Statistics Canada Statistique Canada

SURVEY METHODOLOGY

A JOURNAL OF STATISTICS CANADA

VOLUME 12, NUMBER 1
JUNE 1986

Canada

SURVEY METHODOLOGY

A JOURNAL OF STATISTICS CANADA

JUNE 1986

Published under the authority of
the Minister of Supply and
Services Canada

©Minister of Supply
and Services Canada 1986

November 1986
8-3200-501

Price: Canada, $10.00, $20.00 a year
Other Countries, $11.50, $23.00 a year

Payment to be made in Canadian funds or equivalent

Catalogue 12-001, Vol. 12, No. 1

ISSN 0714-0045

Ottawa

SURVEY METHODOLOGY

A Journal of Statistics Canada

The Survey Methodology Journal is abstracted in The Survey Statistician, Statistical Theory and Methods Abstracts and is referenced in the Current Index to Statistics.

EDITORIAL POLICY

The Survey Methodology Journal will publish articles dealing with various aspects of statistical development relevant to a statistical agency, such as design issues in the context of practical constraints, use of different data sources and collection techniques, total survey error, survey evaluation, research in survey methodology, time series analysis, seasonal adjustment, demographic studies, data integration, estimation and data analysis methods, and general survey systems development. The emphasis will be on the development and evaluation of specific methodologies as applied to actual data collection or the data themselves. All papers will be refereed. However, the authors retain full responsibility for the contents of their papers and opinions expressed are not necessarily those of the Editorial Board or of Statistics Canada.

Submission of Manuscripts

The Survey Methodology Journal is published twice a year. Authors are invited to submit their manuscripts in either of the two official languages, English or French to the Editor, Dr. M.P. Singh, Social Survey Methods Division, Statistics Canada, 4th Floor, Jean Talon Building, Tunney's Pasture, Ottawa, Ontario, Canada K1A 0T6. Two nonreturnable copies of each manuscript prepared following the guidelines given in the Journal are requested.

Subscription Rates

The price of the Survey Methodology Journal (Catalogue No. 12-001) is $10.00 per copy, $20.00 per year in Canada, $11.50 per copy, $23.00 per year for other countries (payment to be made in Canadian funds or equivalent). Subscription order should be sent to: Publication Sales, Statistics Canada, Ottawa, Ontario, Canada K1A 0T6. (A reduced price is available to members of some statistical organizations. Please check with and subscribe through your organization.)

SURVEY METHODOLOGY

A Journal of Statistics Canada
Volume 12, Number 1, June 1986

Special Edition - Missing Data in Surveys

CONTENTS

Preface 0

G. KALTON and D. KASPRZYK
The Treatment of Missing Survey Data 1

R. PLATEK and G.B. GRAY
On the Definitions of Response Rates 17

V.P. GODAMBE and M.E. THOMPSON
Some Optimality Results in the Presence of Nonresponse 29

D.B. RUBIN
Basic Ideas of Multiple Imputation for Nonresponse 37

P. GILES and C. PATRICK
Imputation Options in a Generalized Edit and Imputation System 49

M.S. SRIVASTAVA and E.M. CARTER
The Maximum Likelihood Method for Nonresponse in Sample Surveys 61

M.A. HIDIROGLOU and J.-M. BERTHELOT
Statistical Editing and Imputation for Periodic Business Surveys 73

V. TREMBLAY
Practical Criteria for Definition of Weighting Classes 85

S. CHEUNG and C. SEKO
A Study of the Effects of Imputation Groups in the Nearest Neighbour Imputation Method for the National Farm Survey 99

PREFACE

This issue is devoted to papers presented at the Methodology Symposium on Missing Data in Surveys held at Statistics Canada in Ottawa, April 16-17, 1986. The symposium was jointly sponsored by Statistics Canada's Methodology Research Committee and the Laboratory for Research in Statistics and Probability at Carleton University. Concern about missing data in surveys (due to non-response or unusable responses) has been increasing in recent years. The symposium provided a forum for more than 200 professionals from universities, government organizations and the private sector in Canada and the United States to exchange information concerning recent theoretical and applied developments.

The symposium was opened by the Chief Statistician of Canada, Dr. Ivan Fellegi. He spoke about the international community's concern about the growing gap between theoretical and applied statistics and commended the organizers for bringing together specialists from both fields. While stating that the primary purpose of the conference was to make headway in the chosen topic, Dr. Fellegi also noted that the underlying theme was the extent to which statistical agencies should be involved in model-building.

The symposium included four sessions. The first session "General Issues and Organizational Experiences" was chaired by L. Kish of the University of Michigan and included presentations by G. Kalton (University of Michigan), G.B. Gray (Statistics Canada), D.W. Chapman (U.S. Bureau of the Census) and L.R. Curtin (U.S. National Center for Health Statistics). The chairman of the afternoon session of April 16, "Design and Estimation" was M. Hansen of Westat Inc. Papers were presented by P.S.R.S. Rao (University of Rochester), S. Michaud (Statistics Canada), C.E. Särndal (University of Montreal), G. Lazarus (Statistics Canada) and V.P. Godambe (University of Waterloo).

The morning session of April 17, "Item Non-Response and Imputation" was chaired by M. Moore of the University of Montreal. This session included contributions by D. Rubin (Harvard University), P. Giles (Statistics Canada), M.S. Srivastava (University of Toronto) and M.A. Hidiroglou (Statistics Canada). The chairman of the final session, "Case Studies", was J.N.K. Rao of Carleton University. Papers were presented by S. Hinkins (U.S. Internal Revenue Service), V. Tremblay (University of Montreal) and S. Cheung (Statistics Canada). The symposium was closed with a general discussion of developments concerning missing data in surveys led by J.N.K. Rao (chairman) and a panel including G. Kalton, L. Kish, D. Rubin, and I. Sande (Statistics Canada).

Nine of the symposium papers are included in this issue of the Journal. Additional symposium papers accepted for publication will appear in the next issue.

Survey Methodology, June 1986
Vol. 12, No. 1, pp. 1–16
Statistics Canada

The Treatment of Missing Survey Data

GRAHAM KALTON and DANIEL KASPRZYK[1]

ABSTRACT

Missing survey data occur because of total nonresponse and item nonresponse. The standard way to attempt to compensate for total nonresponse is by some form of weighting adjustment, whereas item nonresponses are handled by some form of imputation. This paper reviews methods of weighting adjustment and imputation and discusses their properties.

KEY WORDS: Nonresponse; Item nonresponse; Weighting adjustments; Imputation.

1. INTRODUCTION

Surveys typically collect responses to a large number of items for each sampled element. The problem of missing data occurs when some or all of the responses are not collected for a sampled element or when some responses are deleted because they fail to satisfy edit constraints. It is common practice to distinguish between total (or unit) nonresponse, when none of the survey responses are available for a sampled element, and item nonresponse, when some but not all of the responses are available. Total nonresponse arises because of refusals, inability to participate, not-at-homes, and untraced elements. Item nonresponse arises because of item refusals, "don't knows", omissions and answers deleted in editing.

This paper reviews the general-purpose methods available for handling missing survey data. The distinction between total and item nonresponse is useful here since different adjustment methods are used for these two cases. In general the only information available about total nonrespondents is that on the sampling frame from which the sample was selected (e.g., the strata and PSUs in which they are located). The important aspects of this information can usually be readily incorporated into weighting adjustments that attempt to compensate for the missing data. Hence as a rule weighting adjustments are used for total nonresponse. Methods for making weighting adjustments are reviewed in Section 2.

In the case of item nonresponse, however, a great deal of additional information is available for the elements involved: not only the information from the sampling frame, but also their responses for other survey items. In order to retain all survey responses for elements with some item nonresponses, the usual adjustment procedure produces analysis records that incorporate the actual responses to items for which the answers were acceptable and imputed responses for other items. Imputation methods for assigning answers for missing responses are reviewed in Section 3.

In general the choice between weighting adjustments and imputation for handling missing survey data is fairly clearcut; there are cases, however, when the choice is not so clear. These are cases of what may be termed partial nonresponse, when some data are collected for a sampled element but a substantial amount of data is missing. Partial nonresponse can arise, for instance, when a respondent terminates an interview prematurely, when data are not obtained for one or more members of an otherwise cooperating household (for household level analysis), or when a sampled individual provides data for some but not all waves of a panel survey. Discussions of the choice between weighting and imputation to compensate for wave nonresponse in a panel survey are given by Cox and Cohen (1985) and Kalton (1986).

[1] Graham Kalton, Survey Research Center, University of Michigan, Ann Arbor, Michigan, 48106-1248 and Daniel Kasprzyk, Population Division, U.S. Bureau of the Census, Washington, D.C., 20233. The authors would like to thank the referees for their helpful comments.

Although weighting adjustments and imputation are treated as separate approaches in the discussion below, they are in fact closely related. The relationship and differences between the two approaches are briefly discussed in Section 4, which also mentions some alternative ways of handling missing survey data.

2. WEIGHTING ADJUSTMENTS

Weighting adjustments are primarily used to compensate for total nonresponse. The essence of all weighting adjustment procedures is to increase the weights of specified respondents so that they represent the nonrespondents. The procedures require auxiliary information on either the nonrespondents or the total population. The following four types of weighting adjustments are briefly reviewed below: population weighting adjustments, sample weighting adjustments, raking ratio adjustments, and weights based on response probabilities. More details are provided in Kalton (1983).

2.1 Population Weighting Adjustments

The auxiliary information used in making population weighting adjustments is the distribution of the population over one or more variables, such as the population distribution by age, sex and race available from standard population estimates. The sample of respondents is divided into a set of classes, termed here weighting classes, defined by the available auxiliary information (e.g., White males aged 15-24, non-White females aged 25-34, etc.). The weights of all respondents within a weighting class are then adjusted by the same multiplying factor, with different factors in different classes. The adjustment is carried out in such a way that the weighted respondent distribution across the weighting classes conforms to the population distribution.

This type of adjustment is often termed poststratification. That term is avoided here, however, because although population weighting resembles poststratification, there is an important difference between the two. Like population weighting, poststratification weights the sample to make the sample distribution conform to the population distribution across a set of classes (or strata). However, the standard textbook theory of poststratification is concerned only with the sampling fluctuations that cause the sample distribution to deviate from the population distribution, not with the more major deviations that can arise from varying response rates across the classes. Poststratification adjustments are more like a fine tuning of the sample, resulting generally in only small variations in the weights across strata. In consequence, provided that the strata are not small, poststratification leads to lower standard errors for the survey estimates. In contrast, population weighting adjustments may involve more major adjustments and result in higher standard errors.

Population weighting adjustments attempt to reduce the bias created by nonresponse and coverage errors. Consider the estimation of a population mean $\bar{Y}$ from a sample in which the elements are selected with equal probability. Suppose that the population is divided into a set of weighting classes, with a proportion W_h of elements in class h. Assume that respondents always respond and that nonrespondents never do. Let R_h and M_h be the proportions of respondents and nonrespondents respectively in class h, and let $\bar{R} = \Sigma W_h R_h$ be the overall response rate. Then, following Thomsen (1973), the bias of the unadjusted respondent mean $(\bar{y})$ can be expressed as

$$B(\bar{y}) = \bar{R}^{-1}\sum W_h(\bar{Y}_{rh} - \bar{Y}_r)(R_h - \bar{R}) + \sum W_h M_h(\bar{Y}_{rh} - \bar{Y}_{mh}) = A + B \qquad (1)$$

where $\bar{Y}_{rh}$ and $\bar{Y}_{mh}$ are the means for respondents and nonrespondents in class h respectively, and $\bar{Y}_r$ is the population mean for the respondents. The use of the population weighting adjustment leads to the weighted sample mean, $\bar{y}_p = \Sigma W_h \bar{y}_{rh}$, where $\bar{y}_{rh}$ is the respondent sample mean in class h. The bias of $\bar{y}_p$ is simply the second term in $B(\bar{y})$, that is, $B(\bar{y}_p) = B$.

If A and B are of the same sign, the population weighting adjustment reduces the absolute bias in the estimate of $\bar{Y}$ by $|A|$. If $\bar{Y}_{rh} = \bar{Y}_{mh}$, as occurs in expectation when the nonrespondents are missing at random within the weighting classes, then $B = 0$. In this case, the population weighting adjustment eliminates the bias. The term A is a covariance-type term between the class response rates and the class respondent means. It is zero if either the response rates or the respondent means do not vary between classes. In either of these cases, the population weighting adjustment has no effect on the bias of the estimator. It may be noted that population weighting adjustments may increase the absolute bias of the estimate of $\bar{Y}$. This will occur when A and B are of opposite signs and $|A| < 2|B|$.

Population weighting adjustments require external data on the population distributions for the variables to be used. Care is needed to ensure that the data on which the population distributions are based are exactly comparable with the survey data; otherwise, inappropriate weights will result. Since the procedure weights up to population distributions, it does more than just attempt to compensate for nonresponse. It also compensates for coverage errors and makes a poststratification adjustment.

2.2 Sample Weighting Adjustments

As with population weighting adjustments, with sample weighting adjustments the sample is divided into weighting classes; varying weights are then assigned to these classes in an attempt to reduce the nonresponse bias. The essential difference between the two procedures lies in the auxiliary information used. As described above, population weighting adjustments are based on externally obtained population distributions. No data are needed for the sample nonrespondents. In contrast, sample weighting adjustments employ only data internal to the sample and require information about the nonrespondents.

With sample weighting adjustments, the nonresponse adjustment weights for the weighting classes are made proportional to the inverses of the response rates in the classes. In order to compute these response rates, the numbers of respondents and nonrespondents in the classes must be determined. It is therefore necessary to know to which class each respondent and nonrespondent belongs. Since typically very little information about the nonrespondents is available, the choice of weighting class is usually severely restricted. It is often limited to general sample design variables (e.g., PSUs and strata), characteristics of those variables (e.g., urban/rural, geographical region), and sometimes some additional variables available on the sampling frame. On occasion it may also be possible to collect information on one or two variables for the nonrespondents, for instance by interviewer observation.

As population weighting adjustments resemble poststratification, so sample weighting adjustments resemble two-phase sampling. The first phase sample is the total sample of respondents and nonrespondents; the second phase sample is the subsample of respondents, selected with different sampling fractions (response rates) in different strata (weighting classes). The sample weighted mean can be represented by $\bar{y}_s = \Sigma w_h \bar{y}_{rh}$, where w_h is the proportion of the total sample in weighting class h. Assuming no coverage errors, $E(w_h) \doteq W_h$, the population proportion in class h, as used in the population weighted estimator

$\bar{y}_p = \Sigma W_h \bar{y}_{rh}$. The bias of $\bar{y}_s$ is the same as that of $\bar{y}_p$, namely $B(\bar{y}_s) = B$ as given in equation (1); hence the effect of the sample weighting adjustment on the bias of the survey estimate is the same as that of the population weighting adjustment. Since sample weighting adjustments use only data for the sample, they do not compensate for coverage errors (unlike population weighting adjustments).

Population and sample weighting adjustments have different data requirements, and hence address different potential sources of bias. In practice the two forms of adjustment are used in combination. Generally sample weighting adjustments are applied first, and then population weighting adjustments are applied afterwards. A common approach is initially to determine the sample weights needed to compensate for unequal selection probabilities, next to revise these weights to compensate for unequal response rates in different sample weighting classes (e.g., urban/rural classes within geographical regions), and finally to revise the weights again to make the weighted sample distribution for certain characteristics (e.g., age/sex) conform to the known population distribution for those characteristics. The use of this approach in the U.S. Current Population Survey is described by Bailar *et al.* (1978).

As with population weighting adjustments, the aim of sample weighting adjustments is to reduce the bias that nonresponse may cause in survey estimates. An effect of sample weighting adjustments is, however, to increase the variances of the survey estimates. There is therefore a trade-off to be made between bias reduction and variance increase.

An indication of the amount of increase in variance from weighting can be obtained by considering the situation where the element variances within the weighting classes are all the same and the variances between the class means are negligible compared to the within-class variances. In this situation, the loss of precision from weighting is approximately the same as that arising from the use of disproportionate stratified sampling when proportionate stratified sampling is optimum; Kish (1965, Section 11.7C; 1976) discusses this latter case.

Under the above conditions, weighting increases the variance of a sample mean by approximately $L = (\Sigma W_h k_h)(\Sigma W_h / k_h)$, where W_h is the proportion of the population and k_h is the weight for class h. An alternative expression for L is $(\Sigma n_h)(\Sigma n_h k_h^2)/(\Sigma n_h k_h)^2$, where n_h is the sample size in class h. The factor L becomes large when the variance of the weights is large.

A large variance in the weights can arise from segmenting the sample into many weighting classes with only a few sampled elements in each. When the weighting classes are small, their response rates are unstable, and this gives rise to a large variation in the weights. To avoid this effect, it is common practice to limit the extent to which the sample is segmented. Even so, there may still be some weighting classes that require large weights. Sometimes these weighting classes are handled by collapsing them with adjacent ones and sometimes their weights are cut back to some acceptable maximum value (see Bailar *et al.* 1978 and Chapman *et al.* 1986, for examples). These procedures avoid the increase in variance associated with the use of extreme weights, but they may lead to increased bias; their effect on the bias is, however, unknown.

In some cases it seems desirable to use several auxiliary variables in forming the weighting classes for population or sample weighting adjustments. However, if the classes are formed by taking the full crossclassification of the variables, there will be a large number of weighting classes. Unless the sample is very large, the sample sizes in the resultant weighting classes will be small, and the instability in the response rates will lead to a large variance in the weights and loss of precision in the survey estimates. One way to deal with this problem is to cut down on the number of classes by collapsing cells, for instance by discarding some of the auxiliary variables or using coarser classifications. Another way is to base the weights on a model, as is done in raking ratio weighting discussed below.

2.3 Raking Ratio Adjustments

When weighting classes are taken to be the cells in the crossclassification of the auxiliary variables, population weighting adjustments make the joint distribution of the auxiliary variables in the sample conform to that in the population. Similarly, sample weighting adjustments make the joint distribution of the auxiliary variables in the respondent sample conform to that in the total sample. As noted above, however, this crossclassification approach may have the undesirable effect of creating many small, and hence unstable, weighting classes. Also, it is not always possible to employ this approach with population weighting adjustments: in many cases the population marginal distributions, and perhaps some bivariate distributions, of the auxiliary variables are available, but the full joint distribution is unknown.

An alternative approach is to develop weights that make the marginal distributions of the auxiliary variables in the sample conform to marginal population distributions (with population weighting) or marginal total sample distributions (with sample weighting), without ensuring that the full joint distribution conforms. The method of raking ratio estimation, or raking, may be used to obtain weights that satisfy these conditions. Raking corresponds to iterative proportional fitting in contingency table analysis (see, for instance, Bishop *et al.*, 1975).

Consider the use of raking in the simple case of two auxiliary variables. Let W_{hk} be the proportion of the population in the (h, k)-th cell of the crossclassification, and let $\tilde{w}_{hk}$ be the proportion assigned to that cell by the raking algorithm. Conditional on the total and respondent sample sizes in the cells (and assuming all cells have at least one respondent), the bias of the raking ratio adjusted sample mean $\bar{y}_q = \Sigma\Sigma\tilde{w}_{hk}\bar{y}_{hk}$ is

$$B(\bar{y}_q) = \sum\sum W_{hk}M_{hk}(\bar{Y}_{rhk} - \bar{Y}_{mhk}) + \sum\sum(\tilde{W}_{hk} - W_{hk})(\bar{Y}_{rhk} - \bar{Y}_{rh.} - \bar{Y}_{r.k} + \bar{Y}_r)$$

where $\tilde{W}_{hk} = E(\tilde{w}_{hk})$. The first term in this bias corresponds to the bias term B in equation (1) for the population and sample weighting adjustments. It is zero in expectation if the cell nonrespondents are random subsets of the cell populations. The second term is zero if either $\tilde{W}_{hk} = W_{hk}$ or there is no interaction in the $\bar{Y}_{rhk}$ for this classification.

Underlying the raking ratio weighting procedure is a logit model for the cell response rates. With the model $\ln[R_{hk}/(1 - R_{hk})] = \alpha_h + \beta_k$ for the response rates in a two-way classification, $\tilde{W}_{hk} = W_{hk}$. Thus, under this model, the second term in $B(\bar{y}_q)$ is zero.

Further discussion of raking ratio weighting is given by Oh and Scheuren (1978a,1978b, 1983). Oh and Scheuren (1978a) also provide a bibliography on raking.

2.4 Weighting with Response Probabilities

Although a number of methods for weighting with response probabilities have been proposed, this approach has not been widely adopted as an adjustment procedure. The basis of the approach is to assume that all population elements have probabilities (usually required to be non-zero) of responding to the survey. Some method is used to estimate the response probabilities for responding elements. These elements are then given nonresponse adjustment weights that are in inverse proportion to their estimated response probabilities.

An early application of this approach is the well-known procedure of Politz and Simmons (1949, 1950). A single (evening) call is made to each selected household, and during the course of the interview respondents are asked on how many of the previous five evenings they were at home at about the same time. Their response probabilities are then taken to be the fraction of the six evenings (including the one of the interview) that they were at home, and the inverses of these probabilities are used in the analysis. Note that the procedure does not deal with those who were out on all six evenings and those who refused.

Another approach for estimating response probabilities is to regress response status (1 for respondents, 0 for nonrespondents) on a set of variables available for both respondents and nonrespondents, using a logistic or probit regression. The predicted values from the regression for the respondents are then taken to be their response probabilities, and weights in inverse proportion to these predicted values are used in the analysis. A special case is when the predictor variables are dummy variables that identify a set of classes. The predicted response probabilities are then the class response rates, and the method reduces to a sample weighting adjustment. The method is most appropriate for situations where a good deal of information is available for the nonrespondents, as for instance when the nonrespondents are losses after the first wave of a panel survey. Little and David (1983) discuss the application of the method for panel nonresponse. It should be noted that if the regression is highly predictive of response status, the resultant weights will vary markedly, leading to a substantial loss in the precision of the survey estimates.

Drew and Fuller (1980, 1981) describe an approach for estimating response probabilities from the number of respondents secured at successive calls. In their model, the population is divided into classes. Within each class, every element is assumed to have the same response probability which remains the same at each call. The model also allows for a proportion of hard-core nonrespondents that is assumed constant across classes. Under these assumptions, the response probabilities for each class and the proportion of hard-core nonrespondents can be estimated, and hence weighting adjustments can be made. Thomsen and Siring (1983) adopt a similar approach using a more complex model.

Finally, mention should be made of a related approach that compensates for nonresponse by weighting up difficult-to-interview respondents. Bartholomew (1961), for instance, proposed making only two calls in a survey, and weighting up the respondents at the second call to represent the nonrespondents. The assumption behind this approach is that the nonrespondents are like the late respondents. This assumption seems questionable, however, and empirical evidence from an intensive follow-up study of nonrespondents in the U.S. Current Population Survey does not support it (Palmer and Jones 1966; Palmer 1967).

3. IMPUTATION

A wide variety of imputation methods has been developed for assigning values for missing item responses. The aim here is to provide a brief overview of the methods, the basic differences between them, and some of the issues involved in imputation. A fuller treatment is provided by Kalton and Kasprzyk (1982).

Imputation methods can range from simple *ad hoc* procedures used to ensure complete records in data entry to sophisticated hot-deck and regression techniques. The following are some common imputation procedures:

(a) *Deductive imputation*. Sometimes the missing answer to an item can be deduced with certainty from the pattern of responses to other items. Edit checks should check for consistency between responses to related items. When the edit checks constrain a missing response to only one possible value, deductive imputation can be employed. Deductive imputation is the ideal form of imputation.

(b) *Overall mean imputation*. This method assigns the overall respondent mean to all missing responses.

(c) *Class mean imputation*. The total sample is divided into classes according to values of the auxiliary variables being used for the imputation (comparable to weighting classes). Within each imputation class the respondent class mean is assigned to all missing responses.

(d) *Random overall imputation*. A respondent is chosen at random from the total respondent sample, and the selected respondent's value is assigned to the nonrespondent. This method is the simplest form of hot-deck imputation, that is an imputation procedure in which the value assigned for a missing response is taken from a respondent to the current survey.

(e) *Random imputation within classes*. In this hot-deck method, a respondent is chosen at random within an imputation class, and the selected respondent's value is assigned to the nonrespondent.

(f) *Sequential hot-deck imputation*. The term sequential hot-deck imputation is used here to describe the procedure used with the labor force items in the U.S. Current Population Survey (Brooks and Bailar 1978). The procedure starts with a set of imputation classes. A single value for the item subject to imputation is assigned for each class (perhaps taken from a previous survey). The records in the survey's data file are then considered in turn. If a record has a response for the item in question, its response replaces the value stored for the imputation class in which it falls. If the record has a missing response, it is assigned the value stored for its imputation class.

The hot-deck method is similar to random imputation within classes. If the order of the records in the data file were random, the two methods would be equivalent, apart from the start-up process. The non-random order of the list generally acts to the benefit of the hot-deck method since it gives a closer match of donors and recipients provided that the file order creates positive autocorrelation. The benefit is, however, unlikely to be substantial.

The sequential hot-deck suffers the disadvantage that it may easily make multiple uses of donors, a feature that leads to a loss of precision in survey estimates. Multiple use of a donor occurs when, within an imputation class, a record with a missing response is followed by one or more other records with missing responses. The number of imputation classes that can be used with the method also has to be limited in order to ensure that donors are available within each class.

Useful discussions of the sequential hot-deck method are provided by Bailar *et al.* (1978), Bailar and Bailar (1978, 1983), Ford (1983), Oh and Scheuren (1980), Oh *et al.* (1980), and Sande (1983).

(g) *Hierarchical hot-deck imputation*. The above disadvantages of the sequential hot-deck are avoided in the hierarchical hot-deck method, a form of hot-deck imputation developed for the items in the March Income Supplement of the Current Population Survey. The procedure sorts respondents and nonrespondents into a large number of imputation classes from a detailed categorization of a sizeable set of auxiliary variables. Nonrespondents are then matched with respondents on a hierarchical basis, in the sense that if a match cannot be made in the initial imputation class, classes are collapsed and the match is made at a lower level of detail. Coder (1978) and Welniak and Coder (1980) provide further details on the hierarchical hot-deck procedure.

(h) *Regression imputation*. This method uses respondent data to regress the variable for which imputations are required on a set of auxiliary variables. The regression equation is then used to predict the values for the missing responses. The imputed value may either be the predicted value, or the predicted value plus some residual. There are several ways in which the residual may be obtained, as discussed later.

(i) *Distance function matching*. This hot-deck method assigns a nonrespondent the value of the "nearest" respondent, where "nearest" is defined in terms of a distance function for the auxiliary variables. Various forms of distance function have been proposed (e.g., Sande 1979; Vacek and Ashikago 1980), and the function can be constructed to reduce the multiple use of donors by incorporating a penalty for each use (Colledge *et al.* 1978).

Although at first sight these may appear a diverse set of procedures, they can nearly all be fitted within a single unifying framework. The methods can all be described, at least approximately, as special cases of the general regression model

$$\hat{y}_{mi} = b_{ro} + \sum b_{rj}z_{mij} + \hat{e}_{mi} \tag{2}$$

where $\hat{y}_{mi}$ is the imputed value for the ith record with a missing y value, z_{mij} are values reflecting the auxiliary variables for that record, b_{ro} and b_{rj} are the regression coefficients for the regression of y on x for the respondents, and $\hat{e}_{mi}$ is a residual chosen according to a specified scheme for the particular imputation method.

Equation (2) represents the regression imputation method in an obvious way. If the $\hat{e}_{mi}$'s are set at zero, then the imputed value is the predicted value from the regression; otherwise a residual of some form may be added. The equation also represents class mean imputation by defining the z_j's to be dummy variables that represent the classes, and setting $\hat{e}_{mi} = 0$. The regression equation then reduces to $\hat{y}_{mi} = \bar{y}_{rh}$, the class mean. Random imputation within classes is obtained by adding a residual to the class mean, where the residual is the deviation from the class mean for one of the respondents. Then $\hat{y}_{mi} = \bar{y}_{rh} + e_{rhk}$, where e_{rhk} is the deviation for respondent k in class h; this reduces to $\hat{y}_{mi} = y_{rhk}$, the value for that respondent. The sequential and hierarchical hot-deck methods resemble the random within class method. The overall mean and random overall imputation methods are degenerate cases of the class mean and random within class methods that use no auxiliary information.

An important consideration in the choice of imputation method is the type of variable being imputed. All the above methods can be applied routinely with continuous variables, but some of them are not suitable for use with categorical or discrete variables (such as being a member of the labor force (1) or not (0), and the number of completed years of education). Overall mean, class mean, and regression imputations impute values like 0.7 for being a member of the labor force (i.e., a 70% chance) and 10.7 for the number of completed years of education. These values are not feasible for individual respondents, and rounding them to whole numbers leads to bias. For this reason, these imputation methods do not work well for categorical and discrete variables. A notable advantage of all hot-deck methods is that they always give feasible values since the values are taken from respondents.

There are two major distinguishing features of the above imputation methods that deserve elaboration: whether or not a residual is added and, if one is, the form of the residual; and whether the auxiliary information is used in dummy variable form to represent classes or whether it is used straightforwardly in the regression. These features are discussed in the next two subsections. Other issues arising with the use of imputation are then discussed in subsequent subsections.

3.1 Choice of Residuals

Imputation methods may be classified as deterministic or stochastic according to whether the $\hat{e}_{mi}$'s are set at zero or not. For each deterministic imputation method, there is a stochastic counterpart. Let $\hat{y}_{mid}$ be the value imputed by the deterministic method and $\hat{y}_{mis} = \hat{y}_{mid} + \hat{e}_{mi}$ be that imputed by the corresponding stochastic method. Then $E_2(\hat{y}_{mis}) = \hat{y}_{mid}$, where E_2 denotes expectation over the sampling of residuals given the initial sample, provided that $E_2(\hat{e}_{mi}) = 0$ (as generally applies).

The choice between a deterministic and the corresponding stochastic imputation method depends on the form of survey analysis to be conducted. Consider first the estimation of the population mean of the y-variable using the sample mean of the respondents' values and the nonrespondents' imputed values. As Kalton and Kasprzyk (1982) show, given that $E_2(\hat{y}_{mis}) = \hat{y}_{mid}$, it follows that the expectation of the sample mean is the same whether the deterministic method or the corresponding stochastic method is used. Thus both methods have the same effect on the bias of the estimate. However, the addition of random residuals in the stochastic method causes a loss of precision in the sample mean. Although this loss can be controlled by the choice of a suitable method of sampling residuals (Kalton and Kish 1984), nevertheless some loss in precision occurs. For this reason a deterministic scheme is preferable for the purpose of estimating the population mean.

