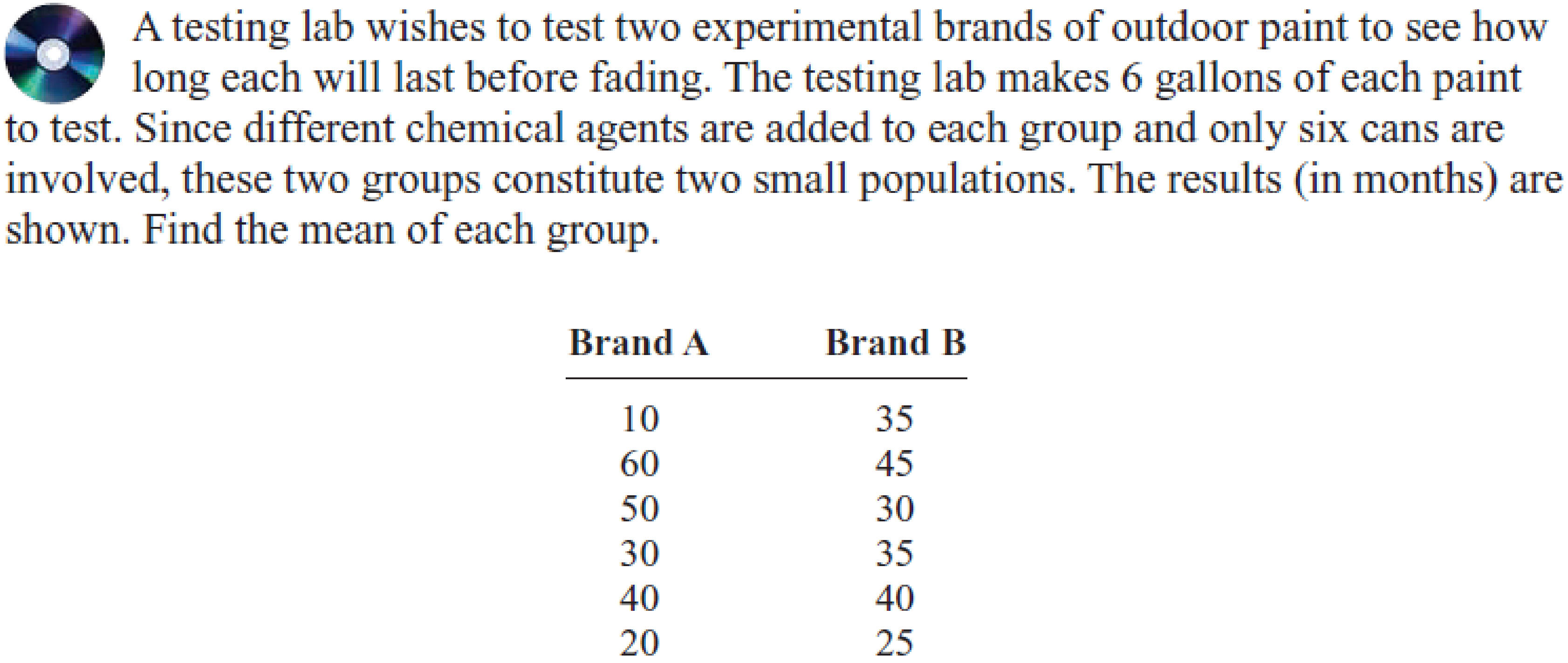
PROBABILITY & STATISTICS

BS 1402

Contents

- Measures of Variation
  - Range
  - Variance
  - Standard Deviation
- Measures of Position
- Exploratory Data Analysis (EDA)

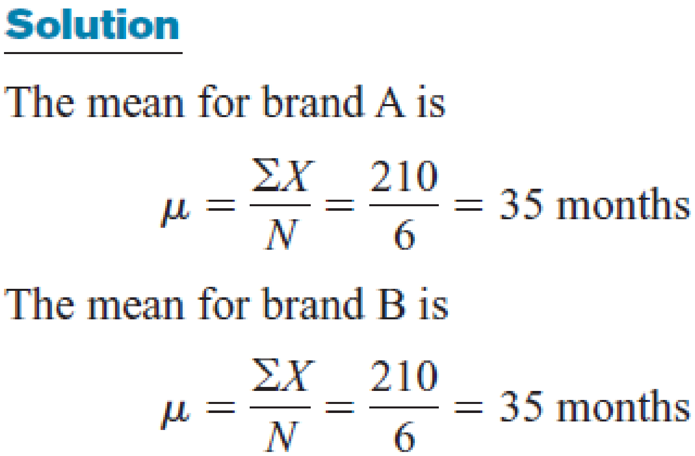
Example 1: Comparison of Outdoor Paint



A testing lab wishes to test two experimental brands of outdoor paint to see how long each will last before fading. The testing lab makes 6 gallons of each paint to test. Since different chemical agents are added to each group and only six cans are involved, these two groups constitute two small populations. The results (in months) are shown. Find the mean of each group.

Brand A	Brand B
10	35
60	45
50	30
30	35
40	40
20	25

Example 1 (cont.):



Solution

The mean for brand A is

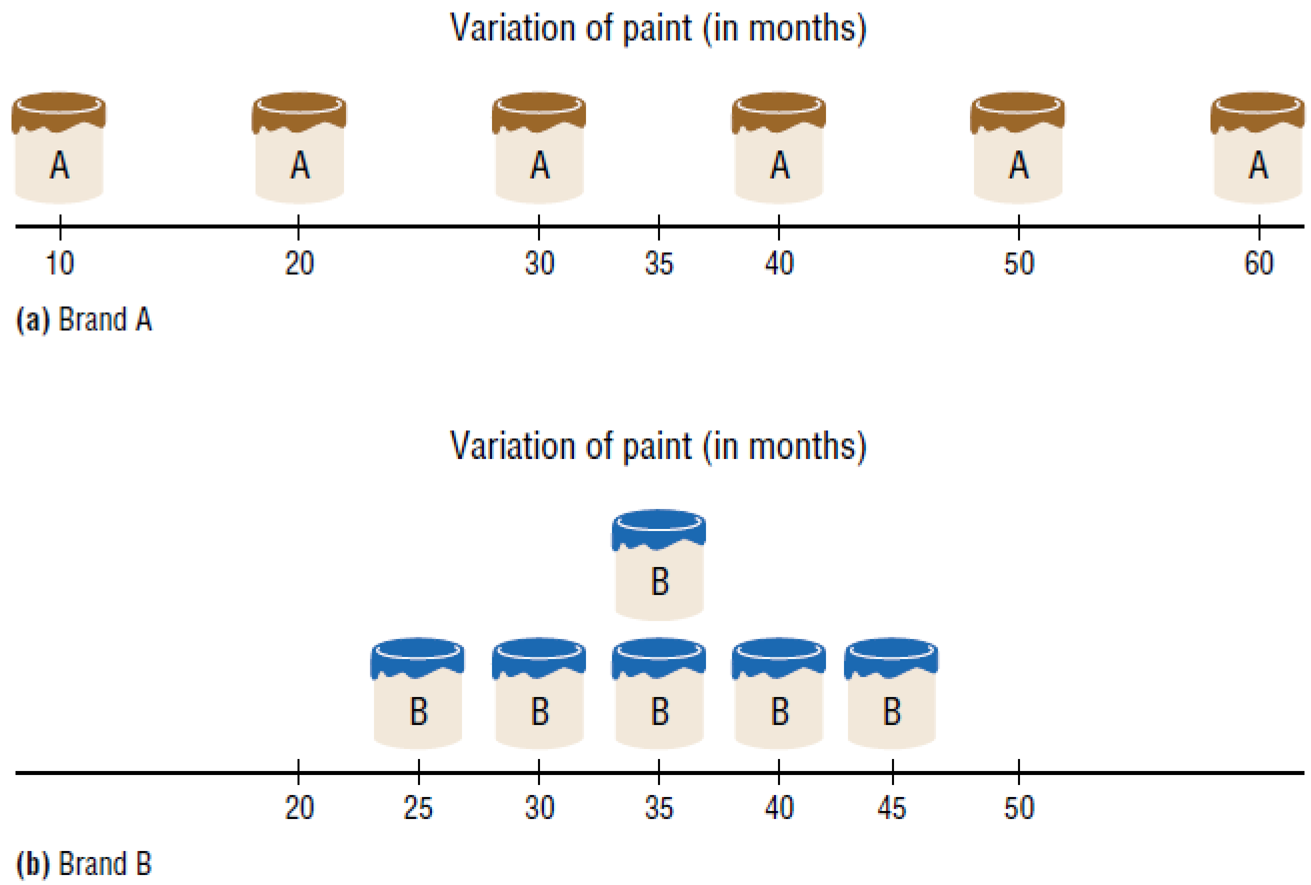
$$\mu = \frac{\Sigma X}{N} = \frac{210}{6} = 35 \text{ months}$$

The mean for brand B is

$$\mu = \frac{\Sigma X}{N} = \frac{210}{6} = 35 \text{ months}$$

Since the means are equal in Example above, you might conclude that both brands of paint last equally well. However, when the data sets are examined graphically, a somewhat different conclusion might be drawn. See Figure 1. As Figure 1 shows, even though the means are the same for both brands, the spread, or variation, is quite different. Figure 1 shows that brand B performs more consistently; it is less variable.

Example 1 (cont.):



Variation of paint (in months)

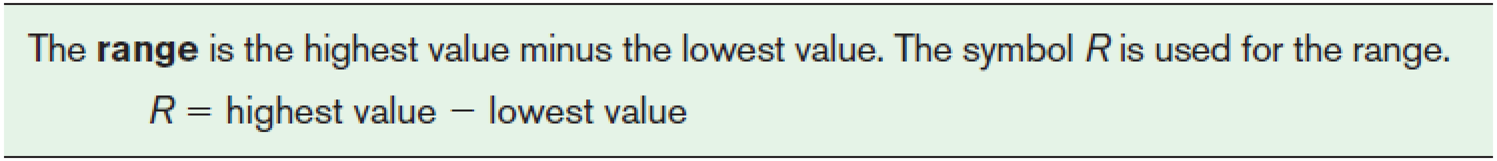
(a) Brand A

Variation of paint (in months)

(b) Brand B

For the spread or variability of a data set, three measures are commonly used: *range*, *variance*, and *standard deviation.*

Measures of Variance– Range

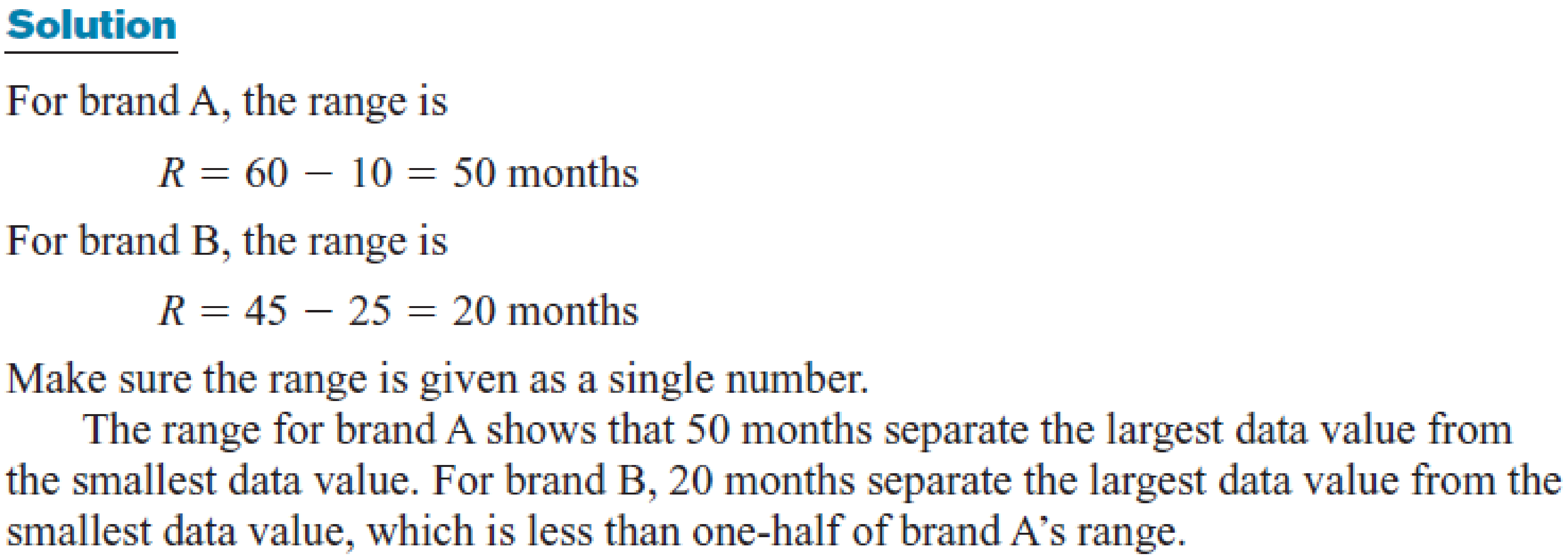


The **range** is the highest value minus the lowest value. The symbol *R* is used for the range.

$$R = \text{highest value} - \text{lowest value}$$

Example 2:

Find the ranges for the paints in Example 1.



Solution

For brand A, the range is

$$R = 60 - 10 = 50 \text{ months}$$

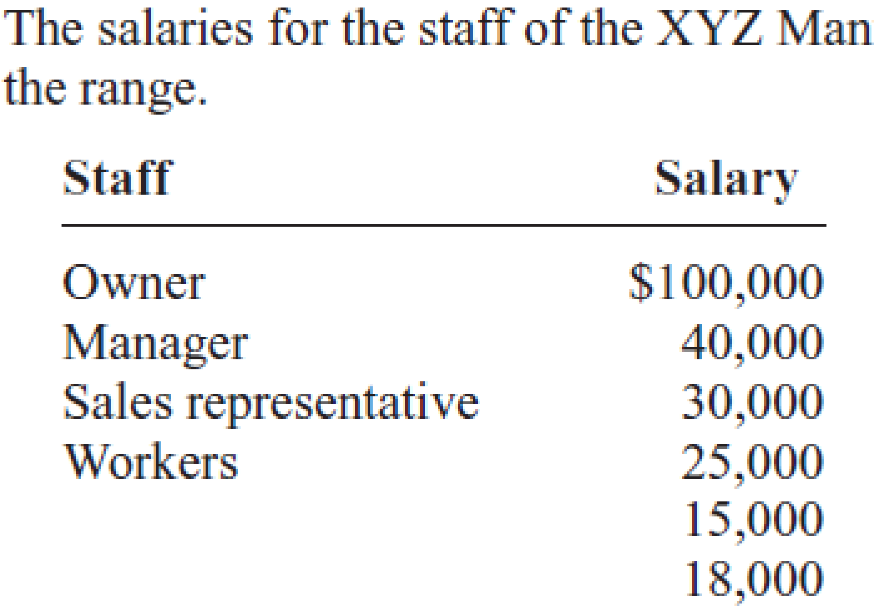
For brand B, the range is

$$R = 45 - 25 = 20 \text{ months}$$

Make sure the range is given as a single number.

The range for brand A shows that 50 months separate the largest data value from the smallest data value. For brand B, 20 months separate the largest data value from the smallest data value, which is less than one-half of brand A's range.

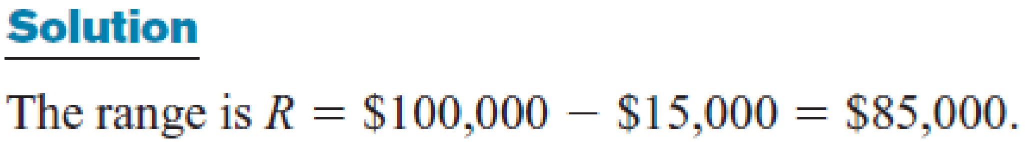
Example 3:



The salaries for the staff of the XYZ Manufacturing Co. are shown here. Find the range.

Staff	Salary
Owner	$100,000
Manager	40,000
Sales representative	30,000
Workers	25,000
	15,000
	18,000

Example 3: (cont.)



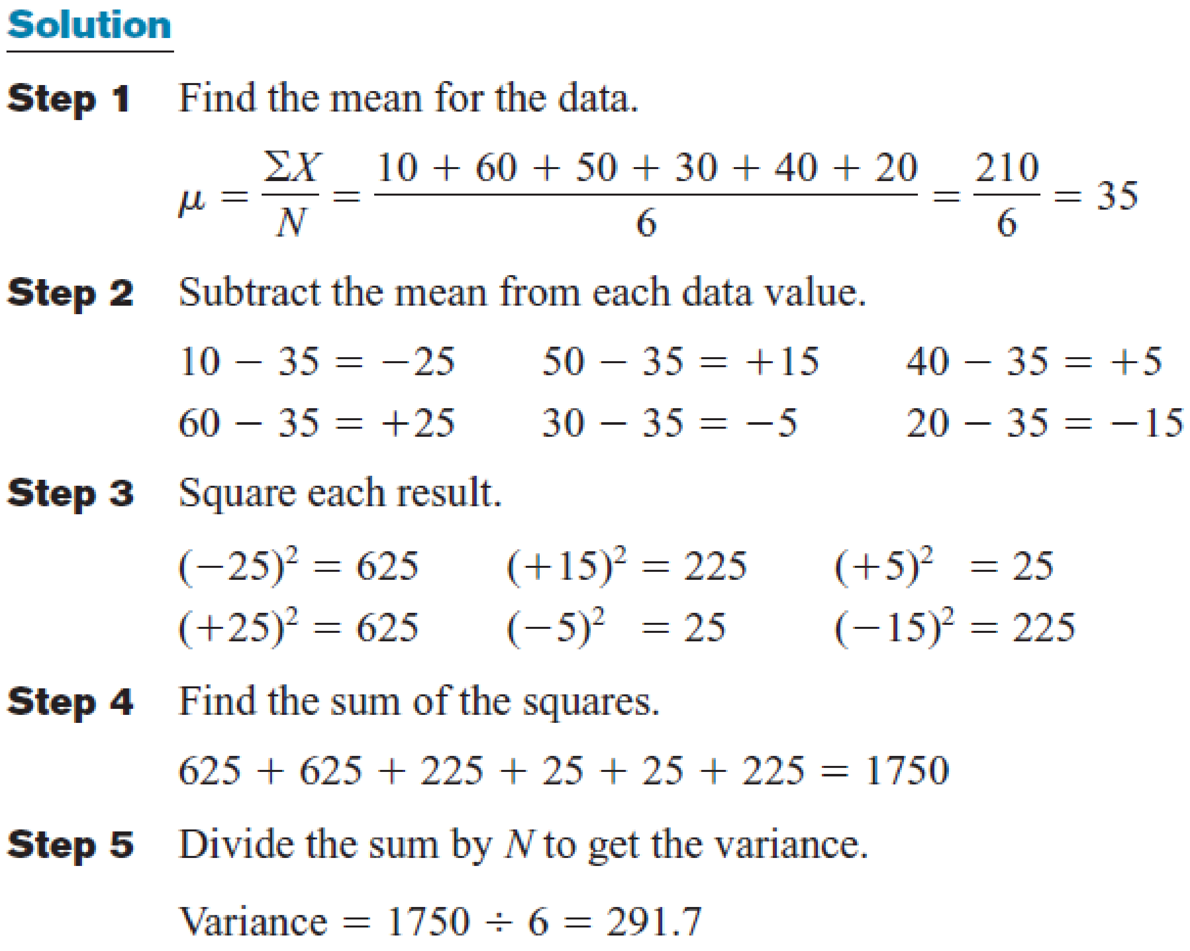
Solution

The range is $R = \$100,000 - \$15,000 = \$85,000$.

Since the owner's salary is included in the data for Example 3, the range is a large number. To have a more meaningful statistic to measure the variability, statisticians use measures called the *variance* and *standard deviation.*

Example 4:

Find the variance and standard deviation for the data set for brand A given in Example 1.



Solution

Step 1 Find the mean for the data.

$$\mu = \frac{\Sigma X}{N} = \frac{10 + 60 + 50 + 30 + 40 + 20}{6} = \frac{210}{6} = 35$$

Step 2 Subtract the mean from each data value.

$$10 - 35 = -25 \qquad 50 - 35 = +15 \qquad 40 - 35 = +5$$
$$60 - 35 = +25 \qquad 30 - 35 = -5 \qquad 20 - 35 = -15$$

Step 3 Square each result.

$$(-25)^2 = 625 \qquad (+15)^2 = 225 \qquad (+5)^2 = 25$$
$$(+25)^2 = 625 \qquad (-5)^2 = 25 \qquad (-15)^2 = 225$$

Step 4 Find the sum of the squares.

$$625 + 625 + 225 + 25 + 25 + 225 = 1750$$

Step 5 Divide the sum by N to get the variance.

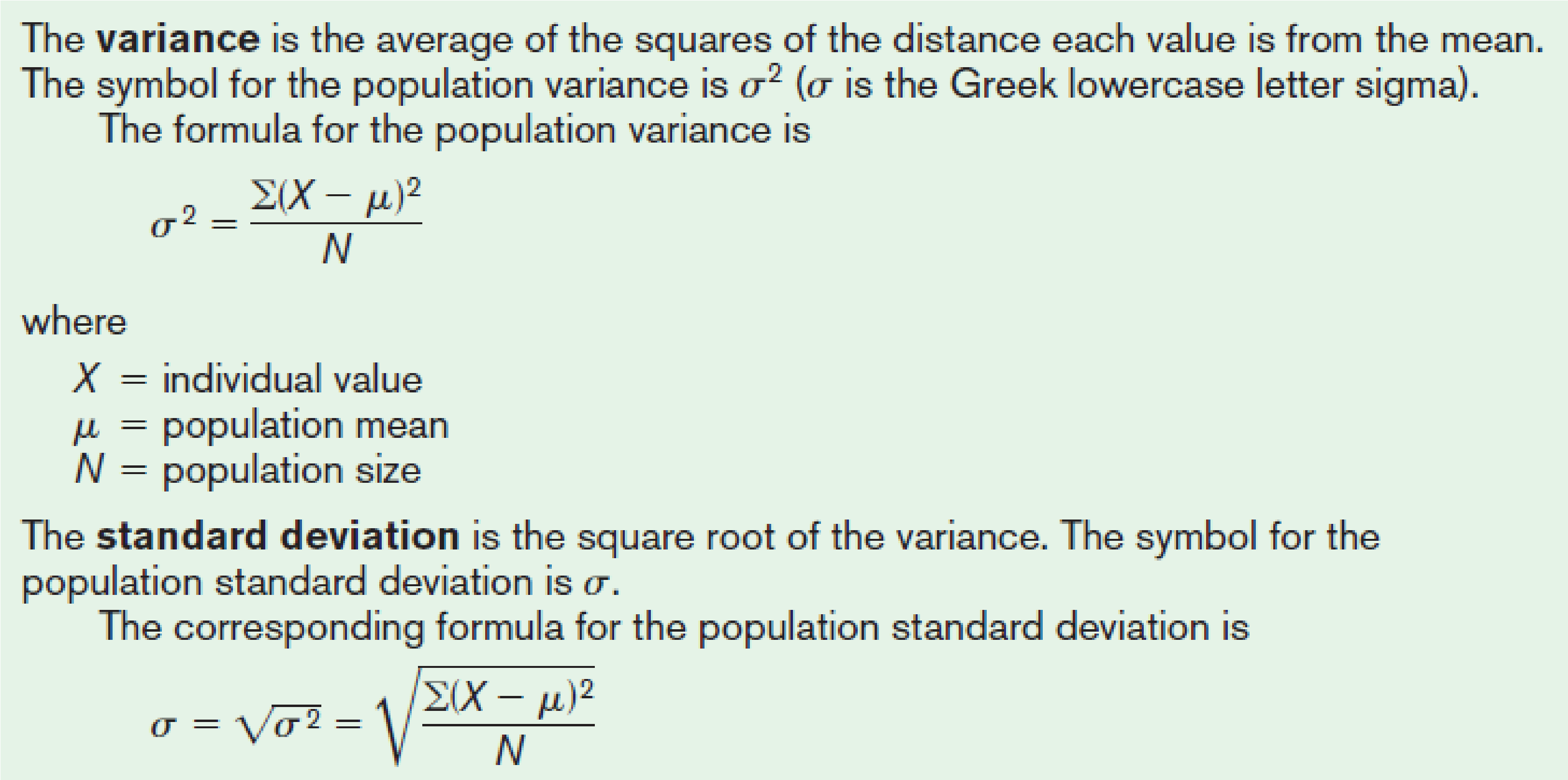
$$\text{Variance} = 1750 \div 6 = 291.7$$

Step 6 Take the square root of the variance to get the standard deviation. Hence, the standard deviation equals $\sqrt{291.7}$, or 17.1. It is helpful to make a table.

