

11-613E

no. 88-11

c. 2

Statistics  
Canada

Statistique  
Canada



Methodology Branch

Social Survey  
Methods Division

Direction de la méthodologie

Division des méthodes  
d'enquêtes sociales

Canada

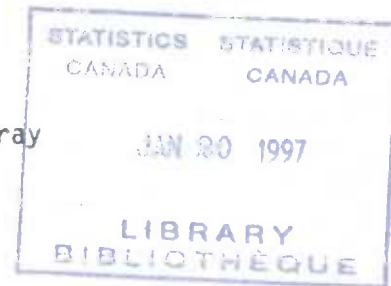
WORKING PAPER NO. SSMD-88-011 E

METHODOLOGY BRANCH

*Z-186E*  
Response Probability Approach to Missing Survey Data *6.2*

SSMD-88-011E

R. Platek and G.B. Gray



### Abstract

In surveys and censuses each selected unit may or may not respond. The probability that a particular unit responds may be regarded as dichotomous; i.e., 0 or 1 or more plausibly, it may respond with some probability between 0 and 1, with different probabilities pertaining to every unit. Unlike selection probabilities that are usually known in advance of the survey, response probabilities are not known for every individual unit. The various assumptions, including equal response probabilities are not known for every individual unit. The various assumptions, including equal response probabilities in cell, are reviewed from both theoretical and practical viewpoints along with the model and concepts pertaining to response probabilities.

### Résumé

Dans les enquêtes et recensements, chaque unité choisie peut répondre ou ne pas répondre. La probabilité qu'une unité particulière réponde peut être considérée comme dichotomique, c'est-à-dire 0 ou 1. Il est toutefois plus plausible qu'elle réponde avec une certaine probabilité comprise entre 0 et 1, cette probabilité étant différente pour chacune des unités. Contrairement aux probabilités de sélection qui sont habituellement connues avant l'enquête, les probabilités de réponse ne sont pas connues pour une unité individuelle. Les diverses hypothèses, y compris les probabilités de réponse égales dans une case, sont examinées d'un point de vue à la fois théorique et pratique, de même que le modèle et les concepts liés aux probabilités de réponse.

## Response Probability Approach to Missing Survey Data

R. Platek and G.B. Gray

### 1. Introduction

Close examination of the behaviour of response rates for a variety of surveys under different survey conditions leads to the natural assumption that in the population, each unit has some probability of responding (if selected) according to the survey conditions such as type of interview, quality of interviewer, subject matter, length of questionnaire and many others. These conditions may interact with the different units in different ways so that the probability of response for any given unit is not simply 0 or 1 nor are the response probabilities the same for every unit in the population. Thus, for most situations, the population may be defined as a collection of potential respondents some of which are respondents and others nonrespondents, depending upon the effect of the survey conditions on the different units. For example, the presence or absence of a household is related to the interviewing schedule of an interviewer as he/she may or may not find a responsible respondent at a selected unit. Thus, the concept of response probability is implicit in the Politz-Simmons (1949) technique of adjusting the survey weights by the estimated probability of finding someone at home to respond.

Response probability may be defined as the probability that a selected unit provides a response completely or partially to a survey questionnaire. In the case of the Politz-Simmons technique, the response probability is correlated with the time of day that an attempt for an interview is made. The probability may be also correlated with such characteristics such as age, sex, income, size of unit, location as well as the subject and interest of the survey to the respondent. Response probabilities will also depend upon the method of taking a survey and are usually higher in personal interview than in telephone or mail surveys.

In the above, we have dealt with the concept of unit response probabilities. There is also the problem of item nonresponse when a unit responds to some but not all questions. The item response probability is related to the complexity of the question, the format of the questionnaire as well as the interest and comprehension of the item on the part of the respondent. As in the case of unit response/nonresponse, the potential unit respondents should be regarded as a collection of potential respondents to each applicable item of a questionnaire rather than a dichotomy of item respondents and item nonrespondents. The probability is related to the ability of the interviewer to explain or clarify the questions to the respondent. As we shall see later when item response is examined in detail, the concept of item response probability is more difficult to define than unit response probabilities as missing data for an item may be due to response error just as much as due to an incomplete questionnaire.

