This is “Descriptive Statistics”, chapter 2 from the book Beginning Statistics (index.html) (v. 1.0).

Chapter 2

Descriptive Statistics

As described in Chapter 1 "Introduction", statistics naturally divides into two branches, descriptive statistics and inferential statistics. Our main interest is in inferential statistics, as shown in Figure 1.1 "The Grand Picture of Statistics" in Chapter 1 "Introduction". Nevertheless, the starting point for dealing with a collection of data is to organize, display, and summarize it effectively. These are the objectives of descriptive statistics, the topic of this chapter.

2.1 Three Popular Data Displays

LEARNING OBJECTIVE

1. To learn to interpret the meaning of three graphical representations of sets of data: stem and leaf diagrams, frequency histograms, and relative frequency histograms.

A well-known adage is that "a picture is worth a thousand words." This saying proves true when it comes to presenting statistical information in a data set. There are many effective ways to present data graphically. The three graphical tools that are introduced in this section are among the most commonly used and are relevant to the subsequent presentation of the material in this book.

Stem and Leaf Diagrams

Suppose 30 students in a statistics class took a test and made the following scores:

86	80	25	77	73	76	100	90	69	93
90	83	70	73	73	70	90	83	71	95
40	58	68	69	100	78	87	97	92	74

How did the class do on the test? A quick glance at the set of 30 numbers does not immediately give a clear answer. However the data set may be reorganized and rewritten to make relevant information more visible. One way to do so is to construct a **stem and leaf** diagram as shown in Figure 2.1 "Stem and Leaf Diagram". The numbers in the tens place, from 2 through 9, and additionally the number 10, are the "stems," and are arranged in numerical order from top to bottom to the left of a vertical line. The number in the units place in each measurement is a "leaf," and is placed in a row to the right of the corresponding stem, the number in the tens place of that measurement. Thus the three leaves 9, 8, and 9 in the row headed with the stem 6 correspond to the three exam scores in the 60s, 69 (in the first row of data), 68 (in the third row), and 69 (also in the third row). The display is made even more useful for some purposes by rearranging the leaves in numerical order, as shown in Figure 2.2 "Ordered Stem and Leaf Diagram". Either way, with the data reorganized certain information of interest becomes apparent immediately. There are two perfect scores; three students made scores under 60; most students scored

in the 70s, 80s and 90s; and the overall average is probably in the high 70s or low 80s.

