2

An Introduction to Descriptive Statistics: Data Description and Graphical Representation

2.1. DESCRIPTIVE STATISTICS

In the previous chapter we defined statistics as a science to help us collect, organize, analyze, and interpret empirical data. We also introduced a four-stage process for applying statistics in empirical research, and we discussed the first two phases of problem definition and data collection. In this chapter, we deal with the third stage of this process, which is concerned with the summarization of data. The summarization of data comprises an important part of statistics that can be further divided into two main branches—descriptive and inferential statistics. This and the next chapter attend to some essential features of descriptive statistics. Later chapters of the book will be concerned with aspects related to inferential statistics.

Descriptive statistics represents a set of methods and activities that permit the description of a given body of data, without making inferences about another set of possible observations (from a larger group, e.g., a population). Descriptive statistics is also used when a whole population of subjects (or aggregates of such) is observed, such as in a population census. For example, if census data are available on income levels of adults living in the United States, we can obtain various descriptors of income, e.g., broken down by gender, race, ethnicity, geographical location, etc. Descriptive statistics enables us to represent potentially vast amounts of data in ways that permit the reader, consumer, or user of the results to obtain relatively quickly an informative summary and even provide a graphical presentation of the data. When an entire population cannot be studied exhaustively, a sample is drawn from it, and descriptive statistics helps summarize and present the sample data collected.

Descriptive statistics includes methods that accomplish (a) various graphical displays providing valuable insights to the scientist or user, and (b) numerical data descriptions via summary indexes that can contain important information about research questions of interest. In this chapter we discuss methods that fall under category (a), and we will be concerned with methods in category (b) in the next chapter. Because graphical displays differ depending on the number of variables involved, we cover first the case of just a single variable. Situations involving several variables will be attended to in later sections of the book. As far as numerical indexes are concerned, we note that we will familiarize ourselves with several of them in Chapter 3, where we will also see in detail how they summarize important features of sets of scores in given samples (or populations).

2.2 GRAPHICAL MEANS OF DATA DESCRIPTION FOR A SINGLE VARIABLE

A *main principle* to follow when graphically presenting collected data from studied samples (populations) is that data should be arranged in any display in such a way that *any single measurement falls within only one possible category*. Thereby, we need in particular to make sure that the boundaries of adjacent categories do not overlap and are not identical in any possibly misleading way. We demonstrate this potential problem on examples used later in this section.

We begin by considering the following example study:

Example 2.1. Suppose we have obtained data on the number of faculty that are employed in a college of education at a particular state university. Let us also assume that the observations given in Table 2.1 represent the obtained faculty data. As indicated earlier, we need to make sure that each single measurement falls within only one possible category. In other words, we need to ensure either that there are no faculty members with a joint appointment in another department, or that any faculty member with such an appointment is only counted in one of the departments he or she is affiliated with. If we ignore

Table 2.1 Number of faculty from five departments in a College of Education at a state university (Example 2.1).

Department	*Number of Faculty*
Educational Psychology	12 faculty
Teacher Education	13 faculty
School Psychology	9 faculty
Measurement, Evaluation, and Statistics	4 faculty
Educational Administration	15 faculty

such a situation (joint appointment), then the actual total number of faculty counted will be incorrect.

Since statistics is a science that deals by definition with potentially sizable amounts of data, it is usually easier to achieve the goals of statistics via the use of a statistical package such as R rather than doing things manually. However, in order to be able to use R for conducting descriptive or other statistical types of analyses, it is necessary that we first communicate to R the data set we wish to use. Once this communication has occurred, R will be in a position to analyze the data set on our behalf accordingly. We attend next to these initial issues.

2.2.1. Reading data into R

The process of communicating a given data set to R typically represents the first activity that a researcher becomes involved with in the process of applying descriptive statistics. This process is often simply referred to as the "reading" of the data set into R. This is a routine activity that is in fact the first step of preparing the data in a form amenable to any kind of analysis using statistical software such as R.

