TABLE 1.1
A Sample of $n = 25$ Job CPU Times (in seconds) Selected from Appendix IV

1.17	1.61	1.16	1.38	3.53
1.23	3.76	1.94	0.96	4.75
0.15	2.41	0.71	0.02	1.59
0.19	0.82	0.47	2.16	2.01
0.92	0.75	2.59	3.07	1.40

TABLE 1.2
Calculating Class Relative Frequencies for the Data of Table 1.1

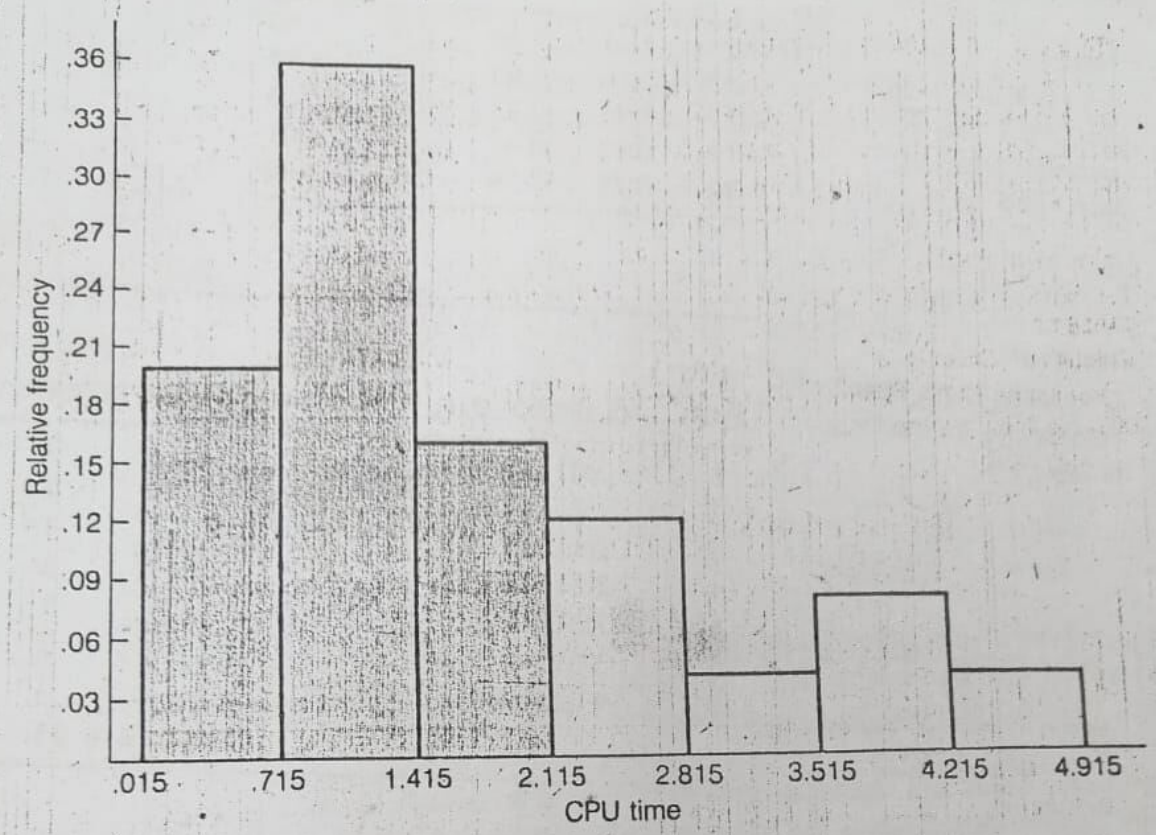
CLASS	CLASS INTERVAL	DATA TABULATION	CLASS FREQUENCY	CLASS RELATIVE FREQUENCY
1	.015–.715	𝍸	5	.20
2	.715–1.415	𝍸 ||||	9	.36
3	1.415–2.115	||||	4	.16
4	2.115–2.815	|||	3	.12
5	2.815–3.515	|	1	.04
6	3.515–4.215	||	2	.08
7	4.215–4.915	|	1	.04
		Totals	$n = 25$	1.00

