work for a living and may record them as not working. The reason is that the interviewer feels that the right place for a woman is in the home where she should do housework. The responses obtained through such an interviewer will be greatly in error.

(ix) The interviewer may ask the question correctly, but may make an error in recording the response. If the reported age is 25, the recorded age may be 35 because of an error in recording. During processing of the data, this item may be punched as 85 as an error on the part of the punch operator in deciphering this figure.

The above discussion shows that many different kinds of errors of response creep into the data when information is collected in a sample survey (or complete census). The modern theory of survey sampling is adequate if we assume that accurate measurements are obtained from the units selected in the sample. Rarely, if ever, do we obtain data which is free from error. Hence, there is a need for more research by which response errors can be reduced, measured, and properly taken care of in
1.3. Mathematical-Statistical Models for Response Errors

The examples in Section 1.2 suggest that a variety of mathematical-statistical models may be needed to describe adequately the types of response errors relevant to the situation at hand. In order to cope with different situations, a number of terms may have to be introduced in the model. But this will make the analysis more intricate and cumbersome. It is, therefore, advisable to use the simplest type of mathematical-statistical model that will reasonably describe the facts observed. Let $y_{jt}$ represent the observed value of the $y$-characteristic of $U_j$ (the $j^{th}$ unit in the population) at the $t^{th}$ trial, and $Y_j$ the true value of characteristic $y$ for $U_j$. It is assumed that the observation made on a unit is a random variable following a certain distribution. This distribution is determined by the essential conditions of the survey under which the measurements are made. We shall keep these conditions fixed. It is now convenient to write

$$y_{jt} = Y_j + e_{jt} \quad (1.3.1.)$$

where $e_{jt}$ is the deviation of the observation on the character $y$ (made at the $t^{th}$ trial) from the true value of $y$ for the unit $U_j$. The simplest model is one in which

$$E(e_{jt} | j) = 0 \quad (1.3.2.)$$

$$\text{Cov}(e_{jt}, e_{jt'}) = 0$$
\[
V(e_{jt} | j) = \sigma^2 \\
\text{Cov}(e_{jt}, e_{j't} | j, j') = 0
\]

In this case, the average of the deviations (of the responses from the true value of \( y \) on the unit) is zero and the variability of the responses is \( \sigma^2 \), which is the same for all units. When the same unit is observed on two different times \( t \) and \( t' \), the responses or response deviations are uncorrelated. Also, the responses obtained on two different units are uncorrelated. This model is likely to hold when a sample of persons living in different areas of a country are asked to answer a set of simple questions (unaided by the interviewer) in which there is no intention on the part of respondents to understate or overstate the facts involved. In case it is considered more difficult to elicit information from certain segments of the population, for example, the rural people as compared with the urban people, \( \sigma_j^2 \) for a segment may be different from that in other segments. In this situation, a more appropriate model is

\[
\begin{align*}
E(e_{jt} | j) &= 0 \quad (1.3.3.) \\
\text{Cov}(e_{jt}, e_{j't}) &= 0 \\
V(e_{jt} | j) &= \sigma_j^2 \\
\text{Cov}(e_{jt}, e_{j't} | j, j') &= 0
\end{align*}
\]

There are certain situations in which models (1.3.2.) or (1.3.3.) will not apply as such. For example, experience shows that when women aged 25 are asked to give their age, the average of the responses is lower
than 25, say 22. There is a tendency on the part of the young women to
understate their age with the result that \( E(e_{jt} | j) = \mu_j < 0 \). Similarly,
when a group of old people, say aged 70, are asked about their age, the
average of the responses is found to be higher than 70 and consequently
\( E(e_{jt} | j) = \mu_j > 0 \). In such situations, the appropriate change to make
in the model is to write

\[
y_{jt} = Y_j + \mu_j + e_{jt} \tag{1.3.4.}
\]

\[
E(e_{jt} | j) = 0
\]

Sometimes, it is not possible to isolate \( \mu_j \) from \( Y_j \). This is true when
no satisfactory method of determining the true value exists. In that
case, we may write \( Y_j + \mu_j = Y'_{jt} \) and the model becomes

\[
y_{jt} = Y'_{jt} + e_{jt} \tag{1.3.5.}
\]

\[
E(e_{jt} | j) = 0.
\]

What covariances to assume for the response deviations \( e_{jt} \) will depend
on the situation under consideration. It is sometimes found that the
responses obtained on the same unit, by two different enumerators, are
correlated. For example, a young girl may understate her age when asked
by an enumerator, and repeat the same figure to the other enumerator.
Or, the two enumerators may record a woman as "not working for a living"
simple because both feel that women are not supposed to go out for work.
In such cases, \( e_{jt} \) and \( e_{jt'} \) are correlated. Experienced samplers have
found that there are correlations within interviewer assignments, in
which case \( e_{jt} \) and \( e_{jt'} \) are correlated. As stated before, the introduc-
tation of several correlation terms into the model complicates the sub-
sequent analysis.

1.4. Scope of the Thesis

We shall consider a mathematical-statistical model that includes the
bias term and assumes that errors are correlated. The model is of the
form:

\[ y_{jt} = Y_j + \epsilon_{jt} \quad (1.4.1) \]
\[ E(\epsilon_{jt}|j) = 0 \]
\[ V(\epsilon_{jt}|j) = \sigma_j^2 \]
\[ \text{Cov}(\epsilon_{jt}, \epsilon_{j't}|j, j') = \rho \sigma_j \sigma_{j'} \]

This model is referred to as the simple correlation model. We will
discuss particular cases of this model when \( \mu_j = 0, \sigma_j^2 = \sigma^2 \), or
\[ \text{Cov}(\epsilon_{jt}, \epsilon_{j't}|j, j') = 0. \]
In this connection, it is worth pointing out that the model assumed in most textbooks on sampling (except for the
chapter on non-sampling errors) is

\[ y_{jt} = Y_j + \epsilon_{jt} \quad (1.4.2) \]
\[ E(\epsilon_{jt}|j) = 0 \]
\[ V(\epsilon_{jt}|j) = 0 \]
\[ \text{Cov}(\epsilon_{jt}, \epsilon_{j't}|j, j') = 0 \]

Another model that we will consider is given by

\[ y_{jt} = Y_j + \epsilon_{jt} \quad (1.4.3) \]
\[ \begin{align*} 
E(e_{jt} | j) &= 0 \\
V(e_{jt} | j) &= \sigma^2_j \\
E(e_{jt} e_{j't}) &= \rho_{jt} E(e_{jt}^2) 
\end{align*} \]

for \( j \neq j' \), and both \( j, j' \) in the sample. This model will be called the intrasample correlation model.

We will discuss in this thesis the effect of measurement errors on the bias, variance and estimators of the variance of different estimators. In Chapters 3 and 4, we consider the usual estimators in equal and unequal probability sampling, and the ratio estimator under simple random sampling without replacement and the Midzuno's scheme. In Chapter 5, we consider two-stage sampling and stratified sampling. Finally, in Chapter 6 we compare the performance of different estimators in an empirical study.
2. REVIEW OF LITERATURE

It is well-known that there may be errors in the data collected when a sample survey or a census is taken. For example, Rice (1929) gave details of a social study of 2,000 men. In this study, twelve investigators interviewed homeless applicants to determine the cause of their sad plight. The results obtained by two investigators, A and B, are given below.

Table 2.1. Percentage distribution of applicants by cause of destitution

<table>
<thead>
<tr>
<th>Investigator</th>
<th>Major cause given as</th>
<th>Minor cause given as</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Liquor</td>
<td>Industrial</td>
</tr>
<tr>
<td>A</td>
<td>62</td>
<td>7</td>
</tr>
<tr>
<td>B</td>
<td>22</td>
<td>39</td>
</tr>
</tbody>
</table>

It appears from this table that while investigator A found liquor as the cause of destitution in a high proportion of cases, investigator B saw industrial causes as the explanation of their sad plight. After the tabulation, inquiry showed that A was an ardent believer in prohibition while B was regarded by his associates as a socialist. This shows that the bias in the mind of the interviewer was communicated by some process of suggestion to the mind of the interviewed and was then reproduced while questioning by the former.

Similar situations in other fields have been reported in the
literature. Wood (1939) presented data showing discrepancies between duplicate reports of the occupations of 4,500 workers. One report came from the worker himself or from some member of the household and the other from the worker's employer. About a quarter of the reports were found to be in disagreement when the occupations were classified into nine major groups.

An extensive study carried out by Palmer (1943) showed large variations in response on items such as age, education and employment status.

Gray (1955) carried out a study in England in which a sample of 433 employees in a government office was administered a questionnaire. Information was obtained about the annual leave and sick leave taken by them during the last five months. The number of days of leave taken and the month in which it was taken, were to be stated from memory. Comparison with office records of leave on a case-by-case basis indicated large differences in the two sets of reports. Of the 205 who had taken no sick leave, 192 gave the correct answer, while of the 228 who had taken some, only 74 gave completely correct answers.

Belloc (1954) compared data on hospitalization as reported in household interviews with the hospital records of the individuals. It was found that the days of hospitalization per person per annum, the average length of stay, and the percent of cases with surgery as obtained from hospital records were all slightly higher than those obtained from the household survey.

Raj (1972) reported an agricultural study carried out in Greece.
A cadastral survey of five communes was undertaken. In this survey, all land was measured and the name of the holder was recorded. This information was matched with the data collected in the census by interview. The important findings were:

1. In the five communes investigated, 9 percent of the resident farmers did not report their land at all, 22 percent of the parcels were not declared, and 7 percent of the agricultural area was not reported.

2. As for all land in the five communes, 10 percent of the agricultural area and 23 percent of the parcels were not reported.

The results of such studies of errors of measurement carried out by workers in the field prompted a number of writers to warn others of the pitfalls in data collection. Kendall (1942) pointed out that respondent bias and questionnaire construction are the outstanding problems toward which statistical research must be directed. Deming (1944) listed several sources of error in surveys and called for more research in the area of non-sampling errors.

One of the earliest attempts in designing a sample survey in which there was some assurance that the response errors were under control (and therefore the errors did not vitiate the results obtained from the survey), came from Mahalanobis (1946). In his work at the Indian Statistical Institute, Mahalanobis developed the method of interpene­trating subsamples and used it regularly in the surveys conducted by the Institute. In this method, the total sample of n units is divided
at random into \( m \) groups, and the groups are allocated at random to the \( m \) investigators. Hence, there is a group of \( \bar{n} = n/m \) units to be enumerated by each investigator. When the data on a characteristic \( y \) have been collected, the sample means \( \bar{y}_1, \bar{y}_2, \ldots, \bar{y}_m \) are calculated. If these means agree, apart from sampling errors, the survey is considered to be under statistical control. In such a case, an estimator of the population mean \( \bar{y}_N \) is given by

\[
\bar{y} = \frac{1}{m} \sum_{k=1}^{m} \bar{y}_k
\]

the variance of the estimator being

\[
V(\bar{y}) = \frac{1}{m} \cdot V(\bar{y}_k)
\]

and an unbiased estimator of the variance being

\[
\hat{V}(\bar{y}) = \frac{1}{m(m-1)} \sum_{i}^{m} (\bar{y}_i - \bar{y})^2
\]

It is obvious that if an investigator has not understood the instructions properly, this will show up in his work in that the sample mean \( \bar{y}_k \) produced by him will differ from the others. But if all investigators have made a systematic error in the same direction, it will not be detected by this method. Sukhatme and Seth (1952) criticized this method on the grounds that only very large non-sampling errors can be detected through this method, and that too much travel is involved on the part of enumerators, thus making this method expensive. Sukhatme and Seth (1952) also suggested methods of measuring the components of non-sampling errors based on a statistical model believed
by them to be general enough to cover the conditions commonly encountered in agricultural and socio-economic surveys. In their initial model, they assume that \( m \) enumerators participate in the survey and the \( j^{th} \) enumerator makes \( n_{ij} \) observations on the \( i^{th} \) unit. Denoting the observation by \( y_{ijk} \), they formulated

\[
y_{ijk} = x_i + \alpha_j + \delta_{ij} + \varepsilon_{ijk}
\]

where

- \( x_i \) = true value of the \( i^{th} \) unit
- \( \alpha_j \) = bias of the \( j^{th} \) enumerator
- \( \delta_{ij} \) = interaction of the \( j^{th} \) enumerator with the \( i^{th} \) unit
- \( \varepsilon_{ijk} \) = random deviation

Later, Sukhatme and Seth (1952) developed the simpler model

\[
y_{ij} = x_i + \alpha_j + e_{ij}
\]

where

\[
E(e_{ij}) = 0, \quad V(e_{ij}) = \sigma^2_e \quad \text{and} \quad Cov(e_{ij}, e_{ij'}) = 0 \quad \text{for} \quad j \neq j'.
\]

Sukhatme further assumed that \( m \) enumerators have been selected at random from a hypothetical population of \( M \) enumerators and that each enumerator has been assigned to a group of \( n = n/m \) units selected at random from the \( n \) units. It was shown that the sample mean does not provide an unbiased estimate of \( \bar{x}_N \) (the true population mean), unless the individual biases of the enumerators average to zero over the population of \( M \) enumerators. Also, the variance becomes inflated by the
variability of the biases ($\alpha_j$'s) of the enumerators. Hence, there is
a need to design the survey such that $\bar{\alpha}$ is about zero and the variability
of the biases is as small as possible.

Probably the most significant advancement in response error research
was made at the U.S. Bureau of the Census by Hansen, Hurwitz, Marks
and Mauldin (1951). They proposed a very general model for the study
of response errors. Three criteria were proposed by them for the
definition of the true value of an individual:

1. The true value must be uniquely defined.

2. The true value must be defined in such a manner that the
purposes of the survey are met.

3. Where it is possible to do so consistently with the first two
criteria, the true value should be defined in terms of
operations which can be actually carried through (even
though it might be difficult or expensive to perform the
operations).

The general model proposed by them assumes:

a. a population of N individuals and a population of M inter­
viewers, both of which, for convenience, are assumed to be
large;

b. a true value associated with each individual;

c. a set of essential survey conditions which determine for a
particular individual and interviewer the expected value of
the random variable;

d. zero correlation between the random components of responses
for two different individuals with two different interviewers;

e. the order of interviewing respondents by an interviewer either
randomly determined or not affecting the responses.

Hansen, Hurwitz and Bershad (1960) and Hansen, Hurwitz and Pritzker
(1964) made a detailed examination of the mean square error of a survey
estimate when response errors are present. A summary of the results
obtained by them follows.

Suppose the population contains N units and it is desired to
estimate the proportion

\[ \bar{U} = \frac{1}{N} \sum_{j} U_j \]

where \( U_j \) takes the value 1 if the \( j^{th} \) unit is a member of a particular
class, and the value zero if it is not. An observation on the \( j^{th} \) unit
in the sample (selected by simple random sampling with replacement) is
denoted by \( x_{jt} \) and takes the value 1 if, at the \( t^{th} \) trial, the \( j^{th} \) unit
is assigned to the particular class under consideration, and takes the
value 0 otherwise. An estimate of \( \bar{U} \) from the survey is

\[ \hat{P}_t = \frac{1}{n} \sum_{j} x_{jt} \]

Let \( E(x_{jt} | j) = P_j \). Then \( e_{jt} = x_{jt} - P_j \) is called a response deviation.

Let \( E(P_j) = \frac{1}{N} \sum_{j} P_j = \hat{P} \). Then \( \Delta_j = P_j - \hat{P} \) is called a sampling deviation.

Writing \( \hat{P} = \frac{1}{n} \sum_{j} P_j \), the variance of \( \hat{P}_t \) is given by

\[
V(\hat{P}_t) = E[(\hat{P}_t - \hat{P})^2] + E((\hat{P} - P)^2)
= E(\hat{P}_t - \hat{P})^2 + E(\hat{P} - P)^2 + 2E(\hat{P}_t - \hat{P})(\hat{P} - P)
\]
\[ \sigma^2 = \frac{1}{n} \left[ 1 + (n-1)\rho \right] + E(\hat{P} - P)^2 + 2E[(p_t - \hat{P})(\hat{P} - P)] \]

where \( \sigma^2 \) is the variance of the individual response deviations averaged over all possible trials, and \( \rho \) is the intraclass correlation among the response deviations in a survey or trial.

The first term on the right hand side of the last expression is called the response variance, the second term, the sampling variance and the third, the covariance of response and sampling deviations. The quantity \( \frac{\sigma^2}{n} \) is known as the simple response variance while the term \( \frac{n-1}{n} \rho \sigma^2 \) is the correlated component of the total response variance.

By adding the term \( (P - \bar{P})^2 \) to the variance of \( p_t \), we obtain the mean square error of \( p_t \).

It has been found that if there are important contributions to response variance, they are likely to arise from the factors involving correlated response deviations. For example, an interviewer's misunderstanding of instructions or a tendency to use his own personal whims may cause his results to differ from those of other interviewers. In such a case, his results would be a source of correlated response deviations. An important problem is to isolate the simple response variance from the total response variance. In case the response deviations are not correlated, the simple response variance of \( p_t \) is \( \frac{1}{n} \sum_{j} \frac{N P_j (1 - P_j)}{N} \). In this case, the simple response variance has an upper limit, namely \( P(1 - P)/n \). Since a large value of the simple response variance should indicate greater inconsistency of classifi-
cation (unreliability of the measurement process), the quantity
\[
I_d = \frac{\sigma^2}{P(1 - P)}
\]
is defined as an index of inconsistency. It is also noted that the
sampling variance in this case is
\[
\frac{1}{n} \sum_{j} \frac{(P_j - P)^2}{N}
\]
Also, the sum of the sampling variance and response variance is the
same as the upper limit of the response variance.

In the literature, considerable attention has been devoted to the
estimation of the simple response variance (for calculating the index
of inconsistency) and the correlated response variance (for assessing
the effects of interviewers etc.). Basically, the two components of
response variance are estimated by the method of replication and the
method of interpenetration. The method of replication consists of re­
peated observation of some units in a sequence of trials.

Fellegi (1964, 1974) has made good use of the methods of replication
and interpenetration to obtain estimates of response variance components.
A scheme utilizing both the methods follows. Suppose a simple random
sample of nk units is selected from a population and the sample is
partitioned into k subsamples of n units each. Let the subsamples be
denoted by $S_1, S_2, \ldots S_k$. We form a Latin square of the letters
$S_1, S_2, \ldots S_k$, randomize it, and take the first two rows. One inter­
viewer is assigned at random to each column. The first assignment of
each interviewer constitutes the original survey and the second assign­
ment the repeat survey. This design was used by Fellegi (1964) in
connection with the 1961 Population Census of Canada. He found that the uncorrelated response variance dominated the total response variance in case of characteristics such as age, sex and marital status. For sensitive items such as ethnic origin and mother tongue, the correlated response variance dominated the total response variance. Similar results were obtained by Pritzker and Hansen (1962) in connection with the 1960 Population Census of the United States.