In order to read data into R, it is convenient to enter the data set initially into a universal file format, such as a "plain text" or "text only" file, often referred to as an "ASCII file."* This can be achieved if one enters the data into a window opened when starting a standard text editor like Notepad (or WordPad, either of which is available on a Windows-based PC under "Accessories" after clicking the "Start" button on the computer screen, followed by selecting "All Programs"). Once we are in that window, the data must be typed in such a way that *each unit of analysis represents a single row* (department in this example—although in many cases within social and behavioral studies it will be individual subject data that comprise the unit of analysis). Therefore, it will be very beneficial if we ensure that within each row the consecutive scores on the measured (collected or recorded) variables are *separated by at least one blank* space. This format of data entry is often referred to as "free format" and usually is the most convenient format to work with subsequently. It is also typical to give names to each column of the resulting data file, in its top row, which names are those of the variables they represent. This practice is highly recommended in order to keep track of what the considered data set represents.

With this in mind, for our above Example 2.1 we first create an ASCII file

* ASCII stands for "American Standard Code for Information Interchange" and is a character-encoding scheme for representing text to computers and digital devices that use text.

that is similar to the one presented in the main body of Table 2.2, using also for simplicity the abbreviation "dept" for "department" (cf. Table 2.1). We note that the first column of Table 2.2 contains the abbreviation of the name of the department (initials), and the second column is the number of faculty in each department. We also notice that the first row of this file contains the names of the two variables—'department' and 'faculty'. We save this data file using an easy-to-remember file name, such as CH2_EX21.dat.

Now in order to read this small data set into R, we need to use a particular command. Before we give it, however, we note that for ease of presentation we adopt the convention of using the Courier font to represent both the command submitted to the R software and the output sections produced by the statistical analysis of the program itself. This same convention for commands and output will be followed throughout the book. We also begin the representation of each used command with the conventional R prompt, which is the sign ">" (we emphasize that when one is actually using a computer to type a command, there is no need to precede any R command with the prompt, as it is generated automatically). With this in mind, the command to read the data from Example 2.1 into R (as provided in Table 2.2) is 'read.table', which we use here as follows:

```
> d = read.table("C://data/CH2_EX21.dat", header = T)
```

We note that one needs to provide with this command the path leading to the specific subdirectory where the ASCII/plain text file resides on the computer used, placed in inverted commas. (In this case, the file is located in the computer subdirectory "data" in drive C; see the Note at the end of this chapter.) The subcommand 'header = T' instructs R to read the first line not as numbers but as variable names (i.e., as "character strings"—sequences of characters or letters; see Note to this chapter).

When the R command 'read.table' is executed by the software, it creates an *object* with the name 'd'. The object named 'd' represents the data that have

Table 2.2 Contents of data file for number of faculty per department (Example 2.1).

dept	*faculty*
ep	12
te	13
sp	9
ms	4
ea	15

been read in by the program. We observe at this point that R is a case-sensitive software package, and we note in passing a simple rule to keep in mind: in a sense everything in R is an object. (For example, an output from an R command is also an object.) To be in a position to analyze data on any of the variables in the data set under consideration, we need next to make it accessible to R. This is accomplished with the command 'attach' as is illustrated in the command line below (for the data set CH2_EX21.dat that was made available in the created object 'd'):

```
> attach(d)
```

For the remainder of the book, whenever an example is considered for analysis, we will assume that the data from the study under consideration have already been read into R and made accessible to R in the aforementioned specific way.

2.2.2. Graphical representation of data

There are numerous ways available within R to represent graphically data on a variable of interest. There are also a multitude of sources available that provide detailed descriptions of each of the different graphical ways. It is beyond the scope of this chapter and book to discuss all of them. In this chapter we only present detailed descriptions of a few important graphical ways for two main types of variables—qualitative and quantitative. Qualitative variables, also frequently referred to as categorical variables, have "scores" (or values that they take) that differ from one another in kind rather than in quantity. That is, the data values on a qualitative variable are in fact just classifications or categories, which is the reason that data stemming from such variables are at times referred to as categorical data. The categories of a qualitative variable are also often referred to as levels (or even labels) of the variable. Example 2.1 illustrates such a variable—the variable 'department' is a qualitative (categorical) variable.

