Problems

1. Based on a sample of 100 individuals, the values 1, 2, 3, 4, 5 are observed with relative frequencies 0.2, 0.3, 0.1, 0.25, 0.15. Compute the mean, variance and standard deviation.

2. Fifty individuals are rated on how open minded they are. The ratings have the values 1, 2, 3, 4 and the corresponding relative frequencies are 0.2, 0.24, 0.4, 0.16, respectively. Compute the mean, variance and standard deviation.

3. For the values 0, 1, 2, 3, 4, 5, 6 the corresponding relative frequencies based on a sample of 10,000 observations are 0.015625, 0.093750, 0.234375, 0.312500, 0.234375, 0.093750, 0.015625, respectively. Determine the mean, median, variance, standard deviation and mode.

4. For a local charity, the donations in dollars received during the last month were 5, 10, 15, 20, 25, 50 having the frequencies 20, 30, 10, 40, 50, 5. Compute the mean, variance and standard deviation.

5. The values 1, 5, 10, 20 have the frequencies 10, 20, 40, 30. Compute the mean, variance and standard deviation.

3.2 Histograms

An important feature of the data previously used to illustrate plots of relative frequencies is that a few values occur many times. In figure 3.2, for example, the responses are limited to five values. But when there are many values, with most values occurring a small number of times, plots of relatively frequencies can be rather uninteresting. If each value occurs only once, a plot of the relatively frequencies would consist of n spikes, each having height $1/n$. A histogram is one way of trying to deal with this problem. It is similar to plots of relatively frequencies, the main difference being that values are binned together to get a more useful plot. That is, a histogram simply groups the data into categories and plots the corresponding frequencies.

Example 1

A histogram is illustrated with data from a heart transplant study conducted at Stanford University between October 1, 1967 and April 1, 1974. Of primary concern is whether a transplanted heart will be rejected by the recipient. With the goal of trying to address this issue, a so-called T5 mismatch score was developed by Dr. C. Bieber. It measures the degree of dissimilarity between the donor and the recipient tissue with respect to HL-A antigens. Scores less than 1 represent a good match and scores greater than 1 a poor match. The

Table 3.2 T5 mismatch scores from a heart transplant study

0.00	0.12	0.16	0.19	0.33	0.36	0.38	0.46	0.47	0.60	0.61	0.61	0.66	0.67	0.68
0.69	0.75	0.77	0.81	0.81	0.82	0.87	0.87	0.87	0.91	0.96	0.97	0.98	0.98	1.02
1.06	1.08	1.08	1.11	1.12	1.12	1.13	1.20	1.20	1.32	1.33	1.35	1.38	1.38	1.41
1.44	1.46	1.51	1.58	1.62	1.66	1.68	1.68	1.70	1.78	1.82	1.89	1.93	1.94	2.05
2.09	2.16	2.25	2.76	3.05										

Table 3.3 Frequencies and relative frequencies for grouped T5 scores, $n = 65$

Test score (x)	Frequency	Relative frequency
−0.5–0.0	1	$1/65 = .015$
0.0–0.5	8	$8/65 = .123$
0.5–1.0	20	$20/65 = .308$
1.0–1.5	18	$18/65 = .277$
1.5–2.0	12	$12/65 = .138$
2.0–2.5	4	$4/65 = .062$
2.5–3.0	1	$1/65 = .015$
3.0–3.5	1	$1/65 = .015$

T5 scores, written in ascending order, are shown in table 3.2 and are taken from Miller (1976). Suppose we group the T5 values into eight categories: (1) values between −0.5 and 0.0, (2) values greater than 0.0 but less than or equal to 0.5, (3) values greater than 0.5 but less than or equal to 1.0, and so on. The beginning and end of each interval are called *boundaries* or *class interval* and the point midway between any to boundaries is called the *class mark* or *midpoint*. So here, the first interval has boundaries −0.5 and 0.0 and the corresponding class mark or midpoint is $(-0.5+0)/2 = -0.25$. Similarly, the second interval has boundaries 0.0 and 0.5, so the class mark is $(0.0+0.5)/2 = 0.25$. Note that all of the categories have the same length, which is a feature routinely used. The frequency and relative frequency associated with each of these intervals is shown in table 3.3. For example, there are eight T5 mismatch scores in the interval extending from 0.0 to 0.5 and the proportion of all scores belonging to this interval is 0.123. Figure 3.3 shows the resulting histogram.