Kish (1962) describes two studies on blue collar industrial workers in which the sample respondents were randomized among the interviewers and were asked many questions involving factual and attitudinal items about their jobs and related matters. He found an intraclass correlation of 0 to 0.07 in the first study and of 0 to 0.05 in the second study.

In the area of health surveys, Koons (1973) made estimates of non-sampling errors based on the reinterview program. Special studies were designed for the estimation of the interviewer contribution to non-sampling variance.

On the methodological side, Raj (1970) used the double sampling technique by which the response bias is reduced and the response variance can be estimated from the sample. In the method of double sampling, a simple random sample of n* units is selected and information collected on the variate y. Let $x_{jt}$ be the response obtained from the $j^{th}$ unit, the true value being $y_j$. A subsample of n units is selected at random without replacement and true values of y are obtained by examining the records (say, of establishments regarding the number of persons on the payroll). The population mean $\mu$ is estimated by
\[ \hat{u} = \frac{1}{n} \sum_{j} y_j - \frac{1}{n} \sum_{j} x_{jt} + \frac{1}{n'} \sum_{j} x_{jt} \]

Under the model

\[ x_{jt} = y_j + e_{jt}, \quad E(e_{jt} | j) = a_j \]

\[ \text{Var}(e_{jt} | j) = \sigma^2_{ej}, \quad \text{Cov}(e_{jt}, e_{j't} | j, j') = 0 \]

it was found that

\[ E(\hat{u}) = u \]

\[ \text{MSE}(\hat{u}) = \left( \frac{1}{n} - \frac{1}{N} \right) \sigma^2_y + \left( \frac{1}{n} - \frac{1}{n'} \right) \left[ \sigma^2_a + \frac{1}{N} \sum_j \sigma^2_{ej} \right] \]

where

\[ \sigma^2_a = \frac{1}{N-1} \sum_j (a_j - \bar{a})^2 \]

An unbiased estimator of \( \text{MSE}(\hat{u}) \) is given by

\[ \left( \frac{1}{n} - \frac{1}{N} \right) \frac{1}{n-1} \sum_j (y_j - \bar{y})^2 + \left( \frac{1}{n} - \frac{1}{n'} \right) \frac{1}{n-1} \sum_j \left[ y_j - \bar{y} - (x_{jt} - \bar{x})^2 \right]^2 \]

Chakrabarty (1977) determined the response bias of the ratio estimator in census, when both \( y \) and \( x \) are measured with error.

Let

\[ y_{it} = Y_i + e_{it} \]

\[ x_{it} = X_i + u_{it} \]

where \( y_{it}, x_{it} \) are the values of the characteristics \( y \) and \( x \) for the \( i^{th} \) unit in the \( t^{th} \) repetition; \( Y_i, X_i \) denote the true values, and \( e_{it}, u_{it} \) the errors in reporting in the \( t^{th} \) repetition. Assume
Define

\[ E(e_{it} | Y_t) = \mu_{1i}, \quad E(u_{it} | X_t) = \mu_{2i} \]

\[ \overline{y_R} = \frac{\overline{y_N}}{\overline{x_N}} \cdot \overline{x_N} \]

where

\[ \overline{y_N} = N^{-1} \sum_{i}^{N} y_{it}, \quad \overline{x_N} = N^{-1} \sum_{i}^{N} x_{it}, \quad \overline{x_N} = N^{-1} \sum_{i}^{N} x_{i}. \]

Also, define

\[ \overline{e_t} = N^{-1} \sum_{i}^{N} e_{it}, \quad \overline{u_t} = N^{-1} \sum_{i}^{N} u_{it}, \quad \overline{\mu_1} = N^{-1} \sum_{i}^{N} \mu_{1i}, \]

\[ \overline{\mu_2} = N^{-1} \sum_{i}^{N} \mu_{2j}, \quad c_1^2 = \frac{\sigma(e_t)}{\mu_1^2}, \quad \text{and} \quad c_2^2 = \frac{\sigma(u_t)}{\mu_2^2}. \]

Let \( \rho^* \) be the correlation coefficient between \( \overline{e_t} \) and \( \overline{u_t} \). The relative bias of \( \overline{y_R} \) as an estimator of \( \overline{y_N} \) is given by

\[ B = \frac{\mu_1}{\overline{y}} - \frac{\mu_2}{\overline{x}} + \frac{\mu_2^2}{\overline{x}^2} (1 + c_2^2) - \frac{\mu_1 \mu_2}{\overline{y} \overline{x}} (1 + \rho^* c_1 c_2) \]

For an excellent bibliography on non-sampling errors in sample surveys, the reader is referred to Dalenius (1977a, 1977b, 1977c).
3. EFFECTS OF MEASUREMENT ERROR IN EQUAL AND UNEQUAL PROBABILITY SAMPLING

3.1. Introduction

Consider a finite population \( U \) containing \( N \) distinct units

\[ U_1, U_2, \ldots, U_N \quad (3.1.1) \]

For example, \( U \) may be a population of households in a geographic region, or a population of fields in a commune, or a population of industrial establishments in an area. Each unit of the population possesses a real-valued characteristic \( y \) such as the number of persons in the household, or the area under a crop in the field, or the sales of an establishment. Let

\[ \mathbf{Y} = (Y_1, Y_2, \ldots, Y_N) \quad (3.1.2) \]

be the vector of the \( y \)-values of the units in the population. The basic problem is to estimate some parameter (a function of the vector \( \mathbf{Y} \)) by taking a probability sample from the population \( U \). Although any function could be considered for estimation, interest in sample surveys has generally centered on the following parameters

\[ Y = \sum_{i=1}^{N} Y_i \quad \text{population total} \]

\[ \overline{Y}_N = \frac{1}{N} \sum_{i=1}^{N} Y_i \quad \text{population mean} \]

\[ R = \frac{\sum_{i=1}^{N} Y_i}{\sum_{i=1}^{N} X_i} = \frac{Y}{X} \quad \text{population ratio} \]
where \( x \) is another real-valued characteristic defined over the population \( U \).

Quite often it is found that auxiliary information on some characteristic which provides the \( X \)-values is available for the different units in the population. For example, we may know the number of persons in each block from a previous census of the population, or the area of each field in the commune from a previous census. In such a case, it becomes important for the investigator to make use of this auxiliary information to improve the precision of the estimate of the parameter. A number of procedures are available for using the auxiliary information on the characteristic \( x \). For example, the units in the population may be allocated to strata on the basis of \( x \), or the sample is selected with probability proportional to size (pps sampling), or \( x \) is used in ratio estimation or regression estimation. In this chapter, we will consider pps sampling as well as equal probability sampling.

A complication that arises in actual survey sampling work is the presence of measurement errors. Usually, we do not obtain the true \( y \)-values of the units selected in the sample, but observe the true values together with measurement errors. Measurement errors may occur due to interviewer's biases, defects in measuring instruments, coding and editing, and other sources. Various models for errors of measurement have been considered in the literature. Some of those have been stated in the introduction of the thesis. We shall consider the model given by equations (1.4.1.), i.e.,
\[ y_{jt} = Y_j' + e_{jt} \]
\[ Y_j' = Y_j + \mu_j \]
\[ E(e_{jt} | j) = 0 \]
\[ V(e_{jt} | j) = \sigma^2_j \]
\[ \text{Cov}(e_{jt}, e_{j't} | j, j') = \rho \sigma_j \sigma_{j'}, \text{ for } j \neq j' \] (3.1.3.)

The quantity \( Y_j \) is the true value of the \( j^{th} \) unit and \( \mu_j \) the bias associated with the \( j^{th} \) unit. The errors in measuring two units \( U_j \) and \( U_j' \) have correlation \( \rho \). We shall call this model the simple correlation model. The effect of measurement errors under the simple correlation model will be discussed in this chapter.

Section 3.2 considers the usual estimator in probability proportional to size with replacement (ppswr) sampling. The bias and mean square error (MSE) of the estimator are derived under the simple correlation model. Also, the effect of measurement error on the variance estimator is studied. Section 3.3 discusses unequal probability sampling without replacement. In particular, the Horvitz-Thompson (1952) estimator is studied in detail. Finally, in Section 3.4, we examine the simple mean under simple random sampling both with and without replacement. This may be viewed as a particular case of unequal probability sampling. The results for the no-measurement error case are obtained as a special case.
3.2. Unequal Probability Sampling with Replacement
   Under the Simple Correlation Model

Consider the situation in which the values

\[ X_1, X_2, \ldots, X_N \]  \hspace{1cm} (3.2.1.)

of the \( x \)-characteristic of the units \( U_1, U_2, \ldots, U_N \) are known at the time of designing the survey. A sample of \( n \) units is selected according to probability proportional to size (pps) \( x \) with replacement, and observations are made on the \( y \)-characteristic of the selected units. We denote the sampling scheme by ppswr. The probability of selecting the \( j^{th} \) unit at each trial is \( p_j = X_j / \sum X_j \). Let the sample be

\[
\left( y_1, y_2, \ldots, y_n \right) \\
\left( p_1, p_2, \ldots, p_n \right)
\]  \hspace{1cm} (3.2.2.)

The subscript \( t \) is dropped when we are concerned with a single survey. We prove the following:

Theorem 3.2.1.

Under ppswr and the simple correlation model (3.1.3.), the bias of the estimator

\[ \hat{\bar{Y}_N} = \frac{1}{nN} \sum_{j=1}^{N} \frac{y_j}{p_j} \]  \hspace{1cm} (3.2.3.)

for estimating the population mean \( \bar{Y}_N \) is given by
and the variance of $\overline{Y}_N$ is given by

$$\hat{V}(\overline{Y}_N) = \frac{1}{n} \sum_{j} p_j \left( \frac{Y^*_j}{n p_j} - \overline{Y}_N \right)^2 + \frac{1}{n N} \sum_{j} \frac{\sigma^2_j}{p_j}$$

$$+ \frac{(n-1)}{n N^2} \sum_{j} \sigma^2_j + \frac{(n-1)}{n N^2} \rho \sum_{j \neq j'} \sigma_j \sigma_{j'}.$$  

(3.2.5)

Proof

Let $t_j$ be the number of times the $j^{th}$ unit appears in the sample.

Then the vector $(t_1, t_2, ..., t_N)$ follows a multinomial distribution.

Thus, $E(t_j) = n p_j$, $V(t_j) = n p_j (1 - p_j)$, and $Cov(t_j, t_{j'}) = -n p_j p_{j'}$.

The expectation of $\overline{Y}_N$ is

$$E(\overline{Y}_N) = \frac{1}{n} \sum_{j} \frac{n}{p_j} \frac{Y_j}{p_j} = \frac{1}{n} \sum_{j} \frac{n}{p_j} E_2 \left( \frac{Y_j}{p_j} \right)$$

$$= \frac{1}{n} \sum_{j} \frac{n}{p_j} \frac{Y_j}{p_j} = \frac{1}{n} \sum_{j} \frac{n}{p_j} Y_j = \overline{Y}_N + \overline{\mu}_N$$

where $E_2$ denotes the expectation over repeated measurements when the selected sample is held fixed, and $E_1$ the expectation over all possible samples. Note that $E = E_1 E_2$. Hence, the bias of $\overline{Y}_N$ is

$$B(\overline{Y}_N) = \overline{\mu}_N.$$
To obtain the variance, we have

\[ \hat{V}(\bar{Y}_N) = \hat{V}_1 + \hat{V}_2 \]

where \( \hat{V}_2 \) denotes the variance over repeated measurements when the selected sample is held fixed and \( \hat{V}_1 \) denotes the variance over all possible samples. Now

\[
\hat{V}_2(\bar{Y}_N) = \hat{V}_2(\frac{1}{nN} \sum_j y_j)
\]

\[
\quad = \hat{V}_2(\frac{1}{nN} \sum_j y_j)
\]

\[
\quad = \frac{1}{nN} \sum_j \frac{N}{2} t_j^2 + \sum_{j \neq j^*} \frac{t_j t_{j^*}}{P_j P_{j^*}} \text{Cov}_2(y_j, y_{j^*})
\]

\[
\quad = \frac{1}{nN} \sum_j \frac{N}{2} t_j^2 \cdot \sigma_j^2 + \sum_{j \neq j^*} \frac{t_j t_{j^*}}{P_j P_{j^*}} \rho \sigma_j \sigma_{j^*}
\]

where \( \text{Cov}_2(y_j, y_{j^*}) \) denotes the covariance between \( y_j \) and \( y_{j^*} \), (both \( U_j \) and \( U_{j^*} \), in the sample) over repeated measurements when the selected sample is held fixed. Thus

\[
\hat{E}_1 V_2(\bar{Y}_N) = \frac{1}{nN} \sum_j \frac{N}{2} \frac{(1 - P_j) + n^2 P_j^2}{P_j} \sigma_j^2
\]

\[
\quad + \sum_{j \neq j^*} \frac{n(n-1) P_j P_{j^*}}{P_j P_{j^*}} \rho \sigma_j \sigma_{j^*}
\]
Also

\[ \hat{E}_2(\overline{Y}_N) = \frac{1}{nN} \sum_j \frac{Y_j}{P_j} = \frac{1}{nN} \sum_j \frac{Y_j}{P_j} \]

\[ \hat{V}_1 \hat{E}_2(\overline{Y}_N) = \frac{1}{n} \sum_j P_j \left( \frac{Y_j}{NP_j} - \overline{Y}_N \right)^2 \]

Hence

\[ \hat{V}(\overline{Y}_N) = \hat{V}_1 \hat{E}_2(\overline{Y}_N) + \hat{E}_1 \hat{V}_2(\overline{Y}_N) \]

\[ = \frac{1}{n} \sum_j P_j \left( \frac{Y_j}{NP_j} - \overline{Y}_N \right)^2 + \frac{1}{nN^2} \sum_j \frac{\sigma_j^2}{P_j} \]

\[ + \frac{(n-1)}{nN^2} \sum_j \sigma_j^2 + \frac{(n-1)}{nN^2} \rho \sum_{j \neq j}^N \sigma_j \sigma_j \]

\[ \Sigma \]

**Remark 1.** The first term on the right hand side of (3.2.5.) is the sampling variance. The sum of the second and third terms is called the simple response variance, and the fourth term is called the correlated response variance. The sum of the second, third and fourth terms is consequently called "response variance."

**Remark 2.** In case measurement errors are absent, we find from (3.2.5.) by putting \( \mu_j = 0, \sigma_j = 0, j = 1, 2, \ldots, N, \) that
\[
V(\bar{Y}_N) = \frac{1}{n} \sum_{j}^{N} p_j (\frac{y_j}{NP_j} - \bar{Y}_N)^2
\]

which coincides with the expression given in textbooks (e.g., Raj, 1968; Sukhatme and Sukhatme, 1970; Cochran, 1977).

We shall now study the effect of measurement errors on the variance estimator \( \hat{V}(\bar{Y}_N) \) given as

\[
\hat{V}(\bar{Y}_N) = \frac{1}{n(n-1)} \sum_{j}^{n} \frac{y_j^2}{NP_j} - \frac{\hat{V}_N^2}{n} \tag{3.2.6}
\]

We have

\[
\hat{V}(\bar{Y}_N) = \frac{1}{n(n-1)} \left[ \sum_{j}^{n} \frac{y_j^2}{NP_j} - \frac{\hat{V}_N^2}{n} \right]
\]

Thus

\[
E\hat{V}(\bar{Y}_N) = \frac{1}{n(n-1)} \left[ E \sum_{j}^{n} \frac{y_j^2}{NP_j} - n E \left( \frac{\hat{V}_N^2}{n} \right) \right]
\]

\[
= \frac{1}{n(n-1)} \left[ E \sum_{j}^{n} \frac{y_j^2}{NP_j} - n \bar{Y}_N^2 - n V(\bar{Y}_N) \right]
\]

Now

\[
E \sum_{j}^{n} \frac{y_j^2}{NP_j} = E_1 \sum_{j}^{n} \frac{y_j^2}{NP_j}
\]

\[
= E_1 \sum_{j}^{n} \frac{1}{NP_j} \cdot (y_j^2 + \sigma_j^2)
\]
Hence

\[
E \hat{V(Y_N)} = \frac{1}{(n-1)} \left[ \sum_{j} \left( \frac{1}{N^2 P_j} (Y_j^2 + \sigma_j^2) - \bar{Y}_N^2 \right)^2 - \hat{V(Y_N)} \right]
\]

\[
= \frac{1}{(n-1)} \left[ \sum_{j} \frac{N Y_j^2}{N^2 P_j} - \bar{Y}_N^2 + \sum_{j} \frac{\sigma_j^2}{N^2 P_j} \right] - \hat{V(Y_N)}
\]

\[
= \frac{1}{n-1} \left[ n \hat{V(Y_N)} - \frac{1}{N^2} \sum_{j} \sigma_j^2 - \frac{(n-1)}{N^2} \sum_{j} \sigma_j^2 \right] - \hat{V(Y_N)}
\]

\[
= \hat{V(Y_N)} + \frac{1}{N^2} \sum_{j} \sigma_j^2 - \frac{1}{N^2} \rho \sum_{j \neq j'} \sigma_j \sigma_{j'}
\]

by using (3.2.5.). This shows that the customary variance estimator is biased if there are errors in the data. The bias of the estimator \(\hat{V(Y_N)}\) is

\[
\frac{1}{N^2} \sum_{j} \sigma_j^2 - \frac{1}{N^2} \rho \sum_{j \neq j'} \sigma_j \sigma_{j'}
\]

Hence, this estimator has a positive bias if
3.3. Unequal Probability Sampling Without Replacement
Under the Simple Correlation Model

We will now consider the situation in which a sample of $n$ units is selected without replacement according to an unequal probability sampling scheme. Let $\pi_i$ be the probability that the unit $U_i$ is selected in the sample. The probability $\pi_i$ is called the inclusion probability of unit $i$. Let $\pi_{ij}$ be the probability that both $U_i$ and $U_j$ are in the sample. Various sampling schemes are available in the literature for achieving these probabilities. Define

$$u_j = 1 \text{ if } j^{th} \text{ unit is in the sample}$$
$$= 0 \text{ otherwise.}$$

Then

$$E(u_j) = \pi_j, \ E(u_i u_j) = \pi_{ij}, \ E(u_j^2) = \pi_j.$$
Theorem 3.3.1.

Consider the estimator

\[ T = \sum_{j} w_j u_j y_{jt} \]

where \( w_j \)'s are constants. Under the simple correlation model given by (3.1.3.) we have

\[ \mathbb{E}(T) = \sum_{j} w_j \pi_j y_{j} \]

(3.3.1.)

and the variance of \( T \) is

\[ \text{Var}(T) = \sum_{j} w_j^2 \pi_j (1 - \pi_j) y_{j}^2 + \sum_{j \neq j'} w_j w_{j'} (\pi_{jj'} - \pi_j \pi_{j'}) y_{j} y_{j'} \]

\[ + \sum_{j} w_j^2 \pi_j \sigma_j^2 + \sum_{j \neq j'} w_j w_{j'} \pi_{jj'} \rho \sigma_j \sigma_{j'} \]

(3.3.2.)

