CAUSES OF INCOMPLETE DATA, ADJUSTMENTS AND EFFECTS

RICHARD PLATEK[1]

STATISTICS CANADA

1. INTRODUCTION

During the past several years, the number of surveys, as a means of providing estimates on a variety of subjects, has greatly increased in most countries, including Canada. The reliability of survey estimates is governed by many factors, one of which is the effect of nonresponse and inconsistent or incomplete data. Any survey, whatever its type and whatever the method of collecting data will suffer from nonresponse for the following reasons; a) not all units of the population were included in the frame; b) units selected and classified as eligible could not be found or c) they refused to participate in the survey. Apart from nonresponse, there are records which are either partly completed or contain invalid responses.

A question has been frequently raised whether, ignoring nonresponse and incomplete data, survey estimates based on the information provided by the responding units only would satisfy the purposes for which a survey was designed. For example, in estimating an item believed to account for, say, 15 percent of a population, what would be the possible effect on the estimate if the nonresponse rate was 15 or 20 percent? To what extent could the potential bias due to 15 or 20 percent nonresponse outweigh the error due to sampling?

Most practicing statisticians or data analysts recognize measures of nonresponse and incompleteness in data as an important indication of quality of data since it affects the estimates by introducing both a possible bias and an increase in sampling variance due to a reduction in the effective sample size. The relationship between the bias and the size of nonresponse is less obvious since it depends on both the magnitude of nonresponse and the differences in the characteristics between respondents and nonrespondents.

[1] R. Platek, Director, Census and Household Survey Methods Division, Statistics Canada.

It is generally believed that in many large-scale surveys the errors due to nonresponse and incomplete data, if measured, would greatly exceed those due to sampling, at least for not too detailed disaggregations. But in most surveys only sampling error is identified and the other components are inadequately recorded and analyzed. Yet they are potential sources of biases and to disregard their effect on the estimates could lead to survey results of unacceptable quality. Therefore, the reduction of the effects of nonresponse and incomplete or invalid responses is very important and it should be undertaken at various stages including survey design, data collection, processing and estimation stage.

One way of dealing with nonresponse, after the data collection, is through methods of imputation or adjustment of weights at the processing and estimation stage. While adjustments for nonresponse may be more or less effective in reducing bias, well designed data collection operations will keep nonresponse at an acceptable level and at reasonable cost, thus minimizing the necessity for the application of adjustments.

In this paper, nonresponse, incomplete data including invalid response in household surveys will be discussed with respect to their origin, various methods of reducing them,as well as adjusting for them in the final estimates.

2. NONRESPONSE

Nonresponse may be defined as a failure to obtain a usable report from a reporting unit which legitimately falls into the sample in a particular survey and it may be one of two kinds:

a) Unit nonresponse where survey questionnaire is not obtained for a designated unit.

b) Item nonresponse, where a survey questionnaire is obtained for a unit, but responses for one or more questions are not obtained.

Nonresponse occurs because of operational difficulties, time and cost restraints, a lack of co-operation from respondent, the inability or unwillingness of interviewer to track down missing respondents, or for some other reason. The severity of nonresponse problems, are measured by nonresponse rates which are calculated as the percentage of nonrespondent households among all sampled households.

Nonresponse rates vary considerably from survey to survey. In some surveys they are as high as 40 percent or more, in other surveys they may only be about 4 percent or so. Whether nonresponse rates are too high or too low depends on the purposes of the survey. If the objective of a survey is to estimate an item which accounts for 15 percent in the population then, depending on the reliability with which the 15% is to be estimated, even a 5% nonresponse can have a major impact on the estimate, particularly if the characteristic of nonrespondents is correlated with the important variables. On the other hand it does not necessarily mean that a survey with a high nonresponse rate may not provide useful information. For example, suppose that the objective of a survey is to find out whether 15 percent of a population would buy a particular product. If in the survey 20 percent responded that they would buy it, 30 percent would not and 50 percent were nonrespondents, then the objective of the survey has been met even if all the nonrespondents were in the category of those who would not buy the product. In general, however, the higher the nonresponse, the higher the possible bias in the estimate and the less likely it is that the objectives of a survey can be satisfied.

The size of nonresponse cannot be simply resolved by starting with an excess sample to allow for the potential nonresponse since in the presence of nonresponse, the sample is no longer a probability sample. The difficulty lies in the fact that the nonrespondents in some ways and to varying degrees are different from those who respond. We can assume that every individual (if selected) is a potential respondent i.e. the individual i, when in the sample, will respond with probability δ_i and will be nonrespondent with probability $1 - \delta_i$. If the response probability δ_i is the same for all i, the situation is easily remedied by adjusting (inflating) the weights of the respondents. But the probability of response may depend on the characteristic of interest and adjustment of respondent weights to account for the nonresponse will give rise to nonresponse bias. The magnitude of the nonresponse bias will depend on the relationship between the characteristic of interest and the response probability. The problem can be illustrated by the following simple examples under different sample designs:

a) Single Stage Simple Random Sampling:

From a population of N individuals, n are selected with SRSWOR and the subject of inquiry is some quantitative variable Y (e.g. income, no. of accidents, etc.). Suppose that the probability of obtaining a response decreases as the value of y increases, had there been no nonresponse, the total $Y = \sum_{t=1}^{N} y_t$ would be estimated unbiasedly by $\hat{Y} = \frac{N}{n} \sum_{i=1}^{n} y_i$, but Y is observed for n* (<n) and $\tilde{Y} = \frac{N}{n^*} \sum_{i=1}^{n^*} y_i$, is used to estimate the total Y. Thus $\tilde{Y}$ would have been unbiased if the response probability did not depend on y values, but in this situation, since response probabilities are low for large y values Y underestimates the total Y; i.e., we have a negative correlation between response probability and characteristic, resulting in a negative nonresponse bias.

b) Single-Stage Probability Proportional to Size (PPS) Sampling: From a population of N units (Firms) n are selected with PPS to estimate $Y = \sum_{t=1}^{N} y_t$, where y is some quantitative attribute (e.g. production, etc.) and p_t is the measure of size for unit t, t=1, 2 ..., N; $\sum_{t=1}^{N} p_t = 1$, where $\Pi_i = np_i$ = Probability that unit i is in the sample.

y_i = observed value for unit i if selected and responding. In the absence of nonresponse $\hat{Y} = \sum_{i=1}^{n} \frac{y_i}{\Pi_i}$ is an unbiased estimate of Y. Suppose that y values are observed for n^* (<n) out of n selected, then $\tilde{Y} = (\frac{n}{n^*}) \sum_{i=1}^{n^*} \frac{y_i}{\Pi_i}$ is used to estimate Y, and will be unbiased if the response probabilities are correlated with the ratio Y_i/P_i. The bias will likely be small compared with the total Y if the ratios Y_i/P_i have small variability. Usually PPS sampling design is employed when there is a large positive correlation between the Y_i values and the sizes P_i's of the units. Often this high correlation results in a small variation in the ratios Y_i/P_i when the regression of Y_i on P_i passes through the origin.

c) Multi-Stage Stratified PPS Sampling

In household surveys, the households are selected using multi-stage stratified PPS sampling design and the design is usually self-weighting. In some surveys, the response can be obtained from onemember of the household about all the eligible members in the household. It is conceivable that the nonresponse will tend to be higher in a single person household as compared to the households with two or more members. Under such conditions, if the nonresponse adjustment in the weight fails to take into account the size of household, the resulting estimate of the population will be biased upward. In this context consider another example. Suppose that the subject of inquiry is a qualitative variable i.e. presence or absence of a particular characteristic. Suppose further that the probability of response is high if the characteristic is absent and low if the characteristic is present. If an estimate of the number of persons with the characteristic is obtained by adjusting the weights of the respondents, then the result is an underestimate i.e., we have a negative correlation between response probabilities and the characteristics.