In contrast to qualitative variables, a quantitative variable has scores (or values) that differ from one another in quantity. That is, their scores—the actual data—are "legitimate" numbers, i.e., as close as possible to real numbers in meaning. For example, the following mathematics test–taking anxiety (MTA) study illustrates the inclusion of such a quantitative variable, the anxiety score.

Example 2.2 (MTA study): A mathematics test–taking anxiety study was carried out with $n = 36$ elementary school students, using an established measuring instrument referred to as the MTA scale. The resulting data are provided in Table 2.3 and saved under the name CH2_EX22.dat. In Table 2.3, 'id' denotes a

Table 2.3 Data from $n = 36$ students in a study of mathematics test–taking anxiety (the first row represents the names of the two variables involved: id = person identifier, y = anxiety score).

id	y
1	15
2	17
3	19
4	21
5	23
6	32
7	18
8	19
9	22
10	23
11	25
12	21
13	22
14	25
15	25
16	24
17	23
18	27
19	19
20	17
21	21
22	29
23	27
24	31
25	19
26	19
27	23
28	25
29	25
30	27
31	29
32	29
33	21
34	24
35	23
36	28

subject identifier, and ‘y’ the score for each student obtained using the mathematics test–taking anxiety measuring instrument (i.e., the student MTA score). For ease of presentation, we adopt here the convention that when representing subject data in this book, we will typically use the first column as the person (case) identifier.

Next we use the two previously discussed examples (see Tables 2.2 and 2.3) to pictorially represent the data on both the qualitative and quantitative variables involved. This is done in order to obtain a first idea of the relationships between the values (scores or categories) that the subjects (the units of analysis) take on them. Two very popular yet distinct methods for graphically representing data on studied variables are discussed next. First, some methods that can be used to graphically represent qualitative variables are presented. These are followed by an illustration of methods for graphically representing quantitative variables.

2.2.2.1. Pie charts and barplots

For a given qualitative variable of interest, the *pie chart* and *bar chart* (sometimes also referred to as *barplot* or bar graph) are two common methods that can be used to graphically represent the frequencies of its obtained values in a particular data set. The charts display graphically (correspondingly either in terms of pie slices or bars) the observed frequencies. Specifically, the size of the pie or the height of the bar represents the frequency of the pertinent data value. Usually, the pies are presented in relative frequency—in other words, their sizes can be interpreted as percentages. The heights of the bars often represent “raw” frequencies but have the same shape when considered for relative frequencies. The relative frequencies are defined as the ratios of observed or raw frequencies to the total number of units of analysis (i.e., usually subjects in this text) in the group studied.

To illustrate this discussion, suppose we wish to obtain the pie chart for the ‘faculty’ variable in Example 2.1 (see data in Table 2.2). A pie chart graphically represents the data in the form of slices of a circle, which are typically filled in with different colors, and their sizes reflect the relative frequencies with which the variable in question takes its values. To obtain a pie chart graph, we use the R command ‘pie’ that we state at the R prompt. We use thereby as an “argument”—i.e., an entity we place within parentheses following the actual command—the name of the variable that needs to be graphed, which is here the variable ‘faculty’:

```
> pie(faculty)
```

This command yields the pie chart displayed in Figure 2.1—this pie chart is shown by R in a separate graphics device window opened by the software after one completes the command (i.e., after one submits the command to R).

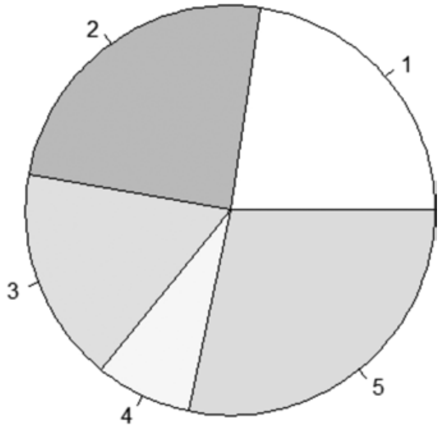

FIGURE 2.1.
Pie chart of the number of faculty per department.

