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

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

INTERPRETATION OF THE MEAN WHEN THE VARIABLE IS CODED

We often have data files where a quantitative variable is not given in its original form, but coded into a small number of categories. For instance, the variable Respondent's Income could be given in the form shown in Table 3.2.

Table 3.2 **Example of a quantitative variable that is coded into 21 categories, with a 22nd category for those who refused to answer**

Category	Code
Less than $1000	1
$1000–2999	2
$3000–3999	3
$4000–4999	4
$5000–5999	5
$6000–6999	6
$7000–7999	7
$8000–9999	8
$10,000–12,499	9
$12,500–14,999	10
$15,000–17,499	11
$17,500–19,999	12
$20,000–22,499	13
$22,500–24,999	14
$25,000–29,999	15
$30,000–34,999	16
$35,000–39,999	17
$40,000–49,999	18
$50,000–59,999	19
$60,000–74,999	20
$75,000 and more	21
Refused to answer	22

Thus, we would not know the exact income of a respondent. We would only know the category he or she falls into.

This kind of measuring scale poses a challenge. If we compute the mean with SPSS, we will not get the mean income. We will get the mean code, because it is the codes that are used to perform the computations. There is a data file that comes with SPSS where the income is coded in this way. This data file contains information about 1500 respondents, including information on the income bracket they fall into, coded as shown in Table 3.2. When we exclude the 22nd category, which consists of the people who refused to answer this question, the computation of the mean with SPSS produces the following result:

Mean = 12.35

What is the use of this number? It is not a dollar amount! If we look at Table 3.2, we see that the code 12 stands for an income of between $17,500 a year and $20,000 a year (with that last number excluded from the category). To interpret this number, we should first translate it into a dollar amount (it can be done with a simple rule). But even without transforming it into the dollar amount it corresponds to, we could use the mean code for comparisons. For instance, we will see in Lab 3 that if we compute the mean income separately for men and women, we get

Mean income for men: 13.9

Mean income for women: 10.9

(excluding the category of people who refused to answer).

Although the mean code does not tell us exactly the mean income for men and women, it still tells us that there is a big difference between men and women for that variable. Table 3.2 tells us that the code 13 corresponds to the income bracket $22,500–25,000, while the code 10 represents the income bracket $12,500–15,000. We can conclude that the difference in income between men and women, for that sample, is roughly around $10,000 a year.

We see that that when the variables are coded, the interpretation of the mean requires us to translate the value obtained into what it stands for. For quantitative variables coded this way, it may also be useful to find the frequencies of the various categories, as we did for nominal variables. For the example at hand, we would get Table 3.3 as shown.

The conclusion of the preceding discussion is that when we have an ordinal variable with few categories, or even a quantitative variable that has been recoded into a small number of categories, it may be useful to compute the frequency table of the various categories, in addition to the mean and other descriptive measures.

Weighted Means

Consider the following situation: you want to find the average grade in an exam for two classes of students. The first class averaged 40 out of 50 in the exam, and the second class averaged 46 out of 50. If you put the two classes together, you *cannot* conclude that the average is 43. This is so because the classes may have different numbers of students. Suppose the first class has 20 students, and the second one 40 students. In other words, we have the data shown in Table 3.4.

To compute the average grade for the two classes taken together, we do not need to know the individual scores of each student. Indeed, we have seen before that a sum of n scores is equal to its average times n. We will use this to obtain the formula shown below for weighted means.

The mean for the two classes taken together can be written as

Table 3.3 **Frequencies of the various income categories for the variable Income**

Respondent's income	Frequency	Valid Percent
LT $1000	26	2.6
$1000–2999	36	3.6
$3000–3999	30	3.0
$4000–4999	24	2.4
$5000–5999	23	2.3
$6000–6999	23	2.3
$7000–7999	15	1.5
$8000–9999	31	3.1
$10,000–12,499	55	5.5
$12,500–14,999	54	5.4
$15,000–17,499	64	6.4
$17,500–19,999	58	5.8
$20,000–22,499	55	5.5
$22,500–24,999	61	6.1
$25,000–29,999	84	8.5
$30,000–34,999	83	8.4
$35,000–39,999	54	5.4
$40,000–49,999	66	6.6
$50,000–59,999	38	3.8
$60,000–74,999	23	2.3
$75,000+	44	4.4
Refused to answer	47	4.7
Total	994	100.0
Missing	506	
Grand Total	1500	

Table 3.4 **Two classes of different size and the mean grade in each**

	Average Grade out of 50	Number of Students
Class A	40	20
Class B	46	40

$$\frac{\text{Sum of all scores in class A} + \text{Sum of all scores in class B}}{60}$$