Since the bars in a relative frequency histogram are of equal width, the area of a particular bar is proportional to the corresponding class relative frequency. If we let the total area of the bars equal 1, then the area of a particular bar is *equal* to its corresponding class relative frequency. Furthermore, if we select one observation from among the $n = 25$ observations in Table 1.1, then the probability that the observation will fall in a particular class is equal to the relative frequency of that class. The probability that the observation will fall in one of two or more specific classes is equal to the sum of their respective relative frequencies and is proportional to the total area of the bars corresponding to those classes. For example, the probability that the observation will be a CPU time less than 2.115 seconds is equal to .72, the sum of the relative frequencies for classes 1, 2, and 3 of Table 1.2. This probability is proportional to the shaded area of Figure 1.4 on page 14.

[Note: Most graphical descriptions of large data sets, whether qualitative or quantitative, are performed on a computer by one of several easy-to-use statistical computer program packages. In Chapter 2 we will show you how to enter data into a computer for use with each of four such packages: BMDP, Minitab, SAS, and SPSS*. In Sections 2.6 and 2.7, we will show you how to use the programs to construct a relative frequency histogram.]

Another graphical method for describing quantitative data is the **stem and leaf display**, which is widely used in exploratory data analysis when the data set is small. The next box contains the steps to follow in constructing a stem and leaf display for the 25 CPU times of Table 1.1.

ency
the $n = 25$
Table 1.1



STEPS TO FOLLOW IN CONSTRUCTING A STEM AND LEAF DISPLAY

Step 1. Divide each observation in the data set into two parts, the stem and the leaf. We will designate the first digit of the CPU time (i.e., the digit to the left of the decimal point) as its stem; we will call the last two digits its leaf. For example, the stem and leaf of the CPU time 2.41 are 2 and 41, respectively:

STEM	LEAF
2	41

Although this assignment is arbitrary, the stem and leaf display will yield more information if you define the stems and leaves so that you obtain a smaller number of stems for a small amount of data and a larger number of stems for large data sets. Using the assignment specified here, we obtain a total of five stems: 0, 1, 2, 3, and 4.

Step 2. List the stems in order in a column, starting with the smallest stem and ending with the largest (see Figure 1.5).

Step 3. Proceed through the data set, placing the leaf for each observation in the appropriate stem row. The completed stem and leaf display for the data of Table 1.1 is shown in Figure 1.5.

Notice that if you rotate the stem and leaf display on its side, you obtain the same type of bar graph as provided by the (relative) frequency distribution. The stem and leaf display of Figure 1.5 partitions the data set into five classes corresponding to the five stems. The number of leaves in each class gives the class frequency.

GURE 1.5
em and Leaf Display
the $n = 25$ CPU
nes of Table 1.1

STEMS	LEAVES	FREQUENCY	RELATIVE FREQUENCY
0	15 19 92 82 75 71 47 96 02	9	.36
1	17 23 61 16 94 38 59 40	8	.32
2	41 59 16 01	4	.16
3	76 07 53	3	.12
4	75	1	.04
	TOTALS	$n = 25$	1.00

One advantage of a stem and leaf display over a frequency distribution is that the original data are preserved. That is, you can look at the display and resurrect the exact values of the data. A stem and leaf display also arranges the data in an orderly fashion and makes it easy to determine certain numerical characteristics to be discussed in the following sections. The third advantage is that the classes and the numbers falling in them are quickly determined once we have selected the digits that we want to use for the stems and leaves.

A disadvantage of the stem and leaf display is that there is sometimes not much flexibility in choosing the stems. For the data of Table 1.1, two stem and leaf options are possible. We could define the stems and leaves as shown in Figure 1.5. Or, we could let the first two digits represent the stem, in which case the number 2.41 would have the stem 24 and the leaf 1:

STEM	LEAF
24	1

The associated stem and leaf display for the data of Table 1.1 would then contain a total of 48 stems, ranging from 00 to 47, and most of these stem rows would contain no leaves. Clearly, this choice of stems and leaves would not provide as much information about the data as does the display of Figure 1.5. Consequently, we are left with the option of using a stem and leaf display that produces five stems (and thus, five classes for the frequency distribution) or one that produces 48 stems.*

Frequency distributions or relative frequency distributions are most often used in scientific publications to describe quantitative data sets. They are better suited to the description of large data sets and they permit a greater flexibility in the choice of class width.