A	B	C
Values X	$X - \mu$	$(X - \mu)^2$
10	−25	625
60	+25	625
50	+15	225
30	−5	25
40	+5	25
20	−15	225
		1750

Column A contains the raw data X. Column B contains the differences $X - \mu$ obtained in step 2. Column C contains the squares of the differences obtained in step 3.

Measures of Variation– Variance and Standard Deviation



The **variance** is the average of the squares of the distance each value is from the mean. The symbol for the population variance is σ^2 (σ is the Greek lowercase letter sigma).

The formula for the population variance is

$$\sigma^2 = \frac{\Sigma(X - \mu)^2}{N}$$

where

X = individual value
μ = population mean
N = population size

The **standard deviation** is the square root of the variance. The symbol for the population standard deviation is σ.

The corresponding formula for the population standard deviation is

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\Sigma(X - \mu)^2}{N}}$$

The preceding computational procedure of Example 4 reveals several things. First, the square root of the variance gives the standard deviation; and vice versa, squaring the standard deviation gives the variance. Second, the variance is actually the average of the square of the distance that each value is from the mean. Therefore, if the values are near the mean, the variance will be small. In contrast, if the values are far from the mean, the variance will be large.

One might wonder why the squared distances are used instead of the actual distances. One reason is that the sum of the distances will always be zero. To verify this result for a specific case, add the values in column B of the table in Example 4. When each value is squared, the negative signs are eliminated.

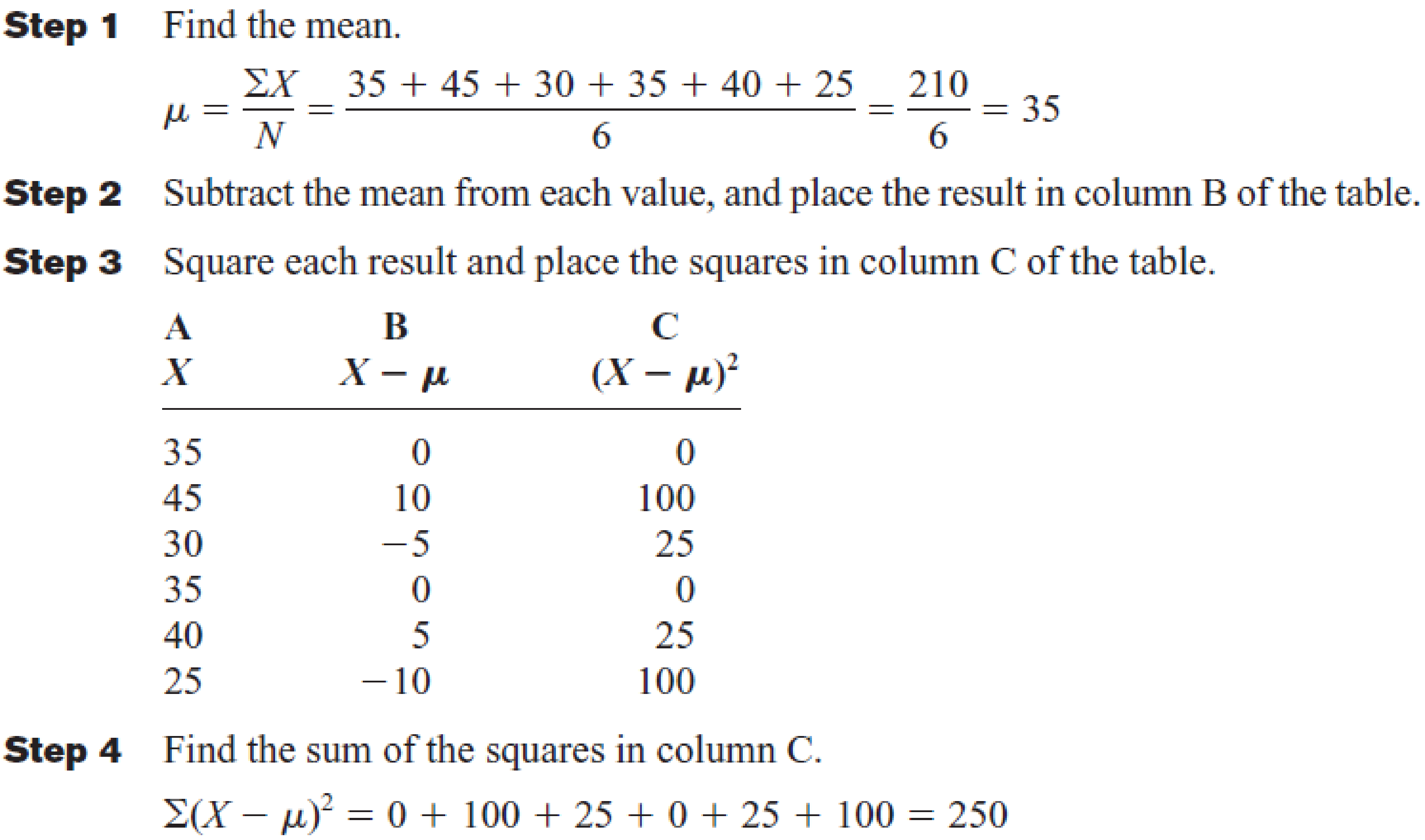
Finally, why is it necessary to take the square root? The reason is that since the distances were squared, the units of the resultant numbers are the squares of the units of the original raw data. Finding the square root of the variance puts the standard deviation in the same units as the raw data.

When you are finding the square root, always use its positive or principal value, since the variance and standard deviation of a data set can never be negative.

Example 5:

Find the variance and standard deviation for the data set for brand B given in Example 1.

Example 5 (cont.):



Step 1 Find the mean.

$$\mu = \frac{\Sigma X}{N} = \frac{35 + 45 + 30 + 35 + 40 + 25}{6} = \frac{210}{6} = 35$$

Step 2 Subtract the mean from each value, and place the result in column B of the table.

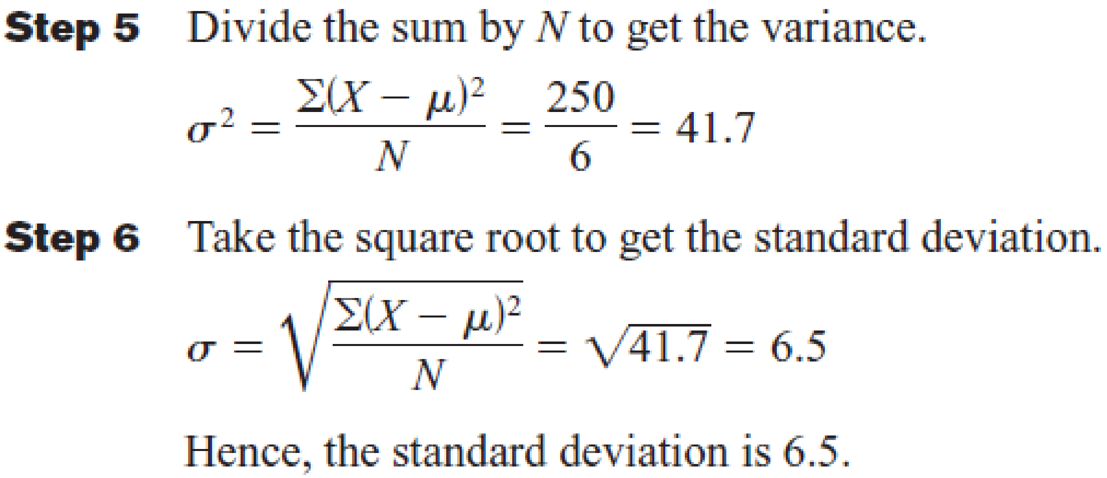
Step 3 Square each result and place the squares in column C of the table.

A	B	C
X	$X - \mu$	$(X - \mu)^2$
35	0	0
45	10	100
30	−5	25
35	0	0
40	5	25
25	−10	100

Step 4 Find the sum of the squares in column C.

$$\Sigma(X - \mu)^2 = 0 + 100 + 25 + 0 + 25 + 100 = 250$$

Example 5 (cont.):



Step 5 Divide the sum by N to get the variance.

$$\sigma^2 = \frac{\Sigma(X-\mu)^2}{N} = \frac{250}{6} = 41.7$$

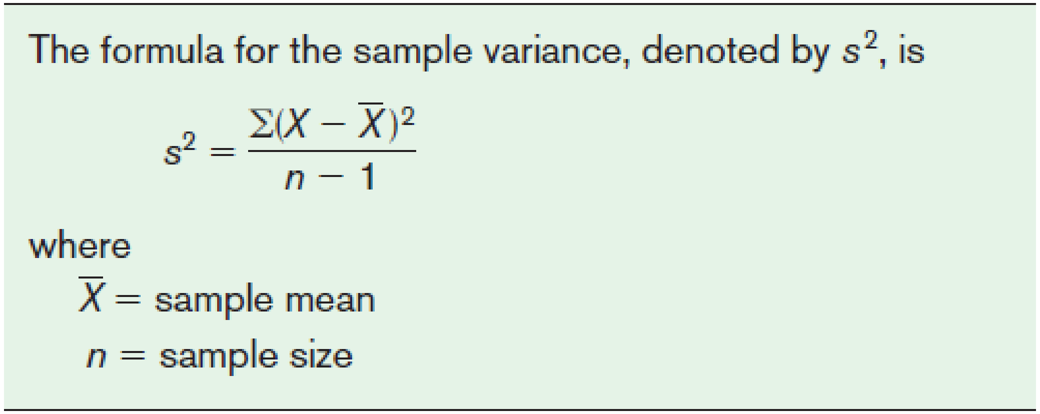
Step 6 Take the square root to get the standard deviation.

$$\sigma = \sqrt{\frac{\Sigma(X-\mu)^2}{N}} = \sqrt{41.7} = 6.5$$

Hence, the standard deviation is 6.5.

Since the standard deviation of brand A is 17.1 (see Example 4) and the standard deviation of brand B is 6.5, the data are more variable for brand A. *In summary, when the means are equal, the larger the variance or standard deviation is, the more variable the data are.*

Sample Variance and Standard Deviation



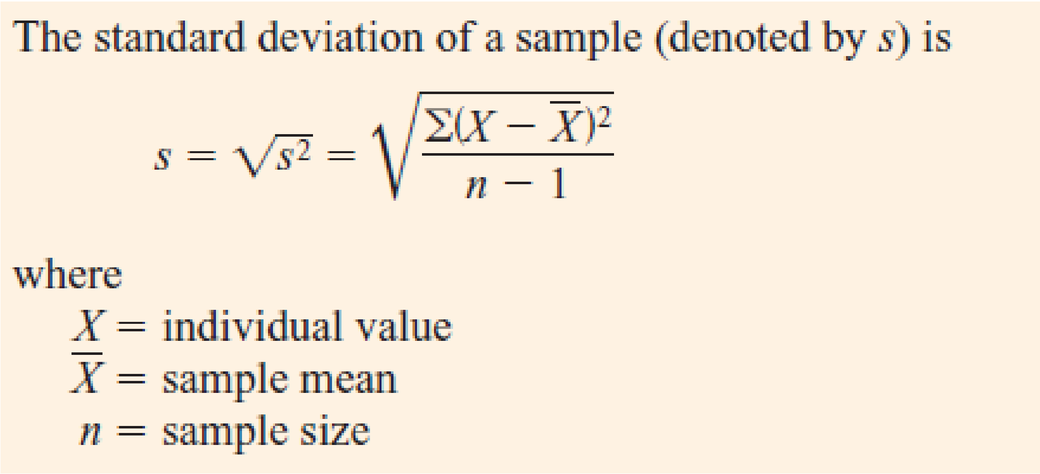
The formula for the sample variance, denoted by s^2, is

$$s^2 = \frac{\Sigma(X - \overline{X})^2}{n - 1}$$

where

$\overline{X}$ = sample mean

n = sample size



The standard deviation of a sample (denoted by s) is

$$s = \sqrt{s^2} = \sqrt{\frac{\Sigma(X - \overline{X})^2}{n - 1}}$$

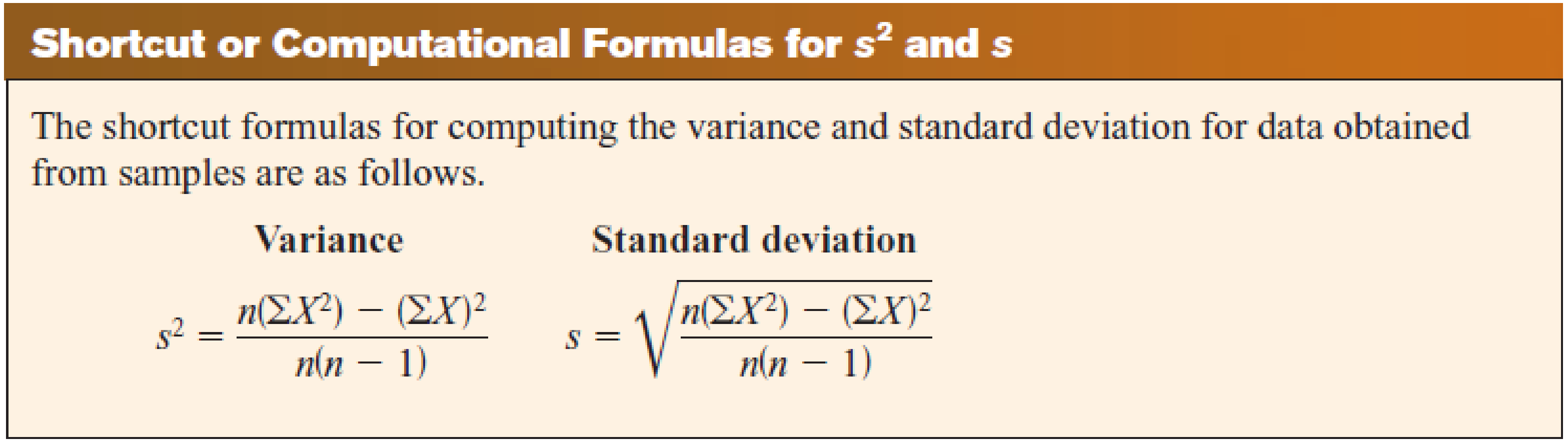
where

X = individual value

$\overline{X}$ = sample mean

n = sample size

Sample Variance and Standard Deviation



Shortcut or Computational Formulas for s^2 and s

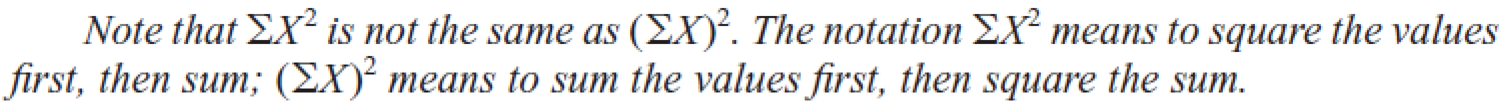
The shortcut formulas for computing the variance and standard deviation for data obtained from samples are as follows.

Variance

$$s^2 = \frac{n(\Sigma X^2) - (\Sigma X)^2}{n(n-1)}$$

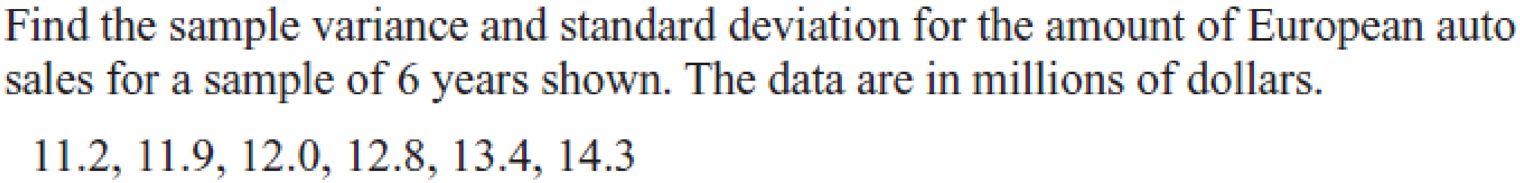
Standard deviation

$$s = \sqrt{\frac{n(\Sigma X^2) - (\Sigma X)^2}{n(n-1)}}$$



Note that ΣX^2 is not the same as $(\Sigma X)^2$. The notation ΣX^2 means to square the values first, then sum; $(\Sigma X)^2$ means to sum the values first, then square the sum.

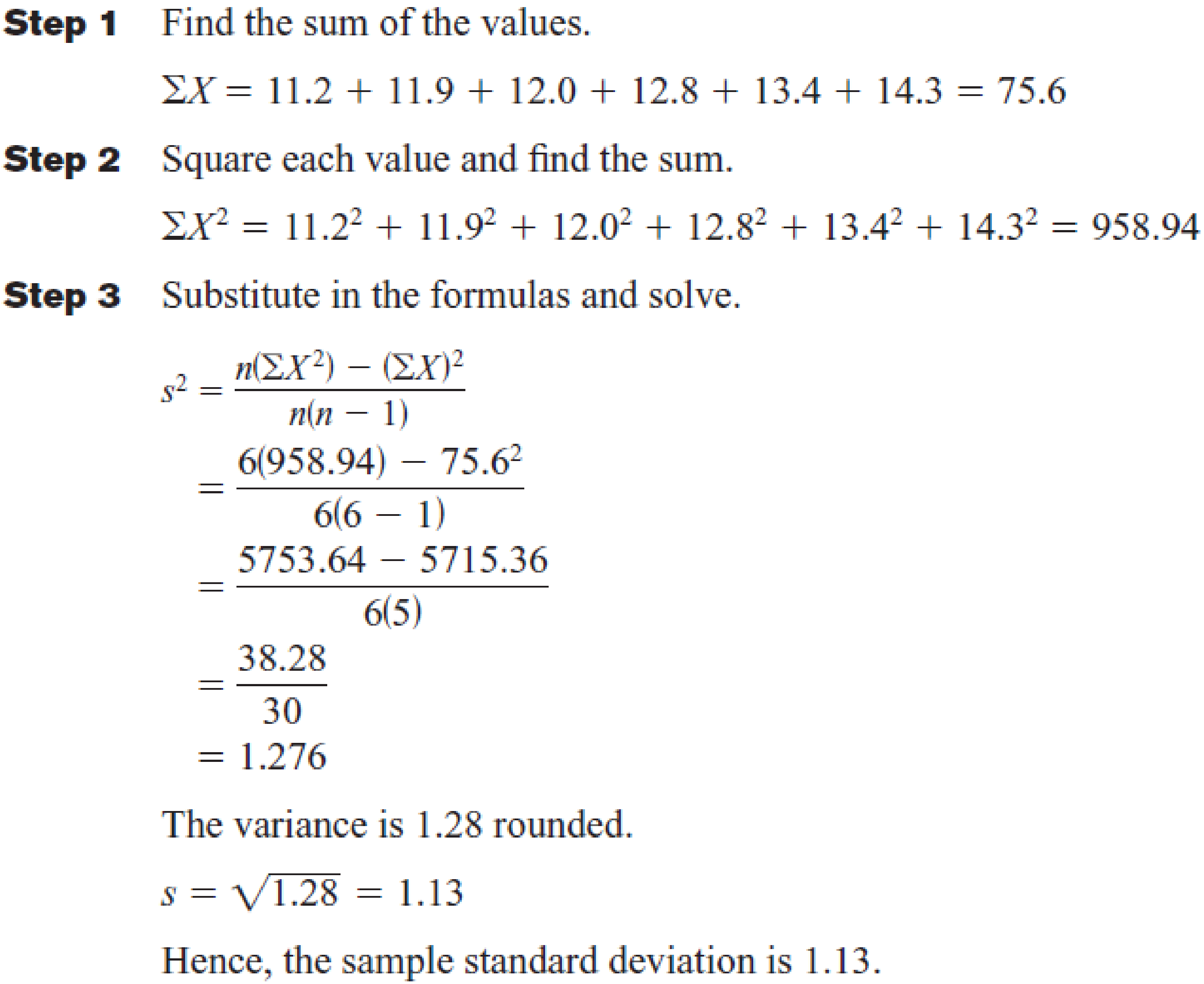
Example 6:



Find the sample variance and standard deviation for the amount of European auto sales for a sample of 6 years shown. The data are in millions of dollars.

11.2, 11.9, 12.0, 12.8, 13.4, 14.3

Example 6: (cont.)