While Figure 2.1 readily presents a rough idea about the relations between the number of faculty members in each department, the labeling of the displayed slices is not immediately informative. In particular, the specific numbers attached to the slices here by R are simply those of the consecutive rows of Table 2.1. Such a graphical display does not provide a clear enough picture of the data and would require one to refer back to the raw data in order to interpret the relative sizes of the slices presented. To deal with this limitation of the displayed figure, it would be best to add the names of the departments (in the abbreviation notation used in Table 2.2). This is possible in R by using the subcommand 'labels', and we attach a title to the figure using the subcommand 'main', leading to the following extended command:

```
> pie(faculty, labels = dept, main = "Piechart of Faculty by Department")
```

As can be seen from this command line, we set the subcommand 'labels' equal to the name of the variable containing those of the departments, and set the subcommand 'main' equal to the title we want to use, which we place in quotation marks. To simplify matters throughout the remainder of the book, we will always make a reference to any R commands and subcommands within the main body of the text by using quotation marks. Using the above pie-chart command produces the output displayed in Figure 2.2.

The pie chart as a graphical device is very useful when one is interested in displaying qualitative data in the form of slices in a circle. Another very popular alternative and equivalent data presentation method is the so-called bar chart (barplot). In its simplest form, in a barplot the levels of a variable under consideration are presented successively (following the order from top to bot-

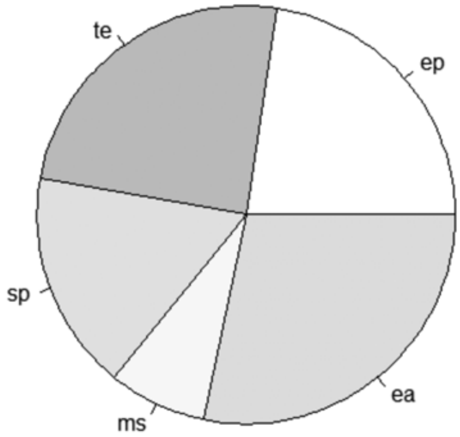

FIGURE 2.2.
Pie chart of faculty by department, using slice labeling and figure title.

tom in the data file), and the frequencies (relative frequencies) with which the variable takes its values are represented by the height of bars—vertical rectangles positioned above these levels. For our currently considered example, we can obtain the barplot of the 'faculty' variable in R with the command 'barplot' as follows:

```
> barplot(faculty)
```

This command yields the graph displayed in Figure 2.3. Along its horizontal axis, the departments are represented from left to right as they follow from top to bottom in the original data file. Above them are the bars, whose heights reflect the number of faculty members per department.

While we can easily obtain from Figure 2.3 an idea about the relations between the number of faculty across each department—which numbers are represented along the vertical axis—it is unclear from the figure alone which department is associated with which of the bars. To assign the names of the departments to the horizontal axis, we use the subcommand 'names.arg', and as above use the subcommand 'main' to attach a title to the figure. The resulting, somewhat extended command we need for these aims, is now as follows:

```
> barplot(faculty, names.arg = dept, main = "Relative Frequency of
Faculty by Department")
```

This extended command furnishes the bar chart displayed in Figure 2.4.

This bar chart is now far more informative than the one displayed earlier

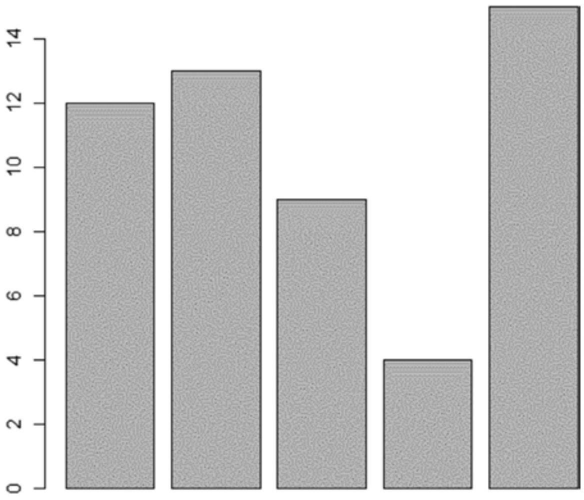

FIGURE 2.3.
Barplot of faculty by department.

