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

3 UNIVARIATE DESCRIPTIVE STATISTICS

This chapter explains how data concerning one variable can be summarized and described, with tables and with simple charts and diagrams.

After studying this chapter, the student should know:

- the basic types of univariate descriptive measures;
- how the level of measurement determines the descriptive measures to be used;
- how to interpret these descriptive measures;
- how to read a frequency table;
- the differences in the significance and the uses of the mean and the median;
- how to interpret the mean when a quantitative variable is coded;
- how to describe the shape of a distribution (symmetry; skewness);
- how to present data (frequency tables; charts);
- what are weighted means and when to use them.

Data files contain a lot of information that must be summarized in order to be useful. If we look for instance at the variable **age** in the data file **GSS93 subset** that comes with the SPSS package, we will find 1500 entries, giving us the age of every individual in the sample. If we examine the ages of men and women separately, we cannot determine, by looking simply at the raw data, whether men of this sample tend to be older than women or whether it is the other way around. We would need to know, let us say, that the average age of men is 23 years and of women 20 years to make a comparison. The average is a descriptive measure.

Descriptive statistics aim 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. It summarizes it with the help of numbers, usually organized into tables, but also with the help of charts and graphs that give a visual representation of the distributions.

In this chapter, we will be looking at one variable at a time. Measures that concern one variable are called **univariate** measures. We will examine **bivariate** measures, those measures that concern two variables together, in Chapter 8.

There are three important types of univariate descriptive measures:

- measures of central tendency,
- measures of dispersion, and
- measures of position.

Measures of central tendency (sometimes called *measures of the center*) answer the question: What are the categories or numerical values that represent the *bulk* of the data in the best way? Such measures will be useful for comparing various groups within a population, or seeing whether a variable has changed over time. Measures of central tendency include the *mean* (which is the technical term for average), the *median*, and the *mode*.

Measures of dispersion answer the question: How spread out is the data? Is it mostly concentrated around the center, or spread out over a large range of values? Measures of dispersion include the standard deviation, the variance, the range (there are several variants of the range, such as the interquartile range) and the coefficient of variation.

Measures of position answer the question: How is one individual entry positioned with respect to all the others? Or how does one individual score on a variable in comparison with the others? If you want to know whether you are part of the top 5% of a math class, you must use a measure of position. Measures of position include percentiles, deciles, and quartiles.

Other measures. In addition to these measures, we can compute the *frequencies* of certain subgroups of the population, as well as certain *ratios* and *proportions* that help us compare their relative importance. This is particularly useful when the variable is qualitative, or when it is quantitative but its values have been grouped into categories.

The various descriptive measures that can be used in a specific situation depend on whether the variable is qualitative or quantitative. When the variable is quantitative, we can look at the *general shape of the distribution*, to see whether it is *symmetric* (that is, the values are distributed in a similar way on both sides of the center) or *skewed* (that is, lacking symmetry), and whether it is rather flat or rather peaked (a characteristic called *kurtosis*).

Finally, we can make use of charts to convey a visual impression of the distribution of the data. It is very easy to produce colorful outputs with any statistical software. It is important, however, to choose the *appropriate* chart, one that is *meaningful* and that *conveys* the most important properties of the data. This is not always easy, and you will have to pay attention to the way an appropriate chart is chosen, a choice that depends on the level of measurement of the variable.

It is very important to realize that the statistical measures used to describe the data pertaining to a variable depend on the level of measurement used. If a variable is measured at the nominal scale, you can compute certain measures and not others. Therefore you should pay attention to the *conditions* under which a measure could be used; otherwise you will end up computing numerical values that are meaningless.

Measures of Central Tendency

For Qualitative Variables

The best way to describe the data that corresponds to a qualitative variable is to show the **frequencies** of its various categories, which are a simple count of how many individuals fall into each category. You could then work out this count as a **percentage** of the total number of units in the sample. When you ask for the frequencies, SPSS automatically calculates the percentages as well, and it does it twice: the percentage with respect to the total number of people in the sample, and the percentage with respect to the valid answers only, called **valid percent** in the SPSS outputs. Let us say that the percentage of people who answered Yes to a question is 40% of the total. If only half the people had answered, this percentage would correspond to 80% of the valid answers. In other words, although 40% of the people answered Yes, they still constitute 80% of those who answered. SPSS gives you both percentages (the total percentage and the valid percentage) and you have to decide which one is more significant in a particular situation.

For instance, Table 3.1 summarizes the answers to a question about the legalization of marijuana, in a survey given to a sample of 1500 individuals.