How many bins should be used when constructing a histogram and how should the length of the bins be chosen? The general goal is to choose the number of bins so as to get an informative plot of the data. If we have one bin only, this tells us little about the data, and too many bins suffer from the same problem. There are simple rules for choosing the number of bins. One is called

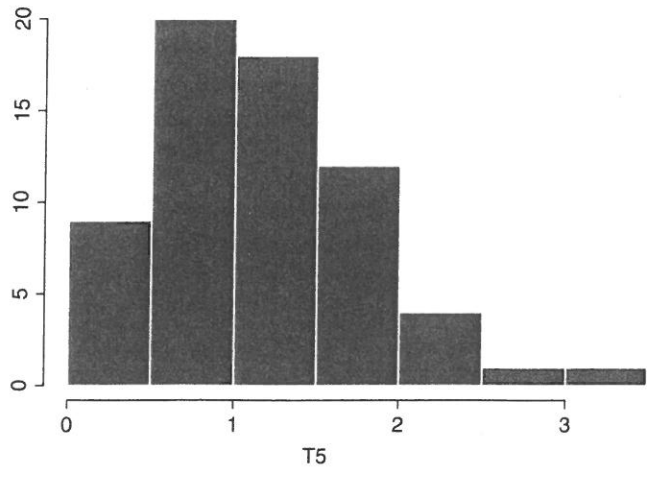

Figure 3.3 A histogram of the heart transplant data in table 3.5.

Sturges's rule, which is commonly used by statistical software, but no details are given here. The main point is that standard methods for choosing the number of bins can result in a rather unsatisfactory summary of the data, as will be illustrated. The good news is that substantially better methods are now available, some of which are outlined in the final section of this chapter.

What do histograms tell us?

Like so many graphical summaries of data, histograms attempt, among other things, to tell us something about the shape of the data. One issue of some concern is whether data are reasonably symmetric about some central value. In figure 3.2, we see exact symmetry, but often data are highly skewed, and this can be a serious practical problem when dealing with inferential techniques yet to be described. The left panel of figure 3.4 shows data that are not symmetric, but rather *skewed to the right*. The right panel shows data that are *skewed to the left*. In recent years, skewness, roughly referring to a lack of symmetry, has been found to be a much more serious problem than once thought for reasons that are best postponed for now. But one important point that should be stressed here is that when distributions are skewed, generally the mean, median and 20% trimmed mean will differ. In some cases they differ by very little, but in other situations these measures of location can differ substantially, as was illustrated in chapter 2. Moreover, even when these measures of location are virtually identical, subsequent chapters will demonstrate that often the choice for a measure of location can make a practical difference when addressing common problems yet to be described.

Example 2

In an unpublished study (by M. Earleywine) was performed that generally dealt with the effects of consuming alcohol. A portion of the was concerned with

