DREAM OF INNOVATION

6 Random Sampling and Data Description

CHAPTER OUTLINE

LEARNING OBJECTIVES

After careful study of this chapter you should be able to do the following:

1. Compute and interpret the sample mean, sample variance, sample standard deviation, sample median, and sample range
2. Explain the concepts of sample mean, sample variance, population mean, and population variance
3. Construct and interpret visual data displays, including the stem-and-leaf display, the histogram, and the box plot
4. Explain the concept of random sampling
5. Construct and interpret normal probability plots
6. Explain how to use box plots and other data displays to visually compare two or more samples of data
7. Know how to use simple time series plots to visually display the important features of time-oriented data.

1. MEASURES OF CENTRAL TENDENCY

6-1 Numerical Summaries

Definition: Sample Mean

If the n observations in a sample are denoted by $x_1, x_2, \ldots, x_n$, the **sample mean** is

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n} \qquad (6\text{-}1)$$

6-1 Numerical Summaries

Example 6-1

Let's consider the eight observations collected from the prototype engine connectors from Chapter 1. The eight observations are $x_1 = 12.6$, $x_2 = 12.9$, $x_3 = 13.4$, $x_4 = 12.3$, $x_5 = 13.6$, $x_6 = 13.5$, $x_7 = 12.6$, and $x_8 = 13.1$. The sample mean is

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{8} x_i}{8} = \frac{12.6 + 12.9 + \cdots + 13.1}{8}$$

$$= \frac{104}{8} = 13.0 \text{ pounds}$$

A physical interpretation of the sample mean as a measure of location is shown in the dot diagram of the pull-off force data. See Figure 6-1. Notice that the sample mean $\bar{x} = 13.0$ can be thought of as a "balance point." That is, if each observation represents 1 pound of mass placed at the point on the x-axis, a fulcrum located at $\bar{x}$ would exactly balance this system of weights.

6-1 Numerical Summaries

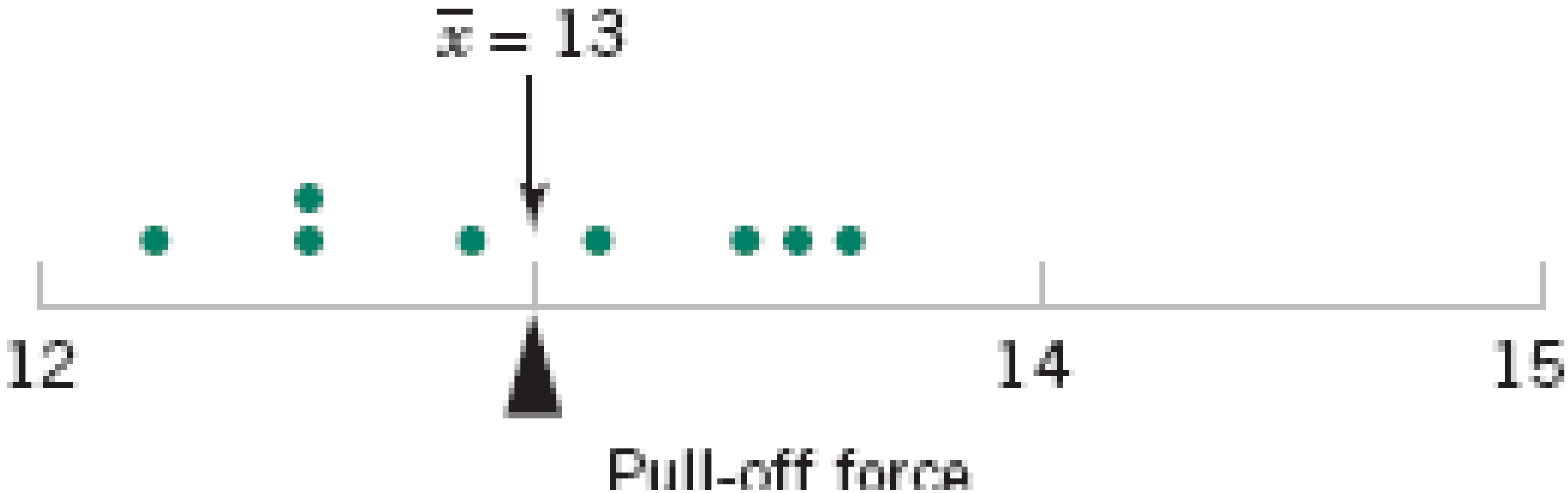


Figure 6-1 The sample mean as a balance point for a system of weights.

6-1 Numerical Summaries

Population Mean

For a finite population with N measurements, the mean is

$$\mu = \sum_{i=1}^{N} x_i f(x_i) = \frac{\sum_{i=1}^{N} x_i}{N} \qquad (6-2)$$

The sample mean is a reasonable estimate of the population mean.

Example

Let's consider the weight of the eight observations collected from the prototype engine connectors: 12.6, 12.9, 13.4, 12.3, 13.6, 13.5, 12.6 and 13.1

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{12.6 + 12.9 + \ldots + 13.1}{8} = 13.0$$

Sample median

(1) The value that lies in the middle of the data when the data set is ordered.

(2) Measures the center of an ordered data set by dividing it into two equal parts.

(3) If the data set has an

(a) even number of entries: median is the mean of the two middle data entries.

(b) odd number of entries: median is the middle data entry.

Example

The prices (in dollars) for a sample of roundtrip flights from Chicago, Illinois to Cancun, Mexico are listed. Find the median of the flight prices.

872 432 397 427 388 782 397

First order the data.

388 397 397 427 432 782 872

The median price of the flights is $427.

Sample mode

(1) The data entry that occurs with the greatest frequency.

(2) If no entry is repeated the data set has no mode.

(3) If two entries occur with the same greatest frequency, each entry is a mode (bimodal).

Example

At a political debate a sample of audience members was asked to name the political party to which they belong. Their responses are shown in the table. What is the mode of the responses?

Political Party	Frequency, f
Democrat	34
Republican	56
Other	21
Did not respond	9

The mode is Republican (the response occurring with the greatest frequency). In this sample there were more Republicans than people of any other single affiliation.

2. MEASURES OF VARIATION

6-1 Numerical Summaries

Definition

If the n observations in a sample are denoted by $x_1, x_2, \ldots, x_n$, the **sample range** is

$$r = \max(x_i) - \min(x_i) \tag{6-6}$$

Sample range

- The difference between the maximum and minimum data entries in the set.

- The data must be quantitative.

- If the n observations in a sample are denoted by $x_1, x_2, \ldots, x_n$, the sample range is

$$r = max(x_i) - min(x_i)$$

Definition: Sample Variance

If $x_1, x_2, \ldots, x_n$ is a sample of n observations, the **sample variance** is

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1} \tag{6-3}$$

The **sample standard deviation,** s, is the positive square root of the sample variance.

6-1 Numerical Summaries

How Does the Sample Variance Measure Variability?

