to back with horizontal bars pointing to opposite directions, where each bar represents a five-year span. This type of histogram is called a **population pyramid**. In a population pyramid, the last class is left open. Usually it is the '80 years and more' class, as shown in Figure 3.10.

### FREQUENCY POLYGONS AND DENSITY CURVES

If we join all the midpoints at the top of the columns in a histogram, we get what is called a **frequency polygon**. The polygon shows the general pattern of the distribution. Imagine now a frequency polygon drawn on a histogram with very large number of columns. We could redraw it as a smooth curve, called a **density curve** (Figure 3.11). A density curve is drawn in such a way that its surface is equal to 1. And if we look at the surface under the curve between any two values, it tells us the exact proportion of data that falls within these two values. We can now be more specific about the definition of the mode for a quantitative variable.

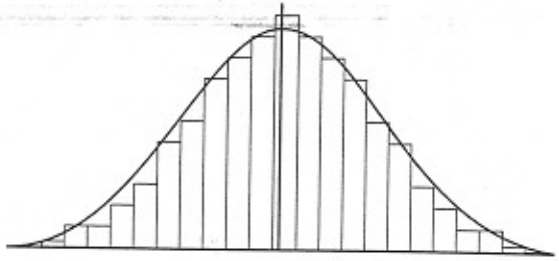

Figure 3.11 **A density curve can be thought of as the curve resulting from joining the midpoints at the top of the various bars of a histogram with a large number of classes**

If the variable is represented by a histogram, the **mode** is *the class with the highest frequency*. If it is represented by a density curve, the mode is *the x-value that corresponds to the highest point on the density curve*.

### HISTOGRAM OR BAR CHART?

When we have a quantitative variable that has been grouped into a small number of categories, we can represent it either by a histogram or by a bar chart. But which of the two representations is better? It depends on what we want to convey. To explain this point, consider a situation where we have the variable Age represented by seven categories as shown in Figure 3.4. If we want to convey how the ages of the sample studied are distributed over the whole range of ages, the histogram shown in Figure 3.9 is better. But if we want to show how the various age groups are divided among men and women, or among married vs. unmarried individuals, a clustered bar chart allows us to do that, as shown in Figure 3.12. A histogram would not permit us to juxtapose corresponding categories of age groups for men and women.

In Figure 3.12, we see the distribution separately for men and women, and we can determine that women are more represented in the older categories, as they tend to live



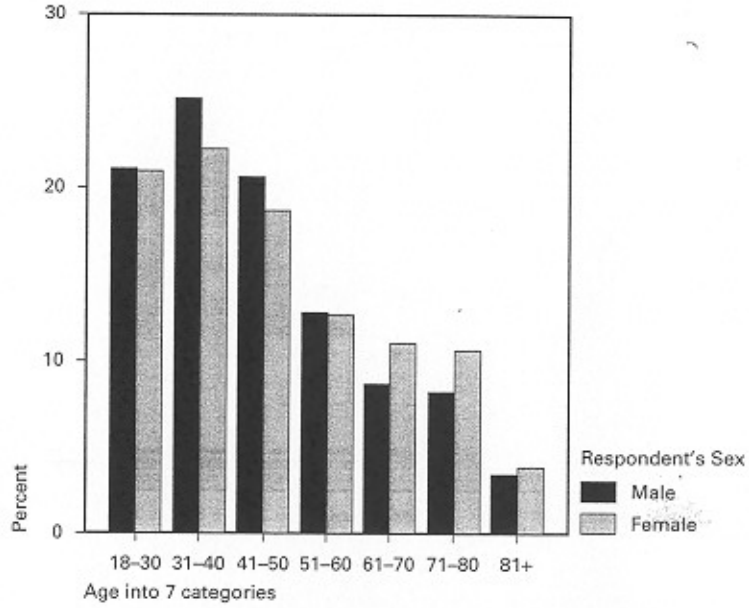


Figure 3.12 **A clustered bar chart allows us to show the pattern of ages separately for men and for women. It is appropriate for a quantitative variable grouped into a small number of categories**

longer than men. Notice that the vertical axis represents the percentages, not the frequencies. If we made it represent the frequencies instead, we would still get the same general shape, but we would not be able to determine whether men or women are more represented in a given class, as the overall number of women in this sample is greater than the overall number of men. In almost every age category, we would therefore find more women, not because a higher percentage of women (as opposed to men) fall into that category, but because there are more women in the sample as a whole.