Consider now the estimation of the element standard deviation and distribution of the y-variable. Deterministic imputation methods fare badly for these purposes, since they cause an attenuation in the standard deviation and they distort the shape of the distribution. This may be simply illustrated in terms of the class mean imputation method. By assigning the class mean to all the missing values in a class, the shape of the distribution is clearly distorted with a series of spikes at the class means. The standard deviation of the distribution is attenuated because the imputed values reflect only the between-class and not the within-class variance. The appeal of the stochastic imputation methods is that the residual term captures the within-class (or residual) variance, and hence avoids the attenuation of the element standard deviation and the distortion of the distribution.

Since some survey analyses are likely to involve the distributions of the variables, stochastic imputation methods like the hot-deck methods are generally preferred. Once a decision is made to use a stochastic method, the question of how to choose the residuals arises. If the standard regression assumptions are accepted, the residuals could be chosen from a normal distribution with a mean of zero and a variance equal to the residual variance from the respondent regression. However, this places complete reliance on the model. An alternative that avoids the normality assumption is to choose the residuals randomly from the empirical distribution of the respondents' residuals. Another alternative is to select a residual from a respondent who is a "close" match to the nonrespondent, measuring "close" in terms of similar values on the auxiliary variables. This attractive alternative avoids the assumption of homoscedasticity and guards against misspecification of the distribution of the residual term. In the limit, the closest respondent is one who has the same values of all the auxiliary variables as the nonrespondent. In this case, the nonrespondent is given one of the matched respondents' values. This case arises with hot-deck methods, where nonrespondents and respondents are matched in terms of the auxiliary variables, and nonrespondents are assigned values from matched respondents.

A further consideration in the choice of residuals is to make the imputed values feasible ones. As noted above, deterministic methods may impute values for categorical and discrete variables that are not feasible. Some stochastic methods solve this problem through the allocation of the residuals. In particular, the use of respondents' residuals with the random within class and the sequential and hierarchical hot-deck methods ensures that the imputed values are feasible ones.

3.2 Imputation Class or Regression Imputation

As noted earlier, both imputation class and regression imputation methods fall within the imputation model given by equation (2). The difference between them lies in the ways in which they employ the auxiliary variables.

Imputation class methods divide the sample into a set of classes. For this purpose, continuous auxiliary variables have to be categorized. There is complete flexibility in the way the classes are formed, and the symmetrical use of the auxiliary variables in different parts of the sample is not required. Thus, for instance, in imputing for hourly rate of pay in a sample of employees, the sample might first be divided into two parts, union members and nonmembers; then the imputation classes for the members might be formed in terms of age and occupation whereas those for nonmembers might be formed in terms of sex and industry. As a rule, the aim is to construct classes of adequate size that explain as much of the variance in the variable to be imputed as possible. When the classes are formed by a complete crossclassification of the auxiliary variables, the underlying model contains all main effects and all interactions for the crossclassification. The limitation of imputation class methods is that the number of classes formed has to be constructed to ensure that there is some minimum number of respondents in each class. The hierarchical hot-deck method attempts to extend the amount of auxiliary data used, but even with this method matches of respondents and nonrespondents often cannot be made at the finer levels of detail. Coupled with the use of a random respondent residual within a class, imputation class methods have the valuable property that imputed values are feasible ones: that is, the imputed values are actual respondents' values.

Regression imputation methods have an advantage over imputation class methods in the number and in the level of detail of the auxiliary variables they can employ. Age can, for instance, be taken as a continuous variable rather than being categorized into a few classes. The regression model allows more main effects to be included in the model, but at the price of fewer interactions. Regression models can, of course, include some interactions, but they need to be specified. The models can also include polynomial terms and employ transformations, but again they need to be specified. The regression model has the potential of providing better predictions for the imputed values, but to achieve this careful modelling is required. Careful imputation modelling is unrealistic for all the variables in a survey, but it may be feasible for one or two major ones (and especially so for continuous surveys). Without careful modelling, there is a serious risk of poor imputations, although as noted earlier, this risk can be reduced by the allocation of random residuals from "close" respondents.

If a regression imputation assigns the residual from a respondent with exactly the same values of the auxiliary variables, the imputed value is necessarily a feasible one. If, however, there is even a small difference between the respondent's and nonrespondent's values on the auxiliary variables, the imputed value may not be feasible. A variant of regression imputation that avoids this problem, termed predictive mean matching, is described by Little (1986b) (Little attributes the method to Rubin). With predictive mean matching, the nonrespondent is matched to the respondent with the closest predicted value. Then, instead of adding the respondent's residual to the nonrespondent's predicted value, the nonrespondent is assigned the respondent's value. The method is thus a hot-deck method, and is similar to distance function matching.

The choice between imputation class and regression imputation methods should in part depend on the efforts made to develop the regression model. Unless adequate resources are devoted to the development of a regression model, the imputation class methods may be safer. The choice should also in part depend on the sample size. With large samples, hot-deck methods are likely to be able to use enough classes to take advantage of all the major predictor variables; however, with small samples this may not hold, and regression methods may have greater potential. David *et al.* (1986) describe an interesting study that compares regression models for imputing wages and salary in the U.S. Current Population Survey with hierarchical hot-deck imputations. Despite the extensive efforts made to develop the regression models, the hot-deck imputations were not found to be inferior in this large sample.

3.3 Effect of Imputation on Relationships

Although most of the literature on imputation deals with its effect on univariate statistics such as means and distributions, a large part of survey analysis is concerned with bivariate and multivariate relationships. Here the analysis of relationships can be considered in broad terms to include crosstabulation, correlation or regression analysis, comparisons of subclass means or proportions, and any other analysis involving two or more variables. As will be illustrated below, imputation can have harmful effects on all analyses of relationships, often attenuating the associations between variables. Discussions of the effects of imputations on relationships are provided by Santos (1981), Kalton and Kaspryzk (1982) and Little (1986a).

The general nature of the effect of imputation on relationships can be seen by considering its effect on the estimate of the sample covariance in the simple situation where the y-variable has missing responses that are missing at random over the population and the x-variable has no missing data. The sample covariance, s_{xy}, is calculated in the standard way, based on the actual values for respondents and the imputed values for nonrespondents, as an estimate of the population covariance S_{xy}. Using the fact that $E_2(\hat{y}_{mis}) = \hat{y}_{mid}$ as above, it can be readily shown that the expected value of s_{xy} under a deterministic imputation method is the same as that under the corresponding stochastic method.

As Santos (1981) shows, the relative bias of s_{xy} when the mean overall or random overall imputation methods are used is approximately $-\bar{M}$, where $\bar{M}$ is the nonresponse rate. This occurs because the imputed y-values are unrelated to their x-values, and hence the cases with imputed values attenuate the covariance towards zero. This attenuation is decreased in magnitude by imputation methods that use auxiliary variables. With class mean imputation or random imputation within classes, the relative bias is approximately $-\bar{M}(S_{xy.z}/S_{xy})$, where $S_{xy.z} = \Sigma W_h S_{xyh}$ is the average within-class covariance for classes formed by the auxiliary variables z, S_{xyh} is the covariance within class h, and W_h is the proportion of the population in class h. With predicted regression imputation or regression imputation with a random residual, both with a single auxiliary variable z, the relative bias is approximately $-\bar{M}[1 - (\rho_{xz}\rho_{yz}/\rho_{xy})]$, where ρ_{uv} is the correlation between u and v.

The disturbing feature of these results is that, unless $\bar{M}$ is small, s_{xy} calculated with imputed values under any of these imputation methods may be subject to substantial bias even under the missing at random model. The estimates s_{xy} computed with imputed values obtained under the imputation class and regression methods are unbiased only if the partial covariance $S_{xy.z}$ is zero. In general, there is no reason to assume uncritically that $S_{xy.z}$ is zero. However, there is an important case when $S_{xy.z} = 0$. This occurs when $x = z$, that is when x is used as an auxiliary variable in the imputation procedure. In this case, the sample covariance is unbiased under the missing at random model. This result suggests that if the relationship between x and y is to form an important part of the survey analysis, x should be used as an auxiliary variable in imputing for missing y-values.

The above theory assumes that only the y-variable was subject to missing data. In practice the x-variable will often also be incomplete. If so, the sample covariance may be attenuated because of the imputations for both variables. A special feature occurs when x and y are both missing for a record. If the two values are imputed separately, the covariance is attenuated, but if they are imputed jointly, using the same respondent as the donor of both values, the covariance structure is retained. This suggests that when a record has several missing related values, they should be taken from the same donor. Coder (1978) describes the use of joint imputation from the same donor in the March Income Supplement of the Current Population Survey.

As an illustration of how the above arguments about the attenuation of covariances apply to other forms of relationships, we will give a simple numerical example of the effect of imputation on the difference between two proportions. Let the variable of interest be whether an individual has a particular attribute or not, and suppose that one half of the respondents fail to answer this question. The missing responses are imputed by a random within class imputation method using two classes, A and B. The objective is now to compare the

Table 1

Number of Respondents with the Attribute, and Number of Sampled Persons by Class, Sex and Response Status

	Class A			Class B		
	M	F	Total	M	F	Total
Respondents with the attribute	80	40	120	60	20	80
Total respondents	100	100	200	100	100	200
Nonrespondents	100	100	200	100	100	200
Total sample	200	200	400	200	200	400

percentages of men and women with the attribute. The data are displayed in Table 1. Since 60% of the total respondents in class *A* have the attribute, 60 of the 100 male and 60 of the 100 female nonrespondents in that class will be imputed to have the attribute. Similarly, in class *B* 40% of the total respondents have the attribute, and so 40 male and 40 female nonrespondents will be imputed to have the attribute. The proportion of actual and imputed males with the attribute is thus $(80 + 60 + 60 + 40)/400 = 0.6$ or 60%. For females the corresponding proportion is $(40 + 60 + 20 + 40)/400 = 0.4$, or 40%. The difference between these two percentages is 20%.

Had sex also been taken into account in forming the imputation classes, the percentages of males and females with the attribute would have been 70% and 30%, differing by 40%. The failure to include sex as an auxiliary variable in the imputation has thus caused a substantial attenuation in the measurement of the relationship between sex and having the attribute.

3.4 Multiple Imputations

Ideally the analyst using a data set with imputed values should be able to obtain valid results for any analyses by applying standard techniques for complete data. However, as noted in the last section, imputation can distort measures of the relationships between variables. It also distorts standard error estimation.

All imputation methods except deductive imputation fabricate data to some extent. The extent of fabrication depends on how well the imputation model predicts the missing values. If the imputation model explains only a small proportion of the variance in the variable among the respondents, the amount of fabrication in each imputed value is likely to be substantial. If the imputation model explains a high proportion of the respondent variance, the amount of fabrication is likely to be less serious. However, it needs to be recognized that the fit of the imputation model for the respondents is not necessarily a good measure of the fit for the nonrespondents.

Standard errors computed in the standard way from a data set with imputed values will generally be underestimates because of the fabrication involved in the imputed values. Rubin (1978, 1979) has advocated the method of multiple imputations to provide valid inferences from data sets with imputed values (see also Herzog and Rubin 1983; Rubin and Schenker 1986). When multiple imputations are used for the purpose of standard error estimation, the construction of the complete data set by imputing for the missing responses is carried out several (say m) times using the same imputation procedure. The sample estimates z_i $(i = 1, 2, \ldots, m)$ of the population parameter of interest Z are computed from each of the replicate data sets, and their average $\bar{z}$ is calculated. A variance estimator for $\bar{z}$ is then given by $\hat{V} = \hat{W} + [(m + 1)/m]\hat{B}$, where $\hat{W}$ is the average of the within-replicate variance of $\bar{z}$ and $\hat{B} = \Sigma(z_i - \bar{z})^2/(m - 1)$ is the between-replicate variance. Even with the inclusion of the between-replicate variance component, however, the coverages of confidence intervals for Z based on $\hat{V}$ are still overstated, with the amount of overstatement increasing with the level of nonresponse.

This overstatement of the confidence levels can be addressed by modifying the imputation procedure, as described by Rubin and Schenker (1986). Their treatment considers the random overall imputation method, and one of their modifications allows for uncertainty about the population mean and variance in the following way. With the standard random overall imputation method, the conditional expected mean and variance of the imputed values are the sample respondents' mean and variance. With the modification, the expected mean and variance of the imputed values for a replicate are drawn at random from appropriate distributions. The imputed values are then a random selection of respondents' values, modified for the randomly-chosen mean and variance. When estimating the population mean, the effect of the changing expected mean and variance between replicates is to increase the between-replicate variance component in $\hat{V}$. This increase gives improved coverage for the resultant confidence intervals.

A major problem with the use of multiple imputations is the additional computer analysis needed, which increases as the number of replicates, m, increases. For this reason, a small value of m, such as $m = 2$, may be preferred. A small value of m may, however, result in a low level of precision for the variance estimator. Even with small m, it is questionable whether the multiple imputation approach is feasible for routine analyses. It may be best reserved for special studies, such as that described by Herzog and Rubin (1983).

In addition to providing appropriate standard errors, another advantage of multiple imputations from the same imputation procedure is that it reduces the loss of precision in survey estimates arising from the random selection of respondents to act as donors of imputed values (see Section 3.1). This loss is reduced with multiple imputations by averaging over the replicates. A small number of replicates serves well for this purpose. As noted earlier, Kalton and Kish (1984) describe alternative ways of selecting the sample of respondents to achieve this end.

A second major potential application of multiple imputations is to generate the imputations for the several replicates by different imputation procedures, making different assumptions about the nonrespondents. Suppose, for instance, that hourly rates of pay are to be imputed for some earners in the sample. One procedure that might be used is the random within class imputation method, which is based on an assumption that nonrespondents are missing at random within the classes. If it is thought that the nonrespondents might in fact come more heavily from those with higher rates of pay in each class, a simple modification to the random within class method might be to impute values that are, say, 50 cents above the donors' values. Other imputation procedures - for instance, using different imputation classes - could also be tried. Comparison of the survey estimates obtained from the data sets in which the different imputation procedures are applied then provides a valuable indication of the sensitivity of the estimates to the values imputed. If the estimates turn out to be very similar, they can be accepted with greater confidence; if they differ markedly, the estimates need to be treated with considerable caution.

4. CONCLUDING REMARKS

Weighting and imputation have been presented as two distinct methods for handling missing survey data, but in fact there is a close relationship between them. This may be illustrated by considering any imputation method that assigns respondents' values to the nonrespondents. For univariate analyses, this process is equivalent to dropping the nonrespondents' records and adding the nonrespondents' weights to those of the donor respondents (Kalton 1986).

The differences between weighting and imputation emerge when one considers the multivariate nature of survey data. It is possible to impute for the responses of a total nonrespondent by taking all the responses from a single donor; however, weighting is generally simpler in this case and it avoids the loss of precision arising from the sampling of respondents to serve as donors. It is not practicable to use weighting to handle item nonresponse since it would result in different sets of weights for each item; this would cause serious difficulties for crosstabulations and other analyses of the relationships between variables.

Weighting is a single global adjustment that attempts to compensate for the missing responses to all the items simultaneously. Imputation, on the other hand, is item-specific. This difference has consequences for the way that the auxiliary data are used. In forming weighting classes, the focus is on determining classes that differ in their response rates. The choice of auxiliary variables to use in imputation, however, is primarily made in terms of their abilities to predict the missing responses.

An assumption underlying all the procedures reviewed in this paper is that once the auxiliary variables have been taken into account the missing values are missing at random. Thus, for instance, the nonrespondents are assumed to be like the respondents within weighting and imputation classes. This assumption can be avoided by using stochastic censoring models, as has been done by Greenlees *et al.* (1982) in imputing wages and salaries in the Current Population Survey. However, as Little (1986b) observes, these models are highly sensitive to the distributional assumptions made.

An alternative approach for handling missing survey data is to leave the values missing in the data set and let the analyst incorporate appropriate missing data models into the analysis (Little 1982). This approach has much to commend it, but the labor and computing time needed to implement it effectively preclude its use as a general purpose strategy. Rather, the approach seems best suited for a small range of special analyses. In order to permit the analyst to adopt this approach, it is essential that all imputed values be flagged to indicate they are not actual responses, so that they can then be dropped from the analysis.

Finally, we should note that all methods of handling missing survey data must depend upon untestable assumptions. If the assumptions are seriously in error, the analyses may give misleading conclusions. The only secure safeguard against serious nonresponse bias in survey estimates is to keep the amount of missing data small.

REFERENCES

BAILAR III, J.C., and BAILAR, B.A. (1978). Comparison of two procedures for imputing missing survey values. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 462-467.

BAILAR, B.A., and BAILAR III, J.C. (1983). Comparison of the biases of the hot-deck imputation procedure with an "equal-weights" imputation procedure. In *Incomplete Data in Sample Surveys, Volume 3, Proceedings of the Symposium*, (Eds. W.G. Madow and I. Olkin), New York: Academic Press, 299-311.

BAILAR, B.A., BAILEY, L., and CORBY, C.A. (1978). A comparison of some adjustment and weighting procedures for survey data. In *Survey Sampling and Measurement*, (Ed. N.K. Namboodiri), New York: Academic Press, 175-198.

BARTHOLOMEW, D.J. (1961). A method of allowing for 'not at home' bias in sample surveys. *Applied Statistics*, 10, 52-59.

BISHOP, Y.M.M., FIENBERG, S.E., and HOLLAND, P.W. (1975). *Discrete Multivariate Analyses*. Cambridge, Mass: The MIT Press.

BROOKS, C.A., and BAILAR, B.A. (1978). *An Error Profile: Employment as Measured by the Current Population Survey*. Statistical Policy Working Paper 3. U.S. Department of Commerce. Washington, D.C.: U.S. Government Printing Office.

CHAPMAN, D.W., BAILEY, L., and KASPRZYK, D. (1986). Nonresponse adjustment procedures at the U.S. Census Bureau. *Survey Methodology*, forthcoming.

CODER, J. (1978). Income data collection and processing from the March Income Supplement to the Current Population Survey. *The Survey of Income and Program Participation Proceedings of the Workshop on Data Processing*, February 23-24, 1978, (Ed. D. Kasprzyk), Chapter II. Washington, D.C.: U.S. Department of Health, Education and Welfare.

COLLEDGE, M.J., JOHNSON, J.H., PARE, R., and SANDE, I.G. (1978). Large scale imputation of survey data. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 431-436.

COX, B.G., and COHEN, S.B. (1985). *Methodological Issues for Health Care Surveys*. New York: Marcel Dekker.

DAVID, M., LITTLE, R.J.A., SAMUHEL, M.E., and TRIEST, R.K. (1986). Alternative methods for CPS income imputation. *Journal of the American Statistical Association*, 81, 29-41.

DREW, J.H., and FULLER, W.A. (1980). Modelling nonresponse in surveys with callbacks. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 639-642.

DREW, J.H., and FULLER, W.A. (1981). Nonresponse in complex multiphase surveys. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 623-628.

FORD, B.L. (1983). An overview of hot-deck procedures. In *Incomplete data in Sample Surveys, Volume 2, Theory and Bibliographies*, (Eds. W.G. Madow, I. Olkin and D.B. Rubin), New York: Academic Press, 185-207.

GREENLEES, W.S., REECE, J.S., and ZIESCHANG, K.D. (1982). Imputation of missing values when the probability of response depends on the variable being imputed. *Journal of the American Statistical Association*, 77, 251-261.

HERZOG, T.N., and RUBIN, D.B. (1983). Using multiple imputation to handle nonresponse in sample surveys. In *Incomplete data in Sample Surveys, Volume 2, Theory and Bibliographies*, (Eds. W.G. Madow, I. Olkin and D.B. Rubin), New York: Academic Press, 209-245.

KALTON, G. (1983). *Compensating for Missing Survey Data*. Ann Arbor: Survey Research Center, University of Michigan.

KALTON, G. (1986). Handling wave nonresponse in panel surveys. *Journal of Official Statistics*, 2, forthcoming.

KALTON, G., and KASPRZYK, D. (1982). Imputing for missing survey responses. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 22-31.

KALTON, G., and KISH, L. (1984). Some efficient random imputation methods. *Communications in Statistics - Theory and Methods*, 13(16), 1919-1939.

KISH, L. (1965). *Survey Sampling*. New York: Wiley.

KISH, L. (1976). Optima and proxima in linear sample designs. *Journal of the Royal Statistical Society*, Ser. A, 139, 80-95.

LITTLE, R.J.A. (1982). Models for nonresponse in sample surveys. *Journal of the American Statistical Association*, 77, 237-250.

LITTLE, R.J.A. (1986a). Survey nonresponse adjustments for estimates of means. *International Statistical Review*, 54, 139-157.

LITTLE, R.J.A. (1986b). Missing data in Census Bureau surveys. *Proceedings of the Second Annual Census Bureau Research Conference*, 442-454.

LITTLE, R.J.A., and DAVID, M.H. (1983). Weighting adjustments for non-response in panel surveys. Working Paper, Washington, D.C.: U.S. Bureau of the Census.

OH, H.L., and SCHEUREN, F. (1978a). Multivariate raking ratio estimation in the 1973 Exact Match Study. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 716-722.

OH, H.L., and SCHEUREN, F. (1978b). Some unresolved application issues in raking ratio estimation. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 723-728.

OH, H.L., and SCHEUREN, F. (1980). Estimating the variance impact of missing CPS income data. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 408-415.

OH, H.L., and SCHEUREN, F. (1983). Weighting adjustment for unit nonresponse. In *Incomplete data in Sample Surveys, Volume 2, Theory and Bibliographies*, (Eds. W.G. Madow, I. Olkin and D.B. Rubin), New York: Academic Press, 143-184.

OH, H.L., SCHEUREN, F., and NISSELSON, H. (1980). Differential bias impacts of alternative Census Bureau hot deck procedures for imputing missing CPS income data. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 416-420.

PALMER, S. (1967). On the character and influence of nonresponse in the Current Population Survey. *Proceedings of the Social Statistics Section, American Statistical Association*, 73-80.

PALMER, S., and JONES, C. (1966). A look at alternate imputation procedures for CPS noninterviews. Washington, D.C.: U.S. Bureau of the Census memorandum.

POLITZ, A., and SIMMONS, W. (1949). I. An attempt to get the 'not at homes' into the sample without callbacks. II. Further theoretical considerations regarding the plan for eliminating callbacks. *Journal of the American Statistical Association*, 44, 9-31.

POLITZ, A., and SIMMONS, W. (1950). Note on an attempt to get the 'not at homes' into the sample without callbacks. *Journal of the American Statistical Association*, 45, 136-137.

RUBIN, D.B. (1978). Multiple imputations in sample surveys: a phenomenological Bayesian approach to nonresponse. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 20-34.

RUBIN, D.B. (1979). Illustrating the use of multiple imputations to handle nonresponse in sample surveys. *Bulletin of the International Statistical Institute*, 48(2), 517-532.

RUBIN, D.B., and SCHENKER, N. (1986). Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *Journal of the American Statistical Association*, 81, 366-374.

SANDE, G. (1979). Numerical edit and imputation. Paper presented to the International Association for Statistical Computing, 42nd Session of the International Statistical Institute.

SANDE, I.G. (1983). Hot-deck imputation procedures. In *Incomplete Data in Sample Surveys, Volume 3, Proceedings of the Symposium*, (Eds. W.G. Madow and I. Olkin), New York: Academic Press, 339-349.

SANTOS, R.L. (1981). Effects of imputation on regression coefficients. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 140-145.

THOMSEN, I. (1973). A note on the efficiency of weighting subclass means to reduce the effects of nonresponse when analyzing survey data. *Statistisk Tidskrift*, 4, 278-283.

THOMSEN, I., and SIRING, E. (1983). On the causes and effects of nonresponse: Norwegian experiences. In *Incomplete Data in Sample Surveys, Volume 3, Proceedings of the Symposium*, (Eds. W.G. Madow and I. Olkin), New York: Academic Press, 25-29.

VACEK, P.M., and ASHIKAGA, T. (1980). An examination of the nearest neighbor rule for imputing missing values. *Proceedings of the Statistical Computing Section, American Statistical Association*, 326-331.

WELNIAK, E.J., and CODER, J.F. (1980). A measure of the bias in the March CPS earnings imputation system. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 421-425.

Survey Methodology, June 1986
Vol. 12, No. 1, pp. 17–27
Statistics Canada

On the Definitions of Response Rates

R. PLATEK and G.B. GRAY[1]

ABSTRACT

In this paper, different types of response/nonresponse and associated measures such as rates are provided and discussed together with their implications on both estimation and administrative procedures. The missing data problems lead to inconsistent terminology related to nonresponse such as completion rates, eligibility rates, contact rates, and refusal rates, many of which can be defined in different ways. In addition, there are item nonresponse rates as well as characteristic response rates. Depending on the uses, the rates may be weighted or unweighted.

KEY WORDS: Eligibility; Completion; Contact; Refusal; Response Rates.

1. INTRODUCTION

The census or sample survey data are gathered by any one of such procedures as personal interview, telephone, or mail. It sometimes happens that some units may not respond for such reasons as "not at home", "away on vacation", "units closed", "respondent refusal", "unit vacant" or "demolished", etc. Other units may respond only partially, e.g. some but not all persons within a dwelling may respond or the units may respond to some but not all questions. Furthermore, units may respond to questions but provide incorrect or inaccurate responses.

Thus, any survey, whatever its type and method of data collection, will suffer from missing data due to nonresponse. Nonresponse has been generally recognized as an important measure of the quality of data since it affects the estimates by introducing a possible bias in the estimates and an increase in sampling variance because of the reduced sample. The relationship between sampling variance and the nonresponse rate is fairly straightforward. However, the relationship between the bias and the size of nonresponse while perhaps more important is less obvious since it depends on both the magnitude of nonresponse and the differences in the characteristics between respondents and nonrespondents. One can speculate that the nonresponse bias is proportional to the nonresponse rate. For a given response rate, the percentage bias would then be independent of sample size. However, the sampling variance is affected by the sample size and is inversely proportional to the responding sample size. Thus, the nonresponse bias may not be nearly so serious relative to the sampling errors for small samples as it is for large samples. The apparent confidence interval may cover the true value in the case of small samples but may not in the case of large samples in the presence of nonresponse bias. If we measure the "seriousness" of the nonresponse bias by the ratio of the nonresponse bias to the coefficient of sampling variation, then the "seriousness" of the nonresponse bias is proportional to the square root of the responding sample size times the nonresponse rate.

In a more practical way, the size of response/nonresponse may indicate the operational problems and provide an insight into the reliability of survey data. However, different types of response/nonresponse rates are used for these two purposes, depending upon whether or

[1] R. Platek, formerly, Director, Census and Households Survey Methods Division, Statistics Canada, G.B. Gray, Social Survey Methods Division, Statistics Canada, Ottawa, Ontario K1A 0T6.

or not a contact has been made with a designated unit. One can therefore distinguish between "contact" and "no contact" of types. One type such as "no one at home" or "temporarily absent" is in fact a "no contact" problem and is primarily operationally oriented. The other type is the true nonresponse problem, where contact has been made with the selected unit but no response or acceptable response is obtained.

In an interview process itself an interviewer may find units in the sample that should not be there (ineligible for the sample). Also, there will be units with questionnaires only or partially completed as well as units with all questionnaires completed. Each of these events may be defined as a rate, i.e. eligibility rate, item response rate, completion rate, etc. The distinction between the "true" nonresponse and other causes affecting the total size of nonresponse rate may give rise to different interpretations.

The interpretation of response/nonresponse rates is particularly difficult when one deals with complex survey designs since the concentration of nonresponse may be higher in one area or class than in another. Still, response rates have been used as proxies for data quality by almost all survey statisticians. That is why the interest in collecting data on nonresponse and the evaluation of it has usually been part of survey taking. However, only the measures of bias, variance, and the resultant mean square error from all sources of sampling and nonsampling errors can provide an informed basis for evaluating survey results.

Recently, nonresponse has been increasing in many surveys in Canada and elsewhere. Consequently, there is a greater need than ever before to monitor nonresponse rates, to make comparisons between surveys, countries, survey organizations, and to ensure some degree of comparability. There have been attempts to standardize the definition of response rate and its complement, the nonresponse rate; see for example, Kviz (1977), Cannell (1978). Problems of inconsistent definitions of response rates related to telephone surveys are described by Wiseman and McDonald (1980).

There are also problems of inconsistent terminology with regard to response/nonresponse in surveys. Terms such as completion rate, contact rate, and under-coverage rate have been used in different contexts in reports and articles dealing with data collection. While these terms may be readily distinguished in an individual report, they may be confusing and subject to conflicting interpretations, when studying different reports.

To consider response/nonresponse problems, a distinction must be made between unit and item nonresponse rates. Unit nonresponse rates generally pertain to the level at which survey data are gathered during the first contact. Examples of the level could be a dwelling, individual, store or establishment. However, in the case of multi-stage sampling, there may be nonresponse of all units within clusters or even primary sampling units (psu) so that unit nonresponse could apply to a selected cluster or psu as well as a dwelling or individual.

Item nonresponse usually pertains to the questionnaires, where information has been provided for some questions but not to all that should have been provided. However, if a unit fails to respond, it automatically fails to respond to any item. Hence, unit nonresponse and item nonresponse are distinct events that should be dealt with separately.

