Chapter 3
Data

Learning Objectives

After reading this chapter you should understand:

- How to explain what kind of data you use.
- The differences between qualitative and quantitative data.
- What the unit of analysis is.
- When observations are independent and when they are dependent.
- The difference between dependent and independent variables.
- Different measurement scales and equidistance.
- The differences between primary and secondary data.
- Reliability and validity from a conceptual viewpoint.
- How to set up several different sampling designs.
- How to determine acceptable sample sizes.

Keywords Case · Construct · Data · Equidistance · Item · Measurement scaling · Observation · Operationalization · Primary and secondary data · Qualitative and quantitative data · Reliability · Validity · Sample sizes · Sampling · Scale development · Variable

Introduction

Data lie at the heart of conducting market research. By *data* we mean a collection of facts that can be used as a basis for analysis, reasoning, or discussions. Think, for example, of the answers people give to surveys, existing company records, or observations of shoppers' behaviors.

In practice, "good" data are very important because they are the basis for good market research findings. In this chapter, we will discuss some of the different types of data. This will help you describe what data you use and why. Subsequently, we discuss strategies to collect data in Chap. 4.

E. Mooi and M. Sarstedt, *A Concise Guide to Market Research*,
DOI 10.1007/978-3-642-12541-6_3, © Springer-Verlag Berlin Heidelberg 2011

Types of Data

Before we start discussing data, it is a good idea to introduce some terminology. Data are present in the form of variables and cases for quantitative data and in the form of words and pictures for qualitative data (we will further discuss this distinction later in this chapter). A *variable* is an attribute whose value can change. For example, the price of a product is an attribute of that product and typically varies over time. If the price does not change, it is a *constant*. A *case* (or *observation*) consists of all the observed variables that belong to an object such as a customer, a company or a country. The relationship between variables and cases is that within one case we usually find multiple variables. In Table 3.1, you can see typical examples of data. In the columns, you find the variables. There are six variables; *type of car bought*, *age* and *gender*, as well as *brand_1*, *brand_2*, and *brand_3* which capture statements related to brand trust. In the lower rows, you can see the four observations we have.

Although marketers often talk about variables, they also use the word *item*, which usually refers to a survey question put to a respondent. Another important term that is frequently used in market research is *construct*, which refers to a variable that is not directly observable. For example, while *gender* and *age* are directly observable, this is not the case with concepts such as customer satisfaction, loyalty or brand trust. More precisely, a construct can be defined as a concept which cannot be directly observed, and for which there are multiple referents, but none all-inclusive. For example, *type of car bought* from Table 3.1 is not a hypothetical construct because, despite variation in brands and categories, there is an agreed upon definition for a car with specific characteristics that distinguish a car from, for example, a plane. Furthermore, the type of car can be directly observed (e.g., BMW 328i, Mercedes C180K). On the other hand, constructs consist of groups of functionally related behaviors, attitudes, and experiences and are psychological in nature. Researchers cannot directly measure satisfaction, loyalty, or brand trust. However, they can measure indicators or manifestations of what we have agreed to call satisfaction, loyalty, or trust, for example, in a brand, product or company. Consequently, we generally have to combine several items to form a so called multi-item scale which can be used to measure a construct.

This aspect of combining several items is called *scale development*, *operationalization*, or, in special cases, *index construction*. Essentially, it is a combination of theory and statistical analysis, such as factor analysis (discussed in Chap. 8) aiming at developing appropriate measures of a construct. For example, in Table 3.1, *brand_1*, *brand_2*, and *brand_3* are items that belong to a construct called brand trust (as defined by Erdem and Swait 2004). The construct is not an individual item that you see in the list, but it is captured by calculating the average of a number of related items. Thus, for brand trust, the score for customer 1 is $(6 + 5 + 7)/3 = 6$.

But how do we decide which and how many items to use when measuring specific constructs? To answer these questions, market researchers make use of scale development procedures which follow an iterative process with several steps

Table 3.1 Quantitative data

Variable name	type of car bought	age	gender	brand_1	brand_2	brand_3
Description	Name of car bought	Age in years	Gender (male = 0, female = 1)	This brand's product claims are believable (1 = fully disagree, 7 = fully agree)	This brand delivers what it promises (1 = fully disagree, 7 = fully agree)	This brand has a name that you can trust (1 = fully disagree, 7 = fully agree)
Customer 1	BMW 328i	29	1	6	5	7
Customer 2	Mercedes C180K	45	0	6	6	6
Customer 3	VW Passat 2.0 TFSI	35	0	7	5	5
Customer 4	BMW 525ix	61	1	5	4	5

and feedback loops. For example, DeVellis (2003) and Rossiter (2010) provide two different approaches to scale development. Unfortunately, these procedures require much (technical) expertise. Describing each step goes beyond the scope of this book. However, for many scales you do not need to use this procedure, as existing scales can be found in scale handbooks, such as the *Handbook of Marketing Scales* by Bearden and Netemeyer (1999). Furthermore, marketing and management journals frequently publish research articles that introduce new scales, such as for the reputation of non-profit organizations (Sarstedt and Schloderer 2010) or brand relevance (Fischer et al. 2010). Nevertheless, we introduce two distinctions that are often used to discuss constructs in Box 3.1.

