1992

Variance estimation under random imputation

Margot Helena Tollefson
Iowa State University

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

Recommended Citation
Tollefson, Margot Helena, "Variance estimation under random imputation " (1992). Retrospective Theses and Dissertations. 10352.
https://lib.dr.iastate.edu/rtd/10352

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 manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.
Variance estimation under random imputation

Tollefson, Margot Helena, Ph.D.

Iowa State University, 1992
Variance estimation under random imputation

by

Margot Helena Tollefson

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

Major: Statistics

Approved:

Signature was redacted for privacy.

In Charge of Major Work

Signature was redacted for privacy.

For the Major Department

Signature was redacted for privacy.

For the Graduate College

Iowa State University
Ames, Iowa

1992
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>LIST OF PRINCIPLE RESULTS</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>v</td>
<td>v</td>
</tr>
</tbody>
</table>

## 1. INTRODUCTION

| 1.1. The problem of missing data | 1 |
| 1.2. Adjustments in the presence of nonresponse | 2 |
| 1.3. Approaches to the nonresponse problem | 3 |
| 1.4. Our approach to estimation | 4 |
| 1.5. The contents of the chapters | 6 |

## 2. A REVIEW OF THE LITERATURE ON NONRESPONSE

| 2.1. Introduction to the problem of nonresponse | 7 |
| 2.2. The size of the nonresponse problem | 17 |
| 2.3. Three approaches to the nonresponse problem | 25 |
| 2.4. Some solutions posed in the literature for variance estimation with missing values | 39 |

## 3. RANDOM IMPUTATION FOR SIMPLE RANDOM SAMPLING

| 3.1. Missing less than responding | 82 |
| 3.1.1. The population and sample | 82 |
| 3.1.2. The expectation and variance of the imputed mean under Little and Rubin | 84 |
| 3.1.3. The expectation and variance of the imputed mean under Kalton | 86 |
| 3.1.4. Kalton's expression for the variance of the imputed mean under the MAR assumption | 93 |
| 3.1.5. The expectation and variance of the imputed mean following our approach | 95 |
| 3.1.6. Estimation of the variance of the imputed mean | 99 |
| 3.1.7. A comparison of our estimators to Kalton's estimator | 102 |

| 3.2. The general case | 104 |
| 3.2.1. Introduction to the general case | 104 |
| 3.2.2. Notation and imputation for the general case | 105 |
| 3.2.3. The expectation and variance of the imputed mean under Little and Rubin | 107 |
| 3.2.4. The expectation and variance of the imputed mean under Kalton | 108 |
| 3.2.5. The expectation and variance of the imputed mean under our method | 111 |
3.2.6. Estimation of the variance of the imputed mean under the general model

4. RANDOM IMPUTATION FOR A STRATIFIED RANDOM SAMPLE

4.1. The finite population and sample
4.2. The imputation method
4.3. The expectation of the imputed mean over the finite population
4.4. The superpopulation structure
4.5. The expectation of the imputed mean under the superpopulation model
4.6. The variance of the imputed mean under the superpopulation model
4.7. Estimation of the variance of the imputed mean

5. IMPUTATION UNDER GENERAL SAMPLING SCHEMES

5.1. The general model
5.1.1. The finite population and sample
5.1.2. The missingness structure and imputation
5.1.3. The expectation of the imputed total for a finite population
5.1.4. The superpopulation structure
5.1.5. The expectation of the imputed total under the superpopulation model
5.1.6. The variance of the imputed total under the superpopulation model
5.1.7. Estimation of the variance of the imputed total minus the finite population total
5.1.8. Estimation of the variance of the imputed total
5.1.9. The superpopulation structure and sample properties for two characteristics
5.1.10. Imputation methods for two variables
5.1.11. The covariance of two imputed totals under the superpopulation model
5.1.12. Estimation of the covariance of two imputed totals minus their respective finite population totals

5.2. The model in PC CARP

5.2.1. Introduction to the estimation of the variance of the imputed total and the covariance of two imputed totals for PC CARP
5.2.2. The population and sample structure in PC CARP
5.2.3. The expectation and variance of the imputed total for PC CARP
Page 294
5.2.4. Estimation of the variance of the imputed total minus the finite population total for PC CARP
Page 296
5.2.5. The covariance in PC CARP
Page 301
5.2.6. The covariance of two imputed totals minus their respective finite population totals for PC CARP
Page 302
5.2.7. Estimation of the covariance of two imputed totals minus their respective finite population totals for PC CARP
Page 303

6. IMPUTATION FOR THE SOIL CONSERVATION SERVICE'S 1987 NATIONAL RESOURCES INVENTORY
Page 306
6.1. The problem
Page 306
6.2. A description of the variables
Page 307
6.3. A description of the data from Indiana
Page 309
6.4. A description of the data in Kansas
Page 314
6.5. Three imputation methods
Page 319
6.6. Evaluation of the imputation methods using the data sets
Page 334
REFERENCES
Page 348
ACKNOWLEDGEMENTS
Page 349
## LIST OF PRINCIPLE RESULTS

<table>
<thead>
<tr>
<th>Theorem or Expression</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expression 3.43</td>
<td>Expected value of estimator of mean under simple random sampling without replacement and the number missing less than the number responding (p. 96)</td>
</tr>
<tr>
<td>Expression 3.48</td>
<td>Variance of estimator of mean under simple random sampling without replacement and the number missing less than the number responding (p. 99)</td>
</tr>
<tr>
<td>Expression 3.49</td>
<td>First estimator of variance of mean under simple random sampling without replacement and the number missing less than the number responding (p. 99)</td>
</tr>
<tr>
<td>Expression 3.54</td>
<td>Second estimator of variance of mean under simple random sampling without replacement and the number missing less than the number responding (p. 101)</td>
</tr>
<tr>
<td>Expression 3.84</td>
<td>Expected value of estimator of mean under simple random sampling without replacement and the number missing arbitrary (pp. 111–112)</td>
</tr>
<tr>
<td>Expression 3.89</td>
<td>Variance of estimator of mean under simple random sampling without replacement and the number missing arbitrary (p. 114)</td>
</tr>
<tr>
<td>Expression 3.91</td>
<td>First estimator of variance of mean under simple random sampling without replacement and the number missing arbitrary (p. 115)</td>
</tr>
<tr>
<td>Expression 3.92</td>
<td>Second estimator of variance of mean under simple random sampling without replacement and the number missing arbitrary (p. 117)</td>
</tr>
<tr>
<td>Expression 3.95</td>
<td>Third estimator of variance of mean under simple random sampling without replacement and the number missing arbitrary (p. 118)</td>
</tr>
<tr>
<td>Theorem 4.1</td>
<td>Expected value of estimator of mean given the finite population under stratified random sampling (p. 128)</td>
</tr>
</tbody>
</table>
### LIST OF PRINCIPLE RESULTS (CONT’D)

<table>
<thead>
<tr>
<th>Theorem or Expression</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>Theorem 4.2</td>
<td>Expected value of estimator of mean using the superpopulation model under stratified random sampling (p. 130)</td>
</tr>
<tr>
<td>Theorem 4.3</td>
<td>Variance of estimator of mean using the superpopulation model and conditioning on the sample sizes in imputation classes and the number of respondents in stratum–imputation class subgroups under stratified random sampling (p. 137)</td>
</tr>
<tr>
<td>Theorem 4.4</td>
<td>Variance of the estimator of the mean conditioning on the sample sizes and the number of respondents in the imputation classes under stratified random sampling (p. 141)</td>
</tr>
<tr>
<td>Theorem 4.5</td>
<td>Variance of the estimator of the mean conditioning on the sampling pattern and the response pattern and the imputation pattern under stratified random sampling (p. 143)</td>
</tr>
<tr>
<td>Theorem 4.7</td>
<td>First estimator of the variance of the estimator of the mean under stratified random sampling (p. 155)</td>
</tr>
<tr>
<td>Theorem 4.8</td>
<td>Second estimator of the variance of the estimator of the mean under stratified random sampling (p. 159)</td>
</tr>
<tr>
<td>Theorem 4.9</td>
<td>Third estimator of the variance of the estimator of the mean under stratified random sampling (p. 161)</td>
</tr>
<tr>
<td>Theorem 5.1</td>
<td>Expected value of the estimator of the population total given the finite population, the sample, and the sample sizes and number of respondents in each imputation class under general sampling designs (p. 169)</td>
</tr>
<tr>
<td>Theorem 5.2</td>
<td>Expected value of the estimator of the population total under general sampling designs (p. 176)</td>
</tr>
<tr>
<td>Theorem 5.3</td>
<td>Variance of the estimator of the population total under general sampling designs (p. 179)</td>
</tr>
</tbody>
</table>
### LIST OF PRINCIPLE RESULTS (CONT’D)

<table>
<thead>
<tr>
<th>Theorem or Expression</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>Theorem 5.4</td>
<td>Variance of the estimator of the population total minus the finite population total under general sampling designs (p. 195)</td>
</tr>
<tr>
<td>Theorem 5.6</td>
<td>Estimator of the variance of the estimator of the population total minus the finite population total under general sampling designs (p. 205)</td>
</tr>
<tr>
<td>Expression 5.131</td>
<td>Estimator of the variance of the estimator of the population total under general sampling designs (p. 216)</td>
</tr>
<tr>
<td>Theorem 5.9</td>
<td>Covariance between estimators of two population totals under general sampling designs (p. 245)</td>
</tr>
<tr>
<td>Theorem 5.10</td>
<td>Covariance between the differences between two estimators of the population totals and the respective finite population totals under general sampling designs (p. 281)</td>
</tr>
<tr>
<td>Theorem 5.13</td>
<td>Estimator of the covariance between the differences between two estimators of population totals and the respective population totals under general sampling designs (p. 287)</td>
</tr>
<tr>
<td>Theorem 5.14</td>
<td>Estimator of the variance of the difference between the estimator of the population total and the finite population total under a design in PC CARP (p. 296)</td>
</tr>
<tr>
<td>Theorem 5.15</td>
<td>Estimator of the covariance between the differences between two estimators of population totals and the respective finite population totals under a design in PC CARP (p. 304)</td>
</tr>
</tbody>
</table>
1. INTRODUCTION

1.1. The problem of missing data

When data are collected for a survey, generally, the data collectors are unable to collect data from some of the units, or on some of the items within the units. The problem of missing data in survey sampling is called the problem of nonresponse. In this work, we address the problem of variance estimation in the presence of nonresponse. We look at variance estimation under the nonresponse adjustment of random imputation.

Nonresponse causes possible biases in estimators of population characteristics if no adjustments are made to compensate for the nonresponse. Biases occur when the nonrespondents are different from the respondents with respect to the variables of interest in the survey. Adjustments for nonresponse have the primary purpose of reducing the nonresponse bias in estimators. We consider the situation in which no information about the nonrespondents is available from the survey. Therefore, the assumptions behind the adjustments for nonresponse are unverifiable from the survey. The assumptions behind the adjustments for nonresponse are made on the basis of past experience and (or) auxiliary surveys.

Since the effective sample size is reduced in the presence of nonresponse, estimators of population characteristics are less precise when nonresponse occurs. The estimation of variances under adjustments for nonresponse is a current area of research in survey sampling.
1.2. Adjustments in the presence of nonresponse

Commonly used adjustments for nonresponse fall into two categories, weighting adjustments, and imputation adjustments. Weighting adjustments, which are usually a weighting up of blocks of data from the observed sample size to the intended sample size, are often done in the presence of unit nonresponse. Unit nonresponse occurs when no information is gathered from a unit in the intended sample. Imputation adjustments, which are the placing of artificial data into the data set in the place of the missing observations, are usually made when the nonresponse is item nonresponse. Item nonresponse occurs when some, but not all, of the desired information is collected from a unit in the intended sample. Nonresponse adjustments can be a mixture of weighting and imputation adjustments.

Both weighting and imputation adjustments are usually made within classes, called, respectively, weighting classes and imputation classes. Classes are formed on the basis of knowledge from outside the sample and (or) information taken in the sample. In forming classes, the analyzer of the data generally believes that the variables of interest are somewhat homogeneous within the classes. Generally, homogeneity is unverifiable in the presence of nonresponse.

Random imputation is a form of imputation in which the values to be used in the imputation are chosen out of the set of respondents by a probability method. Random imputation is done either within the entire sample or within imputation classes.
1.3. Approaches to the nonresponse problem

There are two basic approaches to modeling nonresponse. The first approach divides the population into two strata, those that will always respond and those that will never respond. No probability distribution is put on the responses and the nonresponses. For the first approach, the estimators of population characteristics are valid only for the response strata. The first approach has the advantage that biases in estimators are explicitly modeled. However, it may not be realistic to assume that a unit in the population will either always respond or never respond on a given item in the survey.

The second approach assumes a probability distribution for the respondents and the nonrespondents. Under the second approach, each member of the population has a probability of responding on each item in the survey. The second approach has the advantage that, given a model for the probabilities, estimators of the population characteristics are valid for the entire population. The second approach has the disadvantage that the response probabilities have to be modeled in order to estimate the population parameters. Models for response probabilities are generally unverifiable in the presence of nonresponse.

A common model for response probabilities is the missing at random model. Under the missing at random model, the assumption is made that the probabilities of response are constant. The missing at random model has two forms. Observations are missing completely at random if the probabilities of response are constant throughout the population. Observations are missing at random if the probabilities of response are constant within imputation classes. Missing completely at random and missing at random are defined in Section 2.3.
1.4. Our approach to estimation

The estimators that are given in this paper are intended to be easy to implement using available data analysis routines. The estimators for the means and totals are the standard estimators for the means and totals for a complete data set. In the estimators for the means and totals, the complete data set is replaced by the imputed data set.

The variances that are found in this paper are composed of two terms. The first term is the variance of the standard complete sample estimator of the population mean or total. The second term is a correction term that inflates the variance of the complete sample estimator to account for the imputation and the fact that the effective sample size with missing values is less than the full sample size.

The estimators for the variances are composed of three terms. The first term is the standard complete sample estimator of the variance of the standard complete sample estimator of the population mean or total, computed using a data set with imputed values. The second term is a correction term to account for the fact that the sample used in the first term is an imputed sample rather than a complete sample. The sum of the first term and the second term unbiasedly estimates the variance of the standard complete sample estimator of the population mean or total. The third term is an unbiased estimator of the second term in the variance, that is, of the correction term for imputation and the loss of sample size.

The second term in the estimators of the variances is generally small relative to the first and third terms. For simple random sampling without replacement, with the missing values missing completely at random and the imputation done as
described in Chapter 3 within one imputation class, the expectation of the second
term, conditional on the number of respondents, is

\[(n-1)^{-1}n^{-2}(2n - r - r[r^{-1}n]) S^2, \tag{1.4}\]

where \(r\) is the number of respondents, \(n\) is the sample size, \(S^2\) is the population
variance, and \([x]\) is the largest integer less than or equal to \(x\). The smallest value
that \(r\) can take on if the variance is to be estimated is 2. The full sample
variance is

\[n^{-1} S^2, \tag{1.5}\]

where we ignore the finite population correction factor. The second term in the
variance, conditional on the number of respondents, is

\[n^{-2}(2n - r - r[r^{-1}n]) S^2. \tag{1.6}\]

We can see that the term is \((n-1)\) times expression (1.4). We can see that the
second term goes to zero at least \(n\) times the rate that the full sample variance
goes to zero. We can see that the minimum value, for \(r \geq 2\), of the expectation of
the estimator of the full sample variance in relation to the actual full sample
variance is (for \(r=2\) and \(n=3\)) \(3^{-1} 2^{-1} n^{-1}\) and that as \(n\) increases, the expected
value of the full sample variance estimator quickly approaches the full sample
variance, even for \(r=2\).

The estimate of the population total or mean and the first term in the
estimate of the variance can usually be calculated using the imputed data set in standard data analysis routines. For most sample designs, the second term in the estimator of the variance of the population mean or total can be ignored if \( n \) is reasonably large. For most sample designs, the third term in the estimator of the variance of the population mean or total can also be calculated using available computer routines.

1.5 The contents of the chapters

In Chapter 2, the literature on nonresponse is reviewed. In Chapter 3, estimators are found for simple random sampling designs. The work of Little and Rubin (1987) and Kalton (1983) in the area of random imputation under simple random sampling is covered in Chapter 3. In Chapter 4, estimators are found for stratified random sampling designs. Three variance estimators, taken to three levels of conditioning, are presented. In Chapter 5, estimators are found for any sampling design where the Horvitz—Thompson estimator of the population total is applicable, that is, any design for which the probability of inclusion in the sample is greater than zero for every member of the population. In Chapters 4 and 5, a superpopulation structure is assumed, since the expected values of the estimators of the population mean or total are biased when the expectation is taken over the finite population with no underlying superpopulation. In Chapter 6, the results in Chapter 5 are applied to an imputation problem that the Soil Conservation Service is facing with its 1992 National Resources Inventory.
2. A REVIEW OF THE LITERATURE ON NONRESPONSE

In this chapter, we review some of the literature on nonresponse. In Section 2.1, we look at general works on nonresponse by Cochran (1977), Kalton and Kasprzyk (1986), and Platek (1980). In Section 2.2, we review papers by Murthy (1979), Thomsen and Siring (1979), and Bailar (1989) that describe the size of the problem of nonresponse. In Section 2.3, we describe different approaches to the problem of nonresponse as described by Little and Rubin (1987), Kalton (1983), and Platek and Grey (1979). In Section 2.4, we consider specific solutions to the problem of variance estimation for nonresponse as presented by Hansen, Hurwitz, and Madow (1953), Kalton and Kish (1981), Kalton and Kish (1984), Särndal and Swensson (1987), Rubin (1987), Schafer (1990), and Rao and Shao (1992).

2.1. Introduction to the problem of nonresponse

Cochran (1977), in his classic textbook *Sampling Techniques*, devotes a chapter to the sources of errors in surveys. Part of the chapter is an introduction to the problem of nonresponse. Cochran (1977) describes nonresponse as a failure to measure some of the units in the sample. Cochran (1977) attributes the failure to oversight, difficulties in locating the unit, or a refusal by the subject of the survey to answer questions in the survey. Cochran (1977) describes an approach to the problem of nonresponse as a division of the population into two strata, the first containing all subjects that respond and the second containing all subjects that do not respond. Cochran (1977) also suggests that a more complete specification of the problem would be to assign to each element of the population a probability of response.
Cochran (1977) points out that no information is available from the sample for points that are nonrespondents. Cochran (1977) gives the warning that one cannot assume that the nonrespondents have the same characteristics as the respondents. Cochran (1977) then gives an elucidating example taken from answers to a mail survey by fruit growers in North Carolina. The proportion responding to the mail survey increased as the size of the farm increased.

Using the two strata approach, Cochran (1977) finds the bias in the estimator of the mean of a population, where the estimator is found by the usual methods using only the respondents. The bias is directly proportional to the number of nonrespondents in the population and, to the difference between the population mean of the respondents and the population mean of the nonrespondents. Cochran (1977) warns the reader that one cannot get a good feel for the size of the difference between the population mean of the respondents and the population mean of the nonrespondents from the respondents alone. For the special case of proportions, Cochran (1977), gives a method for finding a conservative confidence interval for the mean of the population.

Cochran (1977) divides the reasons for nonresponse into four classes. Noncoverage is the first of the four classes. Cochran (1977) describes noncoverage as a failure to locate or visit the sample point. The failure could be due to problems with aerial sampling units, incomplete lists, weather, or poor transportation. Cochran (1977) suggests a revisit to the area with a recount as a solution to problems with aerial photography. Cochran (1977) suggests that an area sample be taken in conjunction with the list for noncoverage due to an incomplete list.

The second reason that Cochran gives for nonresponse is not—at—homes. Cochran (1977) describes a not—at—home as a person who lives at home but who is
out when the data recorder calls. Cochran (1977) writes it is easier to get a response if any adult in the household can answer the questions. Cochran (1977) also discusses optimality in terms of the number of call backs and the cost and the efficiency of the estimators. Cochran (1977) describes two methods for collecting and analyzing data that minimize the problem of not—at—homes. The third reason that Cochran (1977) gives for nonresponse is unable—to—answer. Cochran (1977) writes the a person is unable—to—answer if the person either does not have or is unwilling to give the information requested. The fourth reason that Cochran (1977) gives for nonresponse is that the person being questioned is "hardcore". The "hardcore" includes persons that refuse to answer any question, persons that are incapacitated, and persons that are away from home during the entire period of the sample.

Cochran (1977) does not mention imputation in his discussion of the nonresponse problem, but he does give a good feel for the problem and provides practical advise on dealing with nonresponse.

Kalton and Kasprzyk (1986) describe nonresponse and cover a variety of methods for dealing with nonresponse. Kalton and Kasprzyk (1986) differentiate between item and unit nonresponse. Kalton and Kasprzyk (1986) note that most surveys are made up of a number of items per sampling unit (for example, a questionnaire with ten questions measures ten items per unit). With unit nonresponse, no items are measured for the given unit. Kalton and Kasprzyk (1986) write that unit nonresponse is the result of a refusal by the unit to participate, the inability of the unit to participate, not—at—homes, or untraced units. With item nonresponse, contact is made with the sample unit, but some of the items on the survey are not completed or are inconsistent. Kalton and Kasprzyk (1986) attribute
item nonresponse to items that the person does not want to give an answer to, items a person does not know the answer to, items the the data recorder omits, or items that the editor of the data does not find acceptable. Kalton and Kasprzyk (1986) write that weighting adjustments are usually used for unit nonresponse, where no information is available from the sample for the nonrespondents, but that weighting adjustments are impractical for item nonresponse. For item nonresponse, Kalton and Kasprzyk (1986) write that imputation is usually used to adjust for nonresponse. Kalton and Kasprzyk (1986) also describe a type of nonresponse that falls between item and unit nonresponse. Kalton and Kasprzyk (1986) write that partial nonresponse occurs when some of the data is collected for a given unit, but most is not. Kalton and Kasprzyk (1986) describe partial nonresponse as the result of a respondent terminating an interview prematurely or data that is collected from some but not all of a household or data from an individual that responds on some but not all waves of a panel survey. Kalton and Kasprzyk (1986) give references to papers that deal with partial nonresponse.

Kalton and Kasprzyk (1986) devote a section of their paper to weighting adjustments for unit nonresponse. Kalton and Kasprzyk (1986) write that weighting adjustments are closely related to imputation.

Kalton and Kasprzyk (1986) describe nine methods of imputation. The first method that Kalton and Kasprzyk (1986) describe is deductive imputation. Deductive imputation is the the ideal form of imputation. In deductive imputation, the editor of the data fills in a value that deductively should be the correct value (for example, an editor would deduce that a person with the name Mary is a female). Overall mean imputation is the second method that Kalton and Kasprzyk (1986) describe. In overall mean imputation, the mean of the respondents becomes
the value that replaces all of the missing points. The third method that Kalton and Kasprzyk (1986) describe is class mean imputation. With class mean imputation, the processor of the data divides the sample into imputation classes on the basis of what is known about the population or sample. It is assumed that there is some similarity between the respondents and the nonrespondents within the imputation classes. The respondent mean within an imputation class is the value that replaces all of the missing points within that imputation class. The fourth imputation method that Kalton and Kasprzyk (1986) discuss is random overall imputation. In random overall imputation, the missing values are filled in by randomly chosen respondents. The fifth method that Kalton and Kasprzyk (1986) describe is random imputation within classes. For random imputation within classes, the sample is divided into imputation classes and, within an imputation class, the missing values are replaced by randomly chosen respondents from the imputation class.

The sixth imputation scheme that Kalton and Kasprzyk (1986) describe is sequential hot-deck imputation. For sequential hot-deck imputation, first the data set is divided into imputation classes. An initial value for start up is stored for each imputation class. Each record is checked sequentially. If the record has a response, the response replaces the stored value. If the record is missing, the stored value is placed in the record. If the data are in random order, then sequential hot-deck imputation is a random imputation method. If the data are in some sequential order, then sequential hot-deck imputation improves the efficiency of the estimator. The seventh imputation scheme that Kalton and Kasprzyk (1986) describe is hierarchical hot-deck imputation. For hierarchical hot-deck imputation, the data is divided into highly detailed imputation classes on the basis of auxiliary information. A search is made through the imputation class for a match to the
missing data point. If no match is found, the imputation classes are collapsed into larger imputation classes and another search is made. The process continues until a match is found.

The eighth method of imputation that Kalton and Kasprzyk (1986) describe is regression imputation. For regression imputation, a value for the missing point is found by regressing on auxiliary variables available for the missing point. Sometimes a random error term is added to the product of the regression to give good distributional properties to the imputed values. The last imputation method that Kalton and Kasprzyk (1986) describe is distance function matching. For distance function matching, the closest match to the missing point is found, based on auxiliary information, and replaces the missing point.

Kalton and Kasprzyk (1986) divide the imputation methods into two categories, deterministic and stochastic. Overall mean imputation, class mean imputation, and regression imputation are deterministic methods unless an error term is artificially added to the imputed value. Deterministic imputation produces spikes in the distribution of the imputed data set. The rest of the imputation methods listed above are stochastic methods. They contain an error component consistent with the inherent scatter of the distribution out of the randomness of the selection procedure. When stochastic methods are used, precision is lost in the estimator of the simple mean.

Kalton and Kasprzyk (1986) compare the imputation class approach to regression imputation. Regression imputation must be modeled carefully or large errors can occur. Regression imputation is potentially better than any of the imputation class approaches if the sample is small, but surveys with many variables can be hard to model.
Kalton and Kasprzyk (1986) note that when imputation is done, correlations are biased towards zero. Kalton and Kasprzyk (1986) mention Rubin's multiple imputation technique as a method for finding a good estimator for the variance of the estimator of a population characteristic. Kalton and Kasprzyk (1986) also write that bias in the estimator of the population characteristic can only be controlled by keeping the number of missing values small, since the sample contains only partial information about the nonrespondents.

The Kalton and Kasprzyk (1986) paper provides a good overview of techniques that are commonly used to adjust for nonresponse. The comprehensive approach to comparing the various imputation methods is an aid for choosing a method for imputation.

Platek (1980) gives a good general coverage of the nonresponse problem, including discussion of techniques for prevention of nonresponse and techniques for adjustments when nonresponse occurs. Platek (1980) describes nonresponse as being a result of an incomplete frame, units that cannot be found, refusal of the person being questioned to participate, partially complete records, or invalid responses. Platek (1980) suggests that ignoring the nonresponse and analyzing the responses as if they are a full sample is a poor idea, since nonresponse introduces a possible bias. Also, nonresponse reduces the effective sample size and, thus, increases the sampling variance. Platek (1980) writes that it is believed that for many large scale surveys, the unknown error due to nonresponse exceeds the known error due to sampling. Platek (1980) also writes that the errors due to sampling are the only ones that are usually identified.

Platek (1980) suggests that an overall approach should be taken to reduce the effect of nonresponse, starting with the survey design and including data
collection, processing, and estimation. Platek (1980) writes that at the processing stage, the sample surveyor can use imputation or weighting adjustments to adjust for nonresponse. Platek (1980) notes that the adjustments are somewhat effective in reducing bias. Platek (1980) also notes that well designed data collection reduces nonresponse and the need for adjustments.

Platek (1980) defines nonresponse as "a failure to obtain a usable report from a reporting unit which legitimately falls into the sample in a particular survey." (p. 94). Platek (1980) differentiates between unit nonresponse and item nonresponse. Platek (1980) suggests that nonresponse is a result of operational difficulties, time and cost constraints, respondents not cooperating, and interviewers not tracking down missing respondents. Platek (1980) notes that the higher the rate of nonresponse, the higher the possible bias and the less likely it is that the objectives of the survey will be satisfied.

Platek (1980) makes the observation that in the presence of nonresponse the sample ceases to be a probability sample because nonrespondents tend to be different from respondents. Platek (1980) then introduces a model that assigns a probability of response to the units. Platek (1980) notes that if the probability of response varies according to the size of the characteristic, then weighting adjustments on the total sample do not work. If sampling is simple random sampling without replacement, then the estimator of the mean has a large bias. For probability proportional to size sampling, the estimator will be unbiased if the probability of response is also proportional to size.

Platek (1980) devotes a section to methods of dealing with nonresponse. Platek (1980) writes that the experienced survey designer will usually have a good idea of what response rate to expect for a given design. The survey designer will be
interested in the effect of nonresponse on the mean square error. The survey designer will need to balance survey cost with other factors to keep nonresponse small enough to meet the goals of the survey. For example, with a smaller sample size, one can use more intense collection methods. Platek (1980) divides the factors that are to be taken into account in the final design into three groups. In the first group are sample size, stratification, degree of clustering, sample allocation, and method of selection. In the second group are sample frame, method of interviewing, selection, training, and control of staff, length and wording of the questionnaire, sensitivity of the question, type of area in which the survey is taken, feasibility and cost of call backs, and publicity. In the third group are editing, imputation, estimation, and variance estimation. The nonresponse rate and the non-sampling components of the mean square error are affected by the second group of factors.

In data collection, Platek (1980) distinguishes between refusals and no contacts. No contacts are operational and need sufficient resources and good planning to minimize. To decrease refusals, it is important to provide motivation to respond. Persons will not respond if they have difficulty understanding questions, do not want to spend the time to answer the survey, feel that the survey invades their privacy, are indifferent to the goals of the survey, have difficulty recalling information, or find the questions are embarrassing. Persons are motivated to respond in they are interested in the survey, have a willingness to help, out of a sense of duty, or understand the importance of the survey results. Platek (1980) suggests introductory letters, examples of uses of data, brochures describing the objective and authority of the survey, assurances of privacy, and compensation as motivation methods to improve response.

Platek (1980) gives some specific methods for dealing with nonresponse. The
first of the methods is substitution for the nonrespondent in the field (for example, if a person is not at home, interview the closest neighbor). Substitution in the field is a form of imputation and can bias estimates. The second of the methods is call backs, which can be costly. The third method is to use proxies within the household for the nonresponding member of the household. Unfortunately, proxies do not always provide accurate information. The fourth method is imputation of the missing data points.

Platek (1980) writes that there is no known unbiased method of imputing. Unit nonresponse is usually weighted up to the full sample size. Imputation can be deductive (deterministic), from past data, hot deck, or by regression. Platek (1980) writes that it is often assumed that the probability of response is constant and possibility of bias ignored.

Platek (1980) describes several imputation methods. The first of these methods is weighting in adjustment cells. Adjustment cells are formed for the purpose of weighting the sample. It is assumed the the probability of response is constant within a cell. The sample weights are inflated by the inverse of the observed response rate in the cell. If the probability of response is not constant within the cells, then a bias occurs in the estimator. The second method that Platek (1980) describes is the duplication method. In the duplication method, some or all of the respondents are used to fill in the missing values (are duplicated). The third method that Platek (1980) describes is the hot deck method. The hot deck method is a type of duplication method where the closest member on a list is used to fill in the missing value. The variance is difficult to find for the hot deck method. The cold deck method uses a deterministic (deductive) choice of donor. The fourth method that Platek (1980) suggests is historical data substitution. In
historical data substitution, an external source from a previous time is used to fill in the missing values. The fifth method that Platek (1980) describes is the zero substitution method. The zero substitution method is the method of ignoring the nonresponse and treating the data as if there were no nonresponse.

Platek (1980) gives a description of the variance problem encountered with nonresponse. Platek (1980) then describes the application of the methods described in the paper to the household surveys at Statistics Canada.

2.2. The size of the nonresponse problem

In this section, we look at some papers that cover the size of the nonresponse problem. Bailar (1989), in a paper that appeared in Panel Surveys, looks at the problem of coverage and nonresponse as studied by the U.S. Census Bureau. Murthy (1979), has studied the size of the problem of nonresponse in Asia and the Pacific. Thomsen and Siring (1979) present data on the growing nonresponse problem in Norway. This section is not a comprehensive coverage of the size of the nonresponse problem, but a review of a few papers that cover the size of the nonresponse problem for specific surveys and and (or) places in the world.

Bailar (1989), writes that the survey industry is growing and that the industry assumes that asking questions of people will give satisfactory results. Bailar (1989) lists the premises behind the assumption as being that the sample surveyor knows who and what to ask and that those persons asked will answer and answer accurately. Bailar (1989) describes various types of information that are needed and the types of surveys that give the specific kinds of information needed. Bailar (1989) covers three types of errors; coverage and nonresponse errors, recall errors, and time-in-sample errors. Bailar (1989) writes that coverage is related to
the frame used and nonresponse is related to a failure to respond by a person in the survey, but that the two are related when the frame is developed during the survey, which is the case in most household surveys. For example, if the sample unit is a housing unit and the target population is all individuals within housing units, then, if the interviewer collects data on the number of persons in the household at the time of the interview and a person in the household is not listed as in the household, then undercoverage occurs due to nonresponse.

Undercoverage due to nonresponse accounts for most of the undercoverage in the Current Population Survey and the Survey of Income and Program Participation. Using data from the 1980 Census, the Census Bureau found that 93% of the persons 14 years old and older that were supposed to be there were found by the Current Population Survey. The same percentage held for white males. For white females the figure was 95%, for black males the figure was 84%, for black females the figure was 91%, for other race males the figure was 96%, and for other race females the figure was 97%. Bailar (1989) notes that unit nonresponse rates are lower for blacks and for males in the U.S..

For longitudinal surveys attrition is a problem. Bailar (1989) writes that one can find surveys where the response rate is down to 50% after many waves. In the Survey of Income and Program Participation conducted by the Census Bureau, in the first wave in 1984, interviewers were unable to collect data at 5% of the eligible units. On the second wave, this figure had risen to 8% and by the eighth wave, the figure had risen to 16%. Also, by the second wave, 1% of the households had moved and by the eighth wave 6% of the households had moved.

Bailar (1989) assesses how nonresponse varies with the type of survey. Single time surveys and repeated surveys without overlap have the lowest nonresponse
rates. Repeated surveys with partial overlap and longitudinal surveys with rotating panels and few visits are next. Longitudinal surveys with no rotation have the highest nonresponse rates.

Bailar (1989) lists some techniques that have been suggested for reducing the nonresponse problem. The techniques are to work on improving coverage of the whole housing unit, to check for errors in the classification of nonrespondents during the survey, to intensively try to get answers from refusals, to weight data to reduce bias, and, in longitudinal studies, to put resources into intensively trying to follow individuals. Bailar (1989) then assesses these techniques, based on her experience at the Census Bureau. The Census Bureau experience has been that over half of the coverage problem results from persons living in a household but not listed as in the household. Therefore improving coverage of whole housing units would not help reduce undercoverage much. For coverage checks, interviewers, when on an interview, classify houses that are nonrespondents into three groups; the household is a legitimate part of the survey, the house is vacant or the household is not a legitimate part of the survey, and the house is no longer a household. In a coverage check, another person checks to see if the household has been classified correctly. Interviewers tend to misclassify to reduce the number of legitimate households they cannot get answers from. In a coverage check done in the Current Population Survey in October, 1966, the Census Bureau found 1.4% missed persons as opposed to .4% for the preceding six months. For refusal rates, Bailar (1989) states that in the Survey of Income and Program Participation, a more sensitive interviewer goes out when there is a refusal and about 30% of the refusals are converted into responses. However, Bailar (1989) notes that, while there is little hard data, there is a feeling that the quality of the data gathered from the converted refusals is low.
For example, in a 1960 mail survey, internal checks of the data indicated that as the number of mailings required for a response increased, the quality of the data decreased. For weighting, Bailar (1989) notes weighting has been shown in several studies to be effective in reducing nonresponse bias.

Bailar (1989) includes a section on problems with respondent recall of data. Respondents tend to do a poor job of recalling data. Bailar (1989) also includes a section on time-in-sample bias for panel surveys. Bailar (1989) writes that more research is needed on all of the problems that she covers in her paper.

Murthy (1979) covers nonresponse levels in Asia and the Pacific. At the time the paper was written in 1979, Murthy (1979) wrote that the problem of incomplete data was just beginning to be studied intensively. Murthy (1979) provides a framework for studying the problem and data on the levels of nonresponse in Asia and the Pacific. Murthy (1979) notes that in developing countries the problems with survey data tend to be more in accuracy of the data and inadequate frames than in nonresponse, that completion rates are on the order of >95%.

Murthy (1979) describes the stages of a survey as conception (with specification errors), sample selection (with specification errors), data collection (with ascertainment errors), and data processing and inference (with processing errors). Murthy (1979) distinguishes between the target population, the frame population, the survey population, and the inference population. Murthy (1979) describes the target population as the population about which information is wanted. The target population should take into account the objectives of the survey and, also, operational considerations. The frame population is the "group of all units recorded in or approached through a frame" (p. 3). For the frame population,
the sample surveyor must clarify differences from the target population and make appropriate adjustments. The survey population is all units that could be selected in a given survey scheme. The survey population should coincide with the target population, but differences occur through coverage problems, nonresponse, and missing data. The inference population is "the conceptual population of units for which data would have become available at the tabulation stage after survey and processing under a specified survey scheme, if selected, after adjustments" (p. 4). The inference population differs from the survey population due to editing changes, lost data, and procedures to correct for problems with the survey data. Murthy (1979) writes that appropriate weighting and adjustments can bring the inference population close to the target population. The degree of closeness is a measure of the success of a survey. Murthy (1979) discusses differences between the types of populations and what can be done about the differences.

Murthy (1979) attributes nonresponse to missing data due to oversight, households temporarily absent, not-at-homes, refusals of a hardcore nature, and incapacity to respond. Murthy (1979) discusses how to reduce the different kinds of nonresponse. For the inference population, Murthy (1979) lists nine ways of making adjustments.

For the Asian and Pacific region, Murthy (1979) writes that the rates of nonresponse tend to be low, but that there has been an increase in the rates, particularly for establishment surveys and in urban areas. To correct for nonresponse, substitution in the field and post stratification have often been used. In the 1978 Labour Force Survey of Hong Kong, 95.6% responded, 2.9% could not be contacted, and 1.6% did not respond. Of the 626 missing, 520 were left self-administered forms and 120 of the self-administered forms were returned. In
the 1976 By—census of Hong Kong, 97.6% responded and 2.4% were not contacted. Of the 2529 that were not contacted, 1451 returned self—administered forms. In typical household samples in Malaysia, the nonresponse rate is about 1%, but coverage tends to be bad. In the Socio—economic Sample Survey of Households in Malaysia in 1967–68, the nonresponse rate was 2%, varying from 1% to 3.8% between regions. In Papua New Guinea, the response rates vary from 60% to 100%.

In the Philippines, the nonresponse rate is fairly high. There are problems with incomplete reports and access. In the Integrated Survey of Households in the Philippines, the nonresponse rate was <5% in 7 regions and >5% in 6 regions. In the Household Expenditure Survey (1972–73) in Singapore, for the preliminary study, the nonresponse rate was 12.5% and in the main study 16.0% could not be analyzed. For the World Fertility Survey, for households in the survey, in Fiji, the nonresponse rate was 5.4%, in Malaysia, .2%, in Nepal, 5.2%, but the survey included vacant homes and most of the missing were not—at—homes, in Sri Lanka, .3%, in Thailand, 3.7% overall and 9.1% in Bangkok, but the survey in included vacant homes and most of the missing were not—at—homes. For the World Fertility Study, for eligible respondents in responding households, for Fiji, the nonresponse rate was 2.5%, in Malaysia, .8%, in Nepal, 2.1%, in Sri Lanka, .6%, in Thailand, 4.0% overall and 19.2% in Bangkok. In the Additional Rural Income Survey (1968–69 to 1970—71) in India conducted by the National Council of Applied Economic Research, in the first year the response rate was 98.6%, in the second year, 92.8%, in the third year, 88.5%. Another 8% were rejected at processing, leaving 80.5% available for analysis.

Thomsen and Siring (1979) write that, in Norway response rates had been decreasing in the 10 years before 1979 in surveys conducted by the Central Bureau
of Statistics of Norway. The Central Bureau of Statistics of Norway has studied the reasons why. Thomsen and Siring (1979) list the reasons for nonresponse as operational difficulties, time and cost constraints, lack of cooperation, not tracking down missing respondents, and other reasons.

Thomsen and Siring (1979) write that it is difficult to compare nonresponse rates between surveys since surveys differ with respect to the target population, collection methods, work load, and other factors affecting the nonresponse rate. In a list of political opinion surveys from 1969 to 1977, the nonresponse rate was 9.9% for the Election Survey of 1969, 12.6% for the Municipal Election Survey 1971, 19.2% for the Advisory Referendum on Norway's Accession to the ECE of 1972, 19.4% for the Election Survey of 1973, and 21.6% for the Election Survey of 1977. In the Family Expenditure Surveys from 1967 to 1976, the nonresponse rates were 21.8% in 1967, 28.6% in 1973, with 39% of the nonrespondents listed as refusals, 32.6% in 1974, with 46% of the nonrespondents listed as refusals, 19% as not—at—homes, and 35% as missing for other reasons, 32.3% in 1975, with 42% of the nonrespondents listed as refusals, 22% as not—at—homes, and 36% as missing for other reasons, and 31.0% in 1976, with 44% of the nonrespondents listed as refusals, 25% as not—at—homes, and 31% as missing for other reasons.

For the Survey of Housing Conditions, the nonresponse rate was 9.9% in 1967, with 47% refusals and was 22.9% in 1973, with 55% refusals. In the quarterly Labor Force Surveys from 1972 to 1979, the average nonresponse rate in the first four quarters listed was 7.6% and the average nonresponse rate for the last four quarters listed was 10.2%. Refusals averaged 45.3% of the nonrespondents, not—at—homes averaged 37.8% of the nonrespondents, and missing for other reasons averaged 16.8% for the nonrespondents, over the seven year period.
Thomsen and Siring (1979) note that completion rates showed a serious decline in the 10 years before 1979 for interview surveys conducted by the Central Bureau of Norway and the refusals had accounted for about 50% of the nonresponse independently of the completion rates. Thomsen and Siring (1979) also note that increased efforts had been made to improve response rates in the later years of the 1970's. Without the extra efforts, Thomsen and Siring (1979) believe the nonresponse rates would have been even higher in the 1970's.

Thomsen and Siring (1979) write that the factors affecting the nonresponse rates are the contents of the survey, the data collection methods, and the attitudes among the respondents. The contents of the survey and the data collection methods are partially under the control of the sample surveyor, but the attitudes among respondents can only be indirectly influenced by the sample surveyor. Thomsen and Siring (1979) list selection and training of interviewers, general working conditions for interviewers, use of introduction letters, use of incentives, use of proxy—interviews, public relations, general instructions, respondent burden, and the number of visits per respondent as controllable variables. Variables that are indirectly controllable are qualifications of interviewers, motivation of interviewers, availability of respondents, and motivation of respondents. The dependent variables are total nonresponse rates, refusals, temporarily absents, not—at—homes, and other nonresponse. The Central Bureau of Statistics of Norway found that interviewer age, sex, years experience, and qualifications did not significantly affect the response rates, but that the size of the assignment did. Working conditions were not assessed in the studies. The interviewer work load in Norway is about 200 hours per year and has been stable over time. In two studies, the Central Bureau of Statistics of Norway did not find much effect on nonresponse in relation to the respondent
burden. Thomsen and Siring (1979) write that the best method for reducing nonresponse is to increase the number of call backs. Increasing the number of call backs reduces the number of not—at—homes and, if specially trained interviewers are used for the call backs, refusals are reduced. It is not clear, from the studies at the Central Bureau of Statistics of Norway, if incentives improve the response rates, but incentives do improve the interviewers' attitudes towards their work.

Thomsen and Siring (1979) found that two of the most important factors influencing the nonresponse rates are the attitudes of the public towards the usefulness of the survey and the confidentiality of the information given to the interviewer. In three studies, Thomsen and Siring (1979) found that the distributions of variables, for which there was outside data available, were more like the total population for persons that refused to be interviewed than for persons that were missed for other reasons.

Thomsen and Siring (1979) finish their paper by presenting a model for analyzing data with missing values based on post stratification and on the number of call backs that are made to solicit a response.

2.3. Three approaches to the nonresponse problem

In this section, we give three approaches to the nonresponse problem as given by Platek and Gray (1979), Kalton (1983), and Little and Rubin (1987).

Platek and Gray (1979) write that all surveys have some problems with nonresponse, the size of which depends on the questions asked and the way data is collected. Nonresponse can be controlled by careful controls at both the planning and implementation stage of the survey. After the survey is taken, imputation adjustments can be made in the presence of nonresponse.
Platek and Gray (1979) cover four imputation procedures; making no adjustment, weighting, weighting with historical data substituted for some of the missing observations, and random imputation. The random imputation scheme duplicates the respondents as many times as the respondents will fit into the full sample and fills up the rest of the missing values with respondents chosen by simple random sampling without replacement from the set of respondents.

The imputation is done within balancing classes defined by geographical areas, where a balance is struck between geographical areas that are so large that little homogeneity exists within the balancing class and geographical areas that are so small that some of the balancing classes contain almost no respondents. The imputation may be done within weighting classes if partial response occurs.

Weighting classes are defined on the basis of information available both from the survey and outside the survey. Platek and Gray (1979) do not distinguish between balancing classes and weighting classes.

For a single characteristic, $X$, associated with a single unit in the sample, Platek and Gray (1979) model $X$ with both response error and imputation error. Let $X_i$ be the true value of characteristic $X$ for unit $i$. Let $x_i$ be the observed or imputed value of characteristic $X$ for unit $i$. Then the value of $x_i$ in the imputed data set is

$$x_i = X_i + \delta_i \epsilon_{Ri} + (1-\delta_i) \epsilon_{NRi}, \quad (2.1)$$

where $\epsilon_{Ri}$ is a random error term associated with measurement error and has a possibly nonzero expected value, $\epsilon_{NRi}$ is a random error term associated with imputation error and has a possibly nonzero expected value, and $\delta_i$ is a random
indicator variable that takes on the value 1 if unit \( i \) responds and takes on the value 0 if unit \( i \) does not respond. Then, \( \delta_i \) can have a different expectation for each \( i \). The measurement error and the imputation error are assumed to be independent of each other, but neither is assumed to be independent between units.

Platek and Gray (1979) work with estimators of the population total within balancing or weighting classes. The estimator of the finite population total is the sum over the balancing or weighting classes of the estimators of the population totals within the classes. The variance of the finite population total is then the double sum over the covariances between the population totals. Covariances between different classes may exist due to the sampling scheme.

The estimator that Platek and Gray (1979) propose for the within class total is

\[
\hat{X}_c = \sum_{i=1}^{n_c} \left[ \delta_i w_i x_i + (1-\delta_i) \delta_{Hi} w_{Hi} x_{Hi} \right] \pi_i^{-1},
\]

where \( \delta_i \) is as defined before, \( \delta_{Hi} \), takes on the value of 1 if historical data is used for imputation and takes on the value 0 otherwise, \( w_i \) is the imputation weight on unit \( i \) if unit \( i \) responds, \( w_{Hi} \) is the imputation weight on unit \( i \) if unit \( i \) does not respond and historical data has been used to fill in the missing value, \( x_i \) is the observed value of \( X \) if unit \( i \) responds, \( x_{Hi} \) is the historical record if unit \( i \) does not respond and an historical record has been substituted for the missing value of unit \( i \), \( \pi_i \) is the probability that unit \( i \) is in the sample, and \( n_c \) is the sample size in the class. Thus, Platek and Gray (1979) work with a modified Horvitz–Thompson estimator.
For the method of imputation that makes no adjustments, $w_i$ equals 1 and $\delta_{Hi}$ equals 0. For the weighting adjustment imputation method, $w_i$ equals the sample size divided by the number of respondents and $\delta_{Hi}$ equals 0. For the weighting with historical data substitution adjustment imputation method, $w_i$ and $w_{Hi}$ equal the sample size divided by the sum of the number of respondents and the number of nonrespondents for which historical data has been substituted and $\delta_i$ and $\delta_{Hi}$ equal 1. For the random imputation method of adjustment, $w_i$ equals

$$\left\lfloor \frac{\text{the number of respondents}}{\text{the sample size}} \right\rfloor$$

for

$$\left( \frac{\text{the number of respondents}}{\text{the sample size}} \right) - \left( \frac{\text{the number of respondents}}{\text{the sample size}} \right) + \left\lfloor \frac{\text{the number of respondents}}{\text{the sample size}} \right\rfloor \times \left( \frac{\text{the number of respondents}}{\text{the sample size}} \right)$$

and equals

$$\left\lfloor \frac{\text{the number of respondents}}{\text{the sample size}} \right\rfloor + 1$$

for the rest of the respondents and $\delta_{Hi}$ equals 0.

Platek and Gray (1979) give the expected value and variance of the
estimator of the within class total for each of the imputation schemes. Platek and Gray (1979) note that for the random imputation method, the variance is the variance for the weighting method plus a term due to the random imputation. Platek and Gray (1979) write that the term due to the random imputation has been worked out for simple random sampling without replacement if the number of respondents and the size of the sample are fixed and has been worked out for the case of probability proportional to size without replacement sampling with variable sample size and variable number of respondents. Platek and Gray (1979) do not find the term due to the random imputation for the general case. Also, Platek and Gray (1979) note that their formulas are not useful for actually estimating the population total, since many of the variables in the formula are not known for every member of the sample. Platek and Gray (1979) suggest that information about the unknown variables might be available from past surveys or from special smaller surveys designed to estimate the unknown variables.

Platek and Gray (1979) give the formulas for the covariances of the estimated within class totals between balancing or weighting classes. The covariance formulas are simple extensions of the variance formula.

Platek and Gray (1979) compare biases in the estimators under the different imputation schemes. The imputation scheme that makes no adjustment is the most biased. The imputation scheme that weights the data and the random imputation scheme have the same bias. The weighting scheme with historical data substitution has the smallest bias as long as the historical data is a good match to the missing records. The imputation scheme that makes no adjustment for nonresponse underestimates the population total unless the measurement error is a large enough positive value to compensate for the downward bias due to imputation. The other
imputation methods do not have a definite direction in the bias, either for the measurement error or the imputation error, but the imputation error biases for the other methods are usually much smaller than with the method that does nothing.

Platek and Gray (1979) apply the general formulas to a two stage sampling design and find special case formulas. Platek and Gray (1979) then set up an hypothetical example using the two stage sampling design. Platek and Gray (1979) make assumptions about the error variances, the response mechanism, and the error biases in the hypothetical example. Platek and Gray (1979) found that the nonsampling variances were largest under the method that makes no adjustments to the data and smallest under the weighting with historical data method. The weighting method had smaller nonsampling variance that the random imputation method.

Kalton (1983) distinguishes between three types of nonresponse; noncoverage, unit nonresponse, and item nonresponse. Kalton (1983) writes that the distinction between noncoverage, unit nonresponse, and item nonresponse is the amount of information available about the nonrespondents, with the least amount of information available for noncoverage and the most for item nonresponse. For noncoverage, external sources of information are needed to compensate for the nonresponse. For unit nonresponse, the sampling unit and stratum characteristics are know, as are interviewers observations and the information from the sampling frame. For item nonresponse, considerable information is available from responses on other items. Kalton (1983) suggests post stratification weighting adjustments for noncoverage, post stratification or response rate weighting adjustments for unit nonresponse, and imputation for item nonresponse.

Kalton (1983) notes that most surveys do not adjust for nonresponse. Kalton
(1983) gives biases in estimators of population characteristics if no adjustment for nonresponse is made. Kalton (1983) gives biases for a simple random sampling design under the assumption that the units in the population either always respond or never respond, that is, that response is a population parameter. Kalton (1983) notes that for a single variable there is no distinction between item nonresponse and unit nonresponse. Kalton (1983) refers the reader to Platek and Gray (1979) for a development of a probability model for response and nonresponse.

The estimator of the population mean that Kalton (1983) gives is the mean of the respondents. The bias in the estimator of the population mean is proportional to the nonresponse rate in the population and the difference between the population means of the respondents and the nonrespondents.

Kalton (1983) gives the total of the respondents in the sample inflated by the population size divided by the sample size as the estimator for the population total. The bias in the estimator of the population total is the population total minus the population total for the respondents, and is small only if the population response rate is large or the mean of the missing values is close to zero.

Kalton (1983) studies the sample variance of the respondents as the estimator of the population variance. The bias in the estimator of the population variance is a sum of two terms; the population nonresponse rate times the difference between the population variances for the respondents and the nonrespondents and the population response rate times the population nonresponse rate times the difference between the population means of the nonrespondents and the respondents squared. The population variance is under estimated if the population variance is the same for the respondents and the nonrespondents unless the population mean is the same for the respondents and the nonrespondents.
Kalton (1983) also gives the bias for an estimator of the covariance between two characteristics. The bias can be either negative or positive.

Kalton (1983) discusses possible assumptions that can be made in the presence of nonresponse. If no adjustments have been made to the responding sample, one can make the assumption that the population mean of the respondents is approximately the same as the population mean of the nonrespondents or one can claim that the results from the survey are only valid for the portion of the population that responds.

If one is going to adjust for the nonresponse, Kalton (1983) writes that explicit assumptions about the nonresponse have to be made. Kalton (1983) gives, as two possible assumptions, that the mean of the responding population is the same as the mean of the missing population or that the nonrespondents are missing at random from the population. Kalton (1983) suggests two possible missing at random mechanisms, the first being that the probability of response is constant throughout the sample. The second missing at random mechanism assumes a superpopulation from which the respondents in the population and the nonrespondents in the population are both drawn. Response then remains a population parameter. Kalton (1983) uses the second missing at random assumption. Kalton (1983) writes that the second missing at random assumption is less stringent that the assumption the the responding and nonresponding population means are equal in that one only needs the expected value over the superpopulation of the population means of the respondents and the nonrespondents to be equal. Kalton (1983) writes that the second missing at random assumption is more stringent than the assumption that the responding and nonresponding population means are equal in that the distributions of the respondents and the nonrespondents
Kalton (1983) notes that usually none of the assumptions given above are valid for the entire sample. Kalton (1983) writes that to get approximate validity for any of the assumptions given above, the sample or population must be divided into classes. Within the classes, the assumptions more likely to hold. Under the assumptions, the adjustments equate the respondents with the nonrespondents after controlling for other variables in the nonresponse model (variables known for the sample).

Kalton (1983) gives weighting adjustments and imputation adjustments as two types of adjustments that give reasonable estimators in a wide range of survey applications. Kalton (1983) notes that weighting adjustments and imputation adjustments are often combined and are related. Kalton (1983) writes that it is better to use weighting adjustments if only a little information is available on the nonrespondents. If too much information is lost in forming weighting classes, then it is better to use imputation.

Kalton (1983) gives three criteria for choosing an adjustment scheme. The first criteria is that the adjustment scheme gives good precision for the estimator. The second criteria is that it is possible to estimate the standard error of the estimate under the adjustment scheme. The third criteria is that the adjusted sample is suitable for producing estimates for a variety of different parameters.

Kalton (1983) first looks at the precision of estimators under the equal population means for the respondents and the nonrespondents and the missing at random assumptions for a variety of sampling schemes. Kalton (1983) also looks at the ability to compute standard errors of estimators under the two assumptions under the same sampling schemes as with the overview of the precision of the
estimators. In all cases, under the criteria of precise estimators, weighting is favored over imputation. Kalton (1983) notes that under imputation if the variance estimation is done as if the imputed data set were the real data set, then the variances are underestimated. The variances are underestimated because the sample size is taken to be the intended sample size rather than the observed sample size and because there is an additional inflation in the variance due to the imputation.

Under the assumption that the nonrespondents are missing at random, for simple weighting of a simple random sample, the estimate of the population variance is unbiased for the population variance of the respondents. With weighting classes, the estimate of the population variance is unbiased for the population variance of the respondents for large sample sizes. Under the missing at random assumption, simple random sampling, and random imputation, the estimator of the population variance is approximately unbiased.

For the criteria of general applicability, Kalton (1983) writes that the assumption that the population mean of the respondents is equal to the population mean of the nonrespondents is inadequate for estimating parameters other than the population mean or total.


Kalton (1983) lists the advantages of imputation over weighting as 1) imputation makes analysis easier and simpler to present, 2) imputation gets rid of complex algorithms to estimate population parameters, 3) imputation gives consistency for estimators. The dangers of imputation are 1) imputation does not
guarantee that bias will be reduced, bias reduction depends on the validity of the assumptions, 2) imputation may distort relationships between variables, 3) random subsampling reduces the effective sample size, and 4) in subgroups the number missing may be large.

Kalton (1983) suggests that in the imputed data set, the data analyzer should flag imputed values and keep track of the number of times an item is used as a donor and the number of attempts for a successful imputation into a given missing record.

Kalton (1983) writes that imputation is a good general purpose approach to adjusting for nonresponse. Imputation is good for many purposes, but is not necessarily optimal for any one. In forming classes, categorical variables are easier to work with than continuous variables. With continuous variables, one might want to try a different approach than imputation classes. Kalton (1983) describes deductive imputation, hot deck imputation, random imputation, flexible matching imputation, distance function matching imputation, and regression imputation.

Under random imputation, Kalton (1983) derives a variance estimator under simple random sampling and imputing within imputation classes. The derivation of this estimator is described in Chapter 3 of this paper.


Little and Rubin's (1987) book describes approaches for dealing with missing data in survey samples as well as in other types of samples. Little and Rubin (1987)
describe a sample as a rectangular data set where, in the case of survey sampling, the rows are made up of units in the sample and the columns are the items of interest. Little and Rubin (1987) call data missing if the unit–item is in the sample but no acceptable data was collected for the unit–item. Little and Rubin (1987) indicate that the fact that data are missing sometimes gives information about the nonrespondent.

Little and Rubin (1987) describe four methods for dealing with nonresponse. The first method is to only analyze complete records. If nonresponse is large, then the first method can lead to serious biases. The second method is to use imputation and calculate estimators with standard data analysis routines for a full sample. Under the second method, valid inference can only be made if adjustments are made to the variance estimates for the difference between the real and imputed data. The third method is to use weighting, where all of the design weights in an estimator are replaced by design weights multiplied by the inverses of estimated probabilities of response for the unit–items. The fourth method that Little and Rubin (1987) describe is the method of using model based procedures. For the model based procedures, the mechanism for nonresponse is modeled and estimators are found using methods like maximum likelihood. The variances of estimators can then be found from the information matrix of the model for nonresponse. Under the model based procedures, inference is valid if the model holds. Also, model based procedures are flexible and allow evaluation of the assumptions in the model for the nonresponse.

Little and Rubin (1983) describe different types of patterns that may appear in a rectangular data set, such as a monotone pattern. A monotone pattern occurs in a data set if, letting 1 symbolize response and 0 symbolize nonresponse, the
data matrix can be put in order such that the 1's are followed by the 0's in each column and row.

Little and Rubin (1987) discuss mechanisms for missing data. Some mechanisms are under the control of the sampler. For example, in survey sampling one might consider the entire population excluding the sample as missing data. For the population excluding the sample, only the sample design variables are known. Then the missing mechanism is under the control of the sampler. With nonresponse, the mechanism is not under the control of the sampler.

Little and Rubin (1987) discuss univariate nonresponse, multivariate nonresponse when the nonresponse only occurs in one variable, and multivariate nonresponse.

Little and Rubin (1987 p. 19) define missing at random and missing completely at random. We use Little and Rubin's (1987) definitions throughout this paper. We give the definitions here.

**Definition 2.1:** If missing values are *missing completely at random*, then the respondents are a random subsample of the full sample.

**Definition 2.2:** If missing values are *missing at random*, then the respondents within classes are random subsamples from the intended samples within the respective classes, where the classes are formed using information known for the full sample.

Little and Rubin (1987) discuss differences between the randomization approach to survey sampling, where the only random quantities are in the sampling mechanism,
and the model based approach to survey sampling, where the variables of interest are random as well as the sampling mechanism. Little and Rubin (1987) write that the randomization based approach generally requires that the sampling mechanism is known and that the probability that a given unit in the population is in the sample is greater than zero for all members of the population. Little and Rubin (1987) suggest that the objective of the randomization approach is usually to estimate population quantities using sample quantities. The model based approach estimates the unobserved variables of interest by using information from the sample combined with a model for the variables of interest.

When observations are missing, under the randomization approach the probability distribution on which sample members are observed is unknown. Under the randomization model, two approaches to nonresponse are described by Little and Rubin (1987). The first is to make modeling assumptions about the nonrespondent part of the data. The second is to assume a response mechanism for the missing data.

Little and Rubin (1987) describe two types of weighting adjustments. Little and Rubin (1987) also describe mean imputation, hot deck imputation, substitution imputation, cold deck imputation, regression imputation, stochastic regression imputation, composite methods of imputation, and multiple imputation. Under hot deck imputation (random imputation), Little and Rubin (1987) find the variance of the estimator of the population mean for simple random sampling and the missing units missing completely at random. In Chapter 3, we cover Little and Rubin’s (1987) variance estimation for random imputation.

Little and Rubin (1987) give a short discussion of variance estimation in the presence of nonresponse for the randomization approach to survey sampling.
Under the model based approach to survey sampling, Little and Rubin (1987) discuss the Bayesian approach, ignorable nonresponse, multiple imputation (which is described in this paper in Section 2.4 under Rubin (1987)), and ignorable nonresponse.

2.4. Some solutions posed in the literature for variance estimation with missing values

In this section, we review some of the solutions that have been proposed for estimating the variances of estimators of population characteristics in a finite population when data is missing. We first look at the work of Hansen, Hurwitz, and Madow (1953). We then look at two papers by Kalton and Kish, one from 1981 and one from 1984. Next, we review a paper by Särndal and Swensson (1987). We then look at Rubin's (1987) multiple imputation technique. Finally, we look at papers by Schafer (1990) and Rao and Shao (1992).

Hansen, Hurwitz, and Madow (1953), in their book *Sample Survey Methods and Theory, Vol. 2*, cover the simplest case of random imputation, simple random sampling with one imputation class and the number of missing less than the number responding. Hansen, Hurwitz, and Madow (1953) start with a simple random sample without replacement of $r$ units and find the variance of the estimator of the mean if $m$ of the $r$ units are chosen by simple random sampling and duplicated in the sample. The variance of the estimator for the mean that Hansen, Hurwitz, and Madow (1953) find is correct for a sample with some observations missing and filled in by random imputation if the missing observations are missing completely at random from the sample. Hansen, Hurwitz, and Madow (1953) find that
\begin{equation}
V(\bar{y}^*) = \left[1 - \frac{(r + m)}{N}\right] \frac{S_y^2}{r + m} + \frac{2m}{(r + m)^2} S_y^2,
\end{equation}

where $\bar{y}^*$ is the imputed sample mean, $N$ is the size of the population, and $S_y^2$ is the population variance of the $Y$'s. In terms of the nonresponse problem, $r$ is the number of respondents, $m$ is the number missing, and $r + m$ is the original sample size.

Kalton and Kish (1981, 1984), in two papers, the later one being a revision of the earlier one, present a number of imputation schemes and find the variance of the estimator of the population mean for several imputation methods. The Kalton and Kish (1981, 1984) papers were written out of an interest in finding efficient random imputation schemes, that is, schemes with small increase in the variance of the estimator of the mean of a population, when estimated from the sample of the respondents. Kalton and Kish (1981, 1984) mention both weighting and imputation as methods of adjusting for missing data. Both weighting and imputation give the same biases in the estimator, but imputation from the sample adds a source of sampling variation that Kalton and Kish (1981, 1984) call imputation variance. Kalton and Kish (1981, 1984) mention both total (unit) and item nonresponse. For total nonresponse, weighting is sufficient and one does not have to worry about imputation variance. For item nonresponse, the weighting process is too complex, so imputation is generally done.

classes are based on characteristics that are known for the respondents and the nonrespondents. Sample surveyors form imputation classes because sample surveyors want to use the assumption that the nonrespondents are missing at random within a class and because sample surveyors believe that there is some homogeneity among the characteristics of the units within a class. Kalton and Kish (1981, 1984) write that the assumption of data missing at random within an imputation class is needed, but not generally believed in the survey industry and the homogeneity of the characteristics within a class is generally only relative.