in Figure 2.3, and in addition may be seen as more informative than the pie chart presented in Figure 2.2. The reason is in particular the fact that by referring to the vertical axis one can see the values that the ‘faculty’ variable takes across departments. These are the above-mentioned raw frequencies that equal the number of faculty per department. If we wish to obtain a bar chart

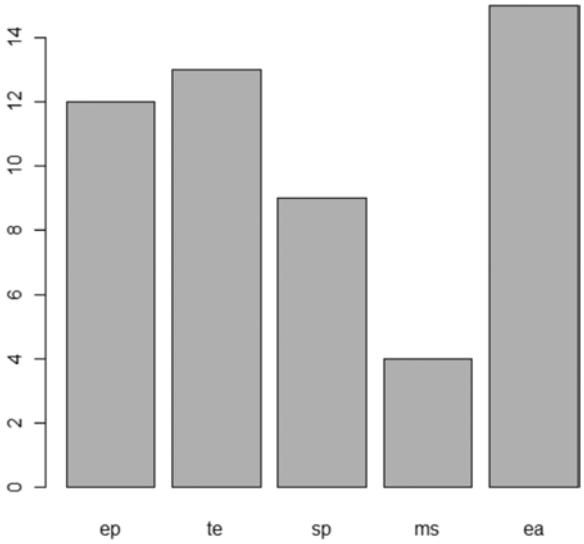

FIGURE 2.4.
Barplot of faculty by department.

not of the raw frequencies as in Figure 2.4, but of the relative frequencies, we can divide the raw frequencies by the total number of cases in the data set—here the total number of faculty at the college in question. In general, as mentioned before, the relative frequency for a given variable category ('department' in this example) is defined as the ratio of observed frequency, denoted for clarity as *f*, to the total number of cases studied—denoted *n*—which is the sample size in a study under consideration:

$$r = f\,/\,n\,, \tag{2.1}$$

where *r* is used to denote relative frequency. That is, to obtain the relative frequency of faculty per department in our example, for each department we need to divide their number by the total sample size. In most studies one would know the sample size already at the beginning of the analysis, but at times it may be necessary to obtain it alternatively (e.g., by having the R software do it for us). We can simply obtain sample size by summing the values of the variable 'faculty' across all departments, which is accomplished with the R command 'sum':

```
> sum(faculty)
```

This command yields the following result for our example:

```
[1] 53
```

In this output, the first number presented in brackets, [1], indicates that immediately following is the first element of the output produced by R. Since in this case all we need from R is a single number (i.e., the studied group size, or sample size), it is presented right after the symbol "[1]." Thus, 53 is the total number of faculty in the college under consideration. We will often encounter this simple pattern of output presentation in the remainder of the book.

Once we know the group size, we can request from R the barplot of relative frequencies per department using the above command 'barplot', where we now divide the variable in question by this size. We achieve this by using the division sign '/' as follows:

```
> barplot(faculty/53, names.arg = dept, main = "Barplot of Faculty by
Department")
```

We stress that the first argument in this command (the first entry in its parentheses, from left to right) is formally the ratio of the variable in question to studied group size (sample size). The last presented, extended 'barplot' command produces the relative frequencies bar chart displayed in Figure 2.5. From this barplot, it can now be readily observed that the largest percentage of faculty are employed in the Educational Administration Department, about a quarter of all faculty are in the Teacher Education Department, less than 10% are in the Measurement, Evaluation, and Statistics Department, and about 17% of all faculty are employed in the School Psychology Department.