Box 3.1 Type of constructs

Reflective vs. *formative constructs*: When considering reflective constructs, there is a causal relationship from the construct to the items. In other words, the items reflect the construct. This is also the case in our previous example, with three items reflecting the concept of brand trust. Thus, if a respondent changes his assessment of brand trust (e.g., because the respondent heard negative news concerning that brand), it is going to be reflected in the answers to the three items. Erdem and Swait (2004) use several (interrelated) items to capture different aspects of brand trust, which generally yields more exact and stable results. Moreover, if we have multiple items, we can use analysis techniques that inform us about the quality of measurement such as factor analysis or reliability analysis (discussed in Chap. 8). On the other hand, formative constructs consist of a number of items that define a construct. A typical example is socioeconomic status, which is formed by a combination of education, income, occupation, and residence. If any of these measures increases, socioeconomic status would increase (even if the other items did not change). Conversely, if a person's socioeconomic status increases, this would not go hand in hand with an increase in all four measures. This distinction is important when operationalizing constructs, as it requires different approaches to decide on the type and number of items. Moreover, reliability analyses (discussed in Chap. 8) cannot be used for formative measures. For an overview of this distinction, see Diamantopoulos and Winklhofer (2001) or Diamantopoulos et al. (2008).

Multi items vs. *single items*: Rather than using a large number of items to measure constructs, practitioners often opt for the use of single items to operationalize some of their constructs. For example, we may use only "This brand has a name that you can trust" to measure brand trust instead of all three items. While this is a good way to make measurement more efficient and the questionnaire shorter, it also reduces the quality of your measures. There is no general rule when the use of single items is appropriate but, in general, if a construct is concrete (which means that respondents across the board describe the construct's definition almost identically), the use of a single item to measure that construct is justifiable. For example, purchase intention is a rather concrete construct whereas corporate reputation is rather abstract. Thus, the former can be measured by means of a single item whereas multiple items should be used to measure the latter. See Bergkvist and Rossiter (2007) and Sarstedt and Wilczynski (2009) for a discussion. Fuchs and Diamantopoulos (2009) offer guidance on when to use multi items and single items.

Primary and Secondary Data

Generally, we can distinguish between two types of data: *primary* and *secondary data*. While *primary data* are data that a researcher has collected for a specific purpose, secondary data have already been collected by another researcher for another purpose. An example of secondary data is the US Consumer Expenditure Survey (http://www.bls.gov/cex/home.htm), which makes data available on what people in the US buy, such as insurances, personal care items, or food. It also includes the prices people pay for these products and services. Since these data have already been collected, they are secondary data. If a researcher sends out a survey with various questions to find an answer to a specific issue, the collected data are primary data. If these primary data are re-used to answer another research question, they become secondary data.

Secondary data can either be *internal* or *external* (or a mix of both). Internal secondary data are data that an organization or individual already has collected, but wants to use for (other) research purposes. For example, we can use sales data to investigate the success of new products, or we can use the warranty claims people make to investigate why certain products are defective. External secondary data are data that other companies, organizations, or individuals have available, sometimes at a cost.

Secondary and primary data have their own specific advantages and disadvantages. These are summed up in Table 3.2. Generally, the most important reasons for

Table 3.2 The advantages and disadvantages of secondary and primary data

	Secondary data	Primary data
Advantages	Tend to be cheaper	Are recent
	Sample sizes tend to be greater	Are specific for the purpose
	Tend to have more authority	Are proprietary
	Are usually quick to access	
	Are easier to compare to other research that uses the same data	
	Are sometimes more accurate	
Disadvantages	May be outdated	Are usually more expensive
	May not completely fit the problem	Take longer to collect
	There may be errors in the data	
	Difficult to assess data quality	
	Usually contains only factual data	
	May not be available	
	No control over data collection	
	May not be reported in the required form (e.g., different units of measurement, definitions, aggregation levels of the data)	

using secondary data are that they tend to be cheaper and quick to obtain access to (although there can be lengthy processes involved). For example, if you want to have access to the US Consumer Expenditure Survey, all you have to do is point your web browser to http://www.bls.gov/cex/csxmicro.htm and pay 145 USD. Furthermore, the authority and competence of some of these research organizations might be a factor. For example, the claim that Europeans spend 9% of their annual income on health may be more believable if it comes from Eurostat (the statistical office of the European Community) than if it came from a single survey conducted through primary research.

However, important drawbacks of secondary data are that they may not answer your research question. If you are, for example, interested in the sales of a specific product (and not in a product or service category), the US Expenditure Survey may not help much. In addition, if you are interested in reasons why people buy products, this type of data may not help answer your question. Lastly, as you did not control the data collection, there may be errors in the data. Box 3.2 shows an example of inconsistent results in two well-known surveys on Internet usage.

