

---

## 2.1 A Step Not to Be Taken Lightly

Building a reliable and relevant database is a key aspect of any statistical study. Not only can misleading information create bias and mistakes, but it can also seriously affect public decisions if the study is used for guiding policy-makers. The first role of the analyst is therefore to provide a database of good quality. Dealing with this can be a real struggle, and the amount of resources (time, budget, personnel) dedicated to this activity should not be underestimated.

There are two types of sources from which the data can be gathered. On one hand, one may rely on pre-existing sources such as data on privately held companies (employee records, production records, etc.), data from government agencies (ministries, central banks, national institutes of statistics), from international institutions (World Bank, International Monetary Fund, Organization for Economic Co-operation and Development, World Health Organization) or from non-governmental organizations. When such databases are not available, or if information is insufficient or doubtful, the analyst has to rely instead on what we might call a homemade database. In that case, a survey is implemented to collect information from some or all units of a population and to compile the information into a useful summary form. The aim of this chapter is to provide a critical review and analysis of good practices for building such a database.

The primary purpose of a statistical study is to provide an accurate description of a population through the analysis of one or several variables. A variable is a characteristic to be measured for each unit of interest (e.g., individuals, households, local governments, countries). There are two types of design to collect information about those variables: census and sample survey. A census is a study that obtains data from every member of a population of interest. A sample survey is a study that focuses on a subset of a population and estimates population attributes through statistical inference. In both cases, the collected information is used to calculate indicators for the population as a whole.

Since the design of information collection may strongly affect the cost of survey administration, as well as the quality of the study, knowing whether the study should be on every member or only on a sample of the population is of high importance. In this respect, the quality of a study can be thought of in terms of two types of error: sampling and non-sampling errors. Sampling errors are inherent to all sample surveys and occur because only a share of the population is examined. Evidently, a census has no sampling error since the whole population is examined. Non-sampling errors consist of a wide variety of inaccuracies or miscalculations that are not related to the sampling process, such as coverage errors, measurement and nonresponse errors, or processing errors. A coverage error arises when there is non-concordance between the study population and the survey frame. Measurement and nonresponse errors occur when the response provided differs from the real value. Such errors may be caused by the respondent, the interviewer, the format of the questionnaire, the data collection method. Last, a processing error is an error arising from data coding, editing or imputation.

Before deciding to collect information, it is important to know whether studies on a similar topic have been implemented before. If this is to be the case, then it may be efficient to review the existing literature and methodologies. It is also critical to be clear on the objectives, especially on the type of information one needs (individuals and organizations involved, time period, geographical area), and on how the results will be used and by whom. Once the process of data collection has been initiated or a fortiori completed, it is usually extremely costly to try and add new variables that were initially overlooked.

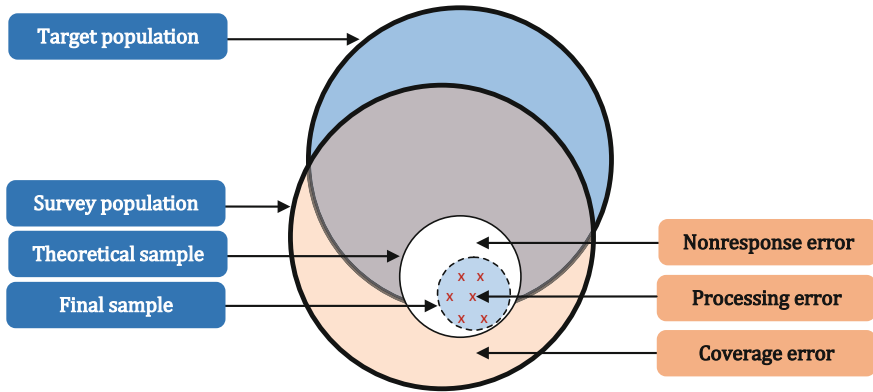
The construction of a database includes several steps that can be summarized as follows. Section 2.2 describes how to choose a sample and its size when a census is not carried out. Section 2.3 deals with the various ways of conceiving a questionnaire through different types of questions. Section 2.4 is dedicated to the process of data collection as it details the different types of responding units and the corresponding response rates. Section 2.5 shows how to code data for subsequent statistical analysis.

---

## 2.2 Choice of Sample

First of all, it is very important to distinguish between the target population, the sampling frame, the theoretical sample, and the final sample. Figure 2.1 provides a summary description of how these concepts interact and how the sampling process may generate errors.

The target population is the population for which information is desired, it represents the scope of the survey. To identify precisely the target population, there are three main questions that should be answered: who, where and when? The analyst should specify precisely the type of units that is the main focus of the study, their geographical location and the time period of reference. For instance, if the survey aims at evaluating the impact of environmental pollution, the target population would represent those who live within the geographical area over which



**Fig. 2.1** From the target population to the final sample

the pollution is effective or those who may be using the contaminated resource. If the survey is about the provision of a local public good, then the target population may be the local residents or the taxpayers. As to a recreational site, or a better access to that site, the target population consists of all potential users. Even at this stage carefulness is required. For instance, a local public good may generate spill-over effects in neighboring jurisdictions, in which case it may be debated whether the target population should reach beyond local boundaries.

Once the target population has been identified, a sample that best represents it must be obtained. The starting point in defining an appropriate sample is to determine what is called a survey frame, which defines the population to be surveyed (also referred to as survey population, study population or target population). It is a list of all sampling units (list frame), e.g., the members of a population, which is used as a basis for sampling. A distinction is made between identification data (e.g., name, exact address, identification number) and contact data (e.g., mailing address or telephone number). Possible sampling frames include for instance a telephone directory, an electoral register, employment records, school class lists, patient files in a hospital, etc. Since the survey frame is not necessarily under the control of the evaluator, the survey population may end up being quite different from the target population (coverage errors), although ideally the two populations should coincide.

For large populations, because of the costs required for collecting data, a census is not necessarily the most efficient design. In that case, an appropriate sample must be obtained to save the time and, especially, the expense that would otherwise be required to survey the entire population. In practice, if the survey is well-designed, a sample can provide very precise estimates of population parameters. Yet, despite all the efforts made, several errors may remain, in particular nonresponse, if the survey fails to collect complete information on all units in the targeted sample. Thus, depending on survey compliance, there might be a large difference between the theoretical sample that was originally planned and the final sample. In addition to these considerations, several processing errors may finally affect the quality of the database.

A sample is only a portion of the survey population. A distinction is consequently made between the population parameter, which is the true value of the population attribute, and the sample statistic, which is an estimate of the population parameter. Since the value of the sample statistic depends on the selected sample, the approach introduces variability in the estimation results. The computation of a margin of error  $\pm e$  is therefore crucial. It yields a confidence interval, i.e. a range of values, which is likely to encompass the true value of the population parameter. It is a proxy for the sampling error and an important issue with sampling design is to minimize this confidence interval.

How large should a sample be? Unfortunately, there is no unique answer to this question since the optimal size can be thought of in terms of a tradeoff between precision requirements ( $\pm e$ ) and operational considerations such as available budget, resources and time. Yet, an indicative formula provides the minimum size of a sample. It is based on the calculation of a confidence interval for a proportion. As an illustration, assume that one wishes to estimate the portion of a population that has a specific characteristic, such as the share of males. The true population proportion is denoted  $\pi$  and the sample proportion is denoted  $p$ . Since  $\pi$  is unknown, we can only use the characteristics of the sample to compute a confidence interval. Assume for instance that we find  $p = 45\%$  (i.e. 45 percent of the sample units are male) and calculate a margin of error equal to  $e = 3\%$ . The analyst can specify a range of values  $45\% \pm 3\%$  in which the population parameter  $\pi$  is likely to belong, i.e. the confidence interval is  $[42\%, 48\%]$ . Statistical precision can thus be thought of as how narrow the confidence interval is.

