Compensating for Missing Survey Data

Graham Kalton

Survey Research Center
Institute for Social Research
The University of Michigan
The Research Report Series of the Institute for Social Research is composed of significant reports published at the completion of a research project. These reports are generally prepared by the principal research investigators and are directed to selected users of this information. Research Reports are intended as technical documents that provide rapid dissemination of new knowledge resulting from ISR research.

Compensating for Missing Survey Data

Graham Kalton

Survey Research Center
Institute for Social Research
The University of Michigan

1983
TABLE OF CONTENTS

Summary .................................................. 1

1. Nonresponse in Surveys .............................. 3
   1.1 Introduction ....................................... 3
   1.2 Types of Nonresponse ............................. 4
   1.3 Nonresponse Bias ................................... 6
   1.4 Nonresponse in the 1978 ISDP Research Panel ......... 10

2. General Issues for Missing Data Compensation Procedures ................. 14
   2.1 Assumptions ....................................... 14
   2.2 Compensation Procedures ......................... 17
   2.3 Three General Criteria in the Choice of Compensation Procedures ..... 20
      2.3.1 Precision of Estimators ..................... 21
      2.3.2 Standard Error Estimation .................. 28
      2.3.3 General Applicability ...................... 32
   2.4 Concluding Remarks .............................. 39

3. Weighting Adjustments .................................... 42
   3.2 Population Weighting Adjustments .................. 43
   3.4 Raking Ratio Estimation ........................... 55
   3.5 Weighting Using Response Probabilities ............. 59
   3.6 Concluding Remarks ................................ 63

4. Imputation ............................................... 65
   4.1 Introduction ....................................... 65
   4.2 Imputation Methods ................................ 67
   4.3 Two Simulation Studies ............................. 80
      4.3.1 Imputation of Hourly Rate of Pay .......... 82
      4.3.2 Imputation of Quarterly Earnings ........... 91
   4.4 Other Imputation Issues ........................... 104

5. Miscellaneous Topics .................................... 112
   5.1 Multiple Imputations .............................. 112
   5.2 Weighted Data ..................................... 123

Appendix .................................................. 131
   Bibliography ........................................... 137
The simulation studies described in Section 4.3 were carried out with Robert L. Santos, who is a joint author of this section of the report. Robert Santos compiled the bibliography on nonresponse and missing data compensation techniques and provided a great deal of valuable assistance in the preparation of the rest of the report. His contributions are greatly appreciated. I should also like to record my thanks to Daniel Kasprzyk who provided many helpful comments on a draft version of the report. In addition I should like to thank Barbara Bailar, Brenda Cox, Brian Greenberg, James Fagan and Frederick Scheuren for drawing my attention to some inaccuracies in the draft version and making other useful comments on it.

This report was made pursuant to contract HEW-100-79-0127 with the Department of Health and Human Services. It was prepared for the Income Survey Development Program by the Survey Development Research Center in Nonresponse and Imputation which was located at the Survey Research Center, University of Michigan.
SUMMARY

This report is concerned with issues of handling missing data in surveys, with a focus on the needs of the 1978 and 1979 Research Panel surveys of the Income Survey Development Program (ISDP).

The report comprises five chapters. The first chapter discusses the types of nonresponse that give rise to missing data in surveys, in particular distinguishing between total (unit) and item nonresponse. It also outlines the bias that nonresponse can cause in survey estimates of means, totals, variances and covariances. A brief account of the total and item nonresponse encountered in the 1978 ISDP Research Panel is provided.

Chapter 2 discusses the assumptions commonly made in compensating for nonresponse, and proposes some criteria for assessing compensation procedures. These criteria include the precision of the resulting survey estimators, the ease of estimating the standard errors of these estimators, and the ability of the procedures to generate reasonable estimators for a range of population parameters. The similarities between two general classes of compensation procedures, weighting adjustments and imputation, are examined.

Weighting adjustments, which are primarily employed to compensate for total nonresponse, are reviewed in Chapter 3. The methods covered include: population weighting adjustments, weighting the sample to some external source such as the Decennial Census (as in post-stratification); sample weighting adjustments, weighting different classes of the sample by the inverses of their response rates; raking ratio estimation; and weighting according to estimated response probabilities.

Chapter 4 describes a range of imputation methods used to assign values for missing item nonresponses. The methods reviewed include deductive imputation, cold-deck and hot-deck imputation, mean-value and random imputation within classes, flexible matching imputation, distance function matching and regression imputation. The results of two simulation studies, using data from the 1978 ISDP Research Panel, are presented to illustrate and compare the properties of some of these
procedures with regard to the estimation of univariate statistics. (A companion report by Santos, 1981a, employs these simulation studies to examine the effects of imputation on multivariate statistics). The chapter also discusses several issues arising with the use of imputation, including the effect of imputation on measures of the relationships between variables, the need for imputed values to satisfy edit checks, problems of handling records with several missing responses, and the choice of control variables for forming imputation classes.

The final chapter, Chapter 5, discusses the use of multiple imputations as advocated by Rubin (1978, 1979b), and the problems of applying missing data compensations to weighted data sets, where the sample units were selected with unequal probabilities. A version of a multiple imputation procedure, termed a Repeated Replication Imputation Procedure, is described in an Appendix by Kish.

A bibliography on nonresponse and nonresponse adjustments is provided at the end of the report.

1. Nonresponse in Surveys

1.1 Introduction

The purpose of this report is to review procedures for handling missing data in surveys, with a focus on the needs of the national panel surveys of the Income Survey Development Program (ISDP).

The ISDP is a program of the Department of Health and Human Services (DHSS). In collaboration with the Bureau of the Census it conducted four major field studies between October, 1977 and August, 1980 - two national household panel surveys and two studies which were confined to a number of local areas. The two national surveys, to which this report relates, are known as the 1978’ and 1979 Research Panels. In both panels persons were interviewed at three month intervals; they were followed to new addresses if they moved. There were five waves of data collection in the 1978 Panel and six in the 1979 Panel. Both panels collected detailed intra-year information on household composition, money and in-kind income, assets, liabilities, program eligibility criteria and program participation, labor force participation, taxes and selected topics of policy interest. The sample size for the 1978 Panel was 2,358 households and that for the 1979 Panel about 11,300 households. Further details on the ISDP and the Research Panels are provided by Ycas and Lininger (1980).

The extensive amount and sensitivity of the data collected in the ISDP Research Panels, along with the longitudinal aspect of their designs, gave rise to a greater degree of missing data than would be encountered in a single one-time survey. In consequence, the issue of compensation for missing data is a special concern for the analysis of the Panels’ data. This issue is being examined by a team from the Survey Research Center at the University of Michigan in conjunction with DHSS. This report is part of that examination.

The report is organized as follows. The next two sections of the present chapter discuss the types of nonresponse that give rise to missing data in surveys and the bias that nonresponse can cause in survey estimates. The chapter concludes with a brief account of the nonresponse encountered
in the 1978 Research Panel. Chapter 2 discusses the assumptions made when compensating for nonresponse, and develops some desirable criteria for compensation procedures. It brings out the similarities and differences between two general classes of compensation procedures: weighting adjustments and imputation. Weighting adjustments, which alter the weights of respondents to compensate for nonrespondents, are then discussed in detail in Chapter 3 and imputation up in Chapter 4. Multiple imputations and compensations when sampled units are selected with unequal probabilities are discussed in Chapter 5.

1.2 Types of Nonresponse

For subsequent discussion it is important to distinguish between three types of nonobservation: noncoverage, unit nonresponse and item nonresponse.

Noncoverage denotes a failure to include some units of the survey population in the sampling frame; in consequence, they have no chance of appearing in the sample. Although not strictly a type of nonresponse, we will loosely classify it as such for convenience of exposition throughout this report. With an area sample design, noncoverage can occur because the defined areal units fail to cover all the survey population or because the listings made within the final stage segments are incomplete.

Unit (or total) nonresponse occurs when no information is collected from a sample unit. It may be caused by a refusal, by a failure to contact the unit (not at home), by the inability of the unit to cooperate (perhaps because of illness or a language barrier), by the unit not being found (for instance, movers in a panel survey), or by completed questionnaires being lost (see Kish, 1965, Section 13.4).

Item nonresponse occurs when the unit cooperates in the survey but fails to provide answers to some of the questions. It may arise because the informant lacks the information necessary to answer the question, including his failure to make the effort required to ascertain the information by retrieving it from his memory or by consulting his records - usually described as "Don't knows"; because he refuses to give an answer, perhaps on the grounds that he finds the question sensitive, embarrassing, or considers it irrelevant to his perception of the survey's objectives; because the interviewer fails to record the answer; or because the response is subsequently rejected at an edit check on the grounds that it is inconsistent with other responses - this category may include an inconsistency arising from a coding or punching error occurring in the transfer of the response to the computer data file.

With regard to compensation for nonresponse, the importance of the distinction between noncoverage, unit and item nonresponse resides in the amount of information available about the nonrespondents. The survey itself can provide no information about units not covered by its sampling frame; if compensations are to be made for noncoverage, they must be based on some external source, such as the Decennial Census. The survey can provide some information about unit nonrespondents, but the amount is usually fairly limited. At the least the sampling unit in which the nonrespondent is located will be known, together with the stratum characteristics of that sampling unit (e.g. region, urbanicity); additional information may be available from interviewers' observations (type of dwelling unit, perhaps race of head of household) and from the sampling frame (for instance, administrative information from the records of Supplemental Security Income - SSI - recipients for the SSI samples in the 1978 and 1979 Research Panels). With item nonresponse the responses to all the items that the unit has answered provide a considerable amount of information about the unit. As will be discussed in subsequent sections, the amount of information available for a particular type of nonresponse has a strong bearing on the form of compensation employed: in brief, noncoverage is commonly treated by some form of post-stratification weighting adjustment, unit nonresponse by some form of post-stratification or response rate weighting

'\It should be noted that in employing an external source for making noncoverage adjustments, consideration needs to be given to the external source's own noncoverage problems."
adjustment, and item nonresponse by some form of imputation procedure.

While the distinction between unit and item nonresponse in terms of the amount of information available on the nonrespondents is a useful broad generalization, it is one of degree and not always clearcut. Sometimes a great deal may be known about unit nonrespondents from the sampling frame, in which case compensations for unit nonresponse may be better made by imputation than by weighting. On the other hand, a respondent may break off an interview at an early stage, or may refuse to answer most of the questions, in which case weighting may be more appropriate than imputation.

1.3 Nonresponse Bias

In most surveys no compensation is made for missing data. This section examines the bias resulting from confining the survey analyses to the available data. For simplicity we will consider a simple random sample of size n drawn from a population of size N, and we will concentrate on a single variable y. With only one variable there is no need to distinguish between the types of nonresponse discussed in the previous section; units not covered by the sampling frame, unit nonrespondents and item nonrespondents will, therefore, be termed simply nonrespondents. We will further assume that, on conceptually repeated applications of the survey under the same essential survey conditions, certain units in the population always respond and the remainder never respond. Thus, the population can be thought of as comprising two strata - respondents and nonrespondents. In reality this division of the population into two strata is an oversimplification because, for some of the units at least, chance plays a part in determining whether they respond or not. The simplified model is attractive, however, because its tractability leads to some informative results. For this reason it will be employed throughout this report. The reader is referred to Platek et al. (1978) for the development of a probability response model for the study of the effects of nonresponse.

Let R be the number of respondents and M be the number of nonrespondents (M for missing) in the population, with R + M = N; the corresponding sample quantities are r and m, with r + m = n. Let R = R/N and M = M/N be the proportions of respondents and nonrespondents in the population and let \( \bar{r} = r/n \) and \( \bar{m} = m/n \) be the response and nonresponse rates in the sample. The population total and mean are given by \( Y = Y_r + Y_m = R\bar{Y}_r + M\bar{Y}_m \) and \( \bar{Y} = R\bar{Y}_r + M\bar{Y}_m \), where \( \bar{Y}_r \) and \( \bar{Y}_m \) are the total and mean for respondents and \( \bar{Y}_m \) and \( \bar{Y}_m \) are the same quantities for the nonrespondents. The corresponding sample quantities are \( y = y_r + y_m = r\bar{y}_r + m\bar{y}_m \) and \( \bar{y} = r\bar{y}_r + m\bar{y}_m \).

If no compensation is made for nonresponse, the respondent sample mean \( \bar{y}_r \) is used to estimate \( \bar{Y} \). Its bias is given by \( B(\bar{y}_r) = E(\bar{y}_r) - \bar{Y} \). The expectation of \( \bar{y}_r \) can be obtained in two stages, first conditional on fixed r and then over different values of r, i.e., \( E(\bar{y}_r) = E_1(E_2(\bar{y}_r) \mid r) \) where \( E_2 \) is the conditional expectation for fixed r and \( E_1 \) is expectation over different values of r. Thus

\[
E(\bar{y}_r) = E_1(E_2(y_{r1})/r) = E_1(\bar{y}_r) = \bar{Y}_r.
\]

Hence the bias of \( \bar{y}_r \) is given by

\[
B(\bar{y}_r) = \bar{Y}_r - \bar{y} = \bar{R}(\bar{Y}_r - \bar{Y}_m).
\] (1.3.1)

Equation (1.3.1) shows that \( \bar{y}_r \) is approximately unbiased for \( \bar{Y} \) if either the proportion of nonrespondents \( \bar{M} \) is small or the mean for nonrespondents, \( \bar{Y}_m \), is close to that for respondents, \( \bar{Y}_r \). Since the survey analyst usually has no direct empirical evidence on the magnitude of \( (\bar{Y}_r - \bar{Y}_m) \), the only situation in which he can have confidence that the bias is small is when the nonresponse rate is low. However, in practice, even with moderate \( \bar{M} \) many survey results escape sizeable biases because \( (\bar{Y}_r - \bar{Y}_m) \) is fortunately often not large.

Consider now the estimation of the population total \( Y \).

With complete response the simple inflation estimator \( F\bar{y} \), where the inflation factor \( F = N/n \) is the inverse of the sampling fraction \( f \), is an unbiased estimator of \( Y \). The application of
this inflation factor to the sample total for respondents gives an estimator \( \hat{Y} = \hat{F}Y_r \), with expectation

\[
E(\hat{Y}) = FE_1Z_2(r\hat{Y}_r) = FE_1(r\hat{Y}_r) = \hat{F}rY_r = Y_r. \tag{1.3.2}
\]

Its bias as an estimator of \( Y \) is thus

\[
B(\hat{Y}) = Y_r - Y = -Y_m = -M\hat{Y}_m.
\]

This bias is small only if the total for the nonrespondents is small. If the \( y \)-variable is a non-negative quantity, as is generally the case with survey variables, \( Y_m \) is small only if \( M \) is small, or the \( Y_{mi} \) values are zero or close to zero.

Next consider the estimation of the difference between two subclass means, \( \hat{Y}_a - \hat{Y}_b \), with the earlier notation extended to \( \hat{Y}_a = \bar{a}_{ra} + \bar{a}_{ma} \) and \( \hat{Y}_b = \bar{b}_{rb} + \bar{b}_{mb} \). Then the bias of \( \hat{Y}_{ra} - \hat{Y}_{rb} \) is

\[
B(\hat{Y}_{ra} - \hat{Y}_{rb}) = B(\hat{Y}_{ra}) - B(\hat{Y}_{rb})
= \bar{a}_{ra}(\bar{\bar{Y}}_a - \bar{\bar{Y}}_{ma}) - \bar{b}_{rb}(\bar{\bar{Y}}_b - \bar{\bar{Y}}_{mb}). \tag{1.3.3}
\]

In favorable circumstances, the biases in the individual estimators \( \hat{Y}_{ra} \) and \( \hat{Y}_{rb} \) may tend to cancel. In particular, this may happen in comparing the means for two rounds of a survey when the rounds have similar response rates.

Finally, consider the effect of missing data on the estimation of variances and covariances. The expectation of the respondent sample variance \( s^2_r \) is

\[
E_1E_2(s^2_r) = E_1(S^2_r) = S^2_r \tag{1.3.4}
\]

where \( S^2_r = \Sigma(Y_{ri} - \bar{Y}_r)^2/(r - 1) \). The bias of \( S^2_r \) as an estimator of \( S^2 = \Sigma(Y_i - \bar{Y})^2/(N - 1) \) is thus

\[
B(S^2_r) = S^2_r - S^2. \tag{1.3.5}
\]

Now

\[
\Sigma(Y_i - \bar{Y})^2 = \Sigma(Y_{ri} - \bar{Y}_r)^2 + \Sigma(Y_{mi} - \bar{Y}_m)^2 + R(\bar{Y}_r - \bar{Y})^2 + M(\bar{Y}_m - \bar{Y})^2
= (R - 1)S^2_r + (M - 1)S^2_m + RM(\bar{Y}_r - \bar{Y}_m)^2.
\]

Thus, approximating \( R - 1 = R, M - 1 = M \) and \( N - 1 = N \),

\[
s^2 = \bar{R}S^2_r + \bar{M}S^2_m + \bar{R}\bar{M}(\bar{Y}_r - \bar{Y}_m)^2.
\]

The bias of \( s^2_r \) is therefore

\[
B(s^2_r) = \bar{M}(S^2_r - S^2_m) - \bar{R}\bar{M}(\bar{Y}_r - \bar{Y}_m)^2. \tag{1.3.6}
\]

The first term of this bias is comparable to the bias for a mean in (1.3.1). The assumption that \( S^2_r \) and \( S^2_m \) are similar may, perhaps, be more realistic than the assumption that \( \bar{Y}_r \) and \( \bar{Y}_m \) are similar. Under this assumption, the first term is negligible. The second term reflects the effect of differences in the respondent and nonrespondent means on the estimator. If respondents and nonrespondents have the same variance \( (S^2_r = S^2_m) \), \( s^2_r \) will underestimate \( s^2 \) unless \( \bar{Y}_r = \bar{Y}_m \).

For the covariance another variable, \( x \), needs to be introduced. For this case respondents will be defined as those who provide both \( x \) and \( y \) values. A "\( \cdot \)" is added to all symbols to draw attention to the fact that they refer to this set of respondents; thus, for instance, \( r' \) is the number of sample units providing both \( x \) and \( y \) values, and \( \bar{Y}_{ra} \) is their sample mean for the \( y \)-variable. If \( x \) values are obtained for all sample units or if \( x \) is missing only for sample units for which \( y \) is missing, \( r' = r \); otherwise \( r' < r \). The expectation of the sample covariance \( s'_{rx} = \Sigma(x'_{ri} - \bar{x}'_r)(y'_{ri} - \bar{Y}_r)/(r' - 1) \) is

\[
E_1E_2(s'_{rx}) = E_1(S'_{rx}) = S'_{rx} \tag{1.3.7}
\]

where

\[
S'_{rx} = \Sigma(x'_{ri} - \bar{x}'_r)(y'_{ri} - \bar{Y}_r)/(r' - 1).
\]

Using the same approach as for the variance,
\[ S_{xy} = \tilde{R}'S_{rxy} + \tilde{R}'S_{mxy} + \tilde{R}'M'(\tilde{x}_r - \tilde{x}_m)(\tilde{y}_r - \tilde{y}_m). \]

Thus the bias of \( s'_{rxy} \) as an estimator of \( S_{xy} \) is

\[ B(s'_{rxy}) = \tilde{R}'(S_{rxy} - S_{mxy}) - \tilde{R}'M'(\tilde{x}_r - \tilde{x}_m)(\tilde{y}_r - \tilde{y}_m). \]  

The interpretation of this bias in the covariance follows that for the variance given above, except that \( s'_{rxy} \) does not necessarily underestimate \( S_{xy} \) if \( S_{rxy} = S_{mxy} \). If this condition applies, the sign of \( B(s'_{rxy}) \) depends on the signs of \( (\tilde{x}_r - \tilde{x}_m) \) and \( (\tilde{y}_r - \tilde{y}_m) \); if these terms have the same sign, \( B(s'_{rxy}) \) will be negative, but if they have different signs \( B(s'_{rxy}) \) will be positive.

The purpose of examining the impact of nonresponse on several statistics in the above discussion is to show that nonresponse can have different effects on different statistics. In examining the properties of various compensation procedures in the following chapters, attention will need to be given to their effectiveness in counteracting these diverse effects.

1.4 Nonresponse in the 1978 ISDP Research Panel

Heeringa (1980) has carried out a detailed examination of the unit and item nonresponse encountered in the 1978 ISDP Research Panel focusing primarily on the first two waves (April and July 1978) and on missing data for the income items. Briefer accounts are provided by Kalton et al. (1980, 1981). A few of the findings will be summarized here in order to give an impression of the magnitude of the missing data problem, and to reveal some of its facets.

Interviews were conducted with the 1978 Research Panel on five occasions, at quarterly intervals, the first wave of interviews taking place in April 1978 and the last in April 1979. The Panel was selected in 60 Census Primary Sampling Units (PSU’s) with an initial sample of 2,358 sample units, consisting of 1,947 area probability sample housing units and 411 persons drawn from a list of Supplemental Security Income (SSI) recipients. The area sample was defined as the households occupying sampled residential units and the SSI sample was defined as the households in which the sampled persons were found to be living. The initial sample of persons for the April 1978 wave of the Panel was defined as all those aged 16 and over living in the sampled households. All members of the initial sample were included in the Panel. Those who moved to new addresses were followed and interviewed; however, for cost reasons, those moving beyond fifty miles of the Panel’s PSU’s were not followed. Persons who moved in with members of the initial sample (or vice-versa) became part of the sample.

In discussing missing data in the 1978 Research Panel, various levels of nonresponse need to be distinguished. At any one wave of the Panel, nonresponse occurred at: (1) the household level, with no data being collected for any household member; (2) the person level, with no data being collected for one or more household members in an otherwise cooperating household; and (3) the item level, when some but not all data are missing for a sample person. The panel aspect of the design adds another dimension to the household and person levels of nonresponse: a household or person may provide data for all waves of the Panel, may provide data for some waves but not for others, or may fail to provide data for any wave.

The household level nonresponse at the first (April) wave of the Panel was 6.5%, three quarters of which were refusals. Some 2.5% to 3.0% of households were lost subsequently at each successive quarter through adults moving beyond fifty miles of the Panel’s PSU’s and through other movers who could not be traced. Among households remaining at their sample addresses or moving within the Panel’s sample areas, there was an increase in nonresponse through the life of the Panel, but at a declining rate; refusals were the dominant cause of nonresponse at each wave. By the last wave of the Panel in April 1979 the nonresponse rate among these households had risen to 15.0%.

The wave to wave change in nonresponse rates is the net effect of obtaining responses in a later wave from households that failed to provide data in an earlier one together with the loss of some households cooperating in an earlier wave but not providing data in a later one. Attempts were made in the 1978 Research Panel to secure responses in a later wave from households that did not respond in an earlier one. The only
evidence available to date on the effect of these attempts comes from the first two waves of the Panel: only 20% of the 140 nonresponding households in the April wave provided data for the July wave (of the 103 households refusing in April, 15 responded in July).

There was little nonresponse at the person level within cooperating households: in the April 1978 wave, no data were provided for only about 1.5% to 2.0% of adults in such households. This low level of person nonresponse was achieved through the use of a proxy interview with a knowledgeable family member when an adult was unavailable. (Data on 31.7% of sampled adults were provided by proxy informants in the April 1978 wave of the Panel.) Person level nonresponse thus occurred only when the adult refused to cooperate or when he was absent and no other family member felt able to respond on his behalf.

The level of item nonresponse varied considerably across items. Almost complete information was obtained for many of the easier items, including the basic demographics, but there was a fair amount of nonresponse for some of the more difficult income items. Thus, for example, in the April 1978 wave some 10.8% of area sample salaried respondents paid the same amount each payday failed to provide the amount of their paycheck either from their records or as an estimate, and some 9.7% of area sample respondents paid hourly failed to provide their regular hourly rate of pay. One of the highest levels of item nonresponse occurred with an item asking for interest received during the last three months from savings accounts. In the April 1978 wave 47.6% of area sample respondents with such accounts failed to provide this information. For many items the item nonresponse rate was found to be considerably higher when the data were collected from proxy informants than from self reporters.

As the preceding discussion indicates, the simple textbook dichotomy of nonresponse into unit and item nonresponse is inadequate for a household panel survey like the 1978 Research Panel, and in fact the distinction between these two types of nonresponse becomes extremely blurred with such a survey. The inclusion of all members of selected households in the sample and the panel aspect of the design both contribute to this situation. If one person in an otherwise cooperating household refuses to be interviewed, should this be viewed as unit or item nonresponse? For analyses conducted at the household level, this type of nonresponse may be treated as a series of item nonresponses in the household record. For person level analyses, however, it may be appropriate to treat the person as a unit nonrespondent (but noting that data for the household and for other household members may provide useful information about the nonrespondent). If a household - or person - fails to provide data on one wave of the panel, should this be viewed as unit or item nonresponse? For longitudinal analyses, this case may be treated as a series of item nonresponses in the household's longitudinal record, while for cross-sectional analyses it may be treated as unit nonresponse (but again noting that the household's data from other waves may provide valuable information about its missing responses).

The inclusion of all members of selected households in the sample also introduces complexity with regard to compensation for person-level nonresponse. Apart from a few cases where interviewers may have failed to determine whether selected housing units were occupied by households, the number of nonresponding households is known. In consequence, nonresponse weighting adjustments based on the inverse of household response rates can be readily implemented. The situation for person-level nonresponse is, however, complicated by the fact that the number of persons in nonresponding households may often not be obtained. As a result person-level response rates cannot be directly computed, but have to be estimated employing some assumption about the size of nonresponding households. The effectiveness of weighting adjustments based on person-level response rates thus depends on the appropriateness of this assumption, as well as that of the other assumptions underlying the procedure.
2. General Issues for Missing Data Compensation Procedures

2.1 Assumptions

Whenever a survey experiences missing data, assumptions must be made in order to produce estimates for the population of interest. This statement holds whether compensation procedures are used or not. Researchers sometimes justify making no compensation for missing data - that is simply analyzing the responses - on the grounds that they thereby avoid the subjective assumptions required in all compensation procedures, and that their results are in consequence more defensible. If valid, this justification would be a powerful one, for objectivity of survey results is a real concern, especially for surveys conducted by government agencies. However, on examination, the justification turns out to be less persuasive than it might at first appear.

To illustrate this point, consider the use of the respondent mean \( \bar{Y}_r \) to estimate the population mean \( \bar{Y} \), as discussed in Section 1.3. The bias of \( \bar{Y}_r \) was shown in equation (1.3.1) to be \( M(\bar{Y}_r - \bar{Y}_m) \). If the nonresponse rate \( M \) is sizeable, the suitability of \( \bar{Y}_r \) as an estimator of \( \bar{Y} \) thus depends on the assumption that the respondents and nonrespondents have similar means. In other words, in a survey with missing data even an estimation procedure which does not compensate for nonresponse incorporates an implicit assumption when estimating the overall population mean. This assumption may be technically avoided by claiming that the results apply only to the response stratum, but this begs the issue. Since the response stratum is not the population of interest, the user of the survey results is left to make his own rough adjustments. Lacking detailed information about the characteristics of nonrespondents, the user cannot be expected to adjust for the missing data as well as the original analyst would have done. Thus, although compensation procedures must depend on assumptions, it seems preferable to adopt them using as realistic assumptions as possible (and fully documenting the assumptions made and procedures used), rather than simply to analyze the respondent data, a procedure which itself depends on an often highly questionable assumption.

Every compensation procedure needs to make an explicit assumption about the characteristics of the nonrespondents. In estimating means or totals, the simplest assumption is that the respondents' means are equal to those of the nonrespondents \( \{\bar{Y}_r = \bar{Y}_m\} \). An alternative assumption is that the missing data are missing at random. Two processes can be envisaged for generating data missing at random. One is to postulate a single superpopulation from which the finite populations of respondents and nonrespondents are drawn at random; in this process, respondents always respond over conceptually repeated applications of the survey and nonrespondents never do. The second process is to consider every unit in the finite population as having the same chance of responding; in this process a respondent on one application of the survey may be a nonrespondent on another. (See, for instance, Platek et al., 1978, Platek and Gray, 1979, 1980).

The assumption that data are missing at random is in one sense somewhat less exacting than the assumption \( \bar{Y}_r = \bar{Y}_m \). With the first process for generating data missing at random, the finite population means for respondents and nonrespondents may differ because of sampling variability in sampling from the superpopulation; however, their \( \xi \)-expectations over sampling from the superpopulation are equal, \( E_\xi(\bar{Y}_r) = E_\xi(\bar{Y}_m) \). With the second process, a complete enumeration of the population would yield different means for respondents and nonrespondents, but the expectations of these means over repeated applications are equal. Providing that the (expected) population sizes for respondents \( R \) and nonrespondents \( M \) are much larger than the sample size \( n \), the sampling variability from the process of generating the data missing at random can be ignored, i.e. the respondent and nonrespondent population means may be treated as approximately equal.

In another sense, the assumption that data are missing at random is more stringent than the assumption \( \bar{Y}_r = \bar{Y}_m \). The latter assumption specifies nothing about the distribution of the nonrespondents' \( y \)-values around the mean, whereas the missing at random assumption implies the equality in expectation of the respondents' and nonrespondents'
distributions, and hence also of the variances and other parameters of the distributions.

The simple assumptions that $\bar{Y}_r = \bar{Y}_m$ or that the data are missing at random are rarely tenable in practice. Numerous surveys have provided ample evidence to demonstrate that response rates vary across population subgroups, and the survey variables are often associated with the characteristics of these subgroups. A natural extension of the simple assumption is first to divide the population into defined subgroups, and then to apply the simple assumption in one form or the other within each subgroup. This extension to an assumption of similarity of respondents and nonrespondents within subgroups is the basis of the most widely used compensation procedures. It is the underlying assumption of both the weighting adjustments discussed in Chapter 3 and the hot-deck imputations discussed in Chapter 4.

A problem arises when the researcher wants to take account of several characteristics in forming the subgroups within which he is prepared to assume the similarity of respondents and nonrespondents. In this case the subgroups formed as the cells of the crosstabulation of the characteristics may be too numerous to permit stable compensations to be made separately within each subgroup. This situation may be avoided by relaxing the assumption of similarity in subgroups to one which equates respondents and nonrespondents in terms of an additive model for the characteristics (i.e. assuming no interaction effects). A similar approach may also be used to handle characteristics which are continuous rather than classificatory variables. It may be preferable to assume a linear model for such variables rather than the alternative of categorizing them into a relatively small number of classes. Raking ratio estimation and regression imputation derive from these kinds of assumptions.

Compensations for missing data in surveys are generally based on equating the respondents and nonrespondents in some way after controlling for other variables. The specification of equality for the two groups is, however, not essential; one could instead specify the differences between them. Rubin (1977) and Ericson (1967), for instance, develop a Bayesian analysis incorporating subjective notions of differences between respondents and nonrespondents. The difficulty of specifying such differences in a survey context with numerous variables and a multitude of analyses severely limits the utility of this approach for general purpose survey work. It will therefore not be considered further in this report.

2.2 Compensation Procedures

The statistical literature on missing data problems can be broadly divided into that which deals with the estimation of specific parameters from a simple random sample with a known density and that which is of a general purpose nature. The concern of the present report is with the latter type of problem. We will therefore not review the sizeable literature on estimating specific parameters from incomplete data; however, some references are provided in the bibliography at the end of the report (for instance, Afifi and Elashoff, 1966, 1967, 1969; Anderson, 1957; Beale and Little, 1975; Dempster et al., 1977; Woodbury and Siler, 1966).

The compensation procedures with which we are concerned are based on the general purpose philosophy of aiming to provide reasonable estimates for many forms of analysis rather than the best estimates for particular parameters. A strong case can be made for adopting this approach in survey work. In particular, given the multitude of analyses to which a survey data set is subjected by many different researchers of varying degrees of statistical sophistication, it is unrealistic to believe that separate efficient compensation procedures could be developed for each analysis; and, even if they could, it would almost certainly be uneconomic to do so. The use of different compensation procedures may also lead to inconsistencies in the results obtained.