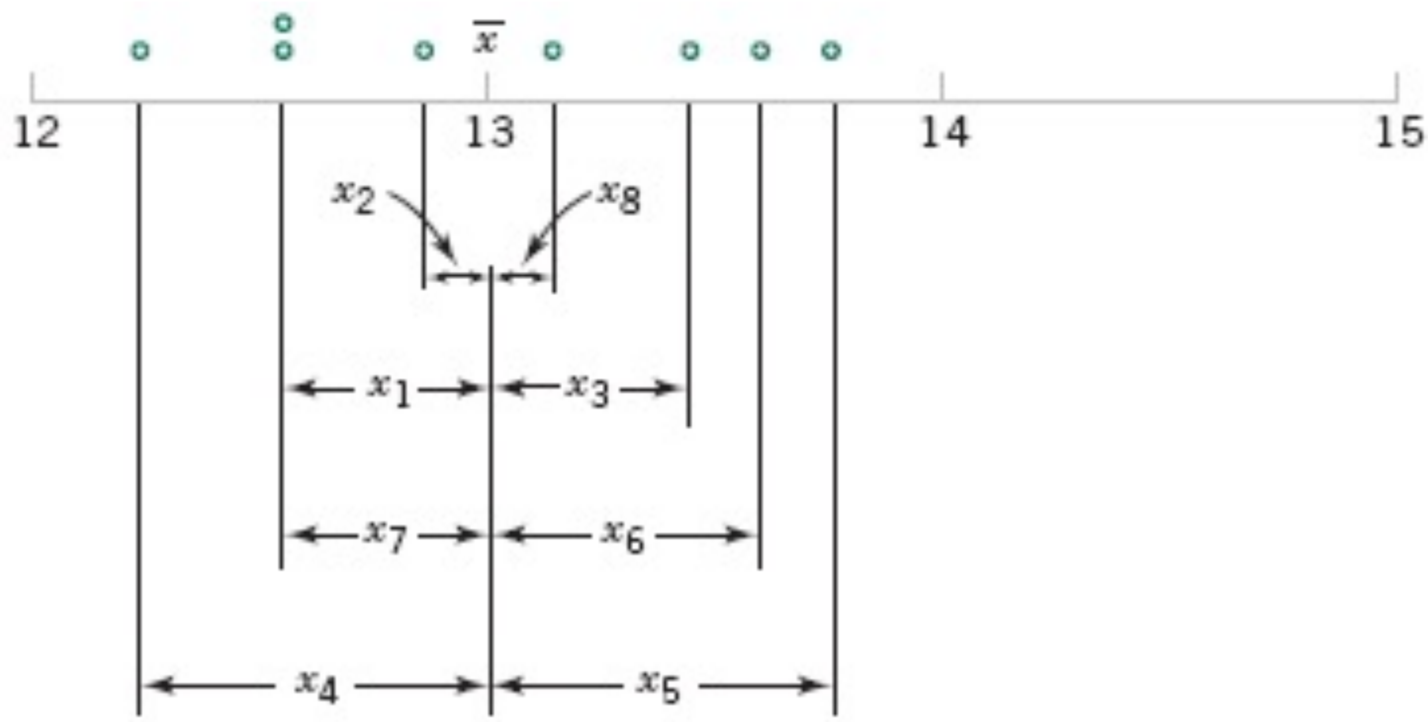


Figure 6-2 How the sample variance measures variability through the deviations $x_i - \bar{x}$.

6-1 Numerical Summaries

Example 6-2

Table 6-1 displays the quantities needed for calculating the sample variance and sample standard deviation for the pull-off force data. These data are plotted in Fig. 6-2. The numerator of s^2 is

$$\sum_{i=1}^{8} (x_i - \bar{x})^2 = 1.60$$

so the sample variance is

$$s^2 = \frac{1.60}{8 - 1} = \frac{1.60}{7} = 0.2286 \text{ (pounds)}^2$$

and the sample standard deviation is

$$s = \sqrt{0.2286} = 0.48 \text{ pounds}$$

6-1 Numerical Summaries

Table 6-1 Calculation of Terms for the Sample Variance and Sample Standard Deviation

i	x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
1	12.6	−0.4	0.16
2	12.9	−0.1	0.01
3	13.4	0.4	0.16
4	12.3	−0.7	0.49
5	13.6	0.6	0.36
6	13.5	0.5	0.25
7	12.6	−0.4	0.16
8	13.1	0.1	0.01
	104.0	0.0	1.60

6-1 Numerical Summaries

Computation of s^2

$$s^2 = \frac{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}{n-1} \tag{6-4}$$

6-1 Numerical Summaries

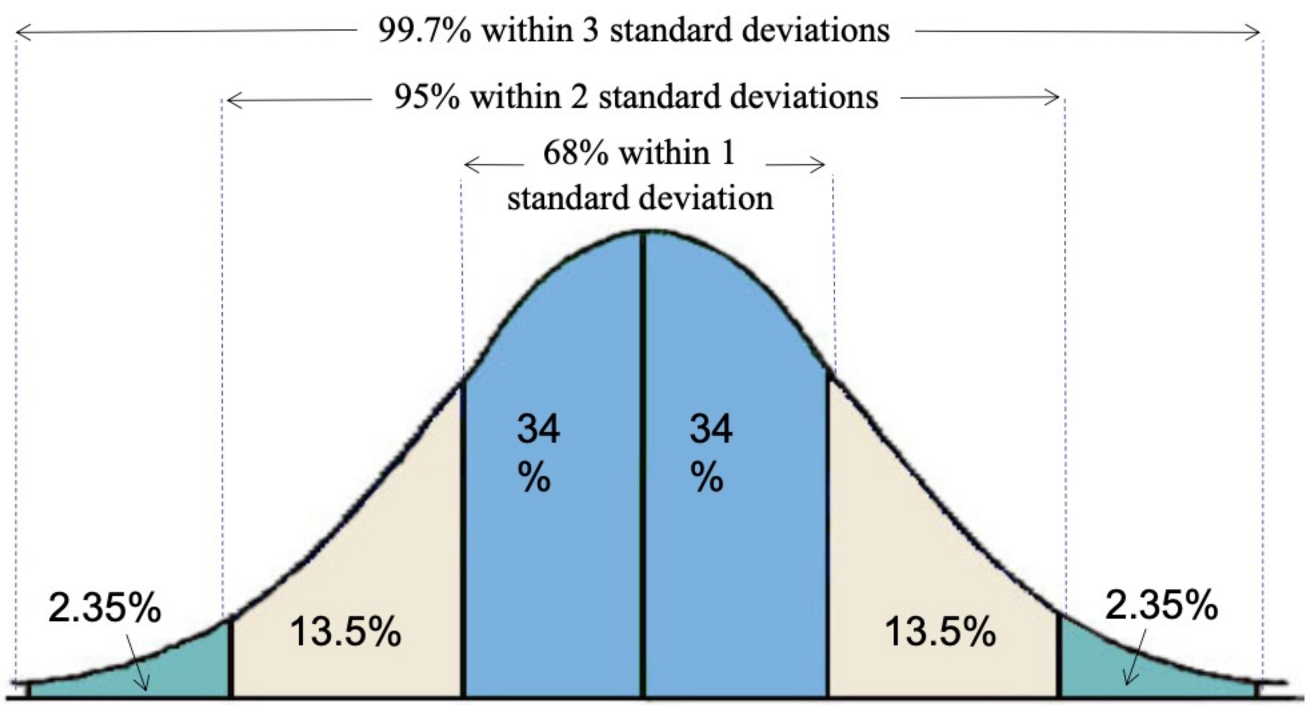
Population Variance

When the population is finite and consists of N values, we may define the population variance as

$$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N} \qquad (6\text{-}5)$$

The sample variance is a reasonable estimate of the population variance.