The formula for calculating a margin of error for a proportion is:

$$e = z_{\alpha} \times \sqrt{\frac{p(1-p)}{n_0}}$$

Three main factors determine the magnitude of the confidence interval. First, the higher is the sample size  $n_0$ , the lower is the margin of error  $e$ . At first glance, one should then try to maximize the sample size. However, since the margin of error decreases with the square root of the sample size, there is a kind of diminishing returns to increasing sample size. Concurrently, the cost of survey administration is likely to increase linearly with  $n_0$ . There is consequently a balance to find between those opposing effects. Second, a sample should be as representative as possible of the population. If the population is highly heterogeneous, the possibility of drawing a non-representative sample is actually high. In contrast, if all members are identical, then the sample characteristics will perfectly match the population, whatever the selected sample is. Imagine for instance that  $\pi = 90\%$ , i.e. most individuals in the population are males. In that case, if the sample is randomly chosen, the likelihood of selecting a non-representative sample (e.g., only females) is low. On the contrary, if the gender attribute is equally distributed ( $\pi = 50\%$ ), then this likelihood is high. Since the population variance  $\pi(1-\pi)$  is unknown, the sample variance  $p(1-p)$  will serve as a proxy for measuring the heterogeneity in the

population. The higher is  $p(1 - p)$ , the lower is the precision of the sample estimate. Third, the  $z_\alpha$  statistic allows to compute a margin of error with a  $(1 - \alpha)$  confidence level, which corresponds to the probability that the confidence interval calculated from the sample encompasses the true value of the population parameter. The sampling distribution of  $p$  is approximately normally distributed if the population size is sufficiently large. The usually accepted risk is  $\alpha = 5\%$  so that the confidence level is 95%. The critical value  $z_{5\%} = 1.96$  is computed with a normal distribution calculator.

Let us now consider the formula for the margin of error from a different perspective. Suppose that instead of computing  $e$ , we would like to determine the sample size  $n_0$  that achieves a given level of precision, hence keeping the margin of error at the given level  $e$ . The equation can be rewritten:

$$n_0 = z_{5\%}^2 \times \frac{p(1 - p)}{e^2}$$

Table 2.1 highlights the relationship between the parameters. For instance, when the proportion  $p$  is 10% and the margin of error  $e$  is set to 5%, the required sample size is  $n_0 = 138$ . If we want to reach a higher precision, say  $e = 1\%$ , then we have to survey a substantially higher number of units:  $n_0 = 3457$ . Of course, the value of  $p$  is unknown before the survey has been implemented. Yet, the maximum of the sample variance  $p(1 - p)$  is obtained for  $p = 50\%$ . For that value of the proportion, and in order to achieve a level of precision  $e = 1\%$ , one should survey at least  $n_0 = 9604$  units, and  $n_0 = 384$  to achieve  $e = 5\%$ .

The sample size also depends on the size of the target population, denoted  $N$  hereafter. Below approximately  $N = 200,000$ , a finite population correction factor has to be used:

**Table 2.1** Sample size for an estimated proportion

Proportion		Margin of error			
p (%)	1-p (%)	0.5%	1%	5%	10%
10	90	13,830	3457	138	35
20	80	24,586	6147	246	61
30	70	32,269	8067	323	81
40	60	36,879	9220	369	92
50	50	38,416	9604	384	96
60	40	36,879	9220	369	92
70	30	32,269	8067	323	81
80	20	24,586	6147	246	61
90	10	13,830	3457	138	35

$$e = z_{5\%} \times \sqrt{\frac{p(1-p)}{n}} \times \sqrt{\frac{N-n}{N-1}}$$

Solving for  $n$  yields:

$$\begin{aligned} n \times \frac{N-1}{N-n} &= z_{5\%}^2 \times \frac{p(1-p)}{e^2} \\ n \times \frac{N-1}{N-n} &= n_0, \\ n &= \frac{n_0 N}{n_0 + N - 1}. \end{aligned}$$

For instance, while we were previously suggesting a sample size of  $n_0 = 384$  to ensure a margin of error of 5%, now, with the new formula, and if the population size is  $N = 500$ , we have:

$$n = \frac{384 \times 500}{384 + 500 - 1} \approx 217$$

Table 2.2 provides an overview of the problem. Those figures provide a useful rule of thumb for the analyst. For a desired level of precision  $e$ , the lower is the population size  $N$ , the lower is the number  $n$  of units to survey. Those results, however, have to be taken with caution. What matters at the end is common sense. For instance, according to Table 2.2, if  $N = 1000$  the analyst should survey  $n = 906$  units to ensure a margin of error of 1%. In that case, sampling would virtually be equivalent to a census, in statistical terms but also in budget and organizational terms. Moving to a less stringent 5% margin of error would provide a much more relevant and tractable number of units to survey.

In practice, most polling companies survey from 400 to 1000 units. For instance, the NBC News/Wall Street Journal conducted in October 2015 a public opinion poll

**Table 2.2** Target population and sample size

Population	Margin of error			
N	0.5%	1%	5%	10%
50	50	50	44	33
100	100	99	80	49
500	494	475	217	81
1000	975	906	278	88
2000	1901	1655	322	92
5000	4424	3288	357	94
10,000	7935	4899	370	95
100,000	27,754	8763	383	96

relating to the 2016 United States presidential election (a poll is a type of sample survey dealing mainly with issues of public opinions or elections). A number of 1000 sampling units were interviewed by phone. Most community satisfaction surveys rely on similar sample sizes. For instance, in 2011, the city of Sydney, Australia, focused on a series of  $n = 1000$  telephone interviews to obtain a satisfaction score related to community services and facilities. Smaller cities may instead focus on  $n = 400$  units. At a national level, sample sizes reach much larger values. To illustrate it, in 2014, the American Community Survey selected a sample of about 207,000 units from an initial frame of 3.5 million addresses. According to our rule of thumb, this would yield a rather high precision, approximately  $e = 0.2\%$ .

The choice of sample size also depends on the expected in-scope proportion and response rate. First, it is possible that despite all efforts coverage errors exist and that a number of surveyed units do not belong to the target population. On top of these considerations, the survey may fail to reach some sampling units (refusals, noncontacts). To guarantee the desired level of precision, one needs therefore to select a sample larger than predicted by the theory, using information about the expected in-scope and response rates. More specifically, the following adjustment can be implemented:

$$\text{Adjusted sample size} = \frac{n}{\text{Expected response rate} \times \text{Expected in-scope rate}}$$

Suppose for instance that the in-scope rate estimated from similar surveys or pilot tests is 91%. Assume also that the expected response rate is 79%. When  $n = 1000$ , the adjusted sample size is:

$$\text{Adjusted sample size} = \frac{1000}{0.91 \times 0.79} = 1391$$

A crucial issue here is that once the expected in-scope and response rates have been defined ex ante, their values should serve as a target during the data collection process. A response rate or in-scope rate lower than the desired values will result in a sample size that does not ensure anymore the precision requirement. For instance, in the case of the American Community Survey, if we fictitiously assume an ex-post response rate of 25% and in-scope rate of 85%, which can be realistic in some cases (if not in this particular one), then the margin of error increases from  $e = 0.2\%$  to  $0.5\%$ .

To conclude, whether one chooses a higher or lower sample size (or equivalently, a higher or lower precision) mainly depends on operational constraints such as the budget, but also the time available to conduct the entire survey and the size of the target population. First, there are direct advantages and disadvantages to using a census to study a population. On the one hand, a census provides a true measure of the population but also detailed information about sub-groups within the population, which can be useful if heterogeneity matters. On the other hand, a sample generates lower costs both in staff and monetary terms

and is easier to control and monitor. Second, the time needed to collect and process the data increases with the sample size. Thus, with a sample survey of realistic size, the results are generally available in less time and can still be representative of the population. Third, the population size is also a determinant factor. If the population is small, a census is always preferable. In contrast, for large populations, accurate results can be obtained from reasonably small samples. In any case, the next step now consists in conceiving the questionnaire that will be proposed to respondents.

---

### 2.3 Conception of the Questionnaire

A questionnaire is a set of questions designed to elicit information upon a subject, or sequence of subjects, from a respondent. Given its impact on data quality, the questionnaire design plays a central role. The purpose of a survey is to obtain sincere responses from the respondent. One main principle applies in this matter: one should start on the basis that most people do not want to spend time on a survey, and if they do, it could be that they actually are not satisfied with the policy under evaluation, which may be non-representative of the population as a whole. Nonresponses should be minimized as much as possible. This can be done by explaining why the survey is carried out, by keeping it quick and by telling the respondents that the results will be communicated once finalized. Those three rules are even truer nowadays since people are frequently required to participate in surveys in many fields.

An important aspect of questionnaire design is the type of response formats. There are two categories of questions: open-ended versus close-ended. Close-ended questions request the respondent to choose one or several responses among a predetermined set of options. While they limit the range of respondents' answers on the one hand, they require less time and effort for both the interviewer and the participant on the other hand. In contrast, open-ended questions do not give respondents options to choose from. Thereby, they allow them to use their own words and to include more information, including their feelings and understanding of the problem.

