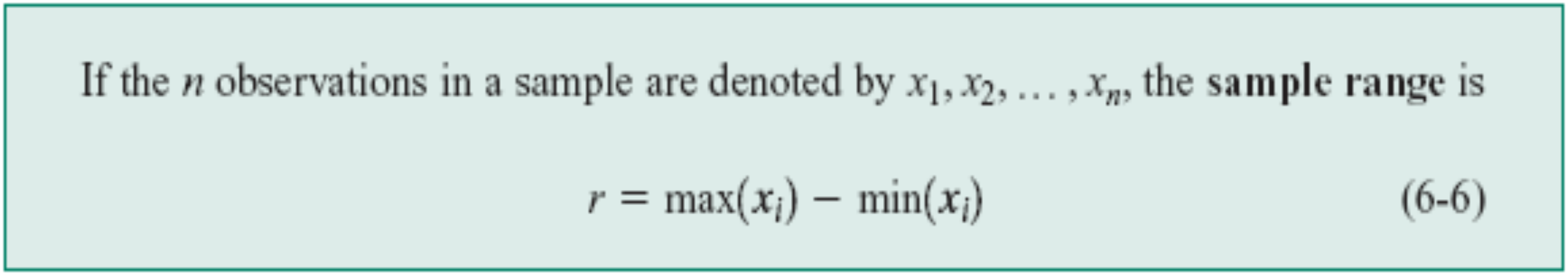
6 Random Sampling and Data Description

CHAPTER OUTLINE

6-1 Numerical Summaries

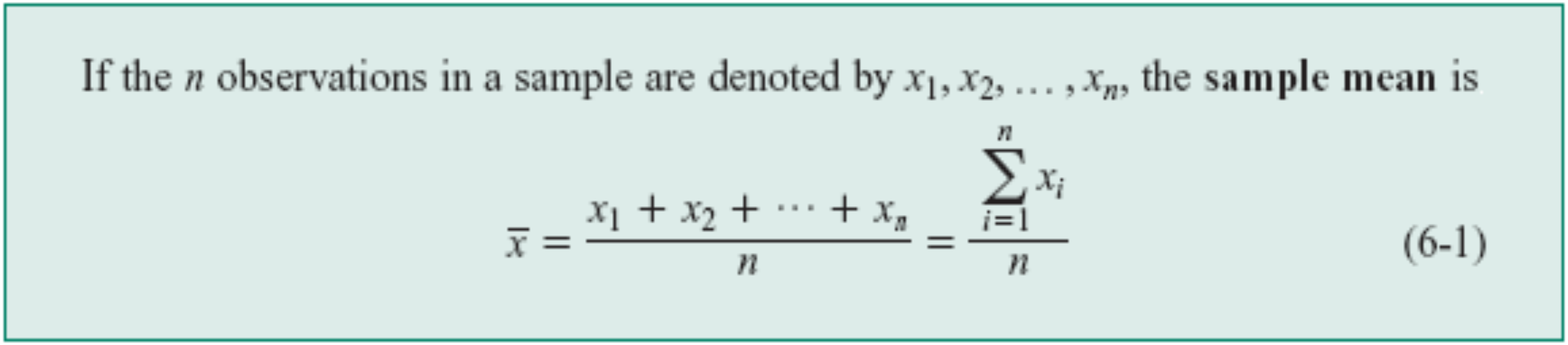
Definition: Sample Range



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

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

Definition: Sample Mean

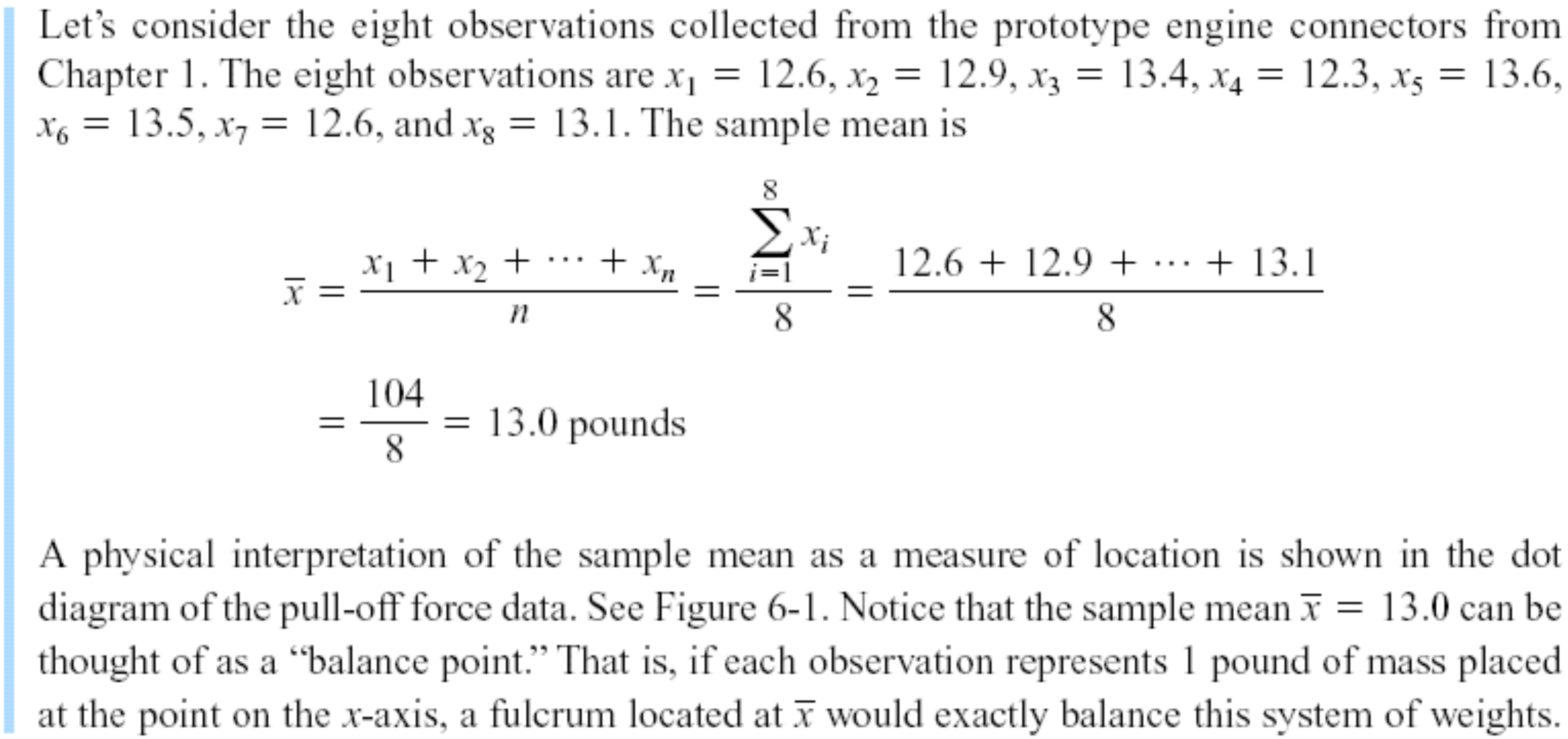


If the n observations in a sample are denoted by $x_1, x_2, \ldots, x_n$, the **sample mean** is

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

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

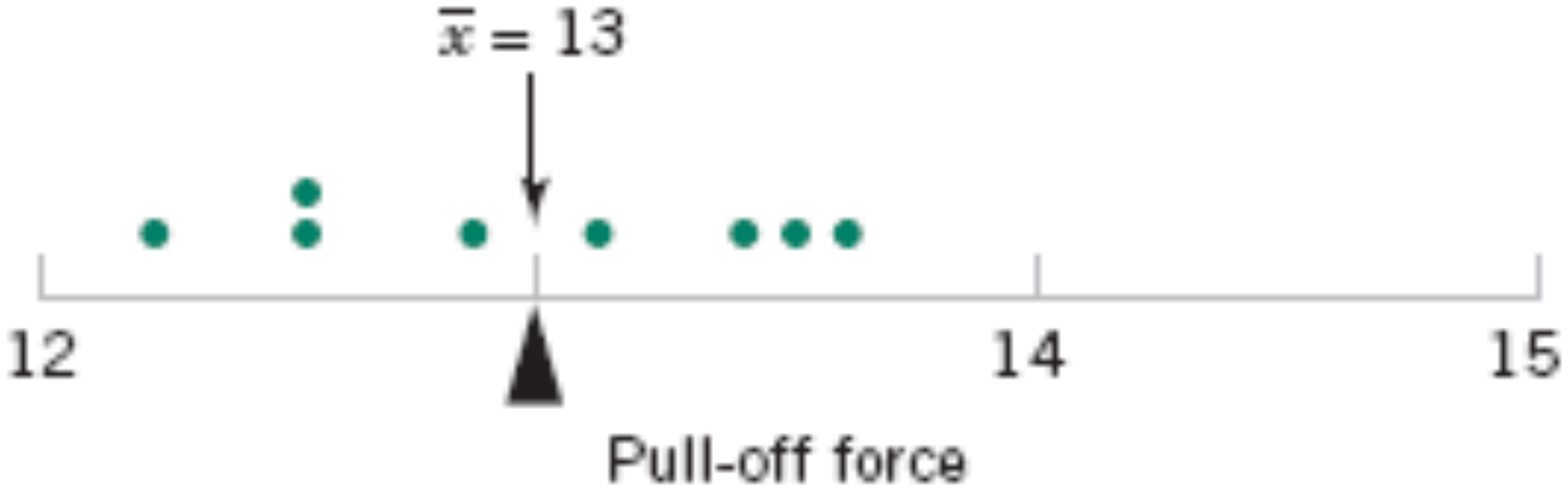


Figure 6-1 The sample mean as a balance point for a system of weights.

