MA026IU
PROBABILITY, STATISTIC AND RANDOM PROCESS

Part 2 A
Descriptive Statistics

Numerical Summaries of Data

- Data summaries and displays are essential to good statistical thinking
- It is useful to describe data features **numerically**
- Characterizing the **location** or **central tendency** in the data is an example of a numerical summary
- Data are often a **sample** of observations that have been selected from some larger **population** of observations
  - This type of population is called a **conceptual** or **hypothetical** population because it does not

Sample Mean

The location or central tendency in the data can be characterized by the **arithmetic average** or the **sample mean.**

Sample Mean

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

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n} \qquad (6.1)$$

For a finite **population** with N equally likely values, the probability mass function is $f(x_i) = 1/N$ and the mean is

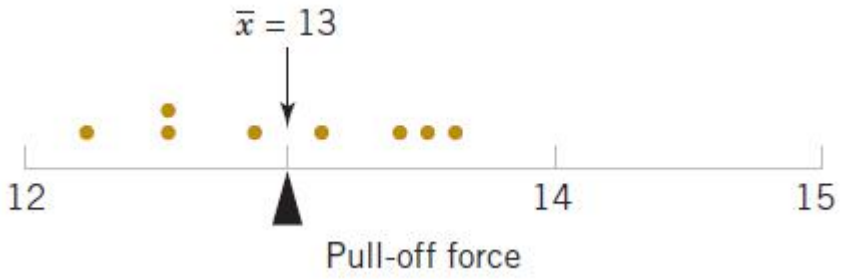
$$\mu = \sum_{i-1}^{N} x_i f(x_i) = \frac{\sum_{i=1}^{N} x_i}{N}$$

Example 1 | Sample Mean

Consider 8 observations (xi) of pull-off force from engine connectors as shown in the table.

$$\bar{x} = \text{average} = \frac{\sum_{i=1}^{8} x_i}{8} = \frac{12.6 + 12.9 + \ldots + 13.1}{8}$$

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



i	x_i
1	12.6
2	12.9
3	13.4
4	12.3
5	13.6
6	13.5
7	12.6
8	13.1
	13.00
= AVERAGE($B2:$B9)	

Figure 1 The sample mean is the balance point.

Sample Variance and Standard Deviation

The variability or scatter in the data may be described by the **sample variance** or the **sample standard deviation.**

The units of measurement for the sample variance are the square of the original units of the variable, while the standard deviation measures variability in the original units.

Sample Variance and Standard Deviation

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

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} \qquad (6.3)$$

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

Example 2 | Sample Variance

The table displays the quantities needed for calculating the sample variance and sample standard deviation.

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

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

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
Total	104.0	0.0	1.60

Computation of s^2

The prior calculation is definitional and tedious. A shortcut is derived here and involves just 2 sums.

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} = \frac{\sum_{i=1}^{n} \left(x_i^2 + \bar{x}^2 - 2\bar{x}x_i\right)}{n-1} = \frac{\sum_{i=1}^{n} x_i^2 + n\bar{x}^2 - 2\bar{x}\sum_{i=1}^{n} x_i}{n-1}$$

and because $\bar{x} = (1/n)\sum_{i=1}^{n} x_i$, this last equation reduces to

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

Example 3 | Shortcut Calculation for

For Example 2, we calculate the sample variance and standard deviation using the shortcut method.

$$s^2 = \frac{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}{n-1} = \frac{1353.6 - \frac{(104)^2}{8}}{7}$$

$$= \frac{1.60}{7} = 0.2286 \text{ (pounds)}^2$$

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

The meaning of $n-1$ in the denominator

- The population variance is calculated with N, the population size. Why isn't the sample variance calculated with n, the sample size?
- The true variance is based on data deviations from the true mean, μ.
- The sample calculation is based on the data deviations from $\bar{x}$, not μ.
- $\bar{x}$ is an estimator of μ; close but not the same.
- So the $n-1$ divisor is used to compensate for the error in the mean estimation.

Degrees of Freedom

- When the sample variance is calculated with the quantity $n-1$ in the denominator, the quantity $n-1$ is called the **degrees of freedom**
- Origin of term:
  - There are n deviations from the $\bar{x}$ in the sample
  - The sum of the deviations is zero
  - $n-1$ of the observations can be freely determined but the n^{th} observation is fixed to maintain the zero sum

Sample Range

In addition to the sample variance and sample standard deviation, the sample range is a useful measure of variability.