The sum of all scores in class A can be replaced by the average score (40) times 20, since there are 20 students in this class. And the sum of all scores in class B can be replaced also by its average score (46) times 40, since this class includes 40 students. The equation for the mean becomes:

$$\frac{(40 \times 20)}{60} + \frac{(46 \times 40)}{60}$$

This can now be written as:

$$\text{mean of the two classes combined} = 40 \times (\mathbf{20/60}) + 46 \times (\mathbf{40/60})$$

or again as:

$$\text{mean of the two classes combined} = 40 \times (\mathbf{1/3}) + 46 \times (\mathbf{2/3})$$

The last formula is important: we see that the average grade of class A is multiplied by the **weight** of class A, which is its relative importance in the total population. Class A forms 1/3 of the total population (20 students out of 60) and class B 2/3 of the total (40 students out of 60). The underlying formula is:

$$\text{Average grade for the two classes: } 40 \times w_1 + 46 \times w_2$$

The $\mathbf{w_i}$'s are called the **weights** of the various classes. In this case, the weight is an expression of the number of people in each class compared to the total population of the two classes.

The general formula is as follows.

If you have n values	$x_1, x_2, x_3, \ldots$ etc.,
each having the corresponding weights:	$w_1, w_2, w_3, \ldots$ etc.,
the **weighted mean** is given by	$x_1 w_1 + x_2 w_2 + x_3 w_3 + \cdots + x_n w_n$

The weights are positive numbers and must add up to 1. That is:

$$w_1 + w_2 + w_3 + \cdots + w_n = 1.$$

The weights are not always a reflection of the size of the various groups involved. If you are computing the weighted average of your grades during your college studies, the weights could be proportional to the credits given to each course, or they could be an expression of the importance of the course in a given program of studies. A Faculty of Medicine may weight the grades of its candidates by giving a bigger weight to Chemistry and Biology than Art History, for instance.

Example

A buyer wants to evaluate several houses she has seen. She attributes a score out of ten to each house on each of the following items: size, location, internal design, and quality of construction. Any house having a score less than 5 on any item would not be acceptable. The resulting scores for three houses that are seen as acceptable on all grounds are recorded in Table 3.5. The buyer does not

attribute the same importance to each item. The size of the house is the most important quality. The quality of the construction is also very important, but not as important. The buyer attributes a weight to each item, which reflects the importance of that item for her. The weights are given in the last column.

Table 3.5 **Scores given to three houses on four items, and their weights**

Item	House A	House B	House C	Weight of item
Size	9	7	6	0.4
Location	5	9	10	0.1
Internal design	6	5	8	0.2
Quality of construction	7	9	7	0.3

We can now calculate the weighted average score for each house, using the formula for weighted means given above.

For house A: weighted mean score: $10 \times 0.4 + 5 \times 0.1 + 6 \times 0.2 + 7 \times 0.3 = 7.8$
For house B: weighted mean score: $7 \times 0.4 + 9 \times 0.1 + 5 \times 0.2 + 9 \times 0.3 = 7.4$
For house C: weighted mean score: $6 \times 0.4 + 10 \times 0.1 + 8 \times 0.2 + 7 \times 0.3 = 7.1$

We see that house A obtained the highest weighted score. The total, unweighted score of house C is higher than that of house A. But because the items do not all have the same importance, house A ended up having a higher weighted score.

THE MEDIAN AND THE MODE

The **median** is another measure of central tendency for quantitative variables. It is defined as the value that sits right in the middle of all data entries when they are listed in ascending order. If the number of entries is odd, there will be one data entry right in the middle. If the number of entries is even, we will have *two* data entries in the middle, and the median in this case will be their average. Here are two examples.

Case 1: variable X 2, 3, 4, 4, 5, 5, **5**, 6, 7, 8, 11, 13, 13
Case 2: variable Y 2, 3, 4, 4, 5, **5**, **6**, 7, 8, 11, 13, 13

For the variable X we have 13 entries. The value 5 sits in the middle, with six entries equal or smaller than it, and six entries equal or larger. The median for X is thus 5. But for variable Y, we have 12 entries. There are therefore two entries in the middle of the ordered list, not just one. The median will be the average of the two, that is $(5 + 6) \div 2 = \mathbf{5.5}$.

The median is not sensitive to extreme values. Suppose, for instance, that the entries for variable X were: 2, 3, 4, 4, 5, 5, 5, 6, 7, 8, 11, 13, 60. Although the last

entry is very large compared to the others, it does not affect the median, which is still 5. The mean, however, would have been affected (compute it yourself for the two situations and see how different it would be). For this reason, the median is a better representative of the center when there are extremely large values on one side of it. But the mean is more useful for statistical computations, as we will see in the coming sections.