Examples of close-ended and open-ended questions are provided in Fig. 2.2. Dichotomous questions (also referred to as two-choice questions) are the simplest version of a close-ended question. They propose only two alternatives to the respondent. Multiple choice questions propose strictly more than two alternatives and ask the respondent to select one single response from the list of possible choices. Checklist questions (or check-all questions) allow a respondent to choose more than one of the alternatives provided. Forced choice questions are similar to checklist questions, although the respondent is required to provide an answer (e.g., yes–no) for every response option individually. Partially closed questions provide an alternative “Other, please specify”, followed by an appropriately sized answer box. This type of question is useful when it is difficult to list all possible alternatives or when responses cannot be fully anticipated. Last, open-ended questions can be of two forms, either text or numerical.



**a**

**Dichotomous question**

Q1. In 2004, did you or did anyone in your household make a call requesting emergency assistance from the Police Department?

Yes	No
<input type="checkbox"/>	<input type="checkbox"/>

(Source: Santa Monica resident survey, 2005)

**Multiple choice question**

Q2. How many years have you lived in Novato?

Less than 2 years	2-5 years	6-10 years	11-20 years	More than 20 years
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

(Source: city of Novato citizen survey, 2013)

Q3. Do you live in a unit, house, townhouse or semi?

Unit	House	Townhouse or semi
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

(Source: Sydney community satisfaction survey, 2012)

**Checklist question**

Q4. Which, if any, of these events did you or a member of your household attend?

Keeping Tradition Alive Jam session	Red, White & Lewisville fireworks	Sounds of Lewisville summer concerts	Western Days	Holiday Stroll
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

(Source: City of Lewisville resident satisfaction survey, 2014)

**Forced choice question**

Q5. Do you receive any of the following benefits?

	Yes	No
Sickness benefit (are on sick leave)	<input type="checkbox"/>	<input type="checkbox"/>
Old age pension, early retirement (AFP) or survivor pension	<input type="checkbox"/>	<input type="checkbox"/>
Rehabilitation/reintegration benefit	<input type="checkbox"/>	<input type="checkbox"/>
Disability pension (full or partial)	<input type="checkbox"/>	<input type="checkbox"/>
Unemployment benefits during unemployment	<input type="checkbox"/>	<input type="checkbox"/>
Social welfare benefits	<input type="checkbox"/>	<input type="checkbox"/>
Transition benefit for single parents	<input type="checkbox"/>	<input type="checkbox"/>

(Source: Tromsø Health Survey, 2001)

**Partially closed question**

Q6. How did you contact the City of Sydney?

Telephone	In person	In writing	Email/Website	Fax	OTHER
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	.....

(Source: Sydney community satisfaction survey, 2012)

**b**

**Text question**

Q7. Now, what would you say are the one or two most important issues facing the City of Santa Monica today?

.....

.....

(Source: Santa Monica resident survey, 2005)

**Numerical question**

Q8. Last month, what was the cost of gas for this house, apartment, or mobile home? \$

.....

(Source: American community survey, 2015)

**Fig. 2.2** Close-ended and open-ended questions. (a) close-ended questions and (b) open-ended questions

Another widely used format is the scale question, which asks the respondent to grade the response on a given range of options (see Fig. 2.3). These questions can be grouped into two subcategories: ranking questions and rating questions. Ranking questions offer several options and request the respondent to rank them from most important to least important on a ranking scale (where 1 is the most important, 2 is the second most important, and so on) or a bipolar scale (where respondents have to rate the intensity of their preference). Respondents thus compare each item to each other. A ranking scale has the inconvenient to force the respondent to make one item worse or better than another, when they actually could be indifferent between them. They also require a significant cognitive effort. Pairwise comparisons overcome these problems through the use of bipolar scales. When the number of

**a**

**6-point ranking scale**

Q9. Please rank the following issues in order of their importance to you. 1 stands for the most important and 6 for the least important.

International tensions (terrorism, war)	.....
Economic concerns (unemployment, inflation)	.....
Environmental concerns (waste, air pollution)	.....
Health concerns (Bird flu, AIDS)	.....
Social issues (poverty, discrimination)	.....
Personal safety (crime, theft...)	.....

(Source: OECD Studies on environmental policy and household behavior, 2011)

**10-point bipolar scale**

Q10. Which of the policy options described below would you be most in favour of? Please indicate your preferences using the scale below.

[description of policy options]											
Strongly prefer Choice A	1	2	3	4	5	6	7	8	9	10	Strongly prefer Choice B
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	

(Source: Economic valuation with stated preference techniques, Bateman et al., 2002)

**b**

**4-point rating scale**

Q11. Generally speaking, are you satisfied or dissatisfied with the job the City of Thousand Oaks is doing to provide city services?

Very satisfied	Somewhat satisfied	Somewhat dissatisfied	Very dissatisfied
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

(Source: Thousand Oaks community satisfaction survey, 2015)

**5-point rating scale**

Q12. To what extent do you agree or disagree that the City of Miami Beach government is open and interested in hearing the concerns or issues of residents? Would you say...?

Strongly agree	Somewhat agree	Neither agree nor disagree	Somewhat disagree	Strongly disagree
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

(Source: Miami Beach resident satisfaction survey, 2007)

**10-point rating scale**

Q13. On a scale from 1 to 10 can you indicate how satisfied you are with the life you lead at the moment? A score of 1 refers to completely dissatisfied and a 10 to completely satisfied.

1	2	3	4	5	6	7	8	9	10
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

(Source: Measuring well-being, Statistics Netherlands, 2014)

**5-point semantic differential scale**

Q14. How would you describe the quality of your community's water as it relates to its effect on household piping, fixtures, and water-using appliances? Please, place a check on one of the five lines of the scale for each effect.

Example					
Rusty	___	___	___	✓	___
Corrosive	___	___	___	___	___
Leaves No Scale	___	___	___	___	___
Stains Fixtures	___	___	___	___	___

(Source: Colorado water resources research institute, 1993)

**Fig. 2.3** Scale questions. (a) ranking questions and (b) rating questions

alternatives is large, it is possible to ask the respondents to choose one single item through a multiple choice question, for instance:

*Let's assume for a moment that the Santa Monica Police Department hired another officer and assigned that officer to your neighborhood. Which of the following five items should be the single highest priority for a new police officer assigned to your neighborhood?*

1. Working with local kids to prevent gangs and youth crime,
2. Patrolling on foot in your local neighborhood,
3. Working with local residents and neighborhood groups to help prevent crime,
4. Patrolling in police cars in your local neighborhood,

### 5. *Patrolling near the schools in your neighborhood.*

(Source: Santa Monica Resident Survey, 2005)

The other category of scale question is the rating question, which requires the respondents to rate their answer, independently of other options, on a rating scale (also referred to as a Likert scale). Usually, this type of scale contains equal numbers of positive and negative positions, which creates a less biased measurement. Often, it is preferable not to propose a neutral position in the middle, as otherwise the respondents could choose this category to save time or hide their preference. Last, semantic differential scales ask the respondents to choose between two opposite positions, with bipolar adjectives at each end. Such a scale allows to include several dimensions in a single question, but also demands higher cognitive effort from the respondent.

The sequencing of questions is as important as the questions themselves. It should be designed to encourage the respondent to complete and maintain interest in the questionnaire. It is usually advised to follow the following sequence. First, an introductory section should give the title of the survey and introduce the authority under which the survey is conducted, its purpose, and the general contents of the questionnaire. What is included is crucial in securing the participation of respondents. This section usually contains general instructions for the interviewer and respondents, provides reassurances about confidentiality and states the expected length of the survey. It requests the respondent's cooperation and stresses the importance of his/her participation. It explains how the survey data will be used and includes contact information. Finally, this section may include the signature of the person in charge of the authority under which the survey is conducted.

The sequence of questions should be as logical as possible. For instance, the first questions should be easy to answer. Sensitive questions should not be placed at the beginning of the questionnaire, but introduced at a point where the respondent is more likely to feel comfortable answering them. The first questions are generally about things respondents do or have experienced, the so-called behavior questions. Knowledge questions can be included to better assess whether the respondent knows the topic. Those types of question are then followed by opinion questions, which ask what the respondents think about a specific item. Motive questions require the respondents to evaluate why they behave in a particular manner. Personal and confidential questions as well as questions about socio-economic status are located at the end of the questionnaire. One should not forget to include an open-ended question at the end, so that the respondents have the possibility to express themselves, as well as an acknowledgement to thank the respondent.