Although the literature on compensating for missing data in surveys is mostly concerned with general purpose techniques, the theoretical results available mainly relate to the impact of the procedures on simple descriptive statistics like means and totals. This emphasis will be reflected in this report but the general purpose nature of the compensation procedures should be constantly borne in mind. The companion report by Santos (1981a) considers the effects of compensation procedures.
on more complex statistics, especially those obtained from regression analyses.

General purpose compensation procedures can be separated into two types, weighting adjustments and imputation techniques. In practice these two types of procedures are often used in combination to handle different kinds of missing data. Weighting adjustments increase the weights of respondents in various subgroups of the sample in order to compensate for the subgroups' differing response rates, while imputation techniques insert values for missing responses. The dominant factor determining the choice between these two types of procedures for handling a particular kind of missing data is the amount of information available on the units involved, often broadly reflected by whether the missing data arose from noncoverage, unit response, or item nonresponse. When missing data compensations are made, noncoverage and unit nonresponse are generally treated by weighting adjustments while item nonresponse is treated by imputation.

The rationale for this situation is as follows. When little information on the unit with missing data is available - as with unit nonresponse - all the information can be used in forming the subgroups for which the weighting adjustments are made. In essence, units with missing data are matched with sets of respondents with the same characteristics on the available data. Making subgroup weighting adjustments to the respondents to compensate for the nonrespondents thus retains all the known information about the nonrespondents. The procedure of matching respondents and nonrespondents on the data available for nonrespondents becomes, however, more demanding the more that is known about the nonrespondents. At some point, it will become impossible to find even one respondent who corresponds exactly with a nonrespondent in terms of the data available on the latter (this typically applies with item nonresponse). Two alternative approaches are possible when this occurs: either to discard some of the less important data about the nonrespondent until one or more matching respondents can be found, and use a weighting adjustment; or to use an imputation procedure. If only a minimal amount of uninformative data needs to be discarded, it may be appropriate to use a weighting adjustment (as, for instance, when a respondent answers only one or two of the initial questions on a questionnaire). However, when the loss of information incurred from discarding data is substantial, as it often would be with item nonresponse, the alternative approach of imputation is preferable. As discussed in the next section, imputation has a number of disadvantages compared with weighting adjustments, but it does routinely retain all the available information on units with missing data.

While the distinction between weighting adjustments and imputation is used as the major classification for the organization of this report, it should however not be assumed that the procedures are unrelated. In fact, they both operate in the same basic way of building up the respondents' data to represent the missing data. Indeed there exists an equivalent simple imputation scheme for any weighting adjustment that employs integer weights. Consider, for instance, a weighting scheme in which a respondent is chosen at random within a subgroup and is given additional weight to represent a nonrespondent from that subgroup. Providing the subgroup characteristics reflect all the information known about the nonrespondent, this scheme is equivalent to an imputation scheme in which a nonrespondent is assigned all the responses from a respondent chosen at random in the same subgroup. The two schemes will produce the same estimates from the resultant data sets. The difference between them is simply that with the weighting adjustment the records for the respondent and the nonrespondent are merged into a single record with increased weight, whereas with the imputation procedure the two records, although identical, remain separate.

An equivalence can also be established in the more typical situation when noninteger weighting adjustments are used. Suppose, for example, that in an equal probability sample the r respondents in a subgroup are each assigned the weight $n/r$ to represent the $m = (n - r)$ nonrespondents. An equivalent imputation procedure would be to split each nonrespondent record into r parts, each with a weight of $1/r$, and assign the records from each respondent in turn to each of the r parts. The complete data set generated from this procedure would have
(r + mr) records, r respondent records with unit weights and mr nonrespondent records with weights of 1/r each. Analysis of this data set would produce the same estimates as those obtained from the data set with nonresponse weighting adjustments. In certain circumstances, there exist variants of the preceding imputation scheme which are also equivalent to the weighting adjustment. If, say, r and m have a common factor r = kr' and m = km', where k, r' and m' are integers, the respondents and nonrespondents in the subgroup could be divided into k groups each of size (r' + m'), and then the preceding procedure could be applied separately within each group; that is, within each group, the nonrespondents' records are divided into r' parts each with a weight of 1/r', and each part is then assigned a respondent record from one of the r' respondents in that group. This procedure generates a data set with (r + mr') records, a smaller number than the (r + mr) in the basic version.

The similarities between weighting adjustments and imputation procedures are worth keeping in mind for they may suggest ways of devising and treating imputation procedures that overcome some of the current disadvantages to imputation as a method of compensating for missing data.

2.3 Three General Criteria in the Choice of Compensation Procedures

The previous section has discussed the basic issue of when weighting adjustments and when imputation methods should be used to compensate for missing data. This section introduces three other considerations involved in making an appropriate choice of compensation procedure. These considerations, which apply whether weighting adjustments or imputation techniques are used, are concerned with: (1) the precision of the resulting estimates; (2) the estimation of standard errors of the estimates; and (3) the suitability of the compensated data set for producing estimates for a variety of different parameters. For simplicity, the section will consider only the simplest type of compensation procedure, one applied to the sample taken as a whole. The extension to the more general situation where compensations are carried out separately for different subgroups will be taken up in the following chapters.

Throughout this chapter it is assumed that the sample of n units is drawn from the population by simple random sampling.

2.3.1 Precision of Estimators

For simplicity we will consider in this section only the estimation of the population total Y from incomplete data. As shown in Section 1.3, the simple inflation estimator \( \hat{Y} = \frac{n}{n} \sum_{i=1}^{n} y_i \) has a negative bias of \( \frac{M}{n} \) as an estimator of Y. Its use is tantamount to assuming that \( \hat{Y}_m = 0 \), usually an unrealistic assumption. In most cases, a more appropriate assumption would be either that \( \hat{Y}_m = \hat{Y}_r \) or that the missing data are missing at random, as discussed in Section 2.1. There are a variety of methods in which compensations for the missing data can be carried out in line with these assumptions, both by weighting adjustments and imputation. For the present discussion these methods may be usefully classified into those which impose the condition \( \hat{Y}_m = \hat{Y}_r \) and those which relax this condition to one in which the expected value of \( \hat{Y}_m \) over repeated applications of the method, \( E_2(\hat{Y}_m) \), is equal to \( \hat{Y}_r \).

The well-known estimator \( \hat{Y}_1 = \frac{n}{n} \sum_{i=1}^{n} y_{ri} \) represents the obvious form of weighting adjustment in which \( \hat{Y}_r = \hat{Y}_m \). This condition can be clearly seen if \( \hat{Y}_1 \) is expressed as \( \hat{Y}_r + \hat{Y}_m \) as an estimator of \( \hat{Y}_r + \hat{Y}_m \). It can be expressed in the form of a simple inflation estimator using a weighting adjustment as \( \hat{Y}_1 = \sum_{i=1}^{n} w_i \hat{y}_{ri} \) where \( \hat{y}_w = \sum_{i=1}^{n} w_i \hat{y}_{ri} \) and \( w_i = \frac{n}{r} \). This adjustment is the simplest form of weighting adjustment in which all respondents are assigned the same weight. In passing, it may be noted that when applied in estimating means this adjustment has no effect, yielding the same estimate as obtained without any compensation. This is readily seen from

\[
\hat{y}_w = \frac{\sum_{i=1}^{n} w_i \hat{y}_{ri}}{\sum_{i=1}^{n} w_i} = \frac{[(n/r) \sum_{i=1}^{n} y_{ri}]/[(n/r) n]}{\sum_{i=1}^{n} w_i} = \hat{y}_r.
\]

It is possible to conceive of alternative weighting adjustment schemes satisfying the condition \( \hat{Y}_r = \hat{Y}_m \) assigning larger weights to a subset of respondents and perhaps different weights to different respondents, but such schemes may be difficult to operationalize.

A straightforward imputation procedure satisfying the condition \( \hat{Y}_r = \hat{Y}_m \) is to impute values of \( \hat{Y}_r \) for all the
nonrespondents. More generally, however, the imputed values could be any set of numbers that sum to $m_{\hat{Y}_r}$.

The relaxation of the condition from $\hat{Y}_m = \hat{Y}_r$ to $E_2(\hat{Y}_m) = \hat{Y}_r$ introduces a degree of flexibility which permits a wide range of alternative compensation procedures. The condition may be satisfied by a variety of weighting adjustment schemes in which weights are assigned to an equal probability (epsem) sample of respondents to compensate for the nonrespondents. There are two main aspects to the choice of weighting scheme, the sample design to be used and the size of sample to be selected. Assuming $r > m$, one obvious choice of sample size would be to take a sample of size $m$, and increase the weight of each selection by 1 to compensate for a nonrespondent. The sample could, for instance, be selected: by unrestricted sampling (simple random sampling with replacement); by SRS (simple random sampling without replacement); by proportionate stratified sampling with the respondents stratified in some way, perhaps by their $y$-values; or by systematic sampling from a list of the respondents ordered in some way. Another choice for the sample size might be $cm$, with the weight of each selection being increased by $1/c$; any epsem scheme could then be used to select the sample. Note that the simple inflation estimator $\hat{Y}_1$ comes from the special case of this scheme when $cm = r$ and sampling is without replacement.

With $r > m$, an imputation scheme satisfying $E_2(\hat{Y}_m) = \hat{Y}_r$ can be obtained by selecting any epsem sample of $m$ respondents, and assigning their values to the nonrespondents. The possible sampling schemes include any of those mentioned above. Sampling can be carried out with or without replacement, and techniques like stratification and systematic sampling can be used.

The case of $m > r$ implies that over 50% of the data on the variable under study are missing. This is seldom likely to occur in practice for the total sample and, if it were to occur, it would raise serious doubts about whether the data were usable. Nonresponse adjustments are, however, made within subgroups of the sample; the case $m > r$ may well occur in some of the subgroups and for this reason it deserves attention.

When $m > r$, an epsem without replacement sampling scheme for choosing respondents to be assigned increased weight or choosing respondents as donors of imputed values cannot be directly implemented. To cope with this problem, we generalize the without replacement sampling schemes as follows. Let $m = kr + t$, where $k$ is a non-negative integer and $0 \leq t < r$; when $r > m$, $k = 0$ and $m = t$. Then, with weighting adjustments, $t$ respondents are selected by the epsem sampling scheme to be given weights of $(k + 2)$ and the remaining $(r - t)$ respondents are given weights of $(k + 1)$. With imputation, all respondents serve as donors on $k$ occasions, with an epsem sample of $t$ of them serving as donors on one more occasion. The versions of these schemes in which SRS is used will be referred to as the SRS weighting and imputation schemes.

In this section we will derive and compare variances of the estimators $\hat{Y}$ under different compensation procedures. In these derivations, the numbers of respondents and nonrespondents in the population, $R$ and $M$, will be treated as fixed. This procedure is appropriate if respondents are viewed as always responding and nonrespondents are viewed as never responding over repeated applications of the survey. If, instead, all units in the population are viewed as having the same response probability, the variances obtained are conditional ones, conditional on a given set of $R$ respondents and $M$ nonrespondents. All the variances are derived conditional also on a fixed number, $r$, of sampled respondents.

The criterion of precision of the estimator $\hat{Y}$ favors compensation procedures that require $\hat{Y}_m = \hat{Y}_r$ over those that require only that $E_2(\hat{Y}_m) = \hat{Y}_r$. To see this, note that the estimator obtained from any compensation procedure that requires $\hat{Y}_m = \hat{Y}_r$ can be shown to be equal to $\hat{Y}_1 = N_{\hat{Y}_r}$, and that the conditional expectation of the estimator obtained from any compensation that requires $E_2(\hat{Y}_m) = \hat{Y}_r$ can be shown to be equal to $\hat{Y}_1$. The variance of any $\hat{Y}$ can be expressed as

$$V(\hat{Y}) = E_1V_2(\hat{Y}) + V_1E_2(\hat{Y}) \quad (2.3.1.1)$$
where $E_2$ and $V_2$ are conditional expectation and variance over the compensation subsampling scheme, and $E_1$ and $V_1$ are the corresponding operators for the initial sampling. Thus

$$V(\hat{Y}) = E_1 V_2(\hat{Y}) + V_1(\hat{Y}_1). \quad (2.3.1.2)$$

The first term is zero when $\hat{Y} = \hat{Y}_1$, a constant, as occurs when $\hat{y}_m = \hat{y}_r$; in this case $V(\hat{Y}) = V_1(\hat{Y}_1) = V(\hat{Y}_1)$. Otherwise it is positive, indicating the larger variance of compensation procedures requiring only that $E_2(\hat{y}_m) = \hat{y}_r$. We will investigate the proportionate increase in the variance of $\hat{Y}$ associated with various compensation subsampling schemes, where the proportionate increase is defined by

$$I = [V(\hat{Y}) - V(\hat{Y}_1)]/V(\hat{Y}_1) = E_1 V_2(\hat{Y})/V(\hat{Y}_1).$$

The magnitude of $I$ depends on the form of subsampling employed: it is, for instance, larger for unrestricted sampling than SRS and likely to be larger for SRS than for proportionate stratified sampling. The denominator of $I$ is the variance of the estimator of the population total for all compensation schemes with $\hat{y}_m = \hat{y}_r$. Conditional on $r$, this variance is

$$V(\hat{Y}_1) = N^2 V(\hat{y}_r) = N^2[1 - (r/R)]S_r^2/r \quad (2.3.1.3)$$

where

$$S_r^2 = \sum(\hat{y}_{ri} - \hat{y}_r)^2/(r - 1).$$

(a) Subsampling by SRS. Consider first the SRS weighting and imputation schemes. With either scheme, the estimator of $Y$ can be expressed as

$$\hat{Y}_2 = N[(k + 1)r\hat{y}_r + t\hat{y}_t]/n \quad (2.3.1.4)$$

where $\hat{y}_t$ is the mean of the SRS of $t$ respondents chosen to have an additional weight or to be used as a donor an extra time. The variance of $\hat{Y}_2$ can be obtained from (2.3.1.2), with the second component of that equation being given by (2.3.1.3). The first component is

$$E_1 V_2(\hat{Y}_2) = E_1[N^2 t(1 - P)S_r^2/n^2] = N^2 t(1 - P)S_r^2/n^2 \quad (2.3.1.5)$$

where $S_r^2 = \sum(y_{ri} - \hat{y}_r)^2/(r - 1)$, and $P = t/r$ is the sampling fraction for the subsample of size $t$. Then

$$V(\hat{Y}_2) = V(\hat{Y}_1) + N^2 S_r^2 P(1 - P)(k + 1 + P)^{-2}. \quad (2.3.1.6)$$

If the finite population correction $[1 - (r/R)]$ can be ignored, the proportionate increase in variance from using the SRS weighting or imputation scheme over the simple equal reweighting scheme (or another compensation procedure that satisfies $\hat{y}_m = \hat{y}_r$) is

$$I_2 = P(1 - P)/(k + 1 + P)^2. \quad (2.3.1.7)$$

For fixed $k$ the maximum increase occurs when $P = (k + 1)/(2k + 3)$ and clearly the increase is greater the smaller the value of $k$. When $k = 0$, i.e. the number of nonrespondents is less than the number of respondents and $t = m$, the maximum increase occurs when $P = m/r = 1/3$; in this case the proportionate increase in variance is $1/8$.

As Kish (1965, Section 11.7; 1976) demonstrates, the SRS weighting scheme described above produces a more precise estimator than one with greater variability in the weights, as for instance would occur if, say, a subsample of $t/2$ respondents were assigned weights of $(k + 3)$ rather than $t$ respondents being assigned weights of $(k + 2)$, or the equivalent with donors in an imputation scheme.

(b) Subsampling by unrestricted sampling. The SRS weighting and imputation schemes also have lower variances than if unrestricted sampling had been used. For the latter scheme, suppose that $m$ respondents are chosen with equal probability with replacement to serve as donors for imputation or to receive unit increases in weight. The estimator of $Y$ may then be expressed as $\hat{Y}_3 = N(y_r + y_m)/n$, with

$$E_1 V_2(\hat{Y}_3) = E_1[N^2 t(1 - P)S_r^2/n^2] = N^2 t(1 - P)S_r^2/n^2 \quad (2.3.1.5)$$

where $S_r^2 = \sum(y_{ri} - \hat{y}_r)^2/(r - 1)$, and $P = t/r$ is the sampling fraction for the subsample of size $t$. Then

$$V(\hat{Y}_3) = V(\hat{Y}_1) + N^2 S_r^2 P(1 - P)(k + 1 + P)^{-2}. \quad (2.3.1.6)$$

If the finite population correction $[1 - (r/R)]$ can be ignored, the proportionate increase in variance from using the SRS weighting or imputation scheme over the simple equal reweighting scheme (or another compensation procedure that satisfies $\hat{y}_m = \hat{y}_r$) is

$$I_2 = P(1 - P)/(k + 1 + P)^2. \quad (2.3.1.7)$$

For fixed $k$ the maximum increase occurs when $P = (k + 1)/(2k + 3)$ and clearly the increase is greater the smaller the value of $k$. When $k = 0$, i.e. the number of nonrespondents is less than the number of respondents and $t = m$, the maximum increase occurs when $P = m/r = 1/3$; in this case the proportionate increase in variance is $1/8$.
Thus, ignoring the f.p.c. \(1 - (r/R)\), the proportionate increase in variance over compensation schemes in which \(Y_m = Y_r\) is

\[
I_3 = m(r - 1)/n^2 = \tilde{m}(1 - \tilde{m}) \quad (2.3.1.9)
\]

if \(r\) is large so that \((r - 1) = r\). The maximum increase in this case occurs when the nonresponse rate \(\tilde{m}\) is 0.5; the proportionate increase in variance is then \(1/4\). From (2.3.1.5) and (2.3.1.8), the increase in variance through using unrestricted sampling over SRS in a weighting or imputation scheme is

\[
V(\hat{Y}_3) - V(\hat{Y}_2) = V(\hat{Y}_1)[kr(r - 1) + t(t - 1)]/n^2. \quad (2.3.1.10)
\]

This difference is always nonnegative. It is zero only when \(k = 0\) or \(r = 1\) and \(t = 0\) or 1.

(c) Subsampling by proportionate stratified sampling. The increase in variance (2.3.1.7) through using the SRS weighting or imputation scheme can be substantially reduced by replacing SRS by proportionate stratified sampling. For the SRS scheme, from (2.3.1.5)

\[
V_2(\hat{Y}_2) = N^2t(1 - P)s_r^2/n^2.
\]

The conditional variance of the estimator of the population total, \(\hat{Y}_4\), with a proportionate stratified sampling scheme is

\[
V_2(\hat{Y}_4) = N^2t(1 - P)s_{r^w}^2/n^2
\]

where \(s_{r^w}^2\) is the average within-stratum variance. The ratio of these two variances is thus

\[
E_1V_2(\hat{Y}_3) = E_1[N^2V_2(\hat{Y}_m)/n^2] = E_1[N^2m(r - 1)s_r^2/rn^2]
\]

\[
= N^2m(r - 1)s_r^2/rn^2. \quad (2.3.1.8)
\]

(d) Subsampling with a larger sample. Finally, mention should be made of another method of reducing the increases in variance arising from the SRS weighting and imputation schemes: this involves increasing the subsample size from \(t\) to, say, \(ct\) \((ct < r)\), with a counterbalancing adjustment to the additional weight to \(1/c\), or with splitting each nonrespondent into \(c\) parts, each with weight \(1/c\), and using different donors for each part. The estimator for the total for this scheme may be expressed as

\[
\hat{Y}_5 = N[(k + 1)r\tilde{y}_r + t\tilde{y}_{ct}]/n \quad (2.3.1.12)
\]

with

\[
E_1V_2(\hat{Y}_5) = E_1[N^2t(1 - cP)s_r^2/cn^2] = N^2t(1 - cP)s_r^2/cn^2.
\]

In this case, the proportionate increase in variance over schemes in which \(\hat{Y}_m = \hat{Y}_r\) is

\[
d = V_2(\hat{Y}_5)/V_2(\hat{Y}_2) = s_{r^w}^2/s_r^2. \quad (2.3.1.11)
\]

If effective stratification is employed, \(s_{r^w}^2\) will be small compared with \(s_r^2\) in which case the losses of subsampling will be substantially reduced.

The respondents can in fact be stratified highly effectively in terms of their \(y\)-values. One simple scheme is to divide them into, say, \(s\) equal-sized strata, with the first stratum containing the \(r/s\) respondents with the largest \(y\)-values, the second stratum containing the \(r/s\) respondents with the largest \(y\)-values among the remainder, etc., and then to select \(t/s\) respondents from each stratum (assuming for simplicity that \(r/s\) and \(t/s\) are integers). With this kind of stratification, which is closely related to problems of matching and grouping, the value of \(d\) rapidly becomes small as \(s\), the number of strata, increases: for \(s = 5\) the value of \(d\) is around 0.05 to 0.10 for many distributions of \(y\). (See, for instance, Aigner et al., 1975; Anderson et al., 1980.) The maximum value that \(s\) can take is \(s = t\). A systematic sample from a list of respondents ordered by their \(y\)-values should give similar gains, and this scheme is very easy to implement.
\[ P(1 - cP)/c(k + 1 + p)^2. \]  

(2.3.1.13)

Comparison of this increase with that for the SRS scheme (2.3.1.7) shows that this increase is smaller by the factor \((1/c)\) and also by a factor \((1 - cP)/(1 - P)\). The first factor occurs because of the increased sample size and the second because of the f.p.c. term. The use of the above scheme with imputation has been proposed by Kish (1979), who has termed it a repeated replication imputation procedure (RRIP). A note he has written on the procedure is included as an appendix to this report. See also Section 5.1 on multiple imputations.

Kalton and Kish (1981) provide further discussion of the variances examined in this section. Other investigations of the effect of missing data compensation procedures on the precision of survey estimators are reported by Bailar et al. (1978), Ernst (1980), Ford (1980), and Scheuren (1980). The loss of precision from random duplication of records is also discussed by Roshwalb (1953) and in the texts of Hansen et al. (1953, Vol. II, pp. 139-141) and Kish (1965, pp. 427-428).

2.3.2 Standard Error Estimation

A second criterion for choice between alternative compensation procedures is whether they enable reasonable variance estimates to be computed in a straightforward way. At least for the simple case considered in this chapter, this criterion tends to favor weighting adjustments rather than imputation. To establish this point, note that the estimator \( \hat{Y} = \frac{N\sum w_i Y_i}{n} \) serves to represent the estimator of the population total for all the main compensation procedures mentioned in this chapter. This is obvious for all weighting adjustments; for imputation procedures involving the assignment of respondent values to nonrespondents, the respondent values are in effect weighted by \( w_i = (1 + g_i) \), where respondent \( i \) serves as donor to \( g_i \) nonrespondents; and the case where all \( m \) nonrespondents are assigned the respondent mean \( \bar{Y}_r \) can be viewed as giving all respondents the weight of \( n/r \).

Consider now the estimation of the variance of \( \hat{Y} \). We assume that the population is large so that the f.p.c. may be ignored and the SRS of sampled units may be treated as independently selected. We also assume that the weights \( w_i \) associated with different respondents are uncorrelated. This assumption could be achieved by giving all respondents the same weight, by imputing the value \( \bar{Y}_r \) to all nonrespondents, or by allocating weights or selecting donors by unrestricted sampling. As an approximation it also covers SRS weighting and imputation schemes where the sampling fraction for the subsample, \( t/r \), is small. With these assumptions the quantities \( U_i = w_i \bar{Y}_i \) are uncorrelated. Hence the variance of \( \hat{Y} = \frac{N\sum u_i}{n} \), where the summation runs over the \( r \) respondents, is

\[ V(\hat{Y}) = \frac{N^2 \sigma_u^2}{n^2}. \]  

(2.3.2.1)

where \( \sigma_u^2 \) is the population variance of the \( U_i \).

Providing the sample values \( u_i \) are available, \( \sigma_u^2 \) may be readily estimated by \( s_u^2 = \frac{1}{r-1} \sum (u_i - \bar{u})^2 \) to produce the variance estimator

\[ v(\hat{Y}) = \frac{N^2 s_u^2}{n^2}. \]  

(2.3.2.2)

With weighting procedures, all the respondent records automatically contain both \( w_i \) and \( \bar{Y}_i \), so that the \( u_i \) are obtained straightforwardly. This is, however, not the case with imputation procedures. What would be needed would be to log on the respondent's record a variable to indicate the number of times he was used as a donor \( (g_i) \) and to attach identifying flags to the nonrespondents (so that they could be eliminated for variance estimation purposes). These steps add to the complexity of the imputation procedure and, because each survey variable has to be treated separately, substantially increase the length of the data records for sampled units; since they are not required for general analyses, they are often not deemed necessary. Although imputed values are frequently flagged, this is not a universal practice; the logging of the use of respondents as donors is seldom carried out.

In passing, it is worth noting that flagging of nonrespondents and logging of the use of respondents as donors can be handled by a single variable. A simple scheme for epsem samples is to define a weight variable \( w_i = (1 + g_i) \) for
respondents and \( w_i = 0 \) for nonrespondents. The scheme can be extended to non-epsem samples by defining \( w_i = a_i + \Sigma a_j \) for respondents, where \( a_i \) is the initial weight of respondent \( i \) and \( \Sigma a_j \) is the sum of the weights of the nonrespondents for which respondent \( i \) acts as donor, and \( w_i = 0 \) for nonrespondents.

Often when imputation is used the researcher is confronted with a full data set of \( n \) values without being able to distinguish between actual and imputed values. A natural path to take in this situation is to treat the imputed values as actual values, but the resultant variance estimator for \( \hat{V}(\hat{y}) = n^2 s^2 / n \), clearly underestimates \( V(\hat{y}) \). An obvious reason for the underestimation is that the sample size is taken to be \( n \) rather than \( r \), but there is also another factor involved. To illustrate this, we will examine the situation with two alternative imputation procedures.

First consider the imputation procedure which assigns the respondent mean to all nonrespondents. The sample estimate of the element variance from the full data set is then

\[
s^2 = \frac{\Sigma (y_i - \bar{y}_r)^2}{(n - 1)} = \frac{\Sigma[(\Sigma y_{ri} - \bar{y}_r) + \Sigma(y_{mi} - \bar{y}_r)^2]}{(n - 1)}
\]

where \( y_{mi} = \bar{y}_r \). Hence

\[
s^2 = \frac{\Sigma[(\Sigma y_{ri} - \bar{y}_r) + \Sigma(y_{mi} - \bar{y}_r)^2]}{(n - 1)} \quad (2.3.2.3)
\]

Thus, since \( E(s^2) = s^2_r \),

\[
E[v(\hat{y})] = (r - 1)N^2 s^2_r / n(n - 1) = (r - 1) \bar{V}(\hat{y}) / n(n - 1).
\]

\[
E[v(\hat{y})] = (r - 1)N^2 s^2_r / n(n - 1) \quad (2.3.2.4)
\]

The component \( (r/n) \) is accounted for by the correction from the inflated sample size of \( n \) to the true one of \( r \), while the second component \( (r - 1)/(n - 1) \) derives from the underestimation of \( s^2 \) for \( s^2_r \).

Secondly, consider the SRS imputation procedure. In this case \( s^2 \) is an almost unbiased estimator of \( s^2_r \), as is shown below, but there is another factor involved in making \( v(\hat{y}) \) an underestimate of \( V(\hat{y}) \). To show that \( s^2 \) is almost unbiased for \( s^2_r \), first note that

\[
\Sigma^n(y_i - \hat{y}_r)^2 = (k + 1)\Sigma^f(y_{ri} - \bar{y}_r)^2 + \Sigma^f(y_{ti} - \bar{y}_r)^2 - (t^2/n)(\bar{y}_t - \bar{y}_r)^2
\]

where \( y_{ti} \) is the \( y \)-value for a respondent chosen by SRS to act as donor for an extra time and \( \bar{y}_t = \Sigma y_{ti} / t \). Then

\[
E_2[\Sigma^n(y_i - \hat{y}_r)^2] = (k + 1)(r - 1)s^2_r + [t(r - 1)s^2_r] - (t^2/n)[1 - (t/r)]s^2_r/t
\]

\[
= s^2_r[(k + 1)r - t - 1] - k - (t/r) - (t/n)[1 - (t/r)]
\]

Hence

\[
E_1E_2(s^2) = s^2_r[1 - (n - 1)^{-1}][k + P + (t/n)(1 - P)] \quad (2.3.2.5)
\]

with \( P = t/r \). In the common situation when \( k = 0 \), i.e. \( r > m \), this becomes \( E_1E_2(s^2) = s^2_r[1 - 2m/(n - 1)] \). In any case, if \( r \) is large so that terms of order \( (1/r) \) may be neglected in (2.3.2.5),

\[
E_1E_2(s^2) = s^2_r. \quad (2.3.2.6)
\]

However, although \( s^2 \) is almost unbiased for \( s^2_r \), the variance estimator \( v(\hat{y}) = N^2 s^2 / n \) underestimates \( V(\hat{y}) \), which in this case is given by equation (2.3.1.6), that is

\[
V(\hat{y}) = N^2 s^2_r[1 + P(1 - P)(k + 1 + P)^{-2}] / r.
\]

Hence

\[
E[v(\hat{y})] = (r/n)[1 + P(1 - P)(k + 1 + P)^{-2}]^{-1}V(\hat{y}). \quad (2.3.2.7)
\]

The first factor \( (r/n) \) reflects the effect of treating the sample as if it were one of size \( n \) rather than \( r \), while the second factor reflects the increase in variance through the SRS
selection of \( t \) of the \( r \) respondents to act as donors on the additional occasion.

These results show that, when the amount of nonresponse is appreciable, the treatment of imputed values for nonrespondents as if they were actual values leads to a substantial overestimation of the sample's precision. A method of obtaining suitable variance estimators when imputed values are assigned to nonrespondents, employing the idea of multiple imputation, has been proposed by Rubin (1978, 1979b, 1980). Multiple imputations are discussed in Section 5.1.

In concluding this section, it should be noted that the variance estimation procedure outlined for use with weighting adjustments applies only to the simple case in which the weighting adjustment procedure is applied across the whole sample, that sample being selected by SRS. Weighting adjustment procedures applied within subgroups of the sample are considered later in the report.

2.3.3 General Applicability

The first two criteria for choice of compensation procedure — the precision and estimable precision of the estimator — have been illustrated by the estimation of a single parameter, the population total. The aim of the general purpose compensation procedures discussed in this report is, however, to produce data sets that can be used legitimately for many forms of analysis. The third criterion is, therefore, that after compensation the data set should satisfy the requirements for general analyses. In order to examine how the various compensation procedures fare with regard to this criterion we will assume that the missing data are missing at random; the alternative assumption \( \bar{y} = \bar{y}_m \) considered earlier becomes inadequate when interest extends to the estimation of parameters other than the mean and total.

After the mean and total, the two most obvious statistics to examine for univariate analyses are the distribution of the variable and its variance, the former because survey reports often contain categorized distributions and the latter because element variances play a role in many forms of analysis. The effect of compensation procedures on relationships can be most simply investigated in terms of the covariance between two variables, \( S_{xy} = \Sigma (x_i - \bar{x})(y_i - \bar{y})/(N - 1) \). We will examine various compensation procedures in turn to see their effects on the estimation of these three statistics. For this purpose we will employ the following estimators of the element variance and covariance from a weighted sample of size \( n \):

\[
\hat{s}_w^2 = \frac{n}{n - 1} \left[ \Sigma w_i (y_i - \bar{y})^2 / \Sigma w_i \right] \\
= \frac{n}{n - 1} \left[ \Sigma (w_i y_i^2 / \Sigma w_i) - \bar{y}^2 \right] \quad (2.3.3.1)
\]

and

\[
\hat{s}_{wxy} = \frac{n}{n - 1} \left[ \Sigma w_i (x_i - \bar{x})(y_i - \bar{y}) / \Sigma w_i \right] \\
= \frac{n}{n - 1} \left[ \Sigma w_i x_i y_i / \Sigma w_i - \bar{x} \bar{y} \right]. \quad (2.3.3.2)
\]

These estimators are commonly used, as are also the alternative estimators in which the term \( n/(n - 1) \) is dropped; the difference between these two forms of estimator is negligible for large \( n \).