6-1 Numerical Summaries

Population Mean

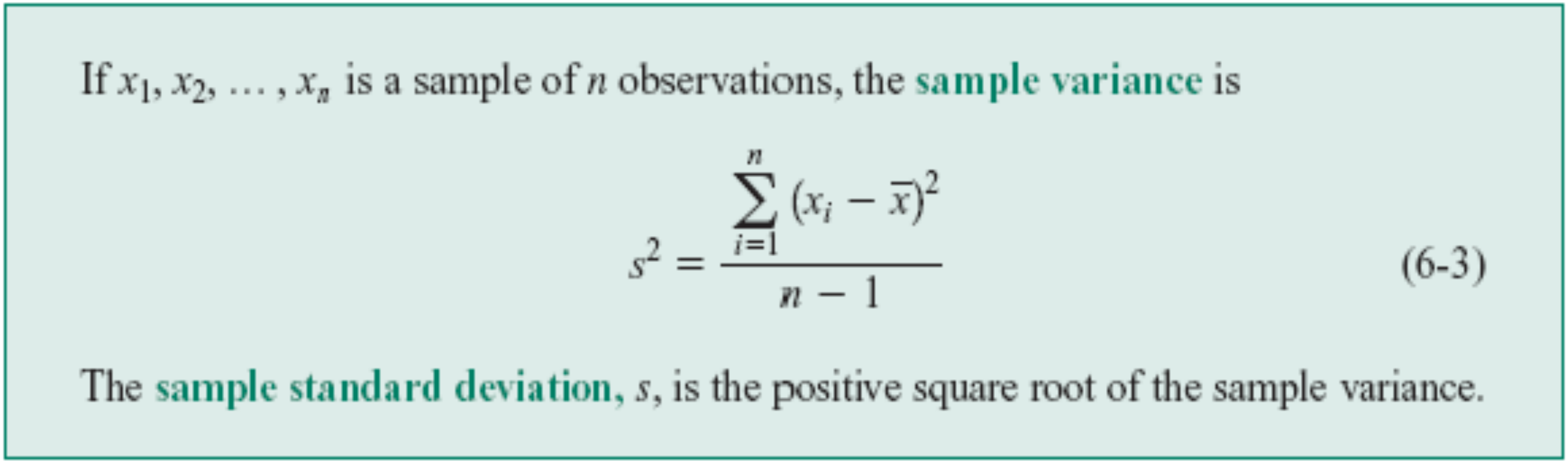
For a finite population with *N* (equally likely) measurements, the mean is

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

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

6-1 Numerical Summaries

Definition: Sample Variance



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

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

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

- n-1 is referred to as the **degrees of freedom**.

6-1 Numerical Summaries

How Does the Sample Variance Measure Variability?

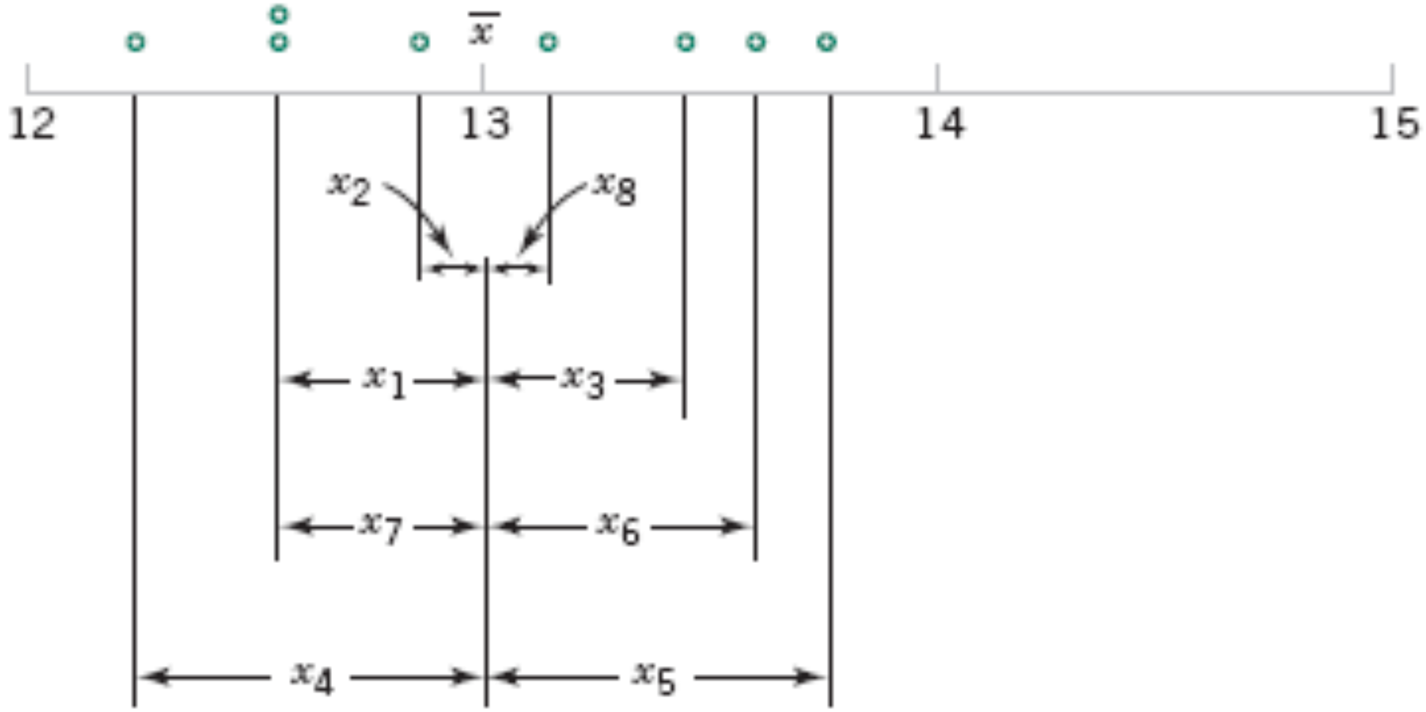


Figure 6-2 How the sample variance measures variability through the deviations $x_i - \overline{x}$.

6-1 Numerical Summaries

Example 6-2

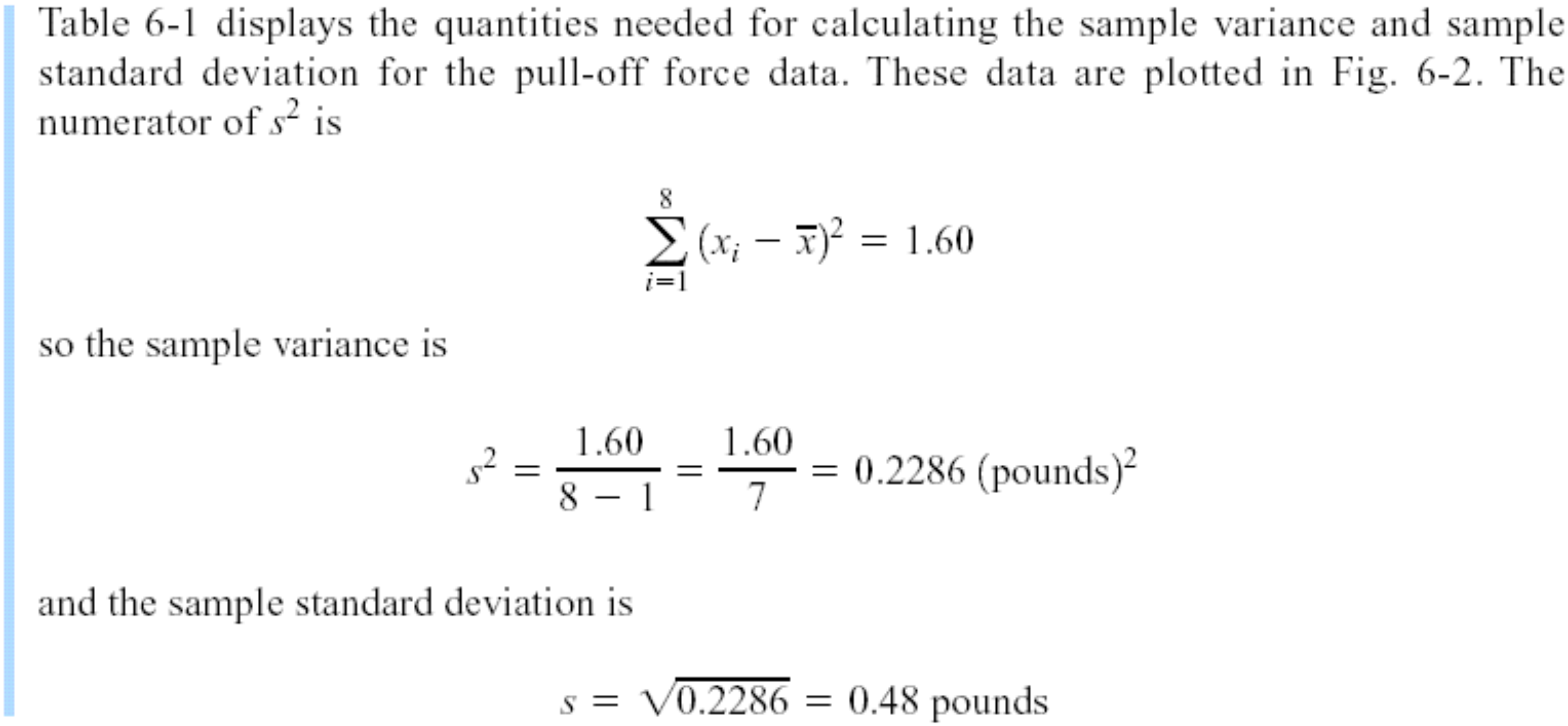


Table 6-1 displays the quantities needed for calculating the sample variance and sample standard deviation for the pull-off force data. These data are plotted in Fig. 6-2. The numerator of s^2 is

