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.

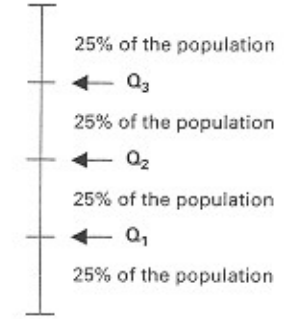

Figure 3.1 **The quartiles are obtained by ordering the individuals in the population by increasing rank, and then splitting it into four equal parts. The quartiles are the values that separate these four parts**

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**

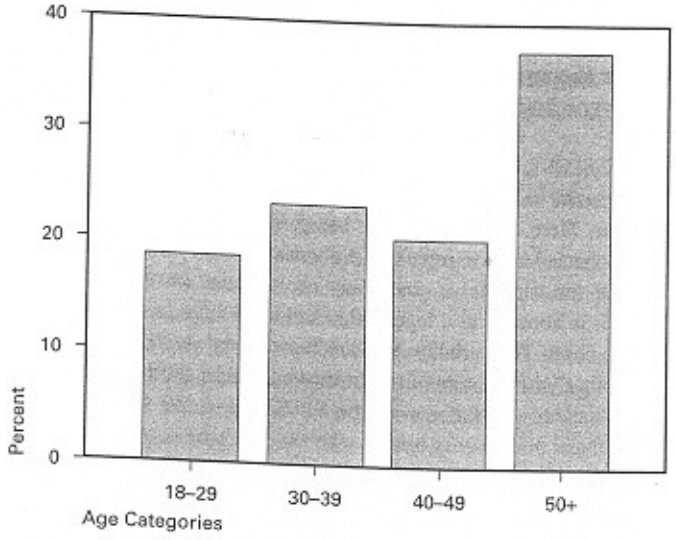

Figure 3.3 **A bar chart representing the various age categories in percentages of the whole sample**

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.

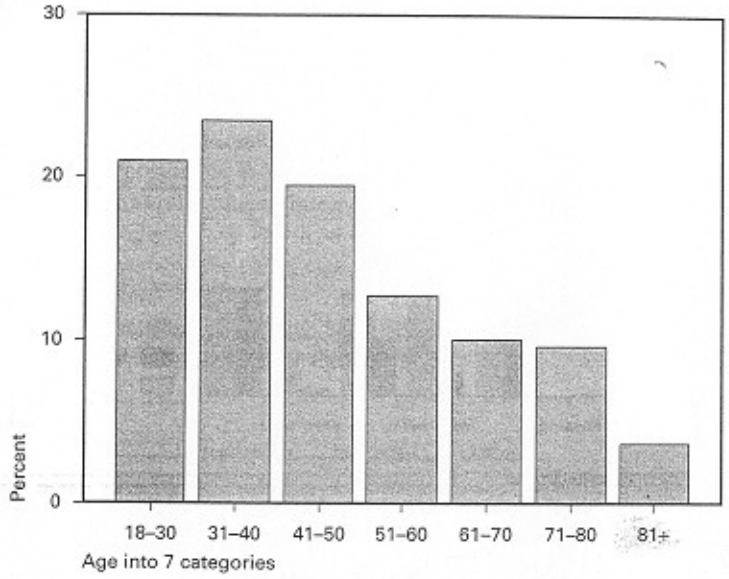

Figure 3.4 **A bar chart where the category 50+ years has been broken down into four categories**

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

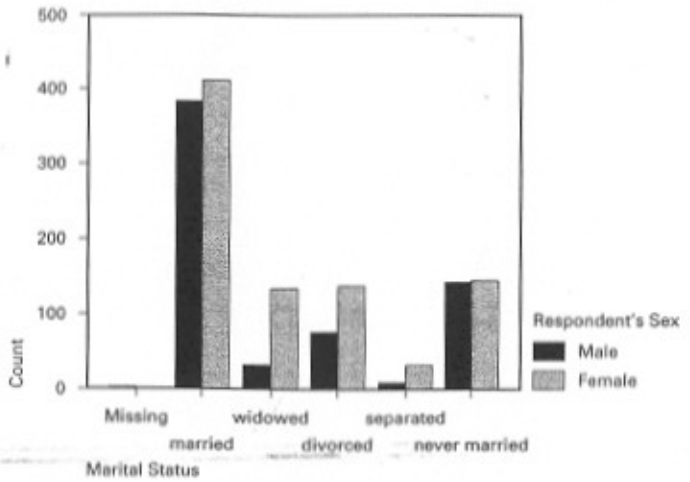

Figure 3.5 **A clustered bar chart where the height of the columns represents the number of people in each category**

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.