(a) Equal weighting scheme. The effects of the simple weighting procedure in which all respondents are assigned the same weight, \( n/r \), to adjust for the nonrespondents are readily determined. The procedure clearly retains the respondents' distribution for the \( y \)-variable. The weighted sample variance estimator is

\[
\hat{s}_w^2 = \frac{r}{r - 1} \left[ \Sigma w_i (y_{ri} - \bar{y}_r)^2 / \Sigma w_i \right] \\
= \Sigma (y_{ri} - \bar{y}_r)^2 / (r - 1) = s_r^2 \quad (2.3.3.3)
\]

which is unbiased for \( s_r^2 \). Similarly the weighted sample covariance estimator is equal to \( s_{rxy} \), which is unbiased for \( s_{rxy} \).

(b) SRS weighting scheme. Results for the SRS weighting procedure will be derived as a special case of any epsm subsampling scheme for assigning the weights to compensate for nonrespondents. First, consider the distribution of the \( y \)-variable in the weighted sample. Let \( d_{ij} = 1 \) if \( y_i \) falls in class \( j \), i.e. \( y_{j-1} < y_i < y_j \), and \( d_{ij} = 0 \) otherwise. The weighted estimate of the proportion of the population in class
\[ j \text{ is then } P_j \doteq \frac{\Sigma w_i d_{ij}}{n}, \text{ and its conditional expectation over} \]
\[
\text{ways of subsampling respondents or assigning the weights is}
\]
\[ E_2(P_j) = \frac{\Sigma E_2(w_i) d_{ij}}{n} = \frac{\Sigma (n/r) d_{ij}}{n} = \frac{\Sigma d_{ij}}{r}. \quad (2.3.3.4) \]
Thus the conditional expectation of the weighted estimate is the unweighted respondent estimate, which is unbiased for the respondent proportion, and hence for the population proportion under the missing at random assumption.

The conditional expectation of \( s^2_w \) for any epsem subsampling weighting procedure in which \( t \) respondents are given a weight of \((k + 2)\) and \((r - t)\) are given a weight of \((k + 1)\) is
\[
E_2[(r/(r - 1))\left\{\Sigma (w_i y^2_i/n) - \bar{y}^2\right\}]
= \frac{r}{(r - 1)}\left\{\frac{(\Sigma y^2_i)}{r} - E_2(\bar{y} - \bar{y}_r)^2 - \bar{y}_r^2\right\]
= s^2_r - \frac{r}{(r - 1)}t^2V_2(\bar{y}_r)/n^2. \quad (2.3.3.5)
\]
For any nonclustered epsem subsampling scheme for the \( t \) respondents to be assigned the weight of \((k + 2)\), \( V_2(\bar{y}_r) \) is of order \((1/t)\). Hence the second term in \( (2.3.3.5) \) is of order \((t/n^2)\). Thus, providing \( n \) is large, \( s^2_w \) is an approximately unbiased estimator of \( s^2_r \), which is an unbiased estimator of \( s^2_r \) (see also Kish, 1965, Section 2.8C).

The same approach can also be applied with the weighted covariance to give
\[ E_2(s^2_{wxy}) = s^2_{rxy} - \frac{r}{(r - 1)}t^2C_2(x_r, \bar{y}_r)/n^2 \quad (2.3.3.6) \]
where \( C_2(x_r, \bar{y}_r) \), the conditional covariance of \( x_r \) and \( \bar{y}_r \), is of order \((1/t)\). For large \( n \), \( s^2_{wxy} \) is thus an approximately unbiased estimator of \( s^2_{rxy} \) for any epsem non-clustered subsampling scheme for assigning the weights.

Results for the SRS weighting scheme are obtained by substituting
\[ V_2(\bar{y}_r) = \left(\frac{1}{t}\right)s_r^2/t \]
and
\[ C_2(x_r, \bar{y}_r) = \left(\frac{1}{t}\right)s_{rxy}/t \]
in \( (2.3.3.5) \) and \( (2.3.3.6) \) respectively. For the conditional expectation of \( s^2_w \) this gives
\[ E_2(s^2_w) = s^2_r\left[1 - \frac{(r - 1)/t}{(r - t)/n^2}\right] \]
Approximating \( (r - 1) = r \), with \( P = t/r \) being the subsampling fraction and \( \bar{r} = r/n \) being the response rate,
\[ E_2(s^2_w) = s^2_r\left[1 - \{\bar{r}P(1 - P)/n\}\right]. \quad (2.3.3.7) \]
A corresponding result may be obtained in the same way for the covariance.

(c) Mean-value imputation. Turning to imputation procedures, consider first the one in which the mean value for respondents \( y_r \) is assigned to all nonrespondents. This procedure distorts the \( y \)-distribution by concentrating all the nonrespondents at the one value. The proportion of the total sample in class \( j \) is
\[ P_j = \frac{\Sigma d_{ij}}{n} = \frac{r_j}{n} \quad \text{if } \bar{y}_r < y_j \quad \text{or } \bar{y}_r \geq y_j \]
and
\[ P_j = \frac{(r_j + m)}{n} \quad \text{if } y_{j-1} \leq \bar{y}_r < y_j', \]
where \( r_j \) is the number of respondents falling in class \( j \). Thus, \( P_j \) overestimates the population proportion in the class containing \( \bar{y}_r \), and underestimates the proportions in all other classes. It may also be noted that if \( y \) is a discrete variable (e.g. number of children in the household), the imputed value \( \bar{y}_r \) will generally be an impossible non-integer individual value (e.g. 1.7). The extreme case is when \( y \) is a 0-1 variable and \( \bar{y}_r \) is the proportion of respondents with a particular attribute. By concentrating the nonrespondents at the mean, this procedure also leads to an underestimation of the element variance: as already shown in equation \( (2.3.2.3) \), \( s^2_r = \Sigma (y_i - \bar{y})^2/(n - 1) \) underestimates \( s_r^2 \) by a factor \((r - 1)/n - 1)\).

This imputation procedure can be extended to cover two (or more) variables by substituting \( x_r \) and \( \bar{y}_r \) for missing \( x \) and \( y \)-values. If there are \( r' \) respondents providing both \( x \) and \( y \)
The extension of the SRS or other epsem imputation scheme to pairs of variables introduces a number of possible outcomes depending on the pattern of missing data and the manner of applying the imputation scheme. One simple case is when both \( x \) and \( y \) are reported or both are missing. If each nonrespondent receives both imputed values from the same donor, the effect on the covariance is a straightforward generalization of that for the element variance (2.3.3.8), i.e.

\[
E_2(s_{xy}) = [n/(n-1)][(r-1)/r]E_2(s_{swy})
\]  

(2.3.3.9)

Since, for large \( n \), \( s_{swy} \) is an approximately unbiased estimator of \( S_{rxy} \) \( s_{xy} \) is also.

An alternative possibility is to assign the imputed values of \( x \) and \( y \) independently. When the \( x \) and \( y \) are missing in pairs as above, the effect of this imputation scheme on the covariance can be seen as follows:

\[
\begin{align*}
\Sigma_2(x_i - \bar{x})(y_i - \bar{y}) &= \Sigma_2(x_i - \bar{x}_r)(y_i - \bar{y}_r) + \Sigma_2(x_{mi} - \bar{x}_r)(y_{mi} - \bar{y}_r) - n(\bar{x}_r - \bar{x})(\bar{y}_r - \bar{y}).
\end{align*}
\]

(2.3.3.10)

With epsem subsampling and independent assignments of the \( x \) and \( y \) values, \( E_2(\bar{x}) = \bar{x}_r \), \( E_2(\bar{y}) = \bar{y}_r \), \( E_2(\bar{x}y) = \bar{x}_r\bar{y}_r \) and \( E_2(x_{mi} - \bar{x}_r)(y_{mi} - \bar{y}_r) = 0 \). Hence

\[
E_2(s_{xy}) = [(r-1)/(n-1)]S_{rxy}.
\]  

(2.3.3.11)

Thus, since \( E(s_{rxy}) = S_{rxy} \), \( s_{xy} \) computed under this imputation procedure underestimates \( S_{rxy} \) by a factor \((r-1)/(n-1)\). This result is the same as that obtained for the covariance when the imputation procedure assigned the respondent mean to all nonrespondents. However, unlike that case, the factor \((r-1)/(n-1)\) applies here only to the covariance and not to the variances. There is therefore no cancellation of the factor between numerator and denominator of the correlation or regression coefficients: both are underestimated by the factor \((r-1)/(n-1)\).
A third possibility is when data on one of the variables, \( x \), is complete, and only \( y \) has missing values. If the \( y_{mi} \)'s are assigned independently of the \( x_{mi} \) values (the reported \( x \) values for sampled elements with missing \( y \) values),

\[
E_2(x_{mi} - \bar{x})(y_{mi} - \bar{y}) = 0.
\]

In addition, since \( \bar{x} \) does not depend on the subsampling scheme, it follows that:

\[
E_2(n(\bar{x}_r - \bar{x})(\bar{y}_r - \bar{y})) = 0.
\]

Hence, from (2.3.3.10),

\[
E_2(\sum x_i (y_i - \bar{y})) = (r - 1) s_{rxy}.
\]

Thus again \( s_{xy} \) underestimates \( S_{rxy} \) by the factor \((r - 1)/(n - 1)\).

A fourth possibility is that both \( x \) and \( y \) are subject to missing data, and that imputations are made independently for each variable. In the partition of the sum of products in (2.3.3.10), let the first summation relate to the respondents who provide values for both \( x \) and \( y \) (i.e. replace \( r \) by \( r' \)), let the second summation relate to the sample members who provide partial data or no data on \( x \) and \( y \), let \( \bar{x}_r \) denote the sample mean for all sample members providing \( x \) values, and let \( \bar{y}_r \) denote the sample mean for all sample members providing \( y \) values. Then, with independent imputations for sample members providing no data,

\[
E_2(x_{mi} - \bar{x}_r)(y_{mi} - \bar{y}_r) = 0
\]

and

\[
E_2(n(\bar{x}_r - \bar{x})(\bar{y}_r - \bar{y})) = 0.
\]

Thus

\[
E_2(\sum x_i (y_i - \bar{y})) = \Sigma r'(x_{ri} - \bar{x}_r)(y_{ri} - \bar{y}_r) + r'(\bar{x}_r - \bar{x})(\bar{y}_r - \bar{y})
\]

where \( \bar{x}_r \) and \( \bar{y}_r \) are the means of the \( x \) and \( y \) values for the \( r' \) respondents providing both values. Hence

\[
E_2(s_{xy}) = [(r' - 1)/(n - 1)]s_{rxy}
\]

and

\[
E_2(n(\bar{x}_r - \bar{x})(\bar{y}_r - \bar{y})) = [r'( \bar{x}_r - \bar{x})(\bar{y}_r - \bar{y})]/(n - 1)
\]

where \( s_{rxy} \) relates to the subset of respondents providing both \( x \) and \( y \) values. If the missing \( x \) and \( y \) values are independently missing at random, the second term is zero and \( s_{rxy} \) estimates \( S_{xy} \). Then \( s_{xy} \) underestimates \( S_{xy} \) by the factor \((r' - 1)/(n - 1)\).

2.4 Concluding Remarks

There is a close relationship between weighting and imputation as means of making compensations for missing data, but there are also differences between the two types of procedures. Weighting is based on matching nonrespondents to respondents in terms of the data available on the nonrespondents, and increasing the weights of the matched respondents to account for the nonrespondents. If the matching is done on an individual basis, one respondent for each nonrespondent, the weighting procedure could be equivalently carried out by assigning the matched respondent's values to the nonrespondent, i.e. by imputation. The difference between the two procedures is that, when weighting, the data set consists of \( r \) respondent records with weights adding to \( n \), whereas with the imputation procedure the data set consists of \( n \) records, one for each respondent and nonrespondent.

Weighting need not, however, be carried out by individual matching, that is by adding integer weights to the respondents. The individual matching procedure in fact introduces an unnecessary stage of subsampling to choose the respondents for use as matched cases, and this gives rise to increased sampling error in the survey estimates. The simple weighting adjustment procedure of increasing all respondents' weights (in a defined subgroup) by the same, probably fractional, factor is a more...
efficient procedure because it avoids this subsampling stage. Before many computer programs had been developed to handle weighted data, a case could be made for using individual matching and the imputation scheme described above in order to avoid weights; nowadays, however, this consideration no longer applies, and the simple weighting procedure is the obvious choice when this type of adjustment is being made. The simple weighting procedure retains all the properties of the respondent sample, and hence leads to good estimates of other univariate statistics as well as means and totals and also to good estimates of multivariate statistics (under the assumption that the missing data are missing at random). It is for these reasons that the simple weighting procedure—applied separately with different weights within subgroups of the sample—is widely used to compensate for unit nonresponse.

Where there is a choice between weighting and imputation procedures, the various considerations discussed favor weighting. Thus weighting is generally employed for unit nonresponse, while imputation is reserved for item nonresponse (for which weighting is impracticable). Weighting also probably has an advantage over imputation in terms of economy of analysis of the resulting data set; it is likely to be more economical to analyze a weighted data set of \( r \) cases than an unweighted one of \( n \) cases. However, weighting adjustments require the matching of nonrespondents with respondents in terms of their available data, and this matching becomes impossible when a sizeable amount of data is available for the nonrespondents, as occurs in particular with item nonresponse. In such cases, the best that can be done is to match nonrespondents and respondents on a selection of major variables, and then to assign a matched respondent's values for the missing values in a nonrespondent's record. This imputation procedure is the same as the imputation scheme corresponding to the individual weighting scheme described above, except that not all the respondent's values are assigned to the nonrespondent, but only those for the variables for which the nonrespondent has missing values.

The use of imputation raises problems about sampling error estimation and also about the estimation of relationships between variables; these problems are discussed further in Chapter 4. The subsampling in the choice of donor adds another sampling error component, but this may be reduced by the use of multiple imputations and/or by the choice of subsampling scheme.
3. Weighting Adjustments

3.1 Types of Weighting Adjustment

The simple weighting procedure examined in the previous chapter has little practical utility since its application does not alter the estimator of the population mean \( \bar{y}_w = \bar{y}_r \) and the population total can be more easily estimated by \( Ny_r \) than by \( Fy_w (Ny_r = Fy_w) \). The procedure, however, forms the basis of more general weighting procedures, in which it is applied separately to different subgroups of the sample. These subgroups will henceforth be called weighting classes or simply classes; some researchers call them adjustment cells.

In Chapter 2 the weighting factor applied with the simple weighting adjustment procedure was the inverse of the response rate. This scheme can be extended to weighting classes by applying the inverse of the response rate in a class as the weighting factor for that class. The effect of this kind of adjustment is to make the distribution of the weighted sample of respondents across the weighting classes the same as that for the total sample of respondents and nonrespondents. For this reason, the weights will be called sample weights. Sample weighting adjustments are discussed in Section 3.3.

Sometimes the population totals for the weighting classes are known, in which case the weighting factors can be based on them rather than on the sample data. These weights will be called population weights, and population weighting adjustments are discussed in Section 3.2.

When the weighting classes are formed from several characteristics, the number of classes may be large. The application of weighting adjustments independently from one class to another can then result in substantial variability in the weights used and hence in sizeable increases in variances for the survey estimates. The technique of raking ratio estimation discussed in Section 3.4 can be useful in this situation.

It may sometimes be reasonable to assume that nonrespondents have similar characteristics to those respondents whose cooperation in the survey was difficult to secure. If so, the nonresponse weighting adjustments may be restricted to this kind of respondent. Weighting procedures of this type are discussed in Section 3.5.

The final section of the chapter discusses some general considerations in weighting adjustments.

3.2 Population Weighting Adjustments

The application of population weighting adjustments requires that information be available on the distribution of the survey population over the weighting classes. This information may, for instance, come from the previous Census. The data from the respondent sample is then weighted so that its weighted distribution over the classes conforms to that from the external source. As an example, an adjustment is employed with the Current Population Survey (CPS) to bring the survey into line with independent estimates of the U.S. population by age, sex and race, these estimates being obtained by updating the previous Census taking account of births, deaths, aging and migration (Bailar et al., 1977). It should be recognized that the data from the external source will themselves be imperfect in some degree. To the extent that this is so the weighted sample data will not truly represent the distribution of the survey population over the weighting classes. Population weighting adjustments bear a close resemblance to post-stratification (stratification after selection), and are sometimes called post-stratification adjustments. Since there are some differences between the two procedures we prefer to employ the term population weighting adjustments when the respondent sample is weighted to conform to data from an external source.

Two differences between population and sample weighting adjustments deserve mention. First, there is a difference in the data requirements. As noted, population weighting adjustments require knowledge of the population distribution over the weighting classes, but sample weighting adjustments do not. On the other hand, sample weighting adjustments require that both respondents and nonrespondents are divided into weighting classes, whereas population weighting adjustments require only that respondents are so divided. In other words, the characteristics defining the weighting classes need to be
available for respondents and nonrespondents for sample weighting adjustments, but they need be available only for the respondents for population weighting adjustments. The second difference between the two forms of weighting adjustment relates to the type of nonresponse for which they compensate: population weighting adjustments compensate for both noncoverage and unit nonresponse, while sample weighting adjustments compensate only for unit nonresponse. Even when sample weighting adjustments are being used, a good case can be made for employing population weighting adjustments as well in order to deal with noncoverage.

To illustrate the effect of population weighting adjustments on non-response bias, we will consider - following Thomsen (1973) - the estimation of a population mean from an epsem sample in which different classes have differing response rates. Let the population be divided into H classes with \( N_h \) units in class \( h \), let \( N_h = R_h + M_h \), where \( R_h \) and \( M_h \) are the numbers of respondents and nonrespondents respectively in class \( h \) (with the assumption that respondents always respond and nonrespondents never do), and let \( \bar{R}_h = R_h/N_h \) and \( \bar{M}_h = M_h/N_h \) be the expected response and nonresponse rates for class \( h \). The population mean for respondents is

\[
\bar{Y}_r = \frac{\sum Y_{rh}}{R_h},
\]

and that for nonrespondents is

\[
\bar{Y}_m = \frac{\sum Y_{mh}}{M_h}.
\]

The overall population mean, the quantity to be estimated, is

\[
\bar{Y} = \frac{\sum Y_{rh}}{N_h} = \frac{\sum Y_h}{N_h} = \frac{\sum Y_{rh} + \sum Y_{mh}}{N_h} = \bar{Y}_r + \bar{Y}_m.
\]

If the unweighted sample mean \( \bar{Y}_r = \frac{\sum Y_{rh}}{R_h} \) is used to estimate \( \bar{Y} \), its bias is

\[
B(\bar{Y}_r) = \bar{Y} - \bar{Y} = \frac{\sum Y_{rh}}{R_h} - \frac{\sum Y_{rh} + \sum Y_{mh}}{N_h} = \frac{\sum Y_{rh}}{R_h} - \frac{\sum Y_{rh}}{\bar{R}_h} + \frac{\sum Y_{mh}}{N_h} - \frac{\sum Y_{mh}}{\bar{M}_h}.
\]

Since \( \sum Y_{h} (\bar{R}_h - \bar{R}) = 0 \), this bias may also be expressed as

\[
B(\bar{Y}_r) = \left[ \sum Y_{h} (\bar{R}_h - \bar{R}) (\bar{R}_h - \bar{R})/\bar{R} \right] + \sum Y_{h} (\bar{Y}_{rh} - \bar{Y}_{mh})
\]

(3.2.3)

(3.2.4)

The estimator \( \bar{Y}_p \) of \( \bar{Y} \) using the population weighting adjustment is

\[
\bar{Y}_p = \frac{\sum Y_{rh}}{R_h} - \bar{R}_h.
\]

This estimator may be expressed as a weighted estimator

\[
\bar{Y}_p = \frac{\sum w_{hi} Y_{rh}}{\sum w_{hi}},
\]

where the weights \( w_{hi} = w_h = w_{rh}/\bar{R}_h \) compensate for the variation in response rates across the classes. Conditional on \( r_h > 0 \) for all \( h \), the bias of \( \bar{Y}_p \) is given by

\[
B(\bar{Y}_p) = \sum w_{hi} (\bar{Y}_{rh} - \bar{Y}_{mh}).
\]

Using (3.2.1) it can be seen that

\[
B(\bar{Y}_p) = \sum w_{hi} (\bar{Y}_{rh} - \bar{Y}_{mh}) = B.
\]

(3.2.5)

The unweighted mean \( \bar{Y}_r \) is unbiased for \( \bar{Y} \) if \( \bar{Y}_r = \bar{Y}_m \), while the weighted mean \( \bar{Y}_p \) is unbiased if \( \bar{Y}_{rh} = \bar{Y}_{mh} \). Equation (3.2.2) partitions the bias of \( \bar{Y}_r \) into two components, one arising from the variation in response rates across classes (A) and the other arising from the differences between respondent and nonrespondent means within classes (B). The use of \( \bar{Y}_p \) removes the first component of bias, leaving only B. If the components A and B are of the same sign, \( \bar{Y}_p \) cannot have greater absolute bias than \( \bar{Y}_r \). However, if A and B are of different signs, it is possible for \( \bar{Y}_p \) to have a greater absolute bias than \( \bar{Y}_r \); when A and B have different signs, the population weighting adjustment reduces the bias only when |A| > |B|.

The maximum amount of bias reduction through using \( \bar{Y}_p \) rather than \( \bar{Y}_r \) is \(|A|\). As expressed in (3.2.3), A can be seen to be a covariance-type term between class means and response rates. The reason for expressing A in this form is to bring out the two conditions needed for it to be non-zero: (1) that the response rates should differ between classes and (2) that the respondent means should differ between classes.

The resemblance of \( \bar{Y}_p \) to a post-stratified mean \( \bar{Y}_{ps} \) is a useful connection to make, and it will be employed below in deriving the variance of \( \bar{Y}_p \). There are, however, two important differences between the two estimators that should be noted. In the first place, with post-stratification the strata means \( \bar{Y}_h \) are unbiased estimators of \( \bar{Y}_h \) whereas, with population weighting adjustments, the class sample means \( \bar{Y}_h \) are unbiased.
estimators of \( \bar{Y}_{rh} \), the respondent class means; the class sample means are only unbiased estimators of the overall class means if the expected values of the respondent and non-respondent class means are equal. Secondly, the post-stratification weighting adjustment serves to correct for random sampling variability in the spread of the sample over the strata, whereas the population weighting adjustment corrects both for this random sampling variability and for the variation in response rates, and is primarily concerned with the latter. In most cases, post-stratification involves relatively minor weighting adjustments and may be thought of as a fine tuning. The weighting adjustments in \( \hat{Y}_p \) will often be much more variable, and hence have a greater impact on the estimator.

This second difference has consequences for the variances of the two estimators. With a SRS of size \( n \), the variance of the post-stratified mean is approximately

\[
V(\hat{Y}_{ps}) = (1 - f)[\Sigma \frac{E_1(n_h)}{E(n_h)}]/n \tag{3.2.6}
\]

(Kish, 1965, p. 90). If the expected sample sizes from the strata \( E(n_h) = nW_{rh} \), are large - say greater than 10 or 20 - the term in \( (1 - W_{rh})/E(n_h) \) is negligible so that \( V(\hat{Y}_{ps}) \) is approximately the same as that of the mean based on a proportionate stratified sample, and with large strata this latter variance cannot exceed that of the mean from a SRS of the same size. Under similar conditions, \( V(\hat{Y}_p) \) may, however, exceed \( V(\hat{Y}_r) \), as is shown below.

Since, conditional on \( r_h \), \( E_1(\hat{Y}_p) = \Sigma \bar{Y}_{rh} \), a constant, \( V_1E_2(\hat{Y}_p) = 0 \) so that

\[
V(\hat{Y}_p) = E_1V_2(\hat{Y}_p) = E_1[\Sigma \frac{f_r^2}{E_1(n_h)}]/r_h \tag{3.2.7}
\]

Ignoring the f.p.c. term, conditional on \( r \), and using the approximation (Stephan, 1945)

\[
E_1(1/r_h) = (1/rW_{rh}) + (1 - W_{rh})/r^2W_{rh}^2 \tag{3.2.8}
\]

\[
V(\hat{Y}_p) = (\Sigma \frac{f_r^2}{E_1(n_h)}W_{rh}^2) + \Sigma (1 - W_{rh})(\bar{Y}_{rh}/W_{rh})^2S_{rh}^2/r^2 \tag{3.2.9}
\]

Substituting \( W_{rh}/W_{rh} = \bar{R}/\bar{R}_{rh} \),

\[
V(\hat{Y}_p) = \Sigma \frac{f_r^2}{E_1(n_h)}W_{rh}^2S_{rh}^2/r + \Sigma (1 - W_{rh})(\bar{R}/\bar{R}_{rh})^2S_{rh}^2/r^2 \tag{3.2.10}
\]

This variance can then be expressed as

\[
V(\hat{Y}_p) = \Sigma \frac{f_r^2}{E_1(n_h)}W_{rh}^2S_{rh}^2(1 + (1 - W_{rh})/E_1(r_h))/r \tag{3.2.11}
\]

on substituting \( r = E_1(r_h)/W_{rh} \) in the second term of (3.2.10). The form of \( V(\hat{Y}_p) \) in (3.2.6) corresponds to that of \( V(\hat{Y}_{ps}) \) in (3.2.11). In the same way, then, the second term in the parentheses may be neglected if the expected numbers of respondents, \( E_1(r_h) \), are large for all classes, to give the approximate result:

\[
V(\hat{Y}_p) = \Sigma \frac{f_r^2}{E_1(n_h)}W_{rh}^2S_{rh}^2/r \tag{3.2.12}
\]

or equivalently

\[
V(\hat{Y}_p) = \Sigma w_{rh}(\bar{R}/\bar{R}_{rh})rS_{rh}^2/\bar{R}_{rh} \tag{3.2.13}
\]

In order to compare \( V(\hat{Y}_p) \) with \( V(\hat{Y}_r) \), the latter may be expressed, without the f.p.c., as

\[
V(\hat{Y}_r) = S_{rh}^2/r = \Sigma (\bar{Y}_{rh} - \bar{Y}_r)^2 + \Sigma W_{rh}S_{rh}^2/\bar{R}_{rh}^2 \tag{3.2.14}
\]

Hence

\[
V(\hat{Y}_r) - V(\hat{Y}_p) = \Sigma (\bar{Y}_{rh} - \bar{Y}_r)^2/r - \Sigma W_{rh}S_{rh}^2(\bar{R}/\bar{R}_{rh}^2 - 1)/r \tag{3.2.15}
\]

No general results on the sign of the difference \( C - D \) can be established, but some special cases are of interest. First, suppose that the response rate is constant across classes, \( \bar{R}_h = \bar{R} \). Then \( D = 0 \), and since \( C \) is nonnegative the difference
is nonnegative, that is $\tilde{y}_p$ is more precise than $\tilde{y}_r$: this case
 corresponds to the post-stratified sampling result. Secondly, suppose that
 the element variances within the classes are equal, say $S^2_{rh} = S_r^2$. Then
\[ D = S_r^2[(\Sigma w_{rh}(\bar{r}/\bar{r}_h))^2 - 1]/r = S_r^2[(\Sigma w_{rh}^2)/r_h - 1]/r. \]
Now let $w_h = w_{rh} + a_h$, with $\Sigma w_h = \Sigma w_{rh} = 1$ and $\Sigma a_h = 0$. Then
\[ \Sigma w_{rh}^2/w_{rh} = \Sigma w_{rh} + 2\Sigma a_h + \Sigma a_h^2/w_{rh} \]
\[ = 1 + \Sigma a_h^2/w_{rh} \geq 1. \]
Thus $D \geq 0$. In this case, both $C$ and $D$ are nonnegative. It is
entirely possible for $D$ to exceed $C$, as when classes differ
substantially in their response rates but only slightly in their means; when this occurs, the weighted mean has larger
variance than the unweighted one. Thirdly, suppose that the
class means and element variances are equal, $\tilde{y}_{hr} = \tilde{y}_r$ and
\[ S^2_{hr} = S_r^2. \]
Then $C = 0$ and, if the response rates differ, the
difference is negative, that is $\tilde{y}_p$ is less precise than $\tilde{y}_r$.
This result shows that the population weighting adjustment is
harmful to precision when the class means are equal.

Since both $\tilde{y}_p$ and $\tilde{y}_r$ are biased estimators of $\tilde{y}$, it is
more appropriate to compare their mean square errors (MSE's) than
their variances. From earlier results, it then follows that
\[ \text{MSE}(\tilde{y}_r) - \text{MSE}(\tilde{y}_p) = C - D + (A + B)^2 - B^2. \]
The sign and magnitude of this difference depend on the several
parameters involved, so again no general conclusions can be
reached. Note, however, that when the class means are equal,
\[ \tilde{y}_{hr} = \tilde{y}_r, \quad A = C = 0, \]
so that
\[ \text{MSE}(\tilde{y}_p) - \text{MSE}(\tilde{y}_r) = D. \]
If the class element variances are also equal, $D \geq 0$, as shown
above, so the population weighted estimator has larger MSE than
the unweighted estimator in this case.

The precision of $\tilde{y}_p$ obtained from a particular sample may
be measured by its variance conditional on the $r_h$ for that
sample, i.e. by $V_2(\tilde{y}_p)$ as given in parentheses in (3.2.7). An
unbiased estimator for $V_2(\tilde{y}_p)$ may then be calculated from
\[ v(\tilde{y}_p) = \Sigma w_{rh}^2(1 - f_{rh})*s_{rh}^2/r_h, \]
where $s_{rh}^2 = \Sigma(y_{rh} - \bar{y}_{rh})^2/(r_h - 1)$. Since $V(\tilde{y}_p) = E_1V_2(\tilde{y}_p)$,
it then follows that $v(\tilde{y}_p)$ is also unbiased for $V(\tilde{y}_p)$.

A variance estimator for a weighted sample was obtained in
Section 2.3.2 by defining quantities $u_i = w_i/\bar{y}_r$, and treating
the $u_i$ as independent. In the present case, the $u_i$ are not
independent, but the approach may nevertheless still yield a
reasonable variance estimator; if so, it would provide a
simpler variance estimator than above. To investigate
this possibility, let the sample mean be
\[ \bar{y}_p = \Sigma w_{hi}y_{rh}/r = \Sigma u_{hi}/r = \bar{u}, \]
where the weights
\[ w_{hi} = r_{wh}/r_h = u_h \]
sum to $r$, with an average weight of 1. The quantities $u_h$ are independent, $V(\bar{u}) = \sigma^2_u/r$ may be estimated
by $s_u^2/r$ where $s_u^2 = \Sigma(u_{hi} - \bar{u})^2/(r - 1)$. Now
\[ \Sigma(u_{hi} - \bar{u})^2 = \Sigma(u_{hi} - \bar{u}_h)^2 + \Sigma_r(r_{hi} - \bar{u}_h - \bar{u})^2 \]
\[ = r^2\Sigma w_{rh}^2(y_{rh} - \bar{y}_r)^2/r_h^2 + \Sigma_r(w_{rh}^{1/2}y_{rh} - \bar{y}_p)^2. \]
Approximating $r = (r - 1)$, $r_h = (r_h - 1)$, and ignoring the
f.p.c. term,
\[ s_u^2/r = v(\tilde{y}_p) + \Sigma_r(w_{rh}^{1/2}y_{rh} - \bar{y}_p)^2/r^2 \]
\[ = v(\tilde{y}_p) + (G/r). \]
where $G = \Sigma_r(w_{rh}^{1/2}y_{rh} - \bar{y}_p)^2/r$ is the average squared deviation of
the quantities $w_{rh}^{1/2}y_{rh}$ around their mean. If the quantities $w_{rh}^{1/2}y_{rh}$
very considerably, $s_u^2/r$ may substantially overestimate $V(\tilde{y}_p)$.
The use of $s_u^2/r$ cannot therefore be generally recommended.