Between each part of the questionnaire it is important to use transitional statements to explain that a new topic will be examined. In addition, several rules have to be obeyed with respect to question writing. Spelling, style and grammar should be carefully checked, otherwise it would devalue the organization that implements or orders the study. It is also recommended to minimize the length of the questionnaire. The greater is the number of questions, the less time the respondents spend, on average, answering each question. There is a point at

which survey completion rates start to drop off, usually after 5–8 min (i.e. 15–20 questions—one web page—one sheet of paper). Do not ask open-ended questions unless necessary. Use the same scales over the questionnaire. Regroup similar questions as follows:

*Now, please rate each of the following possible problems in Santa Monica on a scale of 1 to 5. Use a 1 if you feel the problem is NOT serious at all, and a 5 if you feel it is a VERY serious problem in Santa Monica:*

- |  |  |
|--|--|
| 1. <i>Traffic congestion</i>   | <input type="checkbox"/> 1 <input type="checkbox"/> 2 <input type="checkbox"/> 3 <input type="checkbox"/> 4 <input type="checkbox"/> 5 |
| 2. <i>The affordability of housing for low income families and seniors</i> | <input type="checkbox"/> 1 <input type="checkbox"/> 2 <input type="checkbox"/> 3 <input type="checkbox"/> 4 <input type="checkbox"/> 5 |
| 3. <i>Gang violence</i>  | <input type="checkbox"/> 1 <input type="checkbox"/> 2 <input type="checkbox"/> 3 <input type="checkbox"/> 4 <input type="checkbox"/> 5 |
| 4. <i>The number of homeless people in the city</i>                        | <input type="checkbox"/> 1 <input type="checkbox"/> 2 <input type="checkbox"/> 3 <input type="checkbox"/> 4 <input type="checkbox"/> 5 |
| 5. <i>Lack of parking</i>  | <input type="checkbox"/> 1 <input type="checkbox"/> 2 <input type="checkbox"/> 3 <input type="checkbox"/> 4 <input type="checkbox"/> 5 |

(Source: 2005 Santa Monica resident survey)

Another point is to define and choose carefully the time horizon. For instance, depending on the context, the question “How many times per year do you take the bus” may not be enough specific and “per year” should be replaced by “per week”. Avoid using terms such as “regularly” or “often”, which do not convey the same meaning for all respondents. Instead, an appropriate time horizon should be offered, e.g.:

*How often do you suffer from headaches?*

1. *Rarely or never*
2. *Once or more a month*
3. *Once or more a week*
4. *Daily*

(Source: 2001 Tromsø Health Survey)

Perhaps it is obvious, but simple and clear questions are better than long questions, with complex words, abbreviations, acronyms, or sentences that are difficult immediately to understand. Define the technical terms if necessary. Do not ask negatively worded questions like “Should the City not invest in energy efficiency for municipal buildings?” Avoid double-barreled questions that ask two or more questions in a row. Do not use confusing terms or vague concepts. For instance, when asked “how much do you pay per year in taxes?” respondents may not know what is meant by “taxes”, whether it is income taxes, property taxes, national or local taxes. Finally, there is always the risk of a framing effect when phrasing a question. For instance, questions like “Don’t you think that the city needs to cut the grass around our schools?” may induce yea-saying bias. Prefer instead a question like “to what extent do you agree or disagree that. . .”. Such a question should also specify the cost and/or additional increase in taxes. Check also

that the response options do not force people to answer in a way that they do not wish to. Questions must propose all the relevant options. One should open the question with an item “other”, if one is not sure about the exhaustiveness of the options. Last, each item should be totally independent from the others.

Not only respondents but also public decision-makers or experts in the field should be consulted to provide insight into the type of information that is required. Meetings and focus-groups can help identify issues and concerns that are important. Whether it is a new questionnaire or a set of questions that have been used before, it is also essential to test it before the survey is implemented. This stage represents an opportunity to check whether the interviewers and respondents understand the questions, whether the survey retains the attention of respondents and whether it is sufficiently short. In a first step, an informal pilot test can be implemented using a number of colleagues. While they may be familiar with the questionnaire and will tend to answer the questions more quickly, they will also be more likely to pick up errors than the respondents themselves. The next stage for the questionnaire writer is to implement a larger scale pilot test on a subsample of the target population, but also on specific subgroups of the population that may have difficulties with particular questions. A pilot test of 30–100 cases is usually sufficient to discover the major problems in a questionnaire. The questionnaire should be administered in the same manner as planned for the main survey. A minimum of 30 observations also yields the possibility for the questionnaire writer to implement a preliminary statistical analysis, in order to assess whether the survey is suitable to achieve the objectives of the study.

---

## 2.4 Data Collection

Data collection is any process whose purpose is to acquire information. When it has been decided that a census is not preferable over a sample survey, the first stage consists in selecting a subset of units from the population. There are two kinds of methods in this respect: non-probability and probability sampling. Whether one chooses the first or the second mainly depends on the availability of a survey frame, i.e. a list of each unit in the population. If a survey frame is not available, then one can implement a probability sampling, i.e. select randomly a sample from that list. By definition, probability sampling is a procedure in which each unit of the population has a fixed probability of being selected from the sample. Reliable inferences can then be made about the population. If a survey frame is not available, then one has to rely instead on subjective and personal judgment in the process of sample selection, i.e. on non-probability sampling. The procedure is usually simpler and cheaper to implement, but also more likely to be subject to bias. Hence, whether one chose an approach or another depends on the availability of a survey frame and how one values the sampling error against the cost of survey administration.

Common methods of probability sampling are simple random sampling, systematic sampling, stratified sampling and cluster sampling. We shall consider them successively. With simple random sampling, each unit is chosen randomly using a

random number table or a computer-generated random number. Such sampling is done without replacement, i.e. the procedure should avoid choosing any unit more than once. Systematic sampling is a method that selects units at regular intervals. In a first step, all units in the survey frame are numbered from 1 to  $N$ . Second, a periodic interval  $k = N/n$  is calculated, where  $n$  represents the desired sample size. Third, a starting point is randomly selected between 1 and  $k$ . Fourth, every  $k$ th unit after the random starting point is selected. For instance, assume that the survey frame contains  $N = 10,000$  units and that we would like to sample  $n = 400$  units. The sampling interval is  $k = N/n = 25$ . Then a random number between 1 and 25, say 12, is selected. The units that are selected are 12,  $12 + 25 = 37$ ,  $37 + 25 = 62$ , etc. Stratified sampling is a method by far superior to simple random and systematic sampling because it may significantly improve sampling precision and reduce the costs of the survey. It is used when the survey frame can be divided into non-overlapping subgroups, called strata, according to some variable whose information is available *ex ante* (e.g., males/females, age categories, income categories). The approach consists in drawing a separate random sample from each stratum and then to combine the results. Specifically, the population  $N$  is divided into  $m$  groups with  $N_i$  units in group  $i$ ,  $i = 1, \dots, m$ . If the desired sample size is  $n$  and for a proportional ( $N_i/N$ ) allocation of units between groups, one should then survey  $nN_i/N$  units in each group  $i$ . Systematic or simple random sampling is then used to select a sufficient number of units from each stratum. Finally, cluster sampling randomly selects subgroups of the population. In contrast with stratified sampling, the subgroups are not based on the population attributes, but rather on independent subdivisions, or clusters, such as geographical areas, districts, factories, schools. Clusters  $i$ ,  $i = 1, \dots, M$  of size  $N_i$  must be mutually exclusive and together they must encompass the entire population:  $\sum_{i=1}^M N_i = N$ . The first step amounts to drawing randomly  $m$  clusters amongst the  $M$ . Then two possibilities arise. Either one surveys all units in each selected cluster, in which case the method is referred to as “one-stage cluster sampling”, or one selects a random sample from each cluster, which is the “two-stage cluster sampling”. One advantage of the procedure is that it may significantly reduce the cost of collection for instance if personal interviews are conducted and the geographical zones are spread out. One difficulty, however, is that the selected clusters may be non-representative of the population.