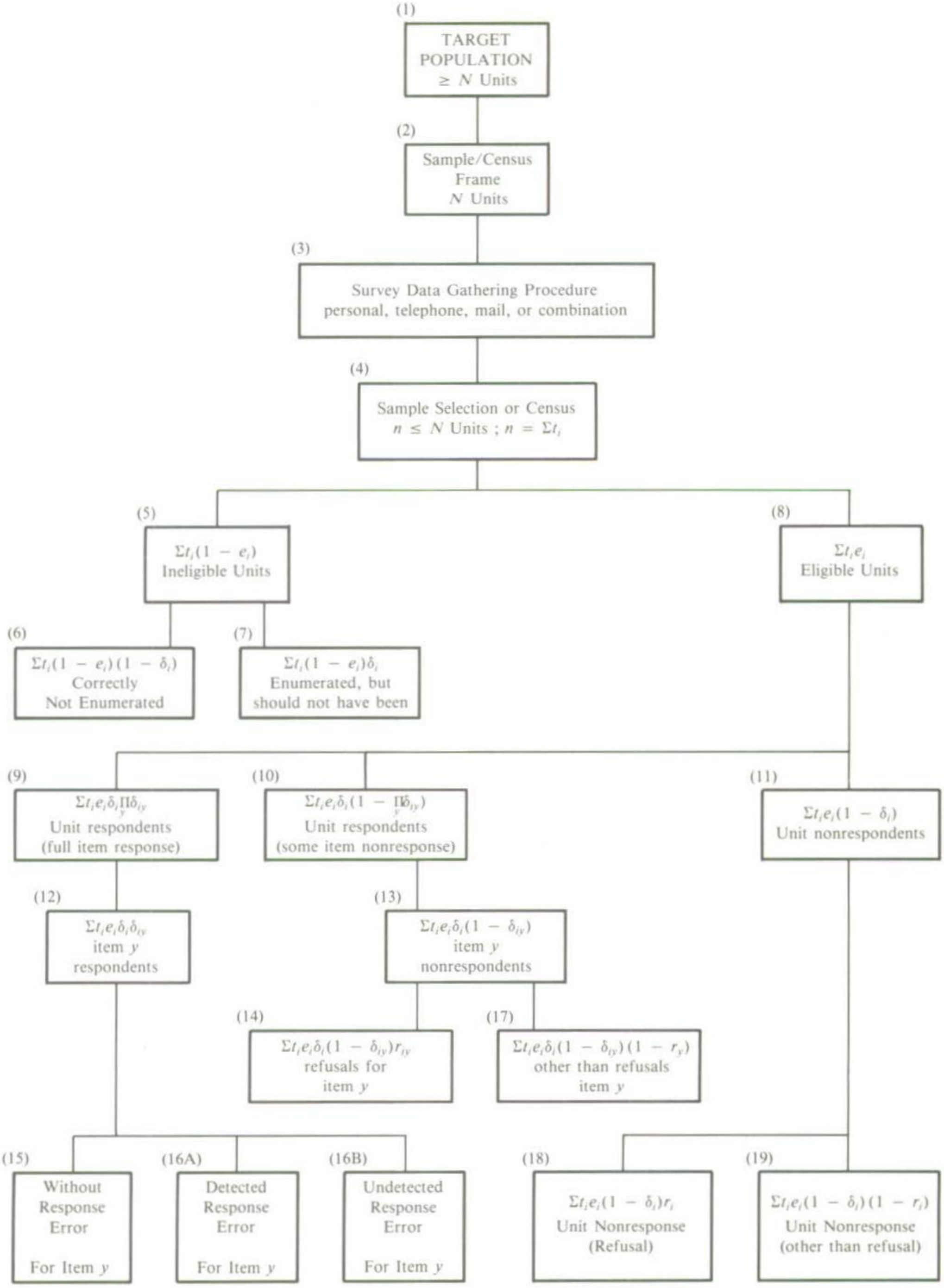
The response rates may pertain to the whole sample and part of a sample such as design-dependent areas or they may apply to administrative areas such as an interviewer assignment, or a group of assignments overseen by a supervisor or field office.

2. RESPONSE/NONRESPONSE COMPONENTS

In order to define various response rates and discuss their uses and applications, it is necessary to split up the target population for the sample or census into the various components, by type of response/nonresponse. Table 1 accomplishes this very purpose, indicating most of the important components of the whole survey that will be used in the rates. Once a target population (Box 1) is defined for a survey, a survey frame of N units (Box 2) is then determined.

Table 1

Response/Nonresponse Components


(1)
TARGET POPULATION $\geq N$ Units
(2)
Sample/Census Frame N Units
(3)
Survey Data Gathering Procedure personal, telephone, mail, or combination
(4)
Sample Selection or Census $n \leq N$ Units ; $n = \Sigma t_i$
(5)
$\Sigma t_i(1 - e_i)$ Ineligible Units
(8)
$\Sigma t_i e_i$ Eligible Units
(6)
$\Sigma t_i(1 - e_i)(1 - \delta_i)$ Correctly Not Enumerated
(7)
$\Sigma t_i(1 - e_i)\delta_i$ Enumerated, but should not have been
(9)
$\Sigma t_i e_i \delta_i \prod_y \delta_{iy}$ Unit respondents (full item response)
(10)
$\Sigma t_i e_i \delta_i (1 - \prod_y \delta_{iy})$ Unit respondents (some item nonresponse)
(11)
$\Sigma t_i e_i (1 - \delta_i)$ Unit nonrespondents
(12)
$\Sigma t_i e_i \delta_i \delta_{iy}$ item y respondents
(13)
$\Sigma t_i e_i \delta_i (1 - \delta_{iy})$ item y nonrespondents
(14)
$\Sigma t_i e_i \delta_i (1 - \delta_{iy}) r_{iy}$ refusals for item y
(17)
$\Sigma t_i e_i \delta_i (1 - \delta_{iy})(1 - r_y)$ other than refusals item y
(15)
Without Response Error For Item y
(16A)
Detected Response Error For Item y
(16B)
Undetected Response Error For Item y
(18)
$\Sigma t_i e_i (1 - \delta_i) r_i$ Unit Nonresponse (Refusal)
(19)
$\Sigma t_i e_i (1 - \delta_i)(1 - r_i)$ Unit Nonresponse (other than refusal)

e_i = 1,0 (unit eligible/ineligible)

t_i = 1,0 (selected/not selected)

δ_i = 1,0 (unit response/nonresponse)

δ_{iy} = 1,0 (item y response/nonresponse)

r_i = 1,0 according as unit refused or not

For r_i = 0, mainly "Not at Home"
or "Temporarily Absent"

It should be mentioned that as a result of possible under- and over-coverage of units the frame may not correspond exactly to the target population. Since under- coverage is usually more prevalent than over-coverage in practice, the actual target population usually contains more than N units.

For the survey to be taken, a data gathering procedure (Box 3) and an appropriate design are decided upon, by or census $n = \Sigma t_i$ units are selected, where:

t_i = 1 or 0 according as unit i is selected or not,

$\sum$ = summation over all N units in the survey frame.

Often, in a sample frame, N may not be precisely known but rather can only be estimated from the sample. This is often the case in multi-stage probability samples with area sampling at earlier stages of selection.

Out of the sample of n units, $\Sigma t_i e_i$ are eligible (Box 8) and $\Sigma t_i(1-e_i)$ are ineligible (Box 5) for the survey, where

e_i = 1 or 0 according as unit i is eligible or not.

Sometimes the eligibility criterion may not be determined if the unit cannot be contacted while at other times the eligibility criterion is obvious from the physical appearance, such as vacant/non-vacant dwellings in a household survey.

The $\Sigma t_i(1 - e_i)$ ineligible units of (Box 5) may be split up between $\Sigma t_i(1 - e_i)(1 - \delta_i)$ units not interviewed just as they should not have been (Box 6) and $\Sigma t_i(1 - e_i)\delta_i$ units incorrectly interviewed (Box 7). One hopes that the number of such units in Box 7 is non-existent or at least very small. However, if such units are discovered, they should be deleted from the sample. In the above and in the breakdowns that follow, δ_i = 1 or 0 according as unit i responded or did not respond.

The $\Sigma t_i e_i$ eligible units (Box 8) may be split up between $\Sigma t_i e_i \delta_i$ unit respondents (Box 9 + Box 10) and $\Sigma t_i e_i(1 - \delta_i)$ unit nonrespondents (Box 11), i.e. they provided no usable survey data and little, if anything, is known about the units, except perhaps their geographic location.

The $\Sigma t_i e_i \delta_i$ units respondents may be split up first between $\Sigma t_i e_i \delta_i \underset{y}{\Pi}(\delta_{iy})$ units, free of item nonresponse, but with possible response errors (Box 9) and $\Sigma t_i e_i \delta_i[1 - \underset{y}{\Pi}(\delta_{iy})]$ units with item nonresponse in at least one characteristic but not in all characteristics (Box 10). Here δ_{iy} = 1 or 0 according as responding unit i responds or does not respond to item or characteristic y. In (Box 9), $\delta_{iy} = 1$ for unit i and all items while in (Box 10), $\delta_{iy} = 0$ for one or more items but not for all of them. For a particular item y, some of the $\Sigma t_i e_i \delta_i \delta_{iy}$ item y respondents (Box 12) come from those unit respondents, free of item nonresponse in (Box 9) while the remainder come from those unit respondents with some item nonresponse among one or more items other than item y. The $t_i e_i \delta_i(1 - \delta_{iy})$ item y nonrespondents of (Box 13) come from those unit respondents with some item nonresponse of (Box 10) that include item y.

The item y respondents of (Box 12) may be decomposed into three components, (i) those units with item y free of response error, (ii) those with a detected response error for item y, and (iii) those with an undetected response error for item y, in Boxes 15, 16A, and 16B respectively.

The $\Sigma\, t_i e_i \delta_i (1 - \delta_{iy})$ item y nonrespondents (Box 13) all come from the unit respondents, i.e. $\delta_i = 1$, $\delta_{iy} = 0$. These item nonrespondents may be decomposed into 2 components, viz., (i) those who refused to reply to question y or those who terminated the interview prior to item y (Box 14) and (ii) those who failed to reply to supply data for item y because of misunderstanding by either the respondent or interviewer or because of other reasons such as failure to follow the proper path in the questionnaire.

Finally, the unit nonrespondent (Box 11) may be split up among refusals (Box 18) and other than refusals (Box 19) mainly non-contacts with reasons such as not at home or temporarily absent. Here, $r_i = 1$ for refusal and $r_i = 0$ for cases of ''other than refusal''. The cases of ''other than refusals'' pertain mainly to ''not at Home'' or ''Temporarily absent.''

In order to count the respondents and nonrespondents according to type and reason, careful records must be kept of every sampled unit. This is essential if a probability sample is not to deteriorate into a quota sample, for example, because of ad hoc treatment of nonresponse, such as arbitrary substitution of other units for the nonrespondents. In the case of quota samples, it is sometimes difficult or impossible to distinguish substituted units from originally selected units when survey takers try to reach the quota with easy-to-obtain survey data from co-operative respondents rather than attempt call-backs of nonrespondents.

Even in probability samples with units carefully labelled and monitored according to plan, it is sometimes difficult to determine precisely the reason for nonresponse among the units that failed to be contacted. The problem is usually most straightforward in the case of personal interviews. However, even in that case, it may be difficult to distinguish ''no one at home'' from ''temporarily absent'' or ''refusals'' from ''non-contacts'' when persons are obviously at home but refuse to answer the door. In the case of telephone interviews, ''no answer'' or ''busy signal'' reveals nothing about the lack of contact of the selected unit although ''refusals'' of contacted units by telephone may be evident. In the case of mail surveys, when the mail is not returned, the reason could be ''refusal'' just as easily as ''temporarily absent''. The ''not at home (unit)'' in the usual context of nonresponse studies as distinguished from ''away from home (unit)'' does not apply to mail surveys. In mail surveys, the reason for nonresponse usually must be determined by personal or telephone follow-up of the unit, often by sub-sampling nonrespondents, some of which may become respondents while others may remain nonrespondents for reasons that may be determined.

The eligibility of selected units is usually evident in the case of personal interviews although failure to contact the units may result in an interviewer's inability to screen out undesirable types of units for a particular survey. No phone answers or busy signals may result in a complete failure to determine either the eligibility or type of nonresponse of the unit. Disconnected telephone numbers or ineligible telephone respondents in a screening survey will provide some measures of ineligibility in a telephone survey. In the case of mail surveys, some returned mail or addresses non- existent among selected units may yield clues about some types of ineligibility while other types may be discernable only by means of personal or telephone follow-up.

3. DEFINITIONS OF VARIOUS RATES

The sample of $n = \Sigma t_i$ units decomposed in Table 1 in section (2) into eligibible units, unit respondents/nonrespondents, refusals, item respondents/non-respondents, etc. leads to many different types of rates which are defined below. For each rate, the numerator is a particular subset of the denominator. Wherever possible, the rate is defined in terms of the counts of units as broken down in Table 1.

(a) *Eligibility Rate*

The eligibility rate is given by:

$$\bar{e} = \sum_i t_i e_i / \sum_i t_i, = (\text{Box } 8)/(\text{Box } 4). \tag{3.1}$$

Wiseman and McDonald (1980) used the term ''incidence rate'' but applied the term only to selected persons of telephone samples that actually answered (responded) at the screening phase to determine their eligibility for the survey.

The eligibility rate, as in (3.1), demonstrates the quality of the survey design in selecting eligible units from a frame, where the eligibility may not be readily determinable without some cursory contact or observation. The rate provides, at the screening stage, information to determine how many eligible units will result at the survey data gathering stage. Thus, the rate may be employed at the design stage if data on eligibility are available from earlier studies. Depending upon the nature and procedure of the survey, the eligibility of units may not be determinable among non-contact or even refusable units. There are two alternatives to the definition of eligibility rate and response rates (which will be defined later) pertaining to eligible units. One can assume, for conservative estimates of data quality and the quality of the procedure for gathering survey data that all non-contacts and refusals would be eligible even though realistically the proportion of eligible units among such nonrespondents is often lower than among respondents and non-respondents for which the eligibility criteria are known. Under the above assumption a lower bound for the response rate and an upper bound for eligibility rate would be obtained. Alternatively, one can assume the same proportion of eligible units among units whose eligibility cannot be determined as among those whose eligibility are known. Under that assumption we would likely have a slight over-estimate of eligibility rate and some of the other rates.

(b) *Response and Completion Rates*

(i) According to one of two alternative definitions provided by the U.S. Federal Committee on Statistical Methodology (1978), the response rate is the percentage of the eligible sample for which information (survey data) is obtained. Thus the response rate is defined as:

$$R_{(1)} = \sum_i t_i e_i \delta_i / \sum_i t_i e_i \tag{3.2}$$

$$= [(\text{Box } 9) + (\text{Box } 10)]/(\text{Box } 8).$$

The above is the most commonly employed response rate in practice as it yields the percent of the sample for which some useful survey data are obtained once the ineligible units are deleted. All types of non- respondents of eligible units are included in the denominator.

The inverse of the above rate at an adjustment cell is frequently used as a weight adjustment to compensate for missing data of nonresponding units, for example, such rates are frequently use in the Canadian LFS for weight adjustments (see Platek and Gray 1985).

The above rate or its complement, the nonresponse rate, is frequently used for administrative and operational assessments of survey organizations. The rates are also used to assess interviewer's ability to contact respondents and to elect this co-operation to provide usable survey data, e.g., response/nonresponse rates by interview assignment. The nonresponse rate includes both refusals, which may be controlled by good public relations and diplomacy, and non-contacts, which may be beyond the control of the interviewer. Hence, whereever possible, the nonresponse rates are frequently split up by reasons. The overall response rate in LFS is abour 95% in most months. Out of the 5% nonresponse about 1% are refusals.

A similar rate to the above was defined as a completion rate by Kviz (1977), who included the whole sample in the denominator. Such a rate may provide a more conservative estimate of quality that (3.2) in that ineligible units such as vacants are included in the denominator. For example, in the LFS, the completion rate by Kviz's definition would drop from 95% according to 3.2 to about 85%.

(ii) Another definition by the above-mentioned committee is the percentage of times an interviewer obtains interviews at sample addresses, where contacts are made given by:

$$R_{(2)} = \sum_i t_i \delta_i / \sum_i t_i [\delta_i + (1 - \delta_i) r_i], \tag{3.3}$$

where unit i refused or did not refuse according as $r_i = 1$ or 0 respectively. The above was defined as a completion rate by O'Neill Groves, and Cannell (1979). If as in (3.3) the eligibility of all units that are contacted can be determined, then another and perhaps superior (known or estimated) definition of the above rate pertaining to eligible units can be given by

$$\begin{aligned} R_{(3)} &= \sum_i t_i \delta_i e_i / \sum_i t_i e_i [\delta_i + (1 - \delta_i) r_i] & (3.4) \\ &= [(\text{Box } 9) + (\text{Box } 10)] / [(\text{Box } 9) + (\text{Box } 10) + \text{Box } 18)] \end{aligned}$$

where e_i, the eligibility criterion is defined after Table 1.

The above rates (3.3) and (3.4) may be useful in personal and telephone surveys where nonrespondents may include non-contacts and refusals. The rates are not practival in mail surveys unless there is a telephone or peronal follow-up of nonrespondents since in most pune mail surveys, the survey organization is forced with either response or nonresponse with unknown reasons. Where the above rates may be useful, however, they measure the ability of a data collection method to elect co-operation of responsible respondents at selected units, given that they are contacted. The non-contacts, that may be beyond the control of interviewers in some survey procedures are removed from the rates entirely.

The response rate in (3.4) was also defined as completion rates by Klecka and Tuchfarber (1979), who assumed, perhaps unrealistically, that all refusals were eligible for the survey. The completion rate would then have ben a conservative estimate for the measure of performance of the data collection method in eliciting the co-operation of eligible units. Alternatively, one may assume the eligibility among refusals to be the same proportion among refusals as among completed and other limits whose eligibility criteria is known.

(c) *Contact Rates*

A "contact rate", defined by Hauck (1974) is the percentage of sample units that are contacted as:

$$R_{(4)} = \frac{\text{Completed interviews} + \text{Refusals (contacted)}}{\text{Completed interviews} + \text{Refusals (contacted} + \text{Noncontacts)}}$$

where the "Noncontacts" were assumed to be eligible for a conservative estimate of the success in contacting sampled units. The "Refusals" may include "Terminations" or "Incomplete Interviews" that are essentially "Refusals" for some items as in (Box 10) of Table 1.

The algebraic expression for the contact rate is given by:

$$R_{(4)} = \frac{\sum_i t_i\delta_i e_i + \sum_i t_i(1-\delta_i)r_i\hat{e}_i}{\sum_i t_i\delta_i e_i + \Sigma t_i(1-\delta_i)r_i\hat{e}_i + \sum_i t_i(1-\delta_i)(1-r_i)\hat{e}_i} \tag{3.5}$$

$$= \frac{(\text{Box } 9) + (\text{Box } 10) + (\text{Box } 18)}{(\text{Box } 9) + (\text{Box } 10) + (\text{Box } 18) + (\text{Box } 19)}, \text{ where}$$

$\hat{e}_i = e_i = 1$ or 0 if eligibility criterion is known,

and, for non-contacts,

$\hat{e}_i = 1$ according to Hauck definition,

or $\hat{e}_i = \bar{e}$, the average eligibility rate among those units whose eligibility criteria are known.

The contact rate measures the ability of the survey organization or interviewers to contact respondents whether or not they succeeded in eliciting their co-operation. In the LFS, the contact rate among non-vacant dwellings is around 96% each month.

(d) *Refusal Rate* (*Non-refusal Rate*)

Two definitions of refusal rates are given by Hauck (1974) and Wiseman and McDonald (1980) respectively as:

$$F_1 = \frac{\text{number of refusals}}{\text{number of completed interviews and refusals}}$$

$$= \sum_i t_i\hat{e}_i(1-\delta_i)r_i / [\sum_i t_i e_i\delta_i + \sum_i t_i\hat{e}_i(1-\delta_i)r_i] \tag{3.6}$$

$$= (\text{Box } 18)/[(\text{Box } 9) + (\text{Box } 10) + (\text{Box } 18)] = 1 - R_{(3)}.$$

and

$$F_2 = \frac{\text{number of refusals}}{\text{number of all selected units}}$$

$$= \sum_i t_i(1-\delta_i)r_i / \sum_i t_i \tag{3.7}$$

$$= (\text{Box } 18)/(\text{Box } 4).$$

With the eligibility criteria taken into account, the refusal rate in (3.7) may be given by:

$$F_3 = \sum t_i\hat{e}_i(1-\delta_i)r_i / \sum t_i\hat{e}_i \tag{3.8}$$

$= (\text{Box } 18)/(\text{Box } 8)$, where $\hat{e}_i$ is defined after (3.5).

The refusal rate measures the extent of the inability of the survey organization or the interviewer to elicit the co-operation of units to provide usable survey data, relative to all contacted units (3.6), relative to the whole sample (3.7) or relative to the eligible sample (3.8). In (3.6), one may wish to determine a "pure" refusal rate without non-contacts that are often beyond the interviewers' control in order to study the efficiency of a questionnaire or effect of the survey topic on the co-operation of contacted units. Alternatively, in (3.7) amd (3.8), one may prefer to examine the refusals rate as one, of several components of overall nonresponse.

(e) *Item Response/Nonresponse Rates*

Complex questionnaire design may result in item nonresponse of specific questions for reasons other than refusals, as noted in Box 17. A controversial or personal question or termination of the interview may result in a refusal to provide data for a specific item as in (Box 14).

Thus, one may measure the overall item nonresponse rate for item y, relative to all responding units, given by:

$$R_y = \frac{(\text{Box } 13)}{(\text{Box } 9) + (\text{Box } 10)}$$

or if item y is relevant only for some units (questionnaires) but not for all of them, one may measure the item nonresponse relative to only those responding units for which item y is relevant (eligible). Consequently, one may define a whole set of item response/nonresponse/eligibility rates, analogous to the unit rates replacing in the rates the number of units (eligible/ineligible)/(responding/refusing, etc.) with the number of responding units (eligible or relevant for item y, irrelevant, responding for item y/refusing for item y etc.) respectively. Most of the rates pertaining to units other than contact rates should have their item y counterparts readily defined by making the proper substitutions in the expressions. However, it may be more difficult to record the reasons for item nonresponse, compared with unit nonresponse, as frequently the item nonresponse is detected only through an edit and imputation routine.

(f) *Weighted Rates and Characteristic Rates*

In the case of sample with different sample weights Π_i^{-1}'s for the units as in probability proportional to size (*pps*) sampling, all of the above rates may be defined as weighted rates by applying the sample weight Π_i^{-1} with the sample selection indicator variable t_i in all the expressions. In the case of self-weighting samples in an area or class for which the rates are calculated the sample weights are redundant. In *pps* sampling at the final stage, however, the usual tendency is for large units to respond more readily than small ones so that weighted response rates, with smaller sample weights applied to the large units than for small units, tend to be smaller than unweighted rates based on the counts of units as in Table 1.

The weighted response rates estimate the proportion of the population that would have responded to the survey under similar survey conditions while the unweighted response rates provide a measure of data collection performance only for the sample or sub-sample pertaining to a specified area or class.

By estimating the nonresponse rate for the entire population rather than for the sample as the unweighted rates do, the weighted rate may provide misleading information on the quality of the data since it may distort the distribution of characteristics in the sample. The advantage of the weighted rates, however, is that the units are added to population levels rather than sample levels so that one obtains an estimate of the rate that would prevail at census levels under similar conditions of gathering survey data. The weighted response rates may under some circumstances be used as weight adjustment factors to inflate the respondents to the full sample in adjustment cells.

When defining characteristic response rates factors include the observed response y_i among item respondents, the imputed value z_{iy} for item nonresponse and the imputed value for z_i for unit nonresponse, which is usually the mean of the respondents in an adjustment cell. If some auxiliary value X_i is known for all units, whether or not they respond, then a characteristic x response rate may be readily calculated and used as a weight adjustment when x is highly correlated with y. The characteristic y response rate, weighted by Π_i^{-1} or unweighted, may be useful in studying the potential nonresponse bias by comparing the charcteristic y response rates with the weighted or unweighted response rates based on counts of units.

4. FINAL REMARKS

Standardization of the definitions of the rates appears to be difficult, owing to the variety of uses and studies of nonresponse and owing to the careful record keeping demanded of survey takers. As long as the rates are unambiguously defined and appropriately applied in their analysis standard definitions for all types of surveys and survey data gathering procedures, may not be all that important. However, in each particular case, the rate should be carefully defined with clear demonstration of the purpose for which it is intended and the reason why it is adopted.

Another issue of standardization dealing with the topic of response/non- response rates is the standard of what is expected from past experience for given surveys, type of survey, subject matter and interview procedure. For example, the response rate, according to 3.2, in the LFS, is expected to be in the 93 to 95% range, with slightly lower rates in the summer months. Out of the 5 to 7% nonresponse, 1% or so may be expected to be refusals. The overall rates have been remarkably consistent for the history of the survey.

It has been observed (see Platek 1977) that finance-oriented surveys tend to have lower response (higher nonresponse) rates than surveys dealing with other topics. The finance surveys appear to be around 25% nonresponse while most of the others centre around 10 to 15%. Also, telephone surveys appear to have a slightly higher nonresponse rate (by about 2 to 3%) than personal surveys for similar subject matter. Thus, from experience, one can determine a standard objective for surveys of a given subject and interview procedure.

It has been observed in publications such as Wiseman and McDonald (1980) that there are many opinions of the way nonresponse should be defined and measured. Thus, it appears that one must grapple with the alternative definitions and terms and obtain relationships between them under various survey conditions. We have attempted to focus on the problems of the various definitions, terms and standards of response rates but have not solved the problems. A proper study can really be undertaken only with a thorough evaluation of survey records, which is possible only when good records are kept. Often, particularly in the case of quota samples, in telephone and mail surveys, nonrespondents are set aside and other units are substituted for them and treated like the originally selected units. The result is a higher observed quality of survey than is the case in reality because of the hidden nonresponse bias. Consequently, the way of treating nonrespondents and the evaluation of nonresponse, completion, etc. must be planned in advance of the survey data gathering in order to deal with it properly rather than during or after the survey.

REFERENCES

CANNELL, CHARLES (1978). Discussion of response rates. Health Survey Research Methods Conference, DHEW Publication No. (PHS) 79-3207.

HAUCK, MATTHEW (1974). Planning field operations. In *Handbook of Marketing Research* (Robert Ferber), New York: McGraw-Hill, 147-159.

KALTON, GRAHAM (1981). Compensating for missing survey data. Survey Research Center, Institute for Social Research, University of Michigan, Annarbor, MI.

KLECKA, W.R. and A.J. TUCHFARBER (1979). Random digit dialing: A comparison to personal surveys. *Public Opinion Quarterly* (Spring), 105-114.

KVIZ, FREDERICK J. (1977). Toward a standard definition of response rate. *Public Opinion Quarterly* (Summer), 265-267.

LINDSTRÖM, HAKAN (1983). Non-response errors in sample surveys. Urval, Nummer 16, Skriftserie utgiven av Statistika Centralbyran, Statistics Sweden, Stockholm.

O'NEILL, MICHAEL J., GROVES, ROBERT M., and CANNELL, CHARLES F. (1979). Telephone interview introductions and refusal rates: Experiments in increasing respondent cooperation. Paper presented at the 1979 Meeting of the American Statistical Association, Washington, D.C.

PLATEK, RICHARD (1977). Some factors affecting non-response. Paper presented at the International Statistical Institute, New Delhi, December.

PLATEK, RICHARD and GRAY, G.B. (1985). Some aspects of nonresponse adjustment. *Survey Methodology*, 11, 1-14.

WISEMAN, FREDERICK, and PHILIP McDONALD (1978). The nonresponse problem in consumer telephone survey. Report No. 78-116, Marketing Science Institue, Cambridge, Mass.

WISEMAN, FREDERICK, and PHILIP McDONALD (1980). Toward the development of industry standards for response and nonresponse rates. Report no. 80-101, Marketing Science Institute, Cambridge, Mass.

Survey Methodology, June 1986
Vol. 12, No. 1, pp. 29–36
Statistics Canada

Some Optimality Results in the Presence of Nonresponse

V.P. GODAMBE and M.E. THOMPSON[1]

ABSTRACT

Using the optimal estimating functions for survey sampling estimation (Godambe and Thompson 1986), we obtain some optimality results for nonresponse situations in survey sampling.

KEYWORDS: Optimum estimating function; Nonresponse.

1. INTRODUCTION AND BACKGROUND

A typical survey sampling set-up consists of a survey population **P** of N labelled individuals i; $\mathbf{P} = \{i: i = 1, ..., N\}$. With each individual i is associated a real value y_i. The vector $\mathbf{y} = (y_1, ..., y_i, ..., y_N)$ is called the population vector. Any subset s of **P** is called a sample. Let $S = \{s\}$. Any probability distribution p on S is called a sampling design. A sample s is drawn using a sampling design p, and the values y_i: $i \epsilon s$ are ascertained through a survey.

Thus the data here are χ_s where

$$\chi_s = \{s, (i,y_i): i \in s\}. \tag{1.1}$$

On the basis of the data x_s one tries to estimate a survey population parameter θ_N, that is a specified real function of the population vector **y**; $\theta_N = \theta_N(y)$.

In relation to the above estimation problem we assume a superpopulation model under which $y_1, ..., y_N$ are independent and for certain known covariate values x_i, $i = 1, ..., N$,

$$\epsilon(y_i - \theta x_i) = 0, \; i = 1, ..., N, \tag{1.2}$$

ϵ being the expectation with respect to the model. In the model (1.2), θ is the usual unknown regression parameter, the expectation being taken holding x_i fixed. The usual intercept term of the regression model is not mentioned in (1.2), for this term can often be eliminated by an appropriate stratification (Godambe 1982). Note the model (1.2) does not specify the variance function.

Following Godambe and Thompson (1986), for some *specified* numbers α_i, $i = 1, ...,$ N, we define the survey population parameter θ_N as the solution of the equation

$$\bar{g} = \sum_{i=1}^{N} (y_i - \theta x_i)\alpha_i = 0. \tag{1.3}$$

[1] V.P. Godambe and M.E. Thompson, Department of Statistics and Actuarial Science, University of Waterloo, Ontario, Canada, N2L 3G1.

That is,

$$\theta_N = \sum_{i=1}^{N} y_i\alpha_i / \sum_{i=1}^{N} x_i\alpha_i. \tag{1.4}$$

The parameter θ_N is related to the model (1.2) through the equation

$$\epsilon\bar{g} = 0. \tag{1.5}$$

Any real function h of the data χ_s in (1.1) and the parameter θ is called an *unbiased estimating function* for both the parameters θ_N and θ if

$$E(h - \bar{g}) = 0 \text{ for all } \mathbf{y} \text{ and } \theta \tag{1.6}$$

'E' being the expectation under the sampling design p employed to draw the sample s. Because of (1.5) and (1.6) we say the solution of the equation

$$h(\chi_s, \theta) = 0,$$

for the given data χ_s, estimates both the parameters and θ and θ_N, given by (1.2) and (1.4) respectively. For the function $\bar{g}$ in (1.4), under the sampling design p, let $H_{(p)}$ be the class of all unbiased estimating functions h. That is

$$H(p) = \{h\colon E(h - \bar{g}) = 0 \text{ for all } \mathbf{y} \text{ and } \theta\}. \tag{1.7}$$

Now we say an *estimating function* $h^* \in H(p)$ is *optimum* if

$$\epsilon E(h^*)^2 \leq \epsilon E(h)^2, \text{ for all } h \in H(p) \tag{1.8}$$

(Godambe and Thompson 1986). Further, when the inequality (1.8) is satisfied,

$$h^* = 0 \tag{1.9}$$

is said to be the *optimum estimating equation* for estimating the parameter θ_N given by (1.3) and (1.4).