### 3.3 Sample Weighting Adjustments

With sample weighting adjustments, the total sample of
both respondents and nonrespondents is divided into weighting
classes, and the sample weight for a class is made proportional
to the inverse of the class response rate. The choice of characteristics for defining the weighting classes is limited by the requirement that information on these characteristics must be available for nonrespondents as well as for respondents. One set of characteristics routinely available for both respondents and nonrespondents comprises those used in the sample design, such as the various sampling stage units and strata; weighting classes are often defined in terms of combinations of these characteristics. Thus, for instance, weighting classes can readily be formed in terms of geographical region, degree of urbanicity, etc. Platek et al. (1978) term weighting adjustments based on sample design characteristics "design-dependent balancing factors." In addition, other information may be available about the nonrespondents, perhaps from the sampling frame or perhaps collected by interviewers' observations; with careful training, interviewers can for instance collect almost complete information on the race of a nonresponding household.

With population weighting adjustments, the population mean is estimated by \( \hat{Y}_p = \sum w h \bar{Y}_{rh} \). With sample weighting adjustments the \( W_h \) are unknown, so they are estimated from the total sample as \( w h = n_h/n \). Thus the estimator of the population mean is \( \hat{Y}_s = \sum w h \bar{Y}_{rh} \), which may be alternatively expressed as \( \hat{Y}_s = \Sigma w h \bar{Y}_{rh} / \Sigma w h \) where \( w hi = w h/r h \) rather than \( w hi = w h/r h \) as with population weighting adjustments.

The expected value of \( \hat{Y}_s \) is given by

\[
E(\hat{Y}_s) = E_1 E_2(\hat{Y}_s) = E_1(\Sigma w h \bar{Y}_{rh}) = \Sigma w h \bar{Y}_{rh}
\]

where \( E_2 \) is the expectation conditional on the \( r h \) (assuming \( r h > 0 \) for all classes). Since \( E(\hat{Y}_s) = E(\hat{Y}_p) \), the bias of \( \hat{Y}_s \) is also given by (3.2.5), i.e., \( B(\hat{Y}_s) = B \), and the comparison made between the biases of \( \hat{Y}_p \) and \( \hat{Y}_r \) also applies for the biases of \( \hat{Y}_s \) and \( \hat{Y}_r \).

As population weighting adjustments may be likened to post-stratification, so sample weighting adjustments may be likened to two-phase (or double sampling) stratified estimation. The total sample of size \( n \) can be viewed as the first-phase sample, and the \( r \) respondents as the second-phase sample. The two-phase stratified mean corresponding to \( \hat{Y}_{r} \) is also of the form \( \Sigma w h \bar{Y}_{rh} \), where \( w h = n_h/n \) is an estimate of \( W_h \) based on a large initial sample of size \( n_h \) and where \( \bar{Y}_{rh} = \Sigma w h \bar{Y}_{rh} / r h \) is the mean of the second phase SRS of size \( r_h \) selected from the \( n_h \) units in stratum \( h \) in the first-phase sample. One difference between sample weighting adjustments and two-phase stratified estimation is that the \( r_h \) units are drawn at random in stratum \( h \) with the two-phase sample design, but they are the self-selected group of respondents within the weighting class with the sample weighting adjustment procedure.

In consequence, with the two-phase design, \( \hat{Y}_{r} \) is unbiased for \( \bar{Y}_{r} \), but within the weighting class \( \bar{Y}_{rh} \) is unbiased only for \( \bar{Y}_{rh} \). A second difference is that the sample designer is able to specify the subsampling rate \( r_h/n_h \) with the two-phase design, but ordinarily he is unable to choose the response rate within a weighting class.

The correspondence between sample weighting adjustments and two-phase stratified estimation may be used in deriving the variance of \( \hat{Y}_s \). This derivation, unlike earlier ones, will not be made conditional on \( r \); since \( r = \Sigma r h \) depends on the \( n_h \) which are subject to sampling variability and which appear in the estimator, it is inappropriate to take \( r \) as fixed. A 's' will be used in \( v^s(\hat{Y}_s) \) to indicate the unconditional variance. Three levels of expectation are used \( E = E_1 E_2 E_3 ; E_3 \) is expectation conditional on \( r h \) and \( n_h \); \( E_2 \) is expectation conditional on \( n_h \), and \( E_1 \) is expectation over \( n_h \). Then

\[
v^s(\hat{Y}_s) = E_1 E_2 V_3(\hat{Y}_s) + E_1 V_2 E_3(\hat{Y}_s) + V_1 E_2 E_3(\hat{Y}_s) \quad (3.3.1)
\]

The three terms in (3.3.1) may be evaluated as follows. First, by treating the \( r_h \) respondents in class \( h \) as a SRS from the \( R_h \) population units in that class, it follows that

\[
V_3(\hat{Y}_s) = \Sigma w h \bar{Y}_{rh}^2 / r h
\]

ignoring the f.p.c. term \( [1 - (r h/R_h)] \). Now, approximately from (3.2.8),
\[ E_2(1/r_h) = (1/n_h \bar{R}_h) + (1 - \bar{R}_h)/n_h^2 \bar{s}_h^2 \]  
\[ (3.3.2) \]

in which \( w_h \) may be substituted for \( n_h \). Thus

\[ E_2 V_3(\tilde{y}_s) = (\Sigma w_h s^2 / R_h n) + [\Sigma (1 - \bar{R}_h) s^2 / \bar{R}_h n^2] \]

and

\[ E_1 E_2 V_3(\tilde{y}_s) = (\Sigma w_h s^2 / R_h n) + [\Sigma (1 - \bar{R}_h) s^2 / \bar{R}_h n^2]. \]  
\[ (3.3.3) \]

Next consider the second term in \( (3.3.1) \). Since \( E_3(\tilde{y}_s) = \Sigma w_h \bar{y}_r'h \), a constant when the \( n_h \) are fixed, \( V_2 E_3(\tilde{y}_s) = 0 \), so that

\[ E_1 V_2 E_3(\tilde{y}_s) = 0. \]  
\[ (3.3.4) \]

Finally, \( E_2 E_3(\tilde{y}_s) = \Sigma w_h \bar{y}_r'h \) so that

\[ V_1 E_2 E_3(\tilde{y}_s) = \Sigma w_h \bar{y}_r'h \bar{v}(w_h) + 2 \Sigma v_{ij} r_h r_j c(w_h, w_j), \]

where \( C(w_h, w_j) = w_h w_j/n \) is the covariance of \( w_h \) and \( w_j \) for the initial SRS of size \( n \), ignoring the f.p.c. term. Thus

\[ V_1 E_2 E_3(\tilde{y}_s) = \Sigma w_h (\bar{y}_r'h - \bar{y}_s)^2/n. \]  
\[ (3.3.5) \]

where \( \bar{y}_s = \Sigma w_h \bar{y}_r'h \).

Combining \( (3.3.3) \), \( (3.3.4) \) and \( (3.3.5) \) gives

\[ V^*(\tilde{y}_s) = (\Sigma w_h s^2 / R_h n) + [\Sigma (1 - \bar{R}_h) s^2 / \bar{R}_h n^2] \]
\[ + \Sigma w_h (\bar{y}_r'h - \bar{y}_s)^2/n. \]  
\[ (3.3.6) \]

As discussed earlier in the case of post-stratification, the second term may be neglected if the \( r_h \) are large in every class; neglecting this term is equivalent to dropping the second term in \( (3.3.2) \), i.e. approximating \( E_2(r_h^{-1}) \) by \( (n_h \bar{R}_h)^{-1} \).

To compare \( V^*(\tilde{y}_s) \) with \( V^*(\tilde{y}_p) \), the unconditional variance of \( \tilde{y}_p \) is needed. This unconditional variance may be obtained with two levels of expectation \( E^* = E_1 E_2 \), where \( E_2 \) is conditional on \( r_h \) and \( E_1 \) is expectation over \( r_h \) without the restriction that \( \Sigma r_h = r \) is a constant. Since \( E_2(\tilde{y}_p) = \tilde{y}_p \) is a constant, \( V_1 E_2(\tilde{y}_p) = 0 \). Thus \( V^*(\tilde{y}_p) = E_1 V_2(\tilde{y}_p) \), where \( V_2(\tilde{y}_p) = \Sigma w_h s^2 / r_h \), ignoring the f.p.c. term. Thus using the approximation \( E_1(1/r_h) = n w_h \bar{R}_h \) for large \( r_h \),

\[ V^*(\tilde{y}_p) = \Sigma w_h s^2 / \bar{R}_h n \]

which is the first term in \( (3.3.6) \). To the order of approximation used, the second term in \( (3.3.6) \) is negligible, so that \( V^*(\tilde{y}_s) \) exceeds the unconditional variance of \( \tilde{y}_p \) by the (nonnegative) third term. This result corresponds to one obtained for two-phase sampling in texts such as Raj (1968) and Cochran (1977). Since \( \tilde{y} \) and \( \bar{y} \) have the same bias in estimating \( y \), but \( V^*(\tilde{y}_s) \gtrsim V^*(\tilde{y}_p) \), it follows that MSE* \( (\tilde{y}_s) \) is greater than MSE* \( (\tilde{y}_p) \). There is therefore a greater likelihood that MSE* \( (\tilde{y}_s) \) is larger than MSE* \( (\tilde{y}_r) \) than is the case with MSE* \( (\tilde{y}_p) \) (see the previous section).

To compare \( V^*(\tilde{y}_s) \) with \( V^*(\tilde{y}_r) \), the unconditional variance of \( \tilde{y}_r \) is needed. To derive \( V^*(\tilde{y}_r) \) let \( E^* = E_1 E_2 \), where \( E_2 \) is conditional on \( r \) and \( E_1 \) is expectation over \( r \). Since \( E_2(\tilde{y}_r) = \tilde{y}_r \) is a constant, \( V_1 E_2(\tilde{y}_r) = 0 \). Thus \( V^*(\tilde{y}_r) = E_1 E_2(\tilde{y}_r) = E_1 (S^2_r / r) \), ignoring the f.p.c. term. Using the approximation \( E_1(1/r) = 1/nR \) and writing

\[ S^2_r = \Sigma w_h (\bar{y}_r'h - \bar{y}_r)^2 + \Sigma w_h s^2 \bar{r}_h, \]

\[ V^*(\tilde{y}_r) \] may be expressed as

\[ V^*(\tilde{y}_r) = [\Sigma w_h (\bar{y}_r'h - \bar{y}_r)^2 + \Sigma w_h s^2 / R_h n]. \]  
\[ (3.3.7) \]

As was the case with the comparison of \( V(\tilde{y}_p) \) and \( V(\tilde{y}_r) \), no general conclusions on the sign or magnitude of \( V^*(\tilde{y}_s) - V^*(\tilde{y}_r) \) can be reached.
In the special case that \( Y_2 \) and \( S_2 \) are constant for all classes, \( \tilde{Y}_r = \tilde{Y}_r = \hat{Y}_s \) and \( S_{2r} = S_2 \), and the sample is large so that the second term in \( V^*(\hat{Y}_s) \) in (3.3.6) can be ignored,

\[
V^*(\hat{Y}_s) - V^*(\tilde{Y}_r) = S_{2r}^2 \Sigma[(W_{rh}/R_h) - W_{rh}/R]/n = S_{2r}^2 T/n.
\]

This difference may be shown to be nonnegative by considering the summation term \( T \), which may be expressed as

\[
T = \Sigma[(N_{rh}^2/R_n) - (R_n/N^2)] = (\Sigma N_{rh}^2/R_h) - (N/R).
\]

Since, by the Cauchy-Schwarz inequality,

\[
(\Sigma N_{rh}^2/R_h) \geq (\Sigma N_{rh})^2
\]

i.e.,

\[
(\Sigma N_{rh}^2/R_h) \geq N^2/R,
\]

it follows that \( T \) is nonnegative. Thus, not surprisingly under these particular conditions, the use of variable weights in \( \tilde{Y}_s \) leads to the result that \( V^*(\tilde{Y}_s) \geq V^*(\hat{Y}_s) \).

In another special case, where the response rate is constant across classes, \( \tilde{R}_r = \tilde{R} \), \( W_{rh} = W_h \) and \( \tilde{Y}_r = \hat{Y}_s \), so that the difference in variances of \( \tilde{Y}_r \) and \( \hat{Y}_s \) reduces to

\[
V^*(\tilde{Y}_r) - V^*(\hat{Y}_s) = [(n\tilde{R})^{-1} - n^{-1}] E_{rh} (\tilde{Y}_{rh} - \tilde{Y}_r)^2
\]

which corresponds to the gains of two-phase sampling with proportionate stratification at the second phase.

The variance of \( \hat{Y}_s \) may be estimated by the two-phase stratified sample variance estimator. A good approximate estimator for large samples is

\[
V(\hat{Y}_s) = (E\Sigma W_{rh}^2 S_{2r}^2 / R_h) + E\Sigma W_h (\tilde{Y}_h - \hat{Y}_s)^2 / n \tag{3.3.8}
\]

(see, for instance, Kish, 1965, equation 12.1.4). It is readily apparent that

\[
E\Sigma W_{rh}^2 S_{2r}^2 / R_h = E\Sigma W_{rh}^2 S_{2r}^2 / R_h = V(\hat{Y}_s)
\]

so that this term is an unbiased estimator of \( E_1 E_2 V(Y_{r2}) \). The second term in (3.3.8) is a reasonable estimator of the other component of \( V(\hat{Y}_s) \), that is of

\[
V(1 - E_2) V(\hat{Y}_s) = \Sigma W_h (\tilde{Y}_{r2} - \hat{Y}_s)^2 / n.
\]

This may be demonstrated as follows:

\[
E_1 E_2 V(Y_{r2}) = E_1 E_2 [\Sigma W_{rh} (\tilde{Y}_{rh} - \hat{Y}_{r2}) - V(\hat{Y}_s) - \tilde{Y}_{r2}^2]
\]

\[
= E_1 E_2 [\Sigma W_{rh} (\tilde{Y}_{rh} - \hat{Y}_{r2})^2] - V(\hat{Y}_s) - \tilde{Y}_{r2}^2
\]

\[
= \Sigma W_h (\tilde{Y}_{r2} - \hat{Y}_s)^2 - V(\hat{Y}_s) + E_1 E_2 [\Sigma W_{rh} S_{2r}^2 / R_h]
\]

\[
= \Sigma W_h (\tilde{Y}_{r2} - \hat{Y}_s)^2 - V(\hat{Y}_s) + \Sigma S_{2r}^2 / R_h / n
\]

approximating \( E_2 (1/r_h) = w_h n \tilde{R}_h \). For large \( n \), the last two terms are negligible compared with the first, thus demonstrating that \( \Sigma W_h (\tilde{Y}_{r2} - \hat{Y}_s)^2 \) provides an approximate estimator of \( \Sigma W_h (\tilde{Y}_{r2} - \hat{Y}_s)^2 \). Hence \( V(\hat{Y}_s) \) in (3.3.8) serves as a reasonable estimator of \( V^*(\hat{Y}_s) \).

3.4 Raking Ratio Estimation

Raking ratio estimation, or raking, can be a useful weighting adjustment procedure when the weighting classes are defined in terms of the crosstabulation of a number of characteristics. The underlying basis of the technique is to make the weighted marginal distributions of the weighting class characteristics in the sample conform to those in the population, without ensuring that the weighted sample and population joint distributions conform. Raking was first introduced by Deming and Stephan (1940) — see also Deming (1943). It has subsequently been widely used under the name of iterative proportional fitting (IPF) in contingency table analysis, and recently has been applied by Purcell and Kish (1980) in small domain estimation.

There are two main reasons for using raking rather than separate class weighting adjustments. First, when many classes are formed by the crosstabulations of characteristics, the resulting separate class weighting adjustments can be extremely
variable and unstable because of the sparse spread of the sample over the classes; raking avoids this problem. Secondly, separate class weighting adjustments require knowledge of the joint distribution of the weighting class characteristics for the population, whereas raking only requires knowledge of the population marginal distributions. Many cases occur where population marginal distributions are known, but the joint distribution is unknown, so that raking is possible while separate weighting adjustments are not.

Raking can be applied with many characteristics but for ease of exposition we will consider just two, one with H and the other with K categories. The population and sample joint distributions for these two characteristics are set out in the two bivariate tables below. The population joint distribution over the two characteristics is given by \( w_{hk} \) with marginal distributions \( w_h \) and \( w_{.k} \), and the corresponding sample distributions are \( q_{hk} \) and \( q_{h} \) and \( q_{.k} \). With an epsem sample \( q_{hk} = n_{hk}/n \). The objective of raking is to introduce weights to make the weighted sample marginal distributions equal to \( w_h \) and \( w_{.k} \).

<table>
<thead>
<tr>
<th>Population</th>
<th>1</th>
<th>2</th>
<th>...</th>
<th>K</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>( w_{11} )</td>
<td>( w_{12} )</td>
<td>...</td>
<td>( w_{1K} )</td>
<td>( w_1 )</td>
</tr>
<tr>
<td>2</td>
<td>( w_{21} )</td>
<td>( w_{22} )</td>
<td>...</td>
<td>( w_{2K} )</td>
<td>( w_2 )</td>
</tr>
</tbody>
</table>
| ...        | ... | ... | ... | ... | ...
| H          | \( w_{H1} \) | \( w_{H2} \) | ... | \( w_{HK} \) | \( w_H \) |
| Total      | \( w_1 \) | \( w_2 \) | ... | \( w_K \) | 1     |

<table>
<thead>
<tr>
<th>Sample</th>
<th>1</th>
<th>2</th>
<th>...</th>
<th>K</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>( q_{11} )</td>
<td>( q_{12} )</td>
<td>...</td>
<td>( q_{1K} )</td>
<td>( q_1 )</td>
</tr>
<tr>
<td>2</td>
<td>( q_{21} )</td>
<td>( q_{22} )</td>
<td>...</td>
<td>( q_{2K} )</td>
<td>( q_2 )</td>
</tr>
</tbody>
</table>
| ...        | ... | ... | ... | ... | ...
| H          | \( q_{H1} \) | \( q_{H2} \) | ... | \( q_{HK} \) | \( q_H \) |
| Total      | \( q_1 \) | \( q_2 \) | ... | \( q_K \) | 1     |

Raking employs an iterative procedure to meet this objective; successive modifications are made to the weights until the procedure stabilizes. One algorithm for the procedure is as follows. First, weight each cell in row \( h \) \((h = 1,2,\ldots,H)\) by the factor \( w_h/q_h \), so that the weighted column marginal distribution becomes \( w_{.h} \); as a result of these weights, the sample marginal row distribution has become \( \Sigma q_{hk} w_h/q_h = \Sigma q_{hk}' = q_{.h} \). Next, weight the sampled units in each cell in column \( k \) \((k = 1,2,\ldots,K)\) by the factor \( w_{.k}/q_{.k}' \), so that the weighted row distribution becomes \( w_{.k}' \). As a result of this second weighting, the column distribution is no longer \( w_h \), but has been changed to \( \Sigma w_{hk} w_{.k}/q_{.k} = \Sigma w_{hk}' = q_{.h}'' \). The procedure then returns to the first step applied to the \( q_{hk}' \) values, and continues to iterate until the process converges.

A variant is to continue the process for a small number of cycles, say from one to four, and then to terminate it regardless of whether it has converged to a specified level. The resulting distribution, which will be denoted by \( \hat{w}_{hk} \), is then used in the estimation of the population mean as \( \hat{y} = \Sigma \Sigma \hat{w}_{hk} y_{hk} \). This estimator may be alternatively expressed as \( y = \Sigma \Sigma \Sigma \hat{w}_{hk} y_{hk} / \Sigma \Sigma \hat{w}_{hk} \), where \( w_{hk} = \hat{w}_{hk}/R_h \) and \( R_h \) is the sample size in cell \((h,k)\).

The assumption underlying raking ratio estimation is that the pattern of response rates across all the cells of the table conforms to the following logit model:

\[
\ln \left( \frac{R_{hk}}{1 - R_{hk}} \right) = a_h + b_k
\]

This may be loosely illustrated by the application of the raking algorithm to the respondents' expected joint distribution across the cells, \( w_{rhk} \), under the special case of the logit model when \( R_{hk} = R_h R_k \). The initial two steps in the algorithm are then first to weight the \((h,k)\) cell by \( w_h/q_h = R_h/R_k \), and then by \( w_{.k}/q_{.k} \). Since \( w_{rhk} = w_{hk}' R_h/R_k \), after these two steps the weighted value for cell \((h,k)\) is

\[
w_{rhk} \left[ w_{.k}/(\Sigma w_{rhk} R_h/R_k) \right] = \left( w_{hk}' R_h/R_k \right) \left( R_h/R_k \right) \left( \Sigma \Sigma \hat{w}_{hk} R_h/R_k \right)
\]

\[
= w_{hk} \left( \hat{R}_{hk}/R_h \right) \left[ w_{.k}/(\Sigma \hat{w}_{hk} R_h/R_k) \right]
\]

\[
= w_{hk} \left( \hat{R}_{hk}/R_h \right) \left[ w_{.k}/(\Sigma \hat{w}_{hk} R_h/R_k) \right]
\]
Under the model \( \hat{R}_{hk} = \hat{R}_{jrk} \), it follows that the weighted
value converges immediately to \( \hat{w}_{hk} = \hat{w}_{hk} \), as required, and no
further iterations are needed.

The bias of the raking ratio estimator of the population
mean is given by

\[
B(\hat{y}_q) = E_1E_2(\Sigma \hat{w}_{hk} \hat{y}_{rhk}) - \bar{y} = \Sigma E_1(\hat{w}_{hk}) \hat{y}_{rhk} - \bar{y}
\]

Letting \( E_1(\hat{w}_{hk}) = \hat{w}_{hk} \), and employing (3.2.1),

\[
B(\hat{y}_q) = \Sigma \hat{w}_{hk} \bar{y}_{rhk} - \Sigma \hat{w}_{hk} \bar{y}_{mhk} (\bar{y}_{rhk} - \bar{y}_{mhk})
\]

i.e.,

\[
B(\hat{y}_q) = \Sigma \hat{w}_{hk} \bar{y}_{rhk} - \Sigma \hat{w}_{hk} \bar{y}_{mhk} + \Sigma (\hat{w}_{hk} - \hat{w}_{hk}) \bar{y}_{rhk}
\]

(3.4.1)

The form of the bias in (3.4.1) brings out its dependence on
two factors: the magnitude of the first term, the bias of a
population or a sample weighting estimator applied to the cells
of the crossclassification, depends on the differences between
respondent and nonrespondent means in the cells; the second
term depends on the closeness of the model estimates of the
population proportions in the cells, \( \hat{w}_{hk} \), and the true
population proportions in the cells, \( \hat{w}_{hk} \). Further insight into
the second term can be gained from expressing \( \hat{y}_{rhk} \)
as

\[
\hat{y}_{rhk} = \bar{y}_{r} + (\bar{y}_{rh} - \bar{y}_{r}) + (\bar{y}_{r.k} - \bar{y}_{r}) + (\bar{y}_{r.hk} - \bar{y}_{r.h} + \bar{y}_{r}).
\]

Then, noting that \( E(\hat{w}_{hk}) = \hat{w}_{hk} \) and \( \Sigma \hat{w}_{hk} = \hat{w}_{hk} \), with the raking
procedure, the second term in (3.4.1) reduces to

\[
\Sigma (\hat{w}_{hk} - \hat{w}_{hk}) (\hat{y}_{rhk} - \bar{y}_{rh} - \bar{y}_{r.k} + \bar{y}).
\]

(3.4.2)

Thus, even if \( \hat{w}_{hk} \) and \( \hat{w}_{hk} \) were different, this term would be
zero if there were no interaction in the \( \hat{y}_{rhk} \) for the
classifications involved.

The variance of \( \hat{y}_q \) conditional on the \( r_{hk} \)'s is obtained
straightforwardly as

\[
V_2(\hat{y}_q) = \Sigma \hat{w}_{hk}^2 (1 - f_{hk}) \Sigma \hat{y}_{rhk}^2 / \hat{y}_{rhk}
\]

where \( f_{hk} = 1 - (\hat{f}_{hk} / \hat{f}_{rk}) \). The unconditional variance is
complex even for SRS (see, for instance, Brackstone and Rao,
1980).

Given the joint distribution of the characteristics to be
included in a population adjustment procedure, the analyst has
the choice between separate cell-by-cell adjustments and
raking. An aspect of that choice includes the possibility that
some cells may be collapsed with the population adjustment
procedure to give greater stability of weights. The
possibility also exists of using raking for some cells, for
instance those with few cases, and using population weighting
adjustments for the remainder. The choice involves
consideration of a trade-off between bias and variance, and no
general rules can be laid down. The best choice depends on the
circumstances and, in particular, on the suitability of the
raking model for the data being analyzed.

Although primarily described here in terms of marginal
population weighting adjustments, raking can also be applied
with sample weighting adjustments: the marginal sample
distributions of the defining characteristics are determined
for the total sample of respondents and nonrespondents
combined, i.e. \( n_i/n \) and \( n_j/n \), and then the raking algorithm
is applied as above.

Raking can be employed for marginal post-stratification
adjustments as well as for nonresponse weighting adjustments.
References on raking include Arora and Brackstone (1977a,b),
Bankier (1978), Oh and Scheuren (1978a,b), Brackstone and Rao

3.5 Weighting Using Response Probabilities

Several different approaches have been suggested for
employing information about the difficulty of securing
responses from respondents in nonresponse weighting
adjustments. The well-known Politz-Simmons procedure is one
element of the Politz-Simmons method (Politz and Simmons, 1959).

The essence of this approach is to group respondents into classes according to
their response probabilities, and then to weight the classes by
the inverses of these probabilities. The procedure was
designed to obviate the need for more than one call on each
sample unit. During the interview respondents are asked on how
many of the previous five evenings they were home at about that
time. Those at home every evening were then assigned response
probabilities of 1, those home on 4 of the 5 evenings were
assigned probabilities of 5/6, etc., and those home on none of
the previous 5 evenings were assigned probabilities of 1/6 (the
evening they were interviewed being their one chance in six of
being interviewed). On the grounds that only a fraction t/6 of
those in t/6 evenings will be found at home on one call, this
class is then weighted by the factor 6/t. Note that the
procedure fails to account for those in none of the six
evenings, since none of them are interviewed; they should,
however, constitute a small proportion of the total sample. It
should also be observed that the Politz-Simmons procedure is
designed to deal with not-at-homes, not refusals. However, as
shown in Chapter 1, the main unit nonresponse problem facing
the 1978 Research Panel was refusals.

The Politz-Simmons procedure requires respondents to
provide the information on their own response probabilities. An
extension to this procedure avoids the need to collect this
information. It involves making call-backs on nonrespondents at
earlier calls, modelling the response probabilities at
successive calls, and estimating the parameters of the model
from the observed pattern of responses at successive calls. This
approach can be carried out within different classes of
the sample to allow for differing parameters across the
classes.

Drew and Fuller (1980, 1981) describe one application of
this approach. Suppose that data can be assumed to be missing
at random within classes of a characteristic, but at different
rates between classes, so that the characteristic would be
useful for weighting adjustments. If the proportion of
the population in each class, \(W_h\), was known, a population weighting
adjustment could be used, estimating the population mean by
\(\hat{Y}_S = \sum W_h \hat{Y}_h\). If neither of these sets of
information is available, the Drew and Fuller model provides
means of estimating \(\hat{W}_h\), \(\hat{Y}_h\), which can then be used in a
weighting adjustment as \(\sum \hat{W}_h \hat{Y}_h\). The Drew and Fuller model
assumes that each unit in class \(h\) has the same response
probability, \(q_h\), and that \(q_h\) remains the same for each
successive call; it also allows for a proportion of hard-core
nonrespondents, \(\gamma\), which is assumed constant across classes.
Given these assumptions, the number of responses obtained at
each successive call follows a multinomial distribution. The
parameters \(q_h\), \(\gamma\) and \(W_h\) can then be estimated by fitting the
model to the observed data of the number of responses at each
call for each class of the characteristic.

Thomsen and Siring (1979) adopt a similar approach to Drew
and Fuller, but with a more complex model. They assume that
within a class a unit has a probability of \(p_h\) of responding at
the first call, and of \(\delta_{h2}p_h\) at subsequent calls; they expect \(\delta_{h2}\)
to exceed 1 because interviewers can make appointments, etc.,
for second and subsequent calls. They also allow for a
proportion of hard-core nonrespondents. They give an
application in which the weighting class characteristic is
itself the variable of interest: they estimate the mean number
of live births per woman, and form classes of women with 0, 1,
..., 6 or more births. This can be viewed as an interesting
special case of the general procedure with \(\hat{Y}_h = (h - 1)\).

Chapman (1976) describes the use of regression analysis to
estimate response probabilities. A response variable (1 for
respondents, 0 for nonrespondents) is regressed on a set of
variables available for both respondents and nonrespondents,
and the predicted values of the response variable are computed
for the respondents. These predicted values estimate the
response probability for units with the same values on the
independent variables, and their inverses are then used as
weights for the respondents. As described, this approach makes
strong assumptions about the linearity and additivity of the
effects of the independent variables on the response
probabilities. It would, however, be possible to modify the
model to take account of some nonlinearities and interactions.
Another type of approach using information on the differential difficulty of securing responses from different respondents is to assume that difficulty is correlated with the survey variables. This leads to two possible procedures for handling nonresponse.

One is to assume that the nonrespondents have similar values on the survey variables to those respondents who were hardest to interview—commonly taken as those who were interviewed only after a number of call-backs. Bartholomew (1961), for instance, has suggested making only two calls in a survey, and weighting up the responses obtained at the second call to compensate for the nonrespondents. This kind of procedure would be effective in reducing bias if the basic assumption holds. However this is often questionable; in particular the nonrespondents comprise two groups, not-at-homes and refusals, and it is not obvious that the late respondents can properly represent the refusals. Based on an intensive follow up study of nonrespondents in the Current Population Survey, Palmer and Jones (1966) and Palmer (1967) found no support for the assumption. Another problem with this procedure is that there is typically only a small number of late respondents, who then need to be assigned sizeable weights if they are to compensate for the nonrespondents. This factor considerably increases the variances of the survey estimates.

The second possible approach for handling nonresponse taking account of a relationship between the survey variables and difficulty in securing responses is to plot the means of the survey variables against difficulty, probably measured by number of calls needed. Any trend observed could then be extrapolated to estimate the mean for the nonrespondents. This procedure assumes that nonrespondents follow along the trend after the last respondents, often a dubious assumption for the reasons given above. Chapman (1976) reports that in his investigation of such trends for many items in the Health and Nutrition Examination Survey there was no general pattern to the trends observed, and in most cases the trends were not clear enough to be used for projecting the means for nonrespondents.

3.6 Concluding Remarks

So far the important issue of how to choose the weighting classes has been mentioned only in passing. The first consideration in making this choice is to ascertain what variables are available for use, that is what variables satisfy the data requirements for a weighting procedure. Population weighting adjustments require knowledge of the population distribution of the variable and information on the variable for survey respondents. Sample weighting adjustments require information on the variable to be available for both respondents and nonrespondents.