Box 3.2 Contradictory results in secondary data

The European Interactive Advertising Association (EIAA) (http://www.eiaa.net) is a trade organization for media companies such as CNN Interactive and Yahoo! Europe focused on interactive business. The Mediascope study issued yearly by EIAA provides insight into the European population's media consumption habits. For example, according to their 2008 study, 47% of all Germans are online every single day. However, this contradicts the results from the well-known German General Survey (ALLBUS) issued by the Leibniz Institute for Social Sciences (http://www.gesis.org), according to which merely 26% of Germans used the Internet on a daily basis in 2008.

In contrast, primary data tend to be highly specific because the researcher (you!) can influence what the research comprises. In addition, primary research can be carried out when and where it is required and cannot be accessed by competitors. However, gathering primary data often requires much time and effort and, therefore, is usually expensive compared to secondary data.

As a rule, start looking for secondary data first. If they are available, and of acceptable quality, use them! We will discuss ways to gather primary and secondary data in Chap. 4.

Quantitative and Qualitative Data

Data can be *quantitative* or *qualitative*. Quantitative data are presented in values, whereas qualitative data are not. Qualitative data can therefore take many forms such as words, stories, observations, pictures, or audio. The distinction between qualitative and quantitative data is not as black-and-white as it seems, because quantitative data are based on qualitative judgments. For example, the questions on brand trust in Table 3.1 take the values of 1–7. There is no reason why we could not have used other values to code these answers, but it is common practice to code answers of a construct's items on a range of 1–5 or 1–7.

In addition, when the data are "raw," we often label them qualitative data, although researchers can code attributes of the data, thereby turning it into quantitative data. Think, for example, of how people respond to a new product in an interview. We can code this by setting neutral responses to 0, somewhat positive responses to 1, positives ones to 2, and very positive ones to 3. We have thus turned qualitative data contained in the interview into quantitative data. This is also qualitative data's strength and weakness; they are very rich but can be interpreted in many different ways. However, the process of how we interpret qualitative data is also subjective. To reduce some of these problems, qualitative data are often coded by highly trained researchers.

Most people think of quantitative data as being more factual and precise than qualitative data, but this is not necessarily true. Rather, what is important is how well qualitative data have been collected and/or coded into quantitative data.

Unit of Analysis

The unit of analysis is the level at which a variable is measured. Researchers often ignore this, but it is crucial because it determines what we can learn from the data. Typical measurement levels include individuals (respondents or customers), stores, companies, or countries. It is best to use data at the lowest possible level, because this provides more detail and if we need these data at another level, we can *aggregate the data*. Aggregating data means that we sum up a variable at a lower level to create a variable at a higher level. For example, if we know how many cars all car dealers in a country sell, we can take the sum of all dealer sales, to create a variable measuring countrywide car sales. This is not always possible, because, for example, we may have incomplete data.

Dependence of Observations

A key issue for any data is the degree to which observations are related. If we have exactly one observation from each individual, store, company, or country, we label the observations *independent*. That is, the observations are completely unrelated. If, however, we have multiple observations from each individual, store, company, or country, we label them *dependent*. The dependence of observations occurs if the number of observations is larger than the number of respondents (or objects) and if each individual respondent's (or object's) responses are related. For example, we could ask respondents to rate a type of Cola, then show them an advertisement, and again ask them to rate the same type of Cola. Although the advertisement may influence the respondents, it is likely that the first and second response will still be related. That is, if the respondents first rated the Cola negatively, the chance is higher that they will still rate the Cola negatively rather than positively after the advertisement. If the data are dependent, this often impacts what type of analysis we should use. For example, in Chap. 6 we discuss the difference between the independent samples t-test (for independent observations) and the paired samples t-test (for dependent observations).

Dependent and Independent Variables

An artificial distinction that market researchers make is the difference between dependent and independent variables. Independent variables are those that are used to explain another variable, the dependent variable. For example, if we use the amount of advertising to explain sales, then advertising is the independent variable and sales the dependent.

This distinction is artificial, as all variables depend on other variables. For example, the amount of advertising depends on how important the product is for a company, the company's strategy, and other factors. However, the distinction is frequently used and helpful, as it indicates what is being explained, and what variables are used to explain it.

Measurement Scaling

Not all data are equal! For example, we can calculate the average age of the respondents of Table 3.1 but it would not make much sense to calculate the average gender. Why is this? The values that we have assigned male (0) or female (1) respondents are arbitrary; we could just as well have given males the value of 1 and female the value of 0, or we could have used the values of 1 and 2. Therefore, choosing a different coding would result in different results.

The example above illustrates the problem of *measurement scaling*. Measurement scaling refer to two things: the measurements we use for measuring a certain construct (see discussion above) and the level at which a variable is measured which we discuss in this section. This can be highly confusing!

There are four levels of measurement: the *nominal scale*, the *ordinal scale*, the *interval scale*, and the *ratio scale*. These relate to how we quantify what we measure. It is vital to know the scale on which something is measured because, as the gender example above illustrates, the measurement scale determines what analysis techniques we can, or cannot, use. For example, it makes no sense to calculate the average of a variable indicating the gender of the respondent. We will come back to this issue in Chap. 5 and beyond.