Ideally to remove the nonresponse bias, one would simply weight up each sample response by the inverse of the product of the selection and the true response probabilities of each responding unit. This is however, an impossible task since the true response probability is unknown for each unit. In practice we employ average response probabilities estimated by response rates to adjust the weights in the estimates. The effect of this procedure on the reduction of nonresponse bias will depend upon the degree of relationship between response probabilities and response rates.

The most desirable way of dealing with the effect of nonresponse is to minmize the size of it. However, any systematic attempt to control the size of nonresponse must be based on clear understanding of any it arises. Basically, the causes and the size nonresponse are related to (i) type of survey, (ii) data collection methods, and (iii) sample design. But even for a given survey design the magnitude of nonresponse will be influenced by factors such as type of area, type of nonresponse, etc. To illustrate this, a brief examination of nonresponse will be provided for the Canadian Labour Force Survey.

The Canadian Labour Force Survey is carried out as a monthly probability sample of dwellings. Households within the selected dwellings are interviewed once a month for six consecutive months. In one particular week (called survey week) each month 56,000 dwellings throughout Canada are contacted by approximately 1,100 interviews. Information is collected by the interviewers on the demographic characteristics of labour force activities of the civilian, non-institutional population 15 years of age and over who are members of households belonging to these dwellings.

In the Canadian Labour Force Survey a detailed record is kept of total nonresponse which may be broken down into a number of components each of which has a different cause and requires a different treatment. For example, in LFS or in any household survey, one can recognize the following components (a) Household Temporarily absent (b) No one at home (c) Refusal (d) No interviewer available (e) Bad weather conditions. etc. The size of nonresponse due to the latter two being of minor general significance.

Table 1 provides an example of non-response rates in LFS followed by a brief discussion of various components of total nonresponse rates.

TABLE 1

Nonresponse Rates (%) According to Tenure of Households in the LFS (July 1977 to June 1978)

Number of Months in Survey	Nonresponse Rates (%)			
	Total Non-response	Refusal	No one at home	Temporarily absent
1	8.04	1.43	2.96	2.94
2	5.09	1.21	1.44	1.99
3	4.71	1.32	1.10	1.90
4	4.65	1.46	1.09	1.79
5	4.62	1.51	0.99	1.77
6	4.45	1.52	0.78	1.73

On the basis of the results shown on Table 1 the following comments can be made:

(a) The total nonresponse rate was highest during the first month, presumably because interviewers had more difficulty finding people at home having not yet determined the best time to call as one may observe in the higher "No one at home" rate, for example. The rate then decreased sharply in the second month and continued to decrease through the third and fourth months.

(b) The refusal rate decreased in the second month, increased gradually through the third, fourth and fifth months and levelled off in the sixth month.

(c) The "No one at home" rate decreased sharply from the first month to the second month by roughly 50 percent. It continued to decrease from the second month to the third month but decreased very gradually through the fourth and fifth months. A larger decrease then occurred in the sixth month. The behaviour of the "No one at home" rate over the six month tenure of households in the survey was most probably due to the fact that the longer a household is in the survey the more familiar the interviewer becomes with knowing when the respondent is most likely to be at home.

(d) The "Temporarily absent" rate decreased through all six months, particularly from the first to second month. It is difficult to explain this phenomenon since the "Temporarily absent" rate should not be expected to depend on how long a household remains in the survey. One can hypothesize that interviewers may have confused "No one at home" and "Temporarily absent" types of nonresponse.

TABLE 2

Nonresponse Rates (%) by Type of Area

(Monthly Average : 1978)

Type of Area	Approximate proportion of sample	Total Nonresponse	Refusal	No one at Home	Temporarily absent
NSRU[1]	0.48	5.0	1.2	1.3	2.0
- urban[2]	0.18	5.3	1.0	1.4	2.4
- rural[2]	0.30	4.9	1.3	1.3	1.8
SRU[3]	0.51	5.7	1.7	1.6	2.0
- built-up[4]	0.37	5.5	1.6	1.5	2.0
- fringe[4]	0.10	4.8	1.6	1.2	1.8
- apartment[5]	0.04	9.8	2.7	3.6	3.0

Within SRU's built-up areas had a higher total nonresponse rate than fringe areas due to higher "No one at home" and "Temporarily absent" components. Thus, it appears that people living in the core areas of cities tend to be more difficult to contact than people living in the fringe areas; the differences, however, were not large.

SRU apartments had a higher total nonresponse rate than any other area shown in Table 2. In fact, the total nonresponse rate in the SRU apartment sample was almost twice the rate in the SRU non-apartment sample (consisting of both built-up and fringe areas). The refusal,

[1] Non-Self-Representing Units are the areas outside SRU's and contain rural and small urban centres.

[2] Every primary sampling unit in an NSRU is divided into an urban and a rural portion.

[3] Self-Representing Units are cities whose population exceeds 15,000 persons or whose unique characteristics demanded their establishment as SRU's. Every SRU is selected with certainty.

[4] SRU's are stratified into sub-units, and sub-units are classified as "built-up" or "fringe" on the basis of their potential for future growth. Generally speaking, SRU fringe households belong to the fringe or suburban areas.

[5] In seventeen large cities across Canada there is a separate frame of apartments in buildings having at least five storeys and thirty or more units.

"No one at home" and "Temporarily absent" components were also highest among apartments. As yet this category is small but it is growing quite rapidly. The "No one at home" rate was almost three times higher in the apartment sample than in the non-apartment sample. This large difference may be due to the different lifestyles of apartment and non-apartment dwellers. Apartment households usually consist of single persons or small families who tend to be more mobile and difficult to find at home, while non-apartment households are more likely to contain larger families with children.

Within NSRU's the total nonresponse rate was higher in the urban portion due to higher "Temporarily absent"[1] rates among NSRU urban households. The "No one at home"[2] rates in the urban and rural portions were roughly the same, but the refusal rates of 1.3% averaged 30% higher in NSRU rural areas than 1.0% as observed in NSRU urban areas.

In this section I have dealt mainly with total or unit nonresponse. The problem of item nonresponse, defined at the beginning of the section are discussed briefly under Data Collection and in the next sections. Incomplete and invalid responses are discussed in the section on Imputation.

3. DEALING WITH NONRESPONSE

Survey Design basically consists of three steps a) Sample Design b) Data Collection c) Estimation.

None of these steps can be undertaken on purely technical grounds or on purely practical grounds. The survey design is decided upon in the light of what is practically feasible and theoretically desirable in order to meet users' requirements. The importance of nonresponse through its effect on survey estimates is an integral part of survey design and can hardly be left to chance. Provisions must be made at every possible stage of a survey in order to control the size of nonresponse and to minimize the effect of the final nonresponse.

[1] The household was absent for the entire survey week.

[2] The occupant could not be contacted after several attempts.

Although the actual nonresponse occurs during the data collection stage its size can be greatly influenced at the planning stage by examining the possible effect that various design factors may have on nonresponse.

Also, careful and appropriate preparation for data collection with respect to methods of interviewing and motivation of respondents and interviewers will considerably affect the magnitude of nonresponse.

At the processing and estimation stage an attempt is made to minimize the effect of nonresponse on the final estimates by imputing for missing values.

Let us consider how the effect of nonresponse can be influenced at each of these stages.