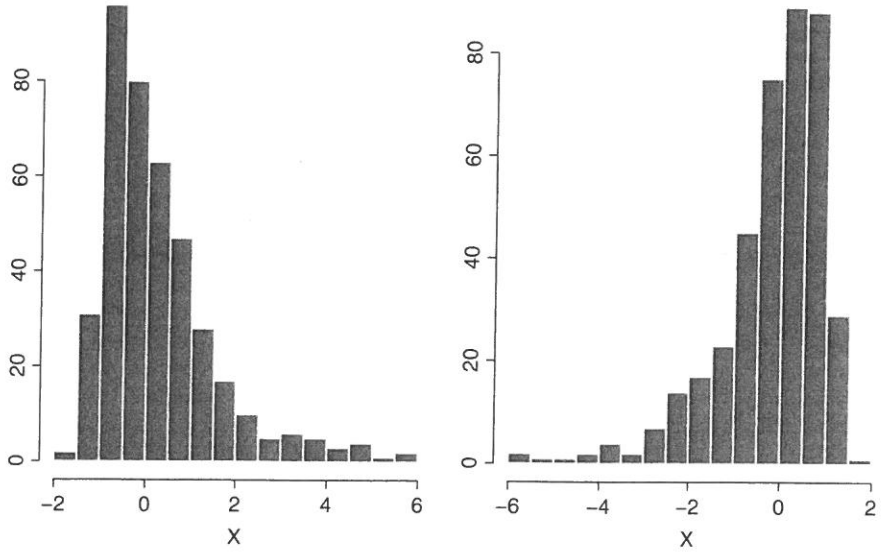

Figure 3.4 The left panel is an example of a histogram that is said to be skewed to the right. In the right panel, it is skewed to the left.

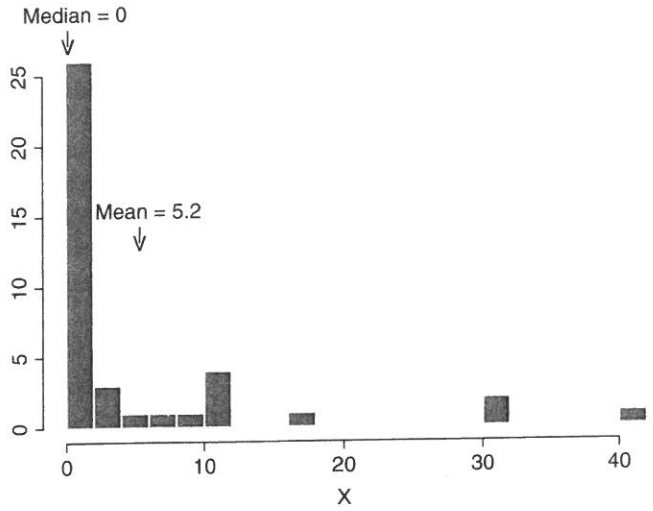

Figure 3.5 A histogram of the measurement of hangover symptoms.

measuring hangover symptoms after consuming a specific amount of alcohol in a laboratory setting. The resulting measures, written in ascending order, were

0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 2 2 2 3 3 3 6 8 9 11 11 11 12 18 32 32 41.

Figure 3.5 shows a histogram based on these data. As is evident, the histogram is skewed to the right with a fairly large difference between the median and mean. Note that the rightmost portion indicates three bins that are separated from the rest of the plot. This might suggest that any value greater than 16 should be flagged an outlier, but a boxplot rule indicates that only values greater than or equal to 32 are outliers. Although histograms can provide a useful summary of data, as an outlier detection method, it is not very satisfactory. One fundamental problem is that, when using a histogram, no precise rule is available for telling us when a value should be declared an outlier. Without a precise rule, there can be no agreement on the extent to which masking (described in chapter 2) is avoided.[1] (Also, examples will be seen where extremely unlikely values occur, yet a histogram does not suggest that they are unusual.) As a general rule, histograms might suggest that certain values are outliers, but when making a decision about which values are outliers, it is better to use methods specifically designed for this task, such as the boxplot rule.

Populations, samples, and potential concerns about histograms

Histograms are routinely taught in an introductory statistics course and in some cases they provide useful information about data. But like so many methods covered in this book, it is important to understand not only when histograms perform well,

1. Outlier detection methods are typically designed to have other properties not covered here. Again, without a precise rule for deciding what constitutes an outlier, it is impossible to determine whether a specific method achieves the properties desired.

but also when, and in what sense, they might be highly misleading. Without a basic understanding of the relative merits of a method, there is the potential of drawing erroneous conclusions, as well as missing interesting results. The immediate goal is to illustrate a fundamental concern about histograms, and in the final section of this chapter, methods aimed at correcting known problems are briefly indicated.

