

## Describing incompleteness

This chapter is concerned with describing the extent and patterns of missing values in a dataset. The nonresponse process is introduced as a nuisance accompaniment of the sampling process and the ground is prepared for a discussion of the several common schemes for addressing data incompleteness in Chapter 3.

### 2.1 The problem of incompleteness

Most surveys rely on subjects' cooperation. We should therefore consider their perspective. As survey designers, interviewers or analysts, we may also be survey subjects. We can easily point to settings or circumstances in which we might respond negatively to an invitation to complete a questionnaire by an unsolicited phone call, or to a request for a face-to-face interview. There are dozens of activities more inspiring, entertaining, stimulating and rewarding than responding to a survey. So, the spectrum of our responses ranges from pretending total incomprehension, through polite or rude refusal, with or without a credible excuse, to reluctant cooperation that may be terminated as soon as we experience some further inconvenience, discomfort or some other perceived unpleasantness. We bear any intrusion with reluctance and jealously protect our privacy.

As a commodity, information is expensive but perishable — what is valuable today is discarded tomorrow or, at best, next week. Although the extent of missing data may be reduced by repeated calls and other time-consuming measures, a survey and its analysts and clients cannot always afford to wait until these measures have run their course.

However carefully a survey may be designed, the best plan is merely an ideal because complete cooperation of all the subjects, an assumption implied by the plan, is but an unattainable ideal. A simplified stereotype of a plan may be to collect the values of  $K$  variables from a random sample of  $n$  subjects drawn from a specified population. We may fail to identify the population

precisely because of migration, changes of status, errors in the sampling frame, and the like. Further, we may fail to contact some of the selected subjects or fail to enlist their cooperation. The cooperation may be interrupted during the interview or completion of the questionnaire, some questionnaire items may remain not responded because of oversight, deliberate omission, inability to respond, nonexistence of an appropriate response, and the like. Further losses can occur in the process of transcribing the collected responses to the computer (due to illegible hand-writing and clerical errors).

Such missing data is visible, easy to detect by inspecting the constructed database, if it is constructed appropriately. The data field reserved for a particular value (of a variable for a subject) is either empty or contains a symbol that indicates that the value has not been recorded. There may be several such symbols, one for each kind of missing value: for ‘do not know’, ‘not willing to tell’, for an apparently inadvertent omission, and the like.

What about subjects from whom we elicited no information? Why should the database be burdened by their records, full of missing values? The implied viewpoint tends to prevail at present. We will argue against it, and against the practice it encourages, on the grounds illustrated by the following comparison. Suppose two surveys, A and B, are conducted in the same population using the same sampling design, instruments and methods of data collection. Survey A has sample size 7650 and complete information is elicited from each selected subject, so that no data is missing. Survey B has sample size 10 000, but only 76.5% of the selected subjects cooperate with the survey completely, and no information is elicited from the remainder. In survey A, the planned and realised designs coincide, whereas in survey B they differ. If the analysis does not reflect the difference between the two surveys, or between the planned and realised sampling designs, we should seek fault with the method applied, not our intuition.

The sampling design is important because the claimed properties of the estimators used are contingent on the sampling design, assuming that it is implemented perfectly. One element of such perfection is that there is no missing data. The purpose of the sampling design is to extract the maximum information with the resources available for the survey. In practice, this is interpreted as ensuring good representation of the population — that the sample is a faithful image, in the miniature, of the population or, more formally, that the sample (empirical) distribution function of the values of any variable is an unbiased estimator of its population counterpart.

The sampling design might ensure this, but not if it is infiltrated with nonresponse. If the sample drawn is representative, and is then reduced by nonresponse, the remainder of the original sample (the respondents) may no longer be representative. For example, if non-responding subjects tend to be wealthier than the respondents, our conclusion about the wealth of the population is distorted. If we regard the complete respondents as the (original) sample, we have no means of detecting such a distortion. Non-respondents are the subjects who tell us nothing, so we have no means of knowing that their

absence from the database spoils the good representation that was arranged by the sampling design. We should strive to overcome this problem, even if in some circumstances it may appear prudent to defuse it by focussing on the population of respondents. Although with apparently greater competence, we would then solve a less relevant problem, because the original inferential task relates to the complete population, not to any of its opportunistically defined subpopulations.

The lack of any evidence that nonresponse causes a problem does not justify ignoring it, because the appropriate interpretation of no evidence is ‘do not know’. For instance, no evidence may arise as a result of no inquiry. To justify ignoring nonresponse, evidence is required that it causes no problem.