## 2. Selection and Response Probabilities

The notion of probability plays an important part in the development of sampling and non-sampling errors. Just as selection probabilities are essential in the derivation of biases and variances of such estimates as the means and totals, the response probability approach makes it possible to derive these statistics when missing data occur. One begins by the definition of indicator variables for the selection/non-selection of a particular unit  $i$ . Similarly, one defines an indicator variable for response/nonresponse of unit  $i$ , given that it is selected. Finally, one defines an indicator variable for item "y" response/nonresponse for a selected and responding unit  $i$ . From the definitions of the indicator variables, one may define various expected value, variance and covariance operators. Below is a summary of defined indicator variables and operators that are employed in the methodology developments pertaining to the response probability approach.

$t_i$  - the event of selection ( $t_i=1$ )/non-selection ( $t_i=0$ ) for unit  $i$ ,

$\delta_i$  - the event of response ( $\delta_i=1$ )/nonresponse ( $\delta_i=0$ ) for selected unit  $i$

The selection probability of unit  $i$  is given by  $E_1(t_i)=\pi_i$ .

The response probability of selected unit  $i$  is given by  $E_2(\delta_i)=\alpha_i$ .

$E_1$  - expected value over all possible samples of units from the population of  $N$  units.

A sample of  $n < N$  units from the whole population of  $N$  units may be given by a vector as follows:

$$\underline{t} = (t_1, t_2, \dots, t_i, \dots, t_N) \quad 2.1$$

In the above,  $n = \sum t_i$ .

$E_2$  - expected value over all possible patterns of nonresponse (subsample of responding units), given the sample  $\underline{t}$ .

In a similar manner as  $\underline{t}$  defines a sample of units from the whole population, the vector  $\underline{\delta}$  consisting of  $n$   $\delta_i$ 's defines a subsample of responding units. Thus, the missingness pattern is defined by:

$$\underline{\delta} = (\delta_1, \delta_2, \dots, \delta_i, \dots, \delta_n) \quad 2.2$$

The analogous expressions for the joint selection and joint response probabilities are described briefly.

The joint inclusion probability of units  $i$  and  $j$  according to some sample design is given by:

$$E_1(t_i t_j) = \pi_i \pi_j + \text{Cov}_1(t_i, t_j) = \pi_{ij} \quad 2.3$$

In a similar manner, the joint response probability of units  $i$  and  $j$ , given that both are selected; i.e.,  $t_i = t_j = 1$ , is given by:

$$E_2(\delta_i \delta_j) = \alpha_i \alpha_j + \text{Cov}_2(\delta_i, \delta_j) = \alpha_{ij} \quad 2.4$$

Just as clustering may result in a positive covariance between the events of selection,  $\text{Cov}_1(t_i, t_j)$ , so a clustering of interview assignments may result in a positive covariance between the events of responding/not responding for a pair of units, i.e., a positive value of  $\text{Cov}_2(\delta_i, \delta_j)$ .

Thus, the mean square error, which includes the nonresponse and response bias (See Platek and Gray, 1983 for definitions) as well as the sampling and non-sampling variance components, may be derived for any statistic such as means or totals under

the variable response probability approach. In fact, the components have been so derived for Horvitz-Thompson estimates under any multi-stage sample design by Platek and Gray (1979 and 1983) and J. Lessler (1979).

The next section deals with the concept of response probability, using the variables defined above.

### 3. Response Probability and Response Rates

The concept of response probabilities and their relationship to response rates is best demonstrated by an example illustrating the correspondence between the approach based on the differences in characteristics among respondents and nonrespondents and the approach which takes into account differences in response probabilities for a particular characteristic. The effect of response errors is ignored in this demonstration. The development that follows for sample surveys is based on that of Platek, Singh and Tremblay (1977), Section 4, for a census. Let us consider a population of  $N$  units, of which  $N_p$  have characteristic "y" and  $N_q$  or  $N(1-p)$  do not. A simple random sample of  $n$  units is drawn to estimate the total  $Y=Np$ . At the time of collecting survey data, only  $m < n$  units respond and  $(n-m)$  do not. The  $m$  respondents may be split up between  $m\hat{p}_R$  with the characteristic and  $m\hat{q}_R$  not having the characteristic. The  $(n-m)$  nonrespondents may be split up between  $(n-m)\hat{p}_{NR}$  with the characteristic and  $(n-m)\hat{q}_{NR}$  not having the characteristic. However, we only know  $m\hat{p}_R$  and  $m\hat{q}_R$ .

In the above,  $\hat{p}_R$  and  $\hat{q}_R$  denote the estimated proportions having and not having respectively the characteristic among the  $m$  respondents. Similarly,  $\hat{p}_{NR}$  and  $\hat{q}_{NR}$  are the proportions pertaining to the  $(n-m)$  nonrespondents.

The above responding and nonresponding sample totals and corresponding response rates are summarized in Table 1.

