ONE

THE BASIC LANGUAGE OF STATISTICS

This chapter is an introduction to statistics and to quantitative methods. It explains the basic language used in statistics, the notion of a data file, the distinction between descriptive and inferential statistics, and the basic concepts of statistics and quantitative methods.

After studying this chapter, you should know:

- **the basic vocabulary of statistics and of quantitative methods;**
- **what an electronic data file looks like, and how to identify cases and variables;**
- **the different uses of the term 'statistics';**
- **the basic definition of descriptive and inferential statistics;**
- **the type of variables and of measurement scales;**
- **how concepts are operationalized with the help of indicators.**

Introduction: Social sciences and quantitative methods

Social sciences aim to study social phenomena – that is, human collective behavior, the culture that sustains it, the relationships between people that make it possible, and the organizations and institutions that regulate it – as rigorously as possible. This involves describing some aspect of social reality, analyzing it to see whether causal or explanatory links can be established between its various parts, and, whenever possible, predicting future outcomes, or at least a range of possible outcomes.

The general objective of such studies is to understand the patterns of individual or collective behavior, the constraints that affect it, the causes and explanations that can help us understand our societies and ourselves better and predict the consequences of certain situations. Such studies are never entirely objective, as they are inevitably based on certain assumptions and beliefs that cannot be demonstrated. Our perceptions of social phenomena are themselves subjective to a large extent, as they depend on the *meanings* we attribute to what we observe. Thus, we *interpret* social and human phenomena much more than we describe them, but we try to make that interpretation as objective as possible.

Some of the phenomena we observe can be *quantified*, which means that we can translate some aspects of our observations into numbers and make use of their properties. For instance, we can quantify population change: we can count how many babies are born every year in a given country, how many people die, and how many people migrate in or out of the country. Such figures allow us to estimate the present size of the population, and maybe even to predict how this size is going to change in a near future. We can quantify psychological phenomena such as the degree of stress or the rapidity of response to a stimulus; demographic phenomena such as population sizes or sex ratios (the ratio of men to women); geographic phenomena such as the average amount of rain over a year or over a month; economic phenomena such as the rate of employment; we can also quantify social phenomena such as the changing patterns of marriage or of unions, and so on.

When a social or human phenomenon is quantified in an appropriate way, we can ground our analysis of it on figures, or statistics. This allows us to describe the phenomenon with some accuracy, to establish whether there are links between some of the variables, and even to predict the evolution of the phenomenon. If the observations have been conducted on a sample (i.e. a group of people smaller than the whole population), we may even be able to generalize to the whole population what we have found on that sample.

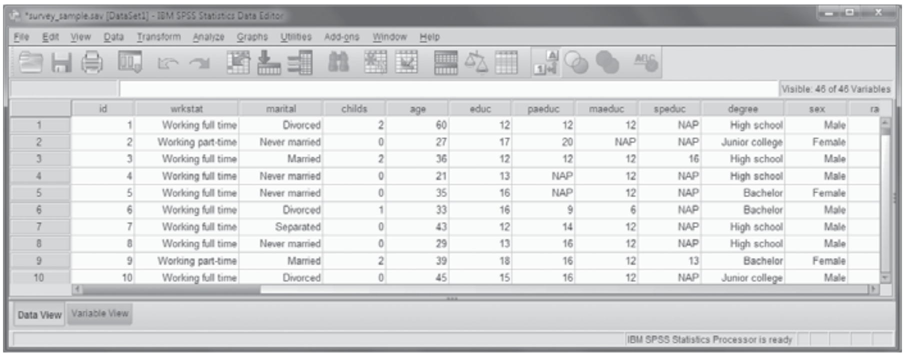
When we observe a social or human phenomenon in a systematic, scientific way, the information we gather about it is referred to as *data*. In other words, **data** is information that is collected in a systematic way, and organized and recorded in such a way that it can be interpreted correctly. Data is not collected haphazardly, but in response to some questions that the researchers would like to answer. Sometimes, we collect information (i.e. data) about a character or a quality that has no numerical value, such as the mother tongue of a person. Sometimes, the data is measurable with numbers, such as a person's age. In both cases, we can treat this data numerically: for instance, we can count how many people speak a certain language, or we can find the average age of a group of people. The procedures and techniques used to analyze data numerically are called *quantitative methods*. In other words, **quantitative methods** are procedures and techniques used to analyze data numerically; they include a study of the valid methods used for collecting data in the first place, as well as a discussion of the *limits of validity* of any given procedure (i.e. an understanding of the situations when a given procedure yields valid results), and of the ways the results are to be interpreted.

This book constitutes an introduction to quantitative methods for the social sciences. The present chapter covers the basic vocabulary of quantitative methods. This vocabulary should be mastered by the student if the remainder of the book is to be understood properly.

Data files

One of the first objects we deal with, in quantitative methods, is a **data file**. This is an electronic file that contains all the data, organized in a systematic way, often using numeric codes to refer to the various observations. When conducting research, we must distinguish between **primary data**, that is, data that is produced by the researcher or by a research agency, and **secondary data**, that is, data which is cited in an academic publication but which has been produced by some other researcher, or some agency, or which has been manipulated and summarized. The term **raw data** designates data that has not been subjected to any kind of statistical treatment, such as grouping, recoding, or selecting.

Figure 1.1 illustrates what an electronic data file looks like when we open it with the SPSS program. We can see in this figure the data window and the menu bar that appears on the top of your screen when you open SPSS.


*survey_sample.sav [DataSet1] - IBM SPSS Statistics Data Editor
File
Edit
View
Data
Transform
Analyze
Graphs
Utilities
Add-ons
Window
Help
Visible: 46 of 46 Variables
Data View
Variable View
IBM SPSS Statistics Processor is ready

	id	wrkstat	marital	childs	age	educ	paeduc	maeduc	speduc	degree	sex	ra
1	1	Working full time	Divorced	2	60	12	12	12	NAP	High school	Male	
2	2	Working part-time	Never married	0	27	17	20	NAP	NAP	Junior college	Female	
3	3	Working full time	Married	2	36	12	12	12	16	High school	Male	
4	4	Working full time	Never married	0	21	13	NAP	12	NAP	High school	Male	
5	5	Working full time	Never married	0	35	16	NAP	12	NAP	Bachelor	Female	
6	6	Working full time	Divorced	1	33	16	9	6	NAP	Bachelor	Male	
7	7	Working full time	Separated	0	43	12	14	12	NAP	High school	Male	
8	8	Working full time	Never married	0	29	13	16	12	NAP	High school	Male	
9	9	Working part-time	Married	2	39	18	16	12	13	Bachelor	Female	
10	10	Working full time	Divorced	0	45	15	16	12	NAP	Junior college	Male	

Figure 1.1 The Data View window in IBM SPSS 19 and the Menu bar (PC version)

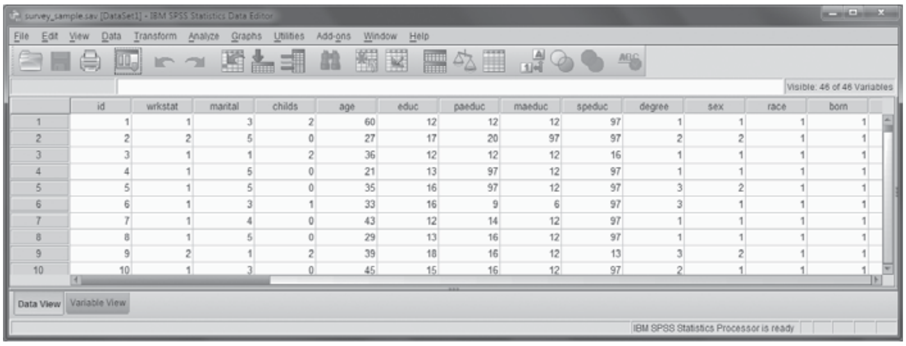
This data file was created by version 19 of the statistical software SPSS (Statistical Package for the Social Sciences), now called IBM SPSS. This software is available for both Windows and Mac. Exercise 1 at the end of this chapter will introduce you to SPSS, but we can look at some of its features right away.

On the top of the window, you can read the name of the data file **survey_sample.sav**.

When we open an SPSS data file, two views can be displayed: the Data View, or the Variable View. Both views are part of the same file, and one can switch from one view to the other by clicking on the tab at the bottom center of the window.