Table 3.1 **A frequency table, showing the frequencies of the various categories, as well as the percentage and valid percentage they represent in the sample**

Should Marijuana Be Made Legal

		Frequency	Percent	Valid Percent
Valid	Legal	211	14.1	22.7
	Not legal	719	47.9	77.3
	Total valid	**930**	**62.0**	**100.0**
Missing		570	38.0	
Total		**1500**	**100.0**	

Table 3.1 tells us that the sample included 1500 individuals, but that we have the answers to that question for 930 individuals only. The percentage of positive answers can be calculated either out of the total number of people in the sample, giving 14.1% as shown in the Percent column, or out of the number of people for whom we have answers, giving 22.7% as shown in the Valid Percent column. Which percentage is the most useful? It depends on the reason for the missing answers. If people did not answer because the question was asked of only a subset of the sample, the valid percentage is easier to interpret. But if 570 people abstained because they do not want to let their opinion be known, it is more difficult to interpret the resulting figures. A good analysis should include a discussion of the missing answers when their proportion is as important as it is in this example.

Table 3.1 comes from the SPSS output. When we write a statistical report, we do not include all the columns in that table. Most of the time, you would choose either the valid percentage (which is the preferred solution) or the total percentage, but rarely both, unless you want to discuss specifically the difference between these two percentages. The cumulative percentage is only used for ordinal or quantitative variables, and even then is included only if you plan to discuss it.

To describe the *center* of the distribution of a qualitative variable, you must determine which category includes the biggest concentration of data. This is called the mode. *The* ***mode*** *for a qualitative variable is the category that has the highest frequency* (sometimes called **modal category**).

The modal category could include more than 50% of the data. In this case we say that this category includes the **majority** of individuals. If the modal category includes less than 50% of the data, we say that it constitutes a **plurality**. We can illustrate this by the following situations concerning the votes in an election.

First situation:	Party A	54% of the votes
	Party B	21% of the votes
	Party C	25% of the votes,

Here we could say that Party A won the election with a *majority*. Compare with the following situation.

Second situation:	Party A	44% of the votes
	Party B	31% of the votes
	Party C	25% of the votes,

Here we can say that Party A won the election with a *plurality* of votes, but without a majority. If Parties B and C formed a coalition, they could defeat Party A. For this reason, some countries include in their electoral law a provision that, should the winning candidate or a winning party get less than the absolute majority of votes (50% + 1), a second turn should take place among those candidates who are at the top of the list, so as to end up with a winner having more than 50% of the votes.

A good description of the distribution of a qualitative variable should include a mention of the modal category, but it should also include a discussion of the pattern always easy, and you will have to pay attention to the way an appropriate chart is chosen, a choice that depends on the level of measurement of the variable.

It is very important to realize that the statistical measures used to describe the data pertaining to a variable depend on the level of measurement used. If a variable is measured at the nominal scale, you can compute certain measures and not others. Therefore you should pay attention to the *conditions* under which a measure could be used; otherwise you will end up computing numerical values that are meaningless.

Measures of Central Tendency

For Qualitative Variables

The best way to describe the data ~~that~~ corresponds to a qualitative variable is to show the **frequencies** of its various categories, which are a simple count of how many individuals fall into each category. You could then work out this count as a **percentage** of the total number of units in the sample. When you ask for the frequencies, SPSS automatically calculates the percentages as well, and it does it twice: the percentage with respect to the total number of people in the sample, and the percentage with respect to the valid answers only, called **valid percent** in the SPSS outputs. Let us say that the percentage of people who answered Yes to a question is 40% of the total. If only half the people had answered, this percentage would correspond to 80% of the valid answers. In other words, although 40% of the people answered Yes, they still constitute 80% of those who answered. SPSS gives you both percentages (the total percentage and the valid percentage) and you have to decide which one is more significant in a particular situation.

For instance, Table 3.1 summarizes the answers to a question about the legalization of marijuana, in a survey given to a sample of 1500 individuals.

Table 3.1 **A frequency table, showing the frequencies of the various categories, as well as the percentage and valid percentage they represent in the sample**

Should Marijuana Be Made Legal

		Frequency	Percent	Valid Percent
Valid	Legal	211	14.1	22.7
	Not legal	719	47.9	77.3
	Total valid	**930**	**62.0**	**100.0**
Missing		570	38.0	
Total		**1500**	**100.0**	

Table 3.1 tells us that the sample included 1500 individuals, but that we have the answers to that question for 930 individuals only. The percentage of positive answers can be calculated either out of the total number of people in the sample, giving 14.1% as shown in the Percent column, or out of the number of people for whom we have answers, giving 22.7% as shown in the Valid Percent column. Which percentage is the most useful? It depends on the reason for the missing answers. If people did not answer because the question was asked of only a subset of the sample, the valid percentage is easier to interpret. But if 570 people abstained because they do not want to let their opinion be known, it is more difficult to interpret the resulting figures. A good analysis should include a discussion of the missing answers when their proportion is as important as it is in this example.

Table 3.1 comes from the SPSS output. When we write a statistical report, we do not include all the columns in that table. Most of the time, you would choose either the valid percentage (which is the preferred solution) or the total percentage, but rarely both, unless you want to discuss specifically the difference between these two percentages. The cumulative percentage is only used for ordinal or quantitative variables, and even then is included only if you plan to discuss it.