Half the population has a score that is lower than or equal to the median, and the other half has a score larger than the median or equal to it. This way of formulating the median is very useful in situations where the distribution is skewed (such as the distribution of income) or in situations where time is involved, especially when processes have not been completed by everybody, as illustrated below.

Examples of the use of the median

- We are told that the average age at first marriage for a population is 22 years for women, and 25 for men. The median for women is 21, and for men it is 24. This means that by the time they reached 21 years of age, half the women in this population were married. For men, half of them were married by the age of 24.
- In a research on the time taken by immigrants to find a job, 500 new immigrants who arrived at least three years ago are interviewed. The mean can not be found because some of them have not found a regular or full-time job yet. But it is found that the median time taken for them to find a regular, full-time job was 18 months for men, and 5 months for women. This means that by the 18th month after arrival, 50% of the men had found a job. Women were faster in finding regular full-time jobs: 50% had a job within 5 months of their date of arrival.

Because the median involves only the *ordered* list of data entries, it can be used if the quantitative variable is measured at the ordinal level. But if the number of categories is small, the median is not very useful.

The **mode** can also be used for quantitative variables. When the values are grouped into classes, the mode is defined as it is for qualitative variables: it is the class that has the highest frequency. But the mean and median remain the best descriptive measures for quantitative variables. If the variable is continuous and the values have not been grouped into classes, the mode is the value at which a *peak* occurs in the graph representing the distribution.

COMPARISON OF THE MEAN AND THE MEDIAN

Both the mean and the median are measures of central tendency of a distribution, that is, they give us a central value around which the other values are found. They are therefore very useful for comparing different samples, or different populations,

or samples with a population, or a given population at different moments in time to see how it has evolved. However, each of the mean and the median has its advantages and its drawbacks.

The mean takes into account every single value that occurs in the data. Therefore, it is sensitive to every value. A single very large value can boost the mean up if the number of entries is not very large. For instance, if one worker in a group of 20 workers won a $1 million lottery ticket, the average wealth of those 20 would look artificially high. The median is not sensitive to every single value. In a distribution where the largest value is changed from 60 to 600, the median would not change. The mean would.

It follows from these remarks that the mean is a more sophisticated measure, because it takes every value into account. Indeed, it is the mean that is used to compute the standard deviation, which a measure of dispersion that will be seen below. However, in situations where the distribution is not very symmetric, and where there are some extreme values on only one side of the distribution, the mean will tend to be shifted towards the extreme values, whereas the median will stay close to the bulk of the data. Therefore, whenever the distribution is highly skewed, the median is a better representative of the center of the distribution than the mean. This is true for variables such as income or wealth, where the distribution among individuals in a country, and also worldwide, is highly skewed. For such a variable, the median is a more accurate representative of the central tendency of the distribution.

Measures of Dispersion

For Qualitative Variables

There are not many measures of dispersion for qualitative variables. One of the measures we can compute is the **variation ratio**. It tells us whether a large proportion of data is concentrated in the modal category, or whether it is spread out over the other categories. The variation ratio is defined as

$$\text{variation ratio} = \frac{\text{number of entries not in the modal class}}{\text{total number of entries}}$$

It is a positive number smaller than one. If this ratio is close to zero, it indicates a great homogeneity, almost every unit being in the modal class. The farther it is from zero, the greater the dispersion of the data over the other categories. Like many other measures, this one is easy to interpret when doing comparisons. For instance, if we compare the sizes of the various linguistic groups in two cities where several languages are spoken, we can use the variation ratio to assess the degree of heterogeneity in each city. Here is an example.

City	Linguistic groups	Percentage
City A	French speaking	30%
	English speaking	34%
	Chinese speaking	20%
	Other	16%
	Total	**100%**
City B	French speaking	28%
	English speaking	40%
	Chinese speaking	20%
	Other	12%
	Total	**100%**

The variation ratio for city A would be $(30 + 20 + 16)/100 = 0.66$, and for city B it would be $(28 + 20 + 12)/100 = 0.60$, showing that city A is a little more heterogeneous than city B.

For Quantitative Variables

There are many ways of measuring the dispersion for quantitative variables. The simplest is the range, but we also have various forms of restricted range, we have the deviation from the mean, the standard deviation, the variance and finally the coefficient of variation. Let us go through these measures one at a time.