Methods of non-probability sampling encompass convenience sampling, judgment sampling, volunteer sampling, quota sampling, and snowball sampling. Convenience sampling, also referred to as haphazard sampling, is the most common approach. As can be deduced from the name, it consists in selecting a sample because it is convenient to do so. Typical examples include surveying people in a street, at a subway stop, at a crowded place. The approach is based on the assumption that the population is equally distributed from one geographical zone to the other. If not, then some bias may occur. Judgment sampling selects the sample based on what is thought to be a representative sample. For instance, one may decide to draw the entire sample from one “typical” city or “representative”

street. The approach may result in several biases, and is generally used for exploratory studies only. Volunteer sampling selects the respondent on the basis of their willingness to participate voluntarily in the survey. Here again, the approach is subject to many bias. In particular, self-selection may produce a sample of highly motivated (pro or against the project) individuals and neglect average or less contrasting views. It is however often used when one needs to survey people with a particular disease or health condition. Quota sampling is usually said to be the non-probability equivalent of stratified sampling. In both cases, one has to identify representative strata that characterize the population. Information about the true population attributes (available from other sources such as a national census) can be used to guarantee that each subgroup is proportionally represented. Then convenience, volunteer, or judgment sampling is used to select the required number of units from each stratum. The procedure may save a lot of time as one would typically stop to survey people with a particular characteristic once the quota has been reached. For instance, assume one would like to survey  $n = 400$  units. If we have an equal share of males and females in the population, one should survey only 200 males and 200 females. Last, snowball sampling is recommended when one needs to survey people with a particular but not frequent characteristic. The approach identifies initial respondents who are then used to refer on to other individuals. Again, it may generate several biases. It is generally used when one wants to survey hard-to-reach units at a minimum cost, such as the deprived, the socially stigmatized, or the users of a specific public service.

Once the sampling procedure has been selected, one has to start the collection of data. The basic methods are self-enumeration, telephone interview, and personal interview. The characteristics of the target population and whether a frame is easily available strongly influence the choice of the method, which can be paper or computer based. Self-enumeration requires the respondents to answer the questionnaire without the assistance of an interviewer. This method of data collection is easy to administer and is typically suited to large samples or when some questions are highly personal or sensitive and easier to complete in private. Respondents should be sufficiently motivated and educated, so that they do not skip or misinterpret information. The response rate can be very low, and one may have to contact several times the respondents to remind them to complete the questionnaire. Personal interview requires the respondents to answer the questionnaire with the assistance of an interviewer, at home, at work, or at a public place. The method yields high response rates but it can however be expensive and thereby more suited to smaller sample sizes. Another issue is that the interviewers may have to reschedule the interview until the respondent is present or has time. Last, telephone interviews offer good response rates at reasonable costs since the interviewers do not need to travel, and the interview can be rescheduled more easily than with personal interviews. It is also easier to control the quality of the interviewing process if it is recorded.

The type of questions may strongly influence the choice of the collection method. If complex questions are asked, then personal or telephone interviews are preferable. In contrast, if questions concern highly personal or sensitive issues,

self-enumeration is preferable. The nature of the sample units is also important. For instance, if people need assistance (e.g., children or distressed people), personal interviews are more relevant. For example, in the case of child’s health condition, the sample unit can be the child’s family. Within this sample unit, one may distinguish between the unit of reference (the child who provides the information) and the reporting unit (one of the parents carrying out the information).

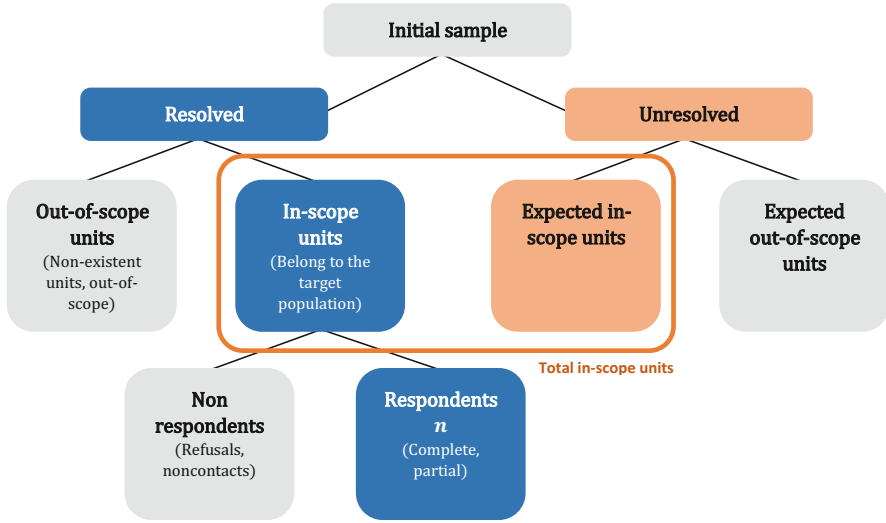
When personal or telephone interviews are chosen as methods for data collection, it is important to prepare the interviewers. They should be informed that the questionnaire has been carefully prepared to minimize potential biases, that they should not improvise, nor influence the respondents. Every question should be asked, in the order presented, exactly as worded. They should be provided with a manual that contains guidelines. These guidelines should also contain answers to the most common questions that respondents may ask, as displayed in Table 2.3. Interviewers must be honest about the length of the interview. Questions that are misunderstood or misinterpreted should be repeated. Personal interviewers should have official badges or documents in case a respondent ask them to prove they are a legitimate representative of the public sector. Last, if a person still refuses to answer the questionnaire, it should be recorded as “refusal”.

It is important to assess the performance of data collection during the survey process itself. In this matter, many rates can be computed. Figure 2.4 provides an

**Table 2.3** Examples of questions and answers during interviews

Usual question of the respondent	Standard answer by the interviewer
Why did you pick me?	<i>By selecting a few people like you, we are able to reduce the costs associated with collecting information because we do not have to get responses from everybody. On average, the data collected will be representative of the population because respondents have been selected randomly.</i>
Who is going to see my data?	<i>All information collected is highly confidential and will be seen only by the survey staff. Your answers will be used only for the production of anonymous statistics.</i>
Why should I participate? How will you use my answers?	<i>The purpose of this survey is to find out your views on _____. Your input in this study will provide useful information and help improving public services.</i>
I do not have the time right now	<i>The questionnaire consists of ____ short questions and will not take more than ____ minutes of your time. Your responses are very important. If you are very busy now, please tell me when I can reach you again.</i>
I do not see how I can help you; I really don’t know the topic.	<i>We are interested in your opinions and experiences, not in what information you may or may not have. In a study of this type, there are no right or wrong answers to questions.</i>
Who is behind this?	<i>This study is supervised by _____. The purpose is to collect information that will be helpful in improving public services.</i>





**Fig. 2.4** From the initial sample to the responding units

illustration of these concepts. A first rate is based on the proportion of resolved records:

$$\text{Resolved rate} = \frac{\text{Number of resolved units}}{\text{Initial sample}}$$

This rate is defined as the ratio of the number of resolved units to the total number of sampling units. A unit is categorized as resolved if it has a determinate status, i.e. if the unit is either in-scope (complete, partial, refusal, noncontact) or out-of-scope.

A crucial issue is that some units may not belong to the target population so that they are out-of-scope. The following indicator estimates the extent of the phenomenon:

$$\text{In-scope rate} = \frac{\text{Number of in-scope units}}{\text{Number of resolved units}}$$

Using this proportion, it is also possible to approximate the expected number of in-scope units among the resolved and unresolved units:

$$\text{Expected number of in-scope units} = \text{In-scope rate} \times \text{Initial sample}$$

The assumption underlying this expectation is that the in-scope rate can be extrapolated to the whole sample.

Another indicator of interest is the response rate, namely the number of respondents (either complete or partial response) divided by the total number of sample units that are in-scope (resolved and unresolved) for the survey. Since the latter is unknown during the collection process, the previous formula is used for the denominator:

$$\text{Response rate} = \frac{\text{number of responding units}}{\text{expected number of in-scope units}}$$

Once the data has been collected, it is common to provide the following information at the beginning of a survey study: (1) the sampling design and data collection method, (2) the number of sampling units, (3) the number of in-scope units, (4) the number of responding units, and (5) the margin of error, as illustrated in Fig. 2.5.

In Fig. 2.5, a sample of 1000 units has been gathered via stratified sampling and computer-assisted personal interviewing. Assume that after one week of data collection, we have 600 resolved units among which 300 units are in-scope. This yields a resolved rate of  $600/1000 = 60\%$  and an in-scope rate equal to  $300/600 = 50\%$ . The expected total number of in-scope units is thus  $1000 \times 50\% = 500$ . Suppose now that among the 300 units that are in scope, 200 units responded to the survey (either complete or partial response). Then the response rate is  $200/500 = 40\%$ . Now imagine that survey completion occurs after 3 weeks. This means that one finally gets 1000 resolved units. Among these units, suppose that 700 units are in-scope and that 500 units responded to the survey. If the target population size is  $N = 10,000$ , the margin of error can be obtained using the formula described in Sect. 2.2:

$$e = 1.96 \times \sqrt{\frac{50\%(1 - 50\%)}{500}} \times \sqrt{\frac{10000 - 500}{10000 - 1}}$$

This yields a margin of error of approximately 4.27%.