For the sampling design p, used to draw a sample s, let π_i, $i = 1, \ldots, N$ be the inclusion probabilities. That is

$$\pi_i = \sum_{s \ni i} p(s), \ i = 1, \ldots, N, \tag{1.10}$$

where $s \ni i$ indicates all samples s which include the individual i. We assume

$$\pi_i > 0, \ i = 1, \ldots, N. \tag{1.11}$$

Theorem 1.1. (Godambe and Thompson 1986). For any sampling design p satisfying (1.11), under the model (1.2), in the class of all unbiased estimating functions $H(p)$ in (1.7), the optimum h^*, that is h^* satisfying (1.8), is given by

$$h^* = \sum_{i \in s} (y_i - \theta x_i)\alpha_i / \pi_i, \tag{1.12}$$

π_i being the inclusion probability given by (1.10). Thus the optimum estimating equation here is

$$\sum_{i \in s} (y_i - \theta x_i)\alpha_i / \pi_i = 0. \tag{1.13}$$

The estimate $\hat{\theta}_s$ of the survey population parameter θ_N in (1.4) and the superpopulation parameter θ in (1.2) is given by

$$\hat{\theta}_s = \frac{\sum_{i \in s} y_i \alpha_i / \pi_i}{\sum_{i \in s} x_i \alpha_i / \pi_i}. \tag{1.14}$$

This estimate was previously put forward by Brewer (1963) and Hájek (1971) on some ''plausibility'' considerations.

To explain the relationships of Theorem 1.1 above with earlier optimality results (e.g. Godambe 1982) we put $\alpha_i \equiv 1$ in (1.3) and therefore in (1.2). Further, we consider a superpopulation model obtained from (1.2) by letting $\theta = \theta_0$, a specified value. Now for any sampling design with inclusion probabilities π_i satisfying (1.11), in the class of all design unbiased estimates of θ_N (in (1.4) with $\alpha_i = 1$, $i = 1, \ldots, N$), the superpopulation expectation of the design variance is minimized for the estimate

$$e = \frac{1}{X}\left\{ \sum_{i \in s} \frac{y_i - \theta_0 x_i}{\pi_i} + \theta_0 \sum_{i=1}^{N} x_i \right\} \tag{1.15}$$

where $X = \Sigma_1^N x_i$. This ''optimality'' of the estimate e at $\theta = \theta_0$ carries over to all values of θ if the sampling design is such that

$$\text{Probability}\left\{ s{:}\left(\sum_{i \in s} \frac{x_i}{\pi_i} - \sum_{i=1}^{N} x_i \right) = 0 \right\} = 1. \tag{1.16}$$

Now when the sampling design satisfies condition (1.16), then $\hat{\theta}_s$ in (1.14) is equal to e in (1.15). Thus all the earlier optimality results are covered by Theorem 1.1, and it does a great deal more: in many situations, such as for designs with $\pi_i \propto x_i$, the condition (1.16) implies a *fixed sample size* design. In contrast the ''optimality'' in Theorem 1.1 holds regardless of the fixed sample size design condition. That is, the ''optimality'' is available for *random sample* size designs, which are common in the nonresponse situations discussed subsequently.

2. NONRESPONSE AND OPTIMALITY

Suppose a sample s is drawn from the survey population $\mathbf{P}$, using a sampling design p. Suppose because of nonresponse the variate values y_i are available only for the subset $s' \subset s$; $s - s'$ are the non-respondents. Thus now the data instead of χ_s in (1.1) are

$$\chi_{s,s'} = (s, s', \{(i,y_i): i \in s'\}). \tag{2.1}$$

We may now consider two problems of estimation:

(I) If there were no nonresponse, that is if all the data χ_s in (1.1) where available, we would have estimated the survey population parameter θ_N in (1.4) by solving the optimum estimating equation given by (1.12), namely $h^* = 0$. When the hypothetical data χ_s are replaced by $\chi_{s,s'}$ in (2.1), one may try to estimate h^* with some function $h'(\chi_{s,s'})$. This is in line with a suggestion of Rubin (1976). Following (1.7) we define the class of unbiased estimating functions h' (for h^*, given the sample s) as

$$H'(p,.,s) = \{h': E(h' - h^*|s) = 0, \text{ for all } \mathbf{y} \ \& \ \theta\}; \tag{2.2}$$

the '.' in H' indicates that the class H' would be specified only after the *response mechanism* is specified. Again we define h'^* as the optimum estimating function in H' in (2.2), if $h'^* \in H'$ and if under the model (1.2), $\epsilon E(h'^*)^2 \leq \epsilon E(h'^*)^2$ for all $h' \in H'$.

(II) Alternatively we could try to estimate the survey population parameter θ_N directly, that is without estimating h^* as in (I) above, from the data $\chi_{s,s'}$. In line with (1.7) we define the class of unbiased estimating functions $h''(\chi_{s,s'})$:

$$H''(p,.) = \{h'': E(h'' - \bar{g}) = 0, \text{ for all } \mathbf{y} \ \& \ \theta\}; \tag{2.3}$$

as before the '.' in H'' indicates that the class H'', for its specification, requires the specification of the *response mechanism*. Again h''^* is called the *optimum estimating function* in H'' if $h''^* \in H''$ and if under (1.2), $\epsilon E(h''^*)^2 \leq \epsilon E(h'')^2$ for all estimating functions $h'' \in H''$.

In $H'(p,.,s)$ and $H''(p,.)$ of (2.2) and (2.3) we have left the response mechanism '.' unspecified. Now we specify it.

RESPONSE MECHANISM: If the individual 'i' of the survey population $\mathbf{P}$ were included in the sample s drawn,

'i' would respond with *known* probability q_i
and would fail to respond with probability $1 - q_i$, (2.4)

$i = 1, \ldots, N$; we assume $q_i > 0$, $i = 1, \ldots, N$.

The response mechanism $\mathbf{q} = (q_1, \ldots, q_N)$ in (2.4) completely characterizes the class $H'(p,.,s)$ in (2,2) as $H'(p, \mathbf{q}, s)$ and $H''(p,.)$ in (2.3) as $H''(p, \mathbf{q})$.

The case (I) above is implemented by the following Theorem 2.1 and the remaining Theorems 2.2, 2.3 and 2.4 implement the case (II).

Theorem 2.1. For any sampling design p satisfying (1.11), and for any sample s, in the class of estimating functions $H'(p, \mathbf{q}, s)$ in (2.2) under the superpopulation model (1.2) $\epsilon E\{h')^2 | s\}$ is minimized for $h' = h'^*$ where

$$h'^* = \sum_{i \in s'} (y_i - \theta x_i)\alpha_i/\pi_i q_i; \tag{2.5}$$

that is h'^* is the optimum estimating function in $H'(p, \mathbf{q}, s)$. □

Proof. As was emphasized in Section 1, the optimality of h^* in (1.12) obtains even for *random sample size* designs and for any values of α_i, $i = 1, \ldots, N$ in (1.3). Thus the proof of Theorem 2.1 is accomplished by replacing, in Theorem 1.1, the population 'P' by 's' and α_i by α_i/π_i, $i \in s$ and noting that now the inclusion probabilities are q_i, $i \in s$. □

Theorem 2.2. Let $\bar{H}''$ be the subclass of H'' in (2.3) such that any estimating function $h''(\chi_{s,s'})$ in $\bar{H}''$ depends on (s,s') only through s'. Then for any sampling design p satisfying (1.11), in the class $\bar{H}''(p, \mathbf{q})$, under the superpopulation model (1.2), $\epsilon E\{(h'')^2\}$ is minimized for $h'' = h''^*$ where

$$h''^* = \sum_{i \epsilon s'} (y_i - \theta x_i)\alpha_i/\pi_i; \tag{2.6}$$

that is h''^* is the optimum estimating function in $\bar{H}''(p, \mathbf{q})$. □

Proof. This follows directly from Theorem 1.1, by replacing in it s by s' and the inclusion probabilities by π_i by $\pi_i q_i$, $i = 1, \ldots, N$.

Theorem 2.3. The estimating function h''^* in (2.6) is the optimum estimating function in the entire class $H''(p, \mathbf{q})$ given by (2.3). That is the result of the Theorem 2.2 is valid without the restriction to the subclass $\bar{H}''$ of H''. □

Proof. For any given response probabilities $\mathbf{q}$ in (2.4) and the sampling design p, the statistic $(\{i, y_i\}: i \in s')$ is *sufficient* for the population vector $\mathbf{y}$. More specifically, referring to (1.1) and (2.1), we have the conditional probability Prob$(\chi_{s,s'} \mid \chi_{s'}, \mathbf{y})$ independent of $\mathbf{y}$. Hence for any estimating function $h'' \in H''(p, \mathbf{q})$ in (2.3) we have the estimating function $E(h'' \mid \chi_{s'}) = \bar{h}'' \in \bar{H}''$ and $\epsilon E(\bar{h}'')^2 \leq \epsilon E(h'')^2$. This proves Theorem 2.3.

When $s \equiv s'$, that is when there are no nonrespondents, do we still estimate h^* by $h'^* = h''^*$? The obvious negative answer to this question is obtained, as shown by Godambe (1986), by an appropriate *conditioning*. The same reservation tends to be felt for cases where there are only a few nonrespondents, and again appropriate conditioning holds some promise of a resolution. In summary the formal optimality of $h'^* = h''$ suggests that it is useful, and is likely to give good estimation when nonresponse is considerable and the relative values of the q_i are known. However, it can clearly be improved upon in situations when nonresponse is rare; improved versions will have natural conditional interpretations. Appropriate conditioning becomes even more important in the case of unknown response probabilities, as will be seen next.

Now we assume that the survey population $\mathbf{P}$ is divided into k strata $\mathbf{P}_j$, of sizes N_j, $j = 1, \ldots, k$. Further suppose that the response probabilities are constant within each stratum. That is

$$q_i = q^{(j)} \text{ for all } i \in \mathbf{P}_j; \; j = 1, \ldots, k. \tag{2.7}$$

Unlike in (2.4), where the response probalities were assumed to be known, now we assume that in (2.7), the response probabilities $q^{(j)}$, $j = 1, \ldots, k$ are *unknown*. Let p_0 denote the stratified sampling design, consisting of drawing from the stratum $\mathbf{P}_j$, a simple random sample (without replacement) of size n_j, $j = 1, \ldots, k$. Now as in (2.3) we define the class of unbiased estimating functions $h_1(\chi_{s,s'})$

$$H_i(p_0) = \{h_i\text{: } E(h_1 - \bar{g}) = 0 \text{ for all } \mathbf{y}, \theta \text{ and } q^{(j)}, j = 1, \ldots, k\}, \tag{2.8}$$

where $q^{(j)}$ are as in (2.7). Let $s_j' = s' \cap \mathbf{P}_j$ and $|s_j'| = n_j'$, that is the size of the sample of respondents from the stratum $\mathbf{P}_j$, $j = 1, \ldots, k$.

Theorem 2.4. For the sampling design p_0, in the class of estimating functions $H_i(p_0)$ in (2.8), under the superpopulation model (1.2), $\epsilon E(h_1^2)$ is minimized for $h_1 = h_1^*$ where

$$h_1^* = \sum_{j=1}^{k} \sum_{i \in s_j'} (y_i - \theta x_i)\alpha_i / (\frac{n_j'}{N_j}); \tag{2.9}$$

that is h_1^* is the optimum estimating function in $H_i(p_0)$.

Proof. The sampling distribution of the data $\chi_{s,s'}$ in (2.1) depends, in addition to the unknown population vector y, on the unknown (parameter) $q^{(j)}$, $j = 1, \ldots, k$. Now for every fixed $\mathbf{y}$, the statistic n_j', $j = 1, \ldots, k$ is *completely sufficient* for the parameter $q^{(j)}$, $j = 1, \ldots, k$. Hence for a fixed $\mathbf{y}$ and θ, in (2.8),

$$[E(h_1 - \bar{g}) = 0, \text{ for all } q^{(j)}, j = 1, \ldots, k]$$

$$\Rightarrow E\{(h_1 - \bar{g}) | n_j', j = 1, \ldots, k\} = 0, \tag{2.10}$$

ignoring sets of '0' measure. Further, *conditional* on the number of respondents n_j' from the stratum P_j, the probability of $i \in s_j'$ is $(n_j/N_j)(n_j'/n_j) = (n_j'/N_j)$. Hence for any estimating function $h_1 \in H_1$ in (2.8) we have from Theorem 2.3.

$$\epsilon E((h_1^*)^2 | n_j', j = 1, \ldots, k\} \leq \epsilon E\{(h_1)^2 | n_j', j = 1, \ldots, k\}, \tag{2.11}$$

h_1^* being given by (2.9). Theorem 2.4 is proved by taking the expectations of both sides of (2.11) for the variations of n_j', $j = 1, \ldots, k$.

The optimum estimating function h_1^* in (2.9) has the following intuitive interpretation. If in (2.7), the response probabilities $q^{(j)}$, $j = 1, \ldots, k$ were *known*, by Theorem 2.3, the optimum estimating function, for the sampling design p_o, would be given by

$$h'' = \sum_{j=1}^{k} \sum_{i \in s_j'} (y_i = \theta x_i)\alpha_i / (\frac{n_j}{N_j} q^{(j)}).$$

Now when $q^{(j)}$ are unknown (which is the case in Theorem 2.4), we *estimate* them by (n_j'/n_j), $j = 1, \ldots, k$. Substituting these estimates for $q^{(j)}$ in h'' yields the estimating function h_1^* of (2.9).

These estimates obtained by solving the equations $h'^* = 0$, $h''^* = 0$ and $h_1^* = 0$ in (2.5), (2.6) and (2.9) respectively have previously been proposed, on plausibility considerations, by several authors. A good reference in this connection in Cassel et al. (1983). The assumption (2.4) of "response probabilities" seems to have evolved gradually in the literature. An interesting early reference in this connection is Hartley (1946).

3. OPTIMAL INCLUSION PROBABILITIES

It should be emphasized here that the "optimality" of the estimating function $h''*$ in (2.6) was established under the superpopulation model (1.2), which does *not* specify the variance function. However the specification of the variance function in the model (1.2) would be required to obtain the "optimal" inclusion probabilities. We assume

$$\epsilon(y_i - \theta x_i)^2 = \sigma^2 f(x_i),\ i = 1, \ldots, N, \tag{3.1}$$

where f is a *known* function of x, and σ^2 can be unknown. Now for the estimating function $h''*$ in (2.6), (3.1), we have

$$\epsilon E(h''*)^2 = \sum_{i=1}^{N} \frac{\epsilon(y_i - \theta x_i)^2 \alpha_i^2}{\pi_i q_i} = \sigma^2 \sum_{i=1}^{N} \frac{f(x_i)\alpha_i^2}{\pi_i q_i} \tag{3.2}$$

In (3.2), the response probabilities q_i as said in (2.4) are given (fixed) numbers. However, (a sampling design with) the optimal inclusion probabilities can be obtained by minimizing $\epsilon E(h''*)^2$ in (3.2) under a restriction, either (A) or (B).

$$(A): \sum_{i=1}^{N} \pi_i = \text{constant},$$

$$(B): \sum_{i=1}^{N} \pi_i q_i = \text{constant} \tag{3.3}$$

In (A) we hold the average size of the sample s fixed, for $E|s| = \Sigma_i^N \pi_i$. In (B) we hold fixed the average size of the effective sample s', for $E|s| = \Sigma_i^N \pi_i q_i$. Now since the q_i are fixed numbers we have for minimizing $\epsilon E(h''*)^2$ in (3.2), respectively,

$$(A): \pi_i \propto \{\frac{f(x_i)}{q_i}\}^{1/2}\alpha_i,$$

$$(B): \pi_i \propto \frac{(f(x_i))^{1/2}}{q_i}\alpha_i. \tag{3.4}$$

Denoting by n' the size of the effective sample s', that is $|s'| = n'$, we have from (B) in (3.4),

$$\pi_i = \frac{(f(x_i))^{1/2}\alpha_i}{\{\Sigma_1^N (f(x_i))^{1/2}\alpha_i\}} \quad \frac{E(n')}{q_i},\ i = 1, \ldots, N. \tag{3.5}$$

Further for a fixed sample size design such that

$$\text{Probability } \{s: |s| \neq n\} = 0,$$

we have from (3.5).

$$\sum_{i=1}^{n} \pi_i = n = \sum_{i=1}^{N} \frac{(f(x_i))^{½}\alpha_i}{\{\Sigma_1^N (f(x_i))^{½}\alpha_i\}} \quad \frac{1}{q_i} E(n'). \tag{3.6}$$

As a special case, when all the response probabilities q_i, $i = 1, \ldots, N$ are equal, $q_i = q$ say, $i = 1, \ldots, N$, in (3.6),

$$n = E(n')/q; \tag{3.7}$$

for instance if $q = 1/2$, the sample size of the (initial) sample s should be double the expectation of the effective sample (s') size!

Now we assume the survey population $\mathbf{P}$ to be divided into strata P_j, $= 1, \ldots, k$ so that the response probabilities in each stratum are constant, that is they satisfy (2.7). For a stratified sampling design consisting of drawing a sample of size n_j from the stratum $\mathbf{P}_j$, $j = 1, \ldots k$ we have from (3.5).

$$n_j = \frac{E(n')}{q^{(j)}} \; \frac{\sum_{i \in \mathbf{P}_j} (f(x_i))^{1/2}\alpha_i}{\sum_{i \in \mathbf{P}} (f(x_i))^{1/2}\alpha_i,} \, , j = 1, \ldots, k.$$

If $(f(x_i))^{1/2}\alpha_i$ are constant for $i = 1, \ldots, N$, it is clear from (3.8) that optimal allocation implies drawing a relatively larger sample from the stratum with smaller response probability. Actually in this situation

$$E(n_j') = E(n')/k$$

where n_j' is the size of the effective sample s_j' from the stratum $\mathbf{P}_j$, $j = 1, \ldots, k$.

REFERENCES

BREWER, K.R.W. (1963). Ratio estimation in finite populations: some results deducible from the assumption of an underlying stochastic process. *Australian Journal of Statistics*, 5, 93-105.

CASSEL, C.M., SARNDA, C.E., and WRETMAN, J.H. (1983). Some uses of statistical models in connection with the nonresponse problems. In *Incomplete Data in Sample Surveys*, Vol. 3, (Eds. W.G. Madow and Ingram Olkin), New York: Academic Press, 143-160.

GODAMBE, V.P. (1982). Estimation in survey sampling: Robustness and optimality. *Journal of the American Statistical Association*, 77, 393-403.

GODAMBE, V.P. (1986). Quasi-score function, quasi-observed Fisher information and conditioning in survey sampling (unpublished).

GODAMBE, V.P. and THOMPSON, M.E. (1986). parameters of superpopulation and survey population: Their relationships and estimation. *International Statistical Institute Review* (to Appear).

HAJEK, J. (1971). Contribution to discussion of paper by D. Basu. In *Foundations of Statistical Inference*, (Eds. V.P. Godambe and D.A. Sprott), Toronto: Holt, Rinehart and Winston, 236.

HARTLEY, H.O. (1946). Discussion of paper by F. Yates. *Journal of the Royal Statistical Society* Series A, 109, 37.

RUBIN, D.B. (1976). Inference and missing data. *Biometrika*, 63, 581-589.

Survey Methodology, June 1986
Vol. 12, No. 1, pp. 37–47
Statistics Canada

Basic Ideas of Multiple Imputation for Nonresponse

DONALD B. RUBIN[1]

ABSTRACT

Multiple imputation is a technique for handling survey nonresponse that replaces each missing value created by nonresponse by a vector of possible values that reflect uncertainty about which values to impute. A simple example and brief overview of the underlying theory are used to introduce the general procedure.

KEY WORDS: Survey nonresponse; Proper imputation methods; Multiple imputation.

1. INTRODUCTION

Any statistician with experience in the field of surveys knows that essentially every survey suffers from some nonresponse. That is, in practical surveys, some items in the survey instrument are not answered by all units included in the survey. Commonly, the items likely to be unanswered are the more sensitive ones, such as those concerning personal income. Because nonresponse creates missing values, the complete-data statistics that would have been used in the absence of nonresponse can no longer be calculated. An obvious desire of both the data collector and the data analyst is to get rid of the missing values and thereby restore the ability to use standard complete-data methods to draw inferences.

1.1 Imputation

It is not surprising, therefore, that a very common method of handling the missing values created by nonresponse is to fill them in, or impute them. That is, when using imputation to handle nonresponse each missing value is replaced with a real value. Many different procedures have been proposed for imputation, for instance, filling in the respondents' mean for that variable or a value predicted from the modelling of the missing variable given observed variables using respondent data; as a specific example, when the missing value is personal income, a linear regression model predicting log(income) from demographic characteristics such as age, sex, education and occupation might be regarded as reasonable.

1.2 Advantages and Disadvantages of Single Imputation

In addition to the obvious advantage of allowing complete-data methods of analysis, imputation by the data collector (e.g. the Census Bureau) also has the important advantage of being able to utilize information available to the data collector but not available to an external data analyst such as a university social scientist analyzing a public-use file. This information may involve detailed knowledge of interviewing procedures and reasons for nonresponse that are too cumbersome to place in public-use files, or may be facts, such as street addresses of dwelling units, that cannot be placed on public-use files because of confidentiality constraints. This kind of information, even though inaccessible to the user of a public-use file, can often narrow the possible range of imputed values.

[1] Donald B. Rubin, Department of Statistics, Harvard University, Science Center, One Oxford Street, Cambridge, Massachusetts, 02138, U.S.A.

Just as there are obvious advantages to imputing one value for each missing value, there are obvious disadvantages of this procedure arising from the fact that the one imputed value cannot itself represent any uncertainty about which value to impute: If one value were really adequate, then that value was never missing. Hence, analyses that treat imputed values just like observed values generally systematically underestimate uncertainty, even assuming the precise reasons for nonresponse are known. Equally serious, single imputation cannot represent any additional uncertainty that arises when the reasons for nonresponse are not known.

1.3 Multiple Imputation to the Rescue

Multiple imputation, first proposed in Rubin (1977, 1978), retains the two major advantages of single imputation and rectifies its major disadvantages. As its name suggests, multiple imputation replaces each missing value by a vector composed of $M \geq 2$ possible values. The M values are ordered in the sense that the first components of the vectors for the missing values are used to create one completed data set, the second components of the vectors are used to create the second completed data set and so on. The first major advantage of single imputation is retained with multiple imputation, since standard complete-data methods are used to analyze each completed data set. The second major advantage of imputation, that is, the ability to utilize data collectors' knowledge in handling the missing values, is not only retained but actually enhanced. In addition to allowing data collectors to use their knowledge to make point estimates for imputed values, multiple imputations allow data collectors to reflect their uncertainty as to which values to impute. This uncertainty is of two types: sampling variability assuming the reasons for nonresponse are known, and variability due to uncertainty about the reasons for nonresponse. Under each posited model for nonresponse, two or more imputations are created to reflect sampling variability under that model; imputations under more than one model for nonresponse reflect uncertainty about the reasons for nonresponse. The multiple imputations within one model are called repetitions and can be combined to form a valid inference under that model; the inferences under different models can be contrasted to reveal sensitivity of answers to posited reasons for nonresponse.

Before reviewing some more general results in Section 3, Section 2 illustrates essential ideas in a highly artificial example used in Rubin (1986a), which is a comprehensive treatment of multiple imputation. Other references on multiple imputation include Rubin (1979, 1980, 1986b), Herzog and Rubin (1983), Li (1985), Schenker (1985), Rubin and Schenker (1986), and Heitjan and Rubin (1986).

2. AN ARTIFICIAL EXAMPLE ILLUSTRATING MULTIPLE IMPUTATION

Suppose we have taken a simple random sample of $n = 10$ units from a large population. The objective of the survey is to estimate $\bar{Y}$ the mean of Y in the population. We know the mean value of a covariate X in the population, and the survey attempts to record both X and Y for each of the n units included in the sample.

Table 1 presents the observed values of (Y, X) for the ten units in the sample where the question marks indicate missing Y data due to nonresponse.

2.1 Multiply Imputing for the Missing Values

Suppose the missing values in Table 1 are to be multiply imputed using two values drawn under each of two models (i.e. two repetitions per model). In general, any number of models can be used with any number of repetitions within each model. Model 1 is an "ignorable" model for nonresponse; ignorable is defined precisely in Rubin (1976), but essentially it means that a nonrespondent is only randomly different from a respondent with the same value of *X*. Model 2 is a nonignorable model and posits a systematic difference between respondents and nonrespondents with the same value of *X*. The repeated imputations under each model are based on a simple procedure closely related to the hot-deck, which can be improved upon but is useful to illustrate ideas.

For each nonrespondent, the two closest matches among the respondents are found, where the distance for matching is defined by the values of *X*. For the first nonrespondent, unit 2, the two closest matches are units 1 and 3, and for the second nonrespondent, unit 4, the closest matches are 3 and 5. The repeated imputations are created by drawing at random from the two closest matches. For the ignorable model, we simply impute the value *Y* provided by the matching respondent: the first two columns of Table 2 give the result. For the nonignorable model, we suppose that the nonresponse bias is such that a nonrespondent will tend to have a value of *Y* 20% higher than the matching respondent's value of *Y*: the last two columns of Table 2 give the result where the *Y* values have been rounded to the nearest integer. The repeated imputations within each model allow the user to draw a valid inference under that model. The use of two models, an ignorable one and a nonignorable one, allows the display of sensitivity of inference to assumptions about nonresponse. Generally such assumptions are untestable using the data at hand.

Table 1

Observed Data

Unit	*Y*	*X*
1	10	8
2	?	9
3	14	11
4	?	13
5	16	16
6	15	18
7	20	6
8	4	4
9	18	20
10	22	25

Table 2

Multiple Imputations for Data of Table 1

	Model 1 Repetition		Model 2 Repetition	
	1	2	1	2
Unit 2	10	14	12	17
Unit 4	16	14	19	17

2.2 Analyzing the Resultant Multiply-Imputed Data Set

Each set of imputations, that is each column of Table 2, can be used with the incomplete data in Table 1 to create a completed data set. Since there are four sets of imputations, four completed data sets can be created; these are displayed in Tables 3 to 6. Each completed data set is analyzed just as if there had been no nonresponse.

Assume that with complete data, the ratio estimator $\bar{X}\bar{y}/\bar{x}$ would be used with associated variance SE^2, where $\bar{X}$ is the known mean of X in the population, say 12, $\bar{y}$ and $\bar{x}$ are the means of Y and X in the random sample of n units, and

$$SE^2 = \sum (Y_i - X_i\bar{y}/\bar{x})^2/[n(n-1)]$$

Table 3

Complete Data Set 1 (Model 1, Rep. 1)
For Multiply Imputed Data Set of Tables 1 and 2

Unit	Y	X
1	10	8
2	10	9
3	14	11
4	16	13
5	16	16
6	15	18
7	20	6
8	4	4
9	18	20
10	22	25
means	14.5	13

Table 4

Complete Data Set 2 (Model 1, Rep. 2)
For Multiply Imputed Data Set of Tables 1 and 2

Unit	Y	X
1	10	8
2	14	9
3	14	11
4	14	13
5	16	16
6	15	18
7	20	6
8	4	4
9	18	20
10	22	25
means	14.7	13

Table 5

Complete Data Set 3 (Model 2, Rep. 1)
For Multiply Imputed Data Set of Tables 1 and 2

Unit	Y	X
1	10	8
2	12	9
3	14	11
4	19	13
5	16	16
6	15	18
7	20	6
8	4	4
9	18	20
10	22	25
means	15	13

Table 6

Complete Data Set 4 (Model 2, Rep. 2)
For Multiply Imputed Data Set of Tables 1 and 2

Unit	Y	X
1	10	8
2	17	9
3	14	11
4	17	13
5	16	16
6	15	18
7	20	6
8	4	4
9	18	20
10	22	25
means	15.3	13

Table 7

Ratio Estimates and Associated Variances of Estimates
for the Complete Data Sets of Tables 3-6

	Model 1 Repetition		Model 2 Repetition	
	1	2	1	2
Estimate	13.38	13.57	13.85	14.12
Variance	2.96	3.19	3.38	3.84

Table 8

Combined Estimates and Variances for the Multiply Imputed Data Sets of Tables 1 and 2

	Model 1	Model 2
Estimate	13.48	13.98
Variance	3.10	3.66

where the sum is over the units in the sample. Table 7 presents the estimates and variances associated with each of the four completed data sets given in Tables 3-6.

The two answers obtained under the same model can be combined to obtain one inference for $\bar{Y}$ under each model. The results are displayed in Table 8: the estimate is the average of the estimates and the variance associated with this estimate has two components: (i) the average within-imputation variance associated with the estimate and (ii) the between-imputation variance of the estimate. Thus, under Model 1, the estimate is $(13.38 + 13.57)/2 = 13.48$; the associated estimated average within variance is $(2.96 + 3.19)/2$, and the associated estimated between variance is $[(13.38 - 13.48)^2 + (13.57 - 13.48)^2)]$. The estimated variances are combined as: (estimated total variance) = (estimated average within variance) + $(1 + M^{-1})$ × (estimated between variance), where the factor $(1 + M^{-1})$ multiplying the usual unbiased estimate of between variance is an adjustment for using a finite number of imputations. The associated 95% interval estimate for $\bar{Y}$ is (10.0, 16.9) under Model 1 and (10.2, 17.7) under Model 2. In practice, better intervals can be formed by calculating degrees of freedom as a simple function of the variance components and using the 95% points appropriate to the corresponding t-distribution; when either M is large or the between variance component is small relative to the total variance (as in this artificial example), the degrees of freedom will be large and thus the normal 95% points will be used. Details are given in Section 3.

The essential feature to notice in this illustrative example is that only complete-data methods of analysis are needed. We merely have to perform the complete-data analysis that would have been used in the absence of nonresponse on each of the completed data sets created by the multiple imputations. The resultant answers under each model are then easily combined to give one inference under each model. Although not illustrated here, diagnostic analyses using complete-data techniques can be applied to each completed data set; Heitjan and Rubin (1986) provides several examples.

3. GENERAL PROCEDURES

The example in Section 2 illustrated methods for creating multiple imputations and analyzing the resultant multiply-imputed data set in a special case. We now outline the methods needed for general practice.

3.1 Proper Imputation Methods

Multiple imputations ideally should be drawn according to the following general scheme. For each model being considered, the M imputations of the missing values, Y_{mis}, are M repetitions from the posterior predictive distribution of Y_{mis}, each repetition being an independent drawing of the parameters and missing values under an appropriate Bayesian model for the posited response mechanism. In practice, implicit models such as illustrated in Section 2 can often be used in place of explicit models. Both types of models are illustrated in Herzog and Rubin (1983), where repeated imputations are created using an explicit regression model and an implicit matching model, which is a modification of the Census Bureau's hot-deck.