The **Data View** displays the data itself, and the information is organized in rows and columns. Each row refers to a **case**, that is, all the information pertaining to one individual. Each column refers to a **variable**, i.e. a character or quality that was observed and recorded using codes to refer to the various values of the variable. For instance, the second column is a variable called *wrkstat*, and the third is a variable called *marital*. Looking at the contents of each of these two columns, we see that the first variable concerns the work status, and the second the marital status of each individual.

It is important to understand that the data is not stored the way it appears in Figure 1.1. Rather it is stored in codes that take less memory space in a computer, making computations much faster. Thus, instead of recording 'Married', the program will just store a simple code, such as '1'. There is a way of showing the codes instead of the value labels. This is done by ticking off the command **Value Labels** under the **View** menu, in the data file. The Data View window now looks as shown in Figure 1.2.



survey_sample.sav [DataSet1] - IBM SPSS Statistics Data Editor

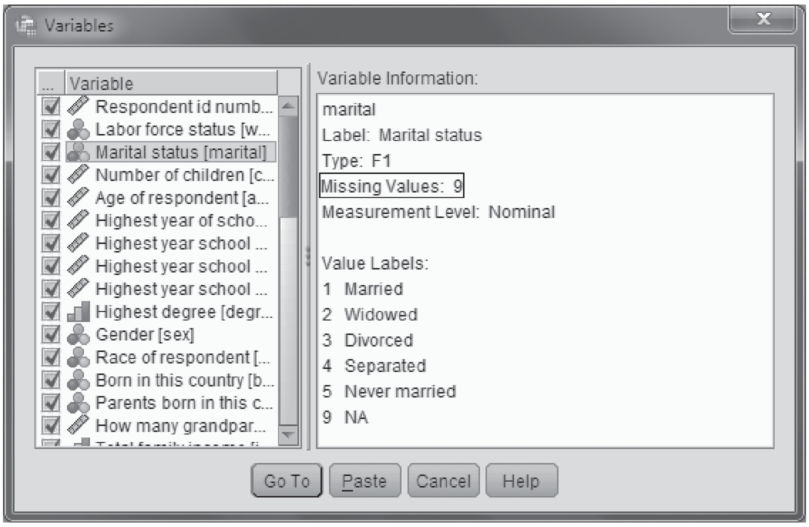
Visible: 46 of 46 Variables

	id	wrkstat	marital	childs	age	educ	paeduc	maeduc	speduc	degree	sex	race	born
1	1	1	3	2	60	12	12	12	97	1	1	1	1
2	2	2	5	0	27	17	20	97	97	2	2	1	1
3	3	1	1	2	36	12	12	12	16	1	1	1	1
4	4	1	5	0	21	13	97	12	97	1	1	1	1
5	5	1	5	0	35	16	97	12	97	3	2	1	1
6	6	1	3	1	33	16	9	6	97	3	1	1	1
7	7	1	4	0	43	12	14	12	97	1	1	1	1
8	8	1	5	0	29	13	16	12	97	1	1	1	1
9	9	2	1	2	39	18	16	12	13	3	2	1	1
10	10	1	3	0	45	15	16	12	97	2	1	1	1

IBM SPSS Statistics Processor is ready

Figure 1.2 The Data View window when the Value Labels command is ticked off the View menu

Here, the codes are displayed rather than the value labels. The meanings of these codes can be seen by clicking on **Variables...** under the **Utilities** menu. The resulting window is shown in Figure 1.3.



Variables

- ☑ Respondent id numb...
- ☑ Labor force status [w...
- ☑ Marital status [marital]
- ☑ Number of children [c...
- ☑ Age of respondent [a...
- ☑ Highest year of scho...
- ☑ Highest year school ...
- ☑ Highest year school ...
- ☑ Highest year school ...
- ☑ Highest degree [degr...
- ☑ Gender [sex]
- ☑ Race of respondent [...
- ☑ Born in this country [b...
- ☑ Parents born in this c...
- ☑ How many grandpar...

Variable Information:

marital
Label: Marital status
Type: F1
Missing Values: 9
Measurement Level: Nominal

Value Labels:

1	Married
2	Widowed
3	Divorced
4	Separated
5	Never married
9	NA

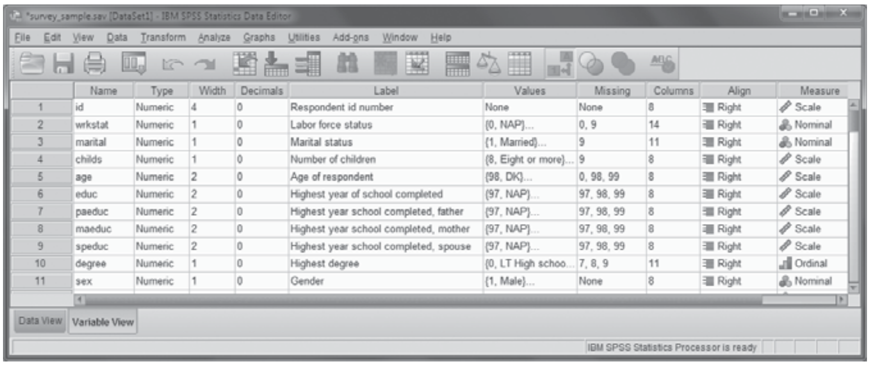
Figure 1.3 The Variable and Value labels for the variable *marital*

You can see here that the variable is designated as 'marital', that its full label is Marital status, and the various codes are also given:

1	stands for	Married
2	stands for	Widowed
3	stands for	Divorced
etc.		

This information is part of the **codebook**, which includes a complete listing of all the variables, their labels, their values (the codes) and their value labels. In SPSS, the codebook is referred to as the **data file information**. You can see it by selecting **File → Display Data File Information**. You will get a new window, called the Output window, that contains all the information about the variables, and all the codes and their meanings.

The information about the variables can also be seen by clicking on the **Variable View** tab at the bottom center of the data window. You then get the window shown in Figure 1.4.


*survey_sample.sav [DataSet1] - IBM SPSS Statistics Data Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
Data View
Variable View
IBM SPSS Statistics Processor is ready

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	id	Numeric	4	0	Respondent id number	None	None	8	Right	Scale
2	wrkstat	Numeric	1	0	Labor force status	{0, NAP}...	0, 9	14	Right	Nominal
3	marital	Numeric	1	0	Marital status	{1, Married}...	9	11	Right	Nominal
4	childs	Numeric	1	0	Number of children	{8, Eight or more}...	9	8	Right	Scale
5	age	Numeric	2	0	Age of respondent	{98, DK}...	0, 98, 99	8	Right	Scale
6	educ	Numeric	2	0	Highest year of school completed	{97, NAP}...	97, 98, 99	8	Right	Scale
7	paeduc	Numeric	2	0	Highest year school completed, father	{97, NAP}...	97, 98, 99	8	Right	Scale
8	maeduc	Numeric	2	0	Highest year school completed, mother	{97, NAP}...	97, 98, 99	8	Right	Scale
9	speduc	Numeric	2	0	Highest year school completed, spouse	{97, NAP}...	97, 98, 99	8	Right	Scale
10	degree	Numeric	1	0	Highest degree	{0, LT High schoo...	7, 8, 9	11	Right	Ordinal
11	sex	Numeric	1	0	Gender	{1, Male}...	None	8	Right	Nominal

Figure 1.4 The Variable View window

Here every line represents a variable and provides some of its features:

Name This is the brief name of the variable. It must be short, with no spaces. Variable names that are defined by users must begin with a letter. The detailed rules for forming variable names can be found in the SPSS Help menu.

Type This column specifies how the variable is written: either a numeric code, or a string of characters, or a date or a currency, or a special character.

Width This column specifies how much space is devoted to each variable, i.e. the maximum number of characters allowed for writing down the observations relating to the variable. If the variable is Sex and the codes are 1: Male; 2: Female, only one space is needed because only the codes are recorded in the SPSS data file. The codebook tells us what these codes stand for.

Decimals We need to specify whether the numerical values in the data file are decimal numbers or not, and with how many decimals. This is indicated in this column.

Labels These are the long names of the variables. These names will appear in the tables produced in the SPSS output. They must be explicit enough to allow a correct reading of the tables, but preferably succinct. If the data file comes from a survey, the labels will be the questions of the survey, maybe in a shorter form.

Values This column indicates what the codes used in the data file stand for.

Missing Indicates which of the codes must be treated as missing. Missing values are not taken into account when performing computations (such as the mean, or the valid frequencies, as will be seen in subsequent chapters).

Columns The numbers shown in this column refer to the visual aspect of the data file, i.e. its appearance: they indicate the width of the columns in the Data View.