The first step in dealing with nonresponse, or controlling its impact, is a survey of the damage. For missing data, this amounts to describing the extent of missing values, classified according to a suitable nomenclature. Although the party in charge of data collection has incentives to present the problem with as little fanfare as possible, there are ample long-term rewards for honesty and integrity. Suppose a survey has 20% of total nonresponse (non-contacts and outright refusals), and it is much lower than in surveys of similar populations and with similar content and protocol. The relatively high response rate does not justify ignoring the problem of nonresponse altogether. We should, at least informally, play the devil’s advocate and contemplate what impact the 20% of the subjects might have had on the planned or intended inferences, had they all responded. It is easy to construct scenarios in which as little as 5% nonresponse results in a substantial distortion of the inferences. For instance, if in a country with low unemployment rate most of the unemployed do not respond to a survey that inquires about their employment status, and most employed and other subjects do, the estimate of the unemployment rate is bound to be problematic. The percentages (rates) of nonresponse (for each variable recorded in the survey), although easy to establish, are but one aspect of the problem. The impact on the planned inferences is what matters because the survey and its analysis have been undertaken specifically for the purpose of drawing the inferences.

The choice of the method for dealing with missing data is informed not only by the extent but also by the *pattern* of the missing data — whether subjects tend to omit responses to isolated questions or to whole sections (contiguous blocks) of questions, whether some questions are (almost) always or never responded when some other questions are responded or not, or whether nonresponse is associated with the values of one or a set of variables.

We conclude this section on a note of pedantry. Although the literature commonly refers to ‘nonresponse’, a more precise term for missing values is ‘no record’. That is, a subject might have responded to a particular questionnaire item, but the processes that followed led to a missing code (blank) being entered, appropriately or not, in the corresponding location in the database. Although we prefer the phrase ‘value not recorded’, it is impossible to avoid

the term nonresponse, as there is no alternative single word for the failure to record a value.

## 2.2 The extent of missing data and the response pattern

In this section, we define a terminology for missing data. We deal only with its *visible features* that can be summarised in one word as frequency of various combinations of missing and recorded values in the records. Section 2.3 discusses invisible features of missing data; they are properties of the process of nonresponse (missingness). We introduce first the general setting, some notation and related conventions.

We consider a fixed (frozen) population, so that there is no ambiguity about any entity whether it is a member of the population or not. The size of the population (number of its elements) is denoted by  $N$ . In general, we use capitals for population quantities and (the corresponding) lowercases for sample quantities, although this notation is difficult to adhere to consistently. For example, boldface capitals are also used for matrices and boldface lowercases for their rows.

A population quantity can be derived with precision only if the relevant data items are available for the entire population or its a priori defined subset. Sampling has no impact on a population quantity. A sample quantity depends on the sample — it is a random variable prior to sampling, and a constant thereafter. Its value can be established when the sampling process is executed perfectly. A sampling-process quantity depends on the sampling process. That is, its value could be established with a specified precision if the sampling design were executed sufficiently many times. Most sample quantities are estimators; most sampling-process quantities describe estimators. For example, the population mean is a population quantity, it is estimated by the sample mean, a sample quantity, and the bias and sampling variance of the sample mean are sampling-process quantities. The sampling variance may be estimated; the estimator is a sample quantity.

Let  $n^*$  be the planned sample size of a survey; it may be a random variable. Suppose the survey collects the values of  $K$  variables. The *complete data* is the hypothetical dataset that was planned to be collected by the survey. Anticipating nonresponse, the planners may have resigned themselves to obtaining this dataset with some of its values missing, but they issued instructions and implemented measures aiming to collect every item of the complete dataset. In most instances, it is a  $n^* \times K$  rectangular array. The collected data is referred to as the observed data, and it may be characterised as *incomplete*; the missing data is, in this context, defined as the difference between the complete and incomplete datasets.

We can define the terms ‘complete’, ‘incomplete’ and ‘missing’ for subsets of data. These subsets can be formed by keeping only some of the variables, only some of the subjects (reducing our attention to a subpopulation), and

by the combination of these two ways of reducing the data. A variable is said to be recorded completely if its value is recorded for every subject; that is, if the dataset reduced to the single column is complete. Similarly, the record of a subject is complete if the value of each variable for the subject is recorded. Otherwise, the record is called incomplete. A record is called *empty* if all its values are missing. A record that is neither empty nor complete is called *partial*. We could use the term ‘empty’ also for a variable that has not been recorded for any subject.

The extent of missing data can be summarised by the numbers or percentages of empty and incomplete records. More detailed summary is provided by these numbers or percentages for various important subsets of variables, such as blocks of questionnaire items. Further, the number or percentage of missing values can be given for each variable.

A rather coarse classification of the nonresponse is to unit and item nonresponse. Unit nonresponse refers to an empty record, when the unit (subject) has provided no data. Item nonresponse refers to a missing item — the subject concerned cooperated with the survey, but only partially. More detail can be introduced by distinguishing parts, or sections, of the survey. Not cooperating with a section can be called section nonresponse. For example, the section may be the questionnaire administered at a given time point in a longitudinal study.