TABLE 1: SAMPLE AND RESPONDING SAMPLE TOTALS

	Respondents	Nonrespondents	Total	Response Rate
With the Characteristic	$m\hat{p}_R$	$(n-m)\hat{p}_{NR}$	$n\hat{p}$	$m\hat{p}_R/n\hat{p}$
Without the Characteristic	$m\hat{q}_R$	$(n-m)\hat{q}_{NR}$	$n\hat{q}$	$m\hat{q}_R/n\hat{q}$
TOTAL	$m$	$(n-m)$	$n$	$m/n$

The sample total  $n\hat{p}$ , inflated by the inverse sampling ratio  $N/n$ , would produce an unbiased estimate of the unknown total  $Np$ . The sample total itself is unknown because

we do not know the characteristics of nonrespondents. The sample total  $n\hat{p}$  may be expressed also as  $(m\hat{p}_R)(n\hat{p})/(m\hat{p}_R)$ .

The response rate in the whole sample is given by  $m/n$ . In terms of the values pertaining to individual units, the rate is given by  $m/n = \sum t_i \delta_i / \sum t_i$ , where  $t_i = 1$  when unit  $i$  is selected and  $t_i = 0$  when unit  $i$  is not selected;  $\delta_i = 1$  when the selected unit responded and  $\delta_i = 0$  when the selected unit did not respond. The above demonstration and table may apply also to a sub-sample such as a cell, which may be a design-dependent area such as primary sampling unit, stratum, or group of strata or it may be a weighting class such as a post-stratum cell of units belonging to a particular class (eg., age-sex group). In that case,  $N$  is the population represented by the cell, which may not be known even if the total population is known. The expected value of the response rate equals  $E(m/n)$  and apart from the ratio estimation bias may be expanded as follows:

$$E(m/n)$$

$$= E_1 E_2 (m/n)$$

$$= E_1 E_2 (\sum t_i \delta_i / \sum t_i)$$

$$= E_1 (\sum t_i \alpha_i / \sum t_i)$$

$$= \sum \pi_i \alpha_i / \sum \pi_i$$

$$= \sum w_i \alpha_i \text{ where } w_i = \pi_i / \sum \pi_i$$

3.1  
= weighted average of the response probabilities of the units in the cell or area for which the response rate is calculated (defined by  $\bar{\alpha}$ ).

The response rate among the units with characteristic "y" is given by  $(m\hat{p}_R)/(n\hat{p})$  and would be an unbiased estimate of the average response probabilities among the persons with the characteristic if we could obtain  $(n\hat{p})$ , which is not known. A similar expression of the response rate pertaining to persons not having the characteristic is given by  $(m\hat{q}_R)/(n\hat{q})$ . These response rates equal the overall known response rate  $(m/n)$  only if the proportions  $\hat{p}_R$  and  $\hat{p}_{NR}$  are identical. Less rigidly, the expected value of the two rates, defined by the operator  $E$  (taken over all possible samples and all possible subsamples of respondents or missingness patterns (defined in Section 2) equals the expected value of the overall response rate  $(m/n)$  only if the expected value of the proportions,  $E(\hat{p}_R)$  and  $E(\hat{p}_{NR})$ , are identical. The extent to which  $E(\hat{p}_R)$  and  $E(\hat{p}_{NR})$  deviate from  $E(\hat{p})$  and consequently the extent to which  $E[(m\hat{p}_R)/(n\hat{p})]$  and  $E[(m\hat{q}_R)/(n\hat{q})]$  deviate from  $E(m/n)$  determine the magnitude and direction of the nonresponse bias of the estimate  $\bar{Y} = N\hat{p}_R$ , based on only  $m$  respondents. One would like this estimate to be as close as possible to  $\bar{Y} = N\hat{p}$ , based on the full sample, or at least have the same expected value.

The estimate  $\bar{Y}$ , based on the  $m$  respondents in the sample of  $n$  units assumes that respondents and nonrespondents have the same characteristic. The bias of the estimate may be easily shown to be proportional to (a), the difference between the expected values of the proportions with the characteristic among respondents and nonrespondents or (b), the difference between the expected values of the response rates pertaining to the units with the characteristic and those without the characteristic.

The bias of the estimate  $\bar{Y}$  is, by definition,

$$B(\bar{Y}) = N[E(\hat{p}) - p] = N[E(\hat{p}_R - \hat{p})],$$

where  $E$  - expected value in 2 stages; i.e.,  $E = E_1 E_2$ .