Interpreting standard deviation: For data with a bell-shaped distribution



3. DISTRIBUTION

6-2 Stem-and-Leaf Diagrams

A **stem-and-leaf diagram** is a good way to obtain an informative visual display of a data set $x_1, x_2, \ldots, x_n$, where each number x_i consists of at least two digits. To construct a stem-and-leaf diagram, use the following steps.

Steps for Constructing a Stem-and-Leaf Diagram

(1) Divide each number x_i into two parts: a **stem,** consisting of one or more of the leading digits and a **leaf,** consisting of the remaining digit.

(2) List the stem values in a vertical column.

(3) Record the leaf for each observation beside its stem.

(4) Write the units for stems and leaves on the display.

6-2 Stem-and-Leaf Diagrams

Example 6-4

To illustrate the construction of a stem-and-leaf diagram, consider the alloy compressive strength data in Table 6-2. We will select as stem values the numbers 7, 8, 9, . . . , 24. The resulting stem-and-leaf diagram is presented in Fig. 6-4. The last column in the diagram is a frequency count of the number of leaves associated with each stem. Inspection of this display immediately reveals that most of the compressive strengths lie between 110 and 200 psi and that a central value is somewhere between 150 and 160 psi. Furthermore, the strengths are distributed approximately symmetrically about the central value. The stem-and-leaf diagram enables us to determine quickly some important features of the data that were not immediately obvious in the original display in Table 6-2.

6-2 Stem-and-Leaf Diagrams

Table 6-2 Compressive Strength (in psi) of 80 Aluminum-Lithium Alloy Specimens

105	221	183	186	121	181	180	143
97	154	153	174	120	168	167	141
245	228	174	199	181	158	176	110
163	131	154	115	160	208	158	133
207	180	190	193	194	133	156	123
134	178	76	167	184	135	229	146
218	157	101	171	165	172	158	169
199	151	142	163	145	171	148	158
160	175	149	87	160	237	150	135
196	201	200	176	150	170	118	149

6-2 Stem-and-Leaf Diagrams

Figure 6-4 Stem-and-leaf diagram for the compressive strength data in Table 6-2.

Stem	Leaf	Frequency
7	6	1
8	7	1
9	7	1
10	5 1	2
11	5 8 0	3
12	1 0 3	3
13	4 1 3 5 3 5	6
14	2 9 5 8 3 1 6 9	8
15	4 7 1 3 4 0 8 8 6 8 0 8	12
16	3 0 7 3 0 5 0 8 7 9	10
17	8 5 4 4 1 6 2 1 0 6	10
18	0 3 6 1 4 1 0	7
19	9 6 0 9 3 4	6
20	7 1 0 8	4
21	8	1
22	1 8 9	3
23	7	1
24	5	1

Stem : Tens and hundreds digits (psi); Leaf: Ones digits (psi)

6-2 Stem-and-Leaf Diagrams

In some data sets, it may be desirable to provide more classes or stems. One way to do this as follows:

Divide the stem 5 into two new stems: 5L and 5U. The stem 5L has leaves 0,1,2,3,4 and 5U has leaves 5,6,7,8,9.

We could increase the number of original stems by four by defining five new stems: 5z with leaves 0 and 1, 5t with leaves 2 and 3, 5f with leaves 4 and 5, 5s with leaves 6 and 7, 5e with leaves 8 and 9.

6-2 Stem-and-Leaf Diagrams

Example 6-5

Figure 6-5 illustrates the stem-and-leaf diagram for 25 observations on batch yields from a chemical process. In Fig. 6-5(a) we have used 6, 7, 8, and 9 as the stems. This results in too few stems, and the stem-and-leaf diagram does not provide much information about the data. In Fig. 6-5(b) we have divided each stem into two parts, resulting in a display that more adequately displays the data. Figure 6 -5 (c) illustrates a stem – and – leaf display with each stem divided into five parts.

There are too many stem in this plot , resulting in a display that does not tell us much about the shape of the data.

6-2 Stem-and-Leaf Diagrams

Stem	Leaf
6	1 3 4 5 5 6
7	0 1 1 3 5 7 8 8 9
8	1 3 4 4 7 8 8
9	2 3 5

(a)

Stem	Leaf
6L	1 3 4
6U	5 5 6
7L	0 1 1 3
7U	5 7 8 8 9
8L	1 3 4 4
8U	7 8 8
9L	2 3
9U	5

(b)

Stem	Leaf
6z	1
6t	3
6f	4 5 5
6s	6
6e	
7z	0 1 1
7t	3
7f	5
7s	7
7e	8 8 9
8z	1
8t	3
8f	4 4
8s	7
8e	8 8
9z	
9t	2 3
9f	5
9s	
9e	

(c)

Figure 6-5 Stem-and-leaf displays for Example 6-5. Stem: Tens digits. Leaf: Ones digits.

6-2 Stem-and-Leaf Diagrams

Figure 6-6 Stem-and-leaf diagram from Minitab.

Character Stem-and-Leaf Display

Stem-and-leaf of Strength
N = 80 Leaf Unit = 1.0

1	7	6
2	8	7
3	9	7
5	10	1 5
8	11	0 5 8
11	12	0 1 3
17	13	1 3 3 4 5 5
25	14	1 2 3 5 6 8 9 9
37	15	0 0 1 3 4 4 6 7 8 8 8 8
(10)	16	0 0 0 3 3 5 7 7 8 9
33	17	0 1 1 2 4 4 5 6 6 8
23	18	0 0 1 1 3 4 6
16	19	0 3 4 6 9 9
10	20	0 1 7 8
6	21	8
5	22	1 8 9
2	23	7
1	24	5

6-2 Stem-and-Leaf Diagrams

Data Features

The **median** is a measure of central tendency that divides the data into two equal parts, half below the median and half above. If the number of observations is even, the median is halfway between the two central values.

From Fig. 6-6, the 40th and 41st values of strength as 160 and 163, so the median is (160 + 163)/2 = 161.5. If the number of observations is odd, the median is the *central* value.