As mentioned in chapter 1, there is an important distinction between samples of individuals or things versus a population of individuals or things. Samples represent a subset of the population under study. Consider, for example, the last example dealing with hangover symptoms. There were 40 participants who represent only a small proportion of the individuals who might have taken part in this study. Ideally, the available participants will provide a reasonably accurate reflection of the histogram we would obtain if all participants could be measured. In some cases, histograms satisfy this goal. But an important practical issue is whether they can be highly unsatisfactory. And if they can be unsatisfactory, is there some strategy that might give substantially better results? In turns out that they can indeed by unsatisfactory, and fundamental improvements are now available (e.g., Silverman, 1986). The only goal here is to illustrate what might go wrong and provide information about where to look for better techniques.

First we consider a situation where the histogram tends to perform tolerably well. Imagine that the population consists of one million individuals and that if we could measure everyone, the resulting histogram would appear as in figure 3.6. (Here, the y-axis indicates the relative frequencies.) Now imagine that 100 individuals are selected from the one million individuals in the population, with every individual having the same probability of being chosen. An issue of fundamental importance is the extent to which a histogram based on a sample of only 100 individuals will reflect the histogram we would get if all individuals could be measured. Mimicking this process on a computer resulted in the histogram shown in figure 3.7. So in this particular case, the histogram provides a reasonable reflection of the population histogram, roughly capturing its bell shape. A criticism of this illustration is that maybe we just got lucky. That is, in general, perhaps with only 100 individuals, the histogram will not accurately reflect the population.

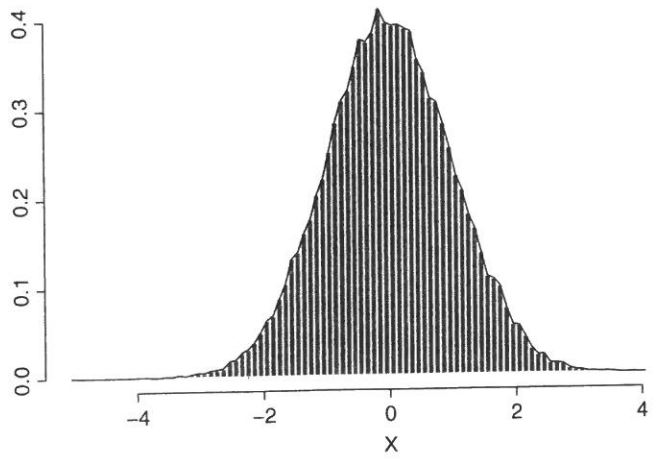

Figure 3.6 A histogram of an entire population that is approximately symmetric about 0 with relatively light tails, meaning outliers tend to be rare.

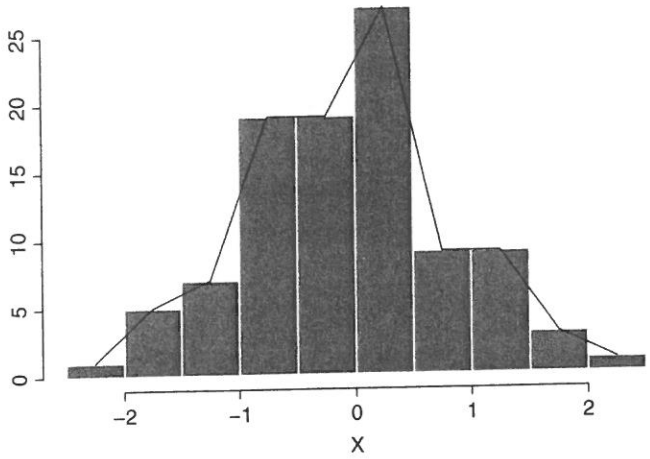

Figure 3.7 A histogram based on a sample of 100 observations generated from the histogram in figure 3.6.

This might indeed happen, but generally it gives a reasonable sense of the shape of the population histogram.