Kalton and Kish (1981, 1984) describe two imputation procedures that are often used, mean value imputation and hot deck imputation. Kalton and Kish (1981, 1984) describe mean value imputation and hot deck imputation in the same manner as Kalton and Kasprzyk (1986) and Platek (1980). If the list of observations is in random order, then hot deck imputation is equivalent to sampling the respondents by simple random sampling with replacement to find donors and to assigning donors to the missing units at random. Mean value imputation produces spikes in the distribution and the element variance of the distribution is artificially reduced with mean value imputation. Kalton and Kish (1981, 1984) give a variance result for mean value imputation and use this variance result as a basis to compare random imputation methods to, since mean value imputation does not result in an imputation variance. In looking at imputation variances, Kalton and Kish (1981, 1984) assume that an equal probability sample of size \( n \) has been taken from a population, with \( r \) of the units responding on item \( Y \) and \( m \) of the units not responding on item \( Y \). Kalton and Kish (1981, 1984) assume that there is only one imputation class and that the population is made up of two strata, the units that will respond and the units that will not respond. Kalton and Kish (1981, 1984)
present their work conditional on $r$ and $m$ being fixed. Let

$$m = n^{-1}m$$  \hspace{1cm} (2.4)$$

and

$$\bar{y}^* = n^{-1}(r \bar{y}_r + m \bar{y}_m)$$, \hspace{1cm} (2.5)$$

where $\bar{y}_r$ is the mean of the respondents and $\bar{y}_m$ is the mean of the imputed values. Then

$$V(\bar{y}^*) = V_1E_2(\bar{y}^*) + E_1V_2(\bar{y}^*)$$, \hspace{1cm} (2.6)$$

where, at level 2, the respondents are held constant and, at level 1, only $r$ and $m$ are held constant. For mean value imputation $\bar{y}_m = \bar{y}_r$. In any scheme in which $\bar{y}_m = \bar{y}_r$,

$$V_2(\bar{y}^*) = 0$$, \hspace{1cm} (2.7)$$

so

$$V(\bar{y}^*) = V_1(\bar{y}_r)$$. \hspace{1cm} (2.8)$$

For any scheme that finds the $m$ values to be used as donors by equal probability sampling
E_2(\bar{y}_m) = \bar{y}_r \quad (2.9)

so

E_2(\bar{y}^*) = \bar{y}_r \quad (2.10)

Also,

V_2(\bar{y}^*) = m^2 V_2(\bar{y}_m) \quad (2.11)

so

V(\bar{y}^*) = V_1(\bar{y}_r) + m^2 E_1 V_2(\bar{y}_m) \quad (2.12)

Kalton and Kish (1981, 1984) define

I_c = [V_1(\bar{y}_r)]^{-1} m^2 E_1 V_2(\bar{y}_m) \quad (2.13)

as the proportional increase in variance for imputation. In order to compare values of \( I_c \) for different imputation schemes, Kalton and Kish (1981, 1984) assume that the original sample is a simple random sample without replacement and that the finite population correction factor can be ignored. Under Kalton and Kish's (1981, 1984) assumptions,

\[ V_1(\bar{y}_r) = r^{-1} S_r^2 \]
The first imputation scheme that Kalton and Kish (1981, 1984) present in the scheme of choosing the donors by simple random sampling with replacement from the respondents. This scheme corresponds to hot deck imputation if the deck is in random order. Kalton and Kish (1981, 1984) find that, for the simple random sampling with replacement scheme,

\[ I_c = \bar{m}(1 - \bar{m}) - \frac{\bar{m}}{n} \geq \bar{m}(1 - \bar{m}) \quad (2.15) \]

and

\[ V(\bar{y}^*) = \frac{S_r^2}{n} + \frac{S_r^2}{(1 - \bar{m})} \bar{m}(2 - \bar{m}(1 + \frac{1}{m})) \quad (2.16) \]

The maximum value for \( \bar{m}(1 - \bar{m}) \) is \( 4^{-1} \) at \( \bar{m} = 2^{-1} \).

The second imputation scheme that Kalton and Kish (1981, 1984) describe is the scheme of choosing the donors by simple random sampling without replacement from the respondents. Let

\[ m = kr + t \quad (2.17) \]

where \( k \) and \( t \) are nonnegative integers and \( t < r \). If \( m > r \), then \( k > 0 \), and the entire set of respondents is replicated \( k \) times for use as donors, while an
additional \( t \) respondents are chosen by simple random sampling without replacement from the respondents to fill in the remaining missing values. For the simple random sampling without replacement scheme, Kalton and Kish (1981, 1984) find that

\[
I_c = [\overline{m}(1 + k) - k][(k + 1) - (k + 2)\overline{m}]
\]  

(2.18)

and

\[
V(\overline{y}^*) = \frac{S_r^2}{n} + \frac{S_r^2}{(1 - \overline{m})n} [(k + 1)^2\overline{m}(2 - \overline{m}) - (k + 1)(\overline{m}^2 + k)]
\]

\[
= \frac{S_r^2}{n} + \frac{S_r^2}{n^2} [k n + (k + 2)t].
\]  

(2.19)

The maximum of \( I_c \), for \( \overline{m} < 1/2 \), is \( 8^{-1} \), when \( \overline{m} = 4^{-1} \). Table 2.1 comes from Kalton and Kish (1981, 1984).

<table>
<thead>
<tr>
<th>( \overline{m} )</th>
<th>SRS w R</th>
<th>SRS w/o R</th>
</tr>
</thead>
<tbody>
<tr>
<td>5%</td>
<td>.0475</td>
<td>.045</td>
</tr>
<tr>
<td>15%</td>
<td>.1275</td>
<td>.105</td>
</tr>
<tr>
<td>25%</td>
<td>.1875</td>
<td>.125</td>
</tr>
<tr>
<td>35%</td>
<td>.2275</td>
<td>.105</td>
</tr>
<tr>
<td>50%</td>
<td>.2500</td>
<td>.000</td>
</tr>
<tr>
<td>65%</td>
<td>.2275</td>
<td>.015</td>
</tr>
<tr>
<td>75%</td>
<td>.1875</td>
<td>.000</td>
</tr>
<tr>
<td>85%</td>
<td>.1275</td>
<td>.005</td>
</tr>
</tbody>
</table>
The third imputation method that Kalton and Kish (1981, 1984) describe is the scheme of choosing donors by proportional stratified sampling from the respondents. If there is more than one imputation class, the proportional stratified sampling is done within the imputation classes. In proportional stratified sampling, the sample is divided into strata based on the measurements for the Y's found during the data collection. The strata are based on the magnitudes of the Y's that responded. All of the respondents are replicated k times. An extra t respondents, chosen using proportionate stratified sampling out of the entire set of respondents, are replicated one more time. The ratio of the conditional imputation variance for proportional stratified sampling to the conditional imputation variance for simple random sampling without replacement is

\[ d = \frac{s^2_{rw}}{s^2_r} \]

where \( s^2_{rw} \) is the average sample within stratum variance of the respondents and where \( s^2_r \) is the simple variance of the respondents.

Kalton and Kish (1981, 1984) describe a proportionate stratified sampling scheme for S strata where \( S^{-1}r \) and \( S^{-1}t \) are integers. Divide the respondents into S equal size strata on the basis of the sizes of the Y values. A simple random sample without replacement of size \( S^{-1}t \) is then taken from the respondents in each strata. The entire sample is replicated k times. Then
d = (r - 1)/(r - 6) [1 - B], \hspace{1cm} (2.21)

where

\[ B = \frac{v(\bar{y}_{rh})}{v(y_{rhi})}, \hspace{1cm} (2.22) \]

\[ v(\bar{y}_{rh}) = \sum_{h=1}^{S} S^{-1} (\bar{y}_{rh} - \bar{y}_r)^2, \hspace{1cm} (2.23) \]

\[ v(y_{rhi}) = \sum_{h=1}^{S} \sum_{i=1}^{r} S^{-1} r^{-1} (y_{rhi} - \bar{y}_r)^2, \hspace{1cm} (2.24) \]

\( \bar{y}_r \) is the mean of the respondents, \( \bar{y}_{rh} \) is the mean of the respondents in stratum \( h \), \( h \) indexes the strata, and \( i \) indexes the individuals within the strata. Kalton and Kish (1981, 1984) call \( B \) the relative explanatory power. Table 2.2 gives values of \( B \) for several distributions and values for \( S \). Table 2.2 comes from the Kalton and Kish (1981, 1984) papers.

Table 2.2. Values of the relative explanatory power

<table>
<thead>
<tr>
<th>( S )</th>
<th>Uniform</th>
<th>Normal</th>
<th>Exponential</th>
<th>Beta(1/2,1/2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>.25</td>
<td>.36</td>
<td>.52</td>
<td>.19</td>
</tr>
<tr>
<td>3</td>
<td>.11</td>
<td>.21</td>
<td>.35</td>
<td>.09</td>
</tr>
<tr>
<td>4</td>
<td>.06</td>
<td>.14</td>
<td>.26</td>
<td>.05</td>
</tr>
<tr>
<td>5</td>
<td>.04</td>
<td>.10</td>
<td>.21</td>
<td>.03</td>
</tr>
<tr>
<td>6</td>
<td>.03</td>
<td>.08</td>
<td>.18</td>
<td>.02</td>
</tr>
<tr>
<td>7</td>
<td>.02</td>
<td>.06</td>
<td>.15</td>
<td>.02</td>
</tr>
</tbody>
</table>
If \( r \) is large, so that \((r - s)^{-1}(r - 1)\) is close to one, it can be seen from Table 2.2 that proportionate stratified sampling can reduce the imputation variance to virtually zero. Kalton and Kish (1981, 1984) note that within imputation classes, \( r \) can be rather small, so that the reduction in the imputation variance might not be so great if imputation classes are used. Kalton and Kish (1981, 1984) note that the sampling scheme can be generalized to the case where \( S^{-1}r \) and \( S^{-1}t \) are not integers.

The fourth imputation method that Kalton and Kish (1981, 1984) describe is the fractional imputation technique. Under the fractional imputation technique, each nonrespondent record becomes \( c \) nonrespondent records with weights \( c^{-1} \), where \( c \) is a positive integer. Then \( cm \) respondents are chosen to fill in the \( cm \) missing records. Kalton and Kish (1981, 1984) mention an alternative method where all of the respondents are replicated \( c \) times and then \( cm \) respondents are chosen by simple random sampling with replacement to be used as donors. Kalton and Kish (1981, 1984) write that the alternative method has a disadvantage in that the deck of variables is larger, but has an advantage in that self-weighting within the sample is maintained. These two adjustment techniques reduce the imputation variance by decreasing the ratio of the weights of the imputed values to the weights of the respondents from \( 2/1 \) to \( (c+1)/c \).

Kalton and Kish (1981, 1984) compare their multiple imputation method to Rubin's multiple imputation method. Kalton and Kish (1981, 1984) write that Rubin has three goals for his multiple imputation method; measuring the total variance of the estimators, assessing the sensitivity of the estimators to the imputation method, and reducing the size of the imputation variance. Kalton and
Kish (1981, 1984) are only interested in reducing the size of the imputation variance.

In order to assess the efficacy of their fractional imputation method, Kalton and Kish (1981, 1984) start with the assumptions that \( r \geq cm \) (which is equivalent to \( \bar{m} \leq (c+1)^{-1} \)) and that \( cm \) donors have been found by simple random sampling without replacement from the respondents. Then

\[
I_c = c^{-1} \bar{m} [1 - \bar{m}(1 + c)] ,
\]

which takes on a minimum of \( [4c(1+c)^{-1}] \) at \( \bar{m} = [2(1+c)]^{-1} \).

For the general case,

\[
cm = ckr + ct .
\]

Let

\[
ct = ar + u ,
\]

where \( a \) and \( u \) are nonnegative integers and \( u < r \). Kalton and Kish (1981, 1984) describe two imputation methods that are equivalent with respect to estimating the population mean. In the first method \( r-u \) respondents donate \( ck+a \) times and \( u \) respondents donate \( ck+a+1 \) times. In the second method, \( kr \) of the missing values are imputed once and \( t \) of the missing values are imputed \( c \) times. Then all of the respondents donate at least \( k+a \) times and \( u \) of the respondents donate \( t \) of the missing values are imputed \( c \) times. Then all of the respondents donate at
least $k+a$ times and $u$ of the respondents donate $k+a+1$ times. With the second method, the size of the file is reduced from $[r(1+ck) + ct]$ to $[r(1+k) + ct]$. For either method,

$$I_c = c^{-2} n^{-2} ru (1 - r^{-1} u).$$  \hspace{1cm} (2.28)

If

$$g = c(1 + k) + a, \hspace{1cm} (2.29)$$

then the maximum value of $I_c$ for a given $g$ occurs when $r^{-1} u = (1+2g)^{-1} g$. The absolute maximum value for $I_c$ occurs when $k=0$, $a=0$, $c=1$, and $m=4^{-1}$.

Table 2.3 gives values for $I_c$ for different values of $c$. If $c=1$, the fractional imputation technique is equivalent to the single random sampling technique given

<table>
<thead>
<tr>
<th>Table 2.3. Values of $I_c$ for fractional imputation</th>
</tr>
</thead>
</table>
| \begin{tabular}{|c|c|c|c|}
| $\bar{m}$ | $c=1$ | $c=2$ | $c=3$ | \\
| 5\% | .045 | .021 | .013 | \\
| 10\% | .080 | .035 | .020 | \\
| 15\% | .105 | .031 | .020 | \\
| 20\% | .120 | .040 | .013 | \\
| 25\% | .125 | .031 | .000 | \\
| 30\% | .120 | .015 | .011 | \\
| 35\% | .105 | .007 | .011 | \\
| 40\% | .080 | .002 | .000 | \\
| 45\% | .045 | .017 | .008 | \\
| \end{tabular} |
as method 2. Table 2.3 comes from Kalton and Kish (1984).

In their conclusion, Kalton and Kish (1981, 1984) note that they have shown that simple random sampling without replacement, proportionate stratified sampling, and multiple imputation have much smaller imputation variances compared to simple random sampling with replacement. The hot deck method of imputation is closest to simple random sampling with replacement, but Kalton and Kish (1981, 1984) note that data sets are not usually in random order and that sequential ordering reduces the imputation variance. Kalton and Kish (1981, 1984) give the qualification that the imputation problem is more difficult if there are more than one variable. In the case where more than one variable is missing, Kalton and Kish (1981, 1984) suggest replacing all missing items for a given unit from one donor.

Särndal and Swensson (1987) present results on two stage sampling, where the sample designs are arbitrary at both stages. Särndal and Swensson (1987) then extend their results to the case of unit nonresponse. In their extension, the first stage sample is the original sample and the second stage sample is the portion of the original sample that responded. Särndal and Swensson (1987) note that for two stage sampling, the probabilities of selection are known for the second stage, but for the case of treating nonresponse as two stage sampling, the probabilities of selection are not known for the second stage. Throughout their paper, Särndal and Swensson (1987) present estimators for the population total under two stage sampling.

Let the first stage sample be denoted by \( s \), where \( s \) is a subset of the finite population. Let \( n_s \) be the size of \( s \), which is not necessarily fixed. Let the sample design be \( P_a(\cdot) \), where \( P_a(s) \) is the probability of choosing \( s \) from the population. Let
\[ \pi_{ak} = \sum_{k \in s} P_a(s) \]  
(2.30)

be the probability of choosing unit \( k \) from the population. Let

\[ \pi_{akl} = \sum_{k, l \in s} P_a(s) \]  
(2.31)

be the probability of choosing both units \( k \) and \( l \) from the population. Define

\[ \pi_{akk} = \pi_{ak} \]  
(2.32)

Assume that \( \pi_{ak} > 0 \) for all \( k \) and \( \pi_{akl} > 0 \) for all \( k \) and \( l \).

Let the second stage sample be denoted by \( r \), where \( r \) is a subset of \( s \).

Let \( m_r \) be the size of \( r \), where \( m_r \) is not necessarily fixed. Let the sample design for the second stage be \( P(.|s) \), where \( P(r|s) \) is the probability of choosing the second stage sample \( r \) from \( s \), given \( s \). Then \( P(.|s) \) can depend on \( s \).

Let

\[ \pi_{k|s} = \sum_{k \in r} P(r|s) \]  
(2.33)

be the probability that element \( k \) in the first stage sample is in the second stage sample, given the first stage sample. Let

\[ \pi_{kl|s} = \sum_{k, l \in r} P(r|s) \]  
(2.34)
be the probability that elements $k$ and 1 in the first stage sample are in the second stage sample, given the first stage sample. Define

$$\pi_{kk|s} = \pi_{k|s} .$$

Assume for any $s$ that $\pi_{k|s} > 0$ for all $k \in s$ and $\pi_{kl|s} > 0$ for all $k, l \in s$.

Särndal and Swensson present four estimators for the population total under arbitrary designs at stage one and stage two. The first of these estimators is what Särndal and Swensson call the expanded sum estimator. Let

$$\hat{t}^* = \sum_{k} \pi_k^{-1} \pi_{k|s} y_k$$

be the expanded sum estimator. Then $\hat{t}^*$ is design unbiased for the population total. The design variance of $\hat{t}^*$ is

$$V(\hat{t}^*) = \sum_{\text{population}} \frac{\pi_{kl} - \pi_{ak} \pi_{al}}{\pi_{ak} \pi_{al}} y_k y_1$$

$$+ E_a \left\{ \sum_s \pi_{kl|s} \pi_{k|s} \pi_{l|s} \pi_{ak} \pi_{al} y_k y_1 \right\} ,$$

where $E_a \{,\}$ is the expected value over the first stage sampling scheme. A design unbiased estimator of $V(\hat{t}^*)$ is

$$\hat{V}(\hat{t}^*) = \sum_{r} \frac{\pi_{kl} - \pi_{ak} \pi_{al}}{\pi_{ak} \pi_{al} \pi_{kl|s} \pi_{ak} \pi_{al}} y_k y_1$$
The expanded sum estimator, \( t^* \), is not the Horvitz–Thompson estimator, since \( \pi_k | s \) and \( \pi_{kl} | s \) are conditional on the sample, and, in general, will not equal the unconditional probability that element \( k \) of the second stage sample is in the second stage sample. Often \( \pi_k | s \) depends on information collected in the first stage sample.

Särndal and Swensson (1987) present two special cases for the second stage sample design. Start with a first stage sample of arbitrary design. After the first stage sample is taken, divide the first stage sample into \( H_s \) strata denoted by \( s_h \), of sizes \( n_h \), \( h=1,...,H_s \). In each of the \( H_s \) strata \( s_h \), \( h=1,...,H_s \), take a subsample, \( r_h \), of size \( m_h \). Then

\[
s = \bigcup_h s_h ,
\]

\[
r = \bigcup_h r_h ,
\]

\[
n_s = \sum_h n_h ,
\]

and

\[
m_r = \sum_r m_h .
\]
The division of the first stage sample into strata is based on information gathered at the first stage.

The two special cases for the second stage sample are 1) the elements of \( r_h \) are drawn by simple random sampling without replacement from the elements of \( S_h \) and 2) the elements of \( r_h \) are drawn by Bernoulli sampling from the elements of \( S_h \). The first case is of general interest and the second case is of interest with respect to the nonresponse problem.

Särndal and Swensson (1987) assume that, in repeated sampling from the population, if the same sample, \( s \), is drawn, then the same strata will be found and will have the same sampling fraction. If a different sample is drawn, then the strata can be different in both number and rationale. For the first case, where the second stage sample is by simple random sampling without replacement within the strata,

\[
\hat{t}^* = \sum_{h=1}^{H} f_h^{-1} \sum_{i} \pi^{-1} y_{ik},
\]

(2.42)

where \( f_h = r_h^{-1} m_h \). Also

\[
V(\hat{t}^*) = \sum_{\text{population}} \sum \pi_{akl} - \pi_{ak} \pi_{al} y_k y_l
\]

\[
+ E_a\left\{ \sum_{h=1}^{H} n_h^2 (1 - f_h) m_h^{-1} s_{\text{shy/}a}^2 \right\},
\]

(2.43)

where \( s_{\text{shy/}a}^2 \) is the sample variance of the \( [\pi_{ak}^{-1} y_k]^s \) in \( s_h \). If \( m_h \geq 2 \) for all \( h \), then
where \( S^2_{rhy/\pi a} \) is the sample variance of the \( \left[ \frac{1}{\pi a y_k} \right] \)'s in \( r_h \). Särndal and Swensson (1987) give an example where the first stage sample is found by simple random sampling without replacement.

For the second case, where the second stage is taken by Bernoulli sampling within the strata with a constant probability of selection, \( \theta_{hs} \), in strata \( h \), \( h=1,\ldots,H_s \), one can assume either that the \( \theta_{hs} \)'s are known or that the \( \theta_{hs} \)'s are not known. Särndal and Swensson (1987) assume the \( \theta_{hs} \)'s are unknown. For Bernoulli sampling, the subsample size, \( m_h \), is random. In order to take into account the random nature of the subsample sizes, Särndal and Swensson (1987) add an extra level of conditioning. At the lowest level of conditioning, Särndal and Swensson (1987) hold \( m = (m_1,\ldots,m_{H_s}) \), the vector of realized counts, constant.

Let

\[
\pi_k | s, m = P(k \epsilon r | s, m). \tag{2.45}
\]

Let

\[
\pi_{kl} | s, m = P(k,l \epsilon r | s, m). \tag{2.46}
\]
Given the sample, \( n = (n_1, \ldots, n_{H_s}) \) is fixed. If \( m \) is also fixed, then the subsample in strata \( h \) is a simple random sample without replacement of \( m_h \) units from a population of size \( n_h \). It follows that \( \pi_{k|h,s,m} = \pi_{kk|h,s,m} = n_h^{-1}m_h = f_h, \quad \pi_{kl|h,s,m} = f_h(n_h-1)^{-1}(m_h-1) \) if \( k \neq l \) and \( k \) and \( l \) are in the same strata, and \( \pi_{kl|h,s,m} = f_h f_g \) if \( k \) and \( l \) are in different strata, where \( h \) and \( g \) denote the strata. Let

\[
A_{1s} = \{ r \text{ such that } m_h \geq 1 \text{ for all } h \}. \tag{2.47}
\]

If \( r \) is an element of \( A_{1s} \) after the second stage sample is taken, then the conditional expanded sum estimator is

\[
\hat{^t}_{c\pi} = \sum_{h=1}^{H_s} f_h^{-1} \sum_{r} \pi^{-1} y_k
\]

and

\[
V(\hat{^t}_{c\pi}) = \sum_{\text{population}} \sum_{\pi} \frac{\pi_{akl} - \pi_{ak} \pi_{al}}{\pi_{ak} \pi_{al}} y_k y_l
\]

\[
+ E_a E_m \left\{ \sum_{h=1}^{H_s} n_h^2 (1-f_h) m_h^{-1} S_{shy/\pi a}^2 \right\}, \tag{2.49}
\]

where \( E_m(.) \) is the expected value over all realizations of \( m \) given the sample and that \( r \) is an element of \( A_{1s} \). Let

\[
A_{2s} = \{ r \text{ such that } m_h \geq 2 \text{ for all } h \}. \tag{2.50}
\]
If the sampled $r$ is an element of $A_2s$, then it is possible to estimate the variance of $t_{c\pi}^*$. The

$$
\hat{V}(t_{c\pi}^*) = \sum_{r} \frac{\pi_{akl} - \pi_{ak} \pi_{al}}{\pi_{kl} | s, m \pi_{ak} \pi_{al} \gamma_k \gamma_l}
$$

$$
+ \sum_{h=1}^{H_s} \frac{n_h^2 (1-f_h)}{m_h} \frac{1}{s_{hy}/\pi_a}. \tag{2.51}
$$

Särndal and Swensson (1987) note that the estimator for the stratified random sampling case and the Bernoulli sampling case are the same, but that the variances are different for the two cases. The estimators are the same because

$$
\pi_{k|s|SRS} = \pi_{k|s|m_{BRS}} \quad \text{and} \quad \pi_{kl|s|SRS} = \pi_{kl|s|m_{BRS}}.
$$

Särndal and Swensson (1987) write that, if the probability that $m_h$ equals zero is close to zero, then the confidence intervals found under Bernoulli sampling will be okay for most cases.

Särndal and Swensson (1987) present three estimators for the case where there are auxiliary variables available for the analysis. Särndal and Swensson (1987) describe three cases; 1) the auxiliary variable, $x_k$ is recorded for all units $k$ in the sample, 2) the auxiliary variable $x_k$ is recorded for all units $k$ in the population, 3) the auxiliary variable $x_k$ is recorded for all units $k$ in the sample and the auxiliary variable $z_k$ is recorded for all units $k$ in the population. Here

$$
x_k = (x_{1k}, \ldots, x_{qk}), \tag{2.52}
$$

and
Särndal and Swensson (1987) start with situations one and two. Särndal and Swensson (1987) assume the existence of a regression relationship between $y_k$ and $x_k$ in the finite population. A regression relationship is assumed on the basis of a scatter plot of the $x_k$'s versus the $y_k$'s in the entire finite population showing what looks like a reasonable fit to a model $\xi$, where

$$E_\xi(y_k) = x_k' \beta, \quad (2.54)$$

$$V_\xi(y_k) = \sigma_k^2, \quad (2.55)$$

and the $y_k$ are independent. The model $\xi$ is a tool to express the relationship between the $x_k$'s and the $y_k$'s in the finite population and is not assumed to be true in any underlying sense. The model is not used for describing and testing the statistical properties of the estimators.

If all of the $N$ points $(y_k, x_k)$ were available then one could find

$$B = \left( \sum_{\text{population}} \sigma_k^{-2} x_k x_k' \right)^{-1} \sigma_k^{-2} \left( \sum_{\text{population}} x_k y_k \right) \quad (2.56)$$

and

$$E_k = y_k - x_k' B \quad (2.57)$$
for all \( k \). Since \( y_k \) is only available for the second stage sample and, in the second stage sample, unit \( k \) has weight \( \pi_{ak}^{-1} \pi_{k|s}^{-1} \), Särndal and Swensson (1987) estimate \( B \) by

\[
b = \left[ \Sigma_r \sigma_k^{-2} \pi_{ak} \pi_{k|s} \right]^{-1} \Sigma_r \sigma_k^{-2} \pi_{ak} \pi_{k|s} y_k y_k.
\]

Särndal and Swensson (1987) define

\[
\hat{y}_k = x_k'b
\]

for \( k \in s \) and

\[
e_k = y_k - \hat{y}_k
\]

for \( k \in r \).

For situation one, Särndal and Swensson (1987) give

\[
\hat{t}_{1\text{reg}} = \Sigma_s \pi_{ak} \hat{y}_k + \Sigma_r \pi_{ak} \pi_{k|s} (y_k - \hat{y}_k)
\]

as an approximately design unbiased estimator of the population total. The approximate variance of \( \hat{t}_{1\text{reg}} \) is

\[
AV(\hat{t}_{1\text{reg}}) = \Sigma_s \Sigma \frac{\pi_{akl} - \pi_{ak} \pi_{al}}{\pi_{ak} \pi_{al}} y_k y_l
\]
Särndal and Swensson (1987) give

\[
\hat{\text{V}}(\hat{t}_{1\text{reg}}) = \Sigma \Sigma \frac{\pi_{kl} - \pi_{ak} \pi_{al}}{\pi_{ak} \pi_{al}} y_k y_l
\]

\[
+ \Sigma \Sigma \frac{\pi_{kl} - \pi_{k} \pi_{l}}{\pi_{kl} \pi_{k} \pi_{l}} e_k e_l
\]

(2.63)

as an estimator of the variance of \(\hat{t}_{1\text{reg}}\).

For the second situation, Särndal and Swensson (1987) give

\[
\hat{t}_{2\text{reg}} = \Sigma \text{population} \hat{y}_k + \Sigma \pi_{ak} \pi_{al}^{-1} \pi_{k}^{-1} (y_k - \hat{y}_k)
\]

(2.64)

as an approximately design unbiased estimator of the population total. Särndal and Swensson (1987) give the approximate variance of \(\hat{t}_{2\text{reg}}\) as

\[
\text{AV}(\hat{t}_{2\text{reg}}) = \Sigma \Sigma \frac{\pi_{akl} - \pi_{ak} \pi_{al}}{\pi_{ak} \pi_{al}} E_k E_l
\]

\[
+ E_a [\Sigma \Sigma \frac{\pi_{kl} - \pi_{k} \pi_{l}}{\pi_{kl} \pi_{k} \pi_{l}} E_k E_l]
\]

(2.65)

and an estimator of the variance of \(\hat{t}_{2\text{reg}}\) as
For situation three, Särndal and Swensson (1987) define some new expressions. Let

\[ B_1 = \left( \sum_{\text{population}} \sigma_{1k}^{-2} z_k z_k' \right)^{-1} \left( \sum_{\text{population}} \sigma_{1k}^{-2} z_k y_k \right), \]  

\[ b_1 = \left( \sum_{s} \sigma_{1k}^{-2} \pi_{ak}^{-1} z_k z_k' \right)^{-1} \left( \sum_{s} \sigma_{1k}^{-2} \pi_{ak}^{-1} z_k y_k \right), \]  

\[ E_{1k} = y_k - z_k' B_1, \]  

\[ e_{1k} = y_k - z_k' b_1, \]  

for \( k \in \mathbb{R} \), and

\[ y_{1k} = z_k' b_{1k}, \]  

where \( \sigma_{1k}^2 = V(y_k) \) under the model that the explanatory variable \( z_k \) predicts \( y_k \).
and where

\[ y_k^* = \hat{y}_k + \pi_{k|s}^{-1} (y_k - \hat{y}_k) \]

for \( k \in \mathbb{R} \), \( Y_K = \hat{Y}_K \)

\[ y_k^* = \hat{y}_k \]

for \( k \in \mathbb{R} \).

Let \( B, b, E_k, e_k \), and \( y_k \) be as defined for situations one and two.

Then Särndal and Swensson (1987) give

\[ \hat{t}_{3reg} = \sum_{\text{population}} \hat{y}_{1k} + \sum_s \pi_{ak}^{-1} (\hat{y}_k - \hat{y}_{1k}) + \sum_{e_k} \pi_{ak}^{-1} \pi_{k|s}^{-1} (y_k - \hat{y}_k) \]

(2.73)

as an approximately design unbiased estimator of the population total. Särndal and Swensson (1987) give the approximate variance of \( \hat{t}_{3reg} \) as

\[ \text{AV}(\hat{t}_{3reg}) = \sum_{\text{population}} \frac{\pi_{ak1} - \pi_{ak} \pi_{al}}{\pi_{ak} \pi_{al}} \text{E}_{1k} \text{E}_{1l} + E_a \left\{ \sum_s \frac{\pi_{k1|s} - \pi_{k|s} \pi_{1|s}}{\pi_{k|s} \pi_{1|s}} \text{E}_k \text{E}_l \right\} \]

(2.74)

and an estimator of the the variance of \( \hat{t}_{3reg} \) as
Särndal and Swensson (1987) then give the results for stratified random sampling and Bernoulli random sampling in the second stage of the sample as a special case.

After developing the results given above, Särndal and Swensson (1987) apply the results to the nonresponse problem. For an arbitrary sampling design, the intended sample can be considered a first stage sample and the respondents in the intended sample as the final sample. Then, using the notation developed above, \( r \) is the response set and \( s-r \) is the nonresponse set. One can postulate that \( r \) is realized from \( s \) by some unknown probability mechanism, \( P(r|s) \). This approach is a more recent approach to the nonresponse problem and is a stochastic approach. The statistician then must make explicit assumptions about \( P(r|s) \).

Särndal and Swensson (1987) write that a widely used technique to reduce bias in the estimator of the population total given the presence of unit nonresponse is the adjustment group technique. The sample is divided into adjustment groups and then a weight of \( m_h^{-1} n_h \) is applied to the respondents within group \( h \) in addition to the sampling weight. The adjustment group technique is based on the assumption that the population is made up of a fixed set of disjoint subpopulations with the probability of response constant within subpopulations. Särndal and Swensson (1987) think that the response mechanism should be looked at with respect to the sample rather than the population.
Särndal and Swensson (1987) give, as a model for nonresponse, that after the intended sample \( s \) is drawn from the population, the sample is divided into \( H_s \) groups, \( s_h, h=1,...,H_s \). Then \( s_h \) is of size \( m_h \) with \( r_h \) respondents and \( m_h \) missing values. It is assumed that the response probability is constant within \( s_h \) for all \( h \). The \( s_h \) are called "response homogeneity groups" and the response mechanism is assumed independent for each unit in a response homogeneity group. Then

\[
P(k \in r \mid s) = \pi_{k\mid s} = \theta_{hs}
\]  

(2.76)

for \( k \in s_h \) and

\[
P(k,l \in r \mid s) = \pi_{kl\mid s} = P(k \in r \mid s) P(l \in r \mid s)
\]  

(2.77)

for \( k \neq l \). Särndal and Swensson (1987) emphasize that the number of groups \( H_s \) and the definition of the groups may change with the sample drawn.

The model that Särndal and Swensson (1987) hypothesize is an exact copy of the Bernoulli case for second stage sampling that Särndal and Swensson (1987) described. The difference between the Bernoulli case for second stage sampling and the nonresponse model is that for the Bernoulli case the probability mechanism at the second stage is known, while for the nonresponse model, the probability mechanism is assumed. By applying the results for the Bernoulli case for two stage sampling, Särndal and Swensson (1987) find that
\[ \hat{t}_{c1} = \sum_{h=1}^{H_h} \left[ \sum_{h=1}^{r_h} \sum_{k=1}^{n_h} \pi_{ak} \Delta y_k \right] + \sum_{h=1}^{r_h} \sum_{k=1}^{n_h} \pi_{ak} (y_k - \hat{y}_k), \]  

(2.79)

\[ \hat{t}_{c2} = \sum \hat{y}_k + \sum_{h=1}^{H_h} \sum_{k=1}^{n_h} \pi_{ak} (y_k - \hat{y}_k), \]  

(2.80)

and

\[ \hat{t}_{c3} = \sum_{k=1}^{n} \hat{y}_{1k} \]

\[ + \sum_{h=1}^{r_h} \sum_{k=1}^{n_h} \pi_{ak} (\hat{y}_k - \hat{y}_{1k}) + \sum_{h=1}^{r_h} \sum_{k=1}^{n_h} \pi_{ak} (y_k - \hat{y}_k). \]  

(2.81)

If the response homogeneity group structure holds, then the four estimators are, at least, approximately unbiased for \( t \).

For any of the \( \hat{t}'s \), the variance estimator of the \( \hat{t} \) can be written

\[ \hat{V}(\hat{t}) = \hat{V}_1(\hat{t}) + \hat{V}_2(\hat{t}). \]  

(2.82)

The \( \hat{V}_1(\hat{t}) \) is the variance due to the randomization involved in choosing the intended sample. The \( \hat{V}_2(\hat{t}) \) is the variance due to nonresponse under the response homogeneity group model. Then, Särndal and Swensson (1987) have shown that
\[ \hat{V}_1(t) = \Sigma \Sigma \frac{\pi_{akl} - \pi_{ak} \pi_{al}}{\pi_{akl} \pi_{kl|s,m} \pi_{ak} \pi_{al} \pi_{k|s,m} \pi_{l|s,m}} a_k a_l \]  
\[ (2.83) \]

and

\[ \hat{V}_2(t) = \Sigma_{h=1}^{H_s} \frac{H_s n_h^2 (1 - f_h) m_h^{-1} S_{rha/\pi a}}{n_h^2 (1 - f_h) m_h^{-1} S_{rha/\pi a}}. \]
\[ (2.84) \]

where \( a \) stands for arbitrary and varies between the estimators. The approximate 100(1-\( \alpha \))% confidence intervals, \( t \pm z_{1-\alpha/2} \hat{V}(t)^{1/2} \), take the nonresponse into account if the response homogeneity group structure holds. Särndal and Swensson (1987) warn that the assumed response model is only an assumption and that the choice of the assumed response model affects the sizes of the point estimators and the confidence intervals.

Särndal and Swensson (1987) give results from a simulation study in which the assumed response model was misspecified. The estimators \( \hat{t}_{c\pi}^* \) and \( \hat{t}_{c1\text{reg}} \) were compared. In the study, repeated simple random samples of size 400 were drawn from a real population of 1227 Swedish households. The samples were each produced using a response homogeneity group model with four groups (household types). Särndal and Swensson (1987) tried three assumed response models on the generated data, the first being the same as the true model, the second using two response homogeneity groups, and the third using one response homogeneity group. Särndal and Swensson (1987) found that for the true model, \( \hat{t}_{c\pi}^* \) and \( \hat{t}_{c1\text{reg}} \) were essentially unbiased, that the variance estimators were essentially unbiased, that the variance of \( \hat{t}_{c1\text{reg}} \) was much smaller than the variance of \( \hat{t}_{c\pi}^* \), and that the
confident intervals looked appropriate. For the false models, Särndal and
Swensson (1987) found that \( \hat{t}_{c\pi}^* \) and \( \hat{t}_{c1\text{reg}} \) were biased, but the bias was
smaller for \( \hat{t}_{c1\text{reg}} \) than for \( \hat{t}_{c\pi}^* \). The confidence intervals were better for \( \hat{t}_{c1\text{reg}} \) than for \( \hat{t}_{c\pi}^* \).

In conclusion, Särndal and Swensson (1987) note that using powerful
auxiliary variables greatly reduces nonresponse bias and helps the variance of the
the estimator involved.

Särndal and Swensson (1987) do not look at item nonresponse or imputation
in their paper. Instead, weighting adjustments for unit nonresponse are carefully
assessed.

Rubin (1987) presents his method of multiple imputation and illustrates the
method with many examples. Rubin (1987) writes that nonresponse is common in
surveys of individual people, households, or businesses. Rubin (1987) writes that if
data are missing from the sample, the estimators of the population parameters
based on the reduced sample are less efficient and possibly biased. Also, standard
complete data methods cannot be used to analyze data with some items missing.
Rubin (1987) notes that the precise reasons for nonresponse are not usually known.

Multiple imputation creates a larger data set by imputing \( m \) values for each
missing value. Every respondent is also replicated \( m \) times. This operation
creates \( m \) sets of data of size \( n \) from the original \( n \) observations. The data sets
can be analyzed by usual methods. An estimate can be computed for each of the \( m \)
subsets.

The objectives of multiple imputation are: 1) to allow the use of complete
data methods, 2) to permit calculation of correct standard errors for estimators, and
3) to increase efficiency relative to single imputation. By repeated use of multiple
imputation it is possible to study the sensitivity of the estimators to the assumptions behind the adjustments that are made to the data.

Rubin (1987) writes that single imputation is the most common form of adjustment in the presence of item nonresponse. Single imputation has the advantages that standard complete-data methods can be used on the imputed data set and that the imputation only has to be done once. Rubin (1987, p. 12) writes that single imputation and standard single data set analysis have the disadvantages that the donor value does not reflect either the sampling variability for a given model of nonresponse or the uncertainty inherent in the choice of the nonresponse model. The variances of the population estimators are underestimated if standard complete-data methods are used to analyze a single imputed data set.

Multiple imputation and the associated analyses, Rubin (1987, p. 15) writes, retains the advantages of single imputation while correcting the disadvantages of single imputation. Also, multiple imputation can allow the survey sampler to measure the sensitivity of the estimators to different models for nonresponse. The disadvantages of multiple imputation are that the resources needed to do the imputation and to store the results are larger than for single imputation. Also, it is necessary to run multiple analyses in order to estimate the variance.

Rubin (1987) presents a model for the nonresponse problem. Rubin (1987) assumes that the indices on the units contain information, also that the covariates contain information for the entire population. Rubin (1987) assumes that the outcome variables (the variables being measured) are measured without error and that the outcome variables are population characteristics. Rubin (1987) defines an inclusion indicator that indicates what members of the population with what outcome variables are included in the sample. Rubin (1987) assumes that the
inclusion indicator is known for every member of the population for each outcome variable. Rubin (1987) also defines a response indicator for the population, that indicates whether each of the outcome variables responds for each of the units in the population. The response indicator is only known for the units in the sample. Rubin (1987) assumes that whether a unit will respond on a given outcome variable is a population characteristic.

Rubin (1987) assumes that the inclusion indicators and the response indicators have a probability distribution on them. Rubin (1987) writes that the probability distributions are necessary for inference. Rubin (1987) defines the sampling mechanism as the probabilities for the inclusion indicators given the covariates, the outcome variables, and the response indicators. The sampling mechanism is said to be unconfounded if the probabilities for the inclusion indicators only depend on the covariates. The sampling mechanism is said to be unconfounded with the response indicators if the probabilities for the inclusion indicators depend only on the covariates and the outcome variables. The sampling mechanism is said to be unconfounded with the outcome variables if the probabilities for the inclusion indicators depend only on the covariates and the response indicators. Rubin (1987) writes that a probability sampling mechanism is any sampling mechanism for which the probability of inclusion on all of the outcome variables for every member of the population is greater than zero.

Rubin (1987) defines the response mechanism as the probabilities for the response indicators given the covariates and the outcome variables. The response mechanism is said to be unconfounded if the probabilities for the response indicators depend only on the covariates. It is not usually assumed that the response mechanism is unconfounded. A response mechanism is a probability response
mechanism if the probability of the response indicators is greater than zero for all of the outcome variables for the entire population.

With the randomization approach to survey sampling, there is no distribution on the covariates or the outcome variables. Rubin (1987) prefers the Bayesian approach in which the covariates and the outcome variables are given a distribution. Rubin (1987) writes that an ignorable sampling mechanism is a sampling mechanism for which the posterior distribution of the unobserved outcome variables does not involve the sampling mechanism. Rubin (1987) writes that the response mechanism is ignorable if and only if the response mechanism given all of the outcome variables in the population is the same as the response mechanism given only the outcome variables in the sample. An ignorable response mechanism is equivalent to the missing at random assumption for the missing values.

We now give Rubin's (1987) basic result. Let $Q$ be the population quantity of interest in the study, where $Q$ is a row vector. Let $\hat{Q}$ be an estimator of $Q$ given a full sample. Let $U$ be the usual variance estimator of $Q - \hat{Q}$ given a fully observed sample. Let $c$ be the number of repeated imputations for each missing value. The quantity $Q$ would be a fixed parameter under a classical model and a vector random variable under the Bayesian model. Let

$$Q_{*1}, ..., Q_{*c}$$

be the $Q$'s for the $c$ imputed data sets. Let the estimators

$$U_{*1}, ..., U_{*c}$$

(2.86)
be the U's for the c imputed data sets. Let

\[ Q_c = \sum_{i=1}^{c} c^{-1} \hat{Q}_{*i}. \] (2.87)

Let

\[ \bar{U}_c = \sum_{i=1}^{c} c^{-1} U_{*i}. \] (2.88)

Let

\[ \hat{B}_c = \sum_{i=1}^{c} (c - 1)^{-1} (\hat{Q}_{*i} - Q_c) (\hat{Q}_{*i} - Q_c). \] (2.89)

Then the estimated total variance of \((Q - \bar{Q}_c)\) is

\[ \hat{V}(Q - \bar{Q}_c) = T_c = \bar{U}_c + (1 + c^{-1}) \hat{B}_c. \] (2.90)

For scalar \(Q\), one compares \((T_c)^{-1/2}(Q - \bar{Q}_c)\) to Student's \(t\) with \(\nu\) degrees of freedom, where

\[ \nu = (c - 1) \left[ 1 + \frac{\bar{U}_c}{(1 + c^{-1}) \hat{B}_c} \right]^2. \] (2.91)

Multiple imputation is more efficient than single imputation because the variance due to random selection of donors is reduced by replication. Rubin (1987) writes that the fraction of information lost due to nonresponse relative to a full sample is
if multiple imputation is done.

Rubin (1987) derives his result based on a Bayesian model for the covariates and the outcome variables. Rubin (1987) also bases his results on the assumption that the assumed response mechanism is correct, with both the analyst of the data and the person doing the imputation agreeing on the response mechanism.

Rubin (1987) shows that, under the Bayesian model, the limit as $c$ goes to infinity of $Q_c$ equals the expected value of $Q$ given the observed data, the sampling mechanism, the covariates, the assumed response mechanism, and the assumed distributions on the covariates and the outcome variables. Also, the limit as $c$ goes to infinity of $U_c$ equals the expected value of $U$ given the observed data, the sampling mechanism, the covariates, the assumed response mechanism, and the assumed distributions on the covariates and the outcome variables. Also, the limit as $c$ goes to infinity of $B_c$ equals the variance of $Q$ given the observed data, the sampling mechanism, the covariates, the assumed response mechanism, and the assumed distributions on the covariates and the outcome variables. Rubin (1987) writes that, under the model, $B_c$ is an unbiased estimator of the limit as $c$ goes to infinity of $B_c$.

Rubin (1987) writes that his results are valid if the imputed values are drawn out of the posterior distribution for the missing outcome variables. The posterior distribution is the distribution of the missing outcome variables given the observed data, the sampling mechanism, the covariates, the assumed response mechanism, and the assumed distributions on the covariates and the outcome variables. Rubin
(1987) also shows that \( T_{c}^{-1/2}(Q - Q_{c}) \) is distributed approximately \( t_{v} \) if 1) the number of imputations is large, so that a central limit theorem for identically and independently distributed posterior means holds or 2) the covariates and the outcome variables have normal distributions.

Rubin (1987) checked his model with randomization—based Monte Carlo studies, using unconfounded sampling mechanisms and assuming the covariates and the outcome variables were fixed and also assuming specific response mechanisms. Rubin (1987) found that if \( Q \) and \( U \) are valid for inference for a complete sample and if the imputation method is proper, where proper imputation methods are defined below, then, for large samples, repeated imputation leads to valid inference for small \( c \), where valid is with respect to the random response randomization—based perspective.

An imputation method is proper if

1) for the covariates, the outcome variables, and the sampling mechanism fixed, given the response mechanism, the conditional distribution of \( Q_{\infty} \) given the covariates, the outcome variables, and the sampling mechanism is \( N(Q, B) \) and the conditional distribution of \( B_{\infty} \) given the covariates, the outcome variables, and the sampling mechanism is \( (B, \epsilon B) \), where

\[
B = V(Q_{\infty}|\text{the covariates, the outcome variables, and the sampling mechanism}),
\]

\[
Q_{\infty} = \lim_{c \to \infty} \tilde{Q}_{c}, \tag{2.93}
\]

and
\[ B_\infty = \lim_{C \to \infty} B_C, \quad (2.94) \]

and the notation \( \ll B \) denotes the fact that the variance of \( B \) is much less than \( B \),

2) for the covariates, the outcome variables, and the sampling mechanism fixed, the conditional distribution of \( \bar{U}_w \) given the covariates, the outcome variables, and the sampling mechanism is \( (U, \ll B) \), where

\[ \bar{U}_w = \lim_{C \to \infty} \bar{U}_C, \quad (2.95) \]

3) for the covariates and the outcome variables fixed, the conditional distribution of \( B \) given the covariates and the outcome variables is \( (B_0, \ll U_0) \),

where

\[ B_0 = E(B|\text{the covariates and the outcome variables}) \quad (2.96) \]

and

\[ U_0 = \text{the population variance of } Q. \quad (2.97) \]

Rubin (1987) notes that for sample random sampling without replacement and random imputation, multiple imputation is not valid, since \( (B_\infty|\text{the covariates, the outcome variables, and the sampling mechanism}) \) is not distributed \( (B, \ll B) \). The underestimation of variance for random imputation comes out of the fact that the missing values are only drawn out of the sample rather than the hypothesized
population distribution. The error in estimating the population parameters is not taken into account.

Rubin (1987) finds values to be used for imputation by first drawing estimated parameters for the model on the outcome variables from the posterior distribution of the parameters and then drawing the values for imputation from the posterior distribution of the missing variables given the drawn, posterior parameters.

Schafer (1990) solves a variance estimation problem posed by the United States Census Bureau's 1990 Post—Enumeration Survey. The parameter that the Census Bureau estimated is a function of sums of weighted binary data. Some of the binary data were missing and were imputed using a logistic regression function on covariates. Schafer (1990) finds an expression for the approximate expectation and variance of the estimated parameters when some form of mean value is used to impute for the missing values. Schafer (1990) applies his results to the Census Bureau estimator.

Schafer (1990) follows a Bayesian approach. Let $T$ be a vector of sums, where each of the sums in $T$ is indexed by $p$ and denoted $T_p$ and where the weights on the binary characteristic vary from sum to sum. Let the estimator of interest be some function of $T$ with two derivatives. Let the missing observations of the binary characteristic be independent Bernoulli random variables. Index the characteristics, and the unknown, but modeled Bernoulli probabilities of the characteristics by $j$. Index the weights by $pj$.

Let the Bernoulli probabilities be modeled as a function of a $k$—dimensional parameter $\theta$, where the probabilities are denoted $\pi_j(\theta)$ and where $\theta$ is also assumed to be a random variable. Assume that the $k$—dimensional parameter, $\theta$,
can be estimated consistently from the covariates and the observed data. Let \( \hat{\theta} \) be the posterior mean of the parameter and let \( \Gamma \) be the posterior variance of the parameter. Fill in the missing values using the model for the Bernoulli random variables with the estimated parameter using the mean value from the model.

Assume that the population size is going to infinity, the sample size is going to infinity, and the ratio of the sample size to the population size is going to zero. Let \( N \) be a constant proportional to the size of the population. Let \( n \) be a constant proportional to the size of the sample. Assume that the number of observed and missing values for the characteristic is \( O(n) \). Assume that the maximum size of the weights is \( O(n^{-1} N) \). Assume that \( n^{-1/2} (\theta - \bar{\theta}) = Z + o_p(1) \), where \( Z \) is a \( k \)-dimensional random variable with zero mean and finite variance.

Given the above definitions and conditions,

\[
E(T_p | \text{the covariates, the observed sample, the imputed values})
= T_p (\text{the covariates, the observed sample, the imputed values}) + E[R_1],
\]

where \( R_1 = O_p(n^{-1}) \), \( E(X | \text{the covariates, the observed sample}) \) is the Bayesian estimator of \( X \) based on the covariates and the observed sample, and \( X(\text{the covariates, the observed sample, the imputed values}) \) is the estimator of \( \theta \) found using the covariates, the observed sample, and the imputed values. Here, \( X \) is the full sample estimator of a population parameter. Also,
\[
\text{Cov}(T_p, T_q \mid \text{the covariates, the observed sample}) = N^{-2} \sum_{j \in \text{missing}} w_j q_j \pi_j(\hat{\theta}) [1 - \pi_j(\hat{\theta})] + N^{-2} D_p(\hat{\theta})^T \Gamma D_q(\hat{\theta}) + E[R_2],
\]

where \( R_2 = O_p(n^{-3/2}) \), \( \pi_j(\hat{\theta}) \) is the imputed value for characteristic \( j \), and \( D_p \) is the sum over the missing values of the weights times the lagrangian of \( \pi_j(\theta) \) with respect to the \( k \) elements of \( \theta \) evaluated at \( \hat{\theta} \). Also,

\[
\text{V}(T_p \mid \text{the covariates, the observed sample}) = N^{-2} \sum_{j \in \text{missing}} w_j^2 \pi_j(\hat{\theta}) [1 - \pi_j(\hat{\theta})] + N^{-2} D_p(\hat{\theta})^T \Gamma D_p(\hat{\theta}) + E[R_3],
\]

where \( R_3 \) in \( O_p(n^{-3/2}) \).

For \( f(T) \),

\[
E[f(T) \mid \text{the covariates, the observed sample}]
\]

\[
= f(T(\text{the covariates, the observed sample, the imputed values})) + E[R_4],
\]

where \( R_4 = O_p(n^{-1}) \), and
\begin{equation}
V[f(T)|\text{the covariates, the observed sample}]
= N^{-2} \sum_{p} \sum_{q} \frac{\partial_i (\hat{T})}{\partial T_p} \frac{\partial f(\hat{T})}{\partial T_q} \sum_{j \in \text{missing}} w_p w_q \pi_j(\hat{\theta}) [1 - \pi_j(\hat{\theta})]
+ N^{-2} \sum_{p} \sum_{q} \frac{\partial_i (\hat{T})}{\partial T_p} \frac{\partial f(\hat{T})}{\partial T_q} D_p(\hat{\theta})^T \Gamma D_q(\hat{\theta}) + E[R_5],
\end{equation}

where \( R_5 = O_p(n^{-3/2}) \) and \( \hat{T} \) stands for \( T(\text{the covariates, the observed sample, the imputed values}) \).

Schafer (1990) writes that for the full sample estimator, \( f(T) \), with full sample variance estimator, \( W \), the posterior variance of \( f(T) \) based on the covariates and the observed sample is

\[ E(W|\text{the covariates, the observed sample}) + V(f(T)|\text{the covariates, the observed sample}). \]

Given the results above, \( f(T) \) can be estimated with small error for large samples by \( f(\hat{T}) \) and the variance of \( f(T) \) can be estimated with small error for large samples by

\[ W(\text{the covariates, the observed sample, the imputed values}). \]
Schafer (1990) notes that

\[ N^{-2} \sum_{p} \sum_{q} \left( \frac{\partial f(\hat{T})}{\partial \hat{T}^p} \frac{\partial f(\hat{T})}{\partial \hat{T}^q} \right) \sum_{j \in \text{missing}} w_p j w_q j \pi_j(\hat{\theta}) [1 - \pi_j(\hat{\theta})] \tag{2.103} \]

is the portion of the variance due to estimating the parameter for the Bernoulli random variable and is very small in relation to the size of the rest of the variance estimator, and so is not included in the estimator.

Schafer (1990) also finds the approximate expectation of the usual variance estimator for a complete sample linear estimator found using a sample with imputed values over a stratified probability proportional to size cluster sample. Schafer (1990) then extends the result to functions with two derivatives of complete sample linear estimators found using a sample with imputed values. Schafer (1990) uses the method of Taylor linearization.

Rao and Shao (1992) discuss the advantages and disadvantages of Rubin and Schenker's (1986) Approximate Bayesian Bootstrap technique and Burns' (1990) stratified multistage survey jackknife technique. Rao and Shao (1992) write that Rubin and Schenker's (1986) Approximate Bayesian Bootstrap technique works well if the imputation does not cross sample clusters. However, if the imputation does cross sample clusters, Rao and Shao (1992) note that Fay (1991) showed that the Approximate Bayesian Bootstrap technique does not always lead to consistent estimators, even with many multiple imputations.
Rao and Shao (1992) write that Burn's (1990) stratified multistage survey jackknife technique gives a larger imputation variance as compared to multiple imputation. Also, Rao and Shao (1992) write that for simple random sampling, the stratified multistage survey jackknife technique can seriously overestimate the variance of the estimator of the imputed total.

Rao and Shao (1992) give consistent variance estimators for functions of imputed population means and totals found using a modified jackknife method. The modification is that, before the jackknife is done, the imputed values are adjusted by a function of the respondents. Rao and Shao (1992) write that the naive jackknife estimator underestimates the variance. Rao and Shao (1992) specify that the random imputation must be a with replacement method, and that for the multistage samples, the first stage must be a with replacement method. Rao and Shao (1992) give formulas for consistent variance estimators of functions of estimators of population means and totals for simple random sampling and for multistage sampling.
3. RANDOM IMPUTATION FOR SIMPLE RANDOM SAMPLING

In this chapter, we look at random imputation when the sampling design is simple random sampling without replacement. The imputation is done by a specific equal probability method, where the values for imputation are chosen out of the set of responding units. This chapter addresses the work of Little and Rubin (1987), and Kalton (1983), and extends that work. In the first part, we look at the case where the number missing is less than the number responding. In the second part, we allow the number missing to be greater than the number responding.

3.1. Missing less than responding

3.1.1. The population and sample Let there be a population of \( N \) units. Let each of the \( N \) units have a characteristic associated with it, denoted by \( Y_i \), \( i=1,\ldots,N \). We define two parameters,

\[
\bar{Y} = N^{-1} \sum_{i=1}^{N} Y_i, \tag{3.1}
\]

the population mean, and,

\[
S^2 = (N - 1)^{-1} \sum_{i=1}^{N} (Y_i - \bar{Y})^2, \tag{3.2}
\]

the population variance.

Suppose a simple random sample without replacement of size \( n \) is taken from
the population of \( N \) units, and that an attempt is made to measure the
characteristic, \( Y_i \), for each sampled unit. Suppose that \( r \) of these sampled units \( (r \leq \ n) \) provide a response for \( Y_i \), and that for \( m \) of these sampled units \( (r+m = \ n) \) we
are unable to get a response for \( Y_i \). Let the sample be denoted by \( \{Y_i\}_{i=1,...,n} \),
where \( Y_1,...,Y_r \) are respondents and \( Y_{r+1},...,Y_n \) are nonrespondents. Let the
augmented sample be denoted by \( \{Y_i^*\}_{i=1,...,n} \), where \( Y_i^* = Y_i \) for \( i = 1,...,r \), and
\( Y_i^* \) is an imputed value for \( i = r+1,...,n \).

Define the quantities,

\[
\bar{y} = \frac{1}{n} \sum_{i=1}^{n} Y_i,
\]

the sample mean,

\[
\bar{y}_r = \frac{1}{r} \sum_{i=1}^{r} Y_i
\]

the mean of the respondents,

\[
\bar{y}_m = \frac{1}{m} \sum_{i=r+1}^{n} Y_i^*
\]

the mean of the imputed values,

\[
\bar{y}^* = \frac{1}{n} \sum_{i=1}^{n} Y_i^*
\]
the mean of the augmented sample,

$$s^2 = (n-1)^{-1} \sum_{i=1}^{n} (Y_i - \bar{Y})^2,$$

(3.7)

the sample variance,

$$s^2_r = (r-1)^{-1} \sum_{i=1}^{r} (Y_i - \bar{Y}_r)^2,$$

(3.8)

the sample variance of the respondents, and

$$s^{*2} = (n-1)^{-1} \sum_{i=1}^{n} (Y_i^* - \bar{Y}^*)^2,$$

(3.9)

the augmented sample variance.

The statistics $\bar{Y}_r$ and $s^2_r$ can be computed directly from the data. The statistics $\bar{Y}^*$, $s^{*2}$ can be computed after the imputation is done. The quantities $\bar{Y}$ and $s^2$ are unobservable if data are missing.

3.1.2. The expectation and variance of the imputed mean under Little and Rubin

Let us now look at the problem of estimating the population mean, $Y$, and the variance of the estimator of the population mean following the approach of Little and Rubin (1987).

First of all, we assume that $r \geq m$, although Little and Rubin (1987) do not make this assumption. In the second section of this chapter we relax this assumption.
An estimate of the population mean, $Y$, is

$$
\bar{y}^* = \frac{\bar{y}_r + m \bar{y}_m}{n}.
$$

(3.10)

Little and Rubin (1987) give the expected value and variance of $\bar{y}^*$ under two assumptions. The first of these assumptions is that the nonresponding units are missing completely at random from the sample, where missing completely at random is defined in Section 2.3. The second assumption is that the imputed values have been found by taking a simple random sample without replacement of size $m$ from the $r$ respondents and the donors have been assigned to the missing values at random.

Let $E(\hat{\tau} | Y, I, R)$ denote the conditional expectation of the statistic $\hat{\tau}$, conditioning on $Y$, the vector of population values for the characteristic, on $I$, the original sampling mechanism for taking the sample from the population, and on $R$, the response mechanism. Following Little and Rubin (1987) we treat $n$, $N$, and $r$ as fixed.

Under the two assumptions, the imputed values are a simple random sample from the respondents, and

$$
E(\bar{y}_m | Y, I, R) = \bar{y}_r.
$$

(3.11)

It follows that

$$
E(\bar{y}^* | Y, I, R) = \left( \frac{r \bar{y}_r + m \bar{y}_m}{n} \right) n^{-1} = \bar{y}_r.
$$

(3.12)
and, since, under the missing completely at random assumption, the respondents are a simple random sample from the population,

\[ E(\bar{y}^* | Y) = E(\bar{y}_r | Y) = Y. \]  

(3.13)

The variance of the estimator (3.10) is

\[ V(\bar{y}^* | Y) = V\{ E(\bar{y}^* | Y, I, R) | Y \} + E\{ V(\bar{y}^* | Y, I, R) | Y \} \]

\[ = V(\bar{y}_r | Y) + E\{ \frac{m^2}{n^2} (1 - \frac{m}{r}) \frac{s_r^2}{m} | Y \} \]

\[ = S^2 \left( \frac{1}{r} - \frac{1}{N} \right) + \frac{m}{n} \left( 1 - \frac{m}{r} \right) \frac{s_r^2}{n} \]

\[ = S^2 \left( \frac{1}{n} - \frac{1}{N} + \frac{2m}{n^2} \right), \]  

(3.14)

where \( S^2 \) is defined in (3.2). If we ignore the finite population correction factor, we get

\[ V(\bar{y}^* | Y) = S^2 \left( \frac{1}{n} + \frac{2m}{n^2} \right). \]  

(3.15)

3.1.3. The expectation and variance of the imputed sample mean under Kalton

Kalton (1983) follows a different approach in deriving the properties of estimator (3.10). First, Kalton (1983) assumes that whether a unit responds or not is a population parameter. That is, under repeated sampling a unit either will
always respond or will never respond. Under this assumption, Kalton (1983) defines the following population parameters,

\[ Y_R = \text{population mean of the respondents} \quad (3.16) \]

\[ R = \text{the number of respondents in the population} \quad (3.17) \]

\[ S^2_R = \text{the population variance of the respondents.} \quad (3.18) \]

To do random imputation, Kalton (1983) divides the sample into \( L \) imputation classes (indexed by \( h \)). In what follows we will subscript our parameters and statistics by \( h \) when referring to characteristics of an imputation class. Kalton (1983) discusses the assumption that the population mean is the same for the responding and missing groups within each class and compares that assumption to the assumption that the missing units are missing at random within each class. Kalton (1983) discusses the advantages and disadvantages of approaches based on the two assumptions and concludes that usually neither approach is adequate unless the data are divided into imputation classes. In Kalton's (1983) derivation of the estimator's mean and variance, Kalton (1983) never uses either assumption. Instead Kalton (1983) subscripts all of the variables and parameters with \( R \) to indicate that the results are a function of the respondents. When Kalton (1983) evaluates the biases of the estimators, Kalton (1983) brings in the underlying assumptions to show decreased biases for the particular technique he is describing.

Kalton's (1983) random imputation technique is to create imputed values for the missing units by doing simple random sampling without replacement within
each imputation class. Here the respondents within each class are sampled and the sampled values become the imputed values for nonrespondents in that class.