Among the potential variables for use in forming weighting classes, the ones that are most effective in reducing nonresponse bias are those that are highly correlated both with the survey variables and with the \((0 - 1)\) response variable (see the discussion in Section 3.2). Since a survey is concerned with numerous survey variables, which may have markedly different correlations with any potential weighting class variable, this aspect of the choice may be confusing unless a few closely-related survey variables can be identified as the main concern. The correlation between a potential weighting class variable and the response variable is a single criterion that applies no matter which survey variable is being analyzed. In consequence considerable importance may be attached to this correlation in making the choice.

Weighting classes can be defined in terms of a combination of variables, and this is indeed common practice. The classes may be simply defined as the cells of a crossclassification of the variables, but there is no need for this balanced approach. Instead some variables may be used to define weighting classes in one part of the sample while others can be used in other parts. Also, if the crossclassification approach is used, some cells may be collapsed to form weighting classes of adequate size. The position corresponds to that involved in choosing strata for stratified sampling, where the strata may be defined in any way providing the required information is available.

The number of variables that can be used in forming weighting classes is limited by the sizes of the resulting classes. If the sample sizes of respondents in the classes are
too small, the effect of the resulting unstable weighting will be to substantially inflate the variances of the survey estimates, more than would be compensated for by any bias reduction. Rules of a minimum of 20 or 25 respondents per class have been used (Chapman, 1979). Another condition sometimes imposed is that any class that requires a weighting factor $n_h/k_h$ greater than a specified level should be combined with an adjacent class; this condition also serves to limit the number of weighting classes. In the Current Population Survey (CPS), for instance, adjacent weighting classes are combined if at least one of the classes requires a weighting factor greater than 2.0. In addition, if after combining two adjacent classes in the CPS the weighting factor exceeds 3.0, it is trimmed down to 3.0 (Bailar et al., 1978; Brooks and Bailar, 1978). These procedures in effect accept some potential loss in bias reduction in order to avoid the increase in variance associated with large variability in weights.

For convenience of exposition, this chapter has discussed population and sampling weighting adjustments separately, but it should be noted that they can be—and often are—applied together. First, sample weighting adjustments may be employed to bring the achieved sample into line with the total sample, primarily in terms of the sample design variables, and then population weighting adjustments can be applied, partly to handle noncoverage and total nonresponse and partly to serve as post-stratification.

4. Imputation

4.1 Introduction

As noted in Chapter 1, most surveys face problems of item nonresponse as well as unit nonresponse. Item nonresponse occurs when no answer is recorded for a question that a respondent should have answered, and when an answer is eliminated by an edit check because it is inconsistent with other answers. Imputation is the process of assigning values for the missing values to produce a complete data set.

There are several benefits to imputation. In the first place, like weighting adjustments, the aim of imputation is to reduce biases in survey estimates. Secondly, imputation makes analysis easier and the results simpler to present. Complex algorithms to estimate population parameters in the presence of missing data are not required, and hence much processing time is saved. There is neither a need to determine the different sets of cases with missing data that have to be deleted from different analyses, nor a need to provide details of the extent and treatment of missing data with each set of results. Thirdly, the results obtained from different analyses are bound to be consistent with one another, a feature which need not apply to results of analyses from an incomplete data set. In particular, the practice of deleting cases with missing values on any of the classifications involved in forming a tabulation from an incomplete data set can give rise to inconsistencies in marginal distributions between tabulations having some classifications in common.

On the other hand, imputation has its dangers. There can be no guarantee that the results obtained after imputation will be less biased than those based on the incomplete data set, and indeed the biases could be greater. It all depends on the suitability of the assumptions built into the imputation procedures used. Even if the biases of univariate statistics are reduced, the relationship between variables may be distorted (see the companion report by Santos, 1981a, which examines this issue). Rockwell (1975) shows, for instance, how this point applies to relationships between the variables subject to imputation and the variables employed in making the
imputations. Another danger with imputation is that researchers may falsely treat the completed data set for \( n \) respondents as if it were from a straightforward sample of size \( n \). Imputation merely fabricates data; it does not increase the amount of information available, and indeed the use of a random subsampling of donors serves to reduce, not to increase, the effective sample size. It is sometimes argued that little risk is involved when imputation is needed for only a small fraction of cases; however, even in this situation, some subgroups of the sample may have sizeable proportions of missing data, and hence their results may significantly depend on the imputed values. (See Banister, 1980, for some concerns about the widespread use of imputation with census data.)

Undoubtedly, when used, imputation should be applied cautiously and analysts of the completed data set should be fully warned of the potential dangers created by the imputation. At a minimum, the imputed values should be flagged, so that the careful analyst can assess the effect that imputations may have on the analysis. I.G. Sande (1979a) has suggested a number of additional aspects of the imputation process that need to be audited in order to keep track of the impact of imputation; these include such aspects as the number of times a particular record is used as a donor of imputed values in a particular stage, the number of attempts made to achieve successful imputations for a particular record with missing data, and information on which donors contribute which fields to which recipients. A second reason for flagging imputed values is to enable an analyst to replace them with his own estimates, or to treat the missing data otherwise in his analyses. Imputation is intended to be a general-purpose strategy, good for many forms of analysis but not necessarily optimum for anyone (in this sense it is like Tukey’s description of the jackknife as a technique of variance estimation). An analyst who has the resources may prefer to ignore the imputed values and instead treat the incomplete data set in a way that is tailor-made to his particular problem.

Given that imputed values are to be substituted for missing responses, there are a variety of methods in which the imputed values may be determined. A selection of these methods is reviewed in the next section, and the results of a simulation study applying some of them to a data set derived from the April 1978 wave of the 1978 ISDP Research Panel are presented in Section 4.3.

Many of the imputation methods divide the sample into imputation classes corresponding to the weighting classes discussed in Chapter 3. The range of choice for imputation classes is far greater than that for weighting classes because many more data are available for the units with item nonresponses. A discussion of methods for choosing imputation classes is given in Section 4.4. That section also discusses a number of other issues that occur with imputation. These include issues arising from using respondents as donors on several occasions, the need to ensure that imputed values are consistent with the unit’s other responses, and the considerations involved when a record has several missing items.

4.2 Imputation Methods

The purpose of this section is to review a range of methods of imputing for missing values. Since there are several variants possible for many of the basic methods, the review does not attempt an exhaustive coverage but rather aims to describe some of the main approaches. For simplicity, the descriptions will be given in terms of missing responses for a single item; the complexities introduced by records having several missing values are taken up in Section 4.3.

The development of imputation methods is of recent origin, and in consequence no universally accepted terminology has yet been established. It will therefore be useful to start by defining some of the terms employed in the following discussion. First, most of the methods start with the division of the sample into classes, which we will call imputation classes. Imputation classes are equivalent to the weighting classes discussed in the previous chapter. The variables used to define the imputation classes, or otherwise used in determining the imputed values, will be termed control or matching variables. Many methods assign the value from a record with a response on the item in question to a record with a missing value on that item. These records will be called the
donor and recipient respectively. Any method that takes the donors from within the same sample as the recipients is often called a "hot-deck" method; however, for the present discussion the term hot-deck is reserved for a particular method using this approach, a sequential method as used, for instance, with the labor force items in the Current Population Survey (Brooks and Bailar, 1978).

It is useful to make a distinction between qualitative and quantitative control variables. The formation of imputation classes in terms of qualitative control variables is in principle straightforward, but when quantitative variables are used, they first need to be categorized into a number of classes, necessarily involving some loss of information. The loss of information is in fact relatively slight providing a reasonable number of well-chosen classes is used, say 5 or more (see, for instance, Cochran, 1968; Aigner et al., 1975; Kish and Anderson, 1978). However, the categorization can be avoided by the use of alternative imputation methods not involving imputation classes, such as methods based on minimizing a distance function between donor and recipient in terms of the control variables or methods using regression for determining the imputed values. These methods are discussed in (g) and (h) below.

We will now turn to descriptions of a selection of methods available for making imputations for missing values. Some of these methods have in fact already been introduced in Chapter 2, where they have been considered in relation to the single imputation class comprising the sample taken as a whole. These methods will be repeated here briefly for the sake of completeness.

(a) Deductive imputation. In its pure form this method of imputation can be applied in situations in which a missing response can be deduced with certainty, or with high probability, from other information on the record. Thus, for instance, if a respondent's sex is missing, but the person has a female name and is known to be married to a male, the sex of the respondent may be deduced to be female. If a record should contain a series of amounts together with a total but one of the quantities is missing, the missing value can be deduced from the others. This method essentially depends on some redundancy in the information collected so that edit constraints can determine the missing values on some items.

Deductive imputation is the special case of other imputation methods in which the control variables determine the missing values with negligible error; it merges into other types of imputation as the amount of error increases. This may be illustrated by another example, taken from the 1978 ISDP Research Panel. Wage earners were asked to report their quarterly earnings from their records (E) and were also asked to give their hourly rates of pay (P), their usual numbers of hours worked (H) and the numbers of weeks worked in the quarter (W). In cases where they were unable or unwilling to consult their records, their missing earnings could be estimated by deductive imputation to be \( E' = (P \times H \times W) \). This deductive process is, however, an imperfect one, as is evidenced by the discrepancies found between E and E' for those wage earners who provided both sets of data (discussed further below). Deductive imputation ignores these discrepancies; if they are too substantial to be ignored, other imputation methods are needed to deal with them.

(b) Cold-deck imputation. The cold-deck imputation method was an early form of computerized imputation which has now been superseded by hot-deck methods. It is mentioned here mainly for the sake of completeness.

The essence of the method involves the specification of imputation classes, often defined as the cells of a cross-classification of the control variables, and then, for each class, the provision of a set of values for the responses to the item in question. These values come from some past data set for the same population; with a periodic survey, they may be obtained from the respondents within the imputation classes on the previous round of the survey. Once this cold-deck has been established, the current survey records are considered in turn. If a record has a missing response for the item, it is assigned a value from the cold-deck. Thus a complete data set is readily produced.

The evident objection to the cold-deck method is the potential lack of comparability between the past values...
inserted into the cold deck and the present survey responses. This lack of comparability might arise both because of obsolescence and because of differences in procedures and definitions used. The obvious solution to this problem is to employ a method based on current values rather than past values for the imputation process. This leads to the hot-deck method of imputation as described in (d) below.

(c) Mean-value imputation. The mean-value imputation procedure consists of dividing the sample into imputation classes, computing the respondent mean within each class, and assigning the class mean value to all records in the class with missing data for the item. This procedure has already been discussed in Chapter 2, where it was shown to be a good procedure for the estimation of the item mean. However its distortion of the distribution of values makes it unsuitable for many other forms of analysis. These conclusions also apply to the more general situation of several imputation classes now being considered.

(d) Hot-deck imputation. Contrary to some other usages, the term 'hot-deck' imputation will be used here only for the sequential kind of procedure used with the labor force items in the Current Population Survey (CPS), sometimes known as the traditional hot-deck procedure. This procedure begins with the specification of imputation classes, and with the assignment of a single cold-deck value for each class to provide a starting point for the process. The records of the current survey are then treated sequentially. If a record has a response for that item, that response replaces the value stored for its imputation class at that point; if the record has a missing response it is assigned the value stored for its imputation class.

This hot-deck procedure was developed to remedy the disadvantage of the cold-deck by using the survey's respondents as donors, while at the same time enabling the imputations to be made from a single pass through the data file. Another argument that is made in favor of the sequential nature of the procedure is that the imputation of a value to a recipient from the last donor in his imputation class serves as an additional degree of matching, which will be beneficial if the file is ordered in a way that creates positive autocorrelation.

The hot-deck procedure has, however, some serious disadvantages. One is its dependence on the order of the file without a probability mechanism: this feature makes it impossible to develop a model-free theoretical evaluation of the procedure. Secondly, the procedure may easily give rise to multiple uses of donors. This event occurs whenever within a given imputation class a record with missing data is followed by one or more other records with missing data. All these records are assigned the value then being stored for the imputation class, namely the value of the last respondent. Thirdly, the number of imputation classes has to be limited to ensure that one or more donors will be found for each class containing a record with a missing value.

Useful discussions of the hot-deck procedure are provided by Bailar et al. (1978), Bailar and Bailar (1978, 1979), Ford (1980), Oh and Scheuren (1980), Oh et al. (1980) and I.G. Sande (1979a,b).

(e) Random imputation. Random imputation is used here to denote a range of procedures in which donors are selected by some form of probability sampling within an imputation class. A number of these procedures have already been described in Chapter 2 for a single imputation class. When there is more than one imputation class, the first step is to identify the \( n_h \) sampled units in imputation class \( h \), and to separate them into \( r_h \) potential donors (units with responses to the item) and \( m_h \) recipients (units with missing responses to the item). Then one of the procedures described in Chapter 2 can be applied within each imputation class.

To recap, there are two main features to be determined in arriving at a probability sampling scheme for donors. One is the sample size and the other is the sample design. The natural choice for sample size is \( m_h \), allocating to each recipient the value from one donor. Given this sample size, a simple sample design would be to select one donor at random from the \( r_h \) potential donors for each recipient, that is to sample donors by unrestricted sampling. This scheme of sampling with replacement gives rise, however, to an undesirable variability in the use of donors. This variability can be controlled by simple random sampling, sampling without
replacement, using the procedure described in Chapter 2 of

taking each donor \( k_h \) times and a SRS of \( t_h \) donors for an extra

occasion, where \( m_h = k_h + t_h \) and \( k_h \) is an integer (if, as

often applies, \( m_h < t_h \), \( k_h = 0 \)). Providing \( m_h > 1 \), the sample

design can be further improved by the use of stratification (in

particular stratifying potential donors by their item

responses) or by systematic sampling from a list of potential

donors ordered by their item responses. The maximum number of

strata - or implicit strata with systematic sampling - that can

be employed is \( t_h \).

The donor sample size for random imputation (or indeed for

other imputation procedures) need not be set at \( m_h = k_h + t_h \),

with \( (r_h - t_h) \) respondents acting as donors to \( k_h \) recipients

each and \( t_h \) respondents acting as donors to \( (k_h + 1) \) recipients. A more general approach is to take a sample of

\( c_h \) respondents to serve as donors on the additional

occasion, with \( c_h \) a positive integer. Then \( k_h \) of the \( m_h \)

recipients are given values from a single donor, and the other

\( t_h \) are given values from \( c_h \) donors each. The multiple donor

cases may be handled by splitting the records of each of the

recipients involved into \( c_h \) parts, allocating one of the \( c_h \)

donors' values to each part, and assigning a weight of \( 1/c_h \) to

each part. (Note that if several parts for one recipient

happen to receive the same value, they can be recombined with

their weights added.) The advantage of using this type of

multiple imputation is the reduction in variance increase

associated with using a sample of respondents as donors for an

additional occasion (see equation 2.3.1.13).

A variant of the above procedure that has been proposed in

the literature is to select a sample of donors for each

recipient and to assign to each recipient the mean value for

his donor sample. This procedure has the same effect for

estimating the mean and total of an item as the weight-

splitting procedure described above. The spread of the
distribution of imputed values obtained from this procedure is,

however, reduced since it is a distribution of means, and hence

the overall distribution of respondents' values and imputed

values will be distorted. The limiting case of this procedure

is the mean-value imputation procedure described in (c) above.

Its disadvantages are the same as those of mean-value

imputation, only in a reduced form.

Consider now the bias and variance of a sample mean \( \bar{y}_k \)

when random imputation has been used to fill in the missing

values. We assume that the original sample is an epsem one and

that \( \bar{y}_k = \frac{\sum y_{hi}}{n} = \frac{\sum y_{rh} + \sum y_{mh}}{n} \), where \( y_{hi} \) is the

\( y \)-value for responding unit \( i \) in imputation class \( h \) and \( y_{mh} \) is

the imputed \( y \)-value for a non-responding unit \( i \) in imputation

class \( h \). The expectation of \( \bar{y}_k \) may be expressed as \( E_1 E_2 (\bar{y}_k) \)

where \( E_2 \) is the expectation conditional on \( n_h \), \( r_h \) and the

responding units' values \( y_{rh} \), and \( E_1 \) is expectation over

variations in these values. Then

\[
E(\bar{y}_k) = E_1 E_2 (\frac{\sum y_{rh} + \sum y_{mh}}{n}).
\]

For any epsem sampling scheme for choosing donors,

\( E_2 (y_{mh}) = \bar{y}_{rh} \), so that

\[
E(\bar{y}_k) = E_1 (\frac{\sum y_{rh}}{n} + \frac{\sum \bar{y}_{rh}}{n})
= E_1 (\frac{\sum \bar{y}_{rh}}{n}) = E_1 (\bar{y}_s)
\]

where \( \bar{y}_s \) is the mean with sample weighting adjustments. The

expectation of \( \bar{y}_s \), as obtained in Section 3.3, is \( \Sigma_h n_h \bar{y}_{rh} \), so that

\[
E(\bar{y}_k) = \Sigma_h n_h \bar{y}_{rh}.
\]

The bias of \( \bar{y}_k \) is the same as that of \( \bar{y}_s \) and \( \bar{y}_p \), the mean with

population weighting adjustments. See Section 3.2 for a

discussion of this bias.

The unconditional variance of \( \bar{y}_k \) may be obtained from

\[
V^*(\bar{y}_k) = E_1 V_2 (\bar{y}_k) + V_1 E_2 (\bar{y}_k)
= E_1 V_2 (\bar{y}_k) + V_1 (\bar{y}_s).
\]

The second term in (4.2.1) is the variance of the mean with

simple weighting adjustments, \( V^*(\bar{y}_s) \). This variance is derived

for an initial SRS of \( n \) units in Section 3.3, with \( V^*(\bar{y}_s) \) being

given by (3.3.6). The first term in (4.2.1) is the additional
variance arising from the sampling of respondents to be used as donors: the conditional variance \( V_2(\tilde{y}_k) \) for various sampling schemes has been considered in Section 2.3.1. One such scheme, for example, is the SRS imputation scheme (within each class). In this case

\[
V_2(\tilde{y}_k) = \sum_{h=1}^{n} (t_h - \bar{r}_h)^2 / n^2
\]

(4.2.2)
as may be readily obtained as a generalization of (2.3.1.5). As noted above, this variance may be reduced by employing a stratified or systematic sampling scheme and by increasing the sample size.

A reasonable estimator of \( V^*(\tilde{y}_k) \) may be readily derived as follows. First, as shown in Section 3.3, \( v(\tilde{y}_s) \) in equation (3.3.8) serves as a good estimator of \( V^*(\tilde{y}_s) \). Secondly, \( V_2(\tilde{y}_k) \) can be computed directly from (4.2.2). Hence

\[
v(\tilde{y}_k) = V_2(\tilde{y}_k) + v(\tilde{y}_s)
\]

may be used to estimate \( V^*(\tilde{y}_k) \).

(f) Flexible matching imputation. The term flexible matching imputation has been coined here to denote the modified hot-deck procedure now being used by the Bureau of the Census for the March Income Supplement of the CPS. The essence of this sophisticated procedure is a sorting and matching of potential donors and recipients on a sizeable number of control variables. The matching is done on a hierarchical basis, in the sense that if no donor can be found to match a recipient on all the control variables, some of the control variables considered less important are dropped or their detail of classification is reduced, and the match is made at a lower level. Three hierarchical levels are defined for this purpose, with the lowest level being such that a matching donor can always be found. This flexible matching procedure enables far closer matches to be secured for many recipients than does the conventional hot-deck procedure. The procedure also avoids the multiple use of donors in classes where the number of donors is not less than the number of recipients. Further details of the implementation and evaluation of the procedure with the March Income Supplement of the CPS are given by Coder (1978) and Welniak and Coder (1980).

(g) Distance function matching. Matching on quantitative control variables may be carried out by categorizing the variables and forming imputation classes. This approach can, however, lead to some less desirable matches: for instance, a recipient with a value on the control variable close to the upper bound of a category may be matched with a donor close to the category's lower bound, whereas he might have been better matched with a donor near the lower bound of the next higher category. The step function effect created by fixing arbitrary boundaries for categories can be avoided by the use of some form of distance function to measure the "closeness" of the match.

When a single quantitative control variable is involved, the merged set of recipients and potential donors can be ordered by it, and the "nearest" neighboring donor to a recipient could be taken. "Nearest" can be defined in terms of the absolute difference between recipient's and donor's values on the control variable, or alternatively in terms of the absolute difference on some transformation of the control variable; for example, using a logarithmic transformation in effect measures "nearest" in terms of proportionate differences. The distance function can be constructed to include a component to reduce the multiple use of donors. Colledge et al. (1978), for instance, define the distance between a potential donor and a recipient to be \( D(1 + p d) \), where \( D \) is the basic distance, \( d \) is the number of times the donor has already been used, and \( p \) is a proportional penalty for each usage (they give \( p = 0.02 \) as an example). A variation of the distance function procedure is to assign the recipient the average value for a set of nearest neighbors, a simple example of which is to average the values of the neighboring donors on either side of the recipient (Ford, 1976). As with other averaging procedures, this procedure suffers the disadvantage of distorting distributions.

When several control variables are used, but only one is quantitative, the above procedure may be applied for the quantitative variable within imputation classes constructed
from the qualitative variables. When there is more than one quantitative variable, the distance function can be generalized. In this case, attention needs to be paid to the metrics and distributions of the variables. Transformations of the variables are almost certainly needed. In particular, distances in the tails of a distribution where observations are sparse are likely to dominate the overall distance function unless a suitable transformation is made. One way of eliminating this effect is to transform all the quantitative control variables to ranks, and define distances in terms of ranks (suggested by G. Sande), or equivalently to transform the control variables to a uniform distribution. One form of distance function between recipient r and potential donor d might then be

$$\text{Sup}_k \{ w_k (R_{rk} - R_{dk}) \}$$

where $R_{rk}$ and $R_{dk}$ are the ranks of the recipient and donor on variable $k$, and $w_k$ is a weight representing the relative importance of variable $k$ in the distance function (see I.G. Sande, 1979b). Vacek and Ashikaga (1980) suggest an approach based on the Mahalanobis distance. Qualitative variables can also be incorporated into a distance function, as an alternative to controlling for them by means of imputation classes.

(h) Regression imputation. With this procedure the item for which imputations are needed (y) is regressed on the control variables ($x_1, x_2, \ldots, x_p$) for the units providing a response on y. The control variables may be quantitative or qualitative, the latter being incorporated into the regression model by means of dummy variables. The y-variable is qualitative, log-linear or logistic models may be employed. The missing values may then be imputed in one of two basic ways: one is to use the predicted value from the model given the values on the control variables for the record with a missing item response; the other is to use this predicted value plus some type of randomly chosen residual.

It will be helpful at this point to compare regression imputation with imputation procedures based on imputation classes. For this purpose, we will consider a situation with two quantitative control variables, each with a few discrete scores so that no categorization is needed for forming imputation classes (one variable, for instance, might be number of children in the household, ranging from, say, 1 to 4). The model underlying procedures based on imputation classes is

$$Y_{ijk} = \bar{Y}_{ij} + e_{ijk}$$

(4.2.3)

where $Y_{ijk}$ is the value for unit k in the class represented by the intersection of the i-th level of control variable $x_1$ and the j-th level of control variable $x_2$, $\bar{Y}_{ij}$ is the population class mean and $e_{ijk}$ is a residual representing the departure of the values of unit k from the class mean with $E(e_{ijk} | ij) = 0$. The distribution of errors is not specified, and it may differ between classes. With this model, the predicted value of $Y_{ijk}$ is

$$\hat{Y}_{ijk} = \bar{Y}_{ij} + \hat{e}_{ijk},$$

that is the mean-value imputation as discussed in (c) above. To retain the distributional properties of the $Y_{ijk}$ it is necessary to include the residual term, i.e. to estimate $(\bar{Y}_{ij} + e_{ijk})$. A natural approach to estimating $e_{ijk}$ is to select a potential donor at random, say donor $k'$, from class $ij$ and to compute

$$\hat{e}_{ijk} = \hat{Y}_{ijk} - \hat{Y}_{ij}.$$ 

From the model this estimator is

$$\hat{Y}_{ijk} = \bar{Y}_{ij} + e_{ijk} = \bar{Y}_{ij} + \hat{e}_{ijk} = \bar{Y}_{ij} + \hat{e}_{ijk} = \bar{Y}_{ij} + \hat{e}_{ij}$$

and hence is not a good estimator of $e_{ijk}$. However,

$$\hat{Y}_{ijk}^* = \bar{Y}_{ij} + \hat{e}_{ijk} = \bar{Y}_{ij} + \hat{e}_{ij} + e_{ijk} = \bar{Y}_{ij} + e_{ijk}$$

does estimate $Y_{ijk}$. Thus $\hat{Y}_{ijk}^*$ is a suitable value to impute for $Y_{ijk}$. In practice, $\hat{Y}_{ijk}$ need not be computed from $\bar{Y}_{ij} + e_{ijk}$, because this quantity is simply equal to $Y_{ijk}$. In other words the recipient can be simply assigned the value of a randomly-chosen donor. This is random imputation, as discussed in (e) above, within class $ij$.

The model (4.2.3) can be reexpressed as
\[ y_{ijk} = \bar{y} + a_i + b_j + c_{ij} + e_{ijk} \quad (4.2.4) \]

where \( a_i = (\bar{y}_i - \bar{y}) \) and \( b_j = (\bar{y}_j - \bar{y}) \) are the main effects and \( c_{ij} = (\bar{y}_{ij} - \bar{y}_i - \bar{y}_j + \bar{y}) \) is the interaction effect of control variables \( i \) and \( j \) in the usual analysis of variance representation. Since no assumption is made about the distribution of the residuals \( e_{ijk} \), this formulation makes clear the lack of assumptions made by the imputation class model. However many control variables are used in defining classes, all interactions of all levels are included in the model. The only assumption involved is the necessary one that recipients are missing at random.

The cost of the generality of the imputation class model is that less good estimators may be obtained than would come from a model in which more assumptions were made—providing those assumptions held. The restrictions imposed on the model by making extra assumptions may also permit the model to be extended in other ways, in particular by including more control variables in it.

Several types of assumptions could be made. One, for instance, would be to assume that the distribution of residuals is the same in all classes, perhaps even that it is a normal distribution; we will return to this later when we consider regression. Another is to assume that there are no interaction effects, for instance \( c_{ij} = 0 \) in (4.2.4); in this case (4.2.4) can be viewed as a dummy variable regression. (Alternatively, one could specify that just certain interactions may be present.) In passing it is worth noting the relationship between the no-interaction model and raking ratio estimation discussed in Section 3.4.

In addition to the assumption of no interactions (additivity), linear relationships between the dependent variable and the control variables might be assumed. The model then becomes a multiple regression model

\[ y_k = \hat{B}_0 + \hat{B}_1 x_{1k} + \hat{B}_2 x_{2k} + \epsilon_k \]

where the single subscript \( k \) now runs over the whole sample. Furthermore, the residuals may be assumed to be homoscedastic and even normally distributed.

With this background, we now return to the issue of how to impute for missing values with regression imputation. The use of the predicted value

\[ \hat{y}_k = \hat{B}_0 + \hat{B}_1 x_{1k} \]

corresponds to the mean-value imputation in the restricted model, and hence has the same undesirable distributional properties. A good case therefore exists for including an estimated residual. There are various ways in which this could be done depending on the assumptions made about the residuals. The following are some of the more obvious possibilities:

(i) Assume that the errors are homoscedastic and normally distributed, \( N(0, \sigma^2_e) \). Then \( \sigma^2_e \) could be estimated by the residual variance from the regression, \( s^2_e \) and the residual for a recipient could be chosen at random from \( N(0, s^2_e) \).

(ii) Assume that the errors are heteroscedastic and normally distributed, with \( \sigma^2_{e_j} \) being the residual variance in some group \( j \). Estimate the \( \sigma^2_{e_j} \) by \( s^2_{e_j} \), and choose a residual for a recipient in group \( j \) from \( N(0, s^2_{e_j}) \).

(iii) Assume that the residuals all come from the same, unspecified, distribution. Then estimate \( y_k \) by \( \hat{y}_k + \hat{e}_k \), where \( \hat{e}_k \) is the estimated residual for a random-chosen donor.

(iv) The assumption in (iii) accepts the linearity and additivity of the model. If there are doubts about these assumptions, it may be better to take not a random-chosen donor but instead one close to the recipient in terms of his \( x \)-values. 'Close' may be defined by a distance function as in (g) above. In the limit, if a donor with the same set of \( x \)-values is found, this procedure reduces to assigning that donor's \( y \)-value to the recipient.
An important special case of the regression model with one control variable occurs when \( B_0 \) is set equal to zero. In this case the model reduces to the ratio model \( y_k = B_1 x_k + e_k \). The ratio model may often be appropriate for panel surveys, with \( x \) representing the same variable as \( y \) measured on the previous wave of the panel. Ford et al. (1980) investigate the use of a ratio model for imputation.

Through the careful development of an explicit model, regression imputation has the potential to produce imputed values closer to the true but missing values than the rough-and-ready, assumption-free, imputation class approaches. The construction and assessment of a good regression model is, however, a time-consuming operation, and it seems unrealistic to consider its application for all the items with missing values in a survey. If it is used at all, it seems best to reserve it for a few major variables for which effective linear models with high R\(^2\)'s can be developed. In using regression imputation, attention also needs to be given to problems of estimating several missing items on the same record. This point is discussed further in Section 4.4.

Applications of regression and categorical data models for imputation are described by Schieber (1978), Herzog and Lancaster (1980), and Herzog (1980).

4.3 Two Simulation Studies

Two small-scale simulation studies have been carried out using data collected in the 1978 ISDP Research Panel to illustrate and compare some of the imputation procedures described in the previous section. One study compares methods of imputing for missing data on the variable hourly rate of pay for those jobs paid at an hourly rate, as reported in the July 1978 questionnaire. The other concerns the variable quarterly earnings, as derived from income records in the April 1978 questionnaire. Both studies are confined to the area probability sample of the 1978 ISDP Research Panel.

The approach used in both studies consisted of first constructing data sets with missing values by taking all the cases with valid data for the variable concerned and deleting some of the recorded values, and then applying a selection of imputation procedures to determine how well they reconstructed the deleted values. The data sets were constructed in a way that attempted to mirror the actual patterns of missing data observed in the full sample. Subgroups of the sample with differing rates of missing data were identified, and values were deleted from cases with valid data at the corresponding rates within these subgroups. It was also necessary to remove some cases with valid data entirely in order to produce data sets with the same distributions over the subgroups as that of the full sample. Within subgroups, the cases whose values were deleted were selected at random. As a protection against the selection of an odd sample of deleted cases, the process was replicated ten times producing ten simulation data sets for each study. Many of the measures evaluating the quality of the imputation schemes are averaged across the ten data sets. The choice of subgroups with differing rates of missing data for the two variables were based on the results of nonresponse analyses of the 1978 ISDP Research Panel reported by Heeringa (1980).

Once the data sets were constructed, a variety of imputation procedures were applied to fill in the deleted values. The control variables used in these procedures were selected on the basis of Hendrix's (1980) analyses of effective predictors for the two variables. Imputation procedures with stochastic components were implemented ten times. Many of the evaluation measures reported below are averaged over these ten replications as well as over the ten simulation data sets.

While these simulation studies provide useful test-beds for investigating the various imputation procedures, their limitations should be recognized. In constructing the data sets attempts were made to reflect the actual patterns of missing data by deleting values at differing rates in various subgroups. However, ultimately within subgroups the cases for deletion were chosen at random, a feature which departs from reality to some, unknown, extent. The studies reported below

\footnote{For reasons of confidentiality geographical data were deleted by the Department of Health and Human Services from our data set. For this reason geographical variables were not available for use as imputation control variables.}
directly examine how the imputation procedures perform in dealing with the missing data model employed, and only indirectly suggest how they might perform in actual practice. In particular, note that if the subgroups from which the values were deleted at differing rates were themselves used as imputation classes with the simulated data set, the estimates of means and totals of the variables would automatically be unbiased estimates of the corresponding population parameters. (Through the method of construction, the data are missing at random within these classes.) There can be no guarantee, however, that this also applies for the actual sample.

