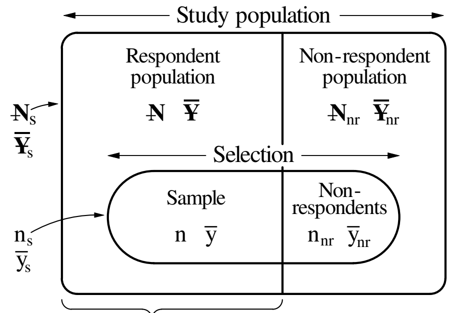
Figure 3.4c. SAMPLE SURVEY DESIGN/EXECUTION: Managing Non-Response

Many data-based investigations encounter the difficulty that not all the data called for in the Plan are acquired – this is the general problem of *missing data*. In dealing with *human* populations, this matter is usually referred to as *non-response* and it can impose *severe* limitations on Answers.

A framework to manage non-response, as indicated in the diagram at the right, *we* consider the *study* population to be made up of the *respondent* population and the *non-respondent* population. The set of units selected from the study population is the *selection*, and comprises he *sample* (from the respondent population) and the *non-respondents* (from the *non*-respondent population). The diagram has *two* categories of symbols:

- the $\cancel{\mathbf{N}}$s and ns refer to *numbers of elements or units*;
- the $\bar{\cancel{\mathbf{Y}}}$s and $\bar{y}$s are *averages* of a response variate $\cancel{\mathbf{Y}}$ of the elements or units.



Statistical theory, particularly of survey sampling, is developed mainly in the context of the *respondent* population, often with*out* recognizing it explicitly.

The relationships among the numbers of elements or units are:

$$\begin{array}{ccc} \text{Study population} & = & \text{Respondent population} + \text{Non-respondent population} \\ \cancel{N}_s & = & \cancel{N} \quad + \quad \cancel{N}_{nr} \end{array}$$

$$\begin{array}{ccc} \text{Selection} & = & \text{Sample} + \text{Non-respondents} \\ n_s & = & n \quad + \quad n_{nr} \end{array}$$

Also, suppose that a selection of n_s units is obtained from the study population by equiprobable selecting (EPS) to estimate $\bar{\cancel{Y}}_s$, the *study* population average; we have:

$$\bar{\cancel{Y}}_s = \frac{\text{study population total}}{\text{study population size}} = \frac{\text{respondent population total + non–respondent population total}}{\text{respondent population size + non–respondent population size}} = \frac{\cancel{N}\bullet\bar{\cancel{Y}} + \cancel{N}_{nr}\bullet\bar{\cancel{Y}}_{nr}}{\cancel{N}+\cancel{N}_{nr}} \qquad \text{-----(3.4c.1)}$$

where, in the denominator: $\cancel{N} + \cancel{N}_{nr} = \cancel{N}_s$.

The direct information available to us about the response variate $\cancel{Y}$ comes from the *sample* of n *respondent* population units, which yield $\bar{Y}$ as an estimator (and $\bar{y}$ as an estimate) of $\bar{\cancel{Y}}$; we know from Figure 2.3 that $\bar{Y}$ is *un*biased for $\bar{\cancel{Y}}$ – *i.e.*, $E(\bar{Y}) = \bar{\cancel{Y}}$. Hence, using quation (3.4c.1), the *non-responding bias*, when using $\bar{Y}$ as an estimator of $\bar{\cancel{Y}}_s$, is given by:

$$E(\bar{Y}) - \bar{\cancel{Y}}_s \equiv \bar{\cancel{Y}} - \bar{\cancel{Y}}_s = \bar{\cancel{Y}} - \frac{\cancel{N}\bullet\bar{\cancel{Y}} + \cancel{N}_{nr}\bullet\bar{\cancel{Y}}_{nr}}{\cancel{N}+\cancel{N}_{nr}} = \frac{\cancel{N}_{nr}}{\cancel{N}+\cancel{N}_{nr}}(\bar{\cancel{Y}} - \bar{\cancel{Y}}_{nr}). \qquad \text{-----(3.4c.2)}$$