Step 1 Find the sum of the values.

$$\Sigma X = 11.2 + 11.9 + 12.0 + 12.8 + 13.4 + 14.3 = 75.6$$

Step 2 Square each value and find the sum.

$$\Sigma X^2 = 11.2^2 + 11.9^2 + 12.0^2 + 12.8^2 + 13.4^2 + 14.3^2 = 958.94$$

Step 3 Substitute in the formulas and solve.

$$\begin{aligned} s^2 &= \frac{n(\Sigma X^2) - (\Sigma X)^2}{n(n-1)} \\ &= \frac{6(958.94) - 75.6^2}{6(6-1)} \\ &= \frac{5753.64 - 5715.36}{6(5)} \\ &= \frac{38.28}{30} \\ &= 1.276 \end{aligned}$$

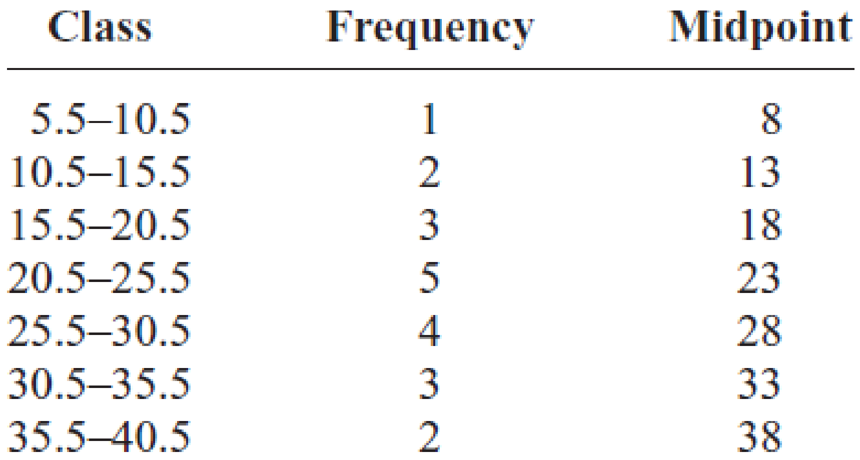
The variance is 1.28 rounded.

$$s = \sqrt{1.28} = 1.13$$

Hence, the sample standard deviation is 1.13.

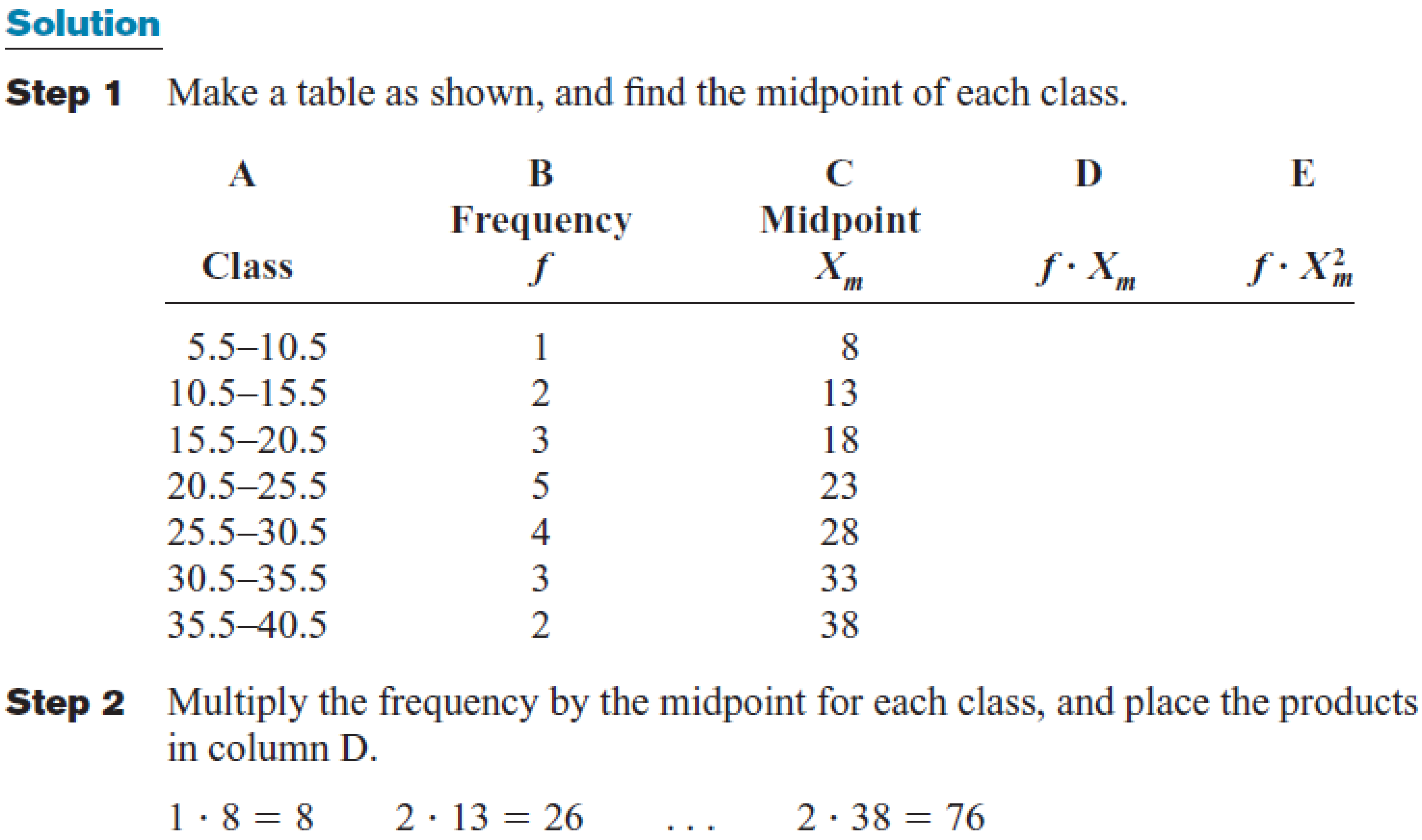
Example 7: Variance and Standard Deviation for Grouped Data

Find the variance and the standard deviation for the frequency distribution of the data for miles run per week



Class	Frequency	Midpoint
5.5–10.5	1	8
10.5–15.5	2	13
15.5–20.5	3	18
20.5–25.5	5	23
25.5–30.5	4	28
30.5–35.5	3	33
35.5–40.5	2	38

Example 7 (cont.):



Solution

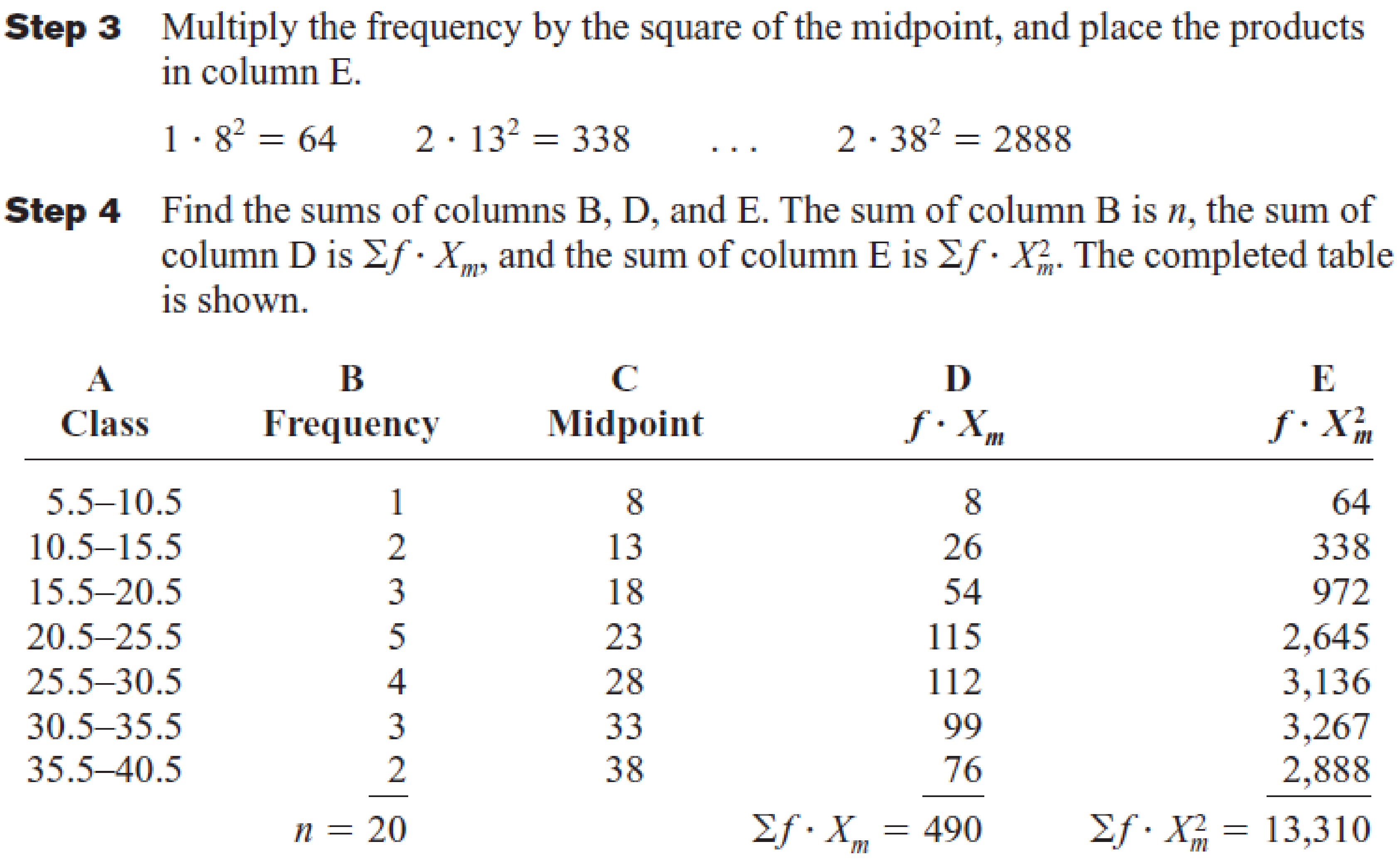
Step 1 Make a table as shown, and find the midpoint of each class.

A	B	C	D	E
	Frequency	Midpoint		
Class	f	X_m	$f \cdot X_m$	$f \cdot X_m^2$
5.5–10.5	1	8		
10.5–15.5	2	13		
15.5–20.5	3	18		
20.5–25.5	5	23		
25.5–30.5	4	28		
30.5–35.5	3	33		
35.5–40.5	2	38		

Step 2 Multiply the frequency by the midpoint for each class, and place the products in column D.

$1 \cdot 8 = 8 \qquad 2 \cdot 13 = 26 \qquad \ldots \qquad 2 \cdot 38 = 76$

Example 7 (cont.):



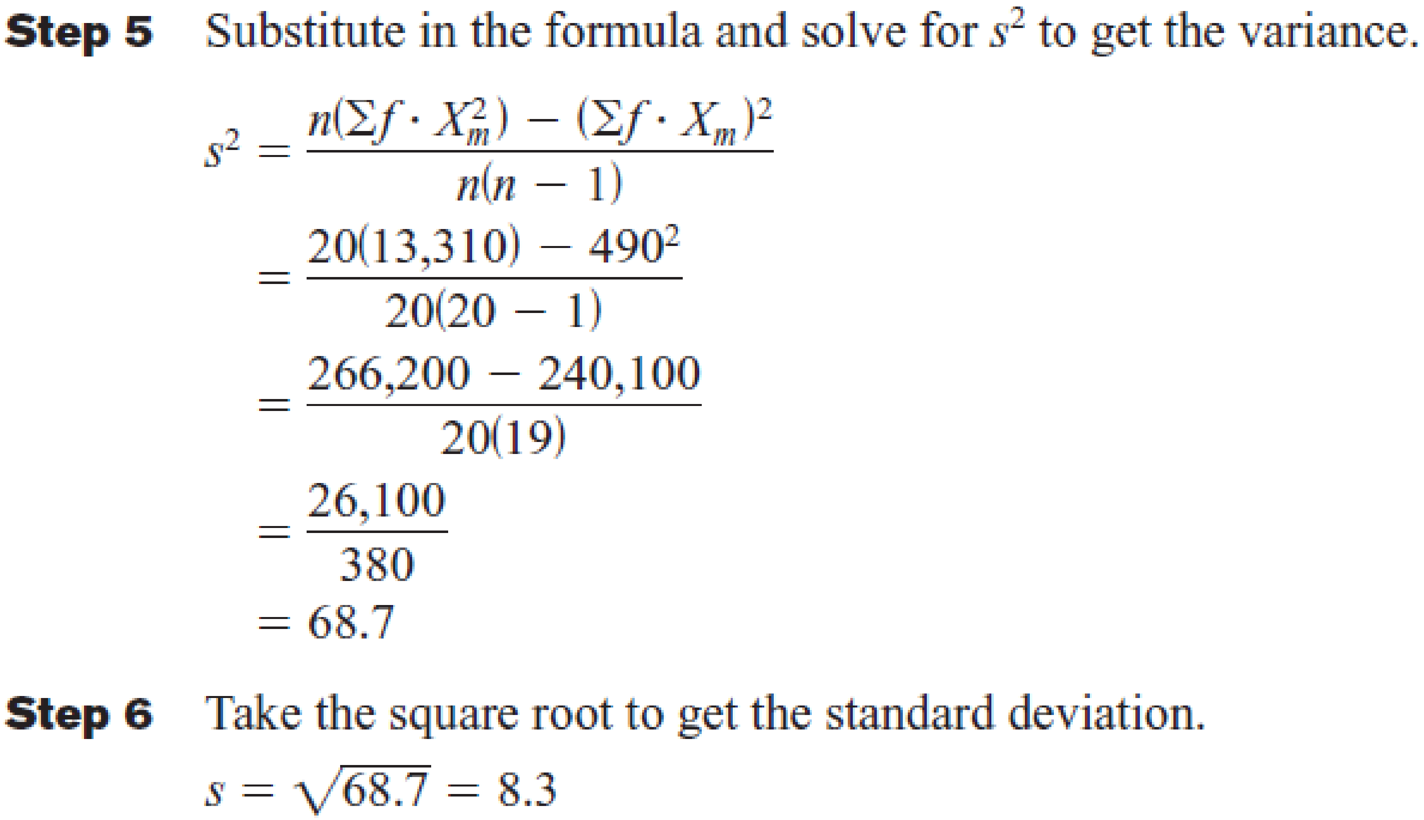
Step 3 Multiply the frequency by the square of the midpoint, and place the products in column E.

$1 \cdot 8^2 = 64 \qquad 2 \cdot 13^2 = 338 \qquad \ldots \qquad 2 \cdot 38^2 = 2888$

Step 4 Find the sums of columns B, D, and E. The sum of column B is n, the sum of column D is $\Sigma f \cdot X_m$, and the sum of column E is $\Sigma f \cdot X_m^2$. The completed table is shown.

A	B	C	D	E
Class	Frequency	Midpoint	$f \cdot X_m$	$f \cdot X_m^2$
5.5–10.5	1	8	8	64
10.5–15.5	2	13	26	338
15.5–20.5	3	18	54	972
20.5–25.5	5	23	115	2,645
25.5–30.5	4	28	112	3,136
30.5–35.5	3	33	99	3,267
35.5–40.5	2	38	76	2,888
	$n = 20$		$\Sigma f \cdot X_m = 490$	$\Sigma f \cdot X_m^2 = 13{,}310$

Example 7 (cont.):



Step 5 Substitute in the formula and solve for s^2 to get the variance.

$$\begin{aligned} s^2 &= \frac{n(\Sigma f \cdot X_m^2) - (\Sigma f \cdot X_m)^2}{n(n-1)} \\ &= \frac{20(13{,}310) - 490^2}{20(20-1)} \\ &= \frac{266{,}200 - 240{,}100}{20(19)} \\ &= \frac{26{,}100}{380} \\ &= 68.7 \end{aligned}$$

Step 6 Take the square root to get the standard deviation.

$$s = \sqrt{68.7} = 8.3$$

Procedure Table

Finding the Sample Variance and Standard Deviation for Grouped Data

Step 1 Make a table as shown, and find the midpoint of each class.

A	B	C	D	E
Class	Frequency	Midpoint	$f \cdot X_m$	$f \cdot X_m^2$

Step 2 Multiply the frequency by the midpoint for each class, and place the products in column D.

Step 3 Multiply the frequency by the square of the midpoint, and place the products in column E.

Step 4 Find the sums of columns B, D, and E. (The sum of column B is *n*. The sum of column D is $\Sigma f \cdot X_m$. The sum of column E is $\Sigma f \cdot X_m^2$.)

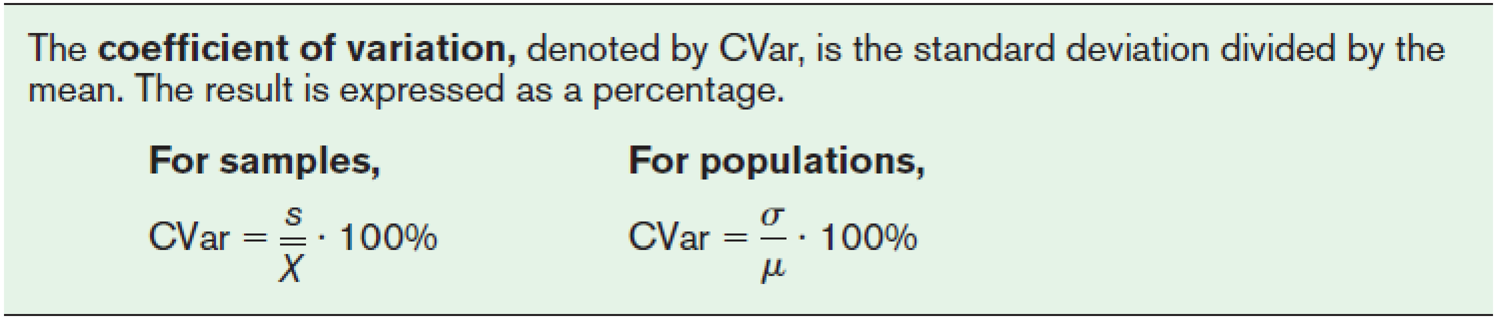
Step 5 Substitute in the formula and solve to get the variance.

$$s^2 = \frac{n(\Sigma f \cdot X_m^2) - (\Sigma f \cdot X_m)^2}{n(n - 1)}$$

Step 6 Take the square root to get the standard deviation.

Coefficient of Variation

A statistic that allows you to compare standard deviations when the units are different, is called the *coefficient of variation.*



The **coefficient of variation,** denoted by CVar, is the standard deviation divided by the mean. The result is expressed as a percentage.

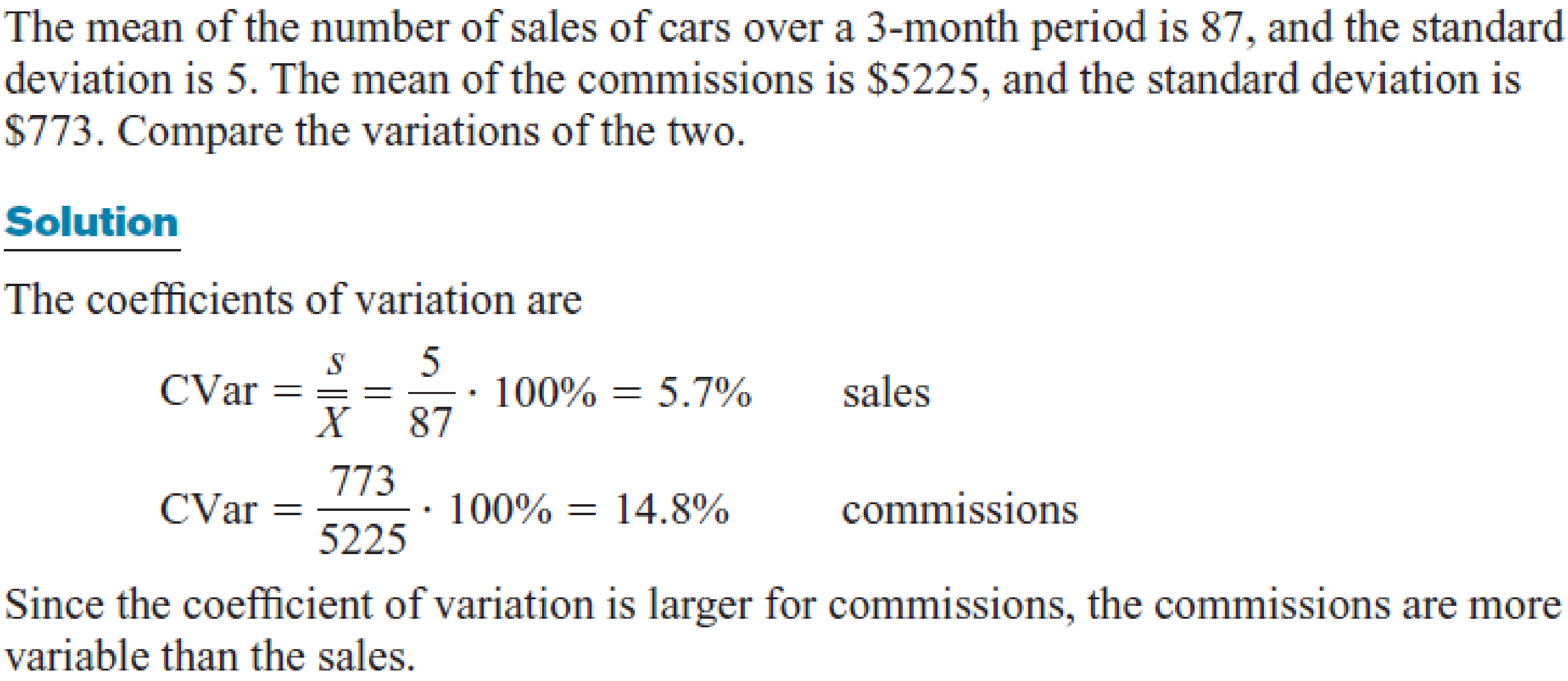
For samples,

$$\text{CVar} = \frac{s}{\overline{X}} \cdot 100\%$$

For populations,

$$\text{CVar} = \frac{\sigma}{\mu} \cdot 100\%$$

Example 8: Sales of Automobiles



The mean of the number of sales of cars over a 3-month period is 87, and the standard deviation is 5. The mean of the commissions is $5225, and the standard deviation is $773. Compare the variations of the two.