Kalton's (1983) estimator of $\bar{Y}$ is

$$\bar{y}^* = \sum_{h=1}^{L} \frac{n_h}{n} \left( r_h \bar{y}_{rh} + m_h \bar{y}_{mh} \right) \frac{1}{n_h}, \quad (3.19)$$

where the $h$ refers to the imputation class and the other the definitions are the same as in equation (3.10). Thus $\bar{y}_{mh}$ is the mean of the imputed values in imputation class $h$. This estimator is the extension of estimator (3.10) from one imputation class to $L$ imputation classes. Kalton (1983) uses four levels of conditioning. The lowest level, denoted by level 4, holds the sample of respondents, $y_r$, the number of respondents in each class, $r_h$, and the sample size in each class, $n_h$, constant. Level 3 holds the number of respondents in each class and the sample size in each class constant. Level 2 holds the sample size in each class constant. At level 1 nothing is held constant. Using these levels of conditioning and Kalton's (1983) assumption of a fixed set of respondents,

$$\mathbb{E}(\bar{y}^*) = \mathbb{E}_1\{ \mathbb{E}_4(\bar{y}^* | y_r, n_h's, r_h's) \}$$

$$= \mathbb{E}_1\{ \sum_{h=1}^{L} \frac{n_h}{n} \bar{y}_{rh} \} = \sum_{h=1}^{L} \frac{n_h}{N} \bar{Y}_{Rh}, \quad (3.20)$$

where $\bar{Y}_{Rh}$ is the population mean of respondents in imputation class $h$. Let

$$\bar{Y}_s = \sum_{h=1}^{L} \frac{N_h}{N} \bar{Y}_{Rh}. \quad (3.21)$$
The variance of $\bar{y}^*$ can be found using the conditioning,

$$V(\bar{y}^*) = E_1 V_2 (\bar{y}^*) + V_1 E_2 (\bar{y}^*)$$

$$= E_1 [E_2 V_3 (\bar{y}^*) + V_2 E_3 (\bar{y}^*)] + V_1 E_2 E_3 (\bar{y}^*)$$

$$= E_1 E_2 [E_3 V_4 (\bar{y}^*) + V_3 E_4 (\bar{y}^*)]$$

$$+ E_1 V_2 E_3 E_4 (\bar{y}^*) + V_1 E_2 E_3 E_4 (\bar{y}^*)$$

$$= E_1 E_2 E_3 V_4 (\bar{y}^*) + E_1 E_2 V_3 E_4 (\bar{y}^*)$$

$$+ E_1 V_2 E_3 E_4 (\bar{y}^*) + V_1 E_2 E_3 E_4 (\bar{y}^*)$$

$$= (3.22)$$

Looking at the individual components of expression (3.22), we have for the first term

$$V_4(\bar{y}^* | y_r, n_h's, r_h's) = \frac{L}{\sum_{h=1}} m_h \left( 1 - \frac{m_h}{r_h} \right) s_{rh}^2 n^{-2},$$

(3.23)

and,

$$E_3[V_4(\bar{y}^* | y_r, n_h's, r_h's) | n_h's, r_h's]$$
Also,

\[
V_3[ E_4 \left( \bar{y}^* \mid y_r, n_{h's}, r_{h's} \right) \mid n_{h's}, r_{h's} ]
\]

\[= V_3[ \frac{L}{n} \frac{n_h}{n} \bar{y}_r \mid n_{h's}, r_{h's} ]\]

\[= \frac{L}{n} \frac{n_h^2}{n^2} \frac{S_{Rh}^2}{n} \frac{r_h^{-1}}{r_h^{-1}} (1 - \frac{r_h}{R_h}). \quad (3.25)\]

Now,

\[
E_3[ E_4 \left( \bar{y}^* \mid y_r, n_{h's}, r_{h's} \right) \mid n_{h's}, r_{h's} ]
\]

\[= \frac{L}{n} \frac{n_h}{n} V_{Rh}. \quad (3.26)\]

Since all of the quantities in equation (3.26) are constants with respect to level two conditioning,

\[
V_2 \{ E_3[ E_4 \left( \bar{y}^* \mid y_r, n_{h's}, r_{h's} \right) \mid n_{h's}, r_{h's} ] \mid n_h \}
\]
Next,

\[ E_2 \left\{ E_3 \left\{ E_4 \left( \bar{y}^* \mid y_x, n_h's, r_h's \right) \mid n_h's, r_h's \right\} \mid n_h's \right\} = \sum_{h=1}^{L} \frac{n_h}{n} Y_{Rh}, \]  

so,

\[ V_1 \left[ E_2 \left\{ E_3 \left\{ E_4 \left( \bar{y}^* \mid y_x, n_h's, r_h's \right) \mid n_h's, r_h's \right\} \mid n_h \right\} \right] = V_1 \left( \sum_{h=1}^{L} \frac{n_h}{n} Y_{Rh} \right) \]

\[ = \frac{1}{n^2} \left[ \sum_{h=1}^{L} Y_{Rh}^2 V(n_h) \right. \]

\[ + \sum_{h=1}^{L} \sum_{g=1}^{L} Y_{Rh} Y_{Rg} \text{Cov}(n_h, n_g) \]  

For a hypergeometric distribution with variables \( n_h, h=1,...,L \) and parameters \( N_h, h=1,...,L \), \( n \), and \( N \),
Kalton (1983) ignores the finite population correction factor \((1 - n N^{-1})\) and divides \(N_h\) by \(N\) rather than \(N-1\). Kalton (1983) presented expressions (3.23), (3.25) (without the fpc factor), (3.26), (3.27), (3.28), and (3.32) (without the fpc factor). Combining equations (3.24), (3.25) and (3.32) we get,

\[
V(\bar{y}^*) = \frac{L}{\sum_{h=1}^{L}} \left\{ E_1 E_2 \left[ \frac{1}{n^2} \left( \frac{\sum y_h}{\sum_{h=1}^{L}} - \frac{\sum_{h=1}^{L} y_h}{\sum_{h=1}^{L} n_h} \right) + \frac{m_h}{r_{h-1}} \left( 1 - \frac{m_h}{r_h} \right) \right] S_{R_h}^2 \right\} \\
+ \frac{N_h}{N-1} \left( \bar{Y}_{R_h} - \bar{Y}_s \right)^2 n^{-1} \left( 1 - \frac{n}{N} \right)
\]

so,

\[
V(\bar{y}^*) = \sum_{h=1}^{L} \left\{ E_1 E_2 \left[ \frac{1}{n^2} \left( \frac{\sum y_h}{\sum_{h=1}^{L}} - \frac{\sum_{h=1}^{L} y_h}{\sum_{h=1}^{L} n_h} \right) + \frac{m_h}{r_{h-1}} \left( 1 - \frac{m_h}{r_h} \right) \right] S_{R_h}^2 \right\} \\
+ \frac{N_h}{N-1} \left( \bar{Y}_{R_h} - \bar{Y}_s \right)^2 n^{-1} \left( 1 - \frac{n}{N} \right)
\]
Kalton (1983) also gives

\[ \hat{V}(\bar{y}^*) = \sum_{h=1}^{L} \left[ \frac{n_h^2}{n^2} s_{rh}^2 \frac{1}{r_h} + \frac{n_h}{n^2} (\bar{y}_{rh} - \bar{y}_s)^2 \right. \]

\[ \left. + m_h \left( 1 - \frac{m_h}{r_h} \right) \frac{s_{rh}^2}{n^2} \right], \]  

(3.34)

where \( \bar{y}_s = \sum_{h=1}^{L} \frac{n_h}{n} \bar{y}_{rh} \), as an estimator for the variance (3.33) (without the corrections for the finiteness of the population).

3.1.4. Kalton’s expression for the variance of the imputed mean under the MAR assumption

We consider expression (3.33) under the alternative assumption that data are missing at random within imputation classes. Under this assumption,
\[ E(\bar{y}_{rh}) = Y_h, \quad (3.35) \]

so,

\[ E(\bar{y}^*) = \sum_{h=1}^{L} \frac{N_h}{N} Y_h = Y, \quad (3.36) \]

and,

\[ V(\bar{y}_{rh}) = \frac{S_h^2}{r_h} \left( 1 - \frac{r_h}{N_h} \right). \quad (3.37) \]

It follows that the unconditional variance from expression (3.33) becomes

\[
V(\bar{y}^*) = \sum_{h=1}^{L} \left[ \left( \frac{N_h}{N} - \frac{1}{n} \right) - \frac{E(n_{h}^2)}{n^2 N_h} \right] S_h^2 \\
+ \frac{N_h}{N-1} \left( \bar{Y}_h - \bar{Y} \right)^2 \left( 1 - \frac{n}{N} \right) \\
+ \sum_{h=1}^{L} \frac{2 E(m_h)}{n^2} S_h^2. \quad (3.38) 
\]

Since \( n_h \) has a hypergeometric distribution with variables \( n_h, h=1,...,L \) and parameters \( N_h, h=1,...,L, n, \) and \( N, \)

\[ E(n_{h}^2) = \frac{n N_h}{N(N-1)} \left[ N - n + N_h(n - 1) \right]. \quad (3.39) \]
Substituting expression (3.39) into expression (3.38), we get

\[
V(\bar{y}^*) = \frac{1}{n} \left[ \sum_{h=1}^{L} \frac{N_h-1}{N-1} S_h^2 + \frac{N_h}{N-1} (Y_h - \bar{Y})^2 \right] \left( 1 - \frac{n}{N} \right) \\
+ \sum_{h=1}^{L} \frac{2}{n^2} \frac{E(m_h)}{S_h^2} \\
= \frac{S^2}{n} \left( 1 - \frac{n}{N} \right) + \sum_{h=1}^{L} \frac{2}{n^2} \frac{E(m_h)}{S_h^2} .
\] (3.40)

If we ignore the finiteness of the population, expression (3.40) is

\[
V(\bar{y}^*) = \frac{S^2}{n} + \sum_{h=1}^{L} \frac{2}{n^2} \frac{E(m_h)}{S_h^2} .
\] (3.41)

### 3.1.5. The expectation and variance of the imputed mean following our approach

We assume that within imputation classes the missing units are missing at random. We derive the mean and variance of \( \bar{y}^* \), as defined in equation (3.19), under alternative assumptions than those used by Kalton (1983). We can write equation (3.19) as

\[
\bar{y}^* = \sum_{h=1}^{L} \gamma_{hi} \left( 1 + \delta_{hi} \right) Y_{hi}
\] (3.42)

where \( \gamma_{hi} \) is a 0–1 random variable that takes on the value 1 if unit \( hi \) responds and
takes on the value 0 if unit $hi$ does not respond, and $E(\gamma_{hi} | n_h, r_h) = r_h n_h^{-1}$. Also, $\delta_{hi}$ is a 0–1 random variable that takes on the value 1 if respondent $hi$ is used for imputation and takes on the value 0 if respondent $hi$ is not used for imputation, with $E(\delta_{hi} | \gamma_{hi} = 1, n_h, r_h) = m_h r_h^{-1}$ and $\sum_{i=1}^{n_h} \delta_{hi} = m_h$.

We also have $P(\delta_{hi} = 1 | \gamma_{hi} = 0) = 0$. When we write the estimator of the population mean as in (3.42), we include the missing at random response mechanism as part of the estimator through the variable $\gamma_{hi}$. Using (3.42) we derive the variance of $\bar{y}^*$ in a manner different from that of Kalton (1983) and in a manner different from that of Little and Rubin (1987). To find the expectation and variance, we condition on the sample, $y = \{ Y_i \}_{i=1}^{n}$, the sample size in each imputation class, $n_h$, and the number of respondents in each imputation class, $r_h$.

We have

$$E(\bar{y}^*) = E\{ E(\bar{y}^* | y, n_h, r_h) \}$$

$$= E\left\{ \sum_{h=1}^{L} \sum_{i=1}^{n_h} \left( \frac{r_h}{n_h} \right) (1 + \frac{m_h}{r_h}) \frac{Y_{hi}}{n} \right\}$$

$$= E\left\{ \sum_{h=1}^{L} \sum_{i=1}^{n_h} \frac{Y_{hi}}{n} \right\} = E\{ \bar{y} \} = \bar{y},$$

where the notation $E(\bar{y}^* | y, n_h, r_h)$ means the expectation is taken with the sample sizes, the numbers of respondents, and the elements of the samples in the imputation classes fixed. The
\[ V(\bar{y}^*) = E\{ V(\bar{y}^* | y, n_h's, r_h's) \} \]

\[ + \ V\{ E(\bar{y}^* | y, n_h's, r_h's) \}. \]  \hspace{1cm} (3.44)

Let us look at the conditional variance in the first term in (3.44),

\[ V(\bar{y}^* | y, n_h's, r_h's) = E\{ \left( \sum_{h=1}^{L} \sum_{i=1}^{n_h} \frac{\gamma_{hi} (1 + \delta_{hi})}{n} Y_{hi} \right)^2 - \bar{y}^2 \} \]

\[ | y, n_h's, r_h's \} \]

\[ = E\{ \left( \sum_{h=1}^{L} \sum_{i=1}^{n_h} \sum_{g=1}^{n_g} \sum_{j=1}^{g} \frac{\gamma_{hi} (1 + \delta_{hi}) \gamma_{gj} (1 + \delta_{gj})}{n^2} Y_{hi} Y_{gj} \right)^2 \}

\[ - 2 \left( \sum_{h=1}^{L} \sum_{i=1}^{n_h} \frac{\gamma_{hi} (1 + \delta_{hi})}{n} Y_{hi} \bar{y} \right) \]

\[ + [\bar{y}^2] | n_h's, r_h's, y \}

\[ = [\bar{y}^2 + \sum_{h=1}^{L} \frac{2 m_h}{n^2} s_h^2] - 2 [\bar{y}^2] + [y^2] \]
\[
E(h_i (1 + \delta_{hi}) | y, n_{h's}, r_{h's}) = \frac{r_h}{n_h} + \frac{m_h}{n_h} = 1 \tag{3.46}
\]

and

\[
E(h_i (1 + \delta_{hi}) \gamma_{gj} (1 + \delta_{gj}) | y, n_{h's}, r_{h's}) = 1
\]

if \( h \neq g \)

\[
= E(h_i (1 + \delta_{hi}) | y, n_{h's}, r_{h's}) E(h_j (1 + \delta_{hj}) | y, n_{h's}, r_{h's})
\]

\[
= \frac{r_h}{n_h} \frac{r_h - 1}{n_h - 1} (1 + 2 \frac{m_h}{r_h} + \frac{m_h}{r_h} \frac{m_h - 1}{r_h - 1})
\]

if \( h = g \) and \( i \neq j \)

\[
= E(h_i^2 (1 + \delta_{hi})^2 | y, n_{h's}, r_{h's})
\]
\[
\begin{align*}
&= E( \gamma_{hi} (1 + 3 \delta_{hi}) | y, n_h \delta_i, r_h \delta_i ) \\
&= \frac{r_h}{n_h} (1 + 3 \frac{m_h}{r_h}) . \quad (3.47)
\end{align*}
\]

if \( h = g \) and \( i = j \).

So,

\[
V(\bar{y}^*) = E\left( \sum_{h=1}^{L} \frac{2 m_h}{n^2} s_h^2 \right) + V(\bar{y})
\]

\[
= \sum_{h=1}^{L} \left( \frac{2 E(m_h)}{n^2} s_h^2 + \frac{S^2}{n} \right) . \quad (3.48)
\]

We ignore the finite population correction factor, \( (1 - n N^{-1}) \), in the variance of \( \bar{y} \).

3.1.6. Estimation of the variance of the imputed mean

We propose two estimators of the variance of \( \bar{y}^* \) as given in equation (3.48). The first of these is

\[
\hat{V}_1(\bar{y}^*) = \sum_{h=1}^{L} \left( \frac{1}{n^2} + \frac{1}{n^2(n-1)} \right) 2 m_h \frac{s^2_{rh}}{n}
\]

\[+ \frac{s^*}{n} , \quad (3.49)\]

where \( s^* \) is the sample variance for the entire sample using both the imputed and the actual values for the data points and \( s^2_{rh} \) is the sample variance for imputation.
class h using only the values for the units that responded. The multiplier 
\( n^{-2(n-1)^{-1}} \) is added to the first term to correct for the bias in \( s^{*^2} \).

Under our assumption that the missing units are missing at random, \( s_{rh}^2 \) is unbiased for \( S_h^2 \). As was mentioned above, \( s^{*^2} \) is biased for \( S^2 \). Let us find this bias. We can write

\[
s^{*^2} = (n-1)^{-1} \sum_{h=1}^{L} \sum_{i=1}^{n_h} (Y_{hi}^* - \bar{Y}^*)^2
\]

\[
= \frac{1}{n-1} \sum_{h=1}^{L} \sum_{i=1}^{n_h} (Y_{hi}^* - \bar{Y})^2 - \frac{n}{n-1} (\bar{Y}^* - \bar{Y})^2. \tag{3.50}
\]

Now, for all units \( h = \) in the imputed sample, \( Y_{hi}^* \) is a randomly drawn observation from the population, so \( E(Y_{hi}^* - \bar{Y})^2 = (N-1) N^{-1} S^2 \) for all \( h \). Also, \( \bar{Y}^* = \bar{y}^* \). So, the expected value of \( s^{*^2} \) is

\[
E(s^{*^2}) = \frac{N-1}{N} \frac{n}{n-1} S^2 - \frac{n}{n-1} E\{ E(\bar{Y}^*^2 | y, n_h \text{'s}, r_h \text{'s}) \} - 2 E(\bar{y}^* | y, n_h \text{'s}, r_h \text{'s}) \bar{Y} + \bar{Y}^2 \}
\]

\[
= \frac{N-1}{N} \frac{n}{n-1} S^2
\]

\[
- \frac{n}{n-1} E\{ \bar{Y}^2 + \sum_{h=1}^{L} \frac{2 m_h}{n^2} s_h^2 \} - 2 [ \bar{y} \bar{Y} ] + [ \bar{Y}^2 ] \}
\]

\[
= [ \frac{N-1}{N} \frac{n}{n-1} - (\frac{1}{n} - \frac{1}{N}) \frac{n}{n-1} ] S^2 - \sum_{h=1}^{L} \frac{E(2 m_h)}{n (n-1)} S_h^2
\]
Therefore an unbiased estimator of $S^2$ is

$$s^*^2 + \frac{1}{n} \sum_{h=1}^{L} \frac{m_h}{n(n-1)} s_{rh}^2.$$  

(3.52)

Also

$$\frac{L}{\Sigma} \frac{2}{n^2} E(\frac{m_h}{n}) S_h^2 = E[\frac{L}{\Sigma} \frac{2}{n^2} m_h s_{rh}^2].$$  

(3.53)

It follows that $\hat{V}_1(\bar{y}^*)$ of (3.49) is an unbiased estimator of $V(\bar{y}^*)$ (ignoring the finite population correction factor).

The second estimator of the variance of $\bar{y}^*$ is

$$\hat{V}_2(\bar{y}^*) = \frac{L}{\Sigma} \sum_{h=1}^{n} \left( \frac{1}{n^2} + \frac{1}{n^2(n-1)} \right) \xi_{hi} (Y_{hi} - \bar{y}_r)^2$$

$$+ \frac{s^*^2}{n}.$$  

(3.54)

where

$$\xi_{hi} = 0$$  

if unit hi is a respondent
\[ \xi_{hi} = 2 \frac{r_h}{(r_h - 1)} \] if unit hi is imputed. \hfill (3.55)

Now \( \{ Y_{hi}^* \}_{i=r_h+1, \ldots, n_h} \) is a simple random sample without replacement from the respondents in class h, so

\[ E[( Y_{hi}^* - \bar{y}_{rh} )^2 | y, n_h's, r_h's ] = \frac{r_h - 1}{r_h} s_{rh}^2 \] \hfill (3.56)

if unit hi is imputed. Then,

\[ E[ \xi_{hi} ( Y_{hi}^* - \bar{y}_{rh} )^2 ] = 2 S_h^2 \] if unit hi is imputed \hfill (3.57)

\[ E[ \xi_{hi} ( Y_{hi}^* - \bar{y}_{rh} )^2 ] = 0 \] if unit hi is a respondent.

Therefore (3.54) is also an unbiased estimator for the \( V( \bar{y}^* ) \) as given in (3.48). Expression (3.54) has an advantage over (3.49) in that it is easier computationally, but expression (3.54) also has a larger variance than expression (3.49).

### 3.1.7 A comparison of our estimators to Kalton's estimator

We now compare our estimators (3.49) and (3.54) to Kalton's (1983) estimator (3.34). Kalton (1983) uses a different set of assumptions that we do. Kalton's (1983) estimator estimates a different parameter that ours does. To do this comparison, we superimpose our assumption of missing at random data on Kalton's (1983) estimator so that his sample statistics estimate the same parameters as ours. Under
this condition $E(s_{rh}^2) = S_h^2$, $E(\bar{y}_{rh}) = \bar{y}_h$, and $E(\bar{y}_s) = \bar{y}$, where $\bar{y}_s$ is defined after expression (3.34). Also, the variance of $\bar{y}^*$ is given by expression (3.41), which is the same as the variance of $\bar{y}$ that we found in expression (3.48). Kalton (1983) estimates $\Sigma_{h=1}^L 2 E(m_h) n^{-2} S_h^2$ by $\Sigma_{h=1}^L 2 m_h n^{-2} s_{rh}^2$. This is the same estimator that we use in expression (3.49). The estimator is unbiased under the assumption of missing at random data. In expression (3.54), we use an estimator for $\Sigma_{h=1}^L 2 E(m_h) n^{-2} S_h^2$ that is also unbiased, but that has a larger variance. Kalton (1983) estimates $S^2$ using

$$S^2 = \Sigma_{h=1}^L \frac{n_h}{n} \left[ s_{rh}^2 + (\bar{y}_{rh} - \bar{y}_s)^2 \right].$$  

(3.58)

In an infinite population split into imputation classes with overall mean $\mu$ and variance $\sigma^2$ and class means and variances $\mu_h$ and $\sigma_h^2$, the overall variance is

$$\sigma^2 = \Sigma_{h=1}^L p_h \left( \sigma_h^2 + (\mu_h - \mu)^2 \right),$$  

(3.59)

where $p_h$ is the probability of a unit falling into class $h$. In a finite population,

$$S^2 = \Sigma_{h=1}^L \left\{ \frac{N_h - 1}{N - 1} S_h^2 + \frac{N_h}{N - 1} (\bar{Y}_h - \bar{Y})^2 \right\}.$$  

(3.60)

If we ignore the finiteness of the population, then $\Sigma_{h=1}^L n_h n^{-1} s_{rh}^2$ is unbiased for the first term in (3.60). However,
104

\[
E[ \sum_{h=1}^{L} \frac{n_h}{n} (\bar{y}_{rh} - \bar{y}_s)^2 ]
\]

\[
= E[ ( \sum_{h=1}^{L} \frac{n_h}{n} \bar{y}_{rh}^2 ) - \bar{y}_s^2 ]
\]

\[
= \left[ \sum_{h=1}^{L} \frac{n_h}{N} (V(\bar{y}_{rh}) + \bar{y}_h^2) \right] - \left[ V(\bar{y}_s) + \bar{y}_s^2 \right]
\]

\[
= \left[ \sum_{h=1}^{L} \frac{N_h}{N} \right] \left[ (\bar{y}_h - \bar{y})^2 + (V(\bar{y}_{rh}) - V(\bar{y}_s)) \right],
\]

(3.61)

so \( \sum_{h=1}^{L} n_h \frac{n}{n} (\bar{y}_{rh} - \bar{y}_s)^2 \) is not unbiased for \( \sum_{h=1}^{L} N_h \frac{N}{N} (\bar{y}_h - \bar{y})^2 \). There is a bias of \( \sum_{h=1}^{L} N_h \frac{N}{N} [V(\bar{y}_{rh}) - V(\bar{y}_s)] \) in Kalton's estimator of \( S^2 \) (assuming an infinite population). Since \( V(\bar{y}_{rh}) \) and \( V(\bar{y}_s) \) go to zero as \( r \) and the \( r_h \)'s get large, Kalton's (1983) estimator will be approximately unbiased in large samples if the number of respondents is reasonably large. In expressions (3.49) and (3.54) we estimate \( S^2 \) unbiasedly. Our estimators have an advantage over Kalton's in that they are unbiased and are easier to compute. We make no attempt to compare the reliability of the estimators.

3.2. The general case

3.2.1. Introduction to the general case In this section, we relax the restriction that the number of missing units be less than the number of respondents. For a sample that is not broken up into imputation classes, a sample surveyor would be hesitant to analyze the data using imputation if the number of respondents were
less that the number of missing. When a sample is broken up into imputation classes, however, it is not uncommon for some of the classes to have the number of missing within that class greater than the number responding. It is for this reason that we present the general case. We will look at the work of Little and Rubin (1987), and Kalton (1983), and also what we have done.

3.2.2. Notation and imputation method for the general case

We introduce some notation. As before, let \( r (r_h) \) be the number of respondents (in class \( h \)), and \( m (m_h) \) be the number missing (in class \( h \)), and \( n (n_h) \) be the total sample size (in class \( h \)). Let \( k = \lfloor m \cdot r_h^{-1} \rfloor \), where \( \lfloor x \rfloor \) is the largest integer less than or equal to \( x \). Let \( t = m - k \cdot r_h \). Then

\[
m = k \cdot r_h + t
\]

(3.62)

and

\[
n = (k + 1) \cdot r_h + t
\]

(3.63)

If the sample is divided into imputation classes, we define

\[
k_h = \lfloor m_h \cdot r_h^{-1} \rfloor,
\]

(3.64)

\[
t_h = m_h - k_h \cdot r_h,
\]

(3.65)
\[ m_h = k_h r_h + t_h, \quad (3.66) \]

and,

\[ n_h = (k_h + 1) r_h + t_h. \quad (3.67) \]

If \( r > m \) (\( r_h > m_h \)) then \( k = 0 \) (\( k_h = 0 \)) and \( m = t \) (\( m_h = t_h \)).

The imputation technique used by Little and Rubin (1987), Kalton (1983), and ourselves is to assign the entire sample of respondents (the sample of respondent in imputation class \( h \)) \( k \) (\( k_h \)) times over to \( k r \) (\( k_h r_h \)) of the missing units and to do simple random sampling without replacement from the respondents (respondents in class \( h \)) to choose the remaining \( t \) (\( t_h \)) units for imputation. The notation for the sample remains the same. The notation for the augmented sample changes to

\[
\begin{align*}
Y_{i}^* &= Y_i \quad i = 1, \ldots, r \\
Y_{i}^* &= Y_{i-r} \quad i = r+1, \ldots, 2r \\
&\quad \vdots \\
Y_{i}^* &= Y_{i-kr} \quad i = kr+1, \ldots, (k+1)r \\
Y_{i}^* \text{ is random} &= Y_{i-(k+1)r+1, \ldots, n}. \quad (3.68)
\end{align*}
\]

We define \( \bar{y}_t = t^{-1} \sum_{i=(k+1)r+1}^{n} Y_{i}^* \), the mean of the randomly chosen imputed values. For imputation classes, we include the subscript \( h \).
3.2.3. The expectation and variance of the imputed mean under Little and Rubin

Little and Rubin (1987) give as an estimator for $Y$,

$$y^* = \frac{1}{n} \{ (k+1)r \, \bar{y}_r + t \, \bar{y}_t \}$$

(3.69)

If $r > m$, then expression (3.69) reduces to (3.10). Following the same assumptions, conditioning, and logic as in Section 3.1.2,

$$E(\bar{y}_r \mid Y, I, R) = \bar{y}_r,$$

(3.70)

so,

$$E(\bar{y}^* \mid Y, I, R) = \frac{1}{n} \{ (k+1)r \, \bar{y}_r + t \, \bar{y}_t \}$$

$$= \bar{y}_r$$

(3.71)

and

$$E(\bar{y}_r \mid Y) = Y,$$

(3.72)

which is the same as (3.13). The variance of $\bar{y}^*$ is

$$V(\bar{y}^* \mid Y) = V\{ E(\bar{y}^* \mid Y, I, R) \mid Y \}$$
If we ignore the finite population correction factor then (3.73) is

\[ V(\bar{y}^* | Y) = \left( \frac{1}{n} + \frac{k n + (k+2) t}{n^2} \right) S^2. \]  

(3.74)

If \( r > m \) then (3.74) reduces to (3.15).

3.2.4. The expectation and variance of the imputed mean under Kalton

Kalton's (1983) estimator of \( \bar{Y} \) in the general case is

\[ \bar{y}^* = \frac{L}{\sum_{h=1}^{L} \frac{n_h}{n} \frac{(k_h+1)}{r_h} \bar{y}_{rh} + t_h \bar{y}_{th}}. \]  

(3.75)

If \( r_h > m_h \) in all \( L \) classes, then (3.75) reduces to (3.19). Using the same
assumptions, conditioning, and logic as in the first section, we find the expectation and variance of $\bar{y}^*$,

$$E_4(\bar{y}_{th} \mid y_r, n_h's, r_h's) = \bar{y}_{rh},$$  \hspace{1cm} (3.76)

so,

$$E_4(\bar{y}^* \mid y_r, n_h's, r_h's)$$

$$= \sum_{h=1}^{L} \frac{n_h}{n} \frac{(k_h+1) r_h \bar{y}_{rh} + t_h \bar{y}_{rh}}{n_h}$$

$$= \sum_{h=1}^{L} \frac{n_h}{n} \bar{y}_{rh},$$  \hspace{1cm} (3.77)

and,

$$E_1 \{ E_4(\bar{y}^*) \mid y_r, n_h's, r_h's \} = E_1\{ \sum_{h=1}^{L} \frac{n_h}{n} \bar{y}_{rh} \}$$

$$= \sum_{h=1}^{L} \frac{N_h}{N} Y_{Rh} = Y_s.$$  \hspace{1cm} (3.78)

Expression (3.78) is the same as expression (3.20). The only changes to the derivation of the variance of $\bar{y}^*$ done in expressions (3.23) to (3.32) are in (3.23) and (3.24). In the general model,
If \( m < r \) in all \( L \) classes, then (3.79) reduces to (3.23). With the change in (3.23), expression (3.33) becomes

\[
\hat{V}(\bar{y}) = \sum_{h=1}^{L} \{ (\frac{N_h}{N} - \frac{1}{n} - \frac{E(n_h^2)}{n^2R_h} ) S_{Rh}^2 \\
+ \frac{N_h}{N-1}(\bar{Y}_{Rh} - \bar{Y}_s)^2(1 - \frac{n}{N}) \}
+ \sum_{h=1}^{L} \frac{E(k_h n_h + (k_h+2) t_h)}{n^2} S_{Rh}^2
\]

Expression (3.34), Kalton's (1983) estimator of the variance of \( \bar{y} \), changes to

\[
\hat{V}(\bar{y}) = \sum_{h=1}^{L} \left[ \frac{n_h^2}{n^2} s_{Rh}^2 \frac{1}{r_h} + \frac{n_h}{n^2} (\bar{Y}_{Rh} - \bar{Y}_s)^2 \right. \\
+ t_h \left( 1 - \frac{t_h}{r_h} \right) \frac{s_{Rh}^2}{n^2} \] ,

Expressions (3.80) and (3.81) are the same as (3.33) and (3.34) if \( m < r \) in all \( L \) classes.
If we impose the alternate assumption that the data are missing at random on Kalton’s (1983) work in the general case, we get, instead of (3.41),

\[
V(\bar{y}^*) = \frac{S^2}{n} + \frac{L}{\sum_{h=1}^{L}} \frac{E(\sum_{i=1}^{n_h} (k_h + \delta_{hi}) Y_{hi}^2)}{n^2} S_h^2. \tag{3.82}
\]

Expression (3.82) reduces to (3.41) if \( m_h < r_h \) in all \( L \) classes.

### 3.2.5. The expectation and variance of the imputed mean under our method

In the general case, Our form of the estimator of the population mean is

\[
\bar{y}^* = \sum_{h=1}^{L} n_h \frac{1}{n-1} \gamma_{hi} (k_h + 1 + \delta_{hi}) Y_{hi}, \tag{3.83}
\]

where \( \gamma_{hi} \) stays the same as before and \( \delta_{hi} \) is a 0–1 random variable that takes on the value 1 if unit \( hi \) is randomly chosen for imputation and takes on the value 0 if unit \( hi \) is not one of the \( t_h \) units chosen randomly for imputation. Some properties of \( \delta_{hi} \) are \( E(\delta_{hi} | \gamma_{hi}=1, n_h \text{'s, } r_h \text{'s}) = t_h r_h^{-1} \), and \( \sum_{i=1}^{n_h} \delta_{hi} = t_h \), and

\[P(\delta_{hi} = 1 | \gamma_{hi} = 0) = 0.\]

If \( m_h < r_h \) in all \( L \) classes then expression (3.83) reduces to (3.42). Using the same assumptions, conditioning, and logic as in the first section, we find the expectation and variance of \( \bar{y}^* \). The expectation of \( \bar{y}^* \) is

\[
E(\bar{y}^*) = E\{E(\bar{y}^* | \bar{y}, n_h \text{'s, } r_h \text{'s}) \}
\]

\[
= E\{ \frac{L}{\sum_{h=1}^{L} n_h} \sum_{i=1}^{n_h} \left( \frac{r_h}{n_h} \right) (k_h + 1 + \frac{t_h}{r_h}) \frac{Y_{hi}}{n} \}
\]
Expression (3.84) is the same as expression (3.43). The variance of \( \bar{y}^* \) is

\[
V(\bar{y}^*) = E\{ V(\bar{y}^* | y, n_h's, r_h's) \}
+ \{ E(\bar{y}^* | y, n_h's, r_h's) \}
\]

Let us look at \( V(\bar{y}^* | y, n_h's, r_h's) \),

\[
V(\bar{y}^* | y, n_h's, r_h's) = \{ (\bar{y}^2 + \sum_{h=1}^{L} \frac{k_h n_h + (k_h+2) t_h}{n^2} s_h^2 ) - 2 \bar{y}^2 + \bar{y}^2 \}
= \sum_{h=1}^{L} \frac{k_h n_h + (k_h+2) t_h}{n^2} s_h^2, \tag{3.86}
\]

since

\[
E(\gamma_h (k_h + 1 + \delta_h) | y, n_h's, r_h's )
\]
\[
(8.9.3) \quad \left( \frac{\eta_x}{\eta_1} \varepsilon + \frac{\eta_x}{\eta_3} \eta_x \zeta + \eta \right) \frac{\eta_u}{\eta_x} =
\]

\[
\left( s, \eta_x, s, \eta_u, \zeta \mid ( \frac{\eta_y}{\eta_9} + \frac{\eta_x}{\eta_7} ( \eta + \frac{\eta_7}{\eta} ) \varepsilon + \frac{\eta_x}{\eta_3} ( \eta + \frac{\eta_7}{\eta} ) \frac{\eta_y}{\eta_9} ) \right) \mathcal{E} =
\]

\[
\left( s, \eta_x, s, \eta_u, \zeta \mid \frac{\eta_y}{\eta_9} + \frac{\eta_x}{\eta_3} ( \eta + \frac{\eta_7}{\eta} ) \frac{\eta_y}{\eta_9} \right) \mathcal{E} =
\]

\[
\left( \frac{\eta_x}{\eta_3} \right) \mathcal{E} = \eta \mathcal{E}.
\]

\[
( s, \eta_x, s, \eta_u, \zeta \mid ( \frac{\eta_y}{\eta_9} + \frac{\eta_x}{\eta_3} ( \eta + \frac{\eta_7}{\eta} ) \frac{\eta_y}{\eta_9} ) \mathcal{E} =
\]

\[
I = ( s, \eta_x, s, \eta_u, \zeta \mid ( \frac{\eta_y}{\eta_9} + \frac{\eta_x}{\eta_3} ( \eta + \frac{\eta_7}{\eta} ) \frac{\eta_y}{\eta_9} ) \mathcal{E} \times
\]

\[
( s, \eta_x, s, \eta_u, \zeta \mid ( \frac{\eta_y}{\eta_9} + \frac{\eta_x}{\eta_3} ( \eta + \frac{\eta_7}{\eta} ) \frac{\eta_y}{\eta_9} ) \mathcal{E} =
\]

\[
( s, \eta_x, s, \eta_u, \zeta \mid ( \frac{\eta_y}{\eta_9} + \frac{\eta_x}{\eta_3} ( \eta + \frac{\eta_7}{\eta} ) \frac{\eta_y}{\eta_9} ) \mathcal{E}
\]

\[
\text{and}
\]

\[
(8.8.3) \quad I = \left( \frac{\eta_x}{\eta_3} + \frac{\eta_x}{\eta_1} \right) + \frac{\eta_u}{\eta_x} =
\]

\[\text{111}\]
if \( h = g \) and \( i = j \).

The variance of \( \bar{y}^* \) is,

\[
V(\bar{y}^*) = \sum_{h=1}^{L} \frac{E( k_h n_h + (k_h+2) t_h )}{n^2} S_h^2 + \frac{S^2}{n}.
\]

(We ignore the finite population correction factor in the variance of \( \bar{y} \).) If \( m_h < r_h \) in all \( L \) classes, then (3.89) reduces to (3.48).

### 3.2.6. Estimation of the variance of the imputed mean under the general model

We now look at the generalizations of the two estimators for the variance of \( \bar{y}^* \) that we presented in Section 3.1.6. We first look at the expectation of the generalized version of \( s^2 \),

\[
E( s^2 ) = E \left[ \sum_{h=1}^{L} \frac{n_h}{n-1} \left( \frac{Y_{hi}^* - \bar{Y}}{n-1} \right)^2 \right] - \frac{n}{n-1} \frac{N-1}{N} S^2
\]

\[= \frac{n}{n-1} \frac{N-1}{N} S^2 - \frac{n}{n-1} E \left[ \sum_{h=1}^{L} \frac{k_h n_h + (k_h+2) t_h}{n^2} S_h^2 \right] - 2 \left[ \bar{Y} \bar{Y} \right] \]

\[\left\{ \begin{array}{l}
\left[ \bar{Y}^2 \right] + \left[ Y^2 \right]
\end{array} \right\}
\[ S^2 - \sum_{h=1}^{L} \frac{1}{n-1} \frac{E(k_h n_h + (k_h+2) t_h)}{s^2} \]

If \( m_h < r_h \) in all \( L \) classes, then (3.90) reduces to (3.51). The generalized estimator corresponding to our first estimator of the variance of \( \bar{y}^* \), expression (3.49), is

\[ \hat{V}(\bar{y}^*) \]

\[ = \sum_{h=1}^{L} \left( 1 + \frac{1}{n-1} \right) \frac{k_h n_h + (k_h+2) t_h}{s^2} + \frac{s^2}{n} \]

Expression (3.91) is unbiased for (3.90) and reduces to (3.49) if \( m_h < r_h \) in all \( L \) classes. In Figure 3.1, we give the \( V(\bar{y}^* | r) \), \( V(\bar{y}^*) \), and \( E(n^{-1} s^2 | r) \) for a simple random sample without replacement of size 100 with \( s^2 = 100 \) and only one imputation class.

The generalized estimator corresponding to our second estimator of the variance of \( \bar{y}^* \) in Section 3.1.6, expression (3.54), can be written two different ways. This estimator requires that the number of times that a respondent has been used for imputation is recorded. Let us call this number \( a \), so that \( a_{hi} \) is the number of times that unit \( hi \) has been used for imputation. If unit \( hi \) is a nonrespondent, then \( a_{hi} \) is the number of times that the unit used as the imputed value for unit \( hi \) has been used for imputation. The first form of the estimator is
Figure 3.1. The $V(\bar{y}^* - Y \mid r)$ and $E(s^*^2 \mid r)$ for one imputation class: $n=100$, $S^2=100$
\[ \hat{V}(\bar{y}^*) = \sum_{h=1}^{n_h} \sum_{i=1}^{r_h} \frac{1}{n^2} - \left( 1 + \frac{1}{n-1} \right) \xi_{hi} \left( Y_{hi}^* - \bar{y}_{rh} \right)^2 \]
\[ + \frac{s^2}{n}, \quad (3.92) \]

where,

- \( \xi_{hi} = 0 \) if unit hi is a nonrespondent.

- \( \xi_{hi} = a_{hi} (a_{hi}+1) r_h (r_h-1)^{-1} \) if unit hi is a respondent. \( \quad (3.93) \)

Let us look at the expected value of \( \sum_{i=1}^{n_h} \xi_{hi} \left( Y_{hi}^* - \bar{y}_{rh} \right)^2 \),

\[ E\{ \sum_{i=1}^{n_h} \xi_{hi} \left( Y_{hi}^* - \bar{y} \right)^2 \} \]
\[ = E\{ \sum_{i=1}^{r_h} a_{hi} (a_{hi}+1) \frac{r_h}{r_h-1} \left( Y_{hi}^* - \bar{y}_{rh} \right)^2 \} \]
\[ = E\{ k_h (k_h+1) r_h \sum_{i=1}^{r_h} \frac{r_h}{r_h-1} \left( Y_{hi}^* - \bar{y}_{rh} \right)^2 \}
\[ + \left[ (k_h + 1) (k_h + 2) - k_h (k_h + 1) \right] \]
\[
\sum_{i=(k_h+1)r_h+1}^{n_h} E\left[ \frac{r_h}{r_h-1} \left( Y_{hi}^* - \bar{y}_{rh} \right)^2 \right | y, n_h, s_h, r_h \}
\]

\[
= E\left\{ k_h (k_h+1) r_h s_{rh}^2 + (2k_h + 2) t_h s_{rh}^2 \right\}
\]

\[
= E\left\{ k_h n_h + (k_h+2) t_h \right\} s_{h}^2.
\]  

(3.94)

Given this expectation, expression (3.94) is unbiased for (3.89). If \( m_h < r_h \) in all L classes, then \( k_h = 0 \) for all \( h \), and \( \sum_{i=1}^{n_h} \xi_{hi} \left( Y_{hi}^* - \bar{y}_{rh} \right)^2 \) becomes

\[
\sum_{i=r_h+1}^{n_h} 2 r_h (r-1)^{-1}(Y^* - \bar{y})^2,
\]

which is what we have in our estimator (3.54). Let us look at the second way of writing this estimator,

\[
\hat{V}(Y^*) = \sum_{h=1}^{L} \sum_{i=1}^{n_h} \frac{1}{n^2} \left( 1 + \frac{1}{n-1} \right) a_{hi} \frac{r_h}{r_h-1} \left( Y_{hi}^* - \bar{y}_{rh} \right)^2 + \frac{s_{*}^2}{n^2}
\]  

(3.95)

The variable \( a_{hi} \) takes on the value \( k_h \) for \( r_h-t_h \) observations and the value \( k_h+1 \) for \( t_h \) observations. The entire set of observations in class \( h \) appears \( (k_h+1) \) times in the augmented data set. The randomly chosen set appears once. In the randomly chosen set \( a_{hi} = k_h + 1 \). So,

\[
\sum_{i=1}^{n_h} a_{hi} \frac{r_h}{r_h-1} \left( Y_{hi}^* - \bar{y}_{rh} \right)^2 = (k_h+1) k_h \sum_{i=1}^{r_h} \frac{r_h}{r_h-1} \left( Y_{hi} - \bar{y}_{rh} \right)^2
\]
This is the same sum that we found for estimator (3.92), so estimator (3.92) and estimator (3.95) are the same, and what we say about (3.92) also holds for (3.95).
4. RANDOM IMPUTATION FOR A STRATIFIED RANDOM SAMPLE

Stratified random sampling is a common form of sampling. Stratified random sampling involves dividing the population to be sampled into subpopulations, called strata. Within each subpopulation, a random sample is taken. Between the subpopulation samples there is independence. The samples from the subpopulations are used to estimate parameters of the total population. In this chapter, we present a method for random imputation of missing values when the sampling design is stratified random sampling.

4.1. The finite population and sample

Let there be a population of \( N \) units divided into \( M \) strata (indexed by \( m \)) of size \( N_m, \ m = 1, \ldots, M \). Also let the population be divided into \( H \) imputation classes (indexed by \( h \)) of size \( N_{mh}, h=1,\ldots,H \). Let \( N_{mh} \) denote the number of units that are in both stratum \( m \) and imputation class \( h \). Assume \( N \) and \( N_m, m=1,\ldots,M \) are known. Let there be a characteristic associated with each unit, \( \{Y_{mhi}\}_{m=1,\ldots,M, h=1,\ldots,H, i=1,\ldots,N_{mh}} \). Let

\[
Y_{mh} = \frac{1}{N_{mh}} \sum_{i=1}^{N_{mh}} Y_{mhi} \tag{4.1}
\]

be the finite population mean of the characteristic for the units in both stratum \( m \) and imputation class \( h \), let

\[
Y = \sum_{m=1}^{M} \sum_{h=1}^{H} \frac{N_{mh}}{N} Y_{mh} \tag{4.2}
\]
be the overall finite population mean, and let

\[ S_m^2 = \frac{1}{N_m - 1} \sum_{h=1}^{H} \sum_{i=1}^{N_{mh}} (Y_{mhi} - \frac{1}{N_{mh}} \sum_{g=1}^{H} \frac{N_{mg}}{N_m} Y_{mg})^2 \]  

be the finite population variance in stratum \( m \). Let us take a simple random sample without replacement within each stratum. Let the size of the sample be \( n_m \) in stratum \( m \). Assume that we know the imputation class into which each unit in the sample falls. Let \( n_{mh} \) denote the number of units in the sample that fall in imputation class \( h \) and stratum \( m \). Let \( n_h \) denote the sum over the strata of \( n_{mh} \), that is, the number of units in the sample that fall in imputation class \( h \). Suppose that within stratum–imputation class cell \( m \ h \) we are able to measure the characteristic for \( r_{mh} \) of the units and unable to measure it for \( m_{mh} \) of the units. Then \( r_{mh} + m_{mh} = n_{mh} \), \( r_{mh} \geq 0 \), and \( m_{mh} \geq 0 \). The number of units responding and missing in imputation class \( h \) are, respectively,

\[ r_h = \sum_{m=1}^{M} r_{mh} \]  

and

\[ m_h = \sum_{m=1}^{M} m_{mh}. \]  

Also \( r_h + m_h = n_h \). The portion of the sample from stratum \( m \) in imputation
class h is written \( \{ Y_{mhi} \}_{i=1}^{n_{mh}} \), where \( Y_{mh1}, \ldots, Y_{mh_{r_{mh}}} \) are associated with units that responded and \( Y_{mh, r_{mh}+1}, \ldots, Y_{mhn_{mh}} \) are associated with units for which there was no response. The augmented sample in stratum m and imputation class h is \( \{ Y_{mhi}^* \}_{i=1}^{n_{mh}} \), where \( Y_{mhi}^* = Y_{mhi} \) for \( i = 1, \ldots, r_{mn} \) and \( Y_{mhi}^* \) is an imputed value for \( i = r_{mh}+1, \ldots, n_{mh} \). Let

\[
\bar{y}_{m}^* = \frac{1}{n_{m}} \sum_{h=1}^{H} \sum_{i=1}^{n_{mh}} Y_{mhi}^*
\]

be the stratum sample mean computed using both the respondents and the imputed values, let

\[
\bar{y}_{m} = \frac{1}{n_{m}} \sum_{h=1}^{H} \sum_{i=1}^{n_{mh}} Y_{mhi}
\]

be the stratum sample mean of the complete sample, let

\[
\bar{y}_{r,h} = \frac{1}{r_{h}} \sum_{m=1}^{M} \sum_{i=1}^{r_{mh}} Y_{mhi}
\]

be the imputation class sample mean of the respondents, let

\[
s_{m}^2 = \frac{1}{n_{m} - 1} \sum_{h=1}^{H} \sum_{i=1}^{n_{mh}} \left( Y_{mhi}^* - \bar{y}_{m}^* \right)^2
\]
be the stratum sample variance computed using both the respondents and the imputed values, let

\[ s^2_{m.} = \frac{1}{n_{m.} - 1} \sum_{h=1}^{H} \sum_{i=1}^{n_{mh}} (Y_{mhi} - \bar{y}_{m.})^2 \]  

(4.10)

be the stratum sample variance for the complete sample, and let

\[ s^2_{r.h} = \frac{1}{r_{h} - 1} \sum_{m=1}^{M} \sum_{i=1}^{r_{mh}} (Y_{mhi} - \bar{y}_{r,h})^2 \]  

(4.11)

be the imputation class sample variance of the respondents.

Assume that within each imputation class the units that are nonrespondents are missing at random from the sample, where missing at random is defined in Section 2.3. Since sampling in stratum \( m \) is simple random sampling without replacement and since the nonrespondents in imputation class \( h \) are missing at random, \( \{Y_{mhi}\}_{i=1}^{r_{mh}} \) is a simple random sample without replacement from the portion of the population that is in stratum \( m \) and imputation class \( h \).

4.2. The imputation method

The imputation is done within imputation classes, ignoring strata. Let

\[ k_{.h} = \left\lfloor \frac{m_{h}}{r_{.h}} \right\rfloor \]  

(4.12)

where \([x]\) is the largest integer less than or equal to \( x \). Let
Then each respondent in imputation class $h$ is used at least $k_h$ times for imputation and $t_h$ of the respondents in class $h$ are used $k_h + 1$ times. The $t_h$ of the respondents that are used $k_h + 1$ times are chosen by simple random sampling without replacement from the respondents in imputation class $h$. The donors are assigned to the missing values randomly.

4.3. The expectation of the imputed mean over the finite population

Let an estimator of the finite population mean be

$$
\bar{y}_* = \frac{1}{N} \sum_{h=1}^{H} \sum_{m=1}^{M} \sum_{i=1}^{n_{mh}} \gamma_{mhi} \left[ w_m + \sum_{p=1}^{P} w_p \left( k_{mhi}^p + \delta_{mhi}^p \right) \right] y_{mhi} \tag{4.14}
$$

This estimator can be written

$$
\bar{y}_* = \frac{1}{N} \sum_{h=1}^{H} \sum_{m=1}^{M} \sum_{i=1}^{n_{mh}} \gamma_{mhi} \left[ w_m + \sum_{p=1}^{P} w_p \left( k_{mhi}^p + \delta_{mhi}^p \right) \right] y_{mhi} \tag{4.15}
$$

where $\gamma_{mhi}$ is a 0–1 random variable that takes on the value 1 if unit $mhi$ responds and takes on the value 0 if unit $mhi$ does not respond, $w_m$ is the stratum weight for stratum $m$ and equals $N_m^{-1} n_{m-1}^{-1} k_{mhi}^p$ is a random variable with a
hypergeometric distribution that indicates whether $Y_{mhi}$ was used for non-random imputation in stratum $p$, and if it was, how many times it was used, where non-random imputation is defined to be imputation out of the set of donors that come from the $k_h$ replications of the set of respondents, and $\delta_{mhi}^P$ is a 0--1 random variable that takes on the value 1 if unit $mhi$ was used for random imputation in stratum $p$ and takes on the value 0 otherwise, where random imputation is defined to be imputation out of the set of the $t_h$ donors that are chosen by simple random sampling from the set of respondents.

Lemma 4.1: Let the assumptions in Section 4.1 hold. Assume that the imputation is done by using all of the respondents in imputation class $h$ $k_h$ times and using $t_h$ of the respondents in imputation class $h$ $k+h$ times, where the $t_h$ of the respondents used $k+h$ times are chosen by simple random sampling without replacement from the respondents in imputation class $h$. Assume that the donors are assigned to the missing values at random. Then

$$E(\gamma_{mhi} | r_{mh's}, n_{mn's}, t_{mh's}) = \frac{r_{mh}}{n_{mh}}, \quad (4.16)$$

$$E(\delta_{mhi}^P | \gamma_{mhi} = 1, n_{mh's}, r_{mh's}, t_{mh's}) = \frac{t_{ph}}{r_h}, \quad (4.17)$$

$$E(k_{mhi}^P | \gamma_{mhi} = 1, n_{mh's}, r_{mh's}, t_{mh's}) = \frac{k_{ph}}{r_h}, \quad (4.18)$$

where the conditioning holds the sample size, the number of respondents, the number of missing units, the number of randomly imputed units, and the number of
non—randomly imputed units constant within the imputation class—stratum subgroups.

**Proof:** For expression (4.16),

\[ E(\gamma_{mhi} \mid r_{mh}'s, n_{mh}'s, t_{mh}'s) \]

\[ = P(\gamma_{mhi} = 1 \mid r_{mh}'s, n_{mh}'s, t_{mh}'s) \]  

and the probability that unit mhi responds, given that there are \( r_{mh} \) respondents out of a sample size of \( n_{mh} \) in the \( mh \)—th stratum — imputation class subgroup, is \( \frac{n_{mh} - 1}{r_{mh}} \).

For expression (4.17), first

\[ \sum_{p=1}^{M} \delta_{mhi}^p = 1 \text{ or } 0 , \]

since unit mhi can be only used at most once for random imputation. Secondly,

\[ P(\sum_{p=1}^{M} \delta_{mhi}^p = 1 \mid \gamma_{mhi} = 1, n_{mh}'s, r_{mh}'s, t_{mh}'s) \]

\[ = \frac{t_{.h}}{r_{.h}} , \]

since expression (4.21) is the probability of unit mhi being chosen for random imputation where \( t_{.h} \) are chosen out of \( r_{.h} \) by simple random sampling without
replacement. Thirdly,

\[ P(\delta_{mhi}^P = 1 | \sum_{q=1}^{M} \delta_{mhi}^q = 1, \gamma_{mhi} = 1, n_{mh's}, r_{mh's}, t_{mh's}) \]

\[ (4.22) \]

is equal to the proportion of the randomly chosen missing values in stratum \( p \) and imputation class \( h \), or \( t_{h}^{-1} t_{ph} \). Fourthly,

\[ P(\delta_{mhi}^P = 1 | \sum_{q=1}^{M} \delta_{mhi}^q = 1, \gamma_{mhi} = 1, n_{mh's}, r_{mh's}, t_{mh's}) \]

\[ \times P(\sum_{q=1}^{M} \delta_{mhi}^q = 1 | \gamma_{mhi} = 1, n_{mh's}, r_{mh's}, t_{mh's}) \]

\[ = P(\delta_{mhi}^P = 1 \text{ and } \sum_{q=1}^{M} \delta_{mhi}^q = 1 | \gamma_{mhi} = 1, n_{mh's}, r_{mh's}, t_{mh's}) \]

\[ = P(\sum_{q=1}^{M} \delta_{mhi}^q = 1 | \delta_{mhi}^P = 1, \gamma_{mhi} = 1, n_{mh's}, r_{mh's}, t_{mh's}) \]

\[ \times P(\delta_{mhi}^P = 1 | \gamma_{mhi} = 1, n_{mh's}, r_{mh's}, t_{mh's}) \]

\[ = 1 \times P(\delta_{mhi}^P = 1 | \gamma_{mhi} = 1, n_{mh's}, r_{mh's}, t_{mh's}) \]

\[ = E(\delta_{mhi}^P | \gamma_{mhi} = 1, n_{mh's}, r_{mh's}, t_{mh's}) . \]

\[ (4.23) \]

Expression (4.17) follows from expression (4.23).
For expression (4.18), $k_{mh}^P$ is an hypergeometric random variable with sample size equal to $k_{.h}$, subpopulation size equal to $k_{ph}$, and population size equal to $k_{..r.h}$. See Tsao (1965, pp. 75–78) for properties of the hypergeometric.

Theorem 4.1: Let a stratified random sample of stratum sizes $n_1, \ldots, n_m$ be given. Let random imputation based on $H$ imputation classes be used for the missing observations. Let $\{Y_{mhi}\}_{i=1}^{r_{mh}}$ be the portion of the observed sample that is in stratum $m$ and imputation class $h$. Assume that the missing observations are missing at random within imputation classes. Then the expectation of estimator (4.14) minus the population mean is

$$E\{ ( \bar{y}^*_{st} - \bar{Y} ) | FP \}$$

$$= E\left\{ \sum_{h=1}^{H} \sum_{p=1}^{M} m_{ph} W_p \left( \sum_{m=1}^{M} \frac{r_{mh}}{r_{..h}} Y_{mh} - Y_{ph} \right) | FP \right\},$$

where the expectation is over all possible samples and all possible response patterns for a fixed finite population and $FP$ stands for the fixed finite population.

Proof: We have, from Lemma 4.1 and the assumption that the responding units in the stratum $m$ — imputation class $h$ subgroup are a simple random sample without replacement from the portion of the population in stratum $m$ and imputation class $h$, that
\[ E(\bar{Y}^*_st \mid FP) = E\{ E(\bar{Y}^*_st \mid n_{mh}'s, r_{mh}'s, t_{mh}'s) \mid FP \} \]

\[ = E\{ \sum_{h=1}^{H} \sum_{m=1}^{M} r_{mh} ( W_m + \sum_{p=1}^{M} W_p \frac{m_{ph}}{r_{h}} ) Y_{mh} \mid FP \} \]

\[ = E\{ \sum_{h=1}^{H} \sum_{m=1}^{M} n_{mh} W_m Y_{mh} \]

\[ + \sum_{p=1}^{M} m_{ph} W_p ( \sum_{m=1}^{M} \frac{r_{mh}}{r_{h}} Y_{mh} - Y_{ph} ) \} \mid FP \} \]

\[ = Y + E\{ \sum_{h=1}^{H} \sum_{p=1}^{M} m_{ph} W_p ( \sum_{m=1}^{M} \frac{r_{mh}}{r_{h}} Y_{mh} - Y_{ph} ) \} \mid FP \} . \quad (4.25) \]

From (4.24), if \( Y_{mh} = Y_h \) for all \( m=1,...,M \), then \( E(\bar{Y}^*_st \mid FP) = Y \).

Otherwise \( \bar{Y}^*_st \) is biased for \( Y \). It is unlikely that for a finite population \( Y_{mh} \) would equal \( Y_h \) for all \( m=1,...,M \).

4.4. The superpopulation structure

The assumption that the missing units are missing at random within imputation classes is not strong enough to give us unbiasedness for our estimator of the finite population mean for a fixed finite population. We now add an assumption of an underlying superpopulation for the finite population. Assume that the superpopulation is made up of \( H \times M \) stratum–imputation class subgroups. Random samples of size \( N_{mh} \), \( m=1,...,M \), \( h=1,...H \) from the subgroups make up the...
finite population. Assume that the units in imputation class $h$ are distributed identically and independently with mean $\mu_h$ and variance $\sigma_h^2$. Let

$$\bar{\mu} = \frac{\sum_{h=1}^{H} \frac{N_h}{N} \mu_h}{H} ,$$

be the overall superpopulation mean. Note that the $N_{mh}$ are fixed in this formulation.

The assumption of identically and independently distributed observations within an imputation class might be justified on the basis of the formation of imputation classes. Sample surveyors, in choosing imputation classes, look not only for uniform response rates but also for uniformity in the known characteristics of that class. They hope to find uniformity in the unknown characteristic if the known characteristics are uniform.

4.5. The expectation of the imputed mean under the superpopulation model

We now find the expected value of $\bar{y}_*^{st}$ over the superpopulation.

Theorem 4.2: Let the stratified sampling assumptions of Theorem 4.1 hold. Let the missing values be missing at random within the imputation classes. Assume that the finite population is a sample from a superpopulation in which elements are independently and identically distributed within imputation classes. Then

$$E(\bar{y}_*^{st}) = \bar{\mu} ,$$

(4.27)
where \( \mu \) is the superpopulation mean, and the expectation is over all possible finite population selections from the superpopulation and over all possible samples, response patterns, and imputation patterns. Also, the expectation of the finite population mean is

\[
E( \bar{Y} ) = \mu. \tag{4.28}
\]

**Proof:** To find the expectation of our estimator (4.14) under the superpopulation model, we use two levels of conditioning. At the lower level, 2, we hold everything constant but the \( Y_{mhi} \)’s and take expectations with respect to the superpopulation distribution. That is, at level 2 we hold the sampling design, the pattern of response, and the pattern and choice of imputed values constant. At level 1 we take expectations and variances over the \( n_{mh} \)’s, the \( r_{mh} \)’s, the \( t_{mh} \)’s, the \( \gamma_{mhi} \)’s, the \( k_{mhi} \)’s, and the \( \delta_{mhi} \)’s. We have, for (4.27), that

\[
E(\bar{y}_{st}^\ast) = E_1[ E_2( \sum_{h=1}^{H} \sum_{m=1}^{M} \sum_{i=1}^{n_{mh}} Y_{mhi} \gamma_{mhi} [ W_m
\]

\[+ \sum_{p=1}^{M} W_p ( k_{mhi}^p + \delta_{mhi}^p ) ] n_{mh} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma_{mhi} \gamma}_{

Noting that

\[ \mathcal{H} \sum_{h=1}^{M} \mu_h \left[ \sum_{m=1}^{M} \sum_{i=1}^{n_{mh}} \gamma_{mhi} W_m + \sum_{p=1}^{M} \sum_{m=1}^{M} \sum_{i=1}^{n_{mh}} \gamma_{mhi} (k_{mhi}^p + \epsilon_{mhi}^p) \right] \]

\[ = \mathcal{H} \sum_{h=1}^{M} \sum_{m=1}^{M} n_{mh} W_m \mu_h, \quad (4.29) \]

expression (4.29) becomes

\[ E(\tilde{y}_{st}^*) = E_1 \left[ \mathcal{H} \sum_{h=1}^{M} \sum_{m=1}^{M} n_{mh} W_m \mu_h \right] \]

\[ = \mathcal{H} \sum_{h=1}^{M} \sum_{m=1}^{M} N_{mh} \frac{n_{m}}{N_m} \mu_h \]

\[ = \mathcal{H} \sum_{h=1}^{M} \frac{N_h}{N} \mu_h \]

\[ = \bar{\mu}. \quad (4.31) \]

We have for (4.28) that
Since $\bar{y}_{st}^*$ and $\bar{Y}$ both estimate the same quantity, it follows that,

$$E(\bar{y}_{st}^* - \bar{Y}) = 0.$$  (4.33)

4.6. The variance of the imputed mean under the superpopulation model

We present three formulas for the variance of $\bar{y}_{st}^* - \bar{\mu}$. Each of the formulas represents the expectation of the variance taken to a different level of conditioning. The lowest level (level 4) of conditioning holds all of the variables but the $Y_{mhi}$'s constant. The expectations and variances are taken over to the superpopulation at this level. Theorem 4.5 presents the expression for the variance for level 4 conditioning. At the next level of conditioning (level 3), expectations and variances are taken over the choices for imputation (over the $k_{mhi}^P$'s and the $\delta_{mhi}^P$'s). The sample sizes and number of respondents in the imputation class–stratum subgroups are held constant. Theorem 4.3 presents the expression for the variance for level 3 conditioning. At the next level of conditioning (level 2), expectations and variances are taken over choices for the number of respondents in imputation class–stratum subgroups (over the $r_{mh}$'s). The sample sizes within the imputation class–stratum...
subgroups and the sample sizes within the imputation classes are held constant for level 2 conditioning. Theorem 4.4 presents the expression of the variance for level 2 conditioning. A the top level (level 1) expectations and variances are taken over the sample sizes in the imputation class—stratum subgroups, over the number of respondents within the imputation classes and over the sample sizes within the imputation classes. We never evaluate the variance of $\bar{y}_{st}^*$ at level 1 conditioning.