4.3.1 Imputation of Hourly Rate of Pay

The dependent variable for this study is hourly rate of pay (SC 1335-47-59) on the July questionnaire of the 1978 ISDP Research Panel. The analyses are conducted at the job level, so that, if a respondent has more than one hourly paid job, each job is taken individually (however, few employees in fact have more than one job). Only respondents in the area probability sample of the Research Panel who report having one or more hourly paid jobs are included in the analyses.'

The two characteristics most highly associated with the item nonresponse rate for hourly rate of pay are the person's sex and whether the interview was conducted in person or with a proxy informant. Table 1 gives the item nonresponse rates for the four cells comprising the cross-classification of these two characteristics.

The study's ten data sets were constructed in two steps from the 1128 cases with reported hourly rates of pay. First 92 cases were removed, randomly within cells, so that the distribution of the remaining cases conformed to that of the second column of Table 1. Next the missing data rates in the last column of Table 1 were applied to delete values at random within each of the four subgroups. This process was replicated ten times, producing the ten simulation data sets. The sex and

<table>
<thead>
<tr>
<th>Total hourly paid jobs</th>
<th>No.</th>
<th>%</th>
<th>No. of jobs with hourly rate of pay reported</th>
<th>Percent of missing data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male, self report</td>
<td>339</td>
<td>27.0</td>
<td>327</td>
<td>3.5%</td>
</tr>
<tr>
<td>Male, proxy report</td>
<td>345</td>
<td>27.5</td>
<td>285</td>
<td>17.4%</td>
</tr>
<tr>
<td>Female, self report</td>
<td>432</td>
<td>34.4</td>
<td>400</td>
<td>7.4%</td>
</tr>
<tr>
<td>Female, proxy report</td>
<td>139</td>
<td>11.1</td>
<td>116</td>
<td>16.5%</td>
</tr>
<tr>
<td>Total'</td>
<td>1255</td>
<td>100.0</td>
<td>1128</td>
<td>10.1%</td>
</tr>
</tbody>
</table>

'There were also 9 cases with missing interview status, 8 of which had valid data on hourly rate of pay. These cases were excluded from the simulation study reported here.'

<table>
<thead>
<tr>
<th>Total hourly paid jobs</th>
<th>Hourly rate retained</th>
<th>Hourly rate deleted</th>
<th>Percent of deleted cases</th>
</tr>
</thead>
<tbody>
<tr>
<td>No.</td>
<td>%</td>
<td>No.</td>
<td>Mean rate ($)</td>
</tr>
<tr>
<td>Male, self</td>
<td>280</td>
<td>27.0</td>
<td>270</td>
</tr>
<tr>
<td>Male, proxy</td>
<td>285</td>
<td>27.5</td>
<td>235</td>
</tr>
<tr>
<td>Female, self</td>
<td>356</td>
<td>34.4</td>
<td>330</td>
</tr>
<tr>
<td>Female, proxy</td>
<td>115</td>
<td>11.1</td>
<td>96</td>
</tr>
<tr>
<td>Total</td>
<td>1036</td>
<td>100.0</td>
<td>931</td>
</tr>
</tbody>
</table>
interview status distribution of each of the data sets so constructed is given in Table 2. The table also gives hourly rates of pay averaged across the ten simulation data sets.

Of the 1036 cases in each constructed data set, 105 have values of the hourly rate of pay deleted. Several alternative imputation procedures have been applied to assign values for these deleted values. The imputation control variables employed in these procedures were obtained from a SEARCH analysis conducted by Hendrix (1980, Section 3.1) to identify good predictors of hourly rate of pay from the April wave data. SEARCH, the successor to the well-known Automatic Interaction Detector (AID) program, determines the most effective predictors of a variable through a series of binary splits of the sample, each split partitioning the sample into two parts on some predictor variable in such a way that the maximum proportion of variance is explained (for further details see Sonquist et al., 1974). The results of the analysis may be conveniently presented in a tree diagram. The tree diagram from Hendrix's analysis is reproduced in Figure 1.

The following imputation procedures were implemented:

1. Grand mean imputation (GM) assigning the respondent mean for all the missing values.

2. Class mean-value imputation using 8 imputation classes (CM8) defined by the cross-tabulation of the first three variables appearing in the SEARCH analysis, namely union membership (member v. not a member and membership not known), occupation (professional, technical and crafts v. others) and industry (agriculture, forestry, fisheries, mining and construction, transportation, communications and other public utilities v. others, i.e. Industry codes 017-077 and 407-479 v. others). The distribution of cases with retained and deleted values across these eight cells for the first simulation data set is given in Table 3.

3. Class mean-value imputation using 10 imputation classes (CM10) defined by the SEARCH analysis, that is the groups with 's' in Figure 1, groups numbered 9 to 19, except for group 12. The distributions of cases with retained and deleted values across these ten cells are given in Table 4.
4. Random imputation within the 8 imputation classes (RC8) defined in procedure 2 above. The donors were chosen within classes by SRS.

5. Random imputation within the 10 imputation classes (RC10) defined by the SEARCH analysis as described for procedure 3 above. The donors were chosen within classes by SRS.

6. Multiple regression imputation using the predicted values from the regression equation (MI). The following independent variables were used in the regression: union membership as one dummy variable (member v. not a member or membership not known); occupation as 6 categories or 5 dummy variables (professional, technical, managers and administrators; sales and clerical; craftsmen; operatives; laborers and farm managers; occupation not known); industry as 6 categories or 5 dummy variables (agriculture, forestry, fisheries, and farm managers; occupation not known); manufacturing; wholesale and retail trades; finance, insurance, real estate, public administration; professional and related services; and industry not known); age as a grouped continuous variable (in four groups: 16-24; 25-44; 45-64; 65 and over); sex as a dummy variable; and hours worked per week as a continuous variable. The regression analysis of hourly rate of pay on these independent variables applied to all 931 cases in the first simulation data set with hourly rate of pay retained had an $R^2$ of 0.53.

7. Multiple regression imputation using the predicted value from the regression plus a random residual chosen from a normal distribution (MN). The procedure employs the same regression analysis as procedure 6. The random residuals are chosen from $\mathcal{N}(0, \hat{s}_e^2)$, where $\hat{s}_e^2$ is the estimated residual variance for the respondents. The residual standard deviation for the first simulation data set was $\hat{s}_e = 1.501$.

8. Multiple regression imputation using the predicted value from the regression plus a randomly chosen respondent residual (MR). This procedure employs the same regression analysis as that described in procedure 6. The values imputed for the deleted cases are the predicted values plus...
Three measures for evaluating the effectiveness of these eight imputation procedures in recapturing the deleted values are presented in Table 5. They are:

(a) The mean deviation defined as \( \frac{\sum (\hat{y}_{mi} - y_{mi})}{m} \), where \( \hat{y}_{mi} \) is the imputed value and \( y_{mi} \) is the actual, but deleted, value for case \( i \) (\( i = 1,2, \ldots, m \)).

(b) The mean absolute deviation defined as \( \frac{\sum |\hat{y}_{mi} - y_{mi}|}{m} \).

(c) The root mean square deviation defined as \( \sqrt{\frac{\sum (\hat{y}_{mi} - y_{mi})^2}{m}} \).

These measures are averaged across the ten simulation data sets. The procedures RC8, RC10, MN and MR involve a stochastic component. For each of these procedures, the measures are also averaged over ten replications of the imputations.

**Table 5: Measures of the effectiveness of the eight imputation procedures for imputing for deleted hourly rates of pay**

<table>
<thead>
<tr>
<th>Imputation scheme</th>
<th>Mean deviation (( \dagger ))</th>
<th>Mean absolute deviation (( \dagger ))</th>
<th>Root mean square deviation (( \dagger ))</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. GM</td>
<td>-15.8</td>
<td>190.4</td>
<td>252.7</td>
</tr>
<tr>
<td>2. CM8</td>
<td>-2.7</td>
<td>126.1</td>
<td>173.0</td>
</tr>
<tr>
<td>3. CM10</td>
<td>-1.9</td>
<td>120.3</td>
<td>167.3</td>
</tr>
<tr>
<td>4. RC8</td>
<td>-5.2</td>
<td>172.9</td>
<td>244.6</td>
</tr>
<tr>
<td>5. RC10</td>
<td>-0.3</td>
<td>165.8</td>
<td>236.2</td>
</tr>
<tr>
<td>6. MI</td>
<td>-4.9</td>
<td>119.6</td>
<td>170.0</td>
</tr>
<tr>
<td>7. MN</td>
<td>-4.7</td>
<td>175.3</td>
<td>228.1</td>
</tr>
<tr>
<td>8. MR</td>
<td>-5.6</td>
<td>167.8</td>
<td>227.5</td>
</tr>
</tbody>
</table>

The mean deviation measures the bias in the imputed values. Since the cases with missing values are only part of the total sample, the resulting bias in the estimation of the population mean is only a fraction \( M \) (the nonresponse rate) of the mean deviation. As seen in Table 5, all the mean deviations are negative, indicating that on average the imputed values underestimated the actual (deleted) values. The average underestimation for the grand mean (GM) procedure is almost 16 cents, and this leads to the same bias in the estimate of the population mean as would the procedure of simply computing the respondent mean ignoring the missing values. The average underestimation for the other procedures is less than 6 cents, so that the use of the control variables seems to have been effective in reducing the absolute magnitude of the imputation bias, as compared with the GM procedure which does not make use of any control variables.

The mean absolute deviations and the root mean square deviations measure the "closeness" with which the deleted values are reconstructed. Necessarily, the GM procedure fares worst in terms of these indices, having the largest measures on both of them. Table 5 shows that the three procedures using predicted values without residuals (CM8, CM10, and MI), yield the best measures, all around 120 for the mean absolute deviation and 170 for the root mean square deviation. As expected, the CM10 procedure does slightly better than the CM8 procedure, indicating the gains from increasing the extent of use of the control variables.

The four procedures that incorporate another stage of randomization through the addition of a residual term in some way (RC8, RC10, MN, and MR) yield mean absolute and root mean square deviations that are approaching 40 per cent larger than those for their deterministic counterparts (CM8, CM10 and MI). As the following argument demonstrates, this result is very close to what might be reasonably predicted. Consider the mean square deviation of the RC procedure given by

\[ MSD(RC) = \frac{\sum (\hat{y}_{mhi} - y_{mhi})^2}{m}. \]

Then

\[ m_{MSD}(RC) = \frac{\sum (\hat{y}_{mhi} - \hat{y}_{rh})^2 + \sum (y_{mhi} - \hat{y}_{rh})^2}{m} \]

\[ + \sum (\hat{y}_{mhi} - \hat{y}_{rh})(y_{mhi} - \hat{y}_{rh}) \]

1 I am grateful to Brian Greenberg and James Fagan for drawing my attention to this point.
Taking the conditional expectation over the random imputations within each imputation class, $E(\hat{Y}_{\text{mhi}}) = \hat{Y}_{\text{rh}}$ so that for large $r_h$,

$$mE_2[\text{MSD}(RC)] = \Sigma m^2 r_h + n\text{MSD}(CM)$$

This equation shows that $E_2[\text{MSD}(RC)] \geq \text{MSD}(CM)$. Under the assumption that values are missing at random within imputation classes, $\text{MSD}(CM) = \Sigma m^2 r_h$, so that $E_2[\text{MSD}(RC)] = 2\text{MSD}(CM)$. Thus, under the missing at random within classes assumption, the root mean square deviation of the RC procedure can be expected to be about $\sqrt{2} \approx 1.41$ times as large as the CM imputation procedure that employs the same set of imputation classes. Under appropriate assumptions, the corresponding factor of $\sqrt{2}$ can also be derived for the stochastic and deterministic regression imputation procedures.

The reason for accepting the less accurate estimates from the imputation procedures with the residual terms is that these residuals maintain the distributional properties of the sample. This case is demonstrated in Table 6 which summarizes the distributions of the imputed values for the eight procedures, and provides the distribution of the true, but deleted, values (TV) for comparison. The table relates only to the first simulation data set, and only to one replication of the imputation schemes incorporating a stochastic component.

The first three columns of Table 6 clearly illustrate the distortion to distributions created by mean-value imputation. The GM procedure allocates the respondent sample mean to all the missing cases, thus concentrating all the distribution of missing values at the one value. The number of different imputed values that can occur with the CM procedures is restricted to the number of imputation classes, and the number of cases with any one of these values is then the number of cases with missing data in the given class. As Table 3 shows, the last class in procedure CMS is extremely large, containing over half the sample. Fifty of the cases in this class have missing data, and they are all assigned a value of $3.20$: these cases thus account for the excessive number of cases in the $3$-$4$ interval in the CMS column of Table 6. The distribution for the regression imputation procedure using predicted values, MI, also shows signs of curtailment of spread, but this is less evident than in the case of the CM procedures; the lesser degree of bunching in this distribution is accounted for partly by the greater number of classes used (via dummy variables) and partly by the use of continuous variables in the regression. The MN and MR procedures, which add a residual term to the predicted value from the regression, produce substantial excesses in the numbers of very small hourly rates of pay. This outcome arises because of the inapplicability of the regression assumptions adopted. Both procedures assume that the residuals are homoscedastic, and the failure of this assumption may well account for the excess. Failure of the assumption of normality of residuals may also be a contributory factor in the case of the MN procedure. The two RC procedures generate the best distributional properties among the procedures considered here.

### Table 6: Distributions of the true values (TV) and of the imputed values from the eight imputation procedures for hourly rate of pay of the 105 cases with deleted values in the first simulation data set.

<table>
<thead>
<tr>
<th>Hourly Rate of Pay ($)</th>
<th>TV</th>
<th>GM</th>
<th>CMS</th>
<th>CM10</th>
<th>RC8</th>
<th>RC10</th>
<th>MI</th>
<th>MN</th>
<th>MR</th>
</tr>
</thead>
<tbody>
<tr>
<td>$2$ or less</td>
<td>3</td>
<td>-</td>
<td>-</td>
<td>1</td>
<td>5</td>
<td>2</td>
<td>12</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>$2$</td>
<td>22</td>
<td>-</td>
<td>24</td>
<td>29</td>
<td>24</td>
<td>13</td>
<td>17</td>
<td>15</td>
<td></td>
</tr>
<tr>
<td>$3$</td>
<td>28</td>
<td>50</td>
<td>19</td>
<td>24</td>
<td>25</td>
<td>33</td>
<td>14</td>
<td>21</td>
<td></td>
</tr>
<tr>
<td>$4$</td>
<td>11</td>
<td>105</td>
<td>22</td>
<td>23</td>
<td>10</td>
<td>14</td>
<td>15</td>
<td>16</td>
<td></td>
</tr>
<tr>
<td>$5$</td>
<td>11</td>
<td>-</td>
<td>16</td>
<td>24</td>
<td>15</td>
<td>8</td>
<td>13</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>$6$</td>
<td>10</td>
<td>-</td>
<td>-</td>
<td>7</td>
<td>8</td>
<td>16</td>
<td>15</td>
<td>13</td>
<td></td>
</tr>
<tr>
<td>$7$</td>
<td>10</td>
<td>-</td>
<td>-</td>
<td>9</td>
<td>16</td>
<td>14</td>
<td>16</td>
<td>13</td>
<td></td>
</tr>
<tr>
<td>$9$</td>
<td>3</td>
<td>-</td>
<td>6</td>
<td>6</td>
<td>2</td>
<td>6</td>
<td>4</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>$11$</td>
<td>4</td>
<td>-</td>
<td>-</td>
<td>4</td>
<td>-</td>
<td>3</td>
<td>-</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>$13$ and over</td>
<td>3</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>

| Total                  | 105 | 105 | 105 | 105  | 105 | 105  | 105 | 105 |

*For each imputation procedure the distribution is obtained from a single replication.*

4.3.2 Imputation of Quarterly Earnings

In the 1978 ISDP Research Panel, respondents were asked to provide information on their earnings from records where
possible. If they were paid the same amount each payday, they were asked to report the amount of the paycheck and the frequency of pay. If they were paid different amounts each payday, the amounts were collected for each of the paychecks received in the preceding three months. From these items of information an aggregate quantity, quarterly earnings, can be constructed. This quantity for the period January to March 1978, as reported in the April 1978 wave of the 1978 Research Panel, is the dependent variable for the second study. The data are analyzed at the job level, and the analyses relate to respondents in the area probability sample only.

The designers of the 1978 Research Panel questionnaires anticipated that it would often not prove possible to collect earnings data from records. They therefore included alternative questions in the questionnaire to serve as the basis for estimating earnings. In particular, data were collected on hourly rates of pay, usual numbers of hours worked per week, and numbers of weeks without pay in the last quarter. These data, which can then be used to estimate quarterly wages, are employed below in a form of deductive imputation.

The methods adopted for the study of imputation procedures for quarterly earnings are the same as those used for hourly rates of pay described in the last section. The two characteristics most highly associated with the missing data rate on quarterly earnings reported from records were interview status and the household's monthly income for March (SC 0014). The missing data rates for the four cells defined by the interview status and the household's monthly income for March are given in Table 7. The table shows that in more than half the cases, the respondent did not provide earnings data from records.

The ten simulated data sets used as the basis of the study of imputation procedures were constructed in the manner described in the previous section. The distribution of each of these data sets over the four cells contained in Table 7 is presented in Table 8. The mean quarterly earnings presented in Table 8 are averages over the ten simulation data sets.

The choice of the imputation control variables to be employed were based on a SEARCH analysis of quarterly earnings carried out by Hendrix (1980, Section 3.2). The tree diagram summarizing his results is presented in Figure 2.

Ten different imputation procedures were applied to impute for the deleted values:

1. Grand mean imputation (GM) assigning the respondent mean for all the missing values.

2. Class mean-value imputation using 8 imputation classes (CM8). Seven of these classes were obtained from the crosstabulation of the first three variables in the SEARCH analysis - sex, household's March income (under $900, $900 or more) and work status (full- or part-time). The cell of this crosstabulation containing men part-time workers with household income under $900 contained only two cases with earnings retained but 11 cases with earnings deleted: this cell was therefore collapsed with that for women part-time workers with household income under $900. The eighth cell comprised a small number of cases with one or more of the classification variables missing. The distribution of cases with earnings retained and those with earnings deleted across the eight cells for the first simulation data set is shown in Table 9.

3. Class mean-value imputation using 12 imputation classes (CM12) defined by the SEARCH analysis, that is the groups with 's in Figure 2, groups numbered 8, 11, 12, 14, and 16 to 23. The distributions of cases with retained and deleted values across these 12 classes are given in Table 10.

4. Random imputation within the 8 imputation classes defined in procedure 2 above (RC8). The donors were chosen within classes by the SRS scheme described earlier in the report. As Table 9 demonstrates, in many classes there were more recipients than potential donors. In these classes each case with a retained value served as a donor once, being allocated to recipients at random, and a SRS of cases with

---

1 The household monthly income question was asked early in the interview, and was used for screening purposes. It was not intended to provide an accurate measure of monthly income, and hence cannot be used for calculating quarterly earnings.
Table 7: Missing data rates for quarterly earnings reported from records by interview status and household's March income: Area Probability Sample, April 1978

<table>
<thead>
<tr>
<th>Household's March Income</th>
<th>Interview Status</th>
<th>Total jobs No.</th>
<th>Percent %</th>
<th>No. of jobs with earnings reported</th>
<th>Percent of missing data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Under $900</td>
<td>Self</td>
<td>404</td>
<td>21.7</td>
<td>188</td>
<td>53.5%</td>
</tr>
<tr>
<td>Under $900</td>
<td>Proxy</td>
<td>176</td>
<td>9.5</td>
<td>38</td>
<td>78.4%</td>
</tr>
<tr>
<td>$900 or more</td>
<td>Self</td>
<td>771</td>
<td>41.5</td>
<td>424</td>
<td>45.0%</td>
</tr>
<tr>
<td>$900 or more</td>
<td>Proxy</td>
<td>507</td>
<td>27.3</td>
<td>191</td>
<td>62.3%</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>1858</td>
<td>100.0</td>
<td>841</td>
<td>54.7%</td>
</tr>
</tbody>
</table>

67 cases with either missing interview status or missing household March income are excluded from this table, and from the following simulation study.

Table 8: Distribution of simulated data sets by household's March income and interview status for quarterly earnings imputation study

<table>
<thead>
<tr>
<th>Household's March Income</th>
<th>Interview Status</th>
<th>Total jobs No.</th>
<th>Percent %</th>
<th>Quarterly earnings retained Mean No. ($)</th>
<th>Quarterly earnings deleted Mean No. ($)</th>
<th>Percent of deleted cases</th>
</tr>
</thead>
<tbody>
<tr>
<td>Under $900</td>
<td>Self</td>
<td>87</td>
<td>21.7</td>
<td>40</td>
<td>1384</td>
<td>54.0%</td>
</tr>
<tr>
<td>Under $900</td>
<td>Proxy</td>
<td>38</td>
<td>9.5</td>
<td>8</td>
<td>2351</td>
<td>78.9%</td>
</tr>
<tr>
<td>$900 or more</td>
<td>Self</td>
<td>166</td>
<td>41.5</td>
<td>91</td>
<td>3135</td>
<td>45.2%</td>
</tr>
<tr>
<td>$900 or more</td>
<td>Proxy</td>
<td>109</td>
<td>27.3</td>
<td>41</td>
<td>3152</td>
<td>62.4%</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>400</td>
<td>100.0</td>
<td>180</td>
<td>2715</td>
<td>55.0%</td>
</tr>
</tbody>
</table>
Table 9: Distribution of cases in the first simulation data set with retained and with deleted values by household's March income, work status and sex

<table>
<thead>
<tr>
<th>Household's March income</th>
<th>Full-or part-time</th>
<th>Sex</th>
<th>Values retained</th>
<th>Values deleted</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; $900</td>
<td>F</td>
<td>M</td>
<td>23</td>
<td>27</td>
<td>50</td>
</tr>
<tr>
<td>&lt; $900</td>
<td>F</td>
<td>F</td>
<td>11</td>
<td>22</td>
<td>33</td>
</tr>
<tr>
<td>&lt; $900</td>
<td>P</td>
<td>M</td>
<td>7</td>
<td>17</td>
<td>24</td>
</tr>
<tr>
<td>&lt; $900</td>
<td>P</td>
<td>F</td>
<td>12</td>
<td>17</td>
<td>29</td>
</tr>
<tr>
<td>&gt; $900</td>
<td>F</td>
<td>M</td>
<td>65</td>
<td>79</td>
<td>144</td>
</tr>
<tr>
<td>&gt; $900</td>
<td>F</td>
<td>F</td>
<td>35</td>
<td>33</td>
<td>68</td>
</tr>
<tr>
<td>&gt; $900</td>
<td>P</td>
<td>M</td>
<td>18</td>
<td>12</td>
<td>30</td>
</tr>
<tr>
<td>&gt; $900</td>
<td>P</td>
<td>F</td>
<td>19</td>
<td>17</td>
<td>36</td>
</tr>
<tr>
<td>Missing</td>
<td></td>
<td></td>
<td>2</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>Total</td>
<td>180</td>
<td></td>
<td>220</td>
<td>400</td>
<td></td>
</tr>
</tbody>
</table>

'One or more of household's March income, work status and sex missing.

Table 10: Distribution of cases in the first simulation data set with retained and with deleted values for the twelve SEARCH analysis classes

<table>
<thead>
<tr>
<th>Class</th>
<th>Values retained</th>
<th>Values deleted</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>23</td>
<td>27</td>
<td>50</td>
</tr>
<tr>
<td>11</td>
<td>7</td>
<td>7</td>
<td>14</td>
</tr>
<tr>
<td>12</td>
<td>8</td>
<td>7</td>
<td>15</td>
</tr>
<tr>
<td>14</td>
<td>7</td>
<td>8</td>
<td>15</td>
</tr>
<tr>
<td>16</td>
<td>7</td>
<td>8</td>
<td>15</td>
</tr>
<tr>
<td>17</td>
<td>13</td>
<td>24</td>
<td>37</td>
</tr>
<tr>
<td>18</td>
<td>36</td>
<td>49</td>
<td>85</td>
</tr>
<tr>
<td>19</td>
<td>34</td>
<td>40</td>
<td>74</td>
</tr>
<tr>
<td>20</td>
<td>9</td>
<td>9</td>
<td>18</td>
</tr>
<tr>
<td>21</td>
<td>24</td>
<td>26</td>
<td>50</td>
</tr>
<tr>
<td>22</td>
<td>8</td>
<td>10</td>
<td>18</td>
</tr>
<tr>
<td>23</td>
<td>4</td>
<td>5</td>
<td>9</td>
</tr>
<tr>
<td>Total</td>
<td>180</td>
<td>220</td>
<td>400</td>
</tr>
</tbody>
</table>

'See Figure 2 for definitions of classes.'
9. Mixed deductive and random imputation using 8 imputation 
classes (DI8). For about 60 per cent of cases with 
earnings deleted, information was available to estimate 
their quarterly earnings by multiplying their hourly rate 
of pay by the number of hours usually worked per week by 
the number of weeks worked in the quarter; for these 
cases, the missing values were imputed from this formula. 
The imputed values for the remaining 40 per cent of cases 
were obtained by random imputation with SRS within the 
eight imputation classes defined in procedure 2.

10. Mixed deductive and random imputation using the 12 
imputation classes defined by the SEARCH analysis – see 
procedure 3 (DI12). This procedure is the same as the 
previous one except that the random imputations for the 
approximately 40 per cent of cases were carried out within 
these twelve classes rather than the eight classes used in 
procedure 9.

The mean deviations, mean absolute deviations, and root 
mean square deviations measuring the effectiveness of these ten 
procedures in imputing for the 220 deleted values, are 
presented in Table 11. These measures are averaged across the 
ten simulation data sets and, in the case of the imputation 
procedures involving a stochastic component, also over ten 
replications of the procedure with each data set.

Table 11: Measures of the effectiveness of the ten 
imputation procedures for imputing for quarterly earnings

<table>
<thead>
<tr>
<th>Imputation scheme</th>
<th>Mean deviation ($)</th>
<th>Mean absolute deviation ($)</th>
<th>Root mean square deviation ($)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. GM</td>
<td>24.3</td>
<td>1571</td>
<td>2028</td>
</tr>
<tr>
<td>2. CM8</td>
<td>-29.2</td>
<td>1103</td>
<td>1609</td>
</tr>
<tr>
<td>3. CM12</td>
<td>-16.7</td>
<td>1033</td>
<td>1480</td>
</tr>
<tr>
<td>4. RCS</td>
<td>-28.5</td>
<td>1538</td>
<td>2191</td>
</tr>
<tr>
<td>5. RC12</td>
<td>-21.3</td>
<td>1379</td>
<td>1964</td>
</tr>
<tr>
<td>6. MI</td>
<td>-3.2</td>
<td>1046</td>
<td>1440</td>
</tr>
<tr>
<td>7. MN</td>
<td>-7.1</td>
<td>1520</td>
<td>1965</td>
</tr>
<tr>
<td>8. MR</td>
<td>-1.7</td>
<td>1491</td>
<td>1972</td>
</tr>
<tr>
<td>9. DI8</td>
<td>-149.5</td>
<td>1052</td>
<td>1737</td>
</tr>
<tr>
<td>10. DI12</td>
<td>-97.4</td>
<td>955</td>
<td>1564</td>
</tr>
</tbody>
</table>

The mean deviations in Table 11 vary between the various 
imputation procedures. From theoretical considerations, the 
mean deviations of the three regression procedures (MI, MN and 
MR) should be similar and the class mean-value imputations (CM8 
and CM12) should have similar mean deviations to their 
stochastic counterparts (RC8 and RC12 respectively). The 
results in Table 11 conform well to these expectations. The 
regression procedures produce almost unbiased estimates, while 
the class mean-value procedures have negative biases of around 
$20 to $30. The grand mean procedure, however, yields a 
positive bias. All the mean deviations are relatively small, 
around 1% or less, except for the two mixed deductive and 
random class imputation procedures (DI8 and DI12). The 
relatively larger biases of these latter procedures will be 
discussed later.

Comparison of the mean absolute deviations and root mean 
squared deviations between the deterministic imputation 
procedures (CM8, CM12 and MI) and their stochastic counterparts 
(RC8, RC12 and MN and MR, respectively) shows that these 
evaluation measures are on average over one third larger for 
the stochastic procedures (this is somewhat less than the 41% 
larger predicted by the appropriate missing at random 
assumption as discussed in the Section 4.3.1). Of these 
procedures, the CM10 and MI procedures fare best, with the 
CM8 procedure being a little less good. As a result of its lack of 
use of control variables, the deterministic GM procedure has 
mean absolute and root mean squared deviations of about the 
same size as the stochastic procedures. Although the two mixed 
deductive and random imputation procedures, DI8 and DI12, 
incorporate a stochastic component in imputing for about 40 per 
cent of cases, they have mean absolute and root mean squared 
deviations of about the same size as the corresponding class 
mean-value procedures CM8 and CM12. This finding suggests that 
the use of the deductive imputation with the other 60 per cent 
of cases must yield imputed values close to the actual values, 
a suggestion substantiated in Table 12 where the two sets of 
cases are taken separately.

The deductive imputation in the DI8 and DI12 procedures 
produces a mean absolute deviation of about 490 for the cases
Table 12: Mean deviations and mean absolute deviations for the ten imputation procedures for imputing quarterly earnings separately (a) for the cases for which the deductive imputation based on hourly rate of pay, usual number of hours worked per week and number of weeks worked in the quarter can be made and (b) for the cases for which this imputation cannot be made.

<table>
<thead>
<tr>
<th>Imputation scheme</th>
<th>Mean deviation (a) for deductive cases ($)</th>
<th>Mean absolute deviation (a) for deductive cases ($)</th>
<th>Mean deviation (b) for nondeductive cases ($)</th>
<th>Mean absolute deviation (b) for nondeductive cases ($)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. GM</td>
<td>563</td>
<td>1423</td>
<td>-692</td>
<td>1674</td>
</tr>
<tr>
<td>2. CM8</td>
<td>181</td>
<td>964</td>
<td>-343</td>
<td>1294</td>
</tr>
<tr>
<td>3. CM12</td>
<td>181</td>
<td>934</td>
<td>-253</td>
<td>1127</td>
</tr>
<tr>
<td>4. RC8</td>
<td>187</td>
<td>1342</td>
<td>-361</td>
<td>1773</td>
</tr>
<tr>
<td>5. RC12</td>
<td>196</td>
<td>1254</td>
<td>-267</td>
<td>1511</td>
</tr>
<tr>
<td>6. MI</td>
<td>65</td>
<td>870</td>
<td>-85</td>
<td>1275</td>
</tr>
<tr>
<td>7. MN</td>
<td>68</td>
<td>1394</td>
<td>-67</td>
<td>1689</td>
</tr>
<tr>
<td>8. MR</td>
<td>56</td>
<td>1348</td>
<td>-105</td>
<td>1682</td>
</tr>
<tr>
<td>9. DI8</td>
<td>94</td>
<td>494</td>
<td>-361</td>
<td>1773</td>
</tr>
<tr>
<td>10. DI12</td>
<td>89</td>
<td>489</td>
<td>-267</td>
<td>1511</td>
</tr>
</tbody>
</table>

with which it can be used. This value is considerably lower than the mean absolute deviation for any of the other procedures for these cases. In these terms, the next best procedure is the regression prediction (MI), and its mean absolute deviation is about 77 per cent larger than that for the deductive imputation. The use of deductive imputation, or at least the inclusion of the deduced values (or some or all of their components) in the imputation procedure employed, makes a valuable contribution in imputing for quarterly earnings.