Proof

We define \( E_1, V_1, E_2 \) and \( V_2 \) as in Theorem 3.2.1. Then

\[ \mathbb{E}(T) = \sum_{j} w_j \mathbb{E}(u_j y_{jt}) = \sum_{j} w_j E_1(u_j) E_2(y_{jt}) \]

\[ = \sum_{j} w_j E_1(u_j y_{j}) = \sum_{j} w_j \pi_j y_{j} \]
The derivation of \( V(T) \) is given as follows:

\[
E_2(T) = \sum_{j}^N w_j u_j y_j
\]

\[
V_2(T) = \sum_{j}^N w_j^2 u_j^2 V_2(y_{jt}) + \sum_{j \neq j'}^N w_j w_{j'} u_j u_{j'} \text{Cov}_2(y_{jt}, y_{j't})
\]

where \( \text{Cov}_2(y_{jt}, y_{j't}) \) denotes the covariance between \( y_{jt} \) and \( y_{j't} \) over repeated measurements when the selected sample is held fixed. Thus

\[
V_2(T) = \sum_{j}^N w_j^2 u_j^2 \sigma_j^2 + \sum_{j \neq j'} w_j w_{j'} \sigma_j \sigma_{j'}
\]

Therefore

\[
E_1 V_2 T = \sum_{j}^N w_j^2 \pi_j \sigma_j^2 + \sum_{j \neq j'} w_j w_{j'} \pi_{j'} \rho_{jj'} \sigma_j \sigma_{j'}
\]

and

\[
V_1 E_2 T = \sum_{j}^N w_j^2 V(u_j) Y_j^2 + \sum_{j \neq j'} w_j w_{j'} \text{Cov}(u_j, u_{j'}) Y_j Y_{j'}
\]

\[
= \sum_{j}^N w_j^2 \pi_j (1 - \pi_j) Y_j^2 + \sum_{j \neq j'} w_j w_{j'} (\pi_{j'} - \pi_{j}) \pi_j Y_j Y_{j'}
\]

Hence

\[
V(T) = V_1 E_2 T + E_1 V_2 T
\]

\[
= \sum_{j}^N w_j^2 \pi_j (1 - \pi_j) Y_j^2 + \sum_{j \neq j'} w_j w_{j'} (\pi_{j'} - \pi_j) \pi_j Y_j Y_{j'}
\]
Theorem 3.3.2.

Under unequal probability sampling without replacement and the simple correlation model (3.1.3.), the bias of the Horvitz-Thompson estimator

\[
\hat{Y}_{HT} = \frac{1}{N} \sum_{j} \frac{y_j}{\pi_j}
\]  

(3.3.3.)

for estimating \( \bar{Y}_N \) is

\[
\hat{B}(\hat{Y}_{HT}) = \mu_N
\]  

(3.3.4.)

and its variance is

\[
\hat{V}(\hat{Y}_{HT}) = \frac{1}{N^2} \sum_{j} \left( 1 - \frac{1}{\pi_j} \right) y_j^2 + \frac{1}{N^2} \sum_{j \neq j'} \left( \frac{\pi_{jj'}}{\pi_j \pi_{j'}} \right) y_j y_{j'} + \frac{1}{N^2} \sum_{j} \left( \frac{1}{\pi_j} \sigma_j^2 \right)
\]  

(3.3.5.)

Proof

Substituting \( w_j = (N \pi_j)^{-1} \) in equations (3.3.1.) and (3.3.2.), we have

\[
\hat{E}(\hat{Y}_{HT}) = \frac{1}{N} \sum_{j} y_j = \bar{Y}_N = \bar{Y}_N + \mu_N
\]
\[ V(\hat{Y}_{HT}) = \frac{1}{N^2} \sum_{j} \frac{(1-\pi_j)}{\pi_j} Y_j^2 + \frac{1}{N^2} \sum_{j \neq j'} \frac{\pi_{jj} - \pi_j \pi_{j'}}{\pi_j \pi_{j'}} Y_j Y_{j'} + \frac{1}{N^2} \sum_{j} \sigma_j^2 + \frac{1}{N^2} \sum_{j \neq j'} \frac{\pi_{jj} - \pi_j \pi_{j'}}{\pi_j \pi_{j'}} \rho \sigma_j \sigma_{j'} \]

The results follow.

**Remark.** If the individual variances \( \sigma_j^2 \) are zero, we obtain the customary expression for the variance (due to Horvitz and Thompson, 1952) applied to the values \( Y_j^2 \).

To study the effect of measurement errors on the variance estimator \( v_{HT} \) (due to Horvitz and Thompson) given by

\[ N^2 v_{HT} = \frac{n}{\pi_j^2} \sum_{j} y_j^2 + \frac{n}{\pi_j} \sum_{j \neq j'} \frac{\pi_{jj} - \pi_j \pi_{j'}}{\pi_j} Y_j Y_{j'} \quad (3.3.6) \]

we have

\[ E \sum_{j} \frac{n}{\pi_j^2} y_j^2 = E \sum_{j} \frac{n}{\pi_j} (Y_j^2 + \sigma_j^2) = \sum_{j} \frac{n}{\pi_j} (Y_j^2 + \sigma_j^2) \]

Also

\[ E \sum_{j \neq j'} \frac{\pi_{jj} - \pi_j \pi_{j'}}{\pi_j} \frac{Y_j Y_{j'}}{\pi_j} \pi_{j'} = \sum_{j \neq j'} \frac{\pi_{jj} - \pi_j \pi_{j'}}{\pi_j^2} \pi_{j'} (Y_j Y_{j'} + \rho \sigma_j \sigma_{j'}) \]
and hence

\[ E \sum_{j \neq j'} \frac{\pi_{jj'} - \pi_{j'j}}{\pi_j \pi_{j'}} \frac{Y_j - Y_{j'}}{\pi_{j'} - \pi_{j'}} = \sum_{j \neq j'} \frac{\pi_{jj'} - \pi_{j'j}}{\pi_j \pi_{j'}} (Y_j Y_{j'}' + \rho \sigma_j \sigma_{j'}) \]

Thus

\[ N^2 E(v_{HT}) = \sum_{j} \frac{1 - \pi_j}{\pi_j} Y_j^2 + \sum_{j \neq j'} \frac{\pi_{jj'} - \pi_{j'j}}{\pi_j \pi_{j'}} Y_j Y_{j'}, \]

\[ + \sum_{j} \frac{1 - \pi_j}{\pi_j} \sigma_j^2 + \sum_{j \neq j'} \frac{\pi_{jj'} - \pi_{j'j}}{\pi_j \pi_{j'}} \rho \sigma_j \sigma_{j'}, \]

\[ = N^2 \hat{V}(\hat{Y}_{HT}) - \sum_{j} \sigma_j^2 - \rho \sum_{j \neq j'} \sigma_j \sigma_{j'}, \quad (3.3.7) \]

If the measurement errors are positively correlated, the estimator \( v_{HT} \) underestimates the true variance of \( \hat{Y}_{HT} \). Even when the measurement errors are uncorrelated, the estimator \( v_{HT} \) has a negative bias for estimating \( \hat{V}(\hat{Y}_{HT}) \).

Another estimator of \( \hat{V}(\hat{Y}_{HT}) \) is due to Yates and Grundy (1953) and is given by

\[ v_{YG} = \frac{1}{N^2} \sum_{j \neq j'} \frac{n \pi_{jj'} - \pi_{j'j}}{\pi_j \pi_{j'}} \left( \frac{y_j - y_{j'}}{\pi_j - \pi_{j'}} \right) \quad (3.3.8) \]
It is easy to see that

\[ E(\bar{\nu}_{vG}) = V(\bar{Y}_{HT}) - \frac{1}{N^2} \sum_{j} \sigma_j^2 - \frac{2}{N} \sum_{j \neq j'} \sigma_j \sigma_{j'}. \]

Thus the two variance estimators have the same bias for estimating the variance.

3.4. Simple Random Sampling Under the Simple Correlation Model

When the units in the population do not vary considerably in size, the sample may be selected with equal probabilities. We shall derive results appropriate to this situation by making use of the results obtained in Sections 3.2 and 3.3. In the case of sampling with replacement, we shall substitute \( p_j = N^{-1} \) in the formulas (3.2.3) - (3.2.5). Thus the bias of the estimator

\[ \bar{Y}_N = \frac{1}{n} \sum_{j} y_j \]  

(3.4.1.)

for estimating the population mean \( \bar{Y}_N \) is

\[ B(\bar{Y}_N) = \bar{\mu}_N \]  

(3.4.2.)

and the variance of \( \bar{Y}_N \) is

\[ \hat{\sigma}_N^2 = \frac{1}{n} \sum_{j} \sigma_j^2 - \frac{2}{N} \sum_{j \neq j'} \sigma_j \sigma_{j'}. \]
The expected value of the variance estimator

\[ E(v) = V(\overline{Y}_N) + \frac{1}{2} \sum_{j=1}^{N} \sigma_j^2 - \frac{1}{2} \sum_{j \neq j'} \sigma_j \sigma_{j'}, \quad (3.4.5.) \]

is obtained as

When sampling is carried out without replacement with equal probabilities, we substitute

\[
\pi_j = \frac{n}{N}, \quad \pi_{jj'} = \frac{n(n-1)}{N(N-1)}
\]

\[
\frac{1-\pi_j}{\pi_j} = \frac{N-n}{n}, \quad \frac{\pi_{jj'}}{\pi_j \pi_{j'}} = \frac{n(N-1)}{N(N-1)}
\]

in the formulas (3.3.3.) - (3.3.5.). We have then

\[ \overline{Y}_N = \frac{1}{n} \sum_{j=1}^{n} y_j \]

\[ B(\overline{Y}_N) = \overline{u}_N \]

(3.4.6.)

(3.4.7.)
The expected value of the variance estimator becomes

\[ E(v) = V(\hat{\Sigma}) - \frac{1}{N} \sum_{j=1}^{N} \sigma_j^2 - \frac{1}{n} \sum_{j\neq j'}^{N} \sigma_j^2 \sigma_{j'}^2, \]

Some check on the results so far presented can be made by considering the situation in which

\[ E(e_{jt}|j) = \mu_j = 0 \]

\[ V(e_{jt}|j) = \sigma_j^2 = 0 \]

\[ \text{Cov}(e_{jt}, e_{j't}|j, j') = \rho \sigma_j \sigma_{j'} = 0 \]
In this case, the true values of y are always observed. Substituting these values in the formulas (3.4.1.) - (3.4.8.), we obtain the following results:

**Sampling with replacement**

\[
\hat{\bar{Y}}_n = \frac{1}{n} \sum_{j} y_j
\]

\[
\hat{B}(\hat{\bar{Y}}_n) = 0
\]

\[
\hat{V}(\hat{\bar{Y}}_n) = \frac{1}{nN} \sum_{j} (y_j - \bar{Y})^2
\]

\[
v = \frac{1}{n(n-1)} \sum_{j} (y_j - \bar{y}_n)^2
\]

\[
E(v) = \hat{V}(\hat{\bar{Y}}_n)
\]

**Sampling without replacement**

\[
\hat{\bar{Y}}_n = \frac{1}{n} \sum_{j} y_j
\]

\[
\hat{B}(\hat{\bar{Y}}_n) = 0
\]

\[
\hat{V}(\hat{\bar{Y}}_n) = \left(\frac{1}{n} - \frac{1}{N}\right) \frac{1}{N-1} \sum_{j} (y_j - \bar{Y})^2
\]

\[
v = \left(\frac{1}{n} - \frac{1}{N}\right) \frac{1}{n-1} \sum_{j} (y_j - \bar{y}_n)^2
\]

\[
E(v) = \hat{V}(\hat{\bar{Y}}_n)
\]
These formulas are the same as those given in the textbooks (e.g., Raj, 1968; Cochran, 1977) for the situation when the true values of $y$ are observed.
4. RATIO ESTIMATION IN THE PRESENCE OF MEASUREMENT ERRORS

4.1. Introduction

When auxiliary information is available, it may be used to construct a ratio estimator which is widely used in practice. In the ratio method of estimation, observations are made on both the x and y characteristics of the units in the sample. The investigator is interested in estimating the population mean of y while x, which is correlated with y, is an auxiliary variable. We assume that y is measured with error and x is measured without error and examine the effects of measurement error in ratio estimation. In practice \( X_i \) (the value of x-characteristic of \( U_i \)) is usually the value of \( Y_i \) (the true value of the y-characteristic of \( U_i \)) at some previous time. Consequently, it is reasonable to assume that the x-characteristic has been thoroughly checked and is measured without error. The ratio estimator of the population mean \( \bar{Y}_N \) is

\[
\bar{Y}_N = \frac{\bar{y}_n \cdot \bar{X}_n}{\bar{X}_n}
\]  

(4.1.1.)

We consider two sampling schemes: (1) simple random sampling, and (2) Midzuno unequal probability sampling. Two models for measurement error are considered in this chapter. The first model, called the simple correlation model, is given by
\[ y_{jt} = Y_j^\prime + e_{jt} \]
\[ Y_j^\prime = Y_j + \mu_j \]
\[ E(e_{jt}) = 0 \]
\[ \text{Cov}(e_{jt}, e_{j',t} | j, j') = \sigma_j^2 \quad \text{for} \quad j = j' \]
\[ = \rho \sigma_j \sigma_{j'}, \quad \text{for} \quad j \neq j' \quad (4.1.2.) \]

where \( y_{jt} \) is the observed value of the \( y \)-characteristic of \( U_j \) in the \( t \)th survey (or trial), \( Y_j \) the true value of the \( y \)-characteristic of \( U_j \), and \( \mu_j \) the bias associated with observing the \( y \)-characteristic of \( U_j \). The second model, called the intrasample correlation model, is given by

\[ y_{jt} = Y_j^\prime + e_{jt} \]
\[ Y_j^\prime = Y_j + \mu_j \]
\[ E(e_{jt} | j) = 0 \]
\[ V(e_{jt} | j) = \sigma_j^2 \]
\[ E(e_{jt} e_{j',t}) = \rho \sigma_j \sigma_{j'} E(e_{jt}^2) \quad \text{for} \quad j \neq j' \quad (4.1.3.) \]

and both \( j \) and \( j' \) in the sample.
Let $\tau_j = \sigma_j^2$. We define

\[
\overline{y}_n = n^{-1} \sum_{j} y_j, \quad \overline{x}_n = n^{-1} \sum_{j} x_j
\]

\[
\overline{\tau}_n = n^{-1} \sum_{j} \tau_j, \quad \overline{\sigma}_n = n^{-1} \sum_{j} \sigma_j
\]

\[
\overline{V}_N = N^{-1} \sum_{j} V_j, \quad \overline{X}_N = N^{-1} \sum_{j} X_j
\]

\[
\overline{\tau}_N = N^{-1} \sum_{j} \tau_j, \quad \overline{\sigma}_N = N^{-1} \sum_{j} \sigma_j
\]

\[
S^2_{\chi} = (N-1)^{-1} \sum_{j} (X_j - \overline{X}_N)^2, \quad S^2_{\sigma} = (N-1)^{-1} \sum_{j} (\sigma_j - \overline{\sigma}_N)^2
\]

\[
S^2_{\tau} = (N-1)^{-1} \sum_{j} (\tau_j - \overline{\tau}_N)^2
\]

\[
S_{XY} = (N-1)^{-1} \sum_{j} (X_j - \overline{X}_N)(Y_j - \overline{V}_N)
\]

\[
S_{X\tau} = (N-1)^{-1} \sum_{j} (X_j - \overline{X}_N)(\tau_j - \overline{\tau}_N)
\]

\[
S_{X\sigma} = (N-1)^{-1} \sum_{j} (X_j - \overline{X}_N)(\sigma_j - \overline{\sigma}_N)
\]
\[ C_X = \frac{S_X}{\bar{X}_N} \quad C_{XY} = \frac{S_{XY}}{\bar{X}_N \bar{Y}_N} \]
\[ C_{Xr} = \frac{S_{Xr}}{\bar{X}_N \bar{r}_N} \quad a = (N-n)^{-1} (N-n) \]

The terms \( \bar{\mu}_n, \bar{Y}_n, \bar{\mu}_N, \bar{V}_N, S_Y, S_Y^2, S_r^2, S_{XYr}, C_Y, C_{Yr}, C_{Xr}, C_{XYr}, \) are defined similarly.

Let \( \hat{A}_1 = (\bar{V}_R, \text{srswor}) \) and \( \hat{A}_2 = (\bar{V}_R, \text{Ms}) \) where srswor and Ms denote simple random sampling without replacement and Midzuno's scheme of sampling respectively.

The bias and mean square error of the ratio estimator is derived under simple random sampling without replacement and each of the two measurement error models (4.1.2.) and (4.1.3.) in Section 4.2. In Section 4.3, we derive the bias and mean square error of the ratio estimator under Midzuno's scheme of sampling and each of the two measurement error models. A comparison of the two strategies, \( \hat{A}_1 \) and \( \hat{A}_2 \), under each of the two measurement error models is made in Section 4.4.

The case, when both \( y \) and \( x \) are measured with error, is very complicated though the derivation is similar. We shall not present that case in this thesis.

4.2. Ratio Estimation Under Equal Probability Sampling

Suppose a simple random sample of size \( n \) is taken, and both \( y \) and \( x \) are measured. We have the following theorem.

**Theorem 4.2.1.**

Under simple random sampling without replacement and the simple
correlation model (4.1,2.), the bias of the ratio estimator

\[ \hat{Y}_R = \frac{\bar{y}_n}{\bar{x}_n} \cdot \bar{X}_N \]  \hfill (4.2.1.)

for estimating \( \bar{Y}_N \) is given by

\[ \text{B}(\hat{Y}_R) \approxeq \mu_n + a\bar{Y}_N [C_X^2 - C_{XY^*}] \]  \hfill (4.2.2.)

and its mean square error (MSE) is

\[ \text{MSE}(\hat{Y}_R) \approxeq a\bar{Y}_N^2 [C_Y^2 + C_X^2 - 2C_{XY^*}] \]

\[ + \frac{(1-\rho)}{n} \tau_N [1 - 2ac_{Xr} + 3ac_X^2] \]

\[ + \rho(\sigma_N^2) [1 - 4ac_{Xr} + 3ac_X^2 + ac_{r}^2] \]

\[ + \mu_n^2 + 2a\mu_n \bar{Y}_N [C_X^2 - C_{XY^*}] \]  \hfill (4.2.3.)