Name of the organization: _____	Date : ____/____/____
Sampling design: stratified sampling	
Data collection method: computer-assisted personal interviewing	
Number of sampling units: 1,000	
Number of in scope units: 700	
Number of responding units: 500	
Margin of error: +/- 4.27%	

Fig. 2.5 Typical header for a survey study

## 2.5 Coding of Variables

Coding is the process of converting textual information into numbers or other symbols that can be counted and tabulated. This step is essential as it determines the final variables that will be used for subsequent analyses. To better implement it, one should understand what a database is. In statistics, it is a computer file (e.g., Excel file) made of rows  $i$  and columns  $j$ , where rows stand for the responding units, and columns for the variables. This framework is illustrated in Table 2.4 where each  $x_{ij}$  represents the value assigned by respondent  $i$  to variable  $j$ .

In this section, we propose to explain how a database is coded using the questions from Figs. 2.2 and 2.3. Table 2.5 shows what the final database looks like for a selected set of questions. For closed questions, codes are generally established before the survey takes place. The categories may be split into one or several variables depending on the nature of the questions.

For dichotomous questions, such as question Q1 (“*In 2004, did you or did anyone in your household make a call requesting emergency assistance from the Police Department?*”), the coding involves the creation of one single variable:

$$Q1 = \begin{cases} 1 & \text{if "yes"} \\ 0 & \text{if "no"} \end{cases}$$

Since their values belong to the set  $\{0,1\}$ , such variables are also called binary variables.

For multiple choice questions, two cases arise depending on whether there is a clear ordering of the variables. First, when the options of answer can be ordered, one can build a unique variable using a scale relevant to the investigated topic. For instance, question Q2 (“*How many years have you lived in Novato?*”) can be coded as:

**Table 2.4** A usual database format

Responding unit/individual	Variable 1	Variable 2	...	Variable $j$	...
1	$x_{11}$	$x_{12}$	...	$x_{1j}$	...
2	$x_{21}$	$x_{22}$	...	$x_{2j}$	...
3	$x_{31}$	$x_{32}$	...	$x_{3j}$	...
...	...	...	...	...	...
$i$	$x_{i1}$	$x_{i2}$		$x_{ij}$	...
...	...	...	...	...	...
$n-2$	$x_{n-2,1}$	$x_{n-2,2}$	...	$x_{n-2,j}$	...
$n-1$	$x_{n-1,1}$	$x_{n-1,2}$	...	$x_{n-1,j}$	...
$n$	$x_{n,1}$	$x_{n,2}$	...	$x_{n,j}$	...

**Table 2.5** Examples of coding

(a) Questions Q1–Q7									
Ind	Q1	Q2	Q3 <i>unit</i>	Q3 <i>house</i>	Q4 <i>jam</i>	Q4 <i>fireworks</i>	Q6 <i>centre</i>	Q7 <i>school</i>	Q7 <i>unemp</i>
1	1	4	0	1	0	0	0	0	0
2	0	3	0	0	0	0	0	0	1
3	1	1	1	0	1	0	0	0	0
4	1	3	1	0	1	1	1	1	1
...	0	5	0	0	1	1	0	1	0

(b) Questions Q8–Q14						
Ind	Q8	Q9 <i>International tensions</i>	Q9 <i>Economic concerns</i>	Q11	Q12	Q14 <i>Corrosive</i>
1	30	5	1	4	2	5
2	63	4	3	1	3	4
3	140	3	4	3	5	1
4	37	6	5	2	4	3
...	0	2	3	3	1	2

$$Q2 = \begin{cases} 1 & \text{if "Less than 2 years"} \\ 2 & \text{if "2 – 5 years"} \\ 3 & \text{if "6 – 10 years"} \\ 4 & \text{if "11 – 20 years"} \\ 5 & \text{if "More than 20 years"} \end{cases}$$

Second, when there is no intrinsic ordering to the options, one has to split them into separate variables. This is the case for instance with question Q3 (“Do you live in a unit, house, townhouse or semi?”):

$$Q3|_{unit} = \begin{cases} 1 & \text{if "yes"} \\ 0 & \text{if "no"} \end{cases}, Q3|_{house} = \begin{cases} 1 & \text{if "yes"} \\ 0 & \text{if "no"} \end{cases}, \dots$$

In a similar manner, checklist questions, whether they are forced or not, imply a transformation of each category into one specific variable. For instance, with question Q4 (“Which, if any, of these events did you or a member of your household attend?”), we have:

$$Q4|_{jam} = \begin{cases} 1 & \text{if "yes"} \\ 0 & \text{if "no"} \end{cases}, Q4|_{fireworks} = \begin{cases} 1 & \text{if "yes"} \\ 0 & \text{if "no"} \end{cases}, \dots$$

Note that forced choice questions like Q5 (“Do you receive any of the following benefits?”) can also be treated as a series of dichotomous questions, with for instance:

$$Q5|sickness\ benefit = \begin{cases} 1 & \text{if "yes"} \\ 0 & \text{if "no"} \end{cases}$$

For text and partially closed questions, coding is more difficult as it requires interpretation and personal judgment to recode them into close-ended questions. The survey staff has to detect whether some items appear frequently. If it seems to be the case, then different categories should be created to take into account the heterogeneity in the responses. Afterwards, the coding can be done either manually or by computer. For instance, to question Q6 (*“How did you contact the City of Sydney?”*), the option *“other”* may have been selected frequently by people who wrote down *“response centre”*, which does not belong to the initial set of categories. If the number of occurrences is large enough, this item should be included as a new item among the categories of question Q6:

$$Q6|response\ centre = \begin{cases} 1 & \text{if "yes"} \\ 0 & \text{if "no"} \end{cases}$$

Similarly, to question Q7 (*“Now, what would you say are the one or two most important issues facing the City of Santa Monica today?”*), one has to detect first which items have been frequently raised (e.g., quality of schools, unemployment rate, environmental concerns, crime rate), and create a category for each of them:

$$Q7|schools = \begin{cases} 1 & \text{if "yes"} \\ 0 & \text{if "no"} \end{cases}, \quad Q7|unemployment = \begin{cases} 1 & \text{if "yes"} \\ 0 & \text{if "no"} \end{cases}, \dots$$

Coding is much simpler when it comes to numerical questions. We may directly use the question as it is. To illustrate, for question Q8 (*“Last month, what was the cost of gas for this house, apartment, or mobile home?”*), we have:

$$Q8 = declared\ cost\ of\ gas$$

For ranking order questions, the data set should include a column for each item being ranked. For instance, for question Q9 from Fig. 2.3 (*“Please rank the following issues in order of their importance to you.”*) we have:

$$Q9|international\ tensions = score\ obtained$$

$$Q9|Economic\ concerns = score\ obtained$$

...

For any given respondent, each ranked item has a unique value, and once an item has reached a score, that score cannot be employed anymore. Notice also that this type of question may return different results depending on the completeness and relevance of the list of items being ranked. Thus, these scores should be analyzed with caution.

Instead of using a ranking scale, one may prefer to use a bipolar scale or a rating scale. With a bipolar scale, and any rating question, it is common to code the values using 1, 2, 3, etc. For instance, let us consider question Q10 (*"Which of the policy options described below would you be most in favour of?"*). The coding has been implicitly made since the respondent must choose a value between 1 and 10. For question Q11 (*"Generally speaking, are you satisfied or dissatisfied with the job the City of Thousand Oaks is doing to provide city services?"*), the coding is similar, the only difference being that a 4-point rating scale should be used:

$$Q11 = \begin{cases} 4 & \text{if "Very satisfied"} \\ 3 & \text{if "Somewhat satisfied"} \\ 2 & \text{if "Somewhat dissatisfied"} \\ 1 & \text{if "Very dissatisfied"} \end{cases}$$

For question Q12 (*"To what extent do you agree or disagree that the City of Miami Beach government is open and interested in hearing the concerns or issues of residents?"*), one should use a 5-point ranking scale, where 3 represents the neutral position:

$$Q12 = \begin{cases} 5 & \text{if "Strongly agree"} \\ 4 & \text{if "Somewhat agree"} \\ 3 & \text{if "Neutral"} \\ 2 & \text{if "Somewhat disagree"} \\ 1 & \text{if "Strongly disagrees"} \end{cases}$$

Question Q13 (*"On a scale from 1 to 10 can you indicate how satisfied you are with the life you lead at the moment?"*) already provides the respondent with a 10-point rating scale.