### Example 1

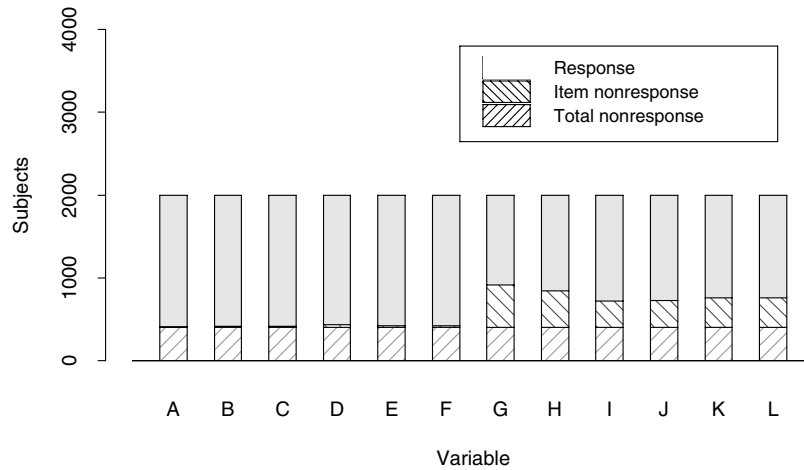
In a survey with a planned sample size of 2000 (human) subjects, 174 were outright refusals and further 229 sampled subjects were not contacted (either not located or not found at home). Further, among the  $2000 - 229 - 174 = 1597$  responding subjects, complete records on the twelve variables on which we focus are available for only 1088 subjects.

The summaries defined for this dataset are: the rate of total nonresponse,  $(1 - 1597/2000) \times 100 \doteq 20\%$ , and the rate of partial (incomplete) cooperation,  $(1 - 1088/2000) \times 100 \doteq 46\%$ . The complements of these rates, 80% and 54%, are the respective rates of at-least-partial and perfect cooperation. Table 2.1 gives the nonresponse rates for each of the twelve variables. Figure 2.1 presents these rates graphically.

These summaries indicate that variables A–F are responded by most of the 1597 cooperating subjects. For example, the response to A is not recorded for only  $407 - 403 = 4$  of them. The nonresponse rates are much higher for variables G–L. The total number of missing values is  $407 + 415 + \dots + 761 = 7225$ , out of  $12 \times 2000 = 24\,000$ , but  $12 \times 403 = 4836$  of them are for unit nonresponse (total non-respondents). The remainder, 2389 values, are distributed among the 509 subjects who have partial records. These subjects have only 89 missing values on the variables A–F. On the other six variables, G–L, they have 2300 missing values, about four-and-a-half per subject. This implies that many of these subjects have five or all six values missing.

**Table 2.1.** The nonresponse rates for the twelve variables, A–L, in a dataset. A fictitious example.

Variable	A	B	C	D	E	F
Number of missing values	407	415	413	431	422	419
Nonresponse rate (%)	20.3	20.7	20.6	21.5	21.1	20.9
Variable	G	H	I	J	K	L
Number of missing values	912	844	717	728	756	761
Nonresponse rate (%)	45.6	42.2	35.8	36.4	37.8	38.0

**Figure 2.1.** A graphical display of the nonresponse rates for variables A–L, given in Table 2.1.

Although offering important insight, Table 2.1 does not contain all the information that might be useful to have. For example, we cannot establish how many subjects have empty records on the sets (segments) of variables A–F and G–L. For A–F, it may be only the 403 total non-respondents, and at most four others, whereas for G–L it could be as many as 314 in addition to the total non-respondents.

For a more detailed description of the nonresponse (or response), we define the *response pattern*. The *indicator of response* is an object of the same shape and size as the complete dataset, in our case a  $n^* \times K$  matrix, in which

**Table 2.2.** Response patterns in an incomplete dataset. A fictitious example.

	Pattern						
	000000	100000	101000	111000	111001	1110111	111111
Subjects	407	6	2	4	3	9	1569

	Pattern						
	000000	001000	001100	001110	001111	011111	111111
Subjects	717	11	28	5	83	68	1088

one symbol is used to indicate that the corresponding value in the dataset is missing, and another that it is recorded. As a convention, 0 is used for a missing and 1 for an available (recorded) item. The indicator of response is denoted by  $\mathbf{R}$ . (We can refer to  $\mathbf{R}$  also as the *indicator of nonresponse*.) An obvious generalisation is to use different symbols for each kind of missing value. For example,  $-1$  may be reserved for ‘no contact’,  $-2$  for ‘refusal’,  $-3$  for ‘do not know’, and similar. For simplicity, we assume that such detail is not given and  $\mathbf{R}$  comprises zeros and unities.

The response pattern for a subject (record) is defined as the corresponding row of  $\mathbf{R}$ . It is a (binary) vector and, by ‘gluing’ its elements together, it can be represented as a sequence of zeros and ones. Thus,  $\mathbf{1} = 11 \dots 1$  represents a complete record,  $\mathbf{0} = 00 \dots 0$  an empty record, and so on. For a small number of variables, the patterns can be summarised by their tabulation. For  $K$  variables, there may be up to  $2^K$  distinct patterns. It is useful to find out whether only a limited set of patterns occur in the data, or whether the vast majority of records have one or a small number of patterns. Table 2.2 gives an example, summarising the patterns of the same dataset as in Table 2.1. The patterns are summarised separately for the sets of variables A–F and G–L, both to conserve space and to get a better insight.