Sample Design (Planning and Development)

At the planning stage, an awareness of the effect of nonresponse on the Mean Square Error (MSE) of survey data may lead to a survey design which would influence the size of nonresponse. One of the important factors in planning a survey is a decision on the tolerance level of nonresponse and an experienced survey designer can estimate fairly accurately the level of response for a particular survey that can be expected under various survey conditions. For example, for national estimates with a large sample size, the effect of nonresponse on sampling and response variance is likely to be unimportant and the bias is the likely predominant component of MSE. However, for subnational estimates the variances are likely to be large so bias might be relatively less important.

The survey cost is another important item which will affect many factors in survey development including nonresponse. It is important to balance the other factors against the cost so as to achieve a nonresponse rate sufficiently low to serve the goals of the survey. It should also be realized that within reasonable limits, it is sometimes better to accept a somewhat smaller sample than originally planned and to transfer the resources to appropriate data collection, follow-up and estimation procedures. This would be particularly advantageous if the survey designer suspected large differences between respondents and nonrespondents in their characteristics.

In survey planning and development a number of factors should be taken into account in arriving at the final design. These factors can be classified into three groups:

Group I
a) sample size
b) stratification
c) degree of clustering
d) sample allocation
e) method of selection

Group II
a) sample frame
b) method of interviewing
c) selection, training and control of staff
d) length of questionnaire and wording
e) sensitivity of questions
f) type of area in which the survey is taken
g) feasibility and cost of call-backs
h) publicity

Group III
a) edit and imputation
b) estimation
c) variance estimation

All of these operations affect the Mean Square Error which provides a measure of reliability of data.

Let us suppose that the Mean Square Error can be decomposed into the following components:

$$MSE = V_S + V_R + V_{CR} + (B_S + B_R)^2$$

where

V_S = sampling variance

V_R = simple response variance

V_{CR} = correlated response variance

B_S = sampling bias

B_R = response bias. (including nonresponse bias)

Sampling variance (V_S) and sampling bias (B_S) are affected by all the factors in Group I, Group III and also by the size of nonresponse. It is important to note that each survey determines its own requirements with respect to design, questionnaire and methods of interviewing. For example, an unclustered survey design may produce a higher nonresponse rate than a clustered design. This may be due to the requirements for extensive travelling in the case of personal interviews where for reasons of cost repeated callbacks must be restricted. The larger the size of nonresponse, the greater the effect it has on sampling variance and nonresponse bias. Non-sampling components of Mean Square Error (V_R, V_{CR}, B_R) components are affected by all the factors in Group II. Furthermore, it is quite evident that all the factors in Group II are a potential source and cause for nonresponse. For example, we have seen that nonresponse rates depend on the type of area in which the survey is taken. Length of and sensitivity of questionnaire will undoubtedly affect the size of nonresponse. Thus if careful attention is paid to the factors in Group II at the design stage, serious nonresponse problems may be avoided.

Data Collection

In discussing approaches to minimizing nonresponse, one can distinguish between two 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 respondent but an acceptable response is not obtained.

The "No contact" type of problem is of course usually attacked with operational solutions. In a telephone or personal interview the time and patterns of calling on the respondent are important. The size of assignment and the time allotted to data collection must be adequate. In a mail survey, ensuring correct addresses on the mailing list, efficient follow-up procedures and convenient materials are all essential. The size of nonresponse due to "No one at home" or a "Temporarily absent" provide an important indication of the operational problems.

The existence of refusals presents a different set of problems. It should be conceded at the outset that refusal rates are not always as straightforward as one might expect. An interviewer may prefer to record a refusal as a "No one at home" or a respondent may simply not answer the door as a means of refusing and thus being recorded as "No one at home". In a mail survey one is not always certain that the respondent received the questionnaire and if he has received it whether he simply neglected to mail it. Thus, the distinction between the true nonresponse and other causes is not easily established. The invalid response presents still a different set of problems since an inexperienced interviewer may not realize that the data is invalid or illogical until an edit routine has discovered it. Also, the interviewer may carelessly code the response in an incorrect location on the questionnaire resulting in invalid data which must be discarded. In any event, regardless of how nonresponse is recorded, the problem seems to be to motivate the appropriate respondent to produce a valid response.

With respect to motivation, let us look upon the respondent as being neutral towards the survey and consider the influences which may motivate him either to respond or not to respond. Such factors as difficulty in understanding questions, use of respondent time, privacy, indifference, difficulties in recalling information, embarrassing or personal questions are all examples of motivation not to respond. On the other hand, examples of motivation to respond are an interest in the survey, willingness to help out, duty, understanding of the importance of survey results, etc.

The problem becomes, how to accentuate the positive motivation and reduce the negative motivation until the balance swings in favour of response. The key element is the respondent and anything which affects his ability and motivation to respond must be of interest and concern to a survey designer.

Introductory letters, examples of the uses of the data, and brochures describing the objectives and authority for the survey, are often excellent means of avoiding hostility and distrust.

Invasions of privacy is related to the content of the questionnaire although the reaction of different respondents is quite variable. Many procedures exist for minimizing the effect on the respondent and the specific procedures should be tailored to the given situation. In some cases, it may be best to allow the respondent to reply in a completely anonymous fashion. This can be accomplished by self-enumeration with no identification whatsoever on the questionnaire. Quite often, though, it is essential to have some area code or sample designation for weighting and estimation purposes of follow-up and in that event care must be taken that the respondent does not perceive this as a means of identifying his replies.

In addition to the assurance of privacy, some forms of compensating the respondents for their time and effort have been practiced by some survey taking organizations.

(a) Substitution In the Field - One method of dealing with nonresponse at the data collection stage is to substitute other previously unselected units in the field. It must be emphasized however that this is still nonresponse and substitution is a means of imputation. There are two basic types of substitution that are used.

a. Selection of a random substitute.
b. Selection of a specifically designated substitute.

With a random substitution method, an additional population unit is selected on a probability basis to replace each nonrespondent. For many random substitution procedures, potential substitutes are selected prior to the data collection in order to avoid any delays and problems that could exist if the substitutes were selected during or after the collection of data.

In a procedure that uses specially designated substitutes (for example, a next-door neighbour), the intent is to find a substitute similar in characteristics to those of the nonrespondent. Unfortunately, this could lead to a sampling bias, especially if the neighbour lives outside the sample frame. While any original unit may be selected with known probability according to the sample design, substitution of other previously unselected respondents to replace uncooperative respondents in some uncontrolled manner or even in a controlled manner will alter the inclusion probabilities. A sampling bias of unknown magnitude could be introduced (since the selection probabilities are unknown). While the sampling variance may be reduced because of an increase in the effective sample size, there would probably be no reduction in either the response or nonresponse bias. Even if the inclusion probabilities could be calculated some nonresponse bias would remain since the uncooperative units essentially have no chance of inclusion. The key question regarding the worth of substitution procedures is whether or not the use of substitution provides better proxy values for nonrespondents than those provided by alternative imputation procedures. Undoubtedly, there are some advantages and disadvantages to the use of substitution procedures. The first advantage is that it is a convenient way of balancing the sample with respect to sample size. The other is the reduction of the sampling variance due to increase in the effective sample size.

One of the major disadvantages of the use of substitution is a tendency to use it rather than making every effort to obtain responses from original units. Thus the use of substitution procedures requires that appropriate control should be taken to ensure that maximum effort is made to obtain responses from the original sample units. Another disadvantage is that there is a tendency to ignore the level and the frequency of substitution when the survey response rate is calculated.

b. Callbacks - In many surveys, callbacks are extensively used in order to reduce nonresponse and the resulting biases. The callbacks may take a variety of forms depending on the type of survey. In mail or interview surveys callbacks may be a letter, a telephone call or a personal interview. In telephone surveys and in interview surveys repeated calls are the normal form of callbacks. There is a need to study various types of callbacks with respect to quality of data, cost and respondent reaction to them.