Sample Range

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

$$r = \max(x_i) - \min(x_i) \qquad (6.6)$$

For Example 3 (pull-off force data), the sample range is $r = 13.6 - 12.3 = 1.3$.

Stem-and-Leaf Diagrams

Steps to Construct 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.

Example 4a | Alloy Strength

- Consider the data in the table. We select as stem values the numbers 7

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

Example 4b | Alloy Strength

- The resulting stem-and-leaf diagram is shown.
- Inspection of the diagram reveals that most of the comprehensive strengths lie between 110 and 200 psi and that a central value is somewhere between 150 and 160 psi.
- The strengths are distributed approximately symmetrically about the central value

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).

FIGURE 6.4

Stem-and-leaf diagram for the compressive strength data in Table 6.2.

Frequency Distributions and Histograms

- A **frequency distribution** is a more compact summary of data than a stem – and – leaf diagram
- To construct, we must divide the range of the data into intervals, which are usually called **class intervals**, **cells**, or **bins**
- Choosing **number of bins** approximately equal to the square root of the number of observations often works well in practice

Frequency Distribution Table

TABLE 6.4 Frequency Distribution for the Compressive Strength Data in Table 6.2

Class	$70 \le x < 90$	$90 \le x < 110$	$110 \le x < 130$	$130 \le x < 150$	$150 \le x < 170$
Frequency	2	3	6	14	22
Relative frequency	0.0250	0.0375	0.0750	0.1750	0.2750
Cumulative relative frequency	0.0250	0.0625	0.1375	0.3125	0.5875
	$170 \le x < 190$	$190 \le x < 210$	$210 \le x < 230$	$230 \le x < 250$	
Frequency	17	10	4	2	
Relative frequency	0.2125	0.1250	0.0500	0.0250	
Cumulative relative frequency	0.8000	0.9250	0.9750	1.0000	

Histograms

- A **histogram** is a visual display of the frequency distribution
- Provides a visual impression of the shape and distribution of the measurements and information about the central tendency and scatter or dispersion in the data
- **Unequal bin widths** will be employed

$$\text{Rectangle height} = \frac{\text{bin frequency}}{\text{bin width}}$$

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.

Histograms

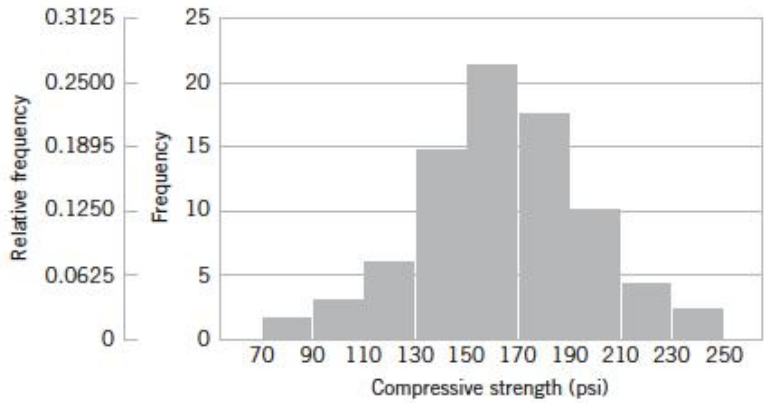


FIGURE 6.7

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

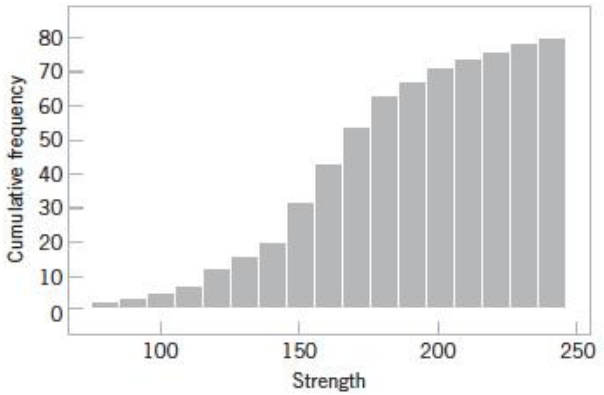


FIGURE 6.10

A cumulative distribution plot of the compressive strength data.