## *Box plots*

Box plots are very useful to show how the values of a *quantitative* variable are distributed. The box plot indicates the minimum and maximum values, and the three quartiles. The central 50% of the data (the 2nd and 3rd quarters) are represented as a shaded solid box, whereas the first and last quarters are represented by thin lines.

The box plot gives automatically the *five-number summary* of the data: the minimum, the 1st quartile, the median (which is the 2nd quartile), the 3rd quartile, and the maximum.

In symbols the **five-number summary** is given by: **Min**, $\mathbf{Q_1}$, **Median**, $\mathbf{Q_3}$, **Max**. The box plot is shown in Figure 3.13.

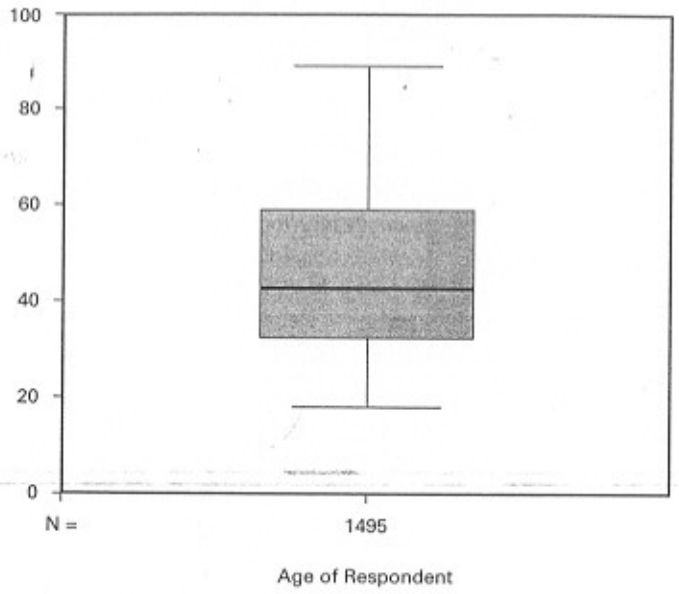


Figure 3.13 **The box plot representing the variable Age of respondent. We can read off directly the five-number summary of the distribution**

Box plots can also be used to represent several similar variables on the same graph, allowing comparisons. You could also split a population into several separate groups (such as men and women) and have a separate box plot for each group, drawn in the same graph, next to each other, to permit comparisons. This is illustrated in Figure 3.14 where five box plots of respondents' income are drawn, for the various groups defined by educational level. This figure illustrates clearly how the income varies with the highest level of education attained. We should note, however, that this data comes from a file where the income is not measured as a continuous scale variable, but is coded into 21 categories, and that the 22nd category is made up, as explained earlier in this chapter. More details are found in Lab 5.

### *Line Charts*

Line charts are most useful to represent the variation of a quantitative variable over time. The X-axis represents the time line, and the Y-axis represents some quantitative variable. For example, the variable could be the number of students enrolled in a given program, or the inflation rate, or the market value of a given portfolio of stocks. The line chart would show how the variable increases or decreases as time goes by. A common mistake sometimes made intentionally consists in not showing

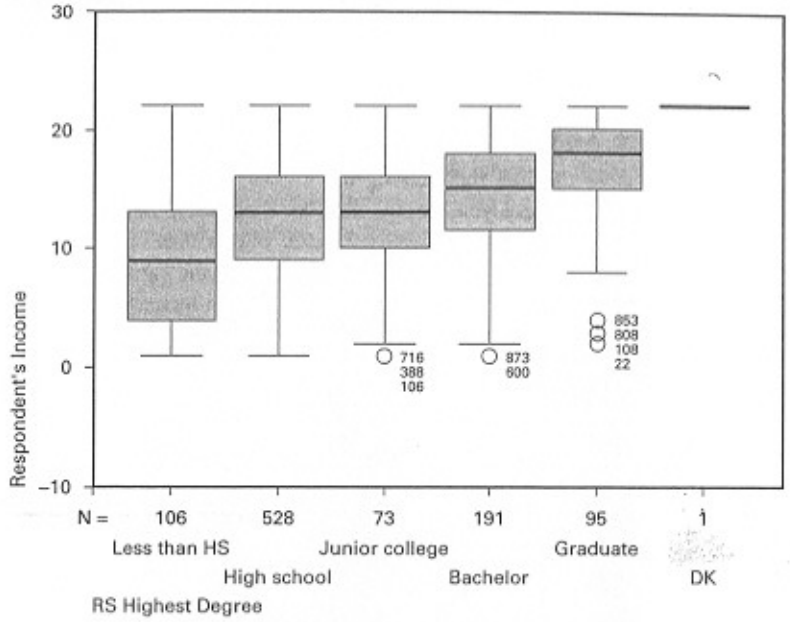


Figure 3.14 **Although the income is coded into 22 categories and not given as a dollar amount, the comparisons of the income for each educational level gives us a good idea of how incomes vary as a function of education**