Callbacks may be used solely to reduce nonresponse or to provide input to an imputation system or to study the effect on quantity.

c. Respondent Rule - In order to avoid any ambiguity as to the eligibility of respondents in a given survey a procedure referred to as "Respondent Rule" should be defined and followed in the field.

For Surveys which involve a designated respondent, two rules are most often used.

1. The designated respondent is to be interviewed and he/she is capable to respond but unavailable, repeated callbacks are made until contact is established.
2. If the designated respondent is not present or not capable to respond (deaf, ill, etc.) a proxy respondent is chosen. Variants of this rule involve different definitions of permissible proxy respondents.

For surveys in which responses for each eligible household member are required one of several possible respondent rules is followed:

1. Every member of the household is to respond personally (self response).
2. One member of the household may answer for every member of the household (proxy response)
3. A mixture between self-response and proxy response i.e. some are self-responses others are proxies depending on the respondents availability at the time of interview or some specific respondent rule e.g. persons unrelated to the head of the household must respond for themselves.

Methodological investigations of the effects of using various respondent rules have focused on two basic areas. The first involves the differences in the number of callbacks needed to contact the desired respondent proposed by Deming and Cochran 1977; the second involves differences in the quality of the data obtained which can be evaluated through a program of reinterviewing of the original respondents.

The use of a proxy respondents diminishes the number of callbacks thus reducing the cost of survey and timeliness of obtaining the data. On the other hand, there may be a disadvantage to the use of proxy in that the data provided by proxy respondents may be less accurate than that obtained from self-responses. The use of a particular respondent rule should be very carefully examined in relation to the type and quality of the data required, cost involved in obtaining the data and timeliness for publications. Those considerations will vary from survey to survey depending on the survey topic, budget and field organization.

Imputation

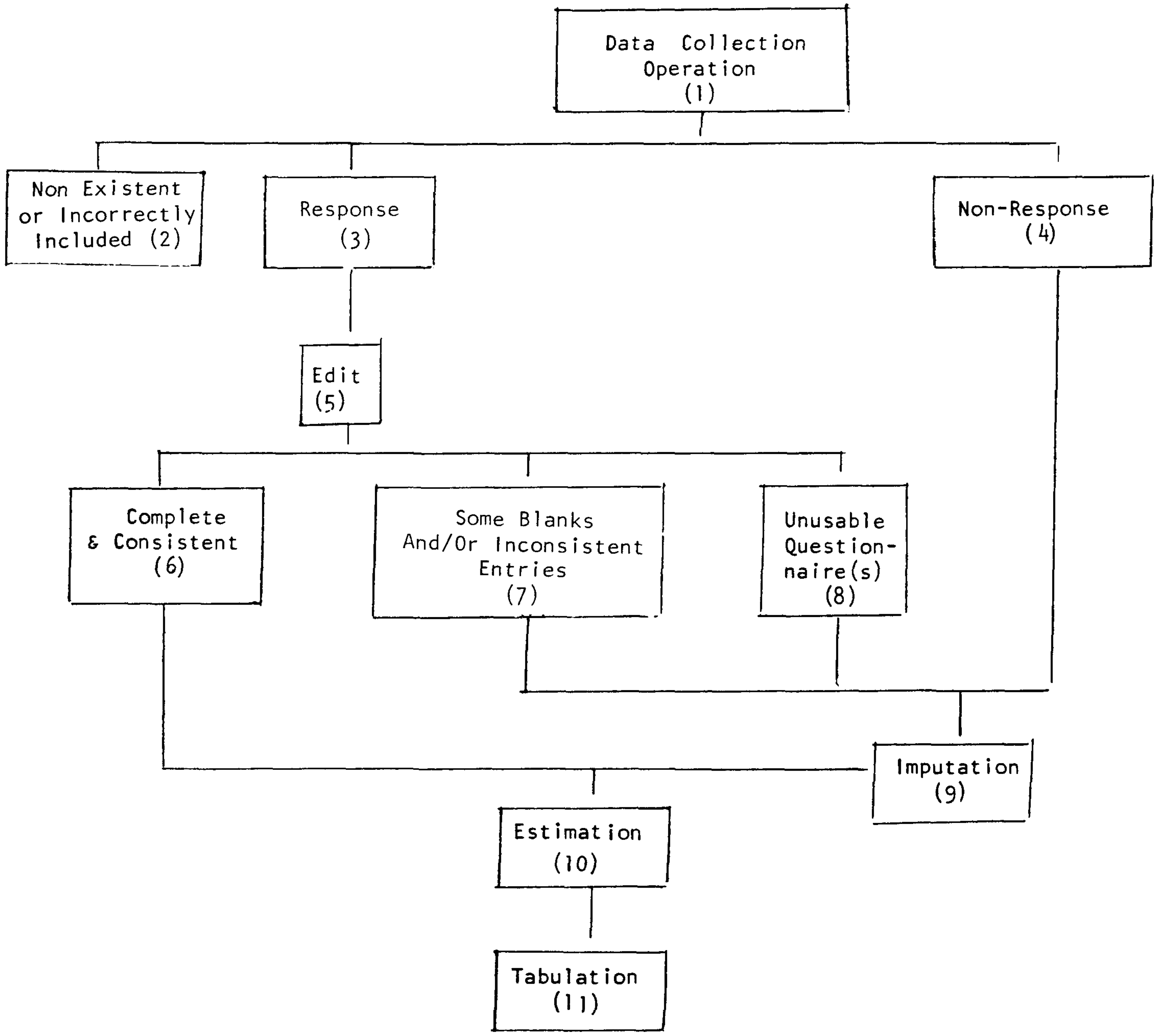
At the Processing and Estimation stage survey data is usually classified according to total nonresponse, partial nonresponse and invalid response.

It is very important to have an effective control system incorporated into the survey design, i.e. to ensure that the selected units (and no others) are interviewed, that non-reporting units are properly classified e.g. nonresponse, non-existent, that of gaps in the frame are identified, that data entry is complete, etc.

There are various ways of dealing with incomplete or invalid responses. Each of them results in assigning a value for the missing or invalid data, unless a decision is made to publish "raw" data. The procedure of assigning the value is called imputation, and some imputed value is assumed to refer to the characteristic of the nonrespondent. Thus a "clean" data set is produced, that is, a value is given to each unit in survey. Before proceeding to discuss various imputation methods let us examine conceptual issues of imputation.

As the information flows from data collection to tabulation, the various types of responses can be identified and are presented as follows in Chart 1.

Chart 1: Flow Chart Pertaining to Each Sampled Unit


Data Collection Operation (1)
Non Existent or Incorrectly Included (2)
Response (3)
Non-Response (4)
Edit (5)
Complete & Consistent (6)
Some Blanks And/Or Inconsistent Entries (7)
Unusable Question-naire(s) (8)
Imputation (9)
Estimation (10)
Tabulation (11)

This is, of course, a highly simplified diagram of the process and it is produced only for the purpose of the discussion of this paper.

From Chart 1, it can be seen that two of the three groups of questionnaires following the edit stage require further action prior to estimation. These are (8) the unusable questionnaire, (7) those containing some blanks and/or inconsistent entries and (4) containing nonresponse. An unusable questionnaire could be classified as total nonresponse or it can be associated with a respondent household with some blank or inconsistent entries. In either case, however, further action denoted by (8) Imputation would be required. Complete nonrespondents are usually weighted up in some manner with the exception for the Census. The deficient questionnaires, on the other hand, fall into two categories such as (7) inconsistent entries or illegitimate blanks.