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

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

Computation of s^2

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

6-1 Numerical Summaries

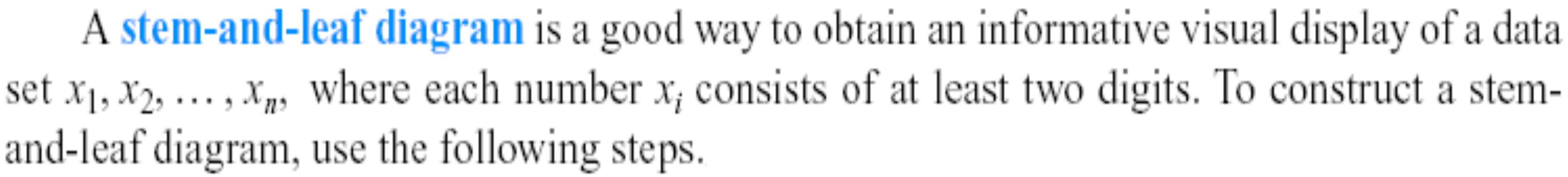
Population Variance

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

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

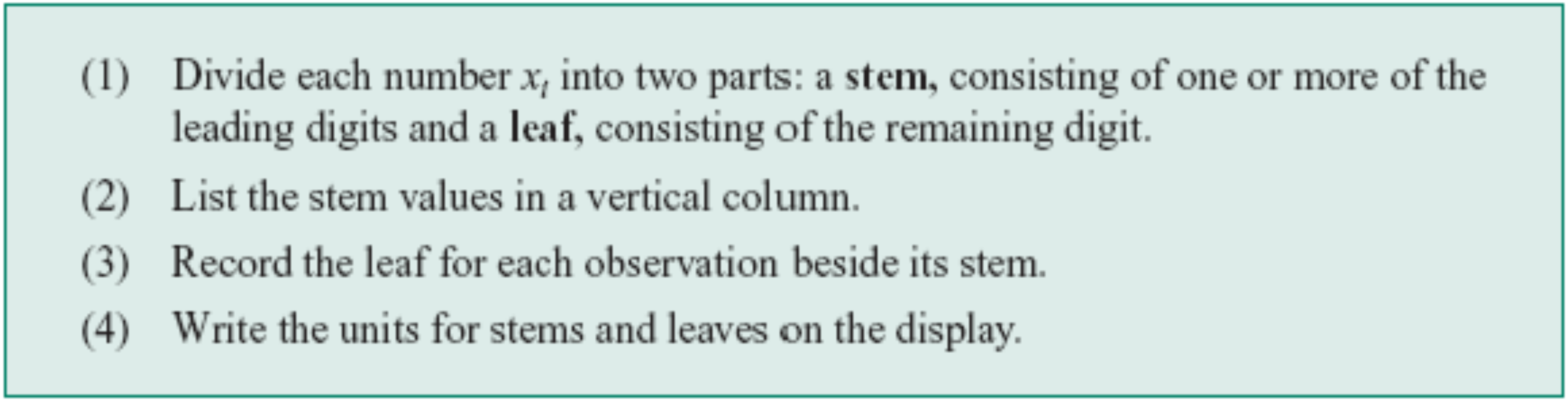
The sample variance is a reasonable estimate of the population variance.

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.

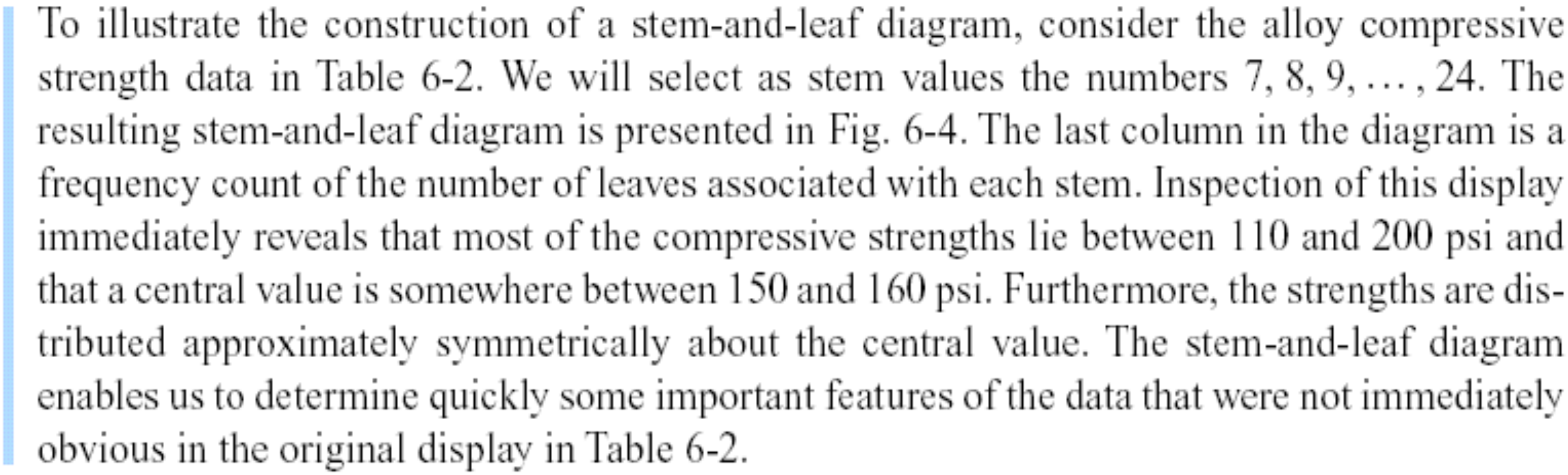
(2) List the stem values in a vertical column.

(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

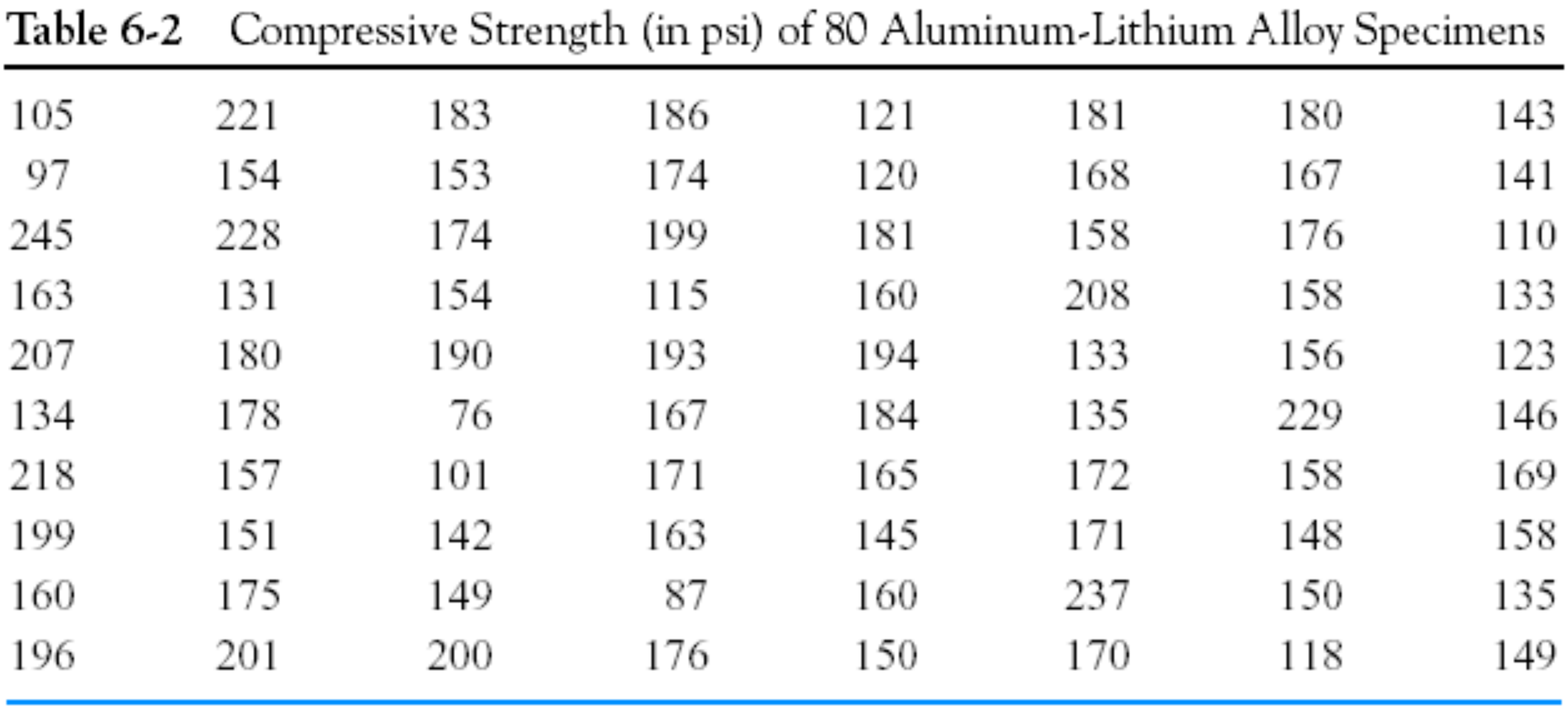
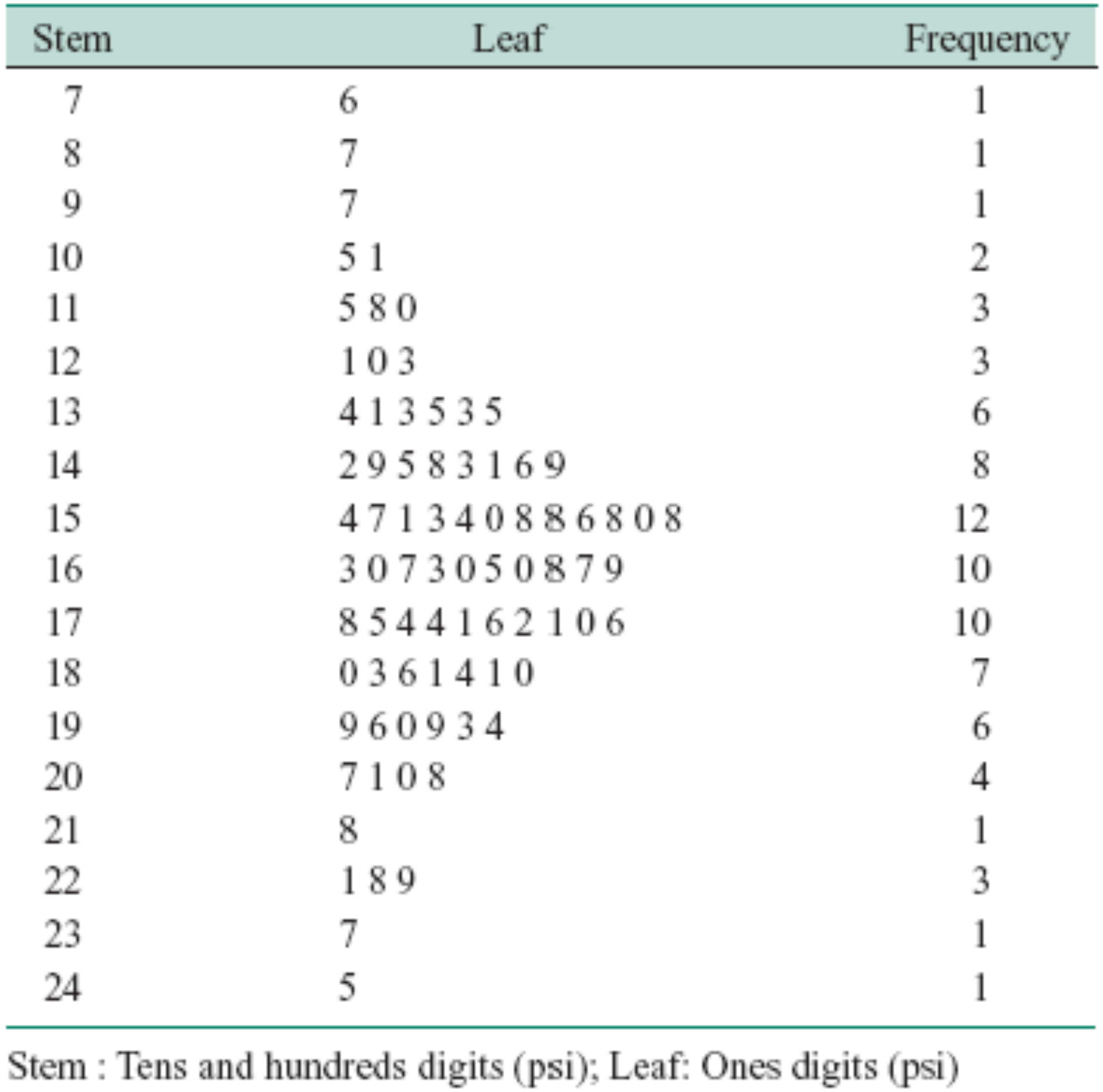


Table 6-2 Compressive Strength (in psi) of 80 Aluminum-Lithium Alloy Specimens

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.