The bias of the estimate  $\bar{Y}$  may be found algebraically using either of following two expressions (3.2) and (3.3).

$$B(\bar{Y}) = N[1 - E(m/n)]E(\hat{p}_R - \hat{p}_{NR}) - N[1 - E(m/n)](p_R - p_{NR}). \quad 3.2$$

In the expression above,  $p_R$  and  $p_{NR}$  define the expected values of the proportions of units having the characteristic among respectively, respondents and nonrespondents. Alternatively, the bias may be found to be:

$$B(\bar{Y}) = NE(n/m)\hat{p}(1-\hat{p})E[(m\hat{p}_R/n\hat{p}) - m\hat{q}_R/n\hat{q}] \quad 3.3$$

$$= NE(n/m)p(1-p)(\bar{\alpha}_p - \bar{\alpha}_q).$$

In the expression above,  $\bar{\alpha}_p$  and  $\bar{\alpha}_q$  are the expected response rates (average response probabilities) among units respectively having and not having the characteristic. For a positive nonresponse bias, for example, the characteristic is more prevalent among respondents than among nonrespondents according to (3.2). Alternatively according to (3.3), the response rate tends to be higher among units with the characteristic than among units without it or units with the characteristic are more likely to respond. The two identical bias expressions hold under any sample design and under any assumption about the response probabilities, variable or dichotomy.

#### 4. Response Probabilities in Imputation

For a successful adjustment for nonresponse by weight adjustment, it is necessary that where the adjustment is made, the characteristics of respondents and nonrespondents are as similar as possible. Often the adjustments are made and discussions of nonresponse bias are undertaken on the assumption of either dichotomous or equal response probabilities. These two assumptions are, however, special cases of a general approach referred to as the "Response Probability Approach" in which variable response probabilities between 0 and 1 are assumed.

The diagram below, Figure 2, shows three different approaches to response probabilities with a discussion of the implications of each assumption on the bias and other survey considerations. The approaches are based on dichotomous, equal or variable response probabilities.

(a) Dichotomy: assumes that the population is divided into two strata of respondents and nonrespondents, where the response probability of any unit is either 1 or 0. Under the dichotomy approach, a nonresponse bias would occur if the characteristic of interest were correlated with the dichotomous response probabilities. The bias reduces to the difference between the means of the units in the two strata. The difference may be detected by sub-sampling the nonrespondents and attempt a follow-up of them. However, in so doing, there is immediately an implied non-zero probability of responding among the so-called nonrespondents, an apparent contradiction to the dichotomy assumption.