Solution

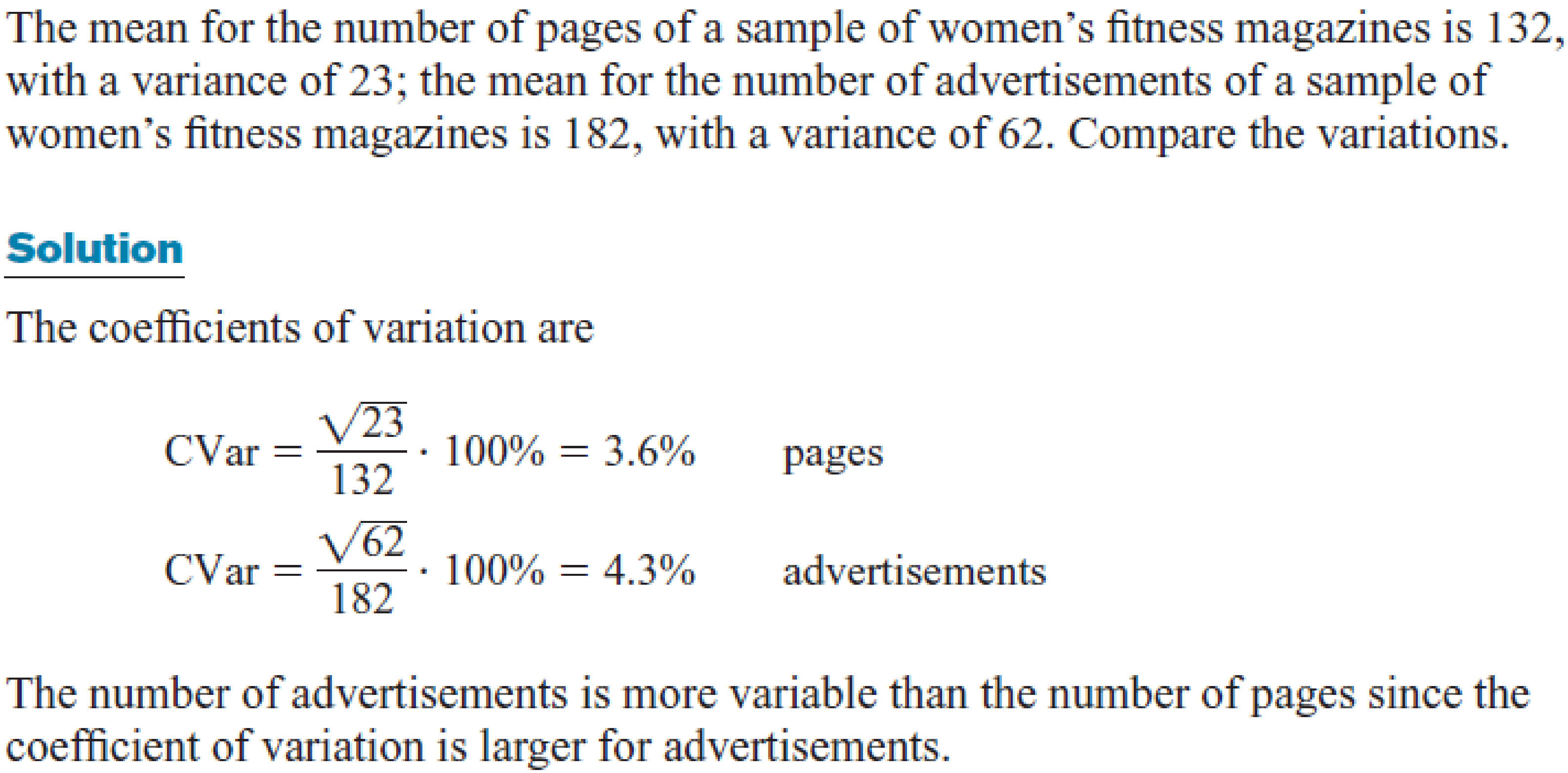
The coefficients of variation are

$$\text{CVar} = \frac{s}{\overline{X}} = \frac{5}{87} \cdot 100\% = 5.7\% \qquad \text{sales}$$

$$\text{CVar} = \frac{773}{5225} \cdot 100\% = 14.8\% \qquad \text{commissions}$$

Since the coefficient of variation is larger for commissions, the commissions are more variable than the sales.

Example 9:



The mean for the number of pages of a sample of women's fitness magazines is 132, with a variance of 23; the mean for the number of advertisements of a sample of women's fitness magazines is 182, with a variance of 62. Compare the variations.

Solution

The coefficients of variation are

$$\text{CVar} = \frac{\sqrt{23}}{132} \cdot 100\% = 3.6\% \qquad \text{pages}$$

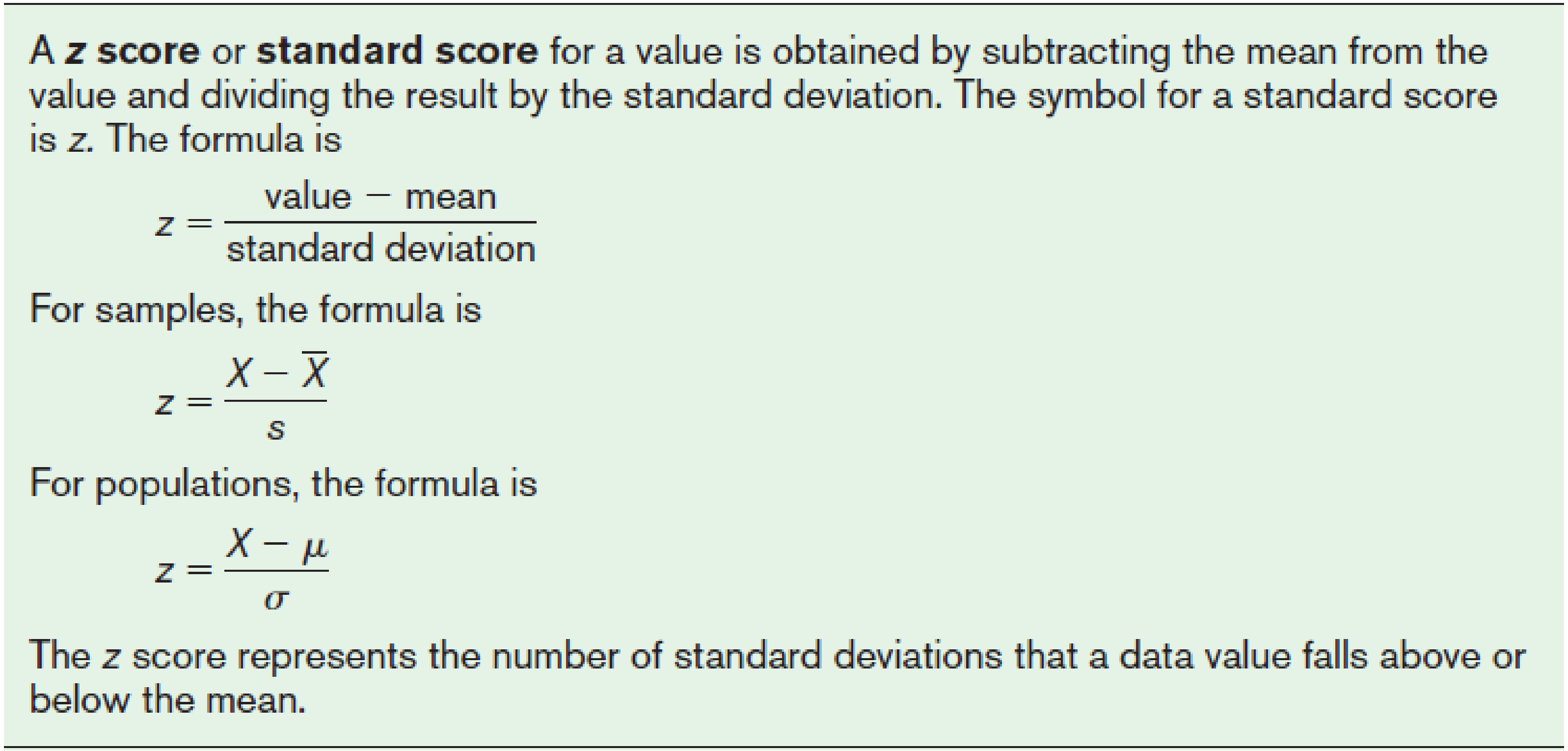
$$\text{CVar} = \frac{\sqrt{62}}{182} \cdot 100\% = 4.3\% \qquad \text{advertisements}$$

The number of advertisements is more variable than the number of pages since the coefficient of variation is larger for advertisements.

Measures of Position

These measures include standard scores, percentiles, deciles, and quartiles. They are used to locate the relative position of a data value in the data set. For example, if a value is located at the 80th percentile, it means that 80% of the values fall below it in the distribution and 20% of the values fall above it. The *median* is the value that corresponds to the 50th percentile, since one-half of the values fall below it and onehalf of the values fall above it.

Measures of Position-z score



A **z score** or **standard score** for a value is obtained by subtracting the mean from the value and dividing the result by the standard deviation. The symbol for a standard score is *z*. The formula is

$$z = \frac{\text{value} - \text{mean}}{\text{standard deviation}}$$

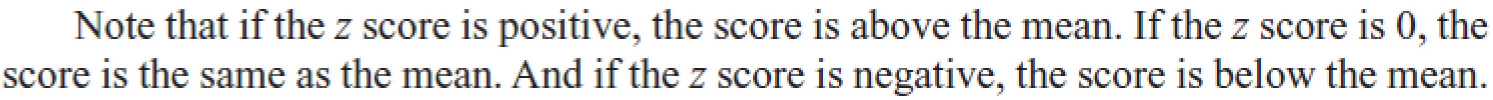
For samples, the formula is

$$z = \frac{X - \overline{X}}{s}$$

For populations, the formula is

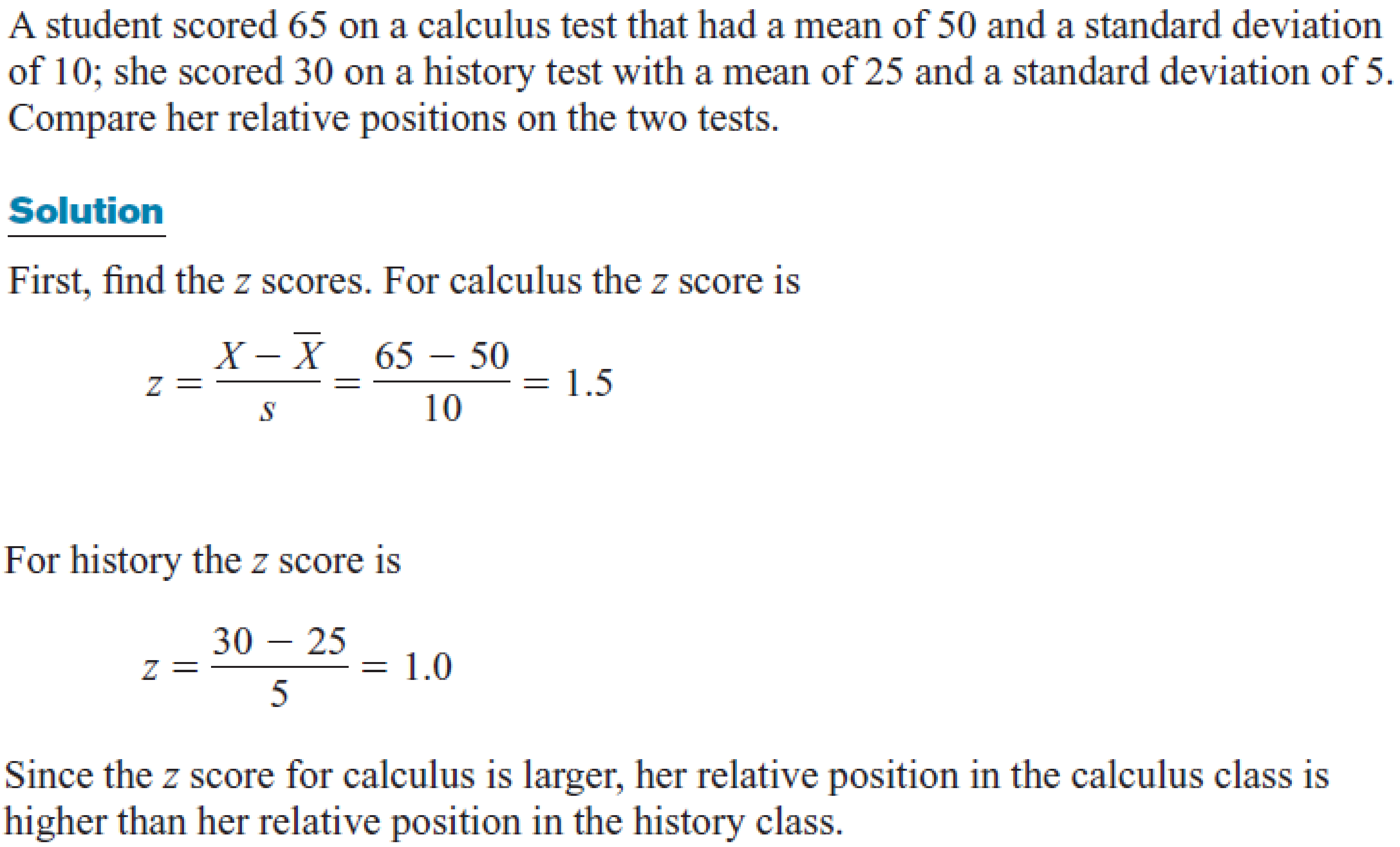
$$z = \frac{X - \mu}{\sigma}$$

The *z* score represents the number of standard deviations that a data value falls above or below the mean.



Note that if the *z* score is positive, the score is above the mean. If the *z* score is 0, the score is the same as the mean. And if the *z* score is negative, the score is below the mean.

Example 10:



A student scored 65 on a calculus test that had a mean of 50 and a standard deviation of 10; she scored 30 on a history test with a mean of 25 and a standard deviation of 5. Compare her relative positions on the two tests.

Solution

First, find the *z* scores. For calculus the *z* score is

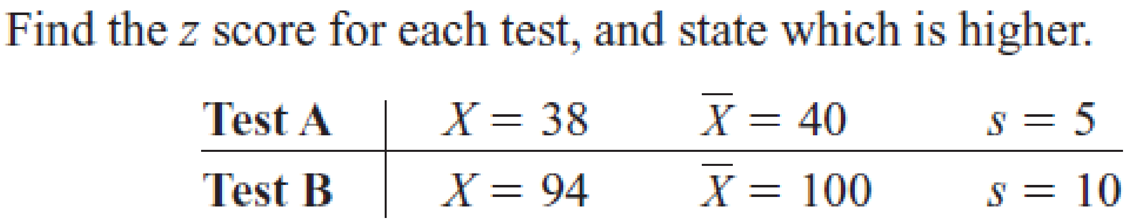
$$z = \frac{X - \overline{X}}{s} = \frac{65 - 50}{10} = 1.5$$

For history the *z* score is

$$z = \frac{30 - 25}{5} = 1.0$$

Since the *z* score for calculus is larger, her relative position in the calculus class is higher than her relative position in the history class.

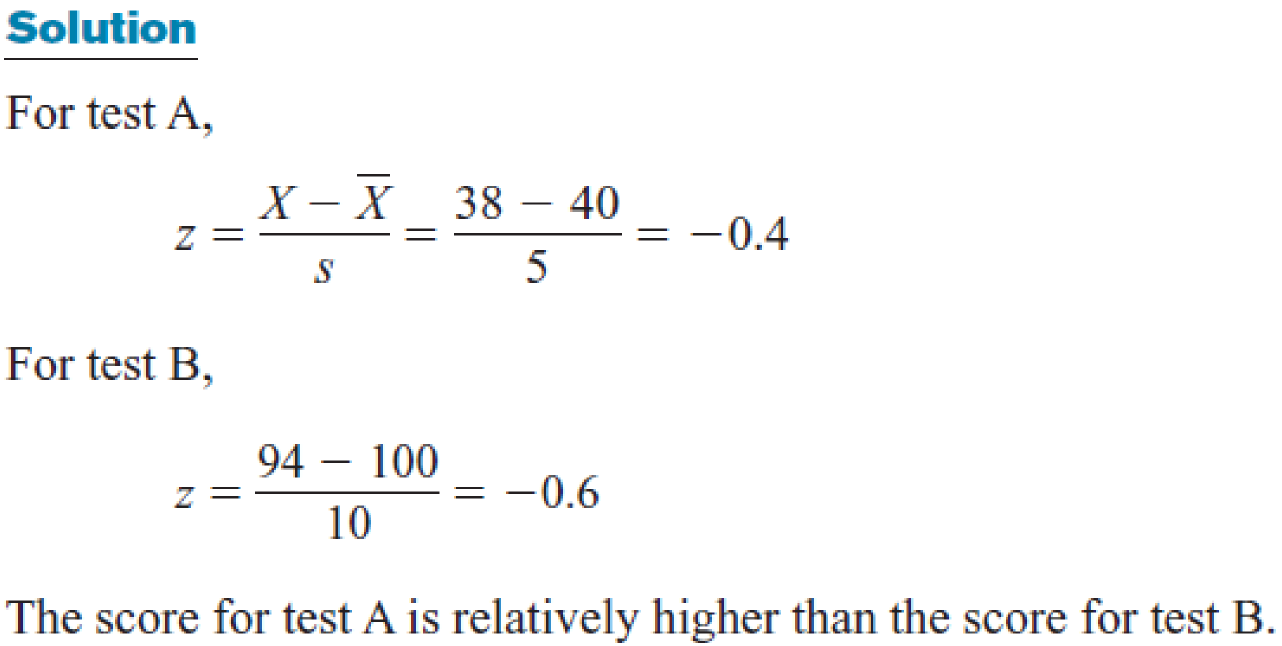
Example 11:



Find the z score for each test, and state which is higher.

Test A	$X = 38$	$\overline{X} = 40$	$s = 5$
Test B	$X = 94$	$\overline{X} = 100$	$s = 10$

Example 11: (cont.)



Solution

For test A,

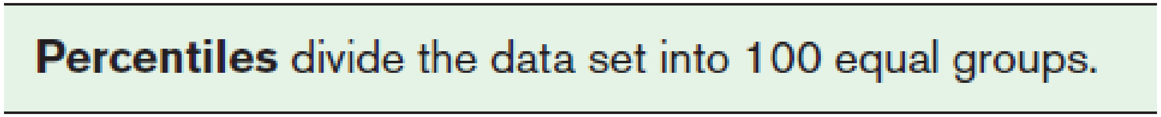
$$z = \frac{X - \overline{X}}{s} = \frac{38 - 40}{5} = -0.4$$

For test B,

$$z = \frac{94 - 100}{10} = -0.6$$

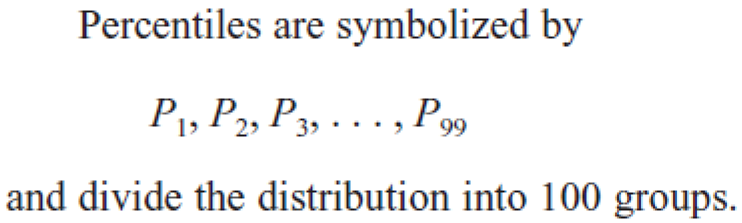
The score for test A is relatively higher than the score for test B.

Measures of Position – Percentiles



Percentiles divide the data set into 100 equal groups.

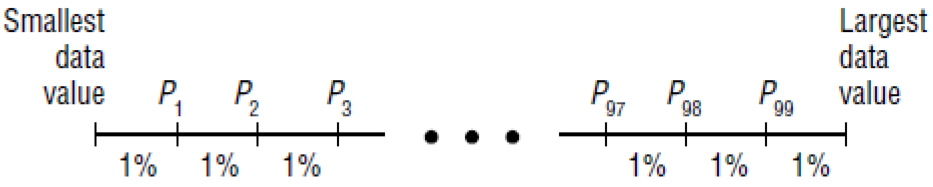
Figure 2 shows percentiles in graphical form of weights of girls from ages 2 to 18. To find the percentile rank of an 11-year-old who weighs 82 pounds, start at the 82-pound weight on the left axis and move horizontally to the right. Find 11 on the horizontal axis and move up vertically. The two lines meet at the 50th percentile curved line; hence, an 11-year-old girl who weighs 82 pounds is in the 50th percentile for her age group.

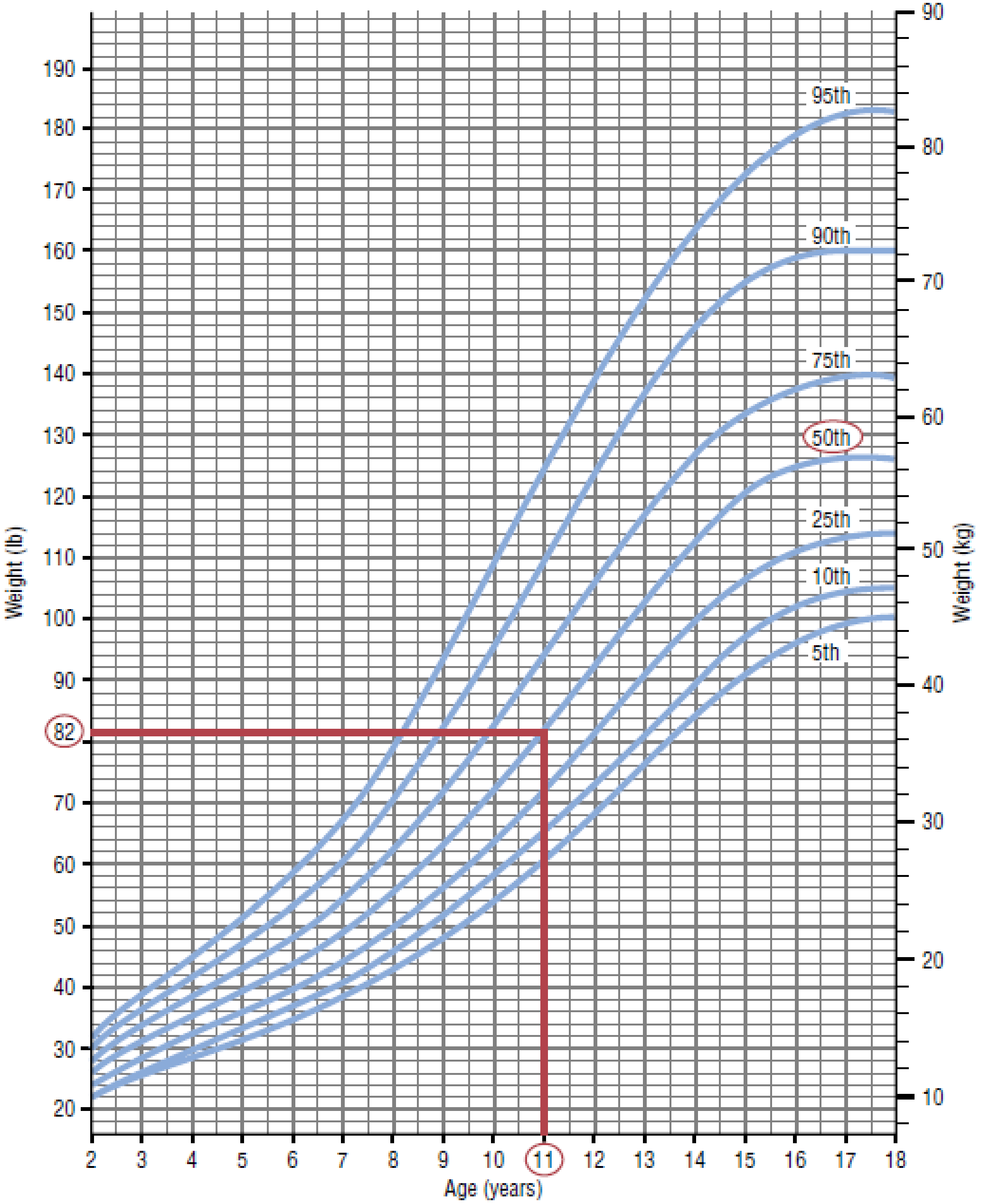


Percentiles are symbolized by

$$P_1, P_2, P_3, \ldots, P_{99}$$

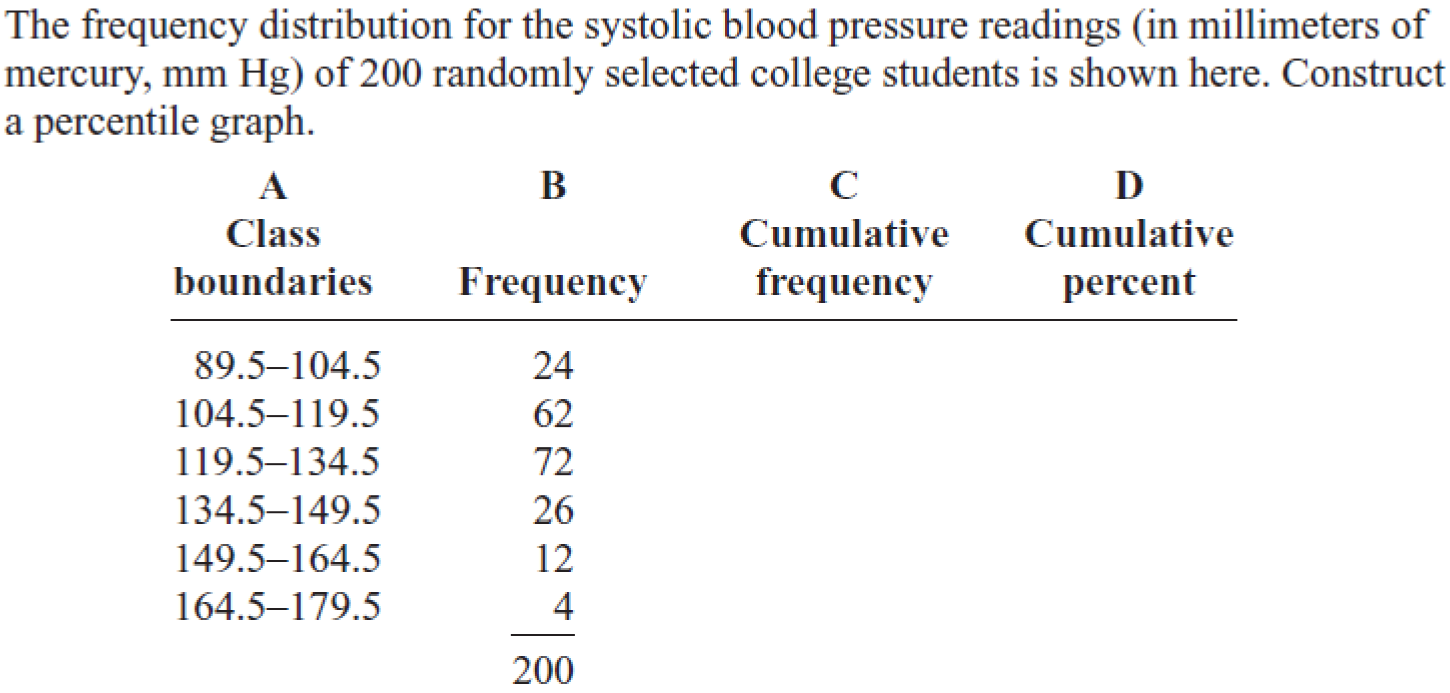
and divide the distribution into 100 groups.