The **range** is a measure of variability that can be easily computed from the ordered stem-and-leaf display. It is the maximum minus the minimum measurement. From Fig.6-6 the range is 245 - 76 = 169.

6-2 Stem-and-Leaf Diagrams

Data Features

When an **ordered** set of data is divided into four equal parts, the division points are called **quartiles.**

The **first** or **lower quartile**, q_1 , is a value that has approximately one-fourth (25%) of the observations below it and approximately 75% of the observations above.

The **second quartile**, q_2, has approximately one-half (50%) of the observations below its value. The second quartile is *exactly* equal to the **median**.

The **third** or **upper quartile**, q_3, has approximately three-fourths (75%) of the observations below its value. As in the case of the median, the quartiles may not be unique.

FPT UNIVERSITY
DREAM OF INNOVATION

6-2 Stem-and-Leaf Diagrams

Data Features

The compressive strength data in Figure 6-6 contains $n = 80$ observations. Minitab software calculates the first and third quartiles as the$(n + 1)/4$ and $3(n + 1)/4$ ordered observations and interpolates as needed.

For example, $(80 + 1)/4 = 20.25$ and $3(80 + 1)/4 = 60.75$.

Therefore, Minitab interpolates between the 20th and 21st ordered observation to obtain $q_1 = 143.50$ and between the 60th and 61st observation to obtain $q_3 = 181.00$.

6-2 Stem-and-Leaf Diagrams

Data Features

- The ***interquartile range*** is the difference between the upper and lower quartiles, and it is sometimes used as a measure of variability. (IQR = $Q_3 - Q_1$)

- In general, the 100*k*th ***percentile*** is a data value such that approximately 100*k*% of the observations are at or below this value and approximately 100(1 - *k*)% of them are above it.

6-3 Frequency Distributions and Histograms

- A **frequency distribution** is a more compact summary of data than a stem-and-leaf diagram.

- To construct a frequency distribution, we must divide the range of the data into intervals, which are usually called **class intervals**, **cells** or **bins**.

Constructing a Histogram (Equal Bin Widths):

(1) Label the bin (class interval) boundaries on a horizontal scale.
(2) Mark and label the vertical scale with the frequencies or the relative frequencies.
(3) Above each bin, draw a rectangle where height is equal to the frequency (or relative frequency) corresponding to that bin.

6-3 Frequency Distributions and Histograms

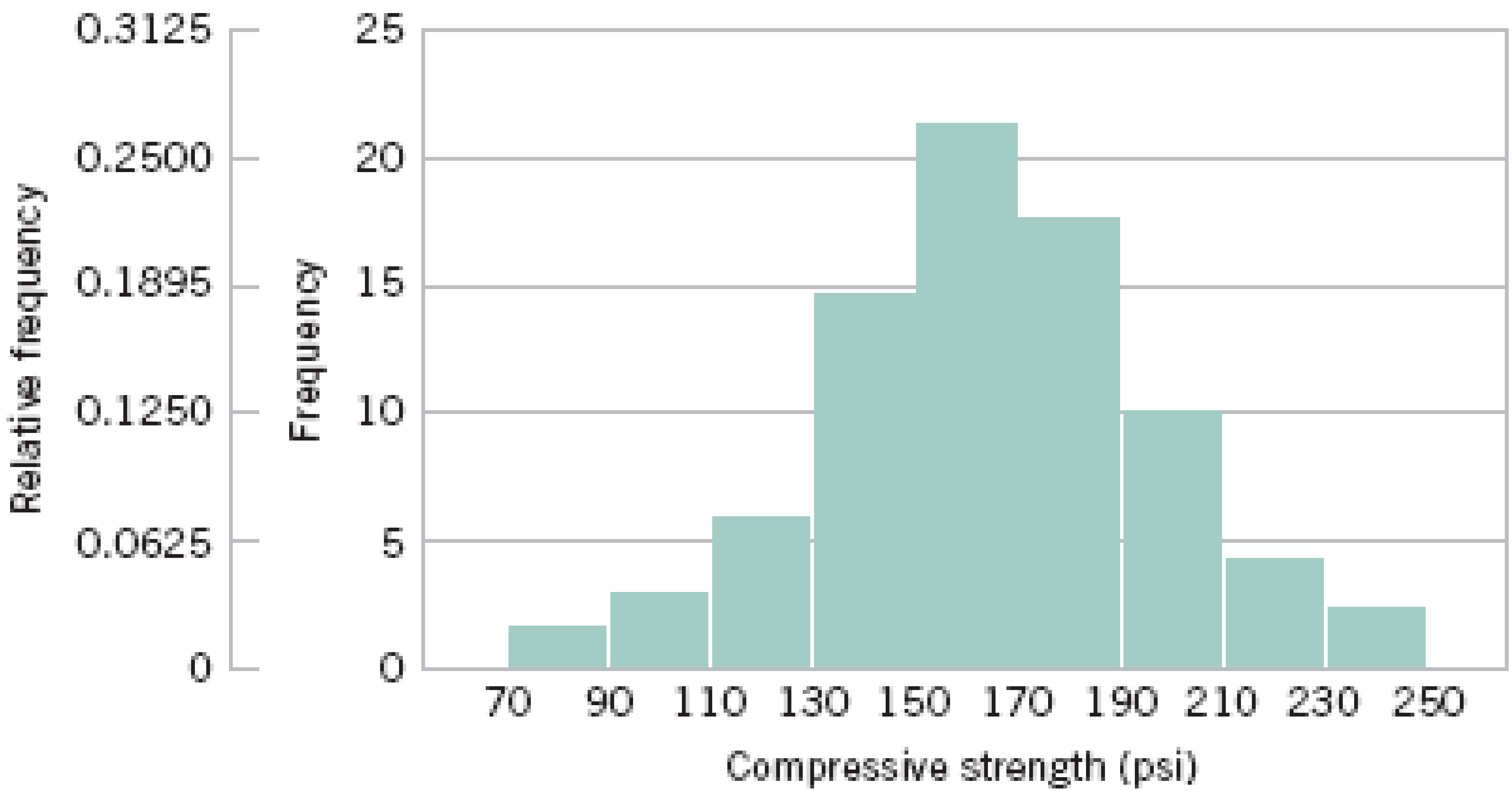


Figure 6-7 Histogram of compressive strength for 80 aluminum-lithium alloy specimens.