The inconsistent entries can be either logical impossibilities or they can be plausible but highly unlikely. It seems natural that if the entries are logical impossibilities and they can be detected as such, they ought to be adjusted even though they may not affect the data to any great extent. The adjustment would eliminate a great deal of embarrassment to subject matter analysts associated with the published reports.

In the case of plausible but highly unlikely entries, one is faced with a difficult choice between leaving what might seem to be an unnatural distribution or removing the extreme values of the distribution which may actually represent the real life situation. Ideally, one ought to opt for one or the other choice on the basis of experience with error mechanisms and the nature of the substantive distribution based on the knowledge of subject matter. In any case, one has to be able to identify the problem cases, i.e. one has to have suitable edit rules whenever one encounters impossible or highly unlikely events and a method of dealing with them (i.e. imputation).

There is a fundamental distinction between editing and imputation. Let us consider the set of all possible code combinations on a questionnaire. Editing can be defined as the division of this set into two mutually exclusive subsets: Those combinations which are judged acceptable and those which are unacceptable, the latter including questionnaires with invalid blanks and inconsistent entries. Thus, editing is basically a diagnosis and operationally it must be defined by a set of rules. Imputation, on the other hand, is more in the nature of a treatment of data.

Imputation may be defined as the assignment of data to empty fields (including total nonresponse) or a replacement of invalid data in fields following a certain set of rules. There is no known unbiased method of imputing but some methods may be more suitable than others.

It is possible that, rather than imputing for nonresponse at the time when survey tabulations are prepared, tabulations could be presented with the amount of nonresponse reported. In this case, the users would have a choice among various methods of imputation from tabulated data. At the first glance, this approach would appear to have some advantages giving the users the opportunity of selecting their own method of imputation. There are, however, some serious disadvantages. Conflicting estimates would be produced by various users due to the different imputation methods employed and, problems in the consistency and integration of data would be created. As well the data collection agency is usually in a more better position, due to its proximity to the sources of the data, to make imputation decisions. For these reasons, imputation is normally carried out by data collecting agencies rather than by individual users. The whole philosophy of imputation is based upon the expectation that an appropriate procedure, whether for nonresponse or for blanks resulting from edit failure, will provide a more logical relationship between cross-classified data and will also lower the mean square error of estimates.

The simplest situation occurs when there is only one possible value which can be imputed for a field in such a way that after the imputation the record will be consistent. This is what is called deterministic imputation. For example, if wife is coded "male" then there is only one possible value to impute for sex to make it consistent with information. Sometimes, there may be more than one value which would make the record consistent. If this is the case, one would choose a particular value which is more predominant in proportion to the total frequency or is more plausible. A good example of this kind can be found in the Labour Force Survey where in the fall to spring months, for 15 and 16 years old persons, if there is no Labour Force characteristic entered, one imputes that they are "attending school", although it is not at all impossible they do not attend school. So long as the proportion of such cases is sufficiently small, the effect of this imputation will be a slight increase in bias, but there will be some reduction in variance.

In other situations where one could reasonably impute a whole range of values, one needs some other criteria. One possible criterion would be to minimize the mean square error of the resulting estimates. The question, however, arises, the mean square error of which estimates? With the continuously increasing demand for micro data tabulated in a number of different and unforeseen ways one really does not know which mean square error one ought to minimize. Furthermore, one would not know all the kinds of aggregates to which a particular record may contribute in different kinds of tabulations. Consequently, one might prefer to use some other criterion which would produce the most appropriate entry for a field in a particular record in relation to the other information in the record. In other words, how can one best predict the value of one field on the basis of knowing the other fields on the record. A good example of this kind of imputation is the use of previous month's data in the Labour Force Survey; for a particular person, one could hardly find a better imputed value, particularly in those cases where demographic characteristics change slowly. If one does not have information based on the past, one would have to resort to such methods of imputation as regression or hot deck.

4. Imputation Procedures

For a number of years various procedures of imputation for missing data due to nonresponse have been used in household surveys and censuses. The use of a particular procedure has been, to my knowledge, mostly justified on the ground of expedience, intuition and experience. It was often assumed that the probabilities of units responding were uniform and the nonresponse bias was largely ignored.

Although variations in response rates have been detected among units according to their characteristics, the effect of individual units responding or not responding upon the bias and variance of the estimates has usually been insufficiently examined.

To facilitate a detailed examination of the effect, Platek and Gray (1980) have developed methodology with respect to the bias and variance pertaining to several imputation procedures. The development of the expression of bias and variance of the estimates is based on a fundamental concept that a unit, if selected, responds or does not respond with a certain response probability attached to that unit. This is an extension of the approach taken by Platek, Singh and Tremblay in 1978 with respect to censuses.

The definition of various imputation procedures involve the following:

(i) the use of cells for imputation; the cells may be either balancing areas or weighting classes, or

(ii) adjustments in weights using estimated response probabilities within the cells.

Balancing areas are frequently referred to as "design-dependent balancing areas" for imputation purposes. A balancing area is a geographic area in which a deficient sample arising from missing data is enlarged to the prescribed level by means of imputation for missing data. Commonly, a balancing area is a stratum, but it could be other design-dependent areas, such as primary sampling unit, cluster, groups or strata or even the entire sample. The balancing areas may be delineated before or after the survey is taken.

Weighting classes are defined by post strata (strata defined after sampling) formed on the basis of information pertaining to respondents and nonrespondents in the sample. The information may be obtained from those partial nonrespondents for whom some characteristics are known even though the particular characteristic being estimated is not known for these units as in the case of item nonresponse for example. The characteritistics used in the post-strata could also be obtained from external sources. From the operational point of view, a weighting class is very similar to a balancing area except that the units having similar characteristics are grouped into classes or post-strata without regard to the original design. The choice of characteristics and the size of balancing areas and weighting classes are important, as the variance and bias of an estimate derived from the sample would depend upon the homogeneity of characteristics between respondents and the nonrespondents within the balancing areas and weighting classes.

The balancing areas and weighting classes are also referred to as cells or balancing units.

(a) "Weighting" in Adjustment Cells

The 'Weighting Method' of imputation applicable in practice to complete nonresponse is one in which the sample weights or inverse inclusion probabilities are inflated by the inverse of the response rate in a cell. Implicitly, the imputed value for the missing data of each nonrespondent is the mean of all responding values in the cell, with some adjustments for selection probabilities.

If a cell "b" contains n_b units in the sample and m_{by} of them responded to a certain question or questions that would determine characteristic y, then the Horvitz Thompson estimate[1] of the total of characteristic y in that cell would be given by (dropping y in m_{by}).

$$\hat{Y}_{b1} = \frac{n_b}{m_b} \sum_{i=1}^{n_b} \delta_{iy} Y_i/\Pi_i = \sum_{i=1}^{n_b} \delta_{iy} Y_i/ (\Pi_i \hat{\alpha}_{iy}) \text{ ; where} \qquad (1)$$

[1] Restricting oneself to this estimator only.

δ_{iy} = 1 or 0 according as unit i responds or does not respond to characteristic 'y'.

$\hat{\alpha}_{iy} = m_b/n_b$ an estimate of the response probability, it may or may not equal the true response probability, given by $E\delta_{iy}$ which is defined by α_{iy}.

Y_i = response for unit i as defined earlier.