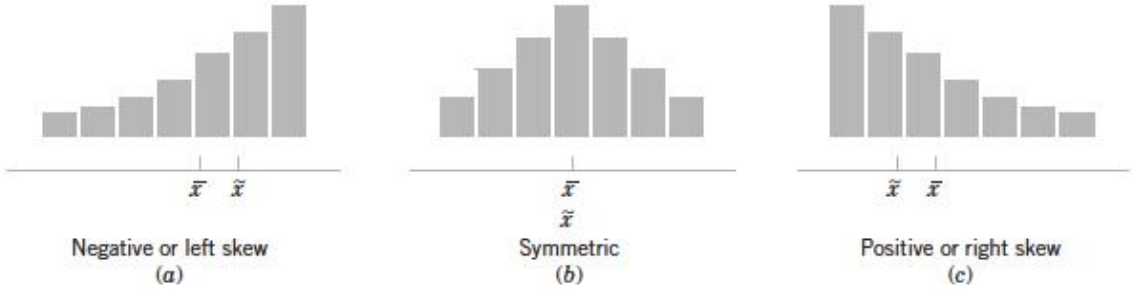


FIGURE 6.11

Histograms for symmetric and skewed distributions.

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 unusual observations or outliers
- Sometimes called *box – and – whisker* plots
- Displays three quartiles
- A line, or **whis**

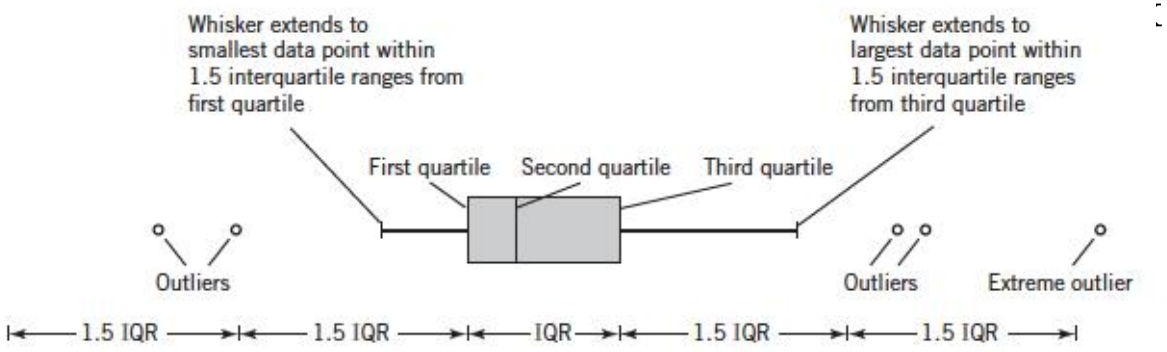


FIGURE 6.13

Description of a box plot.

Box Plots

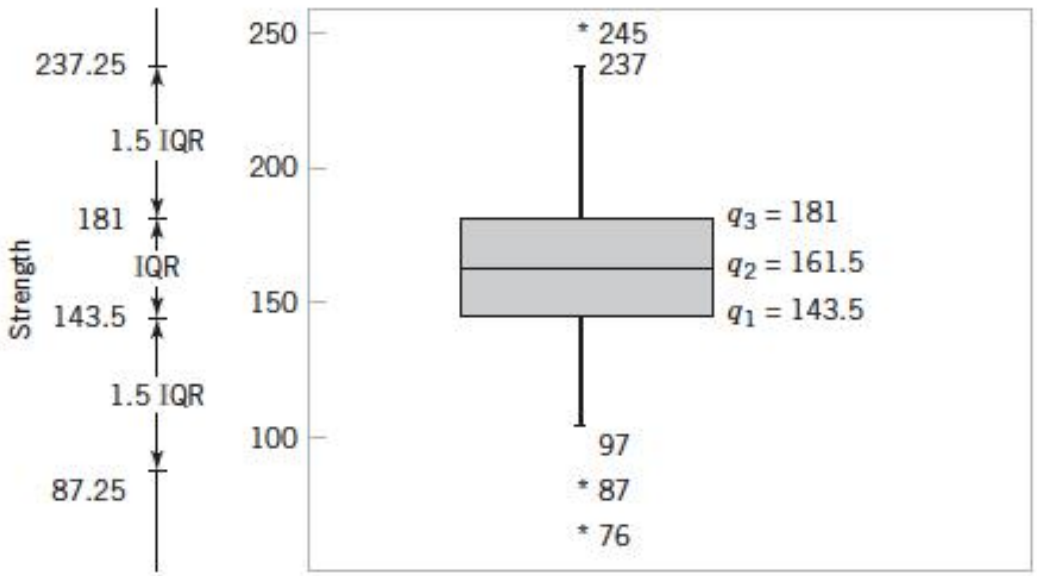


FIGURE 6.14

Box plot for compressive strength data in Table 6.2.