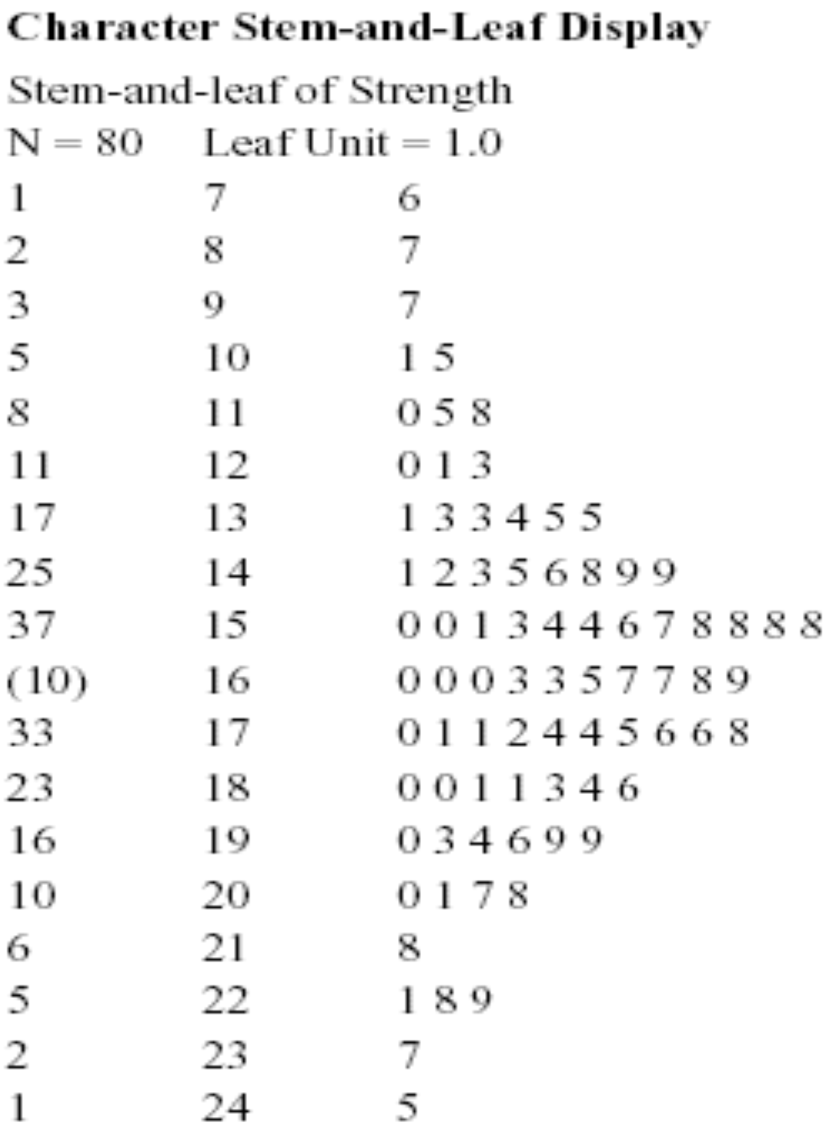


Stem	Leaf	Frequency
7	6	1
8	7	1
9	7	1
10	5 1	2
11	5 8 0	3
12	1 0 3	3
13	4 1 3 5 3 5	6
14	2 9 5 8 3 1 6 9	8
15	4 7 1 3 4 0 8 8 6 8 0 8	12
16	3 0 7 3 0 5 0 8 7 9	10
17	8 5 4 4 1 6 2 1 0 6	10
18	0 3 6 1 4 1 0	7
19	9 6 0 9 3 4	6
20	7 1 0 8	4
21	8	1
22	1 8 9	3
23	7	1
24	5	1

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

6-2 Stem-and-Leaf Diagrams

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.

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.

- 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.

6-4 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

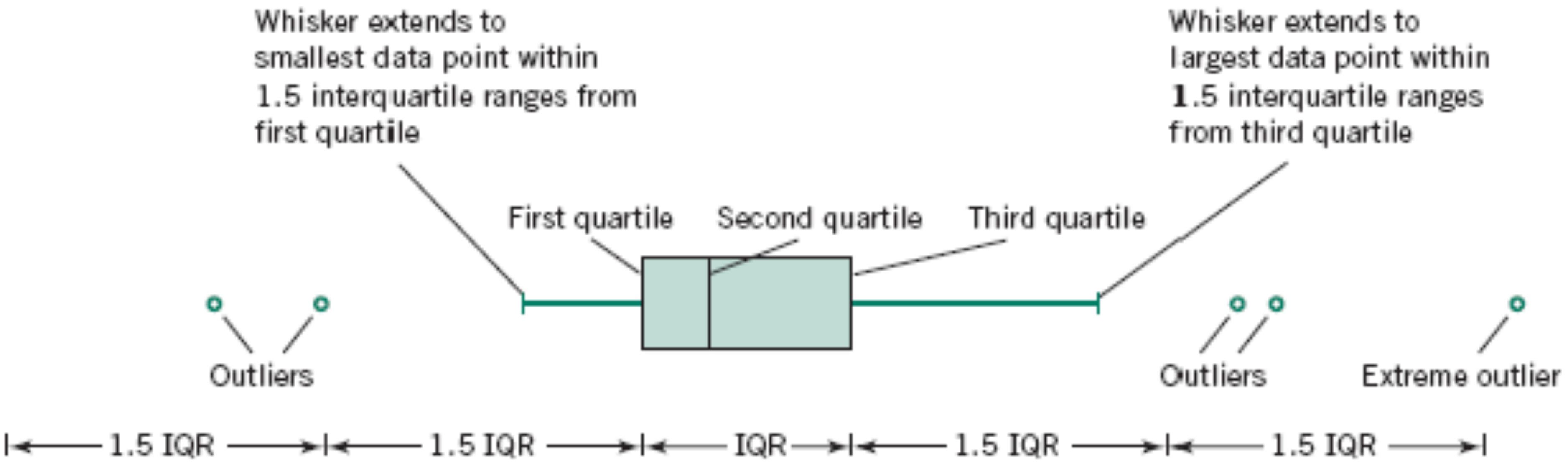

Whisker extends to smallest data point within 1.5 interquartile ranges from first quartile
Whisker extends to largest data point within 1.5 interquartile ranges from third quartile
First quartile
Second quartile
Third quartile
Outliers
Outliers
Extreme outlier
1.5 IQR
1.5 IQR
IQR
1.5 IQR
1.5 IQR

Figure 6-13 Description of a box plot.

6-4 Box Plots

Example: The ordered data in Example 6-1 are
12.3, 12.6, 12.6, 12.9, 13.1, 13.4, 13.5, 13.6

6-4 Box Plots

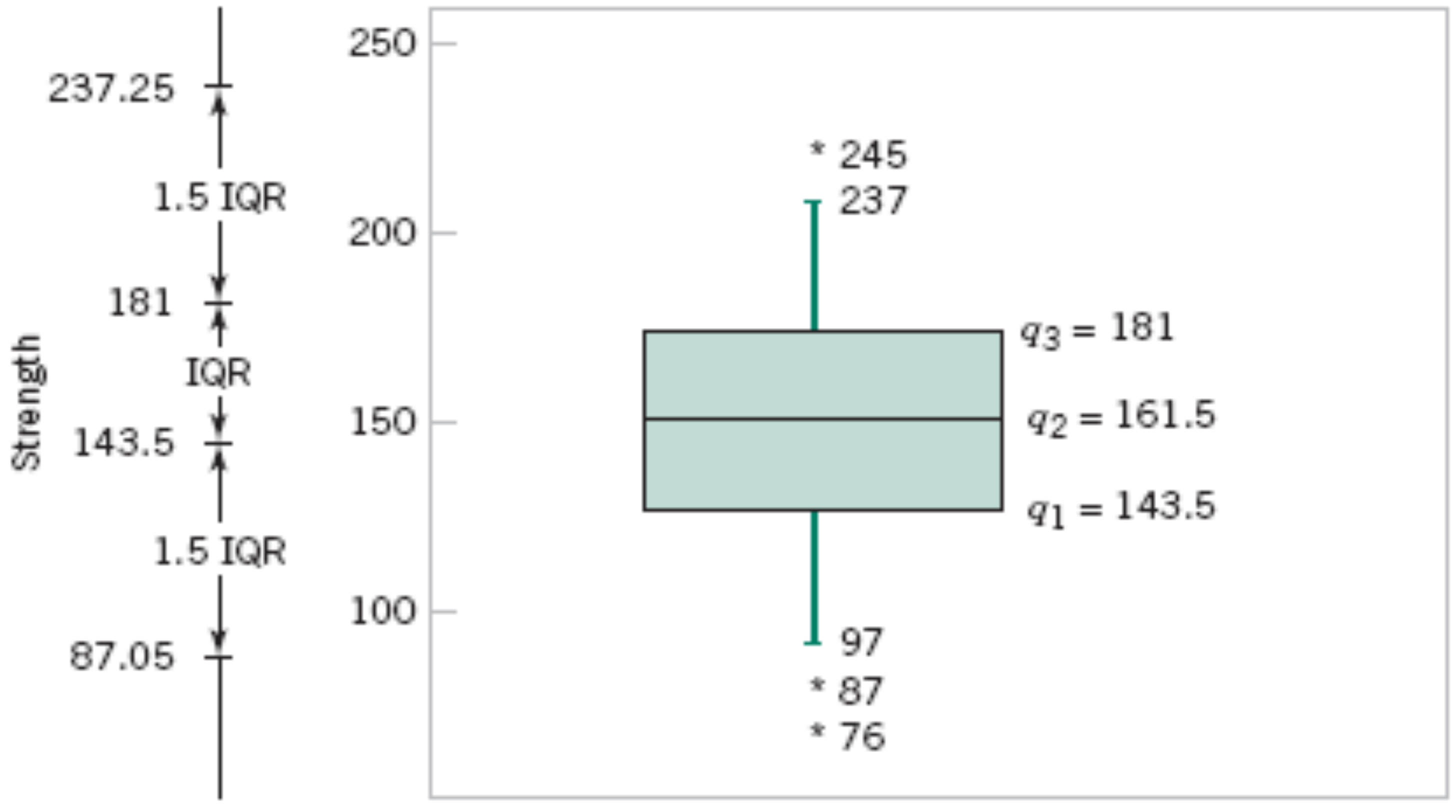


Figure 6-14 Box plot for compressive strength data in Table 6-2.