An area within which adjustments are carried out would consist of mutually exclusive and exhaustive adjustment cells so that the overall estimates of the total with characteristic y would be given by the sum of the estimates over the cells; i.e.,

$$\hat{Y}_1 = \sum_b \hat{Y}_{b1}. \qquad (1)$$

It can be seen from (1) that $\hat{Y}_{b1}$ may be regarded as a weighted up estimate where the weight for unit i includes the inverse selection probability Π_i^{-1} and the estimate of the inverse probability of responding, given by $\hat{\alpha}_{iy}^{-1}$ or $(m_b/n_b)^{-1}$. Since we rarely know individual response probabilities at the time of processing the data, we must employ the best estimate of response probabilities available and, with a proper choice of balancing areas or weighting classes, that estimate is usually the response rate in the cell or the estimated average of the response probabilities of the units.

It can readily be shown that $\hat{Y}_1$ my be subject to response bias and nonresponse or imputation bias given by:

$$\sum_b \bar{\alpha}_{by}^{-1} \sum_{i=1} \alpha_{iy} B_{Riy}$$

$$\text{and} \quad \sum_b \bar{\alpha}_{by}^{-1} \sum_{i=1} (\alpha_{iy} - \bar{\alpha}_{by}) Y_i \qquad (2)$$

respectively.

Here, $\bar{\alpha}_{by}$, is the expected response rate or average response probability in cell 'b', with respect to characteristic y. $\bar{\alpha}_{by} = \sum_i \Pi_i \alpha_{iy} / \sum_i \alpha_i$, with summation taken over all units in cell 'b'.

B_{Riy} is the response bias for unit i with respect to characteristic y, i.e. $Ey_i | (\delta_{iy} = 1) = Y_i + B_{Riy}$.

The imputation bias will exist only if response probabilities α_{iy}'s vary within an adjustment cell and if a correlation exists between the response probabilities, α_{iy}'s, and the characteristic, Y_i.

(b) Duplication Method

The "Duplication Method" of imputation is one in which the deficiency in the sample in a cell due to nonresponse is made up by duplicating all or a subsample of respondents.

In the duplication method, if we consider a weighting class b with n_b selected units and m_b respondents with respect to characteristic y, the estimate in cell 'b' would be given by:

$$\hat{Y}_{b2} = \sum_{i=1}^{n_b} \delta_{iy} W_{iy} Y_i/\Pi_i, \qquad (3)$$

where W_{iy} = number of times unit i is duplicated to account for the $(n_b - m_b)$ nonrespondents incell 'b' so that $\sum_i \delta_{iy} W_{iy} = n_b$. Note that W_{iy} is defined only for respondents, i.e., when $\delta_{iy} = 1$.

There are several ways of duplicating units. One of these and the simplest to treat from the point of view of methodological development is the random selection of units for duplication without replacement.

(c) Hot Deck

Hot Deck procedures are common methods of adjusting data sets for missing values. In general, a Hot Deck procedure is a duplication process -

one reported value from the sample is duplicated to represent a missing value. Thus, the terms "imputation procedure" and "Hot Deck procedure" are not interchangeable. A procedure which imputes the average of all reported values for each missing value is an imputation but not a Hot Deck procedure.

The primary reason for using a Hot Deck or a Cold Deck procedure is to attempt to reduce nonresponse bias. The essential difference between the two procedures lies in the way the information for missing data is specified. A Cold Deck procedure is deterministic, i.e. in each specified condition the same value is substituted for an item nonresponse. In Hot Deck procedures, for each specified condition, the value substituted for item nonresponse is the value of that item which was encountered in the last "acceptable" record. Thus in Hot Deck the substitution is probabilistic, reflecting the frequency values in the items encountered in acceptable records satisfying the same conditions.

As a method of imputation Hot Deck procedures have some attractive features including the following: (a) the procedures result in a relatively easy way of constructing post-strata, [See I. P. Fellegi and Holt], (b) matching of records does not present any special problems and (c) no strong model assumptions need be made in order to estimate the individual values for missing items.

In evaluating Hot Deck procedures one would like to know how the bias and reliability of the principal estimates are affected by the size of classification groups (often referred to as weighting classes), the frequency of missing data, the choice of matching items etc. Some theoretical work, relating to Hot Decks, has been done by (I.P. Fellegi and Holt) (Bailar and Bailar 1978) and (Cox B and Folsom R.E. 1978)

Future theoretical work lies in attempting to generalize the Hot Deck procedures that have been developed and deriving expressions for the bias and variance of estimates based on the procedures. The bias occurs from the deviation of the estimated response probability of an item pertaining to a unit from the actual response probability (which is of course unknown) and the correlation between the value of the item and the response probability. The variance involves additional components beyond those that occur when weighting is applied in a weighting class. Depending upon the methods and restrictions imposed on the Hot Deck procedure, the additional variance terms may become quite complex.

In any case, an extension of the theory pertaining to the variance so far developed by Bailar and Bailar seems to be the most promising direction to follow.

(d) Historical Data Substitution Method

The "Historical Data Substitution Method" is one in which historical or external sources such as Census, earlier survey or adminstrative data are substituted for a unit to replace missing data caused by nonresponse. The following two cases of "Historical Data Substitution Method" procedures may be considered:

Case (i) one type, where historical or external source data are available for all units which have failed to respond, and

Case (ii) another where the external source data is available for some but not all units, and imputation by another method, e.g. "Weighting" must also be applied.

When external source data are available for every unit, then the imputed value of missing data is given by Z^1_{iy} the observed value of characteristic y^1 in the preceding survey, Census, or adminstrative data file. The estimate is then given by:

$$\hat{Y}_3 = \sum_{i=1}^{n} [\delta_{iy}\, y_i/\Pi_i + (1 - \delta_i)\, Z^1_{iy}\ /\Pi_i] \qquad (4)$$

It should be noted that both y_i and Z^1_{iy} are subject to response error (containing both a possible response bias and response variance) relative to the true value for unit i. If historical or external source data are available for every unit that fails to respond with respect to characteristic 'y', then adjustment cells as used in the weighting and duplication methods become redundant.

The bias of the estimate $\hat{Y}_3$ may be shown to be sum of two components.

$\sum_{i=1}^{N} \alpha_{iy}\, B_{Riy}$ ... response bias, due to response errors of current responses,

and $\sum_{i=1}^{N} (1 - \alpha_{iy}) B^1_{Riy}$ ... imputation bias, due to substitution of historical data and introduction of the response bias of historical records relative to the characteristic 'y' in the current survey. Here B^1_{Riy} is the imputation bias pertaining to unit i.

When historical or external source data are only available for some nonrespondents, then another imputation procedure such as weighting must be applied along with the historical data substitution. When weighting is applied in conjunction with the historical data substitution, adjustment cells are required and the estimate is given by:

$$\hat{Y}_4 = \sum_{b} \hat{Y}_{b4}$$

$$\hat{Y}_{b4} = \frac{n_b}{m_b + m_b^1} \sum_{i=1}^{n_b} [\delta_{iy}\, y_i/\Pi_i + (1-\delta_{iy})\, \delta^1_{iy}\, Z^1_{iy}/\, \Pi_i] \qquad (5)$$

In formula (5), δ^1_{iy} = 1 or 0 according as historical or external source data representing characteristic y of unit i are available. The estimated response rate $(m_b + m^1_b) / n_b$ comprises two components; m_b/n_b, the usual response rate for the current survey as defined in (1) and $m_b{}^1/n_b$, the response rate for acceptable historical data among the $(n_b - m_b)$ nonrespondents of the current survey.

The quantity $(m_b + m^1_b)$ equals $\sum_{i=1}^{n_b} [\delta_{iy} + (1-\delta_{iy}) \delta^1_{iy}]$, the number of respondents and nonrespondents with acceptable historical data.