First we note that the number of distinct patterns, seven for both A–F and G–L, is much smaller than what we may have feared —  $2^6 = 64$  for either set of variables. For variables A–F, there are, in addition to the 403 total non-respondents, four subjects with empty records. For variables G–L, there are  $717 - 403 = 314$  such subjects; 195 subjects have one of the five partial response patterns. The most frequent partial patterns are 001111 and 011111. Subjects with this pattern appear to have started their cooperation on the block G–L with delay. (We explore their response pattern on the block A–F below.) A notable feature of the patterns for variables G–L is that the responses are concentrated in contiguous sets of variables, such as I–K for pattern 001110.

Table 2.2 still fails to inform about the pattern for the entire set of variables, A–L. Since the number of patterns does not exceed 49, we could list

**Table 2.3.** The cross-tabulation of the response patterns for the sets of variables A–F and G–L.

Pattern for A–F	Pattern for G–L						
	000000	001000	001100	001110	001111	011111	111111
000000	407	0	0	0	0	0	0
100000	6	0	0	0	0	0	0
101000	2	0	0	0	0	0	0
111000	4	0	0	0	0	0	0
111001	3	0	0	0	0	0	0
111011	9	0	0	0	0	0	0
111111	286	11	28	5	83	68	1088

them, although the two-way table of patterns for the variables A–F and G–L, displayed in Table 2.3, may be easier to digest. All the counts in this table are concentrated in the first column and last row. This indicates that no subjects have partial records on both segments A–F and G–L. The records are either empty (407 records), complete (1088 records), complete for A–F but empty for G–L (286 records), partial for A–F and complete for G–L (24 records), or complete for A–F and partial for G–L (195 records).

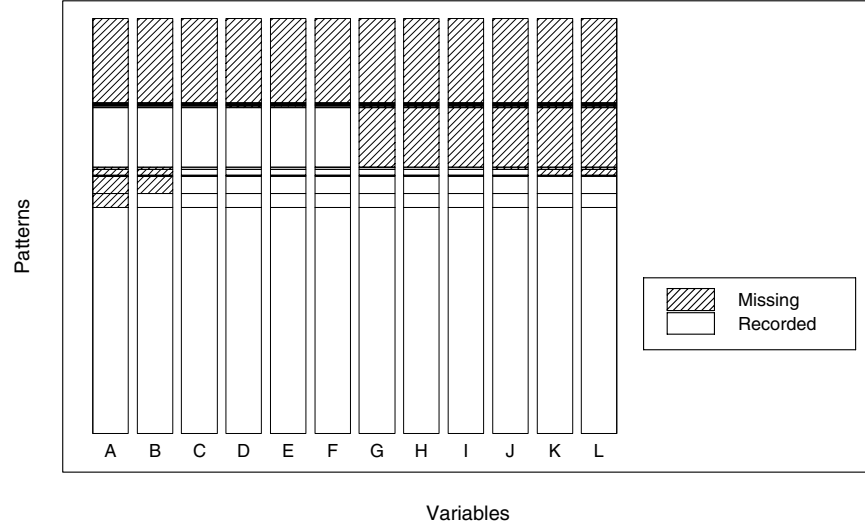
We can define a partial ordering according to the pattern of nonresponse. One variable is said to be recorded more than another if the only response patterns occurring for the two variables are 00, 11, 10; that is, when the first variable is not recorded, neither is the second. Records, or their patterns, can be compared similarly. For example, 111011 represents more response than 111000, although 110011 does not represent more response than 001000.

The patterns can be displayed graphically, by a  $n^* \times K$  array of cells (symbols or squares) with different symbols, colours, shading, or the like, indicating whether the item has been recorded or not. A clearer impression of the distribution of the patterns is created if the subjects are permuted so that records with the same pattern form a contiguous segment. When the sample size  $n^*$  is large, it is practical not to draw the cells but represent a given number of records (rows) by a unit height in the rectangle representing the data. The variables can also be permuted to make the presentation clearer. An example is given in Figure 2.2.

### 2.2.1 Monotone response patterns

The variables in a dataset have a partial ordering according to the extent of their missing values. The response patterns of a dataset are said to be *monotone* if the variables can be permuted so that any variable is recorded



**Figure 2.2.** A graphical summary of the patterns of nonresponse.

more than the following variable; nonresponse to a variable by a subject is followed by the subject's nonresponse to all the subsequent variables.

For vectors  $\mathbf{X}_1$  and  $\mathbf{X}_2$  of the same length, we introduce the notation

$$\mathbf{X}_1 \succeq \mathbf{X}_2$$

if  $\mathbf{X}_1$  exceeds or is equal to the value of  $\mathbf{X}_2$  for every subject. The symbols  $\succ$ ,  $\prec$  and  $\preceq$  are defined similarly. For instance,  $\mathbf{X}_1 \succ \mathbf{X}_2$  if  $\mathbf{X}_1$  exceeds  $\mathbf{X}_2$  for every subject. Variable  $\mathbf{X}_1$  is recorded more than  $\mathbf{X}_2$  if for the corresponding vectors of response indicators we have  $\mathbf{R}_1 \succeq \mathbf{R}_2$ .