Last, question Q14, which uses a semantic differential scale, can be separated into three variables (Corrosive, Leaves No Scale, Stains Fixtures) and recoded on a 5-point scale, for instance:

$$Q14|_{\text{Corrosive}} = \begin{cases} 5 & \text{if "check on 1st line"} \\ 4 & \text{if "check on 2nd line"} \\ 3 & \text{if "check on 3rd line"} \\ 2 & \text{if "check on 4th line"} \\ 1 & \text{if "check on 5th line"} \end{cases}$$

One difficulty with survey methods is that the questionnaire may contain a high number of nonresponses. They can be of two types: item nonresponse, which occurs when the respondent partially answered the questionnaire, and total or unit nonresponse, which occurs when all or almost all data for a sampling unit are missing. While the first type of error can be solved using imputation techniques, the second type generates more severe biases, especially when these nonresponses are correlated to some characteristic of the population (e.g., illiteracy). If the nonresponse

rate is high, that can also impact the sample size and therefore the precision of the analysis.

The problem of total nonresponse can only be tackled ex ante. First, as already stated, if the response rate can be predicted in advance, the initial sample size should be adjusted accordingly. Second, it is possible to improve the response rate by providing higher incentives to participate, for instance by explaining the purpose of the study, by offering coupons, additional services, by using media to let citizens know that their feedbacks have been used after previous surveys.

When faced with item nonresponses, there are two possibilities. Either one excludes the item from the analysis, or replaces the missing value using imputation techniques. In the first case, the value is generally coded with a blank or NA (for Non Available). Using NA is preferable as it does not incur the risk to be mistaken with a value one has forgotten to report. In addition, spreadsheets commonly understand what NA means. For instance, the command AVERAGE from Excel yields 7 when faced with values “10, NA, 4”. It should be stressed that “0” (zero) should never been used to code a missing item. This would be highly confusing since in practice, many variables can reach this value even when the item is not missing. More generally, note that characters like “;” or “/” should be avoided as most statistical packages cannot handle them properly.

Second, if one wants to replace missing values using imputation techniques, the two most common methods are deductive imputation and mean value imputation. Deductive imputation consists in using logic to deduce the missing value. Typical examples are when the sum of percentage items is less than 100%, or when a ranking question has missing values. Assume for instance that question Q9 has been filled in as follows:

*Q9. Please rank the following issues in order of their importance to you. 1 stands for the most important and 6 for the least important.*

1. <i>International tensions (terrorism, war)</i>	<u>3</u>
2. <i>Economic concerns (unemployment, inflation)</i>	<u>2</u>
3. <i>Environmental concerns (waste, air pollution)</i>	<u>1</u>
4. <i>Health concerns (Bird flu, AIDS)</i>	<u>  </u>
5. <i>Social issues (poverty, discrimination)</i>	<u>5</u>
6. <i>Personal safety (crime, theft...)</i>	<u>4</u>

In that case, one may quite safely attribute “6” to the missing item. However, deduction may not always be so straightforward, for instance with two or more missing values.

Mean value imputation replaces the missing value with the mean value for a given class. Assume for instance that a data set contains information about employees in a given industry and that values are missing with respect to their monthly income. Those missing values can be imputed by the average monthly income for respondents who correctly reported their remuneration and who are in the same company or geographic area. This may however reduce the sampling

variance and, as such, artificially increase the sampling precision. The method should thus be used only as a last resort.

There is also always a risk that participants do not pay attention, do not read instructions, or answer randomly. Several methods exist to identify careless responders or inconsistent values. The most popular approach consists in including an attention filter where respondents are required to choose one specific answer option, sometimes regardless of their own preference:

*Reading the instructions carefully is critical. If you are paying attention please choose “7” below.*

1	2	3	4	5	6	7	8	9	10
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

This type of question is used to flag those who do not carefully read the instructions. An alternative method is to use reaction times or the duration of survey completion if the interview is computer based.

Another possibility is to identify outliers, which by definition are values that lie in the tails of a statistical distribution. In this respect, the first thing to do when checking for the quality of a database is to compute minimum and maximum values. This allows to verify whether the collected information is consistent with what one might expect. In Excel, functions *MIN* and *MAX* can be used. A more general approach is to identify values that lie outside the interquartile range. The latter is defined as  $[Q_1, Q_3]$ , where  $Q_1$  is the middle value in the first half of the rank-ordered data set and  $Q_3$  is the middle value in the second half of the rank-ordered data set. In Excel, one may use for instance the function *QUARTILE(array, quart)*, which returns the quartile of a data set. If *quart* equals 1, the function returns the first quartile (25th percentile); if *quart* equals 2, the function returns the median value (50th percentile); If *quart* equals 3, it returns the third quartile (75th percentile).

Respondents who are flagged as outliers can be excluded from subsequent analysis, or inconsistent values be imputed using the previous techniques. One should however be careful as being an outlier is not necessarily synonymous with careless responding. Some respondents may be natural outliers, with preferences rather apart from those of more standard individuals.

Consider for instance Fig. 2.6. It provides a database constructed only for the purpose of illustrating the approach. The data correspond to a survey based on 22 citizens of a city and their satisfaction (on a 4-rating scale) about a public service, say, a response center. The city is divided into three districts whose zip codes are 700, 800 and 900, respectively. *Gender* is coded as 1 for female and 2 for male. The data have been ordered according to age (variable *Age1*). As can be seen, the minimum value for this variable is 1 and the maximum is 861. This quick glance thus points out problems in the database. Using the *quartile* function, we find  $Q_1 = 34$  and  $Q_3 = 74$ . These values correspond to individuals 6 and 16 in the dataset. In theory, one should be suspicious about any value out of this range. For instance, we can eliminate individuals 1, 20 and 21 as their age corresponds to inconsistent values. However, individuals 2, 3, 4, 5, 17, 18, and 19 are



Individual	Age1	Gender	Zip code	Satisfaction	Age2
1	1	2	700	4	55.4
2	20	2	800	2	20
3	20	1	800	4	20
4	21	1	900	3	21
5	29	2	700	2	29
6	34	1	800	2	34
7	36	2	700	3	36
8	45	1	700	4	45
9	48	1	800	4	48
10	56	1	900	2	56
11	56	1	900	1	56
12	56	2	700	3	56
13	65	2	700	2	65
14	66	2	800	4	66
15	68	1	800	4	68
16	74	2	700	4	74
17	76	2	800	3	76
18	77	2	900	1	77
19	98	1	700	3	98
20	455	2	800	4	55.4
21	861	2	900	1	55.4
22	NA	2	700	2	55.4
MIN	1				
MAX	861				
Q1	34				
Q3	74				

**Fig. 2.6** Database: example 1

natural outliers and should not be eliminated. If one wants to keep all the observations for a subsequent analysis, it is possible to replace the missing or inconsistent values either with “NA” or mean values. For instance, the average age for females (individuals 3, 4, 6, 8, 9, 10, 11, 15 and 19) is 49.6, while the average age for males (individuals 2, 5, 7, 12, 13, 14, 16, 17 and 18) is 55.4. These values can be used to recode variable *Age1* into variable *Age2*.

Ideally, a sample survey should cover groups of the target population in proportions that match the proportions of those groups in the population itself. In practice, this may not always be the case. Due to the sampling design, to non-coverage issues or nonresponse, some groups may be over- or under-represented. In such situations, no reliable conclusions can be drawn from the sample, unless it is adjusted using raking techniques (also known as sample-balancing or iterative proportional fitting). The idea is to assign a weight to each responding unit so that the sample better matches the target population. Units that are under-represented are attributed a weight greater than 1, and those that are over-represented are attributed a weight smaller than 1.

Let us first consider a simple example with one single control variable. In Fig. 2.6, we have information about the gender of each respondent: 9 respondents are females, and 13 are males (outliers included). Assume now that we can compare the response distribution of *Gender* with the population distribution, assumed to be equally distributed between males and females:

<b>Sample</b>	Female: 9 (40.9%)	Male: 13 (59.1% )
<b>Population (census)</b>	Female: 2200 (50%)	Male: 2200 (50%)

The population includes 50% of males, while it is 59% in the sample. The males are thus over-represented in our sample. We can solve this representativeness bias by assigning adequate weights to male and female respondents:

$$Weight|_{female} = 50\%/40.9\% = 1.22$$

$$Weight|_{male} = 50\%/59.1\% = 0.85$$

The weights are obtained by dividing the population percentage by the corresponding sample percentage.