We evaluated the $E(\bar{y}_{st}^*)$ at level 1 in Theorem 4.2. We showed, in Theorem 4.2, that

$$E_4(\bar{y}_{st}^* | n_{mh}', r_{mh}', k_{mhi}', \delta_{mhi}')$$

$$= \sum_{h=1}^{L} \mu_h \sum_{m=1}^{M} n_{mh} W_m.$$  \hspace{1cm} (4.34)

Therefore

$$V(\bar{y}_{st}^*) = V_1\{E_2[ E_3[ E_4(\bar{y}_{st}^*) ] ] \}$$

$$+ E_1\{V_2[ E_3[ E_4(\bar{y}_{st}^*) ] ] \} + E_1\{E_2[ V_3[ E_4(\bar{y}_{st}^*) ] ] \}$$

$$+ E_1\{E_2[ E_3[ V_4(\bar{y}_{st}^*) ] ] \}$$

$$= V_1(\sum_{h=1}^{L} \mu_h \sum_{m=1}^{M} n_{mh} W_m)$$

$$+ E_1\{V_2(\sum_{h=1}^{L} \mu_h \sum_{m=1}^{M} n_{mh} W_m | n_{mh}', r_{mh}') \}$$
It follows that we can find our expression for the variance of $\bar{y}_{st} - \bar{\mu}$ by evaluating the fourth term in (4.35) to the desired level of conditioning and then adding the $V_{1}(\sum_{h=1}^{H} \mu_{h} \sum_{m=1}^{M} n_{mh} W_{m})$, since the second and third terms of (4.35) equal zero.

In the next lemma we give the variance of the usual full sample stratified mean in terms of variances associated with our four levels of conditioning.

**Lemma 4.2:** Let the assumptions of Theorem 4.2 hold. Let

$$\bar{y}_{st} = \sum_{h=1}^{H} \sum_{m=1}^{M} \sum_{i=1}^{n_{mh}} W_{m} Y_{mhi}$$

be the usual estimator for the population mean constructed with a complete sample.

Then the variance

$$V(\bar{y}_{st} - \bar{\mu}) = V_{1}(\sum_{h=1}^{H} \sum_{m=1}^{M} n_{mh} W_{m} \mu_{h})$$
Proof: We have

\[ E_4(\bar{y}_{st}) = E_3(\bar{y}_{st}) = E_2(\bar{y}_{st}) \]

\[ = \sum_{h=1}^{H} \sum_{H} n_{mh} W_m \mu_{,h} \]

and

\[ V_4(\bar{y}_{st}) = E_4\{ \sum_{h=1}^{H} \sum_{m=1}^{M} n_{mh} W_m (Y_{mhi} - \mu_{,h}) \}
\times \sum_{g=1}^{H} \sum_{p=1}^{M} n_{pg} W_p (Y_{pgj} - \mu_{,h}) \]

\[ = \sum_{h=1}^{H} \sum_{m=1}^{M} n_{mh} W_m^2 \sigma_{,h}^2 \]

because the \( Y_{mhi} \)'s are independently and identically distributed within imputation classes and independent between classes with respect to the superpopulation. The result follows from the result in (4.35), since \( E(\bar{y}_{st}) = E(\bar{y}_{st}) \) at all of our four levels of conditioning. #
We now give the variance of estimator (4.15) to the level 3 expectation of conditioning.

**Theorem 4.3:** Under the assumptions of Theorem 4.2,

\[
V(\bar{y}_{st}^* - \bar{\mu}) = V(\bar{y}_{st} - \bar{\mu})
\]

\[
+ \frac{H}{\Sigma} \sigma^2_h E\left\{ 2 \sum_{m=1}^{M} \sum_{p=1}^{M} \frac{W_m W_p}{r \cdot h} (r_{mh} m_p + k_{mh} t_{ph}) \right\}
\]

\[
+ \frac{k_{h} - 1}{r \cdot h k_{h} - 1} \left[ (\sum_{m=1}^{M} W_m k_{mh})^2 - \sum_{m=1}^{M} W_m^2 k_{mh} \right].
\] (4.40)

and

\[
V(\bar{y}_{st} - \bar{Y}) = V(\bar{y}_{st}^* - \bar{\mu}) - \sum_{m=1}^{M} \frac{N_m}{N^2} E(S_m^2)
\] (4.41)

**Proof:** For (4.40),

\[
V(\bar{y}_{st}^* - \bar{\mu}) = V_1 \{ E_2 \left[ E_3 \left[ E_4 (\bar{y}_{st}^* - \bar{\mu}) \right] \right] \} + E_1 \{ E_3 \left[ V_4 (\bar{y}_{st}^* - \bar{\mu}) \right] \}
\]

\[
= V_1 \left( \sum_{h=1}^{H} \sum_{m=1}^{M} n_{mh} W_m \mu_{.h} \right)
\]

\[
+ E_1 \{ E_3 \left[ V_4 \left( \sum_{h=1}^{H} \sum_{m=1}^{M} n_{mh} \gamma_{mhi} \right) \right] \}
\]
\[ x \left( W_m + \sum_{p=1}^{M} W_p \left( k_{mhi}^p + \delta_{mhi}^p \right) \right) \left( Y_{mhi} - \mu_h \right) \]

\[ \left| n_{mhi}'s, r_{mhi}'s, t_{mhi}'s, \gamma_{mhi}'s, k_{mhi}'s, \delta_{mhi}'s \right| \left| n_{mhi}'s, r_{mhi}'s \right] \]

\[ = V_1 \left\{ E_2 (\bar{y}_{st} - \bar{\mu}) \right\} \]

\[ + E_1 \{ E_3 \left[ \sum_{h=1}^{H} \sigma_h^2 \left( \sum_{m=1}^{M} \sum_{i=1}^{n_{mhi}} \gamma_{mhi} \right) \right] \}

\[ \times \left( W_m + \sum_{p=1}^{M} W_p \left( k_{mhi}^p + \delta_{mhi}^p \right) \right)^2 \left| n_{mhi}'s, r_{mhi}'s \right] \]

\[ = V_1 \left\{ E_2 (\bar{y}_{st} - \bar{\mu}) \right\} \]

\[ + E_1 \{ \sum_{h=1}^{H} \sigma_h^2 \}

\[ \sum_{m=1}^{M} n_{mhi} W_m^2 + 2 \sum_{m=1}^{M} \sum_{p=1}^{M} \frac{W_m W_p}{r_h} (r_{mhi} m_{ph} + k_{mhi} t_{ph}) \]

\[ + \frac{k_{.h} - 1}{r_h k_{.h} - 1} \left( \sum_{m=1}^{M} W_m k_{mhi} \right)^2 - \sum_{m=1}^{M} W_m^2 k_{mhi} \} \]. \tag{4.42} \]

The result (4.40) follows from (4.37) since, by (4.37),
\[ V(\bar{y}_{st} - \bar{\mu}) = V_1\{ E_2(\bar{y}_{st} - \bar{\mu}) \} + E_1\{ \sum_{h=1}^{H} \sigma_h^2 \Sigma_{m=1}^{M} n_{mh} W_m^2 \} . \] (4.43)

For (4.41), we have that the

\[ V(\bar{y}_{st}^* - \bar{\mu}) = V[ (\bar{y}_{st}^* - \bar{Y}) + (Y - \bar{\mu}) ] \]

\[ = V(\bar{y}_{st}^* - \bar{Y}) + 2 \text{Cov}(\bar{y}_{st}^* - \bar{Y}, Y - \bar{\mu}) + V(Y - \bar{\mu}) , \] (4.44)

so,

\[ V(\bar{y}_{st}^* - \bar{Y}) = V(\bar{y}_{st}^* - \bar{\mu}) - V(Y - \bar{\mu}) \]

\[ - 2 \text{Cov}(\bar{y}_{st}^* - \bar{Y}, Y - \bar{\mu}) . \] (4.45)

We have

\[ \text{Cov}(\bar{y}_{st}^* - \bar{Y}, Y - \bar{\mu}) = E( Y E( \bar{y}_{st}^* | FP ) - Y^2 ) \]

\[ = E\{ \sum_{h=1}^{H} \sum_{p=1}^{M} m_{ph} W_p \Sigma_{m=1}^{M} \frac{r_{mh}}{r} \Sigma_{m=1}^{M} Y_{mh} Y_p \} \]

\[ | n_{mh}'s, r_{mh}'s \} . \]
Also,

\[ V(\overline{y} - \overline{\mu}) = \sum_{m=1}^{M} \frac{N \cdot m - E(S_{m}^{2})}{N^{2}} \]

(4.47)

(see proof of Lemma 4.3). The result in expression (4.41) follows from substituting (4.46) and (4.47) into (4.45).

We give two corollaries that refer to less general conditions.

**Corollary 1:** If the assumptions of Theorem 4.2 hold and if the number of missing observations is less than the number of responding observations in all imputation classes, (4.40) reduces to

\[ V(\overline{y}_{st}^{*} - \overline{\mu}) = V(\overline{y}_{st} - \overline{\mu}) \]

\[ + 2 \sum_{h=1}^{H} \sigma_{h}^{2} \text{E}(r_{h}^{-1}(M_{m=1}^{M} W_{m} r_{mh})(M_{p=1}^{P} W_{p} m_{ph})) \]

(4.48)
Corollary 2: If the conditions of Theorem 4.2 hold and if the number of strata is one, expression (4.40) reduces to

\[ V(\bar{y}_{st}^* - \mu) = V(\bar{y} - \mu) \]

\[ + \sum_{h=1}^{H} \sigma_h^2 E\left( \frac{k_h n_h + (k_h + 2)t_h}{n^2} \right), \quad (4.49) \]

where the notation is that of Chapter 3. Expression (4.49) agrees with expression (3.89) in Section 3.2.5.

We now present the variance of \(\bar{y}_{st}^* - \bar{y}\) taken to level 2 conditioning.

Theorem 4.4: Let the assumptions of Theorem 4.2 hold. Then

\[ V(\bar{y}_{st}^* - \mu) = V(\bar{y}_{st} - \mu) \]

\[ + \sum_{h=1}^{H} \sigma_h^2 E\{ \left[ \frac{k_h n_h + (k_h + 2)t_h}{n_h (n_h - 1)} \right] \} \]

\[ \times \left\{ \left( \sum_{m=1}^{M} n_{mh} W_m \right)^2 - \sum_{m=1}^{M} n_{mh} W_m^2 \right\} \]. \quad (4.50) \]

Proof: To find expression (4.50), we take the expectation of the second term in (4.40) at level 2 conditioning. In order to take the expectation, we need to know the distributions of the \(r_{mh}\)'s, \(m_{mh}\)'s, \(k_{mh}\)'s, and \(t_{mh}\)'s. The distributions are as follows: \(r_{mh}\) has an hypergeometric distribution with sample size equal to
r_{h}', subpopulation size equal to n_{mh}', and population size equal to n_{h}; m_{mh} = n_{mh} - r_{mh}', so m_{mh} has an hypergeometric distribution with sample size equal to m_{h}', subpopulation size equal to n_{mh}', and population size equal to n_{h}; k_{mh} has an hypergeometric distribution with sample size equal to k_{h} r_{h}', subpopulation size equal to m_{mh}, and population size equal to m_{h}'; t_{mh} = m_{mh} - k_{mh}'. It follows that

\[ E_2 \{ \sum_{h=1}^{M} \sigma_{h}^2 \left[ \sum_{m=1}^{M} \frac{W_m W_p}{r_{h} m} (r_{mh} m_{ph} + k_{mh} t_{ph}) \right. \]

\[ + \frac{k_{h} - 1}{r_{h} k_{h} - 1} \left( \sum_{m=1}^{M} W_m k_{mh} \right)^2 - \sum_{m=1}^{M} W_m^2 k_{mh} \} \]

\[ \mid n_{mh}'s, r_{h}'s \}

\[ = \sum_{h=1}^{L} \sigma_{h}^2 \left\{ \frac{2 m_{h}}{n_{h} (n_{h} - 1)} \left[ \left( \sum_{m=1}^{M} n_{mh} W_m \right) - \sum_{m=1}^{M} n_{mh} W_m^2 \right] \right. \]

\[ + \frac{2 k_{h} m_{h} - r_{h} k_{h} (k_{h} + 1)}{m_{h} (m_{h} - 1)} \]

\[ \times E_2 \{ \left( \sum_{m=1}^{M} m_{mh} W_m \right)^2 - \sum_{m=1}^{M} m_{mh} W_m^2 \mid n_{mh}'s, r_{h}'s \} \}

\[ = \sum_{h=1}^{L} \sigma_{h}^2 \left\{ \frac{k_{h} n_{h} + (k_{h} + 2) t_{h}}{n_{h} (n_{h} - 1)} \right. \]

\[ \times \left[ \left( \sum_{m=1}^{M} n_{mh} W_m \right)^2 - \sum_{m=1}^{M} n_{mh} W_m^2 \right] \}. \quad (4.51) \]
The result follows by substituting expression (4.51) into expression (4.49).

We present a corollary for less general conditions.

**Corollary:** Let the assumptions of Theorem 4.2 hold. Let the number of respondents be greater than the number of missing observations in each imputation class. Then

\[
V(\bar{y}_{st}^* - \bar{\mu}) = V(\bar{y}_{st} - \bar{\mu}) + \sum_{h=1}^{L} \sigma^2_{h} \text{E}\left\{ \frac{2 \sum_{m=h}^{M} W_m}{n_{m,h}(n_{m,h} - 1)} \right\} \\
\times \left[ (\sum_{m=1}^{M} n_{m,h} W_m)^2 - \sum_{m=1}^{M} n_{m,h} W_m^2 \right].
\]

(4.52)

In Theorem 4.5 we develop the expression for \( V(\bar{y}_{st}^* - \bar{\mu}) \) at level 4 conditioning. For this form of the variance we need to know which missing value the donor donates to.

**Theorem 4.5:** Let the assumptions of Theorem 4.2 hold. Let

\[
WT_{mhi} = \gamma_{mhi} \left[ W_m + \sum_{p=1}^{M} W_p \left( \delta_{mhi}^p + k_{mhi}^p \right) \right].
\]

(4.53)

Then \( WT_{mhi} \) equals the sum of the weights on unit mhi in \( \bar{y}_{st}^* \) when unit mhi responds and equals zero when unit mhi does not respond. Then
\[
V(\bar{y}_{st}^* - \mu) = V(\bar{y}_{st} - \mu) + \sum_{h=1}^{H} \sigma^2_h E \left[ \sum_{m=1}^{M} \sum_{i=1}^{n_{mh}} (W_{mhi} - W_m)^2 \right] \\
+ 2 W_m (W_{mhi} - W_m)
\]

(4.54a)

\[
= V(\bar{y}_{st} - \mu) + \sum_{h=1}^{L} \sigma^2_h E \left( \sum_{m=1}^{M} \sum_{i=1}^{n_{mh}} W_{mhi}^2 - W_m^2 \right)
\]

(4.54b)

Proof: We have that

\[
\bar{y}_{st}^* = \bar{y}_{st} + (\bar{y}_{st}^* - \bar{y}_{st}),
\]

(4.55)

so that

\[
V(\bar{y}_{st}^* - \mu) = V(\bar{y}_{st} - \mu) + 2 \text{Cov}(\bar{y}_{st}^* - \bar{y}_{st}, \bar{y}_{st} - \mu) \\
+ V(\bar{y}_{st}^* - \bar{y}_{st})
\]

(4.56)

Now

\[
\text{Cov}(\bar{y}_{st}^* - \bar{y}_{st}, \bar{y}_{st} - \mu)
\]

\[
= E\left[ E\left[ \sum_{h=1}^{H} \sum_{m=1}^{M} \sum_{i=1}^{n_{mh}} (W_{mhi} - W_m)(Y_{mhi} - \mu_h) \right] \right]
\]
Since the $Y_{mhi}$'s are identically and independently distributed within imputation classes and independent between classes under the superpopulation model,

$$\text{Cov}(\overline{y}_{st} - \mu, \overline{y}^*_{st} - \overline{y}_{st})$$

$$= \sum_{h=1}^{H} \sigma^2_h \mathbb{E}\left[ \sum_{m=1}^{M} \sum_{i=1}^{n_{mh}} (WT_{mhi} - W_m) (Y_{mhi} - \mu_h) \right].$$

(4.58)

Likewise, since the $Y_{mhi}$'s are identically and independently distributed within imputation classes and independent between classes under the superpopulation model,

$$\text{V}(\overline{y}^*_{st} - \overline{y}_{st})$$

$$= \mathbb{E}\left\{ \mathbb{E}\left[ \sum_{h=1}^{H} \sum_{m=1}^{M} \sum_{i=1}^{n_{mh}} (WT_{mhi} - W_m) (Y_{mhi} - \mu_h) \right] \right\}$$

$$\times \left[ \sum_{g=1}^{H} \sum_{p=1}^{M} \sum_{j=1}^{n_{pg}} (WT_{pgj} - W_p) (Y_{pgj} - \mu_g) \right]$$

$$| WT_{mhi}'s, n_{mh}'s \}.$$
Substituting expressions (4.58) and (4.59) into (4.56) gives (4.54a). Note that

\[
( \text{WT}^\text{mhi} - \text{W}_m )^2 + 2 \text{W}_m ( \text{WT}^\text{mhi} - \text{W}_m ) = \text{WT}^2\text{mhi} - \text{W}_m^2
\]

and that

\[
V_4( \overline{y}_{st} - \overline{\mu} | n_{mh}', r_{mh}', \gamma_{mhi}', k_{mhi}', \delta_{mhi}' )
\]

\[
= \sum_{h=1}^{H} \sigma_h^2 \sum_{m=1}^{M} \sum_{i=1}^{n_{mh}} (\text{WT}^\text{mhi} - \text{W}_m)^2
\]

(4.60)

and that

\[
V_4( \overline{y}_{st} - \overline{\mu} | n_{mh}', r_{mh}', \gamma_{mhi}', k_{mhi}', \delta_{mhi}' )
\]

\[
= \sum_{h=1}^{L} \sigma_h^2 \sum_{m=1}^{M} \sum_{i=1}^{n_{mh}} \text{WT}^2\text{mhi}
\]

(4.61)

so that the expression in (4.54b) is the expression for the \( V( \overline{y}_{st} - \overline{\mu} ) \) evaluated at level 4 conditioning.
4.7. Estimation of the variance of the imputed mean

In Theorem 4.6 we give the expectations of quantities used in constructing estimates of the variance of $\bar{y}_{st}^*$. 

Theorem 4.6: Let the assumptions of Theorem 4.2 hold. Let $s_{m.}^2$, $s_{r.h}^2$, and $s_{m.}^2$ be as defined in (4.9), (4.10), and (4.11). Then

$$E(s_{r.h}^2) = \sigma_{r.h}^2,$$  

$$E(s_{m.}^2) = \frac{1}{n_{m.} - 1} \sum_{h=1}^{H} E\left( n_{m.h} \sigma_{m.h}^2 \right),$$  

$$+ \left( \mu_{m.h} - \sum_{g=1}^{R} \frac{n_{m.g}}{n_{m.}} \mu_{m.g} \right)^2 - \sum_{h=1}^{H} \frac{n_{m.h}}{n_{m.}} \sigma_{m.h}^2,$$  

(4.62)

and

$$E(s_{m.}^2) = E(s_{m.}^2) - \frac{1}{n_{m.}(n_{m.} - 1)} \sum_{h=1}^{H} \sigma_{h}^2 \left[ \frac{r_{m.h}m_{m.h} + k_{m.h}m_{m.h}}{r_{h}} \right]$$  

$$+ \left( k_{m.h}^2 - k_{m.h} \right) \left( \frac{k_{h}^{-1}}{k_{h}r_{h}^{-1}} \right).$$  

(4.63)

$$E(s_{m.}^2) = E(s_{m.}^2) - \frac{1}{n_{m.}(n_{m.} - 1)} \sum_{h=1}^{L} \sigma_{h}^2 \left[ n_{m.h} \left( n_{m.h} - 1 \right) \right]$$  

(4.64)
\[
\frac{k \cdot h \cdot n \cdot h + (k \cdot h + 2) \cdot t \cdot h}{n \cdot h \cdot (n \cdot h - 1)}.
\] (4.65)

**Proof:** For expression (4.62), since \( \{ Y_{mhi} \}_{i=1,...,r_{mh}, m=1,...,M} \) is an independently, identically distributed sample from the portion of the population in imputation class \( h \), \( \mathbb{E}(s_{r,h}^2) = \sigma_{h}^2 \).

To find the expectation of \( s_m^2 \), let us first look at the expectation of \( s_{m}^2 \), the sample variance with no observations missing. For expression (4.63), first

\[
s_m^2 = \frac{1}{n_m - 1} \sum_{h=1}^{H} \sum_{i=1}^{n_{mh}} (Y_{mhi} - \bar{Y}_m)^2
\]

\[
= \frac{1}{n_m - 1} \sum_{h=1}^{H} \sum_{i=1}^{n_{mh}} \left[ (Y_{mhi} - \mu_{h})^2 + (\mu_{h} - \frac{H}{g=1} \frac{n_{mg}}{n_m} \mu_{g})^2 \right]
\]

\[
+ 2 (Y_{mhi} - \mu_{h}) (\mu_{h} - \frac{H}{g=1} \frac{n_{mg}}{n_m} \mu_{g})
\]

\[
- 2 (\mu_{h} - \frac{H}{g=1} \frac{n_{mg}}{n_m} \mu_{g}) (\frac{H}{g=1} \frac{n_{mg}}{n_m} \frac{Y_{mgj} - \mu_{g}}{n_m})
\]

\[
- 2 (Y_{mhi} - \mu_{h}) (\frac{H}{g=1} \frac{n_{mg}}{n_m} \frac{Y_{mgj} - \mu_{g}}{n_m})
\]

\[
+ (\frac{H}{g=1} \frac{n_{mg}}{n_m} \frac{Y_{mgj} - \mu_{g}}{n_m})^2 \] (4.66)
Then, since the $Y_{mhi}$'s are identically and independently distributed within imputation classes and independent between them,

$$
E(s_m^2 | n_m, s) = \frac{1}{n_m - 1} \{ \sum_{h=1}^{H} n_{mh} \left[ \sigma_{h}^2 
+ (\mu_h - \sum_{g=1}^{G} \frac{n_{mg}}{n_m} \mu_g)^2 \right] - \sum_{h=1}^{H} \frac{n_{mh}}{n_m} \sigma_{h}^2 \},
$$

(4.67)

and expression (4.63) follows from (4.67).

Now, let us look at the expectation of $s_m^2$. For expression (4.64), first

$$
\bar{y}_m^* = \frac{1}{n_m} \sum_{h=1}^{H} \left[ \sum_{i=1}^{n_{mh}} \gamma_{mhi} Y_{mhi} 
+ \sum_{p=1}^{P} \sum_{j=1}^{n_{phj}} \gamma_{phj} \left( k_{phj}^m + c_{phj}^m \right) Y_{phj} \right],
$$

(4.68)

and

$$
s_m^2 = \frac{1}{n_m - 1} \sum_{h=1}^{H} \sum_{i=1}^{n_{mh}} \left[ (Y_{mhi} - \bar{y}_m^*)^2 - (\mu_h - \sum_{g=1}^{G} \frac{n_{mg}}{n_m} \mu_g)^2 \right]

+ 2 (Y_{mhi} - \mu_h) (\mu_h - \sum_{g=1}^{G} \frac{n_{mg}}{n_m} \mu_g)
$$
\[-2(Y_{mhi}^* - \mu_{h}) \{ \frac{H}{n_m} \left[ \frac{1}{n_m} \sum_{j=1}^{n_{mg}} \gamma_{mgj} (Y_{mgj} - \mu_{g}) \right] \}

\[+ \sum_{p=1}^{M} \sum_{k=1}^{n_{pg}} \gamma_{pgk} \left( k_{pgk}^{m} + \delta_{pgk}^{m} \right) (Y_{pgk} - \mu_{g}) \}}\]

\[+ \left\{ \frac{H}{n_m} \left[ \frac{1}{n_m} \sum_{j=1}^{n_{mg}} \gamma_{mgj} (Y_{mgj} - \mu_{g}) \right] \right\} \]

\[+ \sum_{p=1}^{M} \sum_{k=1}^{n_{pg}} \gamma_{pgk} \left( k_{pgk}^{m} + \delta_{pgk}^{m} \right) (Y_{pgk} - \mu_{g}) \right\}^{2}\]  \(4.69\)

Since the \(Y_{mhi}\)'s are identically and independently distributed within imputation classes and independent between them,

\[E_4(s_m^{2*} | n_{mh}'s, r_{mh}'s, t_{mh}'s, \gamma_{mhi}'s, k_{mhi}'s, \delta_{mhi}'s)\]

\[= \frac{1}{n_m - 1} \left\{ \frac{H}{n_m} \sum_{h=1}^{n_{mh}} \left[ \sigma_{h}^{2} + (\mu_{h} - \left( \frac{H}{n_m} \sum_{g=1}^{n_{mg}} \mu_{g} \right)^{2} \right] \right\}

- \frac{1}{n_m} \sum_{g=1}^{H} \left[ \sum_{i=1}^{n_{mg}} \gamma_{mgi}^{2} \sigma_{g}^{2} + \sum_{i=1}^{n_{mg}} \gamma_{mgi}^{2} \left( k_{mgi}^{m} + \delta_{mgi}^{m} \right) \sigma_{g}^{2} \right]
so,

\[
E_3( s_{m}^2 | n_{mh}'s, r_{mh}'s, t_{mh}'s ) = \frac{1}{n_m - 1} \left\{ \sum_{h=1}^{H} n_{mh} \left[ \sigma_h^2 + \left( \mu_h - \frac{H}{g=1} n_{mg} \mu_g \right)^2 \right] \right. \\
- \frac{1}{n_m} \sum_{g=1}^{H} n_{mh} \sigma_g^2 + 2 \left( \frac{r_{mg} m_{mg} + k_{mg} t_{mg}}{r_g} \right) \sigma_g^2 \\
+ \left. \left( \frac{k^{2}_{mg} - k_{mg}}{k_{g} r_{g}} \right) \left( \frac{r_g - 1}{k_{g} - 1} \right) \sigma_g^2 \right\}, 
\]  

and expression (4.64) follows from taking the expectation of (4.71).

Expression (4.65) is found by taking the expectation of the second term in expression (4.64) with respect to level 2 conditioning.

\[
E_2\left\{ - \frac{1}{n_m} \left( \frac{n_m - 1}{n_m} \right) \sum_{h=1}^{H} \sigma_h^2 \left[ 2 \frac{r_{mh} m_{mh} + k_{mh} t_{mh}}{r_h} \right] \right. \\
+ \frac{k_{h} - 1}{r_{h} k_{h} - 1} \left( \frac{k^{2}_{mh} - k_{mh}}{k_{h} - 1} \right) \left[ n_{mh}'s, r_{h}'s \right] \right\} \\
= - \frac{1}{n_m} \left( \frac{n_m - 1}{n_m} \right) \sum_{h=1}^{H} \sigma_h^2 n_{mh} \left( n_{mh} - 1 \right)
\]
Expression (4.65) follows from taking the expectation of (4.72) and substituting it into expression (4.64).

We present two corollaries for less general conditions.

**Corollary 1:** Under the assumptions of Theorem 4.2, if the number of missing is less than the number responding in all imputation classes, expression (4.64) becomes

\[
E(\sigma^2_{\text{m.}}) = E(\sigma^2_{\text{m.}}) - \frac{1}{n_m(n_m - 1)} \sum_{h=1}^{H} \sigma^2_h E[2 \frac{m_h m_{mh}}{r_h}].
\]  

(4.73)

\[
= E(\sigma^2_{\text{m.}}) - \frac{1}{n_m(n_m - 1)} \sum_{h=1}^{H} \sigma^2_h E[2 \frac{m_h}{n_h(n_h - 1)}].
\]  

(4.74)

**Corollary 2:** Under the assumptions of Theorem 4.2, if the number of strata equals one, expression (4.64) becomes

\[
E(\sigma^2_{\text{m.}}) = E(\sigma^2_{\text{m.}})
\]
where the notation is that of Chapter 3. Expression (4.75) agrees with expression (3.90) in Section 3.2.6.

We present three estimators of the variance of $\bar{y}_{st}^* - \bar{\mu}$. First we look at the variance of $\bar{y}_{st} - \bar{\mu}$.

Lemma 4.3: Under stratified random sampling and the superpopulation structure of Theorem 4.2,

$$\text{Var}(\bar{y}_{st} - \bar{\mu}) = \sum_{m=1}^{M} \frac{N_m^2}{N^2} \frac{\sigma_m^2}{n_m},$$

where the $\sigma_m^2$ is the variance of the $Y$-values in stratum $m$ of the superpopulation.

Proof: Let

$$\bar{\mu}_m = \sum_{h=1}^{H} \frac{N_{mh}}{N_m} \mu_h$$

be the superpopulation mean of stratum $m$. Let

$$\bar{Y}_m = \sum_{h=1}^{H} \frac{N_{mh}}{N_m} Y_{mh}$$

(4.78)
be the finite population mean of stratum $m$. The variance of $\bar{y}_{st} - \bar{\mu}$ is

$$V(\bar{y}_{st} - \bar{\mu}) = E[V(\bar{y}_{st} | FP)] + V[E(\bar{y}_{st} | FP)]$$

(4.79)

and

$$E[V(\bar{y}_{st} | FP)]$$

$$= \frac{M}{\Sigma m=1} \frac{N_m^2}{N^2} \frac{1}{m} \frac{1}{N} \left( \frac{1}{n_m} - \frac{1}{N} \right),$$

(4.80)

(see Cochran (1977, p. 92)), where $E(S_m^2) = \sigma_m^2$ under the assumption of how the finite population was formed from the superpopulation. Also

$$V[E(\bar{y}_{st} | FP)] = V(Y) = V\left( \frac{M}{\Sigma m=1} \frac{N_m}{N} Y_m \right)$$

$$= E\left[ \frac{M}{\Sigma m=1} \frac{M}{\Sigma p=1} \frac{N_m N_p}{N^2} (Y_m - \bar{\mu}_m) (Y_p - \bar{\mu}_p) \right]$$

$$= \frac{M}{\Sigma m=1} \frac{N_m^2}{N^2} \frac{E(S_m^2)}{N_m}$$

(4.81)

since the $Y_m$ are independent between strata and since $S_m^2$ is an unbiased estimator of the variance in stratum $m$. Substituting (4.80) and (4.81) into (4.79), we get the result. \#
The next three theorems give estimators of the variance of $\overline{y}_{st} - \mu$.

**Theorem 4.7:** Let

\[
\hat{V}_1(\overline{y}_{st} - \mu) = \frac{M}{\Sigma} \frac{N^2}{n_m} \frac{s_m^2}{n_m} + \Sigma_{h=1}^{H} s_{r,h}^2 \left\{ \frac{M}{\Sigma} \frac{W^2}{m} \frac{1}{n_m} - 1 \right\} \left[ 2 \frac{r_{mh}m_{mh} + k_{mh}t_{mh}}{r.h} \right.
\]
\[
\left. + \frac{(k_{mh}^2 - k_{mh})(k_{h} - 1)}{k_{h}r.h - 1} \right\}
\]
\[
+ \left\{ \frac{M}{\Sigma} \frac{M}{\Sigma} \frac{2 W_m}{W_p} \frac{r_{mh}m_{ph} + k_{mh}t_{ph}}{r.h} \right. \]
\[
\left. + \frac{k_{h} - 1}{r.h k_{h} - 1} \left[ \left( \frac{M}{\Sigma} \frac{W_m k_{mh}}{m=1} \right)^2 \right. \right.
\]
\[
\left. \left. - \left( \frac{M}{\Sigma} \frac{W_m k_{mh}}{m=1} \right) \right] \right\}.
\]

(4.82)

Then, under the assumptions of Theorem 4.2, $\hat{V}_1(\overline{y}_{st} - \mu)$ is an unbiased estimator of $V(\overline{y}_{st} - \mu)$. 
Proof: The quantity $s_{m.}^2$ is an unbiased estimator of $S_{m.}^2$ in a finite population and

$$E(s_{m.}^2) = E\left[ s_{m.}^2 - H \sum_{h=1}^{N} s_{r.h}^2 \left\{ M \sum_{m=1}^{M} \frac{1}{n_m} \frac{1}{n_m - 1} \right\} \right]$$

by Theorem 4.6. It follows that the first term in (4.82) plus the first term in curly brackets in the second term in (4.82) is an unbiased estimator of

$$\frac{M}{\Sigma m=1} \frac{N_{m.}^2}{N^2} \frac{E(S_{m.}^2)}{n_m}.$$  

which equals the $V(\bar{y}_{st} - \bar{\mu})$ by Lemma 4.3. By Theorem 4.3, the second term in curly brackets in the second term in (4.82) is an unbiased estimator of the amount to be added to the $V(\bar{y}_{st} - \bar{\mu})$ to correct for imputation.

We present two corollaries. The first of these corollaries gives the estimator of the variance of $\bar{y}_{st}^*$ with the finite population correction factor. The second of these corollaries gives the variance of $\bar{y}_{st}^* - \bar{\mu}$ when the number of respondents is greater than the number of missing in all imputation classes.
these corollaries gives the variance of $\bar{y}_{st}^* - \bar{y}$ when the number of respondents is greater than the number of missing in all imputation classes.

**Corollary 1:** Under the assumptions of Theorem 4.2, we have that

$$V_1(\bar{y}_{st}^* - \bar{y}) = \sum_{m=1}^{M} \frac{N_m^2}{N^2} \frac{s_m^2*}{n_m} \left( 1 - \frac{n_m}{N_m} \right)$$

$$+ \frac{H}{r} \sum_{h=1}^{s_{r.h}} \left[ \left\{ \sum_{m=1}^{M} \sum_{p=1}^{M} \frac{2 W_m W_p}{s_{r.h}} \frac{r_{m.p h} m_{p h} + k_{m h}^t m_{h} + k_{m h} t m_{h}}{r \cdot h} \right) \right.$$  

$$\left. + \frac{(k_{m h}^2 - k_{m h})(k_{h} - 1)}{k_{h} r \cdot h - 1} \right]$$

$$+ \left\{ \sum_{m=1}^{M} \sum_{p=1}^{M} \frac{2 W_m W_p}{s_{r.h}} \frac{r_{m.p h} m_{p h} + k_{m h}^t p h}{r \cdot h} \right) \right.$$  

$$\left. + \frac{k_{h} - 1}{r \cdot h k_{h} - 1} \left[ \left( \sum_{m=1}^{M} W_m k_{m h} \right)^2 \right. \right.$$  

$$\left. \left. - \sum_{m=1}^{M} W_m^2 k_{m h} \right] \right)$$  

(4.85)

is an unbiased estimator of $V(\bar{y}_{st}^* - \bar{y})$. 
Proof: The result follows from the second result in Theorem 4.3, since an unbiased estimator of

\[ V(\overline{y}_{st} - \bar{\mu}) = \sum_{m=1}^{M} \frac{N_{m}}{N^2} E(S_{m}^2) \]  \hspace{1cm} (4.86)

is

\[ \sum_{m=1}^{M} \frac{N_{m}^2}{N^2} \frac{s_{m}^2}{n_{m}} \]

\[ + \sum_{h=1}^{H} \sum_{r,h} s_{r,h} \left\{ \sum_{m=1}^{M} \frac{N_{m}}{N^2} \frac{1}{n_{m}} \left[ \frac{2r_{mh}m_{mh} + k_{mh}m_{mh}}{r_{h}} \right] \right. \]

\[ + \left. \frac{(k_{mh}^2 - k_{mh})(k_{h} - 1)}{k_{h}r_{h} - 1} \right\} \]

\[- \left[ \sum_{m=1}^{M} \frac{N_{m}}{N^2} s_{m}^2 \right] \]

\[ + \sum_{h=1}^{H} \sum_{r,h} s_{r,h} \left\{ \sum_{m=1}^{M} \frac{N_{m}}{N^2} \frac{1}{n_{m}} \frac{1}{n_{m} - 1} \left[ \frac{2r_{mh}m_{mh} + k_{mh}m_{mh}}{r_{h}} \right] \right. \]

\[ + \left. \frac{(k_{mh}^2 - k_{mh})(k_{h} - 1)}{k_{h}r_{h} - 1} \right\} \] . \hspace{1cm} (4.87)
Corollary 2: Under the assumptions of Theorem 4.2 and given that the number of missing observations in less than the number of responding observations in all imputation classes, the estimator

\[
\hat{V}_1(\bar{y}_{st}^* - \bar{\mu}) = \sum_{m=1}^{M} \frac{N_m^2}{N^2}\frac{s_{m.}^2}{n_m}.
\]

\[
+ \frac{H}{\sum_{h=1}^{H}} s_{r.h}^2 \left[ \left\{ \sum_{m=1}^{M} W_m^2 \frac{2}{n_m - 1} \frac{r_{mh}^m}{r_{.h}} \right\} \right] + \left\{ \sum_{m=1}^{M} \sum_{p=1}^{M} W_m W_p \frac{r_{mh}^m}{r_{.h}} \right\} \right]
\]

is unbiased for \(V(\bar{y}_{st}^* - \bar{\mu})\).

Theorem 4.8: Under the assumptions of Theorem 4.2,

\[
\hat{V}_2(\bar{y}_{st}^* - \bar{\mu}) = \sum_{m=1}^{M} \frac{N_m^2}{N^2}\frac{s_{m.}^2}{n_m}.
\]

\[
+ \frac{H}{\sum_{h=1}^{H}} s_{r.h}^2 \left[ \frac{k_h n_h + (k_h + 2) t_h}{n_h (n_h - 1)} \right] \left[ \sum_{m=1}^{M} \frac{n_{mh} (n_{mh} - 1)}{n_m - 1} W_m^2 \right] + \left( \sum_{m=1}^{M} n_{mh} W_m \right)^2 - \sum_{m=1}^{M} n_{mh} W_m^2 \right] \]

is an unbiased estimator of \(V(\bar{y}_{st}^* - \bar{\mu})\).
Proof: By Theorem 4.6 and Lemma 4.3 and since \( s_m^2 \) is an unbiased estimator of \( S_m^2 \) in a finite population,

\[
\begin{align*}
\frac{M}{N} \sum_{m=1}^{N} \frac{N_m^2}{n_m^2} \cdot s_m^2 + \frac{H}{n_h^2} \sum_{h=1}^{k} \frac{k_h n_h + (k_h + 2)t_h}{n_h(n_h - 1)} \\
\times \left[ \frac{M}{n_m} \frac{n_m - 1}{n_m} W_m^2 \right]
\end{align*}
\]

is an unbiased estimator of the \( V(\bar{y}_{st} - \bar{y}) \). By Theorem 4.4 and Lemma 4.3,

\[
\begin{align*}
\frac{H}{n_h^2} \sum_{h=1}^{k} s_{r,h}^2 & \left[ \frac{k_h n_h + (k_h + 2)t_h}{n_h(n_h - 1)} \right] \\
\times \left[ \left( \frac{M}{m=1} n_{mh} W_m \right)^2 - \frac{M}{m=1} n_{mh} W_m^2 \right]
\end{align*}
\]

is an unbiased estimator of the amount to be added to the \( V(\bar{y}_{st} - \bar{y}) \) to correct for the imputation.

The following corollary refers to the case where missing is less than responding in all imputation classes.

**Corollary:** Under the assumptions of Theorem 4.2, if the number of respondents is greater than the number of missing observations in all imputation classes, then
\[ \hat{V}_2(\bar{y}_{st} - \bar{\mu}) = \frac{M}{\Sigma} \frac{N_{m}^{2}}{N^2} \frac{s_{m}^{2*}}{n_{m}}. \]

\[ + \sum_{h=1}^{H} s_{r,h}^{2} \left[ 2 \frac{m_{h}}{n_{h}(n_{h} - 1)} \right] \frac{M}{\Sigma} \frac{n_{mh}(n_{mh} - 1)}{n_{m} - 1} W_{m}^{2} \]

\[ + \left( \frac{M}{\Sigma} \sum_{m=1}^{M} n_{mh} W_{m} \right)^{2} - \left( \frac{M}{\Sigma} \sum_{m=1}^{M} n_{mh} W_{m}^{2} \right) \]

(4.92)

is an unbiased estimator of \( V(\bar{y}_{st} - \bar{\mu}) \).

**Theorem 4.9:** Under the assumptions of Theorem 4.2,

\[ \hat{V}_3(\bar{y}_{st} - \bar{\mu}) = \frac{M}{\Sigma} \frac{N_{m}^{2}}{N^2} \frac{s_{m}^{2*}}{n_{m}}. \]

\[ + \sum_{h=1}^{H} s_{r,h}^{2} \left[ \frac{M}{\Sigma} \sum_{m=1}^{M} \sum_{i=1}^{n_{mh}} 2 W_{m} (W_{m}^{T} W_{m} - W_{m}) \right] \]

\[ + \left( W_{m}^{T} W_{m} - W_{m}^{2} \right) \]

(4.93)

is a biased estimator of \( V(\bar{y}_{st} - \bar{\mu}) \) with a bias of

\[ E\{ \hat{V}_2(\bar{y}_{st} - \bar{\mu}) - V(\bar{y}_{st} - \bar{\mu}) \} \]

\[ = - \sum_{h=1}^{H} \sigma_{h}^{2} \sum_{m=1}^{M} W_{m}^{2} \frac{1}{n_{m} - 1} E[ 2 \frac{r_{mh}^{m_{mh}} + k_{mh}^{t_{mh}}}{r_{h}} ], \]
\[
+ \frac{\left( k_{mh}^2 - k_{mh} \right) \left( k_{,h} - 1 \right)}{k_{,h}^2} \frac{1}{n_{,h} - 1}
\]

\[
= - \sum_{h=1}^{H} \sigma_h^2 \sum_{m=1}^{M} \frac{W_m^2}{n_{m} - 1} \frac{1}{n_{,h} - 1} \frac{k_{,h} n_{,h}}{n_{,h}} + \left( k_{,h} + 2 \right) t_{,h} \]

(4.94)

Proof: By Theorem 4.6 and Lemma 4.3 and since \( s_m^2 \) is an unbiased estimator of \( S_m^2 \) in a finite population, the bias in

\[
\frac{M}{\Sigma} \frac{N_m^2}{N^2} \frac{s_m^2}{n_m}.
\]

(4.95)

as an estimator of \( V(\overline{y}_{st} - \overline{\mu}) \) is expression (4.94). By Theorem 4.5, the second term in (4.93) is an unbiased estimator of the amount to be added to \( V(\overline{y}_{st} - \overline{\mu}) \) to adjust for the increase in due to the imputation.

#

Corollary: Under the assumptions of Theorem 4.9 ,

\[
\hat{V}_4(\overline{y}_{st}^* - \overline{\mu}) = \frac{M}{\Sigma} \frac{N_m^2}{N^2} \frac{s_m^2}{n_m}.
\]
is a biased estimator of $V(\bar{y}_{st} - \bar{\mu})$ with the bias (4.94) given in Theorem 4.9.

**Proof:** Since the $Y^*_{mhi}$'s are individually random draws from the imputation class $h$ in a sample of size $r_h$:

$$E( Y^*_{mhi} - \bar{y}_{r.h} )^2 = E( \frac{r_h - 1}{r_h} ) \sigma^2_h ,$$

and the result follows from Theorem 4.9.

The bias in Theorem 4.9 gets smaller as the sample size gets larger and tends to be very small with moderately large sample sizes in each stratum. Expression (4.89) in Theorem 4.8 is conditional on the number of observations in each imputation class–stratum subgroup being fixed and on the number of respondents in each imputation class being fixed. Expression (4.89) has a smaller variance than expression (4.82) in Theorem 4.7 which is conditional on the sampling pattern, response pattern, and imputation pattern being fixed. In Theorem 4.7, averaging is done over the possible choices for putting imputed values with donors for the fixed imputation pattern. Expression (4.82) has a slightly smaller variance than expression (4.93) in Theorem 4.9, which holds the choice of donors for the imputed
values fixed. Expression (4.96) in the corollary to Theorem 4.9 has a larger variance than expression (4.93) in Theorem 4.9, since $\sigma_h^2$ is estimated less efficiently in the corollary.
5. IMPUTATION UNDER GENERAL SAMPLING SCHEMES

In this chapter we present methods for random imputation applicable to any sampling method that allows every member of the population to be sampled with probability greater than zero. We also show how variance and covariance estimators for estimated population totals can be computed. We develop formulas for the computer program PC CARP. PC CARP is a computer program created at Iowa State University designed to do analyses of survey samples from finite populations. PC CARP is set up to compute variances for stratified cluster samples. Among the estimates that can be calculated with PC CARP are totals and ratios. We look at the estimation of totals in this chapter.

5.1. The general model

5.1.1. The finite population and sample Let there be a finite population of N units, where N is known. Let the population be divided into H imputation classes (indexed by h) of size \( N_h \), \( h = 1, \ldots, H \). Let there be a characteristic associated with each unit, \( \{ Y_{hi} \}_{h=1, \ldots, H} \). Let

\[
Y_h = \sum_{i=1}^{N_h} Y_{hi}
\]  \hspace{1cm} (5.1)

be the finite population total of the characteristic for units in imputation class \( h \), and let
be the overall finite population total for the characteristic.

Let us take a random sample of size \( n \) from the population. Let \( n \) be fixed and let \( \pi_{hi}, \ h=1,...,H, \ i=1,...,N_h \) be the probability that unit \( hi \) is chosen in the random sample, where \( \pi_{hi} > 0 \) for all \( h \) and \( i \). Assume that we know the imputation class into which each unit in the sample falls. Let \( n_h \) denote the number of units in the sample that fall in imputation class \( h \). Suppose that, within imputation class \( h \), we are able to measure the characteristic for \( r_h \) of the units and unable to measure the characteristic for \( m_h \) of the units. Then \( r_h + m_h = n_h \), \( r_h \geq 0 \), and \( m_h \geq 0 \). The portion of the sample that is in imputation class \( h \) is written \( \{Y_{hi}\}_{i=1}^{n_h} \), where \( Y_{h1},...,Y_{hr_h} \) are associated with units that responded and \( Y_{hr_h+1},...,Y_{hn_h} \) are associated with units for which there was no response. The augmented sample in imputation class \( h \) is \( \{Y_{hi}^{*}\}_{i=1}^{n_h} \), where \( Y_{hi}^{*} = Y_{hi} \) for \( i=1,...,r_h \) and \( Y_{hi}^{*} \) is an imputed value for \( i=r_h+1,...,n_h \). It should be clear from the context when we are referring to the sample and when we are referring to the population with our indices.

Let

\[
R = \begin{bmatrix}
    r_1 & m_1 \\
    \vdots & \vdots \\
    r_H & m_H
\end{bmatrix}
\] (5.3)

be the matrix of the number of units that responded and the number of units that are missing within all of the \( H \) imputation classes. Let
be the average for the characteristic in imputation class \( h \) for the full sample. Let

\[
\bar{y}_h = \frac{n_h}{\Sigma} \frac{\sum_{i=1}^{n_h} y_{hi}}{n_h}
\]  

(5.4)

be the usual Horvitz—Thompson estimator of the population total for the full sample and let

\[
\hat{Y} = \frac{1}{\Sigma} \frac{\sum_{h=1}^{H} n_h}{\Sigma} \frac{\sum_{i=1}^{\pi_{hi}} y_{hi}}{\pi_{hi}}
\]  

(5.5)

be the Horvitz—Thompson estimator of the population total using the imputed values.

5.1.2. The missingness structure and imputation

Assume that within each imputation class the units that are nonrespondents are missing at random from the sample, where missing at random is defined in Section 2.3. The imputation is done within imputation classes, ignoring the probability of selection. Let

\[
k_h = \lfloor r_h^{-1} m_h \rfloor,
\]

(5.7)

where \([x]\) is the largest integer less than or equal to \( x \). Let
\[ t_h = m_h - k^*_h r_h . \] (5.8)

Then, under the random imputation scheme, each respondent in imputation class \( h \) is used at least \( k^*_h \) times for imputation, and \( t_h \) of the respondents in imputation class \( h \) are used \( k^*_h + 1 \) times. The \( t_h \) of the respondents that are used \( k^*_h + 1 \) times are chosen by simple random sampling without replacement from the respondents in imputation class \( h \). The donor respondents are assigned to the missing units randomly.

5.1.3. The expectation of the imputed total for a finite population In this section we consider the random imputation procedure using imputation cells. Let

\[ W_{hi} = \pi_{hi}^{-1} \] (5.9)

for all \( h \) and \( i \). Let an estimator of the finite population total be

\[ \hat{Y}^* = \sum_{h=1}^{H} \sum_{i=1}^{n_h} W_{hi} Y_{hi}^* . \] (5.10)

Note that \( \hat{Y}^* \) was also defined in expression (5.6). To find the expected value of expression (5.10) with respect the the finite population, we use four levels of conditioning. Let \( S \) stand for the original sample, \( \{Y_{hi}^*\}_{h=1, \ldots, H, i=1, \ldots, n_h} \).

When the sample is held constant, we hold the choice of which individuals from the population are in the sample constant. We also take the \( Y_{hi}^* \)'s in the sample as
given. Let rp stand for the response pattern. When the response pattern is held constant, the individuals in the sample that responded are given. Let ic stand for the choices for imputation. When the choices for imputation are held constant, we take as given which individuals within the part of the sample that responded were chosen to be used for imputation and how many times a given individual was used for imputation. Let dm stand for the choice of which donors are assigned to which missing values. When the choice of which donors are assigned to which missing values is held constant, the imputed value used to impute for each missing value is fixed. At the lowest level of conditioning (level 4), we hold the sample (S), the response pattern (rp), and the choices for imputation (ic) constant. At level 4, we average over which donor goes with which imputed value (dm). At the next level (level 3), we hold the sample and the response pattern constant. An expectation at level 3 is an average over the choices of elements to be used as donors. At the next level (level 2), we hold the sample constant and the number of observations and respondents in each imputation class constant. An expectation at level 2 is an average over the choices for respondents. At level 1 we average over the number of observations and respondents in each imputation class and over the possible samples from the finite population. We never actually evaluate the expectation at level 1.

Theorem 5.1: Assume that n observations have been sampled out of a population of size N, where every member of the population has a positive probability of selection and the sum of the probabilities of selection over the population equals n. Assume that the population is divided into H imputation classes, and that within the imputation classes the probability of a unit responding is constant. Let $n_h$ sample values fall into the $h$–th class and let $r_h$ of the $n_h$
individuals respond. Assume \( r_h > 0 \). Let the \( m_h \) missing values within the \( h \)-th imputation class be replaced in the following way. Each respondent in imputation class \( h \) is assigned randomly \( k_h \) times to \( k_h r_h \) of the missing units and \( t_h \) of the respondents in imputation class \( h \), where the \( t_h \) units are chosen by simple random sampling without replacement from the \( r_h \) respondents in imputation class \( h \), are assigned randomly to the remaining \( t_h \) missing units. Then

\[
E(\hat{Y}^* | FP, S, R ) = \bar{Y}
\]

\[
+ \sum_{h=1}^{H} \frac{m_h}{n_h} \sum_{i=1}^{n_h} W_{hi} \left( \bar{y}_h - Y_{hi} \right),
\]

where \( FP \) denotes the finite population, \( \bar{y}_h \) is defined in (5.4), \( R \) is defined in (5.3), \( \hat{Y}^* \) is defined in (5.10), and \( \bar{Y} \) is defined in (5.5).

**Proof:** First

\[
\hat{Y}^* = \sum_{h=1}^{H} \left[ \sum_{i=1}^{r_h} W_{hi} Y_{hi} + \sum_{i=r_h+1}^{n_h} W_{hi} Y_{hi}^* \right].
\]

At the lowest level of conditioning, averaging over the donor–missing matches, in imputation class \( h \) there are \( m_h! \) equally probable ways donors can be matched with missing observations and \( (m_h-1)! \) ways a specific donor can be matched with a specific observation, so
At the next level of conditioning, averaging over the choices of donors, in
imputation class \( h \) there are \( r_h! \left[ \frac{t_h! (r_h-t_h)!}{(r_h-1)!} \right]^{-1} \) choices for the donors to be used
more than \( k_h \) times, each equally likely, and each respondent appears \( (r_h-1)! \)
\( \left[ \frac{(t_h-1)! (r_h-t_h)!}{(r_h-1)!} \right]^{-1} \) times in the choices of the donors to be used \( k_h+1 \) times. All of
the respondents are used at least \( k_h \) times. It follows that

\[
E_2( \hat{Y}^* \mid FP, S, rp ) = \sum_{h=1}^{H} \left[ \sum_{i=1}^{r_h} W_{hi} Y_{hi} + \sum_{i=r_h+1}^{n_h} W_{hi} \left( \frac{k_h}{m_h} \sum_{j=1}^{r_h} Y_{hj} + \frac{t_h}{r_h m_h} \sum_{j=1}^{r_h} Y_{hj} \right) \right].
\]  

At the next level of conditioning, averaging over the response patterns, given \( R \),
there are \( n_h! \left[ \frac{r_h! (n_h-r_h)!}{(n_h-1)!} \right]^{-1} \) possible response patterns in imputation class \( h \), each
equally likely. Each sample member in imputation class \( h \) appears in \( (n_h-1)! \)
\( \left[ \frac{(r_h-1)! (n_h-r_h)!}{(n_h-1)!} \right]^{-1} \) of the patterns. Also, each \( W_{hi} Y_{hj} \) appears \( (n_n-2)! \)
\[(r_h - 1)! (n_h - r_h - 1)! \] times in the \( \Sigma_{i=1}^{n_h} \sum_{j=1}^{r_h} W_{hi} Y_{hj} \)'s if \( i \neq j \) and each \( W_{hi} Y_{hi} \) appears zero times in the \( \Sigma_{i=r_h+1}^{n_h} \sum_{j=1}^{r_h} W_{hi} Y_{hj} \)'s, over the set of all possible \( \Sigma_{i=r_h+1}^{n_h} \sum_{j=1}^{r_h} W_{hi} Y_{hj} \)'s. It follows that

\[
E_2(\hat{Y}^* | FP, S, R) = \frac{H}{\Sigma_{h=1}^{H}} \left\{ \frac{r_h}{n_h} \sum_{i=1}^{n_h} W_{hi} Y_{hi} \right\} \\
+ \frac{m_h}{n_h (n_h - 1)} \left\{ \sum_{i=1}^{n_h} W_{hi} \left( \sum_{j=1}^{n_h} Y_{hj} - \sum_{i=1}^{n_h} W_{hi} Y_{hi} \right) \right\} \\
= \frac{H}{\Sigma_{h=1}^{H}} \left\{ \sum_{i=1}^{n_h} W_{hi} Y_{hi} \right\} \\
+ \frac{m_h}{n_h - 1} \left\{ \sum_{i=1}^{n_h} W_{hi} \left( \bar{y}_h - Y_{hi} \right) \right\} .
\]

Alternate Proof: If the sample is held constant, the expected value of the estimator over all possible imputation patterns, given the respondents is

\[
\sum_{i=1}^{r_h} W_{hi} Y_{hi} + \sum_{i=r_h+1}^{n_h} W_{hi} \bar{y}_{rh} ,
\]

where \( \bar{y}_{rh} \) is the mean of the respondents in imputation class \( h \). Because the
probability of a response is equal for all elements within a cell

\[ E \left\{ \sum_{i=1}^{r_h} W_{hi} Y_{hi} \mid FP, S, R \right\} \]

\[ = n_h^{-1} r_h \sum_{i=1}^{n_h} W_{hi} Y_{hi} \]  \hspace{1cm} (5.17)

and

\[ E \left\{ W_{hi} \tilde{y}_{rh} \mid FP, S, R \right\} \]

\[ = W_{hi} (n_h - 1)^{-1} \sum_{j=1}^{n_h} Y_{hj} \]

\[ = W_{hi} (n_h - 1)^{-1} (n_h \tilde{y}_{h} - Y_{hi}) \]  \hspace{1cm} (5.18)

Therefore

\[ E \left\{ \sum_{i=r_h+1}^{n_h} W_{hi} \tilde{y}_{rh} \mid FP, S, R \right\} \]

\[ = n_h^{-1} m_h \sum_{i=1}^{n_h} W_{hi} (n_h - 1)^{-1} (n_h \tilde{y}_{h} - Y_{hi}) \]  \hspace{1cm} (5.19)

and
We present a corollary for less general conditions.

**Corollary:** Let the assumptions of Theorem 5.1 hold. Let the sampling be stratified random sampling as described in Section 4.1. Let

\[
W_m = \frac{N_m}{n_m} \frac{1}{N} .
\]

Then

\[
N^{-1} \hat{Y}^* = \hat{y}_{st}^* = \sum_{m=h}^{M} \sum_{h=1}^{H} \sum_{i=1}^{n_{mi}} W_m Y_{mhi}^* ,
\]

and

\[
E(\hat{y}_{st}^* | FP, S, R) - \bar{y}_{st}
\]
where the notation is that of Section 4.3. The last expression in expression (5.23) agrees with expression (4.24) in Section 4.3.

5.1.4. The superpopulation structure  From Theorem 5.1, we see that the missing at random within imputation classes hypothesis is not strong enough to give us unbiasedness for estimator (5.10) for the finite population total conditioning on the given finite population. This is because 

\[ \mathbb{E}_i \left[ m_h (n_i - 1) \sum_{i=1}^{n_i} W_{hi} (\bar{y}_h - Y_{hi}) \right] \]

is not generally equal to zero. We now add the hypothesis of an underlying superpopulation to our set of assumptions.

Assume that there is a superpopulation made up of \( H \) sub-populations. Assume that equal probability samples of sizes \( N_h, h=1,\ldots,H \) are taken from the sub-populations. These \( H \) samples make up the \( H \) imputation classes within the finite population. Assume that within sub-population \( h, h=1,\ldots,H \), the \( Y \)'s are distributed identically and independently with mean \( \mu_h \) and variance \( \sigma^2_h \). Assume that the \( Y \)'s are independent between the sub-populations.
5.1.5. The expectation of the imputed total under the superpopulation model

We find the expectation of estimator (5.10) under the superpopulation model.

Theorem 5.2: Let the assumptions of Theorem 5.1 hold. Assume that the superpopulation structure described in Section 5.4 holds. Let

$$\bar{\mu} = \sum_{h=1}^{H} N_{h}^{-1} N_{h} \mu_{h}$$

be the superpopulation mean. Then

$$E(\hat{Y}^*) = N \bar{\mu}$$

and

$$E(\hat{Y}^* - Y) = 0,$$

where it is understood that the expectation is over all drawings from the superpopulation and where $Y$ is the finite population total.

Proof: To find the expectation of the estimator (5.10), we use two levels of conditioning. At the lower level (level 2) we hold the choice of which units from the finite population are in the sample (cs), the response pattern (rp), the choices for imputation (ic), and the choices of which donor goes with which missing value (dm) constant. At level 2, we average over draws from the superpopulation.
At the upper level (level 1), we take expectations over everything that is left random after the lower expectation is taken. Then

\[ E_2(\sum_{h=1}^{H} \sum_{i=1}^{n_h} W_{hi} Y_{hi}^* | cs, rp, ic, dm) \]

\[ = \sum_{h=1}^{H} \sum_{i=1}^{n_h} W_{hi} \mu_h, \quad (5.27) \]

since the \( Y_{hi}^* \)'s are identically distributed in the imputation classes. The \( Y_{hi} \)'s are identically distributed within the imputation classes because the \( Y_{hi} \)'s are identically distributed within the imputation classes, the missing values are missing at random within the imputation classes, and the donors are chosen and assigned to the missing values at random within the imputation classes.

Let \( a_{hi} \) be a 0–1 indicator variable that is one if unit hi is in the sample from the finite population and zero otherwise. Then

\[ E_1( a_{hi} ) = W_{hi}^{-1}, \quad (5.28) \]

The

\[ \sum_{h=1}^{H} \sum_{i=1}^{n_h} W_{hi} \mu_h = \sum_{h=1}^{H} \sum_{i=1}^{N_h} a_{hi} W_{hi} \mu_h, \quad (5.29) \]
\[ E_1(\sum_{h=1}^{H} \sum_{i=1}^{n_h} W_{hi} \mu_i) = \sum_{h=1}^{H} \sum_{i=1}^{n_h} E_1(a_{hi}) W_{hi} \mu_i \]
\[ = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \mu_i = \sum_{h=1}^{H} n_h \mu_h = N \bar{\mu}. \quad (5.30) \]

This gives us expression (5.25).

For expression (5.26),

\[ E(\hat{Y}) = E_1(\sum_{h=1}^{H} \sum_{i=1}^{n_h} Y_{hi}) = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \mu_i = N \bar{\mu}, \quad (5.31) \]

since the \( Y_{hi} \)'s in the finite population are an equal probability sample from the portion of the superpopulation in sub-population \( h \). Since \( \hat{Y}^* \) and \( Y \) both have the same expected value,

\[ E(\hat{Y}^* - Y) = 0. \quad (5.32) \]

This completes the proof of Theorem 5.2.

\#

**Alternative Proof:** Under the model, every element in the \( h \)-th cell of the finite population has the same expected value when the expectation is over draws from the superpopulation. Thus, in expression (5.11) we have

\[ E\{ \bar{Y}_h - Y_{hi} \mid cs, R \} = 0 \quad (5.33) \]
5.1.6. The variance of the imputed total under the superpopulation model

We now give the variance for estimator (5.10) under the superpopulation model. The variance of \( \hat{Y}^* - Y \), where \( Y \) is the finite population total, is given in Section 5.1.7.

Theorem 5.3: Let the assumptions of Theorem 5.2 hold. Then

\[
V(\hat{Y}^* - N \bar{\mu}) - V(\hat{Y} - N \bar{\mu}) = \sum_{h=1}^{H} \sigma_h^2 \mathbb{E}\left( \frac{k_h n_h + (k_h + 2) t_h}{n_h (n_h - 1)} \right) \\
\times \left\{ \left( \sum_{i=1}^{n_h} W_{hi} \right)^2 - \sum_{i=1}^{n_h} W_{hi}^2 \right\} \\
= \mathbb{E}\left\{ \sum_{h=1}^{H} \sigma_h^2 (k_h + 1) (m_h + t_h) (W_h^2 - n_h^{-1} S_{Wh}^2) \right\},
\]

(5.34)

where \( \hat{Y} \) is defined in expression (5.5), \( \sigma_h^2 \) is the superpopulation variance of the \( h \)-th imputation cell, \( W_h \) is the mean of the \( n_h \) weights in cell \( h \), and \( S_{Wh}^2 \) is the variance of the \( n_h \) weights in cell \( h \) and where we assume \( \mathbb{E}\{n_h^{-1}(n_h-1)^{-1}\}
(k_h n_h + (k_h + 2) t_h) > 0 \) for all \( h \).
Proof: To find the variance of $\hat{Y}^*$, we use three levels of conditioning. At the lowest level (level 3), we hold the choices for which units are in the sample (cs), the response pattern (rp), the choices for imputation (ic), and the choices of which donor goes with which missing value (dm) constant, and take the expectations with respect to draws from the superpopulation. At level 2 we hold the choice for which units are in the sample (cs) and the $n_h$'s and the $r_h$'s (R) constant and average over the possible response patterns (given R) and over the possible imputation choices and over the possible choices of which donors go with which missing values. At level 1 we average over everything left random in the expression. We never actually evaluate the variance at level 1.