For a dataset  $\mathbf{X}$  with columns  $\mathbf{X}_1, \dots, \mathbf{X}_K$ , and response indicator  $\mathbf{R}$  with columns  $\mathbf{R}_1, \dots, \mathbf{R}_K$ , monotone response patterns are defined by the string of ordering  $\mathbf{R}_1 \succeq \dots \succeq \mathbf{R}_K$ . We can distinguish between 'recorded more than' and 'recorded at least as much as', but this will not be essential at any point.

In Section 2.3, we consider methods for completing the recorded (incomplete) data by substituting a value for each missing item. We assume that the value is well defined, even though it was not recorded. Since we do not know the value, we search for information on which to base our guess. For a subject, the available part of the record is an obvious candidate for this purpose. The more was recorded from the subject, the better our prospects of a good guess. Obviously, fewer missing items have to be filled in for (*imputed*) when more items are recorded. But we also possess more information on which to base the imputation. If we want to use all the information available about a subject we have to devise a different way for each pattern. So, the distribution of the

patterns helps us draft a strategy for this task. In Chapter 4, and Section 4.2 in particular, we find that efficient methods for imputation are much easier to implement when the data has monotone response patterns.

## 2.3 Sampling and nonresponse processes

Why did we fail to collect a particular item of information? At first sight, this question is solely about nonresponse. On reflection, there are two classes of answers: because we did not intend to (as when the subject is not in the sample), and because we failed to elicit a response that, in principle, could have been recorded.

By a (random or stochastic) *process* we refer to one or a collection of random variables that describe a studied phenomenon or some of its ingredients. In a typical survey, we may consider a *data-generating* process describing how the members of the population acquire their values of a random variable or vector, the *sampling* process, describing how some members of the population end up being the subjects in the sample, and the *nonresponse* process, describing how we fail (or succeed) to elicit and record the elements of the planned (complete) dataset. In this chapter, we are not concerned with the data-generating process (for instance, how certain members of the labour force end up being unemployed, at a certain time point), although the purpose of the survey may be to learn about this process. The nonresponse process is formally defined as the conditional distribution of the response indicator  $\mathbf{R}$  given the complete data  $\mathbf{X}^*$ ,

$$(\mathbf{R} \mid \mathbf{X}^*).$$

In formulas, we refer to distributions of random variables or vectors by parentheses  $(\ )$ , to conditioning by the vertical bar  $\mid$ , and to equality in distribution by the symbol  $\sim$ . For example,

$$(\mathbf{R} \mid \mathbf{X}^*) \sim (\mathbf{R})$$

denotes that the conditional distribution of the response indicator  $\mathbf{R}$  given the complete data coincides with the (unconditional) distribution of  $\mathbf{R}$ . That is,  $\mathbf{R}$  and  $\mathbf{X}^*$  are independent. This distributional identity is not true in general. The missing data is denoted by  $\mathbf{X}_{\text{mis}}$ .

A typical survey involves several other processes, such as questionnaire development (piloting) and interviewing. Although the interviewer is meant to be an inert instrument in eliciting responses, different interviewers might have elicited differing responses from the same subject, had such a replication been realised (a repeated interview, separated by a period in which the subject has forgotten the experience of having been interviewed for the first time, and would not recall the responses he or she gave earlier). For instance, the interviewer in a survey has to make an assessment of the need for repairs of the inspected dwelling; different interviewers (surveyors) may come

to different conclusions when inspecting the same dwelling. The underlying *assessment* process has an impact on the quality of the collected data, and consequently on the quality of the inferences. We could consider an *ideal* assessment for each subject, and the assessment by the surveyor or interviewer as its *manifest* (error-prone) version. The ideal assessment is a completely missing variable, but the realised assessment contains a lot of information about it, especially when the assessors make ‘mistakes’ only rarely, and most of them are only minor. This is an example of planned ‘nonresponse’ and it indicates that methods for dealing with missing data may be applicable in some less conventional settings. They are explored in greater detail in Section 4.6.

The nonresponse process describes the momentary influences on the subject’s response. If the subject were asked the same questions about a stable attribute, such as consumption of a food item, he or she may respond differently, depending on the momentary disposition, vagaries of the recall and formulation of the response. In this case, the ideal response is missing for every subject and the recorded response is its manifest version; it informs about the ideal value imperfectly.

The sampling process reduces the information from the population to the (complete) sample. The role of the sampling design is to minimise the loss of information given the resources available for the conduct of the survey. Given adequate resources and perfect implementation, the design ensures that we can make (sample-based) inferences about the population. An imperfect response process reduces the complete sample further, to the incomplete sample (information). The main qualitative difference between the sampling and nonresponse processes is that the former is under our control, by means of the sampling design prescribing the probabilities that a subset of the population forms the sample. The sampling process has a formal description as a function  $\pi$  on  $\exp(\mathcal{P})$ , the set of all subsets of the studied population  $\mathcal{P}$ ; for  $s \in \mathcal{P}$ ,  $\pi(s)$  is the probability that  $s$  forms the sample.