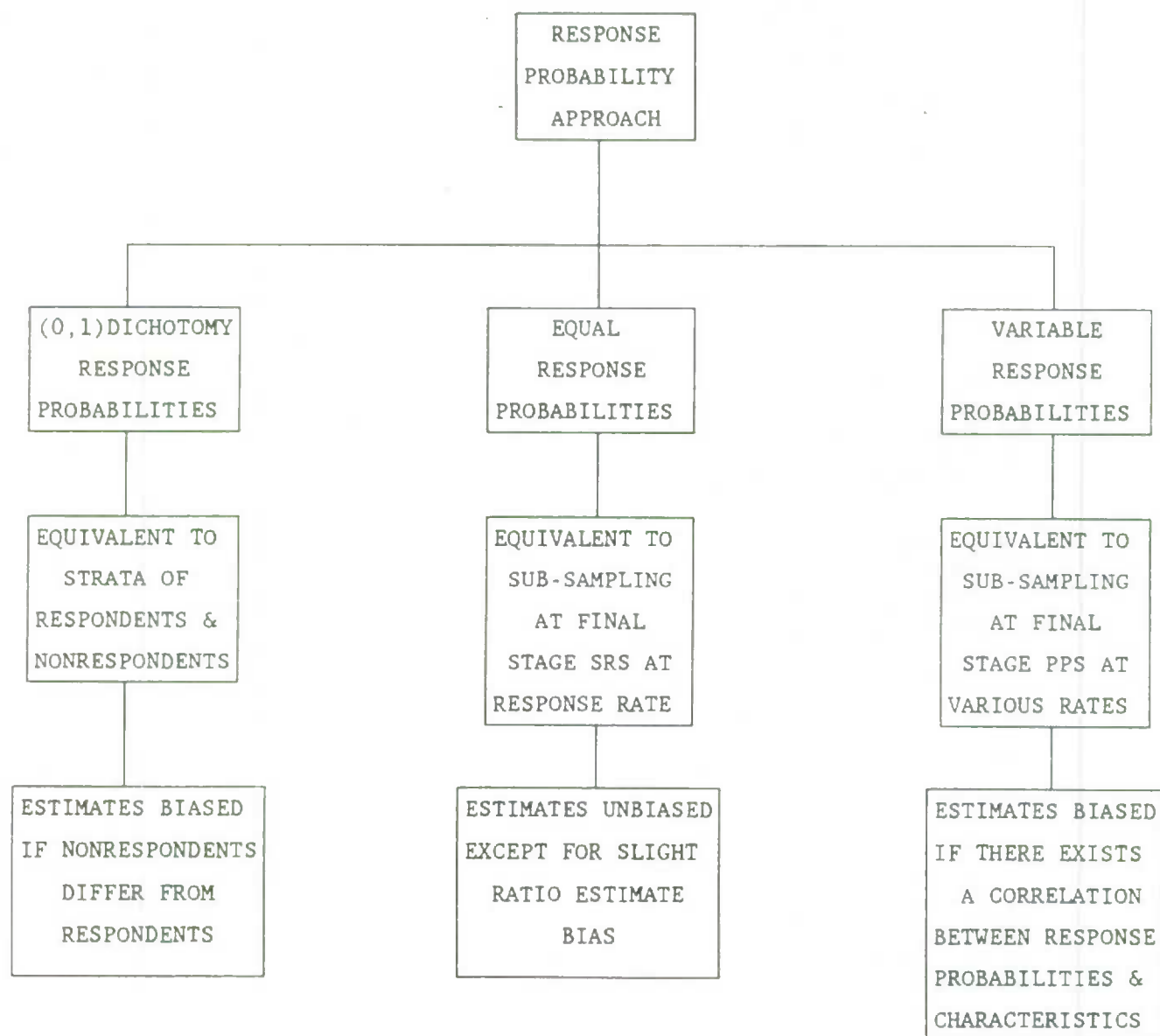
(b) Equal Probability: assumes that each unit has the same probability of responding. In that approach, the respondents form a random sub-sample of the selected units. Upon weighting the sub-sample by the inverse response rate, the resulting estimates are unbiased, except for a small ratio estimate bias when the responding sample size is a variable. This approach is analogous to an additional stage of sampling with equal probability as in Poisson sampling and as such has some cost and variance implications but may be somewhat unrealistic.

(c) Variable Probabilities: assume that the target population consists entirely of potential respondents, such that every unit has some chance other than 0 or 1, but not necessarily the same for every unit, of responding. Under this assumption, when the sub-sample is weighted by the inverse response rate, the nonresponse bias is

proportional to the correlation between the characteristic of interest and the response probability in the cell, where the weight adjustment is made. The nonresponse bias may also be expressed as the difference between the expected values of the mean of the respondents and that of the nonrespondents as in the case of the dichotomy assumption. However, the difference must be interpreted differently. In the variable response probability approach, the expected value of the respondents' mean is obtained by weighting the values of the characteristic of interest by the response probability, as implied in the above-mentioned correlation. Similarly, the expected value of the nonrespondents' mean must be obtained by weighting the values of the characteristic of interest by the nonresponse probabilities, the complements of the response probabilities.

The variable response probability approach is analogous to an additional stage of sampling, where the probabilities vary among the units. The probabilities are unknown and must be estimated from other sources such as, for example, longitudinal and re-interview studies.

FIGURE 2: RESPONSE PROBABILITY APPROACH



The estimates of statistics such as means or totals contain nonresponse bias due to missing data as a result of nonresponse. A common method to compensate for nonresponse is weight adjustment by the inverse response rate in adjustment cells. When response probabilities and characteristics vary within an adjustment cell, there will be a nonresponse bias if the characteristic is correlated with the response probability in that cell. For any given population, the dichotomy assumption results in a nonresponse bias of greater magnitude than the variable response probability since the latter draws the expected values of the two means closer together than in the case of (0,1) probabilities. As the non-zero probabilities become closer together and approach equality in a cell, then there is no bias due to nonresponse. This phenomenon is shown in Appendix III. The variable response probability approach to the derivation of the nonresponse bias is more general than the dichotomy and equal probability approaches. The nonresponse bias and in fact all variance components, derived for statistics under the variable response approach, may be applied to the other approaches as particular cases, but one cannot proceed the other way around.

In many methodological studies of errors due to missing data, the response errors are usually omitted even though the distinction between response errors and item nonresponse is frequently blurred. This occurs, for example, when questions are left blank when they should not have been and, conversely, on the same questionnaire, items are filled in when they should not have been. The first of these appear to be item nonresponse while the second appear to be response errors. In such cases, it is often left to the discretion of the statistician whether to treat such errors as response errors or item nonresponse in the study of and in the development of biases and variances arising from such errors.