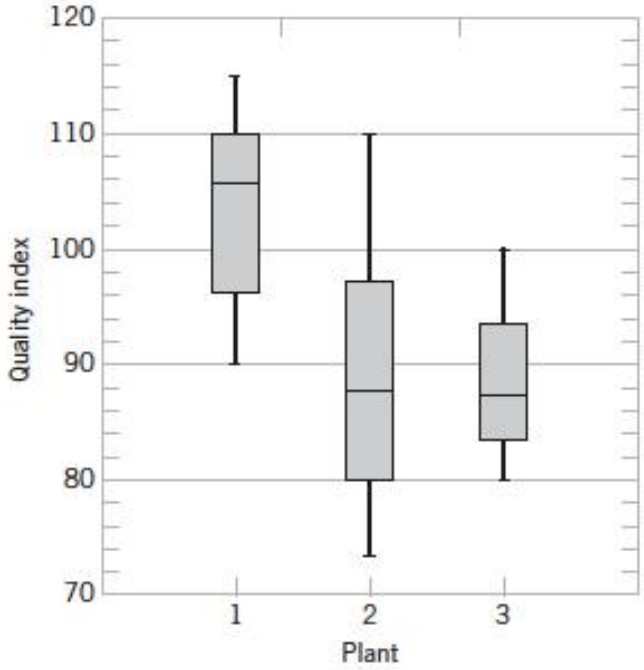


FIGURE 6.15

Comparative box plots of a quality index at three plants.

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 den [illegible] nd the horizont

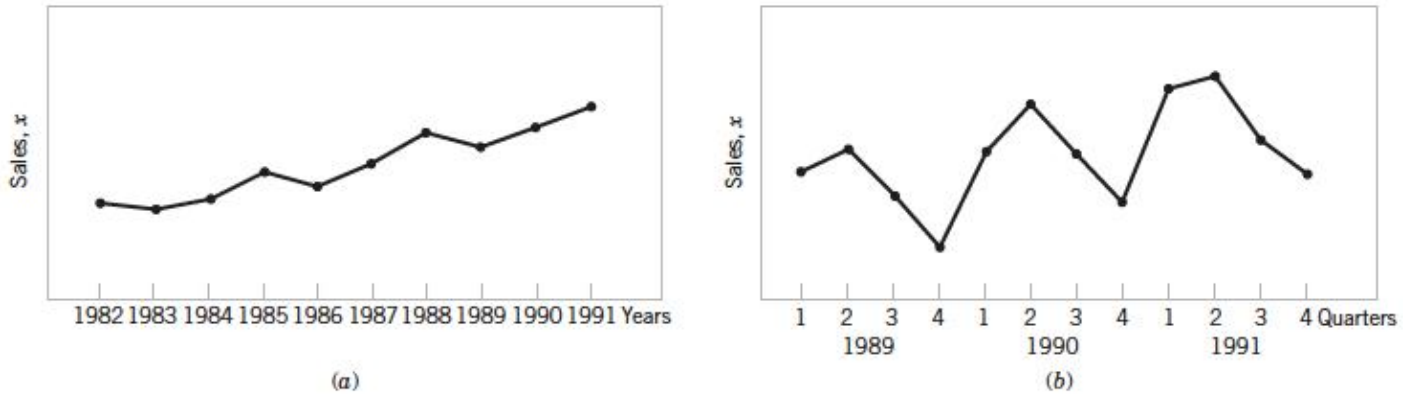


FIGURE 6.16

Company sales by year (*a*). By quarter (*b*).