**Proof**

Let \( E_1, E_2, V_1 \) and \( V_2 \) be defined as in the proof of Theorem 3.2.1., i.e., \( E_2, V_2 \) are the expectation and variance over repeated measurements given the sample, and \( E_1, V_1 \) are the expectation and variance over all possible samples. From equation (10.1), we have on substituting
\( \theta_i = y_i \) and \( \delta_i = x_i \) for every \( i = 1, 2, \ldots, N \), that

\[
E(\hat{Y}_R) = E(\frac{\hat{y}_n}{\hat{x}_n} \cdot \hat{x}_N)
\]

\[
= \hat{x}_N \cdot E(\frac{\hat{y}_n}{\hat{x}_n})
\]

\[
= \hat{x}_N \cdot E(1)\frac{\hat{y}_n}{\hat{x}_n}
\]

\[
= \hat{x}_N \cdot E(\frac{\hat{y}_n}{\hat{x}_n})
\]

\[
\equiv \hat{x}_N \cdot E(\frac{\hat{y}_n}{\hat{x}_n}) \left[ 1 - a \sigma_{XY} + a \sigma^2_X \right]
\]

\[
= \hat{y}_N \left[ 1 - a \sigma_{XY} + a \sigma^2_X \right]
\]

\[
= \hat{y}_N + a \hat{y}_N \left[ \sigma^2_X - \sigma_{XY} \right]
\]

Thus

\[
B(\hat{Y}_R) \equiv \mu_N + a \nu_N \left[ \sigma^2_X - \sigma_{XY} \right]
\]

For the variance, we have
\[ \hat{V}(\bar{Y}_R) = \hat{V}_1 E_2(\bar{Y}_R) + \hat{E}_1 \hat{V}_2(\bar{Y}_R) \]

where \( \hat{V}_1 E_2(\bar{Y}_R) \) is the sampling variance and \( \hat{E}_1 \hat{V}_2(\bar{Y}_R) \) is the response variance. Now

\[ \hat{E}_2(\bar{Y}_R) = E_2(\frac{\bar{y}_n}{X_n} \cdot \bar{X}_n) = \bar{Y}_* \frac{\bar{v}_n}{X_n} \]

Using the standard formula for the variance of the ratio estimate, correct to the first approximation (Cochran, 1977; Sukhatme and Sukhatme, 1970), we have

\[ \hat{V}_1 E_2(\bar{Y}_R) = V_1(\bar{X}_N) \frac{\bar{y}_n}{X_n} \]

\[ \frac{N-n}{N} \left[ \frac{S_Y^2}{X_N} + \frac{2\bar{Y}_*}{X_N^2} \frac{S_X^2}{X_N} - 2 \frac{\bar{Y}_*}{X_N} S_{XY} \right] \]

\[ = a \frac{\bar{Y}_*^2}{N} \left[ c_Y^2 + c_X^2 - 2c_{XY} \right] \]

For the response variance, we have

\[ \hat{E}_1 \hat{V}_2(\bar{Y}_R) = \hat{E}_1 E_2(\bar{Y}_R - E_2 \bar{Y}_R)^2 \]

\[ = E_1 E_2(\frac{\bar{y}_n}{X_n} \cdot \frac{\bar{y}_n}{X_n} - \frac{\bar{y}_n}{X_n})^2 \]
\[= E_1 \frac{\bar{X}_n}{X_n} \left[ \frac{1}{n} \sum_{j} e_{jt} \right]^2 \]

\[= E_1 \frac{\bar{X}_n}{X_n} \cdot \frac{1}{n} \left[ \sum_{j} e_{jt}^2 + \sum_{j \neq k} e_{jt} e_{kt} \right] \]

\[= E_1 \frac{\bar{X}_n}{X_n} \frac{1}{n} \left[ \sum_{j} \sigma_j^2 + \sum_{j \neq k} \rho \sigma_j \sigma_k \right] \]

Now

\[\frac{1}{n^2} \left[ \sum_{j} \sigma_j^2 + \sum_{j \neq k} \rho \sigma_j \sigma_k \right] = \frac{\tau_n}{n} + \frac{\rho}{n} \left( \sum_{j} \sigma_j^2 - \frac{1}{n} \right) \]

\[= \frac{\tau_n}{n} + \rho \left( \frac{\sigma_n}{n} \right)^2 - \frac{\rho}{n} \tau_n \]

\[= \frac{\tau_n}{n} (1 - \rho) + \rho \left( \frac{\sigma_n}{n} \right)^2 \]

Thus

\[E_1 \frac{\bar{X}_n}{X_n} \left( \frac{\bar{X}_R}{X_R} \right) = \frac{\bar{X}_n}{X_n} E_1 \tau_n \left[ \frac{1 - \rho}{n} \right] + \rho \left( \frac{\sigma_n}{n} \right)^2 \]

Using the formulas (10.2.), and (10.4.), we have

\[E_1 \frac{\tau_n}{X_n} = \frac{\tau_n \left[ 1 - 2aC_X + 3a^2 \right]}{X_n} \]
and

$$\frac{(\overline{c}_n)^2}{X_n^2} = \frac{(\overline{c}_N)^2}{X_N^2} \left[ 1 - 4ac_{Xr} + 3ac_x^2 + ac_o^2 \right]$$

Thus, the response variance is

$$E_1 \overline{V}_2(\overline{Y}_R) = \frac{(1-\rho)}{n} \tau_N \left[ 1 - 2ac_{Xr} + 3ac_x^2 \right]$$

$$+ \left( \overline{c}_N \right)^2 \left[ 1 - 4ac_{Xr} + 3ac_x^2 + ac_o^2 \right]$$

Now

$$B(\overline{Y}_R) = \overline{\mu}_N + a\overline{V}_N \left[ c_x^2 - c_{XY} \right]$$

Hence, to the first order of approximation, we have

$$[B(\overline{Y}_R)]^2 = \mu_n^2 + 2a\overline{\mu}_N \overline{V}_N \left[ c_x^2 - c_{XY} \right]$$

Therefore

$$MSE(\overline{Y}_R) = V_1 \overline{E}_2(\overline{Y}_R) + E_1 \overline{V}_2(\overline{Y}_R) + [B(\overline{Y}_R)]^2$$

$$= a\overline{V}_N \left[ c_y^2 + c_x^2 - 2c_{XY} \right]$$

$$+ \left( \frac{1-\rho}{n} \right) \tau_N \left[ 1 - 2ac_{Xr} + 3ac_x^2 \right]$$
Remark 1. The mean square error of \(^\wedge Y_R\) can be written as

\[
\text{MSE}(\hat{Y}_R) = \alpha \bar{Y}_N^2 \left[ C_Y^2 + C_X^2 - 2C_{XY} \right]
\]

\[
+ \frac{\tau N}{n} \left[ 1 - 2\alpha C_X + 3\alpha C_X^2 \right]
\]

\[
+ \rho \left[ \frac{\tau N}{n} (1 - 2\alpha C_X + 3\alpha C_X^2) \right]
\]

\[
- (\sigma_N^2)^2 \left( 1 - 4\alpha C_X + 3\alpha C_X^2 + \alpha^2 \sigma^2 \right)
\]

\[
+ \mu_N \left[ \mu_N + 2\alpha \bar{Y}_N (C_X^2 - C_{XY}) \right]
\]

(4.2.4.)

The first term on the right hand side of (4.2.4.) is the sampling variance, the second term the "simple response variance," the third term the "correlated response variance," and the last term the square of the bias of the estimator.

Remark 2. Let \(\sigma_j = \sigma\) for all \(j = 1, \ldots, N\), then
\[ \bar{\tau}_N = N^{-1} \sum_{j=1}^{N} \sigma_j^2 = \sigma^2, \quad \bar{\sigma}_N = N^{-1} \sum_{j=1}^{N} \sigma_j = \sigma, \quad C = 0, \]

\[ C_X = 0, \text{ and } C_{XY} = 0. \]

Thus

\[ \text{MSEC}_{Y,R} = a \hat{Y}^2 + c_Y^2 + c_X^2 - 2c_{XY}. \]

\[ + \frac{1 - \rho}{n} \sigma^2 \left[ 1 + 3ac_X^2 \right] \]

\[ + \rho \sigma^2 \left[ 1 + 3ac_X^2 \right] \]

\[ + \frac{\mu^2}{n} + 2a \mu \hat{Y} \left[ c_X^2 - c_{XY} \right] \quad (4.2.5.) \]

Now we consider the intrasample correlation model (4.1.3.). Under simple random sampling without replacement, we have

\[ E(e_{jt}^2) = E[E(e_{jt}^2 | j)] \]

\[ = E \tau_j = \sum_{j=1}^{N} N^{-1} \tau_j \]

\[ = \bar{\tau}_N. \]
Theorem 4.2.2.

Under simple random sampling without replacement and the intra-sample correlation model (4.1.3.), the bias of the estimator

\[ \hat{\bar{Y}}_R = \frac{\bar{y}_n}{\bar{x}_n} \]  \hspace{1cm} (4.2.6.)

is given by

\[ B(\hat{\bar{Y}}_R) \equiv \bar{\mu}_N + a\bar{Y}_N \left[ C_X^2 - C_{XY} \right] \]  \hspace{1cm} (4.2.7.)

and its mean square error is given by

\[ \text{MSE}(\hat{\bar{Y}}_R) \equiv a\bar{Y}_N^2 \left[ C_Y^2 + C_X^2 - 2C_{XY} \right] + \frac{\tau_N}{n} \left[ 1 - 2aC_{X_T} + 3aC_X^2 \right] + \frac{(n-1)}{n} \rho \tau_N \bar{\mu}_N \]  \hspace{1cm} (4.2.8.)

Proof

The expression for the bias is obtained as in Theorem 4.2.1., and the result is the same.

To determine the mean square error, we first find the variance.
We have

\[ V(\bar{V}_n) = V_1 E_2(\bar{V}_n) + E_1 V_2(\bar{V}_n) \]

The sampling variance, \( V_1 E_2(\bar{V}_n) \), is the same as in Theorem 4.2.1. We have

\[ V_1 E_2(\bar{V}_n) = \sigma^2 \left[ c_Y^2 + c_X^2 - 2c_{XY} \right] \]

For the response variance, we have

\[ E_1 V_2(\bar{V}_n) = E_1 E_2(\bar{V}_n) - E_2(\bar{V}_n)^2 \]

\[ = E_1 \left\{ E_2 \left( \frac{\bar{Y}}{X} \right) \right\} - E_2(\bar{V}_n)^2 \]

\[ = E_1 \left\{ \frac{\bar{X}}{X} E_2 \left( \frac{\bar{Y}}{X} \right) \right\} - E_2(\bar{V}_n)^2 \]

\[ = \bar{X}^2 \left\{ E_1 \left[ \frac{1}{n} \sum_{j=1}^{n} \left( \sum_{j=1}^{n} e_{jt} \right) \right] \right\} - E_2(\bar{V}_n)^2 \]

\[ = \bar{X}^2 \left\{ E_1 \left[ \frac{1}{n} \sum_{j=1}^{n} \tau_j \right] \right\} + \bar{X}^2 \left\{ E_1 \left[ \frac{1}{n} \sum_{j=1}^{n} e_{jt} \cdot e_{j,t} \right] \right\} \]
Now

\[
E_1[\frac{1}{X_n^2} \cdot (\frac{1}{n^2} \sum_j \tau_j)] = E_1[\frac{\tau_n}{n X_n^2}]
\]

\[
= \frac{\tau_N}{n X_N^2} [1 - 2a c_x + 3a c_x^2]
\]

by equation (10.2.). Also, on expanding \(\bar{X}_n^{-2}\) in Taylor series around \(X_N\), we have

\[
E \left( \frac{1}{n^2} \sum_{j \neq j'} e_j e_j' t \bar{X}_n^{-2} = E \left( \frac{1}{n^2} \sum_{j \neq j'} e_j e_j' t \bar{X}_N^{-2}
\right)
\]

\[
1 - 2 \frac{(\bar{X}_n - \bar{X}_N)}{\bar{X}_N} + \frac{3(\bar{X}_n - \bar{X}_N)^2}{\bar{X}_N^2} + \ldots \]

For large \(n\), \(E[e_j e_j' t (\bar{X}_n - \bar{X}_N)^2]\) will be negligible, and so will be the higher order terms. Also, \(E[e_j e_j' t (\bar{X}_n - \bar{X}_N)] = 0\). Hence,

\[
E \left( \frac{1}{n^2} \sum_{j \neq j'} e_j e_j' t \bar{X}_n^{-2} = E \left( \frac{1}{n^2} \sum_{j \neq j'} e_j e_j' t \bar{X}_N^{-2}
\right)
\]

\[
= \frac{\bar{X}_N^{-2}}{n^2} n(n-1) \rho \sigma \bar{X}_N
\]
Thus

\[ \frac{\bar{x}_{N}^{-2}}{n} (n-1) \rho_w \tau_N \]

Hence

\[ \frac{\tau_N}{n} \left[ 1 - 2ac_{X} + 3a^2c^2 \right] + \frac{(n-1)}{n} \rho_w \tau_N \]

We have, as obtained in Theorem 4.2.1.,

\[ \frac{\bar{a}_{X}}{\bar{Y}_{N}} \left[ b_{XY} \right] = \mu^2 + 2a\bar{\mu}N \bar{Y}_{N} \left[ c_{X}^2 - c_{XY} \right] \]

Therefore

\[ \text{MSE} (\bar{Y}_{r}) = \frac{\bar{a}_{X}^2}{\bar{Y}_{N}} \left[ c_{Y}^2 + c_{X}^2 - 2c_{XY} \right] + \frac{\tau_N}{n} \left[ 1 - 2ac_{X} + 3a^2c^2 \right] \]
4.3. Ratio Estimation Under Midzuno's Scheme of Sampling

In Midzuno's scheme of sampling, the first unit is selected with probability proportional to \( x \), and the remaining \((n-1)\) units are selected according to simple random sampling without replacement. Let \( X_1, X_2, \ldots, X_n \) be the values of the \( x \)-characteristic of the units selected in the sample. Then the probability of selecting the sample is

\[
P(s) = \frac{n\bar{X}_n}{N\bar{X}_N} \cdot \frac{1}{(n-1)}
\]

The following lemma is useful in determining expectations under Midzuno's scheme of sampling.

**Lemma 4.3.1.**

Under Midzuno's sampling scheme

\[
E[f(\bar{X}_n, \bar{Y}_n) \bar{X}_n^{-1}] = \bar{X}_N^{-1} E_{SRS} \{f(\bar{X}_n, \bar{Y}_n)\}
\]

where \( E_{SRS} \) denotes the expectation under simple random sampling without
replacement, and \( f(\overline{x}_n, \overline{y}_n) \) is any function of \( \overline{x}_n \) and \( \overline{y}_n \).

**Proof**

\[
E[f(\overline{x}_n, \overline{y}_n) \overline{x}_n^{-1}] = \sum_{s=1}^{N_n} \frac{\binom{N_n}{s}}{N_n} \frac{f(\overline{x}_n, \overline{y}_n)}{\overline{x}_n} \frac{n\overline{x}_n}{N_n} \frac{1}{\binom{N_n}{s-1}}
\]

\[
= \overline{x}_n^{-1} \frac{1}{\binom{N_n}{n}} \sum_{s=1}^{N_n} f(\overline{x}_n, \overline{y}_n)
\]

\[
= \overline{x}_n^{-1} E_{SRS} \{ f(\overline{x}_n, \overline{y}_n) \} \quad \square
\]

We now prove the following theorem.

**Theorem 4.3.1.**

Under Midzuno's sampling scheme and the simple correlation model (4.1.2.), the bias of the estimator

\[
\hat{B}(\overline{y}_R) = \frac{\overline{y}_n}{\overline{x}_n} \cdot \overline{x}_N
\]

is given by

\[
B(\overline{y}_R) = \overline{\mu}_N
\]

and its variance is given by

\[
V(\overline{y}_R) = \overline{\alpha}_n^2 \left[ c_Y^2 + c_X^2 - 2c_{XY} \right]
\]
The mean square error of the estimator is

\[ \text{MSE}(\hat{Y}) = \text{V}(\hat{Y}) + \mu_N^2 \]  

(4.3.4.)

where \( \text{V}(\hat{Y}) \) is given by (4.3.3.).

Proof

Let \( E_1, E_2, V_1 \) and \( V_2 \) be defined as in the proof of Theorem 4.2.1., then

\[
\text{E}(\hat{Y}_R) = \text{E}(\frac{\bar{Y}_n}{\bar{X}_n}) \\
= \bar{X}_N \text{E}(\frac{\bar{Y}_n}{\bar{X}_n}) \\
= \bar{X}_N \text{E}(\frac{\bar{Y}_n}{\bar{X}_n}) \\
= \bar{X}_N \cdot \frac{1}{\bar{X}_N} \cdot \bar{V}_N \\
= \bar{V}_N
\]
on using Lemma 4.3.1. Thus

\[ \hat{B}(\bar{Y}_R) = \bar{\mu}_N \]

For the variance, we have

\[ \hat{V}(\bar{Y}_R) = V_1 E_2(\bar{Y}_R) + E_1 \hat{V}_2(\bar{Y}_R) \]

Now

\[ V_1 E_2(\bar{Y}_R) = V_1 \frac{\bar{\nu}_n}{\bar{x}_n} \cdot \bar{x}_N \]

Singh (1975) observed that the variance of the ratio estimator under Midzuno's sampling scheme is the same as that under simple random sampling without replacement if terms up to \( O(n^{-1}) \) are considered. Hence

\[ V_1 E_2(\bar{Y}_R) = a \bar{\nu}_n^2 [c_Y^2 + c_X^2 - 2c_{XY}] \]

For the response variance, we have

\[ E_1 \hat{V}_2(\bar{Y}_R) = E_1 E_2(\bar{Y}_R) - E_2(\bar{Y}_R)^2 \]

\[ = E\left[ \frac{\bar{Y}_n}{\bar{x}_n} \cdot \bar{x}_N - \frac{\bar{\nu}_n}{\bar{x}_n} \bar{x}_N \right]^2 \]

\[ = E\left[ \frac{\bar{X}_N}{\bar{x}_n} \cdot \frac{1}{n} \sum_{j=1}^{n} x_j \right]^2 \]
as obtained in the proof of Theorem 4.2.1. Using Lemma 4.3.1. and equation (10.1.), we have

\[
E \frac{\bar{\tau}_n^2}{X_n^2} = \frac{1}{X_N} E_{\text{SRS}} \frac{\bar{\tau}_n^2}{X_n^2} = \frac{\bar{\tau}_N}{X_N^2} [1 - ac_{Xr} + ac_{X}^2]
\]

Also, using Lemma 4.3.1. and equation (10.3.), we have

\[
E (\frac{\bar{\sigma}_n^2}{X_n^2}) = \frac{1}{X_N} E_{\text{SRS}} (\frac{\bar{\sigma}_n^2}{X_n^2}) = \frac{(\bar{\sigma}_N^2)}{X_N^2} [1 - 2ac_{Xo} + ac_{X}^2 + ac_{o}^2]
\]

Thus

\[
E V_2 (\bar{Y}_R) \leq \frac{(1 - \rho)}{n} \bar{\tau}_N [1 - ac_{Xr} + ac_{X}^2]
\]
The result (4.3.4.) is obtained from the definition of mean square error.