Procedures that incorporate appropriate variability among the repetitions within a model are called *proper*, which is defined precisely in Rubin (1986a). The essential idea of proper imputation methods is to properly reflect sampling variability when creating repeated imputations under a model. For example, assume ignorable nonresponse so that respondents and nonrespondents with a common value of X have Y values only randomly different from each other. Even then, simply randomly drawing imputations for nonrespondents' from matching respondents' Y values ignores some sampling variability. This variability arises from the fact that the sampled respondents' Y values at X randomly differ from the population of Y values at X. Properly reflecting this variability leads to repeated imputation inferences that are valid under the posited response mechanism.

In the context of simple random samples and ignorable nonresponse, Rubin and Schenker (1986) study hot-deck imputation (i.e. simply randomly drawing imputed values from respondents), which is *not* proper, and a variety of proper imputation methods based on both explicit and implicit models, including a fully normal model, the Bayesian Bootstrap (Rubin, 1981), and an approximate Bayesian Bootstrap. The Approximate Bayesian Bootstrap (ABB) can be used to illustrate how an intuitive imputation method, such as the simple random hot-deck, can be modified to be proper.

3.2 Example of a Proper Imputation Method with Ignorable Nonresponse - The ABB

Consider a simple random sample of size n with n_R respondents and $n_{NR} = n - n_R$ nonrespondents. The ABB creates M ignorable repeated imputations as follows. For $\ell = 1, \dots, M$, create n possible values of Y by first drawing n values at random with replacement from the n_R observed values of Y, and second drawing the n_{NR} missing values of Y at random with replacement from those n values. The drawing of the n_{NR} missing values from a possible sample of n values rather than the observed sample of n_R values generates appropriate between imputation variability, at least in large samples, as shown by Rubin and Schenker (1986). The ABB approximates the Bayesian Bootstrap by using a scaled multinomial distribution to approximate a Dirichlet distribution.

3.3 Analysis - The Repeated Imputation Inference

The general methods for analyzing a multiply imputed data set implicitly assume proper imputation methods have been used to create the multiple imputations. As illustrated in Section 2, the repeated imputations within each model are analyzed as a collection to create one *repeated-imputation* inference as follows. Each data set completed by imputation is analyzed using the same complete-data method that would be used in the absence of nonresponse. More precisely, let $\hat{\Theta}_\ell, U_\ell, \ell = 1, \dots, M$ be M complete-data estimates and their associated variances for a parameter Θ, calculated from the M data sets completed by repeated imputations under one model for nonresponse. The final estimate of Θ is

$$\bar{\Theta}_M = \sum_{\ell=1}^{M} \hat{\Theta}_\ell / M.$$

The variability associated with this estimate has two components: the average within-imputation variance,

$$\bar{U}_M = \sum_{\ell=1}^{M} U_\ell / M,$$

and the between-imputation component,

$$B_M = \sum (\hat{\Theta}_\ell - \bar{\Theta}_M)^2 / (M-1)$$

where with vector Θ, $(\bullet)^2$ is replaced by $(\bullet)^{\mathrm{T}}(\bullet)$. The total variability associated with $\bar{\Theta}_M$ is then

$$T_M = \bar{U}_M + (1 + M^{-1}) B_M.$$

With scalar Θ, the reference distribution for interval estimates and significance tests is a t-distribution.

$$(\Theta - \bar{\Theta}_M) \, T_M^{-1/2} \sim t_\upsilon,$$

where the degrees of freedom,

$$\upsilon = (M - 1) \{ 1 + [(1 + M^{-1}) B_M / \bar{U}_M]^{-1} \}^2$$

is based on a Satterthwaite approximation (Rubin and Schenker 1986 and Rubin 1986a). The within to between ratio $\bar{U}_M / B_M$ estimates the population quantity $(1 - \gamma) / \gamma$, where γ is the fraction of information about Θ missing due to nonresponse. In the case of ignorable nonresponse with no covariates, γ equals the fraction of data values that are missing.

3.4 Significance Levels for Multicomponent Θ

For Θ with k components, significance levels for null values of Θ can be obtained from M repeated complete-data estimates, $\hat{\Theta}_\ell$, and variance-covariance matrices, U_ℓ, using multivariate analogues of the previous expressions.

A simple procedure described in Li (1985) and Rubin (1986a) that works well for M large relative to k is to let the p-value for the null value Θ_0 of Θ be Prob $\{F_{k,\upsilon} > D_M\}$ where $F_{k,\upsilon}$ is an F random variable and $D_M = (\Theta_0 - \bar{\Theta}_M) \, T_M^{-1} (\Theta_0 - \bar{\Theta}_M)^{\mathrm{T}}$ with υ defined by generalizing $B_M / \bar{U}_M$ to be the average diagonal element of $B_M \bar{U}_M^{-1}$, $\mathrm{trace}(B_M \bar{U}_M^{-1}) / k$. Better procedures are described in Rubin (1986a). Less precise p-values can be obtained directly from M repeated complete-data significance levels; also see Rubin (1986a).

4. DISCUSSION

4.1 Frequency Evaluations

Although repeated imputation inferences are most directly motivated from the Bayesian perspective, they can be shown to possess good frequency properties. In fact, the definition of proper imputation methods means that in large samples infinite-M repeated imputation inferences will be valid. Since the finite-M adjustments are derived using approximations to Bayesian posterior distributions, however, deficiencies can arise with finite M. For example, the large sample relative efficiency of $\bar{\Theta}_M$ to $\bar{\Theta}_\infty$ that is, the efficiency of the finite-M repeated imputation estimator using proper imputation methods relative to the infinite-M estimator in units of standard errors is $(1 + \gamma / M)^{-1/2}$. Even for relatively large γ, modest values of M result in estimates $\bar{\Theta}_M$ that are nearly fully efficient.

4. 2 Confidence Coverage

In large samples the confidence coverage of proper imputation methods using the *t*-reference distribution can be tabulated as a function of M, γ and the nominal level, $1 - \alpha$. Table 9 is from Rubin (1986a) and is also partially reported in Rubin and Schenker (1986) and Schenker (1985). Also included are results for single imputation, where the between component of variance is set to zero, since it cannot be estimated, and the reference distribution is the normal, since υ cannot be estimated without B_M. Even in extreme cases, two or three repeated imputations yield nearly valid confidence coverages; this is in striking contrast to using only one imputation. Even worse coverages for single imputation would have been obtained using best prediction methods, such as "fill in the mean".

Table 9

Coverage probabilities in % of interval estimates based on the *t*-reference distribution as a function of the number of proper repeated imputations, $M \geq 2$, the fraction of missing information, γ, and the nominal level, $1 - \alpha$. Also included for contrast are results based on single imputation $M = 1$, using the normal reference distribution with the between component of variability set to zero.

		γ								
$1-\alpha$	M	.1	.2	.3	.4	.5	.6	.7	.8	.9
	1	46	42	38	34	30	26	22	18	12
	2	50	50	51	51	50	50	50	50	50
50%	3	50	50	50	50	50	50	50	50	50
	5	50	50	50	50	50	50	50	50	50
	∞	50	50	50	50	50	50	50	50	50
	1	75	70	65	60	54	48	41	33	23
	2	80	80	80	79	78	77	76	76	76
80%	3	80	80	80	80	79	79	79	79	79
	5	80	80	80	80	80	80	80	80	80
	∞	80	80	80	80	80	80	80	80	80
	1	86	82	77	72	66	59	51	42	29
	2	90	90	89	88	87	86	85	84	83
90%	3	90	90	90	89	89	88	88	88	88
	5	90	90	90	90	90	90	89	89	89
	∞	90	90	90	90	90	90	90	90	90
	1	92	89	85	80	74	67	59	49	35
	2	95	95	94	93	92	91	89	88	87
95%	3	95	95	95	94	94	93	93	92	92
	5	95	95	95	95	95	94	94	94	94
	∞	95	95	95	95	95	95	95	95	95
	1	98	96	94	91	86	80	72	61	45
	2	99	99	98	98	97	96	95	93	92
99%	3	99	99	99	98	98	98	97	97	96
	5	99	99	99	99	99	99	98	98	98
	∞	99	99	99	99	99	99	99	99	99

4.3 Significance Levels

Work on accurately obtaining significance levels is at an early stage of development. Table 10 is from Rubin (1986a) and is also partially reported in Li (1985). It indicates that if $M > k$ and γ is modest, accurate tests can be obtained using D_M. Better procedures are considered by Li (1985), Rubin (1986a) and in current thesis work by T.E. Raghunathan.

Table 10

Level in % of D_M with $F_{k,\,\upsilon}$ reference distribution as a function of: nominal level, α; number of components being tested, k; number of repeated proper imputations, M; and fraction of missing information, γ.

			α = 1%				α = 5%				α = 10%			
k	M	γ =	.1	.2	.3	.5	.1	.2	.3	.5	.1	.2	.3	.5
	2		1.0	1.2	1.6	2.5	4.9	5.3	5.9	7.5	9.9	10.3	11.0	12.9
	3		1.0	1.0	1.0	1.3	4.9	4.9	5.0	5.5	9.9	9.8	10.0	10.9
2	5		1.0	1.0	1.1	1.2	5.0	5.0	5.1	5.6	10.0	10.0	10.2	10.9
	10		1.0	1.0	1.1	1.2	5.0	5.1	5.3	5.7	10.1	10.2	10.4	11.0
	25		1.0	1.0	1.0	1.0	5.0	5.0	5.0	5.0	10.0	9.9	9.9	10.0
	50		1.0	1.0	1.0	1.0	5.0	5.0	5.0	5.0	10.0	9.9	9.9	10.0
	100		1.0	1.0	1.0	1.0	5.0	5.0	5.0	5.0	10.0	10.0	10.0	10.1
	2		1.0	1.1	1.3	1.7	5.1	5.3	5.6	6.3	10.3	10.6	11.1	12.0
	3		1.0	1.0	1.0	1.0	5.1	5.2	5.3	5.7	10.2	10.5	10.9	12.3
3	5		1.0	1.0	1.1	1.3	5.0	5.2	5.4	6.2	10.1	10.3	10.8	12.2
	10		1.0	1.0	1.1	1.2	5.0	5.2	5.3	5.9	10.1	10.3	10.6	11.6
	25		1.0	1.0	1.1	1.2	5.0	5.1	5.2	5.6	10.1	10.2	10.4	10.9
	50		1.0	1.0	1.0	1.0	5.0	5.0	5.0	5.1	10.0	10.0	10.0	10.2
	100		1.0	1.0	1.0	1.0	5.0	5.0	5.1	5.1	10.0	10.0	10.1	10.2
	2		0.9	0.8	0.8	0.9	5.1	4.8	4.5	4.0	10.5	10.4	10.1	9.2
	3		1.0	1.0	1.0	0.9	5.2	5.5	5.7	6.1	10.5	11.3	12.1	14.4
5	5		1.1	1.1	1.2	1.4	5.2	5.6	6.1	7.7	10.4	11.1	12.2	15.4
	10		1.0	1.1	1.2	1.5	5.1	5.3	5.6	6.9	10.1	10.4	11.1	13.1
	25		1.0	1.0	1.1	1.3	5.0	5.2	5.3	6.0	10.1	10.3	10.6	11.5
	50		1.0	1.0	1.0	1.1	5.0	5.1	5.1	5.4	10.0	10.1	10.2	10.7
	100		1.0	1.0	1.0	1.1	5.0	5.0	5.1	5.2	10.0	10.1	10.1	10.4
	2		0.8	0.5	0.3	0.1	5.1	4.0	2.9	1.5	10.8	10.1	8.5	5.4
	3		1.1	0.9	0.6	0.3	5.6	5.9	5.7	4.9	11.3	12.7	13.8	16.2
10	5		1.1	1.2	1.3	1.4	5.4	6.3	7.4	11.0	10.7	12.4	14.8	22.7
	10		1.1	1.2	1.4	2.2	5.2	5.8	6.8	10.3	10.4	11.4	13.1	19.0
	25		1.0	1.1	1.2	1.6	5.0	5.2	5.6	7.1	10.0	10.4	11.0	13.4
	50		1.0	1.0	1.1	1.3	5.0	5.1	5.4	6.1	10.0	10.2	10.6	11.8
	100		1.0	1.0	1.1	1.2	5.0	5.2	5.3	5.8	10.1	10.2	10.5	11.3

5. CONCLUSION

In conclusion, multiple imputation is a very promising new tool for helping to handle nonresponse in surveys. Although much work remains to be done before it will become a commonplace method, many interesting theoretical and practical results suggest effort expended in its development will be well rewarded by important contributions to applied work.

ACKNOWLEDGEMENTS

This work was supported by the National Science Foundation Grant Number SES-8311428. I wish to thank M.P. Singh and a referee for their helpful editorial comments on the earlier draft of this article.

REFERENCES

HEITJAN, D.F., and RUBIN, D.B. (1986). Inference for coarse data using multiple imputation. *Proceedings of the 18th Symposium on the Interface of Computer Science and Statistics.*

HERZOG, T.N., and RUBIN, D.B. (1983). Using multiple imputations to handle nonresponse in sample surveys. In *Incomplete Data in Sample Surveys , Volume 2: Theory and Bibliography*, New York: Academic Press, 209-245.

LI, K.H. (1985). *Hypothesis Testing in Multiple Imputation - with Emphasis on Mixed-up Frequencies in Contingency Tables.* Ph.D. Thesis, Department of Statistics, University of Chicago.

RUBIN, D.B. (1976). Inference and missing data. *Biometrika*, 63, 581-592.

RUBIN, D.B. (1977). The design of a general and flexible system for handling nonresponse in sample surveys. Unpublished paper prepared for the U.S. Social Security Administration.

RUBIN, D.B. (1978). Multiple imputations in sample surveys - a phenomenological Bayesian approach to nonresponse. *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 20-34. Also in *Imputation and Editing of Faulty or Missing Survey Data*, U.S. Dept. of Commerce, 1-23.

RUBIN, D.B. (1979). Illustrating the use of multiple imputations to handle nonresponse in sample surveys. *Proceedings of the 1979 Meetings of the ISI-IASS, Manila.*

RUBIN, D.B. (1980). *Handling Nonresponse in Sample Surveys by Multiple Imputations.* U.S. Dept. of Commerce, Bureau of the Census Monograph.

RUBIN, D.B. (1981). The Bayesian bootstrap. *Annals of Statistics*, 9, 130-134.

RUBIN, D.B. (1986a). *Multiple Imputation for Nonresponse in Surveys.* New York: John Wiley & Sons.

RUBIN, D.B. (1986b). Statistical matching using file concatenation with adjusted weights and multiple imputations. *Journal of Business and Economic Statistics*, 87-94.

RUBIN, D.B., and SCHENKER, N. (1986). Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *Journal of the American Statistical Association*, 81, 366-374.

SCHENKER, N. (1985). *Multiple Imputation for Interval Estimation from Surveys with Ignorable Nonresponse.* Ph.D Thesis, Department of Statistics, University of Chicago.

Survey Methodology, June 1986
Vol. 12, No. 1, pp. 49–60
Statistics Canada

Imputation Options in a Generalized Edit and Imputation System

P. GILES and C. PATRICK[1]

ABSTRACT

Statistics Canada has undertaken a project to develop a generalized edit and imputation system, the intent of which is to meet the processing requirements of most of its surveys. The various approaches to imputation for item non-response, which have been proposed, will be discussed. Important issues related to the implementation of these proposals into a generalized setting will also be addressed.

KEY WORDS: Modularity; Prototyping; Donor imputation; Regression models.

1. GENERALIZED SYSTEMS

Due to resource constraints imposed on surveys in recent years, especially in the area of development, the idea of generalized software has received considerable support. By generalized software, it is meant a set of computer programs, tied together into one system, which allows the user to select a suitable approach to the problem, from among several alternatives. For example, a user has a data file from which a sample of records is to be selected. A generalized sample selection system would offer the user the choice of various sampling schemes such as simple random or unequal probability sampling (with or without replacement), systematic, stratified, or cluster sampling.

A genuinely generalized system is, almost by definition, a complex object. The concept of modularity is an important device for the reduction of complexity, by allowing the overall task to be split into a number of simpler sub-tasks. Each of the sub-tasks, or functions, is performed sequentially. The user is offered several alternatives for each sub-task. Therefore, not only is the overall task able to be split into smaller, more manageable components, but also each sub-task can be performed in more than one way.

Figure 1 demonstrates how the edit and imputation task can be split into three sub-tasks. These three sub-tasks are editing, identification of fields to impute, and imputation. Each of the boxes, or modules, in a row employ different approaches to that particular sub-task. For example, C1 could employ some type of donor imputation, C2 could employ the imputation of a mean value, and so on. The user would select one of the modules from each of rows A, B, and C.

It should be noted that this representation of a generalized system for edit and imputation is not the only possibility. In fact, the actual proposal for a developmental project actually contains five sub-tasks, as opposed to the three exemplified here. This representation is given only for simplicity.

Each sub-task, or row in the example, would be a clearly defined function. The input files required, and the output files created, must have prespecified formats. This allows the user to concentrate on the choice of modules in each row, knowing that the system can handle the "housekeeping". (This refers to file handling and other mundane details about which the user would prefer not to worry.) Even though the system may accept all possible combinations of choices of modules, some combinations may not be desirable or even logically valid. It is usually the responsibility of the user to ensure that the pieces fit together.

[1] Philip Giles and Charles Patrick, Business Survey Methods Division, Statistics Canada, Tunney's Pasture, Ottawa, Ontario, Canada, K1A 0T6.

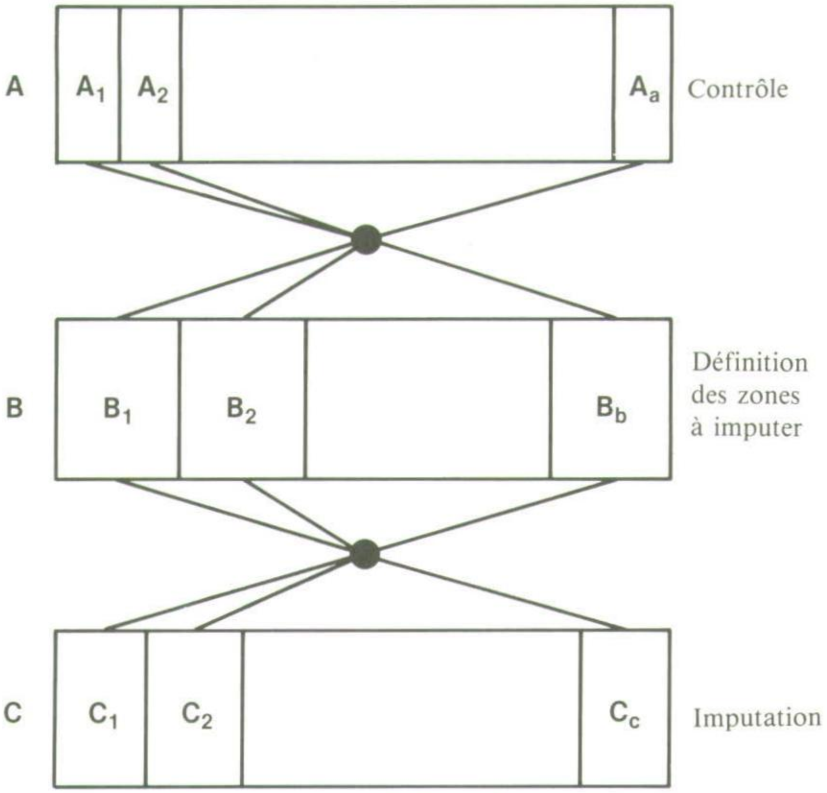


Figure 1. Generalized System Example - Edit and Imputation

A modular approach to the development of a processing system has an important consequence. From a certain point of view, the system is always "under development", since additional modules embodying new approaches and enhancements to "old" modules, can always, and in principle should, be added. This open-endedness also means that the very important concept of prototyping can be easily accommodated. Prototyping is an approach wherein a subset of modules are developed initially. The system would then be available to some of the users. Subsequently, additional modules are developed to meet the requirements of additional users. Thus, the key advantage of prototyping and modularity is that piecemeal improvements to the system are deliberately anticipated and more easily accomplished. A minimal, but imperative, requirement of such an approach is that a framework (as shown in Figure 1) and a host environment (format of data files and programming language) must be carefully defined and specified very early in the overall developmental process.

In addition to the foregoing developmental advantages, others may be gained after the system is in place. The user has considerable flexibility in choosing the path to proceed. If several alternatives seem equally viable, one can use historical data to choose among them, by testing the various alternatives prior to data collection. This can be accomplished without an undue expenditure of effort. Once the generalized system is developed there is a reduction in resource requirements for each of its users, with a corresponding reduction in elapsed time to implementation.

There are some disadvantages to following a generalized route. The utilization of generalized software in a production environment may be less efficient than the corresponding custom-designed system. The initial resource requirement will be higher for a generalized system as compared to a customized system. However, this higher cost must be assessed against the substantially higher costs of repeated custom-designed implementations. Nor is it reasonable to expect a generalized system to satisfy every specific requirement. In this situation, the user has two options. The first option is to develop a user-written module. This would not require the same degree of effort as a complete customization. However, if this occurs frequently, the purpose of the generalized system is defeated. The second option is for the user to modify the specifications in order to fit the generalized system mold. If the system has been well-designed, any required compromise should not result in a serious deterioration of data quality. It should also be recognized that compromises to the original specifications are usually and frequently required during the development of a customized system.

2. BACKGROUND TO IMPUTATION

The term "imputation", in this document, refers to a certain class of procedures for handling non-response. The input is a data captured file. The imputation procedure creates a file with individually "clean" records; a "clean" record being one which has no missing values and which satisfies all the specified edits. In order to create a clean record, a value must be estimated for each missing value.

The edits, specified by the user, are logical constraints on the values that each variable can assume. The set of edits, as a whole, define the acceptance region for the data. For categorical data, an edit is specified as a set of combinations of acceptable data values. The acceptance region can be represented as a set of lattice points in N-space. For numerical data, an edit is a linear equality or inequality. The requirement of linearity is not unduly restrictive, since a non-linear edit can be made linear by either algebraic manipulation or by adding supplementary variables, which are suitably defined non-linear functions of survey variables. The acceptance region for numerical data is a set of convex regions in N-space. The reason that there may be more than one convex region is that conditional edits are possible. Conditional edits are edits which pertain to only a subset of records. For example, the edits which are relevant to a particular record may be very different, depending on whether the variable Sex is recorded as Male or Female.

If one or more edits fail for a particular record, it may not be obvious which variable(s) is/are in error, and, by implication, to be imputed. For example, a failed edit is $A + B \leq C$. The data record under consideration has data values $A = 10$, $B = 5$, $C = 12$. There are seven combinations of variables to change which would result in a clean record. These are A, B, C, A & B, A & C, B & C, and, A & B & C. Without any other information or decision rule, each of these choices is equally valid. The problem of how to decide which variable(s) to impute will not be discussed in this document. It will be assumed that, for each record, the variable(s) to impute have been identified. No distinction is made between variables to impute due to missing values and variables to impute due to edit failures.

3. PROPOSED IMPUTATION TECHNIQUES

This section is comprised of four sub-sections, which define all the proposed imputation techniques. These are Deterministic Imputation, Donor Imputation, Regression Models, and Other Imputation Estimators. The use of regression models and the section on other estimators is restricted to numerical data. The other two sub-sections apply both to numerical and categorical data.

Almost all imputation techniques can be formulated in a prediction framework, described by Rubin (1976), as follows. A joint distribution, $f(X_1, ..., X_N)$, summarizing the statistical behavior of the population of complete records is specified. This can be done whether the individual variables are quantitative or qualitative. Without loss of generality, for a record i which requires imputation, the N variables can be partitioned into $X_1, \ldots, X_{m_i}$, which require imputation, and $X_{m_{i+1}}, \ldots, X_N$, which do not require imputation. A conditional distribution $f(X_1, \ldots, X_{m_i} | x_{m_{i+1}}, \ldots, x_N)$ can be derived. Imputed values, $y_1, \ldots, y_{m_i}$, are chosen for $X_1, \ldots, X_{m_i}$ from the set.

$$\{y_1, \ldots, y_{m_i} : f(y_1, \ldots, y_{m_i} | x_{m_{i+1}}, \ldots, x_N) > 0\}$$

Various selection mechanisms can be employed. However, as stated above, some of these are relevant only to certain types of data variables.

It should be noted that there is nothing new or radically different in these proposals. They are based on work done previously, both in Statistics Canada and outside. The discussion on donor imputation is based on Fellegi and Holt (1976). The model-based approach to determining a value to impute is discussed by Little (1982). Other related papers of interest are Sande (1976), Kalton and Kasprzyk (1982), and Kalton and Kish (1981).

3.1 Deterministic Imputation

The first type of imputation is called deterministic imputation. This occurs when only one value can satisfy the edits. If more than one variable is to be imputed for a particular record, a deterministic solution may be possible for some, or all, variables. The check for determinacy should be done before proceeding to other imputation procedures.

Deterministic imputation may arise in very simple, and easily detectable situations. For example, suppose that there is an edit $A + B = 10$. The record under consideration requires A to be imputed and B has value 6. Obviously, $A = 4$ is the only value which will satisfy the edit. Another example demonstrates this for categorical variables. Suppose an edit is stated as "If the relationship to the household reference person is wife, then sex must be female." If the reference record has "wife" as the value of "relationship to the household reference person", and the variable "Sex" requires imputation, then the only valid imputed value is Sex = Female.

However, a typical survey situation will have several edits, rather than just one. This may mean that an existing deterministic solution may not be apparent. The procedure for checking for deterministic imputation is to find the reduced acceptance region defined by the active edits and the "good" data values. The active edits are defined as the subset of edits in which the variable(s) to be imputed are participant. This can also be expressed in the notation of the prediction framework given at the beginning of Section 3. The conditional distribution $f(X_1, \ldots, X_{m_i} | x_{m_{i+1}}, \ldots, x_N)$ will specify a unique value for some or all of the variables $X_1, \ldots, X_{m_i}$.

An example serves to illustrate the procedure for identifying deterministic imputation. Note that while the example is written with numerical variables, an analogous situation exists for categorical variables.

There are three edits:

$$\begin{aligned} X + Y &\le 16, \\ Y + Z &\le 4, \\ X - 3Z &\le 8. \end{aligned}$$

The reference record has values

$$X = 11 \textit{ and } Y = 3.$$

The variable Z is to be imputed.

It is not apparent whether or not a determinancy exists. This first step is to consider all active edits. In the example, there are two edits which contain the variable Z.

$$Y + Z \leq 4,$$

$$X - 3Z \leq 8.$$

Next, the known values of X and Y are inserted into these edits, and the reduced acceptance region is determined.

$$3 + Z \leq 4,$$

$$11 - 3Z \leq 8.$$

Solving these inequalities gives the following solution.

$$Z \leq 1,$$

$$Z \geq 1.$$

It is now obvious that $Z = 1$ is the only possible valid imputed value.

In most "real-life" situations, the incidence of deterministic imputation should be low. The contrary would indicate that the edits are more restrictive than necessary or desirable, and should lead to a re-examination of the edit specifications. However, in the sense that it reduces the imputation problem, deterministic imputation is a useful first step.

3.2 Donor Imputation

Donor imputation is a method which pairs each record requiring imputation, the candidate record, with one record from a defined donor population. In order to determine the value to impute, one approach is to directly copy the value from the donor record onto the candidate record. For numerical variables, if suitable auxiliary information is available, more complex methods may be used to determine the value to be imputed. Further discussion on imputation estimators for donor imputation is given in Section 3.3.

Usually, the donor population is defined as all records in the current survey which have no variables to be imputed. Referring to the prediction framework described at the beginning of Section 3, then this situation implies that $f(X_1, \ldots, X_N)$ is the empirical probability function. However, other approaches to defining the donor population are possible. For the remainder of the discussion on donor imputation, it will simply be assumed that a donor population has been defined.

Donor-candidate pairs are formed using matching variables. Matching variables are defined as variables which do not require imputation on the candidate record and are "highly correlated" with the variable(s) requiring imputation. Preferably, the matching variables should also have "low correlation" with each other. Two matching variables with "high correlation" would have the same discriminatory power as one alone, but would have the effect of doubling the weight given to one alone.

For categorical variables, a donor record is chosen, using some random process, from amongst potential donor records having the same values for the matching variables to those for the candidate record. Since numerical variables can assume many more values than categorical variables, it is very unlikely that an exact match on matching variables would be possible. Therefore, for numerical data, a distance function is used to define similarity. This distance function is a function of the matching variables on the candidate and potential donor records. The chosen donor is the record with minimum distance from the candidate record. Usually, the matching variables are transformed for the purpose of distance calculations in order to remove the effect of scale in which the variable is recorded. For example, it would be quite worrisome to the user if the formation of the donor-candidate pairs was dependent on whether a length variable was recorded in metres or feet. The proposed transformations and distance functions are discussed below.

The matching variables to be used can be a user input, or determined by an automated procedure. Usually, due to time considerations, all decisions must be made prior to data collection. Therefore, if the determination of matching variables is a user input, the user must specify the matching variables for each pattern of variables to be imputed. If there are N variables on the file, the user must make $(2^{**}N)-2$ input specifications. Obviously, the value of N does not have to be very large in order for this approach to become unmanageable. In order to reduce this number, the matching variables may be specified by stratum. All candidate records in a particular stratum would use the same matching variables. In this situation, it is possible (depending on how careful the user is in specifying the matching variables) that a particular candidate record may have a matching variable which requires imputation. All in all, the user who inputs the matching variable specifications, is warned that this decision may result in a large increase in the work required.

One possible approach for automatically determining the matching variables is proposed. This procedure can be used, analogously, for both categorical and numerical data. Basically, the procedure is as follows. At a minimum, the set of matching variables must contain the variables sharing in the edit rules with the variables to be imputed. As defined earlier, these are the active edits. This approach seems intuitively reasonable, since it is desirable that the matching variables be correlated with the variable(s) to be imputed. The variables in the active edits constrain the range of possible values to be imputed. This implies a type of dependence, or correlation structure.

The use of this matching procedure, together with direct transcription, has one important consequence for categorical variables. All imputed values are guaranteed to pass the edits. This is very important as it is required in order to create a clean record. Without this guarantee, the user must re-edit the records, and possibly adopt a secondary imputation procedure. For numerical data, similarity as defined by a distance function does not guarantee this outcome. However, the closer the distance between the donor and candidate record is to zero, the greater the probability that the imputed values will satisfy the edits.