The magnitude of response and nonresponse errors is greatly influenced by the method of data collection. In the case of personal interviews, the presence of an interviewer face-to-face may help the respondent in his/her answers to a greater extent possible than either telephone or even more so mail surveys. Also, nonresponse errors arising from missing data and the extent of imputations for them may be lower in personal than in telephone surveys because of the greater ability of the interviewer to prevent refusals or convert refusals to cooperative respondents. In mail surveys, it may be even more difficult to convert refusals and also to get in touch with respondents who may be absent during the survey. With mail surveys, the respondent is more likely to discard the questionnaire than in the case of telephone or personal interview surveys, thus resulting in unit nonresponse.

The lower response and nonresponse errors in personal and telephone than in mail surveys may be partly offset by a positive correlated effect in the errors because of the clustering of personal and telephone assignments (See Fellegi, 1964 for the study of correlated response errors; Lessler, 1979 and Platek and Gray, 1979 & 1983 for the study of correlated nonresponse errors). As far as nonresponse errors are concerned, the correlated effects arise from positive values of  $Cov_2(\delta_i, \delta_j)$  in one of the variance components of mean square error arising from missing data when the response probabilities are other than dichotomy; i.e., 0 or 1.

The variance and covariance components of mean square error that arise from missing data depend upon the magnitudes of  $V_2(\delta_i)$  and  $Cov_2(\delta_i, \delta_j)$  respectively which depend upon the approach assumed for response probabilities in an area or cell. Under the dichotomy approach,  $V_2(\delta_i)$  and  $Cov_2(\delta_i, \delta_j)$  are zero so that under that approach there is no variance or covariance component due to missing data, only a potential nonresponse bias. As the response probabilities approach equality in a cell, the

variance and covariance components due to missing data approach the maximum values for a given overall response rate. This occurs because the variance component is proportional to the average values of  $V_2(\delta_i)$  in an area or cell and it is shown in the appendix that average values are the maximum value under the equal probability approach. The higher variance and covariance components are likely more than offset by a lack of nonresponse bias under the equal probability approach and a smaller sampling variance component because of less variation in responding sample size than there would be with a variable and dichotomy response probability assumptions.

The variable response probability approach is more realistic than the equal and dichotomy approaches. Only a small portion of the population would have no chance of responding (hard core refusals or permanently absent units), unless the design of the data gathering procedure is very poor. Many hard core refusals, for example, may be converted into willing respondents by means of response incentives (Gower, 1977), tact, and resourcefulness on the part of interviewers and a good public relation campaign related to the survey. The assumption of equal response probability for all units, the simplest of the three approaches, is unrealistic since in practice, it is difficult to delineate adjustment cells containing units with the same response probabilities. Nevertheless, the application of a fixed weight adjustment in a cell by an inverse response rate implies the assumption of equal probabilities of response for all of the units in the cell. The consequence of the failure of the assumption is a nonresponse bias, proportional to the correlation between the actual response probabilities and the characteristics in the adjustment cell.

Except for hard core refusals, which would normally be few in number in a well-organized survey, the nonresponse bias may be estimated from the sample by estimating the correlation between the response probabilities and the characteristics. The response probabilities, however, are almost never known in advance, so they must also be estimated. In the case of continuous surveys, the probabilities may be estimated from longitudinal surveys (see for example, Lawes and Paul, 1982) of the units sampled on successive occasions (eg., Canadian Labour Force Survey). The units may respond on some occasions and fail to respond on others. Excluding those units that never respond on any occasion that they are in the sample, one may obtain the characteristics of the units and derive an estimate of the correlation from which an estimate of the nonresponse bias may be obtained. This procedure is not available for one-time ad hoc surveys. The Politz-Simmons technique of finding the probability of finding the respondents at home (JASA, 1949) may be applied for any household survey and in such cases, apart from the hard-core refusals as above, the correlation between the characteristic and the estimated probability of response may be obtained. The Politz-Simmons has sometimes been used for weight adjustment by the estimated probability to reduce or eliminate the nonresponse bias.

The study of nonrespondents' characteristics may be based on the design or model approach. The variable and the less realistic equal response probability concepts lend themselves to the design approach since, under either assumption, the event of responding/not responding may be regarded as an additional random event in the sample design, analogous to an additional stage of selection. The model approach may be employed under either the dichotomy or variable response probability approaches. Examples of applied models in the study of nonresponse biases include super-population models by size of the units, assuming the sizes known for every unit and some assumed distribution of their characteristics by their propensity to respond. Under a constant or equal probability approach, there is no model as there is no bias due to nonresponse and the problem is strictly one of design considerations of potential nonrespondents.