RANGE

The **range** is the simplest way of measuring how spread out the data is. You simply subtract the smaller entry from the larger one and add 1, and this tells you the size of the interval over which the data is spread out. For example, you would describe a range of values for the variable **Age** as follows:

> In this sample, the youngest person is 16 years old and the oldest 89, spanning a range of 74 years (89 – 16 + 1).

But we may have extreme values that give a misleading impression about the dispersion of the data. For instance, suppose that a retired person decided to enroll in one of our classes. We could then say that the ages of the students in this class range from 16 years up to 69 years, but that would be misleading, as the great majority of students are somewhere between 17 years old and maybe 23 or 24 years old. For this reason, we can introduce variants of the notion of range.

The $\mathbf{C_{10\text{-}90}}$ **range**, for instance, computes the range of values after we have dropped 10% of the data at each end: the 10% largest entries and the 10% smallest

entries. This statistic gives us the range of the remaining 80% of data entries. We can also compute the **5% trimmed range** by deleting from the computation the 5% of values that are the farthest away from the mean. We will also see in a forthcoming section something called a *box-plot*, that shows us graphically both the full range, and the range of the central 50% of the data after you have disregarded the top 25% and the bottom 25%. This last range is called the **interquartile range**, the distance between the first and third **quartiles**, which are the values that split the data into four equal parts.

These various notions of the range do not use the exact values of *all* the data in their computation. The following measures do.

STANDARD DEVIATION

The most important measure is the standard deviation. To explain what it is we must first define some simpler notions such as the deviation from the mean. For an individual data entry x_i the **deviation from the mean** is the *distance* that separates it from the mean. If we want to write it in symbols, we will have to use two different symbols, depending whether we have a sample or a population.

For a sample, the deviation from the mean is written: $(x_i - \bar{x})$
For a population, the deviation from the mean is written: $(x_i - \mu)$

The list of all deviations of the mean may give us a good impression of how spread out the data is.

Example

Consider the following distribution, representing the grades out of ten of a group of 14 students:

4, 5, 5, 6, 7, 7, 8, 8, 8, 9, 9, 9, 10, 10

Here the mean is given by 105/14 = 7.5. The deviations from the mean are given in Table 3.6.

But that list may be long. We want to summarize it, and end up with a single numerical value that constitutes a measure of how dispersed the data is. We could take the mean of all these deviations. If you perform the computation for the mean deviation, you will get a mean deviation equal to zero (do the computation yourself on the preceding example). This is no accident. Indeed, we can easily show that the mean of these deviations is necessarily zero, as the positive deviations are cancelled out by the negative deviations.

Table 3.6 **Calculation of the deviations from the mean**

Data entry x_i	Deviation from the mean: $(x_i - \bar{x})$
4	4 – 7.5 = –3.5
5	5 – 7.5 = –2.5
5	5 – 7.5 = –2.5
6	6 – 7.5 = –1.5
7	7 – 7.5 = –0.5
7	7 – 7.5 = –0.5
8	8 – 7.5 = 0.5
8	8 – 7.5 = 0.5
8	8 – 7.5 = 0.5
9	9 – 7.5 = 1.5
9	9 – 7.5 = 1.5
9	9 – 7.5 = 1.5
10	10 – 7.5 = 2.5
10	10 – 7.5 = 2.5

The mathematical proof (which is given only for those who are interested and which can be ignored otherwise) goes like this:

Sum of all deviations from the mean =

$$\sum (x_i - \bar{x}) = \sum x_i - \sum \bar{x} = n * \bar{x} - n * \bar{x} = 0$$

(Explanation: Recall that the sum of all entries is equal to n times the mean, and that the mean, in the second summation, is counted n times. This is why we get n times the mean twice, once with a positive sign, and once with a negative sign.)

We thus conclude that the deviations from the mean always add up to zero, and therefore we cannot summarize them by finding their mean. The way around this difficulty is the following: we will square the deviations, and then take their mean. By squaring the deviations, we get rid of the negative signs, and the positive and negative deviations do not cancel out any more. This operation changes their magnitude, however, and gives an erroneous impression about the real dispersion of data, since the deviations are all squared. This distortion will be corrected by taking the square root of the result, which brings it back to an order of magnitude similar to the original deviations. In summary, we end up with the following calculation:

Standard deviation for a population, denoted by the symbol σ

$$\sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{N}}$$

In the case of a sample, μ will be replaced by $\bar{x}$ and N will be replaced not by n, but by $n - 1$. The reason why we write $n - 1$ instead of n is due to some of the mathematical properties of the standard deviation. It can be proven that using $n - 1$ in the formula gives a better prediction of the standard deviation of a population when we know that of the sample.