In contrast, the nonresponse process is usually oblivious to the sampling design — the subject’s reasons for not responding are unrelated to the sampling plan. Because it is outside our control, we should be concerned that the nonresponse process may spoil the representativeness of the sample. We can easily construct scenarios in which the representativeness is severely undermined. As an example, suppose in a survey aimed to estimate the unemployment rate, the unemployed tend to be much more difficult to contact or they are more reluctant to respond to the relevant questionnaire items. More subtle processes may be at play, such as when non-studying young men in urban areas have a lower response rate, or when the response rates among quite finely divided subpopulations do not differ a great deal, but in certain groups unemployed and in others employed subjects are less likely to respond.

This suggests two approaches. To ignore the issue, since the ‘correct’ answer is beyond the realm of possibilities, or to give up on the original goal of estimating the specified population quantities altogether because of a gross

failure in the data collection process. Neither approach is very constructive. Instead, we will speculate about the possible nonresponse processes, draw inferences assuming these processes, and then explore how the inferences change as the assumed process is altered. In this way, we take a risk, but assess, informally, its magnitude. We will also look for means of reducing the risk by searching for insights about the nonresponse process. First we define a typology for the nonresponse processes, starting with the setting of a survey in which a single variable is recorded.

### 2.3.1 The nature of the nonresponse process

Since nonresponse is a process akin to sampling, we could describe it using the terminology from sampling theory. As nonresponse could in principle be described and motivated as a result of certain decisions or actions, it is also called the nonresponse *mechanism*. The ideal, usually not attainable, is to find a complete description of this mechanism, so that, for instance, we could simulate it on a computer.

In the simplest conceivable nonresponse mechanism, subjects who fail to respond are as if selected by simple random sampling (SRS) from all the subjects selected by the sampling process. The nonresponse mechanism is independent of the complete data:

$$(\mathbf{R} \mid \mathbf{X}^*) \sim (\mathbf{R}) . \quad (2.1)$$

With such a mechanism, the data are said to be *missing completely at random* (MCAR). As SRS is a very special sampling process, we cannot expect, without exercising any control over it, that the nonresponse would be MCAR. Usually we have no means of establishing that a nonresponse mechanism is MCAR. A more plausible assumption is that the mechanism belongs to a more general class.

A class of sampling designs more general than SRS is *stratified* simple random sampling (sSRS). In sSRS, the population is classified into strata (subpopulations), and simple random sampling (with stratum-specific probabilities of inclusion) is applied in each stratum. The stratification is given by a categorical variable defined in the surveyed population. Stratification based on a (categorical) variable  $A$  is said to be more detailed than stratification based on  $B$ , if  $B$  can be formed by aggregating (collapsing) some of the categories of  $A$ .

In a more general sampling design, the probability of inclusion is a function of one or several variables, and the inclusions are mutually independent. Such a design can be motivated by defining a sequence of designs with more and more detailed stratification. In the corresponding nonresponse mechanisms, data is said to be *missing at random* (MAR). A key characterisation of MAR is that the response indicator depends on the complete data only through its recorded part:

$$(\mathbf{R} | \mathbf{X}^*) \sim (\mathbf{R} | \mathbf{X}) . \quad (2.2)$$

That is, the missing data  $\mathbf{X}_{\text{mis}}$  contains no information about  $\mathbf{R}$ . Note that missing data contains information about most population quantities. Although much more general than MCAR, it is easy to construct mechanisms that are not MAR. In all such mechanisms, data is said to be *not missing at random* (NMAR). In NMAR, the response indicator depends on the missing data. NMAR contains all manner of ‘strange’ mechanisms, as illustrated in Figure 2.3. In each panel, the histogram of the complete data is composed of the missing values, represented by the shaded bars, and the recorded values by the plain bars above them. In panel MCAR, the probability of missing, equal to 0.25, is the same within every interval (bar). Panel NMAR 1 depicts a mechanism with higher response rates for the smallest and largest values of  $X$ , NMAR 2 a mechanism with lower response rates for the extreme values, and NMAR 3 a mechanism with response rates decreasing with the value of  $X$ . These examples are in no way exhaustive. Any idiosyncratic mechanism is an example of NMAR. MAR mechanisms that are not MCAR are more difficult to represent graphically because they involve at least two variables.

On the one hand, we should be aware of NMAR mechanisms and contemplate how they might affect our inferences. On the other hand, we should be realistic and, while not subscribing to the assumption of a limited class of non-response mechanisms, such as MCAR, restrict our attention to the range of NMAR mechanisms that are plausible. Intelligence about the studied setting that reduces this range is particularly valuable.