The term: $\dfrac{\cancel{N}_{nr}}{\cancel{N}+\cancel{N}_{nr}} = \dfrac{\cancel{N}_{nr}}{\cancel{N}_s}$ is the *non-response rate*; its *value* is (approximately): $\dfrac{n_{nr}}{n+n_{nr}} = \dfrac{n_{nr}}{n_s}$; -----(3.4c.3)

the corresponding *response rate* is: $\dfrac{\cancel{N}}{\cancel{N}+\cancel{N}_{nr}} = \dfrac{\cancel{N}}{\cancel{N}_s} = 1 - \dfrac{\cancel{N}_{nr}}{\cancel{N}_s}$ and its *value* is (approximately): $\dfrac{n}{n+n_{nr}} = \dfrac{n}{n_s}$. -----(3.4c.4)

1. Dealing with Non-Response – General Considerations

Equation (3.4.2) shows that the (*un*attainable) ideal of *zero* non-responding bias, when using $\bar{Y}$ to estimate $\bar{\cancel{Y}}_s$, occurs when:

- the non-response rate is *zero* – *i.e.*, there are *no* non-respondents; **OR**
- $\bar{\cancel{Y}} = \bar{\cancel{Y}}_{nr}$ – *i.e.*, the respondent and *non*-respondent populations have the *same* average for the response variate $\cancel{Y}$.

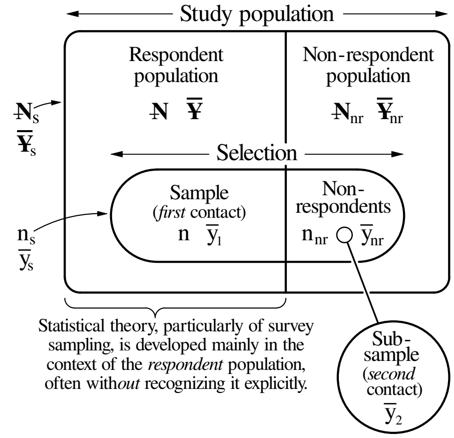
In practice, we are more concerned with the factors that *reduce* (as distinct from eliminate) the magnitude of the non-responding bias when using $\bar{Y}$ to estimate $\bar{\cancel{Y}}_s$; on the basis of our *model*, these (*two*) factors are:

* when $\bar{\cancel{Y}}$ is *close to* $\bar{\cancel{Y}}_{nr}$ – *i.e.*, the respondent and *non*-respondent populations have *similar* averages for the response variate $\cancel{Y}$;
  - this factor is *not* under the control of the investigator(s) because it depends on the distribution of values of the response variate $\cancel{Y}$ over the elements (or units) which make up the respondent and the non-respondent populations;
* when the non-response rate is *low* – *i.e.*, when the *non*-respondent population is small in size *relative to* the respondent population, or when $\cancel{N}_{nr} \ll \cancel{N}$;
  - this factor *is* under the control of the investigator(s) in the sense that it is part of the Plan for a sample survey (as specified in the *sampling protocol* in the Design stage of the FDEAC cycle) – strategies for *reducing* non-response include:
    - good questionnaire design – *e.g.*, clear, succinct, answerable questions;
    - careful interviewer selecting and training;
    - introductory letter or telephone call prior to the interview (a primary approach letter – Figure 3.2, right-hand column, first paragraph);
    - incentives for response;
    - refusal conversion techniques – *e.g.*, randomized response (see assignment question **A1 – 2**);
    - follow-up reminders and call-backs;
    - use of proxy respondents – but see Figure 3.5f;
    - use of stratification.

(*continued overleaf*)

2. Dealing with Non-Response – Call-backs

One technique to reduce non-response is to *call back* (denoted below by an abbreviation *cb*) one or more times on a sub-sample of the non-respondents from the first contact, using the most experienced interviewers and/or added incentives for response. The diagram at the right (adapted from the diagram overleaf on page 3.21) illustrates the case of *one* call-back (or a *second* contact); the notation now involves an average $\overline{y}_1$ for the sample (or *first* contact) and an average $\overline{y}_2$ for the sub-sample of the first contact non-respondents (*i.e.*, *second* contacts or *first* call-backs). In the following symbolic development, it is assumed that *complete* response is obtained from the second contact and that samples are obtained by EPS.