Align This also refers to the appearance of the data file, indicating the alignment of the data that is displayed: left, center or right.

Measure This column indicates whether the codes used for the variable are to be taken as codes that do not indicate size or magnitude (nominal, such as the codes 1 and 2 for the variable Sex), or numbers that indicate a rank (ordinal) or else as numerical values that indicate a magnitude (scale variables, such as age, length, duration, weight, or a score on some numerical scale).

Now that we have seen what a data file looks like and how the data is stored, we can raise a number of questions: How did we come up with this data? What are the rules for obtaining reliable data that can be interpreted easily? How can we analyze this data? Table 1.1 includes a systematic list of such questions. The answers to these questions will be found in the various chapters and sections of this manual.

Table 1.1 Questions arising from the use of quantitative methods

Questions	Chapters
How did we come up with a given data set? What are the questions we are trying to answer? What is the place of quantitative analysis in social research, and how does it link up with the qualitative questions we may want to ask? What is the scientific way of defining concepts and operationalizing them? What are independent and dependent variables?	1. The Basic Language of Statistics
How do we conduct social research in a scientific way? What procedures should we follow to ensure that results are scientific? What are the basic types of research designs? How do we go about collecting the data?	2. The Research Process
Once collected, the data must be organized and described. How do we do that? When we summarize the data, what are the characteristics that we focus on? What kind of information is lost?	3. Univariate Descriptive Statistics
What are the most common types of shapes and distributions we encounter? How do we select the appropriate graphical representation of the data pertaining to a variable?	4. Graphical Representations

Questions	Chapters
Once variables are entered in a data file, can we recode them in a way that is more adequate for our analysis? How do we create new ones? How do we regroup the categories? Can we sort them? etc.	5. Creating New Variables with SPSS
What are the fundamental properties of samples that allow us to build inference procedures? This chapter is a technical one, and it is crucial for understanding the logic of inference.	6. Normal Distributions and Sampling Distributions
What are the procedures for selecting a sample? Are some of them better than others? How do we ensure that our sample is representative and that we can draw general conclusions from it?	7. Sampling Designs
When the data comes from a sample, under what conditions can we generalize our conclusions to the whole population? How can this be done? Is it precise? What are the risks that our conclusions are wrong?	8. Estimation 9. Hypothesis Testing
Sometimes we notice coincidences in the data: for instance, those who have a higher income tend to behave differently on some social variables than those who do not. Is there a way of describing such relationships between variables, and drawing their significance?	10. Correlation and the Regression Line 11. Two-way Tables and the Chi-squared Test 12. *t*-tests and ANOVA
Finally, when we have finished our statistical analyses, how do we report them? What constitutes good practice?	Appendix: Reporting a Quantitative Analysis

The discipline of statistics

The term *statistics* is used in two different senses: it can refer to the *discipline* of statistics, or it can refer to the *actual data* that has been collected.

As a scientific discipline, the object of **statistics** is the numerical treatment of data pertaining to a large quantity of individuals or a large quantity of objects. It includes a general, theoretical aspect which is based on the mathematical study of probability, but it can also include the study of the concrete problems that are raised when we apply the theoretical methods to specific disciplines. The term **quantitative methods** is used to refer to methods and techniques of statistics which are applied to concrete problems. Thus, the difference between statistics and quantitative methods is that the latter include practical concerns such as finding solutions to the problems arising from the collection of real data, and interpreting the numerical results as they relate to concrete situations. For instance, proving that the mean (or average) of a set of values has certain mathematical properties is part of statistics. Deciding that the mean, rather than, say, the median, is an appropriate measure to use in a given situation is part of quantitative methods. But the line between statistics and quantitative methods is fuzzy, and the two terms are sometimes used interchangeably. In practice, the term ‘statistics’ is often used to mean quantitative methods, and we will use it in that way too.

The term **statistics** also has a different meaning, and it is used to refer to the actual data that has been obtained by statistical methods. Thus, we will say, for instance, that the latest statistics published by the Ministry of Labor indicate a decrease in unemployment. In that last sentence, the word *statistics* was used to refer to data published by the Ministry.

Populations, samples, and units

Three basic terms must be defined to explain the subject matter of the discipline of statistics:

- unit (or element, or case),
- population, and
- sample.

A **unit** (sometimes called an **element** or **case**) is the smallest object of study. If we are conducting a study on individuals, a unit is an individual. If our study is about the health system (we may want to know, for instance, whether certain hospitals are more efficient than others), a unit for such a study would be a hospital, not a person.

A **population** is the collection of all units that we wish to consider. If our study is about the hospitals in the UK, the population will consist of all hospitals in the UK. Sometimes, the term **universe** is used to refer to the set of all individuals under consideration, but we will not use it in this manual.

Most of the time, we cannot afford to study each and every unit in a population, due to the impossibility of doing so or to considerations of time and cost. In this case, we study a smaller group of units, called a **sample**. Thus, a sample is *any* subset (or subgroup) of our population.

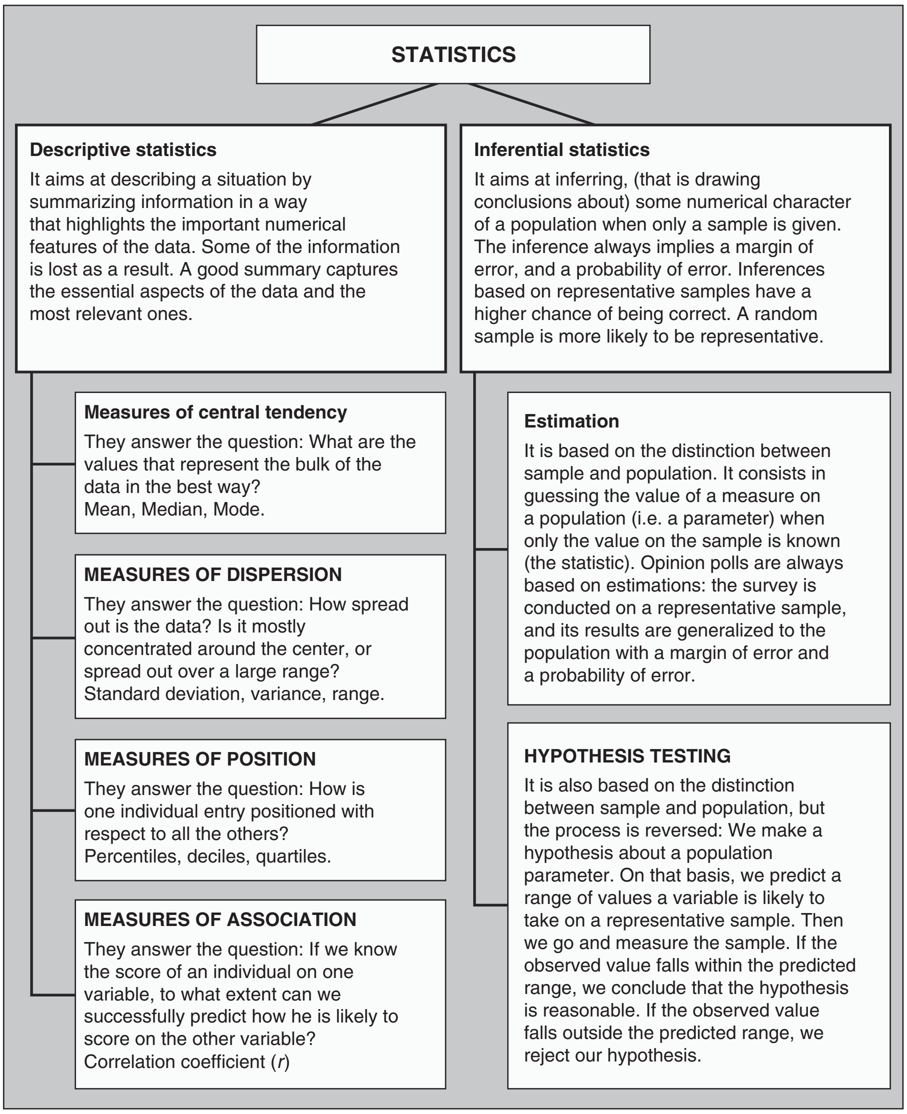
The distinction between sample and population is absolutely fundamental. Whenever you are doing a computation, or making any statement, it must be clear in your mind whether you are talking about a sample (a group of units generally smaller than the population) or about the whole population.