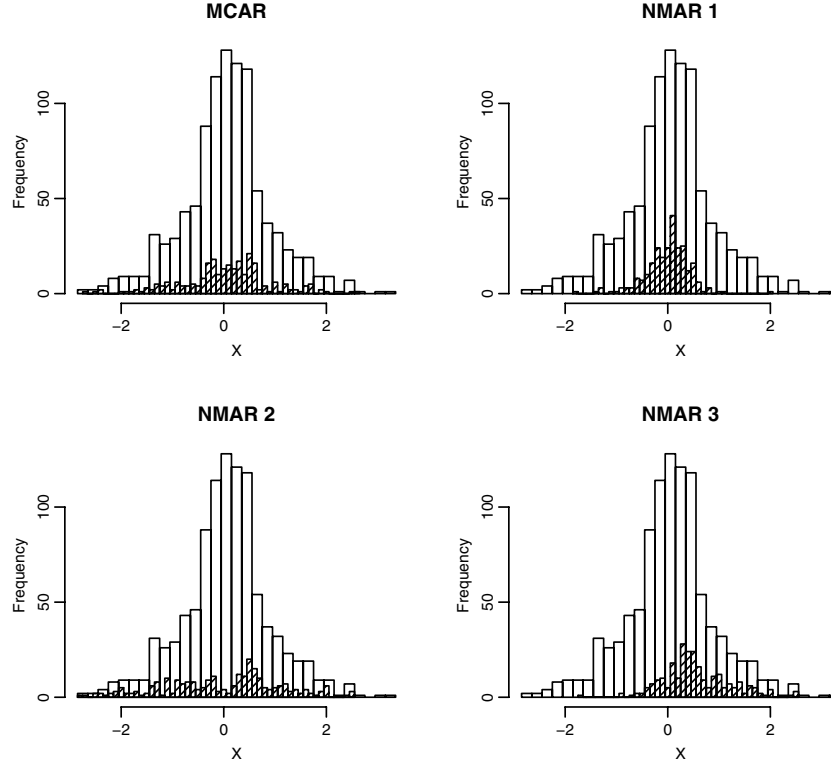
### Example 2

Suppose a survey collects the values of a single categorical variable  $X$  from a sample  $s$ , with a sampling design defined by  $\pi$ . If the nonresponse mechanism is MCAR the probability of response is the same for each value of  $X$ . If the nonresponse mechanism is MAR the probability of response does not depend on the subject’s value of  $X$ , but may depend on the (known) inclusion probability  $\pi_i$ . An example of NMAR arises when the probabilities of response depend on  $X$ . For example, the subjects with a particular value of  $X$  are much more reluctant to respond, and those with other values of  $X$  are more forthcoming.

### Several variables in $\mathbf{X}^*$

Although the definitions of MCAR, MAR and NMAR are easier to interpret for single variables, their definitions in terms of  $(\mathbf{R} | \mathbf{X}^*)$  apply to sets of variables in  $\mathbf{X}^*$ . Simply, the joint distribution of the  $n \times K$  elements of  $\mathbf{R}$  is independent of the complete data  $\mathbf{X}^*$  (MCAR), or depends on it only through the incomplete data  $\mathbf{X}$  (MAR).

**Figure 2.3.** Examples of MCAR and NMAR mechanisms. Missing values are represented by the shaded sections of the bars.



It is useful to separate the survey variables  $\mathbf{X}$  into those that may be recorded incompletely,  $\mathbf{Y}$ , and those that never contain any missing values,  $\mathbf{Z}$ ;

$$\mathbf{X} \sim (\mathbf{Y} \ \mathbf{Z}) .$$

For example, the values of some of the variables in  $\mathbf{Z}$  may be available prior to interviewing and those in  $\mathbf{Y}$  are established by the interview. Variables that describe the circumstances of the interview (whether completed or not, whether conducted at the first appointment, and the like), usually belong to  $\mathbf{Z}$ . We can draw a distinction between completeness of a variable in the process of data collection and in the realised dataset. The former refers to hypothetical replications of the survey. When the sample size is large the distinction is unimportant.

### 2.3.2 The importance of MAR

The importance of MAR stems from a characterisation alternative to (2.2): when MAR applies the joint distribution of  $\mathbf{X}$  for subjects with incomplete records is the same as for subjects with complete records:

$$(\mathbf{x} \mid \mathbf{r} = \mathbf{1}) = (\mathbf{x} \mid \mathbf{r} = \mathbf{r}^*), \quad (2.3)$$

where  $\mathbf{r}^*$  is any response pattern for  $\mathbf{x}$  (a row of  $\mathbf{X}$ ). The characterisation in (2.3) provides an important recipe for dealing with nonresponse. We establish, or estimate, the associations among the variables in  $\mathbf{X}$  for subjects with complete records (pattern  $\mathbf{1}$ ), and then assume that it applies also for the incomplete records. Before doing this, we have to be satisfied that the nonresponse mechanism is MAR. This we can rarely accomplish analytically, but that should not stop us from proceeding by assuming MAR. The best we can do is to reduce as much as possible the error incurred in the inferences that can be attributed to the assumption. Drawing inferences from  $\mathbf{X}$  is much easier under MAR because NMAR includes a range of mechanisms in which some combinations of values of  $\mathbf{X}$  that are infrequent among the complete records are quite frequent among the incomplete records. Whether MAR applies, as well as the departure from it, depends on the variables  $\mathbf{X}$ . If a nonresponse mechanism is MAR it will remain so when variables are added to  $\mathbf{X}$ . However, a NMAR mechanism need not become MAR when variables are added to  $\mathbf{X}$ . The variables considered,  $\mathbf{X}$ , are an important qualification of MAR and NMAR.