Rewriting equation (3.4c.1) in appropriate notation, we have:

$$\overline{\mathbf{Y}}_s = \frac{\mathbf{N}}{\mathbf{N}_s}\overline{\mathbf{Y}} + \frac{\mathbf{N}_{nr}}{\mathbf{N}_s}\overline{\mathbf{Y}}_{nr}. \qquad \text{-----(3.4c.5)}$$

Remembering the Plan's reponsibility to make it reasonable to treat the (real-world) ys as the (model) *y*s, we estimate $\overline{\mathbf{Y}}$ by $\overline{y}_1$ (= $\overline{y}_1$) for the sample of respondents to the first contact, and $\overline{\mathbf{Y}}_{nr}$ by $\overline{y}_2$ (= $\overline{y}_2$) for complete response from the sub-sample of the non-respondents (the corresponding random variables are $\overline{Y}_1$ and $\overline{Y}_2$); hence, our estimator of $\overline{\mathbf{Y}}_s$ is given by:

$$\overline{Y}_{cb} = \frac{\mathbf{N}}{\mathbf{N}_s}\overline{Y}_1 + \frac{\mathbf{N}_{nr}}{\mathbf{N}_s}\overline{Y}_2 \qquad \text{(assuming the response rate – and the } non\text{-response rate – are } known\text{).} \qquad \text{-----(3.4c.6)}$$

Then: $$E(\overline{Y}_{cb}) = \frac{\mathbf{N}}{\mathbf{N}_s}E(\overline{Y}_1) + \frac{\mathbf{N}_{nr}}{\mathbf{N}_s}E(\overline{Y}_2) = \frac{\mathbf{N}}{\mathbf{N}_s}\overline{\mathbf{Y}} + \frac{\mathbf{N}_{nr}}{\mathbf{N}_s}\overline{\mathbf{Y}}_{nr} = \overline{\mathbf{Y}}_s \qquad \text{(under equiprobable selecting);} \qquad \text{-----(3.4c.7)}$$

i.e., under EPS and if *complete* response is obtained at the *final* (here, the first) call-back, $\overline{Y}_{cb}$ is *un*biased with respect to non-response for $\overline{\mathbf{Y}}_s$; the corresponding *estimate* is $\overline{y}_{cb}$.

The foregoing results can be extended to *multiple* call-backs if we think of the non-respondent population as a *set* of *sub*-populations, often of progressively *de*creasing size and requiring an *in*creasing level of coercion to respond; increasing coercion would be represented in practice by factors such as the experience and persistence of the interviewer, the incentive(s) offered for response, and the fact of multiple contacts by the interviewer. Each successive contact after the initial one is usually with a *sub*-sample of the non-respondents from the previous contact.

To accommodate this model of the non-respondent population into the symbolic development given above, we would extend the notation to define $\overline{y}_r$ as the average of the responses from the r*th* contact [or (r – 1)*st* call-back]. Ideally, the call-back process is repeated until *complete* response is obtained from the relevant sub-sample of the previous non-respondents; in practice, the process is continued until either of or some combination of:

- the sub-population of remaining non-respondents is considered too small to introduce appreciable non-responding bias into the estimate of $\overline{\mathbf{Y}}_s$ – this can be thought of as achieving the condition $\mathbf{N}_{nr} \ll \mathbf{N}$ discussed overleaf on page 3.21;
- the use of further resources for call-backs is not considered cost effective in terms of the sample survey answers.

A final matter about our model for the use of call-backs to reduce the effect of non-responding bias when using $\overline{Y}_{cb}$ as an estimator of $\overline{\mathbf{Y}}_s$ is that if the response rate, $\mathbf{N}/\mathbf{N}_s$ [and the *non*-response rate, $\mathbf{N}_{nr}/\mathbf{N}_s$] are *un*known and are estimated by n/n_s (and n_{nr}/n_s) in the expression (3.4c.6) for $\overline{Y}_{cb}$, taking the expectation [as in equation (3.4c.7)] involves $E(N/n_s)$ and $E(N_{nr}/n_s)$ [N and N_{nr} are the respective random variables representing n and n_{nr}], which can be shown to be $\mathbf{N}/\mathbf{N}_s$ and $\mathbf{N}_{nr}/\mathbf{N}_s$, respectively, and similarly for multiple call-backs. Then, the estimator of $\overline{\mathbf{Y}}_s$ is:

$$\overline{Y}_{cb} = \frac{N}{n_s}\overline{Y}_1 + \frac{N_{nr}}{n_s}\overline{Y}_2 \qquad \text{(if the response rate – and the } non\text{-response rate – are } estimated\text{);} \qquad \text{-----(3.4c.8)}$$