**Remark 1.** The variance of the estimate \( \bar{Y}_R \) can be written as

\[
\hat{\gamma} (\bar{Y}_R) = \frac{\sigma^2}{Y_N} [c^2_Y + c^2_X - 2c_{XY}]
\]

\[
+ \frac{(1-\rho)}{n} \tau_N [1 - ac_{X_H} + ac^2_X]
\]

\[
+ \frac{\sigma^2}{Y_N} [1 - 2ac_{X_H} + ac^2_X + ac^2_Y]
\]

\[
+ \frac{\sigma^2}{Y_N} [1 - 2ac_{X_H} + ac^2_X + ac^2_Y]
\]

\[
(4.3.5.)
\]
The first term on the right hand side of (4.3.5.) represents the sampling variance, the second term the simple response variance, and the sum of the third and fourth terms the correlated response variance.

\textbf{Remark 2.}

Let $\sigma_j = \sigma$ for all $j = 1, 2, \ldots, N$, then $\tau_N = N^{-1} \sum_{j=1}^{N} \sigma_j^2 = \sigma^2$, where

$$\tau_j = \frac{1}{N} \sum_{j=1}^{N} \sigma_j^2 = \sigma^2, \quad c_{ij} = 0, \quad c_{XX} = 0, \quad c_{X\sigma} = 0.$$ Thus

$$V(\bar{Y}_R) = \frac{n}{N^2} \left[ c_{Y}^2 + c_{X}^2 - 2c_{XY} \right]$$

$$+ \frac{(1 - p)}{n} \sigma^2 \left[ 1 + a c_{X}^2 \right]$$

$$+ \rho \sigma^2 \left[ 1 + a c_{X}^2 \right]$$

$$= \frac{n}{N^2} \left[ c_{Y}^2 + c_{X}^2 - 2c_{XY} \right]$$

$$+ \frac{\sigma^2}{n} \left( 1 + ac_{X}^2 \right) \left[ 1 + (n-1)p \right] \quad (4.3.6.)$$

\textbf{Theorem 4.3.2.}

Under Midzuno's sampling scheme and the intrasample correlation model (4.1.3.), the bias of the ratio estimator

\[ \text{bias}(\hat{Y}_R) = \frac{N}{n} \left[ \frac{c_{Y}}{c_{X}} - p \right] \frac{1}{\sigma^2} \left[ c_{Y}^2 + c_{X}^2 - 2c_{XY} \right] \]

\[ + \frac{n}{N} \left( 1 + ac_{X}^2 \right) \left[ 1 + (n-1)p \right] \]
is given by

\[ \hat{B}(\bar{Y}_R) = \mu_n \]  

(4.3.8.)

and its variance is given by

\[ \hat{\nu}(\bar{Y}_R) = a\frac{n}{n} \left[ \frac{c^2_Y}{X^2} + c^2_X - 2c_{X,Y} \right] \]

\[ + \frac{\tau_n}{n} \left[ 1 - ac_{X,Y} + ac^2_X \right] \]

\[ + \frac{(n-1)}{n} \rho_{\omega \tau_n} \]

(4.3.9.)

The mean square error of the estimator (4.3.7.) is

\[ \hat{\text{MSE}}(\bar{Y}_R) = \hat{\nu}(\bar{Y}_R) + \mu_n^2 \]  

(4.3.10.)

where \( \hat{\nu}(\bar{Y}_R) \) is given by (4.3.9.).

**Proof**

The expression for the bias is obtained as in Theorem 4.3.1. For the variance, we note that

\[ \hat{\nu}(\bar{Y}_R) = \nu E(\bar{Y}_R) + E(\nu) \hat{\nu}(\bar{Y}_R) \]
The sampling variance, $\hat{V}_{2}(\bar{Y}_R)$, is obtained exactly as in Theorem 4.3.1., and we obtain

$$\hat{V}_{2}(\bar{Y}_R) = a^2 \bar{Y}_n^2 \left[ c_Y^2 + c_X^2 - 2c_{XY} \right]$$

For the response variance, we have

$$E_1 V_2(\bar{Y}_R) = \bar{X}_n^2 \frac{\sum_{j=1}^{n} \left( \frac{1}{n} \sum_{j \neq j'} e_{jj'} \right)}{n}$$

as obtained in the proof of Theorem 4.2.2. Using Lemma 4.3.1. and equation (10.1.), we have

$$E_1 \frac{\tau_n}{\bar{X}_n} = \frac{1}{\bar{X}_N} \frac{\tau_n}{\bar{X}_N}$$

$$= \frac{\tau_n}{\bar{X}_N} \left[ 1 - a c_{Xt} + a c_X^2 \right]$$

Also, on using Lemma 4.3.1. and Taylor's expansion of $\bar{X}_N^{-1}$, we have

$$E \frac{1}{\bar{X}_n^2} \sum_{j \neq j'} e_{jj'} \frac{E_{SRS}}{\bar{X}_n} = \frac{1}{\bar{X}_N} \frac{E_{SRS}}{\bar{X}_N} \left( \sum_{j \neq j'} e_{jj'} \right)$$

$$= \frac{1}{\bar{X}_N} \frac{E_{SRS}}{\bar{X}_N} \left( \sum_{j \neq j'} e_{jj'} \right)$$
For large $n$, $E_{srs}\{e_{jt} e_{jt'} (\bar{X}_n - \bar{X}_N)^2\}$ will be negligible, and so will be the higher order terms. Also, $E_{srs}\{e_{jt} e_{jt'} (\bar{X}_n - \bar{X}_N)\} = 0$. Hence,

$$E_{\bar{X}_n^{-2}} = \sum_{j\neq j'} e_{jt} e_{jt'} = \bar{X}_n^{-2} \sum_{j\neq j'} e_{jt} e_{jt'}$$

$$= \bar{X}_n^{-2} n(n-1) \rho_w \tau_N$$

Thus,

$$V(\hat{Y}_R) = \frac{\tau_N}{n} [1 - aC_{Xr} + aC_{X}^2] + \frac{(n-1)}{n} \rho_w \tau_N$$

Hence,

$$V(\hat{Y}_R) = a\bar{Y}_N^2 \left[C_{Y_+}^2 + C_X^2 - 2C_{XY}\right]$$

$$+ \frac{\tau_N}{n} [1 - aC_{Xr} + aC_{X}^2] + \frac{(n-1)}{n} \rho_w \tau_N$$

Thus, the mean square error is given by

$$MSE(\hat{Y}_R) = V(\hat{Y}_R) + \mu_N^2$$
4.4. Comparison of the Two Strategies

Let $S_1 = (\hat{Y}_R, \text{srsww})$, and $S_2 = (\hat{Y}_R, \text{Ms})$ where \text{srsww} and \text{Ms} denote simple random sampling without replacement and Midzuno's scheme respectively. Singh (1975) noted that in the absence of measurement errors, the two strategies are equally efficient if only terms up to $O(n^{-1})$ are considered. We consider the case when there are measurement errors in observing the $y$-characteristic.

**Theorem 4.4.1.**

Under the simple correlation model (4.1.2.), the difference in the biases of the two strategies $S_1$ and $S_2$ is

$$B(S_1) - B(S_2) = aY_R \left[ c_X^2 - c_{XY} \right]$$  (4.4.1.)

and the difference in the mean square errors is

$$\text{MSE}(S_1) - \text{MSE}(S_2) = \frac{(1 - \rho^2)}{n} \sum_{N} \left[ 2c_X^2 - c_{Xr} \right]$$

$$+ 2 \rho \sum_{N} \left[ c_X^2 - c_{Xr} \right]$$

$$+ 2 \sum_{N} \sum_{N} \left[ c_X^2 - c_{Xr} \right]$$  (4.4.2.)
Proof

The expression (4.4.1.) for the difference in the biases is obtained immediately by subtracting equation (4.3.2.) from equation (4.2.2.). The result (4.4.2.) is obtained by subtracting equation (4.3.4.) from (4.2.3.) after the expression (4.3.3.) has been substituted for \( \hat{V}(\hat{Y}_R) \) in equation (4.3.4.).

Remark 1. We note that the expressions (4.4.1.) and (4.4.2.) both reduce to zero when all \( Xs \) are equal, i.e., \( X_j = \bar{X}_N \forall j = 1, \ldots, N \). This was expected since in this case Midzuno's scheme is the same as simple random sampling without replacement.

Remark 2. In case measurement errors are absent, i.e., \( \rho = 0, \mu_j = 0, \sigma_j = 0 \forall j = 1, \ldots, N \), then from equation (4.4.1.) we have

\[
B(S_1) - B(S_2) = \frac{\bar{Y}}{N} [C_X^2 - C_{XY}]
\]

This was anticipated as, in this case, strategy \( S_2 \) is unbiased, and the bias of the strategy \( S_1 \) is given by

\[
B(S_1) = \frac{\bar{Y}}{N} [C_X^2 - C_{XY}]
\]

(see, e.g., Sukhatme and Sukhatme, 1970). We also note that the difference in mean square errors is zero. This was expected as the mean square errors of the two strategies are the same if only terms up to \( O(n^{-1}) \) are considered.

Remark 3. In case \( \rho = 0 \), i.e., the correlation between errors on two different units is zero, then we have
\[ B(S_1) - B(S_2) = \alpha \bar{Y}^t_N [c_X^2 - c_{XY}^t] \quad (4.4.3.) \]

and

\[ \text{MSE}(S_1) - \text{MSE}(S_2) = \frac{a}{n} \tau_N^2 \left[ 2c_X^2 - c_{XR} \right] \]
\[ + 2a\mu_2 \bar{Y}^t_N [c_X^2 - c_{XY}^t] \quad (4.4.4.) \]

Further, if also \( \sigma_j^2 = 0 \) \( \forall j = 1, ..., N \), i.e., there is only measurement bias, then

\[ \text{MSE}(S_1) - \text{MSE}(S_2) = 2a\mu_2 \bar{Y}^t_N [c_X^2 - c_{XY}^t] \quad (4.4.5.) \]

**Theorem 4.4.2.**

Under the intrasample correlation model (4.1.3.) the difference in
the biases of the two strategies \( S_1 \) and \( S_2 \) is

\[ B(S_1) - B(S_2) = \alpha \bar{Y}^t_N [c_X^2 - c_{XY}^t] \quad (4.4.6.) \]

and the difference in the mean square error is

\[ \text{MSE}(S_1) - \text{MSE}(S_2) = \frac{\tau_N^2}{n} \left[ 2a c_X^2 - a c_{XR} \right] \]
\[ + 2a\mu_2 \bar{Y}^t_N [c_X^2 - c_{XY}^t] \quad (4.4.7.) \]
Proof

The expression (4.4.6.) for the difference in the biases is obtained immediately by subtracting equation (4.3.8.) from equation (4.2.7.).
The result (4.4.7.) is obtained by subtracting equation (4.3.10.) from equation (4.2.8.) after the expression (4.3.9.) has been substituted for $V(\hat{Y}_R)$ in equation (4.3.10.).
5. EXTENSIONS TO TWO-STAGE AND STRATIFIED SAMPLING

5.1. Introduction

In most surveys, considerations of cost often dictate the use of multi-stage sampling. The primary sampling units (psu's) may vary considerably in size. We shall discuss the case when the psu's are selected with unequal probability and measurement errors are present.

Consider a population which is divided into \( N \) primary sampling units (psu's), the \( i \)-th psu having \( M_i \) second stage unit (ssu's). A sample of \( n \) psu's is selected. From the \( i \)-th selected psu, a sample of \( m_i \) ssu's is chosen and the procedure repeated for the selected psu's.

Consider the model

\[
y_{ij} = Y_{ij}^* + e_{ij}
\]

\[
Y_{ij}^* = Y_{ij} + \mu_{ij}
\]

where \( y_{ij} \) is the observed value of the \( (i, j) \)-th unit (\( i \) denotes the psu and \( j \) the ssu); \( Y_{ij} \) the true value of the \( (i, j) \)-th unit and \( \mu_{ij}, e_{ij} \) the bias and error, respectively, associated with the \( (i, j) \)-th unit.

For the given sample assume

\[
E(e_{ij}|i, j) = 0
\]

\[
V(e_{ij}|i, j) = \sigma^2
\]

\[
\text{Cov}(e_{ij}, e_{ij'}|i, j, j') = \rho \sigma^2 \quad j \neq j'
\]
Define

\[ Y_{ij} = y_{ij} \]

\[ M_i = \sum_{i} N \]

\[ v_i = \sum_{j} v_{ij} = \sum_{j} Y_{ij} \]

\[ v = \sum_{i} v_i \]

\[ \bar{v}_i = (M_i)^{-1} v_i \]

\[ s_{v_i}^2 = (M_i - 1)^{-1} \sum_{j} (v_{ij} - \bar{v}_i)^2 \]

\[ Y = \sum_{i} \sum_{j} Y_{ij} \]

\[ \mu = \sum_{i} \sum_{j} \mu_{ij} \]

\[ v = Y + \mu \]

\[ \bar{y}_i = (m_i)^{-1} \sum_{j} y_{ij} \]

\[ \bar{Y}_i = (m_i)^{-1} \sum_{j} Y_{ij} \]

Note that \( \bar{v}_i \) denotes the average over all ssu's in the \( i^{th} \) psu and \( \bar{Y}_i \) denotes the average over the sampled ssu's in the \( i^{th} \) psu.
5.2. Unequal Probability Sampling with Replacement

Consider the following sampling scheme. A sample of \( n \) psu's is selected with replacement, with \( P_i \) as the probability of selecting the \( i^{th} \) psu at any trial. From the \( i^{th} \) selected psu, a simple random sample of \( m_i \) ssu's is chosen. Information on the \( y \)-characteristic of the selected ssu's is obtained by an interviewer allocated at random to that psu. This procedure is repeated for the selected psu's.

For estimating the population total \( Y \) consider the estimator

\[
\hat{Y} = \frac{1}{n} \sum_{i} \frac{1}{P_i} \sum_{j} \frac{m_i}{m} y_{ij}
\]  

(5.2.1.)

Let \( E_3 \), \( V_3 \) denote the expectation and variance over repeated measurements when the sample is held fixed. Then

\[
E_3(\hat{Y}) = \frac{1}{n} \sum_{i} \frac{1}{P_i} \sum_{j} \frac{m_i}{m} Y_{ij}
\]

and

\[
V_3(\hat{Y}) = \frac{1}{2} \sum_{i} \frac{1}{P_i} \sum_{j} \frac{m_i}{m} V(y_{ij} | i, j) + \frac{1}{2} \sum_{i} \frac{1}{P_i} \sum_{j} \frac{m_i}{m} \text{Cov}(y_{ij}, y_{ij}, | i, j, j')
\]

\[
= \frac{1}{2} \sum_{i} \frac{1}{P_i} \sum_{j} \frac{1}{2} \frac{m_i}{m} \left[ m_i \sigma^2 + m_i (m_i - 1) \rho \sigma^2 \right]
\]

\[
= \frac{2}{n} \sum_{i} \frac{1}{P_i} \sum_{j} \frac{m_i^2}{m_i} \cdot \frac{1}{m_i} \left[ 1 + (m_i - 1) \rho \right]
\]
Let $E_2$, $V_2$ represent the expectation and variance over all selections of $m_1, m_2, \ldots, m_n$ second stage units from the psu's which are kept fixed; $E_1$, $V_1$ the expectation and variance over all possible samples of $n$ psu's from the $N$ in the population. Then

\[
E_2 E_3 (\hat{Y}) = \frac{1}{n} \sum_i \frac{V_i}{P_i}
\]

\[
V_1 E_2 E_3 (\hat{Y}) = \frac{1}{n} \sum_i \frac{N}{P_i} \left( \frac{V_i}{P_i} - \nu \right)^2
\]

\[
E_1 V_2 E_3 (\hat{Y}) = \frac{1}{n} \sum_i \frac{N}{m_i P_i} \left( \frac{1}{m_i} - \frac{1}{M_i} \right) S_{\nu i}^2
\]

\[
E_1 E_2 V_3 (\hat{Y}) = \frac{1}{n} \sum_i \frac{N}{m_i P_i} \left( \frac{1}{m_i} - \frac{1}{M_i} \right) S_{\nu i}^2
\]

\[
E_2 V_3 (\hat{Y}) = \sigma^2 \frac{n}{n} \sum_i \frac{M_i^2}{m_i P_i} \left[ 1 + (m_i - 1) \rho \right]
\]

\[
E_1 E_2 V_3 (\hat{Y}) = \sigma^2 \frac{n}{n} \sum_i \frac{M_i^2}{m_i P_i} \left[ 1 + (m_i - 1) \rho \right]
\]

\[
E_1 E_2 E_3 (\hat{Y}) = \sum_i V_i = \nu = Y + \mu
\]

Thus, the estimator in (5.2.1.) is subject to a bias of $\mu$, and its variance is

\[
V(\hat{Y}) = V_1 E_2 E_3 (\hat{Y}) + E_1 V_2 E_3 (\hat{Y}) + E_1 E_2 V_3 (\hat{Y})
\]
\[
\frac{1}{n} \sum_{i} \left( \frac{v_i}{F_i} - v \right)^2 + \frac{1}{n} \sum_{i} \frac{M_i^2}{F_i^2} \left( \frac{1}{m_i} - \frac{1}{M_i} \right) s_{v_i}^2 + \frac{\sigma^2}{n} \sum_{i} \frac{M_i^2}{m_i F_i} \left[ 1 + (m_i - 1) \rho \right]
\]

(5.2.2.)

The first term in this expression is the between-psu sampling variance, the second, the within-psu sampling variance, and the third, the response variance. Because of the presence of response errors, the variance increases by the quantity represented by the last term.

As an estimator of the variance, consider

\[
v = \frac{1}{n(n-1)} \sum_{i} \frac{M_i}{F_i} \left( \frac{y_i}{y} - \hat{y} \right)^2
\]

(5.2.3.)

Now

\[
E \sum_{i} \left( \frac{1}{F_i} \bar{y}_i - \bar{y} \right)^2 = E \left[ \sum_{i} \frac{M_i^2}{F_i^2} \frac{1}{m_i} \left( \sum_{j} y_{ij} \right)^2 + m_i \sigma^2 (1 + \frac{m_i}{m_i - 1} \rho) \right] - n E (\hat{y}^2)
\]

\[
= E \sum_{i} \frac{M_i^2}{F_i^2} \left[ \frac{1}{m_i} \left( \frac{1}{m_i} - \frac{1}{M_i} \right) s_{v_i}^2 \right] + E \sum_{i} \frac{M_i^2}{F_i^2} \sigma^2 \left[ 1 + (m_i - 1) \rho \right]
\]
Hence

\[ -n \, \text{var}(\hat{Y}) - n[\text{E}(\hat{Y})]^2 \]

\[ = n \sum_{i} \frac{M_i^2}{P_i} \left[ \frac{1}{S_i^2} + \left( \frac{1}{m_i} - \frac{1}{M_i} \right) S_i^2 \right] \]

\[ + n \sum_{i} \frac{M_i^2}{P_i} \cdot \frac{\sigma^2}{m_i} \left[ 1 + (m_i - 1)p \right] \]

\[ -n \, \text{var}(\hat{Y}) - n\nu^2 \]

\[ = n \sum \frac{1}{P_i} v_i^2 - n \, \nu^2 + n \sum \frac{M_i^2}{P_i} \left( \frac{1}{m_i} - \frac{1}{M_i} \right) S_i^2 \]

\[ + n\sigma^2 \sum \frac{M_i^2}{m_i P_i} \left[ 1 + (m_i - 1)p \right] - n \, \text{var}(\hat{Y}) \]

\[ = n^2 \, \text{var}(\hat{Y}) - n \, \text{var}(\hat{Y}) \]

\[ = n(n-1) \, \text{var}(\hat{Y}) \]

Hence

\[ \text{E}(\nu) = \text{var}(\hat{Y}) \]

which shows that the variance estimator is unbiased for the two-stage design considered here.