The discipline of statistics includes two main branches: descriptive statistics and inferential statistics. These two sets of procedures and techniques will be explained throughout the book, but for the time being we can only define them broadly:

Descriptive statistics aims to describe a situation by summarizing information in a way that highlights the important numerical features of the data. Some of the information is lost as a result.

Inferential statistics aims to infer (i.e. draw conclusions about) some numerical character of a population when only a sample is given.

Thus, **statistical inference** is a form of reasoning that aims to generalize a statistical result obtained on sample data, to draw a conclusion about a population parameter. Statistical inference usually involves a margin of error and a probability of error. The following sections explain what each branch is about; refer also to Figure 1.5. Some of the terms used in Figure 1.5 may not be clear for now, but they will be explained as we progress.



STATISTICS

Descriptive statistics

It aims at describing a situation by summarizing information in a way that highlights the important numerical features of the data. Some of the information is lost as a result. A good summary captures the essential aspects of the data and the most relevant ones.

Measures of central tendency

They answer the question: What are the values that represent the bulk of the data in the best way?
Mean, Median, Mode.

MEASURES OF DISPERSION

They answer the question: How spread out is the data? Is it mostly concentrated around the center, or spread out over a large range?
Standard deviation, variance, range.

MEASURES OF POSITION

They answer the question: How is one individual entry positioned with respect to all the others?
Percentiles, deciles, quartiles.

MEASURES OF ASSOCIATION

They answer the question: If we know the score of an individual on one variable, to what extent can we successfully predict how he is likely to score on the other variable?
Correlation coefficient (*r*)

Inferential statistics

It aims at inferring, (that is drawing conclusions about) some numerical character of a population when only a sample is given. The inference always implies a margin of error, and a probability of error. Inferences based on representative samples have a higher chance of being correct. A random sample is more likely to be representative.

Estimation

It is based on the distinction between sample and population. It consists in guessing the value of a measure on a population (i.e. a parameter) when only the value on the sample is known (the statistic). Opinion polls are always based on estimations: the survey is conducted on a representative sample, and its results are generalized to the population with a margin of error and a probability of error.

HYPOTHESIS TESTING

It is also based on the distinction between sample and population, but the process is reversed: We make a hypothesis about a population parameter. On that basis, we predict a range of values a variable is likely to take on a representative sample. Then we go and measure the sample. If the observed value falls within the predicted range, we conclude that the hypothesis is reasonable. If the observed value falls outside the predicted range, we reject our hypothesis.

Figure 1.5 Descriptive and inferential statistics

Descriptive statistics

The methods and techniques of descriptive statistics aim to summarize large quantities of data by a limited number of numerical values, in a way that highlights the most important features of the data. For instance, if you say that your grade point average (GPA) in secondary schooling is 3.62, you are giving only one number that gives a pretty good idea of your performance during all your secondary schooling. If you also say that the *standard deviation* (this term will be explained later on) of your grades is 0.02, you are saying that your marks are very consistent across the various courses. A standard deviation of 0.1 would indicate a variability that is 5 times bigger, as we will learn later on. You do not need to give the detailed list of your marks in every exam of every course: the GPA is a sufficient measure in many circumstances. However, an average can sometimes be misleading. When is the average misleading? Can we complement it by other measures that would help us have a better idea of the features of the data we are summarizing? Such questions are part of descriptive statistics.

Descriptive statistics include measures of central tendency, measures of dispersion, measures of position, and measures of association. They also include a description of the general shape of the distribution of the data. These terms will be explained in Chapter 3.

Inferential statistics

Inferential statistics aims to generalize a measure taken on a small number of cases that have been observed, to a larger set of cases that have not been observed. Using the terms explained above, we could reformulate this aim, and say that inferential statistics aims to generalize observations made on a sample to a whole population. For instance, when pre-election polls are conducted, only 1000–2000 individuals are questioned, and on the basis of their answers the polling agency draws conclusions about the voting intentions of the whole population. Such conclusions are not very precise, and there is always a risk that they are completely wrong. More importantly, the sample used to draw such conclusions must be a *representative sample*, that is, a sample in which all the relevant qualities of the population are adequately represented (more detailed definitions will be given in Chapter 7 on sampling). How can we ensure that a sample is representative? Well, we cannot. We can only increase our chances of selecting a representative sample if we select it randomly. We will devote a chapter to sampling methods.

Inferential statistics includes estimation and hypothesis testing, two techniques that will be studied in Chapters 8 and 9.

A few more terms must be defined in order to be able to go further in our study. We need to talk a little about variables and their types.

Variables and measurement

A **variable** is a character or quality that is observed, measured, and recorded in a data file (generally, in a single column). If you need to keep track of the country of birth of the individuals in your population, you will include in your study a variable called *Country of birth*. You may also want to keep track of the nationality of the individuals: you will then have another variable called *Nationality*. The two variables are distinct, since some people may carry the nationality of a country other than the one they were born in. Here are some examples of variables used widely in social sciences:

Socio-demographic variables

- Age
- Sex
- Religion
- Level of education
- Highest degree obtained
- Marital status
- Country of birth
- Nationality
- Mother tongue
- Percentage of people under 30 (the unit here is a population, or perhaps a territory)
- Life expectancy (in a specific population)

Psychological variables

- Level of anxiety
- Stimulus response time
- Score obtained in a personality test
- Score obtained in an aptitude test
- IQ (intelligence quotient, a very controversial variable)

Economic variables

- Working status
- Income
- Value of individual assets
- Average number of hours of work per week
- Gross domestic product (GDP) of a country
- Total monthly sales (of an entreprise)

Variables that refer to units other than the individual

- Number of hospitals in a country
- Percentage of people who can read
- Percentage of people who completed high school
- Total population
- Birth rate
- Fertility rate

- Number of teachers per 1000 people
- Number of doctors per 10,000 people
- Population growth
- Predominant religion
- Unemployment rate

You may have noticed that some of these variables refer to qualities (such as mother tongue) and others refer to quantities, such as the total population of a country. In fact, we can distinguish two basic **types of variables**:

- quantitative variables and
- qualitative variables.

Quantitative variables are characters or features that are best expressed by numerical values, such as the age of a person, the number of people in a household, the size of a building, or the annual sales of a product. Qualitative variables are characters or qualities that are not numerical, such as mother tongue or country of origin. The scores of the individuals of a population on the various variables are called the **values** of that variable.

Example 1.1

Suppose you have the information shown in Table 1.2, concerning five students in your college.

Table 1.2 Examples of qualitative and quantitative variables

Name	Age	Program of Study	Grade Point Average
John	19	Social Science	3.78
Mary	17	Pure and Applied Science	3.89
Peter	18	Commerce	3.67
Colette	19	Office Systems Technology	3.90
Suzie	20	Graphic Design	3.82

There are three variables: *Age* (quantitative), *Program of Study* (qualitative) and *Grade Point Average* (quantitative).

The **values**, or **scores**, taken by the individuals for the variable *Age* are 17, 18, 19 (twice), and 20. The values taken for the variable *Program of Study* are Social Science, Pure and Applied Science, Commerce, Office Systems Technology, and Graphic Design. Qualitative variables are sometimes referred to as **categorical** variables because they consist of categories in which the population can be classified. For instance, we can classify all students in a college into categories according to the program of study they are in.

Careful attention must be given to the way observations pertaining to a variable are *recorded*. We must find a system for recording the data that is very clear, and that can be interpreted without any ambiguity. Consider, for instance, the following characters: age, rank in the family, and mother tongue. The first character is a quantity, the second is a rank, and the third is a quality. The systems used to record our observations about these characters will be organized into three **levels of measurement**:

- measurement at the **nominal** level;
- measurement at the **ordinal** level; and
- measurement at the **numerical scale** level.

Each level of measurement allows us to perform certain statistical operations, and not others.

The **nominal level of measurement** is used to measure **qualitative** variables. It is the simplest system for writing down our observations: when we want to measure a character at the nominal level, we establish a number of categories in such a way that each observation falls into one and only one of these categories. For example, if you want to write down your observations about mother tongue in the Canadian context, you may have the following categories:

- English,
- French,
- Native, and
- Other.

Depending on the subject of your research, you may have more categories to include other languages, or you may want to make a provision for those who have two mother tongues.

It is important to note that when a variable is measured at the nominal level, the categories must be

- exhaustive, and
- mutually exclusive.