(e) Zero Substitution Method

The "Zero Substitution Method" of imputation is one in which the missing data due to nonrespondents are ignored or by implication zero substituted for their missing values.

In some cases, the missing cells are simply labelled as a "not stated" and then it is left to the data analyst to treat the data in the manner that suits his purposes.

5. ESTIMATED RESPONSE PROBABILITIES

The bias in the estimate under each imputation procedure, apart from the response bias, results from the estimated response probabilities differing from the true response probabilities and a correlation between the values of the characteristic and the true response probabilities. By delineating cells for imputation purposes, for example strata or clusters, one attempts to employ estimated reponse probabilities as close to the true values as possible. The most common procedure is to employ response rates which are equivalent to estimates of average response probabilities in adjustment cells. The estimated response probabilities are presented in Table 3 by imputation procedure.

TABLE 3

Estimated Response Probability $\hat{\alpha}_{iy}$ by Imputation Procedure

Procedure	Estimate	Formula no. (in brackets)	Est'd Response Probability
Weighting	$\hat{Y}_{b1}$	(1)	m_b/n_b
Duplication	$\hat{Y}_{b2}$	(3)	$[n_b/m_b]^{-1}$ or $[n_b/m_b +1]^{-1}$
Historical Data Substitution			
Case (i)	$\hat{Y}_3$	(4)	1 (includes probability of available historical data)
Case (ii)	$\hat{Y}_{b4}$	(5)	$(m_b + m_b^1)/n_b$ (includes probability of available historical data among nonrespondents)

The estimated response probabilities are, in all cases, except for $\hat{Y}_3$, obtainea by the response rate in cell 'b'.

6. VARIANCES OF ESTIMATES

With the application of the concept of response probabilities as opposed to strata of respondents and nonrespondents, the development of the variances of the estimates under different imputation procedures results in very complex expressions for both sampling and nonsampling variances. The components of which are listed:

(i) sampling variance

(ii) simple response variance

(iii) correlated response variance

(iv) variance component due to the variation in the even of responding/not responding for each unit

(v) covariance component due to the covariance between the events of responding/not responding for each pair of units

The first three have been dealt with extensively in the case of full response by Fellegi (1964) in the case of srswor and Koch (1973) in the case of ppswor. The last two components have been developed by R. Platek and G. Gray (1978).

It can be shown that the method of duplication will result in a slightly higher variance than simple weight inflation in weighting classes or balancing areas (See Hansen et al 1953) and the increase will apply to all five components. The historical data substitution will almost certainly result in lower variances if the historical data is highly correlated with the current values of the characteristics since the high correlation will result in effectively a larger sample. Again, all five components would likely be reduced. The zero substitution contains a lower sampling variance mainly because of the under-estimate of the total; however, the non-sampling variances are not necessarily lower and the overall mean square error of the zero estimate is most likely larger than that of the other estimates because of its large under-estimate. The above results have been substantiated by a hypothetical example carried out by Platek and Gray [1978].

7. APPLICATION OF IMPUTATION TO HOUSEHOLD SURVEYS IN STATISTICS CANADA

The imputation procedures that are in common use in sample surveys and censuses in Canada include weight adjustment, duplication, substitution of historical or external data and Hot Decks.

Labour Force Survey

One of the major continuous surveys conducted by Statistics Canada is the Labour Force Survey from which monthly estimates of unemployment, employment, and many other characteristics are obtained. The data are published a few weeks after surveying about 56,000 households in approximately the third week of each month. It is impossible to contact every household that should have been contacted because of the stringent schedule in collecting and processing the data. Some households are away for the entire week, absent each time the interviewers call, or else they refuse to be interviewed. There are of course other reasons for nonresponse but they make only a small contribution to the total nonresponse which is maintained around five percent most of the time except for increases to seven or eight percent in the summer months because of 'Temporary absences' due to vacations.

Imputation for complete household nonresponse is carried out according to the following criterion. (i) for about one-third of the nonrespondents substitution of last month's values where it is applicable (with suitable transformations in some fields to update last month's data) or (ii) inflation by the inverse 'response' rate in balancing units ('response' in this case including substituted values for those nonrespondents who actually responded in the preceding survey with applicable data). In the case of (ii), the imputation for the remaining 2/3 of the nonrespondents is implicitly the mean of all respondents, in the balancing unit (primary sampling units, small urban and rural portions of PSU, or subunits in large cities).

Imputation for item nonresponse or edit rejects is carried out in one of three ways, depending upon the item or items with missing of faulty data and depending upon the response status and characteristics of the unit in the previous survey.

(i) the proper item response that has been omitted can be unambiguously deduced from the remainder of the questionnaire (a decision table would ensure a unique and consistent response);

(ii) the substitution of the item response of the previous survey if it is available and if it is appropriate according to a decision table;

(iii) the application of a Hot Deck procedure whereby a similar record is obtained in the same PSU, same path taken (one of six possible) in the sequence of questions, and same age-sex group. Here, collapsing of weighting classes may be required to find a similar type record; usually age-sex categories rather than psu's or paths taken are grouped together for imputation purposes when necessary.

To illustrate the imputation procedure for partially completed questionnaires, consider the following examples, pertaining to the Canadian Labour Force Survey (LFS).

Example 1 (Canadian Labour Force Survey)

Every person 15 years of age and over, within the households selected for the LFS is asked, Question 80 (Q80) "Is he going to school or not?". If he is, then the response is coded '1', if not, then the response is coded '2'. For those not going to school no related questions are asked. However, for those going to school, there are follow-up questions: Question 81 (Q81), "Is he going to school part-time or full-time?". If yes, then Question 82 (Q82) "What kind of school?"? Thus, the following situation may arise.

Q80 ≠ 1 or 2 which is in error.

The required relationship between questions must be of the following kind:

a. if Q80 = 1 i.e. going to school
   then Q81 = 1 or 2 i.e. full-time or part-time
   and Q82 = 1 or 2 or 3 or 4 i.e. the type of school

b. if Q80 = 2 i.e. not going to school
   then Q81 = blank (not applicable)
   and Q82 = blank (not applicable).

There are also entries in other questions that may be related to Q80 and 82 and these are: Q14 and Q36: Attending school as a reason for working < 30 hours, Q58 "Going to school" as an answer to what the respondent was doing immediately before looking for work and "Going to school" as a reason in Q64 for not being able to take a job last week. In each case, the code is '3'.

As an example for specific set of conditions[**], let us suppose the answer to Q80 is not available or is inconsistent with the logic of subsequent questions. The following decision table summarizes the steps followed in imputing the appropriate value for Q80.

	Imputation Rule						
	1	2	3	4	5	6	7
Q81 = 1 or 2	Y	N	N	N	N	N	N
Q82 = 1, 2, 3 or 4	Y	N	N	N	N	N	N
Q14, 36, 58 or 64 = 3	│	Y	N	N	N	N	N
Age = 15 or 16 yrs.	│	│	Y	Y	N	N	N
Survey Month =July or August[*]	│	│	Y	N	N	N	N
Q80 in Previous Month = 1	│	│	│	│	Y	N	N
= 2	│	│	│	│	│	Y	N
≠ 1 or 2	│	│	│	│	│		Y
Then Q80 in Current Month	1	1	2	1	1	2	Hot Deck

Here, Hot Deck means a search within the same PSU, Path, and age-sex group as stated earlier.

Each column represents a separate sequence of steps that must be followed in order to arrive at the imputed value of whether a person is or is not going to school (1 or 2). The Y = Yes and N = No in the main body of the decision table corresponding to the condition statements.