6-4 Box Plots

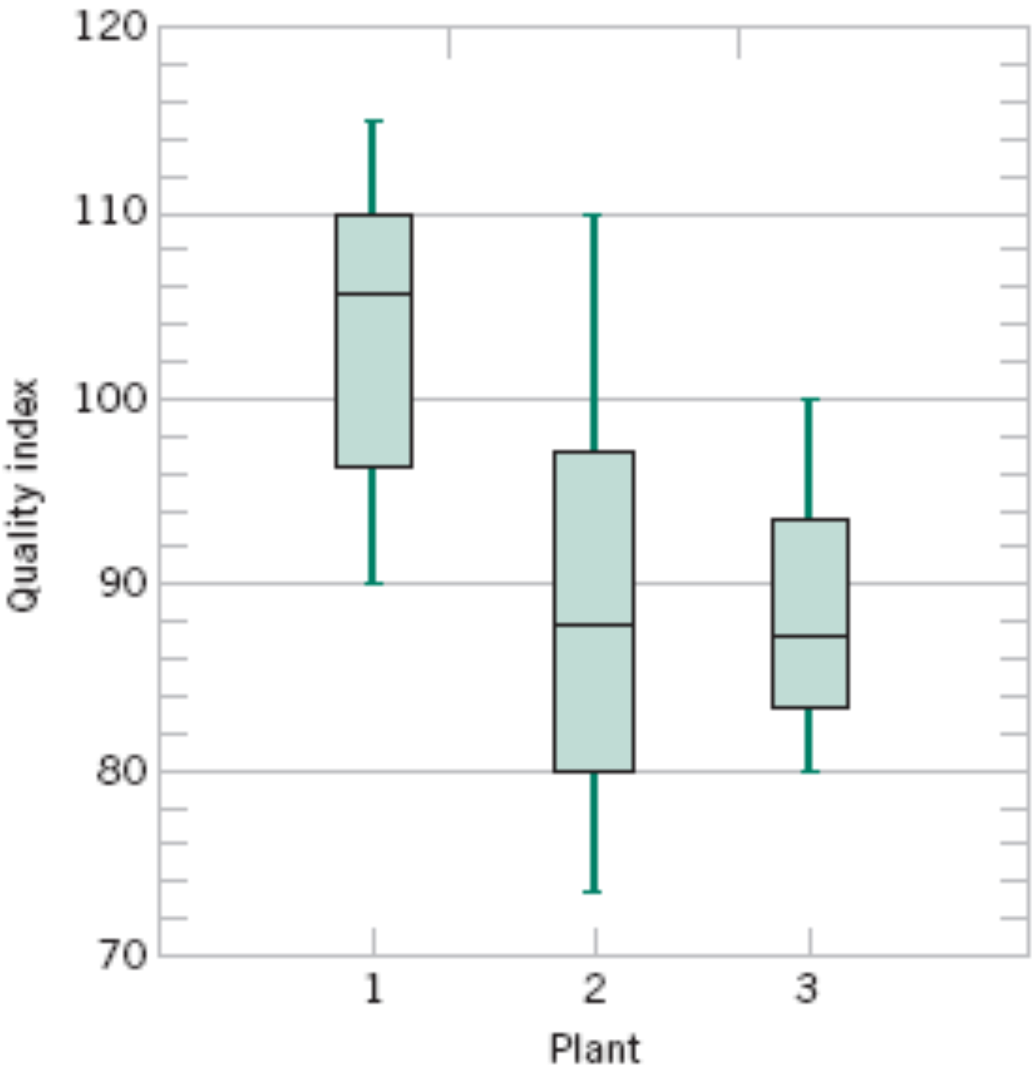


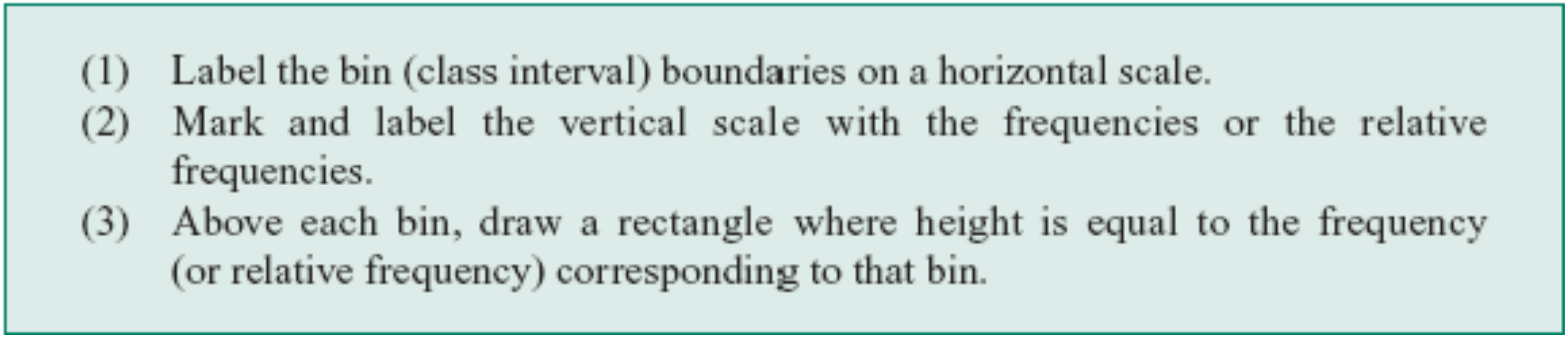
Figure 6-15
Comparative box plots of a quality index at three plants.

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

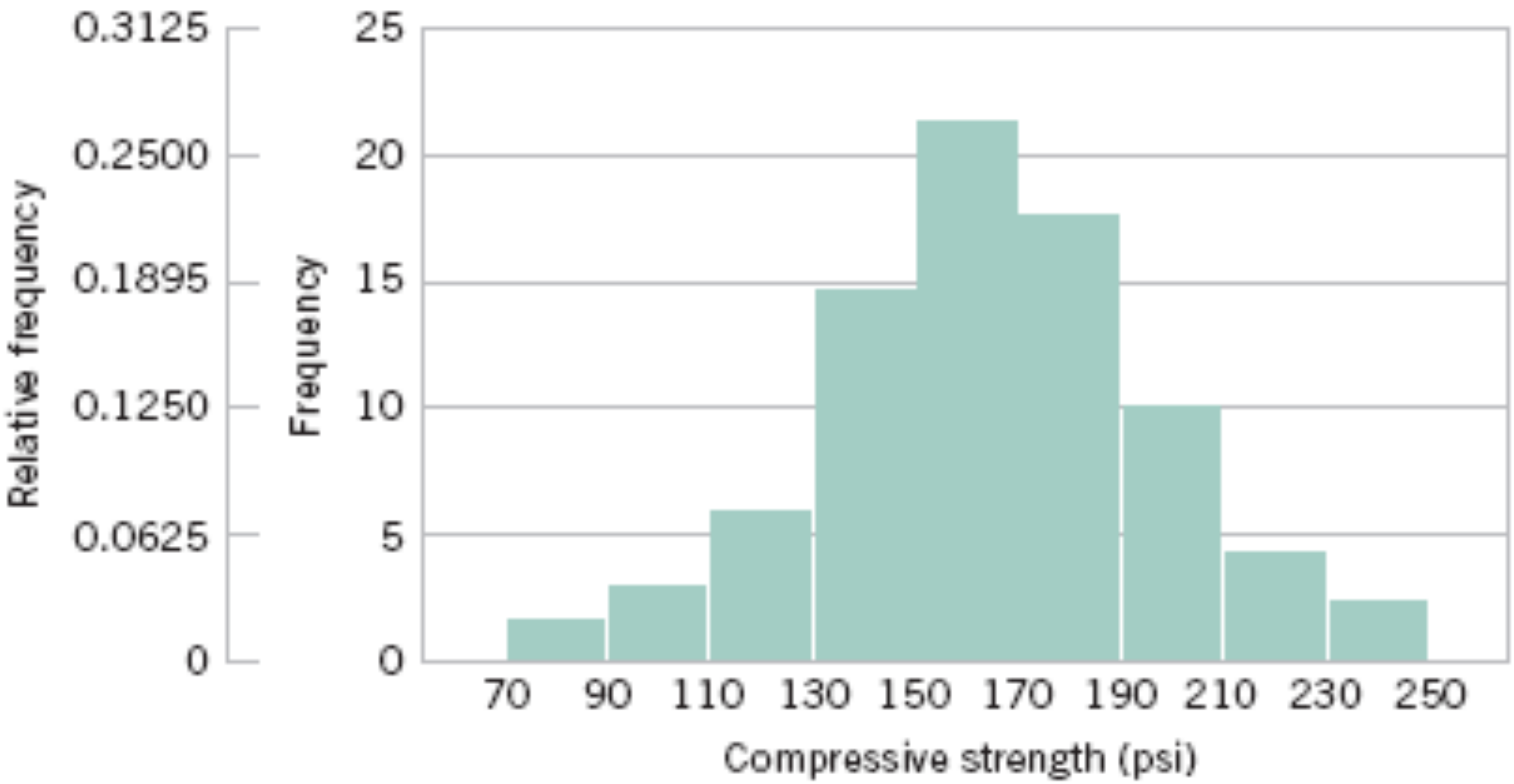


Figure 6-7 Histogram of compressive strength for 80 aluminum-lithium alloy specimens.