6-3 Frequency Distributions and Histograms

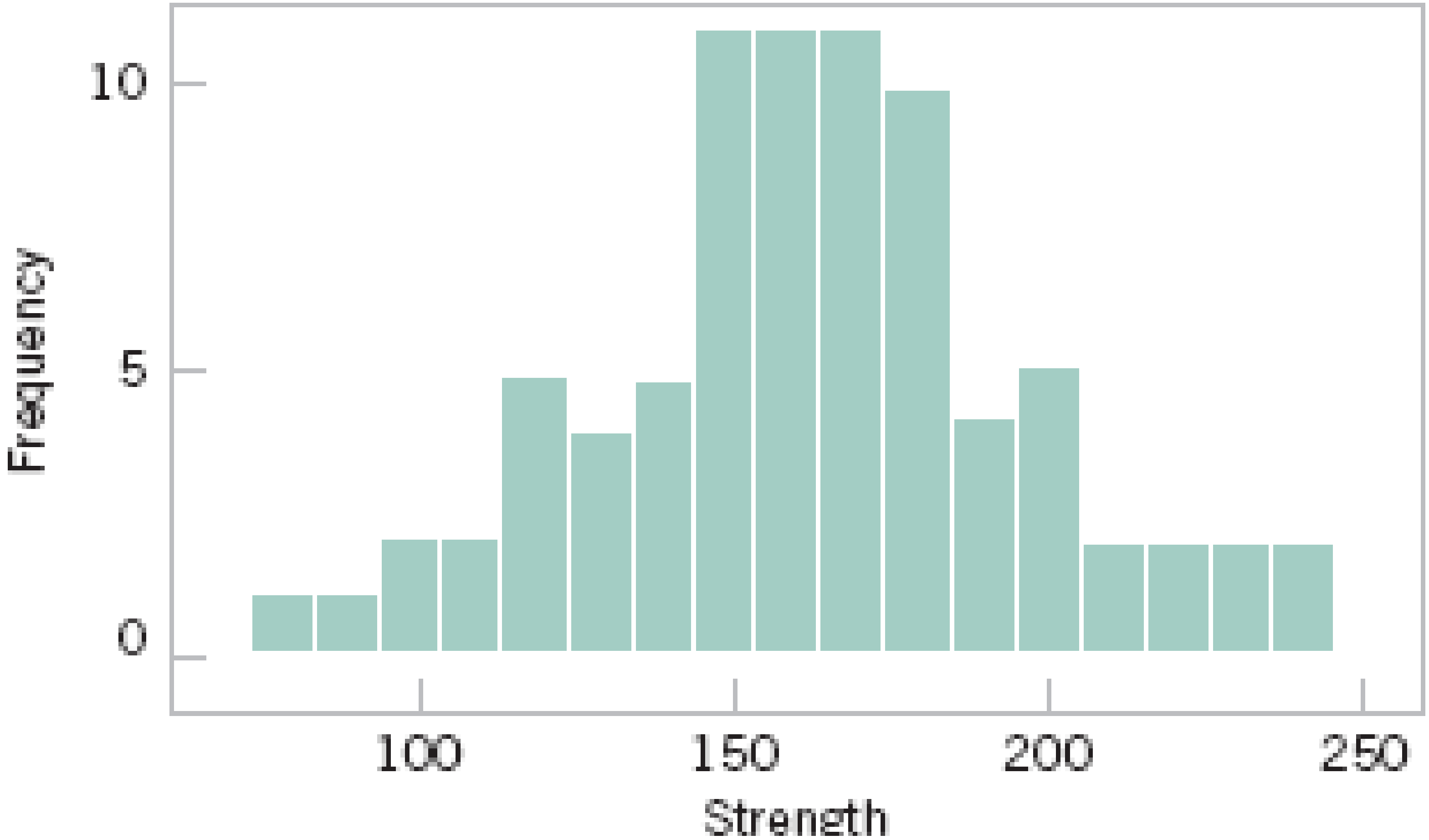


Figure 6-8 A histogram of the compressive strength data from Minitab with 17 bins.

6-3 Frequency Distributions and Histograms

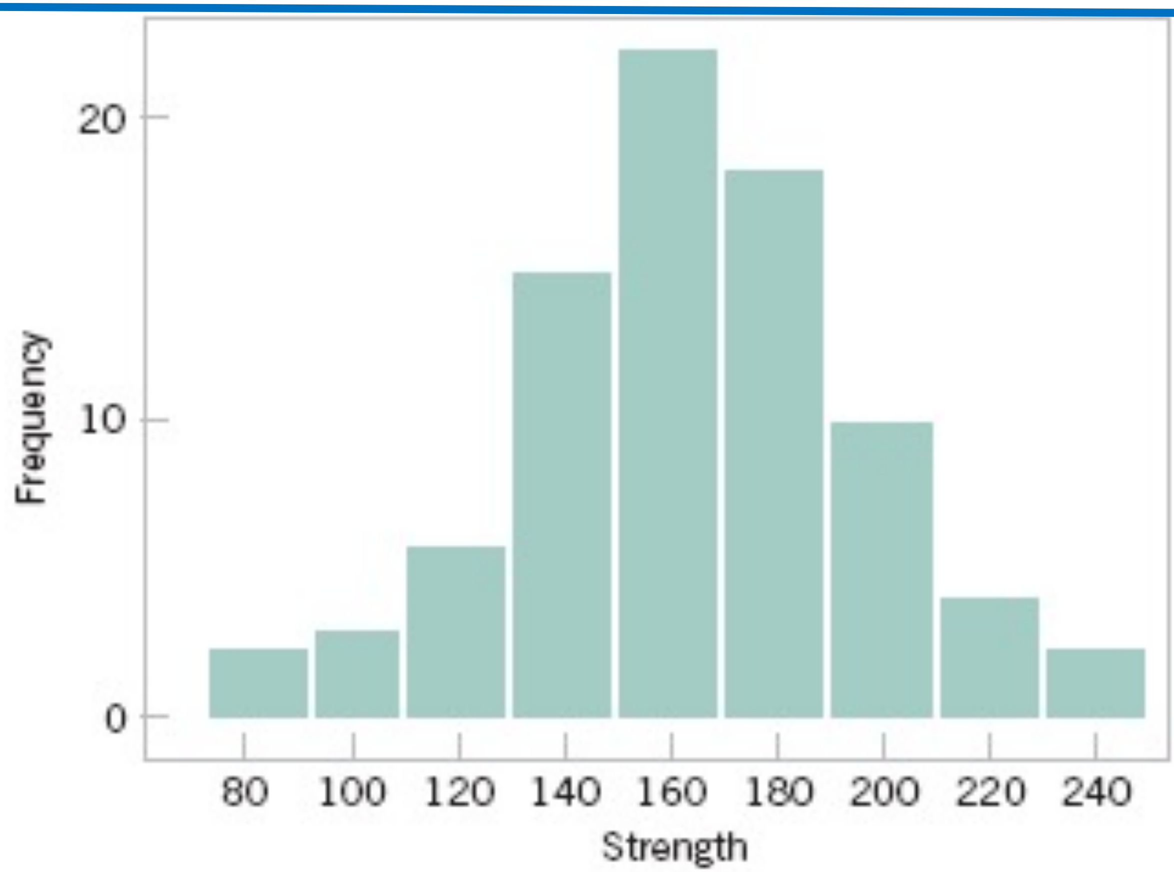


Figure 6-9 A histogram of the compressive strength data from Minitab with nine bins.