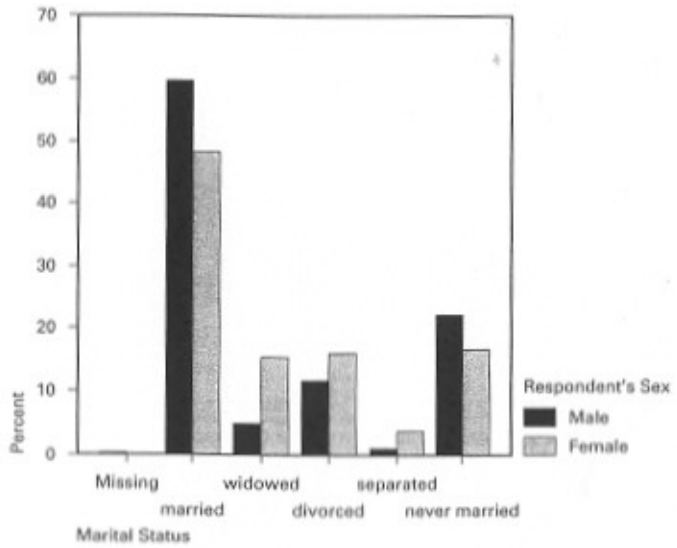

Figure 3.6 **A clustered bar chart displaying the percentages rather than the frequencies**

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.

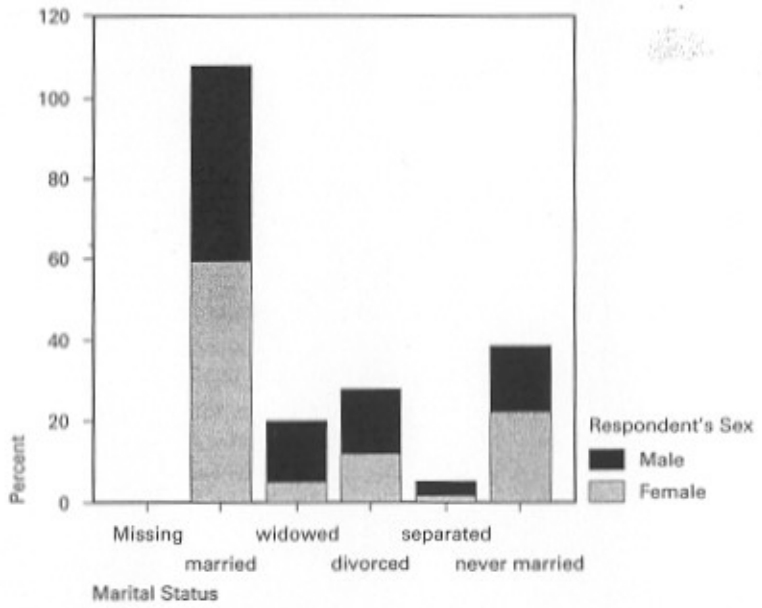

Figure 3.7 **A stacked bar chart**

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.

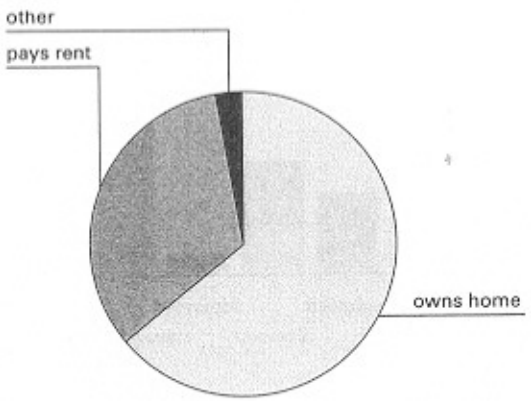

Figure 3.8 **Pie chart illustrating the proportion of people who own a home as compared to those who pay rent. One of the options in the pie chart command allows you to either include or exclude the category of missing answers. In this diagram it has been excluded from the graph**

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

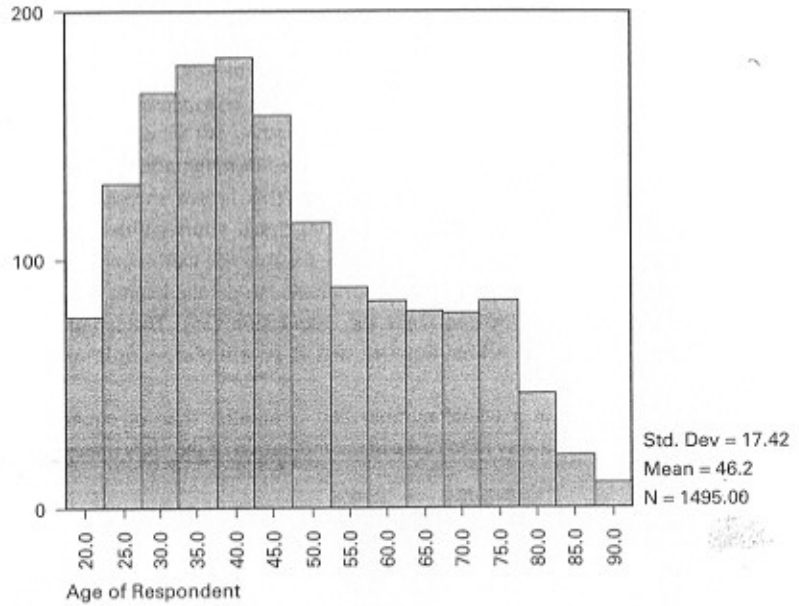

Figure 3.9 **Illustration of the histogram for the variable Age**

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.