The determination of matching variables using this automated procedure can be illustrated by an example.

There are five edits:

I. $A + B \leq \alpha_1$,

II. $B - E \leq \alpha_2$,

III. $C + 2D + 3E \leq \alpha_3$,

IV. $A + C + D \leq \alpha_4$,

V. $A - 2B + C \leq \alpha_5$.

There are five survey variables A, B, C, D, E and α_1, α_2, α_3, α_4, α_5 are known scalars. The candidate record under consideration has variable B only to be imputed.

The first step is to identify the active edits. In this example, there are three active edits. These are edits I, II, and V.

The second step is to determine the active variables. The active variables are defined as all variables which are contained in at least one of the active edits. In the example, there are four active variables: A, B, C, E. Note that, by definition, the active variables contain all variables to be imputed.

The third step is to determine the matching variables, as those active variables which do not require imputation. For this example, the matching variable are A, C, E.

In addition to the determination of matching variables, donor imputation for numerical data requires the choice of a data transformation and the choice of a distance function.

Two types of data transforms are proposed. For both of these, each variable is to be transformed independently. The two proposed transformations are a rank value transform and a location-scale transform.

For the rank value transform, the values for each variable are sorted. Then, the rank values are divided by a suitable constant such that all values are in the range from zero to one. The transformed values are distributed uniformly in that range.

The location-scale transform is of the form,

$$y^T = \frac{1}{b}(y - a),$$

where y^T is the transformed value,

y is the original data value,

a, b are user-specified parameters.

Two popular choices for these constants are, one, that a be the sample mean and b be the sample standard deviation, and, two, that a be the sample minimum and b be the range of values in the sample. Other options may be possible.

In choosing a data transform, there are robustness and outlier considerations. The rank value transform is very robust against changes in data values, and pulls outliers closer to the other data values. This may or may not be desirable. There are no bounds on the transformed values, using the location-scale transform with the mean and standard deviation. These parameters are also sensitive to outliers. The choice of the minimum value and range would restrict the transformed values between zero and one. However, these are very sensitive to extreme values. One very large value could cause all of the transformed values, except one, to be virtually zero.

In considering the choice of distance function, a family of distance functions are proposed. These are the weighted $\mathcal{L}^p$ norms, where p is a user-specified constant. The general form of these functions is

$$D(X, Y) = \left[\sum_{k=1}^{r} w_k |x_k - y_k|^p \right]^{1/p},$$

where x_k, y_k are the r matching variables on the two records,

w_k are user-specified weights,

p is a user-specified constant.

The weights are used if one wishes some of the matching variables to contribute more to the distance calculation than others. The default values are for all weights to be set to one.

Three particular choices of a value for p are of special interest, $p = 1$, $p = 2$, and $p = \infty$. For $p = 1$, this function calculates the city block distance. For $p = 2$, the Euclidean distance is calculated. The limiting case of this function, when $p = \infty$, yields the minimax distance. For this choice of p, the function is written as

$$D(X, Y) = \underset{1 \leq k \leq r}{\text{Max}} \; [w_k |x_k - y_k|].$$

One final point to be discussed about donor imputation is the concept of a "penalty" for donor usage. This penalty would reduce the number of times that a particular donor record is used. For donor imputation of categorical data, a donor record is selected from the donor population without replacement. This strategy has to be modified slightly if the size of the candidate population is greater than the size of the donor population.

For numerical data, the distance function is modified by increasing the distance calculation according to the number of times a particular donor is used. One possible approach is to use $D'(X, Y)$ to calculate distances, where

$$D'(X, Y) = D(X, Y) \times (1 + ud),$$

where u is the "penalty" imposed by the user,

d is the number of times that donor record has been chosen.

An implication of the imposition of a penalty on the distance function, is that the choice of a donor record for each candidate record is now dependent on the order of the candidate records.

3.3 Regression Models

This section discusses imputation estimators which result from the use of regression models. For this discussion, only two models are used. These are:

MODEL I : $y_i = \alpha + \epsilon_i$, $\quad \text{Var}(\epsilon_i) = \sigma^2$,

MODEL II: $y_i = \beta x_i + \epsilon_i$, $\quad \text{Var}(\epsilon_i) = \sigma^2 x_i$.

Note that these models are special cases of the more general formulation of regression models, which has the form

$$\underset{\sim}{y} = \underset{\sim}{X}\underset{\sim}{\beta} + \underset{\sim}{\epsilon},$$

$$\text{where } E(\underset{\sim}{\epsilon}) = \underset{\sim}{0}, \; V(\underset{\sim}{\epsilon}) = \underset{\sim}{V}$$

Model II is used when auxiliary data is available. Otherwise Model I is used. Both models have one parameter to be estimated. Using least-squares, the parameter estimates are:

$$\hat{\alpha} = \bar{y},$$

$$\hat{\beta} = \frac{\bar{y}}{\bar{x}}.$$

Before stating the various proposed estimators, some notation will be introduced.

Let t be the subscript for time t, the present survey,

y_{it} be the variable under study for unit i and time t; this is the value to be imputed for candidate records,

x_{it} be the auxiliary variable (correlated with Y) for unit i and time t,

R be the subscript for all non-respondents at time t (i.e., y_{it} is known),

NR be the subscript for all non-respondents at time t (i.e., y_{it} is to be imputed),

C, D be superscripts which denote either a candidate or donor record, whenever the distinction is required.

Several explanatory notes are required along with the notation. First, R and NR are as defined in the current survey, regardless of the reporting history of each record. Second, the values for the variables $y_{i(t-1)}$, x_{it}, $x_{i(t-1)}$ may themselves have been imputed. The only restriction is that they are not missing. Third, the notation does not include the concept of imputation classes. Imputation classes are essentially post-strata, in that they define sets of records which are judged homogeneous within, and heterogeneous between groups. However, both the notation and the imputation estimators are readily extendible to include imputation classes.

Thus, estimators can be classified according to:

(i) the choice of model, I or II,

(ii) the imputation group, and,

(iii) the variables in the regression used to estimate the parameter.

The data on the records in the specified imputation group are precisely the data used to estimate the parameter(s) in the model. This concept allows considerable flexibility. For example, it could allow the preclusion of outliers from the calculation of the parameter estimate. After the parameter is estimated, it is used for prediction purposes to determine the imputed value. According to the notation, Y_t is always the variable predicted.

Based on the two models, eight imputation estimators are proposed. Even though there are eight proposed estimators, this list can be augmented in the future. These additional estimators could be derived, for example, by choosing other models, possibly incorporating more variables.

Scanning the list of eight, one can see that these are the familiar imputation estimators that have been used traditionally.

Estimator 1: The value from the previous survey for the same unit is imputed. $y_{i(t-1)}$

Estimator 2: The mean value from the previous survey is imputed. $\bar{y}_{(t-1)}$

Estimator 3: The mean value of all respondents to the current survey is imputed. $\bar{y}_{tR}$

Estimator 4: The value is copied directly from the donor record to the candidate record, y_{it}^{D}

Estimator 5: A ratio estimate, using values from the current survey is imputed.

$$\frac{\bar{y}_{tR}}{\bar{x}_{tR}} x_{it}$$

Estimator 6: A ratio estimate, based on values on the donor and candidate records is imputed.

$$\frac{y_{it}^{D}}{x_{it}^{D}} x_{it}$$

Estimator 7: The value from the previous survey for the same unit, with a trend adjustment calculated from an auxiliary variable, is imputed.

$$\frac{y_{i(t-1)}}{x_{i(t-1)}} x_{it}$$

Estimator 8: The value from the previous survey for the same unit, with a trend adjustment calculated from the change in reported values to variable Y, is imputed.

$$\frac{\bar{y}_{tR}}{\bar{y}_{(t-1)R}} y_{i(t-1)}$$

It is interesting to contrast the difference in estimators when one fixes all classification items but one. For example, the difference between estimators one and two is due only to the difference in choice of imputation group, as is also the case for estimators three and four, and, estimators five and six. The difference between estimators one and seven is due only to the choice of model. The same is true for estimators three and five, and, estimators four and six. It should also be noted that estimators four and six are those used in donor imputation, which were discussed in Section 3.2.

3.4 Other Imputation Estimators

The choice of imputation techniques is dependent upon the assumptions made by the user about the non-responding population. When using donor imputation, one assumes that there are some respondents which are similar to each non- respondent. If one imputes the mean from the current survey, the assumption is that the mean value of the respondents is the same as the mean value of the non- respondents. Similarly, one can go through all the estimators and list the implied assumptions. The first estimator proposed in this section tries to ease the somewhat restrictive (and usually untrue) assumptions required in the previous section. It pays for this by being more complex. It is called the chain-link estimator, given by Madow and Madow (1978).

The derivation of this estimator is described. First, by assuming that the rate of change (trend) of the non-responding and responding populations are the same as observed in the previous survey, the population mean of the variable Y for the non-responding population in the current survey is estimated.

$$\bar{y}_{NRt} = \frac{\bar{y}_{NR(t-1)}}{\bar{y}_{R(t-1)}} \bar{y}_{Rt}.$$

One then determines the imputed value according to the auxiliary variable.

$$y_{it} = \frac{\bar{y}_{NRt}}{\bar{x}_{NRt}} x_{it}$$

$$= \frac{\bar{y}_{NR(t-1)}}{\bar{x}_{NRt}} \quad \frac{\bar{y}_{RT}}{\bar{y}_{R(t-1)}} x_{it}$$

Note that this amounts to a more complex application of the Regression Model approach discussed in Section 3.3. First, temporarily impute $y_{it} = \bar{y}_{NRt}$, as given above. Then, use Model II, and define the imputation group as being all non- responding records to the present survey for variable Y. The response variable is Y_t. The regressor variable is X_t. The resulting estimator is as given above.

The second estimator proposed in this section can be used when one has data on variable Y for several previous surveys. It does not use auxiliary variables, or data from other records. The behavior of each non-respondent is considered independently of others. This method is called exponential smoothing. It is a standard econometric forecasting technique. There is one user-specified parameter. It allows the flexibility of changing the relative contribution of the various data values. Algebraically, the estimator is given by

$$y_{it} = \frac{1-A}{1-A^t} \sum_{r=0}^{t-1} A^r \, y_{i(t-r-1)},$$

where $0 < A < 1$, is prespecified.

The closer A is to zero, the more weight is given to recent data. If $t = 1$, this reduces to imputing the value for the previous survey.

4. PAST WORK IN STATISTICS CANADA

Statistics Canada has made efforts in the past to develop a generalized edit and imputation system. Two of these will be highlighted, as they form the basis for the current proposal. These are the CAN-EDIT system and the Numerical Edit and Imputation System (NEIS).

4.1 CAN-EDIT

CAN-EDIT is itself, not a completely generalized system. However, the methodology that it employed is. The system is based on the work by Fellegi and Holt (1976) on imputation for categorical data. It was developed for processing the 1976 and 1981 Canadian Censuses of Population and Housing.

CAN-EDIT adopted a donor imputation approach. The matching variables were determined automatically, using the procedure described in Section 3.2. The CAN-EDIT system employed what it called primary and secondary imputation. If a candidate record could not be imputed in primary imputation, it was sent to secondary imputation.

In primary imputation, all imputed values are taken from the same donor. The matching variables were determined based on all variables to be imputed. A record would fail primary imputation if no donor record had identical values on the matching variables.

In secondary imputation, each of the variables to be imputed are treated independently and sequentially. The procedure for determining the matching variables is the same. However, by considering only one variable at a time, the number of matching variables will, in general, be less than under primary imputation. (There cannot be more, but the number may be the same). This implies that the potential donor population is larger. There are a few disadvantages to secondary imputation, as compared to primary imputation. First, it is possible to choose, as a matching variable, a variable which is to be imputed. There is no value to match on. Second, this approach does not make use of the joint distributions of the variables. The imputed values for two variables may satisfy the edits, each may be a very valid value, but which may occur in the population in combination only rarely.

4.2 Numerical Edit and Imputation System (NEIS)

The NEIS is a first prototype of a generalized E&I system for numerical data. It was written as a set of modules in the PSTAT statistical package. Subsequent prototypes have never been developed. This system was developed by Gordon Sande (1979). It is felt that the methodology is very sound, and should be incorporated in a new system. However, PSTAT may no longer be a suitable software environment. The NEIS was used, in a production environment, by the 1981 Farm Energy Use Survey. The methodology was employed in the development of the 1981 Census of Agriculture processing system.

The NEIS, similar to CAN-EDIT, used a donor imputation approach with matching variables determined automatically using the procedure described in Section 3.2. However, as explained in that section, the determination of matching variables in this fashion for numerical data will not always result in the imputation procedure producing a clean record. The strategy adopted to reduce this problem is to select the closest r donors. If the closest donor does not impute values which satisfy the edits, then the next closest donor is considered, and so on.

The NEIS gave the user no choice of transformation or distance function. It used the rank value transformation and the weighted $\mathcal{L}^\infty$ norm for distance calculations.

5. CONCLUSION

The proposals presented would allow considerable choice to a user of a generalized edit and imputation system. As mentioned, it does not close the door on additional approaches. However, it is felt that a system which is developed with these components would be suitable for a large number of users. It has been the experience of the authours that the ultimate power and usefulness of such a system is not apparent until one starts to use it. As testing proceeds, it becomes clear that there are more capabilities and extensions than first appear.

REFERENCES

FELLEGI, I.P., and, HOLT, D. (1976). A systematic approach to automatic edit and imputation. *Journal of the American Statistical Association*, 71, 17-35.

KALTON, G., and, KASPRZYK, D. (1982). Imputing for missing survey responses. *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 22-31.

KALTON, G., and KISH, L. (1981). Two efficient random imputation procedures. *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 146-151.

LITTLE, R.J.A. (1982). Models for non-response in sample surveys. *Journal of the American Statistical Association*, 77, 237-250.

MADOW, L.H., and MADOW, W.G. (1978). On link relative estimators. *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 534-539.

RUBIN, D.B. (1976). Inference and missing data. *Biometrika*, 63, 581-592.

SANDE, G. (1976). Searching for numerically matched records. Technical Report, Business Survey Methods Division, Statistics Canada.

SANDE, G. (1979). *The Numerical Edit and Imputation Subsystem for PSTAT — A User's Guide.* Research and General Systems Subdivision, Statistics Canada.

Survey Methodology, June 1986
Vol. 12, No. 1, pp. 61–72
Statistics Canada

The Maximum Likelihood Method for Non-Response in Sample Surveys

M.S. SRIVASTAVA and E.M. CARTER[1]

ABSTRACT

The analysis of survey data becomes difficult in the presence of incomplete responses. By the use of the maximum likelihood method, estimators for the parameters of interest and test statistics can be generated. In this paper the maximum likelihood estimators are given for the case where the data is considered missing at random. A method for imputing the missing values is considered along with the problem of estimating the change points in the mean. Possible extensions of the results to structured covariances and to non-randomly incomplete data are also proposed.

KEY WORDS: Incomplete response; Missing at random; Maximum likelihood method; Imputation.

1. INTRODUCTION

Examples of non-response in sample surveys are in abundance. Various attempts with varying degrees of success have been made in the literature to solve this problem. The success of a particular procedure is dependent on the complexity of the problem. For example, when the data is not missing at random, the problem is far from being solved. The recent attempts by Heckman (1976) and Greenlees *et al.* (1982) among others, are highly sensitive to model misspecification. Similarly the hot-deck method has been severely criticized in the literature. However, when the sample size is large, the hot-deck method and a carefully designed regression method yield similar results in imputing the non-response income in Current Population Survey (CPS). See David, Little, Samuhel and Triest (1986).

The regression method is based on the assumption that the non-response is random, but unlike the hot-deck method does not require complete information from a previous census, which in a majority of cases is non-existent. Thus it appears that a carefully designed regression method may be of great help.

In this paper, the situation when the non-response is random is considered. Random non-response arises naturally in many situations. For example, in successive sampling, the sampling starts with a certain number of people from whom certain observations are obtained for a period of time. At the end of this period, some people are dropped from the survey and new people are added. The survey continues in this manner until completion. Examples of this nature are considered by Woolson, Leeper and Clarke (1978) and Woolson and Leeper (1980).

Even when the non-response in not random, the non-random nature of the incomplete data may be accounted for, by using a sufficient number of explanatory variables in the regression model and employing some of the techniques used in the hot-deck method as was done in David *et al.* (1986) for a univariate model. For example, in Section 2.5 a method for imputing the missing values is given.

[1] M.S. Srivastava, Department of Statistics, University of Toronto, Toronto, Ontario, Canada M5S 1A1, and E.M. Carter, Department of Mathematics and Statistics, University of Guelph, Guelph, Ontario, Canada N1G 2W1.

In the course of developing these results, a method will be derived for checking if there have been any changes over time in the response patterns. The models used can also be modified to include error variance-covariance matrices that are structured by the imposition of a time series to the reponse variables. In this paper it is assumed that the data are normally distributed from a simple random sampling scheme and that the data are missing at random. If the normality assumptions is dropped then the estimators can no longer be considered maximum likelihood estimators but may still be considered as good heuristic estimators.

In the next section, the form of the model will be described for the one sample problem.

2. THE ONE SAMPLE PROBLEM

2.1 The Model

The bivariate incomplete data problem is considered first to introduce the general procedure that follows. Let $\underline{y} = (y_1, y_2)'$ be a bivariate random vector with mean vector $\underline{\mu}$ and covariance matrix Σ. Without loss of generality, the missing data in the bivariate situation can be described as follows:

$$\begin{array}{l} y_{11}, \ldots, y_{1n_1}, y_{1,n_1+1}, \ldots, y_{1,n_1+n_2}, - - - - - - - - - - - - - - \\ y_{21}, \quad , y_{2n_1}, - - - - - - - - - - - - y_{2,n_1+n_2+1}, \ldots, y_{2,n_1+n_2+n_3} \end{array} \qquad (1)$$

That is, there are n_1 pairs of observations, n_2 observations on y_1 with the corresponding observation on y_2 missing, and n_3 observations on y_2 with the corresponding observation on y_1 missing. Thus $N = n_1 + n_2 + n_3$ observations are grouped into three subsets. If the complete data set were to be represented as $\underline{y}_1, \ldots, \underline{y}_N$, then the actual observed responses can be defined as

$$\underline{z}_{1j} = B_1\underline{y}_j = \underline{y}_j \ , \text{ for } j = 1, \ldots, n_1,$$

$$z_{2j} = B_2\underline{y}_j = y_{1j} \, , \text{ for } j = n_1 + 1, \ldots, n_1 + n_2,$$

and

$$z_{3j} = B_3\underline{y}_j = y_{2j} \, , \text{ for } j = n_1 + n_2 + 1, \ldots, n_1 + n_2 + n_3,$$

where $B_1 = I_2$, the identity matrix, $B_2 = (1\ 0)$ and $B_3 = (0\ 1)$.

For the general multivariate one sample problem, there will be K subsets of the data containing $n_1, \ldots, n_K$ observations. Note that the maximum number of groups is $2^p - 1$. Also the total sample size is $N = n_1 + \ldots + n_K$. If the k-th subset contains p_k characteristics $i_1, \ldots, i_{p_k}$, then the matrix B_k would be a $p_k \times p$ matrix with a one in the (s, i_s) position for $s = 1, \ldots, p_k$ and zero elsewhere. With this notation the observed vectors of responses can be written as:

$$\underline{z}_{kj} = B_k\underline{y}_{kj}, \ j = 1, \ldots, n_k, \ k = 1, \ldots, K.$$

Hence,

$$E(\underline{z}_{kj}) = B_k \, \underline{\mu},$$

and

$$\text{cov}(\underline{z}_{kj}) = B_k\Sigma B_k' \, , \ j = 1, \ldots, n_k \text{ and } k = 1, \ldots, K.$$

Example 1: (Data)

Wei and Lachin (1984) give the cholesterol levels for a treatment group studied at times 0, 6, 12, 20 and 24 months. For reasons not pertaining to the response variable, certain observations were incomplete. The data can be grouped into $K = 8$ subsets. For the first group of complete data the sample mean and covariance matrix, based on 36 observations, were:

$$\underline{\bar{z}}_1 = \begin{pmatrix} 226.6 \\ 249.6 \\ 252.6 \\ 253.1 \\ 256.7 \end{pmatrix}, \quad S_1 = \begin{pmatrix} 1964 & 1301 & 1151 & 960 & 1008 \\ 1301 & 1715 & 1109 & 1023 & 1199 \\ 1151 & 1109 & 1554 & 697 & 1266 \\ 960 & 1023 & 697 & 1148 & 667 \\ 1008 & 1199 & 1266 & 667 & 2546 \end{pmatrix}.$$

The data for each of the other subsets is given in Table 1 with the imputed values in parenthesis.

The matrices that define the model for the observed values are:

$$B_1 = I_5, \quad B_2 = \begin{pmatrix} 1&0&0&0&0 \\ 0&1&0&0&0 \\ 0&0&1&0&0 \\ 0&0&0&1&0 \end{pmatrix}, \quad B_3 = \begin{pmatrix} 1&0&0&0&0 \\ 0&1&0&0&0 \\ 0&0&0&1&0 \\ 0&0&0&0&1 \end{pmatrix},$$

$$B_4 = \begin{pmatrix} 1&0&0&0&0 \\ 0&1&0&0&0 \\ 0&0&1&0&0 \end{pmatrix}, \quad B_5 = \begin{pmatrix} 1&0&0&0&0 \\ 0&1&0&0&0 \\ 0&0&0&0&1 \end{pmatrix}, \quad B_6 = \begin{pmatrix} 1&0&0&0&0 \\ 0&1&0&0&0 \end{pmatrix},$$

$$B_7 = (1\ 0\ 0\ 0\ 0), \qquad B_8 = (0\ 1\ 0\ 0\ 0).$$

Now that the model is defined, estimation of the parameters and the imputation of the missing data can be performed.

2.2 Estimation of the Population Mean Vector and Covariance Matrix.

For each of the K subsets define the sample mean as

$$\underline{\bar{z}}_k = (n_k)^{-1} \sum_{j=1}^{n_k} \underline{z}_{kj}.$$

Table 1

Observed Cholesterol Levels and Imputed Values

	Variable	1	2	3	4	5
Subset 2:	$n_2 = 7$	224	273	242	274	(231)
		231	252	267	299	(233)
		268	296	314	330	(303)
		284	288	268	261	(300)
		217	231	276	257	(238)
		209	200	269	233	(323)
		200	261	264	300	(279)
Subset 3:	$n_3 = 1$	193	189	(257)	232	211
Subset 4:	$n_4 = 12$	201	219	220	(231)	(172)
		202	186	253	(245)	(328)
		209	207	167	(208)	(194)
		212	253	225	(157)	(194)
		276	326	304	(300)	(376)
		163	179	199	(211)	(224)
		239	243	265	(238)	(246)
		204	203	198	(234)	(171)
		247	211	225	(224)	(215)
		195	250	272	(265)	(231)
		228	228	279	(276)	(259)
		290	264	260	(249)	(325)
Subset 5:	$n_5 = 1$	227	247	(215)	(267)	220
Subset 6:	$n_6 = 5$	250	269	(327)	(250)	(295)
		175	214	(250)	(210)	(210)
		260	268	(327)	(248)	(321)
		197	218	(235)	(251)	(258)
		248	262	(286)	(251)	(271)
Subset 7:	$n_7 = 2$	193	(209)	(219)	(230)	(255)
		256	(277)	(294)	(260)	(281)
Subset 8:	$n_8 = 1$	(284)	327	(287)	(336)	(309)

Note: Total sample size is $N = 65$.

Then

$$E(\underline{\bar{z}}_k) = B_k\underline{\mu},$$

$$\mathrm{cov}(\underline{\bar{z}}_k) = n_k^{-1}(B_k\Sigma B_k'),$$

and the $\underline{\bar{z}}_k$ are independently distributed for $k = 1, \ldots, K$. Applying the least squares theory, we minimize

$$\sum_{k=1}^{K} \mathrm{tr}\; n_k (B_k\Sigma B_k')^{-1}[\underline{\bar{z}}_k - B_k\underline{\mu}][\underline{\bar{z}}_k - B_k\underline{\mu}]'.$$

The solution for a given value of Σ is

$$\hat{\underline{\mu}} = \left[\sum_{k=1}^{K} n_k B_k' (B_k\Sigma B_k')^{-1} B_k\right]^{-1}\left[\sum_{k=1}^{K} n_k B_k' (B_k\Sigma B_k')^{-1}\underline{\bar{z}}_k\right]. \qquad (2)$$

If a normal distribution is assumed, then the least squares estimator is also the maximum likelihood estimator. Little (1982) has suggested the use of the EM algorithm for this problem and claimed that the normal distribution assumption is not necessary. That is, estimators of $\underline{\mu}$ and Σ can be defined as the solution of the normal likelihood equations even if the underlying population is not normal. These estimators cannot then be considered maximum likelihood estimators, but only heuristic estimators that are consistent under certain general conditions. However, if a normal distribution is not assumed, then there is no justification in maximizing the normal likelihood equations to obtain estimators. An alternative heuristic estimator for Σ is given at the end of this section. The maximum likelihood estimator for Σ, assuming normality, are given from Srivastava (1985) as the solution of the following equation:

$$H = \sum_{k=1}^{K} n_k B_k' (B_k\Sigma B_k')^{-1} B_k - \sum_{k=1}^{K} B_k' (B_k\Sigma B_k')^{-1} V_k (B_k\Sigma B_k')^{-1} B_k = 0, \qquad (3)$$

where

$$V_k = (\underline{z}_{k1} - B_k\underline{\mu}, \ldots., \underline{z}_{k,n_k} - B_k\underline{\mu})(\underline{z}_{k1} - B_k\underline{\mu}, \ldots., \underline{z}_{k,n_k} - B_k\underline{\mu})'.$$

Methods for computing the solutions of (2) and (3) are given in Section 3.

Note: Alternate estimators for the covariance matrix can be defined heuristically without the normality assumption. For example $\hat{\Sigma}$ can be defined as the value of Σ that minimizes

$$\sum_{k=1}^{K} n_k^{-1}\,\mathrm{tr}[(B_k\Sigma B_k')^{-1} V_k - n_k I_k]^2 \qquad (4)$$

However, the covariance matrix must be positive definite; therefore any expression that is minimized must yield a positive definite solution. If one of the groups contains complete data, then (4) will be infinite for any singular matrix Σ; hence, there will exist a minimum for (4) in the space of positive definite matrices. A similar argument holds for the maximum likelihood estimators.

2.3 Asymptotic Distribution of $\hat{\underline{\mu}}$.

From (2) it follows that $\hat{\underline{\mu}}$ is asymptotically normally distributed with mean $\underline{\mu}$ and covariance matrix

$$P = [\sum_{k=1}^{K} n_k B_k' (B_k \Sigma B_k')^{-1} B_k]^{-1}, \tag{5}$$

which can be estimated by $\hat{P}$ obtained from P by substituting the $\hat{\Sigma}$ for Σ. Using this asymptotic theory, tests of significance and confidence regions (intervals) for $\underline{\mu}$ or linear combinations of $\underline{\mu}$ can be obtained. Alternatively, the likelihood ratio tests given by Srivastava (1985) may be used for testing the hypothesis $H\colon \underline{\mu} = \underline{0}$ against the alternative $A\colon \underline{\mu} \neq \underline{0}$. The likelihood ratio test rejects the null hypothesis H if

$$\lambda = \prod [| B_k \hat{\Sigma} B_k' | / | B_k \tilde{\Sigma} B_k' |]^{n_k/2} > \chi^2_{p,\alpha},$$

where $\tilde{\Sigma}$ is the MLE of Σ under H and $\chi^2_{p,\alpha}$ is the upper $100\alpha\%$ point of a chi-square distribution with p degrees of freedom.

2.4 Maximum Likelihood Estimates for Example 1

The maximum likelihood estimates for example 1 were obtained as:

$$\hat{\underline{\mu}} = \begin{pmatrix} 226.82 \\ 246.78 \\ 252.02 \\ 255.15 \\ 255.22 \end{pmatrix} \quad \text{and} \quad \hat{\Sigma} = \begin{pmatrix} 1809 & 1220 & 1033 & 873 & 913 \\ 1220 & 1642 & 992 & 1017 & 1121 \\ 1033 & 992 & 1438 & 718 & 1189 \\ 873 & 1017 & 718 & 1233 & 915 \\ 913 & 1121 & 1189 & 915 & 2508 \end{pmatrix} .$$

The estimated covariance matrix for the estimate of the mean vector is

$$P^{-1} = \begin{pmatrix} 28.05 & 18.78 & 15.96 & 13.46 & 14.08 \\ 18.78 & 25.67 & 15.42 & 15.84 & 17.51 \\ 15.96 & 15.42 & 24.19 & 11.24 & 19.31 \\ 13.46 & 15.84 & 11.24 & 23.33 & 15.38 \\ 14.08 & 17.51 & 19.31 & 15.38 & 54.77 \end{pmatrix} .$$

Inference on $\underline{\mu}$ can be made from the asymptotic distribution of the estimators given in Section 2.3.

2.5 Imputation

The imputation of the missing data can be made from the conditional distribution of the unobserved data given the observed data. That is define the matrices C_k for $k = 1, \ldots, K$

to be the complements of B_k. That is for a $p_k \times p$ matrix B_k with ones as the (s, i_s) entries for $s = 1, ..., p_k$ and 0's elsewhere, the matrix C_k is defined as the $(p - p_k) \times p$ matrix with ones in the (t, i_t) position and 0's elsewhere for $i_t \neq i_s$ for all $t = 1, ..., (p - p_k)$ and $s = 1, ..., p_k$. If the response vector $\underline{y}_{kj}$ corresponds to the j-th observation from subset k, then the actual observed response vector is $\underline{z}_{kj} = B_k\underline{y}_{kj}$ and the unobserved vector is $\hat{\underline{u}}_{kj} = C_k\underline{y}_{kj}$. The estimated value for the missing vector is given by

$$\hat{\underline{u}}_{kj} = C_k\hat{\underline{\mu}} + [C_k\hat{\Sigma}B_k'][B_k\hat{\Sigma}B_k']^{-1}(\underline{z}_{kj} - B_k\hat{\mu}) \tag{6}$$

Note that the estimated values for the missing vector have no random error. If the data is to be used at a subsequent analysis, with these imputed values, as if it were a complete data set, then the estimated error covariance matrix will be too small. The problem of underestimating the covariance matrix can be overcome by adding in an appropriate residual $\underline{\epsilon}$ to the estimated value $\underline{\mu}_{kj}$. If the first subset of complete data is sufficiently large then the residual vectors for missing observations in subset k can be randomly drawn from the set of values

$$(C_k\underline{y}_{1i} - C_k\hat{\underline{\mu}}) - [C_k\hat{\Sigma}B_k'][B_k\hat{\Sigma}B_k']^{-1}(B_k\underline{y}_{1i} - B_k\hat{\mu}) \text{ for } i = 1, ..., n_1. \tag{7}$$

Example 1 (continued):

The complete data set, including the imputed values based on (6) and (7) are given in Table 1 for subsets 2–8 with the imputed values in parenthesis.