Now consider the population histogram in figure 3.8. This histogram has the same bell shape as in figure 3.6, but the tails extend out a bit farther. This reflects the fact that for this particular population, there are more outliers or extreme values. Now look at figure 3.9, which is based on 100 individuals sampled from the population histogram in figure 3.8. As is evident, it provides a poor indication of what the population histogram looks like. Figure 3.9 also provides another illustration that the histogram can perform rather poorly as an outlier detection rule. It suggests that values greater than 10 are highly unusual, which turns out to be true based on how the data were generated. But values less than −5 are also highly unusual, which is less evident here. The fact that the histogram can miss outliers limits its ability to deal with problems yet to be described.

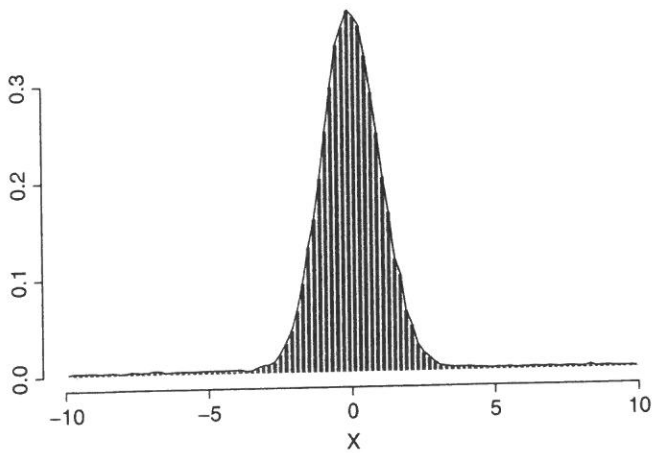

Figure 3.8 A histogram of an entire population that is approximately symmetric about 0 with relatively heavy tails, meaning outliers tend to be common.

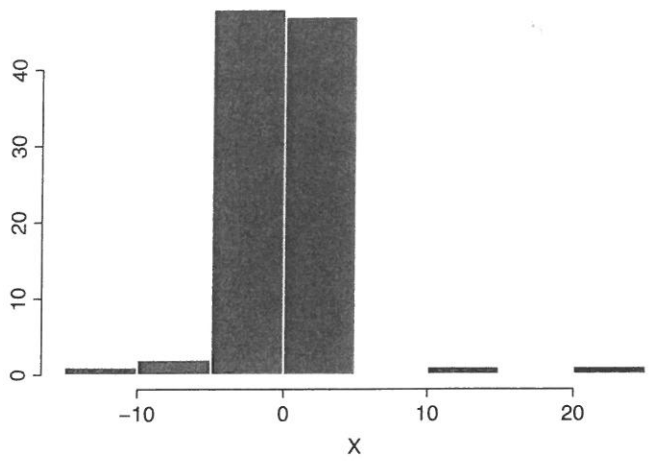

Figure 3.9 A histogram based on a sample of 100 observations generated from the histogram in figure 3.8.

A rough characterization of the examples just given is that when the population histogram is symmetric and bell-shaped, and outliers tend to be rare, it performs tolerably well with 100 observations, in terms of indicating the shape of the population histogram. But when outliers are relatively common, the reverse is true.

Problems

6. For the data in table 2.1, dealing with changes in cholesterol levels, suppose a histogram is to be created with bins defined as follows: −40 −30 −20 −10 0 10 20 30 40. That is, the first bin has boundaries −40 and −30, the next bin contains all values greater than −30 but less than or equal to −20, and so on. Determine the frequencies for each bin and construct a histogram.

7. For the data in table 2.2, suppose a histogram is to be created with bins defined as follows: −50 −40 −30 −20 −10 0 10 20 30 40 50 60 70 80. Determine the frequencies for each bin and construct a histogram.

8. The heights of 30 male Egyptian skulls from 4000 BC were reported by Thomson and Randall-Maciver (1905) to be

   121 124 129 129 130 130 131 131 132 132 132 133 133 134 134 134 134 135 135 136 136 136 136 137 137 138 138 138 140 143.