Before we find the variance for $\hat{Y}^*$, we find the variance for $\hat{Y}$ in terms of our three levels of conditioning. The

$$V(\hat{Y} - \mu) = V_1\{E_2[E_3(\hat{Y})]\} + E_1\{V_2[E_3(\hat{Y})]\}$$

$$+ E_1\{E_2[V_3(\hat{Y})]\}. \tag{5.35}$$

Now,

$$E_3(\hat{Y}) = E_3(\sum_{h=1}^{H} \sum_{i=1}^{n_h} W_{hi} Y_{hi} \mid cs, rp, ic, dm)$$

$$= \sum_{h=1}^{H} \sum_{i=1}^{n_h} W_{hi} \mu_h, \tag{5.36}$$

since the $Y_{hi}$'s are identically distributed within the imputation classes, so
Also,

\[ V \{ \hat{y} \} = V \{ \sum_{i=1}^{N} W_i y_i \} = 0. \]  

and

\[ E \{ V \{ E_i Y(h) \} \} = V \{ \sum_{i=1}^{N} W_i Y_i \} = 0. \]  

It follows that

\[ V \{ E_i \{ E_i Y(h) \} \} = V_i \{ \sum_{h=1}^{N} W_i Y_i(h) \} = 0. \]  

and

\[ E \{ E_i \{ E_i Y(h) \} \} = E \{ \sum_{h=1}^{N} W_i Y_i(h) \} = 0. \]  

(5.40)  

(5.39)  

(5.38)  

(5.37)
since the \( Y_{hi} \)'s are independent between imputation classes and are independently and identically distributed within imputation classes. From (5.35), (5.39), (5.40) and (5.41)

\[
V(\hat{Y} - N\mu) = V_1\{ \sum_{h=1}^{H} \sum_{i=1}^{n_h} W_{hi} \mu_h \} + E_1\{ \sum_{h=1}^{H} \sum_{i=1}^{n_h} W_{hi}^2 \sigma_h^2 \} .
\]  

(5.42)

We now find the variance of \( \hat{Y}^* \). The

\[
V(\hat{Y}^* - N\mu) = V_1\{ E_2[ E_3(\hat{Y}^*)] \} + E_1\{ V_2[ E_3(\hat{Y}^*)] \} + E_1\{ E_2[ V_3(\hat{Y}^*)] \} .
\]  

(5.43)

We know, from Theorem 5.2 expression (5.27), that

\[
E_3\{ \hat{Y}^* | cs, rp, ic, dm \} = E_3\{ \hat{Y} | cs, rp, ic, dm \} ,
\]  

(5.44)

so,

\[
V(\hat{Y}^* - N\mu) = V_1\{ \sum_{h=1}^{H} \sum_{i=1}^{n_h} W_{hi} \mu_h \} .
\]
+ E_1 \{ E_2 [ V_3 ( \hat{Y}^* | \text{cs, rp, ic, dm} ) | \text{cs, R} ] \} . \tag{5.45}

The

E_2 [ V_3 ( \sum_{h=1}^{H} \sum_{i=1}^{n_h} W_{hi} Y_{hi}^* | \text{cs, rp, ic, dm} ) | \text{cs, R} ]

= \sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{n_h} W_{hi} W_{hj} E_2 [ \text{Cov}_3 ( Y_{hi}^*, Y_{hj}^* | \text{cs, rp, ic, dm} ) | \text{cs, R} ] , \tag{5.46}

since the $Y_{hi}^*$'s are independent between the imputation classes. Now the

$\text{Cov}_3 ( Y_{hi}^*, Y_{hj}^* | \text{cs, rp, ic, dm} ) = \sigma_h^2$ \tag{5.47}

for all $h$, since the $Y_{hi}^*$'s are identically distributed within the imputation classes. Also,

$E_2 [ \text{Cov}_3 ( Y_{hi}^*, Y_{hj}^* | \text{cs, rp, ic, dm} ) | \text{cs, R} ]$

= \sum_{i=1}^{n_h} \sum_{j=1}^{n_h} \text{Cov}_3 ( Y_{hi}^*, Y_{hj}^* | \text{cs, rp, ic, dm} ) / \sum_{i \neq j} n_h ( n_h - 1 )$

= \frac{k_h n_h + ( k_h + 2 ) t_h}{n_h ( n_h - 1 )} \sigma_h^2 , \tag{5.48}
for \( i \neq j \), since each of the \( n_h \) elements in imputation class \( h \) in the augmented sample appears at least \((k_h+1)\) times in the augmented sample and \((k_h+2)t_h\) of the elements appear one more time and since we need to subtract out the \( n_h \) of the elements that appear on the diagonal of the cross product. The sum in expression (5.48) is independent of the sample \((\text{given } R)\), the response pattern \((\text{given } R)\), the imputation choices, and the choices of which donors go with which missing values. From expressions (5.46), (5.47), and (5.48), the

\[
E_2[ \text{V}_3( \hat{Y}^* | \text{cs, rp, ic, dm} ) | \text{cs, R} ]
\]

\[
= \sum_{h=1}^{H} \sum_{i=1}^{n_h} W_{hi}^2 \sigma_h^2
\]

\[
+ \sum_{h=1}^{H} \sigma_h^2 \frac{k_h n_h + (k_h + 2)t_h}{n_h (n_h - 1)}
\]

\[
\times \left[ (\sum_{i=1}^{n_h} W_{hi})^2 - \sum_{i=1}^{n_h} W_{hi}^2 \right].
\]  (5.49)

Expression (5.34) follows from expressions (5.42), (5.45), and (5.49).

**Alternative Proof 1:** We have

\[
\text{V}( \hat{Y}^* - N\hat{\mu} ) = \text{V}( \hat{Y} - N\hat{\mu} ) + \text{V}( \hat{Y}^* - \hat{Y} ) + 2 \text{Cov}( \hat{Y}^* - \hat{Y} , \hat{Y} - N\hat{\mu} )
\]
\[ = V(\hat{Y} - N\bar{\mu}) + E\{ V(\hat{Y}^* - \hat{Y} \mid cs, R) \} \]

\[ + V\{ E(\hat{Y}^* - \hat{Y} \mid cs, R) \} + 2E\{ Cov(\hat{Y}^* - \hat{Y}, \hat{Y} - N\bar{\mu} \mid cs, R) \} \]

\[ + 2 Cov\{ E(\hat{Y}^* - \hat{Y} \mid cs, R), E(\hat{Y} - N\bar{\mu} \mid cs, R) \} . \]

(5.50)

Fixing the choice of the original sample corresponds to fixing the weights \( W_{hi} \), \( h=1,\ldots,H, i=1,\ldots,n_h \). Now, because the elements within a cell are assumed to be identically distributed, the imputed values are chosen and donated randomly within a cell, and the missing values are missing at random within a cell,

\[ E(\hat{Y}^* - \hat{Y} \mid cs, R) = \sum_{h=1}^{H} \sum_{i=1}^{n_h} W_{hi} (\mu_h - \mu_h) \]

\[ = 0 . \]

(5.51)

Therefore

\[ V(\hat{Y}^* - N\bar{\mu}) = V(\hat{Y} - N\bar{\mu}) + E\{ V(\hat{Y}^* - \hat{Y} \mid cs, R) \} \]

\[ + 2E\{ Cov(\hat{Y}^* - \hat{Y}, \hat{Y} - N\bar{\mu} \mid cs, R) \} . \]

(5.52)

Now, letting \( i=1,\ldots,r_h \) denote the respondents,
\[ V(\hat{Y} - \hat{Y}^* \mid cs, R) = V\left( \sum_{h=1}^{H} \sum_{i=1}^{n_h} W_{hi} Y_{hi} - \sum_{h=1}^{H} \sum_{i=1}^{n_h} W_{hi} Y_{hi}^* \mid cs, R \right) \]

\[ = V\left\{ \sum_{h=1}^{H} \sum_{i=r_h+1}^{n_h} W_{hi} \left( Y_{hi} - Y_{hi}^* \right) \mid cs, R \right\} , \quad (5.53) \]

where the \( Y_{hi}^* \)'s in the second expression are imputed values and the \( Y_{hi} \)'s in the
second expression are the missing values.

We know that \( r_h - t_h \) of the respondents are used \( k_h \) times as donors and that
\( t_h \) of the respondents are used \( k_h + 1 \) times. A donor element appearing \( k_h + 1 \) times
will create \( (k_h + 1)k_h \) covariance terms. Under the model, every individual in an
imputation cell is equally likely to respond, the individuals in an imputation cell are
identically and independently distributed, and the imputed values are chosen and
assigned randomly. It follows that

\[ V(\hat{Y} - \hat{Y}^* \mid cs, R) = \left\{ 2 \sum_{h=1}^{H} m_h \sum_{i=1}^{n_h-1} W_{hi}^2 \right\} \sigma_h^2 \]

\[ + \sum_{h=1}^{H} \left[ (r_h - t_h) k_h (k_h - 1) + t_h (k_h + 1) k_h \right] \]

\[ \times \left[ W_h^2 - n_h^{-1} S_h^2 \right] \sigma_h^2 , \quad (5.54) \]

where we have used
\[ E\{W_{hi} \cdot W_{hj} | cs, R\} = (n_{h}^{-1} \sum_{i=1}^{n_{h}} W_{hi})^2 - n_{h}^{-1} S_{Wh}^2 \] (5.55)

for \(i \neq j\), where it is understood that we are taking the expectation of the product of the randomly chosen \(W\)'s associated with the respondents, and where

\[ S_{Wh}^2 = (n_{h} - 1)^{-1} \sum_{i=1}^{n_{h}} (W_{hi} - W_{h})^2 \] (5.56)

and

\[ W_{h} = n_{h}^{-1} \sum_{i=1}^{n_{h}} W_{hi} \] (5.57)

We note that, given a random sample of \(m_{h}\) from \(n_{h}\),

\[ E\{(m_{h}^{-1} \sum_{i=1}^{m_{h}} W_{hi})^2 - m_{h}^{-1} S_{Wh}^2 | cs, R\} \]

\[ = W_{h}^2 + (n_{h} - m_{h}) m_{h}^{-1} S_{Wh}^2 - m_{h}^{-1} S_{Wh}^2 \]

\[ = W_{h}^2 - n_{h}^{-1} S_{Wh}^2 \] (5.58)

Also,
\[
\text{Cov}(\hat{Y}^*-\hat{Y}, \hat{Y}-N\mu | \text{cs}, R)
\]

\[
= \text{Cov}\{ \sum_{h=1}^{n_h} \sum_{i=r_h+1}^{r_h} W_{hi} (Y_{hi}^*-Y_{hi}), \sum_{h=1}^{n_h} \sum_{i=1}^{n_h} W_{hi} Y_{hi} | \text{cs}, R \}
\]

\[
= \sum_{h=1}^{n_h} E\{ \sum_{i=r_h+1}^{r_h} W_{hi} W_{hj} Y_{hi}^* Y_{hj} | \text{cs}, R \} - \sum_{h=1}^{n_h} m_h n_h^{-1} \sum_{i=1}^{n_h} W_{hi}^2 \sigma_h^2
\]

\[
= \sum_{h=1}^{n_h} \left[ (r_h-t_h) k_h + t_h (k_h+1) \right] (W_h^2 - n_h^{-1} S_{Wh}^2) \sigma_h^2
\]

\[
- \sum_{h=1}^{n_h} m_h n_h^{-1} \sum_{i=1}^{n_h} W_{hi}^2 \sigma_h^2,
\]

where we have assumed without loss in generality that \(\mu_h=0\). It follows that

\[
\text{V}(\hat{Y}^* - \hat{Y} | \text{cs}, R) + 2 \text{Cov}\{ (\hat{Y}^* - \hat{Y}, \hat{Y}) | \text{cs}, R \}
\]

\[
= \sum_{h=1}^{n_h} \left[ (r_h-t_h) k_h (k_h-1) + t_h (k_h+1) k_h + 2 (r_h-t_h) k_h \right]
\]

\[
+ 2 t_h (k_h+1) \right] (W_h^2 - n_h^{-1} S_{Wh}^2) \sigma_h^2
\]

\[
= \sum_{h=1}^{n_h} (k_h+1) (m_h + t_h) (W_h^2 - n_h^{-1} S_{Wh}^2) \sigma_h^2.
\]

(5.60)
Expression (5.34) follows from substituting (5.60) into (5.52).

Alternative Proof 2: We use a new set of indices on the $Y_{hi}^*$'s. Let the $Y_{hi}^*$'s in imputation class $h$, $h=1,...,H$, be put in order such that the first $(k_h+2)t_h$ elements in class $h$ are the $t_h$ respondents used $k_h+1$ times in the imputation, where each respondent is repeated $k_h+2$ times. The elements that follow the first $(k_h+2)t_h$ elements are the $(r_h-t_h)$ respondents that are used for imputation $k_h$ times, where each respondent is repeated $(k_h+1)$ times. The weights on the $Y_{hi}^*$'s are the weights assigned during the imputation process. Then

$$
\hat{Y}^* = \sum_{h=1}^{H} \left[ \sum_{i=1}^{t_h} \left( \sum_{j=1}^{k_h+2} Y_{hi2i} W_{hi2ij} \right) Y_{hi2i} + \sum_{i=t_h+1}^{r_h} \left( \sum_{j=1}^{k_h+1} Y_{hi1ij} W_{hi1ij} \right) Y_{hi1i} \right],
$$

(5.61)

where

$$
l_{ii} = (k_i + 1)(i - 1) + 1
$$

(5.62)

$$
l_{iij} = (k_i + 1)(i - 1) + j
$$

(5.63)

$$
L_{2i} = (k_i + 2)(i - 1) + 1
$$

(5.64)

$$
L_{2ij} = (k_i + 2)(i - 1) + j.
$$

(5.65)
Let \( \hat{Z} \) be an estimator of the population total. Let \( rp \) be as defined in Section 5.4, \( R \) be as defined in expression (5.3), and \( cs \) be as defined in Theorem 5.2. Then

\[
V(\hat{Z} - N\mu) = V\{ E[ E(\hat{Z} | cs, rp) | cs, R] \}
+ E\{ V[ E(\hat{Z} | cs, rp) | cs, R] \} + E\{ E[ V(\hat{Z} | cs, rp) | cs, R] \}.
\]

(5.66)

Since the \( Y_{h}^{*} \)'s in the sample are identically and independently distributed within the imputation classes and independent between the imputation classes, for \( \hat{Z} = \hat{Y} \),

\[
E(\hat{Z} | cs, rp) = \sum_{h=1}^{H} \sum_{i=1}^{n_h} W_{hi} \mu_h
\]

(5.67)

and

\[
V(\hat{Z} | cs, rp) = \sum_{h=1}^{H} \sum_{i=1}^{n_h} W_{hi}^2 \sigma_h^2 = \sum_{h=1}^{H} \sigma_h^2 \left[ n_h W_{hi}^2 + (n_h - 1) S_{Wh}^2 \right].
\]

(5.68)

Since the \( Y_{hi}^{*} \)'s in the augmented sample are identically distributed within the imputation classes, for \( \hat{Z} = \hat{Y}^* \),

\[
E(\hat{Z} | cs, rp) = \sum_{h=1}^{H} \sum_{i=1}^{n_h} W_{hi} \mu_h
\]

(5.69)
It follows that

\[ V(\hat{Y}^*-N\mu) = V(\hat{Y}-N\mu) + \mathbb{E}\{ \mathbb{E}[V(\hat{Y}^* | cs, rp) | cs, R] \} \]

\[ - \mathbb{E}\{ \sum_{h=1}^{H} \sigma_h^2 [n_h W_h^2 + (n_h - 1) S_{Wh}^2] \}. \quad (5.70) \]

Since the \( Y_{hi} \)'s are identically and independently distributed within the imputation classes and independent between imputation classes, the

\[ \mathbb{E}[V(\hat{Y}^* | cs, rp) | cs, R] \]

\[ = \mathbb{E}\{ \sum_{h=1}^{H} \sum_{i=1}^{t_h} (\sum_{j=1}^{k_h+2} W_{hij})^2 \sigma_h^2 + \sum_{i>t_h+1}^{k_h+1} (\sum_{j=1}^{r_h} W_{h1ij})^2 \sigma_h^2 \} | cs, R \}. \quad (5.71) \]

If we assume that the weights associated with each respondent group in imputation class \( h \) are a simple random sample without replacement from the set of all of the weights in the sample in imputation class \( h \), then it is straightforward to show that

\[ \mathbb{E}[ (\sum_{j=1}^{k_h+x} W_{hlxij})^2 | cs, R ] \]

\[ = (k_h+x)^2 W_h^2 + \frac{n_h - (k_h+x)}{n_h} (k_h+x) S_{Wh}^2, \quad (5.72) \]
where $x$ takes on the value 1 or 2. The assumption that the weights in the respondent groups in imputation class $h$ are a simple random sample without replacement from the set of all of the weights in the sample in imputation class $h$ follows from the assumption that the missing values are missing at random within the imputation classes and that the donors are chosen and assigned randomly to the missing values within the imputation classes.

Using expressions (5.70), (5.71), and (5.72), it is straightforward, though tedious, to show that

$$V(\hat{Y}^* - N\bar{\mu}) = V(\hat{Y} - N\bar{\mu})$$

$$+ \sum_{h=1}^{H} \sigma_h^2 \{ (k_h + 1)(m_h + t_h) [W_h^2 - n_h^{-1} S_{Wh}^2] \}. \quad (5.73)$$

We give a corollary for equal probability random imputation schemes. First we define equal probability random imputation.

**Definition 5.1:** Equal probability random imputation is any form of random imputation for which, within imputation classes, the probability that a respondent donates to a missing value is constant for all respondents and all missing values. Under random imputation, all of the donors within an imputation class come from the set of the respondents within that imputation class.
Corollary 1: Let the assumptions of Theorem 5.2 hold, except that the imputation is any form of equal probability random imputation rather than the form given in Section 5.2. Let \( g_{hi} \) be the number of times that respondent \( hi \) is used for imputation, \( h=1,\ldots,H, i=1,\ldots,r_h \). Then

\[
V(\hat{Y}^* - N\mu) = V(\hat{Y} - N\mu) + \sum_{h=1}^{H} \sigma_{h}^2 E\left\{ \left[ \sum_{i=1}^{r_h} \frac{g_{hi}(g_{hi}+1)}{n_h(n_h-1)} \right] \right\}
\]

\[
\times \left[ \left( \sum_{i=1}^{n_h} W_{hi} \right)^2 - \sum_{i=1}^{n_h} W_{hi}^2 \right].
\]  \hspace{1cm} (5.74)

Proof: From Theorem 5.3, the term in expression (5.48) depends on the number of nonzero covariance terms. The result follows since, for \( h=1,\ldots,H, i=1,\ldots,r_h \), respondent \( hi \) contributes \( g_{hi}(g_{hi}+1) \) covariance terms.

We present a second corollary for less general conditions.

Corollary 2: Let the assumptions of Theorem 5.2 hold. Let the sampling be stratified random sampling as described in Section 4.1. Let \( W_m \) and \( \bar{y}_{st}^* \) be as defined in (5.21) and (5.22). Then

\[
V(\bar{y}_{st}^* - \mu) = V(\bar{y}_{st} - \mu) + \sum_{h=1}^{H} \sigma_{h}^2 E\left\{ \left( k_h n_h + (k_h + 2) t_h \right) \right\}
\]

\[
\times \left( \frac{k_h n_h}{n_h(n_h-1)} \right). \hspace{1cm} (5.75)
\]
\begin{equation}
\times \left[ \left( \sum_{m=1}^{M} n_{mh} W_m \right)^2 - \sum_{m=1}^{M} n_{mh} W_m^2 \right], \tag{5.75}
\end{equation}

where

\begin{equation}
\bar{y}_{st} = N^{-1} \hat{Y}, \tag{5.76}
\end{equation}

and the notation is that of Section 4.6. Expression (5.75) agrees with expression (4.50) in Section 4.6.

Proof: Since \( W_{hi} = NW_m \),

\begin{align*}
\frac{1}{N^2} \left[ \left( \sum_{i=1}^{n_h} W_{hi} \right)^2 - \sum_{i=1}^{n_h} W_{hi}^2 \right] \\
= \frac{1}{N^2} \left[ \left( \sum_{m=1}^{M} \sum_{i=1}^{n_{mh}} N W_m \right)^2 - \sum_{m=1}^{M} \sum_{i=1}^{n_{mh}} N^2 W_m^2 \right] \\
= \left[ \left( \sum_{m=1}^{M} n_{mh} W_m \right)^2 - \sum_{m=1}^{M} n_{mh} W_m^2 \right]. \tag{5.77}
\end{align*}

Expression (5.34) of Theorem 5.3 is an important result because, under fairly general assumptions and a general sampling scheme, expression (5.34) gives an expression for the variance of the population total that can be estimated using available computer programs. If the variance of the full sample estimator is
estimated by the usual method using a sample with imputed values, then the estimator will usually have small bias if the number of respondents is reasonably large. Also, the term in expression (5.34) that corrects for the imputation is easily calculated using available computer routines.

5.1.7. Estimation of the variance of the imputed total minus the finite population total We now find an unbiased estimator of the $V(\hat{Y}^* - Y)$. We start by finding the variance of $\hat{Y}^* - Y$.

Theorem 5.4: Let the assumptions of Theorem 5.2 hold. Then

$$V(\hat{Y}^* - Y) = V(\hat{Y} - Y)$$

$$= \sum_{h=1}^{H} \sigma_h^2 \mathbb{E}\left\{ n_h^{-1} \left( n_h - 1 \right)^{-1} \left[ k_h n_h + (k_h + 2) t_h \right] \right\}$$

$$\times \left\{ \left( \sum_{i=1}^{n_h} W_{hi} \right)^2 - \sum_{i=1}^{n_h} W_{hi}^2 \right\}$$

$$= \sum_{h=1}^{H} \sigma_h^2 \mathbb{E}\left\{ (k_h + 1)(m_h + t_h)(W_{h}^2 - n_h^{-1} s_{Wh}^2) \right\}.$$

(5.78)

Proof: The

$$V(\hat{Y}^* - N\bar{\mu}) = V(\hat{Y}^* - Y + Y - \hat{Y} + \hat{Y} - Y - N\bar{\mu})$$
\[ \begin{align*}
= & \text{V}(\hat{Y}^* - Y) + \text{V}(\hat{Y} - Y) + \text{V}(\hat{Y} - N\bar{\mu}) - 2 \text{Cov}(\hat{Y}^* - Y, \hat{Y} - Y) \\
+ & 2 \text{Cov}(\hat{Y}^* - Y, \hat{Y} - N\bar{\mu}) - 2 \text{Cov}(\hat{Y} - Y, \hat{Y} - N\bar{\mu}) \\
= & \text{V}(\hat{Y}^* - Y) + \text{V}(\hat{Y} - N\bar{\mu}) \\
- & [ \text{V}(\hat{Y}) - 2 \text{Cov}(\hat{Y}^*, Y) + \text{V}(Y) ] . \quad (5.79)
\end{align*} \]

The

\[ \text{Cov}(\hat{Y}^*, Y) = \text{E}(\hat{Y}^* Y) - N^2 \bar{\mu}^2 \]

\[ = \text{E} [ \text{E}(\hat{Y}^* | \text{FP}, S, R) Y ] - N^2 \bar{\mu}^2 \]

\[ = \text{E} \left[ \hat{Y} Y - \sum_{h=1}^{H} \frac{m_h}{n_h - 1} \sum_{i=1}^{n_h} W_{hi} (\bar{y}_h Y - Y_{hi} Y) \right] - N^2 \bar{\mu}^2 \]

\[ = \text{Cov}(\hat{Y}, Y) - \text{E} \left[ \sum_{h=1}^{H} \frac{m_h}{n_h - 1} \sum_{i=1}^{n_h} W_{hi} \text{E}(\bar{y}_h Y - Y_{hi} Y | \text{cs}, R) \right] \]

\[ = \text{Cov}(\hat{Y}, Y) - \text{E} \left[ \sum_{h=1}^{H} \frac{m_h}{n_h - 1} \sum_{i=1}^{n_h} W_{hi} \left( N \bar{\mu}^2 + \sigma_h^2 - N \bar{\mu}^2 - \sigma_h^2 \right) \right] \]

\[ = \text{Cov}(\hat{Y}, Y) . \quad (5.80) \]
So, by (5.79) and (5.80),

\[
V(\hat{Y} - Y)
\]

\[
= V(\hat{Y}^* - N\overline{\mu}) - V(\hat{Y} - N\overline{\mu}) + V(\hat{Y} - Y). \quad (5.81)
\]

Expression (5.78) follows from the result in Theorem 5.3.

Alternative Proof: We have

\[
V(\hat{Y}^* - Y) = V(\hat{Y}^* - \hat{Y} + \hat{Y} - Y)
\]

\[
= V(\hat{Y}^* - \hat{Y}) + 2Cov(\hat{Y}^* - \hat{Y}, \hat{Y} - Y) + V(\hat{Y} - Y). \quad (5.82)
\]

The

\[
\text{Cov}(\hat{Y}^* - \hat{Y}, \hat{Y} - Y) = E[\text{Cov}(\hat{Y}^* - \hat{Y}, \hat{Y} - Y \mid cs, R)]
\]

\[
+ \text{Cov}[E(\hat{Y}^* - \hat{Y} \mid cs, R), E(\hat{Y} - Y \mid cs, R)]
\]

\[
= E[\text{Cov}(\hat{Y}^* - \hat{Y}, \hat{Y} - Y \mid cs, R)]
\]

\[
= E[\text{Cov}(\hat{Y}^* - \hat{Y}, \hat{Y} - N\overline{\mu} \mid cs, R)].
\]
\[-E[\text{Cov}(\hat{Y}^* - \hat{Y}, Y - N \bar{\mu} | cs, R)].\]  \hspace{1cm} (5.83)

Also

\[\nabla(\hat{Y}^* - \hat{Y}) = E\{2 \sum_{h=1}^{H} n_h^{-1} m_h \sum_{i=1}^{n_h} W_{hi}^2\} \sigma_h^2\]

\[+ \sum_{h=1}^{H} \sigma_h^2 E\{[(r_h - t_h) k_h (k_h - 1) + t_h (k_h + 1) k_h]\]

\[\times [W_h^2 - n_h^{-1} S_{Wh}^2]\} \hspace{1cm} (5.84)\]

by expression (5.51) and (5.54) in Theorem 5.3. Now

\[E[\text{Cov}(\hat{Y}^* - \hat{Y}, \hat{Y} - N \bar{\mu} | cs, R)]\]

\[= \sum_{h=1}^{H} \sigma_h^2 E\{[(r_h - t_h) k_h + t_h (k_h + 1)] [W_h^2 - n_h^{-1} S_{Wh}^2]\}\]

\[E\{\sum_{h=1}^{H} n_h^{-1} m_h \sum_{i=1}^{n_h} W_{hi}^2\} \sigma_h^2\] \hspace{1cm} (5.85)

by expression (5.59) in Theorem 5.3. The

\[\text{Cov}(\hat{Y}^* - \hat{Y}, Y - N \bar{\mu} | cs, R)\]
\[ H \sum_{i=r_h+1}^{n_h} \sum_{j=1}^{N_h} W_{hi} \text{Cov}( Y_{hi}^* - Y_{hj} | cs, R) \]

\[ = H \sum_{h=1}^{H} m_h (\sigma^2_h - \sigma^2_h) = 0, \quad (5.86) \]

since the $Y_{hi}^*$'s and the $Y_{hj}$'s are identically distributed within the imputation classes and the $Y_{hj}$'s are independently distributed. It follows from expressions (5.82), (5.85), and (5.86) that

\[ V( \hat{Y}^* - Y) = V( \hat{Y} - Y) \]

\[ + \sum_{h=1}^{H} \sigma^2_h E\{( k_h + 1) (m_h + t_h) (W_h^2 - n_h^{-1} S^2_{Wh}) \}. \quad (5.87) \]

We give a corollary for equal probability random imputation schemes.

**Corollary:** Let the assumptions of Theorem 5.2 hold, except that the imputation is any form of equal probability random imputation rather than the form given in Section 5.1.2. Let $g_{hi}$ be the number of times that respondent $hi$ is used for imputation, $h=1,...,H$, $i=1,...,r_h$. Then
\[ V(\hat{Y} - Y) = V(\hat{Y} - Y) + \sum_{h=1}^{H} \sigma_h^2 E\{ \sum_{i=1}^{r_h} \frac{g_{hi}}{n_h} \frac{g_{hi} + 1}{n_h - 1} \} \]

\[ \times \left[ \left( \sum_{i=1}^{n_h} W_{hi} \right)^2 - \sum_{i=1}^{n_h} W_{hi}^2 \right] . \]

(5.88)

We note that

\[ V(\hat{Y} - Y) = E[V(\hat{Y} - Y \mid FP)] , \]

(5.89)

since

\[ E(\hat{Y} - Y \mid FP) = 0 . \]

(5.90)

Cochran (1977) gives

\[ V(\hat{Y} - Y \mid FP) = \sum_{h=1}^{H} \sum_{i=1}^{N_h} (1 - \pi_{hi}) \pi_{hi}^{-1} Y_{hi}^2 \]

\[ + \sum_{h=1}^{H} \sum_{i=1}^{N_h} \sum_{g=1}^{N_h} \sum_{j=1}^{N_h} \pi_{hi,gj} - \pi_{hi} \pi_{gj} Y_{hi} Y_{gj} \]

(5.91)

as the variance of the Horvitz–Thompson estimator of the population total for a finite population, where \( \pi_{hi} \) is the probability that unit \( hi \) is in the sample and \( \pi_{hi,gj} \) is the probability that both unit \( hi \) and unit \( gj \) are in the sample. Cochran (1977) gives
\[ \hat{V}(\hat{Y} - Y \mid FP) = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \left( 1 - \pi_{hi} \right) \pi_{hi}^{-2} \hat{Y}_{hi} \]

\[ + \sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{g=1}^{H} \sum_{j=1}^{n_g} \frac{\pi_{hi,gj} - \pi_{hi} \pi_{gj}}{\pi_{hi} \pi_{gj} \pi_{hi,gj}} \hat{Y}_{hi} \hat{Y}_{gj} \]  \hspace{1cm} (5.92)

as an unbiased estimator of \( V(\hat{Y} - Y \mid FP) \) for a finite population. Since expression (5.92) is unbiased for the finite population, given expression (5.89), expression (5.92) is unbiased for the superpopulation. We now present a theorem for the bias when estimating \( V(\hat{Y} - Y \mid FP) \) from a sample with imputed values.

**Theorem 5.5:** Let the assumptions of Theorem 5.2 hold. Then

\[ E[\hat{V}(\hat{Y} - Y \mid FP)] = V(\hat{Y} - Y) \]

\[ + \sum_{h=1}^{H} \sigma_h^2 E\left\{ \frac{k_h n_h + (k_h + 2) t_h}{n_h (n_h - 1)} \right\} \]

\[ \times \sum_{i=1}^{n_h} \sum_{j=1}^{n_h} \frac{\pi_{hi,hj} - \pi_{hi} \pi_{hi,j}}{\pi_{hi} \pi_{hj} \pi_{hi,hj}} \]  \hspace{1cm} (5.93)

where

\[ \hat{V}^*(\hat{Y} - Y \mid FP) = \sum_{h=1}^{H} \sum_{i=1}^{N_h} \left( 1 - \pi_{hi} \right) \pi_{hi}^{-2} \hat{Y}_{hi}^2 \]
Proof: We have

\[ \hat{V}(\hat{Y} - Y | FP) = \hat{V}(\hat{Y} - Y | FP) \]

\[ + \sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{g=1}^{N_g} \left( 1 - \pi_{hi} \right) \pi_{hi}^2 (Y_{hi} - \hat{Y}_{hi}) \]

\[ + \sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{g=1}^{N_g} \sum_{j=1}^{n_g} \left( \pi_{hi,gj} - \pi_{hi} \pi_{gj} \right) \pi_{hi} \pi_{gj} \pi_{hi,gj} \]

\[ \times (Y_{hi}^* Y_{gj}^* - Y_{hi} Y_{gj}) . \] (5.95)

Averaging over draws from the superpopulation,

\[ E(\frac{Y_{hi}^* - Y_{hi}^2}{Y_{hi} | cs, R}) = 0 , \] (5.96)

since the \( Y_{hi}^* \)'s and the \( Y_{hi} \)'s are identically distributed within the imputation classes. Also, for \( hi \neq gj \),

\[ E(\frac{Y_{hi}^* Y_{gj}^* - Y_{hi} Y_{gj}}{cs, rp, ic, dm}) = 0 \] (5.97)
if $Y^*_h$ is not the same as $Y^*_g$, and

$$E( Y^*_h Y^*_h - Y^*_h Y^*_h | cs, rp, ic, dm ) = \sigma^2_h$$ (5.98)

if $Y^*_h$ and $Y^*_g$ are the same. Result (5.93) follows from the unbiasedness of $V(Y - Y|FP)$ and because there are $k_h n_h + (k_h + 2) t_h$ out of $n_h (n_h - 1)$ combinations that will give $Y^*_h$ and $Y^*_g$ the same in imputation class $h$ and no combination will give $Y^*_h$ and $Y^*_g$ the same if $h \neq g$.

We now give a corollary for equal probability random imputation schemes.

**Corollary:** Let the assumptions of Theorem 5.2 hold, except that the imputation is done by any equal probability random imputation scheme. Let $g_{hi}$ be as defined in the first corollary to Theorem 5.3. Then

$$E[ \hat{V}^*(\hat{Y} - Y | FP ) ] = V(\hat{Y} - Y) + \sum_{h=1}^{H} \sigma^2_h E\left[ \sum_{i=1}^{r_h} \frac{g_{hi}}{n_h} \left( \frac{g_{hi} + 1}{n_h - 1} \right) \right]$$

$$\times \left\{ \sum_{i=1}^{n_h} \sum_{j=1}^{n_h} \pi_{hi,hj}^{-1} - \pi_{hi}^{-1} \pi_{hj}^{-1} \right\}.$$ (5.99)

**Proof:** In Theorem 5.5, expression (5.93) dependents of the number of nonzero covariance terms. The result follows since each respondent contributes
We now evaluate the importance of the bias in \( \hat{V}^*(\hat{Y} - Y|FP) \). We work within imputation classes. We find the relative importance of the estimator of the bias in \( \hat{V}^*(\hat{Y} - Y|FP) \) compared to an estimator of the increase in variance due to imputing for the incomplete sample. An estimator of \( V(Y^* - Y) \) with no correction for the bias in \( \hat{V}^*(\hat{Y} - Y|FP) \) is

\[
\hat{V}(\hat{Y}^* - Y) = \hat{V}^*(\hat{Y} - Y | FP)
\]

The second term in this expression is an estimator of the increase in variance due to imputation for the incomplete sample. The estimated bias in \( \hat{V}^*(\hat{Y} - Y|FP) \) as an estimator of \( \hat{V}(\hat{Y} - Y|FP) \) is

\[
\frac{H}{\sum_{h=1}^{H} s_{rh}^2} \frac{k_h n_h + (k_h + 2) t_h}{n_h (n_h - 1)} \left( \sum_{i=1}^{n_h} \sum_{j=1}^{n_h} \pi_{hi} \pi_{hj} \right) .
\]

We define the relative importance of the bias in \( \hat{V}^*(\hat{Y} - Y|FP) \) in imputation class \( h \) as
An approach to evaluating the relative importance of the estimated bias in $\hat{V}^*(\hat{Y}-Y|FP)$ is to find

$$\xi_n = \max_{i \neq j} \left( \frac{\pi_{hi} \pi_{hj}}{\pi_{hi,hj}} - 1 \right). \quad (5.103)$$

If $\xi_n$ goes to zero as $n$ goes to infinity, where $n$ is the sample size, then the relative importance of the estimator of the bias in $\hat{V}^*(\hat{Y}-Y|FP)$ goes to zero as $n$ increases. For simple random sampling without replacement, $\xi_n = n^{-1}$. For stratified random sampling, $\xi_n = \max_i (n_i^{-1})$, where $n_i$ is the sample size in stratum $i$. Note that the relative importance of the estimator of the bias in $\hat{V}^*(\hat{Y}-Y|FP)$ does not depend on the number of respondents if one works within the imputation classes.

We now present an estimator for the $V(Y^*-Y)$.

Theorem 5.6: Let the assumptions of Theorem 5.2 hold. Then

$$\hat{V}(\hat{Y}^*-Y) = \hat{V}^*(\hat{Y}-Y|FP).$$
\[ + \sum_{h=1}^{H} s_{rh}^2 \{ n_h^{-1} ( n_h - 1 )^{-1} [ k_h n_h + ( k_h + 2 ) t_h ] \} \]

\[ \times \sum_{i=1}^{n_h} \sum_{j=1}^{n_h} \left\{ \pi_{hi}^{-1} \pi_{hj}^{-1} - \pi_{hi}^{-1} \pi_{hj}^{-1} \pi_{hi,hj}^{-1} ( \pi_{hi,hj} - \pi_{hi} \pi_{hj} ) \right\} \]

(5.104)

where \( s_{rh}^2 \) is the sample variance of the respondents in imputation class \( h \), is an unbiased estimator of \( V(\hat{Y} - Y) \).

Proof: The

\[ E\left[ s_{rh}^2 \{ n_h^{-1} ( n_h - 1 )^{-1} [ k_h n_h + ( k_h + 2 ) t_h ] \} \right] \]

\[ \times \sum_{i=1}^{n_h} \sum_{j=1}^{n_h} \left\{ \pi_{hi}^{-1} \pi_{hj}^{-1} - \pi_{hi}^{-1} \pi_{hj}^{-1} \pi_{hi,hj}^{-1} ( \pi_{hi,hj} - \pi_{hi} \pi_{hj} ) \right\} \]

\[ = \sigma_h^2 E\{ n_h^{-1} ( n_h - 1 )^{-1} [ k_h n_h + ( k_h + 2 ) t_h ] \} \]
since we assume that the units in the sample in imputation class $h$ are
independently and identically distributed and since the missing values are missing
at random within the imputation classes. By Theorem 5.5,

$$\mathbb{E}[\hat{V}^*(\hat{Y} - Y | FP)]$$

$$= \sum_{h=1}^{H} \sigma_h^2 \mathbb{E}\left\{ n_h^{-1} (n_h - 1)^{-1} \left[ k_h n_h + (k_h + 2) t_h \right] \right\}$$

$$\times \sum_{i=1}^{n_h} \sum_{j=1}^{n_h} \left( \pi_{hi} \pi_{hj} \pi_{hi,hj} \left( \pi_{hi,hj} - \pi_{hi} \pi_{hj} \right) \right)$$

$$= \mathbb{V}(\hat{Y} - Y).$$

By Theorem 5.3,

$$\mathbb{V}(\hat{Y} - Y) + \sum_{h=1}^{H} \sigma_h^2 \mathbb{E}\left\{ n_h^{-1} (n_h - 1)^{-1} \left[ k_h n_h + (k_h + 2) t_h \right] \right\}$$
We can see from expression (5.104) that we are replacing \((\pi_{hi} \pi_{hj})^{-1}\) by \((\pi_{hi, hj})^{-1}\) in the correction term when we take into account the bias in \(\hat{\nu}^* (Y - Y|FP)\).

We now give a corollary for equal probability random imputation schemes.

**Corollary 1:** Let the assumptions of Theorem 5.2 hold, except, let the imputation be any form of equal probability random imputation. Let the \(g_{hi}'s\) be as defined in the first corollary to Theorem 5.3. Then an unbiased estimator of \(V(\hat{Y}^* - Y)\) is

\[
V(\hat{Y}^* - Y) = V^*(\hat{Y} - Y|FP)
\]

\[
+ \sum_{h=1}^{H} \sum_{r=1}^{R_h} g_{hi} \left( \frac{g_{hi} - 1}{n_h} \right) \pi^{-1}_{hi,hj} = \sum_{i=1}^{n_h} \sum_{j=1}^{n_h} \pi^{-1}_{hi,hj}.
\]

We now present a second corollary for stratified random sampling.

**Corollary 2:** Let the assumptions of Theorem 5.2 hold. Let the sampling be stratified random sampling as described in Section 4.1. Then, using the
notation of Chapter 4,

\[
N^{-1} \hat{Y}^* = \hat{y}_{st}^* = \sum_{m=1}^{M} \sum_{h=1}^{H} \sum_{i=1}^{n_{mh}} W_m y_{mhi},
\]  

(5.109)

where

\[
W_m = n_{m}^{-1} N^{-1} N_m ,
\]  

(5.110)

\[
\pi_{mhi} = N_{m}^{-1} n_{m} ,
\]  

(5.111)

and

\[
\pi(mhi)(ngj) = N_{m}^{-1} N_{n}^{-1} n_{m} n_{n} .
\]  

(5.112)

for \( m \neq n \)

\[
= N_{m}^{-1} ( N_{m} - 1 )^{-1} n_{m} ( n_{m} - 1 )
\]  

(5.113)

for \( m = n, \ hi \neq gj \).

The

\[
\hat{V}^* ( \bar{y}_{st} - Y | FP )
\]

\[
= \sum_{m=1}^{M} N^{-2} n_{m}^{-2} ( 1 - \frac{n_{m}}{N_m} ) n_{m}^{-1} s_{m} ,
\]  

(5.114)
where \( s_m^2 \) is defined in (4.9) (see Cochran (1977, p.93)) and

\[
\hat{V}(\bar{y}_s^* - \bar{y}) = \sum_{m=1}^{M} \frac{N-m}{N} \left[ 1 - \frac{n_m}{N} \right] n^{-1} s_m^2
\]

\[
+ \sum_{h=1}^{H} s_{r.h}^2 \left[ \frac{k_h n_h + (k_h + 2) t_h}{n_h (n_h - 1)} \right] \left\{ \frac{M}{\sum_{m=1}^{M} n_{mh} W_m} \right\}^2
\]

\[
- \frac{M}{\sum_{m=1}^{M} n_{mh} W_m} \left[ 1 - \frac{n_{mh}}{n_m} \frac{n_{mh}}{n_m} - 1 \right] (1 - \frac{n_{mh}}{n_m}) \}
\]

is an unbiased estimator of \( V(\bar{Y} - \bar{Y}) \). Expression (5.115) agrees with expression (4.89) if we include the finite population correction factor in (4.89).

5.1.8. Estimation of the variance of the imputed total  

We now present an estimator of the variance of \( \bar{Y}^* \) that ignores the finite population correction factor. We first find the \( V(\bar{Y} - \bar{Y}) \) in terms of the \( \pi_h \)'s and the \( \pi_{hi,gj} \)'s. The

\[
V(\bar{Y} - \bar{Y}) = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \pi_{hi} \mu_h + \left( \sum_{h=1}^{H} \sum_{i=1}^{n_h} \pi_{hi} \sigma_h^2 \right)
\]

from expression (5.42) in Theorem 5.3. Using the \( a_{hi} \)'s defined in the first proof of Theorem 5.2,
\[(22.5.21)\]
\[
\left\{ \begin{array}{c}
\eta_N - (\eta_{\rho} + \eta_{\gamma}) \\
\eta_{N, H}
\end{array} \right\} \begin{array}{c}
1 = i \\
1 = \eta
\end{array} + 
\begin{array}{c}
\eta_N \\
\eta_{N, H}
\end{array} = (M_N - \lambda_0) \Lambda
\]

\[(22.5.21)\]

\[
\Pi_{\mu} \left( \Pi_{\mu} - 1 \right) = (\Pi_{\mu}) \Lambda
\]

\[(6.130)\]

\[
\Pi_{\mu} - \Sigma_{\mu} = \left( \Pi_{\mu} - \Sigma_{\mu} \right) \Lambda
\]

\[(6.118)\]

\[
\Pi_{\mu} = (\Pi_{\mu}) \Delta
\]

\[(5.118)\]

\[
\begin{array}{c}
\eta_{\gamma} \\
\eta_{\gamma, H}
\end{array} 
\begin{array}{c}
1 = i \\
1 = \eta
\end{array} + 
\begin{array}{c}
\eta_u \\
\eta_{u, H}
\end{array} = \begin{array}{c}
\eta_{\gamma} \\
\eta_{\gamma, H}
\end{array} 
\begin{array}{c}
1 = i \\
1 = \eta
\end{array}
\]

\[(\mu.11)\]

\[
\begin{array}{c}
\eta_{\gamma} \\
\eta_{\gamma, H}
\end{array} 
\begin{array}{c}
1 = i \\
1 = \eta
\end{array} + 
\begin{array}{c}
\eta_u \\
\eta_{u, H}
\end{array} = \begin{array}{c}
\eta_{\gamma} \\
\eta_{\gamma, H}
\end{array} 
\begin{array}{c}
1 = i \\
1 = \eta
\end{array}
\]

211
We now present an unbiased estimator of the $V(\hat{Y} - \bar{N})$.

**Theorem 5.7:** Let the assumptions of Theorem 5.2 hold. Then an unbiased estimator of the $V(\hat{Y} - \bar{N})$ is

$$
\hat{V}(\hat{Y} - \bar{N}) = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{g=1}^{n_g} \frac{\pi_{h_i,g_j}}{\pi_{h_i}} \frac{\pi_{g_j}}{\pi_{g_j}} Y_{hi} Y_{gj}
$$

$$
+ \sum_{h=1}^{H} \sum_{i=1}^{n_h} \left[ \frac{1 - \pi_{h_i}^2}{\pi_{h_i}} Y_{hi}^2 + \frac{1}{\pi_{h_i}} s_h^2 \right]
$$

$$
- \sum_{h=1}^{H} n_h^{-1} (n_h - 1)^{-1} \left( \sum_{i=1}^{n_h} \pi_{h_i}^{-1} \right) \left[ \sum_{i=1}^{n_h} \sum_{j=1}^{n_h} \pi_{h_i,g_j}^{-1} Y_{hi} Y_{hj} \right]
$$

$$
+ \sum_{h=1}^{H} \sum_{i=1}^{n_h} \left\{ \frac{1 - \pi_{h_i}^2}{\pi_{h_i}} + n_h^{-1} \left( \sum_{j=1}^{n_h} \pi_{h_j}^{-1} \right) \right\} Y_{hi}^2
$$

where $s_h^2$ is the sample variance of the $Y$'s in imputation class $h$. 

(5.123)
Proof: Taking expectations over draws from the superpopulation
and holding the choice of the sample constant

\[
E( \hat{V}( \hat{Y} - N\bar{\mu} ) | \text{cs} ) = \sum_{h=1}^{H} \sum_{i=1}^{N_h} \sum_{g=1}^{N_g} a_{hi} a_{gj} \frac{\pi_{hi,gj}}{\pi_{hi} \pi_{gj} \pi_{hi,gj}} \mu_h \mu_g
\]

\[
+ \sum_{h=1}^{H} \sum_{i=1}^{N_h} a_{hi} \left[ -\frac{1}{\pi_{hi}} (\mu_h^2 + \sigma_h^2) - \frac{1}{\pi_{hi}} \mu_h^2 \right]
\]

(5.124)

since the \( Y_{hi} \)'s are independently and identically distributed within the imputation
classes and independent between imputation classes. Taking the expectation over
the choices for the sample, we get

\[
E\{ E( \hat{V}( \hat{Y} - N\bar{\mu} ) | \text{cs} ) \} = \sum_{h=1}^{H} \sum_{i=1}^{N_h} \sum_{g=1}^{N_g} \sum_{j=1}^{N_g} a_{hi} a_{gj} \frac{\pi_{hi,gj}}{\pi_{hi} \pi_{gj}} \mu_h \mu_g
\]

\[
+ \sum_{h=1}^{H} \sum_{i=1}^{N_h} \sum_{j=1}^{N_g} \left[ -\frac{1}{\pi_{hi}} (\mu_h^2 + \sigma_h^2) - \mu_h^2 \right],
\]

(5.125)

since \( E(a_{hi} a_{gj}) = \pi_{hi,gj} \) for \( hi \neq gj \).

We now find the bias in \( \hat{V}( \hat{Y} - N\bar{\mu} ) \) when the augmented sample is used
instead of the full sample in the estimator.
Theorem 5.8: Let the assumptions of Theorem 2 hold. Let

\[ \hat{V}(\hat{Y} - N\mu) = \sum_h \sum_i \sum_g \sum_j \frac{\pi_{hi,gj} - \pi_{hi} \pi_{gj}}{\pi_{hi} \pi_{gj} \pi_{hi,gj}} Y_{hi} Y_{gj} \]

\[ \sum_h \sum_i \sum_g \sum_j \frac{\pi_{hi,gj} - \pi_{hi} \pi_{gj}}{\pi_{hi} \pi_{gj} \pi_{hi,gj}} Y_{hi} Y_{gj} \]

\[ \sum_h \sum_i \sum_g \sum_j \frac{\pi_{hi,gj} - \pi_{hi} \pi_{gj}}{\pi_{hi} \pi_{gj} \pi_{hi,gj}} Y_{hi} Y_{gj} \]

Then

\[ E[\hat{V}(\hat{Y} - N\mu)] = V(\hat{Y} - N\mu) + \sum_h \sigma^2 \sum_i \sum_j \frac{\pi_{hi,hi} - \pi_{hi} \pi_{hi}}{\pi_{hi} \pi_{hi} \pi_{hi}} \cdot \frac{n_h}{n_h - 1} \cdot \frac{1}{\pi_{hi}} \]

\[ \sum_{i \neq j} \left( \sum_i \sum_j \frac{\pi_{hi,hj} - \pi_{hi} \pi_{hj}}{\pi_{hi} \pi_{hj} \pi_{hi,hj}} \right) \]

Proof: The

\[ \hat{V}(\hat{Y} - N\mu) - \hat{V}(\hat{Y} - N\mu) \]

\[ = \sum_h \sum_i \sum_g \sum_j \frac{\pi_{hi,gj} - \pi_{hi} \pi_{gj}}{\pi_{hi} \pi_{gj} \pi_{hi,gj}} (Y_{hi} Y_{gj} - Y_{hi} Y_{gj}) \]
We know from Theorem 5.5 that

$$E[Y_{hi}^* - Y_{hi}^2 | cs, R] = 0$$

(5.129)

and

$$E[Y_{hi}^* Y_{gj}^* - Y_{hi} Y_{gj} | cs, R] = \frac{k_h n_h + (k_h + 2) t_h}{n_h (n_h - 1)} o^2$$

if $h=g$ and $i \neq j$

(5.130)

if $h \neq g$.

The result follows from the unbiasedness of $\hat{V}(\hat{Y} - \mu)$ and from substituting $E(Y_{hi}^* - Y_{hi}^2 | cs, R)$ and $E(Y_{hi}^* - Y_{hi} Y_{gj} | cs, R)$ into the expectation of expression (5.128). #
We now present an unbiased estimator of \( V(\hat{Y}^* - N\mu) \). Let

\[
\hat{V}(\hat{Y}^* - N\mu) = \hat{V}(\hat{Y} - N\mu) + \sum_{h=1}^{H} s^2_{rh} \frac{k_h n_h}{n_h} + \frac{k_h + 2}{n_h - 1}
\]

\[
\times \left[ \sum_{i=1}^{n_h} \sum_{j=1}^{n_h} \pi_{hi,hj}^{-1} + \sum_{i=1}^{n_h} \pi_{hi}^{-1} \right].
\]

(5.131)

Then, by Theorems 5.3 and 5.8 and since \( E(s^2_{rh} | cs,R) = \sigma^2_h \), expression (5.131) is an unbiased estimator of \( V(\hat{Y}^* - N\mu) \).

5.1.9. The superpopulation structure and sample properties for two characteristics

We now set up the superpopulation structure and sample properties for imputation with two characteristics. We then investigate the properties of the covariance when imputation is done.

Let us assume that we have the population set up described in Section 5.2. Assume that we have two characteristics, \( \{Y_{hi}\}_{h=1,\ldots,H, i=1,\ldots,N_h} \) and \( \{X_{hi}\}_{h=1,\ldots,H, i=1,\ldots,N_h} \). Observe that the same imputation cells are used for characteristic \( Y \) and for characteristic \( X \). Assume that a sample of size \( n \) is taken from the population as described in Section 5.1.1. Assume that an attempt is made to measure both \( Y_{hi} \) and \( X_{hi} \) for units \( hi, h=1,\ldots,H, i=1,\ldots,n_h \). Let \( r_{yh} \) be the number of units in imputation class \( h \) for which we are able to measure the \( Y \) characteristic. Let \( m_{yh} \) be the number of units in imputation class \( h \) for which we are unable to measure the \( Y \) characteristic. Then \( r_{yh} \geq 0 \) and \( m_{yh} \geq 0 \) and \( r_{yh} + m_{yh} < n_h \).
Let \( r_{xh} \) be the number of units in imputation class \( h \) for which we are able to measure the \( X \) characteristic. Let \( m_{xh} \) be the number of units in imputation class \( h \) for which we are unable to measure the \( X \) characteristic. Then \( r_{xh} \geq 0 \) and \( m_{xh} \geq 0 \) and \( r_{xh} + m_{xh} = n_h \). Let \( r_{xyh} \) be the number of units in imputation class \( h \) for which we are able to measure both characteristics. Let \( m_{xyh} \) be the number of units in imputation class \( h \) for which we are unable to measure both characteristics.

Let \( r_{xyh} \) be the number of units in imputation class \( h \) for which we are able to measure \( Y \) but not \( X \). Let \( m_{xyh} \) be the number of units in imputation class \( h \) for which we are able to measure \( X \) but not \( Y \). Then \( r_{xyh} + m_{xyh} + r_{xyh} + m_{xyh} = n_h \) and \( r_{xyh} \geq 0 \), \( m_{xyh} \geq 0 \), \( r_{xyh} \geq 0 \), and \( m_{xyh} \geq 0 \). Also, under the assumption that the nonrespondents in imputation class \( h \) are missing at random, \((r_{xyh}, m_{xyh}, r_{xyh}, m_{xyh})\) is a multinomial random variable with sample size \( n_h \). Let the probabilities of success for \((r_{xyh}, m_{xyh}, r_{xyh}, m_{xyh})\) be \((p_{rrh}, p_{mhh}, p_{rrh}, p_{mhh})\). Let \((p_{pyh}, p_{myh})\) be the marginal probabilities of response and nonresponse for \( Y \). Let \((p_{rxh}, p_{mrxh})\) be the marginal probabilities of response and nonresponse for \( X \). A two way table of the response patterns is given in Table 5.1.

In Table 5.2, we have the unknown population probabilities of response and missingness for \( X \) and \( Y \). Let

\[
R_1 = \begin{bmatrix}
      r_{xy1} & m_{xy1} & r_{xy1} & m_{xy1} \\
      \vdots & \vdots & \vdots & \vdots \\
      r_{xyH} & m_{xyH} & r_{xyH} & m_{xyH}
\end{bmatrix}
\]

be the matrix of the number of respondents and missing for all of the imputation classes. Let

\[
(5.132)
\]
Table 5.1. Response of X versus response of Y in imputation class h

<table>
<thead>
<tr>
<th></th>
<th>responds</th>
<th>missing</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>responds</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$r_{xyh}$</td>
<td>$r_{mxyh}$</td>
<td>$r_{yh}$</td>
</tr>
<tr>
<td>missing</td>
<td>$m_{xyh}$</td>
<td>$m_{xyh}$</td>
<td>$m_{yh}$</td>
</tr>
<tr>
<td></td>
<td>$r_{xh}$</td>
<td>$m_{xh}$</td>
<td>$n_{h}$</td>
</tr>
</tbody>
</table>

Table 5.2. Probabilities of responding and missing in imputation class h

<table>
<thead>
<tr>
<th></th>
<th>responds</th>
<th>missing</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>responds</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$p_{rth}$</td>
<td>$p_{rmh}$</td>
<td>$p_{ryh}$</td>
</tr>
<tr>
<td>missing</td>
<td>$p_{mrh}$</td>
<td>$p_{mmh}$</td>
<td>$p_{myh}$</td>
</tr>
<tr>
<td></td>
<td>$p_{rxh}$</td>
<td>$p_{mxh}$</td>
<td>1</td>
</tr>
</tbody>
</table>

$$R_2 = \begin{bmatrix} r_{y1} & m_{y1} & r_{x1} & m_{x1} \\ \vdots & \vdots & \vdots & \vdots \\ r_{yH} & m_{yH} & r_{xH} & m_{xH} \end{bmatrix}$$

(5.133)

be the matrix of the marginal number of respondents and missing values for all of
the imputation classes.

Assume that the \( Y \) characteristics and the \( X \) characteristics in imputation class \( h, h=1,\ldots,H \) are missing at random, where missing at random is defined in Section 2.3. Furthermore, assume that the \( Y \) characteristics and the \( X \) characteristics are missing with some covariance between the two missingness mechanisms and with possibly different probabilities of response.

Let the superpopulation model described in Section 5.1.4 hold for the two characteristics. Within each sub-population, let the mean of \((X,Y)\) be \((\mu_{xh}, \mu_{yh})\), and let the covariance of \((X,Y)\) be

\[
\Sigma_h = \begin{bmatrix}
\sigma_{xh}^2 & \sigma_{xh} \sigma_{yh} \\
\sigma_{xh} \sigma_{yh} & \sigma_{yh}^2
\end{bmatrix}, \quad h=1,\ldots,H.
\] (5.134)

5.1.10. Imputation methods for two variables We present three methods for random imputation. The first and third methods are for any type of covariance between the missingness mechanism for \( Y \) and the missingness mechanism for \( X \). The second method is applicable when \( X \) is missing only when \( Y \) is missing and vice versa, so that the correlation of the missingness mechanism of \( Y \) with the missingness mechanism of \( X \) is one.

For the first method, let imputation be done independently on the \( X \)'s and on the \( Y \)'s within each imputation class. Let

\[
k_{yh} = \left[ \frac{m_{yh}}{r_{yh}} \right],
\] (5.135)
and let

\[ k_{xh} = \left\lfloor \frac{m_{xh}}{r_{xh}} \right\rfloor, \quad (5.136) \]

where \([x]\) is the largest integer less than or equal to \(x\). Let

\[ t_{yh} = m_{yh} - k_{yh} r_{yh}, \quad (5.137) \]

and let

\[ t_{xh} = m_{xh} - k_{xh} r_{xh}. \quad (5.138) \]

Then

\[ n_h = (k_{yh} + 1) r_{yh} + t_{yh} \]

\[ = (k_{xh} + 1) r_{xh} + t_{xh}. \quad (5.139) \]

For the imputation, each \(Y_{hi}\) donates at least \(k_{yh}\) times to the missing \(Y\) values and \(t_{yh}\) of the \(Y_{hi}\)'s, where the \(t_{yh}\) of the \(Y_{hi}\)'s are chosen by simple random sampling without replacement for the \(r_{yh}\) \(Y\) respondents in imputation class \(h\), donate \(k_h + 1\) times to the missing \(Y\) values in imputation class \(h\). Also, independently of the \(Y\)'s, each \(X_{hi}\) donates at least \(k_{xh}\) times to the missing \(X\) values and \(t_{xh}\) of the \(X_{hi}\)'s, where the \(t_{xh}\) of the \(X_{hi}\)'s are chosen by simple random sampling without replacement from the \(r_{xh}\) \(X\) respondents in imputation class \(h\),
donate \( k_h + 1 \) times to the missing X values in imputation class h. Let the donors be assigned to the missing values at random.

For the second method, let X be missing only when Y is missing and vice versa. We impute the Y variables by the same method as described for method one, but instead of imputing the X variables independently, when we impute a Y variable for a given missing unit, we use the same donor to impute the X variable for that missing unit.

For the third method, if X and Y are both missing, they are imputed simultaneously from the set of observations that responded on both X and Y. If X is missing and Y is not missing or if Y is missing and X is not missing, then the missing observation is imputed from the set of respondents for the missing characteristic.

Let

\[ k_{xyh} = \lfloor \frac{m_{xy}}{r_{xyh}} \rfloor, \tag{5.140} \]

where \( \lfloor x \rfloor \) is the largest integer less than or equal to x. Let

\[ t_{xyh} = m_{xy} - k_{xyh} \cdot r_{xyh}. \tag{5.141} \]

Divide the set of observations in imputation class h into four sets, \( S_{1h}, S_{2h}, S_{3h}, \) and \( S_{4h} \), \( S_{1h} \) being the set for which X and Y are both missing, \( S_{2h} \) being the set for which X is missing and Y responds, \( S_{3h} \) being the set for which Y is missing and X responds, and \( S_{4h} \) being the set for which X and Y both respond. For elements in \( S_{1h} \), we impute the entire set of \( S_{4h} \) into \( S_{1h} \) randomly \( k_{xyh} \) times and then we
choose \( t_{xyh} \) of the elements of \( S_{4h} \) by simple random sampling without replacement to randomly fill in the remaining empty places in \( S_{1h} \). Let

\[
k_{yyh} = \left\lfloor \frac{m_{xyh}}{r_{yh}} \right\rfloor \quad (5.142)
\]

and

\[
k_{xxh} = \left\lfloor \frac{m_{xyh}}{r_{xh}} \right\rfloor , \quad (5.143)
\]

where \( \lfloor x \rfloor \) is the largest integer less than or equal to \( x \). Let

\[
t_{yyh} = m_{xyh} - k_{yyh} \left( \frac{r_{yh}}{r_{xh}} \right) \quad (5.144)
\]

and

\[
t_{xxh} = m_{xyh} - k_{xxh} \left( \frac{r_{xh}}{r_{xh}} \right) . \quad (5.145)
\]

Let all of the \( Y \) observations in \( S_{2h} \cup S_{4h} \) be imputed into \( S_{3h} \) \( k_{yyh} \) times. Choose \( t_{yyh} \) of the \( Y \) observations in \( S_{2h} \cup S_{4h} \) by simple random sampling without replacement to fill in the remaining \( t_{yyh} \) missing \( Y \) values in \( S_{3h} \). Let all of the \( X \) observations in \( S_{3h} \cup S_{4h} \) be imputed into \( S_{2h} \) \( k_{xxh} \) times. Choose \( t_{xxh} \) of the \( X \) observations in \( S_{3h} \cup S_{4h} \) by simple random sampling without replacement to fill in the remaining \( t_{xxh} \) missing \( X \) values in \( S_{2h} \). The observations used for imputation are assigned to the missing observations at random.
5.1.11. The covariance of two imputed totals under the superpopulation model

We write our estimators of the population totals for X and Y in a new form. Let

\[
\hat{Y}^* = \sum_{h=1}^{H} \sum_{i=1}^{n_h} W_{hi} Y_{hi}^*
\]

where \( Y_{hi}^*\) is a 0–1 random variable that takes on the value 1 if \( Y_{hi} \) responds and that takes on the value of 0 otherwise and \( \tau_{yhi,j} \) is a 0–1 random variable that takes on the value 1 if \( Y_{hi,j} \) was imputed using \( Y_{hi} \) and that takes on the value 0 otherwise. Let

\[
\hat{X}^* = \sum_{\gamma=1}^{\hat{H}} \sum_{i=1}^{n_i} W_{hi} X_{hi}^*
\]

where \( X_{hi}^*\) is a 0–1 random variable that takes on the value 1 if \( X_{hi} \) responds and that takes on the value 0 otherwise and \( \tau_{xhi,j} \) is a 0–1 random variable that takes on the value 1 if \( X_{hi,j} \) was imputed using \( X_{hi} \) and takes on the value 0 otherwise.
and let

$$\Gamma = \begin{bmatrix} \gamma_{y11} & \gamma_{x11} \\ \vdots & \vdots \\ \gamma_{yHL} & \gamma_{xHL} \end{bmatrix}$$  \hfill (5.148)

and let

$$T = \begin{bmatrix} \tau_{y11,1} & \tau_{x11,1} \\ \vdots & \vdots \\ \tau_{yHL,L} & \tau_{xHL,L} \end{bmatrix}$$  \hfill (5.149)

Then $\Gamma$ describes the response mechanism and $T$ describes the imputation.