The *nominal scale* is the most basic level at which we can measure something. Essentially, if we use a nominal scale, we substitute a word for a *label* (numerical value). For example, we could code what types of soft drinks are bought as follows: Coca-Cola = 1, Pepsi-Cola = 2, Seven-Up = 3. In this example, the numerical values represent nothing more than a label.

The *ordinal scale* provides more information. If we have a variable measured on an ordinal scale, we know that if the value of that variable increases, or decreases, this gives meaningful information. For example, if we code customers' usage of a product as non-user = 0, light user = 1, and heavy user = 2, we know that if the value of the usage variable increases the usage also increases. Therefore, something measured with an ordinal scale provides information about the *order* of our observations. However, we do not know if the differences in the order are equally spaced. That is, we do not know if the difference between "non-user" and "light user" is the same as between "light user" and "heavy user," even though the difference in the values (0–1 and 1–2) is equal.

If something is measured with an *interval scale*, we have precise information on the rank order at which something is measured and, in addition, we can interpret the magnitude of the differences in values directly. For example, if the temperature is 25°C, we know that if it drops to 20°C, the *difference* is exactly 5°C. This difference of 5°C is the same as the increase from 25°C to 30°C. This exact "spacing" is called *equidistance* and equidistant scales are necessary for certain analysis techniques, such as factor analysis (which we will discuss in Chap. 8). What the interval scale does not give us, is an absolute zero point. If the temperature is 0°C it may feel cold, but the temperature can drop further. The value of 0 therefore does not mean that there is no temperature at all.

The *ratio scale* provides the most information. If something is measured with a ratio scale, we know that a value of 0 means that that particular variable is not present. For example, if a customer buys no products (value = 0) then he or she really buys no products. Or, if we spend no money on advertising a new product (value = 0), we really spend no money. Therefore, the *zero point* or origin of the variable is equal to 0.

While it is relatively easy to distinguish the nominal and interval scales, it is quite hard to understand the difference between the interval and ratio scales. Fortunately, in practice, the difference between the interval and ratio scale is largely ignored and both categories are sometimes called the *quantitative scale* or *metric scale*. Table 3.3 could help you assess the difference between these four scales.

Table 3.3 Measurement Scaling

	Label	Order	Differences	Origin is 0
Nominal scale	✓			
Ordinal scale	✓	✓		
Interval scale	✓	✓	✓	
Ratio scale	✓	✓	✓	✓

Reliability and Validity

In any market research process, it is paramount to use "good" measures. Good measures are those that consistently measure what they are supposed to measure. For example, if we are interested in knowing if existing customers like a new TV commercial, we could show them the commercial and ask, for example, the following two questions afterwards (1) "Did you enjoy watching the commercial?," and (2) "Did the commercial provide the essential information necessary for a purchase decision?" How do we know if either of these questions really measures if the viewers like the commercial? We can think of this as a measurement problem through which we relate what we want to measure – whether existing customers like a new TV commercial – with what we actually measure in terms of the questions we ask. If these relate perfectly, our actual measurement is equal to what we intend to measure and we have no measurement error. If these do not relate perfectly, as is usual in practice, we have *measurement error*.

This measurement error can be divided into a *systematic* and a *random* error. We can express this as follows:

$$X_O = X_T + X_S + X_R$$

X_O stands for the observed score, X_T for the true score, X_S for the systematic error, and X_R for the random error. Systematic error is a measurement error through which we consistently measure higher, or lower, than we actually want to measure. If we were to ask, for example, customers to evaluate a TV commercial and offer them remuneration in return, they may provide more favorable information than they would otherwise have. This may cause us to think that the TV commercial is systematically more, or less, enjoyable than it is in reality. On the other hand, there may be random errors. Some customers may be having a good day and indicate that they like a commercial whereas others, who are having a bad day, may do the opposite.

Systematic errors cause the actual measurement to be consistently higher, or lower, than what it should be. On the other hand random error causes (random) variation between what we actually measure and what we should measure.

The systematic and random error concepts are important because they relate to a measure's validity and reliability. *Validity* refers to whether we are measuring what we want to measure and, therefore, to a situation where the systematic error X_S is zero. *Reliability* is the degree to which what we measure is free from random error and, therefore, relates to a situation where the X_R is zero. In Fig. 3.1, we explain the difference between reliability and validity by means of a target comparison. In this analogy, repeated measurements (e.g., of a customer's satisfaction with a specific service) are compared to arrows that are shot at a target. To measure each true score, we have five measurements (indicated by the