Figure 6-9 A histogram of the compressive strength data from Minitab with nine bins.

6-3 Frequency Distributions and Histograms

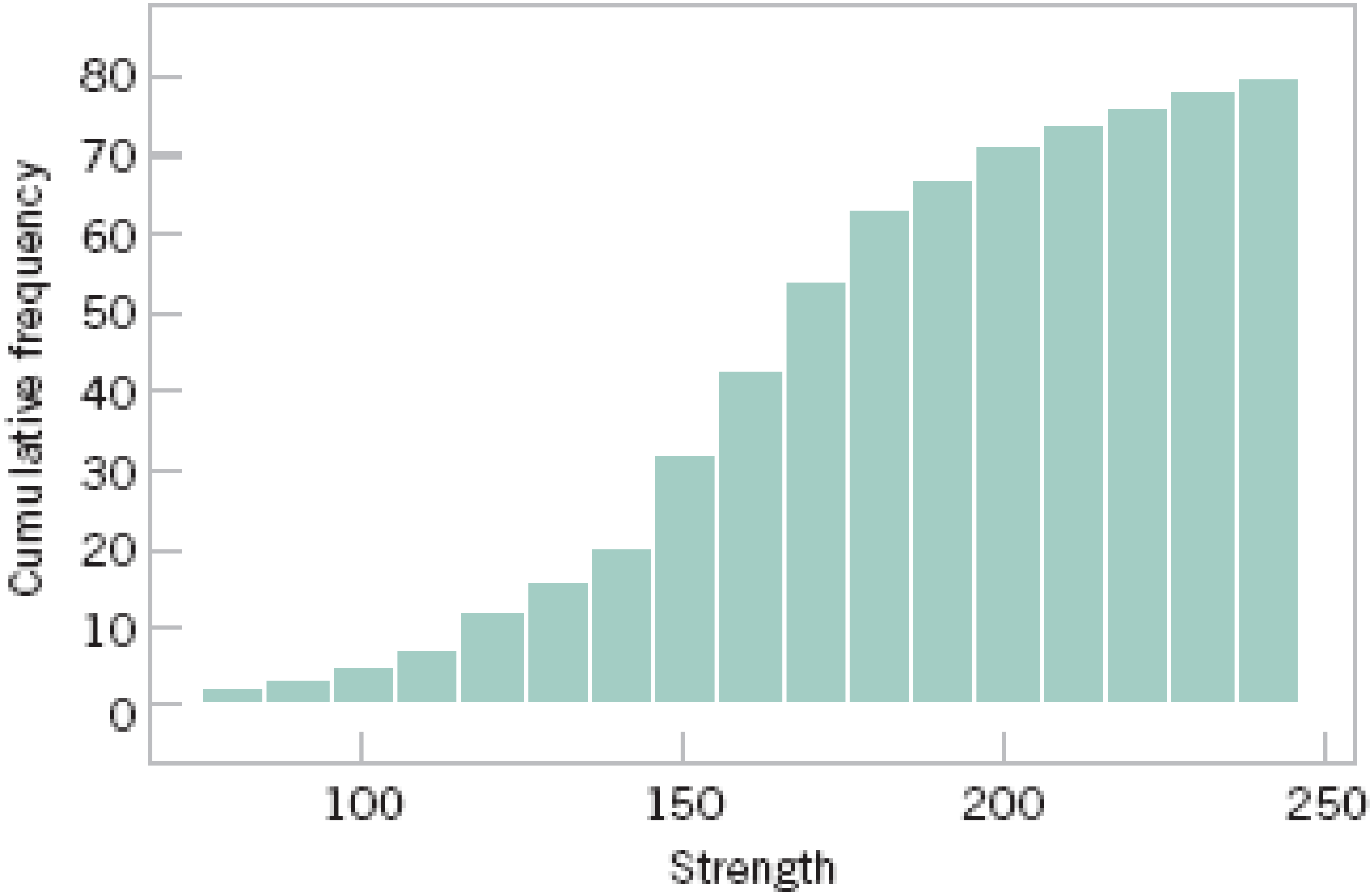


Figure 6-10 A cumulative distribution plot of the compressive strength data from Minitab.

6-3 Frequency Distributions and Histograms

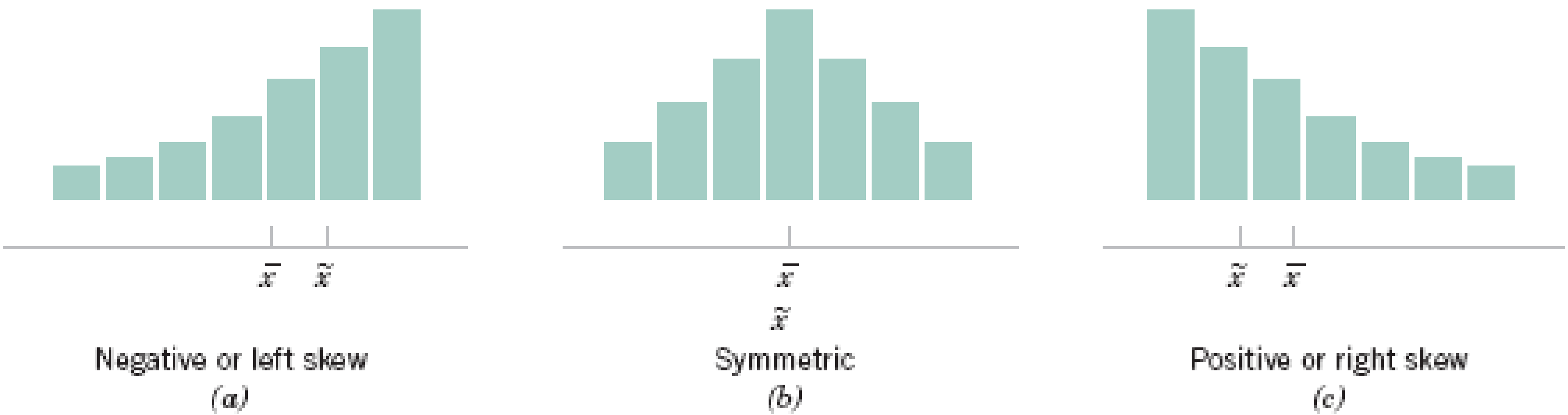


Figure 6-11 Histograms for symmetric and skewed distributions.

6-4 Box Plots

- The **box plot** is a graphical display that simultaneously describes several important features of a data set, such as center, spread, departure from symmetry, and identification of observations that lie unusually far from the bulk of the data.