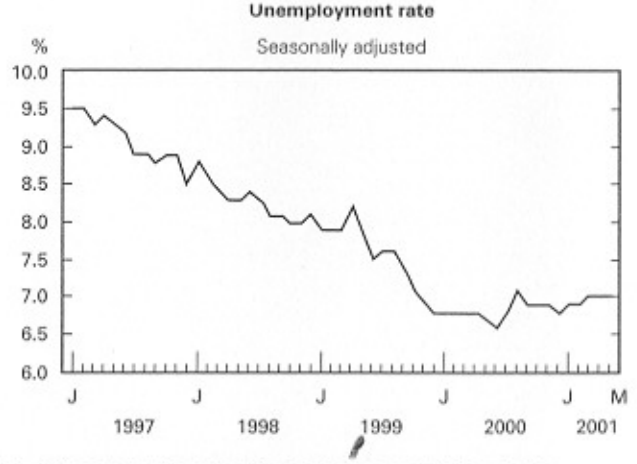


Figure 3.15 **A line chart showing the variation over time of the unemployment rate in Canada. The reader should be aware of the fact that the Y-axis does not start at zero, which may give the impression that the variations are greater than they really are. An awareness of this can guard us against misinterpretations. (Source: Statistics Canada)**

the zero level of the quantity, or in drawing the Y-axis shorter than it should be. This procedure has the effect of giving the impression that the variations are bigger than they really are, but at the same time it allows us to see the variations in the graph in greater detail. When it is necessary to show a shorter Y-axis, this should be indicated by an interruption in the line representing the Y-axis. Figure 3.15 provides an example of a line chart. Here you can see that the Y-axis does not start at zero, giving the impression that the variations are much bigger than they really are. However, this is justified by the fact that it allows us to see the variation in unemployment rates in great detail, and by the fact that this is an increasingly standard practice, which means that readers should be aware of the resulting distortion and interpret what they see accordingly.

## The General Shape of a Distribution

In addition to the measures explained above, we could describe the general shape of the distribution of a quantitative variable by looking at two of its features: *symmetry* and *kurtosis*.

### *Symmetry*

The first characteristic to look at is *symmetry*. A distribution is said to be **symmetric** if the mean splits its histogram into two equal halves, which are mirror images of each other. A typical symmetric distribution is the *normal distribution*. It is a bell-shaped distribution that follows a very specific pattern, and occurs in a wide range of situations. It is represented by the curve of Figure 3.16. It will be studied later on.

In a symmetric distribution, the mean and the median are equal. If the distribution is also unimodal, then the mean, the median, and the mode are all equal. This is true of normal distributions.

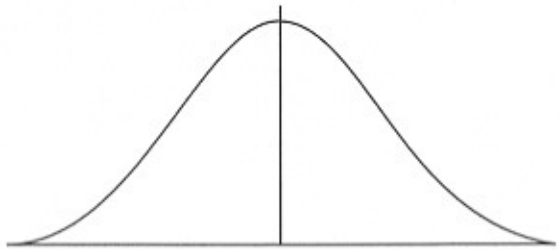

Figure 3.16 **An example of a symmetric distribution. This one is the normal distribution, which will be studied in Chapter 5**

If a distribution is symmetric and unimodal, the mean is a good representative of the center. However, it often happens that a distribution is not symmetric. We then say



that it is **skewed**. That means that one side of the graph of the distribution is stretched more than the other. We say that it is **positively skewed** if it is stretched on the right side, and **negatively skewed** if it is stretched on the left side. Figure 3.17 illustrates the difference between symmetric distributions and skewed distributions. SPSS allows you to compute a statistic called **skewness**, which is a measure of how skewed a distribution is. A normal curve has a skewness of 0. If the skewness is larger than 1, the shape starts to look significantly different from that of a normal curve.