95th
90th
75th
50th
25th
10th
5th
Weight (lb)
190
180
170
160
150
140
130
120
110
100
90
82
70
60
50
40
30
20
Weight (kg)
90
80
70
60
50
40
30
20
10
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
Age (years)

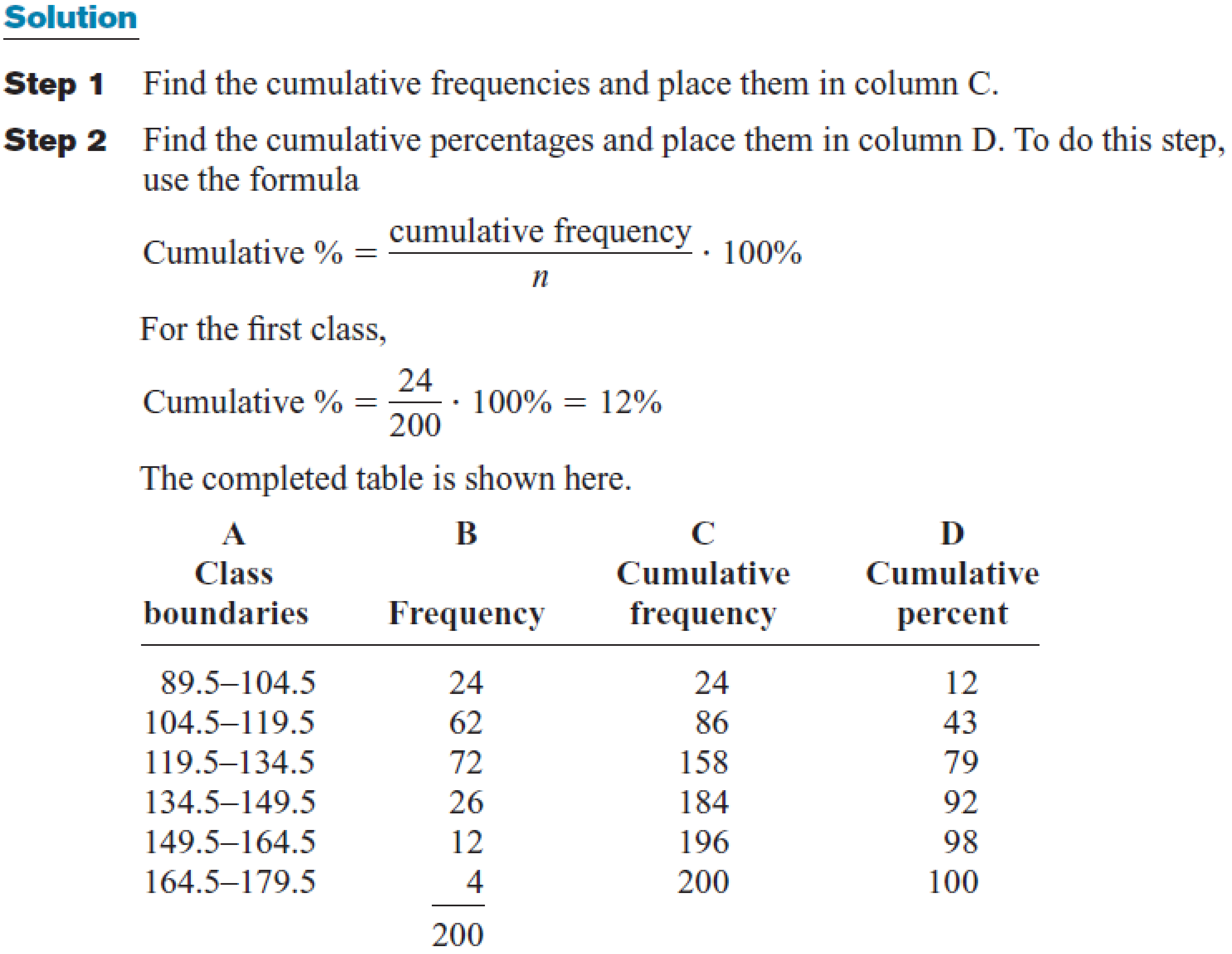
Example 12:



The frequency distribution for the systolic blood pressure readings (in millimeters of mercury, mm Hg) of 200 randomly selected college students is shown here. Construct a percentile graph.

A Class boundaries	B Frequency	C Cumulative frequency	D Cumulative percent
89.5–104.5	24		
104.5–119.5	62		
119.5–134.5	72		
134.5–149.5	26		
149.5–164.5	12		
164.5–179.5	4		
	200		

Example 12: (cont.)



Solution

Step 1 Find the cumulative frequencies and place them in column C.

Step 2 Find the cumulative percentages and place them in column D. To do this step, use the formula

$$\text{Cumulative \%} = \frac{\text{cumulative frequency}}{n} \cdot 100\%$$

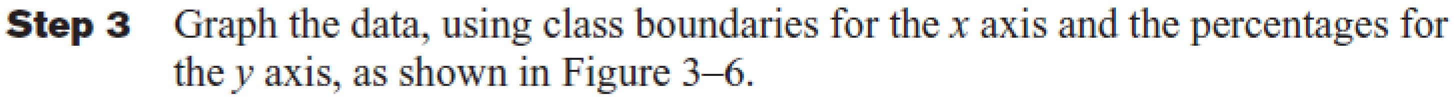
For the first class,

$$\text{Cumulative \%} = \frac{24}{200} \cdot 100\% = 12\%$$

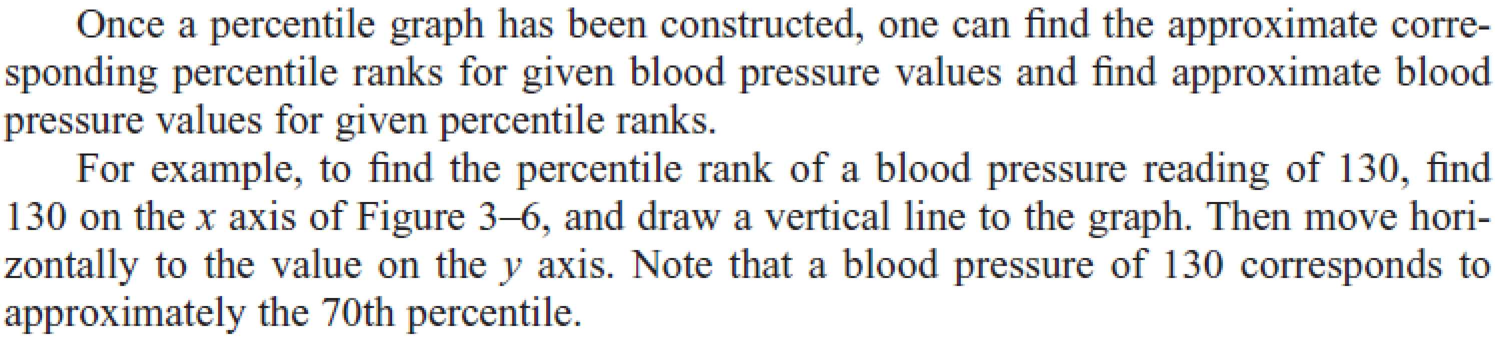
The completed table is shown here.

A	B	C	D
Class boundaries	Frequency	Cumulative frequency	Cumulative percent
89.5–104.5	24	24	12
104.5–119.5	62	86	43
119.5–134.5	72	158	79
134.5–149.5	26	184	92
149.5–164.5	12	196	98
164.5–179.5	4	200	100
	200		

Example 12: (cont.)

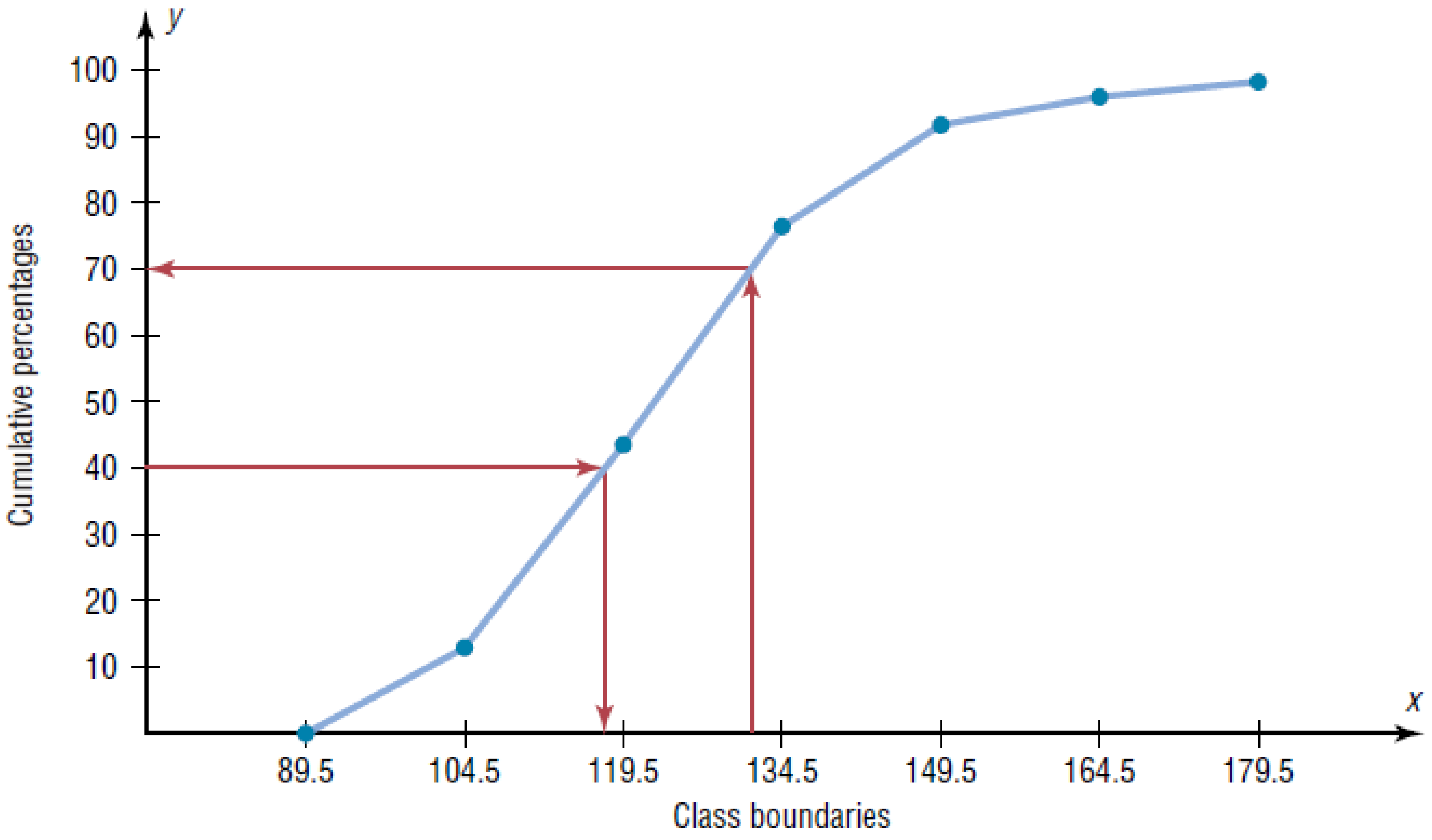


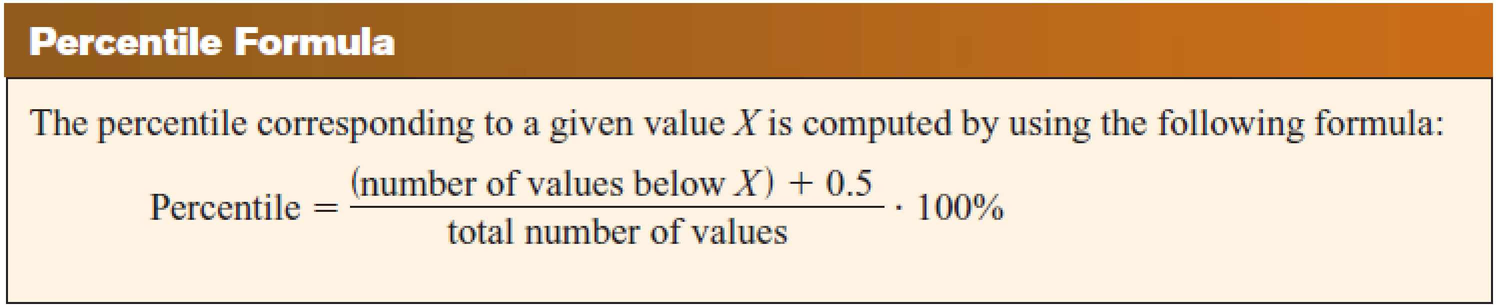
Step 3 Graph the data, using class boundaries for the *x* axis and the percentages for the *y* axis, as shown in Figure 3–6.



Once a percentile graph has been constructed, one can find the approximate corresponding percentile ranks for given blood pressure values and find approximate blood pressure values for given percentile ranks.

For example, to find the percentile rank of a blood pressure reading of 130, find 130 on the *x* axis of Figure 3–6, and draw a vertical line to the graph. Then move horizontally to the value on the *y* axis. Note that a blood pressure of 130 corresponds to approximately the 70th percentile.


y
100
90
80
70
60
50
40
30
20
10
Cumulative percentages
x
89.5
104.5
119.5
134.5
149.5
164.5
179.5
Class boundaries

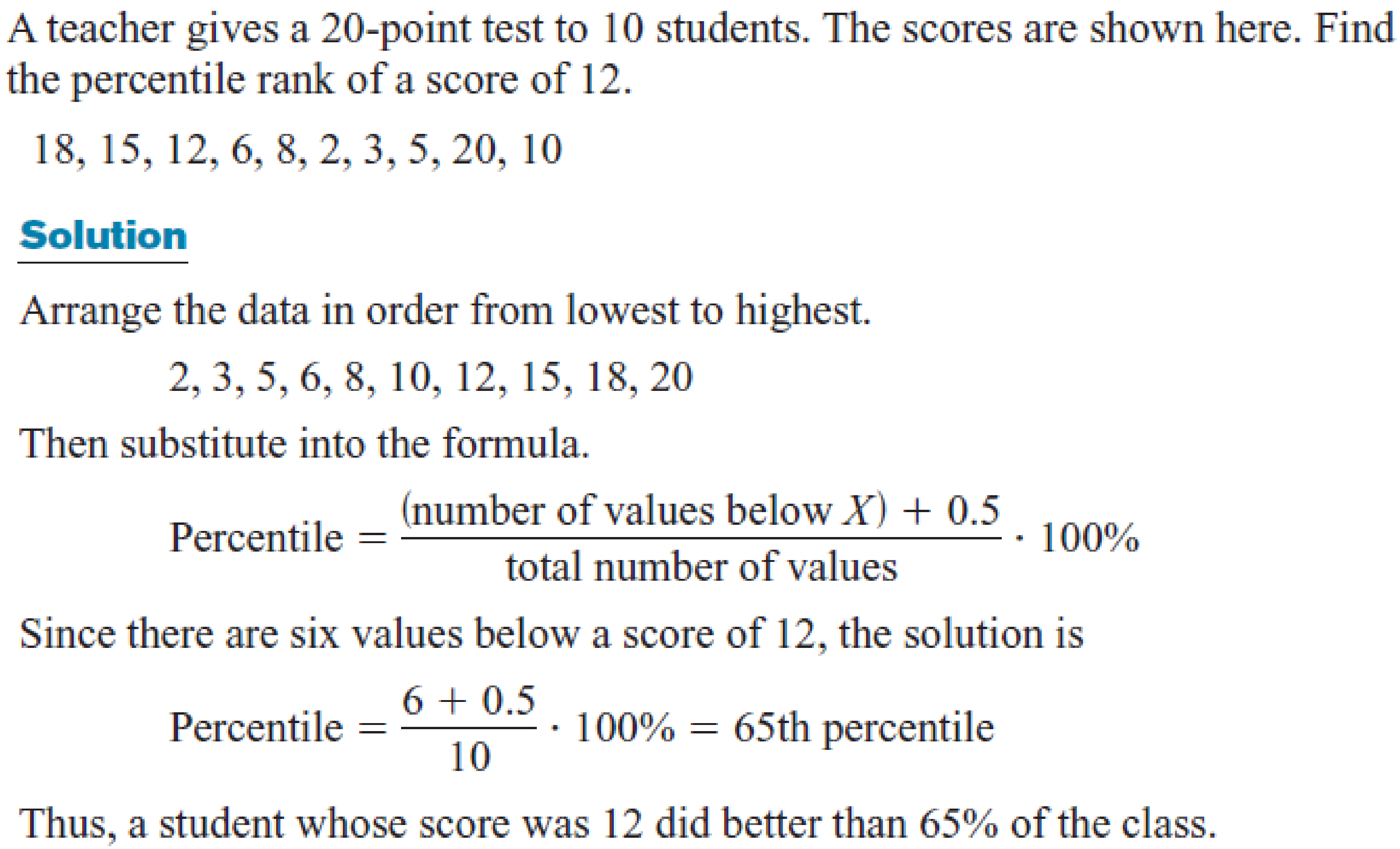


Percentile Formula

The percentile corresponding to a given value *X* is computed by using the following formula:

$$\text{Percentile} = \frac{(\text{number of values below } X) + 0.5}{\text{total number of values}} \cdot 100\%$$

Example 13:



A teacher gives a 20-point test to 10 students. The scores are shown here. Find the percentile rank of a score of 12.

18, 15, 12, 6, 8, 2, 3, 5, 20, 10

Solution

Arrange the data in order from lowest to highest.

2, 3, 5, 6, 8, 10, 12, 15, 18, 20

Then substitute into the formula.

$$\text{Percentile} = \frac{(\text{number of values below } X) + 0.5}{\text{total number of values}} \cdot 100\%$$

Since there are six values below a score of 12, the solution is

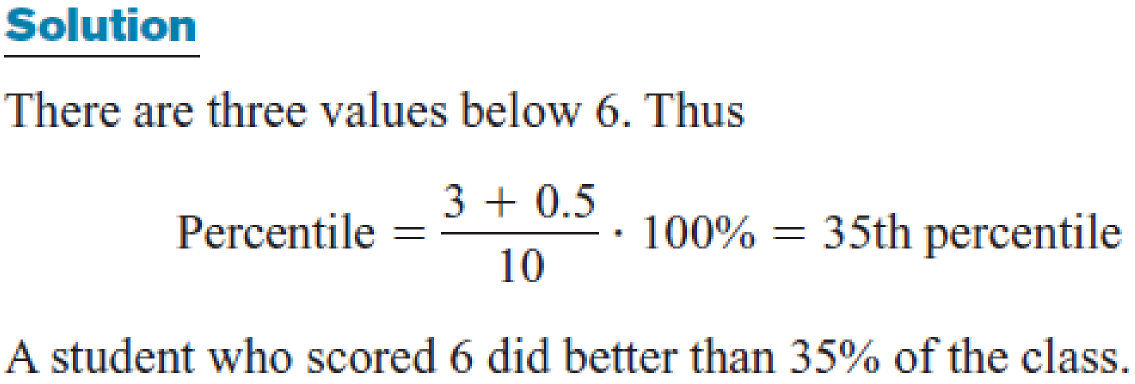
$$\text{Percentile} = \frac{6 + 0.5}{10} \cdot 100\% = 65\text{th percentile}$$

Thus, a student whose score was 12 did better than 65% of the class.

Example 14:

Using the data in Example 13, find the percentile rank for a score of 6.

Example 14: (cont.)