   Create a histogram with bins extending from 120–125, 125–130, and so on. Based on this histogram, does the largest value, 143, appear to be an outlier?

9. For the data in the previous problem, does the boxplot rule (described in chapter 2) indicate that 143 is an outlier?

10. What do the last two problems suggest about using a histogram to detect outliers?

Table 3.4 Word identification scores

58 58 58 58 58 64 64 68 72 72 72 75 75 77 77 79 80 82 82
82 82 82 84 84 85 85 90 91 91 92 93 93 93 95 95 95 95 95
95 95 95 98 98 99 101 101 101 102 102 102 102 102 103 104 104 104 104
104 105 105 105 105 105 107 108 108 110 111 112 114 119 122 122 125 125 125
127 129 129 132 134

3.3 Boxplots and stem-and-leaf displays

A *stem-and-leaf display* is another method of gaining some overall sense of what data are like. The method is illustrated with measures taken from a study aimed at understanding how children acquire reading skills. A portion of the study was based on a measure that reflects the ability of children to identify words. (These data were supplied by L. Doi.) Table 3.4 lists the observed scores in ascending order.

The construction of a stem-and-leaf display begins by separating each value into two components. The first is the *leaf* which, in this example, is the number in the ones position (the single digit just to the left of the decimal place). For example, the leaf corresponding to the value 58 is 8. The leaf for the value 64 is 4 and the leaf for 125 is 5. The digits to the left of the leaf are called the *stem*. Here, the stem of 58 is 5, the number to the left of 8. Similarly, 64 has a stem of 6 and 125 has a stem of 12. We can display the results for all 81 children as follows:

Stems	Leaves
5	88888
6	448
7	22255779
8	0222224455
9	011233355555555889
10	11122222344444555557 88
11	01249
12	22555799
13	24

There are five children who have the score 58, so there are five scores with a leaf of 8, and this is reflected by the five 8s displayed to the right of the stem 5 and under the column headed by Leaves. Two children got the score 64, and one child got the score 68. That is, for the stem 6, there are two leaves equal to 4 and one equal to 8, as indicated by the list of leaves in the display. Now look at the third row of numbers where the stem is 7. The leaves listed are 2, 2, 2, 5, 5, 7, 7 and 9. This indicates that the value 72 occurred three times, the value 75 occurred two times, as did the value 77, and the value 79 occurred once. Notice that the display of the leaves gives us some indication of the values that occur most frequently and which are relatively rare. Like the histogram, the stem-and-leaf display gives us an overall sense of what the values are like.

The leaf always consists of the numbers corresponding to a specified digit. For example, the leaf might correspond to tenths digit, meaning that the leaf is the first number to the right of the decimal, in which case the stem consists of all the numbers to the left of the leaf. So for the number 158.234, the leaf would be 2 and the stem would be 158. If we specify the leaf to be the hundredth digit, the leaf would now be 3 and the stem would be 158.2. The choice of which digit is to be used as the leaf depends in part

on which digit provides a useful graphical summary of the data. But details about how to address this problem are not covered here. Suffice it to say that algorithms have been proposed for deciding which digit should be used as the leaf and determining how many lines a stem-and-leaf display should have (for example, Emerson and Hoaglin, 1983).

Example 1

Chapter 1 mentioned the software S-PLUS. When its version of a stem-and-leaf display is applied to the T5 mismatch scores, the result is

```
Decimal point is at the colon
  0 : z122344
  0 : 556667777788889999
  1 : 00000111111223334444
  1 : 5566777788999
  2 : 0122
  2 : 8
  3 : 0
```

The z in the first row stands for zero. So this plot suggests that the data are reasonably symmetric, with maybe a hint of being skewed to the right. Also, there are no values visibly separated from the overall plot suggesting that there are no outliers. (The boxplot rule, described in chapter 2, also finds no outliers.)

