Antonius Rachard. 2003. *Interpreting Quantitative Data with SPSS*. London: Sage.

1 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, the student 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 at studying social and human phenomena as rigorously as possible. This involves describing some aspect of the social reality, analyzing it to see whether logical links can be established between its various parts, and, whenever possible, predicting future 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 into numbers some aspects of our observations. 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 the 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 unemployment rate; 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 (that is, 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 a 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 (that is, data) about a character or a quality, such as the mother tongue of a person. Sometimes, the data is something 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 (that is, 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 first 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

The first object of analysis in quantitative methods is a **data file**, that is, a set of pieces of information written down in a codified way. Figure 1.1 illustrates what an **electronic data file** looks like when we open it with the SPSS program.

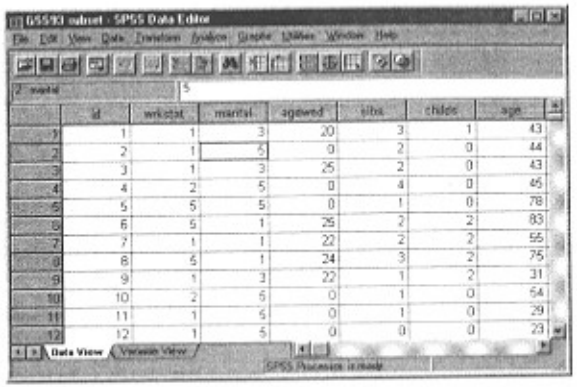

Figure 1.1 **The Data window in SPSS version 10.1. © SPSS. Reprinted with permission.**

This data file was created by the statistical software package SPSS Version 10.1, which will be used in this course. The first lab in the second part of this manual will introduce you to SPSS, which stands for *Statistical Package for the Social Sciences*. On the top of the window, you can read the name of the data file: **GSS93 subset**. This stands for **Subset of the General Social Survey**, a survey conducted in the USA in 1993.

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 left of the window.

The Data View: The information in this data view 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**, that is, a character or quality that was measured in this survey. For instance, the second column is a variable called **wrkstat**, and the third is a variable called **marital**.

But what are the meanings of all these numbers and words? A data file must be accompanied by information that allows a reader to interpret (that is, understand) the meanings of the various elements in it. This information constitutes the **codebook**. In SPSS, we can find the information of the codebook by clicking the word **Variables...** under the **Utilities** menu. We get a window listing all the variables contained in this data file. By clicking once on a variable, we see the information pertaining to this variable:

- the short name that stands on the top of the column;
- what the name stands for (the **label** of the variable);
- the numerical type of the variable (that is, how many digits are used, and whether it includes decimals);
- other technical information to be explained later;
- and the **Value Labels**, that is, what each number appearing in the data sheet stands for.

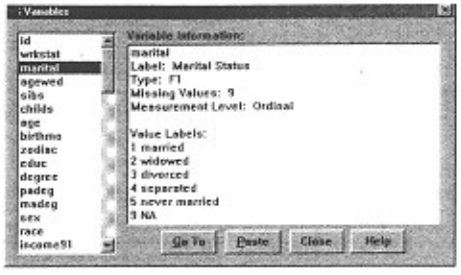

Figure 1.2 **The Variables window in SPSS. The codes and value labels of the variable Marital Status are shown**

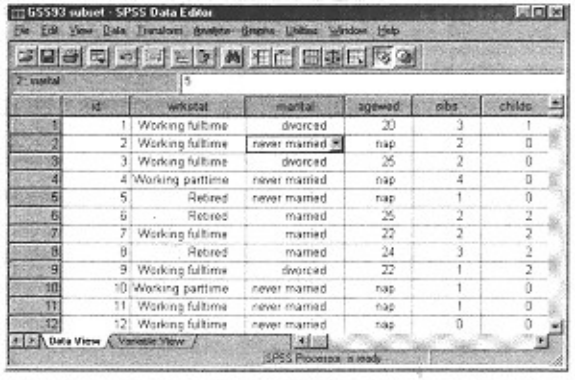

Figure 1.3 **The Data View window in SPSS when the Show Labels command is ticked in the View menu. The value labels are displayed rather than the codes**

Figure 1.2 shows the codes used for the variable Marital Status.

You may have noticed that:

1 stands for married
2 stands for widowed
3 stands for divorced
etc.

The numbers 1, 2, 3, etc. are the **codes,** and the terms married, widowed, divorced, etc. are the **value labels** that correspond to the various codes. The name *marital*, which appears at the top of the column, is the **variable name**. *Marital Status* is the **variable label**: it is a usually longer, detailed name for the variable. When we print tables or graphs, it is the variable labels and the value labels that are printed.

Figure 1.4 **The Variable View window in SPSS. The variables are listed in the rows, and their properties are displayed**

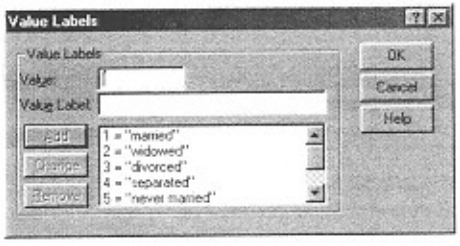

Figure 1.5 **The Value Labels window in SPSS. In this window it is possible to add new codes and their corresponding value labels, or to modify or delete existing ones**

There is a way of showing the value labels instead of the codes. This is done by clicking **Value Labels** under the **View** menu. The Data View window looks now as shown in Figure 1.3.