We present a lemma describing the properties of the $\gamma_{yhi}'s$, $\gamma_{xhi}'s$, $\tau_{yhi,j}'s$, and $\tau_{xhi,j}'s$.

**Lemma 5.1:** Let characteristics $X$ and $Y$ be missing at random within the imputation classes from a sample. Let random imputation from within the imputation classes be done to fill in the missing values. Let $\gamma_{yhi}'$, $\gamma_{xhi}'$, $\tau_{yhi,j}'$ and $\tau_{xhi,j}'$ be as described above in this section. Then, if $i \neq j$ and $i \neq k$,

$$E( \gamma_{yhi} \gamma_{xhi} | cs, R_1 ) = \frac{\tau_{xyh}}{n_h}$$  \hfill (5.150)

and

$$E( \gamma_{yhi} \gamma_{xhi} \tau_{xhi,j} | cs, R_1 )$$
\[ P( \text{\(X_{hi}\)} donates to \(X_{hj}\) | \(Y_{hi}\) and \(X_{hi}\) both respond, \(X_{hj}\) is missing, cs, \(R_1\)) \]

and

\[ E( \gamma_{xhi} \gamma_{yhi} \tau_{yhi,j} \mid cs, R_1) \]

\[ = \frac{r_{xyh}}{n_h} \frac{m_{xh}}{n_h-1} P( \text{\(Y_{hi}\) donates to \(Y_{hj}\) | \(Y_{hi}\) and \(X_{hi}\) both respond, \(Y_{hj}\) is missing, cs, \(R_1\))} \]

and

\[ E( \gamma_{yhi} \gamma_{xhi} \tau_{yhi,j} \tau_{xhi,k} \mid cs, R_1) \]

\[ = \frac{r_{xyh}}{n_h} \left( \frac{m_{yh} m_{xh} - m_{xyh}}{n_h - 1} \right) \left( \frac{n - 2}{n_h - 2} \right) \]

\[ \times P( \text{\(Y_{hi}\) donates to \(Y_{hj}\) and \(X_{hi}\) donates to \(X_{hk}\) | \(Y_{hi}\) and \(X_{hi}\) both respond, \(Y_{hj}\) and \(X_{hk}\) are both missing, cs, \(R_1\))} \]

if \(j \neq k\)
\[ \frac{r_{xyh} m_{xyh}}{n_h (n_h - 1)} \]

\[
\times P(\ Y_{hi} \text{ donates to } Y_{hj} \text{ and } X_{hi} \text{ donates to } X_{hj} \mid Y_{hi} \text{ and } X_{hi} \text{ both respond,}
\]

\[ Y_{hj} \text{ and } X_{hj} \text{ are both missing, cs, } R_1 \) \tag{5.154}

if \( j = k \).

If \( i=j \) and (or) \( i=k \), then expressions (5.151) to (5.154) equal zero.

Also, for the first method of imputation given in Section 5.1.10 (imputing \( X \) and \( Y \) independently), if \( i\neq j \) and \( i\neq k \), then

\[
P(\ X_{hi} \text{ donates to } X_{hj} \mid Y_{hi} \text{ and } X_{hi} \text{ both respond, } X_{hj} \text{ is missing, cs, } R_1 )
\]

\[ = \frac{1}{r_{xh}} \tag{5.155} \]

and

\[
P(\ Y_{hi} \text{ donates to } Y_{hj} \mid Y_{hi} \text{ and } X_{hi} \text{ both respond, } Y_{hj} \text{ is missing, cs, } R_1 )
\]

\[ = \frac{1}{r_{yh}} \tag{5.156} \]

and
P( \( Y_{hi} \) donates to \( Y_{hj} \) and \( X_{hi} \) donates to \( X_{hk} \) | \( Y_{hi} \) and \( X_{hi} \) both respond,

\( Y_{hj} \) and \( X_{hk} \) are both missing, cs, \( R_1 \) )

\[ = \frac{1}{r_{yh}} \frac{1}{r_{xh}}. \]  

(5.157)

Also, for the second method of imputation given in Section 5.1.10 (\( X \) imputed concurrently with \( Y \)), if \( i \neq j \) and \( i \neq k \), then

P( \( X_{hi} \) donates to \( X_{hj} \) | \( Y_{hi} \) and \( X_{hi} \) both respond, \( X_{hj} \) is missing, cs, \( R_1 \) )

\[ = \frac{1}{r_{yh}} \]  

(5.158)

and

P( \( Y_{hi} \) donates to \( Y_{hj} \) | \( Y_{hi} \) and \( X_{hi} \) both respond, \( Y_{hj} \) is missing, cs, \( R_1 \) )

\[ = \frac{1}{r_{yh}} \]  

(5.159)

and

P( \( Y_{hi} \) donates to \( Y_{hj} \) and \( X_{hi} \) donates to \( X_{hk} \) | \( Y_{hi} \) and \( X_{hi} \) both respond,

\( Y_{hj} \) and \( X_{hk} \) are both missing, cs, \( R_1 \) )
For the third imputation method described in Section 5.1.10, if $i \neq j$ and $i \neq k$, then

\[
\Pr( X_{hi} \text{ donates to } X_{hj} \mid Y_{hi} \text{ and } X_{hi} \text{ both respond, } X_{hj} \text{ is missing, cs, } R_1 )
\]

\[
= \frac{m_{xyh}}{m_{xh}} \frac{1}{r_{xyh}} + \frac{rm_{xyh}}{m_{xh}} \frac{1}{r_{xh}}
\]  

(5.162)

and

\[
\Pr( Y_{hi} \text{ donates to } Y_{hj} \mid Y_{hi} \text{ and } X_{hi} \text{ both respond, } Y_{hj} \text{ is missing, cs, } R_1 )
\]

\[
= \frac{m_{xyh}}{m_{yh}} \frac{1}{r_{xyh}} + \frac{m_{rxyh}}{m_{yh}} \frac{1}{r_{yh}}
\]  

(5.163)

and
\[ P( Y_{hi} \text{ donates to } Y_{hj} \text{ and } X_{hi} \text{ donates to } X_{hk} \mid Y_{hi} \text{ and } X_{hi} \text{ both respond,} Y_{hj} \text{ and } X_{hk} \text{ are both missing, cs, } R_1 ) \]

\[ = \frac{r_{xyh}}{m_{xyh}} \frac{m_{xyh}}{m_{xyh} - m_{xyh}} \frac{1}{r_{xh}} \frac{1}{r_{yh}} + \frac{m_{xyh}}{m_{xyh} - m_{xyh}} \frac{r_{xyh}}{r_{xyh}} \frac{1}{r_{xyh}} \frac{1}{r_{xyh}} \]

\[ + \frac{r_{xyh}}{m_{xyh}} \frac{m_{xyh}}{m_{xyh} - m_{xyh}} \frac{1}{r_{xh}} \frac{1}{r_{xyh}} + \frac{m_{xyh}^2}{m_{xyh} - m_{xyh}} \frac{r_{xyh}}{r_{xyh}} \frac{1}{r_{xyh}} \frac{1}{r_{xyh}} \]

\[ \times \frac{k_{xyh}}{r_{xyh}} \frac{m_{xyh} + t_{xyh} - r_{xyh}}{m_{xyh} (m_{xyh} - 1)} \]  

(5.164)

if \( j \neq k \)

\[ = \frac{1}{r_{xyh}}. \]  

(5.165)

if \( j = k \).

Also, if the X's and the Y's are missing independently, then

\[ E( r_{xyh} \mid \text{cs, } R_2 ) = \frac{r_{yh} r_{xh}}{n_h} \]  

(5.166)

and
Also, if the X's are missing only when the Y's are missing, \[ E( r_{xy} | cs, R_2 ) = r_{yh} \] (5.168)

and

\[ E( r_{xy} m_{xy} | cs, R_2 ) = r_{yh} m_{yh} \] (5.169)

Proof: For (5.150),

\[ E( \gamma_{hi} \gamma_{xhi} | cs, R_1 ) = P( Y_{hi} and X_{hi} both respond | cs, R_1 ) \]

\[ = \frac{r_{xy}}{n_h} \] (5.170)

under the multinomial model, which applies since the missing observations are missing at random.

For (5.151),

\[ E( \gamma_{hi} \gamma_{xhi} \tau_{xhi,j} | cs, R_1 ) = P( Y_{hi} and X_{hi} both respond | cs, R_1 ) \times P( X_{hj} is missing | Y_{hi} and X_{hi} both respond, cs, R_1 ) \]
\[ \times P( \tau_{xhi,j} = 1 \mid Y_{hi} \text{ and } X_{hi} \text{ both respond, } X_{hj} \text{ is missing, } cs, R_1 ) \]

\[ = \frac{r_{xyh}}{n_h} \frac{m_{xh}}{n_h - 1} \times P( X_{hi} \text{ donates to } X_{hj} \mid Y_{hi} \text{ and } X_{hi} \text{ both respond, } X_{hj} \text{ is missing, } cs, R_1 ) \]  

(5.171)

if \( i \neq j \)

\[ = 0 \]  

(5.172)

if \( i = j \),

under the multinomial model.

For (5.152), switch the \( X \)'s and \( Y \)'s in (5.151).

For (5.153) and (5.154),

\[ E( \tau_{yhi}^{xhi} \tau_{yhi,j}^{xhi,k} \mid \text{cs, } R_1 ) = P( Y_{hi} \text{ and } X_{hi} \text{ both respond } \mid \text{cs, } R_1 ) \]

\[ \times P( Y_{hj} \text{ and } X_{hk} \text{ are both missing } \mid Y_{hi} \text{ and } X_{hi} \text{ both respond, cs, } R_1 ) \]

\[ \times P( \tau_{yhi,j}^{xhi,k} = 1 \mid Y_{hi} \text{ and } X_{hi} \text{ both respond, } Y_{hj} \text{ and } X_{hk} \text{ are missing, cs, } R_1 ) \]
\[ r_{xy} = \frac{m_{yh} m_{xh} - m_{xyh}}{n_h (n_h - 1) (n_h - 2)} \]

\[ x \cdot P( Y_{hi} \text{ donates to } Y_{hj} \text{ and } X_{hi} \text{ donates to } X_{hk} \mid Y_{hi} \text{ and } X_{hi} \text{ both respond}, \\
Y_{hj} \text{ and } X_{hk} \text{ are missing, } cs, \ R_1 ) \] (5.173)

if \( i \neq j \) and \( i \neq k \) and \( j \neq k \)

\[ = \frac{r_{xy} m_{xyh}}{n_h (n_h - 1)} \]

\[ x \cdot P( Y_{hi} \text{ donates to } Y_{hj} \text{ and } X_{hi} \text{ donates to } X_{hj} \mid Y_{hi} \text{ and } X_{hi} \text{ both respond}, \\
Y_{hj} \text{ and } X_{hj} \text{ are missing, } cs, \ R_1 ) \] (5.174)

if \( i \neq j \) and \( i \neq k \) and \( j = k \)

\[ = 0 \] (5.175)

if \( i = j \) and (or) \( i = k \),

under the multinomial model.

For (5.155), for the first imputation method,

\[ P( X_{hi} \text{ donates to } X_{hj} \mid Y_{hi} \text{ and } X_{hi} \text{ both respond, } X_{hj} \text{ is missing, } cs, \ R_1 ) \]
\[ \begin{align*}
&= P( Y_{hi} \text{ and } X_{hi} \text{ both respond } | \ X_{hi} \text{ donates to } X_{hj}, \ X_{hj} \text{ is missing, cs, } R_1 ) \\
&\quad \times P( X_{hi} \text{ donates to } X_{hj} | \ X_{hj} \text{ is missing, cs, } R_1 ) \\
&\quad \times \{ P( Y_{hi} \text{ and } X_{hi} \text{ both respond } | \ X_{hj} \text{ is missing, cs, } R_1 ) \}^{-1} \\
&= (\frac{r_{xyh}}{r_{xh}}) \left( \frac{1}{n_h - 1} \right) \left( \frac{r_{xyh}}{n_h - 1} \right)^{-1} \\
&= \frac{1}{r_{xh}}, \quad (5.176)
\end{align*} \]

since the missing values are missing at random and the donors are chosen and assigned randomly.

For (5.156), switch the X's and Y's in the proof for (5.155).

For (5.157), the X's and the Y's are imputed independently, so

\[ \begin{align*}
P( X_{hi} \text{ donates to } X_{hk} \text{ and } Y_{hi} \text{ donates to } Y_{hj} \\
| Y_{hi} \text{ and } X_{hi} \text{ both respond, } X_{hk} \text{ and } Y_{hj} \text{ are both missing, cs, } R_1 )
&= P( X_{hi} \text{ donates to } X_{hk} | Y_{hi} \text{ and } X_{hi} \text{ both respond, } X_{hk} \text{ is missing, cs, } R_1 ) \\
&\quad \times P( Y_{hi} \text{ donates to } Y_{hj} | Y_{hi} \text{ and } X_{hi} \text{ both respond, } Y_{hj} \text{ is missing, cs, } R_1 )
\end{align*} \]
For the second imputation method, for (5.158) and (5.159), the probabilities are \( r_{yh}^{-1} \), since all of the missing units have an equal chance of being filled in by each of the respondents, since the missing values are missing at random and the donors are chosen and assigned randomly.

For (5.160), we introduce new notation. Let \( k_{hi,j} \) be a 0–1 random variable that takes on the value of zero if unit hi does not donate to unit hj as a member of the total number of respondents that are imputed \( k_{yh} \) times and that takes on the value of one if unit hi does donate to unit hj as a member of the total number of respondents that are imputed \( k_{yh} \) times. Let \( \delta_{hi,j} \) be a 0–1 random variable that takes on the value of zero if unit hi does not donate to unit hj as a member of the \( t_{yh} \) respondents that are chosen by simple random sampling without replacement from the \( r_{yh} \) respondents in imputation class h and that takes on the value one if unit hi does donate to unit hj as a member of the \( t_{yh} \) respondents that are chosen by simple random sampling without replacement from the \( r_{yh} \) respondents in imputation class h. Call the imputation non–random imputation if the imputed value is one of the \( k_{yh} r_{yh} \) imputed values that are from the replication \( k_{yh} \) times of the set of respondents in imputation class h. Call the imputation random imputation if the imputed value is one of the \( t_{h} \) imputed values chosen by simple random sampling without replacement from the set of respondents in imputation class h. Then

\[
= \frac{1}{r_{xh}} \frac{1}{r_{yh}}.
\]
\[
\tau_{yhi,j} = k_{hi,j} + \delta_{hi,j} \quad (5.178)
\]

and

\[
\tau_{xhi,j} = k_{hi,j} + \delta_{hi,j} \quad (5.179)
\]

If \( j \neq k \) and \( i \neq j \) and \( i \neq k \), then

\[
P( k_{hi,j} k_{hi,k} = 1 \mid Y_{hi} \text{ and } X_{hi} \text{ both respond}, Y_{hj} \text{ and } X_{hk} \text{ are missing, cs, } R_1 )
\]

\[
= P( k_{hi,j} \text{ and } k_{hi,k} \text{ both equal one } \mid Y_{hi} \text{ responds,}
\]

\[
Y_{hj} \text{ and } Y_{hk} \text{ are missing, cs, } R_1 )
\]

\[
= P( Y_{hi} \text{ donates non-randomly to } Y_{hj} \mid Y_{hi} \text{ responds,}
\]

\[
Y_{hj} \text{ and } Y_{hk} \text{ are missing, cs, } R_1 )
\]

\[
\times P( Y_{hi} \text{ donates non-randomly to } Y_{hk} \mid Y_{hi} \text{ donates non-randomly to } Y_{hj},
\]

\[
Y_{hi} \text{ responds, } Y_{hj} \text{ and } Y_{hk} \text{ are missing, cs, } R_1 )
\]

\[
= \frac{k_{yh}}{m_{yh}} \frac{k_{yh} - 1}{m_{yh} - 1}, \quad (5.180)
\]

because all of the respondents are equally likely to donate \( k_{yh} \) times non-randomly
1.25 into a choice of \( m_{yh} \) missing points.

Since \( X \) is imputed simultaneously with \( Y \), if \( j \neq k \) and \( i \neq j \) and \( i \neq k \),

\[
P(k_{hi,j}, \delta_{hi,k} = 1 \mid Y_{hi} \text{ and } X_{hi} \text{ both respond, } Y_{hj} \text{ and } X_{hk} \text{ are missing, cs, } R_1)
\]

\[
= P(k_{hi,j}, \delta_{hi,k} = 1 \mid Y_{hi} \text{ responds, } Y_{hj} \text{ and } Y_{hk} \text{ are missing, cs, } R_1)
\]

\[
= P(Y_{hi} \text{ donates non-randomly to } Y_{hj} \mid Y_{hi} \text{ responds, } Y_{hj} \text{ and } Y_{hk} \text{ are missing, cs, } R_1)
\]

\[
\times P(Y_{hi} \text{ donates randomly} \mid Y_{hi} \text{ donates non-randomly to } Y_{hj}, Y_{hi} \text{ responds, } Y_{hj} \text{ and } Y_{hk} \text{ are missing, cs, } R_1)
\]

\[
\times P(Y_{hi} \text{ donates randomly to } Y_{hk} \mid Y_{hi} \text{ donates randomly, } Y_{hi} \text{ donates non-randomly to } Y_{hj}, Y_{hi} \text{ responds, } Y_{hj} \text{ and } Y_{hk} \text{ are missing, cs, } R_1)
\]

\[
= \frac{k_{yh}}{m_{yh}} \frac{t_{yh}}{r_{yh}} \frac{1}{m_{yh} - 1}, \quad (5.181)
\]

because all of the respondents are equally likely to donate non-randomly \( k_{yh} \) times and equally likely to donate randomly one time into the \( m_{yh} \) missing units and because the chance of a given unit donating randomly is \( r_{yh}^{-1} t_{yh} \).

Since \( X \) is imputed simultaneously with \( Y \), if \( i \neq j \) and \( i \neq k \) and \( j \neq k \), then
\[ P( \delta_{hi,j} \delta_{hi,k} = 1 \mid Y_{hi} \text{ responds, } Y_{hj} \text{ and } Y_{hk} \text{ are missing, cs, } R_1 ) \]

\[ = P( Y_{hi} \text{ donates randomly to } Y_{hj} \text{ and } Y_{hi} \text{ donates randomly to } Y_{hk} \]

\[ \mid Y_{hi} \text{ and } X_{hi} \text{ both respond, } Y_{hj} \text{ and } Y_{hk} \text{ are missing, cs, } R_1 ) \]

\[ = 0 , \quad (5.182) \]

since \( Y_{hi} \) donates randomly at most once.

Using expressions (5.180), (5.181), and (5.182), for \( i \neq j \) and \( i \neq k \) and \( j \neq k \),

\[ P( \tau_{yhi,j} \tau_{xhi,k} = 1 \mid Y_{hi} \text{ and } X_{hi} \text{ both respond, } \]

\[ Y_{hj} \text{ and } X_{hk} \text{ are both missing, cs, } R_1 ) \]

\[ = E( \tau_{yhi,j} \tau_{xhi,k} \mid Y_{hi} \text{ and } X_{hi} \text{ both respond, } \]

\[ Y_{hj} \text{ and } X_{hk} \text{ are both missing, cs, } R_1 ) \]

\[ = E( k_{hi,j} k_{hi,k} + k_{hi,j} \delta_{hi,k} + \delta_{hi,j} k_{hi,k} + \delta_{hi,j} \delta_{hi,k} \]

\[ \mid Y_{hi} \text{ and } X_{hi} \text{ both respond, } Y_{hj} \text{ and } X_{hk} \text{ are both missing, cs, } R_1 ) \]

\[ = P( k_{hi,j} k_{hi,k} = 1 \mid Y_{hi} \text{ and } X_{hi} \text{ both respond}, \]

\[ ]
\[ Y_{hj} \text{ and } X_{hk} \text{ are both missing, cs, } R_1 \]

\[ + P( \delta_{hi,j}\delta_{hi,k} = 1 \mid Y_{hi} \text{ and } X_{hi} \text{ both respond,} \]

\[ Y_{hj} \text{ and } X_{hk} \text{ are both missing, cs, } R_1 \]

\[ + P( \delta_{hi,j}\delta_{hi,k} = 1 \mid Y_{hi} \text{ and } X_{hi} \text{ both respond,} \]

\[ Y_{hj} \text{ and } X_{hk} \text{ are both missing, cs, } R_1 \]

\[ + P( \delta_{hi,j}\delta_{hi,k} = 1 \mid Y_{hi} \text{ and } X_{hi} \text{ both respond,} \]

\[ Y_{hj} \text{ and } X_{hk} \text{ are both missing, cs, } R_1 \]

\[ = \frac{k_{yh}}{m_{yh}} \frac{k_{yh} - 1}{m_{yh} - 1} + 2 \frac{k_{yh}}{m_{yh}} \frac{t_{yh}}{r_{yh}} \frac{1}{m_{yh} - 1} \]

\[ = \frac{k_{yh}n_{h} + k_{yh}(t_{yh} - 2r_{yh})}{r_{yh}m_{yh}(m_{yh} - 1)}. \quad (5.183) \]

For (5.161), since \( X \) is imputed simultaneously with \( Y \), if \( i \neq j \), then

\[ P( \tau_{hi,j}\tau_{hi,j} = 1 \mid Y_{hi} \text{ and } X_{hi} \text{ both respond,} \]

\[ Y_{hj} \text{ and } X_{hj} \text{ are both missing, cs, } R_1 \]
because all of the respondent Y's are equally likely to be used as a donor.

For imputation method three, for (5.162), if i≠j, then,

\[
P(X_{hi} \text{ donates to } X_{hj} \mid Y_{hi} \text{ and } X_{hi} \text{ both respond, } \nonumber
\]

X_{hj} \text{ is missing, cs, } R_1 \nonumber
\]

\[
= P(X_{hi} \text{ donates to } X_{hj} \cap (Y_{hj} \text{ responds } \cup Y_{hj} \text{ is missing }) \nonumber
\]

| Y_{hi} \text{ and } X_{hi} \text{ both respond, } X_{hj} \text{ is missing, cs, } R_1 \nonumber
\]

\[
= P(X_{hi} \text{ donates to } X_{hj} \mid Y_{hj} \text{ responds, } Y_{hi} \text{ and } X_{hi} \text{ both respond, } \nonumber
\]

X_{hj} \text{ is missing, cs, } R_1 \nonumber
\]

× \nonumber

\[
P(Y_{hj} \text{ responds } \mid Y_{hi} \text{ and } X_{hi} \text{ both respond, } X_{hj} \text{ is missing, cs, } R_1 \nonumber
\]

+ \nonumber

\[
P(X_{hi} \text{ donates to } X_{hj} \mid Y_{hj} \text{ is missing, } Y_{hi} \text{ and } X_{hi} \text{ both respond, } \nonumber
\]

X_{hj} \text{ is missing, cs, } R_1 \nonumber
\]
\[ P( Y_{hi} \text{ is missing} | Y_{hi} \text{ and } X_{hi} \text{ both respond, } X_{hj} \text{ is missing, cs, } R_1 ) \]

\[ = \frac{1}{r_{xh}} \frac{r_m}{m_{xyh}} + \frac{1}{r_{xyh}} \frac{m_{xyh}}{m_{xh}}, \quad (5.185) \]

since if \( Y_{hj} \) responds, \( X_{hj} \) is imputed from the set of all responding \( X \) observations in imputation class \( h \), and if \( Y_{hj} \) is missing, \( X_{hj} \) is imputed out of the set where both \( X \) and \( Y \) respond in imputation class \( h \).

For (5.163), switch the \( X \)'s and \( Y \)'s in the proof for (5.162).

For (5.164), if \( i \neq j \) and \( i \neq k \) and \( j \neq k \), then

\[ P( Y_{hi} \text{ donates to } Y_{hj} \text{ and } X_{hi} \text{ donates to } X_{hk} | Y_{hi} \text{ and } X_{hi} \text{ both respond, } Y_{hj} \text{ and } X_{hk} \text{ both are missing, cs, } R_1 ) \]

\[ = P( Y_{hi} \text{ donates to } Y_{hj} \text{ and } X_{hi} \text{ donates to } X_{hk} | Y_{hi} \text{ and } X_{hi} \text{ both respond, } Y_{hj} \text{ and } X_{hk} \text{ both are missing, cs, } R_1 ) \]

\[ \cap ( X_{hj} \text{ and } Y_{hk} \text{ are both respondents } \cup X_{hj} \text{ is a respondent and } Y_{hk} \text{ is missing } \]

\[ \cup X_{hj} \text{ is missing and } Y_{hk} \text{ is a respondent } \cup X_{hj} \text{ and } Y_{hk} \text{ are both missing } ) \]

\[ | Y_{hi} \text{ and } X_{hi} \text{ both respond, } Y_{hj} \text{ and } X_{hk} \text{ both are missing, cs, } R_1 ] \]

\[ = P( Y_{hi} \text{ donates to } Y_{hj} \text{ and } X_{hi} \text{ donates to } X_{hk} | X_{hj} \text{ and } Y_{hk} \text{ are respondents, } \]

\[ Y_{hi} \text{ and } X_{hi} \text{ both respond, } Y_{hj} \text{ and } X_{hk} \text{ both are missing, cs, } R_1 ) \]
\[ * P( X_{hj} \text{ and } Y_{hk} \text{ are respondents} \]

\[ | Y_{hi} \text{ and } X_{hi} \text{ both respond, } Y_{hj} \text{ and } X_{hk} \text{ both are missing, cs, } R_1 ) \]

\[ + P( Y_{hi} \text{ donates to } Y_{hj} \text{ and } X_{hi} \text{ donates to } X_{hk} \]

\[ | X_{hj} \text{ responds and } Y_{hk} \text{ is missing, } \]

\[ Y_{hi} \text{ and } X_{hi} \text{ both respond, } Y_{hj} \text{ and } X_{hk} \text{ both are missing, cs, } R_1 ) \]

\[ * P( X_{hj} \text{ responds and } Y_{hk} \text{ is missing} \]

\[ | Y_{hi} \text{ and } X_{hi} \text{ both respond, } Y_{hj} \text{ and } X_{hk} \text{ both are missing, cs, } R_1 ) \]

\[ + P( Y_{hi} \text{ donates to } Y_{hj} \text{ and } X_{hi} \text{ donates to } X_{hk} \]

\[ | X_{hj} \text{ is missing and } Y_{hk} \text{ responds, } \]

\[ Y_{hi} \text{ and } X_{hi} \text{ both respond, } Y_{hj} \text{ and } X_{hk} \text{ both are missing, cs, } R_1 ) \]

\[ * P( X_{hj} \text{ is missing and } Y_{hk} \text{ responds} \]

\[ | Y_{hi} \text{ and } X_{hi} \text{ both respond, } Y_{hj} \text{ and } X_{hk} \text{ both are missing, cs, } R_1 ) \]
\[ + \frac{1}{r_{xy}} \frac{1}{r_{xh}} \left( \frac{r_m x y h}{m_{xh}} \frac{r_m x y h}{m_{yh}} \frac{m_{xyh}}{m_{xy}} \right) \]

\[ + \frac{1}{r_{xy}} \frac{1}{r_{xyh}} \left( \frac{m_{xyh}}{m_{xh}} \frac{m_{xyh}}{m_{yh}} \frac{m_{xyh}}{m_{xyh}} \right) \]

\[ + \frac{1}{r_{xyh}} \frac{1}{r_{xh}} \left( \frac{r_m x y h}{m_{xh}} \frac{m_{xyh}}{m_{yh}} \frac{m_{xyh}}{m_{xy}} \right) \]

\[ + \frac{k_{xyh}}{r_{xyh}} \frac{m_{xyh}}{m_{xyh}} \left( \frac{t_{xyh} - r_{xyh}}{m_{xyh} - 1} \right) \]

\[ \times \frac{m_{xyh}^2 - m_{xyh}}{m_{xh}m_{yh} - m_{xyh}}. \]  

(5.186)

For (5.165), both \( X_{hj} \) and \( Y_{hj} \) are missing, so the respondents used for donors are taken out of the set of units that responded on both characteristics and the probability that any one of the donors donates to unit hj is \( r_{xyh}^{-1} \).

For the response mechanism of \( X \) independent of the response mechanism of
Y, for (5.166), it is a standard result under the multinomial model that under the independence of X and Y,

\[ E(r_{xyh} \mid cs, R_2) = \frac{r_{xh} r_{yh}}{n_h}, \]  

(5.187)

see Bishop, Fienberg, and Holland (1975, p. 28).

For (5.167),

\[ E(r_{xyh} m_{xyh}) = n_h (n_h - 1) p_{rrh} p_{mnh}, \]  

(5.188)

see Bishop, Fienberg, and Holland (1975, p. 442). Also, under the independence of X and Y,

\[ E(r_{xh} r_{yh} m_{xh} m_{yh}) = E(r_{xh} m_{xh}) E(r_{yh} m_{yh}) \]
\[ = E[(r_{xyh} + m_{xh})(m_{xyh} + m_{xyh})] \]
\[ \times E[(r_{xyh} + m_{xyh})(m_{xyh} + m_{xyh})] \]
\[ = n_h (n_h - 1) [(p_{rrh} + p_{mnh})(p_{mnh} + p_{mnh})] n_h (n_h - 1) \]
\[ \times [(p_{rrh} + p_{mnh})(p_{mnh} + p_{mnh})], \]  

(5.189)

see Bishop, Fienberg, and Holland (1975, p. 442). But, under the independence of X and Y,
\[ p_{rrh} = (p_{rrh} + p_{rhm})(p_{rrh} + p_{mrh}) \quad (5.190) \]

and

\[ p_{mmh} = (p_{mmh} + p_{mrh})(p_{mmh} + p_{rhm}), \quad (5.191) \]

so

\[ E(r_{xh} r_{yh} m_{xh} m_{yh}) = n_h^2 (n_h - 1)^2 p_{rrh} p_{mmh}. \quad (5.192) \]

It follows that

\[ E(r_{xyh} m_{xyh} \mid cs, R_2) = \frac{r_{xh} r_{yh} m_{xh} m_{yh}}{n_h (n_h - 1)}, \quad (5.193) \]

since \((r_{xh}, r_{yh}, m_{xh}, m_{yh})\) is sufficient for \((p_{rrh}, p_{mmh})\) under independence.

For (5.168) and (5.169), under the complete dependence of \(X\) and \(Y\), \(rm_{xyh}\) and \(mr_{xyh}\) are equal to zero, so

\[ E(r_{xyh} \mid cs, R_2) = r_{yh} \quad (5.194) \]

and
We present the Cov(Y -N/Z , X -Nïï ), where fi and Ji are defined in (6.196) below.

Theorem 5.9: Assume that we have a finite population of size N that is a sample from a superpopulation made up of H sub—populations, where within the sub—populations there are two characteristics, X and Y, and the observations (X,Y) are identically and independently distributed. Let (X,Y) have mean (μx,y, μx) and covariance matrix Σ in sub—population h, h=1,...,H. The matrix Σ is defined in (5.97). Assume that equal probability samples of sizes Nh taken from the H sub—populations make up the H imputation classes in the finite population. Assume that a probability sample is taken from the finite population, where the probability that unit hi is in the sample is greater than zero for all hi, h=1,...,H, i=1,...,Nh. Let ( W hi )⁻¹ be the probability that unit hi is chosen for the sample. Assume that the missing values of the Yhi's and the Xhi's are missing at random within the imputation classes and that the missingness mechanism for the Yhi's is possibly correlated with the missingness mechanism for the Xhi's. Assume that some form of random imputation within imputation classes from the observed sample has been done to fill in the missing values. Let

\[ E( r_{xyh} m_{xyh} | cs, R_2 ) = r_{yh} m_{yh} \]
\[
(\bar{\mu}_x, \bar{\mu}_y) = \frac{\sum_{h=1}^{H} N_h}{N} (\mu_{xh}, \mu_{yh}).
\] (5.196)

Let

\[
P_{xh} = P(\tau_{xhi,j} = 1 \mid Y_{hi} \text{ and } X_{hi} \text{ both respond, } X_{hj} \text{ is missing, cs, } R_1),
\] (5.197)

where \(i \neq j\). Let

\[
P_{yh} = P(\tau_{yhi,j} = 1 \mid Y_{hi} \text{ and } X_{hi} \text{ both respond, } Y_{hj} \text{ is missing, cs, } R_1),
\] (5.198)

where \(i \neq j\). Let

\[
P_{1h} = P(\tau_{yhi,j} \tau_{xhi,k} = 1 \mid Y_{hi} \text{ and } X_{hi} \text{ both respond, } Y_{hj} \text{ and } X_{hk} \text{ are missing, cs, } R_1),
\] (5.199)

where \(i \neq j\) and \(i \neq k\) and \(j \neq k\). Let

\[
P_{2h} = P(\tau_{yhi,j} \tau_{xhi,j} = 1 \mid Y_{hi} \text{ and } X_{hi} \text{ both respond, } Y_{hj} \text{ and } X_{hj} \text{ are missing, cs, } R_1),
\] (5.200)
where \( i \neq j \). Let \( \hat{Y} \) be as defined in expression (5.5) and let

\[
\hat{X} = \sum_{h=1}^{H} \sum_{i=1}^{n_h} W_{hi} X_{hi}. \tag{5.201}
\]

Then,

\[
\text{Cov}( \hat{Y}^* - N \bar{\mu}_y, \hat{X}^* - N \bar{\mu}_x) = \text{Cov}( \hat{Y} - N \bar{\mu}_y, \hat{X} - N \bar{\mu}_x)
\]

\[
+ \sum_{h=1}^{H} \sigma_{xyh} \sum_{i=1}^{n_h} \frac{E\{ r_{xyh} - n_h + r_{xyh} m_{xyh} p_{2h} \}}{n_h} \sum_{i=1}^{n_h} W_{hi}^2
\]

\[
+ \frac{r_{xyh}}{n_h} [\frac{m_{xh} p_{xh}}{n_h - 1} + \frac{m_{yh} p_{yh}}{n_h - 1} + \frac{m_{xh} m_{yh} - m_{xyh}}{n_h - 1} p_{1h}]
\]

\[
= \sum_{h=1}^{H} \left[ (\sum_{i=1}^{n_h} W_{hi})^2 - \sum_{i=1}^{n_h} W_{hi}^2 \right] , \tag{5.202}
\]

where we assume that \( E(r_{xyh}^{-1}) \) and \( E(n_{h}^{-1}) \) exists for all \( h \).

**Proof:** To show (5.202), we first find the \( \text{Cov}(\hat{Y} - N \bar{\mu}_y, \hat{X} - N \bar{\mu}_x) \) and then find the \( \text{Cov}(\hat{Y}^* - N \bar{\mu}_y, \hat{X}^* - N \bar{\mu}_x) \) with respect to the \( \text{Cov}(\hat{Y} - N \bar{\mu}_y, \hat{X} - N \bar{\mu}_x) \).

We use three levels of conditioning. At the lowest level (level 3), we hold the choice of the sample (cs) constant, we hold the response pattern constant (\( \Gamma \)), and we hold the choices for imputation and the choices for which donor goes with which missing
values constant (T). We take expectations over draws from the superpopulation. At the second level, we hold cs and $R_1$ constant, and take expectations over response patterns (given $R_1$), the choices for imputation, and the choices for which donors go with which missing values. At the top level (level 1), we take expectations with respect to everything left random at the second level. We never actually evaluate the expectations at the top level. First we look at the $\text{Cov}(\hat{Y} - N\hat{\mu}_y, \hat{X} - N\hat{\mu}_x)$. The

$$\text{Cov}(\hat{Y} - N\hat{\mu}_y, \hat{X} - N\hat{\mu}_x) = \text{Cov}_1\{E_2[ E_3(\hat{Y})], E_2[ E_3(\hat{X})]\}$$

$$+ E_1\{ \text{Cov}_2[ E_3(\hat{Y}), E_3(\hat{X})]\}$$

$$+ E_1\{ E_2[ \text{Cov}_3(\hat{Y}, \hat{X})]\}. \quad (5.203)$$

The

$$E_3(\hat{Y} | cs, \Gamma, T) = \sum_{h=1}^{H} \sum_{i=1}^{n_h} W_{hi} \mu_{yh}, \quad (5.204)$$

from expression (5.36) in Theorem 5.3. Likewise,

$$E_3(\hat{X} | cs, \Gamma, T) = \sum_{h=1}^{H} \sum_{i=1}^{n_h} W_{hi} \mu_{xh}. \quad (5.205)$$

The
\[
\text{Cov}_2( \sum_{h=1}^{n_h} \sum_{i=1}^{n_{hi}} W_{hi} \mu_{yh}, \sum_{h=1}^{n_h} \sum_{i=1}^{n_{hi}} W_{hi} \mu_{xh} | cs, R_1) = 0, \quad (5.206)
\]

and

\[
\text{E}_2[ \sum_{h=1}^{n_h} \sum_{i=1}^{n_{hi}} W_{hi} \mu_{yh} | cs, R_1] = \sum_{h=1}^{n_h} \sum_{i=1}^{n_{hi}} W_{hi} \mu_{yh}, \quad (5.207)
\]

and

\[
\text{E}_2[ \sum_{h=1}^{n_h} \sum_{i=1}^{n_{hi}} W_{hi} \mu_{xh} | cs, R_1] = \sum_{h=1}^{n_h} \sum_{i=1}^{n_{hi}} W_{hi} \mu_{xh}. \quad (5.208)
\]

The

\[
\text{E}_2[ \text{Cov}_3( \sum_{h=1}^{n_h} \sum_{i=1}^{n_{hi}} W_{hi} Y_{hi}, \sum_{h=1}^{n_h} \sum_{i=1}^{n_{hi}} W_{hi} X_{hi} | cs, R_1 | cs, R_1 )]
\]

\[
= \sum_{h=1}^{n_h} \sum_{i=1}^{n_{hi}} W_{hi}^2 \sigma_{xyh}, \quad (5.209)
\]

since the \((X,Y)\)'s are identically and independently distributed with covariance \(\sigma_{xyh}\) in imputation class \(h\) and independent between imputation classes. It follows that

\[
\text{Cov}( \hat{Y} - N \hat{\mu}_y, \hat{X} - N \hat{\mu}_x ) = \text{Cov}( \sum_{h=1}^{n_h} \sum_{i=1}^{n_{hi}} W_{hi} \mu_{yh}, \sum_{h=1}^{n_h} \sum_{i=1}^{n_{hi}} W_{hi} \mu_{xh} )
\]
We now find the $\text{Cov}(Y^* - N\hat{\mu}_y, X^* - N\hat{\mu}_x)$. The

\[
\text{Cov}(Y^* - N\hat{\mu}_y, X^* - N\hat{\mu}_x) = \text{Cov}
\{ E_2[ E_3(\hat{Y}^* )], E_2[ E_3(\hat{X}^* )] \}
\]

\[+ E_1\{ \text{Cov}_2[ E_3(\hat{Y}^* ), E_3(\hat{X}^* )] \}\]

\[+ E_1\{ E_2[ \text{Cov}_3(\hat{Y}^*, \hat{X}^* )] \}. \tag{5.211}\]

We know from Theorem 5.2, expression (5.27), that the

\[
E_3[ \hat{Y}^* | cs, \Gamma, T ] = E_3[ \hat{Y} | cs, \Gamma, T ] \tag{5.212}\]

and that the

\[
E_3[ \hat{X}^* | cs, \Gamma, T ] = E_3[ \hat{X} | cs, \Gamma, T ]. \tag{5.213}\]

It follows that the

\[
\text{Cov}(Y^* - N\hat{\mu}_y, X^* - N\hat{\mu}_x) = \text{Cov}(\sum_{h=1}^{H} \sum_{i=1}^{n_h} W_{hi} \mu_{yh}, \sum_{h=1}^{H} \sum_{i=1}^{n_h} W_{hi} \mu_{xh})
\]
\[ + E_1 \{ E_2[ \text{Cov}_3(\hat{Y}^*, \hat{X}^*)] \} . \] \hspace{1cm} (5.214) 

The

\[ \text{Cov}_3 \{ \frac{1}{H} \sum_{i=1}^{n_h} \gamma_{yhi} \left[ W_{hi} + \sum_{j=1}^{n_h} W_{hj} \tau_{yhi,j} \right] Y_{hi} , \] 

\[ \frac{1}{H} \sum_{i=1}^{n_h} \gamma_{xhi} \left[ W_{hi} + \sum_{j=1}^{n_h} W_{hj} \tau_{xhi,j} \right] X_{hi} | cs, \Gamma, T \} \]

\[ = \frac{1}{H} \sum_{i=1}^{n_h} \sigma_{xyh} \{ \gamma_{yhi} \left[ W_{hi} + \sum_{j=1}^{n_h} W_{hj} \tau_{yhi,j} \right] \} \]

\[ \times \{ \gamma_{xhi} \left[ W_{hi} + \sum_{j=1}^{n_h} W_{hj} \tau_{xhi,j} \right] \} \] \hspace{1cm} (5.215)

since the \((X,Y)\)'s in the sample are independently and identically distributed within the imputation classes and independent between the imputation classes. We use Lemma 5.1 to take the expectation of the \( \text{Cov}_3(\hat{Y}^*, \hat{X}^*) \) at level 2 conditioning.

The

\[ E_2[ \text{Cov}_3(\hat{Y}^*, \hat{X}^*) | cs, \Gamma, T ] | cs, R_1 ] = \frac{1}{H} \sum_{i=1}^{n_h} \sigma_{xyh} \left[ \frac{r_{xyh}}{n_h} W_{hi} \right] \]

\[ + \sum_{i=1}^{n_h} \sum_{j=1}^{n_h} W_{hi} W_{hj} \frac{r_{xyh}}{n_h} \frac{m_{xh}}{n_h-1} P_{xh} \]
\[
(5.216) \quad \{ \int_{\Omega} x_i \frac{1}{\eta} \eta d \eta \frac{\eta_x}{\eta_{xx}} \} + \frac{1}{\eta} \eta \frac{1}{\eta_{xx}} \frac{\frac{\eta_x}{\eta_{xx}}}{\eta_x} \frac{\eta_x}{\eta_{xx}} + \\
\left[ \int_{\Omega} \frac{1}{\eta} \eta \frac{1}{\eta_{xx}} \frac{1}{\eta_x} \eta_x \frac{1}{\eta_{xx}} \right] \eta_x \frac{1}{\eta} \frac{1}{\eta_x} \eta_x \frac{1}{\eta_{xx}} + \\
\left[ \int_{\Omega} \frac{1}{\eta} \eta \frac{1}{\eta_{xx}} \frac{1}{\eta_x} \eta_x \frac{1}{\eta_{xx}} \right] \eta_x \frac{1}{\eta} \frac{1}{\eta_x} \eta_x \frac{1}{\eta_{xx}} + \\
\int_{\Omega} \frac{1}{\eta} \eta \frac{1}{\eta_{xx}} \frac{1}{\eta_x} \eta_x \frac{1}{\eta_{xx}} \frac{1}{\eta_x} \eta_x \frac{1}{\eta_{xx}} = \\
\int_{\Omega} \frac{1}{\eta} \eta \frac{1}{\eta_{xx}} \frac{1}{\eta_x} \eta_x \frac{1}{\eta_{xx}} \frac{1}{\eta_x} \eta_x \frac{1}{\eta_{xx}} + \\
\int_{\Omega} \frac{1}{\eta} \eta \frac{1}{\eta_{xx}} \frac{1}{\eta_x} \eta_x \frac{1}{\eta_{xx}} \frac{1}{\eta_x} \eta_x \frac{1}{\eta_{xx}} + \\
\int_{\Omega} \frac{1}{\eta} \eta \frac{1}{\eta_{xx}} \frac{1}{\eta_x} \eta_x \frac{1}{\eta_{xx}} \frac{1}{\eta_x} \eta_x \frac{1}{\eta_{xx}}
\]
under the missing at random assumption.

Expression (5.202) follows from substituting the expectation of (5.216) into (5.214), since

\[ \text{Cov}(\hat{Y},\hat{X}) = \text{Cov}(\sum_{h=1}^{H} \sum_{i=1}^{n_{h}} W_{hi} \mu_{yh}, \sum_{h=1}^{H} \sum_{i=1}^{n_{h}} W_{hi} \mu_{xh}) \]

\[ + \mathbb{E}(\sum_{h=1}^{H} \sum_{i=1}^{n_{h}} W_{hi}^{2} \sigma_{xyh}) \quad (5.217) \]

Alternative Proof 1: Let the sample be put in order such that, within imputation class \( h \), units \( i=1,\ldots,r_{xyh} \) are the units for which \( X \) and \( Y \) both have responses, units \( i=r_{xyh}+1,\ldots,r_{yh} \) are the units for which there is a response for \( Y \) but not one for \( X \), units \( i=r_{yh}+1,\ldots,r_{xh}+r_{xyh} \) are the units for which there is a response for \( X \) but not one for \( Y \), and units \( i=r_{xh}+r_{xyh}+1 \) to \( n_{h} \) are the units for which there is no response on either \( X \) or \( Y \).

The

\[ \text{Cov}[\left(\hat{Y}^{*} - N\mu_{Y}\right),\left(\hat{X}^{*} - N\mu_{X}\right)] = \text{Cov}(\hat{Y}^{*} - \hat{Y},\hat{X}^{*} - \hat{X}) \]

\[ + \text{Cov}(\hat{Y}^{*} - \hat{Y},\hat{X} - N\mu_{X}) + \text{Cov}(\hat{Y} - N\mu_{Y},\hat{X}^{*} - \hat{X}) \]

\[ + \text{Cov}(\hat{Y} - N\mu_{Y},\hat{X} - N\mu_{X}) \]
\begin{align*}
= \text{Cov}(\hat{Y} - N\bar{y}, \hat{X} - N\bar{x}) + E[\text{Cov}(\hat{Y}^* - \hat{Y}, \hat{X}^* - \hat{X} | cs, R_1)] \\
+ E[\text{Cov}(\hat{Y}^* - \hat{Y}, \hat{X}^* - N\bar{x} | cs, R_1)] \\
+ E[\text{Cov}(\hat{Y} - N\bar{y}, \hat{X}^* - \hat{X} | cs, R_1)] \quad (5.218)
\end{align*}

since, under the superpopulation model,

\begin{align*}
E(\hat{Y}^* - \hat{Y} | cs, R_1) = E(\hat{X}^* - \hat{X} | cs, R_1) = 0. \quad (5.219)
\end{align*}

Now,

\begin{align*}
\hat{Y}^* - \hat{Y} &= H \sum_{h=1}^{r_{xyh} + rm_{xyh}} \sum_{i=r_{yh} + 1}^{Y_{hi}} W_{hi} (Y_{hi} - Y_{hi}) \\
&+ \sum_{i=r_{xyh} + rm_{xyh} + 1}^{n_h} W_{hi} (Y_{hi} - Y_{hi}) \quad (5.220)
\end{align*}

and

\begin{align*}
\hat{X}^* - \hat{X} &= H \sum_{h=1}^{r_{xyh}} \sum_{j=r_{xyh} + 1}^{x_{hj}} W_{hj} (X_{hj} - X_{hj}) \\
&+ \sum_{j=r_{xyh} + rm_{xyh} + 1}^{n_h} W_{hj} (X_{hj} - X_{hj}) \quad . \quad (5.221)
\end{align*}
If we consider the $W_{hi}$'s that fall in any of the response classes as random draws from the sample, then

$$E( W_{hi} \mid cs, R_1 ) = W_h^2 - n_h^{-1} S_{Wh}^2$$

(5.222)

for $i \neq j$, and

$$E( W_{hi}^2 \mid cs, R_1 ) = \left( 1 - \frac{1}{n_h} \right) S_{Wh}^2 + W_h^2 .$$

(5.223)

Also, the choice of the $W_{hi}$'s are independent of the $Y_{hi}$'s, the $X_{hi}$'s, the $Y_{hi}$'s and the $X_{hi}$'s. It follows that, for $i \neq j$,

$$\text{Cov}( W_{hi} ( Y_{hi} - Y_{hi} ) , W_{hj} ( X_{hj} - X_{hj} ) \mid cs, R_1 )$$

$$= E[ W_{hi} W_{hj} \mid cs, R_1 ] E[ ( Y_{hi} - Y_{hi} ) ( X_{hj} - X_{hj} ) \mid cs, R_1 ]$$

$$= [ W_{hi}^2 - n_h^{-1} S_{Wh}^2 ] \text{Cov}( Y_{hi} - Y_{hi} , X_{hj} - X_{hj} \mid cs, R_1 ) ,$$

(5.224)

since

$$E( Y_{hi}^* - Y_{hi} \mid cs, R_1 ) = E( X_{hi}^* - X_{hi} \mid cs, R_1 ) = 0 .$$

(5.225)

Similarly,
\[ \text{and} \]

\[(6.229)\]

\[ \langle \mathcal{I}_H, \mathcal{C}_S | \left( \frac{x_{N}^{*} - x_{X}^{*}}{x_{X}^{*}} \right)^{2} \lambda_{\mathcal{M}}', \left( \frac{y_{N}^{*} - y_{X}^{*}}{y_{X}^{*}} \right)^{2} \lambda_{\mathcal{M}}' \rangle_{\mathcal{O}} = \]

\[ \langle \mathcal{I}_H, \mathcal{C}_S | \left( \frac{x_{N}^{*} - x_{X}^{*}}{x_{X}^{*}} \right)^{2} \lambda_{\mathcal{M}}', \left( \frac{y_{N}^{*} - y_{X}^{*}}{y_{X}^{*}} \right)^{2} \lambda_{\mathcal{M}}' \rangle_{\mathcal{O}} \]

\[(6.228)\]

\[ \langle \mathcal{I}_H, \mathcal{C}_S | \left( \frac{x_{N}^{*} - x_{X}^{*}}{x_{X}^{*}} \right)^{2} \lambda_{\mathcal{M}}', \left( \frac{y_{N}^{*} - y_{X}^{*}}{y_{X}^{*}} \right)^{2} \lambda_{\mathcal{M}}' \rangle_{\mathcal{O}} = \]

\[ \langle \mathcal{I}_H, \mathcal{C}_S | \left( \frac{x_{N}^{*} - x_{X}^{*}}{x_{X}^{*}} \right)^{2} \lambda_{\mathcal{M}}', \left( \frac{y_{N}^{*} - y_{X}^{*}}{y_{X}^{*}} \right)^{2} \lambda_{\mathcal{M}}' \rangle_{\mathcal{O}} \]

\[(6.227)\]

\[ \langle \mathcal{I}_H, \mathcal{C}_S | \left( \frac{x_{N}^{*} - x_{X}^{*}}{x_{X}^{*}} \right)^{2} \lambda_{\mathcal{M}}', \left( \frac{y_{N}^{*} - y_{X}^{*}}{y_{X}^{*}} \right)^{2} \lambda_{\mathcal{M}}' \rangle_{\mathcal{O}} = \]

\[ \langle \mathcal{I}_H, \mathcal{C}_S | \left( \frac{x_{N}^{*} - x_{X}^{*}}{x_{X}^{*}} \right)^{2} \lambda_{\mathcal{M}}', \left( \frac{y_{N}^{*} - y_{X}^{*}}{y_{X}^{*}} \right)^{2} \lambda_{\mathcal{M}}' \rangle_{\mathcal{O}} \]

\[(6.226)\]

\[ \langle \mathcal{I}_H, \mathcal{C}_S | \left( \frac{x_{N}^{*} - x_{X}^{*}}{x_{X}^{*}} \right)^{2} \lambda_{\mathcal{M}}', \left( \frac{y_{N}^{*} - y_{X}^{*}}{y_{X}^{*}} \right)^{2} \lambda_{\mathcal{M}}' \rangle_{\mathcal{O}} = \]

\[ \langle \mathcal{I}_H, \mathcal{C}_S | \left( \frac{x_{N}^{*} - x_{X}^{*}}{x_{X}^{*}} \right)^{2} \lambda_{\mathcal{M}}', \left( \frac{y_{N}^{*} - y_{X}^{*}}{y_{X}^{*}} \right)^{2} \lambda_{\mathcal{M}}' \rangle_{\mathcal{O}} \]
\[
= \left[ 1 - \frac{1}{n_h} \right] S_{Wh}^2 + W_h^2 \right] \text{Cov} \left( Y_{hi} - N\mu_y, X_{hi}^* - X_{hi} \mid cs, R_1 \right).
\]

(5.230)

Using expressions (5.220) and (5.221), the \( \text{Cov}(\hat{Y}^* - \hat{Y}, \hat{X}^* - \hat{X} \mid cs, R_1) + \text{Cov}(\hat{Y}^* - \hat{Y}, \hat{X} - N\mu_x \mid cs, R_1) + \text{Cov}(\hat{Y} - \hat{Y}_y, \hat{X}^* - \hat{X} \mid cs, R_1) \) can be written as the sum of twenty-five terms, five for the \( \text{Cov}(\hat{Y}^* - \hat{Y}, \hat{X}^* - \hat{X} \mid cs, R_1) \), ten for the \( \text{Cov}(\hat{Y}^* - \hat{Y}, \hat{X} - N\mu_x \mid cs, R_1) \), and ten for \( \text{Cov}(\hat{Y} - \hat{Y}_y, \hat{X}^* - \hat{X} \mid cs, R_1) \). Canceling like terms of opposite sign, setting covariance terms where the two expressions are independent to zero, and taking the expected value of terms in \( \text{Cov}(Y_{hi}, X_{hi} \mid cs, R_1) \), we get

\[
\text{Cov}(\hat{Y}^* - \hat{Y}, \hat{X}^* - \hat{X} \mid cs, R_1) + \text{Cov}(\hat{Y}^* - \hat{Y}, X - N\mu_x \mid cs, R_1)
\]

\[
+ \text{Cov}(\hat{Y} - \hat{Y}_y, \hat{X}^* - \hat{X} \mid cs, R_1)
\]

\[
= \sum_{h=1}^{H} \left\{ \left[ W_h^2 - n_h^{-1} S_{Wh}^2 \right] \text{Cov} \left( \sum_{i=r_{yh}+1}^{r_{xh}+r_{xyh}} Y_{hi}, \sum_{j=r_{xyh}+1}^{r_{yh}+1} X_{hi} \mid cs, R_1 \right) \right.
\]

\[
+ \text{Cov} \left( \sum_{i=r_{yh}+1}^{r_{xh}+r_{xyh}} Y_{hi}, \sum_{j=r_{xh}+r_{xyh}+1}^{r_{yh}} X_{hi} \mid cs, R_1 \right) \]

\[
+ \text{Cov} \left( \sum_{i=r_{xh}+r_{xyh}+1}^{n_h} Y_{hi}, \sum_{j=r_{xyh}+1}^{r_{yh}} X_{hi} \mid cs, R_1 \right) \]

\[
+ \text{Cov} \left( \sum_{i=r_{xh}+r_{xyh}+1}^{n_h} Y_{hi}, \sum_{j=r_{xyh}+1}^{r_{yh}+1} X_{hi} \mid cs, R_1 \right) \]

\[
+ \text{Cov} \left( \sum_{i=r_{xyh}+1}^{n_h} Y_{hi}^*, \sum_{j=r_{xyh}+1}^{n_h} X_{hi}^* \mid cs, R_1 \right)
\]

\[
+ \text{Cov} \left( \sum_{i=r_{xyh}+1}^{r_{xyh}} Y_{hi}, \sum_{j=1}^{r_{xyh}} X_{hi} \mid cs, R_1 \right)
\]

\[
+ \text{Cov} \left( \sum_{i=1}^{r_{xyh}} Y_{hi}, \sum_{j=r_{xyh}+1}^{r_{xyh}+1} X_{hi}^* \mid cs, R_1 \right)
\]

\[
+ \text{Cov} \left( \sum_{i=1}^{r_{xyh}} Y_{hi}, \sum_{j=1}^{r_{xyh}+1} X_{hi}^* \mid cs, R_1 \right)
\]

\[
+ \text{Cov} \left( \sum_{i=1}^{r_{xyh}} Y_{hi}, \sum_{j=r_{xyh}+1}^{r_{xyh}+1} X_{hi}^* \mid cs, R_1 \right)
\]

\[
\left( 1 - \frac{1}{n_h} \right) S_{Wh}^2 + W_h^2
\]

\[
\times \left[ \sum_{i=r_{xyh}+1}^{n_h} \text{Cov} \left( Y_{hi}^*, X_{hi}^* \mid cs, R_1 \right) \right] \]

\[
+ (r_{xyh} - n_h) \sigma_{xyh} \right) \} .
\]

Now, using \( S_{1h}, S_{2h}, S_{3h}, \) and \( S_{4h} \) as defined in Section 5.1.10 and shown in Table 5.3,
\[ \text{Cov}(Y_j^*, X_j^*) = \sigma_{xyh} P(\ Y_{hi}^* \text{ and } X_{hj}^* \text{ both come from the same element of } S_{4h} \) | \ Y_{hi} \in S_{3h} \text{ and } X_{hj} \in S_{2h}, cs, R_1 \) \times (\text{the number of possible matches}) \]

\[ \times P(\ Y_{hi}^* \in S_{4h} \| Y_{hi} \in S_{3h}, cs, R_1 ) \]

\[ \times P(\ X_{hj}^* \in S_{4h} \| Y_{hi}^*, X_{hj} \in S_{2h}, cs, R_1 ) \]

\[ \times [\text{the same technique to find the covariances for the rest of the terms in (5.231), we get}] \]

\[ \text{Cov}(\hat{Y}^* - N\bar{\mu}_y, \hat{X}^* - N\bar{\mu}_x) = \text{Cov}(\hat{Y} - N\bar{\mu}_y, \hat{X} - N\bar{\mu}_x) \]

\[ + \sum_{h=1}^{H} \sigma_{xyh} E\left[ \frac{W_h^2}{n_h} - \frac{1}{n_h} S_{Wh}^2 \right] \]

\[ \times [\text{the same technique to find the covariances for the rest of the terms in (5.231), we get}] \]
\[ Y_{hi} \in S_{3h}, X_{hj} \in S_{2h}, cs, R_1 \) m_{xyh} \] 

+ \( P( Y_{hi}^* \text{ and } X_{hj}^* \text{ both come from the same element in } S_{4h} \) 

\[ Y_{hi} \in S_{3h}, X_{hj} \in S_{1h}, cs, R_1 ) m_{xyh} \] 

+ \( P( Y_{hi}^* \text{ and } X_{hj}^* \text{ both come from the same element in } S_{4h} \) 

\[ Y_{hi} \in S_{1h}, X_{hj} \in S_{2h}, cs, R_1 ) m_{xyh} \] 

+ \( P( Y_{hi}^* \text{ and } X_{hj}^* \text{ both come from the same element in } S_{4h} \) 

\[ Y_{hi} \in S_{1h}, X_{hj} \in S_{1h, i\neq j, cs, R_1 ) m_{xyh} (m_{xyh} - 1) \] 

+ \( P( Y_{hi}^* \text{ comes from } S_{4h} \mid Y_{hi} \in S_{3h}, cs, R_1 ) m_{xyh} \) 

+ \( P( Y_{hi}^* \text{ comes from } S_{4h} \mid Y_{hi} \in S_{1h}, cs, R_1 ) m_{xyh} \) 

+ \( P( X_{hj}^* \text{ comes from } S_{4h} \mid X_{hj} \in S_{2h}, cs, R_1 ) m_{xyh} \) 

+ \( P( X_{hj}^* \text{ comes from } S_{4h} \mid X_{hj} \in S_{1h}, cs, R_1 ) m_{xyh} \) 

+ \[ [(1 - \frac{1}{n_h}) S_{W_1}^2 + W_1^2] \]
We now show that expression (5.233) is equivalent to expression (5.202).

First,

\[
W_h^2 - n_h^{-1} S_{Wh}^2
\]

\[
= n_h^{-2} \left( \sum_{i=1}^{n_h} W_{hi} \right)^2 - n_h^{-1} (n_h - 1)^{-1} \sum_{i=1}^{n_h} W_{hi}^2 + n_h^{-2} (n_h - 1) \left( \sum_{i=1}^{n_h} W_{hi} \right)^2
\]

\[
= n_h^{-1} (n_h - 1)^{-1} \left[ \left( \sum_{i=1}^{n_h} W_{hi} \right)^2 - \sum_{i=1}^{n_h} W_{hi}^2 \right]
\]

(5.234)

and

\[
(1 - \frac{1}{n_h}) S_{Wh}^2 + W_h^2
\]

\[
= (n_h - 1) n_h^{-1} (n_h - 1)^{-1} \left[ \sum_{i=1}^{n_h} W_{hi}^2 - n_h W_h^2 \right] + W_h^2
\]
\[ P_{1h} = P( Y_{hk} \text{ donates to } Y_{hi} \text{ and } X_{hk} \text{ donates to } X_{hj} \]

\[ | Y_{hk} \text{ and } X_{hk} \text{ both respond, } Y_{hi} \text{ and } X_{hj} \text{ are missing, } i \neq j, \text{ cs, } R_1 ) \]

\[ = P[ Y_{hi}^* \text{ and } X_{hj}^* \text{ come from a given element of } S_{4h} \]

\[ | Y_{hi} \in S_{3h} \cup S_{1h}, X_{hj} \in S_{2h} \cup S_{1h}, i \neq j, \text{ cs, } R_1 ] \]

\[ = r_{xyh}^{-1} P[ Y_{hi}^* \text{ and } X_{hj}^* \text{ come from the same element of } S_{4h} \]

\[ | Y_{hi} \in S_{3h} \cup S_{1h}, X_{hj} \in S_{2h} \cup S_{1h}, i \neq j, \text{ cs, } R_1 ] \]

(5.236)

Let \( A = \{ Y_{hi}^* \text{ and } X_{hj}^* \text{ come from the same element of } S_{4h} \} \). Let \( B = \{ Y_{hi} \in S_{3h}, X_{hj} \in S_{2h} \} \). Let \( C = \{ Y_{hi} \in S_{3h}, X_{hj} \in S_{1h} \} \). Let \( D = \{ Y_{hi} \in S_{1h}, X_{hj} \in S_{2h} \} \). Let \( E = \{ Y_{hi} \in S_{1h}, X_{hj} \in S_{1h}, i \neq j \} \). Then

\[ P_{1h} = r_{xyh}^{-1} \left\{ P[ B \cup C \cup D \cup E | \text{ cs, } R_1 ] \right\}^{-1} [ P( A | B, \text{ cs, } R_1 ) P( B | \text{ cs, } R_1 ) \]

\[ + P( A | C, \text{ cs, } R_1 ) P( C | \text{ cs, } R_1 ) + P( A | D, \text{ cs, } R_1 ) P( D | \text{ cs, } R_1 ) \]
since B, C, D, and E are disjoint. The

\[ P( B \cup C \cup D \cup E \mid cs, R_1) = \frac{m_{xyh} \cdot m_{xyh} - m_{xyh}}{n_h \left( \frac{n_h}{n_h - 1} \right)} . \] (5.238)

The

\[ P( B \mid cs, R_1) = \frac{m_{xyh} \cdot m_{xyh}}{n_h \left( \frac{n_h}{n_h - 1} \right)} . \] (5.239)

The

\[ P( C \mid cs, R_1) = \frac{m_{xyh} \cdot m_{xyh}}{n_h \left( \frac{n_h}{n_h - 1} \right)} . \] (5.240)

The

\[ P( D \mid cs, R_1) = \frac{m_{xyh} \cdot m_{xyh}}{n_h \left( \frac{n_h}{n_h - 1} \right)} . \] (5.241)

The

\[ P( E \mid cs, R_1) = \frac{m_{xyh} \cdot \left( \frac{m_{xyh}}{n_h} - 1 \right)}{n_h \left( \frac{n_h}{n_h - 1} \right)} . \] (5.242)

The first four terms of expression (5.233) follow from substituting (5.238) through
(5.242) into (5.237) and then substituting (5.237) into (5.202). The other terms in (5.233) can be found by the same method.