2.2.2.2. Graphical representation of quantitative variables

The above section dealing with the pie charts and barplots demonstrated how one can graphically represent a variable that is qualitative. Specifically, we considered some ways to present the frequencies with which the variable ‘department’ takes its values (categories) or levels—as measured by the number of faculty affiliated with each department. That is, the variable ‘department’ had the values (levels) ‘ep’, ‘te’, ‘sp’, ‘ms’, and ‘ea’—for the names of departments—which we might as well consider simply labels for the departments. The frequencies associated with these labels were formally presented in the variable named ‘faculty’. (We note that we could just as well have also called the ‘faculty’ variable ‘frequency’, which it actually is with regard to the levels of the ‘department’ variable). Since the levels (categories) of the variable ‘department’—being the different departments in the college in question—differ from one another only in kind rather than in quantity, ‘department’ is considered a qualitative variable. For such qualitative variables, the pie chart and barplot are very useful and informative methods to graphically represent

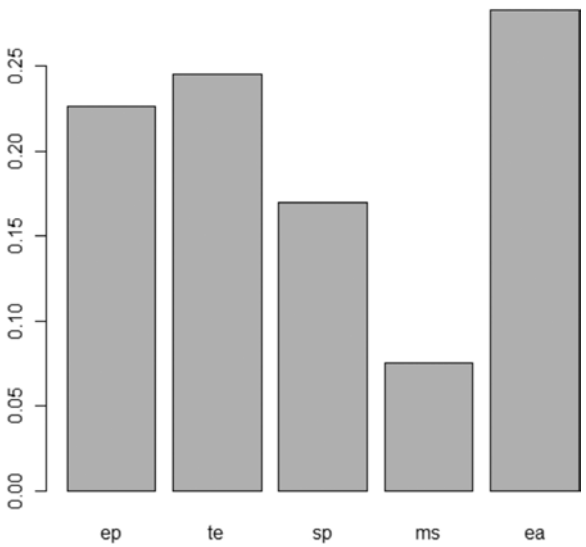

FIGURE 2.5.
Relative frequencies bar chart for faculty by department.

the frequencies or relative frequencies with which these variables take their values or levels.

Although qualitative variables are often encountered in empirical research, data arising from quantitative variables are just as common (or at least from variables that could be treated as quantitative). As mentioned earlier, the values of a quantitative variable differ from one another in quantity rather than in quality as is the case for a qualitative variable. For quantitative variables, it is equally necessary to have methods that can be used to graphically represent their values in a way that provides an informative summary of them.

A very popular method that can be used to graphically represent data from a quantitative variable is the so-called histogram. A *histogram* is basically a series of vertical rectangles that represent the frequency with which scores fall in the interval that is at the bottom of that rectangle. Fortunately, these intervals, including their length and position, are automatically chosen for us by the R software through some built-in, reasonable, and widely applicable defaults. In fact, with R the intervals are defined by default as including the number positioned at their right end, but not at their left end. With this feature, the histogram gives an impression of what the *distribution* of a variable under study actually looks like. We mention in passing that the distribution is the way in which scores on the variable relate to one another, and we will have more to say about this concept in a later section of the book.

To illustrate the construction of a histogram, let us return to Example 2.2 where we were interested in the mathematics test–taking anxiety (MTA) variable in a study with a sample of $n = 36$ students. Suppose we wish to construct the histogram of the variable 'anxiety score' (i.e., the MTA score obtained with an established measuring instrument used in that study); we recall that this variable was denoted as 'y' in Table 2.3. Now using this data file, R can readily produce a histogram for the variable with the command 'hist':

```
> hist(y, main = "Histogram of anxiety scores")
```

We note that in this command the name of the variable for which a histogram is to be obtained is given first (which here is denoted as 'y'). In addition, the subcommand 'main' is used to attach a title to the top of the generated histogram. Figure 2.6 provides the histogram for the MTA scores generated using this command statement.