6-3 Frequency Distributions and Histograms

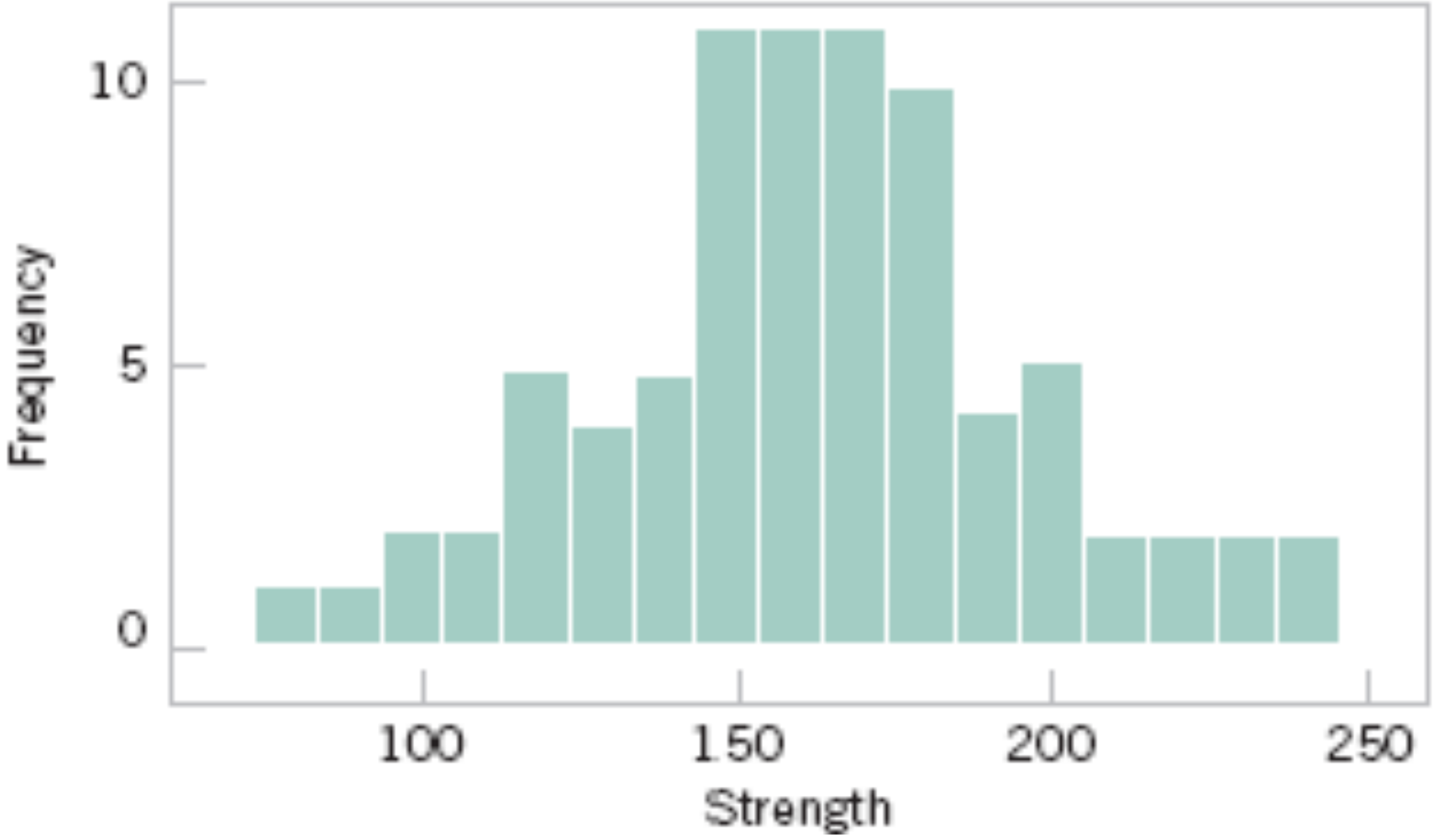


Figure 6-8 A histogram of the compressive strength data from Minitab with 17 bins.

6-3 Frequency Distributions and Histograms

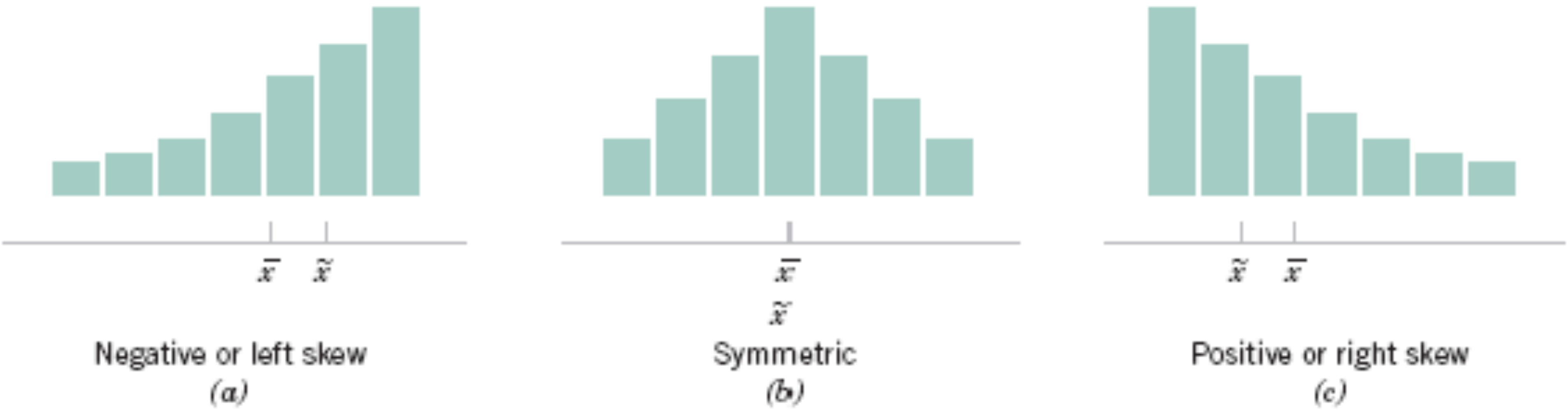


Figure 6-11 Histograms for symmetric and skewed distributions.

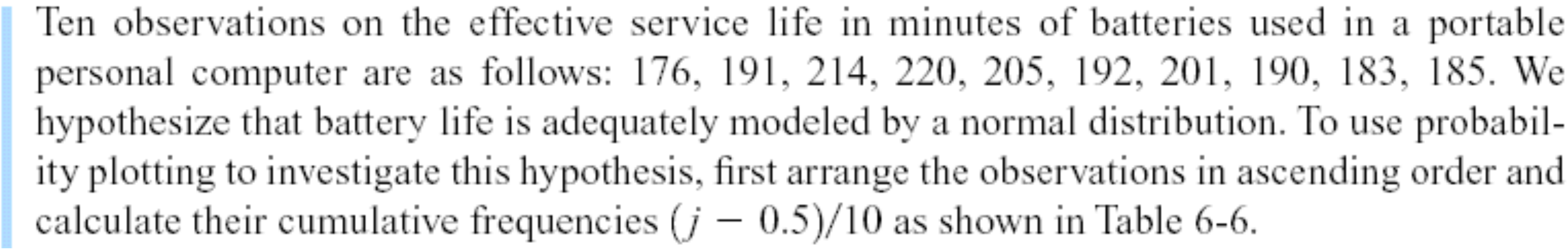
6-6 Probability Plots

- **Probability plotting** is a graphical method for determining whether sample data conform to a hypothesized distribution based on a subjective visual examination of the data.

- Probability plotting typically uses special graph paper, known as **probability paper,** that has been designed for the hypothesized distribution. Probability paper is widely available for the normal, lognormal, Weibull, and various chi-square and gamma distributions.

6-6 Probability Plots

Example 6-7



Ten observations on the effective service life in minutes of batteries used in a portable personal computer are as follows: 176, 191, 214, 220, 205, 192, 201, 190, 183, 185. We hypothesize that battery life is adequately modeled by a normal distribution. To use probability plotting to investigate this hypothesis, first arrange the observations in ascending order and calculate their cumulative frequencies $(j - 0.5)/10$ as shown in Table 6-6.