3. COMPUTATIONAL PROCEDURES

Equations (2) and (3) can be solved iteratively. A procedure using a combined Newton-Raphson and steepest ascent method is given in Carter (1986) for a general case that includes linearly restricted means and covariances. The procedure is a generalization of the one given by Hartley and Hocking (1971). The method can be described as follows. For an initial choice of Σ, say Σ_0, suppose

$$\Sigma = \Sigma_0 + \wedge$$

is a solution. This expression is substituted into (3) and the equation is then expanded in a series involving only the linear terms of $\wedge$. The following approximate solution for $\wedge$ results. Define

$$Q = \sum_{k=1}^{K} (D_k \otimes D_k - D_k \otimes F_k - F_k \otimes D_k),$$

where $A \otimes B$ denotes the kronecker product of two matrices A and B defined by $A \otimes B = (a_{ij}B)$,

$$D_k = B_k' (B_k\Sigma_0B_k')^{-1}B_k,$$

and

$$F_k = B_k' (B_k\Sigma_0B_k')^{-1}V_k(B_k\Sigma_0B_k')^{-1}B_k.$$

For any matrix $A = (\underline{a}_1, \dots, \underline{a}_q)'$, we define $\mathrm{vec}(A) = (\underline{a}_1', \dots, \underline{a}_q')'$. Then (3) can be written as approximnately

$$Q \,\mathrm{vec}(\wedge) = \mathrm{vec}(E),$$

where

$$E = \sum_{k=1}^{K} (D_k - F_k).$$

To insure the nonsingularity of Q, we shall write the solution for vec($\wedge$) as

$$\mathrm{vec}(\wedge) = (Q + \lambda I)^{-1} \,\mathrm{vec}(E), \tag{8}$$

where λ is allowed to vary with the algorithm but is initially set to a very small number. For a given value of Σ, $\hat{\mu}$ is obtained from (2) and then a value of $\wedge$ is obtained from (8) to produce an updated estimate for Σ. The procedure is then iterated until a desired level of convergence is reached.

The above method can be extended to more complex structured covariance matrices; however, the procedure does require the inversion of $Q + \lambda I$. For a large number of variables this matrix will be extremely large. In this instance the alternate method of solving (3) using the EM algorithm is preferable. Again the procedure is iterative, so calculations must be performed using the updated estimates of $\underline{\mu}$ and Σ at each iteration. For an initial choice of Σ say Σ_0, define the complete predicted vector $\underline{\hat{y}}_{kj} = B_k{}'\underline{z}_{kj} + C_k{}'\underline{\hat{\mu}}_{kj}$, where the predicted missing value $\underline{\hat{\mu}}_{kj}$ is given in (6). Then

$$\underline{\hat{\mu}} = (1/N) \sum_{k=1}^{K} \sum_{j=1}^{n} \underline{\hat{y}}_{kj}$$

Define the matrix V by

$$V = \sum_{k=1}^{K} \sum_{j=1}^{n_k} (\underline{\hat{y}}_{kj} - \underline{\hat{\mu}})\ (\underline{\hat{y}}_{kj} - \underline{\hat{\mu}})'.$$

The updated estimate of Σ is then given by

$$\hat{\Sigma} = (1/N)\,[V + \sum_{k=1}^{K} n_k C_k{}' H_k C_k],$$

where H_k is the conditional variance of the incomplete data given the observed data for the k-th class defined by

$$H_k = C_k \Sigma C_k' - (C_k \Sigma B_k')\,(B_k \Sigma B_k')^{-1} (B_k \Sigma C_k').$$

The procedure is then iterated. The EM algorithm is advantageous for those situations where there exists simple closed form solutions for the likelihood equations in the complete data situations. If a Newton-Raphson procedure is necessary to solve the complete data likelihood equations then little is gained from the EM algorithm.

4. A REGRESSION MODEL

4.1 Incomplete Response Variables.

The model discussed in section 2 can be extended to handle the regression situation. The data is again partitioned into K subsets. Then the following regression model is formed:

$$Z_k = B_k' \beta A_k + \epsilon_k, \text{ for } k = 1, \ldots, K,$$

where Z_k is a $p_k \times n_k$ matrix of observed values, β is a $p \times q$ matrix of unknown parameters, B_k is as defined in Section 2, A_k is the design matrix for the matrix Z_k and the columns of ϵ_k are independently distributed with mean $\underline{0}$ and covariance matrix $B_k \Sigma B_k'$. For a given Σ, the least squares estimator of β can be written from Carter (1986) explicitly as

$$\operatorname{vec} \hat{\beta} = P^{-1} \operatorname{vec}(E),$$

where

$$P = \sum_{k=1}^{K} n_k B_k' \, (B_k \Sigma B_k')^{-1} B_k \otimes A_k A_k{}', \tag{10}$$

$$E = \sum_{k=1}^{K} B_k{}' \, (B_k \Sigma B_k')^{-1} Z_k A_k'. \tag{11}$$

The maximum likelihood estimator of Σ is given by the same formula as (3), except that now

$$V_k = [Z_k - B_k \beta A_k][Z_k - B_k \beta A_k]'. \tag{12}$$

The asymptotic distribution of $\hat{\beta}$ can be written in the form

$$\operatorname{vec}(\hat{\beta}) \sim N_{pq}(\operatorname{vec}(\beta), P^{-1}). \tag{13}$$

Inference on the regression parameters can be made from this asymptotic distribution or from the likelihood ratio statistic given in Srivastava (1985).

4.2 Incomplete Explanatory Variables

In Section 3.1, the design matrices were assumed to be known completely. In some instances the explanatory variables can also be incomplete. If the explanatory variables are random, then these missing values can first be imputed for the explanatory variables given the observed data, using the procedure of Section 2 . Once imputed values for the explanatory variables are obtained then the method of Section 3.1 can be applied to estimate the regression parameters and to impute the missing response variables.

4.3 The Likelihood Ratio Test.

The likelihood ratio procedure can be used to determine if the variables in the model are significant. To test the hypothesis

$$H: \beta = \beta_1 F \quad \text{vs} \quad A: \beta \neq \beta_1 F,$$

for F an $m \times q$ matrix of full rank, the estimates of Σ are obtained under the null hypothesis ($\tilde{\Sigma}$) and under the alternate hypothesis ($\hat{\Sigma}$). The null hypothesis is rejected at the α level of significance if

$$-2 \ell\text{n}\, \lambda > \chi^2_{(q-m)p;\, \alpha},$$

where

$$\lambda = \prod_{k=1}^{K} | B_k \hat{\Sigma} B_k' |^{n_k/2} \; / \; | B_k \tilde{\Sigma} B_k' |^{n_k/2}. \tag{14}$$

5. ESTIMATING A CHANGE POINT

Consider a sequence of observations $\underline{y}_j$, $j = 1, \ldots, N$, with expected values $E(\underline{y}_j) = \underline{\mu}_j$. Srivastava and Worsley (1986) have given a procedure for estimating the point of change of the mean vectors $\underline{\mu}_j$. It is first assumed that the change occurs at some point r. Then the following hypothesis is tested.

$$H\colon \underline{\mu}_1 = \ldots = \underline{\mu}_N$$

$$A\colon \underline{\mu}_1 = \ldots = \underline{\mu}_r \neq \underline{\mu}_{r+1} = \ldots - \underline{\mu}_N.$$

The likelihood ratio statistic is then calculated as λ_r, for $r = 1, \ldots, N-1$. The estimated point of change is that value of r that yields the maximum value of λ_r.

The existence of incomplete data poses no problems for estimating the change point. The linear model is set up as for the complete data case, then the observations are grouped into the K subsets. Suppose that the observed portion of $\underline{y}_j$ is $\underline{z}_{ki}$. Then under the alternate hypothesis for a given r, $\hat{\Sigma}$ the estimate for Σ is given from (3) for the regression model defined in (9)–(12), where the parameter matrix β is defined as

$$\beta = (\underline{\mu}_1, \underline{\mu}_2)$$

and the design matrix for the k-th subset is defined by

$$A_k = \begin{pmatrix} 1 & \ldots & 1 & 0 & \ldots & 0 \\ 0 & \ldots & 0 & 1 & \ldots & 1 \end{pmatrix},$$

where the i-th column of A_k has a one in the first row if observation $\underline{z}_{ki}$ corresponds to the vector $\underline{y}_j$ and $j \leq r$ and zero otherwise. Under the null hypothesis the population mean vector is considered the same for all N observations; hence, Σ the estimate for Σ is given from (2) and (3) for the one population mean problem. The likelihood ratio statistic is obtained from (14).

Modifications of this procedure are possible. For example the vectors $\underline{y}_j$ for $j = 1, \ldots, N$ could be sample means for N sampling time points. Multiple change points can be obtained by repeating the procedure on each section of the data. For 50 observations, if the change point occurs at point 20 then the procedure is repeated for points 1-20 and 21-50.

6. STRUCTURED COVARIANCE MATRICES

For longitudinal studies the error vectors over time may not be arbitrary, but may follow a time series model. If such a model can be assumed, then the number of parameters to be estimated is reduced. A stationary time series would assume that the covariance matrix Σ can be written as

$$\Sigma = \sigma^2 \begin{bmatrix} 1 & \rho_1 & \cdots\cdots & \rho_{p-1} \\ \rho_1 & 1 & \rho_1 \cdots\cdots & \rho_{p-2} \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ \rho_{p-1} & \rho_{p-2} & \cdots\cdots\rho_1 & 1 \end{bmatrix} . \tag{15}$$

Further models can be obtained. The correlations ρ_j can be structured. For example ρ_j can be set equal to $\rho^{|j|}$. The likelihood equations can be solved using the Newton-Raphson technique. Carter (1986) considered the case where the covariance matrix can be written as $\text{vec}(\Sigma) = G\underline{\gamma}$ for some matrix G. By defining $\gamma_i = \sigma^2\rho_i$ for $i = 1, \ldots, p - 1$ and $\gamma_p = \sigma^2$, then the covariance matrix for the stationary time series can be expressed in this linearly restricted form. For example for $p = 3$ we have

$$\begin{bmatrix} \sigma_{11} \\ \sigma_{12} \\ \sigma_{13} \\ \sigma_{21} \\ \sigma_{22} \\ \sigma_{23} \\ \sigma_{31} \\ \sigma_{32} \\ \sigma_{33} \end{bmatrix} = \begin{bmatrix} 0\,0\,1 \\ 1\,0\,0 \\ 0\,1\,0 \\ 1\,0\,0 \\ 0\,0\,1 \\ 1\,0\,0 \\ 0\,1\,0 \\ 1\,0\,0 \\ 0\,0\,1 \end{bmatrix} \begin{bmatrix} \gamma_1 \\ \gamma_2 \\ \gamma_3 \end{bmatrix}$$

The estimate of Σ can be solved numerically from the likelihood equation $G'H = 0$, where H is defined in (3). Numerically the Newton-Raphson algorithm from Section 3 can be employed with the modification that the estimate for $\underline{\gamma}$ at each iteration is given by

$$\hat{\underline{\gamma}} = (G'QG + \lambda I)^{-1}G' \ \text{vec}(E).$$

REFERENCES

CARTER, E.M. (1986). The analysis of a generalized multivariate linear model. Technical Report, University of Guelph.

DAVID, M., LITTLE, R.J., SAMUHEL, M.E., and TRIEST, R.K. (1986). Alternative methods for CPS income imputation. *Journal of the American Statistical Association,* 81, 29-41.

DEMPSTER, A.P., LAIRD, N.M., and RUBIN, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, Ser. B, 39, 1-38.

DRAPER and SMITH (1981). *Applied Regression Analysis.* New York: Wiley.

GREENLEES, W.S., REECE, J.S., and ZIESCHANG, K.D. (1982). Imputation of missing values when the probability of response depends on the variable being imputed. *Journal of the American Statistical Association*, 77, 251-261.

HARTLEY, H.O., and HOCKING, R.R. (1971). The analysis of incomplete data (with discussion). *Biometrics*, 27, 783-823.

HECKMAN, J.D. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimation for such models. *Annals of Economic and Social Measurements*, 5, 475-492.

LITTLE, R.J.A. (1982). Models for non-response in sample surveys. *Journal of the American Statistical Association*, 77, 237-250.

RUBIN, D.B., and SZATROWSKI, T.H. (1982). Finding maximum likelihood estimates of patterned covariance matrices by the EM algorithm. *Biometrika*, 69, 657-660.

SRIVASTAVA, M.S. (1985). Multivariate data with missing observations. *Communications in Statistics Theory and Methods,* 14, 775-792.

SRIVASTAVA, M.S., and WORSLEY, K.J. (1986). Likelihood ratio tests for a change in the multivariate mean. *Journal of the American Statistical Association*, 81, 199-204.

WEI, L.J., and LACHIN, J.M. (1984). Two sample asymptotically distribution free tests for incomplete multivariate observations. *Journal of the American Statistical Association*, 79, 653-661.

WOOLSON, R.F., and LEEPER, J.D. (1980). Growth curve analysis of complete and longitudinal data, *Communications in Statistics, Theory and Methods*, 9, 1491-1513.

WOOLSON, R.F., LEEPER, J.D., and CLARKE, W.R. (1978). Analysis of incomplete data from longitudinal and mixed longitudinal studies. *Journal of the Royal Statistical Society*, Ser. A, 141 ,242-252.

Survey Methodology, June 1986
Vol. 12, No. 1, pp. 73–83
Statistics Canada

Statistical Editing and Imputation for Periodic Business Surveys

M.A. HIDIROGLOU and J.-M. BERTHELOT[1]

ABSTRACT

For periodic business surveys which are conducted on a monthly, quarterly or annual basis, the data for responding units must be edited and the data for non-responding units must be imputed. This paper reports on methods which can be used for editing and imputing data. The editing is comprised of consistency and statistical edits. The imputation is done for both total non-response and partial non-response.

KEY WORDS: Periodic survey; Statistical editing; Total/partial non-response; Imputation.

1. INTRODUCTION

Data are routinely collected by large organizations such as Statistics Canada based on properly designed sample surveys. If such data are collected on a periodic basis from the same sampling unit, there are several possibilities which will occur with respect to the data consistency (quality) over a given time period. The sampling unit may report the data faithfully with no dramatic departure in continuity ("smoothness") as time progresses. The data may be reported faithfully, with questionable jumps between two time periods. The sampling unit may not report all the requested data items: this is known as partial non-response. The sampling unit may report data sporadically with breaks of total non-response for some periods. These can occur simultaneously in a periodic survey which collects required data from a large number of sampling units.

The problems which will be addressed in this article are the editing and imputation of data for sampling units that are contacted on a periodic basis by a surveying organization. The methods discussed are general for data of a multivariate nature composed of both quantitative and qualitative variables. The editing will include consistency and statistical edits.

For quantitative data, consistency edits ensure that linear combination of the data fields within a given time period satisfy given requirements. For qualitative data, consistency edits ensure that variables correspond to well defined values.

Statistical edits are used to isolate sampling units which may report some of their quantitative data fields in an inconsistent manner either from time period to time period or within a specific time period. Units with unusually high or low values will be termed "outliers". The identification of "outliers" is extremely important in an ongoing survey for two reasons. First, they influence statistics of the data set which may be for instance totals. This point has been studied by Hidiroglou and Srinath (1981). Second, the imputation of quantitative data for non-response units for periodic business surveys is usually based on trends or means: the removal of outlier units from the computation of these trends or means, will produce statistics that are not contaminated with there observations. For units which have partial non-response, data must be imputed for the missing fields.

For large data sets, where timely release of the summary information is crucial, the editing and the imputation of data should be automatic and computer handled given some well specified rules. This is in agreement with Gentleman and Wilk (1975), and Fellegi and Holt (1976).

[1] M.A. Hidiroglou and J.-M. Berthelot, Business Survey Methods Division, 11th Floor, R.H. Coats Building, Tunney's Pasture, Ottawa, Ontario K1A 0T6.

2. EDITING PERIODIC DATA

2.0 Consistency Edits

For a given unit i and time period t, let $\underset{\sim}{x}_i(t)$ represent the vector of data which is to be collected. The vector $\underset{\sim}{x}_i(t)$ may be decomposed into a series of elementary vectors for which independent editing and imputation are required.

That is, $$\underset{\sim}{x}_i(t) = (\underset{\sim}{x}_i^{(1)}(t), \dots, \underset{\sim}{x}_i^{(P)}(t))$$

where $$\underset{\sim}{x}_i^{(p)}(t) = (x_{il}^{(p)}(t), \dots, x_{ik_p}^{(p)}(t))$$

for $$i=1, \dots, n;\ p=1, \dots, P;\ t=1, \dots, T$$

and k_p is the number of variables in the *p:th* elementary vector.

For each elementary vector $\underset{\sim}{x}_i^{(p)}(t)$, the consistency edits may be represented as

$$\underset{\sim}{A}^{(p)}(\underset{\sim}{x}_i^{(p)}(t))' \leq (\underset{\sim}{c}^{(p)})'$$

where $\underset{\sim}{A}^{(p)}$ is a ℓ_p by k_p matrix representing the rules that the elements of the elementary vector $\underset{\sim}{x}_i^{(p)}(t)$ must obey, and $\underset{\sim}{c}^{(p)}$ is a 1 by ℓ_p vector which represents the constraints. This formulation allows one to define consistency edits for both qualitative and quantitative variables. For qualitative variables, the consistency edits could be used to check if the variables correspond to well-defined values. For quantitative variables, the consistency edits can check if certain variables are not larger (or smaller) than other variables or that a linear combination is equal to (or greater than or less than) a given variable.

2.1 Statistical Edits

Given that data are reported periodically, the problem is to isolate outlying observations within the time series. In the present context, an outlying observation i, will be defined as one whose trend for the current period to a previous period, for given variables of the element vector $\underset{\sim}{x}_i(t)$, differs significantly from the corresponding overall trend of other observations belonging to the same subset of the population. Statistical edits can also be applied within a time period, by comparing the ratios of two correlated variables amongst themselves, within a given subset of the population. In this article, the statistical edit will only be discussed in terms of the trend between time periods. Similar, somewhat imprecise but working definitions of outliers have also been given by other authors, for example:

GRUBBS (1969) says that "An outlying observation, or outlier, is one that appears to deviate markedly from the other members of the sample in which it occurs."

GUMBEL (1960) says: "The outliers are values which seem either too large or too small as compared to the rest of the observations."

KENDALL and BUCKLAND (1957, p. 209), write: "In a sample of n observations it is possible for a limited number to be so far separated in value from the remainder that they give rise to the question whether they are from a different population, or that the sampling technique is at fault. Such values are called outliers. Tests are available to ascertain whether they can be accepted as homogeneous with the rest of the sample."

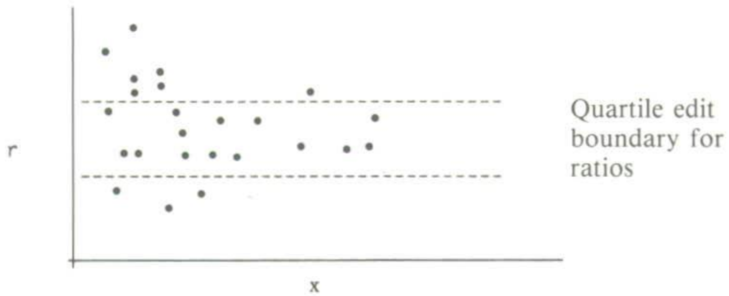
2.1.1 Review of Some Methods Currently Used

Methods for detecting outliers have been proposed by Dixon (1953), Grubbs (1969), Tietgen and Moore (1972), and Prescott (1978) to mention a few. Most of the test procedures for outlier detection proposed by these authors consider the problem as one of hypothesis testing. In the simplest cases, the null hypothesis is that the sample comes from a normal distribution with unspecified mean and variance, while the alternative hypothesis is that one or more of the observations come from a different distribution. Percentage points of a test statistic may be determined under the null hypothesis and compared with computed values of the test statistic in particular applications. Applying these methods to periodic data from large surveys presents problems for the following reasons. First, the assumption of normality of trends from one period to another may not hold. Second, these traditional methods require the existence of tables for determining critical values which define rejection regions. The method which we will propose in Section 2.1.2 does not have the above mentioned disadvantages. It can be easily implemented on the computer, does not require the assumption of normality, and does not make use of tables.

In our specific context, and given elements of the vectors $\underset{\sim}{x}_i(t)$ and $\underset{\sim}{x}_i(t + 1)$, denote as $x_i(t)$ and $x_i(t + 1)$ the responses for two consecutive periods for a given unit, where $i=1, ..., n$. Denote as r_i the ratio of current period data to previous period data. One method which is known as the range edit, is to simply define fixed upper and lower bounds based on experience for comparison purposes. Ratios found outside these bounds are declared as outliers. A major drawback with this method is that the definition of outlier is too subjective and does not make use of the distribution of the ratios.

A method that attempts to make use of the distribution of the ratios is the Chebychev inequality edit. This edit is constructed by computing the lower bound as $\bar{r} - ks_r$ and the upper bound as $\bar{r} + ks_r$ where $\bar{r} = \Sigma_{i=1}^{n} r_i/n$ and $s_r^2 = \Sigma_{i=1}^{n} (r_i-\bar{r})^2/(n - 1)$. This edit has two main drawbacks. First, the choice of k is subjective and can result in having an edit that cannot detect any outliers. This last point has been demonstrated by Wilkinson (1982). Second, "large" outliers may hide "smaller" outliers. This effect is known as the masking effect.

An improvement to this method has been the use of quartiles and interquartile distances rather than the use of mean and standard error to come up with the upper and lower bounds. In this case, the edit is constructed by computing the lower bound as $r_M - k\, D_{r_{Q1}}$ and the up per bound as $r_M + k\, D_{r_{Q3}}$ where r_M is the median of the ratios, $D_{r_{Q1}}$is the distance between the first quartile and the median, and $D_{r_{Q3}}$ is the distance between the third quartile and the median. Since the quartiles are not affected by the tails of the distribution, it greatly alleviates the masking effect problem. However, this method has two drawbacks. First, in some very specific circumstances, it is possible that the outliers on the left tail of the distribution are undetectable. Second this method does not take into account the fact that in most of the periodic business surveys, the variability of ratios for small businesses is larger than the variability of ratios for large businesses (Sugavanam 1983). This fact is expressed by the following graph:



This drawback has the effect of identifying too many small units as outliers and not enough large units. This effect will be referred to as the "size masking effect".

2.1.2 Proposed Procedure

For two occasions t and $t + 1$, the overall trend for the data pair given by

$$(x_i(t),\ x_i(t + 1)),\ i = 1,\ \ldots,\ n$$

is

$$R = \sum_{i=1}^{n} x_i(t + 1) / \sum_{i=1}^{n} x_i(t).$$

Now, R may be expressed as

$$R = \sum_{i=1}^{n} I_i\, r_i$$

where

$$I_i = x_i(t) / \sum_{i=1}^{n} x_i(t)$$

and

$$r_i = x_i(t + 1)/x_i(t).$$

I_i is a measure of the relative importance of the *i:th* unit amongst the n units at time t. The individual trends r_i must be transformed in order to ensure that outliers are detected at both tails of the distribution. This transformation is:

$$s_i = \begin{cases} 1 - r_M/r_i, \text{ if } 0 < r_i < r_M \\ r_i/r_M - 1, \text{ if } r_i \geq r_M \end{cases}$$

where r_M is the median of the ratios.

In order to bring in the magnitute of the data, the following transformation is required (Berthelot 1983):

$$E_i = s_i \{\text{Max } (x_i(t),\ x_i(t + 1))\}^U$$

where $0 \leq U \leq 1$. The E_i's will be referred to as effects and the exponent U in the transformation provides a control on the importance associated with the magnitude of the data. This transformation allows us to place more importance on a small change associated with a "large" unit as opposed to a large change associated with a "small" unit. The values of the median and quartiles as used by Sande (1981) will be applied to the transformed, E_i*'s*, in order to detect potential outliers. Denoting as E_{Q1}, E_M and E_{Q3} as the first quartile, the median and the third quartile respectively, define the following two deviations:

$$d_{Q1} = \text{Max } (E_M - E_{Q1},\ |AE_M|),$$

$$d_{Q3} = \text{Max } (E_{Q3} - E_M,\ |AE_M|).$$

Outliers will be defined as all those units whose associated effect E_i lies outside the interval $(E_M - Cd_{Q1}, E_M + Cd_{Q3})$. The purpose of the AE_M term is to avoid difficulties which arise when $E_M - E_{Q1}$ or $E_{Q3} - E_M$ are very small. That is, the problem which may arise when the effects E_i are clustered around a single value with one or two modest deviations may produce false outliers. The parameter C controls the width of the acceptance interval. The parameter U controls the shape of the curve defining upper and lower boundaries. The effect of increasing U is to attach more importance with fluctuations associated with the larger observations. A value of 0.05 is suggested for A as it has proved to be adequate in practice.

2.1.3 Treatment For Outliers

Once units have been identified as possible outliers, they are flagged as such and brought to the attention of the survey takers. A decision must then be taken on how these abnormal observations are treated. Their existence may have arisen as a result of several factors. These factors include measurement error, incorrect interpretation of the questionnaire by the responding unit, or intrinsic variability of the population being surveyed. For units which have measurement error due to incorrect transcription of the data or incorrect responses, a simple follow-up will clear up the majority of these errors. For units which display intrinsic variability as a result of rapid growth, the reported values are correct but dominate too much the resulting summary tables. For those units, techniques, which reduce the sampling weight as suggested by Hidiroglou and Srinath (1981) or change the values themselves as suggested by Ernst (1980), must be used in order to accomodate (minimize) the effect of outlying observations. For units having unrepresentative data which cannot be verified, their data must be substituted with other data based on imputation techniques. The different kinds of corrective actions taken on outlying units must be flagged as well.

3. IMPUTING PERIODIC DATA

The information collected by periodic business surveys, such as sales and employment are collected via samples using mail questionnaires or telephone interviews. Non-responding units are followed up as much as possible within allotted budgets in order to improve the response rates. The follow-up is usually done by mail in the case of the smaller to medium sizes non-responding companies and by telephone for the larger or dominating companies. Although following up delinquent companies improves response rates for a given reference period, there will be nevertheless, a group of non-responding companies which may be classified into either hard-core or late respondents. Hard-core non-respondents are units which require a great deal of persuation to respond, if at all. Late respondents are units which respond late with respect to the survey's reference period either because they do not mail back their questionnaire on time or because they need to be prompted by a follow-up questionnaire. The non-responding units must therefore be imputed in order to make up for their contribution to the particular estimator being used by the survey. In the case of Monthly Business Surveys, such as the Monthly Retail Trade Survey, totals (e.g., sales) are being estimated. Imputation procedures can also be used to generate values for units declared as outliers. These imputed values can be used in lieu of these outlying observations, if no valid explanation can be provided for their presence.

The units with no response whatsoever, will be termed as total non-respondents and those with some, but not all, required data items, will be termed partial non-respondents. Desirable features of an imputation system should include the following properties (Berthelot and Hidiroglou 1982):

- it must automatically determine the most reasonable imputation procedure possible under the existing circumstances,
- the imputation cell, the level at which the computation of trends and means (medians) is performed, will usually correspond to the finest level of stratification of the sample,
- a minimum number of units must participate in the computation of trends or means (medians), otherwise, the imputation cells are automatically collapsed (using a pre-determined pattern), until the minimum requirement has been satisfied,
- it will recognize through the use of status codes that there are units which must not be imputed. These include seasonal units during the period that they are not operating, units temporarily out of business, or units which are no longer active,
- births which have no previous business history will have their data imputed using the means (medians) of similar responding births,
- units will be re-imputed for a number of periods previous to the current period: this is done in order to improve the strength of the imputations if the previous periods have been updated with data,
- backward imputations will be applied to units which have been continuously imputed using a forward imputation procedure as soon as a good response is obtained for a given period,
- imputation status codes will be associated with imputed units in order to provide a history of the procedure used for imputation,
- the ranking for imputing non-responding units is as follows: trends (monthly, quarterly, annual), means (medians) with the most recent trends being given priority. For instance, in the case of a monthly system, monthly trends are used for units which have data (response or imputed) in the month prior to the one to be imputed. Annual trends are used mostly for units which are seasonal and which fail to provide a response as they emerge from their out of season period and for which a last year value existed for the month to be imputed. Imputations based on the trends are obtained by multiplying the trends by the unit's last month or last year value. In the event that trends cannot be applied, the mean (median) of the cell is used as an imputation.

In order to formalize the preceding paragraphs in a mathematical fashion, let the number of units which are expected to respond for a given cell and given month be n. Let the number of non-respondents with total non-response be n_3, the number of respondents with total response be n_1 and the number of respondents with partial response be n_2. It is assumed that the sample design is stratified with the sampling being simple random without replacement. Let the size for the follow-up sample of the non-respondents be m_3 ($2 \leq m_3 \leq n_3$, with m_3 having been selected from n_3 according to a randomized mechanism). Note that $n_4 = n - \Sigma_{i=1}^{3} n_i$ units are not expected to provide any response to the survey process for a number of possible reasons. At a time t, they may be out of season, inactive, dead, or out of scope to the survey. For these units, the system will automatically associate zero values for all relevant fields in the given period.

The imputation process will then be done in several different ways according to the type of non-response.