but: $$E\left(\frac{N}{n_s}\overline{Y}_1\right) = E_N\left[E_{\overline{Y}_1|N}\left(\frac{N}{n_s}\overline{Y}_1\right)\right] = E_N\left[\frac{N}{n_s}E_{\overline{Y}_1|N}(\overline{Y}_1)\right] = E_N\left[\frac{N}{n_s}\overline{\mathbf{Y}}\right] = E_N\left[\frac{N}{n_s}\right]\overline{\mathbf{Y}} = \frac{\mathbf{N}}{\mathbf{N}_s}\overline{\mathbf{Y}}; \qquad \text{-----(3.4c.9)}$$

similarly: $$E\left(\frac{N_{nr}}{n_s}\overline{Y}_2\right) = \frac{\mathbf{N}_{nr}}{\mathbf{N}_s}\overline{\mathbf{Y}}_{nr}, \qquad \text{-----(3.4c.10)}$$

so that, as in equation (3.4c.7) above where the response rate – and the *non*-response rate – are *known*, under EPS and if *complete* response is obtained at the *final* (here, the first) call-back, $\overline{Y}_{cb}$ is likewise *un*biased with respect to non-response for $\overline{\mathbf{Y}}_s$ when the two rates are *estimated* from the selection.

NOTES: 1. The current context of the *study* population and (*our* terminology of) the *selection*, together with their attributes $\mathbf{N}_s$, $\overline{\mathbf{Y}}_s$, n_s and $\overline{y}_s$, makes equations (3.4c.8) to (3.4c.10) *diffferent* from what we routinely encounter in the usual context of the *respondent* population and (*our* meaning of) the *sample* – in *that* contexr, there is seldom a reason to involve the *non*-respondent population and the non-respondents, again *un*like the current discussion.

Here, n_s *is* under the control of the investigator(s), whereas n and n_{nr} are *not* and, under EPS, we model them as values of the corresponding random variables N and N_{nr}.

(*continued*)

Figure 3.4c. SAMPLE SURVEY DESIGN/EXECUTION: Models for Non-Response (continued 1)

NOTES: (cont.) 1. ● Non-responding bias, the focus of equations (3.4c.8) to, (3.4c.10) and the random variables N and N_{nr}, is peripheral to the key ideas the two diagrams at the upper right on pages 3.22 and 3.21 are intended to convey; this is why, to avoid distraction, roman, *not* italic, letters are used for n and n_{nr} in the two diagrams.

2 The use of conditional expectation in equations (3.4c.9) and (3.4c.10) on the lower half of the facing page 3.22, to show that the expected value of the product of two random variables is equal to the product of their expected values, reminds us of three matters:

- Equation (3.4c.9) is reminiscent of the conditional probability result for events A and B: $\Pr(A \cap B) = \Pr(A) \times \Pr(B|A)$.
- We usually associate multiplicative behaviour of expected values with *probabilistic independence*; however, such independence is a *strong* requirement in that its occurrence imposes substantial constraints on the relevant probability functions.
- Multiplicative behaviour of expected values can occur under *weaker* conditions – here, that $E_{\overline{Y}_1|N}(\overline{Y}_1)$ is *not* a function of N – than probabilistic independence, as illustrated in equation (3.4c.9); we are familiar with this idea from the fact that neither zero covariance nor zero correlation necessarily implies probabilistic independence.

3. Dealing with Non-Response – Stratifying

A second approach to try to reduce non-response is to *stratify* the study population into groups (*i.e.*, strata) within *each* of which respondents and non-respondents are *more* similar with respect to the attribute(s) of interest than are respondents and non-respondents in the study population as a whole.

A summary of some possible reasons for non-response is:

1. Non-contact:
   - inaccurate information about the physical location of a element (or unit);
   - accessibility of elements (or units) – *e.g.*, homemakers *vs.* university students;
   - timing of contact attempt(s) – *e.g.*, weekdays, weeknights, weekends;
   - number of contact attempts;
   - competence of interviewer – *e.g.*, locating people who have moved residence.
2. Refusal to participate in the sample survey:
   - participation is aided by interviewer's level of satisfaction in working with people, intellectual curiosity, professional identification with the sample survey's objectives, etc.
3. Item non-response (some questionnaire items not answered):
   - question is sensitive or personal;
   - respondent does not know the answer to a question.

Blank page