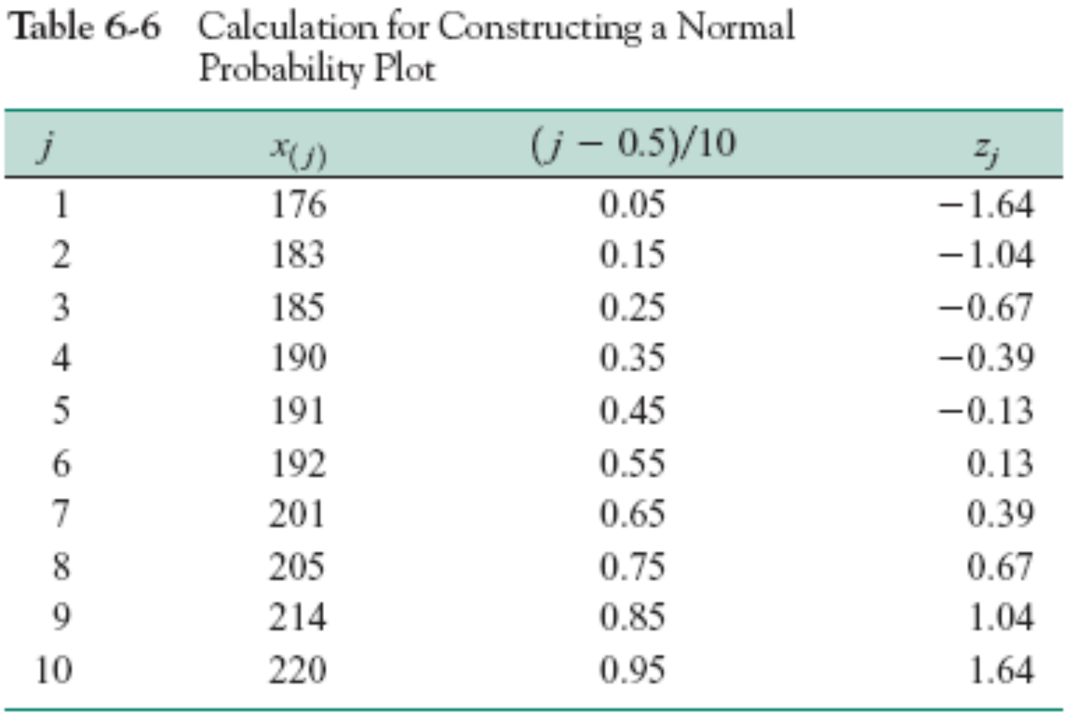
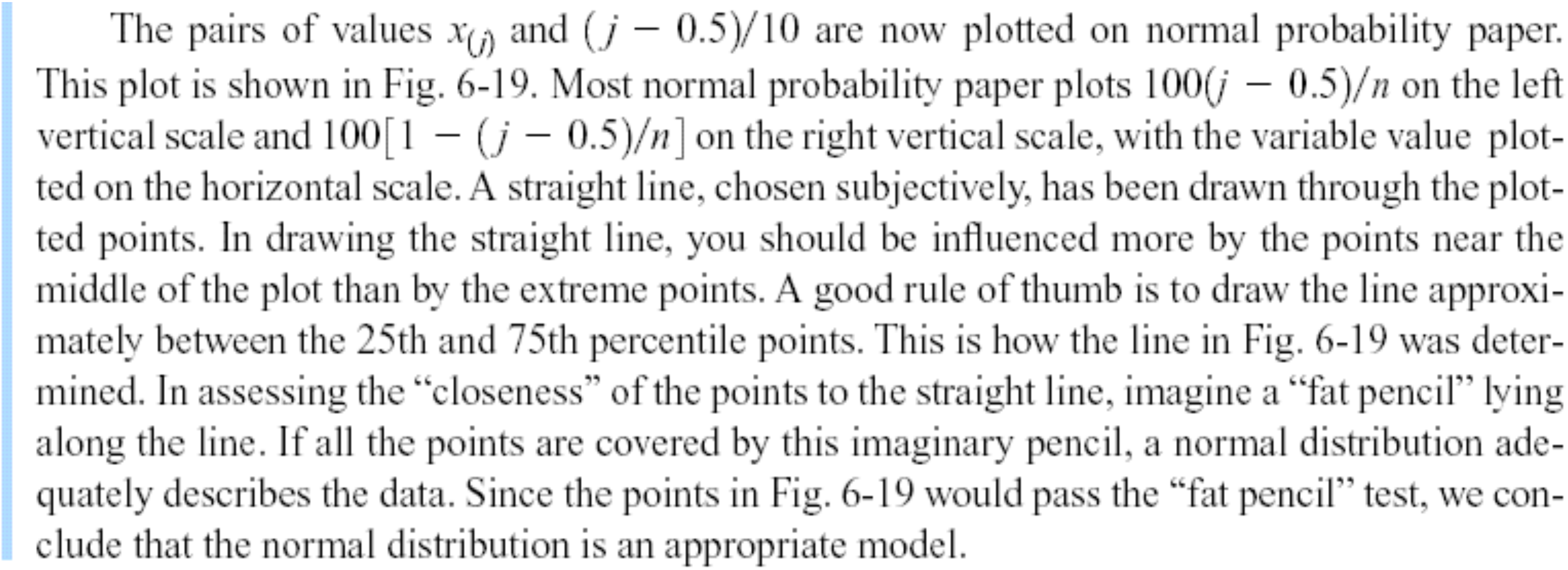


Table 6-6 Calculation for Constructing a Normal Probability Plot

j	$x_{(j)}$	$(j - 0.5)/10$	z_j
1	176	0.05	−1.64
2	183	0.15	−1.04
3	185	0.25	−0.67
4	190	0.35	−0.39
5	191	0.45	−0.13
6	192	0.55	0.13
7	201	0.65	0.39
8	205	0.75	0.67
9	214	0.85	1.04
10	220	0.95	1.64

6-6 Probability Plots

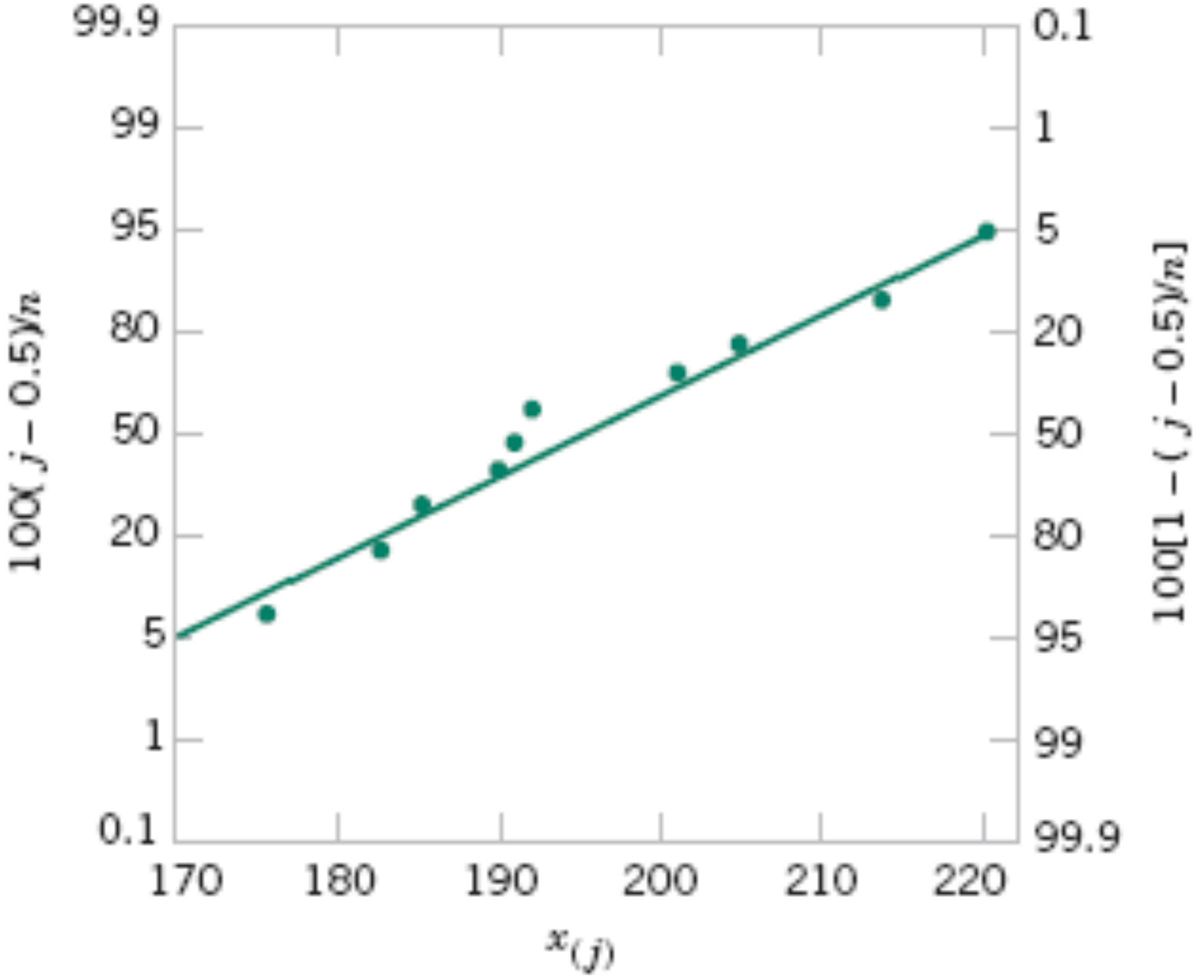
Example 6-7 (continued)



The pairs of values $x_{(j)}$ and $(j - 0.5)/10$ are now plotted on normal probability paper. This plot is shown in Fig. 6-19. Most normal probability paper plots $100(j - 0.5)/n$ on the left vertical scale and $100[1 - (j - 0.5)/n]$ on the right vertical scale, with the variable value plotted on the horizontal scale. A straight line, chosen subjectively, has been drawn through the plotted points. In drawing the straight line, you should be influenced more by the points near the middle of the plot than by the extreme points. A good rule of thumb is to draw the line approximately between the 25th and 75th percentile points. This is how the line in Fig. 6-19 was determined. In assessing the “closeness” of the points to the straight line, imagine a “fat pencil” lying along the line. If all the points are covered by this imaginary pencil, a normal distribution adequately describes the data. Since the points in Fig. 6-19 would pass the “fat pencil” test, we conclude that the normal distribution is an appropriate model.

6-6 Probability Plots

Figure 6-19 Normal probability plot for battery life.