5.3. Unequal Probability Sampling without Replacement

Let us now consider a sampling scheme which is different from the scheme described in the previous section only in that the psu's are
selected without replacement. For estimating the population total $Y$
consider the estimator

$$\hat{Y} = \frac{n}{\sum_{i} \pi_i} \frac{1}{m_i} \sum_{j} \frac{m_i}{Y_{ij}}$$

(5.3.1.)

where $\pi_i$ is the inclusion probability for the $i$th unit. We have

$$E(\hat{Y}) = E_1 E_2 \frac{n}{\pi_i} \frac{1}{m_i} \sum_{j} Y_{ij}$$

$$= E_1 \frac{n}{\pi_i} \frac{1}{m_i} \sum_{j} \pi_j$$

$$= E_1 \frac{n}{\pi_i} \frac{1}{m_i} \pi_i = \nu$$

Thus, the bias in $\hat{Y}$ is $\nu - Y = \mu$. To obtain the variance, we have

$$E_2 E_3(\hat{Y}) = \frac{n}{\sum_{i} \pi_i} \frac{1}{m_i} \pi_i = \frac{n}{\sum_{i} \pi_i} \frac{1}{m_i}$$

$$V_1 E_2 E_3(\hat{Y}) = \frac{N}{\sum_{i} (1 - \pi_i)} \frac{1}{\pi_i} \nu_i^2 + \sum_{i \neq j} ^{N} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} \nu_i \nu_j$$

$$V_2 E_3(\hat{Y}) = \frac{n}{\sum_{i} \pi_i} \frac{1}{m_i} \left(\frac{1}{m_i} - \frac{1}{\pi_i}\right) s^2_{\nu_i}$$

$$E_1 V_2 E_3(\hat{Y}) = \frac{n}{\sum_{i} \pi_i} \frac{1}{m_i} \left(\frac{1}{m_i} - \frac{1}{\pi_i}\right) s^2_{\nu_i}$$

$$V_3(\hat{Y}) = \frac{n}{\sum_{i} \pi_i} \frac{1}{m_i} \frac{1}{2} \left[ m_i \sigma^2 + m_i (m_i - 1) \rho \sigma^2 \right]$$
\[
E_2V_3(\hat{Y}) = \sum_{i} \frac{\sigma_i^2}{\pi_i} \left[ 1 + \frac{(m_i - 1)p}{m_i} \right]
\]

\[
E_1E_2V_3(\hat{Y}) = \sum_{i} \frac{\sigma_i^2}{m_i} \left[ 1 + \frac{(m_i - 1)p}{m_i} \right]
\]

Hence,

\[
V(\hat{Y}) = \left[ \sum_{i} \frac{1}{\pi_i} v_i^2 + \sum_{i \neq j} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} v_i v_j \right]
\] 

\[+ \sum_{i} \frac{\sigma_i^2}{m_i} \left( \frac{1}{m_i} - \frac{1}{M_i} \right) s_i^2 \]

\[+ \sum_{i} \frac{\sigma_i^2}{m_i} \left[ 1 + \frac{(m_i - 1)p}{m_i} \right] \quad (5.3.2.)
\]

The first term in the expression (5.3.2.) is the between-psu sampling variance, the second term, the within-psu sampling variance, and the last term, the response variance.

The expression of between-psu sampling variance in (5.3.2.) can be written in an alternative form using \( \sum \pi_i = n \) and \( \sum \pi_{ik} = (n-1)\pi_i \) (e.g. Raj, 1968, p. 54) as

\[
\frac{1}{2} \sum_{i \neq k} \left( \pi_i \pi_k - \pi_{ik} \right) \left( \frac{v_i}{\pi_i} - \frac{v_k}{\pi_k} \right)^2
\]

Also, an unbiased estimator of

\[
S^2_{v_i} = (M_i - 1)^{-1} \sum_{j} (v_{ij} - \bar{v}_i)^2,
\]

when measurement errors are absent is
\[ s_{y_i}^2 = (m_i - 1)^{-1} \sum_{j}^{m_i} (y_{ij} - \bar{y}_i)^2 \]

So, for an estimate of \( V(\hat{Y}) \) consider

\[
v = \frac{1}{2} \sum_{i \neq k} \frac{n \pi_i \pi_k - \pi_{ik}}{\pi_{ik}} \left( \frac{\hat{v}_i}{\pi_i} - \frac{\hat{v}_k}{\pi_k} \right)^2 + \sum_{i} \frac{M_i^2}{n \pi_i} \left( \frac{1}{m_i} - \frac{1}{M_i} \right) s_{y_i}^2 \tag{5.3.3.}
\]

where \( \hat{v}_i = M_i \bar{y}_i \). To find the expectation of \( v \) we have

\[
E_3 (\frac{\hat{v}_i}{\pi_i} - \frac{\hat{v}_k}{\pi_k})^2 = V_3 (\frac{\hat{v}_i}{\pi_i} - \frac{\hat{v}_k}{\pi_k}) + [E_3 (\frac{\hat{v}_i}{\pi_i} - \frac{\hat{v}_k}{\pi_k})]^2
\]

and

\[
E_2^2 E_3 (\frac{\hat{v}_i}{\pi_i} - \frac{\hat{v}_k}{\pi_k})^2 = E_2 V_3 (\frac{\hat{v}_i}{\pi_i} - \frac{\hat{v}_k}{\pi_k}) + E_2 E_3 (\frac{\hat{v}_i}{\pi_i} - \frac{\hat{v}_k}{\pi_k})^2 + [E_2 E_3 (\frac{\hat{v}_i}{\pi_i} - \frac{\hat{v}_k}{\pi_k})]^2 \tag{5.3.4.}
\]

In order to evaluate each term of (5.3.4.), we have the following results.

\[
E_3 (\overline{Y}_i) = E_3 m_i^{-1} \sum_{j}^{m_i} y_{ij}
\]

\[
= m_i^{-1} \sum_{j}^{m_i} Y_{ij} = \overline{Y}_i
\]

\[
V_3 (\overline{Y}_i) = \frac{1}{m_i} V_3 (\sum_{j}^{m_i} y_{ij})
\]
The first term of (5.3.4) is obtained as follows

\[ V_2 E_3(\overline{y}_i) = E_2(\overline{y}_i) = \overline{v}_i \]

\[ V_2 E_3(\overline{y}_i) = V_2 \overline{v}_i = \left( \frac{1}{m_i} - \frac{1}{M_i} \right) S^2_{v_i} \]

The first term of (5.3.4) is obtained as follows

\[ V_3 \left( \frac{\hat{v}_i}{\hat{v}_k} - \frac{\hat{k}}{\hat{v}_k} \right) = V_3 \left( \frac{\hat{v}_i}{\hat{v}_k} \right) + V_3 \left( \frac{\hat{k}}{\hat{v}_k} \right) - 2 \text{Cov}_3 \left( \frac{\hat{v}_i}{\hat{v}_k} \right) \]

where \( \text{Cov}_3(\hat{v}_i, \hat{v}_k) \) is the covariance under repeated measurements holding the selected sample fixed.

Now

\[ V_3 \left( \frac{\hat{v}_i}{\hat{v}_k} \right) = V_3 \left( \frac{\hat{v}_i}{\hat{v}_i} \right) \]

\[ \frac{\hat{v}_i^2}{\hat{v}_i^2} = \frac{m_i^2}{\hat{v}_i^2} \]

\[ = \frac{1}{2} V_3(\overline{y}_i^2) \]

\[ = \frac{\hat{v}_i^2}{\hat{v}_i^2} \frac{\sigma^2}{\pi_i^2} \left[ 1 + (m_i - 1)\rho \right] \]

and

\[ \text{Cov}_3 \left( \frac{\hat{v}_i}{\hat{v}_i}, \frac{\hat{v}_k}{\hat{v}_k} \right) = 0 \]
Therefore

\[ V_3 \left( \frac{\hat{v}_i}{\eta_i} - \frac{\hat{v}_k}{\eta_k} \right) = \frac{M_i^2}{\eta_i^2} \sigma_i^2 \left[ 1 + (m_i - 1)\rho \right] \]

\[ + \frac{M_k^2}{\eta_k^2} \sigma_k^2 \left[ 1 + (m_k - 1)\rho \right] \]

Since this is a constant with respect to \( E_2 \), we have

\[ E_2 V_3 \left( \frac{\hat{v}_i}{\eta_i} - \frac{\hat{v}_k}{\eta_k} \right) = V_3 \left( \frac{\hat{v}_i}{\eta_i} - \frac{\hat{v}_k}{\eta_k} \right) \]

Next, we consider the second and third terms of (5.3.4.). We have

\[ E_3 \left( \frac{\hat{v}_i}{\eta_i} - \frac{\hat{v}_k}{\eta_k} \right) = \frac{M_i \bar{V}_i}{\eta_i} - \frac{M_k \bar{V}_k}{\eta_k} \]

\[ E_2 E_3 \left( \frac{\hat{v}_i}{\eta_i} - \frac{\hat{v}_k}{\eta_k} \right) = \frac{M_i \bar{V}_i}{\eta_i} - \frac{M_k \bar{V}_k}{\eta_k} \]

Thus, equation (5.3.4.) becomes

\[ E_2 E_3 \left( \frac{\hat{v}_i}{\eta_i} - \frac{\hat{v}_k}{\eta_k} \right)^2 = \frac{M_i^2}{\eta_i^2} \sigma_i^2 \left[ 1 + (m_i - 1)\rho \right] \]
Now, let us consider the second term of (5.3.3.). We have

\[ E_2E_3 s^2_{y_i} = E_2E_3 (m_i - 1)^{-1} \sum_j \left( y_{ij} - \bar{y}_i \right)^2 \]

\[ = E_2E_3 (m_i - 1)^{-1} \left[ \sum_j y_{ij}^2 - m_i \bar{y}_i^2 \right] \]

The expectation of the component terms is obtained as

\[ E_3 \sum_j y_{ij}^2 = \sum_j \left[ y_{ij}^2 + \sigma^2 \right] \]

\[ = \sum_j y_{ij}^2 + m_i \sigma^2 \]

\[ E_2E_3 \sum_j y_{ij}^2 = \frac{m_i}{M_i} \sum_j y_{ij}^2 + m_i \sigma^2 \]

\[ E_2E_3 \left( \bar{y}_i^2 \right) = E_2 \left[ V_3(\bar{y}_i) + \{E_3(\bar{y}_i)\}^2 \right] \]
Hence

\[ \frac{\sigma^2}{m_i} \left[ 1 + (m_i - 1)\rho \right] + \left( \frac{1}{m_i} - \frac{1}{M_i} \right) \mu_i^2 + \nu_i^2 \]

Thus

\[ \sum_{j=1}^{m_i} E_2 \left( y_{ij} - \bar{y}_i \right)^2 = E_2 \left[ \sum_{j=1}^{m_i} y_{ij}^2 - m_i \bar{y}_i^2 \right] \]

\[ = \frac{m_i}{M_i} \sum_{j=1}^{m_i} \left( \frac{1}{M_i} - \frac{1}{M_i} \right) \mu_i^2 + \mu_i \nu_i^2 \]

\[ - \sigma^2 \left[ 1 + (m_i - 1)\rho \right] - m_i \left( \frac{1}{m_i} - \frac{1}{M_i} \right) \sigma_i^2 \]

\[ = \frac{m_i}{M_i} (M_i - 1) \mu_i^2 + (m_i - 1) \sigma_i^2 \]

\[ - \sigma^2 (m_i - 1) \rho - m_i \left( \frac{1}{m_i} - \frac{1}{M_i} \right) \sigma_i^2 \]

\[ = (m_i - 1) \mu_i^2 + (m_i - 1) \sigma_i^2 \]

\[ - \sigma^2 (m_i - 1) \rho \]

\[ = (m_i - 1) \left[ \sigma_i^2 + \sigma^2 (1 - \rho) \right] \]
\[ E^2 \bar{F}^3 \sum_i \frac{1}{\pi_i} \left( \frac{1}{m_i} - \frac{1}{M_i} \right) s^2_{\gamma_i} = \sum_i \frac{1}{\pi_i} \left( \frac{1}{m_i} - \frac{1}{M_i} \right) \left[ s^2_{\gamma_i} + \sigma^2 (1 - \rho) \right] \]

Hence, the expectation of the second term of (5.3.3.) is
\[ E \sum_i \frac{1}{\pi_i} \left( \frac{1}{m_i} - \frac{1}{M_i} \right) s^2_{\gamma_i} = \sum_i \frac{1}{\pi_i} \left( \frac{1}{m_i} - \frac{1}{M_i} \right) \left[ s^2_{\gamma_i} + \sigma^2 (1 - \rho) \right] \quad (5.3.5.) \]

Also, the expectation of the first term of (5.3.3.) can be written as
\[ E \frac{1}{2} \sum_{i \neq k} \frac{\pi_i \pi_k - \pi_{ik}}{\pi_{ik}} \left( \frac{\gamma_i}{\pi_i} - \frac{\gamma_k}{\pi_k} \right)^2 \]
\[ = \frac{1}{2} \sum_{i \neq k} \left( \frac{\pi_i \pi_k - \pi_{ik}}{\pi_{ik}} \right) \left[ \frac{\sigma^2}{\pi_i} \left\{ 1 + (m_i - 1) \rho \right\} \right. \]
\[ + \frac{\sigma^2}{\pi_k} \left\{ 1 + (m_k - 1) \rho \right\} \]
\[ + \frac{\sigma^2}{\pi_i} \left( \frac{1}{m_i} - \frac{1}{M_i} \right) s^2_{\gamma_i} \]
\[ + \frac{\sigma^2}{\pi_k} \left( \frac{1}{m_k} - \frac{1}{M_k} \right) s^2_{\gamma_k} \]
\[ + \left. \left( \frac{\gamma_i}{\pi_i} - \frac{\gamma_k}{\pi_k} \right)^2 \right] \]

Using
\[ \sum_{i \neq k} \left( \frac{\pi_i \pi_k - \pi_{ik}}{\pi_i} \right) \frac{\gamma_i^2}{\pi_i} = \sum_{i \neq k} \frac{\gamma_i^2}{\pi_i} k(\bar{\gamma}_k) \]
\[ \sum_{i \neq k} \left( \frac{\pi_i \pi_k - \pi_{ik}}{\pi_i} \right) \frac{\gamma_i^2}{\pi_i} = \sum_{i \neq k} \frac{\gamma_i^2}{\pi_i} k(\bar{\gamma}_k) \]
\[ \psi_i^2 = \sum_{i} \frac{1}{\pi_i} (n - \pi_i) - \sum_{i} \frac{1}{\pi_i} (n-1)\pi_i \]

\[ = \sum_{i} (1 - \pi_i) \frac{1}{\pi_i} \]

we have

\[ \frac{1}{2} \sum_{i \neq k} \frac{N}{\pi_i \pi_k} (\pi_i \pi_k - \pi_{ik}) \left( \frac{1}{\pi_i} - \frac{1}{\pi_k} \right)^2 \]

\[ = \sum_{i} \frac{1 - \pi_i}{\pi_i} \psi_i^2 - \sum_{i \neq k} \frac{N}{\pi_i \pi_k} (\pi_i \pi_k - \pi_{ik}) \frac{\psi_i \psi_k}{\pi_i \pi_k} \]

\[ = \sum_{i} \frac{1 - \pi_i}{\pi_i} \psi_i^2 + \sum_{i \neq k} \frac{\pi_{ik} - \pi_i \pi_k}{\pi_i \pi_k} \psi_i \psi_k \]

and

\[ \frac{1}{2} \sum_{i \neq k} \frac{N}{\pi_i \pi_k} (\pi_i \pi_k - \pi_{ik}) \left[ \frac{M_k^2}{\pi_i} \sigma_i^2 \{ 1 + (m_i - 1)\rho \} \right. \]

\[ + \frac{M_k^2}{\pi_k} \frac{\sigma_i^2}{m_k} \{ 1 + (m_k - 1)\rho \} \]

\[ + \frac{M_i^2}{\pi_i} \left( \frac{1}{m_i} - \frac{1}{M_i} \right) \psi_i^2 + \frac{M_k^2}{\pi_k} \left( \frac{1}{m_k} - \frac{1}{M_k} \right) \psi_k^2 \]

\[ = \sum_{i} \frac{1 - \pi_i}{\pi_i} \cdot \frac{M_i^2}{m_i} \sigma_i^2 \{ 1 + (m_i - 1)\rho \} \]

\[ + \sum_{i} \frac{1 - \pi_i}{\pi_i} M_i \left( \frac{1}{m_i} - \frac{1}{M_i} \right) \frac{\sigma_i^2}{\psi_i} \]
Hence, the expectation of the first term of (5.3.3.) becomes

\[
E \frac{1}{2} \sum_{i \neq k}^{n} \frac{\pi_i \pi_k - \pi_{ik}}{\pi_{ik}} \left( \hat{v}_i - \hat{v}_k \right)^2
\]

\[
= \sum_{i}^{N} \frac{1 - \pi_i}{\pi_i} \cdot \frac{M_i^2 \sigma_i^2}{m_i} [1 + (m_i - 1) \rho]
\]

\[
+ \sum_{i}^{N} \frac{1 - \pi_i}{\pi_i} M_i^2 \left( \frac{1}{m_i} - \frac{1}{M_i} \right) S_{\hat{v}_i}^2
\]

\[
+ \sum_{i}^{N} \frac{1 - \pi_i}{\pi_i} \nu_i^2 + \sum_{k \neq i}^{N} \frac{\pi_{ik} - \pi_i \pi_k}{\pi_i \pi_k} \nu_i \nu_k
\]

\[(5.3.6.)\]

Substituting (5.3.6.) and (5.3.7.) in the expression (5.3.3.) for \( E(v) \), we finally obtain

\[
E(v) = \sum_{i}^{N} \frac{1 - \pi_i}{\pi_i} \nu_i^2 + \sum_{k \neq i}^{N} \frac{\pi_{ik} - \pi_i \pi_k}{\pi_i \pi_k} \nu_i \nu_k
\]

\[
+ \sum_{i}^{N} \frac{M_i^2}{\pi_i} \left( \frac{1}{m_i} - \frac{1}{M_i} \right) S_{\hat{v}_i}^2 - \sum_{i}^{N} \frac{M_i^2 \sigma_i^2}{m_i} [1 + (m_i - 1) \rho]
\]

\[
+ \sum_{i}^{N} \frac{M_i^2}{m_i} \left( \frac{1}{m_i} - \frac{1}{M_i} \right) S_{\hat{v}_i}^2 + \sum_{i}^{N} \frac{M_i^2 \sigma_i^2}{m_i} [1 + (m_i - 1) \rho]
\]

\[
+ \sum_{i}^{N} \frac{M_i^2}{m_i} \left( \frac{1}{m_i} - \frac{1}{M_i} \right) S_{\hat{v}_i}^2 + \sum_{i}^{N} \frac{M_i^2 \sigma_i^2}{m_i} (1 - \rho)
\]

\[
= \sum_{i}^{N} \frac{1 - \pi_i}{\pi_i} \nu_i^2 + \sum_{k \neq i}^{N} \frac{\pi_{ik} - \pi_i \pi_k}{\pi_i \pi_k} \nu_i \nu_k
\]
Thus, the estimator \( v \) is biased for estimating \( V(\hat{Y}) \), and the bias is given by the last term of (5.3.7.). Since an interviewer obtains information from the units within a PSU, \( \rho \) is usually positive. Then the variance estimator is biased downward, i.e. it has a negative bias. In case \( M_i = 1 \), the sampling scheme considered reduces to one-stage unequal probability sampling without replacement, and the bias in estimating the variance is zero.