Time Sequence Plots

- Combination of stem – and – leaf plot with a time series :

Leaf	Stem
5	24
7	23
189	22
8	21
7108	20
960934	19
0361410	18
8544162106	17
3073050879	16
471340886808	15
29583169	14
413535	13
103	12
580	11
15	10
7	9
7	8
6	7

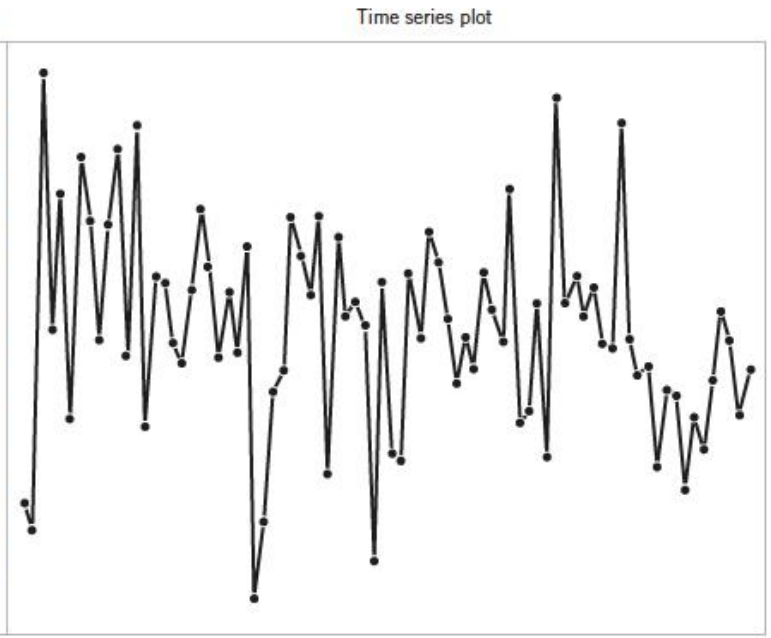


FIGURE 6.17

A digidot plot of the compressive strength data in Table 6.2.

Scatter Diagrams

- **Multivariate:** each observation consists of measurements of several variables
- The **scatter diagram** is a useful way to graphically display the potential relationship between quality and one of the other qualities
- When two or more variables exist, the **matrix of scatter diagrams** may be useful in looking at all of the pairwise relationships between the variables in the sample
- The **sample correlation coefficient** is a quantitative measure of the strength of the linear relationship between two random variables x and y

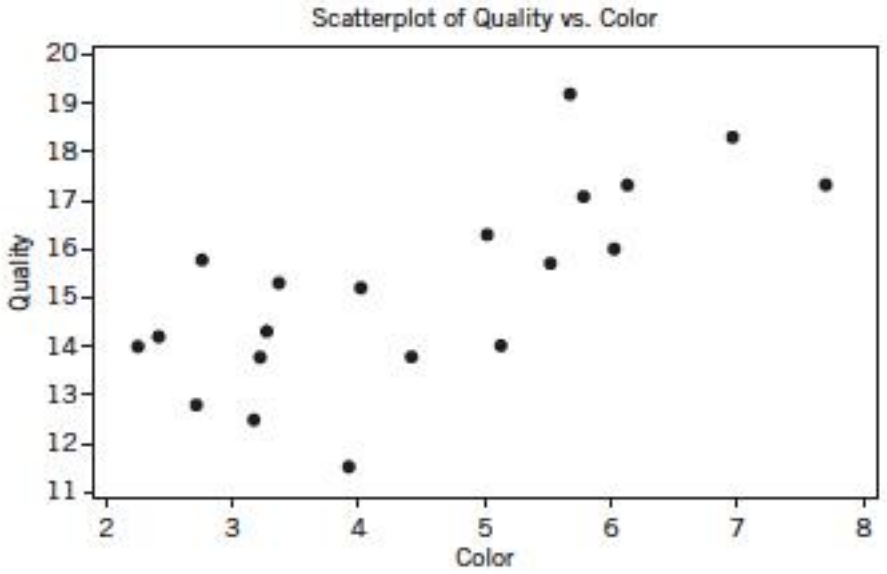


FIGURE 6.19

Scatter diagram of wine quality and color from Table 6.5.

$$r_{xy} = \frac{\sum_{i=1}^{n} y_i(x_i - \bar{x})}{\left[\sum_{i=1}^{n}(y_i - \bar{y})^2 \sum_{i=1}^{n}(x_i - \bar{x})^2\right]^{1/2}}$$

Probability Plots

- A **probability plot** is a graphical method for determining whether sample data conform to a hypothesized distribution based on a subjective visual examination of the data
- To construct a probability plot:
  - Rank the data observations in the sample from smallest to largest: $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$.
  - The observed value $x_{(j)}$ is plotted against the observed cumulative frequency $(j - 0.5)/n$.
  - The paired numbers are plotted on the probability paper of the proposed distribution.
- If the plotted points deviate a straight line, then the hypothesized distribution adequately describes the data.