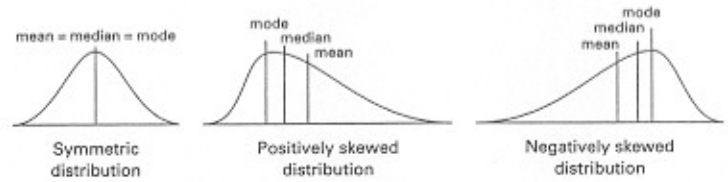


Figure 3.17 **Symmetric and skewed distributions**

How can we know that a distribution is skewed? The first indication is the histogram: the tail end of the histogram is longer on one side than on the other. We can also see that a distribution is skewed through its numerical features: the mean is different from the median. When the distribution is positively skewed, the mean is larger than the median, as it is pushed by the extreme values toward the longer tail. For negatively skewed distributions, the mean is smaller than the median. Therefore, a mean larger than the median tells us that the extreme values on the higher end of the distribution are much larger than the bulk of the data in the distribution, pulling the mean toward the positive side. This is illustrated by the numerical example given in the section on the median, where one extreme value (60) pulls the mean up but does not affect the median. Therefore, when a distribution is highly skewed, the median is usually a better representative of the center of the data than the mean.

### *Kurtosis*

This is a measure of the degree of peakedness of the curve. It tells you whether the curve representing the distribution tends to be very peaked, with a high proportion of data entries clustered near the center, or rather flat, with data spread out over a wide range. A normal distribution has a kurtosis equal to 0. A positive value indicates that the data is clustered around the center, and that the curve is highly peaked. A negative value indicates that the data is spread out, and that the curve is flatter than a normal curve. Figure 3.18 shows three curves with zero, positive, and negative kurtosis respectively.

## Methodological Issues

Although they seem to be simple, descriptive measures can be tricky to use. We would like to point out here some of the pitfalls and difficulties associated with their use.

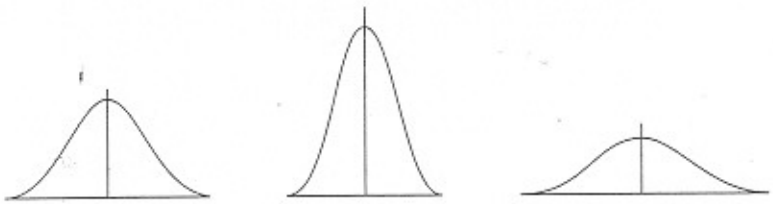


Figure 3.18 **Illustration of zero, positive and negative kurtosis**

### *The Definition of the Categories over which the Counting is Done*

Suppose I say that the passing rate in a given class is 82%. In another college, a colleague tells me that his passing rate is 95%. Before concluding that his passing rate is much higher, I have to make sure that we are defining the passing rate in the same way. I may define the passing rate as the number of students who pass a course compared to those who were registered at the beginning of the semester. If he defines it the same way, we can make meaningful comparisons. But if he defines it as the number of students who pass the course compared to the number registered at the end of the semester, we cannot make a meaningful comparison. This is so because all the students who dropped out would not be taken into account in his calculation, whereas they would be taken into account in mine. A careful definition of the categories used to define a concept is therefore important. Such problems arise when we define the unemployment rate in various countries, or even wealth. The conclusion is that careful attention should be given to the way categories are defined when comparing the statistics that refer to different populations.

### *Outliers*

**Outliers** are values that are unusually large or unusually small in a distribution. They have to be examined carefully to determine if they are the result of an error of measurement, or a typing error, or whether they actually represent an extreme case. For instance, the value 69 in the column of the variable **age** for college students could be a typing error, but it could also represent the interesting case of a retired person who decided to pursue a college program. Even if they represent an extreme case, it may be desirable to disregard extreme values in some of the statistical computations.

When producing a Box Plot diagram, SPSS excludes the outliers from the computation, and prints them above or below the box plot. An option allows users to have the case number printed next to the dot representing the outlier, so as to be able to identify the case and examine it more closely.



## Summary

We have seen in this chapter the various measures used to summarize the data pertaining to a single variable as well as the various types of charts that could be used to illustrate the distribution. You should keep in mind one fundamental point: *the level of measurement used for the variable determines which measures and graphs are appropriate*. It does not make sense, for example, to compute the mean of the variable when the level of measurement is nominal, that is, when the variable is qualitative.

There are three types of univariate descriptive measures:

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

Measures of central tendency, also called measures of the center, tell us the values around which most of the data is found. They give us an order of magnitude of the data, allowing comparisons across populations and subgroups within a population. They include the mean, the median, and the mode. The mean should not be used when the variable is qualitative.

Measures of dispersion are an indication of how spread out the data is. They are mostly used for quantitative data. The most important ones are the range, the interquartile range, the variance, and the standard deviation.

Measures of position tell us how one particular data entry is situated in comparison to the others. The percentile rank is one such measure. Other measures include the quartiles and the deciles.