Boxplot

Proposed by Tukey (1977), a boxplot is a commonly used graphical summary of data, an example of which is shown in figure 3.10. As indicated, the ends of the rectangular box mark the lower and upper quartiles. That is, the box indicates where the middle half of the data lie. The horizontal line inside the box indicates the position of the median. The lines extending out from the box are called *whiskers*.

Boxplots determine whether values are outliers using the boxplot rule described in chapter 2. (See equations 2.4 and 2.5.) Figure 3.11 shows a boxplot with two outliers. The ends of the whiskers are called *adjacent values*. They are the smallest and largest values not declared outliers.

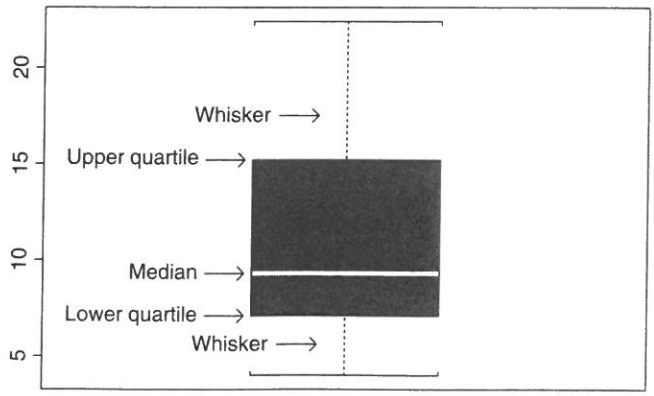

Figure 3.10 An example of a boxlpot with no outliers.

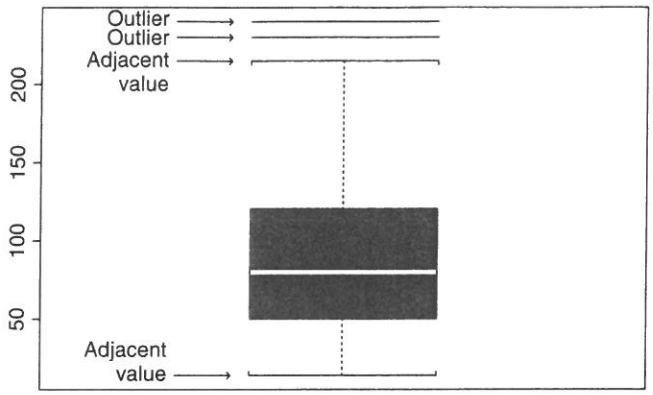

Figure 3.11 An example of a boxplot with outliers.

All of the statistical software mentioned in chapter 1 contain routines for creating boxplots. In case it helps, here is an outline of how they are constructed using the data in figure 3.11. First, compute the lower and upper quartiles using the method in chapter 2 (based on the ideal fourths). This yields $q_1 = 49.8$ and $q_2 = 120.75$, which correspond to the lower and upper ends of the box, respectively. The median is $M = 80$ and determines where the line within the box is placed. Next, using the boxplot rule in chapter 2, determine how small a value must be to be declared an outlier. Here, this value is -56.5. The smallest value not declared an outlier determines the end of the lower whisker, which is 14. Again using the boxplot rule, any value greater than 227 is declared an outlier. The largest value not declared an outlier is 215, so this value marks the end of the upper whisker. There are two values greater than 227, which correspond to the horizontal lines at the top of figure 3.11.

Problems

11. Table 3.5 shows the exam scores for 27 students. Create a stem-and-leaf display using the digit in the ones position as the stem.
12. If the leaf is the hundredths digit, what is the stem for the number 34.679?
13. Consider the values 5.134, 5.532, 5.869, 5.809, 5.268, 5.495, 5.142, 5.483, 5.329, 5.149, 5.240, 5.823. If the leaf is taken to be the tenths digit, why would this make an uninteresting stem-and-leaf display?
14. For the boxplot in figure 3.11, determine, approximately, the quartiles, the interquartile range, and the median. Approximately how large is the largest value not declared an outlier?
15. In figure 3.11, about how large must a value be to be declared an outlier? How small must it be?