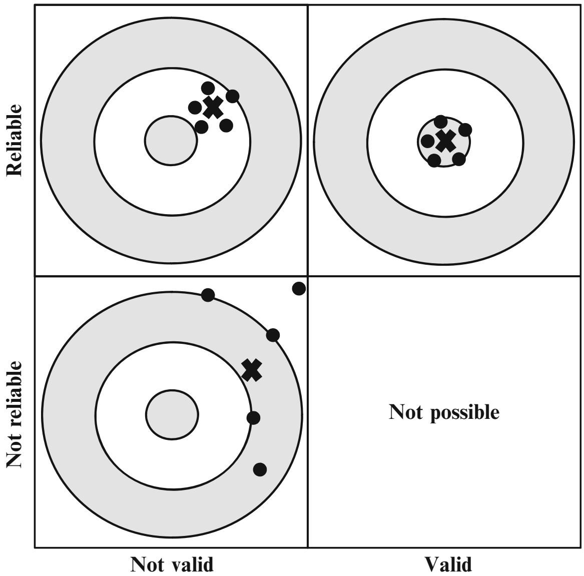

Fig. 3.1 Validity and reliability

black circles) whose average is indicated by a cross. Validity describes the cross's proximity to the bull's eye at the target center. The closer the average to the true score, the higher the validity. If several arrows are fired, reliability is the degree to which the arrows are apart. When all arrows are close together, the measure is reliable, even though it is not necessarily near the bull's eye. This corresponds to the upper left box where we have a scenario in which the measure is reliable but not valid. In the upper right box, both reliability and validity are given. In the lower left box, though, we have a situation in which the measure is neither reliable, nor valid. Obviously, this is because the repeated measurements are scattered quite heavily and the average does not match the true score. However, even if the latter were the case (i.e., if the cross were in the bull's eye), we would still not consider the measure valid. An unreliable measure can never be valid because there is no way we can distinguish the systematic error from the random error. If we repeated the measurement, say, five more times, the random error would likely shift the cross in a different position. Thus, reliability is a necessary condition for validity. This is also why the scenario not reliable/valid (lower right box) does not make sense.

For some variables, such as length or income, we can objectively verify what this true score should be, but for constructs that describe unobservable phenomena, such as satisfaction, loyalty or brand trust, this is virtually impossible (from a philosophical point of view, one could even argue that there is no "true" score for a construct). So how do we know if a measure is valid? Because there is no direct way to know what it is that we are measuring, there are several different aspects of validity, including face, content, construct, and predictive validity. These types of validity help us understand the association between what we should measure and what we actually measure.

Construct validity in a general term that includes the different types of validity discussed next.

Face validity is an absolute minimum requirement for a variable to be valid and refers to whether a variable reflects what you want to measure. Essentially, face validity exists if a measure seems to make sense. For example, if you want to measure trust, asking questions such as “this company is honest and truthful” makes quite a lot of sense whereas “this company is not well known” makes much less sense. Researchers should agree on face validity before starting the actual measurement (i.e., handing out the questions to respondents). Face validity is often determined by using a sample of experts who discuss and agree on the degree of face validity (this is also referred to as *expert validity*).

Content validity is strongly related to face validity but is more formalized. To assess content validity, researchers need to first define what they want to measure and discuss what is included in the definition and what not. For example, trust between businesses is often defined as the extent to which a firm believes that its exchange partner is honest and/or benevolent (Geyskens et al. 1998). This definition clearly indicates what should be mentioned in the questions used to measure trust (honesty and benevolence). After researchers have defined what they want to measure, questions have to be developed that relate closely to the definition. Consequently, content validity is mostly achieved before the actual measurement.

Predictive validity can be tested if we know that a measure should relate to an outcome. For example, loyalty should lead to people purchasing a product, or a measure of satisfaction should be predictive of people not complaining about a product or service. Assessing predictive validity requires collecting data at two points in time and therefore often takes a great deal of time.

Criterion validity is closely related to predictive validity, with the one difference that we examine the relationship between two constructs measured at the same time.[1]

Now that we know a bit more about validity, how do we know if a measure is reliable? Reliability can be assessed by three key factors: stability of the measurement, internal reliability, and inter-rater reliability.

Stability of the measurement means that if we measure something twice (also called *test–retest reliability*) we expect similar outcomes. Stability of measurement requires a market researcher to have collected two data samples and is therefore costly and could prolong the research process. Operationally, researchers administer the same test to the same sample on two different occasions and evaluate

[1]There are other types of validity, such as discriminant validity (see Chap. 8) and nomological validity. See Netemeyer et al. (2003) for an overview.

how strongly the measurements are related (more precisely, they compute correlations between the measurements; see Chap. 5 for an introduction to correlations). For a measure to be reliable, i.e., stable over time, we would expect the two measurements to correlate highly. This approach is not without problems. For example, it is often hard, if not impossible, to survey the same people twice. Furthermore, the respondents may learn from past surveys, leading to "practice effects." For example, it may be easier to recall events the second time a survey is administered. Moreover, test–retest approaches do not work when the survey is about specific time points. If we ask respondents to provide information on their last restaurant experience, the second test might relate to a different restaurant experience. Thus, test–retest reliability can only be assessed for variables that are relatively stable over time.

Internal consistency is by far the most common way of assessing reliability. Internal reliability requires researchers to simultaneously use multiple variables to measure the same thing (think of asking multiple questions in a survey). If these measures relate strongly and positively to one another there is a considerable degree of internal consistency. There are several ways to calculate indices of internal consistency, including split-half reliability and Cronbach's α (pronounced as alpha), which we discuss in Chap. 8.