5.4. Stratified Sampling

Let a population of \( N \) units be divided into \( L \) non-overlapping strata (or subpopulations) of sizes \( N_1, N_2, \ldots, N_L \) such that \( \sum_{h=1}^{L} N_h = N \). We assume that an interviewer enumerates the units in a stratum alone, which is usually the case because of travel costs. So for an estimator

\[
\hat{\bar{y}}_{\text{st}} = \sum_{h=1}^{L} W_h \hat{\bar{y}}_h
\]

(5.4.1.)

where \( \hat{\bar{y}}_h \) is an estimate of the population mean of the \( y \)-values of the units in the \( h^{\text{th}} \) stratum and \( W_h \ (h = 1, 2, \ldots, L) \) are constants, we have
Consider the model

\[ y_{hi} = Y_{hi}^* + e_{hi} \]

where \( y_{hi} \) is the observed value of the \((h, i)\)th unit (\( h \) denotes the stratum and \( i \) the unit within the stratum); \( Y_{hi}^* \) the true value of the \((h, i)\)th unit; and \( \mu_{hi}, e_{hi} \) the bias and error respectively associated with the \((h, i)\)th unit. For samples of \( n_h \) and \( n_\ell \) units from the \( h \)th and \( \ell \)th strata respectively, we assume

\[ \mathbb{E}(e_{hi} | h, i) = 0 \]

\[ \mathbb{V}(e_{hi} | h, i) = \sigma_h^2 \]

\[ \text{Cov}(e_{hi}, e_{hj} | h, i, j) = \rho_h \sigma_h^2 \quad i \neq j \]

\[ \text{Cov}(e_{hi}, e_{\ell j} | h, \ell, i, j) = 0 \quad h \neq \ell \] (5.4.5.)

A simple random sample of \( n_h \) units is selected without replacement from the \( h \)th stratum. Let
\[ \bar{y}_h = n_h^{-1} \sum_{i} n_{hi} \]

\[ \bar{v}_h = N_h^{-1} \sum_{i} y_{hi} \]

\[ S_{y_h}^2 = (N_h - 1)^{-1} \sum_{i} (Y_{hi} - \bar{y}_h)^2 \]

For the estimator

\[ \bar{y}_{st} = \sum_{h} w_h \bar{y}_h \]

we have

\[ V(\bar{y}_{st}) = \sum_{h} w_h^2 V(\bar{y}_h) \]

using (5.4.3.). Under model (5.4.4.) along with the assumptions (5.4.5.), we have, on using (3.4.8.) that

\[ V = V(\bar{y}_{st}) = \sum_{h} w_h^2 \left( \frac{1}{n_h^2} - \frac{1}{N_h^2} \right) S_{y_h}^2 \]

\[ + \frac{c_h^2}{n_h} \left[ 1 + (n_h - 1) \rho_h \right] \]  \hspace{1cm} (5.4.6.)

We now consider the problem of allocation of sample size to strata.

The criterion of determining the vector \((n_1, n_2, \ldots, n_L)\) is either to minimize \(V(\bar{y}_{st})\) for a fixed cost or to minimize cost for a fixed variance.

Let \(c_h\) be the cost of collecting information from a unit in stratum \(h\), and \(c_0\) the overhead cost. Then the total cost of the survey is

\[ C = c_0 + \sum_{h} c_h n_h \]  \hspace{1cm} (5.4.7.)
Consider

\[ V'C' = (V + \sum_{h} \frac{w_h^2 S_{y_h}^2}{n_h} - \sum_{h} \frac{w_h^2 \sigma^2_h}{n_h} \rho_h) \cdot (C - c_0) \]

\[ = \left[ \sum_{h} \frac{w_h^2}{n_h} \left( S_{y_h}^2 + \sigma^2_h - \rho_h \sigma^2_h \right) \right] \cdot \left[ \sum_{h} c_h n_h \right] \]

Using Cauchy-Schwarz inequality, we infer that \( V'C' \) attains its minimum value if and only if

\[ \frac{w_h^2 \left( S_{y_h}^2 + \sigma^2_h - \rho_h \sigma^2_h \right) / n_h}{c_h n_h} = \text{constant} \]

for all \( h \). Hence

\[ \frac{n_h}{n} = \frac{w_h \sqrt{S_{y_h}^2 + \sigma^2_h (1 - \rho_h)}/c_h}{\sum_{h} w_h \sqrt{S_{y_h}^2 + \sigma^2_h (1 - \rho_h)}/c_h} \quad (5.4.8.) \]

We now have the allocation of sample size to strata. Suppose
the sample is chosen to minimize \( V(y_{st}) \) for specified cost, then on
substituting the optimum values of \( n_h \) from (5.4.8.) in the cost function
(5.4.7.), we have

\[ n = \frac{(C - c_0) \sum_{h} w_h \sqrt{S_{y_h}^2 + \sigma^2_h (1 - \rho_h)}/c_h}{\sum_{h} w_h \sqrt{c_h \left( S_{y_h}^2 + \sigma^2_h (1 - \rho_h) \right)}} \]

If \( V \) is fixed, then \( n \) can be found by substituting the optimum values of
\( n_h \) in the equation (5.4.6.).
If \( c_h = c \) for \( h = 1, 2, ..., L \), then the cost is

\[
C = c_0 + cn
\]

and optimum allocation for fixed cost reduces to optimum allocation for fixed \( n \). Then

\[
n_h = \frac{n W_h \sqrt{S_{Y_h}^2 + \sigma_h^2 (1 - \rho_h)}}{\sum W_h \sqrt{S_{Y_h}^2 + \sigma_h^2 (1 - \rho_h)}}
\]

This allocation will be called the modified Neyman allocation. The minimum value of \( V(\bar{y}_{st}) \) when \( n \) is fixed is

\[
\min V(\bar{y}_{st}) = \frac{1}{n} \cdot \left( \sum W_h \sqrt{S_{Y_h}^2 + \sigma_h^2 (1 - \rho_h)} \right)^2
\]

\[
- \sum \frac{W_h^2 S_{Y_h}^2}{N_h} + \sum W_h^2 \sigma_h^2 \rho_h
\]

5.5. Systematic Sampling

Systematic sampling can be viewed as a particular case of cluster sampling. In systematic sampling, a cluster (or psu) of size \( n \) is selected from the \( k \) possible clusters (\( nk = N \)). It follows from Section 5.2 that for systematic sampling an estimator of the population total \( Y \) is

\[
\hat{Y} = k \sum_{j} y_{ij}
\]
where $y_{ij}$ is the observed value of the $y$-characteristic of the $j^{th}$ unit in the $i^{th}$ cluster. Under the model given by (5.1.1.) and (5.1.2.), the bias of the estimator $\hat{Y}$ is

$$B(\hat{Y}) = \sum_{i} \sum_{j} \mu_{ij} = \mu$$

and its variance is

$$V(\hat{Y}) = \frac{1}{k} \sum_{i} \left( \sum_{j} y'_{ij} - Y' \right)^2 + \sigma^2 Nk[1 + (n-1)\rho]$$

where

$$Y' = \sum_{i} \sum_{j} y'_{ij}$$
6. EMPIRICAL STUDY

6.1. Comparison of ppswr Sampling with Simple Random Sampling

Various statisticians, including Cochran (1977), have called for empirical work in the area of response errors. It is in this spirit that one population has been selected and the effect of measurement errors studied under the models considered in this thesis.

The population considered is taken from Kish (1965, Appendix E). The population relates to the 270 blocks in Ward I of Fall River, Massachusetts, and is taken from the column of Block Statistics of the 1950 U.S. Census. The total number of dwellings ($X_i$) and the number of dwellings occupied by renters ($Y_i$) are known for each block. The purpose is to estimate, from the sample, the total number of rented dwellings, or the average number per block. We will assume that the number of dwellings occupied by renters in the $i^{th}$ block, i.e., $Y_i$ as given in Kish (1965, Appendix E), is the true value of $y$. We note that the correlation between $X$ and $Y$ is 0.96, which is typical of the populations considered for the study of the ratio estimator (Royall and Cumberland, 1981).

To study different strategies under measurement errors, response errors will be introduced in the data in the following directions:

1. the bias associated with unit $U_j$, i.e. $u_j$, will be assumed to be at levels $A_1 Y_j$ with $A_1 = \pm 0.05, \pm 0.01, 0.00$;
2. the within-trial variance $\sigma^2$ will be taken as $A_2 \sigma_Y^2$ with $A_2 = 0.00, 0.05, 0.10, 0.3, 1.0$;
3. the correlation coefficient $p$ will be taken as $p = 0.00, 0.01, 0.05$.

The sample sizes to be considered are 30, 45 and 60.

The values of $A_1$, $A_2$ and $p$ are chosen in view of the studies undertaken by Gray (1955) and Kish (1962).

Initially, a comparison will be made among the current estimators in the three situations - sampling with replacement with probability proportional to $X$ (ppswr), simple random sampling with replacement (ssrwr), and simple random sampling without replacement (srswor).

Later on, the method of ratio estimation and the Midzuno scheme will also be considered.

Let

\[ S_1 = \left( \bar{Y}_N, \text{ppswr} \right) \]
\[ S_2 = \left( \bar{y}_n, \text{srswr} \right) \]
\[ S_3 = \left( \bar{y}_n, \text{srswor} \right) \]

where

\[ \bar{Y}_N = (Nn)^{-1} \sum_{j=1}^{n} y_j \frac{P_j}{p_j} \]
\[ \bar{y}_n = n^{-1} \sum_{j=1}^{n} y_j \]
\[ P_j = X_j / X \]

On substituting $\mu_j = A_1 y_j$ and $\sigma_j^2 = A_2 \sigma_Y^2$ for $j = 1, 2, ..., N$, in equations (3.2.5.), (3.2.4.), (3.4.3.), (3.4.2.), (3.4.8.) and (3.4.7.), we have
\[
MSE(S_1) = \frac{(1 + A_1)^2}{n} \left[ \frac{1}{N} \sum_{j} \left( \frac{x_j}{X_j} \right)^2 - \frac{\gamma^2}{N} \right]
+ \frac{(A_2 \sigma_Y^2)}{nN} \left[ \frac{1}{N} \sum_{j} \left( \frac{1}{X_j} \right) + (n-1)N \right]
+ \frac{(n-1)(N-1)}{nN} \cdot \rho \cdot (A_2 \sigma_Y^2) + (\Delta \overline{Y}_N)^2
\]

\[
MSE(S_2) = (1 + A_1)^2 \frac{(N-1)}{nN} \sigma_Y^2 + \frac{(N+n-1)}{nN} (A_2 \sigma_Y^2)
+ \frac{(n-1)(N-1)}{nN} \cdot \rho \cdot (A_2 \sigma_Y^2) + (\Delta \overline{Y}_N)^2
\]

\[
MSE(S_3) = (1 + A_1)^2 \frac{N-n}{Nn} \sigma_Y^2 + \frac{1}{n} \cdot (A_2 \sigma_Y^2)
+ \frac{(n-1)}{n} \cdot \rho \cdot (A_2 \sigma_Y^2) + (\Delta \overline{Y}_N)^2
\]

In the absence of errors, the variances of the three strategies are as given in Table 6.1.

<table>
<thead>
<tr>
<th>Sample size</th>
<th>pps</th>
<th>srswr</th>
<th>srswor</th>
</tr>
</thead>
<tbody>
<tr>
<td>30</td>
<td>0.829</td>
<td>14.269</td>
<td>12.730</td>
</tr>
<tr>
<td>45</td>
<td>0.553</td>
<td>9.512</td>
<td>7.956</td>
</tr>
<tr>
<td>60</td>
<td>0.415</td>
<td>7.134</td>
<td>5.569</td>
</tr>
</tbody>
</table>
We observe that the variance of $S_1$ is considerably smaller than the variance of $S_2$ and $S_3$. This is to be expected, as the $Y_i$'s are highly correlated with the $X_i$'s. We also note that doubling the sample size halves the variance. Since the three strategies are unbiased, the mean square error (MSE) is the same as the variance.

We now retain $A_2 = 0$, $\rho = 0$ but introduce bias in reporting of rented dwellings in the block. Table 6.2 presents the sampling variance and MSE of the three strategies. The response variance of the three strategies is zero. We observe that the sampling variance decreases when the bias is negative and increases when the bias is positive (as compared with the situation in which there is no bias). Since the sampling variance is low in the case of pps sampling, the percentage increase in MSE is much higher in this case, as compared with sampling with equal probabilities. We also observe that the MSE for the measurement error case may be smaller than the MSE for the no-measurement-error case. This would happen when the measurement bias is large and negative and thus the decrease in sampling variance is enough to make the MSE for the measurement error case smaller than the MSE for the no-measurement-error case.

Let us consider the case, when both $A_1$ and $A_2$ are not zero. Table 6.3 gives the sampling variance and MSE of the three strategies for different sample sizes, $A_1$, $A_2 = 0.3$ and $\rho = 0$. In this case the response variances of $S_1$, $S_2$ and $S_3$ are 19.379, 4.758 and 4.296 respectively. By examining Table 6.3, we find that strategy $S_3$ is more efficient than $S_1$ or $S_2$. We also studied the case $A_2 = 0.05$ and
Table 6.2. Variance and MSE for different sample sizes, $A_1, A_2 = 0$ and $\beta = 0$

<table>
<thead>
<tr>
<th>Sample Size</th>
<th>$A_1$</th>
<th>Sampling variance</th>
<th>MSE</th>
<th>ppswr</th>
</tr>
</thead>
<tbody>
<tr>
<td>30</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.748</td>
<td>1.461</td>
<td>(-9.77)$^a$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.813)</td>
<td>(0.841)</td>
<td>(-1.93)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.846</td>
<td>0.875</td>
<td>(2.05)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.914</td>
<td>1.627</td>
<td>(10.25)</td>
</tr>
<tr>
<td>60</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.374</td>
<td>1.087</td>
<td>(-9.88)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.406)</td>
<td>(0.435)</td>
<td>(-2.17)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.423</td>
<td>0.452</td>
<td>(1.93)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.457</td>
<td>1.170</td>
<td>(10.12)</td>
</tr>
</tbody>
</table>

$^a$The figures in parentheses denote the percentage increase over the case when measurement errors are absent.
<table>
<thead>
<tr>
<th>srswr</th>
<th>srswor</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.05</td>
<td>-0.05</td>
</tr>
<tr>
<td>-0.01</td>
<td>-0.01</td>
</tr>
<tr>
<td>+0.01</td>
<td>+0.01</td>
</tr>
<tr>
<td>+0.05</td>
<td>+0.05</td>
</tr>
</tbody>
</table>

<p>| | | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>(-9.76)</td>
<td>(-1.99)</td>
<td>(2.00)</td>
<td>(10.25)</td>
<td>(-9.75)</td>
<td>(-1.99)</td>
<td>(2.01)</td>
<td>(10.25)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(-4.76)</td>
<td>(-1.79)</td>
<td>(2.21)</td>
<td>(15.24)</td>
<td>(-4.16)</td>
<td>(-1.76)</td>
<td>(2.23)</td>
<td>(15.85)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(-9.74)</td>
<td>(-1.99)</td>
<td>(2.02)</td>
<td>(10.26)</td>
<td>(-9.75)</td>
<td>(-1.98)</td>
<td>(2.01)</td>
<td>(10.25)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7.151</td>
<td>7.021</td>
<td>7.306</td>
<td>8.578</td>
<td>5.739</td>
<td>5.487</td>
<td>5.710</td>
<td>6.853</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(0.24)</td>
<td>(1.58)</td>
<td>(2.41)</td>
<td>(20.24)</td>
<td>(3.05)</td>
<td>(1.47)</td>
<td>(2.53)</td>
<td>(23.06)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 6.3. Variance and MSE for different sample sizes, $A_1, A_2 = 0.3$ and $\rho = 0$

<table>
<thead>
<tr>
<th>Sample Size</th>
<th>$A_1$</th>
<th>ppswr</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-0.05</td>
<td>-0.01</td>
</tr>
<tr>
<td>30 Sampling variance</td>
<td>0.748</td>
<td>0.813</td>
</tr>
<tr>
<td>60 Sampling variance</td>
<td>0.374</td>
<td>0.406</td>
</tr>
<tr>
<td>MSE</td>
<td>17.352</td>
<td>16.700</td>
</tr>
<tr>
<td></td>
<td>srswr</td>
<td></td>
</tr>
<tr>
<td>------</td>
<td>-------</td>
<td>----------</td>
</tr>
<tr>
<td></td>
<td>-0.05</td>
<td>-0.01</td>
</tr>
<tr>
<td></td>
<td>6.439</td>
<td>6.992</td>
</tr>
</tbody>
</table>
0.01. The tables, not shown here, indicate that with $A_2$ this small, the within-trial variance is not large enough to make the response variance of $S_1$ large. Hence, the behavior of MSE is similar to the case $A_2 = 0$, and pps sampling is better than equal probability sampling.

We next consider the case when $\rho$ is not zero. Table 6.4 gives the sampling variance and MSE of the three strategies for different sample sizes, $A_1, A_2 = 0.3$ and $\rho = 0.01$. The response variances of $S_1$, $S_2$ and $S_3$ increase to 21.625, 6.004 and 5.542 respectively. After examining Table 6.4, we conclude that $S_3$ is more efficient than $S_1$ or $S_2$, precisely what we inferred from Table 6.3. This is to be expected since the correlated response variance of the three strategies is the same.