Conclusion: the **standard deviation for a sample**, denoted by the symbol **s**, is given by:

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$$

The standard deviation (often written **st.dev.**) is the most powerful measure of dispersion for quantitative data. It will permit us to do very sophisticated descriptions of various distributions. All the calculations of statistical inference are also made possible by the use of the standard deviation.

VARIANCE

Another useful measure is the **variance**, which is defined as the square of the standard deviation. It is thus given by

$$\text{variance of a sample} = s^2$$

or

$$\text{variance of a population} = \sigma^2$$

THE COEFFICIENT OF VARIATION

Finally, we can define the coefficient of variation. To explain the use of this measure, suppose you have two distributions having the means and standard deviations given below:

Distribution 1	mean = 30	st. dev. = 3
Distribution 2	mean = 150	st. dev. = 3

In one case the center of the distribution is 30, indicating that the data entries fall in a certain range *around* the value 30. Their magnitude is around 30. In the other case, the mean is 150, indicating that the data entries fall in a range around the value 150 and have an average magnitude of 150. Although they have the same dispersion (measured by the standard deviation), the relative importance of the dispersion is not the same in the two cases because the magnitude of the data is different. In one case the entries revolve around the value 30, and the standard deviation is equal to 10% of the average value of the entries. In the other case, the entries revolve around the value 150 and the standard deviation is about 3/150, that is, 2% of the average value of the entries, a value which denotes a smaller relative variation.

There is a way to assess the relative importance of the variation among the entries, by comparing this variation with the mean. The measure is called the *coefficient of variation*. The **coefficient of variation** is defined as the standard deviation divided by the mean, and multiplied by 100 to turn it into a percentage. The formula is thus:

$$\text{Coefficient of variation } CV = \frac{\sigma}{\mu} \times 100$$

This measure will only be used occasionally.

Measures of Position

Measures of position are used for quantitative variables, measured at the numerical scale level. They could sometimes be used for variables measured at the ordinal level. They provide us with a way of determining how one individual entry compares with all the others.

The simplest measure of position is the quartile. If you list your entries in an ascending order according to size, *the **quartiles** are the values that split the ranked population into four equal groups*. Twenty-five percent of the population has a score less or equal than the 1st quartile (Q_1), 50% has a score less than the 2nd quartile (Q_2), and 75% has a score less than the 3rd quartile (Q_3). Recall that we have seen earlier a measure of dispersion called the interquartile range, which is the difference between Q_1 and Q_3. Figure 3.1 illustrates the way the quartiles divide the ordered list of units in a sample or in a population.

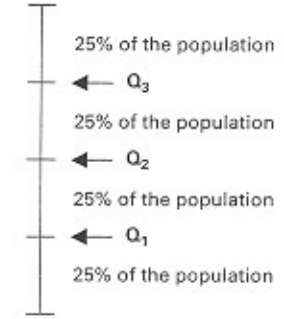

Figure 3.1 **The quartiles are obtained by ordering the individuals in the population by increasing rank, and then splitting it into four equal parts. The quartiles are the values that separate these four parts**

In a similar way, we can define the **deciles**: they split the ranked population into ten equal groups. If a data entry falls in the first decile it means that its score is among the lowest 10%. If it is in the 10th decile it means it is among the top 10%.

The most common measure of position, however, is the *percentile rank*. The data is arranged by order of size (recall it must be quantitative) and divided into 100 equal groups. The numerical values that separate these 100 groups are called **percentiles**. The **percentile rank** of a data entry is the rank of the percentile group this entry falls into. For example, if you are told that your percentile rank in a national exam is 83, this means that you fall within the 83rd percentile. Your grade is just above that of 82% of the population, and just below that of 17% of the population. You will learn in the SPSS session how to display the percentile ranks of the data entries.

You may have realized by now the connection between the median and the various measures of position, since the median divides your ranked population into two equal groups. The median is equal to the 50th percentile. It is also equal to the 5th decile, and of course the 2nd quartile.