The categories are said to be **exhaustive** when they include the whole range of possible observations, that is, they exhaust all the possibilities. This means that every one of the observations can fit in one of the available categories. The categories are said to be **mutually exclusive** if they are not overlapping: every observation fits in only one category. These two properties ensure that the system used to write down the observations is clear and complete, and that there are no ambiguities when recording the observations or when reading the data file. Table 1.3 displays examples of measurements made at the nominal level.

Table 1.3 Examples of variables measured at the nominal level

Variable	Categories used
Sex	Male Female
Place of birth	The country where the survey is conducted Abroad
Work status	Working full-time Working part-time Temporarily out of work Unemployed Retired Homemaker Other

The **ordinal level of measurement** is used when the observations are organized in categories that are *ranked* or *ordered*. We can say that one category precedes another, but we cannot say by how much exactly (or if we can, we do not keep that information). Here too the categories must be exhaustive and mutually exclusive, but in addition you must be able to compare any two categories, and say which one precedes the other (or is bigger, or better, etc.). Table 1.4 displays examples of variables measured at the ordinal level.

Table 1.4 Examples of variables measured at the ordinal level

Variable	Ranked Categories
Rating of a restaurant	Excellent Very good Acceptable Poor Very poor
Rank among siblings	First child Second child etc.
Income	High Medium Low

The scale used to write down an ordinal variable is often referred to as a **Likert scale**. It usually has a limited number of ranked categories: anywhere from three to seven categories, sometimes more. For instance, if people are asked to rate a service as:

❑ Excellent

❑ Very good

❑ Good

❑ Poor

❑ Very poor,

the proposed answers constitute a five-level Likert scale.

Another example of Likert scale, this time with four levels, is provided by the situations where a statement is given, and respondents are asked to say whether they:

❑ Strongly agree

❑ Agree

❑ Disagree

❑ Strongly disagree.

When conducting a survey, a delicate choice must be made between a Likert scale with an even number of categories and an odd number. You choose an even number of categories (e.g. four) when you want to force the respondents to choose between agreement and disagreement, rather than standing on neutral middle ground, which some people may want to do. When you have an even number of categories, you have the choice of grouping them into two categories, thus recoding your variable into a dichotomous variable (i.e. a variable having exactly two categories). This may be desirable in some instances. Some of the statistical techniques we will see later only work for dichotomous variables. Even simple tables are easier to understand and interpret when there are only two categories. But there is no correct answer for this issue: it is a matter of judgment, which depends on the analysis you want to perform with your data.

A variable measured at the ordinal level could be either qualitative or quantitative. In Table 1.4, the variable *Income* is quantitative, and the variable *Rating of the Restaurant* is qualitative, but they are both measured at the ordinal level. For a variable measured at the ordinal level, we can say that one value precedes another, but we cannot give an exact numerical value for the difference between them. For instance, if we know that a respondent is the first child and the other is the second child in the same family, we do not keep track of the age difference between them. It could be 1 year in one case and 5 years in another case, but the values recorded under this variable do not give us this information: they only give us the rank.

Finally, some variables are measured by a **numerical scale**, a system where the numerical codes used denote a magnitude (such as a length or a weight). Every observation is measured against the scale and assigned a numerical value, which measures a quantity. These variables are said to be **quantitative**. Table 1.5 displays examples of numerical scale variables.

Notice that the same variable can be measured by different scales, as shown in Table 1.5. So, when we use a numerical scale, we must determine the units used (e.g. years or months), and the number of decimals used.

Table 1.5 Examples of variables measured at the numerical scale level

Variable	Numerical scale
Annual income	In dollars, without decimals (no cents)
Annual income	In dollars, to the nearest thousand
Age	In years, with no fractions
Age	In years, with one decimal for fractions of a year
Temperature	In degrees Celsius
Time	In years – a starting point must be specified

Numerical scales are sometimes subdivided into **interval scales** and **ratio scales**, depending on whether there is an absolute zero to the scale or not. Thus, *temperature* (when measured in degrees Celsius) and *time* are measured by interval scales, whereas *age* or *number of children* are each measured by a ratio scale. Temperature is a ratio scale when using the Kelvin scale, which has an absolute zero. However, this distinction will not be relevant for most of what we are doing in this course, and we will simply use the term **numerical scale** to talk about this level of measurement. The program SPSS that we are going to use simply uses the term **scale** to refer to such variables.

Most statistical software packages include more specific ways of writing down the observations pertaining to a numerical scale. For instance, SPSS will offer the possibility of specifying that the variable is a currency, or a date.

Moreover, it is also possible to group the values of a quantitative variable into **classes**. Thus, when observing the variable *age*, we can write down the exact age of a person in years, or we can simply write the age group the person falls in, as is done in the following example:

- 18 to 30 years
- 31 to 40 years
- 41 to 50 years
- 50 to 60 years
- over 60.

When we group a variable such as *age* into a small number of categories as we have just done, we cannot perform the same statistical operations as we can when it is ungrouped. For instance, the mean, or average of the variable *age* is best calculated when the ages are *not* grouped. When we group the values, it is because we want to know the relative importance (i.e. the frequency, as a percentage) of one group as compared to another. The information that 50% of the population is under 20 years old in some developing countries is obtained by grouping the ages into 20 *years old or younger* and *over 20 years old*. When we collect the data, it is always better to collect it in actual years, since we can easily group it later on in the data file with the help of a statistical software package. In this case, a new column is added to the data file, and it contains the grouped data of the quantitative variable. For example, in the *Survey_sample* data file that we will use in the exercises, you will find two variables for *age*: one is called *age*, and the other one is called *agecat*. The latter is calculated from the former, by grouping individuals into six age groups. In the column of *agecat*, the specific age of an individual is not recorded: only the age group of the individual is recorded.

When grouping a numerical variable into a limited number of categories, you must decide whether the categories should all be equal in range or not. This depends on the use of the results. In planning municipal services, for instance, you may need to know how many people are minors (below 18), how many are adults, and how many are over 60 or 65. Here the categories will not be equal in range.

Finally, numerical scales can be either *continuous* or *discrete*. A scale is said to be **continuous** if the observations can theoretically take any value over a certain range, including fractions of a unit. For instance, age, weight, length, are continuous variables because they are not limited to specific values, and they can take any value within a certain range. A numerical scale is said to be **discrete** if it includes a limited number of possible values, but not values in between. For instance, the variable *Number of children* is measured by a discrete scale because it can only be equal to a whole number: 0, 1, 2, etc.

Importance of the level of measurement

The level of measurement used for a variable depends on whether it is qualitative or quantitative.

Qualitative variables must be measured at the nominal level or at the ordinal level. They cannot be measured at the numerical scale level, even when their categories are coded with numbers. For instance, we can code the variable *Sex* as follows:

1 Male

2 Female.

In this case, *the numbers 1 and 2 are not used for their numerical value*. They are simply codes. It is shorter to write 1 than *Male*, and we could have assigned the numbers differently. If you ask SPSS to compute the mean (or average) for a variable coded in this way, you *will* get a numerical answer. But you must always keep in mind that such a numerical answer is *totally meaningless* because the level of measurement of that variable is nominal. The numbers used to record the information are simply codes.

Quantitative variables are usually measured by a numerical scale, but they could also be measured at the ordinal level. For instance, if you have the annual income of an individual, you may treat it as a numerical scale, but you could also group the values into low, medium and high income and treat the variable at the ordinal level.

When you perform a statistical analysis of data, it is very important to pay attention to the level of measurement of each variable. Some statistical computations are appropriate only to a given level of measurement, and should not be performed if the variable is measured at a different level.

Concepts, dimensions, and indicators

We often want to observe social phenomena that are too abstract and complex to be expressed by a single variable. Suppose, for instance, that we want to observe and measure the degree of *religious inclination* (or the tendency of a person toward religion) in a given social group. Religious inclination can be manifested in many ways: people may have or not have certain *beliefs* about their religion; they may also perform or not perform certain *rituals* such as attending religious services, fasting, and praying; they may also *seek the advice of the religious leadership* on important decisions, or ignore such leadership; finally, they may seek to look at everything from the point of view of religion, and *apply the teachings* of their religion in their daily lives, or ignore them. All these aspects are not found all the time in all individuals. Some individuals may have strong beliefs, while avoiding the religious services. Others may attend all services while being skeptical about some of the religious dogma. The way to handle this complexity is to subdivide the concept of *religious inclination* into dimensions, which are themselves measured by several indicators. If we were to study religious inclination in the Catholic religion, we would get a set of dimensions and indicators that would look as in Table 1.6 (we are simplifying the issues a little, of course).