Solution

There are three values below 6. Thus

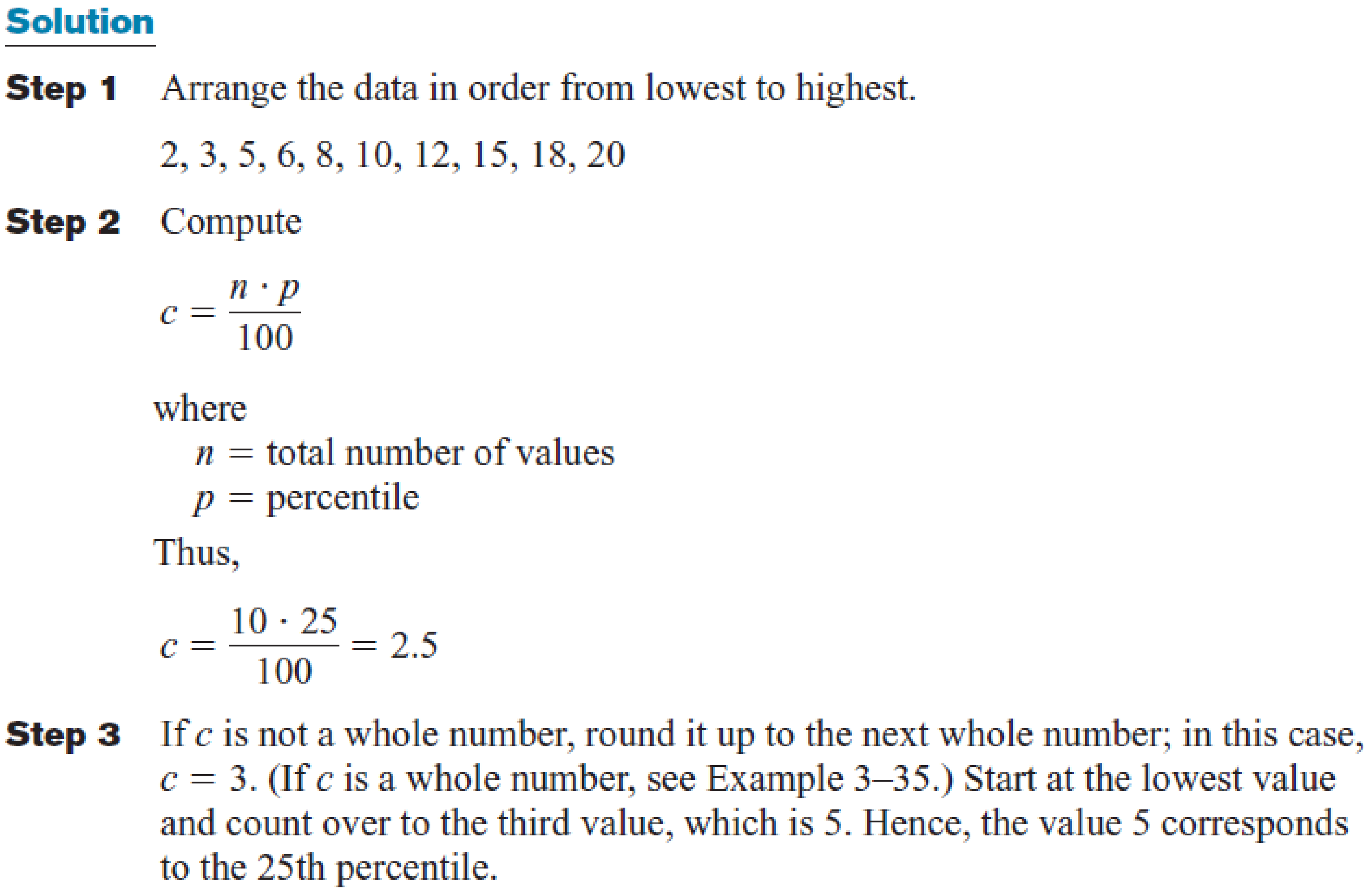
$$\text{Percentile} = \frac{3 + 0.5}{10} \cdot 100\% = 35\text{th percentile}$$

A student who scored 6 did better than 35% of the class.

Examples 15 and 16 show a procedure for finding a value corresponding to a given percentile.

Example 15:

Using the scores in Example 13, find the value corresponding to the 25th percentile.



Solution

Step 1 Arrange the data in order from lowest to highest.

2, 3, 5, 6, 8, 10, 12, 15, 18, 20

Step 2 Compute

$$c = \frac{n \cdot p}{100}$$

where
n = total number of values
p = percentile

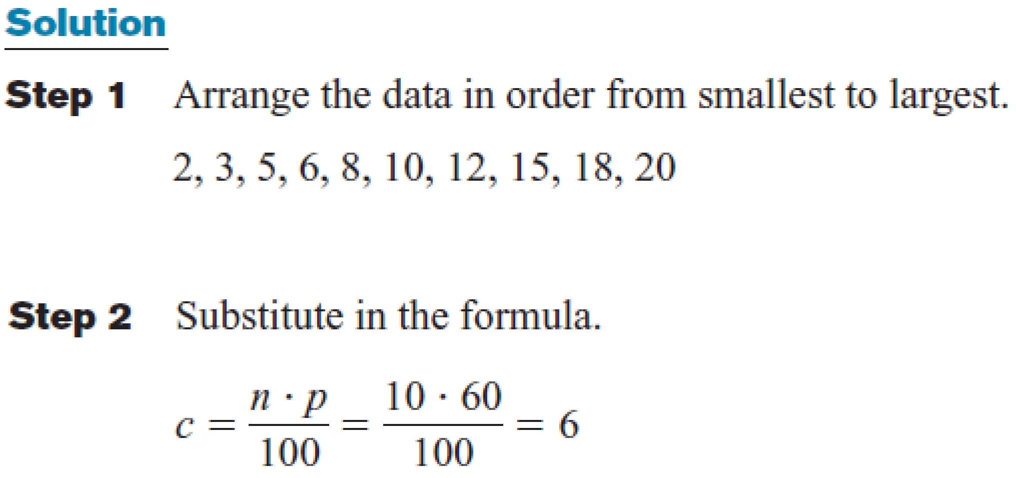
Thus,

$$c = \frac{10 \cdot 25}{100} = 2.5$$

Step 3 If c is not a whole number, round it up to the next whole number; in this case, $c = 3$. (If c is a whole number, see Example 3–35.) Start at the lowest value and count over to the third value, which is 5. Hence, the value 5 corresponds to the 25th percentile.

Example 16:

Using the data set in Example 13, find the value that corresponds to the 60th percentile.



Solution

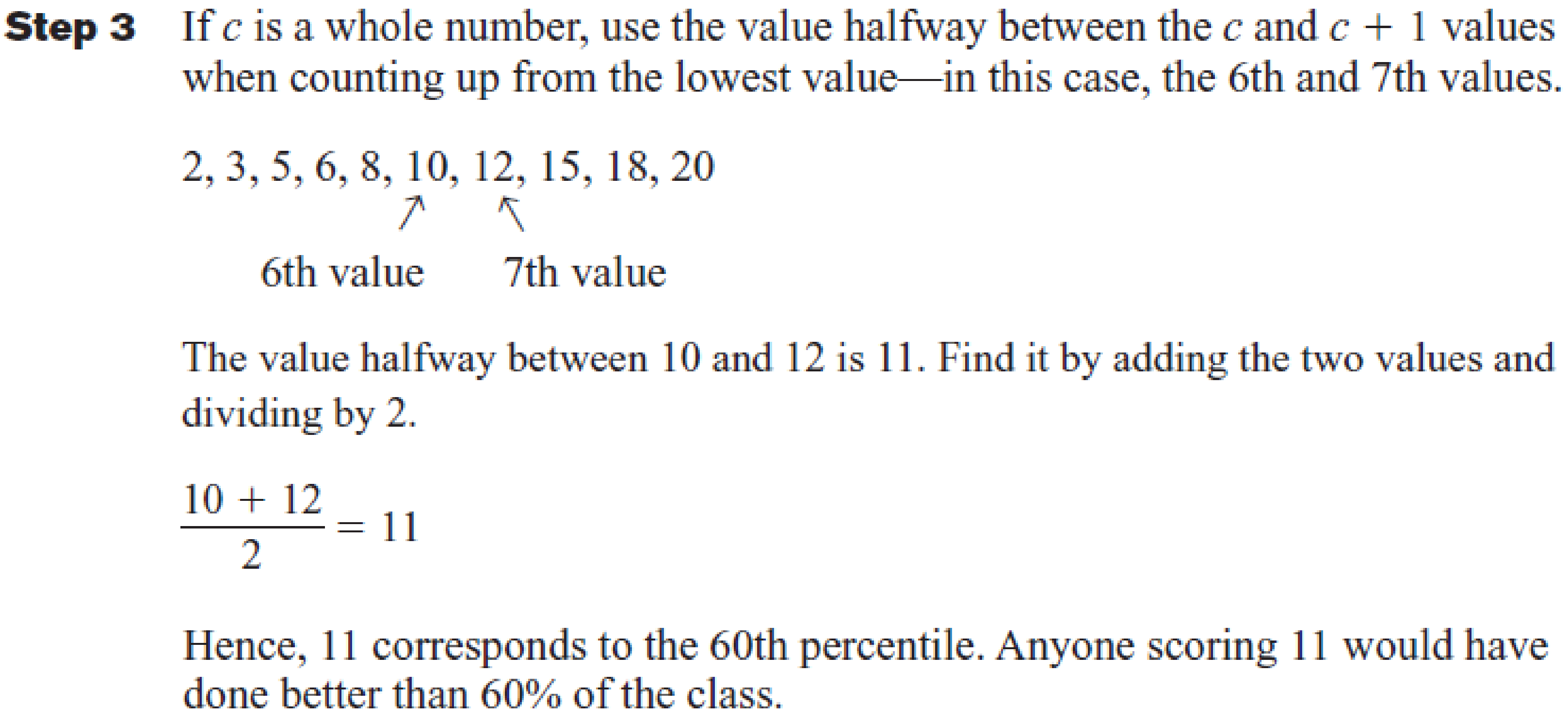
Step 1 Arrange the data in order from smallest to largest.

2, 3, 5, 6, 8, 10, 12, 15, 18, 20

Step 2 Substitute in the formula.

$$c = \frac{n \cdot p}{100} = \frac{10 \cdot 60}{100} = 6$$

Example 16: (cont.)



Step 3 If c is a whole number, use the value halfway between the c and $c + 1$ values when counting up from the lowest value—in this case, the 6th and 7th values.

2, 3, 5, 6, 8, 10, 12, 15, 18, 20

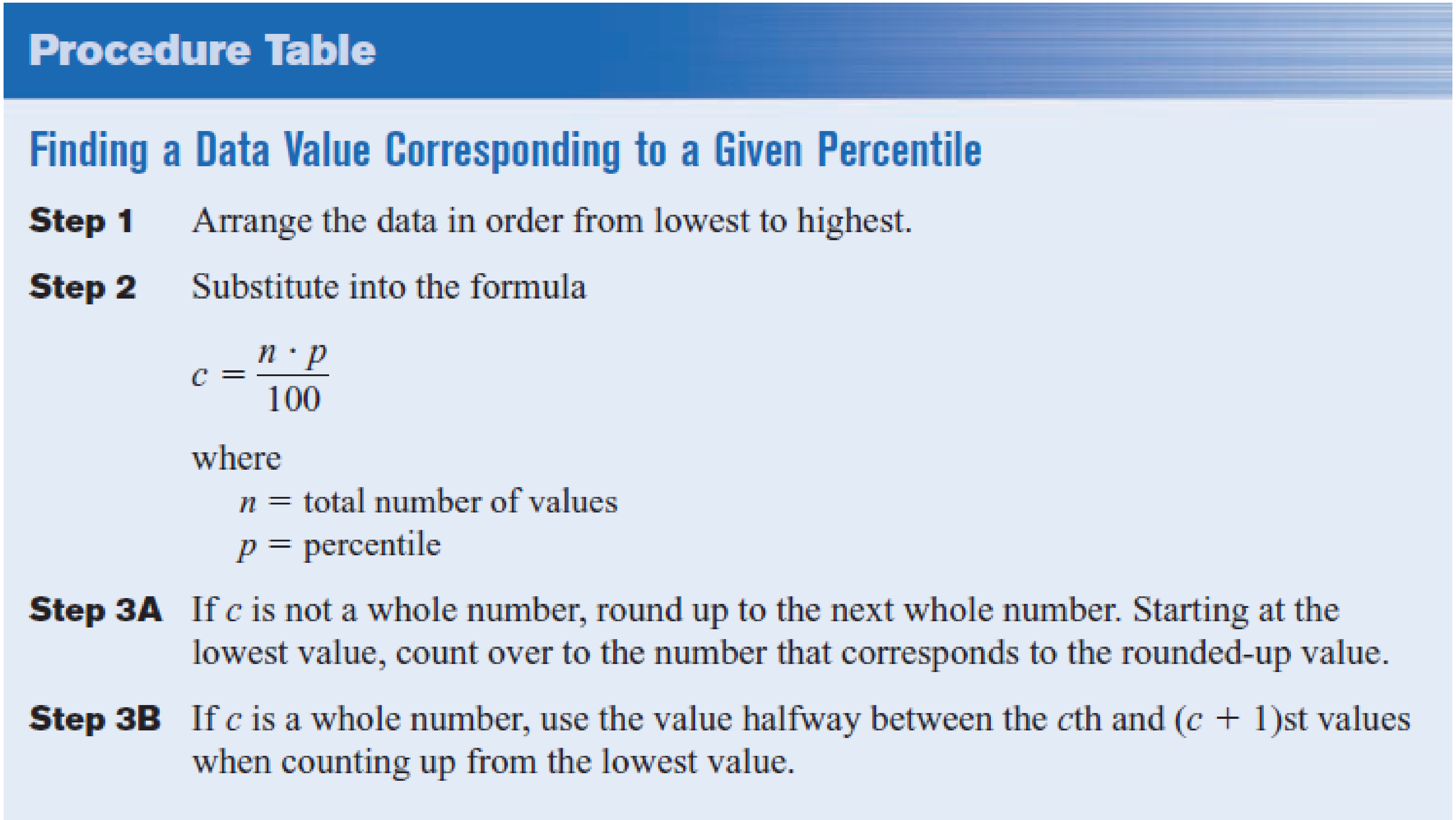
↗ ↖

6th value 7th value

The value halfway between 10 and 12 is 11. Find it by adding the two values and dividing by 2.

$$\frac{10 + 12}{2} = 11$$

Hence, 11 corresponds to the 60th percentile. Anyone scoring 11 would have done better than 60% of the class.



Procedure Table

Finding a Data Value Corresponding to a Given Percentile

Step 1 Arrange the data in order from lowest to highest.

Step 2 Substitute into the formula

$$c = \frac{n \cdot p}{100}$$

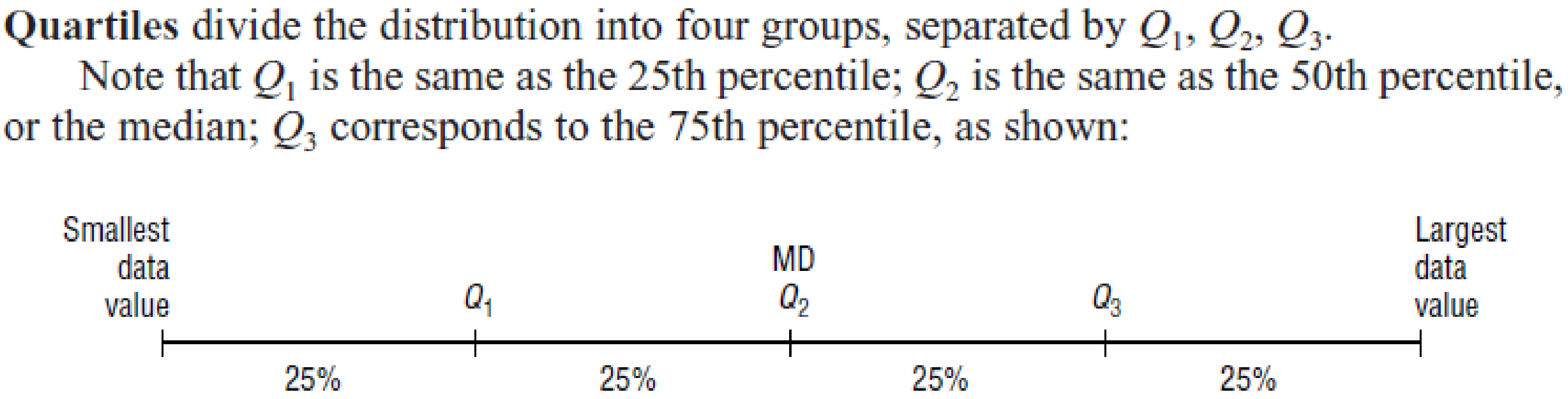
where

n = total number of values
p = percentile

Step 3A If c is not a whole number, round up to the next whole number. Starting at the lowest value, count over to the number that corresponds to the rounded-up value.

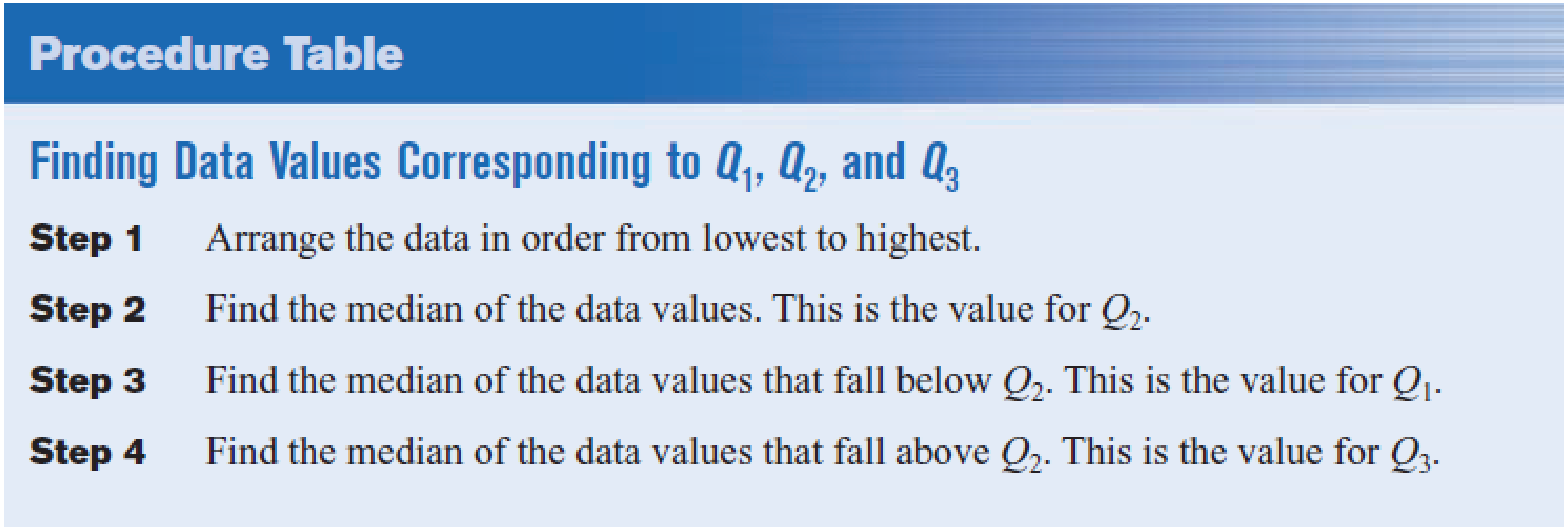
Step 3B If c is a whole number, use the value halfway between the cth and $(c + 1)$st values when counting up from the lowest value.

Measures of Position – Quartiles



Quartiles divide the distribution into four groups, separated by Q_1, Q_2, Q_3.

Note that Q_1 is the same as the 25th percentile; Q_2 is the same as the 50th percentile, or the median; Q_3 corresponds to the 75th percentile, as shown:



Procedure Table

Finding Data Values Corresponding to Q_1, Q_2, and Q_3

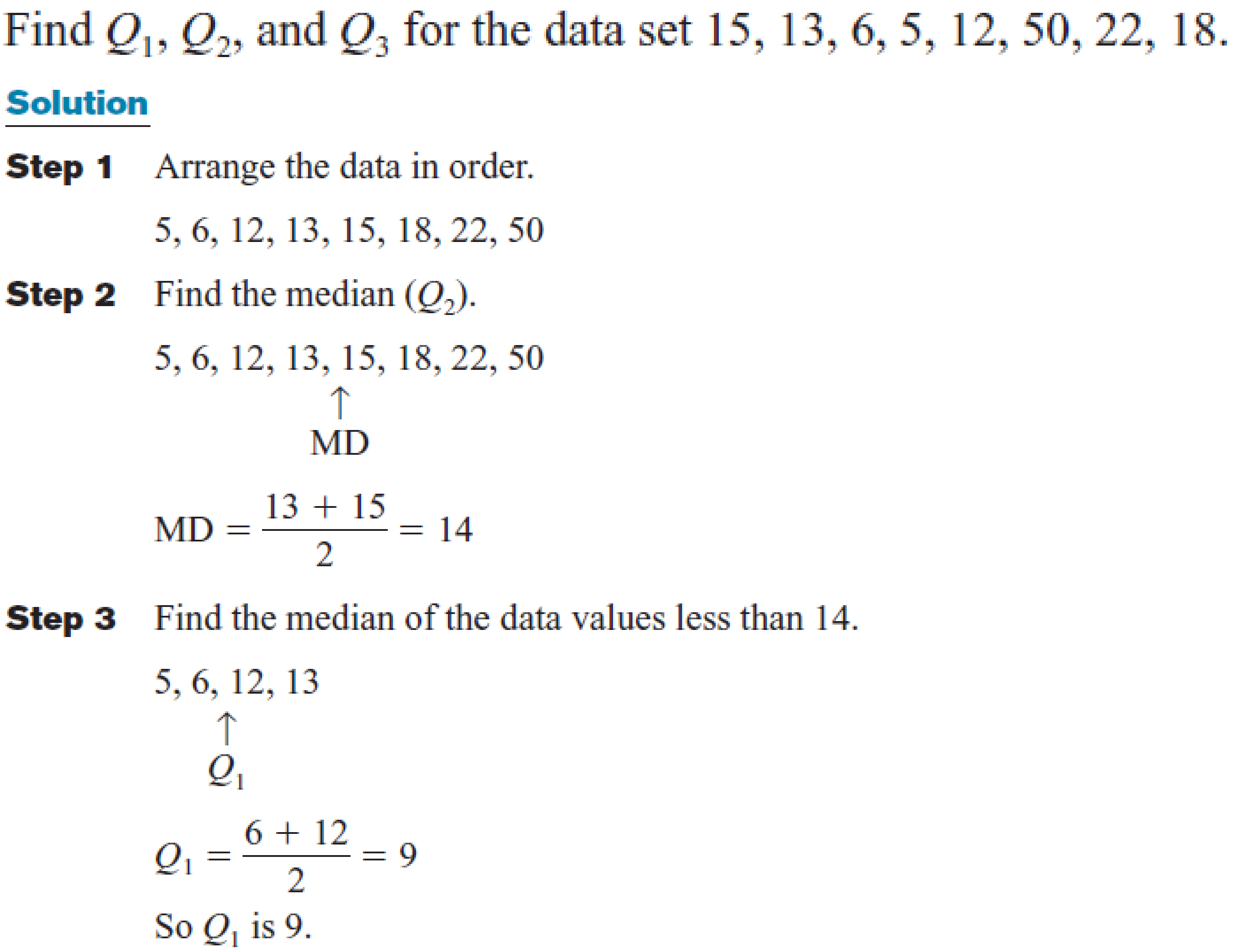
Step 1 Arrange the data in order from lowest to highest.