The separation of deductive and non-deductive cases in Table 12 also sheds further light on the mean deviations for the various procedures. A striking finding is that the mean deviations are positive for the deductive cases and negative for the non-deductive cases for all ten procedures. These fairly sizeable positive and negative measures tend to cancel out in the aggregate, creating the small mean deviations observed in Table 11. For the non-deductive cases, the use of the DI8 and DI12 procedures means employing the RC8 and RC12 procedures respectively, so that the mean deviations of the DI8 and RC8 and of the DI12 and RC12 procedures are the same for these cases (and similar also to the mean deviations for the CM8 and CM12 procedures). The deductive imputation procedures, however, yield lower mean deviations for the deductive cases than their RC (and CM) counterparts. As a result, the cancellation of positive and negative mean deviations between deductive and non-deductive cases that applies in the aggregate with the RC and CM procedures does not occur to the same extent with the DI procedures. This explains the larger mean deviations of the DI procedures seen in Table 11.

The deductive imputation, which consists of multiplying hourly rate of pay by number of hours usually worked per week by number of weeks worked in the quarter, applies only to wage earners for whom these data are available. Thus all the deductive cases are wage earning jobs and, on examination, the non-deductive cases are found to be mostly salaried jobs. Thus it appears that all the imputation procedures implemented in this study tend to overestimate wages and underestimate salaries. Such biases would not occur if the data were missing at random within the imputation classes employed in the CM and RC procedures. The missing data were in fact generated at random within the four subgroups given in Table 7, subgroups created by cross-classifying household’s March income (under $900, $900 or more) and interview status (proxy v. self report). Since the imputation classes used in the CM8 and RC8 procedures included the household’s March income in exactly the same form, and this income was also partially included in defining the SEARCH imputation classes used in the CM12 and RC12 procedures, it seems that the biases have mainly occurred because of the failure of the imputation classes to control on the interview status variable.

The distributions of the imputed values for the ten procedures are compared with the distribution of the actual, but deleted, values in Table 13. As with the first study, the GM procedure concentrates all the imputed values at one point and the CM8 and CM12 mean-value procedures create distributions with spikes at the imputation class means. The procedure using the predicted values from the regression (MI) produces a
bunching of imputed values in the middle of the distribution. The distributions for the RC8 and RC12 procedures fit the distribution of true values reasonably well, as also do the DI8 and DI12 procedures (but their fits seem to be not quite as good).

Table 13: Distributions of the true values (TV) and of the imputed values from the ten imputation procedures for quarterly earnings for the 220 cases with deleted values in the first simulation data set

<table>
<thead>
<tr>
<th>Quarterly earnings</th>
<th>TV</th>
<th>GM</th>
<th>CM8</th>
<th>CM12</th>
<th>RC8</th>
<th>RC12</th>
<th>MI</th>
<th>MN</th>
<th>MR</th>
<th>DI8</th>
<th>DI12</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; $0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>2</td>
<td>19</td>
<td>23</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>$0-1999</td>
<td>50</td>
<td>45</td>
<td>50</td>
<td>52</td>
<td>47</td>
<td>9</td>
<td>26</td>
<td>20</td>
<td>55</td>
<td>52</td>
<td>-</td>
</tr>
<tr>
<td>$2000-3999</td>
<td>48</td>
<td>34</td>
<td>35</td>
<td>42</td>
<td>42</td>
<td>30</td>
<td>39</td>
<td>34</td>
<td>36</td>
<td>53</td>
<td>54</td>
</tr>
<tr>
<td>$4000-5999</td>
<td>220</td>
<td>60</td>
<td>88</td>
<td>40</td>
<td>45</td>
<td>39</td>
<td>33</td>
<td>52</td>
<td>50</td>
<td>54</td>
<td>54</td>
</tr>
<tr>
<td>$6000-7999</td>
<td>25</td>
<td>2</td>
<td>59</td>
<td>30</td>
<td>34</td>
<td>41</td>
<td>42</td>
<td>39</td>
<td>31</td>
<td>38</td>
<td>38</td>
</tr>
<tr>
<td>$8000-9999</td>
<td>20</td>
<td>79</td>
<td>68</td>
<td>28</td>
<td>26</td>
<td>56</td>
<td>29</td>
<td>25</td>
<td>22</td>
<td>17</td>
<td>17</td>
</tr>
<tr>
<td>$10000-11999</td>
<td>15</td>
<td>-</td>
<td>-</td>
<td>12</td>
<td>12</td>
<td>27</td>
<td>21</td>
<td>15</td>
<td>11</td>
<td>11</td>
<td>11</td>
</tr>
<tr>
<td>$12000-13999</td>
<td>9</td>
<td>-</td>
<td>15</td>
<td>9</td>
<td>7</td>
<td>12</td>
<td>7</td>
<td>5</td>
<td>11</td>
<td>11</td>
<td>11</td>
</tr>
<tr>
<td>$14000-15999</td>
<td>3</td>
<td>-</td>
<td>-</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>$16000+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Total: 220 220 220 220 220 220 220 220 220 220 220

For each imputation procedure the distribution is obtained from a single replication.

The main purpose of presenting the results of the two studies in this section is to illustrate the effects of a variety of imputation procedures on some univariate statistics (Santos, 1981a, uses the same simulation studies to investigate the effects on multivariate statistics). The particular applications chosen were not developed in great detail to be the optimum versions of the specific procedures; in consequence, the results should be interpreted with caution, bearing this point in mind. Nevertheless, there are some general conclusions that can be tentatively drawn.

First, the studies illustrate the benefit and cost of using random rather than mean-value imputation within imputation classes. The benefit can be clearly seen from the comparisons of the distributions for the two types of procedures, with the random imputed values having distributions reasonably close to those of the true values, while the mean-value imputations have distorted, spiked, distributions. The cost is seen by comparing the mean absolute deviations or the root mean square deviations for the two types of procedure, where the measures for random imputations are about a third or more higher than those for the mean-value imputations. While the amount of increase in these measures depends on the homogeneity of the imputation classes employed, and on the dependent variable under study, it seems likely that increases of at least this magnitude could often occur in practice.

Secondly, the results indicate the security derived from using a random imputation procedure rather than a regression model of the type employed. The failure of the data to conform to the assumptions made about the distribution of residuals in the stochastic regression models (MN and MR) created some distortions in the distributions of imputed values. The lack of need for specific distributional assumptions in the random imputation procedures enabled these distortions to be avoided. It should be made clear that the argument here is not against amounts. The transformation approach may also often improve the fit of imputed values to highly-skewed distributions of actual values. Serious consideration should be given to the use of transformed variables in imputation when a regression model is contemplated.
all uses of regression or other such models, but rather to point out the need to carefully construct and test out such models if they are to be employed. Unless sufficient resources are put into the development of the models to ensure that they do adequately reflect the distributional properties of the data, it is probably safer to apply an imputation procedure that assigns donors' values to recipients. This is not intended to suggest that such imputation procedures are assumption free, for they do depend on the assumption that data are missing at random within imputation classes. They therefore require an appropriate choice of imputation classes. However, in assigning donors' values to recipients, they automatically preserve the distribution of the donors' values (within chance variations), and hence avoid the need for assumptions specifying a more general distribution form.

Finally, the relative success of the deductive imputation procedure using hourly wage rates, usual numbers of hours worked per week and numbers of weeks worked for estimating quarterly wages brings out the wisdom of anticipating those survey variables that are likely to encounter high levels of missing data, and collecting substitute information which may be used in imputation - whether deductive or some other form of imputation is used.

4.4 Other Imputation Issues

Earlier sections of this report have sought to demonstrate the close relationship between weighting adjustments and imputation. It is now time to emphasize the difference between the two approaches, and hence to identify the added complexities of imputation.

Weighting class adjustments are appropriate when little important information is available for the nonrespondents. That information is incorporated in forming the weighting classes (or in raking ratio estimation), and any other information known about the nonrespondents is ignored. Weighting adjustments can be thought of as imputing complete sets of data from respondents to nonrespondents. Thus, for instance, in the equal weighting adjustment procedure within weighting class $h$ (with $r_h$ respondents) each nonrespondent's record can be thought of as being divided into $r_h$ parts, each with a weight of $1/r_h$. Each part is then assigned the complete set of responses from one respondent. An important property of the resultant data set is that within weighting classes the data retain all the interrelationships between the survey variables that were found for the respondents, as is appropriate under the assumption that the missing data were missing at random within weighting classes (see Section 2.3.3). Moreover, providing the respondents' responses were consistent, satisfying any edit checks, the weighted data set is also bound to be consistent.

The position with imputation is made much more complicated by the fact that a great deal is known about the units with missing data, much more than can be incorporated in a matching of donors and recipients. Considering imputation for missing data on only a single item, it is necessary to distinguish between variables used in the imputation process (usually in forming imputation classes) and other variables. Variables used in the imputation process then need to be subdivided into variables for which full details are used and those for which only partial details are used (e.g. by categorizing a quantitative variable or combining categories of a qualitative variable). The interrelationships between the item with missing data and variables used in full detail in forming imputation classes have the desirable properties of a weighting adjustment procedure. However, the interrelationships between the item and variables not used in the imputation process may be distorted and the imputed values may fail edit checks involving the item and these variables. Similar damaging effects may also occur with variables used only partially in forming imputation classes.

A particular case of the problem of imputed values failing edit checks occurs when the item concerned is a quantitative one, and is related to a number of other quantitative variables. It is usually then impossible to obtain exact matches between donors and recipients in terms of these related variables, and approximate matches may result in the imputed values failing edit checks. An example will clarify this point. Suppose that the item with missing data is length of first marriage for persons known to have been married more than
once, with the time from the date of the first marriage to the date of the second marriage being known. A matching of donors and recipients on the approximate time between first and second marriages may result in the lengths of first marriage imputed to some recipients exceeding the time interval between their two marriages. Thus, for instance, if a recipient with 22 years between marriages is matched with a donor with 25 years between marriages, the recipient may be allocated from the donor a length of first marriage of say, 23 years, an impossible length since it exceeds the interval between the recipient's two dates of marriage. A useful way to avoid this inconsistency is to impute not for the length of the first marriage but for this length as a proportion of the interval between the two marriages. In this way the imputed value is certain to satisfy the edit constraint. Inequalities of the type \( A + B + C \leq D \) are often met with quantitative variables. In such cases a transformation of the type described may be appropriate (see, for instance, I.G. Sande, 1979b).

Another commonly encountered feature of quantitative survey variables is that they do not apply, or have a zero value, for many members of the sample. Thus, for instance, many people do not have investment and property incomes, and do not receive SSI payments. For such variables, it is often useful to consider imputation in two stages: first to impute whether the variable is zero or not; and then, if not zero, to impute the amount. Herzog (1980) uses this two-stage approach with regression imputation for the amount of Social Security benefits received. The "zero-spike" procedure of Ford et al. (1980) also employs this two-stage approach. Sometimes the first-stage answer is known from a filter question; in this case, if the amount is not zero, the imputation should be based on those with non-zero amounts.

An additional complexity with imputation is that every one of the survey variables is potentially subject to missing data. One difficulty this creates is that the variables used for defining the imputation classes for a given item might themselves sometimes be missing. A further issue arising in this case is that when two or more items are missing on a record it is preferable, ceteris paribus, to impute the values for them from the same donor; otherwise the covariance between the items will be attenuated and inconsistent values may be imputed. A way to carry out joint imputations is to employ the same imputation classes for all the items concerned and then to use the same donor for all the missing items for a given recipient. However, this procedure may operate against a choice of imputation classes that maximizes the proportion of variance explained in an individual item. Often a compromise solution is necessary, making joint imputations for a group of closely-related items, but treating different groups of items separately. One approach is a sequential procedure (Coder, 1978; Brooks and Bailar, 1978, Section IVB): first, fill in the "small holes" in basic items which are to be used as initial imputation classes; second, impute for a group of closely-related items using one set of imputation classes; third, impute for another group of variables using a different set of imputation classes (which may or may not be defined to include variables from the first group of variables); and so on.

The almost infinite range of possible patterns of missing data in any survey record makes it impracticable to construct individual procedures for each pattern. Instead, some small number of procedures is needed to cover all the possibilities. Thus, for instance, a single set of imputation classes may be defined for all items in one topic area and then imputations for any missing data for the items in that area are done simultaneously from a single donor. If a record has missing responses for all items in the topic area, this procedure reflects the interrelationships between items in the donor's record. However, if some responses are present and others are absent, the procedure attenuates the covariances between those imputed and those reported, and may possibly give rise to inconsistencies. One way to avoid these outcomes is to incorporate the reported values in the topic area in defining the imputation classes for the missing values. However, as already observed, this solution is usually impracticable because of the many different patterns of responses and nonresponses to the set of items. Another solution is to assign the donor's values for all items to a recipient, irrespective of whether the recipient's responses are totally
or only partially missing. This solution is, of course, harmful for estimating univariate statistics for the various items, but is preferable for estimating the relationships between the items. If there are several items in the topic area, and only one or two responses are missing, this solution may be unattractive; if, however, only one or two responses are present, their deletion may create no serious loss of data.

When regression imputations with predicted values plus donor's residuals are used for several items, it may in the same way be advantageous to take residuals from the same donor for all of a recipient's imputed values. When such regression imputations are used for some items and some form of imputation class procedure is used for other items, there may again be benefits in using the same donor for a given recipient where possible. Procedures could easily be devised to achieve this objective.

The dissimilarities between weighting adjustments and imputation result in different considerations being involved in choosing imputation and weighting classes. As pointed out at the end of Chapter 3, the range of potential control variables from which weighting classes can be constructed is usually strictly limited, mainly to sample design variables and to variables whose distributions are available from an external source such as the Census. To be effective for a particular survey variable, weighting classes need to differ both in response rates and in the mean values of that variable. A typical survey collects data on many variables, each of which has a different set of relationships with the potential control variables, while on the other hand there is a single unit response/nonresponse variable that applies for all of the survey variables. For this reason the response variable plays a major role in determining weighting classes.

The imputation position differs in several respects. First, many more potential control variables are available. In addition to the control variables that may be used in forming weighting classes, all the other survey responses are available for use in forming imputation classes for a particular variable. This wealth of potential control variables highlights the need for an effective method for determining the combination of control variables to be used in defining imputation classes. Secondly, different sets of imputation classes can be considered for each survey variable (although, for the reasons already given, it may be preferable to use the same sets of classes for groups of variables). Thirdly, while the unit nonresponse rate is constant across all the survey variables, the item nonresponse rate changes from one variable to the next. And finally, as noted above, while weighting adjustments automatically retain relationships between items and maintain consistency, imputation does not do so. One consideration in forming imputation classes may therefore be the retention of certain relationships and the avoidance of certain inconsistencies.

As a consequence of these features of imputation, the prime consideration in forming imputation classes is the generation of good estimates of the missing values, estimates that are close to the actual values and that are consistent with other values on the record. There are two requirements for the achievement of this goal: one is that within each imputation class the donors' values should represent the recipients' values - i.e. the data missing at random assumption should hold - and the other is that within each class the donors' values should have little variance. In the ideal imputation classes, the values - both potential donors' and recipients' - have no variance. In this case, the imputation procedure is in effect a type of deductive imputation with the values being imputed without error.

In practice there is usually little information to guide the choice of imputation classes to satisfy the missing at random assumption, and therefore attention is mainly focussed on forming classes within which the potential donors are as homogeneous as possible with regard to the survey variable being considered. The general principle is then to form imputation classes that minimize the variance of the survey variable within classes, or equivalently that maximize the variance between classes. This is the principle behind the SEARCH technique - a successor to the Automatic Interaction Detector (AID) technique - used for determining imputation classes in the last section. In general the SEARCH technique
seems a valuable tool for guiding the choice of imputation classes. However, since the technique capitalizes on chance patterns in the data, it should be applied with caution. The technique requires large samples, and splitting should not be continued to the point that the ultimate classes contain small numbers of cases. (The SEARCH technique can also be used for determining weighting classes. In this case, given the importance attached to the response variable, it may be used as the dependent variable in the analysis. See, for instance, Chapman, 1974).

The requirement that the same set of imputation classes is to be used as the basis for imputing for a group of variables will necessitate some compromise between the classes derived by SEARCH for each of the variables taken individually. If the variables in the group are closely related to each other, however, similar classes are likely to be obtained from the separate SEARCH analyses, and a compromise may not be hard to reach. Another possible approach is to employ a multivariate version of SEARCH, seeking to maximize the variance between classes with respect to a set of variables. Gillo and Shelly (1974) describe a multivariate version of SEARCH known as MAID - M (Multivariate Automatic Interaction Detection - Monitored).

A further consideration in forming imputation classes is that the sample should not be too finely divided into many small classes. The reasoning is the same as that applied to weighting classes, namely that the resultant instability in the compensations gives rise to a greater increase in variance than can be offset by the likely decrease in bias. With weighting, the argument may be framed in terms of variation in the weights, with large weights occurring by chance for some classes. The corresponding argument with imputation may be put in terms of variation in the extent of use of donors, with donors in some classes being used several times. Multiple use of some donors needs to be carefully controlled. Monitoring the usage of donors should be an important component of an imputation auditing process.

Finally, we should note in this chapter the close relationship between editing and imputation (I.G. Sande, 1979a). When edit checks identify inconsistencies, some item responses are deleted and replaced by imputed values. In cases when many interrelated edit constraints are specified, as for instance often occurs with industrial statistics, the choice of which values to delete when inconsistencies are found is a complicated one. A principle, such as minimizing the number of deletions, may be required (Greenberg, 1981; Fellegi and Holt, 1976). Another aspect of the relationship between editing and imputation is the need for the imputed values to satisfy the edit constraints. When many constraints are employed, the choice of imputed values to satisfy them may be severely limited. Hill (1978) discusses automatic edit and imputation in the 1976 Canadian Census with categorical edits. G. Sande (1979b) has developed a procedure for linear edits with continuous variables.
5. Miscellaneous Topics

5.1 Multiple Imputations

A serious risk with the use of imputation is that the analyst may treat the completed data set as if all the data had been collected from respondents and no imputations had been made, and hence be misled into believing that the data are more reliable than they really are. There are two dangers involved in treating the imputed values as actual values. In the first place, imputations made for missing values are dependent on the imputation model employed, whether it is the implicit model involved in assigning donors' values to recipients or whether an explicit model, such as a regression equation, is used. The analyst needs to be conscious of this model dependence and, in particular, should consider how the use of alternative models could affect the results of the analyses. Secondly, although imputation assigns values for missing responses, it does not increase the amount of data collected. Imputation does not raise the sample size from r to n responses for an item; in fact, the effective sample size may be reduced to a number less than r, the extent of the reduction depending on the level of item nonresponse and the particular imputation procedure employed (see Section 2.3.1).

Rubin (1977b, 1978, 1979a,b, 1980) has widely advocated the use of multiple imputations to reduce and quantify these dangers. On the one hand, the replication of a given imputation model can provide information on the effect of the imputation on the stability of survey results. On the other hand, the application of several different imputation models can shed light on the sensitivity of the survey results to the choice of model. A further advantage of multiple imputations, already noted at the end of Section 2.3.1, is the reduction in the variance increase caused by random imputation; see also the note by Kish (1979) included as an Appendix to this report.

In developing the theory of multiple imputations, we start with a very general result, applicable for any statistic, any sample design and any imputation procedure. Following equation (2.3.1.1), the variance of any statistic z can be expressed as

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

(5.1.1)

where \( E_2 \) and \( V_2 \) are conditional expectation and variance over the random imputation scheme, and \( E_1 \) and \( V_1 \) are the corresponding operators for the initial sampling. The first term in (5.1.1) represents the variance from the imputation scheme. An estimator of this term can be readily generated from the use of multiple imputations, through the approach of simple replicated sampling. First, form a complete data set by applying the given imputation scheme to impute values for all missing responses. Repeat this operation independently c times, to produce c complete data sets. Compute the statistic \( z_j \) for data set j, and let \( z = \frac{\sum z_j}{c} \) be the average estimator based on all c data sets. Then \( V_2(z) = \frac{V_2(z_j)}{c} \), where \( V_2(z_j) \) is the common variance of each of the \( z_j \)'s. An unbiased estimator of \( V_2(z_j) \) may be simply obtained from the random sample of \( z_j \)'s as \( v(z_j) = \frac{\sum (z_j - z)^2}{c - 1} \). Since \( E_2[v(z_j)/c] = V_2(z) \), it follows that \( E_1 E_2[v(z_j)/c] = E_1 V_2(z) \), i.e. that \( v(z_j)/c \) is an unbiased estimator of the imputation variance.

The generality of the above result is limited only by the assumption that the replications of the imputations are conducted independently and that the overall estimator is formed as the average of the replicate values. The latter assumption may not be met in practice, for the analyst may opt for the overall estimator formed from pooling the c data sets, \( \tilde{z} \), rather than the average value, \( z \). With linear estimators \( \tilde{z} = z \), but the equality does not hold for nonlinear statistics, and \( \tilde{z} \) may be preferred on the grounds of a smaller bias. However, since \( \tilde{z} \) and \( z \) are generally close, it is common practice with replicated sampling to approximate the variance of \( \tilde{z} \) by \( v(z_j)/c \) (see, for instance, Kalton, 1977). The repeated replication imputation procedure described by Kish in the Appendix and the procedure of partitioning weights described at the end of Section (2.3.1) both create pooled data sets from which the estimator \( \tilde{z} \) would be obtained.

The preceding discussion has been concerned with estimating the imputation variance component \( E_1 V_2(z) \) of the total variance of \( z \). When imputation is used, care is also
needed to obtain a suitable estimator of the sampling variance component, $V_1E_2(Z)$. In order to illustrate the issues involved, we will consider the simple situation of a SRS of size n, from which m y-values are missing. These missing values are then imputed by random imputation using unrestricted sampling within a single imputation class. Multiple imputations are used. The estimator we will consider is the sample mean $\bar{Y} = E_1V_2(\bar{Y}) + V_1E_2(\bar{Y})$, may be readily obtained as follows. First note that

$$\bar{Y} = \sum (r_1 \bar{Y} + m_1 \bar{Y} m_j)/nc = \bar{Y} + (m_1 \bar{Y} m_j/c). \tag{5.1.2}$$

Then $V_2(\bar{Y}) = m_1^2 \sum V_2(\bar{Y} m_j)/c^2$ with independent replications,

$$= \bar{m}^2(r - 1)s^2_{tr}/nc \text{ with unrestricted sampling,}$$

$$= \bar{m}_1^2 \text{ treating } r - 1 = r. \tag{5.1.3}$$

Thus $E_1V_2(\bar{Y}) = m_1^2 \sum V_2(\bar{Y} m_j)/nc$. Now $E_2(\bar{Y}) = \bar{Y}_r$, so that, ignoring the f.p.c. term $[1 - (r/R)]$,

$$V_1E_2(\bar{Y}) = s^2_{tr}/r. \tag{5.1.4}$$

Hence

$$V(\bar{Y}) = (s^2_{tr}/r) + (m_1^2 \sum V_2(\bar{Y} m_j)/nc). \tag{5.1.5}$$

In this simple case a variance estimator for $V(\bar{Y})$ can be easily obtained providing the respondents' values and the imputed values can be distinguished: an unbiased estimator $s^2_{tr} = \sum (Y_{r1} - \bar{Y}_r)^2/(r - 1)$ of $s^2_{tr}$ can be calculated, and substituted in (5.1.5). There is no need to employ the replication variance estimator $v(\bar{Y})/c$ to estimate the imputation variance - the second term in (5.1.5). Using (5.1.2), the replicated variance estimator in fact reduces to

$$v(\bar{Y})/c = \sum (\bar{Y}_j - \bar{Y})^2/c(c - 1)$$

$= m_1^2 \sum (\bar{Y}_m_j - \bar{Y}_m)^2/c(c - 1) = m_1^2 V(\bar{Y} m_j)/c.$

Since $E_2[v(\bar{Y} m_j)] = s^2_{tr}/m$, the conditional expectation of the replicated variance estimator is $m_1^2 s^2_{tr}/nc$, the value that would otherwise be substituted directly in (5.1.5). In this case, the replicated variance estimator is just a less precise estimator of $E_1V_2(\bar{Y})$ than would be obtained by the direct substitution of $s^2_{tr}$ for $s^2_{tr}$ in the second term of (5.1.5).

Suppose now, however, that the respondents' values and imputed values cannot be distinguished, a not infrequent occurrence in practice; in addition, assume that the numbers of respondents ($r$) and nonrespondents ($m$) are unknown. In this situation, the average sample variance across replicates $s^2 = \sum s^2_{tr}/c$, where $s^2_{tr} = \sum (Y_{r1} - \bar{Y}_j)^2/(n - 1)$, provides a reasonable estimator for $s^2_{tr}$ (see Section 2.3.2), but the variance in (5.1.5) cannot be estimated because $r$ and $m$ are unknown. The replication estimator of the imputation variance can, however, still be computed, so that the main problem concerns the estimation of the sampling variance ($s^2_{tr}$). Without knowledge of the extent of missing data, the analyst might be tempted to estimate the sampling variance by $s^2/r$. But this estimator effectively treats the imputed values as actual values and hence overstates the precision achieved.

Rubin (1979b) has proposed the following general estimator of the variance of a statistic $z$ with multiple imputations, an estimator which does not require knowledge of which values are actual ones and which are imputed,

$$v(z) = (\sum V_j/c) + \sum (z_j - \bar{z})^2/(c - 1) \tag{5.1.6}$$

where $V_j$ is the variance of $z_j$ obtained by treating both imputed and real values as real values. Applying this estimator in the current case gives
From Section (2.3.2) \[ E(s_j^2) = S_r^2 \] and from above \[ E[v(\bar{y}.)] = \bar{m}S_r^2/n \], so that
\[ E(v) = S_r^2(1 + \bar{m})/n. \]
Thus
\[ (S_r^2/r - E(v)) = m^2S_r^2/r > 0. \]

This result shows that \( v \) underestimates the sampling variance of \( \bar{y} \), the first term in (5.1.5), let alone the total variance.

To counteract this bias, Rubin (1979) has proposed a modification to the multiple imputation procedure, developing the modification from a Bayesian perspective. We will attempt to explain it from a different viewpoint. First, we recall the missing at random assumption made about the respondents and nonrespondents, an assumption which implies that \( \bar{y}_m = \bar{y}_r \). Not knowing \( \bar{y}_m \), the best estimator of \( \bar{y} \) under this assumption (i.e. the one with smallest variance) is \( \bar{y}_r \), which can be generated by carrying out the imputations so that \( \bar{y}_m = \bar{y}_r \). The condition \( \bar{y}_m = \bar{y}_r \) can be achieved by imputing \( \bar{y}_r \) for all missing values, but, as has been observed, this procedure distorts the distribution of the \( y \)-variable. A random imputation procedure assigning donors' values to recipients avoids this distortion, but at the cost of increasing the variance of the mean. The variance increase results from the relaxation of the condition that \( \bar{y}_m = \bar{y}_r \) to one that \( E_2(\bar{y}_m) = \bar{y}_r \). A further relaxation of the condition to \( E_2(\bar{y}_m) = \bar{y}_r \) can be used to assist with the variance estimation problem under discussion.

The condition \( E_2(\bar{y}_m) = \bar{y}_r \) can be expressed as
\[ E_2(\bar{y}_m) = \bar{y}_r + (\bar{y}_r - \bar{y}_r) = \bar{y}_r + d. \]

The value of \( d \) is of course unknown for a particular sample, but its sampling distribution is known to be normal with a mean of zero and variance of \( S_r^2/r \), i.e. \( N(0, S_r^2/r) \). Estimating \( S_r^2 \) by \( S_r^2 \), Rubin's modification is then to draw a value of \( d \) from \( N(0, S_r^2/r) \) and add this value, \( d_j \), to all imputed values in replicate \( j \). This modification changes the overall sample mean over the \( c \) replicates from \( \bar{y} \) in (5.1.2) to
\[ \bar{y}^* = \frac{\Sigma(\bar{y}_r + \bar{m}d_j)}{c} \]
\[ = \frac{\Sigma(\bar{y}_r + \bar{m}d_j)}{c} + \bar{m}d_j/c. \] (5.1.8)

To establish that \( \bar{y}^* \) is an unbiased estimator of \( \bar{y}_r \), we introduce another level of expectation, \( E_3 \), denoting expectation over random sampling of \( d_j \). Since \( E_3(d_j) = 0 \), \( E_3(\bar{y}^*) = \bar{y} \). Hence \( E(\bar{y}^*) = \bar{y}_r \).

The variance of \( \bar{y}^* \) may be derived from
\[ V(\bar{y}^*) = V_1E_2E_3(\bar{y}^*) + E_1V_2E_3(\bar{y}^*) + E_1E_2V_3(\bar{y}^*) \] (5.1.9)
\[ = V(\bar{y}) + E_1E_2V_3(\bar{y}^*) \]
where the second term represents the loss in precision in using the estimator \( \bar{y}^* \) rather than \( \bar{y} \). From (5.1.8), approximating \((r - 1) = r\),
\[ V_3(\bar{y}^*) = \bar{m}^2S_r^2/rc \] (5.1.10)
and
\[ E_1E_2V_3(\bar{y}^*) = m^2S_r^2/cr. \] (5.1.11)

Thus, using (5.1.5)
\[ V(\bar{y}^*) = (S_r^2/r) + \bar{m}S_r^2/nc + m^2S_r^2/rc \] (5.1.12)
\[ = (S_r^2/r) + \bar{m}S_r^2/nc(1 + (m/r)) \] (5.1.13)
\[ = (S_r^2/r) + \bar{m}S_r^2/rc. \] (5.1.14)
The effect of the modification to the imputation procedure is thus to increase the imputation variance by a factor of \( n/r \) (c.f. equation (5.1.5)).

Turning to the issue of estimating \( \Sigma (\hat{y}^*_r)^2 \), we will first establish that \( s^2 = \Sigma \hat{s}_j^2 \) provides a reasonable estimator of \( S^2_r \). Consider \( (n-1)s_j^2 = \Sigma (y_{ji}^* - \hat{y}_r^*)^2 \).

\[
\Sigma (y_{ji}^* - \hat{y}_r^*)^2 = \Sigma (y_{ri} - \hat{y}_r)^2 + \Sigma (y_{mj1}^* - \hat{y}_r^*)^2 - n(\hat{y}_r^* - \hat{y}_r)^2
\]

= \[ \Sigma (y_{ri} - \hat{y}_r)^2 + \Sigma (y_{mi} + d_j - \hat{y}_r^*)^2 - n(\hat{y}_r^* - \hat{y}_r)^2 \].

Now \( \hat{y}_r^* - \hat{y}_r = m(d_j + \hat{y}_m - \hat{y}_r)/n \), so that, with \( w_i = (1 + g_i) \) where \( g_i \) is the number of times the \( i \)th respondent is used as a donor,

\[
\Sigma (y_{ji}^* - \hat{y}_r^*)^2 = \Sigma w_i (y_{ri} - \hat{y}_r)^2 + m d_j + 2 m d_j (y_m - \hat{y}_r)
\]

= \( (m^2 d_j^2/n) - [2 m d_j (y_m - \hat{y}_r)/n] \)

= \( [m^2 (y_m - \hat{y}_r)^2/n] \).

Hence,

\[
E_3 \Sigma (y_{ji}^* - \hat{y}_r^*)^2 = \Sigma w_i (y_{ri} - \hat{y}_r)^2 + (m^2 d_j^2/n) - (m^2 d_j^2/rn)
\]

= \( [m^2 (y_m - \hat{y}_r)^2/n] \).

and

\[
E_2 E_3 \Sigma (y_{ji}^* - \hat{y}_r^*)^2 = [n(r-1)s_j^2/r] + [m s_j^2 (1 - m)/r]
\]

- \[ m^2 (r - 1) s_j^2 /nr \] = \( [s_j^2 (r - 1)(m - \hat{m})/r] + \hat{m} s_j^2 \)

= \( s_j^2 [n - (n/r) + (m/\hat{m})] \) \hspace{1cm} (5.1.15)

= \( s_j^2 [n - (n/r)] \) if \( r \) is large.