Figure 3.2 **A bar chart representing the size (in percentage) of the various categories for the variable Marital Status**

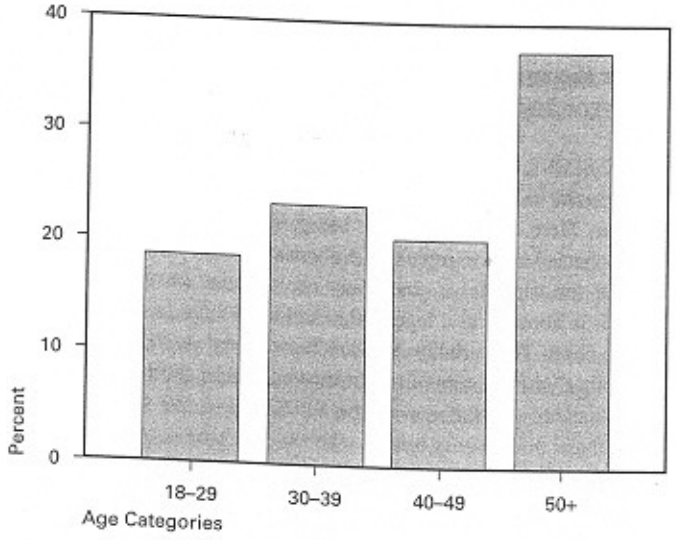

Figure 3.3 **A bar chart representing the various age categories in percentages of the whole sample**

people who speak a given language. You can choose to have the Y-axis represent *percentages* instead of counts. The chart shown in Figure 3.2 represents the percentages of the various marital categories.

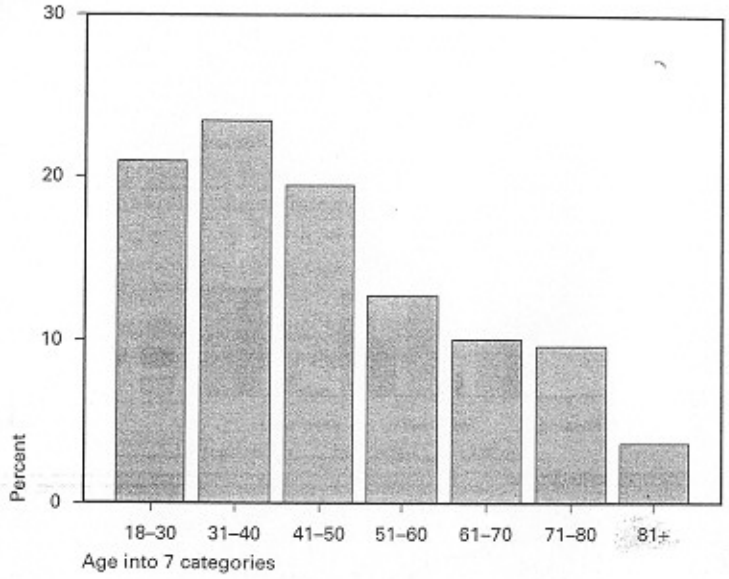

Figure 3.4 **A bar chart where the category 50+ years has been broken down into four categories**

The variable on the X-axis could also be a *quantitative variable that has been grouped into a small number of categories*. For instance, we could have **agecat4** as the variable on the X-axis. The bars would then represent the number of people found in each of the four age categories. In this kind of bar graph, you must be careful about the *range* (that is, the length of the interval) of each of the categories. If the categories are intervals that do not have the same length, you may get the wrong impression that one group is more numerous than the other, such as with the group of people who are 50 years old or more in the chart shown in Figure 3.3.

However, this group (50 years and older) spans a range of ages which is much wider than the other groups: close to 40 years (from 50 years to 89 years exactly). If we regroup the respondents into age categories that are equal or almost equal, we get the chart in Figure 3.4.

This bar chart is a much better representation of the distribution of ages than the previous one.

In a **clustered bar chart**, each column is subdivided in several columns representing the categories of a second variable. For instance, each column could be split in two, for men and for women. Figure 3.5 provides an example of a clustered bar chart where the height of the columns represents the number of people in each category.

In a clustered bar chart, it is generally preferable to display the percentages of the various categories rather than their frequencies. Look for instance at the clustered bar chart displayed in Figure 3.5. We see that in every category, women are more

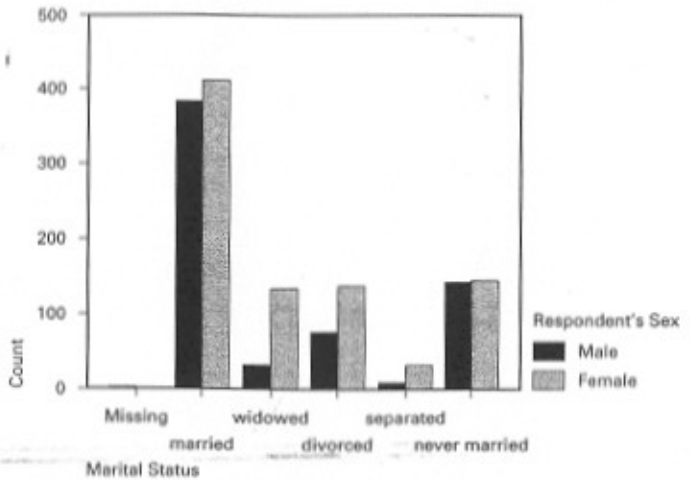

Figure 3.5 **A clustered bar chart where the height of the columns represents the number of people in each category**

numerous than men. This is so because the sample as a whole contains more women. This chart does not allow us to assess how the percentages of men and women compare in each category. If we display the percentages rather than the frequencies (the count), we get the chart illustrated in Figure 3.6.