3.0 Total Non-Response

The imputation process for the total non-respondents will first be discussed. Bearing in mind that either the whole vector $\underset{\sim}{x}_i(t)$ or that some of its elementary vectors as given in

Section 2.0 must be totally imputed, denote as $(x_{i1}(t), \ldots, x_{ip}(t))$ one of the elementary vector within $\underset{\sim}{x}_i(t)$ where the editing and imputation process is independent from other elementary vectors within $\underset{\sim}{x}_i(t)$. Assuming that

$$x_{ip}(t) \geq \sum_{j=1}^{p-1} x_{ij}(t),$$

(which implies that the sum of the first $p-1$ data elements of the elementary vectors are smaller than the *p:th* datum element, the total) $x_{ip}(t)$ will first be imputed as

$$I_{ip}^{(1)}(t) = \sum_{k=1}^{6} [z_{ip}^{(k)}(t)\ \delta_i^{(k)}]$$

where $\delta_i^{(k)}$ refers to the procedure used for imputation and $z_{ip}^{(k)}$is the associated imputed value. One of the six $\delta_i^{(k)}$ values will be one and the other five must be zero ($\Sigma_{k=1}^{6}$ $\delta_i^{(k)} = 1$). The imputed $z_{ip}^{(k)}(t)$ values will be as follows:

$$z_{ip}^{(1)}(t) = [\sum_{r \epsilon s_1} w_r\, x_{rp}(t) / \sum_{r \epsilon s_1} w_r\, x_{rp}(t-1)]\ x_{ip}(t-1),$$

$$z_{ip}^{(2)}(t) = [\sum_{r \epsilon s_2} w_r\, x_{rp}(t) / \sum_{r \epsilon s_2} w_r\, x_{rp}(t-Q)]\ x_{ip}(t-Q),$$

$$z_{ip}^{(3)}(t) = [\sum_{r \epsilon s_3} w_r\, x_{rp}(t) / \sum_{r \epsilon s_3} w_r\, x_{rp}(t-1)]\ x_{ip}(t-1),$$

$$z_{ip}^{(4)}(t) = [\sum_{r \epsilon s_4} w_r\, x_{rp}(t) / \sum_{r \epsilon s_4} w_r\, x_{rp}(t-Q)]\ x_{ip}(t-Q),$$

$$z_{ip}^{(5)}(t) = [\sum_{r \epsilon s_5} w_r\, x_{rp}(t) / \sum_{r \epsilon s_5} w_r],$$

$$z_{ip}^{(6)}(t) = [\sum_{r \epsilon s_6} w_r\, x_{rp}(t) / \sum_{r \epsilon s_6} w_r],$$

w_r = inverse selection probability of unit r for the given cell. The subsets s_i $(i=1, \ldots, 6)$, will be determined by selecting the units which have provided a response for the *p:th* variable at time t and which have passed the edits. The conditions for each subset is

s_1 = all units which have provided edited responses between times t and $t-1$,

s_2 = all units which have provided edited responses between times t and $t-Q$,

s_3 = units in the follow-up subsample which have provided edited responses between times t and $t-1$,

s_4 = units in the follow-up subsample which have provided edited responses between times t and $t-Q$,

s_5 = all units which have provided edited responses at time t,

s_6 = units in the follow-up subsample which have provided edited responses at time t.

The choice of the imputation procedure will be governed by the following considerations.

(i) Procedures 1 (or 2) will be used if there is a response or imputed value at time $t-1$ (or $t-Q$) and that it is believed that the trends for the non-respondents is the same as the one for the respondents, within the given cell,

(ii) Procedures 3 (or 4) will be used if there is a response or imputed value at time $t-1$ (or $t-Q$) and that it is believed that the trends for the non-respondents differs from the one for the respondents within the given cell.

(iii) Procedure 5 will be used if there is no response at either times $t-1$ or $t-Q$ and that is believed that the mean of the non-respondents is equal to the mean of the respondents within the given cell,

(iv) Finally, procedure 6 will be used if there is no response at either times $t-1$ or $t-Q$ and that it is believed that the means of the respondents and non-respondents are different.

The choices between the different procedures can be made using decision tables which determine the conditions and, given the condition, choose the best imputation procedure according to pre-determined rules. Once that $x_{ip}(t)$ has been imputed for an elementary vector, its remaining components can be imputed using the procedures for partial non-response.

3.1 Partial Non-Response

For an elementary vector $(x_{i1}(t), x_{i2}(t), \ldots, x_{ip}(t))$ which is part of $\underset{\sim}{x}_i(t)$, let δ_{ij} be the indicator variable which is equal to 1 if $x_{ij}(t)$ is present and zero otherwise at time t. Some additional notation is introduced at this point in order to ease the development. To this end, define

$$s_{i,R}(t-1) = \sum_{j=1}^{p-1} \delta_{ij}\, x_{ij}(t-1)$$

= the sum of responses at time $t-1$, for which there is a response at time t

$$s_{i,NR}(t-1) = \sum_{j=1}^{p-1} (1-\delta_{ij})\, x_{ij}(t-1)$$

= the sum of responses at time $t-1$, for which there is no response at time t,

$$s_{i,R}(t) = \sum_{j=1}^{p-1} \delta_{ij}\, x_{ij}(t).$$

The partial imputation will be based on the assumptions that $x_{ip}(t) \geq \Sigma^{p-1}_{j=1} x_{ij}(t)$ and that the distribution of the elements within $\underset{\sim}{x}_i(t)$ is similar to the distribution of the elements within $\underset{\sim}{x}_i(t-1)$. Two separate cases will be discussed.

Case 1: Parts of the elementary vector missing and $x_{ip}(t)$ present

Two subcases are possible: $x_{ip}(t) = \sum_{j=1}^{p-1} x_{ij}(t)$ or $x_{ip}(t) > \sum_{j=1}^{p-1} x_{ij}(t)$.

(i) $x_{ip}(t) = \sum_{j=1}^{p-1} x_{ij}(t)$

If all the elements of $\underset{\sim}{x}_i(t)$ excluding $x_{ip}(t)$ are missing, that is $\Sigma_{j=1}^{p-1} \delta_{ij} = 0$, then we must have that $s_{i,NR}(t) = x_{ip}(t)$. If some of the elements of $\underset{\sim}{x}_i(t)$ excluding $x_{ip}(t)$ are missing, that is $\Sigma_{j=1}^{p-1} \delta_{ij} > 0$, then $s_{i,NR}(t) = x_{ip}(t) - s_{i,R}(t)$.

(ii) $x_{ip}(t) > \sum_{j=1}^{p-1} x_{ij}(t)$

If all the elements of $\underset{\sim}{x}_i(t)$ *excluding* $x_{ip}(t)$ are missing, then $s_{i,NR}(t) = s_{i,NR}(t-1)$ $x_{ip}(t)/x_{ip}(t-1)$. If some of the elements of $\underset{\sim}{x}_i(t)$ excluding $x_{ip}(t)$ are missing, the choice of $s_{i,NR}(t)$ is not so obvious. In any event, one must have that $s_{i,R}(t) + s_{i,NR}(t) < x_{ip}(t)$. To this end, four separate possible imputations for $s_{i,NR}(t)$ will be given in order of preference.

(a) $s_{i,NR}(t) = [s_{i,NR}(t-1) + s_{i,R}(t-1)]\, x_{ip}(t)/x_{ip}(t-1) - s_{i,R}(t)$ provided that $s_{i,NR}(t) \geq 0$. Note that the condition $x_{ip} > \Sigma_{j=1}^{p-1} x_{ij}(t)$ is met if $s_{i,NR}(t) \geq 0$.

(b) $s_{i,NR}(t) = s_{i,NR}(t-1)\, [s_{i,R}(t)/s_{i,R}(t-1)]$

(c) $s_{i,NR}(t) = s_{i,NR}(t-1)\, [x_{ip}(t)/x_{ip}(t-1)]$

(d) $s_{i,NR}(t) = x_{ip}(t) - s_{i,R}(t)$.

The preferred imputation will be the first one that does not violate the inequality condition. For all the above cases, the imputed (actual values) will then be

$$\begin{aligned} I_{ij}^{(2)}(t) &= (1-\delta_{ij})\, [s_{i,NR}(t)/s_{i,NR}(t-1)]\, x_{ij}(t-1) \\ &\quad + \delta_{ij}\, x_{ij}(t);\ j=1, \ldots, p-1 \end{aligned}$$

Case 2: Parts of the elementary vector missing and $x_{ip}(t)$ is missing

As in case 1, two subcases are possible:

(i) $x_{ip}(t) = \sum_{j=1}^{p-1} x_{ij}(t)$

If $\Sigma_{j=1}^{p-1} \delta_{ij} = 0$, then $s_{i,NR}(t) = I_{ip}^{(1)}(t)$ where $I_{ip}^{(1)}(t)$ has been obtained using the imputation for total non-response. The imputation $I_{ij}^{(2)}(t)$ is then used. If $\Sigma_{j=1}^{p-1} \delta_{ij} > 0$, $I_{ij}^{(2)}(t)$ will be used provided that $s_{i,NR}(t) = I_{ip}^{(1)}(t) - s_{i,R}(t) \geq 0$. Otherwise, the following imputation must be used

$$\begin{aligned} I_{ij}^{(3)}(t) &= (1-\delta_{ij})\ [s_{i,NR}(t)/s_{i,NR}(t-1)]\ x_{ij}(t-1) \\ &\quad + \delta_{ij}\, x_{ij}(t);\ j=1, \ldots, p-1 \end{aligned}$$

and $I_{ip}^{(1)}(t)$ is replaced by $I_{ip}^{(3)}(t) = \Sigma_{j=1}^{p-1} I_{ip}^{(3)}(t)$

(ii) $x_{ip}(t) > \Sigma_{j=1}^{p-1} x_{ij}(t)$

For this case, the $x_{ip}(t)$ in case 1(ii) is replaced by $I_{ip}^{(1)}(t)$ and the methods given for this case are used, provided that the above inequality condition is satisfied. If the condition cannot be met, $I_{ip}^{(3)}(t)$ must be used and $I_{ip}^{(1)}(t)$ is replaced by $I_{ip}^{(3)}(t) = \Sigma_{j=1}^{p-1} I_{ip}^{(3)}(t)$.

If the assumption, that the distributions of the data elements of vectors $\underset{\sim}{x}_i(t)$ and $\underset{\sim}{x}_i(t-1)$ is similar, does not hold, then each individual element must be imputed using procedures for imputation for total non-response. These imputations must then be adjusted in order to satisfy the inequality requirement $x_{ip} \geq \Sigma_{j=1}^{p-1} x_{ij}$. Hence, for example, for case 1(i), we would have for $\Sigma_{j=1}^{p-1} \delta_{ij} = 0$,

$$I_{ij}^{(4)}(t) = [x_{ip}(t) / \sum_{j=1}^{p-1} I_{ij}^{(1)}(t)] \; I_{ij}^{(1)}(t)$$

and for $\Sigma_{j-1}^{p-1} \delta_{ij} > 0$

$$I_{ij}^{(4)}(t) = (1-\delta_{ij}) \left[\frac{x_{ip}(t) - \Sigma_{j=1}^{p-1} \delta_{ij} \, x_{ij}(t)}{\Sigma_{j=1}^{p-1} \; (1-\delta_{ij}) \; I_{ij}^{(1)}(t)} \right] + \delta_{ij} \, x_{ij}(t); j = 1, \ldots, p-1.$$

Similarly, cases 1(ii) and 2, could be developed using the imputed values $I_{ij}^{(1)}(t)$.

4. CONCLUSION

For periodic business surveys, it is important to have computer systems which can quickly and accurately monitor the flow of in-coming data in terms of its quality. Conversely, for expected data that are not coming in, the system should impute as well as possible for the non-response given some well specified rules.

The editing will cause the flagging of records in possible error. These errors can be termed as critical and non-critical. All errors should be corrected by either reviewing the questionnaires or checking their authenticity with the respondent. If this is not possible on account of time or budgetary constraints, the most critical errors must be corrected. Given that the errors have been taken care of, the next step of the processing is to impute for the non-respondents. Diagnostic summaries of the actions (edits or imputations) taken by the system, should be printed out in order to inform the survey analyst on the status of his data.

REFERENCES

BERTHELOT, J.-M., and HIDIROGLOU, M.A. (1982). Specifications for imputations in the retail trade survey. Technical report, Statistics Canada.

BERTHELOT, J.-M. (1983). Wholesale-retail redesign, statistical edit proposal. Technical Report, Statistics Canada.

DIXON, W.G. (1953). Processing data for outliers. *Biometrics*, 9, 74-89.

Survey Methodology, June 1986
Vol. 12, No. 1, pp. 99–106
Statistics Canada

A Study of the Effects of Imputation Groups in the Nearest Neighbour Imputation Method for the National Farm Survey

SIMON CHEUNG and CRAIG SEKO[1]

ABSTRACT

A new processing system using the nearest neighbour (N-N) imputation method is being implemented for the National Farm Survey (NFS). An empirical study was conducted to determine if the NFS estimates would be affected by using imputation groups based on type of farm. For the specific imputation rule examined, the study showed evidence that the effect might be small.

KEY WORDS: National Farm Survey; Item non-response; Nearest neighbour imputation; Match variable transformation.

1. INTRODUCTION

The National Farm Survey (NFS) is an annual multi-purpose survey of agricultural activity in Canada. The survey uses a 2-frame sample design i.e. a list frame of large farms (based on the quinquennial Census of Agriculture) and an area frame of agricultural land. The largest units in the list frame are sampled with certainty (i.e. with probability one) because of their disproportionate impact on the survey estimates. These units are called specified farms. The remaining farms in the list frame are stratified and sampled. The small farms in the survey population, which are comparatively very large in number, are covered by the area frame and sampled less extensively than the list frame farms. Thus three samples are selected: specified, list and area. The detailed NFS sample design has been described by Davidson and Ingram (1983), and Davidson (1984).

The NFS is processed by a system adopted from predecessor surveys. This system employs the sequential hot-deck imputation method to adjust for unit and item non-response (Philips 1979). A new survey processing system will be implemented in 1987 in order to integrate all the agricultural surveys conducted by Statistics Canada. This system will use the nearest neighbour (N-N) imputation method to adjust for item non-response. The decision to implement the N-N imputation method was based on many reasons, among which there are three important ones: First, the use of the N-N method is theoretically more justified than the exact-matching sequential hot-deck method since the survey collects mostly quantitative data. Second, empirical studies, e.g. Kovar (1982), suggest that the two imputation methods would yield similar estimates for the NFS with the N-N method resulting in fewer outliers i.e. imputed data which have disproportionate contributions to the survey estimates. Third, switching to this new imputation method for the NFS would help standardize the survey methodology of all agricultural surveys, a long term goal of Statistics Canada. Currently, the Census of Agriculture and the Farm Tax Data Survey both use the N-N imputation methodology.

This paper reports on an empirical study which attempts to provide information that will help in a more efficient implementation of the new imputation method. The next section describes briefly the N-N imputation method adopted in our study. Section three presents the study procedure and the main results obtained. Finally, we discuss our preliminary observations drawn from the results in section four.

[1] Simon Cheung and Craig Seko, Business Survey Methods Division, Statistics Canada, 11th Floor, R.H. Coats Building, Tunney's Pasture, Ottawa, Ontario, Canada K1A 0T6.

2. NEAREST NEIGHBOUR IMPUTATION METHOD

The method of donor imputation, in general, is to replace the missing or invalid values of a respondent (recipient) with the valid response of another respondent (donor) who is deemed to have the same characteristics as the recipient. The sequential hot-deck imputation method identifies donors sequentially in the course of processing as those reporting the same values as the recipient in the pre-specified match variables. This method, however, often fails to obtain an exact match when a match variable assumes a large number of possible values. To alleviate this, the range of the match variable is split into intervals and the donor is obtained by matching on the interval code. In nearest neighbour imputation, this problem is solved by selecting a donor based on a multivariate distance measure which represents the degree of similarity between the donor and the recipient as defined by the pre-specified match variables. The more similar two respondents are with respect to the match variables, the smaller the magnitude of the distance. Thus, the best donor for a recipient is the donor candidate which has the smallest distance value from the recipient, i.e. its nearest neighbour in the sense of statistical distance.

The nearest neighbour imputation method used in this study was proposed by Sande (1976, 1981). This method uses the maximum norm based on transformed data as the distance function. The method is described briefly below.

Let $X = (x_1, x_2, x_3, \ldots, x_k)$ be a vector of k match variables. Each match variable x_j is transformed by $t_j = \hat{F}(y)$, where $\hat{F}(y)$ is the empirical distribution function of x_j. Note that t_j follows the uniform distribution over [0, 1]. Then the distance between a given recipient X^r and a donor candidate X^d defined by the maximum norm is

$$d\,(X^r, X^d) = \max_{j} |\, t_j^r - t_j^d \,| ,$$

where t_j^r and t_j^d are the transformed values of the j^{th} match variable x_j in X^r and X^d, respectively. The donor candidate with the smallest d-value will be selected and its response will be copied for the missing item of the recipient. The uniform transformation may be considered as an objective method to scale the match variables regardless of their natural distributions.

3. EMPIRICAL STUDY

3.1 Motivation

In adopting the nearest neighbour imputation method for the NFS, some issues regarding detailed implementation of this method need to be resolved, particularly in regards to transforming match variables. The method of uniform transformation in the N-N imputation could be applied using all the records in the sample or using only subsets of the sample data. A group of unit respondents in which imputation for non-response takes place is called an imputation group. Different imputation groups would yield different transformed values which in turn would result in different selection of donor records.

It was conjectured that transforming match variables within an imputation group defined by a homogeneity criterion which is closely related to the item to be imputed would result in a more correct scaling of the match variables, and hence would yield better imputed data. For example, in the NFS one may expect that match variable tranformation within imputation groups defined by farm type should yield better imputed data and hence better estimates, 'better' being in the sense of bias and variance reduction. Unfortunately, the transformation of match variables is costly in terms of computer resources. If one does not need to transform within homogeneous imputation groups, savings in computer costs can be realized.

The main objective of the study was to answer the following question in an experimental setting: 'Do the two methods of match variable transformation, i.e., transformation using all records vs. within farm type groups, yield substantially different survey estimates? If so, which method yields better estimates?'

3.2 Data Used in the Study

After consultation with the subject matter analysts, the 1984 NFS sample for the province of Alberta was selected for the study. The sample of approximately 2000 farms consists of 50% crop farms, 27% livestock farms and 23% mixed farms. The population percentages of the three farm types were estimated to be 52%, 27% and 21% repectively. Farm types were assigned according to the main source of projected agricultural receipts of a farm. If at least 75% of a farm's projected agricultural receipts came from its livestock inventory, the farm was classified as a livestock farm. A similar rule was used to classify crop farms. The remaining farms were classified as mixed farms.

3.3 Method of the Study

We assumed that the data was 'clean', even though it contained imputed values via the sequential hot-deck imputation procedure. Once the data had been classified by farm type, the following procedure was followed:

i) Ten per cent of the values for each imputation variable was randomly set to a missing value within each farm type. This error generation was done independently for each imputation variable.
ii) The generated non-responses were imputed using the N-N imputation method based on the two sets of imputation groups defined by the whole sample (called 'whole') and by farm type (called 'by-type'). The imputation procedures were carried out using the Numerical Edit/Imputation System (Statistics Canada 1982), as implemented within the P-STAT statistical package (Buhler and Buhler 1978).
iii) The NFS weighted estimates for the variable totals for the province and for each farm type were produced based on each set of imputed data.
iv) These steps were repeated 10 times to get 10 independent replications (i.e., simulations), and the results were averaged over the ten replications for each imputation variable. This average estimate was then compared with the estimate obtained based on the 'clean' file, both at the provincial level and for each farm type.

The whole experiment was repeated for higher non-response rates of 15% and 20% in order to observe the impact of nonresponse rates.

The imputation and match variables used in the study are shown below:

Imputation Variables

UTIL	=	Utility expenses
AUTO	=	Farm vehicle and machinery operating expenses
TAX	=	Property tax

Match Variables

Farm type (exact matching)
FEED = Feed expense
SEED = Seed expense
INCOME = Gross agricultural receipts

In addition, the donor's sample type was restricted by the recipient's. Recall that three types of samples are used in the NFS: specified, list, and area. A specified farm can be imputed by a farm from any of the sample types but can not be a donor to a list or area farm. Similarly, a farm from the list sample can be imputed from a farm in either the list or area samples but can only be a donor to farms that are in the list sample or are specified. Finally, farms in the area sample can only be imputed by another area farm but can serve as a donor to any of the three samples. These restrictions arise from the premise that if a list or specified farm was allowed to impute for an area farm, the imputed value could potentially raise the survey estimates to an unacceptable level because of the higher sampling weights associated with area farms.

3.4 The Empirical Distribution Functions of the Match Variable

Figure 1 shows the unweighted empirical distribution functions of the three match variables which are obtained from the imputation groups defined by the whole sample and by farm type. Note that the differences are substantial and hence could lead to the selection of different donor records for a given recipient.

3.5 Results

The results are tabulated in Table 1. For each imputation variable (UTIL, AUTO or TAX), each of the two sets of imputation groups (whole vs. by-type), and each level of non-response rate (10%, 15% or 20%), the average value of the ten estimates for the variable total was calculated over the ten replications. The bias of this average value is displayed as a percentage of the "clean" estimate. The average *cv* over the ten replicates is also displayed as a percentage.

4. OBSERVATIONS AND DISCUSSION

This study imputed for three farming expense variables. The donor records were selected by exact matching on farm type and by nearest-neighbour matching on three variables: gross agricultural receipts, feed expense and seed expense. The two expense match variables were believed to be of different effectiveness for the three farm types. For example, feed expense was expected to work better for livestock farms but not so for crop farms, etc. The strength of correlation between the match variables and the imputation variables presented in Table 2 seems to support this expectation.

Therefore the homogeneous subsets based on type of farm have differing relationships for the match variables. This might imply that transformations using imputation groups defined by these subsets would perform better than using the entire sample as an imputation group. The results, however, indicate that using these homogeneous subsets as imputation groups does not seem to yield substantially different estimates or lower bias. The bias itself

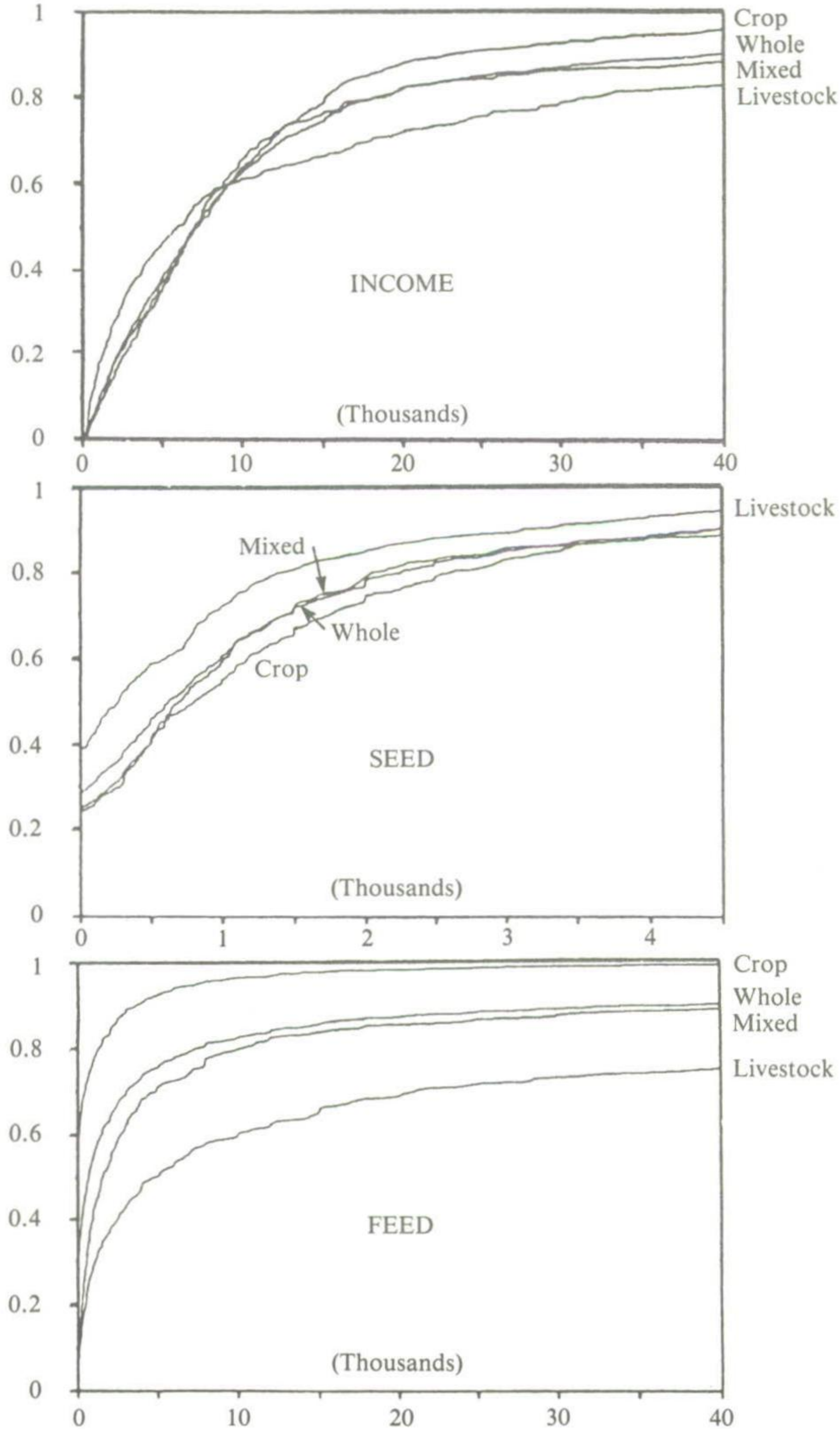


Figure 1: Empirical Distribution Functions of Match Variables

Table 1
Percentage Bias and cv's for the Totals of the Imputation Variables after Imputation

Non-response rate	Imputation group	Imputation Variables					
		UTIL		AUTO		TAX	
		% Bias	% cv	% Bias	% cv	% Bias	% cv
		All Farms in Sample					
clean			3.137		2.831		3.224
10%	by-type	0.176	3.165	-0.004	2.849	0.228	3.260
	whole	0.124	3.143	-0.074	2.840	0.199	3.296
15%	by-type	0.339	3.195	0.604	2.885	0.255	3.275
	whole	0.336	3.131	0.278	2.870	-0.624	3.289
20%	by-type	0.869	3.173	0.023	2.875	-0.715	3.280
	whole	0.554	3.111	-0.150	2.843	-0.877	3.285
		Crop Farms					
clean			4.829		4.092		4.536
10%	bt-type	0.023	4.872	0.516	4.159	0.200	4.574
	whole	-0.221	4.829	0.328	4.155	0.371	4.625
15%	by-type	0.468	4.981	0.611	4.200	0.855	4.695
	whole	0.156	4.863	-0.199	4.231	-0.026	4.672
20%	by-type	0.402	5.008	0.620	4.238	-1.201	4.770
	whole	-0.170	4.944	0.129	4.227	-1.158	4.699
		Livestock Farms					
clean			6.770		5.596		9.527
10%	by-type	0.125	6.798	-0.885	5.575	0.688	9.471
	whole	0.687	6.800	-0.487	5.532	-0.093	9.515
15%	by-type	0.234	6.829	0.156	5.523	0.346	9.325
	whole	0.789	6.797	0.646	5.533	-1.666	9.227
20%	by-type	1.526	6.920	-0.370	5.538	0.654	9.250
	whole	1.136	6.830	-0.051	5.495	-0.354	9.565
		Mixed Farms					
clean			7.433		7.190		6.993
10%	by-type	0.570	7.519	-0.549	7.175	-0.092	7.029
	whole	0.093	7.507	-0.715	7.132	-0.009	7.027
15%	by-type	0.219	7.404	0.957	7.150	-1.437	7.143
	whole	0.115	7.407	1.142	7.107	-1.335	7.152
20%	by-type	0.984	7.541	-1.108	6.984	-0.599	7.010
	whole	1.303	7.595	-0.927	7.001	-0.576	7.050

Table 2

Correlation Coefficients between Match and Imputation Variables[a]

Farm Type	Imputation variable	Match variables		
		FEED	SEED	INCOME
whole	UTIL	0.46	0.39	0.50
	AUTO	0.34	0.18	0.50
	TAX	0.10	0.16	0.27
crop	UTIL	0.13	0.57	0.69
	AUTO	0.25	0.28	0.65
	TAX	0.18	0.19	0.48
livestock	UTIL	0.64	0.25	0.51
	AUTO	0.41	0.47	0.52
	TAX	0.13	0.25	0.28
mixed	UTIL	0.55	0.49	0.76
	AUTO	0.48	0.46	0.73
	TAX	0.24	0.45	0.55

[a] The coefficients are based on unweighted data from the 1984 NFS core sample in Alberta.

seems negligible at low rates of non-response. As the non-response rate rises, the bias grows but is still not substantial. Except for the variable TAX, the differences between the estimates seldom exceed the 95% confidence limits. In the case of TAX, statistical significance, when detected, is usually at the 15% and 20% non-response rates. Unfortunately, the average estimates for the variables UTIL and TAX do show a pattern of consistent, positive bias. No explanation is obvious for this observation and further investigation is warranted to uncover the potential source of bias.

Thus, there is no need to transform match variables by imputation groups defined by farm type for the imputation studied; transforming match variables using the whole sample leads to very similar survey estimates. This may not be the case for other imputation rules and patterns of non-response that are not random. These are topics for future studies. Although the imputed estimates compare well with the clean estimates in practical terms, however,there may still be some unknown sources of bias. These sources, if they exist, may be related to this imputation method, to the imputation rule examined in this study or some other unidentified factor. It is suggested that the presence of bias be confirmed and if confirmed, its source determined. Further study is recommended to this end as well as to aid in determining future imputation rules for the National Farm Survey.

5. ACKNOWLEDGEMENT

The authors would like to thank Nanjamma Chinnappa and Allison Miller for their suggestions and support during the course of the study. We are also thankful for the comments of the referee of this paper.

REFERENCES

BUHLER, S. and BUHLER, R. (1978). *P-Stat* 78 *Users's Manual.* P-Stat Inc., Princeton, N. J., U. S. A.

DAVIDSON, G. (1984). 1983 National Farm Survey. Note on the sample design and estimation procedures. Working Paper, Institution and Agriculture Survey Methods Division, Statistics Canada.

DAVIDSON, G., and INGRAM, S. (1983). Methods used in designing the National Farm Survey. *Proceedings of the Section on Survey Research Methods of the American Statistical Association,* 220-225.

KOVAR, J. (1982). A closer look at the nearest neighbour/hot deck imputation methods: An empirical study. Working Paper, Institution and Agriculture Survey Methods Division, Statistics Canada.

PHILIPS, J. (1979). Imputation techniques used for the F.E.S. Working Paper. Institution and Agriculture Survey Methods Division, Statistics Canada.

SANDE, G. (1976). Searching for numerically matched records. Unpublished manuscript, Business Survey Methods Division, Statistics Canada.

SANDE, G. (1981). Descriptive statistics used in monitoring edit and imputation process. *Proceedings of the* 13*th Symposium on the Interface.* Pittsburgh, Pennsylvania.

STATISTICS CANADA. (1982). *The Numerical Edit and Imputation Subsystem for P-Stat - A User's Guide.* Special Resources Subdivision, Systems Development Division, Statistics Canada.