Table 1.6 An example of breaking down a concept into dimensions and indicators

Concept	Dimensions	Indicators
RELIGIOUS INCLINATION	I. Beliefs	Belief in God Belief in the Holy Trinity Belief in the main dogma etc.
	II. Rituals	Attending services Performing prayers Baptizing children etc.
	III. Guidance	Consulting the priest about important decisions Consulting the official opinions of the church on certain issues such as birth control etc.
	IV. Daily life	Being kind and generous to people Not cheating others in commercial transactions etc.

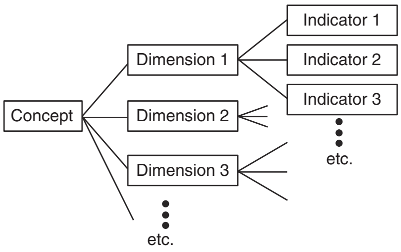


Figure 1.6 The general pattern of concepts, dimensions and indicators

The items listed on the right-hand side of the table are **indicators** of the concept of *religious inclination*. None of them, taken alone, is a measure of religious inclination, but each of them constitutes *one* aspect of it. Indicators that are seen as similar are grouped together to form one *dimension* of the concept. Finally, the various dimensions, taken together, capture the concept as a whole. This way of breaking down a complex concept into dimensions and indicators is called the **operationalization** of the concept. As an illustration, we may want to see how economists operationalize the concept of *cost of living*. They estimate the average cost of most of the standard expenses a family of four is expected to incur. The various expenses are divided into main dimensions such as: food, housing, transportation, education, and leisure. Each dimension is then subdivided into smaller dimensions, themselves subdivided further into indicators. For instance, food is broken down as meat, vegetables, milk products, etc., and each category subdivided into specific items such as tomatoes, lettuce, etc. Finally, for each of these indicators, the increase or decrease in the cost of living is measured against the corresponding cost in some year, called the base year. By combining these indicators, economists are able to measure how the cost of living has changed, on average, for a family of four. The general pattern of concepts, dimensions and indicators is represented in Figure 1.6.

The way a concept is broken down, or operationalized, into dimensions and indicators depends on the theoretical framework adopted for a study. Researchers may not agree on how to operationalize a concept, and you will find in the literature different studies that operationalize concepts in completely different ways, because they rely on different theoretical frameworks. It then becomes necessary to compare the various definitions of the concept, to examine how indicators are defined and which of them are taken into account. Exercise 5 at the end of the chapter discusses one such example.

Validity and reliability

When we define a variable in order to measure some concept, the variable must satisfy two important criteria: it must be a valid measurement of the concept, and it must be reliable.

The **validity** means that the variable must indeed measure adequately what it is supposed to measure. A simple illustration can be given by the concept of *strength* of an individual. If you weigh the individual and use his or her weight as an indication of strength, you will not get a valid measure. Obviously, people can be thin, yet strong, or fat, yet weak. In the domain of social sciences, the validity of a variable is not easy to establish, and sometimes some apparently reasonable proposals turn out to lack validity totally. When psychologists tried to measure intelligence in the nineteenth century, they used the size of the brain as a measure of intelligence. They developed sophisticated metallic instruments to measure the volume of the skull of individuals, including special devices to measure the height of every bump on it, making two assumptions: that the size of the skull is directly linked to the volume of the brain, and that the volume of the brain is a direct measure of intelligence. They soon realized that cows had larger brains than humans, and concluded that the size of the brain is not a good measure. They then modified the definition to consider the ratio of the weight of the brain to the weight of the body, but that too did not turn out to be a valid measure. Eventually they came up with a series of tests that measure certain intellectual skills, such as our command of vocabulary, the associations we can make between different geometric shapes, etc. On this basis, they came up with a measure called *intelligence quotient*, or IQ, that is still used to orient students in various study programs. The IQ is designed to have an average of 100, and a standard deviation of 16. This measure is very controversial. Its use resulted in the fact that students belonging to certain cultural subgroups, or to low-income milieus, were prevented from accessing some prestigious programs of study, which means that it has become a tool of discrimination. It turned out that this measure is culture-specific, in the sense that getting a high score on the IQ depends on your familiarity with a specific subculture: people from poor neighborhoods would have low scores not because they are less intelligent, but because they use words in a different way. When adjusted to specific subcultures, the IQ measures certain intellectual abilities, and can predict academic success, but certainly not intelligence. This is an example of a variable whose validity is highly questionable.

The other essential criterion is that of **reliability**. It is the quality of a measure that is consistent, which means that it gives consistent results when used with repetition on the same subject and in similar circumstances. An illustration of a measure that is not reliable is that of a stretchable tape to measure a length: it would give different results depending on how much it is stretched. For social variables, reliability of a variable means that the measure obtained is stable across the range of conditions in which it is used. Reliability means two things: *consistency* and *replicability*. Consistency means that if you observe a social phenomenon with a given method, you get similar answers when you use it in similar circumstances. Replicability means that if some other researchers repeat the measure in different but similar circumstances, they would get roughly the same answers. In the example of the IQ above, the physical measurements of the size of the skull and brain are very reliable, but they are not valid.

Summary

Quantitative methods are procedures and techniques for collecting, organizing, describing, analyzing, and interpreting data. In this chapter we have learned the basic vocabulary used to talk about quantitative methods. Data is organized into electronic data files with the help of statistical packages. A data file contains the values taken by a number of cases (which are the units of the population under study) over some variables. Every row represents a case, while every column represents a variable. The units in the data file usually form a sample, which is itself a subset of the whole population. Sometimes, the data file refers to the whole population.

The variables can be either qualitative or quantitative. The system used to record the information is called a measurement scale. There are three levels of measurement: nominal, ordinal and numerical (interval or ratio). The level of measurement of a variable will determine what statistical procedures can be performed, and what kind of graphs must be used to illustrate the data. When a concept is complex, it is not measured directly. It is usually broken down into dimensions and indicators, which are then combined to provide a single measure.

The statistical procedures themselves fall into two broad categories: descriptive statistics and inferential statistics. Descriptive statistical techniques aim to describe the data by summarizing it, while inferential statistical techniques aim to generalize to a whole population what has been observed on a sample.

Key words

After studying this chapter, you should be able to define and explain *all* the following terms.

- Data
- Data file
- Primary data
- Secondary data
- Raw data
- Quantitative methods
- Variable
- Variable label
- Value
- Value label
- Variable type
- Quantitative variable
- Qualitative variable
- Case
- Unit
- Sample
- Population
- Level of measurement
- Nominal level
- Ordinal level
- Numerical level (interval or ratio)
- Exhaustive categories
- Mutually exclusive categories
- Likert scale
- Continuous numerical scales
- Discrete numerical scales
- Codebook
- Statistics (the two meanings)
- Descriptive statistics
- Inferential statistics
- Statistical inference
- Validity
- Reliability
- Consistency
- Replicability
- Dimensions of a concept
- Indicators of a concept
- Operationalization of a concept

Exercises

The aim of the following exercises is for you to familiarize yourself with the SPSS windows. You will also need to go through the SPSS tutorial that follows these exercises.

1. Consider the windows shown in Figures 1.1–1.4. The following questions can be answered by inspecting one or other of these windows.

   (a) Give the names of the first three variables and their corresponding labels.

   (b) What are the value labels that are visible in Figure 1.1 for the variable *marital*?

   (c) What are the value labels that you can see for the variable *degree*?

   (d) For each variable shown, say whether it is qualitative or quantitative, and identify the correct level of measurement to be used.

   (e) Here is a description of case 1 based on what we can see in Figures 1.1 and 1.4 (the latter tells us about the variable labels):

   The first individual is a 60-years old man, classified as 'white', who has two children and who is divorced. He has completed 12 years of schooling, and has obtained a high school degree. Both his parents completed 12 years of schooling. At the time of the survey, he was working full-time.

   Write a similar sentence for cases 2 and 5.

2. The following three questions are asked in a questionnaire, and the answers are recorded at the nominal level. Are the categories offered exhaustive and mutually exclusive? Say whether these categories are adequate for the society you live in, and if they are not, propose another set of categories that is more appropriate for the situations that you know.