**Alternative Proof 2:** We use two sets of indices on the augmented sample in this proof. The first set of indices is the set given in the first alternative proof of this theorem. The second set of indices is as follows. Let $b_{hi}$ be the number of times respondent $Y_{hi}$ by the first index set appears in the augmented sample. Let $c_{hi}$ be the number times respondent $X_{hi}$ by the first index set appears with respondent $Y_{hi}$ by the first index set. Let $d_{hi}$ be the number of times respondent $Y_{hi}$ by the first index set appears with respondents $X_{hj}$ by the first index set, where $j 
eq i$. Then the first $r_{xyh}$ units are the first $r_{xyh}$ respondent $Y_{hi}$'s by the first index set each repeated $b_{hi}(c_{hi}+d_{hi})$ times first with the $X_{hi}$'s that appear with the $Y_{hi}$'s then with the $X_{hj}$'s that appear with the $Y_{hi}$'s, where $j 
eq i$. The second set of units is the $r_{xyh}+1$ to $r_{xyh}$ respondent $Y_{hi}$'s by the first index set each repeated $b_{hi}d_{hi}$ times with the $X_{hj}$'s that appear with the $Y_{hi}$'s, where $j 
eq i$. Let

$$p_{hi} = \sum_{j=1}^{i-1} b_{hj} (c_{hj} + d_{hj}) + 1$$

if $hi < r_{xyh}$

$$= \sum_{k=1}^{r_{xyh}} b_{hk} (c_{hk} + d_{hk}) + \sum_{j=r_{xyh}+1}^{i-1} b_{hj} d_{hj} + 1$$

(5.243)
Let \( p_{hi} = p_i - 1 + j \). \( (5.244) \)

Let \( p_{hikw1} \) be one of the indices on the units associated with respondent \( X_{hi} \) by the first index set where respondent \( X_{hi} \) appears with respondent \( Y_{hi} \) by the first index set. Let \( p_{hiX} \) be any of the indices on respondents \( X_{hj} \) by the first data set where the \( X_{hj} \)'s appear with respondent \( Y_{hi} \) by the first data set and where \( i \neq j \). Let \( p_{hikw2} \) be the indices on the weights associated with respondent \( X_{hphix} \) by the second data set.

Let \( \hat{Z} \) be an estimator of the population total for \( Y \). Let \( \hat{T} \) be an estimator of the population total for \( X \). Then

\[
\text{Cov}(\hat{Z} - N\bar{Y}, \hat{T} - N\bar{X}) = \text{Cov}\{E\{E(\hat{Z} | cs, rp, ic, dm) | cs, R_1\}, E\{E(\hat{T} | cs, rp, ic, dm) | cs, R_1\}\}
+ E\{ \text{Cov}[E(\hat{Z} | cs, rp, ic, dm), E(\hat{T} | cs, rp, ic, dm) | cs, R_1] \}
+ E\{ E\{ \text{Cov}(\hat{Z}, \hat{T} | cs, rp, ic, dm) | cs, R_1 \} \}. \tag{5.245} \]
Since the \((X,Y)\)'s are identically and independently distributed within the
imputation classes and independent between the imputation classes, for \(\hat{Z}=\hat{Y}\) and
\(\hat{T}=\hat{X}\),

\[
E(\hat{Z} \mid cs, rp, ic, dm) = \sum_{h=1}^{H} \sum_{i=1}^{n_h} W_{hi} \mu_{yh},
\]

\[
E(\hat{T} \mid cs, rp, ic, dm) = \sum_{h=1}^{H} \sum_{i=1}^{n_h} W_{hi} \mu_{xh},
\]

and

\[
\text{Cov}(\hat{Z}, \hat{T} \mid cs, rp, ic, dm) = \sum_{h=1}^{H} \sum_{i=1}^{n_h} W_{hi}^2 \sigma_{xyh}.
\]

Since the \((X^*,Y^*)\)'s are identically distributed within the imputation classes, for
\(\hat{Z}=\hat{Y}^*\) and \(\hat{T}=\hat{X}^*\),

\[
E(\hat{Z} \mid cs, rp, ic, dm) = \sum_{h=1}^{H} \sum_{i=1}^{n_h} W_{hi} \mu_{yh}
\]

and
\[
E( \hat{\mathcal{T}} \mid cs, rp, ic, dm) = \sum_{h=1}^{H} \sum_{i=1}^{n_h} W_{hi} \mu_{xh} \quad (5.250)
\]

The

\[
\text{Cov}(\hat{Y}^*, \hat{X}^* \mid cs, rp, ic, dm) = \sum_{h=1}^{H} \sum_{i=1}^{r_{xyh}} b_{hi} c_{hi} - \sum_{i=1}^{r_{xyh}} b_{hi} d_{hi} + \sum_{i=1}^{r_{xyh}} b_{hi} d_{hi}
\]

\[
= \sum_{h=1}^{H} \sum_{i=1}^{r_{xyh}} b_{hi} c_{hi} \quad (5.251)
\]

since

\[
\text{Cov}(Y_{hp_{hi}}, X_{hp_{hi}} \mid cs, rp, ic, dm) = 0 \quad (5.252)
\]

for all \(p_{hiX}\) and \(p_{hi}\) and

\[
\text{Cov}(Y_{hp_{hi}}, X_{hp_{hi}} \mid cs, rp, ic, dm) = \sigma_{xyh} \quad (5.253)
\]

for \(hi < r_{xyh}\).
Returning to the first index set

\[
E\{ \sum_{i=1}^{r_{xyh}} \sum_{j=1}^{b_{hi}} \sum_{c_{hi}} E[ \sum_{j=1}^{r_{xyh}} \sum_{i=1}^{h_{hpij}} W_{hp_{hikwl}} | cs, rp ] | cs, R_1 \}
\]

\[
=E\{ \sum_{i=1}^{r_{xyh}} W_{hi}^2 \}
\]

\[
+ \sum_{i=1}^{r_{xyh}} \sum_{j=r_{xyh}+1}^{r_{yh}} W_{hi} W_{hj} P_{xh} + \sum_{i=1}^{r_{xyh}} \sum_{j=r_{xyh}+r_m_{xyh}+1}^{n_{h}} W_{hi} W_{hj} P_{xh}
\]

\[
+ \sum_{i=r_{yh}+1}^{r_{xyh}+r_m_{xyh}} \sum_{j=1}^{r_{xyh}+1} W_{hi} W_{hj} P_{1h} r_{xyh}
\]

\[
+ \sum_{i=r_{yh}+1}^{r_{xyh}+r_m_{xyh}} \sum_{j=r_{xh}+r_m_{xyh}+1}^{n_{h}} W_{hi} W_{hj} P_{1h} r_{xyh}
\]

\[
+ \sum_{i=r_{xh}+r_m_{xyh}+1}^{r_{xyh}} \sum_{j=r_{xyh}+1}^{r_{xyh}+1} W_{hi} W_{hj} P_{1h} r_{xyh}
\]

\[
+ \sum_{i=r_{xh}+r_m_{xyh}+1}^{r_{xyh}+r_m_{xyh}} \sum_{j=r_{xyh}+1}^{r_{xyh}+1} W_{hi} W_{hj} P_{1h} r_{xyh}
\]

\[
+ \sum_{i=r_{xh}+r_m_{xyh}+1}^{r_{xyh}+r_m_{xyh}} \sum_{j=r_{xh}+r_m_{xyh}+1}^{n_{h}} W_{hi} W_{hj} P_{1h} r_{xyh}
\]

\[
+ \sum_{i=r_{xh}+r_m_{xyh}+1}^{r_{xyh}+r_m_{xyh}+1} \sum_{j=r_{xh}+r_m_{xyh}+1}^{n_{h}} W_{hi} W_{hj} P_{1h} r_{xyh}
\]

\[
+ \sum_{i=r_{xh}+r_m_{xyh}+1}^{r_{xyh}+r_m_{xyh}+1} \sum_{j=r_{xh}+r_m_{xyh}+1}^{n_{h}} W_{hi} W_{hj} P_{1h} r_{xyh}
\]

\[
i \neq j
\]
\[ n_h \]
\[ + \sum_{i=1}^{n_h} W_{hi}^2 \left[ \frac{r_{xyh}}{n_h} + \frac{r_{xyh} m_{xyh}}{n_h} \right] P_{2h} \]
\[ + [\left( \sum_{i=1}^{n_h} W_{hi} \right)^2 - \sum_{i=1}^{n_h} W_{hi}^2 ] n_h^{-1} (n_h - 1)^{-1} r_{xyh} \]
\[ \times \{ m_{xh} P_{xh} + m_{xy} P_{xy} \}
\[ + [m_{r_{xyh}} m_{xyh} + m_{r_{xyh}} m_{xyh} + m_{r_{xyh}} m_{xyh} \left( m_{xyh}^2 - m_{xyh} \right)] P_{1h} \} \]
\[ = \sum_{i=1}^{n_h} W_{hi}^2 \left[ -\frac{r_{xyh}}{n_h} + \frac{r_{xyh} m_{xyh}}{n_h} \right] P_{2h} \]
\[ + [\left( \sum_{i=1}^{n_h} W_{hi} \right)^2 - \sum_{i=1}^{n_h} W_{hi}^2 ] n_h^{-1} (n_h - 1)^{-1} r_{xyh} \]
\[ \times \{ m_{xh} P_{xh} + m_{yh} P_{yh} + (m_{xh} m_{yh} - m_{xyh}) \} P_{1h} \}. \quad (5.254) \]

Result (5.202) follows from substituting the expectation of expression (5.254) into the expectation of expression (5.251) and substituting expressions (5.246) through (5.250) and the expectation of expression (5.254) into expression (5.220) for \( \hat{Y}, \hat{X}, \hat{Y}^*, \) and \( \hat{X}^* \) respectively.
Table 5.3. The four sets that make up the observations in imputation class $h$: 1 = response, 0 = nonresponse

<table>
<thead>
<tr>
<th>Y</th>
<th>X</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>$S_{1h}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td># of elements = $m_{xyh}$</td>
</tr>
<tr>
<td>.</td>
<td>.</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>$S_{2h}$</td>
</tr>
<tr>
<td>.</td>
<td>.</td>
<td># of elements = $r_{xyh}$</td>
</tr>
<tr>
<td>.</td>
<td>.</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>$S_{3h}$</td>
</tr>
<tr>
<td>.</td>
<td>.</td>
<td># of elements = $m_{xyh}$</td>
</tr>
<tr>
<td>.</td>
<td>.</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>$S_{4h}$</td>
</tr>
<tr>
<td>.</td>
<td>.</td>
<td># of elements = $r_{xyh}$</td>
</tr>
<tr>
<td>.</td>
<td>.</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

At this point, we present a justification of why expression (5.202) is true.

Let us use the sets $S_{1h}$, $S_{2h}$, $S_{3h}$, and $S_{4h}$ of Section 5.1.10. The sets are described in Table 5.3. Note that

$$r_{xyh} + m_{xyh} = m_{yh} \quad (5.255)$$

and
\[ m_{xyh} + m_{xyh} = m_{xh} \quad (5.256) \]

The \( E_2[Cov_3(Y^*, X^* | cs, R_1, \Gamma, T) | cs, R_1] \) is the sum over the imputation classes of the sum of the expectations, given \( cs \) and \( R_1 \), of the diagonal elements in the cross product of \( Y \) and \( X \) in imputation class \( h \) divided by \( n_h \) plus the sum of the expectations, given \( cs \) and \( R_1 \), of the off diagonal elements in the cross product of \( Y \) and \( X \) in imputation class \( h \) divided by \( n_h \times n_{h-1} \). For the set \( S_{4h} \), it is clear that the diagonal elements contribute \( m_{xyh} \) to the sum, while the off diagonal elements contribute 0 to the sum. For the sets \( S_{3h} \) and \( S_{2h} \), it is clear that the diagonal elements and the off diagonal elements contribute 0 to the sum. For the \( Y \) elements in \( S_{2h} \), it is clear that the cross products with the \( X \) elements in \( S_{1h}, S_{3h}, \) and \( S_{4h} \) contribute 0 to the sum. For the \( X \) elements is \( S_{3h} \), it is clear that the cross product with the \( Y \) elements in \( S_{1h} \) and \( S_{4h} \) contribute 0 to the sum. This leaves the cross product of the \( Y \) elements in \( S_{4h} \) with the \( X \) elements in \( S_{1h} \cup S_{2h} \), the cross product of the \( X \) elements in \( S_{4h} \) with the \( Y \) elements in \( S_{1h} \cup S_{3h} \), and the cross product of the \( Y \) elements in \( S_{1h} \cup S_{3h} \) with the \( X \) elements in \( S_{1h} \cup S_{2h} \). There are \( r_{xyh} m_{xh} \) combinations in the cross product of the \( Y \) elements in \( S_{4h} \) with the \( X \) elements in \( S_{1h} \cup S_{2h} \). Any of the combinations contributes \( \sigma_{xyh} \) to the sum if the \( X \) element associated with the \( Y \) element in \( S_{4h} \) donates to the \( X \) element in \( S_{1h} \cup S_{2h} \) and contributes 0 to the sum otherwise, so the contribution of the cross product is \( r_{xyh} m_{xh} \sigma_{xyh} \times \) times the probability, given \( cs \) and \( R_1 \), that a given \( X \) element in \( S_{4h} \) donates to a given \( X \) element in \( S_{1h} \cup S_{2h} \). This probability is \( P_{xh} \). Similarly, the cross product of the \( X \) elements in \( S_{4h} \) with the \( Y \) elements in \( S_{1h} \cup S_{3h} \) contributes \( r_{xyh} m_{yh} \sigma_{xyh} P_{yh} \) to the sum. For the cross product of the \( Y \) elements in \( S_{1h} \cup S_{3h} \)
with the \( X \) elements in \( S_{1h} \cup S_{2h} \), we first look at the diagonal elements in \( S_{1h} \).

Each one contributes \( \sigma_{xyh} \) to the covariance only if the \( Y \) element and the \( X \) element both come from the same observation in \( S_{4h} \); otherwise the contribution is 0. There are \( r_{xyh} \) possibilities for any given diagonal element in \( S_{1h} \) to contribute \( \sigma_{xyh} \) instead of 0, so the contribution of the diagonal elements in \( S_{4h} \) is

\[
r_{xyh} m_{xyh} \sigma_{xyh}
\]

the probability that a given observation in \( S_{4h} \) donates both the \( Y \) element and the \( X \) element to a given element in \( S_{1h} \). This probability is \( P_{2h} \). Similarly, the contribution of the off diagonal terms of the cross product of the \( Y \) elements in \( S_{1h} \cup S_{3h} \) with the \( X \) elements in \( S_{1h} \cup S_{2h} \) is

\[
r_{xyh} (m_{xh} m_{yh} - m_{xyh}) \sigma_{xyh} P_{1h}.
\]

These contributions give us the result in (5.69).

We present four corollaries for less general conditions.

**Corollary 1:** Let the assumptions of Theorem 5.9 hold. Let the imputation on the \( Y \)'s be independent of the imputation of the \( X \)'s (as described in method one of Section 5.1.10). Then

\[
\text{Cov}\left( \hat{Y}^* - N \mu_y, \hat{X}^* - N \mu_x \right) = \text{Cov}\left( \hat{Y} - N \mu_y, \hat{X} - N \mu_x \right)
\]

\[
+ \sum_{h=1}^{H} \sigma_{xyh} E\left\{ \frac{r_{xyh} m_{xyh}}{\sum_{i=1}^{n_h} W_{hi}^2} \right\} + \frac{r_{xyh}}{n_h} \left[ \frac{m_{xh} + m_{yh} - m_{xyh}}{n_h} \right]
\]
Proof: By Lemma 5.1, under the independence of the imputation,

\[
P_{xh} = \frac{1}{r_{xh}},
\]

(5.258)

\[
P_{yh} = \frac{1}{r_{yh}},
\]

(5.259)

\[
P_{1h} = \frac{1}{r_{xh} r_{yh}},
\]

(5.260)

and

\[
P_{2h} = \frac{1}{r_{xh} r_{yh}}.
\]

(5.261)

By substituting (5.258), (5.259), (5.260), and (5.261) into (5.202), we get (5.257).

#

Corollary 2: Let the assumptions of Theorem 5.9 hold. Let the X's and the Y's be missing independently. Let the imputation of the X's and the Y's be done independently (as described in method one of Section 5.1.10). Then

\[
\text{Cov}(\hat{Y}^* - N \bar{\mu}_y, \hat{X}^* - N \bar{\mu}_x) = \text{Cov}(\hat{Y} - N \bar{\mu}_y, \hat{X} - N \bar{\mu}_x)
\]
Proof: By Lemma 5.1, under the independence of the missingness mechanism for the X's and Y's,

\[
E( r_{xyh} \mid cs, R_2 ) = \frac{r_{xh} r_{yh}}{n_h} \quad (5.263)
\]

and

\[
E( r_{xyh} m_{xyh} \mid cs, R_2 ) = \frac{r_{xh} r_{yh} m_{xh} m_{yh}}{n_h (n_h - 1)} . \quad (5.264)
\]

Under the independence of the imputation, \( P_{xh}, P_{yh}, P_{1h}, \) and \( P_{2h} \) are given by (5.258) to (5.261). By substituting (5.263), (5.264), (5.258), (5.259), (5.260), and (5.261) into (5.202), we get (5.262).

\#

Corollary 3: Let the assumptions of Theorem 5.9 hold. Let the X's be missing only when the Y's are missing, and vice versa. Let the X's be imputed simultaneously with the Y's (as described in method two of Section 5.1.10). Then
\[
\text{Cov}(\hat{Y}^* - N\bar{\mu}_y, \hat{X}^* - N\bar{\mu}_x) = \text{Cov}(\hat{Y} - N\bar{\mu}_y, \hat{X} - N\bar{\mu}_x)
\]

\[
+ \sum_{h=1}^{H} \sigma_{xyh} E\{ \frac{k_{yh} n_h + (k_{yh} + 2) t_{yh}}{n_h (n_h - 1)} \}
\]

\[
\sum_{i=1}^{n_h} \left[ \left( \sum_{i=1}^{W_{hi}} \right)^2 - \sum_{i=1}^{W_{hi}} \right] . \tag{5.265}
\]

**Proof:** By Lemma 5.1, under the dependence of the X's and Y's,

\[
E(r_{xyh} | cs, R_2) = r_{yh} , \tag{5.266}
\]

and

\[
E(r_{xyh} m_{xhy} | cs, R_2) = r_{yh} m_{yh} . \tag{5.267}
\]

Under the dependence of the imputation,

\[
P_{xh} = \frac{1}{r_{yh}} , \tag{5.268}
\]

\[
P_{yh} = \frac{1}{r_{yh}} , \tag{5.269}
\]

\[
P_{ih} = \frac{k_{yh} n_h + k_{yh} (t_{yh} - 2r_{yh})}{r_{yh} m_{yh} (m_{yh} - 1)} , \tag{5.270}
\]

and
Expression (5.265) follows from substituting (5.266) through (5.271) into (5.202).

Corollary 4: Let the assumptions of Theorem 5.9 hold. Let the imputation be done by method three of Section 5.1.10, that is, where X and Y are both missing, the X and Y are imputed simultaneously from the set of observations that have responses for both X and Y, and when either just the X or the Y is missing, the value is imputed from the total set of respondents for that variable.

Then

\[
\text{Cov}(Y^*-N\mu_y, X^*-N\mu_x) = \text{Cov}(\hat{Y}-N\mu_y, \hat{X}-N\mu_x) = \\
+ \sum_{h=1}^{H} \sigma_{xyh} \left( \frac{r_{xyh} - n_h + m_{xyh}}{n_h} \right) \sum_{i=1}^{n_h} W_{hi}^2 \\
+ \frac{1}{n_h} \left( \frac{1}{n_h - 1} \right) \left[ 2 m_{xyh} + r_{xyh} \left( \frac{r_{xyh} m_{xxyh}}{r_{xh} r_{yh}} \right) \right] \\
+ (r_{xyh} + m_{xyh}) \left( \frac{r_{xyh} m_{xxyh}}{r_{xh} r_{yh}} + k_{xyh} \right) + k_{xyh} \left( t_{xyh} - 2 r_{xyh} \right) \\
\times \left[ \left( \sum_{i=1}^{n_h} W_{hi} \right)^2 - \sum_{i=1}^{n_h} W_{hi}^2 \right].
\]

(5.272)
Proof: Under the given imputation method, by Lemma 5.1, expression (5.162),

\[ P_{xh} = \frac{1}{m_{xh}} \left( \frac{m_{xyh}}{r_{xyh}} + \frac{m_{ryh}}{r_{xyh}} \right) \]  
(5.273)

and, by expression (5.163),

\[ P_{yh} = \frac{1}{m_{yh}} \left( \frac{m_{xyh}}{r_{xyh}} + \frac{m_{ryh}}{r_{yh}} \right) \]  
(5.274)

and, by expression (5.164),

\[ P_{1h} = \frac{1}{m_{xh}} \left( \frac{m_{xyh}}{m_{yh}} - m_{xyh} \right) \left[ \frac{m_{xyh}}{r_{xyh}} + \frac{mr_{xyh}}{r_{xyh}} \right] \]
\[ + \frac{mr_{xyh} m_{xyh}}{r_{yh} r_{xyh}} + \frac{rm_{xyh} m_{xyh}}{r_{xyh} r_{xyh}} \]
\[ + \frac{k_{xyh} \left( m_{xyh} - r_{xyh} + t_{xyh} \right)}{r_{xyh}} \]  
(5.275)

and, by expression (5.165),

\[ P_{2h} = \frac{1}{r_{xyh}} \]  
(5.276)
Expression (5.272) follows from substituting (5.273) through (5.276) into (5.202).

#

Alternative Proof: Using expression (5.232) in the first alternative proof of Theorem 5.9,

\[
P( Y_{hi}^* \text{ and } X_{hj}^* \text{ both come from the same element in } S_{4h} \mid Y_{hi} \in S_{3h} \text{ and } X_{hj} \in S_{2h}, cs, R_1 )
\]

= \[ P( Y_{hi}^* \in S_{4h} \mid Y_{hi} \in S_{3h}, X_{hj} \in S_{2h}, cs, R_1 ) \]

\times \ P( X_{hj}^* \in S_{4h} \mid Y_{hi}^* \in S_{4h}, Y_{hi} \in S_{3h}, X_{hj} \in S_{2h}, cs, R_1 )

\times \ P( Y_{hi}^* \text{ and } X_{hj}^* \text{ both come from the same element of } S_{4h} \mid Y_{hi}^* \in S_{4h}, X_{hj}^* \in S_{4h}, Y_{hi} \in S_{3h}, X_{hj} \in S_{2h}, cs, R_1 )

= \frac{r_{xyh}}{r_{yh}} \times \frac{r_{xyh}}{r_{xh}} \times \frac{1}{r_{xyh}} = \frac{r_{xyh}}{r_{yh} \ r_{xh}}. \tag{5.277}

Similarly,

\[
P( Y_{hi}^* \text{ and } X_{hj}^* \text{ both come from the same element in } S_{4h} \mid Y_{hi} \in S_{3h} \text{ and } X_{hj} \in S_{1h}, cs, R_1 )
\]
\begin{align}
\frac{r_{xyh}}{r_{yh}} \times 1 \times \frac{1}{r_{xyh}} = \frac{1}{r_{yh}}, \\
&= \frac{r_{xyh}}{r_{xh}} \times 1 \times \frac{1}{r_{xyh}} = \frac{1}{r_{xh}},
\end{align}

(5.278)

\[ P( Y_{hi}^* \text{ and } X_{hj}^* \text{ both come from the same element in } S_{4h} \mid Y_{hi} \in S_{1h} \text{ and } X_{hj} \in S_{2h}, cs, R_1 ) \]

\begin{align}
&= 1 \times \frac{r_{xyh}}{r_{xh}} \times \frac{1}{r_{xyh}} = \frac{1}{r_{xh}},
\end{align}

(5.279)

and

\[ P( Y_{hi}^* \text{ and } X_{hj}^* \text{ both come from the same element in } S_{4h} \mid Y_{hi} \in S_{1h} \text{ and } X_{hj} \in S_{1h}, cs, R_1 ) \]

\begin{align}
&= 1 \times 1 \times \frac{k_{xyh} \left( k_{xyh} - 1 \right) ( r_{xyh} - t_{xyh} ) + ( k_{xyh} + 1 ) k_{xyh} t_{xyh}}{m_{xyh} \left( m_{xyh} - 1 \right)} \\
&= \frac{k_{xyh} \left( m_{xyh} - r_{xyh} + t_{xyh} \right)}{m_{xyh} \left( m_{xyh} - 1 \right)},
\end{align}

(5.280)

The

\[ P( Y_{hi}^* \text{ comes from } S_{4h} \mid Y_{hi} \in S_{3h}, cs, R_1 ) \]
\[ = \frac{r_{xyh}}{r_{yh}}. \quad (5.281) \]

The

\[ P( Y_{hi}^* \text{ comes from } S_{4h} | Y_{hi} \in S_{1h}, cs, R_1) = 1. \quad (5.282) \]

The

\[ P( X_{hj}^* \text{ comes from } S_{4h} | X_{hj} \in S_{2h}, cs, R_1) \]

\[ = \frac{r_{xyh}}{r_{xh}}. \quad (5.283) \]

The

\[ P( X_{hj}^* \text{ comes from } S_{4h} | X_{hj} \in S_{1h}, cs, R_1) = 1. \quad (5.284) \]

The

\[ P( Y_{hi}^* \text{ and } X_{hi}^* \text{ both come from the same element of } S_{4h} \]

\[ | (Y_{hi}, X_{hi}) \in S_{1h}, cs, R_1) = 1. \quad (5.285) \]

Substituting (5.277) through (5.285) into (5.232) we get the result in (5.272).
5.1.12. Estimation of the covariance of two imputed totals minus their respective finite population totals. We now present an estimator for $\text{Cov}(\hat{Y}^* - Y, \hat{X}^* - X)$. We start by finding $\text{Cov}(\hat{Y}^* - Y, \hat{X}^* - X)$.

Theorem 5.10: Let the assumptions of Theorem 5.9 hold. Then

$$\text{Cov}(\hat{Y}^* - Y, \hat{X}^* - X) = \text{Cov}(\hat{Y} - Y, \hat{X} - X) + \sum_{h=1}^{H} \sigma_{xy} E\left( \frac{n_h}{n} (r_{xy} - \frac{n_h}{n} + r_{xy} \frac{m_{xy} P_{2h}}{n}) \sum_{i=1}^{n_h} W_{hi}^2 \right)$$

$$+ \frac{1}{n_h} r_{xy} \left[ \left( \frac{n_h - 1}{n_h} \right)^{-1} \left( \frac{m_{xy} P_x + m_{xy} P_y + (m_{xy} m_{xy} - m_{xy}) P_{1h}}{n_h} \right) \right]$$

$$\times \left\{ \left( \sum_{i=1}^{n_h} W_{hi} \right)^2 - \sum_{i=1}^{n_h} W_{hi}^2 \right\}.$$ (5.286)

Proof: The

$$\text{Cov}(\hat{Y}^* - Y, \hat{X}^* - X) = \text{Cov}(\hat{Y}^* - \overline{N_y}, \hat{X}^* - \overline{N_x})$$

$$- \text{Cov}(\hat{Y} - \overline{N_y}, \hat{X} - \overline{N_x}) - \text{Cov}(\hat{Y} - Y, \hat{X} - X)$$

$$+ \text{Cov}(\hat{Y}^* - Y, \hat{X} - X) + \text{Cov}(\hat{Y} - Y, \hat{X}^* - X)$$
\[- \text{Cov}( \hat{Y}^* - Y, \hat{X} - N \bar{\mu}_x) - \text{Cov}( \hat{Y} - N \bar{\mu}_y, \hat{X}^* - X) \]
\[+ \text{Cov}( \hat{Y} - N \bar{\mu}_y, \hat{X} - X) + \text{Cov}( \hat{Y} - Y, \hat{X} - N \bar{\mu}_x) \]
\[= \text{Cov}( \hat{Y}^* - N \bar{\mu}_y, \hat{X}^* - N \bar{\mu}_x) - \text{Cov}( \hat{Y} - N \bar{\mu}_y, \hat{X} - N \bar{\mu}_x) \]
\[+ \text{Cov}( \hat{Y}, \hat{X}) - \text{Cov}( \hat{Y}^*, X) \]
\[- \text{Cov}( Y, \hat{X}^*) + \text{Cov}( Y, X). \quad (5.287) \]

Now

\[\text{Cov}( \hat{Y}^* - \hat{Y}, X) = E[ E( \hat{Y}^* - \hat{Y} | \text{FP}) X ] \quad (5.288)\]

since

\[E( \hat{Y}^* - \hat{Y}) = 0 \quad (5.289)\]

under the superpopulation model. The

\[E[ E( \hat{Y}^* - \hat{Y} | \text{FP}) X ] \]
\[= E[ \sum_{h=1}^{H} \frac{m_{yh}}{n_h - 1} \sum_{i=1}^{n_h} W_{hi} E( \bar{Y}_h X - Y_{hi} X | cs, R_1 ) ] \]
It follows that

$$\text{Cov}( \hat{Y}^*, X) = \text{Cov}( \hat{Y}, X).$$  \hspace{1cm} (5.291)

Similarly

$$\text{Cov}( Y, \hat{X}^*) = \text{Cov}( Y, \hat{X}),$$  \hspace{1cm} (5.292)

so

$$\text{Cov}( \hat{Y}^* - Y, \hat{X}^* - X) = \text{Cov}( \hat{Y} - Y, \hat{X} - X)$$

$$+ \text{Cov}( \hat{Y}^* - N \bar{Y}, \hat{X}^* - N \bar{X}) - \text{Cov}( \hat{Y} - N \bar{Y}, \hat{X} - N \bar{X}).$$  \hspace{1cm} (5.293)

Result (5.286) follows from result (5.202) in Theorem 5.9 since (5.293) holds.

We note that
\[ \text{Cov}(\hat{Y} - Y, \hat{X} - X) = E[\text{Cov}(\hat{Y} - Y, \hat{X} - X | FP)] \] (5.294)

since

\[ E(\hat{Y} - Y | FP) = E(\hat{X} - X | FP) = 0. \] (5.295)

We next find the Cov(\( \hat{Y} - Y, \hat{X} - X | FP \)).

**Theorem 5.11:** Let the assumptions of Theorem 5.9 hold. Then

\[
\text{Cov}(\hat{Y} - Y, \hat{X} - X | FP) = \sum_{h=1}^{H} \sum_{i=1}^{N_h} \left[ 1 - \frac{\pi_{hi}}{\pi_i} \right] \frac{\pi_{hi}}{\pi_i} Y_{hi} X_{hi} + \sum_{h=1}^{H} \sum_{i=1}^{N_h} \sum_{g=1}^{N_{hi}} \sum_{j=1}^{N_{gj}} \left( \frac{\pi_{hi} \pi_{gj} - \pi_{hi} \pi_{gj}}{\pi_{hi} \pi_{gj}} \right) Y_{hi} X_{gj}.
\] (5.296)

**Proof:** First,

\[
\hat{Y} = \sum_{h=1}^{H} \sum_{i=1}^{N_h} \pi_{hi}^{-1} a_{hi} Y_{hi}
\] (5.297)

and

\[
\hat{X} = \sum_{h=1}^{H} \sum_{i=1}^{N_h} \pi_{hi}^{-1} a_{hi} X_{hi},
\] (5.298)

where \( a_{hi}, h=1,...,H, i=1,...,N_h \), is a 0–1 random variable that takes on the value
one if $(Y_{hi}, X_{hi})$ is in the sample and takes on the value 0 if $(Y_{hi}, X_{hi})$ is not in the sample, as defined in the first proof of Theorem 5.2. Then

\[
\text{Cov}(\hat{Y} - Y, \hat{X} - X | FP) = E\left[ \sum_{h=1}^{H} \sum_{i=1}^{N_h} (\pi_{hi}^{-1} a_{hi} Y_{hi} - Y_{hi}) \sum_{g=1}^{H} \sum_{j=1}^{N_g} (\pi_{gj}^{-1} a_{gj} X_{gj} - X_{gj}) \right] | FP
\]

\[
= \sum_{h=1}^{H} \sum_{i=1}^{N_h} \sum_{g=1}^{H} \sum_{j=1}^{N_g} Y_{hi} X_{gj}
\]

\[
\times E\left[ \frac{a_{hi}}{\pi_{hi}} \frac{a_{gj}}{\pi_{gj}} - \frac{a_{hi}}{\pi_{hi}} - \frac{a_{gj}}{\pi_{gj}} + 1 \right] | FP,
\]

since $E(\hat{Y} - Y | FP)$ and $E(\hat{X} - X | FP)$ equal zero. Now, if $hi=gj$,

\[
E\left[ \frac{a_{hi}}{\pi_{hi}} \frac{a_{gj}}{\pi_{gj}} - \frac{a_{hi}}{\pi_{hi}} - \frac{a_{gj}}{\pi_{gj}} + 1 \right] | FP
\]

\[
= \frac{\pi_{hi}}{\pi_{hi}^2} - \frac{\pi_{hi}}{\pi_{hi}} - \frac{\pi_{hi}}{\pi_{hi}} + 1 = \frac{1 - \pi_{hi}}{\pi_{hi}}
\]

and, if $hi \neq gj$,

\[
E\left[ \frac{a_{hi}}{\pi_{hi}} \frac{a_{gj}}{\pi_{gj}} - \frac{a_{hi}}{\pi_{hi}} - \frac{a_{gj}}{\pi_{gj}} + 1 \right] | FP
\]
\[
\frac{\frac{\pi_{hi,g,j}}{\pi_{hi} \pi_{g,j}}}{\pi_{hi} \pi_{g,j}} = 1 + \frac{\pi_{hi,g,j}}{\pi_{hi} \pi_{g,j}}
\]

(5.301)

Result (5.296) follows from substituting (5.300) and (5.331) into (5.299).

We now present an estimator of the Cov(\( \hat{Y} - Y, \hat{X} - X \mid FP \)).

**Theorem 5.12:** Let the assumptions of Theorem 5.9 hold. Then

\[
\text{Cov}(\hat{Y} - Y, \hat{X} - X) = \frac{\sum_{h=1}^{H} \sum_{i=1}^{n_h} 1 - \frac{\pi_{hi}}{\pi_{hi}}}{\sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{n_g} \pi_{hi,g,j}} \frac{\pi_{hi,g,j}}{\pi_{hi} \pi_{g,j}} \frac{\pi_{hi}}{\pi_{hi}} Y_{hi} X_{hi}
\]

(5.302)

is an unbiased estimator of Cov(\( \hat{Y} - Y, \hat{X} - X \mid FP \)).

**Proof:** Using the a_{hi}'s of Theorem 5.2, the expected value of expression (5.302) is

\[
\frac{\sum_{h=1}^{H} \sum_{i=1}^{N_h} E( a_{hi} ) 1 - \frac{\pi_{hi}}{\pi_{hi}}}{\sum_{h=1}^{H} \sum_{i=1}^{N_h} \frac{\pi_{hi}}{\pi_{hi}}} Y_{hi} X_{hi}
\]
We now present an unbiased estimator of $\text{Cov}(\hat{Y} - Y, \hat{X} - X)$.

**Theorem 5.13:** Let the assumptions of Theorem 5.9 hold. Let

$$\text{Cov}(\hat{Y} - Y, \hat{X} - X | \text{FP}) = \frac{H}{\Sigma} \sum_{h=1}^{H} \sum_{i=1}^{n_h} \frac{1 - \pi_h}{\pi_h} Y_{hi} X_{hi}$$

$$+ \frac{H}{\Sigma} \sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{g=1}^{n_g} \sum_{j=1}^{n_g} \frac{\pi_{hi,gj} - \pi_h \pi_{gj}}{\pi_h \pi_{gj} \pi_{hi,gj}} Y_{hi} X_{gj}$$

be an estimator of the covariance between $\hat{Y} - Y$ and $\hat{X} - X$ given the finite population.
using the imputed values instead of the full sample. Let

\[ \bar{y}_{xyh} = \sum_{Y_{hi} \in S_{4h}} r_{xyh}^{-1} Y_{hi} \]  

(5.305)

be the sample mean of the Y's for the units that responded on both X and Y in imputation class h, let

\[ \bar{x}_{xyh} = \sum_{X_{hi} \in S_{4h}} r_{xyh}^{-1} X_{hi} \]  

(5.306)

be the sample mean of the X's for the units that responded on both X and Y in imputation class h, and let

\[ s_{xyh} = \sum_{(Y_{hi}, X_{hi}) \in S_{4h}} \frac{(Y_{hi} - \bar{y}_{xyh})(X_{hi} - \bar{x}_{xyh})}{r_{xyh}^{-1}} \]  

(5.307)

be the sample covariance between the X's and Y's that both responded in imputation class h. Then

\[ \tilde{Cov}(\tilde{Y^*} - \tilde{Y}, \tilde{X^*} - \tilde{X}) = \tilde{Cov}(\tilde{Y} - \tilde{Y}, \tilde{X} - \tilde{X} \mid \text{FP}) \]

\[ + \sum_{h=1}^{H} s_{xyh} \left\{ n_{h}^{-1} (r_{xyh} - n_{h} + r_{xyh} m_{xyh} P_{2h}) \right\} \]

\[ \times \sum_{i=1}^{n_{h}} \left[ \frac{n_{hi}^{-2}}{\pi_{hi}^{-2} (1 - \pi_{hi})} \right] \]
\[
+ n_h^{-1} (n_h - 1)^{-1} r_{xyh} \left[ m_{xh} P_{xh} + m_{yh} P_{yh} + (m_{xh} m_{yh} - m_{xyh}) P_{1h} \right] \\
\quad \times \sum_{i=1}^{n_h} \sum_{j=1}^{n_h} \left[ \tau_{hi}^{-1} \tau_{hj}^{-1} - \tau_{hi}^{-1} \tau_{hj}^{-1} \right] \right] \\
\quad \times \sum_{i=1}^{n_h} \sum_{j=1}^{n_h} \left[ \tau_{hi}^{-1} \tau_{hj}^{-1} \tau_{hi,hj} \left( \tau_{hi,hj} - \tau_{hi} \tau_{hj} \right) \right] \\
\quad \times \sum_{i\neq j}^{n_h} \left[ \tau_{hi}^{-1} \tau_{hj}^{-1} \tau_{hi,hj} \left( \tau_{hi,hj} - \tau_{hi} \tau_{hj} \right) \right] \\
\quad + \sum_{h=1}^{H} E\{ E[ s_{xyh} | cs, R_1 ] \}
\]

is an unbiased estimator of the Cov( \( \hat{Y}^* - Y, \hat{X}^* - X \) ).

Proof: The

\[
E[ \text{Cov}( \hat{Y}^* - Y, \hat{X}^* - X ) ] = E\{ \text{Cov}^* ( \hat{Y} - Y, \hat{X} - X | FP ) \}
\]

\[
- \sum_{h=1}^{H} E[ s_{xyh} | cs, R_1 ] \\
\quad \times \left[ n_h^{-1} (r_{xyh} - n_h + r_{xyh} m_{xyh} P_{2h}) \sum_{i=1}^{n_h} \tau_{hi}^{-2} (1 - \tau_{hi}^{-1}) \right]
\]

\[
+ n_h^{-1} (n_h - 1)^{-1} r_{xyh} \left[ m_{xh} P_{xh} + m_{yh} P_{yh} + (m_{xh} m_{yh} - m_{xyh}) P_{1h} \right] \\
\quad \times \sum_{i=1}^{n_h} \sum_{j=1}^{n_h} \left[ \tau_{hi}^{-1} \tau_{hj}^{-1} \tau_{hi,hj} \left( \tau_{hi,hj} - \tau_{hi} \tau_{hj} \right) \right] \\
\quad \times \sum_{i\neq j}^{n_h} \left[ \tau_{hi}^{-1} \tau_{hj}^{-1} \tau_{hi,hj} \left( \tau_{hi,hj} - \tau_{hi} \tau_{hj} \right) \right] \\
\quad + \sum_{h=1}^{H} E\{ E[ s_{xyh} | cs, R_1 ] \}
\]
The expected value of the first term in (5.309) is \( \text{Cov}(Y - Y, X - X) \) since

\[
\mathbb{E}[s_{xyh} | cs, R_1] = \sigma_{xhy}
\]  

(5.310)
\[
\left[ ( \eta \Delta x_w - \eta \Delta x_w \eta \Delta x_w ) + \eta \Delta c \eta \Delta w + \eta \Delta d \eta \Delta w \right] \eta \Delta x_1 \mid_{\eta} \left( I - \eta u \right) \eta_u \mid_{\eta} \} \mathfrak{g} \times
\]

\[
\eta \Delta x_\rho \mid_{\eta} \frac{I=\eta}{\mathfrak{g}} \frac{H}{H}
\]

\[
\left[ ( \mathfrak{g} \eta \eta \eta - \mathfrak{g} \eta \eta \eta \right] \times
\]

\[
( \mathfrak{g} \eta \eta \eta - \mathfrak{g} \eta \eta \eta \right] \times
\]

\[
\text{and}
\]

\[
( \eta \eta - I ) \eta \eta \frac{I=I}{\mathfrak{g}} \frac{H}{u}
\]

\[
( \eta \Delta d \eta \Delta x_w \eta \Delta x_1 + \eta u - \eta \Delta x_1 ) \eta_u \mid_{\eta} \} \mathfrak{g} \times
\]

\[
\left[ ( \eta \eta \eta \eta \eta - \eta \eta \eta \eta \eta \right] \times
\]

\[
\text{and}
\]

\[
( \mathfrak{g} \eta \eta \eta - \mathfrak{g} \eta \eta \eta \right] \times
\]

121
The second term in (5.309) is an unbiased estimator of the amount to be added to the $Cov(\hat{Y} - Y, \hat{X} - X)$ to account for the imputation, as can be seen from expression (5.286) in Theorem 5.10.

5.2. The model in PC CARP

5.2.1. Introduction to the estimation of the variance of the imputed total and the covariance of two imputed totals for PC CARP

In this section, we apply the results derived in the first part of this chapter to construct estimators to be implemented in PC CARP. We describe the type of sampling scheme that PC CARP is set up to analyze. We then describe estimators of the variance and covariance as they might be calculated by PC CARP.

5.2.2. The population and sample structure in PC CARP

Let there be $L$ sampling strata, indexed by $i$. Within stratum $i$, $i=1,...,L$, let there be $C_i$ clusters, indexed by $j$. Within cluster $ij$, $i=1,...,L$, $j=1,...,C_i$, let there be $B_{ij}$ individuals, indexed by $k$. It is assumed that a sample of size $c_i$ is taken by simple random sampling with or without replacement from the $C_i$ clusters within stratum $i$, $i=1,...,L$. Also, it is assumed that a sample of $b_{ij}$ individuals within cluster $ij$ is taken by simple random sampling without replacement from the $B_{ij}$ individuals in cluster $ij$, $i=1,...,L$, $j=1,...,C_i$. This sampling structure is a simplification of the sampling structure permitted in PC CARP. The sampling structure represents a
special case of the general sampling structure given in the first part of this chapter.

Assume that there are $H$ imputation classes within the population, indexed by $h$. Assume the $H$ imputation classes are random samples from a superpopulation as described in Section 5.1.4. Assume that the imputation classes cross the strata, clusters, and individuals. Then the population is made up of

$$\sum_{i=1}^{L} \sum_{j=1}^{C_i} \sum_{h=1}^{H} \sum_{k=1}^{B_{ijh}} 1 = N$$

individuals with characteristic $\{Y_{ijhk}\}_{i=1,\ldots,L,\,j=1,\ldots,C_i,\,h=1,\ldots,H,\,k=1,\ldots,B_{ijh}}$, where $B_{ijh}$ is the number of individuals in imputation class $h$ and cluster $ij$ in the population and where

$$\sum_{h=1}^{H} \sum_{k=1}^{B_{ijh}} 1 = B_{ij}$$

The sample is made up of

$$\sum_{i=1}^{L} \sum_{j=1}^{c_i} \sum_{h=1}^{H} \sum_{k=1}^{b_{ijh}} 1 = n$$

individuals with characteristic $\{Y_{ijhk}\}_{i=1,\ldots,L,\,j=1,\ldots,c_i,\,h=1,\ldots,H,\,k=1,\ldots,b_{ijh}}$, where $b_{ijh}$ is the number of individuals in imputation class $h$ and cluster $ij$ in the sample and where

$$\sum_{h=1}^{H} \sum_{k=1}^{b_{ijh}} 1 = b_{ij}$$
Let the augmented sample be \( \{Y_{ijhk}^*\}_{i=1,\ldots,L_1, j=1,\ldots,c_{ij}, h=1,\ldots,H, k=1,\ldots,b_{ijh}} \). 

Then \( Y_{ijhk}^* = Y_{ijhk} \) if individual \( ijhk \) responded and \( Y_{ijhk}^* \) is an imputed value if individual \( ijhk \) did not respond. Let \( W_{ij} = C_i c_{ij}^{-1} B_{ij} b_{ijh}^{-1} \). Then \( W_{ij} \) is the inverse of the probability that individual \( ijhk \) is in the sample. For one stage sampling at the cluster level, \( b_{ij} b_{ij}^{-1} = 1 \).

5.2.3. The expectation and variance of the imputed total for PC CARP

We now look at the expected value and variance of \( \hat{Y}^* \). We have

\[
\hat{Y}^* = \sum_{i=1}^{L} \sum_{j=1}^{c_{ij}} \sum_{h=1}^{H} \sum_{k=1}^{b_{ijh}} W_{ij} Y_{ijhk}^* ,
\]

and

\[
\hat{Y} = \sum_{i=1}^{L} \sum_{j=1}^{c_{ij}} \sum_{h=1}^{H} \sum_{k=1}^{b_{ijh}} W_{ij} Y_{ijhk} .
\]

Given that the assumptions of Theorem 5.2 hold,

\[
E(\hat{Y}^*) = N \bar{\mu} ,
\]

where

\[
\bar{\mu} = \sum_{i=1}^{L} \sum_{j=1}^{c_{ij}} \sum_{h=1}^{H} B_{ijh} N^{-1} \mu_{i,j,h} .
\]
where \( \mu_{..h.} \) is the superpopulation mean in imputation class \( h \).

Given that the assumptions of Theorem 5.2 hold, by Theorem 5.4,

\[
V(\hat{Y}^* - Y) = V(\hat{Y} - Y) + \sum_{h=1}^{H} \sigma_{..h.}^2 \cdot \mathbb{E}\left\{ \frac{k_{..h.} \cdot n_{..h.}}{n_{..h.} \cdot (n_{..h.} - 1)} \cdot \left( \sum_{i=1}^{L} \sum_{j=1}^{c_i} b_{ijh} W_{ij} \right)^2 \right\}
\]

\[
= \sum_{i=1}^{L} \sum_{j=1}^{c_i} b_{ijh} W_{ij}^2 \}
\]

(5.318)

where

\[
k_{..h.} = \left\lfloor \frac{m_{..h.}}{r_{..h.}} \right\rfloor
\]

(5.319)

(where \( \lfloor x \rfloor \) is the largest integer less than or equal to \( x \)),

\[
t_{..h.} = m_{..h.} - k_{..h.} \cdot r_{..h.}
\]

(5.320)

\( r_{..h.} \) is the number of individuals that responded in imputation class \( h \), \( m_{..h.} \) is the number of individuals that did not respond in imputation class \( h \), \( n_{..h.} \) is the number of individuals in the sample in imputation class \( h \), and \( \sigma_{..h.}^2 \) is the superpopulation variance in imputation class \( h \).
5.2.4. Estimation of the variance of the imputed total minus the finite population total for PC CARP  
We now present an estimator of $V(Y^* - Y)$. Let

$$s^2_{..h.} = \sum_{i=1}^{L} \sum_{j=1}^{c_i} \sum_{k=1}^{b_{ijh}} \gamma_{ijhk} \frac{(Y_{ijhk} - \bar{y}_{..h.})^2}{r_{..h.} - 1},$$  
(5.321)

where $\gamma_{ijhk}$ is a 0–1 random variable that takes on the value one if unit $ijhk$ responds and takes on the value zero if unit $ijhk$ does not respond and

$$\bar{y}_{..h.} = \frac{\sum_{i=1}^{L} \sum_{j=1}^{c_i} \sum_{k=1}^{b_{ijh}} \gamma_{ijhk} Y_{ijhk}}{r_{..h.}}.$$  
(5.322)

Then $\bar{y}_{..h.}$ is the sample mean of the respondents in imputation class $h$ and $s^2_{..h.}$ is the sample variance of the respondents in imputation class $h$.

**Theorem 5.14:** Let the assumptions of Theorem 5.2 hold. Let sampling be as described in Section 5.2.2. Assume that $r_h > 2$ for all $h$. Let

$$\hat{V}(Y^* - Y) = \hat{V}(\hat{Y} - Y | FP)$$

$$+ \sum_{h=1}^{H} s^2_{..h.} \frac{k_{..h.} n_{..h.}}{n_{..h.}} + \left( k_{..h.} + 2 \right) s_{..h.}$$

$$\times \left\{ \sum_{i=1}^{L} \sum_{j=1}^{c_i} b_{ijh} W_{ij} \right\}^2 - \sum_{i=1}^{L} c_i \left( \sum_{j=1}^{b_{ijh}} W_{ij} \right)^2.$$
where \( s^2 \), is defined in (5.321), \( \hat{V}^* (\hat{Y} - Y | FP) \) is the estimator of the \( V(\hat{Y} - Y | FP) \) in PC CARP when an augmented sample is analyzed, and it is understood that \( (b_{ij}^{-1})^{-1} (B_{ij} - 1) = 0 \) for \( b_{ij} = 1 \). Then \( \hat{V}^* (\hat{Y} - Y) \), is an unbiased estimator of \( V(\hat{Y} - Y) \).

Proof: Let \( \pi_{ijhk} \) be the probability the unit \( ijhk \) is in the sample and let \( \pi_{ijhk,mngl} \) be the probability that units \( ijhk \) and \( mngl \) are both in the sample. Then

\[
\pi_{ijhk} = C_i^{-1} c_i B_{ij}^{-1} b_{ij} = W^{-1}_{ij}
\]

(5.324)

for all \( i,j,h,k \) and \( k \)

\[
\pi_{ijhk,mngl} = C_i^{-1} c_i B_{ij}^{-1} b_{ij} C_m^{-1} c_m B_{mn}^{-1} b_{mn} = W^{-1}_{ij} W^{-1}_{mn}
\]

(5.325)

if \( i \neq m \)
Substituting expressions (5.324) through (5.327) into expression (5.104), we get result (5.323), since the estimator for \( V(Y - \hat{Y}|FP) \) given in PC CARP agrees with \( \hat{V}(Y - \hat{Y}|FP) \) given in expression (5.92). #

We now evaluate the relative importance of the estimator of the bias in \( \hat{V}^*(Y - \hat{Y}|FP) \). For different strata,

\[
\frac{\pi_{ijk,mnl}}{\pi_{ijk}} - 1 = 1 - 1 = 0 . \tag{5.328}
\]

For the same strata but different clusters,
\[
\frac{\pi_{ijhk,inhl}}{\pi_{ijhk} \pi_{inhl}} - 1 = \left( 1 - c_i^{-1} c_i \right) (c_i - 1)^{-1}, \tag{5.329}
\]

where \( C_i \) is the number of clusters in the population and \( c_i \) is the number of clusters in the sample in stratum \( i \). For the same strata and the same cluster but different individuals

\[
\frac{\pi_{ijhk,ijhl}}{\pi_{ijhk} \pi_{ijhl}} - 1 = \left[ \frac{c_i}{C_i} \frac{b_{ij}}{b_{ij} - 1} \frac{B_{ij} - 1}{B_{ij}} \right]. \tag{5.330}
\]

We assume that the population cluster sizes are bounded. Since \( C_i^{-1} c_i \) does not necessarily go to zero, the approach to the relative importance of estimator of the bias in \( \hat{V}^* (\hat{Y} - Y | FP) \) give in Section 5.1.7 does not give a definitive answer, so we directly evaluate the relative importance in the estimator of the bias in \( \hat{V}^* (\hat{Y} - Y | FP) \). The relative importance of the estimator of the bias in \( \hat{V}^* (\hat{Y} - Y | FP) \) for two stage sampling is

\[
\frac{\sum_{i=1}^{L} \left\{ \frac{c_i}{C_i} \frac{b_{ij} b_{inh}}{b_{ij} - 1} \frac{B_{ij}}{B_{in}} \right\} \left( 1 - \frac{c_i}{C_i} \right) (c_i - 1)^{-1} }{\sum_{j=1}^{C_i} b_{ij} (b_{ij} - 1) \frac{c_i^2}{C_i^2} \frac{B_{ij}^2}{B_{ij}} \left[ \frac{c_i}{C_i} \frac{b_{ij}}{b_{ij} - 1} \frac{B_{ij} - 1}{B_{ij}} \right]}
\]
for all \( i \) and \( j \). It follows that the relative importance of the estimator of the bias in \( \hat{V}^*(\hat{Y} - Y|FP) \) is less than

\[
\sum_{i=1}^{L} \left\{ (1 - \frac{c_i}{C_i}) (c_i - 1)^{-1} \right\}
\]

\[
+ \left[ \sum_{j=1}^{C_i} b_{ij} \frac{B_{ij}}{b_{ij}} \left( \sum_{n=1}^{C_i} b_{ihn} \frac{B_{in}}{b_{in}} \right) \right]^{-1}
\]

\[
\times \left[ \sum_{j=1}^{C_i} b_{ij} \frac{B_{ij}}{b_{ij}} \left( b_{ijh} - 1 \right) \frac{B_{ij}}{b_{ij}} \left( \frac{c_i}{C_i} b_{ij} \frac{B_{ij}}{b_{ij}} \frac{B_{ij} - 1}{b_{ij} - 1} - 1 \right) \right]
\]

\[
= O[\max_{i,h} (c_{hi}^{-1})],
\]

where \( c_{hi} \) is the number of clusters in the sample in the stratum \( i \) — imputation
class \( h \) subgroup. Then the order of the relative importance of the bias in \( \hat{V}^*(\hat{Y} - Y|FP) \) is at most \( O(\max(c_{h1}^{-1})) \). If we restrict the \( c_i^{-1}C_i \)'s to all be of the same order, then the order of the relative importance of the bias in \( \hat{V}^*(\hat{Y} - Y|FP) \) is \( \max(c_i, n_h^{-1}) \).

5.2.5. The covariance in PC CARP

We now present estimators for the \( \text{Cov}(\hat{Y}^* - Y, \hat{X}^* - X) \) for the sampling set up of PC CARP. Let the superpopulation construction of Section 5.1.9 hold. Let the population structure and sample structure described in Section 5.2.2 hold. Then we have two characteristics for each individual in the population, \( \{Y_{ijhk}\}_{i=1, \ldots, L, j=1, \ldots, C_i, h=1, \ldots, H, k=1, \ldots, B_{ijh}} \) and \( \{X_{ijhk}\}_{i=1, \ldots, L, j=1, \ldots, C_i, h=1, \ldots, H, k=1, \ldots, B_{ijh}} \) and we have two characteristics for each individual in the sample, \( \{Y_{ijhk}\}_{i=1, \ldots, L, j=1, \ldots, c_i, h=1, \ldots, H, k=1, \ldots, b_{ijh}} \) and \( \{X_{ijhk}\}_{i=1, \ldots, L, j=1, \ldots, c_i, h=1, \ldots, H, k=1, \ldots, b_{ijh}} \). Let \( \hat{Y} \) and \( \hat{Y}^* \) be as defined in (5.314) and (5.315). Let

\[
\hat{X} = \sum_{i=1}^{L} \sum_{j=1}^{C_i} \sum_{h=1}^{H} \sum_{k=1}^{B_{ijh}} W_{ij} X_{ijhk}
\]

(5.334)

and let

\[
\hat{X}^* = \sum_{i=1}^{L} \sum_{j=1}^{c_i} \sum_{h=1}^{H} \sum_{k=1}^{b_{ijh}} W_{ij} X_{ijhk}^*
\]

(5.335)

Let \( r_{y..h} \) be the number of individuals that responded on characteristic \( Y \) in
imputation class \( h \). Let \( m_{y..h} \) be the number of individuals that did not respond on characteristic \( Y \) in imputation class \( h \). Let \( r_{x..h} \) be the number of individuals that responded on characteristic \( X \) in imputation class \( h \). Let \( m_{x..h} \) be the number of individuals that did not respond on characteristic \( X \) in imputation class \( h \). Let \( r_{xy..h} \) be the number of individuals in imputation class \( h \) that responded on both \( X \) and \( Y \). Let \( m_{xy..h} \) be the number of individuals in imputation class \( h \) that did not respond on either \( X \) or \( Y \).

5.2.6. The covariance of two imputed totals minus their respective finite population totals for PC CARP

By Theorem 5.10, under the assumptions of Theorem 5.10,

\[
\text{Cov}( \hat{Y}^* - Y, \hat{X}^* - X ) = \text{Cov}( Y - \hat{Y}, X - \hat{X} )
\]

\[
+ \sum_{h=1}^{H} \sigma_{xy}^{h} E\{ \frac{r_{xy..h}}{n_{..h}} \left[ \frac{m_{x..h} P_{xh} + m_{y..h} P_{yh}}{n_{..h}} - \frac{m_{xy..h}}{n_{..h}} \right] \\
+ \frac{( m_{x..h} - m_{xy..h} )}{n_{..h} - 1} P_{1h} \}
\]

\[
\times \left[ \left( \sum_{i=1}^{L} \sum_{j=1}^{C_i} b_{ijh} W_{ij} \right)^2 - \sum_{i=1}^{L} \sum_{j=1}^{C_i} b_{ijh} W_{ij}^2 \right] \\
+ \left[ \frac{r_{xy..h}}{n_{..h}} - \frac{n_{..h}}{n_{..h}} + \frac{r_{xy..h} m_{xy..h}}{n_{..h}} \right] P_{2h} \]
\]
5.2.7. Estimation of the covariance of two imputed totals minus their respective finite population totals for PC CARP

We now present an estimator for the Cov($\hat{Y}^* - Y, \hat{X}^* - X$). Let $\gamma_{ijhk}$ be a 0–1 random variable that takes on the value one if unit $ijhk$ responds on $Y$ and takes on the value zero if unit $ijhk$ does not respond on $Y$. Let $\gamma_{xijhk}$ be a 0–1 random variable that takes on the value one if unit $ijhk$ responds on $X$ and takes on the value zero if unit $ijhk$ does not respond on $X$. Let

$$\bar{y}_{r..h} = \frac{\sum_{i=1}^{L} c_i \sum_{j=1}^{b_{ijh}} \gamma_{yijhk} \gamma_{xijhk} Y_{ijhk}}{r_{xy..h}}.$$  \hfill (5.337)

be the sample mean in imputation class $h$ of the characteristic $Y$ for the individuals that responded for both $X$ and $Y$. Let

$$\bar{x}_{r..h} = \frac{\sum_{i=1}^{L} c_i \sum_{j=1}^{b_{ijh}} \gamma_{yijhk} \gamma_{xijhk} X_{ijhk}}{r_{xy..h}}.$$  \hfill (5.338)

be the sample mean of the characteristic $X$ for individuals that responded for both $X$ and $Y$. Let

$$s_{xyr..h} = \frac{\sum_{i=1}^{L} c_i \sum_{j=1}^{b_{ijh}} \gamma_{yijhk} \gamma_{xijhk}}{r_{xy..h}}.$$
be the sample covariance of $X$ and $Y$ for the individuals that responded on both $X$ and $Y$.

**Theorem 5.15:** Let the assumptions of Theorem 5.9 hold. Let sampling be as described in Section 5.2.2. Then an unbiased estimator of $\text{Cov}(Y^* - Y, X^* - X)$ is

$$
\hat{\text{Cov}}(Y^* - Y, X^* - X) = \hat{\text{Cov}}(\hat{Y} - Y, \hat{X} - X | FP) + \sum_{h=1}^{H} \mathbf{s}_{xyr..h.} \left\{ n_{..h.}^{-1} (r_{xy..h.} - n_{..h.} + r_{xy..h.} m_{xy..h.} P_{2h}) \right. \\
\left. \times \left[ \sum_{i=1}^{L} \sum_{j=1}^{c_i} W_{ij} (W_{ij} - 1) \right] \right. \\
\left. + n_{..h.}^{-1} (1 - n_{..h.})^{-1} r_{xy..h.} \left[ m_{x..h.} P_{xh} + m_{y..h.} P_{yh} \right. \\
\left. + (m_{x..h.} m_{y..h.} - m_{xy..h.}) P_{1h} \right] \right. \\
\times \left[ \left( \sum_{i=1}^{L} \sum_{j=1}^{c_i} b_{ijh} W_{ij} \right)^2 - \sum_{i=1}^{L} c_i \left( \sum_{j=1}^{c_i} b_{ijh} W_{ij} \right)^2 \right].
$$

(5.339)
where $s_{xyr..h.}$ is defined in expression (5.339), $\text{Cov}^*(\hat{Y} - \hat{Y}, \hat{X} - \hat{X}|FP)$ is the estimator for $\text{Cov}(\hat{Y} - Y, \hat{X} - X|FP)$ in PC CARP when an augmented sample is analyzed, and it is understood that $(b_{ij} - 1)^{-1}(B_{ij} - 1) = 0$ for $b_{ij} = 1$.

Proof: The result follows since

$$\Sigma_{i,j=1}^{c_i} \Sigma_{n=1}^{c_i} b_{ijh} b_{inh} (c_i - 1)^{-1} W_{ij} (c_i W_{in} - \frac{B_{in}}{b_{in}})$$

$$- \Sigma_{j=1}^{c_i} b_{ijh} (b_{ijh} - 1) W_{ij} \frac{B_{ij} - 1}{b_{ij} - 1}$$

$$= \left( \frac{C_i^2 B_{ij}^2}{c_i^2 b_{ij}^2} (1 - \frac{c_i b_{ij}}{C_i B_{ij}}) = W_{ij} (W_{ij} - 1) \right),$$

from the result in the proof of Theorem 5.14 on the weights in the correction term, and since the estimator of $\text{Cov}(\hat{Y} - Y, \hat{X} - X|FP)$ in PC CARP agrees with expression (5.302).
6. IMPUTATION FOR THE SOIL CONSERVATION SERVICE'S 1987 NATIONAL RESOURCES INVENTORY

We look at a practical problem that the Soil Conservation Service is facing with its National Resources Inventory and we apply the techniques developed in Chapter 5 as a solution to a portion of the problem. The Soil Conservation Service has taken surveys in the years 1977, 1982, and 1987, and will take a survey in 1992, to determine the condition and trend of the nation's soil, water, and related resources, as mandated by the Rural Development Act of 1972. These surveys are called National Resources Inventories. The surveys cover the entire country and use sophisticated sampling techniques to get a representative sample for the country. We look at a problem presented by the 1987 data.