6-6 Probability Plots

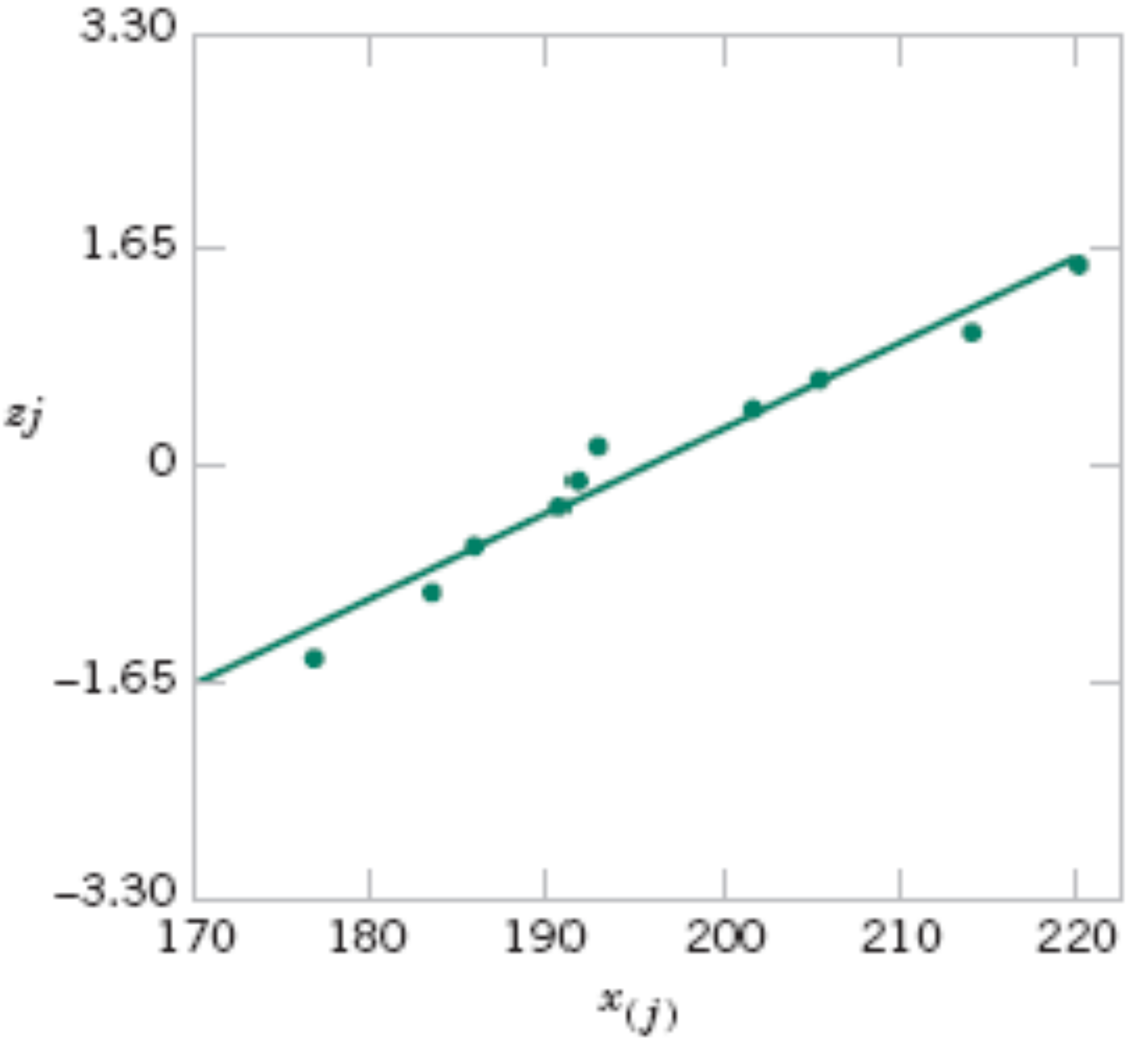


Figure 6-20 Normal probability plot obtained from standardized normal scores.

6-6 Probability Plots

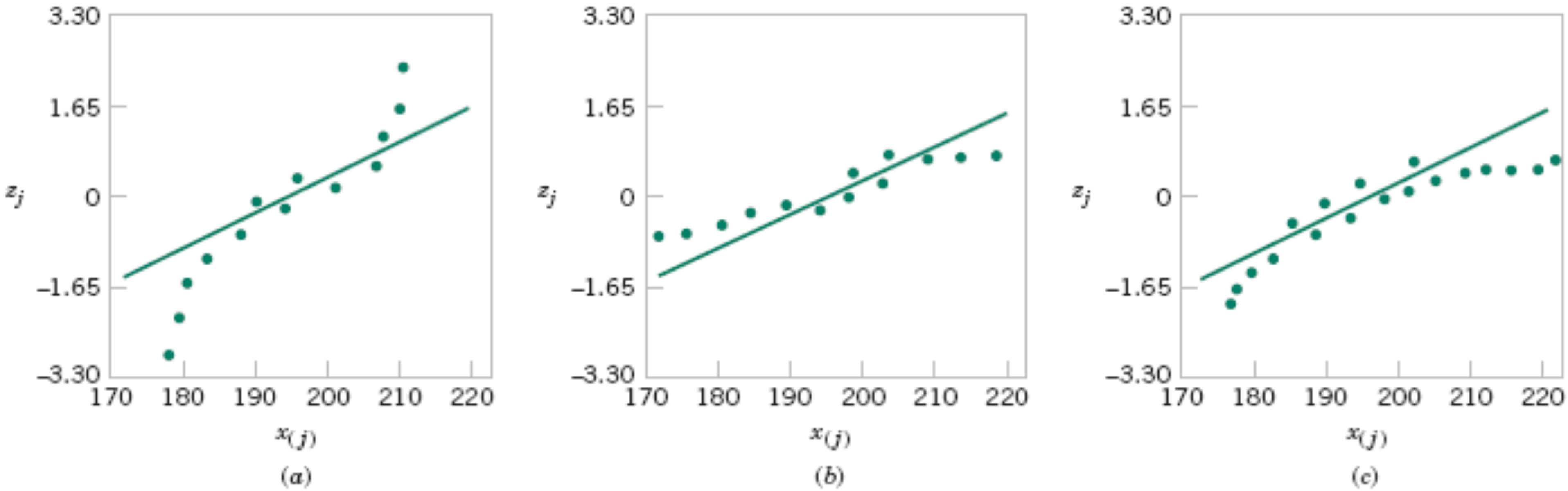


Figure 6-21 Normal probability plots indicating a nonnormal distribution. (a) Light-tailed distribution. (b) Heavy-tailed distribution. (c) A distribution with positive (or right) skew.

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 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

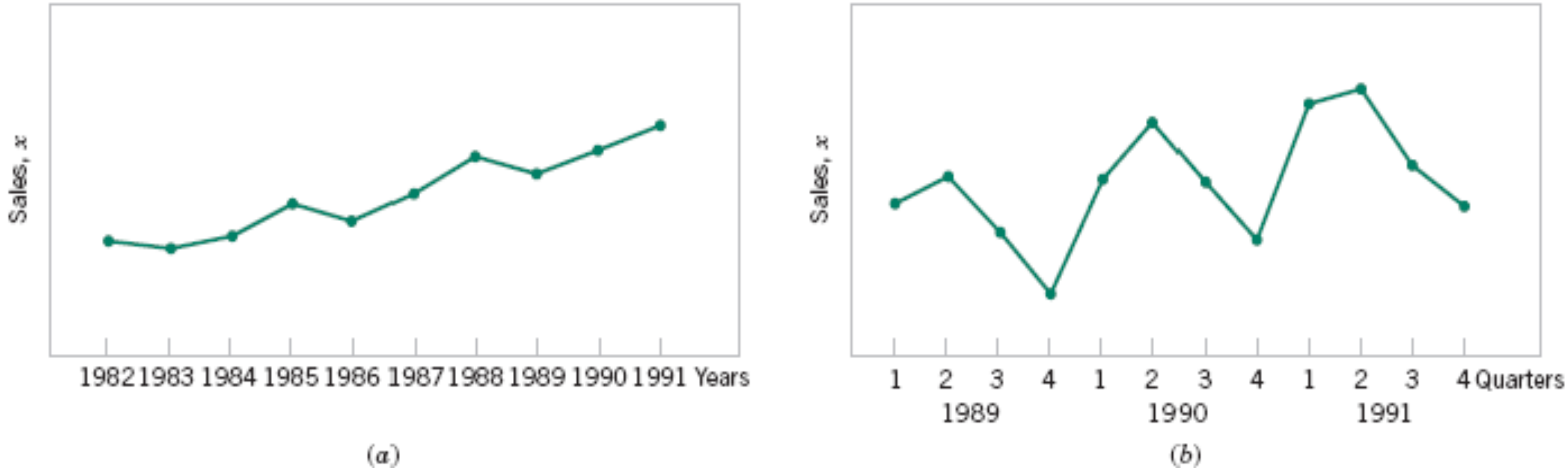
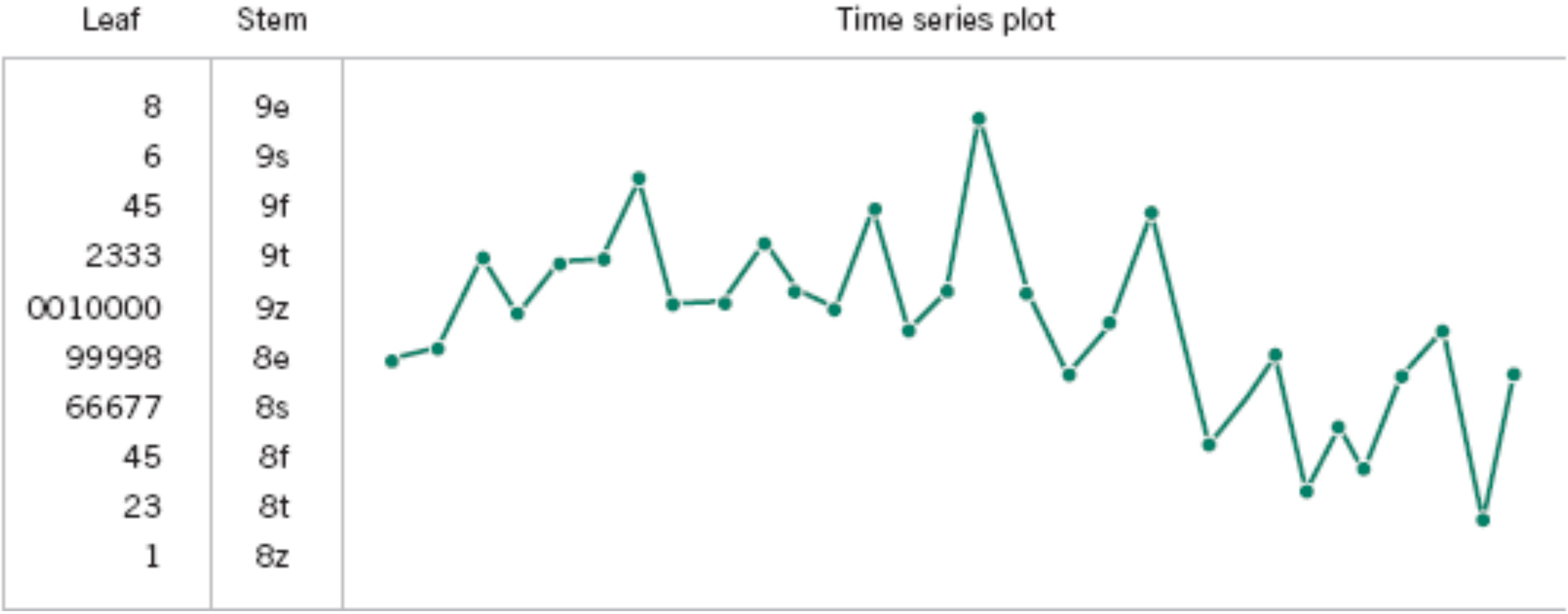


Figure 6-16 Company sales by year (a) and by quarter (b).

6-5 Time Sequence Plots



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.

Some Useful Comments

- Locations: mean and median
- Spreads: standard deviation (s.d.) and IQR
  - Mean and s.d. are **sensitive** to extreme values (**outliers**)
  - Median and IQR are **resistant** to extreme values and are better for skewed distributions
  - Use mean and s.d. for symmetrical distributions without outliers
- Software has defaults, which may not be the best choice
  - How many stems or bins?
  - The reference line in a normal probability plot.