   1. Marital status: Married
   Divorced
   Single

   2. Language spoken at home: English
   French
   Other

   3. What method of transportation do you mostly use to come to school? (choose only one):
   Public transportation
   Private car (alone or shared)
   Bicycle
   Walking
   Other

3. Determine whether the following variables are qualitative or quantitative:

   age
   program of study
   number of children
   value of your house
   height
   country of origin
   income
   religion you were raised in
   marital status
   nationality
   ownership of home (Yes/No)
   GPA in high school

4. Determine the level of measurement used in the following cases.

   (a) Annual Income: $_ _ _ _ _ _ . _ _

   (b) Annual Income: 1. Low (less than $20,000)
   2. Medium (more than $20,000 but less than $50,000)
   3. High (more than $50,000)

   (c) Sex: 1. Male
   2. Female

(d) Language(s) spoken: 1. French
2. English
3. German
4. Other

(e) Number of languages spoken: 1. One language
2. Two languages
3. Three languages
4. More than three

5. **Operationalizing a concept: the example of job satisfaction**. You wish to study satisfaction at work for employees of a large company. You consider dimensions such as: the quality of the social environment, the interest in the tasks to be accomplished, the potential for professional advancement, the pay, etc. As a way of gaining insight into the issue, propose a set of dimensions and indicators to measure this concept.

   Note. There is a large body of literature on this issue, with much lively debate about which indicators to include and which to leave out, and how to define them. A good piece of research on the subject should start with a thorough literature review of the scientific production on this issue. The purpose of this exercise is not to discuss the theoretical issue of defining and measuring job satisfaction, but to use this concept as an illustration of how concepts are broken down into dimensions and indicators, and as a way to acquire an insight into the process of determining indicators. Students should understand that any real research must rely on the scientific literature. In an old but classical study, Wanous and Lawler (1972) identified 23 indicators, which they grouped into six dimensions:

   Dimensions: esteem; growth; security; social; autonomy; pay.

   Indicators: self esteem or respect; opportunity for growth; prestige of job inside company; amount of close supervision; opportunity for independent thought; feeling of security; opportunity for feedback on performance; prestige of job outside company; opportunity to complete work; opportunity to do challenging work; feeling that you know when the job is done well; opportunity to do many things; opportunity to get to know others; chance to do a whole piece of work; freedom on the job; variety on the job; pay for job; feeling of accomplishment; opportunity to help others; opportunity for participation; opportunity for close friendships; opportunity for promotion; amount of respect and fair treatment.

   A more recent study by van Saane et al. (2003) evaluates 29 instruments (i.e. 29 ways of operationalizing the concept) and comes up with 11 dimensions that should be present: autonomy; work content; communication; financial rewards; growth/development; promotion; co-workers; meaningfulness; supervision/feedback/recognition; workload; work demands.

   Both studies are available on the internet should you wish to look into this issue more closely. This example demonstrates that there is no unique way of defining a concept and of breaking it down into indicators that can be measured. There are no right and wrong answers. But the issue is not arbitrary either: the choice of dimensions and indicators depends on the theoretical framework that is used, and on the objectives of the study. This choice is always open to challenges and debate, leading to improvements of the measures.

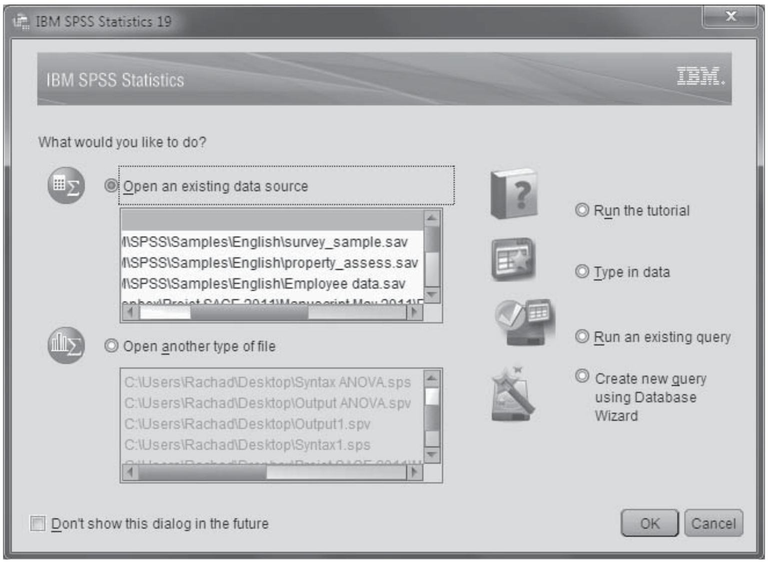
6. **The consumer price index**. The consumer price index is a typical example of breaking down a concept into dimensions and indicators to be able to measure it, and then combining the measures of the various indicators to come up with a single number that constitutes an overall measure of the concept. The cost of living index is calculated by

taking into account several dimensions (housing, food, transportation, health, education, leisure, etc.), measuring indicators for each, then combining all the measures into a single measure. There is a problem, however: the importance of a dimension is not the same across levels of income and across regions in a country. For example, housing carries a larger weight in a small budget than in a bigger one. Theoretically, there should be different cost of living indices for different income levels. Most countries, however, work with averages. In almost every country, this is done by the national statistical institution (such as the UK National Statistics, the US Census Bureau, Statistics Canada, Australian Statistics Bureau, etc.).

Find out how the consumer price index is calculated in the country you live in. Identify the dimensions and indicators that are used, by searching the database of your national statistical institution. What weight is given to each dimension?

SPSS tutorial: Getting started with SPSS

This tutorial is intended to be used when you are sitting at a computer with the SPSS program installed. It will help you become familiar with the basic features of the program. You will learn how the windows of SPSS are organized, how to read the information displayed, and how cases and variables are organized. The tutorials in this book have been written with IBM SPSS 19 (Graduate pack), with permission. They have been produced on a Windows platform, but the windows and dialog boxes are almost identical on a Mac. The dialog boxes in earlier versions look slightly different, but all the analyses shown in this manual can be performed with most of the earlier versions. The difference is much greater in the displays: tables and graphs look much better and are more sophisticated in later versions.



What would you like to do?

● Open an existing data source

\SPSS\Samples\English\survey_sample.sav
\SPSS\Samples\English\property_assess.sav
\SPSS\Samples\English\Employee data.sav

○ Open another type of file

C:\Users\Rachad\Desktop\Syntax ANOVA.sps
C:\Users\Rachad\Desktop\Output ANOVA.spv
C:\Users\Rachad\Desktop\Output1.spv
C:\Users\Rachad\Desktop\Syntax1.sps

○ Run the tutorial

○ Type in data

○ Run an existing query

○ Create new query using Database Wizard

□ Don't show this dialog in the future

Figure 1.7 The opening dialog box in SPSS

Starting SPSS

When you open SPSS you will see the dialog box shown in Figure 1.7.

You will notice that you have several choices: you can open a recent SPSS file, or a recent file of another type (such as Excel, or tab-delimited, or several other types), or you can run the tutorial to know more about SPSS, or you can open a blank file and type in data, or, finally, you can fetch the data in a database (this operation is called a *query*: we will not deal with it in this book, but the Help menu available in SPSS explains clearly how to do it should you have such a need).

If you click on the words **More Files…** in the first box, you can seek the SPSS sample files which are stored in the IBM folder on your computer. On a Mac computer this folder is located in the Applications folder. On a Windows-based computer, it is located in the Programs folder. You then click on the subfolders to open them as follows:

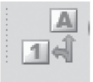
IBM → SPSS → Statistics → 19 → Samples → English

You will find in this folder the samples that are provided with your SPSS 19 program. Take a look at the range of files you have. One of them is labelled **survey_sample.sav**. The extension **.sav** indicates that this file is saved as an SPSS file. We are going to use this file to illustrate the basic features of SPSS.

Navigating in the Data Editor

Every time you open an SPSS data file, you get what is called an **SPSS Data Editor**. It can appear in one of two views: a **Data View** (see Figures 1.1 and 1.2) and a **Variable View** (Figure 1.4). At the bottom center of the SPSS window, there are two tabs that allow you to switch from one view to the other.

When you click on the Data View tab of the **SPSS Data Editor**, you see the data itself, and you can modify it directly in this window. Recall now all the definitions learned at the beginning of this chapter, as we will refer to them in the exploration of SPSS that now follows. Perform the operations indicated below and see their effect on the screen:

- Get SPSS to display the value labels instead of the codes by selecting **Value Labels** under the **View** menu. Alternatively, you could click on the following icon in the icon bar:
  
- Every click switches back and forth between the codes and the labels.
- Read the full name of any variable by pointing the mouse on the short name, at the top of any column.
- Enlarge any column by positioning the mouse right on the edge separating the variable names, and drag slightly to the right without lifting your finger.
- Select **Variables** under the **Utilities** menu. You will get the dialog box shown in Figure 1.3. By scrolling down the list of variables with the mouse, or simply moving from one variable to the next, by using the arrows on your keyboard, you will be able to see the detailed description of each variable, one at a time.