This suggests that when planning a survey we should think not only about recording the *outcome* variables directly connected with the desired inferences, but also variables that promote MAR. Although such *auxiliary* variables may also be recorded incompletely, they may be helpful nevertheless.

#### Example 3

One-week diaries of alcohol consumption in a survey of middle-aged people in the UK are analysed by [166]. Diaries are regarded as the most reliable way of collecting information about the consumption of food and drink, even though incomplete and empty (no-response) diaries are quite frequent. Keeping a diary for a whole week requires a lot of commitment from the subjects. In the survey, almost all subjects completed the diary for the first two days, but many dropped out thereafter. The subjects were also asked to recall how much alcohol they consumed during the previous week and were given a set of four brief questions about problems related to their past alcohol consumption (CAGE, [61]).

The response rate to the four recall questions (about drinking beer, wine, sherry and spirits) was much higher because the questions can be responded within a short time, after a cursory recall. The higher response rate comes at

the price of lower reliability. The CAGE questions (response options Yes/No) also had very little nonresponse. Numerous other variables were collected, such as smoking habit, gender, body mass and height.

An obvious concern with diary data is that subjects may drop out from keeping it because of embarrassment over excessive consumption. Thus, the incomplete diaries would deceive the analyst by stopping at a day preceding excessive consumption. Such deception is an example of NMAR. From the recorded data we cannot infer whether it is present, and to what extent. However, the other variables related to alcohol consumption may provide some insights. Concerns about NMAR would be well supported if there were many subjects who declared substantial consumption in the recall and much lower consumption in their incomplete diaries, even after pro-rating for the number of completed diary days.

The recall variables play the role of auxiliary information that makes MAR more plausible. As an outcome they are not suitable, but as informants about the missing values that would have been derived from the diaries they are ideal. Details of the analysis are discussed in Section 5.2.

## 2.4 Exercises

1. Find or construct examples of nonresponse mechanisms that are MCAR, MAR and NMAR, and examples that are NMAR, without conditioning on a particular variable, but are MAR otherwise.
2. For a given incomplete dataset of at least 1000 subjects and several variables, write a programme (`Splus` function) to summarise the response patterns by a table and graphically. Look for ways of excluding as few subjects as possible to make the patterns monotone.
3. Unit nonresponse is encountered in a survey of a particular human population. The distribution of age within the sexes is known from a census or register. To assess whether the nonresponse presents a problem for the planned (complete-data) analyses, tests are carried out of the hypotheses that the proportions of the sexes and the sample distributions of age within the sexes are compatible with the population distribution. Provide a critique of this approach.
4. Simulate the values of a log-normally distributed variable in a population of at least 20 000 subjects. Regard the variable as the annual income. Define a small number of cut-points for ‘income brackets’ and devise methods for estimating the population mean income from the tabulation according to these income brackets. Draw a large number of samples from the population (according to the same sampling design, say, SRS without replacement with sample size 500), and compare the distributions of the sample means for the income and the income bracket.
5. Formulate the problem of estimating the population mean income from the income bracket data as a case of missing information.



6. In a national crime survey, interviews are conducted with a sample drawn from the country's population of households. An adult member of the household is asked to recall all instances of crime committed against any member of the household. Why does the survey not collect any information about the crimes committed *by* the interviewees and members of their households? What is the likely difference between the records of victimisation from the interviewees and records from the Police?
7. Data about total alcohol consumption in a year in the UK could be obtained from the records of payments of excise duty on alcohol. What information about alcohol consumption may be obtained by surveys of the (adult) population that could not be extracted from the excise duty enumeration?
8. Suppose a survey with sample size 8000 collects information about 30 variables, but item nonresponse occurs by MCAR with the same probability, 0.01, for every item and the event of nonresponse is independent across items. Calculate the expectation and variance of the number of incomplete records. Simulate this setting on a computer and verify the calculation. Alter the nonresponse process so that the events of nonresponse are dependent within subjects.
9. Describe the response patterns obtained in the previous example, by tables and graphs, and relate the within-subject dependence of response to the distribution of the patterns.

Missing Data and Small-Area Estimation  
Modern Analytical Equipment for the Survey Statistician  
Longford, N.T.  
2005, XVI, 360 p. 45 illus., Hardcover  
ISBN: 978-1-85233-760-5