In practice, it is frequent to use several control variables. The computational approach is complex and relies on raking algorithms. To illustrate, assume now that we use both *Gender* (two categories: male, female) and *Zip code* (three categories: 700, 800, 900) to correct the representativeness bias. Combining all possibilities of gender and zip code leads to  $2 \times 3$  different groups. Assume now that we have information about the distribution of *Zip code* within the target population:

<b>Zip code</b>	700	800	900
<b>Sample</b>	9 (40.9%)	8 (36.4%)	5 (22.7%)
<b>Population</b>	2000 (45.45%)	1200 (27.27%)	1200 (27.27%)

How can we use this information to compute the weights? Figure 2.7 illustrates the approach. Figure 2.7a contains information about the total frequencies in the sample. For instance, in our dataset, 2 females live in district 700, 4 in district 800 and 3 in district 900. Last row and column of Fig. 2.7a provide the target to attain. Since the population is equally distributed among males (50%) and females (50%), one should obtain similar proportions in the sample, i.e.  $50\% \times 22 = 11$ . Similarly, since 45.45% of the population lives in district 700, one should have  $45.45\% \times 22 = 10$  units for this category in the sample, and 6 units for districts 800 and 900.

Raking is achieved with successive iterations until one converges to the desired set of proportions. In Fig. 2.7b, the first iteration consists in aiming at 11 for the total of males and females. Value 2.44 is obtained by multiplying the sample frequency (here 2) by  $11/9$ . Similarly, 4.89 is the product of 4 with  $11/9$ , and so on. The second iteration aims at the desired set of proportions for *Zip code*. For instance, value 2.92 is obtained by multiplying 2.44 with  $10/8.37$ . The new values obtained however affect in return the total frequency of males and females, and the process must be reiterated until one reaches convergence. As can be seen from Fig. 2.7e, after four iterations, the values are more stable. The weights are finally obtained by dividing the values of Fig. 2.7e by those of Fig. 2.7a (see Fig. 2.7f).

As can be deduced from the previous example, raking can be laborious, and one may rely instead on statistical software to assess the relevant weights. Figure 2.8

<b>a</b>					
	700	800	900	Total	Target
Female	2	4	3	9	11
Male	7	4	2	13	11
Total	9	8	5	22	
Target	10	6	6		
<b>b</b>					
	700	800	900	Total	Target
Female	$2.44 = 2 \times 11/9$	$4.89 = 4 \times 11/9$	$3.67 = 3 \times 11/9$	11	11
Male	5.92	3.38	1.69	11	11
Total	8.37	8.27	5.36	22	
Target	10	6	6		
<b>c</b>					
	700	800	900	Total	Target
Female	$2.92 = 2.44 \times 10/8.37$	3.55	4.11	10.57	11
Male	$7.08 = 5.92 \times 10/8.37$	2.45	1.89	11.43	11
Total	10.00	6.00	6.00	22.00	
Target	10	6	6		
<b>d</b>					
	700	800	900	Total	Target
Female	3.04	3.69	4.27	11.00	11
Male	6.81	2.36	1.82	11.00	11
Total	9.85	6.05	6.10	22.00	
Target	10	6	6		
<b>e</b>					
	700	800	900	Total	Target
Female	3.08	3.66	4.20	10.95	11
Male	6.92	2.34	1.80	11.05	11
Total	10.00	6.00	6.00	22.00	
Target	10	6	6		
<b>f</b>					
	700	800	900		
Female	$3.08/2 = 1.54$	0.91	1.40		
Male	0.98	0.58	0.89		

**Fig. 2.7** Raking with two variables: example 1. (a) Sample, (b) Iteration 1, (c) Iteration 2, (d) Iteration 3, (e) Iteration 4, (f) weights

provides the coding to be used in R-CRAN. Command `read.table` is used to upload the database, saved afterwards under the name “D”, using the path `C://mydata.csv`, which denotes the location of the file. The file format is `.csv`, with “;” as a separator, and can be easily created with Excel. The command `head` displays the first rows of the dataset. To use the `anesrake` function, all variables must be coded continuously (1, 2, 3, etc.) with no missing values. Variable `Zip.code` has thus been recoded from 1 to 3. On average, the level of satisfaction with respect to the public service under evaluation is 2.81. This value however does not take into account the representativity bias. The package `weights` allows to compute the proportion of males and females using `wpct(D$Gender)`, as well as how the sampling units are distributed among the districts, using `wpct(D$Zip.code)`. The next step is to specify manually the population characteristics:

$$p.gender = c(0.50, 0.50)$$

$$p.zip = c(0.454545455, 0.272727273, 0.272727273)$$

$$targets = list(p.gender, p.zip)$$

$$names(targets) = c("Gender", "Zip.code")$$

```

> D=read.table("C://mydatasampling.csv",head=TRUE,sep=";")
> head(D)
  Individual Age1 Gender Zip.code Satisfaction Age2
1          1    1     2         1          4 55.4
2          2   20     2         2          2 20.0
3          3   20     1         2          4 20.0
4          4   21     1         3          3 21.0
5          5   29     2         1          2 29.0
6          6   34     1         2          2 34.0
> mean(D$Satisfaction)
[1] 2.818182

> library(weights)
> wpct(D$Gender)
      1      2
0.4090909 0.5909091
> wpct(D$Zip.code)
      1      2      3
0.4090909 0.3636364 0.2272727

> p.gender=c(0.50,0.50)
> p.zip=c(0.454545455,0.272727273,0.272727273)
> targets=list(p.gender, p.zip)
> # important: use the same variable names of the dataset
> names(targets)=c("Gender", "Zip.code")

> library(anesrake)
> myrake=anesrake(targets, D, caseid=D$Individual)

> D$myweights=myrake$weightvec
> head(D$myweights)
[1] 0.9845212 0.5817086 0.9182914 1.4061610 0.9845212 0.9182914
> mean(D$myweights*D$Satisfaction)
[1] 2.78211

```

**Fig. 2.8** Raking with R-CRAN: example 1

It is important to use the same variable names as the dataset. The raking algorithm is implemented using the package *anesrake*. One has to specify the desired set of proportions (*targets*), the dataset (*D*), the column that contains the individual numbers (*D\$Individual*). Finally, the weights (*myrake\$weightvec*) are saved into the dataset *D* under the name *D\$myweights*. Those weights are similar to those presented in Fig. 2.7f. For instance, individual 1 is a male from district 1 and as such receives a weight of 0.98. Individual 3 is a female from the second district and gets a weight of 0.91. Weights can be used to compute the average satisfaction *mean(D\$myweights \* D\$Satisfaction)*. We now find 2.78.

Note that raking adjustments imply to know only the population totals of the specific variables, not all cells of a cross-table. The first step is to identify a set of variables likely to be used as control variables, and to compare them with reliable data sources (e.g., census). Some typical variables are age groups, gender, socio economic status, geographical location. When selected variables have categories with less than 5% in the sample, it is recommended to collapse them.

**Bibliographical Guideline** Several definitions in this chapter have been taken and modified from the OECD Glossary of Statistical Terms. This glossary contains additional definitions of key concepts and commonly used acronyms. These definitions are primarily drawn from existing international statistical guidelines and recommendations that have been prepared over the last two or three decades by international organizations (such as the United Nations, International Labor Organization, Organization for Economic Co-operation and Development, Eurostat, International Monetary Fund) working with national statistical institutes and other agencies responsible for the initial compilation and dissemination of statistical data.

Several guides are also available online, such as the “Guidelines for Designing Questionnaires for Administration in Different Modes” proposed by the United States Census Bureau, “Public Opinion Surveys as Input to Administrative Reform” by the Organization for Economic Co-operation and Development, “Designing Household Survey Samples: Practical Guidelines” by the Department of Economic and Social Affairs of the United Nations, “Survey Methods and Practices” by Statistics Canada, the “Handbook of Recommended Practices for Questionnaire Development and Testing in the European Statistical System” by Eurostat.

---

## Bibliography

Eurostat. (2004). *Handbook of recommended practices for questionnaire development and testing in the European Statistical System*.

OECD. (1998). *Public opinion surveys as input to administrative reform* (SIGMA Papers, No. 25). OECD Publishing.

Statistics Canada. (2010). *Survey methods and practices*.

US Census Bureau. (2007). *Guidelines for designing questionnaires for administration in different modes*.

United Nations. (2005). *Designing household survey samples: Practical guidelines*.

Statistical Tools for Program Evaluation

Methods and Applications to Economic Policy, Public  
Health, and Education

Josselin, J.-M.; Le Maux, B.

2017, X, 531 p. 139 illus., 86 illus. in color., Hardcover

ISBN: 978-3-319-52826-7