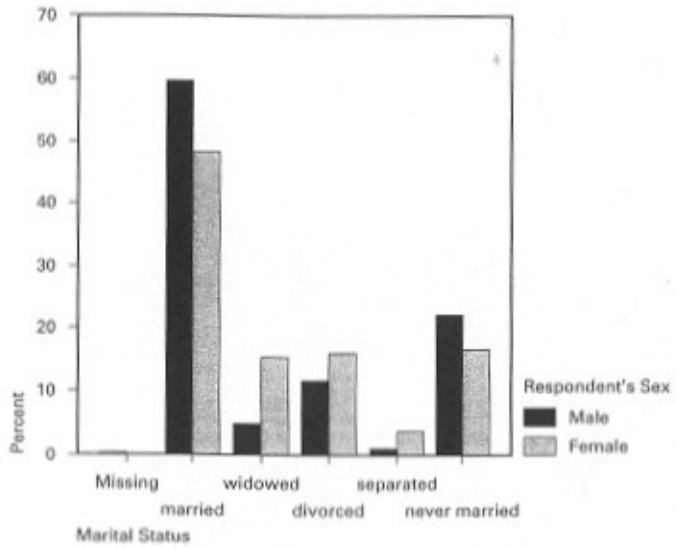

Figure 3.6 **A clustered bar chart displaying the percentages rather than the frequencies**

We see now that percentage-wise, there are a lot more women who are widows than men who are widowers. In that sample, it also happens that the divorced women are slightly more numerous than the divorced men (divorced women whose ex-husband has died are not counted in the Widow category but in the Divorced category). Although the sample used here is not necessarily representative of the whole American population, it does illustrate a social reality: as in many other societies, women tend to live longer than men. Therefore, the percentage of women in the categories Widowed and Divorced is larger than the percentage of men, and consequently lower than the percentage of men in all other categories, even if their numbers are bigger.

In a stacked bar chart, rather than being adjacent, the split columns are stacked one on top of the other, as shown in the Figure 3.7.

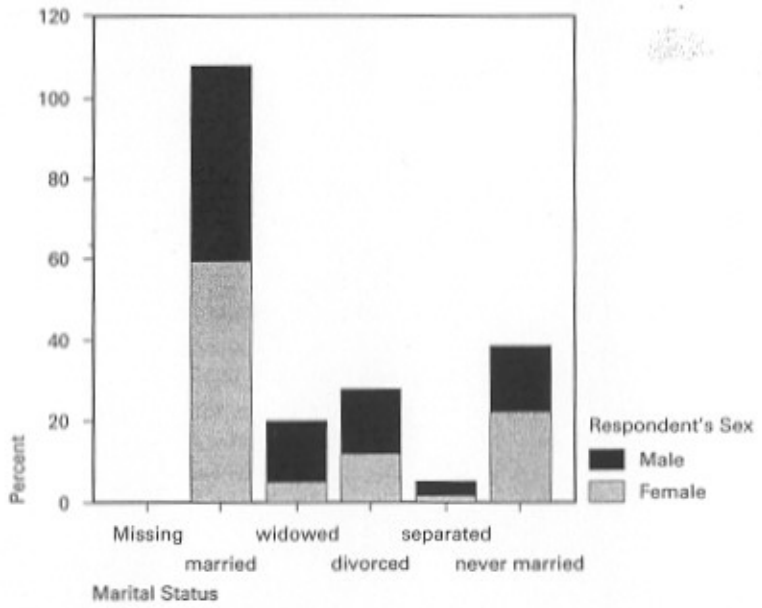

Figure 3.7 **A stacked bar chart**

The advantage of a stacked bar chart, as opposed to a clustered bar chart, is that it shows the overall importance of the categories (married, widowed, etc.), while at the same time showing how they are broken down into the categories of another variable such as Sex.

Bar charts are most adequate when you want to highlight the *quantity* associated with every category on the X-axis. A bar chart where the vertical axis does not start at 0 can be very misleading, for if the columns are truncated at their base, the

differences in height between them can appear to be more important than they really are. Consequently, as a general rule, bar charts should start at zero and should not be truncated from their base.