*By sacrificing some of the simplicity of our procedure, we could define the stems and leaves so that the number of stems falls between 5 and 48. We omit discussion of this topic.

DEFINITION 1.18

An observation y that is unusually large or small relative to the other values in a data set is called an outlier.

The most obvious method for determining whether an observation is an outlier is to calculate its z-score (Section 1.8). For example, if the z-score for a y value is -4.13, we know that it lies more than 4 standard deviations below the mean of the data set. Both the Empirical Rule and Tchebysheff's theorem tell us that almost all the observations in a data set will have z-scores less than 3 in absolute value. Thus, a z-score as small as -4.13 is highly improbable and points to the possibility of an errant observation.

Another procedure for detecting outliers is to construct a box plot of the data. With this method, we construct intervals similar to the $\bar{y} \pm 2s$ and $\bar{y} \pm 3s$ intervals of the Empirical Rule; however, the intervals are based on a quantity called the interquartile range instead of the standard deviation s.

DEFINITION 1.19

The interquartile range, IQR, is the distance between the upper and lower quartiles:

$$\text{IQR} = Q_U - Q_L$$

The procedure is especially easy to use for small data sets because the quartiles and interquartile range can be quickly determined. The steps to follow in constructing a box plot are given in the box on page 30.

A box plot for the 25 CPU times in Table 1.1 is shown in Figure 1.7. From the plot you can see that $Q_L = .82$, $m = 1.38$, $Q_U = 2.16$, and

$$\text{IQR} = Q_U - Q_L = 2.16 - .82 = 1.34$$

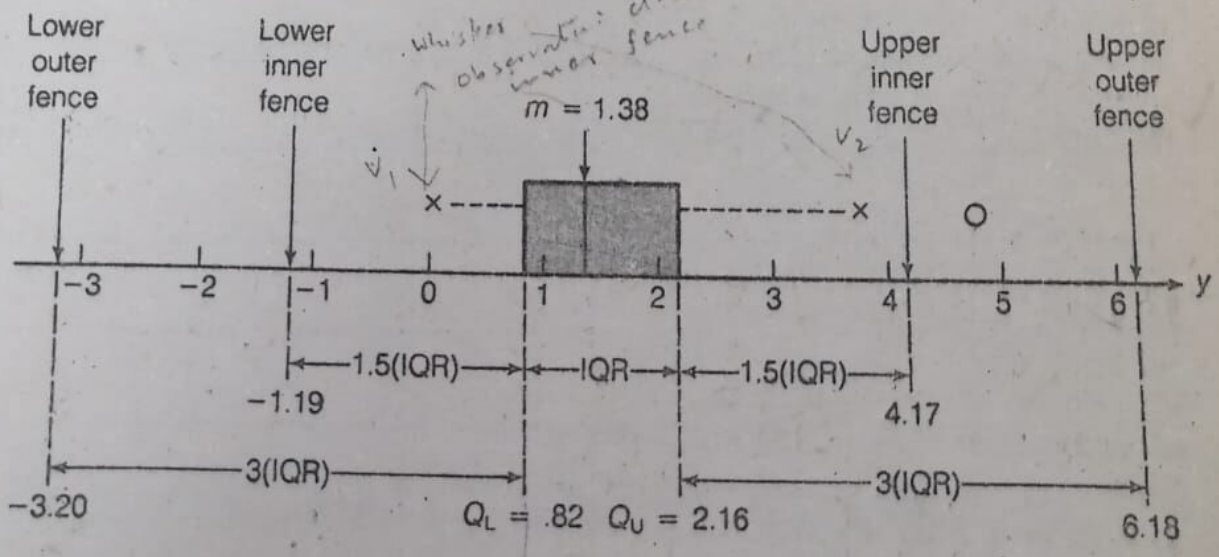


FIGURE 1.7
Box Plot for the $n = 25$ CPU Times of Table 1.1