Table 3.5 Examination Scores

83	69	82	72	63	88	92	81	54
57	79	84	99	74	86	71	94	71
80	51	68	81	84	92	63	99	91

16. Create a boxplot for the data in table 3.1.
17. Create a boxplot for the data in table 3.2.

3.4 Some modern trends and developments

We have seen that in terms of providing information about the shape of the population histogram, a histogram based on 100 observations can be relatively ineffective in certain situations. There is a vast literature on how this problem might be addressed using what are called *kernel density estimators*. There are in fact many variations of this approach, some of which appear to perform very well over a fairly broad range of situations. Some of these methods come with the software R and S-PLUS mentioned in chapter 1. The computational details go well beyond the scope of this book, but an illustration might help motivate their use.

Example 1

Consider again the data used to create the histogram shown in figure 3.9. Recall that the 100 observations were sampled from a population having the symmetric histogram shown in figure 3.8, yet the histogram in figure 3.9 suggests a certain amount of asymmetry. In particular, the right tail differs from the left; values in the right tail appear to be outliers and the values in the left tail seem to have a low relatively frequency, but otherwise there is no sense that they are unusually far from the central values. One of the seemingly better methods for improving on the histogram is called an adaptive kernel density estimator. Figure 3.12 shows a plot of the data in figure 3.9 using this method.[2] The plot in figure 3.12 does not capture the exact symmetry of the population histogram, but typically it does a better job of indicating its shape versus the histogram.

A Summary of Some Key Points

- No single graphical summary of data is always best. Different methods provide different and potentially interesting perspectives. What is required is some familiarity with the various methods to help you choose which one to use.
- However, the choice of method is not always academic. The histogram is a classic, routinely taught method, but it is suggested that kernel density estimators be given serious consideration. Perhaps the most important point to keep in mind is that the histogram performs rather poorly as an outlier detection technique.
- The boxplot is one of the more useful graphical tools for summarizing data. It conveys certain important features that were described and illustrated, but kernel density estimators can help add perspective.
- The stem-and-leaf display can be useful when trying to understand the overall pattern of the data. But with large sample sizes, it can be highly unsatisfactory.

2. The S-PLUS function akerd was used, which belongs to a library of S-PLUS functions mentioned in chapter 1.

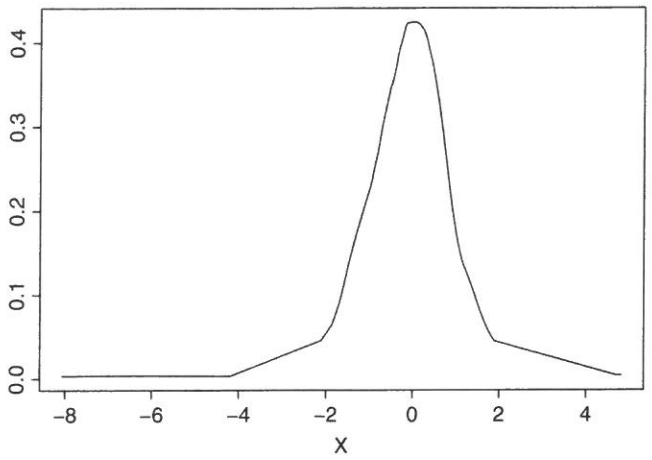

Figure 3.12 An example of a kernel density plot based on the same 100 observations generated for figure 3.8 and used in figure 3.9. Note how the kernel density plot does a better job of capturing the shape of the population histogram in figure 3.8.

Problems

18. Describe a situation where the sample histogram is likely to give a good indication of the population histogram based on 100 observations.
19. Comment generally on how large a sample size is needed to ensure that the sample histogram will likely provide a good indication of the population histogram?
20. When trying to detect outliers, discuss the relative merits of using a histogram versus a boxplot.
21. A sample histogram indicates that the data are highly skewed to the right. Is this a reliable indication that if all individuals of interest could be measured, the resulting histogram would also be highly skewed?