6.1. The problem

A full sample was taken in 1982 for the National Resources Inventory, and a full sample will be taken in 1992. In 1987, however, only about one third of the sample points were sampled. For the points that were not sampled in 1987, aerial photography, taken in 1987, is usually available. The objective is to devise a suitable method for imputing the measurements that were not taken in the 1987 survey.

The missing points can be divided into two categories, those points for which aerial photography is available for 1987 and those points for which no aerial photography is available for 1987. The portion of the missing sample points for which aerial photography is not available for 1987 is very small. In this paper, we look only at the problem of imputing when there is aerial photography available for
the 1987 data. We develop imputation methods for the units that were in the 1982 survey, will be in the 1992 survey, and are missing from the 1987 survey by using the units included in all three of the surveys.

6.2. A description of the variables

For the primary sampling units within the survey, several measurements are taken. Some of these measurements are descriptions of the overall primary sampling unit, others are measurements and descriptions at two or three points within the primary sampling unit. The portion of a major land resource area within a county (MLRÂC) is the important tabulation unit. From ISU Statistical Bulletin number 756, page 121, "Major Land Resource Areas are geographically associated land resource units. They are characterized by a particular pattern of soils, climate, vegetation, water resources, land use, and type of farming." The MLRAC does not change over time. It is a given for all three surveys.

For the point data, which is what we will be imputing, there are many variables of interest to us. Points can be characterized by their current cover or use. We will be imputing data only for those points with agricultural activities or with cover such as forest. No imputation is necessary for urban points or points covered with water. The variables that will be available for all of the three surveys are the SCS—SOI—5 slope class (a rough measure of the slope of the land that does not change over time), whether the point is prime farmland, the type of land use at the point for the three years (such as corn, urban land, farmstead, water, strip mine, peanuts, pasture, forest, etc.), and the land capability class (a variable that describes soil quality and other characteristics at the point). The land use can be determined from the aerial photographs.
The variables that we will be imputing for the 1987 survey for which we will have information from 1982 and 1992 are listed below. Three years of cropping history will be imputed, where the cropping history variables are the land use variables for the three years, 1986, 1985, and 1984. Four variables in the Universal Soil—Loss Equation (USLE) — the C—factor, the P—factor, the slope length, and the percent slope — will be imputed. The four Universal Soil—Loss equation variables describe characteristics of the ground at the given sampling point. The percent slope describes the steepness of the ground. The slope length describes the distance until the grade of the land changes. The P—factor and C—factor are composite factors that describe the erodibility of the land.

Fourteen variables in the Wind Erosion Equation (WEQ) — knoll erodibility, the K—factor for 1987 and the three previous years, the L—factor for 1987 and the three previous years, the V—factor for 1987 and the three previous years, and the length of rotation — will be imputed. The fourteen wind erosion equation variables describe characteristics of the point that affect the potential for soil loss due to wind erosion. The knoll erodibility describes characteristics of knolls if knolls are present at the point. The K—factors, L—factors, and V—factors describe, respectively, roughness of tillage ridges, the unsheltered distance that the wind blows at the point, and the amount of vegetative residue on the surface at the point. The rotation length variable describes the number of years between crop rotation. The length of rotation variable takes on the value zero if 1) there is no standard rotation pattern or 2) the land cover is a perennial cover.

Four variables for conservation practices will be imputed. Some conservation practices are engineering practices, in that, relatively permanent structures, such as terraces or drainage tile, are put in to help with conservation of the soil. Also,
whether conservation treatment is needed and the potential of the land for conversion to crop land will be imputed. Conservation practices that are needed will be imputed but information on conservation practices that are needed is not available for 1982.

6.3. A description of the data from Indiana

We studied imputation methods using a data set from 10 counties in Indiana that contains six hundred and twenty sample points. The data set is a set of measurements for both 1982 and 1987 with no missing observations. The Indiana data set contains no Wind Erosion Equation variables because Wind Erosion Equation variables are not measured in Indiana. Also, the variable for potential of land for conversion to crop land was not included in the Indiana data. We compared the 1982 data with the 1987 data to develop imputation procedures to be used in the final study. In general, about 42% of the land in the Indiana data set is used for growing corn and soybeans, about 7% of the land is in other crops, about 10% of the land is pasture, about 21% of the land is in forest, about 1% of the land is in strip mines and other barren land, and about 13% is built up or water or roads. About 6% of the land shifted between uses between 1982 and 1987. There was an increase in urban land from 1982 to 1987 and crop uses rotated between corn and soybeans and other crops. A small amount of prime farmland shifted out of the prime farmland class from 1982 to 1987 due to urbanization.

For the variables that will be available for 1982, 1987, and 1992, the SCS–SOI–5 slope class is fixed for all three years and the MLRAC is fixed for all three years. For the observations that were given a land capability class in 1987, the land capability class stayed the same from 1982 to 1987. The points that were
given a land capability class in 1982 were not given one in 1987 if their use changed to urban. For land use, rotation between corn and soybeans and other crops occurred, but most of the pasture and hay stayed pasture and hay. Also, most of the forest stayed forest, and built up land, water, and roads stayed built up land, water, and roads. Some of the prime farmland shifted out of prime farmland due to urbanization.

For the variables that will be imputed, strong correlations exist between 1982 and 1987. The Universal Soil—Loss Equation variables are continuous variables, and we plotted the values for 1982 against those for 1987. The percent slope, slope length, and P—factor were essentially the same for 1982 and 1987, with a few changes on crop land. For the C—factor, there was a correlation of .7967 between 1982 and 1987. Most of the values for the C—factor were very small and highly correlated, but for the larger values of the C—factor there was a square shaped pattern to the scatter plot. For the largest values of the C—factor, there was perfect correlation between the 1982 and the 1987 variables. The plot is given in Figure 6.1.

For conservation practices, in 1982 the data recorder could list up to three conservation practices and in 1987 the data recorder could list up to four conservation practices. In actuality, virtually none of the points for which this variable was measured had a fourth conservation practice listed in 1987. We compared the 1982 conservation practices to the 1987 conservation practices by forming two variables that described the three practices in 1982 as a group and the four practices in 1987 as a group. The grouped variables were formed by summing the three or four numerical labels for the conservation practices to form a single, unique, numerical label. We then made a table of the grouped 1982 practices versus
the grouped 1987 practices. For 69% of the points, the conservation practices were the same in 1982 and 1987. However, there was a shift into conservation tillage on cropland from 1982 to 1987 and more farms had drainage systems in 1987.

For the variable that indicates whether conservation treatment is needed, 85% of the land stayed the same from 1982 to 1987, 7% of the land went from needing conservation treatment to not needing it, and 8% of the land went from not needing conservation treatment to needing it.

Looking at the kinds of conservation treatments needed in 1987 versus 1982 cropping practices, for the points that did not list any conservation practice in 1982 (360 points), 19% of the points did not need any conservation treatment in 1987, 22% of the points needed erosion control in 1987, 14% of the points needed drainage in 1987, 8% of the points needed both erosion control and drainage in 1987, 9% of the points needed forage improvement in 1987, and 26% of the points needed timber stand improvement in 1987. For the points that listed conservation tillage and (or) drainage as conservation practices in 1982 (117 points), 16% of the points needed no treatment in 1987, 82% of the points needed erosion control and (or) drainage in 1987, and 2% of the points needed forage improvement in 1987. For the points that did not include conservation tillage and (or) drainage as a conservation practice (65 points), 51% needed no treatment in 1987, 23% need erosion control in 1987, 5% needed drainage 1987, 6% needed both erosion control and drainage in 1987, and 15% needed forage management in 1987.

Overall, the 1982 data give a lot of information about the 1987 data. Regressions of the continuous 1987 variables (the USLE variables) were run on the land capability classes (grouped into 7 groups), the SCS—SOI—5 slope classes (grouped into 3 groups), the prime farmland variable for 1982 and 1987 (grouped
Figure 6.1. Indiana C-factors — 1982 versus 1987
into 2 groups), the USLE variables from 1982, and the 1982 and 1987 land use variables (grouped into 6 groups).

The 1982 percent slope explained 99.96% of the variation in the 1987 percent slope and only the 1982 slope length was significant.

The slope length in 1982 explained 91.56% of the variation in the slope length in 1987. The regression equation is

\[
\text{Est. slope length 1987} = 16.06 + .8978 \text{ (slope length 1982)}.
\]

For the P-factor in 1987, 532 points have a P-factor of 1 and 2 points have a P-factor of .6 out of the 534 points with a 1987 P-factor. Therefore no regression was computed.

The results of the regression of the C-factor from 1987 on the predictor variables are given in Table 6.1. The regression shows that the land use variables from 1982 and 1987 and the C-factor from 1982 are the only variables that are significant at any reasonable level of significance. In Table 6.2, the regression of the C-factor from 1987 on the C-factor from 1982 and the land use variable from 1987 (grouped into 10 classes) is given. The regression indicates that a random imputation method based on the C-factor from 1982 and the land use from 1987 would give good results for the C-factor in 1987.

In Table 6.3, the regression of the C-factor from 1987 on the MLRAC's, the land capability class groups given in Table 6.16, and the exact 1987 land uses is given. The results in Table 6.3 indicate that the MLRAC's, the land capability classes and the 1987 land use are good predictors of the 1987 C-factor in Indiana. So an imputation method based on the MLRAC's, the land capability classes, and
The results of the multiple regression indicate that imputing the percent slope and the slope length from 1982 or 1992 is appropriate. It would also be appropriate to impute the P-factor and the conservation treatment needed from 1982 or 1992, except that the conservation treatment needed and the P-factor should be imputed from the same record as the conservation practices. For the conservation practices, there is a shift from 1982 to 1987, so random imputation is more appropriate for the conservation practices.

6.4. A description of the data in Kansas

We studied imputation methods developed on the Indiana data set using a data set from fifteen counties in Kansas that contains twelve hundred and seventy-five sample points. The Kansas data set contains Wind Erosion Equation variables. In general, about 4% of the land in the Kansas data set is in crops other than sorghum or wheat. About 47% of the land is in sorghum, wheat, or summer fallow. About 2% of the land is in pasture and forest, 26% in range land, and 11% built up or water. About 10% of the land shifted between uses between 1982 and 1987. There was a 2% increase in urban land from 1982 to 1987 and crop uses rotated among sorghum, wheat, summer fallow, and other crops. A small amount (4%) of prime farmland shifted out of the prime farmland class from 1982 to 1987 due to urbanization.

For land use, rotation between sorghum, wheat, other crops, and summer fallow occurred. Most of the pasture, forest, and range stayed pasture, forest, and range. About 3% of the range shifted to crop land from 1982 to 1987. The built up land, water, and roads stayed built up, water, and roads. There is no barren land in
Table 6.1. The regression for the 1987 C-Factor in Indiana on 10 factors

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>Sums of Squares</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>24</td>
<td>8.74</td>
<td>69.52</td>
</tr>
<tr>
<td>Error</td>
<td>509</td>
<td>2.67</td>
<td></td>
</tr>
<tr>
<td>Total(corrected)</td>
<td>533</td>
<td>11.40</td>
<td></td>
</tr>
</tbody>
</table>

$R^2 = .7663$

Regression Variables — F's for fit last in the model

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>F</th>
<th>Pr&gt;F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Land Capability Class—grouped</td>
<td>6</td>
<td>.99</td>
<td>.43</td>
</tr>
<tr>
<td>SCS—SOI—5 Slope Class—grouped</td>
<td>2</td>
<td>.55</td>
<td>.58</td>
</tr>
<tr>
<td>Prime Farmland 1982</td>
<td>1</td>
<td>.05</td>
<td>.83</td>
</tr>
<tr>
<td>Prime Farmland 1987</td>
<td>1</td>
<td>.08</td>
<td>.78</td>
</tr>
<tr>
<td>Percent Slope 1982</td>
<td>1</td>
<td>.04</td>
<td>.84</td>
</tr>
<tr>
<td>Slope Length 1982</td>
<td>1</td>
<td>2.35</td>
<td>.13</td>
</tr>
<tr>
<td>C—factor 1982</td>
<td>1</td>
<td>72.18</td>
<td>&lt;.01</td>
</tr>
<tr>
<td>P—factor 1982</td>
<td>1</td>
<td>.32</td>
<td>.57</td>
</tr>
<tr>
<td>Land Use 1982—grouped</td>
<td>5</td>
<td>3.79</td>
<td>&lt;.01</td>
</tr>
<tr>
<td>Land Use 1987—grouped</td>
<td>5</td>
<td>38.95</td>
<td>&lt;.01</td>
</tr>
</tbody>
</table>
Table 6.2. The regression of the C—Factor in 1987 in Indiana on 2 factors

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>Sums of Squares</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>10</td>
<td>8.66</td>
<td>164.96</td>
</tr>
<tr>
<td>Error</td>
<td>523</td>
<td>2.74</td>
<td></td>
</tr>
<tr>
<td>Total(corrected)</td>
<td>533</td>
<td>11.40</td>
<td></td>
</tr>
</tbody>
</table>

\[ R^2 = .7596 \]

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>C—factor 1982</td>
<td>1</td>
<td>67.18</td>
</tr>
<tr>
<td>Land Use 1987—grouped</td>
<td>9</td>
<td>30.07</td>
</tr>
</tbody>
</table>

Table 6.3. The regression of the C—Factor in 1987 in Indiana on 3 factors

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>Sums of Squares</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>40</td>
<td>8.88</td>
<td>43.41</td>
</tr>
<tr>
<td>Error</td>
<td>493</td>
<td>2.52</td>
<td></td>
</tr>
<tr>
<td>Total(corrected)</td>
<td>533</td>
<td>11.40</td>
<td></td>
</tr>
</tbody>
</table>

\[ R^2 = .7789 \]

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>MLRAC</td>
<td>17</td>
<td>4.89</td>
</tr>
<tr>
<td>Land Use 1987</td>
<td>17</td>
<td>54.62</td>
</tr>
<tr>
<td>Land Capability Classes</td>
<td>6</td>
<td>3.41</td>
</tr>
</tbody>
</table>
the Kansas data set.

For the variables that will be imputed, strong correlations exist between the 1982 values and the 1987 values. Plots were made of the USLE and the WEQ variables, plotting 1987 versus 1982. For the USLE variables, the plots of the percent slope and the slope length show most points on the diagonal. However, there is more scatter in the Kansas plots than in the Indiana plots, probably because there is terracing in Kansas. The plot of the P-factor has all but thirty-five points on the diagonal, with the P-factor taking on values of .3, .5, .6, .7, .8, and 1.0. The plot of the C-factor is similar to the plot of the C-factor for Indiana, but there are more points on the diagonal in Kansas.

For the WEQ variables, the rotation length plot shows most points on the diagonal with nice scatter except for zero rotation length. The rotation length variable can take on integer values from zero to six. When the 1987 rotation length was zero, the 1982 rotation length took on values from zero to six. When the 1982 rotation length was zero, the 1987 rotation length took on values from zero to four. Most of the points are at (0,0).

The knoll erodibility plot shows perfect correlation between 1982 and 1987, with most of the points at (0,0). The four plots for the K-factor show most of the points on the diagonal at (.75,.75) and (1.0,1.0) with (.55,.55), (1.0,.75), and (.75, 1.0) also emphasized. The scatter is square shaped with some extreme points. The L-factor and V-factor plots look similar to the C-factor plots for Kansas and Indiana, with many points on the diagonal.

For the conservation practices, 75% of the practices stayed the same from 1982 to 1987. Farms shifted out of conservation tillage and into terracing from 1982 to 1987. Also, there were more ranches using proper grazing techniques in 1987.
For the variable that indicates whether conservation treatment is needed, 89% of the points stayed the same, 6% went from yes to no, and 5% went from no to yes.

For the kinds of conservation treatments needed, for the points that had no conservation practice listed in 1982 (464 points), 18% did not need any treatment in 1987, 52% needed erosion control in 1987, and 28% needed some kind of forage management in 1987. For the points that listed conservation tillage as a conservation practice in 1982 (285 points), 28% needed no treatment in 1987, 69% needed erosion control in 1987, and 2% needed drainage in 1987. For the points that listed terracing as a conservation practice in 1982 (136 points), 38% needed no treatment in 1987 and 62% needed erosion control in 1987. For the points that listed contour farming as a conservation practice in 1982 (135 points), 43% needed no treatment in 1987 and 57% needed erosion control in 1987. For the points that listed proper grazing techniques as a conservation practice in 1982 (190 points), 44% needed no treatment in 1987, 9% needed erosion control in 1987, and 46% needed some kind of forage management in 1987. For the points that did not list conservation tillage, terracing, contour farming, or proper grazing as a conservation practice in 1982, 47% needed no treatment in 1987, 38% needed erosion control in 1987, 10% needed irrigation management in 1987, and 5% needed some kind of forage management in 1987.

Regressions were run for some of the 1987 USLE and WEQ variables using the C-factor from 1982, the L-factor from 1982, the 1987 land use and the county as the predictor variables. The regressions are presented in Tables 6.4 to 6.8. We can see from the tables that an imputation method based on the 1982 values for the C-factor and L-factor, the 1987 land use, and the county would work well for the
1987 C—factor and the 1987 L—factor, would work moderately well for the 1987 K—factor and the 1987 V—factor. The $R^2$ for the 1987 rotation length was about .36, the lowest of any variable.

Also, regressions were run for some of the 1987 USLE and WEQ variables using the MLRAC’s, the land capability class groups given in Table 6.16, and the 1987 land uses. The regressions are given in Tables 6.9 through 6.13. We can see from the tables that an imputation method based on the MLRAC’s, the land capability classes and the 1987 land uses would work well for the K—factors, the V—factors, and the 1987 C—factor and would work moderately well for the L—factors and the 1987 length of rotation. The length of rotation again has the lowest $R^2$, about .38.

6.5. Three imputation methods

In this section, we look at three random imputation methods that were developed for the Indiana and Kansas data sets. The first of the methods is random imputation within imputation cells. This is the imputation method developed in Chapter 5. We evaluate this method with the Indiana data set only. The second method is a hierarchical method in which a search is made for a donor point through a series of levels. This method is closely related to the cell method with small initial cells that are permitted to expand if the initial small cell is empty. The third method is a mixture of a hierarchical method and a distance matching function method. At each level in the hierarchy, a search is made to find the point with the smallest distance from the point to be imputed.

The variables that are imputed by the random imputation methods are the cropping history, the C—factor, the P—factor, the K—factors, the L—factors, the
Table 6.4. The regression of the 1987 C—Factor in Kansas on 4 factors

### ANOVA

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>Sums of Squares</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>38</td>
<td>6.99</td>
<td>94.62</td>
</tr>
<tr>
<td>Error</td>
<td>1065</td>
<td>2.07</td>
<td></td>
</tr>
<tr>
<td>Total(corrected)</td>
<td>1103</td>
<td>9.06</td>
<td></td>
</tr>
</tbody>
</table>

\[ R^2 = .7715 \]

### Regression Variables — F’s for fit last in the model

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>C—factor 1982</td>
<td>1</td>
<td>406.18</td>
</tr>
<tr>
<td>Land Use 1987</td>
<td>22</td>
<td>39.18</td>
</tr>
<tr>
<td>L—factor 1982</td>
<td>1</td>
<td>9.52</td>
</tr>
<tr>
<td>County</td>
<td>14</td>
<td>25.58</td>
</tr>
</tbody>
</table>

Table 6.5. The regression of the 1987 K—Factor in Kansas on 4 factors

### ANOVA

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>Sums of Squares</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>38</td>
<td>14.68</td>
<td>51.58</td>
</tr>
<tr>
<td>Error</td>
<td>1065</td>
<td>7.98</td>
<td></td>
</tr>
<tr>
<td>Total(corrected)</td>
<td>1103</td>
<td>22.66</td>
<td></td>
</tr>
</tbody>
</table>

\[ R^2 = .6480 \]

### Regression Variables — F’s for fit last in the model

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>C—factor 1982</td>
<td>1</td>
<td>15.34</td>
</tr>
<tr>
<td>Land Use 1987</td>
<td>22</td>
<td>43.43</td>
</tr>
<tr>
<td>L—factor 1982</td>
<td>1</td>
<td>1.08</td>
</tr>
<tr>
<td>County</td>
<td>14</td>
<td>9.51</td>
</tr>
</tbody>
</table>
Table 6.6. The regression of the 1987 L-Factor in Kansas on 4 factors

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>Sums of Squares</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>38</td>
<td>$4.00 \times 10^9$</td>
<td>89.75</td>
</tr>
<tr>
<td>Error</td>
<td>1065</td>
<td>$1.25 \times 10^9$</td>
<td></td>
</tr>
<tr>
<td>Total (corrected)</td>
<td>1103</td>
<td>$5.25 \times 10^9$</td>
<td></td>
</tr>
</tbody>
</table>

$R^2 = .7620$

Regression Variables — F’s for fit last in the model

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>C-factor 1982</td>
<td>1</td>
<td>0.01</td>
</tr>
<tr>
<td>Land Use 1987</td>
<td>22</td>
<td>1.18</td>
</tr>
<tr>
<td>L-factor 1982</td>
<td>1</td>
<td>1850.01</td>
</tr>
<tr>
<td>County</td>
<td>14</td>
<td>19.32</td>
</tr>
</tbody>
</table>
Table 6.7. The regression of the 1987 V—Factor in Kansas on 4 factors

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>Sums of Squares</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>38</td>
<td>$9.4 \times 10^9$</td>
<td>28.36</td>
</tr>
<tr>
<td>Error</td>
<td>1065</td>
<td>$9.3 \times 10^9$</td>
<td></td>
</tr>
<tr>
<td>Total (corrected)</td>
<td>1103</td>
<td>$1.86 \times 10^9$</td>
<td></td>
</tr>
</tbody>
</table>

$R^2 = .5030$

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>C—factor 1982</td>
<td>1</td>
<td>6.54</td>
</tr>
<tr>
<td>Land Use 1987</td>
<td>22</td>
<td>21.33</td>
</tr>
<tr>
<td>L—factor 1982</td>
<td>1</td>
<td>.80</td>
</tr>
<tr>
<td>County</td>
<td>14</td>
<td>12.57</td>
</tr>
</tbody>
</table>
Table 6.8. The regression of the 1987 rotation length in Kansas on 4 factors

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>Sums of Squares</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>38</td>
<td>686</td>
<td>15.64</td>
</tr>
<tr>
<td>Error</td>
<td>1057</td>
<td>1220</td>
<td></td>
</tr>
<tr>
<td>Total (corrected)</td>
<td>1095</td>
<td>1906</td>
<td></td>
</tr>
</tbody>
</table>

\[ R^2 = .3599 \]

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>C-factor 1982</td>
<td>1</td>
<td>10.76</td>
</tr>
<tr>
<td>Land Use 1987</td>
<td>22</td>
<td>5.61</td>
</tr>
<tr>
<td>L-factor 1982</td>
<td>1</td>
<td>.98</td>
</tr>
<tr>
<td>County</td>
<td>14</td>
<td>18.65</td>
</tr>
</tbody>
</table>
Table 6.9. The regression of the 1987 C—Factor in Kansas on 3 factors

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>Sums of Squares</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>53</td>
<td>6.38</td>
<td>47.12</td>
</tr>
<tr>
<td>Error</td>
<td>1050</td>
<td>2.68</td>
<td></td>
</tr>
<tr>
<td>Total(corrected)</td>
<td>1103</td>
<td>9.06</td>
<td></td>
</tr>
</tbody>
</table>

\[ R^2 = .7040 \]

Regression Variables — F's for fit last in the model

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>MLRAC</td>
<td>23</td>
<td>14.29</td>
</tr>
<tr>
<td>Land Use 1987</td>
<td>22</td>
<td>67.57</td>
</tr>
<tr>
<td>Land Capability Class</td>
<td>8</td>
<td>2.68</td>
</tr>
</tbody>
</table>

Table 6.10. The regression of the 1987 K—Factor in Kansas on 3 factors

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>Sums of Squares</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>53</td>
<td>14.68</td>
<td>36.44</td>
</tr>
<tr>
<td>Error</td>
<td>1050</td>
<td>7.98</td>
<td></td>
</tr>
<tr>
<td>Total(corrected)</td>
<td>1103</td>
<td>22.66</td>
<td></td>
</tr>
</tbody>
</table>

\[ R^2 = .6478 \]

Regression Variables — F's for fit last in the model

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>MLRAC</td>
<td>23</td>
<td>5.83</td>
</tr>
<tr>
<td>Land Use 1987</td>
<td>22</td>
<td>51.53</td>
</tr>
<tr>
<td>Land Capability Class</td>
<td>8</td>
<td>1.27</td>
</tr>
</tbody>
</table>
Table 6.11. The regression of the 1987 L–Factor in Kansas on 3 factors

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>Sums of Squares</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>53</td>
<td>$1.92 \times 10^6$</td>
<td>11.44</td>
</tr>
<tr>
<td>Error</td>
<td>1050</td>
<td>$3.33 \times 10^9$</td>
<td></td>
</tr>
<tr>
<td>Total(corrected)</td>
<td>1103</td>
<td>$5.25 \times 10^9$</td>
<td></td>
</tr>
</tbody>
</table>

$R^2 = .3660$

Regression Variables — F’s for fit last in the model

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>MLRAC</td>
<td>23</td>
<td>20.02</td>
</tr>
<tr>
<td>Land Use 1987</td>
<td>22</td>
<td>1.50</td>
</tr>
<tr>
<td>Land Capability Class</td>
<td>8</td>
<td>2.17</td>
</tr>
</tbody>
</table>
Table 6.12. The regression of the 1987 V–Factor in Kansas on 3 factors

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>Sums of Squares</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>53</td>
<td>$0.97 \times 10^9$</td>
<td>21.45</td>
</tr>
<tr>
<td>Error</td>
<td>1050</td>
<td>$0.89 \times 10^9$</td>
<td></td>
</tr>
<tr>
<td>Total(corrected)</td>
<td>1103</td>
<td>$1.86 \times 10^9$</td>
<td></td>
</tr>
</tbody>
</table>

$R^2 = 0.5199$

Regression Variables — F's for fit last in the model

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>MLRAC</td>
<td>23</td>
<td>8.79</td>
</tr>
<tr>
<td>Land Use 1987</td>
<td>22</td>
<td>27.99</td>
</tr>
<tr>
<td>Land Capability Class</td>
<td>8</td>
<td>4.50</td>
</tr>
</tbody>
</table>
Table 6.13. The regression of the 1987 rotation length in Kansas on 3 factors

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>Sums of Squares</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>53</td>
<td>715</td>
<td>11.81</td>
</tr>
<tr>
<td>Error</td>
<td>1042</td>
<td>1190</td>
<td></td>
</tr>
<tr>
<td>Total (corrected)</td>
<td>1095</td>
<td>1906</td>
<td></td>
</tr>
</tbody>
</table>

R² = .3752

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>MLRAC</td>
<td>23</td>
<td>12.15</td>
</tr>
<tr>
<td>Land Use 1987</td>
<td>22</td>
<td>8.07</td>
</tr>
<tr>
<td>Land Capability Class</td>
<td>8</td>
<td>1.26</td>
</tr>
</tbody>
</table>

V-factors, the rotation length, the conservation practices, whether conservation treatment in needed, the potential of the land for conversion to cropland, and the kinds of conservation practices that are needed. For percent slope, slope length, and knoll erodibility, donor points are not found by the random imputation methods, but by an algorithm that chooses the donor from either the 1982 or the 1992 value at the same point.

The algorithm used to find donors for the percent slope, slope length, and knoll erodibility is as follows. The donor point is always either the 1982 or the 1992 point associated with the 1987 observation at the same sample point. If the 1982 land use does not equal the 1992 land use and the 1987 land use is equal to either the 1982 land use or the 1992 land use, then the donor is chosen by matching to the 1987 land use. If the 1987 land use does not equal either the 1982 land use or the
1992 land use, then the donor point is chosen by assigning a probability of .5 to the 
1982 point and a probability of .5 to the 1992 point. If the 1982 land use equals the 
1992 land use, then the donor point is found assigning a probability of .5 to the 1982 
point and a probability of .5 to the 1992 point.

For the random imputation within imputation classes method, classes are 
formed using the 1982 C-factor, the 1987 land use, and the counties. Table 6.14 
gives the eleven land use groups used in the imputation. For Kansas, each 
aricultural group is divided into two groups of the basis of whether irrigation has 
been used. Table 6.15 gives the cells found on the basis of the C-factor within the 
land use groups. The method for assigning the donor points to the missing values is 
similar to the "hot deck" method of imputation and assumes that the observations 
are in random order. We believe that the observations in the actual data set will be 
in close to random order.

The second imputation method is a hierarchical method. The donor points 
for the missing data points are found by randomly choosing respondents within cells. 
If no match can be found to the missing point in a cell, then the cell is enlarged. As 
with the random imputation within cells method, we use a technique similar to the 
"hot deck" imputation method for choosing the donors and assigning them to the 
missing values. We believe the observations will be close to random order in the 
actual data set. A respondent can be used up to two times as a donor at the first 
five levels in the imputation of the actual data set. For our tests described in 
Section 6.7, a respondent can be used as a donor only one time in the first five 
levels. For the two largest cells there is no limit on how many times a given donor 
can be used. The cells are based on the MLRAC or county, the land use, and the 
land capability classes. The land capability classes are grouped. The land
<table>
<thead>
<tr>
<th>Group</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Horticultural crops, including fruit, nut, vineyard, bush fruit, berries</td>
</tr>
<tr>
<td>2</td>
<td>Row crops, including corn, sorghum, soybeans, cotton, peanuts, tobacco, sugar beets, potatoes, vegetable and truck crops, sunflowers</td>
</tr>
<tr>
<td>3</td>
<td>Close grown crops, including wheat, oats, rice, barley, flax</td>
</tr>
<tr>
<td>4</td>
<td>Hayland, including cool season grasses, warm season grasses, legumes, legume-grasses</td>
</tr>
<tr>
<td>5</td>
<td>Other cropland, including summer fallow, aquaculture in a crop rotation, not planted cropland</td>
</tr>
<tr>
<td>6</td>
<td>Pasture, including cool season grasses, warm season grasses, legumes, legume-grass mixed, grass-forbes mixed, grass-forbes-legumes mixed</td>
</tr>
<tr>
<td>7</td>
<td>Rangeland</td>
</tr>
<tr>
<td>8</td>
<td>Forest land, grazed and ungrazed</td>
</tr>
<tr>
<td>9</td>
<td>Other land in farms, including farmsteads and ranch headquarters, windbreaks, commercial feedlots, greenhouses, nurseries, broiler facilities</td>
</tr>
<tr>
<td>10</td>
<td>Barren land, including salt flats, bare rock, strip mines, quarries, gravel pits, borrow pits, beaches, sand dunes, mud flats, river wash, oil wasteland</td>
</tr>
<tr>
<td>11</td>
<td>Other lands, including permanent snow and ice, marshland</td>
</tr>
</tbody>
</table>
Table 6.15. Imputation cells excluding the counties

<table>
<thead>
<tr>
<th>Imp. Cell</th>
<th>Land Use Gr.</th>
<th>C-factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>all</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>&lt;.1</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>≥.1 &amp; &lt;.2</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>≥.2 &amp; &lt;.3</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>≥.3 &amp; &lt;.36</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
<td>≥.36 &amp; &lt;.4</td>
</tr>
<tr>
<td>7</td>
<td>2</td>
<td>≥.4</td>
</tr>
<tr>
<td>8</td>
<td>3</td>
<td>all</td>
</tr>
<tr>
<td>9</td>
<td>4</td>
<td>&lt;.05</td>
</tr>
<tr>
<td>10</td>
<td>4</td>
<td>≥.05</td>
</tr>
<tr>
<td>11</td>
<td>5</td>
<td>all</td>
</tr>
<tr>
<td>12</td>
<td>6</td>
<td>&lt;.005</td>
</tr>
<tr>
<td>13</td>
<td>6</td>
<td>≥.005 &amp; &lt;.015</td>
</tr>
<tr>
<td>14</td>
<td>6</td>
<td>≥.015 &amp; &lt;.1</td>
</tr>
<tr>
<td>15</td>
<td>6</td>
<td>≥.1</td>
</tr>
<tr>
<td>16</td>
<td>7</td>
<td>all</td>
</tr>
<tr>
<td>17</td>
<td>8</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>18</td>
<td>8</td>
<td>≥.001 &amp; ≤.0011</td>
</tr>
<tr>
<td>19</td>
<td>8</td>
<td>&gt;.0011 &amp; ≤.004</td>
</tr>
<tr>
<td>20</td>
<td>8</td>
<td>&gt;.004</td>
</tr>
<tr>
<td>21, 22, 23</td>
<td>9, 10, 11</td>
<td>all</td>
</tr>
</tbody>
</table>
Table 6.16. Land capability class groups

<table>
<thead>
<tr>
<th>Group</th>
<th>Land Capability Classes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1, 2W, 2S</td>
</tr>
<tr>
<td>2</td>
<td>2E</td>
</tr>
<tr>
<td>3</td>
<td>3E</td>
</tr>
<tr>
<td>4</td>
<td>2C, 3C</td>
</tr>
<tr>
<td>5</td>
<td>3W, 3S, 4W</td>
</tr>
<tr>
<td>6</td>
<td>4E, 4S, 4C</td>
</tr>
<tr>
<td>7</td>
<td>5E, 5W, 5S, 5C</td>
</tr>
<tr>
<td>8</td>
<td>6E, 6W, 6S, 6C</td>
</tr>
<tr>
<td>9</td>
<td>7E, 7W, 7S, 7C</td>
</tr>
<tr>
<td>10</td>
<td>8E, 8W, 8S, 8C</td>
</tr>
</tbody>
</table>

Note: The E stands for problems with the soil due to erosion, the W stands for problems with the soil due to wetness, the S stands for problems with the soil due to problems in root development, the C stands for problems with the climate. The order of the problems goes from E, as the most difficult, to W, to S, to C. A point is always classified in the most difficult applicable class. The numbers indicate the quality of the soil, going from 1, the highest quality, to 8, the lowest quality.
Table 6.17. Levels for the strictly hierarchical imputation method

<table>
<thead>
<tr>
<th>Level</th>
<th>Description</th>
</tr>
</thead>
</table>
| 1     | same MLRAC  
      | same exact 1987 Land Use  
      | same Land Capability Class group |
| 2     | same MLRAC  
      | same exact 1987 Land Use |
| 3     | same county  
      | same exact 1987 Land Use  
      | same Land Capability Class group |
| 4     | same county  
      | same exact 1987 Land Use |
| 5     | same county  
      | same 1987 Land Use group from Table 14 |
| 6     | can use a point more than twice at this level  
      | same county  
      | same exact 1987 Land Use |
| 7     | can use a point more than twice at this level  
      | same county  
      | same 1987 Land Use Group from Table 14 |

Note: If no match is found, a message is printed for that point
Table 6.18. Hierarchical levels for hierarchical—distance function matching

<table>
<thead>
<tr>
<th>Level</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>same county</td>
</tr>
<tr>
<td></td>
<td>same exact 1987 Land Use</td>
</tr>
<tr>
<td>2</td>
<td>over 10 nearest counties</td>
</tr>
<tr>
<td></td>
<td>same exact 1987 Land Use</td>
</tr>
<tr>
<td>3</td>
<td>same county</td>
</tr>
<tr>
<td></td>
<td>same 1987 Land Use group from Table 14</td>
</tr>
<tr>
<td>4</td>
<td>over 10 nearest counties</td>
</tr>
<tr>
<td></td>
<td>same 1987 Land Use group from Table 14</td>
</tr>
</tbody>
</table>

Note: If no match is found a message is printed for that point.

capability class groups are given in Table 6.16. The levels of the hierarchical cells are given in Table 6.17.

For the hierarchical method with a distance matching function, a search is made through the hierarchical cell to find the closest match to the missing point in terms of distances based on the 1982 C—factor and the 1982 L—factor. Two distances are found, the closest point to the point to be imputed out of the the points that have not been used as donors in the cell and the closest point to the point to be imputed out of all of the points within the cell. The distance between points i and j is
where

\[ d_{ij} = \left[ k_C \left( C_i - C_j \right)^2 + k_L \left( L_i - L_j \right)^2 \right]^{\frac{1}{5}}, \]

\[ C_i = 1982 \text{ C-factor for } i^{th} \text{ point} \]
\[ C_j = 1982 \text{ C-factor for } j^{th} \text{ point} \]
\[ L_i = 1982 \text{ L-factor for } i^{th} \text{ point} \]
\[ L_j = 1982 \text{ L-factor for } j^{th} \text{ point} \]
\[ k_C = \text{reciprocal of twice the variance of 1982 C-factors} \]
\[ k_L = \text{reciprocal of twice the variance of 1982 L-factors}. \]

The L-factor is not measured for all locations. If the L-factor is measured for the points, then, if the distance for points that have not been used is less than four and the difference between the two distances is less than two, the nearest point in terms of the first distance is used as a donor. If the first distance is greater than four or if the first distance is less than four but the difference between the first distance and the second distance is greater than two, then the point associated with the second distance is used if the second distance is less than four. If the L-factor is not measured for the points, then \( L_i \) and \( L_j \) are set to zero in the formula for \( d_{ij} \) and the comparison values are divided by two. If no match is found, the search continues at the next higher hierarchical cell. The hierarchical cells are given in Table 6.18.

6.6. Evaluation of the imputation methods using the data sets

We evaluated the three imputation methods using the data sets from Indiana and Kansas. In order to evaluate the imputation methods, each observation in each of the data sets was treated as missing and a donor point was selected from the
remaining points in the data set using the imputation methods. Since no 1992 data are available for use in testing the imputation methods, the methods were tested using the 1982 data only. For the strictly hierarchical imputation method, in the simulation a donor could be used only once at levels one through five. Four data sets were created for Indiana, one with the real values and three with imputed values, imputed by the three methods. Three data sets were created for Kansas, one with the real values and two with imputed values found by the second and third methods.

To evaluate how well the imputation methods worked, the means of the differences and the variances of the differences between the real and the imputed values were found. Zero–one variables were created for the categorical variables (the cropping history, the conservation practices, and the kinds of treatment needed). The means and variances of the differences between the 0–1 variables for real values and the 0–1 values for the imputed values were found. All three of the imputation methods appeared to give unbiased results in that the means for the differences are not significantly different from zero.

We estimated the variance of the imputed sample mean, relative to the variance of the sample mean for a full data set. Using the theory in Chapter 5, if we assume that the number missing is exactly two times the number observed in each imputation class, then, assuming all weights equal $n^{-1}$,

$$V(\bar{y}_I) = V(\bar{y}_F) + \sum_{h=1}^{H} \left( n^{-1}n_h \right) 2\left( n^{-1}\sigma_h^2 \right),$$  \hspace{1cm} (6.1)

where $\bar{y}_F$ is the mean of the full sample, $\bar{y}_I$ is the mean of the imputed sample, the $h$ index is for the imputation classes, $\sigma_h^2$ is the superpopulation variance of
the Y's in imputation class h, \( n_h \) is the full sample size in imputation class h, and \( n \) is the full sample size. Under the independence of the Y's within imputation classes,

\[
2 \sigma_h^2 = \mathbb{E}[(Y_{hi} - Y_{hi}^*)^2],
\]

(6.2)

where \( Y_{hi} \) is the real value for observation \( hi \) and \( Y_{hi}^* \) is the imputed value for observation \( hi \). It follows that, under the model,

\[
s^2_{Y-Y^*} = \sum_{h=1}^{H} \sum_{i=1}^{n_h} (n-1)^{-1} [(Y_{hi} - Y_{hi}^*) - (\bar{y}_F - \bar{y})]^2
\]

(6.3)

is a good estimator of \( \sum_{h=1}^{H} n^{-1}_h n_2 \sigma_h^2 \).

To estimate the inflation in variance due to imputation, we computed

\[
\frac{\text{V(1/3 obs, 2/3 imp)}}{\text{V(all obs)}} = \frac{s^2_Y + s^2_{Y-Y^*}}{s^2_Y},
\]

(6.4)

where

\[
s^2_Y = \sum_{h=1}^{H} \sum_{i=1}^{n_h} (n-1)^{-1} (Y_{hi} - \bar{y}_F)^2
\]

(6.5)

is the sample variance of the Y's in the full sample. If the sample were a random sample \( n^{-1} s^2_Y \) would be an unbiased estimator of \( \text{V}(\bar{y}_F) \). Therefore, expression (6.4) is a reasonable estimator of the ratio of the variance of the imputed mean to the variance of the mean for the full sample. We do not expect to see variance
ratios greater than three. A variance ratio equal to three would occur if there were only one imputation class. The actual sampling design in Indiana and Kansas is a stratified cluster sample.

For the hierarchical—distance function method in Kansas, 94% of the imputed values were found at the first hierarchical level. The hierarchical distance function method in Indiana and the strictly hierarchical method for Indiana and Kansas should give comparable percentages. Therefore, the random imputation model from Chapter 5 furnishes a reasonable approximation for the strictly hierarchical method. For the hierarchical—distance function matching method, smaller variation would be expected within the imputation classes than with the random imputation method.

In Tables 6.19, 6.20, and 6.21, we present the results of our experiment. We can see from Table 6.19 that the deterministic algorithm for imputing the slope length, percent slope, and knoll erodibility works well. From Table 6.20, in Indiana all three methods work well for the C—factor, with the methods that use the C—factor in the imputation working a little better. For the cropping history, in Indiana, the strictly hierarchical method works best, followed by the hierarchical method with distance function matching. For the conservation practices, the methods are all comparable.

For Kansas, we can see from Table 6.21 that the hierarchical—distance function matching method is a little superior to the strictly hierarchical method for the C—factor. The K—factors, the V—factors, the cropping histories, and the conservation practices show comparable results for the two methods. For the L—factors, the hierarchical—distance function matching method is greatly superior.

The results from Indiana and Kansas indicate that the second and third
imputation methods are comparable for all of the variables, except for the L—factors, in terms of variance. The hierarchical—distance function matching method performs better for the L—factors. The random imputation within imputation classes method based on the land use groups, the 1982 C—factors, and the 1982 L—factors does not perform as well as the hierarchical distance function matching method, also based on the land uses, the 1982 C—factors, and the 1982 L—factors, partially because, for the hierarchical—distance function matching method, the land use match is based on the exact land use. Also the hierarchical distance function matching method should find a closer match to the missing value than the random imputation method.
Table 6.19. Results for Indiana and Kansas for % slope, slope length, and knoll erodibility

<table>
<thead>
<tr>
<th>Variable</th>
<th>$\bar{y}$</th>
<th>$s_y^2$</th>
<th>$\bar{y}-\bar{y}^*$</th>
<th>$s_{y-y}^2$</th>
<th>$\frac{V\left{2 \text{ imp., } \frac{1}{3} \text{ obs.}\right}}{V(\text{all obs.})}$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Indiana</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Slope length</td>
<td>181</td>
<td>11,813</td>
<td>-2.7</td>
<td>1137</td>
<td>1.10</td>
</tr>
<tr>
<td>% slope</td>
<td>8.4</td>
<td>179.05</td>
<td>.01</td>
<td>.08</td>
<td>1.00</td>
</tr>
<tr>
<td><strong>Kansas</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Slope length</td>
<td>410</td>
<td>103561</td>
<td>-10</td>
<td>11378</td>
<td>1.11</td>
</tr>
<tr>
<td>% slope</td>
<td>2.2</td>
<td>5.335</td>
<td>.001</td>
<td>.040</td>
<td>1.01</td>
</tr>
<tr>
<td>Knoll Erod.</td>
<td>.10</td>
<td>.4930</td>
<td>0</td>
<td>0</td>
<td>1.00</td>
</tr>
</tbody>
</table>
Table 6.20. Results from imputation for Indiana data

<table>
<thead>
<tr>
<th>Variable</th>
<th>$\bar{y}$</th>
<th>$s^2_Y$</th>
<th>Imp meth</th>
<th>$\bar{y}-\bar{y}^*$</th>
<th>$\frac{\text{V(2/3 imp, 1/3 obs)}}{\text{V(all obs)}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>C-factor</td>
<td>.16</td>
<td>.0214</td>
<td>1</td>
<td>0</td>
<td>.0069</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2</td>
<td>0</td>
<td>.0084</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>-.001</td>
<td>.0068</td>
</tr>
<tr>
<td>Crop. Hist. 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Group</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1—corn</td>
<td>.29</td>
<td>.2064</td>
<td>1</td>
<td>.004</td>
<td>.1839</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2</td>
<td>0</td>
<td>.1163</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>-.006</td>
<td>.1332</td>
</tr>
<tr>
<td>2—soybeans</td>
<td>.24</td>
<td>.1826</td>
<td>1</td>
<td>0</td>
<td>.1726</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2</td>
<td>0</td>
<td>.1313</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>0</td>
<td>.1426</td>
</tr>
<tr>
<td>3—hayland</td>
<td>.03</td>
<td>.0291</td>
<td>1</td>
<td>0</td>
<td>.0188</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2</td>
<td>.002</td>
<td>.0206</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>.002</td>
<td>.0206</td>
</tr>
<tr>
<td>4—pasture</td>
<td>.15</td>
<td>.1263</td>
<td>1</td>
<td>.002</td>
<td>.0281</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2</td>
<td>0</td>
<td>.0263</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>.004</td>
<td>.0263</td>
</tr>
<tr>
<td>5—forest</td>
<td>.25</td>
<td>.1864</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Crop. Hist. 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Group</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1—corn</td>
<td>.27</td>
<td>.1973</td>
<td>1</td>
<td>.002</td>
<td>.1745</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2</td>
<td>0</td>
<td>.1163</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>0</td>
<td>.1426</td>
</tr>
</tbody>
</table>
Table 6.20. Results from imputation for Indiana data (cont’d)

<table>
<thead>
<tr>
<th>Variable</th>
<th>$\bar{y}$</th>
<th>$s^2_Y$</th>
<th>Imp meth</th>
<th>$\bar{y} - \bar{y}^*$</th>
<th>$s^2_{Y-Y}^*$</th>
<th>$V\left[\frac{2}{3} \text{ imp., } \frac{1}{3} \text{ obs.}\right] / V(\text{all obs.})$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>.004</td>
<td>.1613</td>
<td>1.85</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2</td>
<td>0</td>
<td>.0976</td>
<td>1.52</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>-.006</td>
<td>.1332</td>
<td>1.70</td>
</tr>
<tr>
<td>2-soybeans</td>
<td>.25</td>
<td>.1893</td>
<td>1</td>
<td>0</td>
<td>.0300</td>
<td>2.03</td>
</tr>
<tr>
<td>2-soybeans</td>
<td></td>
<td></td>
<td>2</td>
<td>.0225</td>
<td>.0319</td>
<td>2.10</td>
</tr>
<tr>
<td>2-soybeans</td>
<td></td>
<td></td>
<td>3</td>
<td>.0225</td>
<td>.0319</td>
<td>2.10</td>
</tr>
<tr>
<td>3-hayland</td>
<td>.03</td>
<td>.0291</td>
<td>1</td>
<td>0</td>
<td>.0263</td>
<td>1.22</td>
</tr>
<tr>
<td>3-hayland</td>
<td></td>
<td></td>
<td>2</td>
<td>.0225</td>
<td>.0319</td>
<td>2.10</td>
</tr>
<tr>
<td>3-hayland</td>
<td></td>
<td></td>
<td>3</td>
<td>.0225</td>
<td>.0319</td>
<td>2.10</td>
</tr>
<tr>
<td>4-pasture</td>
<td>.14</td>
<td>.1196</td>
<td>1</td>
<td>0</td>
<td>.0263</td>
<td>1.22</td>
</tr>
<tr>
<td>4-pasture</td>
<td></td>
<td></td>
<td>2</td>
<td>.0225</td>
<td>.0319</td>
<td>2.10</td>
</tr>
<tr>
<td>4-pasture</td>
<td></td>
<td></td>
<td>3</td>
<td>.0225</td>
<td>.0319</td>
<td>2.10</td>
</tr>
<tr>
<td>5-forest</td>
<td>.25</td>
<td>.1864</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1.00</td>
</tr>
<tr>
<td>5-forest</td>
<td></td>
<td></td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>1.00</td>
</tr>
<tr>
<td>5-forest</td>
<td></td>
<td></td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>.006</td>
<td>.2120</td>
<td>1.99</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2</td>
<td>0</td>
<td>.1614</td>
<td>1.76</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>-.006</td>
<td>.1895</td>
<td>1.89</td>
</tr>
<tr>
<td>2-soybeans</td>
<td>.24</td>
<td>.1836</td>
<td>1</td>
<td>.002</td>
<td>.1857</td>
<td>2.01</td>
</tr>
<tr>
<td>2-soybeans</td>
<td></td>
<td></td>
<td>2</td>
<td>.002</td>
<td>.1332</td>
<td>1.73</td>
</tr>
<tr>
<td>2-soybeans</td>
<td></td>
<td></td>
<td>3</td>
<td>.006</td>
<td>.1557</td>
<td>1.85</td>
</tr>
<tr>
<td>3-hayland</td>
<td>.02</td>
<td>.0238</td>
<td>1</td>
<td>0</td>
<td>.0263</td>
<td>2.11</td>
</tr>
<tr>
<td>3-hayland</td>
<td></td>
<td></td>
<td>2</td>
<td>.0225</td>
<td>.0281</td>
<td>2.18</td>
</tr>
<tr>
<td>3-hayland</td>
<td></td>
<td></td>
<td>3</td>
<td>0</td>
<td>.0225</td>
<td>1.95</td>
</tr>
<tr>
<td>4-pasture</td>
<td>.13</td>
<td>.1141</td>
<td>1</td>
<td>0</td>
<td>.0225</td>
<td>1.20</td>
</tr>
<tr>
<td>4-pasture</td>
<td></td>
<td></td>
<td>2</td>
<td>.0225</td>
<td>.0225</td>
<td>1.20</td>
</tr>
<tr>
<td>4-pasture</td>
<td></td>
<td></td>
<td>3</td>
<td>.004</td>
<td>.0263</td>
<td>1.23</td>
</tr>
</tbody>
</table>
Table 6.20. Results from imputation for Indiana data (cont’d)

<table>
<thead>
<tr>
<th>Variable</th>
<th>( \bar{y} )</th>
<th>( s^2_Y )</th>
<th>Imp meth</th>
<th>( \bar{y}<em>\text{imp} - \bar{y}</em>\text{obs} )</th>
<th>( s^2_{Y - Y} )</th>
<th>( V\left( \frac{2}{3} \text{ imp., } \frac{1}{3} \text{ obs.} \right) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>5—forest</td>
<td>.25</td>
<td>.1874</td>
<td>1</td>
<td>0</td>
<td>.0038</td>
<td>1.02</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2</td>
<td>0</td>
<td></td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>0</td>
<td>.0038</td>
<td>1.02</td>
</tr>
<tr>
<td>Sum Con. Prac.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Group</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1—no practices</td>
<td>.52</td>
<td>.2501</td>
<td>1</td>
<td>.006</td>
<td>.2720</td>
<td>2.09</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2</td>
<td>.002</td>
<td>.2795</td>
<td>2.12</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>.006</td>
<td>.2983</td>
<td>2.19</td>
</tr>
<tr>
<td>2—con. till.</td>
<td>.23</td>
<td>.1766</td>
<td>1</td>
<td>0</td>
<td>.1726</td>
<td>1.98</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2</td>
<td>0</td>
<td>.2026</td>
<td>2.15</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>0</td>
<td>.1538</td>
<td>1.87</td>
</tr>
<tr>
<td>3—drainage</td>
<td>.08</td>
<td>.0726</td>
<td>1</td>
<td>.002</td>
<td>.1144</td>
<td>1.42</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2</td>
<td>0</td>
<td>.0938</td>
<td>2.29</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>0</td>
<td>.1126</td>
<td>2.55</td>
</tr>
<tr>
<td>4—both con. till. &amp; dr.</td>
<td>.08</td>
<td>.0726</td>
<td>1</td>
<td>0</td>
<td>.0976</td>
<td>2.34</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2</td>
<td>0</td>
<td>.0863</td>
<td>2.19</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>0</td>
<td>.0901</td>
<td>2.24</td>
</tr>
<tr>
<td>5—other</td>
<td>.10</td>
<td>.0865</td>
<td>1</td>
<td>-.007</td>
<td>.1238</td>
<td>2.43</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2</td>
<td>-.002</td>
<td>.1257</td>
<td>2.45</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>-.006</td>
<td>.1294</td>
<td>2.50</td>
</tr>
</tbody>
</table>
Table 6.21. Results from imputation for Kansas data

<table>
<thead>
<tr>
<th>Variable</th>
<th>$\bar{Y}$</th>
<th>$s_Y^2$</th>
<th>Imp meth</th>
<th>$\bar{Y} - \bar{Y}^*$</th>
<th>$s_{Y-Y}^2$</th>
<th>$V\left[\frac{2}{3}\text{ imp., } \frac{1}{3}\text{ obs.}\right] V(\text{all obs.})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>C-factor</td>
<td>.12</td>
<td>.0082</td>
<td>2</td>
<td>-.003</td>
<td>.0030</td>
<td>1.36</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>.0001</td>
<td>.0020</td>
<td>1.25</td>
</tr>
<tr>
<td>Crop. Hist. 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Group</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1—sorghum</td>
<td>.09</td>
<td>.0810</td>
<td>2</td>
<td>-.003</td>
<td>.0698</td>
<td>1.86</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>-.0009</td>
<td>.0607</td>
<td>1.75</td>
</tr>
<tr>
<td>2—wheat</td>
<td>.40</td>
<td>.2399</td>
<td>2</td>
<td>.0009</td>
<td>.1115</td>
<td>1.46</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>.002</td>
<td>.1024</td>
<td>1.44</td>
</tr>
<tr>
<td>3—sum. fall.</td>
<td>.08</td>
<td>.0742</td>
<td>2</td>
<td>0</td>
<td>.0435</td>
<td>1.59</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>-.18</td>
<td>.0526</td>
<td>1.71</td>
</tr>
<tr>
<td>4—rangeland</td>
<td>.30</td>
<td>.2105</td>
<td>2</td>
<td>0</td>
<td>.0091</td>
<td>1.04</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>0</td>
<td>.0090</td>
<td>1.04</td>
</tr>
<tr>
<td>Crop. Hist. 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Group</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1—sorghum</td>
<td>.08</td>
<td>.0772</td>
<td>2</td>
<td>0</td>
<td>.0870</td>
<td>2.13</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>.009</td>
<td>.0907</td>
<td>2.17</td>
</tr>
<tr>
<td>2—wheat</td>
<td>.43</td>
<td>.2446</td>
<td>2</td>
<td>-.002</td>
<td>.1378</td>
<td>1.56</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>-.0009</td>
<td>.1514</td>
<td>1.62</td>
</tr>
<tr>
<td>3—sum. fall.</td>
<td>.07</td>
<td>.0649</td>
<td>2</td>
<td>.002</td>
<td>.0544</td>
<td>1.84</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>0</td>
<td>.0617</td>
<td>1.95</td>
</tr>
<tr>
<td>4—rangeland</td>
<td>.31</td>
<td>.2126</td>
<td>2</td>
<td>0</td>
<td>.0218</td>
<td>1.10</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>0</td>
<td>.0181</td>
<td>1.09</td>
</tr>
</tbody>
</table>
Table 6.21. Results from imputation for Kansas data (cont’d)

<table>
<thead>
<tr>
<th>Variable</th>
<th>$\bar{y}$</th>
<th>$s_Y^2$</th>
<th>Imp meth</th>
<th>$\bar{y}-\bar{y}^*$</th>
<th>$s_Y^2-Y^*$</th>
<th>V(\frac{2}{3} \text{ imp., } \frac{1}{3} \text{ obs.})</th>
<th>V(\text{ all obs.})</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Crop. Hist. 3</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Group</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1—sorghum</td>
<td>.08</td>
<td>.0765</td>
<td>2</td>
<td>-.002</td>
<td>.0816</td>
<td>2.07</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>.003</td>
<td>.0789</td>
<td>2.03</td>
<td></td>
</tr>
<tr>
<td>2—wheat</td>
<td>.41</td>
<td>.2415</td>
<td>2</td>
<td>.0009</td>
<td>.1442</td>
<td>1.60</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>0</td>
<td>.1487</td>
<td>1.62</td>
<td></td>
</tr>
<tr>
<td>3—sum. fall.</td>
<td>.10</td>
<td>.0876</td>
<td>2</td>
<td>-.0009</td>
<td>.0571</td>
<td>1.62</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>-.002</td>
<td>.0707</td>
<td>1.81</td>
<td></td>
</tr>
<tr>
<td>4—rangeland</td>
<td>.31</td>
<td>.2133</td>
<td>2</td>
<td>0</td>
<td>.0218</td>
<td>1.10</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>0</td>
<td>.0181</td>
<td>1.08</td>
<td></td>
</tr>
<tr>
<td><strong>Sum Con. Prac.</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Group</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1—no practices</td>
<td>.39</td>
<td>.2372</td>
<td>2</td>
<td>0</td>
<td>.2339</td>
<td>1.99</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>.004</td>
<td>.2538</td>
<td>2.07</td>
<td></td>
</tr>
<tr>
<td>2—con. till.</td>
<td>.20</td>
<td>.1575</td>
<td>2</td>
<td>.0009</td>
<td>.0988</td>
<td>1.63</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>-.002</td>
<td>.1070</td>
<td>1.68</td>
<td></td>
</tr>
<tr>
<td>3—c.f. &amp; ter.</td>
<td>.08</td>
<td>.0719</td>
<td>2</td>
<td>-.008</td>
<td>.1042</td>
<td>2.45</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>-.008</td>
<td>.0861</td>
<td>2.20</td>
<td></td>
</tr>
<tr>
<td>4—prop. graze</td>
<td>.18</td>
<td>.1479</td>
<td>2</td>
<td>.0009</td>
<td>.0879</td>
<td>1.59</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>0</td>
<td>.0888</td>
<td>1.60</td>
<td></td>
</tr>
<tr>
<td>5—others</td>
<td>.16</td>
<td>.1347</td>
<td>2</td>
<td>.006</td>
<td>.1604</td>
<td>2.19</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>.006</td>
<td>.1786</td>
<td>2.33</td>
<td></td>
</tr>
</tbody>
</table>
Table 6.21. Results from imputation for Kansas data (cont'd)

<table>
<thead>
<tr>
<th>Variable</th>
<th>$\bar{y}$</th>
<th>$s_Y^2$</th>
<th>Imp. Method</th>
<th>$\bar{y} - \bar{y}^*$</th>
<th>$s_{Y-Y}^2$</th>
<th>$\frac{V[\frac{2}{3} \text{ imp., } \frac{1}{3} \text{ obs.}]}{V(\text{all obs.})}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Len. Rot.</td>
<td>.72</td>
<td>1.741</td>
<td>2</td>
<td>.0009</td>
<td>1.310</td>
<td>1.75</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>.02</td>
<td>1.601</td>
<td>1.92</td>
</tr>
<tr>
<td>K-factor</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1987</td>
<td>.85</td>
<td>.0205</td>
<td>2</td>
<td>-.0002</td>
<td>.0105</td>
<td>1.51</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>.0005</td>
<td>.0098</td>
<td>1.48</td>
</tr>
<tr>
<td>1986</td>
<td>.85</td>
<td>.0215</td>
<td>2</td>
<td>.0002</td>
<td>.0110</td>
<td>1.51</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>.0009</td>
<td>.0112</td>
<td>1.52</td>
</tr>
<tr>
<td>1985</td>
<td>.85</td>
<td>.0215</td>
<td>2</td>
<td>.0002</td>
<td>.0139</td>
<td>1.65</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>.0000</td>
<td>.0129</td>
<td>1.60</td>
</tr>
<tr>
<td>1984</td>
<td>.85</td>
<td>.0223</td>
<td>2</td>
<td>.0002</td>
<td>.0127</td>
<td>1.57</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>.0006</td>
<td>.0130</td>
<td>1.58</td>
</tr>
<tr>
<td>L-factor</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1987</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1969</td>
<td>4756120</td>
<td>2</td>
<td>-2.5</td>
<td>4512958</td>
<td>1.95</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>17.5</td>
<td>1697017</td>
<td>1.36</td>
</tr>
<tr>
<td>1986</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1971</td>
<td>4745885</td>
<td>2</td>
<td>-2.5</td>
<td>4529563</td>
<td>1.95</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>17.5</td>
<td>1708144</td>
<td>1.36</td>
</tr>
<tr>
<td>1985</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1965</td>
<td>4755499</td>
<td>2</td>
<td>-2.5</td>
<td>4510168</td>
<td>1.95</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>18.4</td>
<td>1704854</td>
<td>1.36</td>
</tr>
<tr>
<td>1984</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1952</td>
<td>4665010</td>
<td>2</td>
<td>-2.5</td>
<td>4469157</td>
<td>1.96</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>19.5</td>
<td>1707770</td>
<td>1.37</td>
</tr>
</tbody>
</table>
Table 6.21. Results from imputation for Kansas data (cont'd)

<table>
<thead>
<tr>
<th>Variable</th>
<th>$\bar{y}$</th>
<th>$s_y^2$</th>
<th>Imp meth</th>
<th>$\bar{y} - \bar{y}^*$</th>
<th>$s_{Y-Y^*}^2$</th>
<th>$V{\left[\frac{2}{3}\text{ imp.}, \frac{1}{3}\text{ obs.}\right]}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>V-factor</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1987</td>
<td>1686</td>
<td>1688908</td>
<td>2</td>
<td>.3</td>
<td>1004757</td>
<td>1.59</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>-.2</td>
<td>918040</td>
<td>1.54</td>
</tr>
<tr>
<td>1986</td>
<td>1723</td>
<td>1665576</td>
<td>2</td>
<td>1.1</td>
<td>1062962</td>
<td>1.64</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>-2.7</td>
<td>969836</td>
<td>1.58</td>
</tr>
<tr>
<td>1985</td>
<td>1687</td>
<td>1674230</td>
<td>2</td>
<td>1.5</td>
<td>1121886</td>
<td>1.67</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>.5</td>
<td>1103835</td>
<td>1.66</td>
</tr>
<tr>
<td>1984</td>
<td>1724</td>
<td>1672625</td>
<td>2</td>
<td>.8</td>
<td>1101742</td>
<td>1.66</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>-2.4</td>
<td>1023463</td>
<td>1.61</td>
</tr>
</tbody>
</table>
REFERENCES


I would like to thank Dr. Wayne A. Fuller for the direction and advice that he has given me over the past three years. I would also like to thank Helen Nelson and Judy Shafer for the secretarial and typing help that they have given me in preparing this paper. I would like to give credit to Joe Croos for writing the programs that actually do the imputations described in Chapter 6 and to Savas Papadopoulos for helping with the reduction of the data from Joe Croos's programs. Lastly, I would like to thank my husband, Clayton Conard, for the patience he has shown and the financial and emotional support he has given me as I have pursued my education here at Iowa State University.

Margot Tollefson
Iowa State University
July 15, 1992