Step 2 Find the median of the data values. This is the value for Q_2.

Step 3 Find the median of the data values that fall below Q_2. This is the value for Q_1.

Step 4 Find the median of the data values that fall above Q_2. This is the value for Q_3.

Example 17:



Find Q_1, Q_2, and Q_3 for the data set 15, 13, 6, 5, 12, 50, 22, 18.

Solution

Step 1 Arrange the data in order.

5, 6, 12, 13, 15, 18, 22, 50

Step 2 Find the median (Q_2).

5, 6, 12, 13, 15, 18, 22, 50

↑

MD

$$\text{MD} = \frac{13 + 15}{2} = 14$$

Step 3 Find the median of the data values less than 14.

5, 6, 12, 13

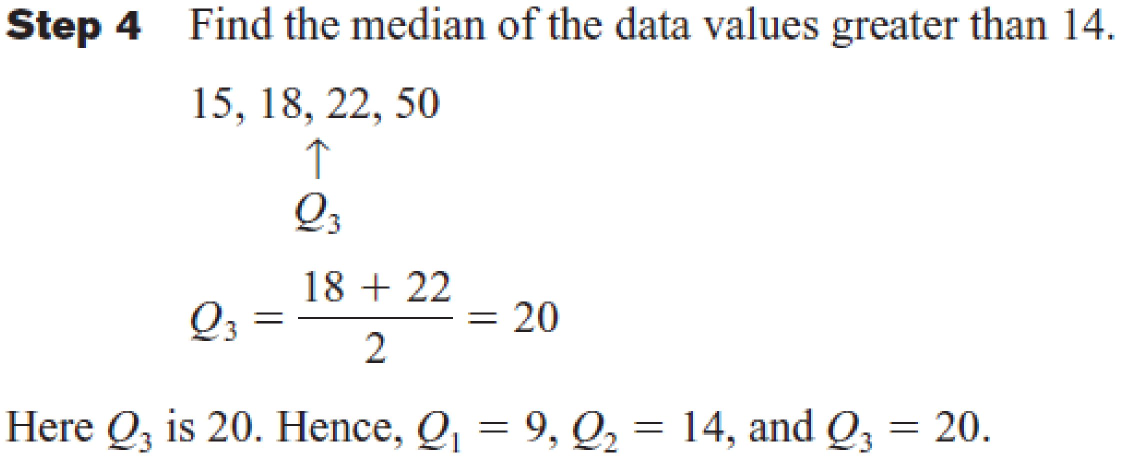
↑

Q_1

$$Q_1 = \frac{6 + 12}{2} = 9$$

So Q_1 is 9.

Example 17: (cont.)



Step 4 Find the median of the data values greater than 14.

$$15,\ 18,\ \underset{\substack{\uparrow \\ Q_3}}{}\ 22,\ 50$$

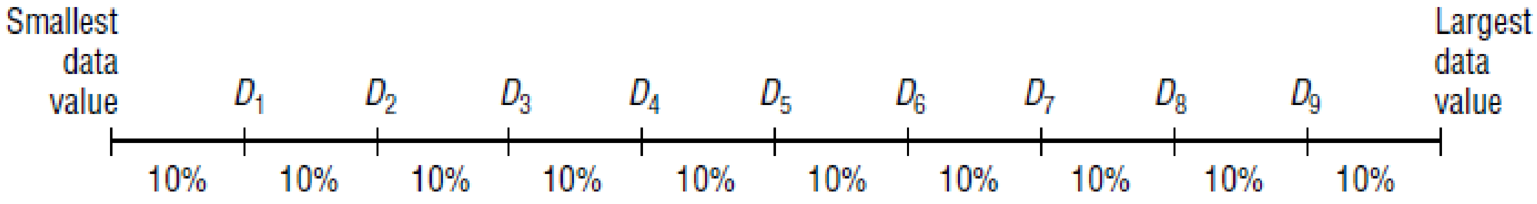
$$Q_3 = \frac{18 + 22}{2} = 20$$

Here Q_3 is 20. Hence, $Q_1 = 9$, $Q_2 = 14$, and $Q_3 = 20$.

In addition to dividing the data set into four groups, quartiles can be used as a rough measurement of variability. The **interquartile range (IQR)** is defined as the difference between *Q*1 and *Q*3 and is the range of the middle 50% of the data.

Measures of Position – Deciles

Deciles divide the distribution into 10 groups, as shown. They are denoted by D_1, D_2, etc.



Deciles are denoted by $D_1, D_2, D_3, \ldots, D_9$, and they correspond to $P_{10}, P_{20}, P_{30}, \ldots, P_{90}$.

Quartiles are denoted by Q_1, Q_2, Q_3 and they correspond to P_{25}, P_{50}, P_{75}.

The median is the same as P_{50} or Q_2 or D_5.

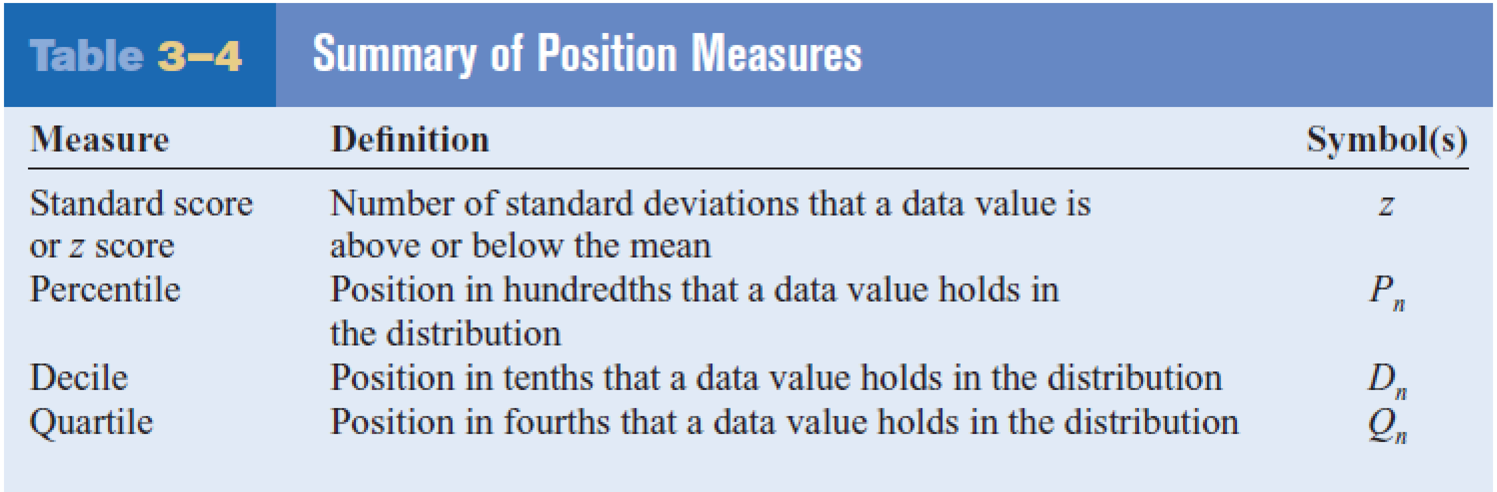
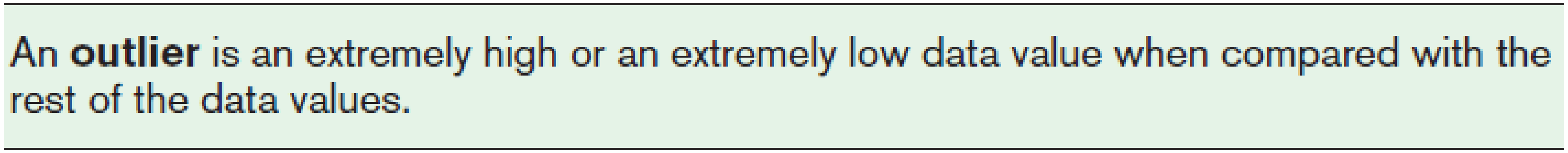


Table 3–4 Summary of Position Measures

Measure	Definition	Symbol(s)
Standard score or z score	Number of standard deviations that a data value is above or below the mean	z
Percentile	Position in hundredths that a data value holds in the distribution	P_n
Decile	Position in tenths that a data value holds in the distribution	D_n
Quartile	Position in fourths that a data value holds in the distribution	Q_n

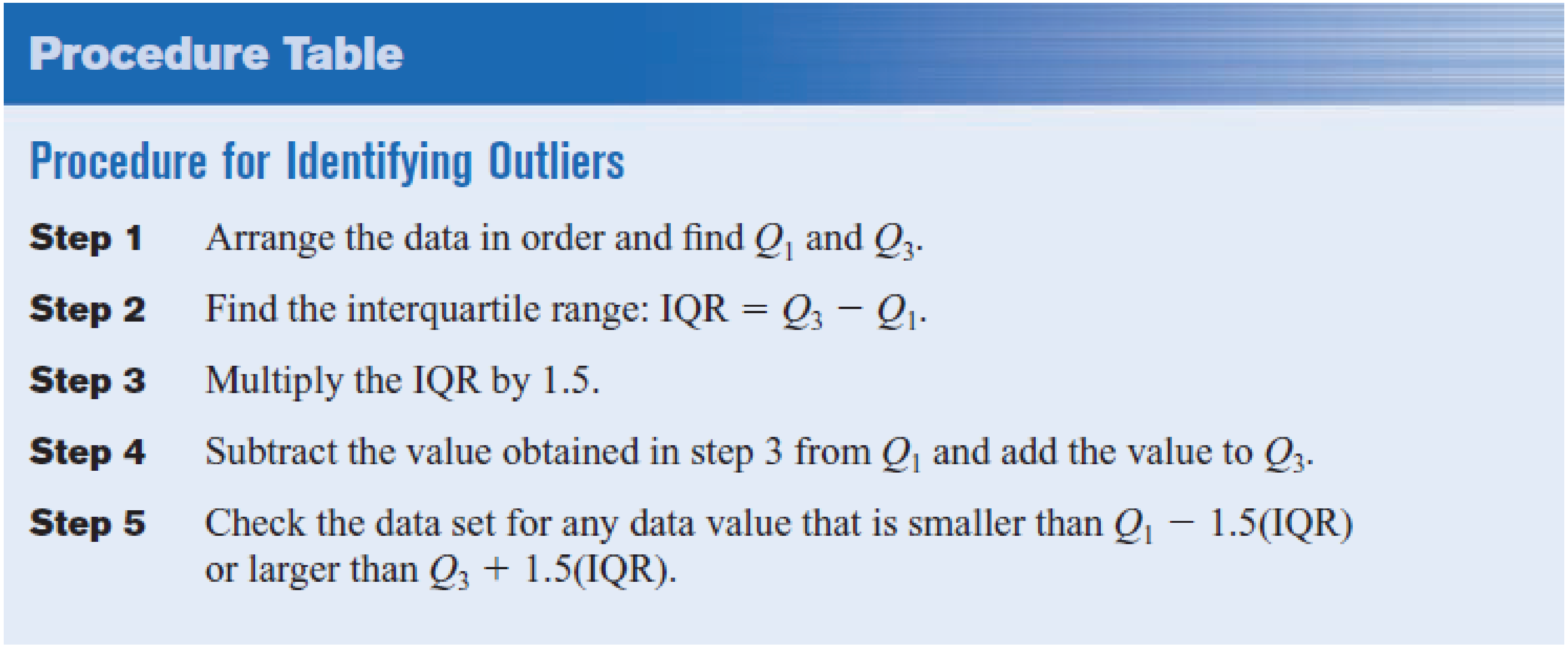
Outliers



An **outlier** is an extremely high or an extremely low data value when compared with the rest of the data values.

An outlier can strongly affect the mean and standard deviation of a variable. For example, suppose a researcher mistakenly recorded an extremely high data value. This value would then make the mean and standard deviation of the variable much larger than they really were. Outliers can have an effect on other statistics as well.

There are several ways to check a data set for outliers. One method is shown in the Procedure Table



Procedure Table

Procedure for Identifying Outliers

Step 1 Arrange the data in order and find Q_1 and Q_3.

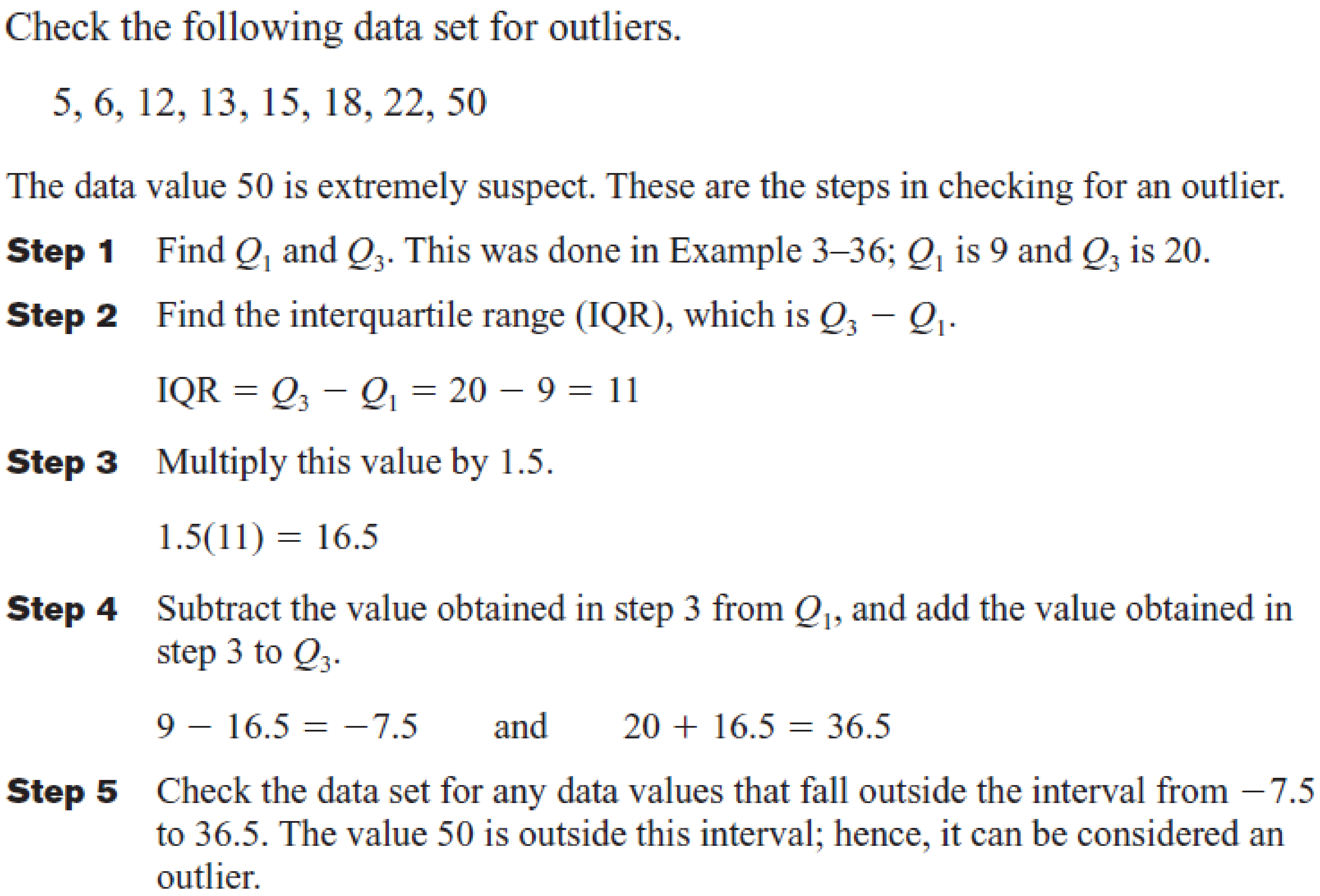
Step 2 Find the interquartile range: IQR = $Q_3 - Q_1$.

Step 3 Multiply the IQR by 1.5.

Step 4 Subtract the value obtained in step 3 from Q_1 and add the value to Q_3.

Step 5 Check the data set for any data value that is smaller than $Q_1 - 1.5(\text{IQR})$ or larger than $Q_3 + 1.5(\text{IQR})$.

Example 18:



Check the following data set for outliers.

5, 6, 12, 13, 15, 18, 22, 50

The data value 50 is extremely suspect. These are the steps in checking for an outlier.

Step 1 Find Q_1 and Q_3. This was done in Example 3–36; Q_1 is 9 and Q_3 is 20.

Step 2 Find the interquartile range (IQR), which is $Q_3 - Q_1$.

$$\text{IQR} = Q_3 - Q_1 = 20 - 9 = 11$$

Step 3 Multiply this value by 1.5.

$$1.5(11) = 16.5$$

Step 4 Subtract the value obtained in step 3 from Q_1, and add the value obtained in step 3 to Q_3.

$$9 - 16.5 = -7.5 \qquad \text{and} \qquad 20 + 16.5 = 36.5$$

Step 5 Check the data set for any data values that fall outside the interval from -7.5 to 36.5. The value 50 is outside this interval; hence, it can be considered an outlier.

Exploratory Data Analysis

In **exploratory data analysis (EDA),** data can be organized using a *stem and leaf plot.* (See Chapter 2.) The measure of central tendency used in EDA is the *median.* The measure of variation used in EDA is the *interquartile range* $Q3$ $Q1$. In EDA the data are represented graphically using a *boxplot* (sometimes called a box-and-whisker plot).

The purpose of exploratory data analysis is to examine data to find out what information can be discovered about the data such as the center and the spread.

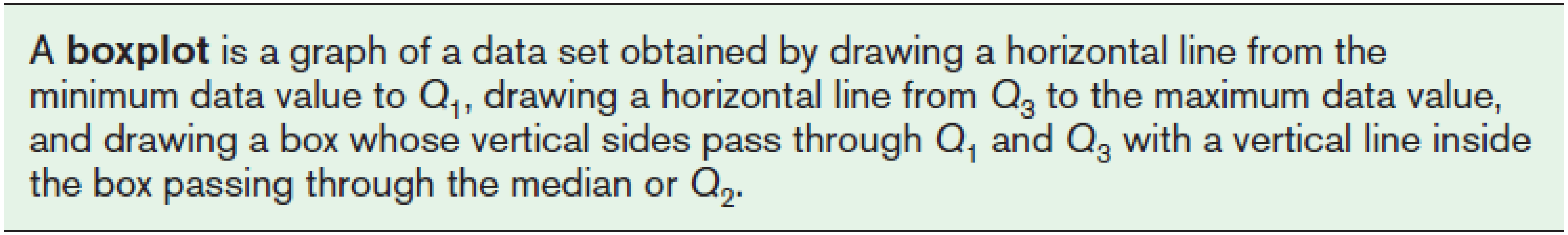
The Five-Number Summary and Boxplots

A **boxplot** can be used to graphically represent the data set. These plots involve five specific values:

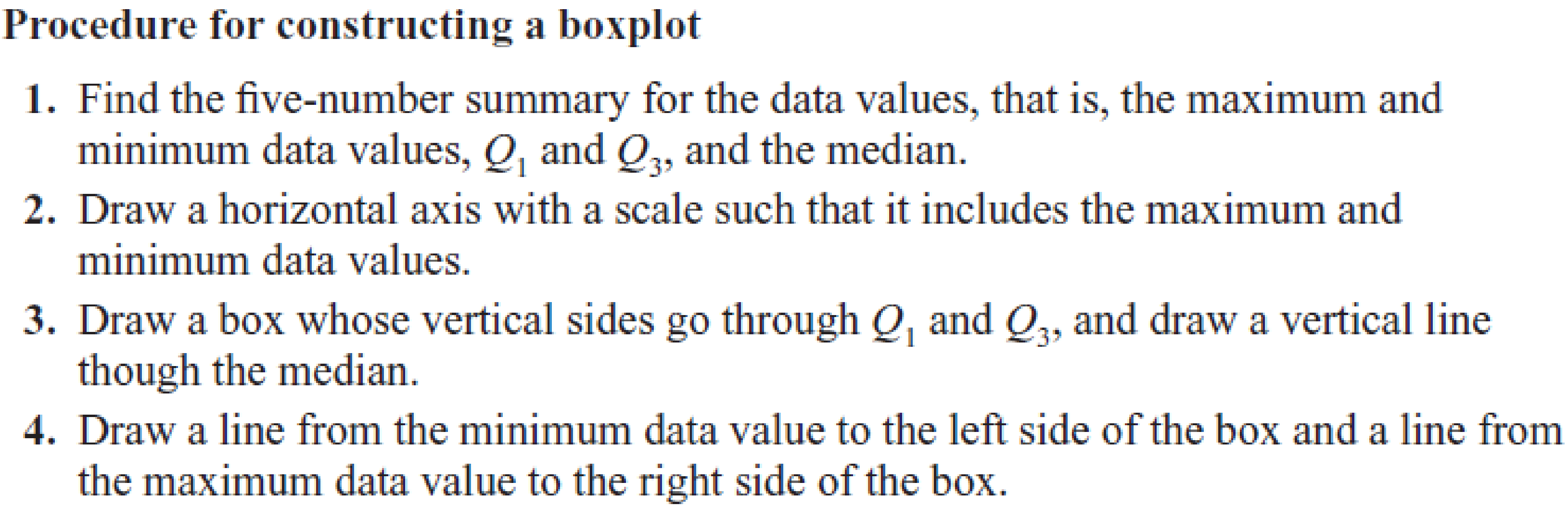
1. The lowest value of the data set (i.e., minimum)
2. Q_1
3. The median
4. Q_3
5. The highest value of the data set (i.e., maximum)

These values are called a **five-number summary** of the data set.