The results of our study indicate that if measurement errors are absent then $S_1$ is more efficient than $S_2$ or $S_3$. But, if measurement errors are present, then $S_3$ may be more efficient than $S_1$ or $S_2$. Also, we observed that the larger the within trial variance, the better the strategies $S_2$ and $S_3$ perform in relation to $S_1$.

6.2. Ratio Estimation

We now consider the two strategies $S_4$ and $S_5$, where $S_4$ denotes the ratio estimator with simple random sampling without replacement and $S_5$, the ratio estimator with Midzuno's scheme. It is well-known that $S_4$ is biased and $S_5$ is unbiased. Table 6.5 gives the bias of $S_4$ for the population given in Section 6.1.
Table 6.4. Variance and MSE for different sample sizes, $A_1, A_2 = 0.3$ and $\rho = 0.01$

<table>
<thead>
<tr>
<th>Sampling Size</th>
<th>$A_1$</th>
<th>-0.05</th>
<th>-0.01</th>
<th>+0.01</th>
<th>+0.05</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Sampling variance</td>
<td>0.748</td>
<td>0.813</td>
<td>0.846</td>
<td>0.914</td>
</tr>
<tr>
<td>30</td>
<td>MSE</td>
<td>22.086</td>
<td>21.466</td>
<td>21.499</td>
<td>22.252</td>
</tr>
<tr>
<td>60</td>
<td>Sampling variance</td>
<td>0.374</td>
<td>0.406</td>
<td>0.423</td>
<td>0.457</td>
</tr>
<tr>
<td></td>
<td>MSE</td>
<td>12.282</td>
<td>11.630</td>
<td>11.647</td>
<td>12.365</td>
</tr>
<tr>
<td></td>
<td>SWSW</td>
<td></td>
<td>SWSW</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-------</td>
<td>------</td>
<td>-------</td>
<td>------</td>
<td>-------</td>
<td></td>
</tr>
<tr>
<td>-0.05</td>
<td>12.877</td>
<td>-0.05</td>
<td>11.489</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-0.01</td>
<td>13.985</td>
<td>-0.01</td>
<td>12.477</td>
<td></td>
<td></td>
</tr>
<tr>
<td>+0.01</td>
<td>14.555</td>
<td>12.986</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>+0.05</td>
<td>15.731</td>
<td>14.035</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>19.594</td>
<td></td>
<td>17.744</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>20.017</td>
<td></td>
<td>18.048</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>20.588</td>
<td></td>
<td>18.577</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>22.448</td>
<td></td>
<td>20.290</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>6.439</td>
<td></td>
<td>5.026</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>6.992</td>
<td></td>
<td>5.459</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>7.278</td>
<td></td>
<td>5.681</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>7.866</td>
<td></td>
<td>6.140</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>11.037</td>
<td></td>
<td>9.155</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>10.906</td>
<td></td>
<td>8.903</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>11.191</td>
<td></td>
<td>9.126</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>12.463</td>
<td></td>
<td>10.269</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 6.5. Bias of $S_4$

<table>
<thead>
<tr>
<th>$n$</th>
<th>30</th>
<th>45</th>
<th>60</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bias</td>
<td>-0.0975</td>
<td>-0.0609</td>
<td>-0.0426</td>
</tr>
</tbody>
</table>

Singh (1975) stated that the two strategies are equally efficient if terms to $O(n^{-1})$ are considered. We study the difference in mean square errors of the two strategies when measurement errors are present.

Table 6.6. Difference in mean square errors of strategies $S_4$ and $S_5$ for $A_2 = 0$ and $\rho = 0$

<table>
<thead>
<tr>
<th>Sample Size</th>
<th>$A_1$</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$n$</td>
<td>-0.05</td>
<td>-0.01</td>
<td>+0.01</td>
<td>+0.05</td>
</tr>
<tr>
<td>30</td>
<td>0.1564</td>
<td>0.0326</td>
<td>-0.0332</td>
<td>-0.1728</td>
</tr>
<tr>
<td>60</td>
<td>0.0684</td>
<td>0.0143</td>
<td>-0.0145</td>
<td>-0.0756</td>
</tr>
</tbody>
</table>

The data used are the same as in the last section. Table 6.6 gives the difference in mean square errors of strategies $S_4$ and $S_5$ when only measurement bias is present. In this case, the response variance of the two strategies is zero, and the sampling variance is the same. Hence, the comparison is between the square of the biases. We conclude that for negative measurement bias, i.e. $A_1$ negative, $S_4$ is less efficient than $S_5$ and, for positive measurement bias, $S_4$ is more efficient than $S_5$. 
Table 6.7. Difference in mean square errors of \( S_4 \) and \( S_5 \) for \( A_1 = 0, A_2 \) and \( \rho = 0 \\

<table>
<thead>
<tr>
<th>Sample Size</th>
<th>( n )</th>
<th>( A_2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.05</td>
<td>0.1</td>
</tr>
<tr>
<td>30</td>
<td>0.0420</td>
<td>0.0840</td>
</tr>
<tr>
<td>60</td>
<td>0.0092</td>
<td>0.0184</td>
</tr>
</tbody>
</table>

Table 6.7 provides the difference in mean square errors for different values of \( A_2, A_1 = 0 \) and \( \rho = 0 \). From this table, we conclude that \( S_4 \) is less efficient than \( S_5 \) as the difference is positive in each case.

Table 6.8. Difference in MSEs. of strategies \( S_4 \) and \( S_5 \) for \( n = 30, \rho = 0 \\

<table>
<thead>
<tr>
<th>( A_1 )</th>
<th>0.00</th>
<th>0.05</th>
<th>0.10</th>
<th>0.30</th>
<th>1.00</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.05</td>
<td>0.1564</td>
<td>0.1983</td>
<td>0.2403</td>
<td>0.4083</td>
<td>0.9961</td>
</tr>
<tr>
<td>-0.01</td>
<td>0.0326</td>
<td>0.0746</td>
<td>0.1166</td>
<td>0.2845</td>
<td>0.8724</td>
</tr>
<tr>
<td>0.00</td>
<td>0.0000</td>
<td>0.0420</td>
<td>0.0840</td>
<td>0.2519</td>
<td>0.8398</td>
</tr>
<tr>
<td>+0.01</td>
<td>-0.0332</td>
<td>0.0087</td>
<td>0.0507</td>
<td>0.2187</td>
<td>0.8065</td>
</tr>
<tr>
<td>+0.05</td>
<td>-0.1728</td>
<td>-0.1308</td>
<td>-0.0888</td>
<td>0.0791</td>
<td>0.6670</td>
</tr>
</tbody>
</table>
Table 6.9. Difference in MSEs of strategies $S_4$ and $S_5$ for $n = 30$, $\rho = 0.01$

<table>
<thead>
<tr>
<th>$A_1$</th>
<th>$A_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.00</td>
</tr>
<tr>
<td>-0.05</td>
<td>0.1564</td>
</tr>
<tr>
<td>-0.01</td>
<td>0.0326</td>
</tr>
<tr>
<td>-0.00</td>
<td>0.0000</td>
</tr>
<tr>
<td>+0.01</td>
<td>-0.0332</td>
</tr>
<tr>
<td>+0.05</td>
<td>-0.1728</td>
</tr>
</tbody>
</table>

Table 6.10. Difference in MSEs of strategies $S_4$ and $S_5$ for $n = 30$, $\rho = 0.05$

<table>
<thead>
<tr>
<th>$A_1$</th>
<th>$A_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.00</td>
</tr>
<tr>
<td>-0.05</td>
<td>0.1564</td>
</tr>
<tr>
<td>-0.01</td>
<td>0.0326</td>
</tr>
<tr>
<td>0.00</td>
<td>0.0000</td>
</tr>
<tr>
<td>+0.01</td>
<td>-0.0332</td>
</tr>
<tr>
<td>+0.05</td>
<td>-0.1728</td>
</tr>
</tbody>
</table>

Table 6.8 presents the difference in mean square errors for $n = 30$ and $\rho = 0$. We observe that as $\sigma^2$ (the within-trial variance) increases, the difference in MSE's becomes larger, which makes $S_5$ more efficient than $S_4$. The results for $n = 60$ were similar, and hence not tabulated.
Tables 6.9 and 6.10 consider the case \( \rho \neq 0 \). We observe that for low levels of \( A_2 (A_2 = 0.00, 0.05 \) and \( 0.10 \), high positive measurement bias \( (A_1 = 0.05) \), and low \( \rho (\rho = 0 \) and \( 0.01 \), \( S_4 \) is better than \( S_5 \). But, as the within-trial variance increases, or \( \rho \) increases, \( S_5 \) performs better than \( S_4 \).

The results for the intrasample correlation model were similar, and hence have not been presented.

The results of our study indicate:

(i) if measurement bias is negative, then \( S_5 \) is better than \( S_4 \), i.e. when ratio estimator is used, Midzuno's scheme is better than simple random sampling without replacement;

(ii) the larger the \( \rho \), the better \( S_5 \) performs in comparison with \( S_4 \); and

(iii) the larger the within-trial variance \( \sigma^2 \), the better \( S_5 \) is in comparison with \( S_4 \).
7. SUMMARY

The effects of measurement error in survey sampling are investigated in this thesis. Two models, viz., the "simple correlation model" and the "intrasample correlation model" are considered.

The usual unbiased estimators in equal and unequal probability sampling with replacement are studied under the simple correlation model. The Horvitz-Thompson estimator and the simple mean are also examined under the simple correlation model. The unbiased estimators in the three sampling schemes, viz., probability proportional to size with replacement, simple random sampling with replacement and simple random sampling without replacement are compared in an empirical study. It is found that if the within-trial variance $\sigma^2$ is large, then simple random sampling can be more efficient than probability proportional to size with replacement sampling scheme.

The ratio estimator is studied under the two sampling schemes, namely, simple random sampling without replacement (srswor) and Midzuno's scheme. The bias and MSE of the ratio estimator under these two schemes and the two measurement error models are derived and compared. Also, a comparison is made in an empirical study. It is found that Midzuno's scheme is more efficient than srswor if: (1) the measurement bias is negative, or (2) the within-trial variance is large, or (3) the correlation coefficient $\rho$ is large. The scheme srswor is more efficient than Midzuno's scheme when measurement bias is positive and both $\rho$ and $\sigma^2$ are small.
The results are extended to two-stage sampling and stratified sampling. The optimum allocation of sample size to strata is obtained in the presence of measurement errors. However, its effect has not been studied thoroughly. Further investigation in this case is needed.
8. BIBLIOGRAPHY


Eckler, A. R., and L. Pritzker. 1951. Measuring the accuracy of 
enumerative surveys. Bull. de Institut. International de 


methods and theory illustrated by the work of the United States 
Bureau of the Census. In N. L. Johnson and H. Smith, Jr. (eds.). 


errors in censuses and surveys. Bull. de Institut. International 

Hansen, M. H., W. N. Hurwitz, and L. Pritzker. 1964. The estimation 
and interpretation of gross differences and the simple response 

Hansen, M. H., W. N. Hurwitz, and L. Pritzker. 1968. Standardization 
of procedures for the evaluation of data: measurement errors and 

Hanson, R. H., and E. S. Marks. 1958. Influence of the interviewers 

variance components in sample surveys. p. 35-43. In survey sampling 

Horvitz, D. G., and D. J. Thompson. 1952. A generalization of sampling 
without replacement from a finite universe. J. Amer. Stat Assoc. 
49:663-685.


9. ACKNOWLEDGEMENTS

I wish to express my sincere appreciation to Dr. Chien-Pai Han for his patience, encouragement and understanding during the preparation of this thesis.

I would like to thank Dr. Herbert A. David and Dr. Wayne A. Fuller for the financial support provided in the form of teaching and research assistantships.

Thanks are due to Frances Bradley for the typing of this thesis.

Finally, I am indebted to my parents for their moral support during my stay in Ames. I am also thankful to my father for numerous discussions about my thesis and several other topics.
10. APPENDIX

Let $\theta_1, \theta_2, ..., \theta_N$ and $\delta_1, \delta_2, ..., \delta_N$ be two sets of real numbers. Define

$$
\overline{\theta}_n = n^{-1} \sum_{i}^{n} \theta_i
$$

$$
\overline{\delta}_n = n^{-1} \sum_{i}^{n} \delta_i
$$

$$
\overline{\theta}_N = N^{-1} \sum_{i}^{N} \theta_i
$$

$$
\overline{\delta}_N = N^{-1} \sum_{i}^{N} \delta_i
$$

$$
S^2_{\theta} = (N - 1)^{-1} \sum_{i}^{N} (\theta_i - \overline{\theta}_N)^2
$$

$$
S^2_{\delta} = (N - 1)^{-1} \sum_{i}^{N} (\delta_i - \overline{\delta}_N)^2
$$

$$
S_{\theta \delta} = (N - 1)^{-1} \sum_{i}^{N} (\theta_i - \overline{\theta}_N)(\delta_i - \overline{\delta}_N)
$$

$$
c_{\theta} = S_{\theta}/\overline{\theta}_N
$$

$$
c_{\delta} = S_{\delta}/\overline{\delta}_N
$$

$$
c_{\theta \delta} = S_{\theta \delta}/(\overline{\theta}_N \overline{\delta}_N)
$$

$$
a = (Nn)^{-1} (N - n)
$$
In this Appendix, we determine the expectation of some functions of \( \overline{\theta}_n \) and \( \overline{\delta}_n \) under simple random sampling without replacement. The expressions are obtained by assuming that the relevant Taylor's expansions are valid, and that we can ignore terms of order higher than \( O(n^{-1}) \).

Let \( \overline{\theta}_n = \overline{\theta}_N + \tau, \overline{\delta}_n = \overline{\delta}_N + \varepsilon \), then \( E(\tau) = E(\varepsilon) = 0 \), \( V(\tau) = E(\tau^2) = a_{\theta}^2 \), \( V(\varepsilon) = a_{\delta}^2 \), \( E(\varepsilon \tau) = a_{\theta\delta} \).

I. \( \overline{\theta}_n = \overline{\theta}_N + \tau, \overline{\delta}_n = \overline{\delta}_N + \varepsilon \), then \( E(\tau) = E(\varepsilon) = 0 \), \( V(\tau) = E(\tau^2) = a_{\theta}^2 \), \( V(\varepsilon) = a_{\delta}^2 \), \( E(\varepsilon \tau) = a_{\theta\delta} \).

\[
\left. \frac{\overline{\theta}_n}{\overline{\delta}_n} \right| \overline{\theta}_N = \overline{\theta}_N + \tau, \overline{\delta}_N = \overline{\delta}_N + \varepsilon, \quad \frac{\overline{\theta}_n}{\overline{\delta}_n} \left| \frac{\overline{\theta}_N}{\overline{\delta}_N} \right| (1 + \frac{\tau}{\overline{\theta}_N})(1 + \frac{\varepsilon}{\overline{\delta}_N})^{-1}
\]

\[
= \frac{\overline{\theta}_N}{\overline{\delta}_N} \left[ 1 + \frac{\tau}{\overline{\theta}_N} - \frac{\varepsilon}{\overline{\delta}_N} - \frac{\tau \varepsilon}{\overline{\theta}_N \overline{\delta}_N} + \frac{\varepsilon^2}{\overline{\delta}_N^2} \right]
\]

\[
\therefore \left. E \left( \frac{\overline{\theta}_n}{\overline{\delta}_n} \right) \right| \overline{\theta}_N = \overline{\theta}_N + \tau, \overline{\delta}_N = \overline{\delta}_N + \varepsilon = \frac{\overline{\theta}_N}{\overline{\delta}_N} \left[ 1 - a_{\theta\delta} + a_{\theta}^2 \right]
\] (10.1.)

II. \( \frac{\overline{\theta}_n^2}{\overline{\delta}_n^2} = \overline{\theta}_N^2 \left( 1 + \frac{\tau}{\overline{\theta}_N} \right) \left( 1 + \frac{\varepsilon}{\overline{\delta}_N} \right)^{-2} \)

\[
= \frac{\overline{\theta}_N^2}{\overline{\delta}_N^2} \left[ 1 + \frac{\tau}{\overline{\theta}_N} - 2 \frac{\varepsilon}{\overline{\delta}_N} - 2 \frac{\tau \varepsilon}{\overline{\theta}_N \overline{\delta}_N} + 3 \frac{\varepsilon^2}{\overline{\delta}_N^2} \right]
\]
\[
\begin{align*}
\therefore E(\frac{\theta_n}{\delta_n}) &= \frac{\theta_n}{\delta_n^2} \left[ 1 - 2a C_{\theta \delta} + 3a C_{\delta}^2 \right] \\
(10.2.) \\

\text{III. } \frac{\theta_n^2}{\delta_n} &= \frac{\theta_n^2}{\delta_n^2} \left( 1 + \frac{r}{\theta_n} \right)^2 \left( 1 + \frac{\xi}{\delta_n} \right)^{-1} \\
&= \frac{\theta_n^2}{\delta_n^2} \left( 1 + 2 \frac{r}{\theta_n} + \frac{r^2}{\theta_n^2} \right) \left( 1 - \frac{\xi}{\delta_n} + \frac{\xi^2}{\delta_n^2} \right) \\
&= \frac{\theta_n^2}{\delta_n^2} \left[ 1 + 2 \frac{r}{\theta_n} - \frac{\xi}{\delta_n} - 2 \frac{\xi r}{\theta_n \delta_n} + \frac{r^2}{\delta_n^2} \right] \\
\therefore E(\frac{\theta_n}{\delta_n}) &= \frac{\theta_n^2}{\delta_n^2} \left[ 1 - 2a C_{\theta \delta} + a C_{\theta}^2 + a C_{\delta}^2 \right] \\
(10.3.) \\

\text{IV. } \frac{\theta_n^2}{\delta_n^2} &= \frac{\theta_n^2}{\delta_n^2} \left( 1 + \frac{r}{\theta_n} \right)^2 \left( 1 + \frac{\xi}{\delta_n} \right)^{-2} \\
&= \frac{\theta_n^2}{\delta_n^2} \left( 1 + 2 \frac{r}{\theta_n} + \frac{r^2}{\theta_n^2} \right) \left( 1 - 2 \frac{\xi}{\delta_n} + 3 \frac{\xi^2}{\delta_n^2} \right) \\
&= \frac{\theta_n^2}{\delta_n^2} \left[ 1 + 2 \frac{r}{\theta_n} - \frac{\xi}{\delta_n} - 4 \frac{\xi r}{\delta_n \theta_n} + 3 \frac{\xi^2}{\delta_n^2} + \frac{r^2}{\delta_n^2} \right] \\
\therefore E(\frac{\theta_n}{\delta_n}) &= \frac{\theta_n^2}{\delta_n^2} \left[ 1 - 4a C_{\theta \delta} + 3a C_{\theta}^2 + a C_{\delta}^2 \right] \\
(10.4.)
\end{align*}
\]