Inter-rater reliability is a particular type of reliability that is often used to assess the reliability of secondary data or qualitative data (discussed later). If you want to measure, for example, which are the most ethical organizations in an industry, you could ask several experts to provide a rating and then check whether their answers relate closely. If they do, then the degree of inter-rater reliability is high.

Population and Sampling

A *population* is the group of units about which we want to make judgments. These units can be groups of individuals, customers, companies, products, or just about any subject in which you are interested. Populations can be defined very broadly, such as the people living in Canada, or very narrowly, such as the directors of large hospitals. What defines a population depends on the research conducted and the goal of the research.

Sampling is the process through which we select cases from a population. When we develop a sampling strategy, we have three key choices: the census, probability sampling, and non-probability sampling. The most important aspect of sampling is that the sample selected is *representative* of the population. With representative we mean that the characteristics of the sample closely match those of the population. How we can determine whether a sample is representative of the population is discussed in Box 3.3.

Box 3.3 Is my sample representative of the population?

Market researchers consider it important that their sample is representative of the population. How can we see if this is so?

- You can use (industry) experts to judge the quality of your sample. They may look at issues such as the type and proportion of organizations in your sample and population.
- To check whether the responses of people included in your research do not differ from non-respondents, you could use the *Armstrong and Overton procedure*. This procedure calls for comparing the first 50% of respondents to the last 50% with regard to key demographic variables. The idea behind this procedure is that later respondents more closely match the characteristics of non-respondents. If these differences are not significant (e.g., through hypothesis tests, discussed in Chap. 6), we find some support that there is little, or no, response bias (See Armstrong and Overton 1977).
- Perhaps the best way to test whether the sample relates to the population is to use a dataset with information on the population. For example, the Amadeus and Orbis databases provide information at the population level. We can (statistically) compare the information from these databases to the sample selected. The database is available at:

http://www.bvdinfo.com/Home.aspx

If we get lucky and somehow manage to include every unit of the population in our study, we have conducted a *census* study. Census studies are rare because they are very costly and because missing just a small part of the population can have dramatic consequences. For example, if we were to conduct a census study among directors of banks in Luxemburg, we may miss out on a few because they were too busy to participate. If these busy directors happen to be those of the very largest companies, any information we collect would underestimate the effects of variables that are more important in large banks. Census studies work best if the population is small, well-defined, and accessible. Sometimes census studies are also conducted for specific reasons. For example, the United States Census Bureau is required to hold a census of all persons resident in the US every 10 years. Check out the US Census Bureau's YouTube channel using the mobile tag or URL in Box 3.4 to find out more about the door-to-door census taking process of the US Census 2010.

Box 3.4 The US Census 2010

http://www.youtube.com/uscensusbureau#p/

If we select part of the population, we can distinguish two types of approaches: probability sampling and non-probability sampling.

Probability Sampling

Probability sampling approaches provide every individual in the population a chance (not equal to zero) of being included in the sample. This is often achieved by using an accurate *sampling frame*. A sampling frame is a list of individuals in the population. There are various sampling frames, such as Dun & Bradstreet's Selectory database (includes executives and companies), the Mint databases (includes companies in North and South Americas, Italy, Korea, The Netherlands, and the UK), or telephone directories. These sampling frames rarely completely cover the population of interest and often include some outdated information, but due to their ease of use and availability they are frequently used. If the sampling frame and population are very similar, we have little *sampling frame error*, which is the degree to which sample frames represent the population.

Starting from a good-quality sampling frame, we can use several methods to select units from the sampling frame. The easiest way is to use *simple random sampling*, which is achieved by randomly selecting the number of cases required. This can be achieved by using specialized software, or using Microsoft Excel. Specifically, Microsoft Excel can create random numbers between 0 and 1 (using the RAND() function to select cases). Next you choose those individuals from the sampling frame where the random value falls in a certain range. The range depends on the percentage of respondents needed. For example, if you wish to approach 5% of the sampling frame, you could set the range from 0.00 to 0.05. This method is easy to implement and works well for homogeneous sampling frames.

Systematic sampling uses a different procedure. We first number all the observations and subsequently select every n^{th} observation. For example, if our sampling frame consists of 1,000 firms and we wish to select just 100 firms, we could select the 1st observation, the 11th, the 21st, etc. until we reach the end of the sampling frame and have our 100 observations.

Stratified sampling and *cluster sampling* are more elaborate techniques of probability sampling, which require dividing the sampling frame into different groups. When we use stratified sampling, we divide the population into several different homogenous groups called *strata*. These strata are based on key sample characteristics, such as different departments in organizations or the area in which consumers live. Subsequently we draw a random number of observations from each strata. While stratified sampling is more complex and requires knowledge of the sampling frame and population, it also helps to assure that the sampling frame's characteristics are similar to those of the sample. Stratified sampling is very useful if the market research question requires comparing different groups (strata).