- **Whisker**
- **Outlier**
- **Extreme outlier**

6-4 Box Plots

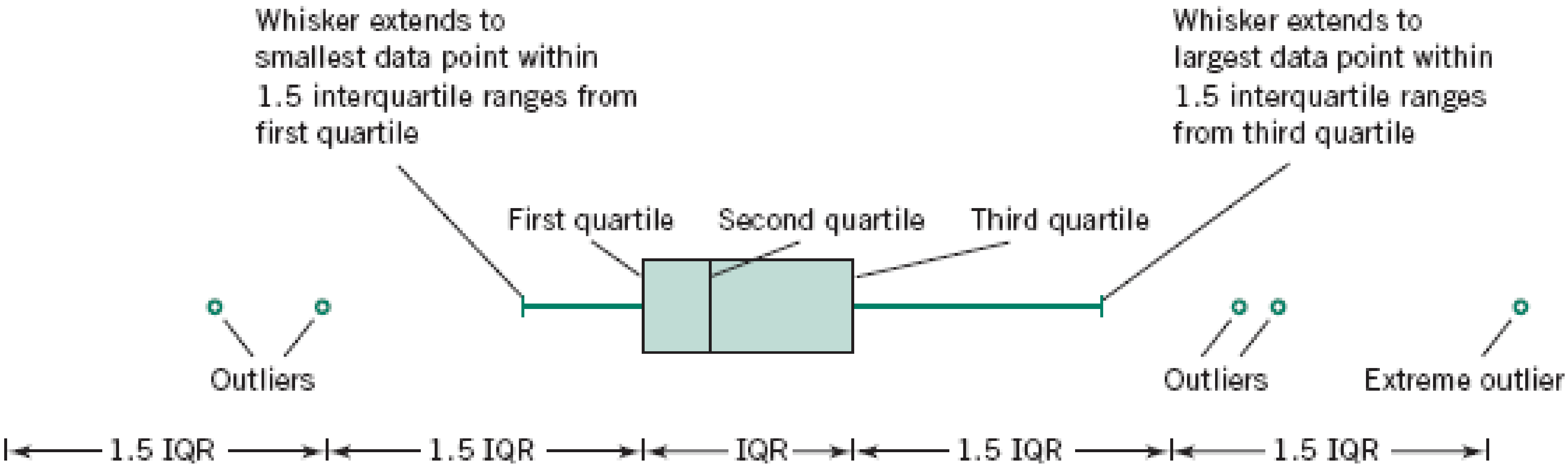


Figure 6-13 Description of a box plot.

6-4 Box Plots

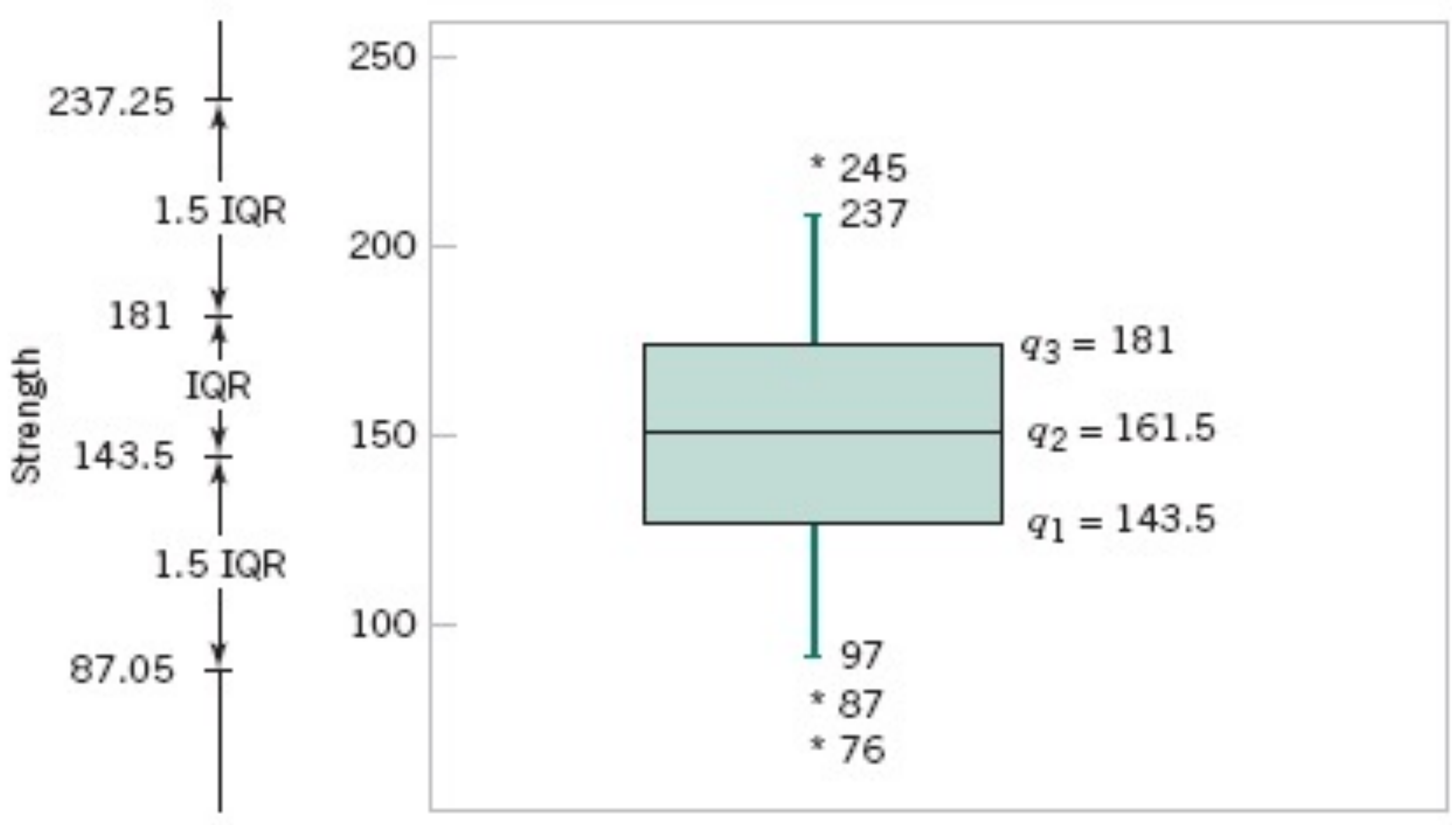


Figure 6-14 Box plot for compressive strength data in Table 6-2.

6-5 Time Sequence Plots

- A **time series** or **time sequence** is a data set in which the observations are recorded in the order in which they occur.

- A **time series plot** is a graph in which the vertical axis denotes the observed value of the variable (say $\boldsymbol{x}$) and the horizontal axis denotes the time (which could be minutes, days, years, etc.).