To describe the *center* of the distribution of a qualitative variable, you must determine which category includes the biggest concentration of data. This is called the mode. *The* **mode** *for a qualitative variable is the category that has the highest frequency* (sometimes called **modal category**).

The modal category could include more than 50% of the data. In this case we say that this category includes the **majority** of individuals. If the modal category includes less than 50% of the data, we say that it constitutes a **plurality**. We can illustrate this by the following situations concerning the votes in an election.

First situation:	Party A	54% of the votes
	Party B	21% of the votes
	Party C	25% of the votes.

Here we could say that Party A won the election with a *majority*. Compare with the following situation.

Second situation:	Party A	44% of the votes
	Party B	31% of the votes
	Party C	25% of the votes.

Here we can say that Party A won the election with a *plurality* of votes, but without a majority. If Parties B and C formed a coalition, they could defeat Party A. For this reason, some countries include in their electoral law a provision that, should the winning candidate or a winning party get less than the absolute majority of votes (50% + 1), a second turn should take place among those candidates who are at the top of the list, so as to end up with a winner having more than 50% of the votes.

A good description of the distribution of a qualitative variable should include a mention of the modal category, but it should also include a discussion of the pattern of the distribution of individuals across the various categories. Concrete examples will be given in the last section of this chapter.

For Quantitative Variables

Quantitative variables allow us a lot more possibilities. The most useful measures of central tendency are the mean and the median. We will also see how and when to use the mode. *The **mean** of a quantitative variable is defined as the sum of all entries divided by their number.*

In symbolic terms,

the mean of a *sample* is written as $\bar{x} = \frac{\sum x_i}{n}$, and

the mean of a *population* is written as $\mu_x = \frac{\sum x_i}{N}$

These symbols are read as follows:

- $\bar{x}$ is read as *x bar*, and it stands for the mean of a sample for variable X.
- μ_x is read as *mu x*, and it stands for the mean of a population. The subscript x refers to the variable X.
- x_i is read as x i. It refers to all the entries of your data that pertain to the variable X, which are labeled x_1, x_2, x_3, etc.
- Σ is read as *sigma*. When followed by x_i, it means: add all the x_i's, letting i range over all possible values, that is, from 1 to n (for a sample) or from 1 to N (for a population).
- n is the size of the sample, that is, the number of units that are in it.
- N is the size of the population.

You may have noticed that we use different symbols for a population and for a sample, to indicate clearly whether we are talking about a population or a sample. We do not always need to write the subscript x in μ_x. We do it only when several variables are involved, and when we want to keep track of which of the variables we are talking about. In such a situation we would use μ_x, μ_y, and μ_z to refer to the mean of the population for the variables x, y, and z respectively. Notice that in the formula for the mean of a population, we have written a capital N to refer to the size of the population rather than the small n used for the size of a sample.

The mean is very useful to compare various populations, or to see how a variable evolves over time. But it can be very misleading if the population is not homogeneous. Imagine a group of five people whose hourly wages are: $10, $20, $45, $60 and $65 an hour. The average hourly wage would be:

$$\bar{x} = \frac{10 + 20 + 45 + 60 + 65}{5} = \$40 \text{ an hour.}$$

But if the last participant was an international lawyer who charged $400 an hour of consultancy, the average would have been $107 an hour (you can compute it yourself), which is well above what four out of the five individuals make, and would be a misrepresentation of the center of the data.

In order to avoid this problem, we can compute the **trimmed mean**: you first eliminate the most extreme values, and then you compute the mean of the remaining ones. But you must indicate how much you have trimmed. In SPSS, one of the procedures produces a **5% trimmed mean**, which means that you disregard the 5% of the data that are farthest away from the center, and then you compute the mean of the remaining data entries.

The mean has a mathematical property that will be used later on. Starting from the definition of the mean, which states that $\bar{x} = \frac{\sum x_i}{n}$, we can conclude, by multiplying both sides by n, that:

$$\bar{x} * n = \sum x_i$$

In plain language, this states that the sum of all entries is equal to n times the mean.

We will discuss all the limitations and warnings concerning the mean in a later section on methodological issues.

THE MEAN OF DATA GROUPED INTO CLASSES

When we are given numerical data that is grouped into classes, and we do not know the exact value of every single entry, we can still compute the mean of the distribution by using the midpoint of every class. What we get is not the exact mean, but it is the closest guess of the mean that is available. If the classes are not too wide, the value obtained by using the midpoints is not that different from the value that would have resulted from the individual data.

Consider one of the intervals i with frequency f_i and midpoint x_i. The exact sum of all the entries in that class is not known, but we can approximate it using the midpoint. Thus, instead of the sum of the individual entries (not known) we will count the midpoint of the class f_i times. We obtain the following formula.

$$\text{Mean for grouped data} = \frac{\sum f_i * x_i}{n}$$

Here, n is the number of all entries in the sample. It is therefore equal to the sum of the class frequencies, that is, the sum of the number of individuals in the various classes. The formula can thus be rewritten as