The practical implications of the response probability approach are inherent in the use of hot deck imputation procedures. In most hot deck imputations, the units are grouped into similar type records of a cell, applying some criterion of "similarity". When missing data occur because of item and sometimes unit nonresponse, a so-called donor record is used to impute the missing data of one or more items or, in the case of unit nonresponse, nearly all items. However, hot decking is more practical for item than for unit nonresponse since item nonresponse implies some knowledge about the unit, i.e., information that may be lacking in the case of unit nonresponse. The selection of a donor record to impute the missing data is either systematic or random, often on the basis of a distance function pertaining to pairs units. Clustering, for example, one of the criteria of "similarity" may be measured by the geographic distance between pairs of units and may be applied to obtain donor records. Thus, the use of hot deck procedures is frequently design dependent even though item and unit response probabilities may not depend upon the sample. However, the item response rate in the cell containing donor records and those with missing items, is an estimate of the average item response probability in the cell.

### 5. Conclusion

The estimation of response probabilities is fundamental in the approach discussed in this paper. The problems of estimating response probabilities, whether on the basis of the design or on the model approach, are real since they can be estimated only as averages in cells from the response rates in the cells. Otherwise, external information from longitudinal data or other surveys and censuses, if available, may be used to estimate the response probabilities.

When external sources of data (eg., administrative data, census records or earlier survey data) are used to impute for missing survey data, the response rate is effectively increased by the availability rate for external source data. The concepts are similar to those of current response probability approaches. However, external source data may be less complete or out-of-date relative to current survey data and may themselves be subject to greater response errors. The mean square error components of means and totals, under the external source data procedures of imputation, are similar to those based on only the current survey data, using the above concepts. The smaller components pertaining to nonresponse and non-availability, however, may be more than offset by a greater response bias and response variance in the estimates.

As nonresponse is always a survey problem, and may become more serious with changing technology of gathering survey data, it is essential to maintain continuous studies on nonresponse, especially those directly to nonresponse bias, in order to monitor survey and interviewer performance and to discuss and analyze quality of survey and published information. The following appendices discuss a few additional issues pertaining to the response probability approach that have been mentioned in this paper.

## Appendices to the Paper

### Appendix I: Weighted Average of Response Probabilities in a Cell

The response rate in a cell or area is defined in TABLE 1 by  $m/n$ . In terms of the values of individual units, the rate is given by  $m/n = \sum \delta_i / \sum t_i$ . The expected value of the rate was shown in Section 2 to be a weighted average of response probabilities of the units in a cell.

If we define  $E^*$  as the operator of taking a weighted average of individual unit parameters in a cell; i.e., using  $\pi_i$ 's as the weights and if we define  $\pi_i / \sum \pi_i$  by  $w_i$ , the relative size of unit  $i$  in the cell, then

$$E_1 E_2 (m/n) = E^*(\alpha_i) = \sum (w_i \alpha_i) = \bar{\alpha}, \text{ say.}$$

### Appendix II: Variance of Response Probabilities in a Cell

Suppose that in a cell of  $N$  units, the response probabilities (unknown) are  $\alpha_1, \alpha_2, \dots, \alpha_N$ . In Appendix I above, it was shown that  $m/n$ , the unweighted response rate is an unbiased estimate (apart from ratio estimate bias) of the average response probabilities (weighted by the selection probabilities) in the cell. A simple expression for the population variance of the response probabilities in a cell are now derived.

Let  $V^*(A_i) = E^*(A_i^2) - [E^*(A_i)]^2$ ; i.e.,  $V^*$  denotes the population variance of unit parameters in a cell, using the relative sizes  $p_i$  as the weights as in Appendix I above.  $A_i$  = any unit  $i$  parameter such as characteristic or response probability.

$$\text{Then } V^*(\alpha_i) = E^*(\alpha_i^2) - \bar{\alpha}^2.$$

$$\text{Now } E^*(\alpha_i^2) = \sum w_i \alpha_i^2$$

Since  $\alpha_i \geq \alpha_i^2$ , with equality occurring only when  $\alpha_i = 1$  or  $0$ ,

$$\text{then, } V^*(\alpha_i) \leq \sum w_i \alpha_i - \bar{\alpha}^2$$

$$\text{or } V^*(\alpha_i) \leq \bar{\alpha}(1 - \bar{\alpha})$$

Hence, one may express the variance of the response probabilities in a cell as some constant times the "binomial variance" of the expected response rate or:

$V^*(\alpha_i) = k \bar{\alpha}(1 - \bar{\alpha})$ , where  $0 \leq k \leq 1$ , with  $k=1$  pertaining to the dichotomy approach to response probabilities;  $k=0$  pertaining to the equal response probability approach (constant probabilities, with no variance) and  $k$  between 0 and 1 pertaining to the variable response probability approach.