You will see:

- The **Variable Label** (which is the long name of the variable).
- The **Type**. This is the format of the numerical values or labels you enter:

F1	means that the format used for that variable is one space and no decimals;
F4.2	means that 4 spaces are reserved to write the entries of that variable, of which 2 are decimals. The decimal separator (the dot) uses one space. Thus, the number 3.62 has the format F4.2, because 4 spaces are needed to write it (the dot uses a space).
F6.1	is a format that uses 6 spaces, including one decimal and the dot. Thus 4527.3 has the format F6.1.

Scroll down the variable list, examine the various types that are written, then compare them with what you see in the Data View window.

- The **Missing Values**. For each variable, some of the answers should not be taken into account in the statistics, such as when somebody refuses to answer, or when the question does not apply to that person. We give codes for the values 'Refuses to answer' and 'Does Not Apply', but we must indicate that these answers are not to be treated like the other answers. We label them missing values. We will see later on how to do that.
- The **Measurement Level**. This refers to what you have seen in this chapter. For instance, the marital status variable is measured by a nominal scale, which means that the various values are not ranked.
- The **Value Labels**, and their corresponding codes. This is what tells you that
  1. stands for Married
  2. stands for Widowed, etc.

SPSS allows you to get all this information, for all the variables, in one single shot. You do that by choosing:

File → Display Data File Information → Working file

This command produces all the information about the variables (i.e. the *codebook*) in a new window, called the Output Viewer window.

Windows in SPSS

SPSS has four kinds of windows. We will examine three of them, and ignore the fourth one (the Script window) as we will not use it in this book.

The **Data Editor** window is what we have seen so far. In this window, you can edit the data by erasing it, modifying it, or adding new data in the form of new cases, new codes and value labels, or new variables. It comes in two views, the Data View and the Variable View.

Notice that next to the name of the file, on the top of the window, you can read: **[Data Set 1]**. This is because this version of SPSS allows you to have several data sets open simultaneously. SPSS keeps track of them by numbering them in this way.

The **Output Viewer** is a different type of window that contains the results of any task performed by SPSS. Tables, charts, and file information are all displayed in the Output Viewer window. The contents of this window can be exported to another file if you wish, by selecting **File → Export**. You can then name the exported file as you wish, select its location (by clicking on the Browse button), and also select its format: Excel, PDF, Word/RTF, or simply plain text.

Perform these operations now: produce the data file information and export it in various formats to see the advantages and disadvantages of each format.

Notice that the Output Viewer is split in two vertically. The part on the left is called the *document map*. When you have produced a large output, the document map allows you to navigate quickly from one element to the other, simply by clicking on the little icon representing the table or graph that you want to see. The tutorials in the subsequent chapters will help you become familiar with these various functions.

The **Syntax** window is where you can write commands to be executed by SPSS, rather than clicking items in the menu. This is a very important feature of SPSS. For one thing, it keeps track of all the statistical computations you have done. If you notice that there was a mistake in the file, or if you want to modify the label of a variable because it is too long or unclear, and you want to perform the analysis again, you just select the commands and make them run, and you have all your results in one shot.

Commands in SPSS must obey certain rules, referred to as syntax rules, by analogy to natural languages. These commands allow us to perform more operations in SPSS than with the menus, and in a much more efficient way. The syntax needed to obtain a frequency table for the marital status variable is, for example:

```
FREQUENCIES VARIABLES=marital
 /ORDER = ANALYSIS.
```

The commands must follow very precise rules that are explained in the SPSS on-line manuals. We will not be working with syntax systematically in this book, but you should know what it is and we will make occasional remarks on the syntax. You should know that the great advantage of syntax is that when you have a long set of commands you can modify any command manually (e.g. we can add another variable next to the word ‘marital’), and you can repeat the same analysis with or without modifications just by selecting all commands and clicking the Run button. Moreover, when you perform certain operations with the menus, by clicking with the mouse, most dialog boxes include a button labelled: **Paste**. Clicking it will paste the commands in the Syntax window, for future use.

Here is a concrete example. We will produce a frequency table for the variable *marital*, and save its command syntax. Select:

Analyze → Descriptives → Frequencies

You will get the dialog box shown in Figure 1.8.

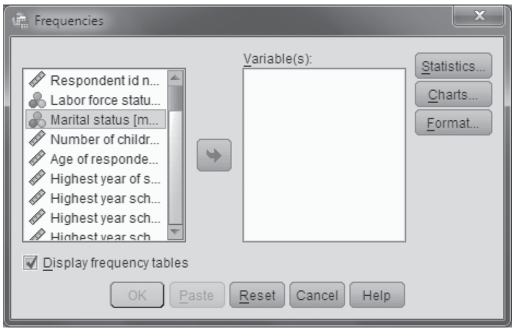
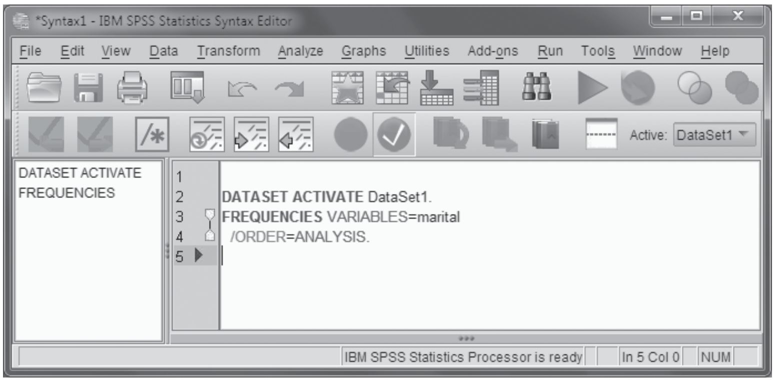


Figure 1.8 The Frequencies dialog box

Click on Marital status, then on the arrow that places it in the Variable(s): box on the right. If you click **OK**, you get the frequency table directly. But if you click **Paste** instead, you get the window shown in Figure 1.9.


*Syntax1 - IBM SPSS Statistics Syntax Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons Run Tools Window Help
Active: DataSet1
DATASET ACTIVATE
FREQUENCIES
IBM SPSS Statistics Processor is ready
In 5 Col 0
NUM

```
DATASET ACTIVATE DataSet1.
FREQUENCIES VARIABLES=marital
  /ORDER=ANALYSIS.
```

Figure 1.9 The Syntax window

Here you can see the Frequencies command, which specifies which variables you want to analyse. Notice also that this syntax has not been saved yet. You can save it as you would save any file. If you select the command and you run it by clicking on the green arrow at the top of the window, you get a frequency table in the Output Viewer. Do it now. Table 1.7 shows what you should see.

Table 1.7 The frequency table for the marital status variable

Marital status

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Married	1346	47.5	47.5	47.5
	Widowed	283	10.0	10.0	57.5
	Divorced	446	15.7	15.8	73.3
	Separated	93	3.3	3.3	76.6
	Never married	663	23.4	23.4	100.0
	Total	2831	100.0	100.0	
Missing	NA	1	.0		
Total		2832	100.0		

This table gives the number of people in each category, their percentage, but also two columns labelled Valid Percent and Cumulative Percent. The first eliminates the missing answers before computing the percentages. The second is automatically calculated by SPSS, but it is not useful here: it is only useful when the measuring scale is ordinal or quantitative. In this case, it makes sense to add together the percentages as you go. Not so when the variables are nominal.

Additional exercises

Now that you know how SPSS functions, take some time to become familiar with the basic features explained in this tutorial. Examine the variables, their labels, the categories used, etc. For instance, you may want to give a full description of indivudual number 20, say, as we did in Exercise 1.e.

The tutorials in the next chapters will take you through the basic aspects of SPSS that are relevant to this book.

Suggestions for further reading

Babbie, Earl R. (2009) *The Practice of Social Research* (12th edn). Belmont, CA: Wadsworth.

Blalock Jr., Hubert M. (1982) *Conceptualization and Measurement in the Social Sciences*. London: Sage.

Field, Andy (2009) *Discovering Statistics Using SPSS* (3rd edn). London: Sage.

Trudel, Robert and Antonius, Rachad (1991) *Méthodes quantitatives appliquées aux sciences humaines*. Montréal: CEC.