As can be readily seen from this histogram, there are for example in total seven scores that are either 23 or 24, since the bar corresponding to the interval (22, 24] has a height of seven units—i.e., the frequency of scores in this interval is seven. (We note that the symbol "(.,.]" is used to denote an interval including the number at its right end but not the number at its left end.) Similarly, there are three scores in the interval (16, 18], as can also be verified by inspecting the data in Table 2.3. (Below we also illustrate an alternative

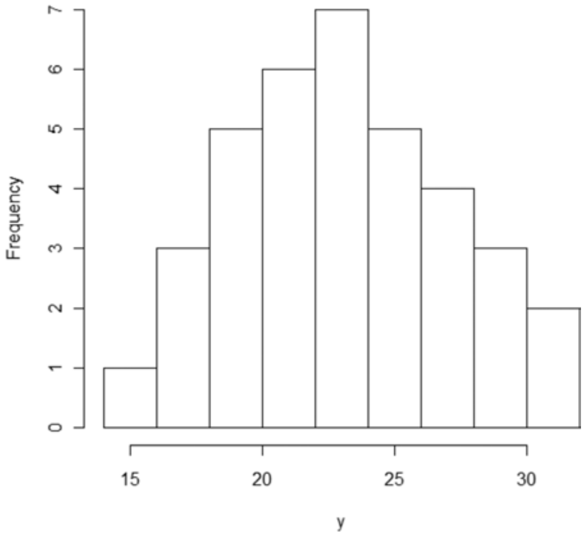

FIGURE 2.6.
Histogram of mathematics test–taking anxiety scores (Example 2.2).

way of achieving the same graphical representation aim using the so-called stem-and-leaf plot.)

When informally presenting a histogram, often one may wish to connect the middle points of the top sides of these rectangles. If one uses thereby corresponding segments of straight lines, the resulting curve is commonly referred to as a frequency polygon. If one connects these middle points by a smooth curve rather than segments, the resulting curve is informally also referred to as variable "distribution" curve (or just distribution). We will discuss this alternative notion for graphical presentation of data in more detail in a later section of the book.

Histograms are very useful devices to present data, but depending on how their class intervals are built, they can sometimes conceal some important information (e.g., Verzani, 2005). For example, such a situation may arise when grouping together fairly different scores (as may happen with variables that have wide ranges of scores). For this reason, it can be particularly informative to examine the actual frequency with which a particular score appears in a data set. This is readily accomplished with the *stem-and-leaf plot*. In it, the *stem* is composed of all numbers apart from their last digit, and the *leaves* are their last digits.

Using the same anxiety score data for Example 2.2 considered above, we can obtain their stem-and-leaf plot (graph) with R using the command 'stem'. Thereby, the subcommand 'scale' used requests that the last digit of the scores be presented as leaves, and all the preceding digits as stem:

```
> stem(y, scale = .5)
```

```
1 | 577899999
2 | 1111223333344
2 | 555557778999
3 | 12
```

FIGURE 2.7.
Stem-and-leaf plot of the anxiety scores (Example 2.2).

The generated stem-and-leaf plot is presented in Figure 2.7.

This plot reveals some additional detail about the considered data set. For example, looking at the first row of the plot, it can be readily determined that there is only one score of 15 in the data set under consideration, two 17's, one 18, and five scores of 19. Similarly, looking at the second row it can be determined that there are four 21's, two 22's, five 23's, and two 24's. Similar details can be revealed by examining the remaining aspects of the stem-and-leaf plot. Such details are readily seen by connecting each stem with each of the leaves on the right from the vertical bar within each row of the plot.

While graphical approaches can at times be excellent visual aids, they can also present data in ways that allow different and possibly even contradictory subjective interpretations. As a consequence, it is recommended that conclusions about populations of interest based only on sample data graphs not be hastily drawn. As it turns out, there are more advanced methods that should be used in order to reach such conclusions in an objective manner. A more detailed discussion concerning these fundamental approaches will be provided in a later section of the book. To begin this discussion, we consider in the next chapter some simple numerical indexes that can be initially used to objectively summarize data.

NOTE

In the remainder of the book, when explicitly including R commands for reading data, we will indicate the corresponding paths where the data sets employed reside on the computer used by the authors. (For most examples, these data sets reside in their subdirectory "C://data.") These paths obviously need to be modified correspondingly when the data sets are utilized by readers.