Now the variance and covariance components due to missing data are approximately proportional to the average values of  $V_2(\delta_i)$  in the area or cell; i.e., proportional to  $E[V_2(\delta_i)]$ .

$$\text{Since } V_2(\delta_i) = \alpha_i(1 - \alpha_i),$$

$$\text{then } E^*[V_2(\delta_i)] = E^*(\alpha_i) - E^*(\alpha_i^2) = \bar{\alpha} - \bar{\alpha}^2 - V^*(\alpha_i).$$

It can be seen that  $E^*[V_2(\delta_i)]$  is a minimum when  $V^*(\alpha_i)$  is a maximum; i.e., under a dichotomy approach and is a maximum when  $V^*(\alpha_i)=0$ ; i.e., under an equal response probability approach. Consequently, the variance components due to missing data approach their maximum value as the response probabilities approach constant values.

#### Appendix III: Correlation between Response Probability and Characteristic (Non-response Bias)

Let  $Y_i$  be the true value of characteristic "y" for unit i and  $y_i$ , the observed value. Then  $y_i - Y_i$  is the response error of the observed characteristic of unit i and  $E_3(y_i) - Y_i$  is the response bias, where  $E_3$  denotes the operator of taking expected values over all possible responses  $y_i$ , given that unit i is selected and responds.

It may be shown that, upon applying a weight adjustment to the sample-weighted survey data by the inverse response rate  $n/m$  in an adjustment cell (See Platek and Gray, 1983) to derive an estimate  $\bar{Y}$ , the imputation or nonresponse bias is given by:

$$B(\bar{Y}) = \bar{\alpha}^{-1} \text{Cov}^*(\alpha_i, Y_i) \\ - \bar{\alpha}^{-1} r^*(\alpha_i, Y_i) \sqrt{V^*(\alpha_i) \cdot V^*(Y_i)},$$

where  $\text{Cov}^*$  denotes the operator of taking covariance between two unit parameters in a cell, using  $p_i$  as the weights, and  $r^*(\alpha_i, Y_i)$  denotes the population correlation coefficient between  $\alpha_i$  and  $Y_i$  in a cell.

As the non-zero response probabilities become closer together, the population variance of the response probabilities in a cell; i.e.,  $V^*(\alpha_i)$ , become smaller and reduce to 0 in the case of equal or constant response probabilities in the cell. In that case, it is obvious from above that  $B(\bar{Y}) = 0$ .



REFERENCES

- Fellegi, I.P. (1964), "Response Variance and its Estimation", Journal of the American Statistical Association 59: pp 1016-1041.
- Gower, A.R. (1977), "The Response Incentives Experiment in the Canadian Labour Force Survey", Survey Methodology Journal 3 (no. 1): pp 84-103.
- Lessler, J.T. (1979), "An Expanded Survey Error Model", paper presented at the Symposium on Incomplete Data, August 10-11 in Washington, D.C., also in Incomplete Data in Sample Surveys, vol. 3, chapter 12, pp 259-296.
- Paul, E.C. and Lawes, M. (1982), "Characteristics of Respondent and Non-Respondent Households in the Canadian Labour Force Survey", Survey Methodology Journal 8(nos. 1 & 2): pp 48-85.
- Platek, R. and Gray, G.B. (1978), "Non-response and Imputation", Survey Methodology Journal 4: (2), pp 144-177.
- Platek, R. and Gray, G.B. (1979), "Methodology and Application of Adjustments for Nonresponse", presented at the 42nd. session of the International Statistical Institute (Manila).
- Platek, R. and Gray, G.B. (1983), "Imputation Methodology: Total Survey Error", Chapters 16-19, "Incomplete Data in Sample Surveys", Vol. 2, Academic Press.
- Platek, R., Singh, M.P. and Tremblay, V. (1977), "Adjustment for Non-response in Surveys", Survey Methodology Journal 3: (1), pp 1-24.
- Politz, A.N. and Simmons, W.R., (1949), "An Attempt to Get the Not at Homes into the Sample without Callbacks", Journal of the American Statistical Association 44: pp 9-31.