- When measurements are plotted as a time series, we often see
  - **trends,**
  - **cycles, or**
  - **other broad features of the data**

6-5 Time Sequence Plots

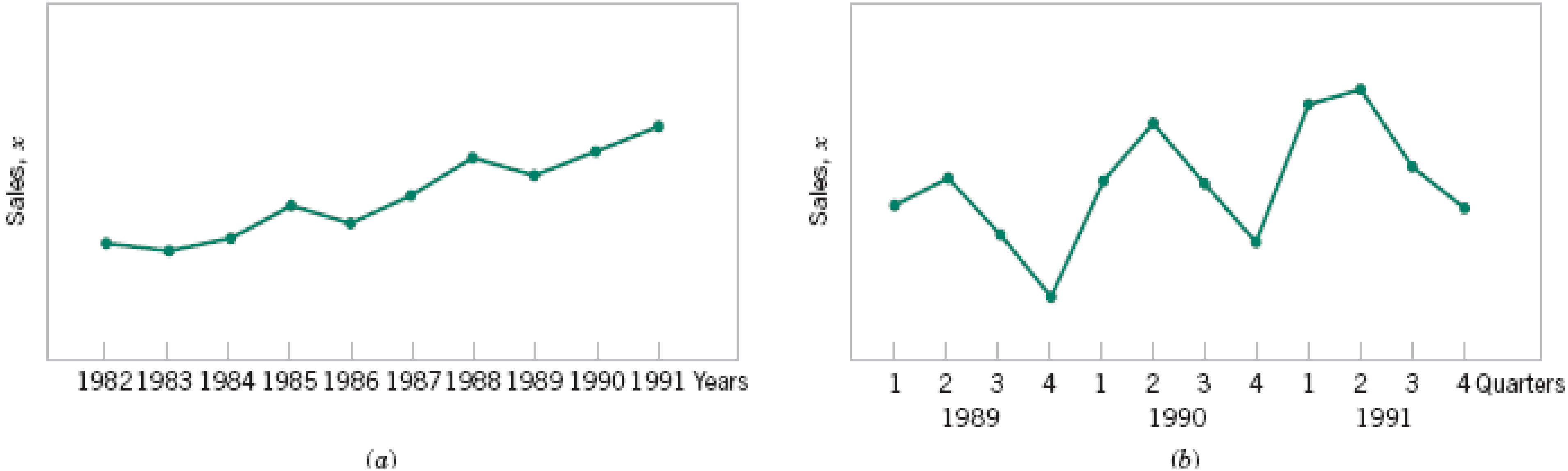


Figure 6-16 Company sales by year (a) and by quarter (b).

6-5 Time Sequence Plots

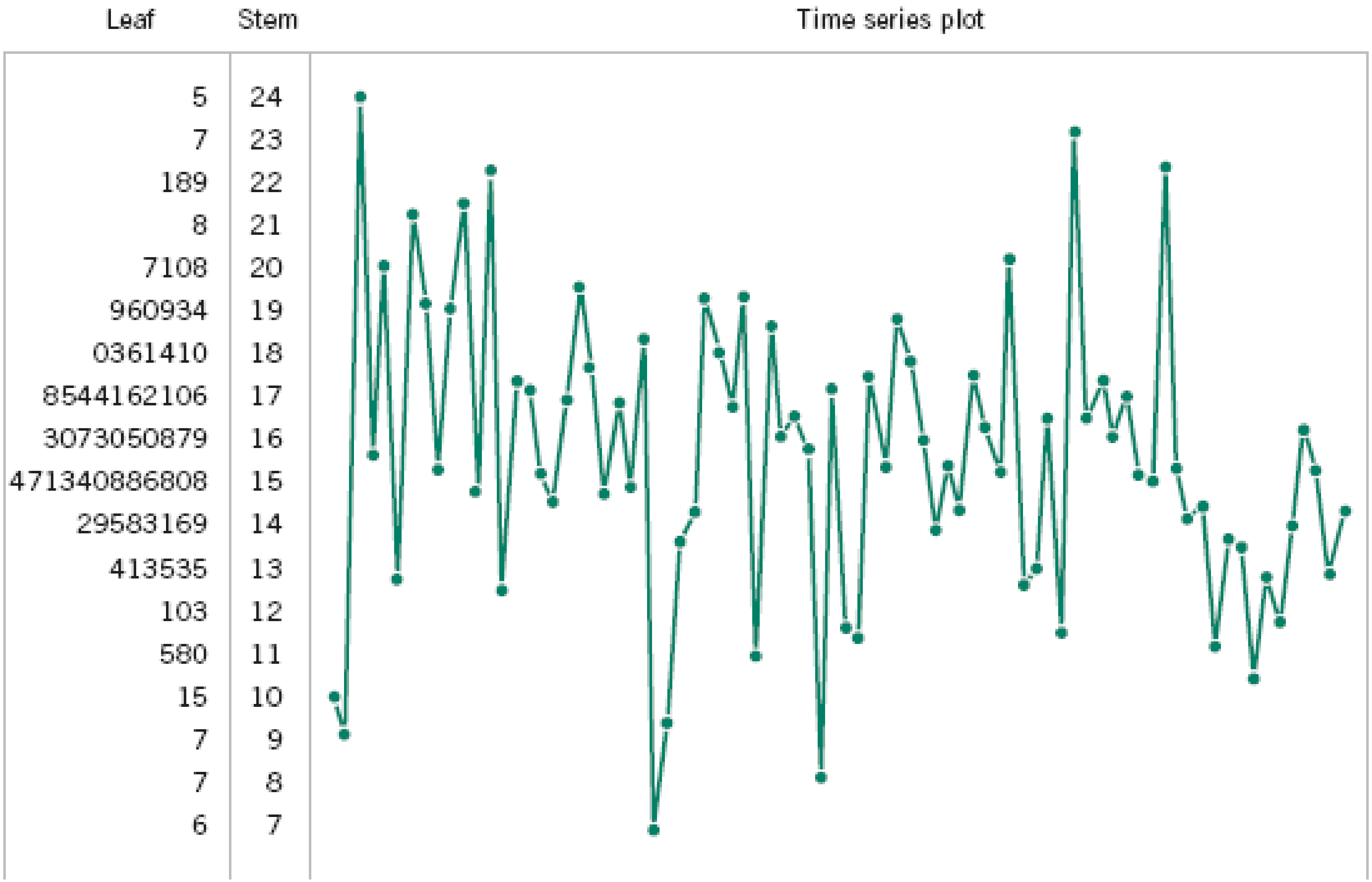


Figure 6-17 A digidot plot of the compressive strength data in Table 6-2.

6-5 Time Sequence Plots

Leaf	Stem
8	9e
6	9s
45	9f
2333	9t
0010000	9z
99998	8e
66677	8s
45	8f
23	8t
1	8z

Figure 6-18 A digidot plot of chemical process concentration readings, observed hourly.

IMPORTANT TERMS AND CONCEPTS

- Box plot
- Frequency distribution and histogram
- Median, quartiles and percentiles
- Multivariable Data
- Normal probability plot
- Pareto chart
- Population mean
- Population standard deviation
- Population variance
- Probability plot
- Relative Frequency Distribution
- Sample mean
- Sample standard deviation
- Sample variance
- Stem-and-leaf diagram
- Time series plots