Example 7 | Battery Life

- The effective service life (X_j in minutes) of batteries used in a laptop are given in the table.
- We hypothesize that battery life is adequately modeled by a normal distribution.
- To test this hypothesis, first arrange the observations in ascending order and calculate their cumulative frequencies and plot them.

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

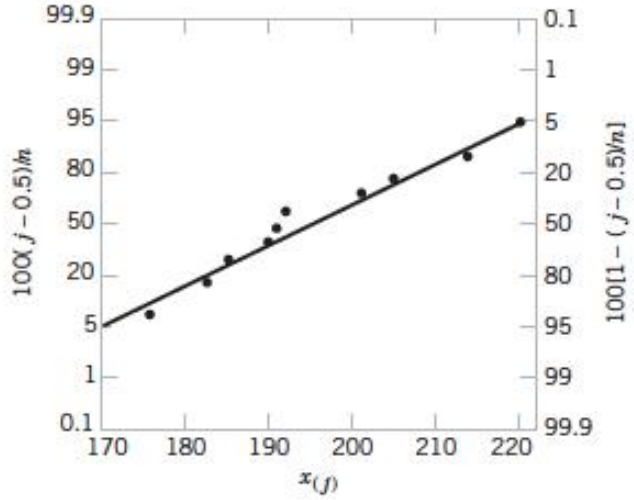


FIGURE 6.22
Normal probability plot for battery life.

Normal Probability Plot

- Can be constructed on ordinary axes by plotting the standardized normal scores z_j against $x(j)$, where the standardized normal scores satisfy

$$\frac{j - 0.5}{n} = P(Z \le z_j) = \Phi(z_j)$$

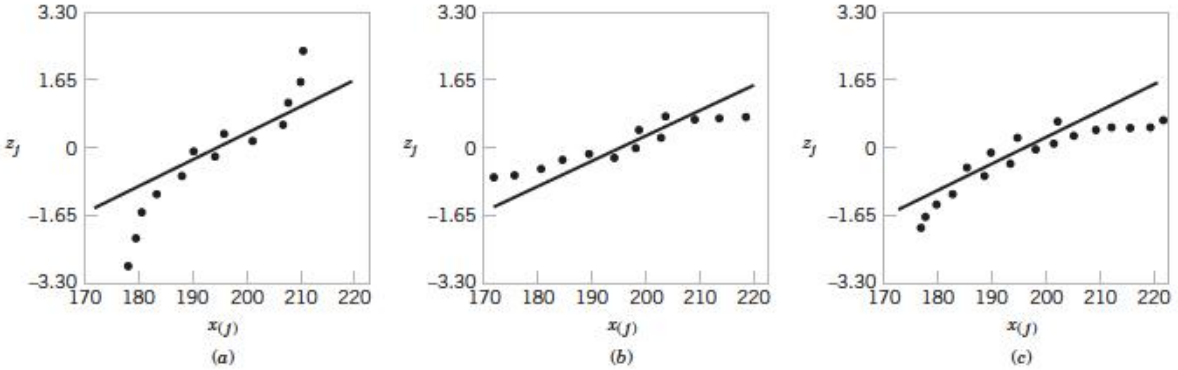


FIGURE 6.24
Normal probability plots indicating a nonnormal distribution. (a) Light-tailed distribution. (b) Heavy-tailed distribution. (c) A distribution with positive (or right) skew.

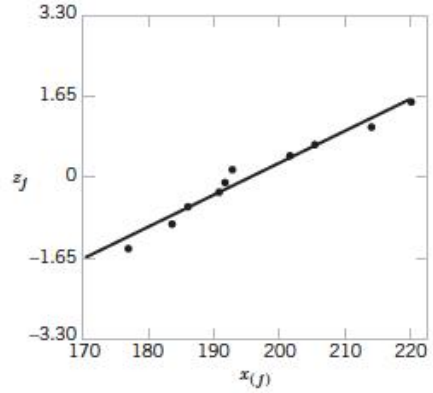


FIGURE 6.23
Normal probability plot obtained from standardized normal scores.