EDA – Boxplots



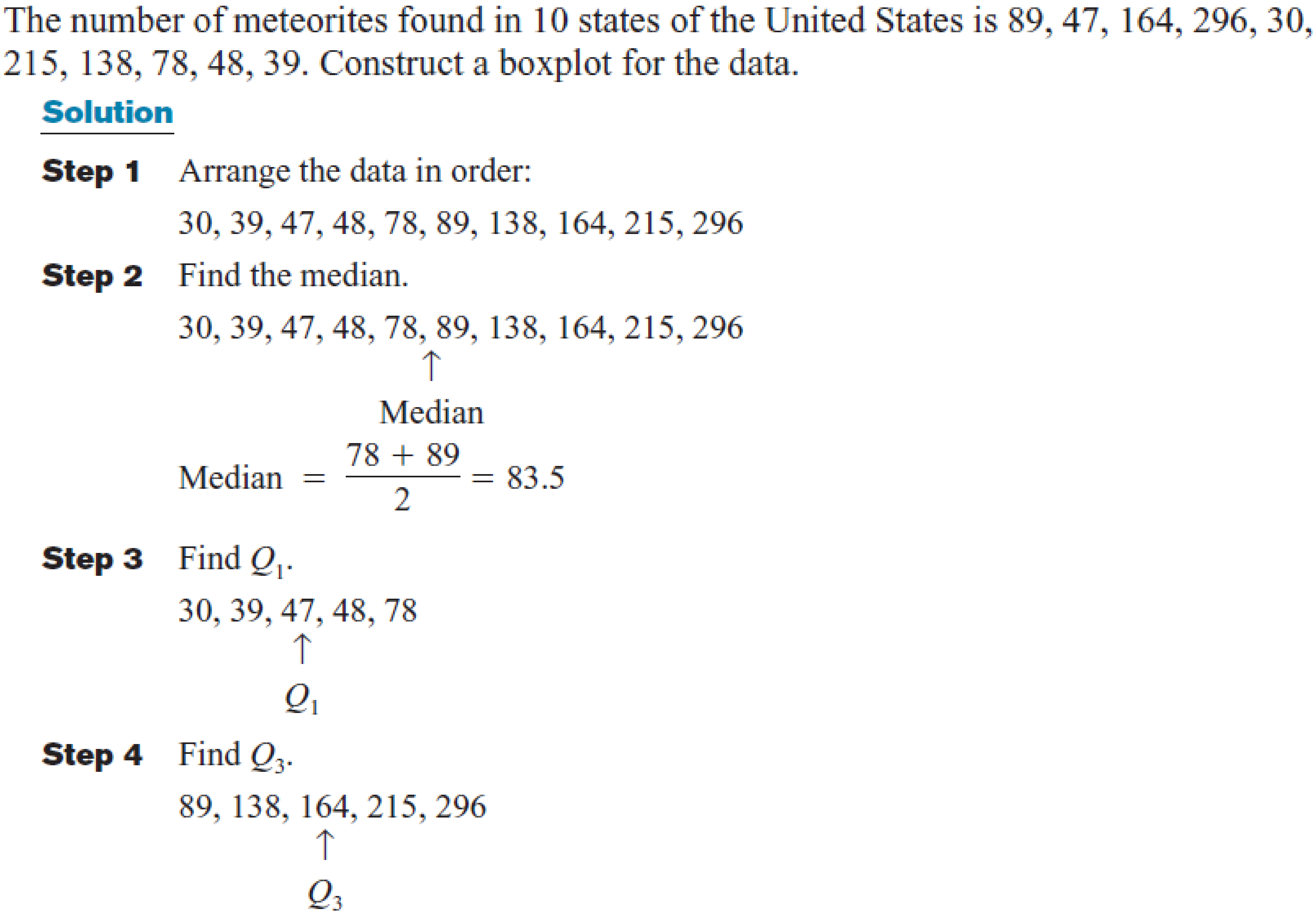
A **boxplot** is a graph of a data set obtained by drawing a horizontal line from the minimum data value to Q_1, drawing a horizontal line from Q_3 to the maximum data value, and drawing a box whose vertical sides pass through Q_1 and Q_3 with a vertical line inside the box passing through the median or Q_2.



Procedure for constructing a boxplot

1. Find the five-number summary for the data values, that is, the maximum and minimum data values, Q_1 and Q_3, and the median.
2. Draw a horizontal axis with a scale such that it includes the maximum and minimum data values.
3. Draw a box whose vertical sides go through Q_1 and Q_3, and draw a vertical line though the median.
4. Draw a line from the minimum data value to the left side of the box and a line from the maximum data value to the right side of the box.

Example 19:



The number of meteorites found in 10 states of the United States is 89, 47, 164, 296, 30, 215, 138, 78, 48, 39. Construct a boxplot for the data.

Solution

Step 1 Arrange the data in order:

30, 39, 47, 48, 78, 89, 138, 164, 215, 296

Step 2 Find the median.

30, 39, 47, 48, 78, 89, 138, 164, 215, 296

↑

Median

$$\text{Median} = \frac{78 + 89}{2} = 83.5$$

Step 3 Find Q_1.

30, 39, 47, 48, 78

↑

Q_1

Step 4 Find Q_3.

89, 138, 164, 215, 296

↑

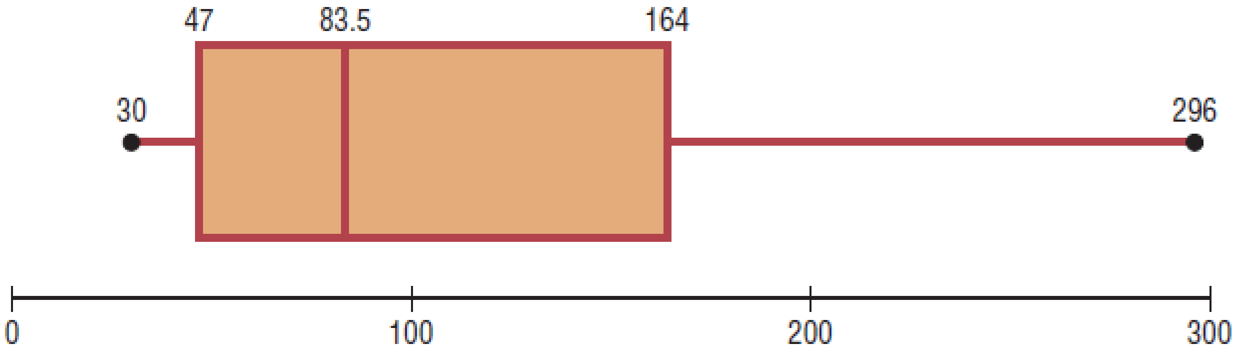
Q_3

Example 19 (cont.):

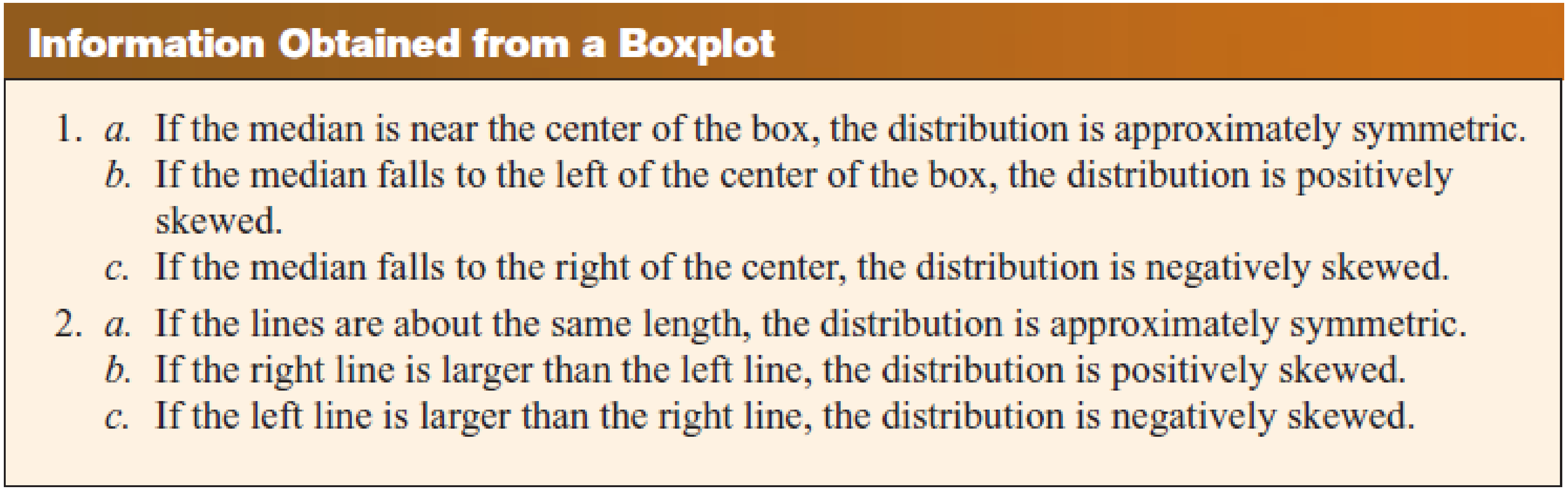
Step 5 Draw a scale for the data on the *x* axis.

Step 6 Located the lowest value, Q_1, median, Q_3, and the highest value on the scale.

Step 7 Draw a box around Q_1 and Q_3, draw a vertical line through the median, and connect the upper value and the lower value to the box. See Figure 3–7.



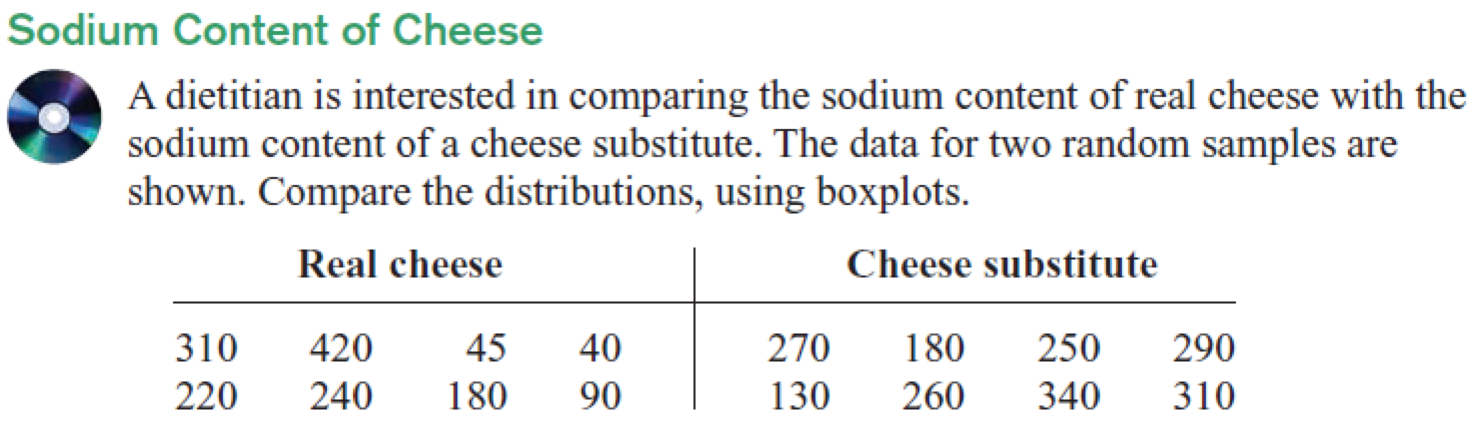
The distribution is somewhat positively skewed.



Information Obtained from a Boxplot

1. *a.* If the median is near the center of the box, the distribution is approximately symmetric.
   b. If the median falls to the left of the center of the box, the distribution is positively skewed.
   c. If the median falls to the right of the center, the distribution is negatively skewed.
2. *a.* If the lines are about the same length, the distribution is approximately symmetric.
   b. If the right line is larger than the left line, the distribution is positively skewed.
   c. If the left line is larger than the right line, the distribution is negatively skewed.

Example 20:

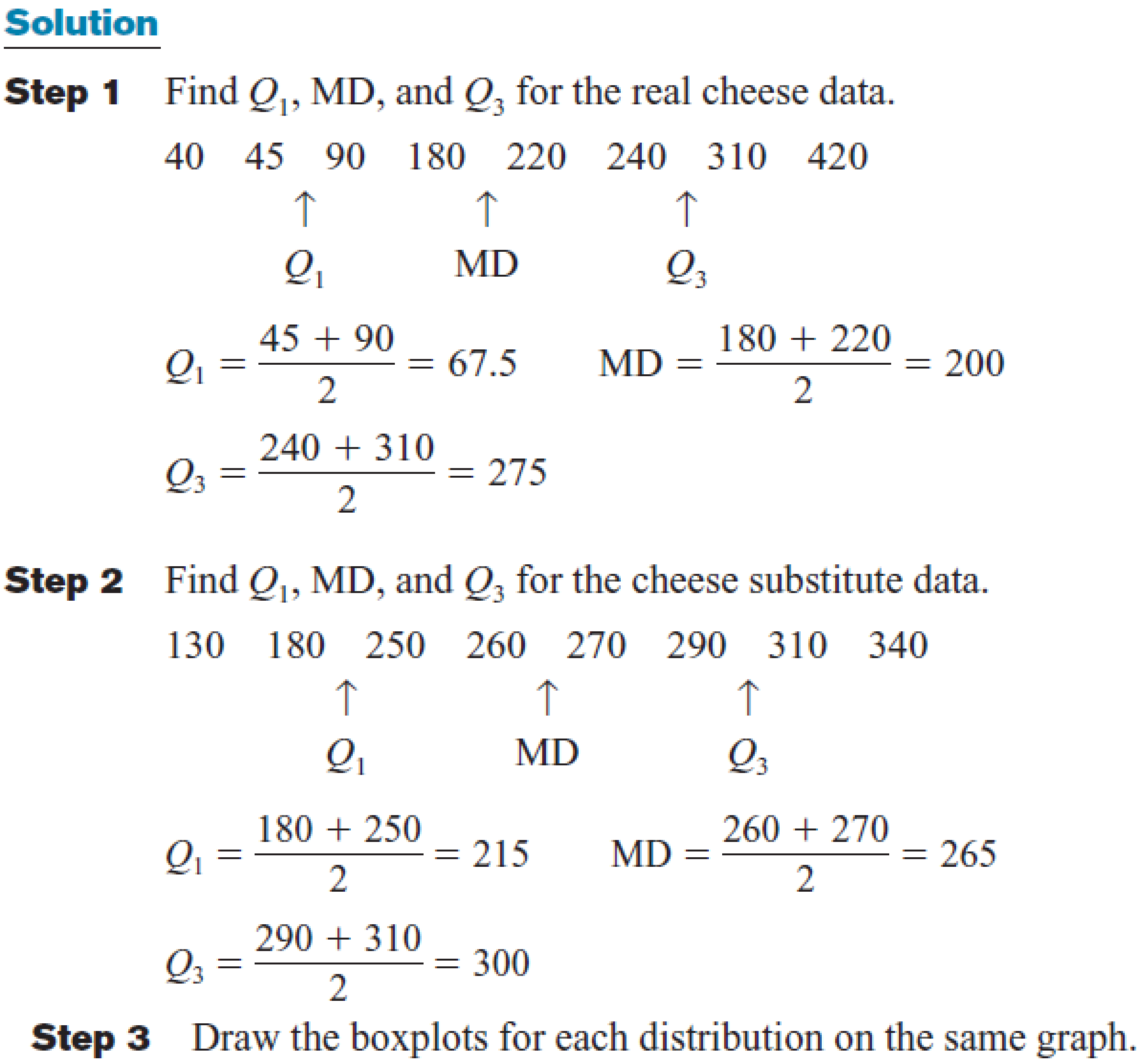


Sodium Content of Cheese

A dietitian is interested in comparing the sodium content of real cheese with the sodium content of a cheese substitute. The data for two random samples are shown. Compare the distributions, using boxplots.

Real cheese				Cheese substitute			
310	420	45	40	270	180	250	290
220	240	180	90	130	260	340	310

Example 20 (cont.):



Solution

Step 1 Find Q_1, MD, and Q_3 for the real cheese data.

$$\begin{array}{ccccccccccc} 40 & 45 & & 90 & 180 & & 220 & 240 & & 310 & 420 \\ & & \uparrow & & & \uparrow & & & \uparrow & & \\ & & Q_1 & & & \text{MD} & & & Q_3 & & \end{array}$$

$$Q_1 = \frac{45 + 90}{2} = 67.5 \qquad \text{MD} = \frac{180 + 220}{2} = 200$$

$$Q_3 = \frac{240 + 310}{2} = 275$$

Step 2 Find Q_1, MD, and Q_3 for the cheese substitute data.

$$\begin{array}{ccccccccccc} 130 & 180 & & 250 & 260 & & 270 & 290 & & 310 & 340 \\ & & \uparrow & & & \uparrow & & & \uparrow & & \\ & & Q_1 & & & \text{MD} & & & Q_3 & & \end{array}$$

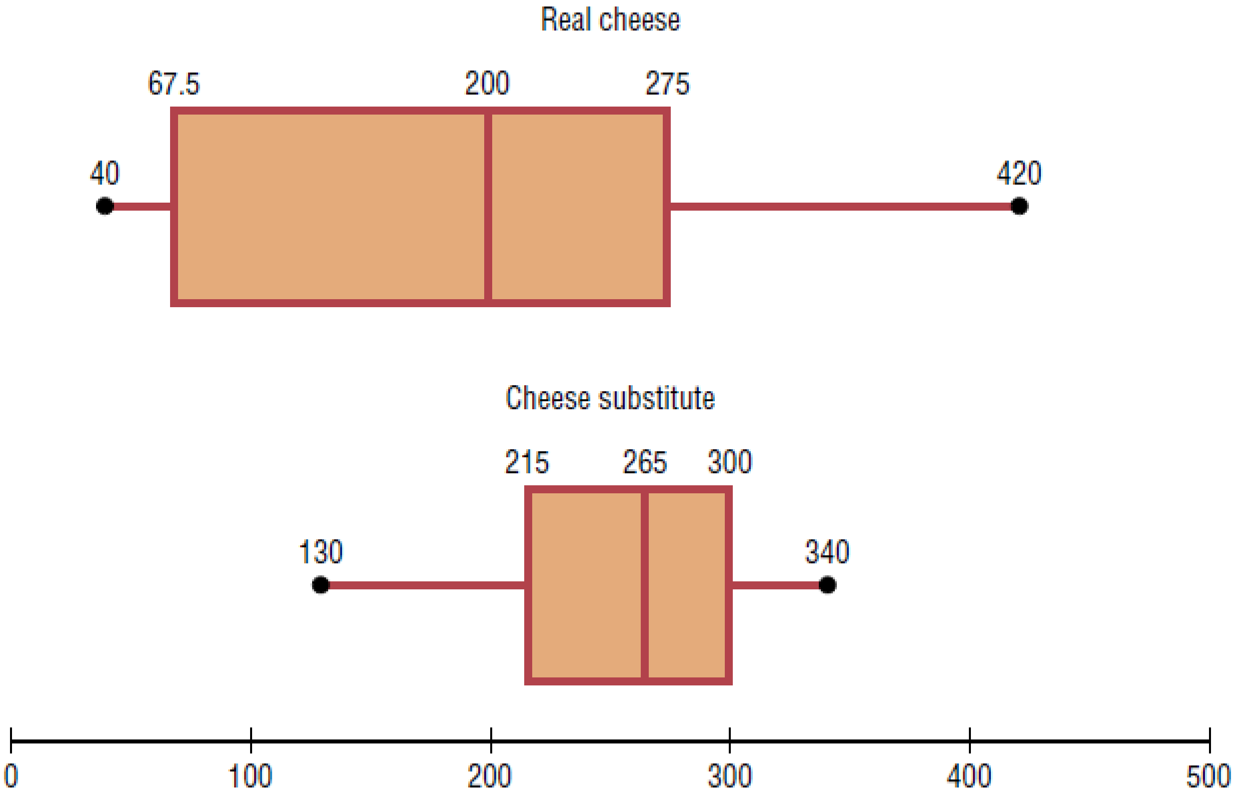
$$Q_1 = \frac{180 + 250}{2} = 215 \qquad \text{MD} = \frac{260 + 270}{2} = 265$$

$$Q_3 = \frac{290 + 310}{2} = 300$$

Step 3 Draw the boxplots for each distribution on the same graph.

Example 20 (cont.):

Step 4 Compare the plots. It is quite apparent that the distribution for the cheese substitute data has a higher median than the median for the distribution for the real cheese data. The variation or spread for the distribution of the real cheese data is larger than the variation for the distribution of the cheese substitute data.



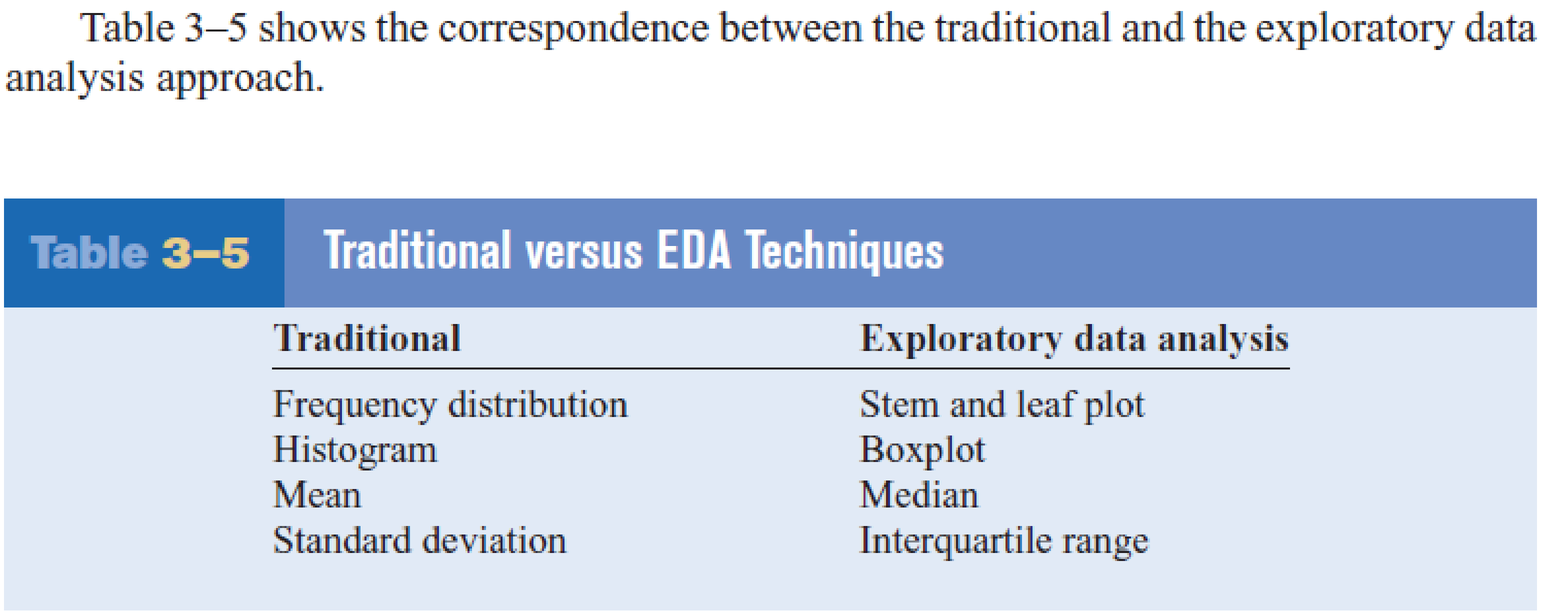


Table 3–5 shows the correspondence between the traditional and the exploratory data analysis approach.

Table 3–5 Traditional versus EDA Techniques

Traditional	Exploratory data analysis
Frequency distribution	Stem and leaf plot
Histogram	Boxplot
Mean	Median
Standard deviation	Interquartile range

Thank You.