We can see that case number 4, for example, is a person who works part time, and who has never been married. To understand the precise meaning of the numbers written in the other cells, we should first read the variable information found in the codebook for each of the variables.

In version 10.0 and version 11 of SPSS, you can read the information pertaining to the variables in the Variable View. By clicking on the tab for Variable View, you get the window shown in Figure 1.4.

In the Variable View, no data is shown. You can see, however, all the information pertaining to the variables themselves, each variable being represented by a line. The various variable names are listed in the first column, and each is followed by information about the corresponding variable: the way it is measured and recorded, its full name, the values and their codes, etc. All these terms will be explained in detail later on. The label, that is, the long name of the variable **marital**, is **Marital Status**. By clicking on the **Values** cell for the variable **marital**, the window shown in Figure 1.5 pops up.

We can see again the meanings of the codes used to designate the various marital statuses. We can now 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 **Some questions that arise when we want to use quantitative methods**

Questions	Chapters
How did we come up with this data? 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?	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? What are the most common types of shapes and distributions we encounter?	3. Univariate Descriptive Statistics 5. Normal Distributions
What are the procedures for selecting a sample? Are some of them better than others?	6. Sampling Designs
Some institutions collect and publish a lot of social data. Where can we find it? How do we use it?	7. Statistical Databases
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?	8. Statistical Association
Sometimes the data comes from a sample, that is, a part of the population, and not the whole population. Can we generalize our conclusions to the whole population on the basis of the data collected on a sample? How can this be done? Is it precise? What are the risks that our conclusions are wrong?	9. Statistical Inference: Estimation 10. Statistical Inference: Hypothesis Testing

The Discipline of Statistics

The term *statistics* is used in two different meanings: 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 that pertain to a large quantity of individuals or a large quantity of objects. It includes a general, theoretical aspect which is very mathematical, 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 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* has also 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 **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 were 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 Quebec, the population will consist of all hospitals in Quebec. 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.

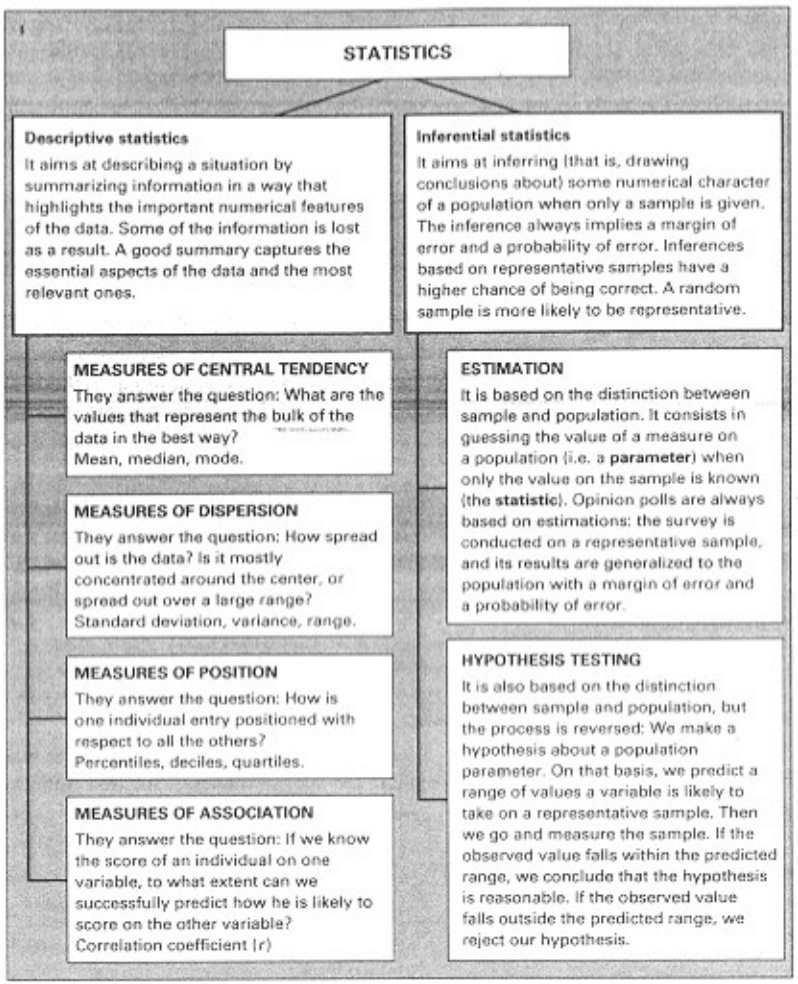

Figure 1.6 **The discipline of statistics and its two branches, descriptive statistics and inferential statistics**

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.

The following paragraphs explain what each branch is about. Refer also to Figure 1.6. Some of the terms used in the diagram may not be clear for now, but they will be explained as we progress.

Descriptive Statistics

The methods and techniques of descriptive statistics aim at summarizing large quantities of data by a few numbers, in a way that highlights the most important numerical features of the data. For instance, if you say that your average GPA (grade point average) 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 average GPA is a sufficient measure in many circumstances. However, the 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 the corresponding chapters.

Inferential Statistics

Inferential statistics aim at generalizing 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 aim at generalizing observations made on a sample to a whole population. For instance, when pre-election polls are conducted, only one or two thousand 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. How can we ensure that a sample is representative? Well, we can't. 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 include estimation and hypothesis testing, two techniques that will be studied in Chapters 9 and 10.