Cluster sampling requires dividing the population into different heterogeneous groups with each group's characteristics similar to those of the population. For example, we can divide the consumers of one particular country into different provinces, counties, or councils. Several of these groups can perhaps be created on the basis of key characteristics (e.g., income, political preference, household composition) that are very similar (representative) to those of the population. We can select one or more of these representative groups and use random sampling to select our observations. This technique requires knowledge of the sampling frame and population, but is convenient because gathering data from one group is cheaper and less time consuming.

Non-probability Sampling

Non-probability sampling procedures do not give every individual in the population an equal chance of being included in the sample. This is a drawback, because it can bias subsequent analyses. Nevertheless, non-probability sampling procedures are frequently used as they are easily executed, and are typically less costly than probability sampling methods.

Judgmental sampling is based on researchers taking an informed guess regarding which individuals should be included. For example, research companies often have panels of respondents who were previously used in research. Asking these people to participate in a new study may provide useful information if we know, from experience, that the panel has little sampling frame error.

Snowball sampling is predominantly used if access to individuals is difficult. People such as directors, doctors, or high-level managers often have little time and are, consequently, difficult to involve. If we can ask just a few of these people to provide names and details of others in a similar position, we can expand our sample quickly and access them. Similarly, if you send out a link to an online questionnaire and ask everyone to forward the link to at least five other persons, this is snowball sampling through referrals to people who would be difficult to access otherwise,

When we select a fixed number of observations from different groups, we use *quota sampling*. Quota sampling is similar to stratified random sampling, but we use a fixed number of observations for each group. For example, we can recruit 100 consumers of Pantene shampoo and 100 consumers of Nivea shampoo. Unlike in random stratified sampling, in which the number of observations for each group is determined by population characteristics, in the quota sampling process, the researcher determines the number for judgmental sampling. Once the required number of observations has been found, no observations are added.

Finally, *convenience sampling* is a catch-all term for methods (including the three non-probability sampling techniques just described) in which the researcher makes a subjective judgment. For example, we can use *mall intercepts* to ask people in a shopping mall if they want to fill out a survey. The researcher's control over who ends up in the sample is limited and influenced by situational factors.

Figure 3.2 provides an overview of the different sampling procedures.

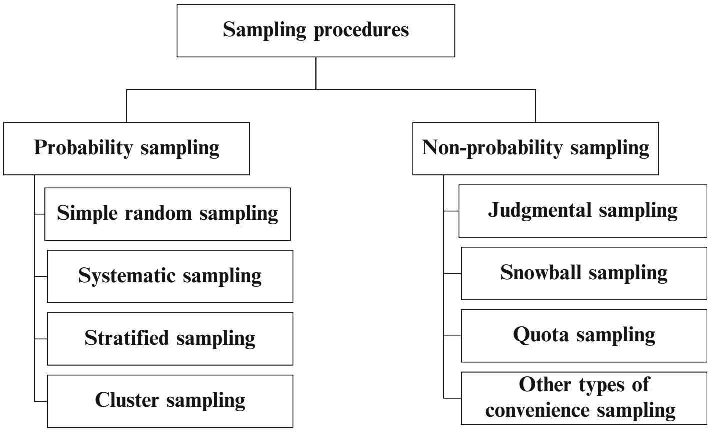

Fig. 3.2 Sampling procedures

Sample Sizes

After determining the sampling procedure, we have to determine the sample size. Larger sample sizes increase the precision of the research, but are also much more expensive to collect. Usually the gains in precision are very marginal if we increase the sample size far above 100–250 observations (in Box 6.3 we discuss the question whether a sample size can be too large in the context of significance testing). It may seem surprising that relatively small sample sizes are precise, but the strength of samples comes from accurately selecting samples, rather than through sample size. Furthermore, the required sample size has very little relation to the population size. That is, a sample of 100 employees from a company with 300 employees can be nearly as accurate as selecting 100 employees from a company with 100,000 employees. There are some problems in selecting sample sizes. The first is that market research companies often push their clients towards accepting large sample sizes. Since the fee for market research services is often directly dependent on the sample size, increasing the sample size increases the market research company's profit. Second, if we want to compare different groups (e.g., in a stratified sample design), we need to multiply the required sample by the number of groups included. That is, if 150 observations are sufficient to measure how much people spend on organic food, 2 times 150 observations are necessary to compare singles and couples' expenditure on organic food.

The figures mentioned above are net sample sizes; that is, these are the actual (usable) number of observations we should have. Owing to non-response (discussed in Chaps. 4 and 5), a multiple of the initial sample size (say, 400 or 800) is normally necessary to obtain the desired sample. Before collecting data, we should have an idea of the percentage of respondents we are likely to reach (often fairly high), a percentage estimate of the respondents willing to help (often low), as well as a percentage estimate of the respondents likely to fill out the survey correctly (often high). For example, if we expect to reach 80% of the identifiable respondents, and if 25% are likely to help, and 75% of those who help are likely to fully fill out the questionnaire, only 15% (0.80·0.25·0.75) of identifiable respondents are in this case likely to provide a usable response. Thus, if we wish to obtain a net sample size of 100, we need to send out $\left(\frac{\text{desired sample size}}{\text{likely usable responses}}\right) = 100/0.15 = 667$ surveys. That is a lot more than the sample size we are likely to use later on. These response issues have led many market researchers to believe that survey research may no longer be a good way to collect representative samples. In Chap. 4, we will discuss how we can increase response rates (the percentage of people willing to help).