In addition to these measures, we have seen the weighted mean. When calculating it, the various entries are multiplied by a weight, which is a positive number between 0 and 1. All the weights add up to 1. The weighted mean is used when the numbers that are averaged have been calculated over populations of unequal size. For instance, if you have the birth rates in all Canadian provinces and you want to find the average birth rate for Canada as a whole, you must weight these numbers by the demographic importance of every province. The weighted mean is also used when you want to increase or decrease the relative importance of the numbers you are averaging, as is done when finding the average grade over exams that do not count for the same percentage in the final grade.

When categories are involved (either because the variable is qualitative, or when quantitative values have been grouped) we can find ratios, percentages, and proportions of the groups corresponding to the categories.

The general shape of a distribution is analyzed in terms of symmetry or skewness, and in terms of kurtosis (the degree to which the curve is peaked).

The comparison of the mean and the median is very useful. Recall the following:

> If the distribution is very skewed, the median is a better representative of the center of the data, as the extreme values tend to pull the mean towards one side of the curve. The median is not affected by extreme values.
>
> If the mean is larger than the median, the distribution is positively skewed. If the mean is smaller than the median, the distribution is negatively skewed.

As for the graphical representation of a distribution, recall again that the level of measurement of the variable determines what kind of chart is appropriate. Bar charts and pie charts are appropriate when the data is qualitative, or measured at the nominal or ordinal levels. Quantitative data (whether measured at the ordinal of numerical scale levels) could also be represented by bar charts or pie charts if the values have been grouped into a small number of categories.

The essential difference between pie charts and bar charts is that in the former, the emphasis is on the relative importance of each category as compared to the other categories, whereas in the latter, the emphasis is on the size of each category. However, there is no clear-cut distinction between the two, and if one is appropriate, the other is usually appropriate also, even if the emphasis is slightly different. The great advantage of bar charts is that it allows making comparisons between the distributions of subgroups, with the help of clustered bar charts.

Quantitative variables are better represented through histograms. A specific type of histogram is the population pyramid, which is a standard tool in demography.

Line charts are most suited to represent the variation of a quantity across time.

In all kinds of charts, truncating the Y-axis is sometimes done to zoom in on the variations of the variables and to represent them in a more detailed way. However, we should be aware of the fact that truncating the Y-axis may also convey a mistaken impression that the variations of the variable are more important than they are in reality.

## Keywords

| | | |
|---|---|---|
| Univariate | Frequencies | Ratios |
| Bivariate | Cumulative frequencies | Proportions |
| Measures of central tendency | Valid percent | Bar graph |
| Measures of dispersion | Range | Clustered bar graph |
| Measures of position | Trimmed range | Pie chart |
| Mean | Interquartile range | Histogram |
| Trimmed mean | Deviation from the mean | Frequency polygon |
| Weighted mean | Standard deviation | Line chart |
| Median | Variation ratio | Box plot |
| Mode | Coefficient of variation | Five-number summary |
| Modal category | Quartiles | Symmetry |
| Majority | Deciles | Skewness |
| Plurality | Percentiles | Kurtosis |
| | Percentile rank | Outliers |



## Suggestions for Further Reading

Devore, Jay and Peck, Roxy (1997) ***Statistics, the Exploration and Analysis of Data*** (3rd Edn). Belmont, Albany: Duxbury Press.

Harnett, Donald H. and Murphy, James L. (1993) ***Statistical Analysis for Business and Economics***. Don Mills, Ontario: Addison-Wesley Publishers.

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

Wonnacott, Thomas H. and Wonnacott, Ronald J. (1977) ***Introductory Statistics*** (3rd edn). New York: John Wiley and Sons.

## EXERCISES

3.1 Complete the following sentences:

(a) Three types of measures are useful to summarize a numerical distribution. They are

_____,
_____ and
_____.

(b) The most frequent value in a distribution is called _____.

(c) When the values of the distribution are grouped into classes, the mode is the _____ with the highest frequency.

(d) When there are two classes that are bigger than the ones immediately next to them, the distribution is called _____.

(e) If the modal class includes more than 50% of the population, we say that it constitutes the _____. Otherwise, we simply talk of a _____.

(f) The median falls _____ of the ordered list of entries. _____% of the data are less than or equal to the median, and _____ % are larger than or equal to it.

(g) The mean of a numerical distribution is equal to the _____ of all entries divided by _____.

(h) The mathematical measure used to find the mean when the entries do not have the same relative importance is called _____.