Thus \( E_2 E_3 (s_j^2) = s_r^2 \), and hence \( E(s^2) = s_r^2 \).

Now consider the replication variance estimator \( \Sigma (\hat{y}^* - \hat{y}_r^*)^2 / (c - 1) \). Being a replication estimator it is unbiased for the imputation variance, i.e. from (5.1.14) \[
E(\Sigma (\hat{y}^* - \hat{y}_r^*)^2 / (c - 1)) = \hat{m} s_r^2 / rc
\]

so that

\[
E(\Sigma (\hat{y}^* - \hat{y}_r^*)^2) = \hat{m} s_r^2 / r.
\]

Combining this result with \( E(s_j^2) = s_r^2 \) gives the expectation of \( \Sigma (\hat{y}^* - \hat{y}_r^*)^2 / (c - 1) \) as

\[
E(\Sigma (\hat{y}^* - \hat{y}_r^*)^2) = \hat{m} s_j^2 / r.
\] (5.1.16)

Comparing \( E(\Sigma (\hat{y}^* - \hat{y}_r^*)^2) \) with \( E(\Sigma (\hat{y}^* - \hat{y}_r^*)^2) \) in (5.1.14) shows that \( \Sigma (\hat{y}^* - \hat{y}_r^*)^2 / (c - 1) \) is an unbiased estimator of the sampling variance. Providing the imputation variance is small relative to the sampling variance, as will occur when response rates are high and/or when the number of multiple imputations is sizeable, \( \Sigma (\hat{y}^* - \hat{y}_r^*)^2 / (c - 1) \) will provide a reasonable estimator of \( V(\hat{y}^*) \). In other cases, the additional term of \( \Sigma (\hat{y}^* - \hat{y}_r^*)^2 / (c - 1) \) may be added to reflect the imputation variance. If this is done the variance estimator becomes

\[
\Sigma (\hat{y}^* - \hat{y}_r^*)^2 / (c - 1)
\]

with \( E(\Sigma (\hat{y}^* - \hat{y}_r^*)^2) = \hat{m} s_r^2 / r \).

An alternative approach is to increase the variance of the \( d_j \) to make \( \Sigma (\hat{y}^* - \hat{y}_r^*)^2 / (c - 1) \) an unbiased estimator of the variance of \( \hat{y}_r^* \). Let \( V(d_j) = k s_r^2 / r \). Then \( V(\Sigma (\hat{y}^* - \hat{y}_r^*)^2) \) in (5.1.10) becomes \( V(\Sigma (\hat{y}^* - \hat{y}_r^*)^2) = (\hat{m}^2 k s^2 / rc) \) for the revised estimator \( \hat{y}^*_1 \). Hence

\[
E(\Sigma (\hat{y}^*_1)^2) = \hat{m} k s^2 / rc
\] (5.1.18)

and

\[
V(\hat{y}^*_1) = (S^2_r / r) + \hat{m} s^2 (\hat{r} + \hat{k} m)/rc.
\] (5.1.19)

Now \( E_2 E_3 (\hat{y}_{ji}^* - \hat{y}_r^*)^2 \) in (5.1.15) becomes

\[
E (\Sigma (\hat{y}^* - \hat{y}_r^*)^2) = \hat{m} s_r^2 / rc
\]

118

119
which remains approximately equal to \((n - 1)S_r^2\) for large \(r\), providing \(k\) is small. Hence, under this proviso, \(s^2\) is still a reasonable estimator of \(S_r^2\). The expected value of \(v^*\) is then

\[
E(v^*) = \left(\frac{s_r^2}{n}\right) + \frac{\bar{m}S_r^2(r + k\bar{m})}{r} \tag{5.1.21}
\]

This is equal to \(V(\tilde{y}_r^*)\) in (5.1.19) if

\[
\left(\frac{1}{r}\right) - \left(\frac{1}{n}\right) = \bar{m}(r + k\bar{m})(c - 1)/rc
\]

i.e. if

\[
k = 1 + \left[1/(\bar{m}(c - 1))\right]. \tag{5.1.22}
\]

With the smallest possible value of \(c = 2\), \(k = 1 + (1/\bar{m})\), in which case the last term in (5.1.20) is \(\bar{m}(k - 1) = \bar{m}\). Hence \(k\) is sufficiently small to make \(s^2\) a reasonable estimator of \(S_r^2\). Thus, if the \(d_j\) are chosen from \(N(0, \sigma^2)\), where

\[
\sigma^2 = \left(\frac{s_r^2}{r}\right)\left[1 + \left[1/(\bar{m}(c - 1))\right]\right],
\]

the estimator \(v^*\) is approximately unbiased for \(V(\tilde{y}_r^*)\).

The reason for incorporating the \(d_j\) term into the imputation procedure is to provide a variance estimator \(v^*\) which reflects the true variance of the sample mean \(\tilde{y}^*\). In deciding whether to include the \(d_j\)'s, this benefit has to be balanced against the loss of precision in \(\tilde{y}^*\) given by (5.1.11) or the larger loss with the amended modification given by (5.1.18). In the case of Rubin's modification, the variance is increased over the simple multiple imputation variance, \(V(\tilde{y})\) in (5.1.5), by the factor

\[
V(\tilde{y}^*)/V(\tilde{y}) = \left[\left(\frac{s_r^2}{r} + (\bar{m}s_r^2/rc)\right)/\left(\frac{s_r^2}{r} + (\bar{m}s_r^2/nc)\right)\right]
\]

\[
= \left[\frac{1 + (\bar{m}/c)\left[1 + (\bar{m}r/c)\right]}{\bar{m}(c - 1)}\right] = 1 + \left[\frac{\bar{m}^2/(c + \bar{m}r)}{c}\right]. \tag{5.1.23}
\]

Thus, for instance, with \(c = 2\) and \(\bar{m} = 0.3\), the factor is 1.04, a 4 per cent increase in the variance of the sample mean. The larger the value of \(c\), the smaller this factor is.

With the amended modification \(V(\tilde{y}_r^*)\) from (5.1.19) becomes

\[
V(\tilde{y}_r^*) = \left(\frac{s_r^2}{r}\right)\left[1 + \left[\frac{\bar{m}(c - 1)}{c}\right]\right]
\]

so that

\[
V(\tilde{y}_r^*)/V(\tilde{y}) = \left[1 + \left[\frac{\bar{m}(c - 1)}{c}\right]\right]/\left[1 + (\bar{m}/c)\right] = 1 + \left[\frac{\bar{m}^2/(c - 1)(c + \bar{m})}{c}\right]. \tag{5.1.24}
\]

With \(c = 2\) and \(\bar{m} = 0.3\), the variance of the mean with the amended modification is thus increased by a substantial factor of 1.13 over that of the mean with the original modification.

A decision to be made with any multiple imputation procedure is the number of replications, \(c\), to employ. The larger the value of \(c\), the smaller is the imputation variance, and hence the greater is the precision of the sample mean, Rubin (1979) shows that, with unrestricted sampling of donors, the choice of \(c = 2\) goes a good way towards removing the imputation variance. With a SRS of donors, the imputation variance with two replications is even smaller (see the Appendix), and it would be smaller still if a proportionate stratified sample of donors - as discussed in Section 2.3.1 - were used (see also Kalton and Kish, 1981). Thus, with regard to the reduction of the imputation variance, most of the gain of multiple imputations can be achieved from a small number of replications, even the minimum of 2.

When multiple imputations are used for the purpose of variance estimation, the issue of the precision of the variance estimate arises. The replication variance estimator \([v(z)/c]\) of the conditional imputation variance \(V_2(z)\) discussed earlier has \((c - 1)\) degrees of freedom. A sizeable number of degrees of freedom (say 20 to 30 or more) is needed if this estimator is to be of adequate precision for analysis purposes. The argument here corresponds to the usual one concerning the precision of replication variance estimators (see Kish, 1965, ...
The concern for precision also applies to the overall variance estimator for $z$, viz given in (5.1.6).

Practical considerations also play an important role in determining the choice of $c$, and indeed in whether multiple imputations should be employed at all. The added complexity and cost of analyzing the data sets resulting from multiple imputations are major hurdles that need to be cleared. The larger the number of replications the larger the total data set and hence the more costly the analyses.

The total data set for multiple imputations has been described above as a set of $c$ independently completed data sets, with any statistic of interest being computed as the average $\bar{z}$ of the $c$ replicate values, $z_j$. For economy of processing the total data might be better constructed along the lines suggested in Chapter 2 of splitting records with missing data into $c$ parts. With this scheme, the total data set comprises the records with no missing data and $c$ parts of each of the records with missing data, each part being assigned a weight of $1/c$ (or $\omega_i/c$ if record $i$ has a weight of $\omega_i$). Each part of a record with missing data is given an index $(1, 2, \ldots, c)$ for variance estimation purposes. The $z_j$ needed for the computation of $v(z_j)$ are obtained by confining calculations to the records with no missing data and the $j$th part of split records, with the latter being given the total weight of their original records. Since the records with no missing data are included in the calculation of all the $z_j$'s, computing economies that calculate their contribution to the $z_j$'s only once can be readily envisaged for many types of statistics.

Although this alternative construction of the data set permits the calculation of the overall estimate as $\bar{z}$, the more natural estimator to use is $\hat{z}$, the single estimator obtained from the total data set. The use of $\hat{z}$ avoids the need to compute the $z_j$'s and for reasons of bias and variance is often preferred to $\bar{z}$. Since the variance estimator used for $\hat{z}$ is the same as that for $\bar{z}$, and this depends on $z_j$, it is of course true that the $z_j$'s have to be calculated for any statistic for which a variance estimate is wanted. However, in complex surveys the number of estimates made exceeds by manyfold the number of variance estimates computed, and the $z_j$'s need not be calculated for estimates whose variance estimates are not required.

The economy achieved through using this alternative construction of the data set depends on the extent and pattern of missing data. If the missing data are concentrated in a few records, the increase to the size of the data set by splitting these records into $c$ parts may be relatively minor. If, however, the missing data are spread across most of the records, with say some records having missing responses for certain items and other records having missing responses for others, the size of the data set will increase substantially. When the situation begins to approach that in which all records have some missing data, it may be simplest to split all records into $c$ parts, and impute for any missing data as necessary; in this case the size of the data set is increased $c$-fold.

To make the $z_j$'s independent, as required by the theory of replicated sampling, the imputations have been described above as being carried out independently from one replication to the next. From the viewpoint of precision of the estimator ($\hat{z}$ or $\bar{z}$), however, it would be preferable to assign the imputed values by some nonindependent procedure, say using stratified or systematic sampling for all $c$ parts simultaneously. This is the standard conflict with replicated sampling: should one opt for an unbiased variance estimator of an estimate with larger variance, or for an upwardly biased variance estimator of an estimate with smaller variance? There is in general no obvious answer to this question. In the present case, however, since the use of even a small number of replications effectively removes most of the imputation variance, replications with independent imputations should generally serve well.

5.2 Weighted Data

Until this point, this report has assumed that the compensations for missing data are made to records which all have the same weight. In the case of weighting adjustments for unit nonresponse, the sampled units were assumed to be selected with equal probabilities. When units are selected with unequal probabilities, selection weights proportional to the inverse of the selection probabilities are needed in the analysis. In this section we consider the effect of unequal selection
weights on nonresponse weighting adjustments. In the case of imputation for item nonresponse, the sample units may have weights both because of unequal selection probabilities and because of weighting adjustments made for unit nonresponse. We will discuss how these weights might affect the method of imputation used. Little attention has been given to the effects of weights on missing data compensation procedures; more research in this area seems needed.

Consider first the effect of selection weights on nonresponse weighting adjustments. The usual extension to procedures already described to take account of selection weights is to reweight according to the sums of selection weights rather than sample sizes. As an illustration, consider the estimator of the population mean with sample weighting adjustment for nonresponse. Let $q_{rh}$ be the selection weight for respondent $i$ in weighting class $h$, and let $q_{rh} = \Sigma q_{rhi}$; let $q_{mh}$ be the equivalent quantities for the nonrespondents; and let $q_h = q_{rh} + q_{mh}$. The adjusted estimator of the population mean is then $\bar{Y}_s = \Sigma q_{rhi}Y_{rhi}/\Sigma q_{rhi}$, where $Y_{rhi} = \Sigma q_{rhi}Y_{rhi}/\Sigma q_{rhi}$. This may be expressed as $\bar{Y}_s = \Sigma q_{rhi}Y_{rhi}/\Sigma q_{rhi}$ with $w_{hi} = q_{rhi}/q_{rh}$. In other words, the selection weight is increased by the factor $q_h/q_{rh}$ to compensate for the nonrespondents. This estimator is a straightforward generalization of the estimator with sample weighting adjustments discussed in Section 3.3.

In general the final weights assigned, i.e. the combined selection weights and sample weighting adjustments ($w_{hi}$), will vary between respondents within a weighting class. There is, however, a special case when all sampled units in a weighting class have the same selection weights, say $q_h$; in this case $w_{hi} = q_{rhi}(n_h/r_h)$. This situation would, for instance, occur with a disproportionate stratified sample in which the strata are used in forming weighting classes. For example suppose that a certain subgroup of the population (e.g. the elderly) were sampled from a separate sampling frame with a higher sampling fraction than used for sampling the rest of the population, with epsem sampling within the two strata. If weighting classes were formed so that no class contained both persons sampled from the elderly stratum and persons sampled from the other stratum, then the condition for the equality of final weights for respondents within weighting classes would hold.

It may be noted that in this special case the nonresponse sample weight adjustment $q_h/q_{rh}$ reduces to $n_h/r_h$. In general, the equality of $q_h/q_{rh}$ and $n_h/r_h$ will not hold exactly. However, if the assumption of data missing at random within the weighting class is valid, the equality should hold approximately, that is within random chance fluctuations. This may be demonstrated by considering the $q$-variable as simply another survey variable. We will condition on the selected sample units in the weighting class and the number of respondents, and assume that data are missing at random. Then with $q_{rh}/r_h$ being the respondent mean weight and $q_{mh}/m_h$ being the nonrespondent weight, it follows from the missing-at-random assumption that

$$E_2(q_{rh}/r_h) = E_2(q_{mh}/m_h) = E_2(q_h/n_h)$$

i.e.

$$E_2(q_h)/E_2(q_{rh}) = n_h/r_h.$$

The implication of this result is that a marked difference between $q_h/q_{rh}$ and $n_h/r_h$ suggests that the missing-at-random assumption may be violated. Checks should be made on this difference, and a general comparison made between the selection weights of respondents and nonrespondents within weighting classes, to examine the appropriateness of the missing-at-random assumption. If marked differences are found, consideration needs to be given to incorporating the information reflected in the selection weights into the construction of the weighting classes. If selection weights are related to response rates and also to the survey variables, biased results will occur unless some nonresponse adjustments based on selection weights are made.

The common practice in making imputations with weighted data is to proceed as described in Chapter 4 without taking account of the weights. Little work has been done to examine
the appropriateness of this procedure, but two observations can be made. First, it is apparent that there is a danger of increased imputation variance if a sizeable variability in selection weights is ignored. For example, recalling the equivalence between imputation and weighting adjustments for estimating a mean or total, imputing a donor's value to a nonrespondent with a large weight is equivalent to increasing the donor's weight by that large weight. A simple method for avoiding this large weight increase would be to partition the records and weights of recipients with sizeable weights, and make separate imputations for the parts along the lines already described at the end of Section 2.3.1 and earlier in this chapter. The added feature here is that the recipient records are partitioned into variable numbers of parts depending on the magnitudes of their weights. A possible elaboration on this method would be to partition both respondents' and nonrespondents' records into multiple parts, each part having the same weight across all sampled units. The equality of weights for all parts can be achieved by fixing the number of parts for a sampled unit proportional to its weight; thus, for instance, with integer weights, units with weights 2, 3, 5 could be divided into 2, 3 and 5 parts respectively. Imputations could then be carried out with the equal-weighted parts. This procedure bears some similarity to the probability-proportional-to-weight scheme of Cox (1980) as described below, except that in her scheme each recipient is assigned a single donor's value whereas here the several parts of a recipient's record may be assigned from different donors.

The second general observation to be made about imputations with weighted data is that, as discussed above in the case of weighting adjustments, the weights may be significantly related both to the pattern of missing data and the survey variables. If so, it may be advisable to incorporate them in some way into the control variables used for forming the imputation classes; such a procedure would in effect be matching recipients and donors in terms of their weights.

Suppose now that imputation classes are formed in a way which leaves variability in the weights within classes, and assume that the data are missing at random within classes. A consequence of this assumption is that the class population mean of respondents is equal to that of nonrespondents,  \( \bar{y}_r = \bar{y}_m \) or  \( E_\xi(\bar{y}_r) = E_\xi(\bar{y}_m) \) as discussed in Section 2.1 (we will drop the imputation class subscript \( h \) throughout this discussion).

Following the line of argument used in Chapter 2, we may then seek an imputation procedure satisfying the condition  \( E_2(\bar{y}_m) = \bar{y}_r \), where  \( E_2 \) is the expectation over repeated applications of the imputation procedure. Cox (1980) has devised a procedure that satisfies this condition with weighted data.

We will use the following notation to describe a simplified version of the procedure. Let  \( w_{ri} \) and  \( y_{ri} \) be the weight and \( y \)-value for respondent 1, with  \( w_{ri} = w_{ri}/w_{ri} \); similarly for recipient 1, let  \( w_{mi} \) be the weight and  \( y_{mi} \) be the imputed value, with  \( w_{mi} = w_{mi}/w_{mi} \). The respondent mean is then  \( \bar{y}_r = \sum w_{ri} y_{ri} \) and the respondent mean is  \( \bar{y}_m = \sum w_{mi} y_{mi} \).

Cox's procedure resembles a PPS sampling scheme from cumulative totals of the respondents' weights, but taking into account the recipients' weights as well. It is perhaps most simply explained in terms of the diagram on the next page. In the diagram the seven respondents and five recipients are listed in two columns, the columns being split up according to the weights of the units. The column of recipients' weights is mapped into the respondents' column as indicated by the dotted lines, and a recipient's imputed value is taken from his part of the respondents' column. For instance, recipient 1's imputed value comes from either respondent 1 or 2, while recipient 2's value comes from one of respondents 2, 3, 4 or 5. The weight for respondent 2 is divided into  \( w_{r2.1} \) and  \( w_{r2.2} \), the first part being associated with recipient 1 and the second part with recipient 2. Similarly, the weight for respondent 5 is divided between recipients 2 and 3. The choice between respondents 1 and 2 as donor of the imputed value for recipient 1 is made by choosing a number at random up to  \( w_{r1} + w_{r2.1} \). If the number is over  \( w_{r1} \), respondent 2 is chosen; otherwise respondent 1 is chosen. In the same way the donor for recipient 2 is chosen by drawing a random number up to  \( w_{r2.2} + w_{r3} + w_{r4} + w_{r5.2} \), and determining the donor from respondents 2, 3, 4 and 5 by a standard PPS selection from the
cumulative totals of these values (see, for instance, Kish, 1965, Chapter 7). This procedure is applied to impute a value for each of the recipients in the class.

We will now show that this procedure satisfies the condition that $E_2(\tilde{y}_m) = \tilde{y}_r$. For this purpose, the recipients' mean $\Sigma w_{mj}^* y_{mj}$ can be expressed as

$$\tilde{y}_m = \Sigma a_i y_{ri}$$

where $a_i = \Sigma w_{mj}^* y_{mj}$ and $d_{ij} = 1$ if respondent $i$ donates his value to recipient $j$ and $d_{ij} = 0$ if not. Then

$$E_2(\tilde{y}_m) = \Sigma E_2(a_i) y_{ri}$$

and

$$E_2(a_i) = \Sigma E_2(d_{ij}) w_{mj}^*.$$ 

Now $E_2(d_{ij} | j) = P(d_{ij} = 1 | j)$. In general, suppose that the mapping of recipient $j$'s weight covers an amount $w_{ri,j}$ of respondent $i$'s weight. Then

$$P(d_{ij} = 1 | j) = w_{ri,j}^* / w_{mj}^*.$$ 

Thus

$$E_2(a_i) = \Sigma w_{ri,j}^*(w_{ri,j}^* / w_{mj}^*) w_{mj}^* = \Sigma w_{mj}^* r_{i,j}.$$ 

Since all of respondent $i$'s weight is associated with some recipient, $r_{i,j} = w_{mj}^*$. Thus $E_2(\tilde{y}_m) = \Sigma w_{ri}^* y_{ri} = \tilde{y}_r$, as required. This result is derived conditional on the particular order chosen, but it clearly also applies unconditionally.

The above procedure fairs well with regard to the avoidance of multiple uses of donors within the restrictions imposed by variable weights, but there is a slight modification which can further improve the position. As indicated in the diagram, there is a chance for a respondent mapped into two recipients to be chosen as the donor for both of them, by the choice of a high random number to determine the donor for the first recipient and a low one for the second. If a single random number between 1 and $K$ is chosen, and scaled by a factor
$w_{m1}/K$ for the first recipient and $w_{m2}/K$ for the second, the use of these two scaled numbers to determine the two donors will minimize the chance of the overlap donor being selected twice. Since this modification does not affect the above theory, it remains true that $E_n(y_m) = y_r$. Cox's procedure, which is a more sophisticated version of that described above, minimizes the multiple use of donors.

APPENDIX

Repeated Replication Imputation Procedure (RRIP)

By: Leslie Kish

This technique obtains dramatic reductions in the increases of the variance caused by imputation for nonresponse. Like other techniques for imputations with simple randomized replications it makes several simplifying implicit assumptions:

(1) The sample is divided into two parts in which the response rates are $R$ and $R(1 - W)$ respectively, and the two parts are treated as if they were strata.

(2) To control for (eliminate?) the differential bias due to the differential response rates the weight of the second group is increased in the ratio $1/(1 - W)$.

(3) This increase could be made with the weight $(1 - W)$ assigned to each case in the group, or by assigning the mean of the $(1 - W)$ portion to a fraction $W$ of imputed cases, or some other way. Instead it is decided to select with simple random sampling without replacement the fraction $W/(1 - W)$ of the cases and duplicate them.

(4) The proportionate increase due to randomized duplication is the second term in the variance $1 + W(1 - W)/(1 + W)^2$.

The last term is a special case of the increase where the portion $(1 - W)$ has the weight $k$ (integer) and portion $W$ (SRS without replacement) has the weight $(k + 1)$. That increase in the variance has been shown (Kish, 1965, Section 11.7B, and 1976, Section 7.5) to be the second term in the ratio of SRS variances, i.e. in $1 + [(W(1 - W)/(k + W)^2)]$.

This ratio decreases rapidly with $k$. The increase has a maximum of $M = 1/4k(k + 1)$ at $W = k/(2k + 1)$. Values of $W$ and $M$ for $k \leq 5$ are given in Table 1.
The procedure (RRIP) advanced here is based on the rapid decreases obtainable by making \( k > 1 \), and \( k \) of 2, 3, 4 will do very well. The procedure is simple. We suppose that the different response portions have been identified and \( W_1 \) is the proportion of duplication needed in the \( i^{th} \) portion.

(1) First make \( k \) (2, 3 or 4) replications of the entire sample of both (or all) portions. This has no effect on simple estimates on their variances, except for the constants of adjustment \( k \) and \( k^2 \).

(2) From one of the \( k \) replicates duplicate \( kW_1 \) in the \( i^{th} \) portion.

(3) If \( kW_1 > 1 \), duplicate one (or more) entire replicates plus \((kW_1 - r) < 1 \) duplicates from the next replicate, with \( r = 1 \) or 2 or 3...

The variance ratio is

\[
1 + \frac{[kW(1 - kW)]/(k + kW)^2}{1 + [W(1 - W)/(1 + W)^2]} - \frac{W(k - 1)/k}{(1 + W)^2}
\]

The second term is the increase due to simple duplication and the third term shows the reduction wrought by the RRIP procedure. Table 2 presents for a few values of \( W \) and for \( k = 1, 2, 3, 4 \) and 5 the net increase in the variance on the top and the reduction on the bottom. The net increase becomes .0000 when \( kW = 1 \), and is shown with a "*". When \( kW > 1 \), \( k \) in the formula becomes \((k + 1)\) in actuality.

---

### Table 1: Maximum proportionate increase in variance, \( M = 1/4(k + 1) \), at \( W = k/(2k + 1) \).

<table>
<thead>
<tr>
<th>( k )</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \frac{1}{3} )</td>
<td>2/5</td>
<td>3/7</td>
<td>4/9</td>
<td>5/11</td>
<td></td>
</tr>
<tr>
<td>( \frac{1}{8} )</td>
<td>1/24</td>
<td>1/48</td>
<td>1/80</td>
<td>1/120</td>
<td></td>
</tr>
</tbody>
</table>

---

### Table 2: Net increase in variance (on top) and reduction due to RRIP (on bottom).

<table>
<thead>
<tr>
<th>( W )</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>.05</td>
<td>.0431</td>
<td>.0204</td>
<td>.0129</td>
<td>.0091</td>
<td>.0068</td>
</tr>
<tr>
<td>0</td>
<td>.0227</td>
<td>.0302</td>
<td>.0340</td>
<td>.0363</td>
<td></td>
</tr>
<tr>
<td>.10</td>
<td>.0744</td>
<td>.0331</td>
<td>.0193</td>
<td>.0124</td>
<td>.0083</td>
</tr>
<tr>
<td>0</td>
<td>.0413</td>
<td>.0551</td>
<td>.0662</td>
<td>.0661</td>
<td></td>
</tr>
<tr>
<td>.15</td>
<td>.0964</td>
<td>.0379</td>
<td>.0208</td>
<td>.0113</td>
<td>.0057</td>
</tr>
<tr>
<td>0</td>
<td>.0567</td>
<td>.0756</td>
<td>.0851</td>
<td>.0907</td>
<td></td>
</tr>
<tr>
<td>.20</td>
<td>.1111</td>
<td>.0417</td>
<td>.0185</td>
<td>.0069</td>
<td>.0000</td>
</tr>
<tr>
<td>0</td>
<td>.0694</td>
<td>.0926</td>
<td>.1042</td>
<td>.1111</td>
<td></td>
</tr>
<tr>
<td>.33</td>
<td>.1250</td>
<td>.0306</td>
<td>.0000</td>
<td>.0078</td>
<td>.0050</td>
</tr>
<tr>
<td>0</td>
<td>.0944</td>
<td>.1250</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

This RRIP technique is closely related to the multiple imputation technique introduced by Rubin (1977, 1978, 1979b). This note was stimulated by reading the large gains he shows for multiple imputation and by working on our own project. But those results coincide neatly with the drastic reduction of the variance increase from replication where the effects due to differences in weights \( k \) and \((k+1)\) are reduced by increasing \( k \) in the denominator of \( W(1 - W)/(k + W)^2 \). This, in turn, is just the special case of dichotomies for the general case of \((\Sigma W_i k_i^2)/(\Sigma W_i)\) for variance increase for a range (1 to \( K \)) of relative weights \( k_i \) (Kish 1965, 1976).

Both techniques depend on reducing the range of weights by averaging multiple duplications. Rubin's technique first duplicates for nonresponses and then replicates that operation. But RRIP first replicates the sample and then duplicates for nonresponse. RRIP seems to me a simpler way to obtain random selection without replacement for the duplicates, and we want this in practice, I assume. Rubin's theory has duplication with replacement only for simplicity; that is why his maximum increase of the variance from duplication is 0.250 (at \( W = 1/2 \)) instead of 0.125 (at \( W = 1/3 \)). That doubling of the variance increase is an extreme, but even for lower values of \( W \) sampling with replacement within the sample size \( n \) would bring
nontrivial increases in the variance, and would be opposed - correctly - in practice.

Several important relevant points must wait for later developments, but should be mentioned here. Most have been noted in passing in Kish (1965, Section 11.7, and 1976). These imputations treat total nonresponse. Item nonresponses need another treatment.

(1) In practical situations the number of groups used for imputations is much larger than two. But the principles used here for measuring gains can be extended readily to more groups.

(2) Duplicating introduces the ratio 1:2 into weights. Triplicating and replicating with higher integral weights 1:k (k = 3,4,...) would increase the variance even more. Hence higher ratios and sampling with replacement should be avoided.

(3) The increases of variance are greater for larger proportions of duplicates (up to W = 1/3). It follows readily that duplication should not be to 100 per cent, but only to the group with the largest response rate to (1 - W min ).

(4) Even the above limit should be lowered somewhat (say by 0.95 or so). Our formula for k = 0 becomes 

W(1 - W)/(0 + W)² = (1 - W)/W, the increase of variance due to elimination of the proportion (1 - W). For small proportions this increase is only a little greater than duplicating the proportion (1 - W).

Suppose for simplicity that there are two strata only, with proportions A and (1 - A) of the entire sample, and with response rates R(1 + W) and R respectively. We can duplicate the proportion W in the second stratum to bring its weight up to R(1 + W) hence increasing the variance by (1 - A)W(1 - W)/(1 + W)². Or we can eliminate the proportion W/(1 + W) from the first stratum to bring its weight down to R. The relative increase, with (1 - W') = W/(1 + W), is W'(1 - W')/(0 + W')² = W; hence this would increase the variance by AW. The two losses stand in the ratio F = [(1 - A)/A][(1 - W)/(1 + W)²]. Eliminating from A is better when A and W are small when F > 1, but duplicating from (1 - A) becomes preferable with larger W and A > 0.5 when

F < 1. However even better than either extreme can be the elimination of a smaller proportion C from A and duplicating for the rest (W - C) in (1 - A).

This technique is feasible also with more strata and we have applied it long ago for modest but nontrivial gains. However with RRIP the gains become larger and this refinement probably trivial.

(5) Matching cases are sometimes imputed instead of random selections from a larger group: the matching case found for each missing case is "nearest" in some (vector of) variable(s). The RRIP technique of multiple imputation may also be applied to matching to decrease the range from 1:2 to 1:k (and k = 2,3,4,...) with these procedures:

(a) Replicate the response deck k times.

(b) From the original deck make k "nearest" matchings for each nonresponse.

(c) Add one of the k matchings to each of the decks.

(6) It is likely that some improvement over simple random selection of k selections can be made with some kind of control, especially with stratified sampling.

Such controls and stratification for the separate imputations for several nonresponses in the same group could further constrain the variance increase from imputations.

(7) We must note that all forms of imputations and weighting not only increase the variances but also complicate their estimation. All inferential statistics, probability statements, such as tests of significance and confidence intervals, are affected. This weighty subject must be treated separately.

In all of the above it is assumed that a sample (random or epsem) of the responses in subclasses is imputed for the nonresponses. It is likely that nonresponses differ from responses even within subclasses. To reduce those biases should be a good field for applying models, subjective and Bayesian methods, and Rubin (1977a, 1978) presents the issues well and with strength. But there are deep problems both of
practical methods and of public policy, and we must avoid them in this brief note.


Bibliography


Heckman, J.J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a sample estimator for such models. Annals of Economic and Social Measurement, 5, 475-492.

Heeringa, S.G. (1980). Nonresponse in the ISDP 1978 Research Panels: An Examination of Rates and Patterns of Nonresponse in the April (ISDP303) and July (ISDP403) Interviews. Survey Research Center, University of Michigan.


Yates, F. (1933). The analysis of replicated experiments when the field results are incomplete. The Empire Journal of Experimental Agriculture, 1, 129-142.

The following Research Reports have been published by ISR. They are available in paperbound editions only. For information on prices and availability, write to the ISR Publishing Division, P.O. Box 1248, Ann Arbor, Michigan 48106.


**Perceptions of Life Quality in Rural America: An Analysis of Survey Data from Four Studies.** Robert W. Marans and Donald A. Dillman, with the assistance of Janet Keller. 1980. 118 pp.