Questions

1. Explain the difference between items and constructs.
2. What is the difference between reflective and formative constructs?
3. What are data? Give a few examples of quantitative and qualitative data.
4. What is the scale on which the following variables are measured?
   – The amount of money spent by a customer on shoes.
   – The country-of-origin of a product.
   – The number of times an individual makes a complaint.
   – The grades of a test.
   – The color of a mobile phone.

5. Try to find two websites that contain secondary data and also discuss what kind of data are described. Are these qualitative or quantitative data? What is the unit of analysis and how are the data measured?
6. What are “good data”?
7. Discuss the concepts of reliability and validity. How do they relate to each other?
8. Please comment on the following statement: “Face and content validity are essentially the same.”
9. What is the difference between predictive and criterion validity?
10. Imagine you have just been asked to execute a market research study to estimate the market for notebooks prices 300 USD or less. What sampling approach would you propose to the client?
11. Imagine that a university decides to evaluate their students’ satisfaction. To do so, employees issue a questionnaire to every tenth student coming to the student cafeteria on one weekday. Which type of sampling is conducted in this situation? Can the resulting sample be representative of the population?

Further Readings

Carmines EG, Zeller RA (1979) Reliability and validity assessment. Sage, Beverly Hills, CA.

Carmines and Zeller provide an in-depth discussion of different types of reliability and validity, including how to assess them.

Churchill GA (1979) A paradigm for developing better measures for marketing constructs. J Mark Res 16(1):64–73.

Probably the most widely cited marketing research article to this day. This paper marked the start of a rethinking process on how to adequately measure constructs.

Diamantopoulos A, Winklhofer HM (2001) Index construction with formative indicators: an alternative to scale development. J Mark Res 38(2):269–277.

In this seminal article the authors provide guidelines how to operationalize formative constructs.

Cochran WG (1977) Sampling techniques. Wiley series in probability and mathematical statistics, 3rd edn, Wiley, New York, NY.

This is a seminal text on sampling techniques which provides a thorough introduction into this topic. However, please note that most descriptions are rather technical and require a sound understanding of statistics.

DeVellis RF (2003) Scale development: theory and applications, 2nd edn. Sage, Thousand Oaks, CA.

This is a very accessible book which guides the reader through the classic way of developing multi-item scales. The text does not discuss how to operationalize formative constructs, though.

Marketing Scales Database at http://www.marketingscales.com/search/search.php

This website offers an easy-to-search database of marketing-related scales. For every scale a description is given, the scale origin and reliability and validity are discussed and the items are given.

Netemeyer RG, Bearden WO, Sharma S (2003) Scaling procedures: issues and applications, Sage, Thousand Oaks, CA.

Like DeVellis (2003), *this book presents an excellent introduction into the principles of scale development of measurement in general.*

References

Armstrong JS, Overton TS (1977) Estimating nonresponse bias in mail surveys. J Mark Res 14 (August):396–403

Bearden WO, Netemeyer RG (1999) Handbook of marketing scales. Multi-item measures for marketing and consumer behavior research, 2nd edn. Sage, Thousand Oaks, CA

Bergkvist L Rossiter JR (2007) The predictive validity of single-item versus multi-item measures of the same construct. J Mark Res 44(2):175–184.

DeVellis RF (2003) Scale development: theory and applications, 2nd edn. Sage, Thousand Oaks, CA

Diamantopoulos A, Riefler P, Roth KP (2008) Advancing formative measurement models. J Bus Res 61(12):1203–1218

Diamantopoulos, A, Winklhofer HM (2001) Index construction with formative indicators: an alternative to scale development. J Mark Res 38(2):269–277

Erdem T, Swait J (2004) Brand credibility, brand consideration, and choice. J Consum Res 31 (June):191–198

Fischer M, Völckner F, Sattler H (2010) How important are brands? A cross-category, cross-country study. J Mark Res 47(5):823–839

Fuchs C, Diamantopoulos A (2009) Using single-item measures for construct measurement in management research. Conceptual issues and application guidelines. Die Betriebswirtschaft 69 (2):195–210

Geyskens I, Steenkamp J-BEM, Kumar N (1998) Generalizations about trust in marketing channel relationships using meta-analysis. Int J Res Mark 15(July):223–248

Netemeyer RG, Bearden WO, Sharma S (2003) Scaling procedures: issues and applications. Sage, Thousand Oaks, CA

Rossiter JR (2010) Measurement for the social sciences. Springer, New York

Sarstedt M, Schloderer MP (2010) Developing a measurement approach for reputation of non-profit organizations. Int J Nonprofit Volunt Sect Mark 15(3):276–299

Sarstedt M, Wilczynski P (2009) More for less. A comparison of single-item and multi-item measures. Die Betriebswirtschaft 69(2):211–227