[*] For persons not 15 or 16 years of age, it does not matter whether the survey month is July or August for imputation rules 5, 6 and 7.

[**] It excludes other conditions, for example if Q81 is not 1 or 2 but Q82 is i, 2, 3 or 4 and vice versa.

The detailed discussion for each imputation rule is given below:

1. The imputed value is based on the internal logic of the related questions. Thus, if Q81 and Q82 have valid responses, then only the value of Q80 consistent with other information is assigned i.e. if Q81 = 1 or 2 and Q82 = 1, 2, 3 or 4, then impute '1' for Q80, i.e. going to school.

2. If the information for Q81 and Q82 are not available, then the next step is to seek relevant information elsewhere within the questionnaire. The other questions are asked depending upon the 'path' a person may have followed. Should he/she have indicated earlier that he/she was going to school, then Q80 is also coded as such, i.e. if Q81 ≠ 1, 2 and Q82 ≠ 1, 2, 3, 4 but Q14, 36, 58 of 64 = 3, then Q80 = 1, i.e. going to school.

3-4. If there is no directly related information available, then use is made of the information that is indirectly related to the question in edit conflict. In this case, it is the age and the month of the survey. If a person is 15 or 16 years old, then during the months of July and August he most likely was not going to school and thus Q80 is coded 2, i.e. not going to school. Whereas for all other months he is coded as going to school, i.e. Q81 = 1, i.e. going to school.

These imputations are examples of rare cases so that the assignment of the "most probable value" is justified, even though hot deck might be theoretically better but possibly more expensive.

5-6. If no information in the current month's data is available and the person is not 15 nor 16 years old, the next recourse is to the previous month's data. Whatever the response is from the previous month, if it is available, it is transferred to the current month's questionnaire.

The use of last month's data is justified largely on the ground that the month to month correlation of characteristics of certain estimates is quite high.

7. If no information is directly related to school attendance is available then a response value is imputed from another similar record. In this case, the similar record is selected from the same PSU, path age-sex group on the basis of availability, i.e., it is the first record on the file that meets the selection criteria. Considering the way in which the records are received for processing, selection of the first available record is assumed to be a close approximation of the random selection method. It should be mentioned that should this search fail, the conditions are relaxed to include the next age-sex group. The questionnaire for respondent households is now complete and internally consistent using the relevant method of imputation.

I have given an example for imputation procedures used in the Canadian Labour Force Survey. Procedures used in other household surveys lean heavily on the methods used in the Labour Force Survey and the Survey of Consumer Finances. The latter defines several imputation strata and the method of stratification is primarily based on a technique developed by Morgan and Sonquist.

8. CONCLUSION

The paper has provided a brief overview of the concept of nonresponse, several sources of nonresponse and of various methods of adjustment for it. The results of the various ways of adjusting for nonresponse is that a "clean data set" is obtained, that is a set of consistent values is available for each unit in the sample.

There are a number of methods for adjustments for nonresponse but there seems to be a lack of sound methodological development for most of them. The development of integrated theory for imputation becomes more and more important with the increasing number of surveys and the difficulties of obtaining full responses. However, the primary importance will always be to control the size of nonresponse in the field in preference to adjusting for it by various techniques.

RESUME

L'article donne un aperçu général des concepts de données incomplètes et de la non-réponse. Il est reconnu que la non-réponse est un indice important de la qualité des données puisqu'elle affecte les estimateurs en y introduisant un biais et une augmentation de la variance à cause d'une réduction de la taille effective de l'échantillon. La relation entre le biais et le taux de non-réponse est moins évidente puisqu'elle dépend de l'ampleur de la non-réponse et aussi de la différence des diverses caractéristiques entre les répondants et les non-répondants.

Le moyen le plus efficace de traiter les effets de la non-réponse est d'en minimiser l'ampleur. Cependant, toute tentative de contrôler l'ampleur de la non-réponse doit être fondée sur une bonne compréhension de ses origines. Les causes de la non-réponse et son ampleur sont fondamentalement liées i) au type d'enquête, ii) aux méthodes de saisie des données et iii) au plan d'échantillonnage. Toutefois, étant donné un plan d'échantillonnage, l'ampleur de la non-réponse sera influencée par des facteurs tels le type de région et le type de non-réponse.

Il y a plusieurs façons de traiter les données incomplètes. Chacune d'elles, en fin de compte, attribue une valeur aux données manquantes ou incorrectes; à moins qu'il ne soit décidé de publier des données "brutes". La procédure d'attribution de valeurs s'appelle imputation et une telle valeur imputée décrit, présumément, la caractéristique du non-répondant.

L'article donne une brève discussion philosophique sur le sujet de la validation et de l'imputation et leurs applications à la méthodologie des diverses procédures d'imputation. Parmi celles-ci, mentionnons la pondération, réplication, "Hot Deck substitution par des données antérieures et remplacement par la valeur zéro. L'application de l'imputation par rapport aux méthodes employées par l'enquête sur la population active au Canada y est aussi discutée. Une table de décision est fournie indiquant les diverses étapes à suivre pour un cas particulier d'un questionnaire de l'EPA partiellement complet.

REFERENCES

[1] Ashraf, A. and Macredie, I., "Edit and Imputation in the Labour Force Survey," Imputation and Editing of Faulty or Missing Survey Data, 114-119, selected papers presented at 1978 American Statistical Association meeting.

[2] Bailar, J.C. III and Bailar, B.A. (1978), "Comparison of Two Procedures for Imputing Missing Survey Values", Imputation and Editing of Faulty or Missing Survey Data, 65-75, selected papers presented at 1978 American Statistical Association meeting.

[3] Fellegi, I.P. (1964) "Response Variance and its Estimation", Journal of the American Statistical Association, Volume 59. 1016-1041.

[4] Ford, Bary, L. (1976) "Missing Data Procedures: A Comparative Study, American Statistical Association, Proceedings of Social Statistics Section.

[5] Gower, A.R. "Non-response in the Canadian Labour Force Survey" Survey Methodology Journal (Statistics Canada) Volume 5, Number 1 (June 1979) 29-55.

[6] Hansen, M.H., Horvitz, W.N., and Madow, W.G. (1953), "Sample Survey Methods and Theory", Volume II, Theory, 139-141, John Willy and Sons, Inc.

[7] Koch, G. (1973), "An Alternative Approach to Multivariate Response Error Model for Sample Survey Data with Applications to Estimators Involving Subclass Means", Journal of the American Statistical Association, Volume 68, 906-913.

[8] Kovar, Mary Grace and Robert P. Wright (1973), "An Experiment with Alternative Respondent Rules in NHIS", Proceedings of Social Statistics Section, American Statistical Association.

[9] Madow, William G., Unpublished Report on Nonresponse (1980).

[10] Morgan, J.A. and Sonquist, J.N. (1964), Monograph 35 Survey Research Centre, University of Michigan.

[11] Platek, R. "Some Factors Affecting Nonresponse" Survey Methodology Journal (Statistics Canada) Volume 3, Number 2 (December 1977), 191-214.

[12] Platek, R., Singh, M.P. and Tremblay, V., "Adjustment for Nonresponse in Surveys", Survey Sampling and Measurement, Academic Press, New York, 1978, 157-175.

[13] Platek, R. and Gray, G.B. "Methodology of Adjustments for Nonresponse", Invited Paper presented at the 42 Session of International Statistical Institute, Manila, Philippines, December 1979.

[14] Statistics Canada, "Imputation for Household Survey in Statistics Canada", paper submitted by Canada at the meeting of European Statisticians, Geneva, March 1978.