Finally, it should be said that bar charts could also be presented horizontally, by interchanging the X- and Y-axes.

Pie Charts

Pie charts (Figure 3.8) are most useful when you want to illustrate proportions, rather than actual quantities. They show the relative importance of the various categories of the variable. In SPSS you have the option of including missing values as a slice in the pie, or excluding them and dividing the pie among valid answers. The details of how to do that are explained in Lab 5. Pie charts are better suited when we want to convey the way a fixed amount of resources is allocated among various uses. For instance, the way a budget is spent over various categories of items is best represented by a pie chart. When the emphasis is on the amount of money spent on each budget item, rather than on the way the budget is allocated, a bar chart is more suggestive. However, both bar charts and pie charts are appropriate to represent the distribution of a nominal variable, and there is no clear-cut line of demarcation that would tell us which of the two is preferable.

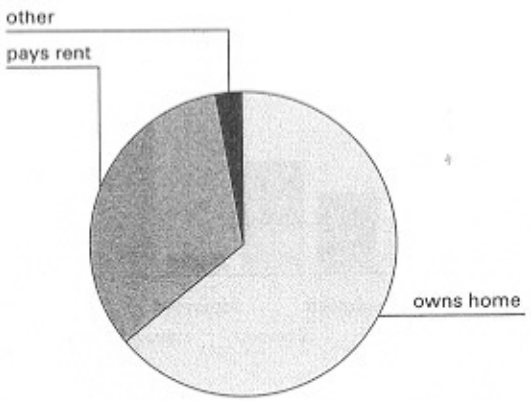

Figure 3.8 **Pie chart illustrating the proportion of people who own a home as compared to those who pay rent. One of the options in the pie chart command allows you to either include or exclude the category of missing answers. In this diagram it has been excluded from the graph**

Histograms

Histograms are useful when the variable is *quantitative*. The data are usually grouped into classes, or intervals, and then the frequency of each class is represented

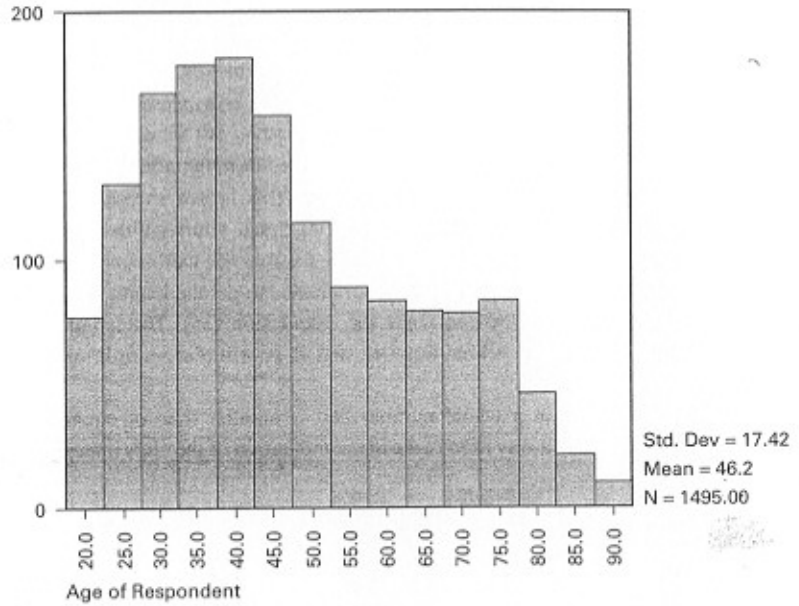

Figure 3.9 **Illustration of the histogram for the variable Age**

by a bar. The bars in a histogram are adjacent, and not separated as in a bar chart, because the numerical values are continuously increasing. For instance, if you draw the histogram of the variable Respondent's Age (Figure 3.9), you will see the pattern of the distribution of the individuals of the sample across the various categories. Contrary to a bar chart, which is used for a qualitative variable, the columns of the histogram cannot be switched around. You can switch around the categories of a variable measured at the nominal level, but not those of an ordinal or quantitative variable.

When producing a histogram with SPSS, the program automatically selects the number of classes (usually no more than 15) and divides the range of values accordingly into intervals of equal size. In the histogram shown in Figure 3.9, the **midpoints** of the classes are shown on the graph. They are:

20, 25, 30, 35, etc.

Therefore, the **class limits** (that is, the cut-point between one class and the next) are the values in between: 22.5, 27.5, 32.5, etc. We can infer that the lower limit of the first class is 17.5 years, and the upper limit of the last class is 92.5 years.