Figure 2.1 *Stem and Leaf Diagram*

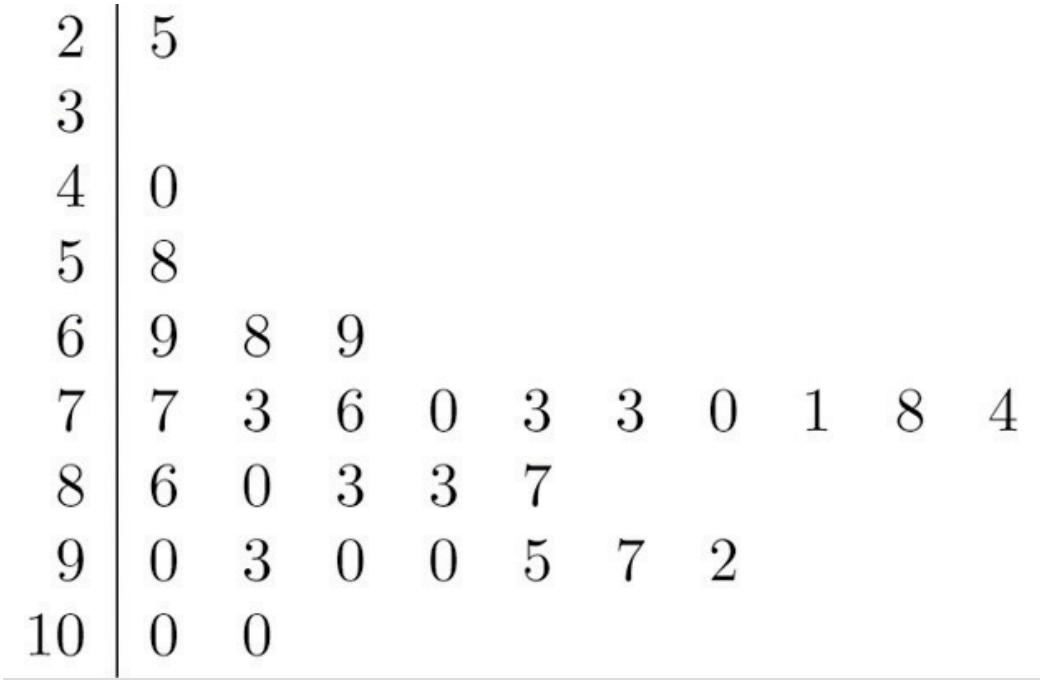
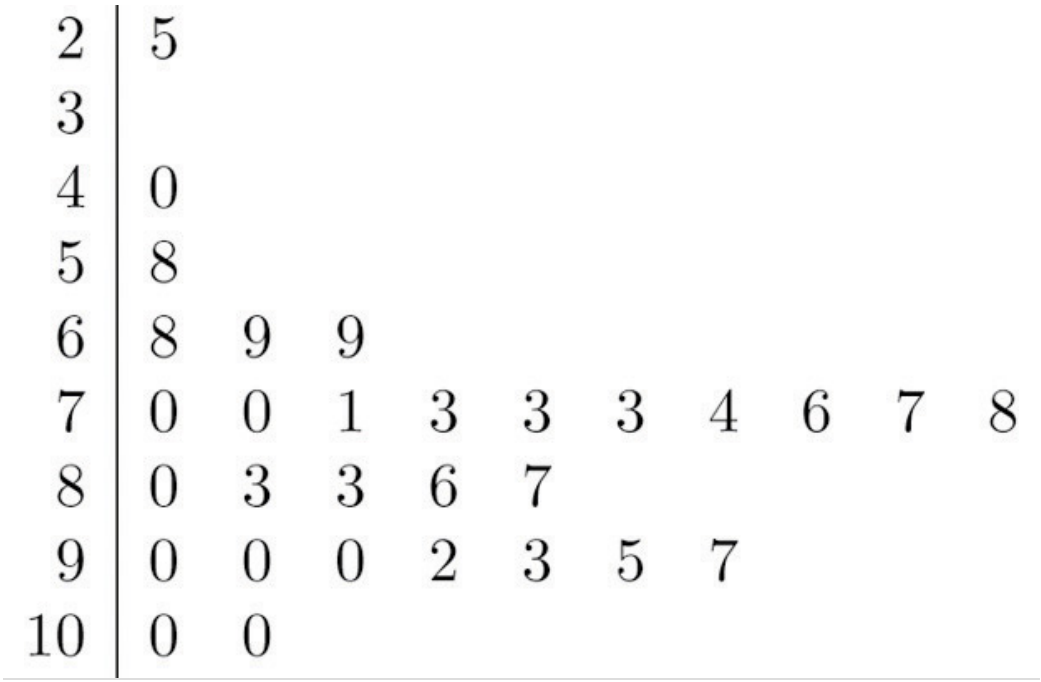


Figure 2.2 *Ordered Stem and Leaf Diagram*



In this example the scores have a natural stem (the tens place) and leaf (the ones place). One could spread the diagram out by splitting each tens place number into lower and upper categories. For example, all the scores in the 80s may be represented on two separate stems, lower 80s and upper 80s:

8	0 3 3
8	6 7

The definitions of stems and leaves are flexible in practice. The general purpose of a stem and leaf diagram is to provide a quick display of how the data are distributed across the range of their values; some improvisation could be necessary to obtain a diagram that best meets that goal.

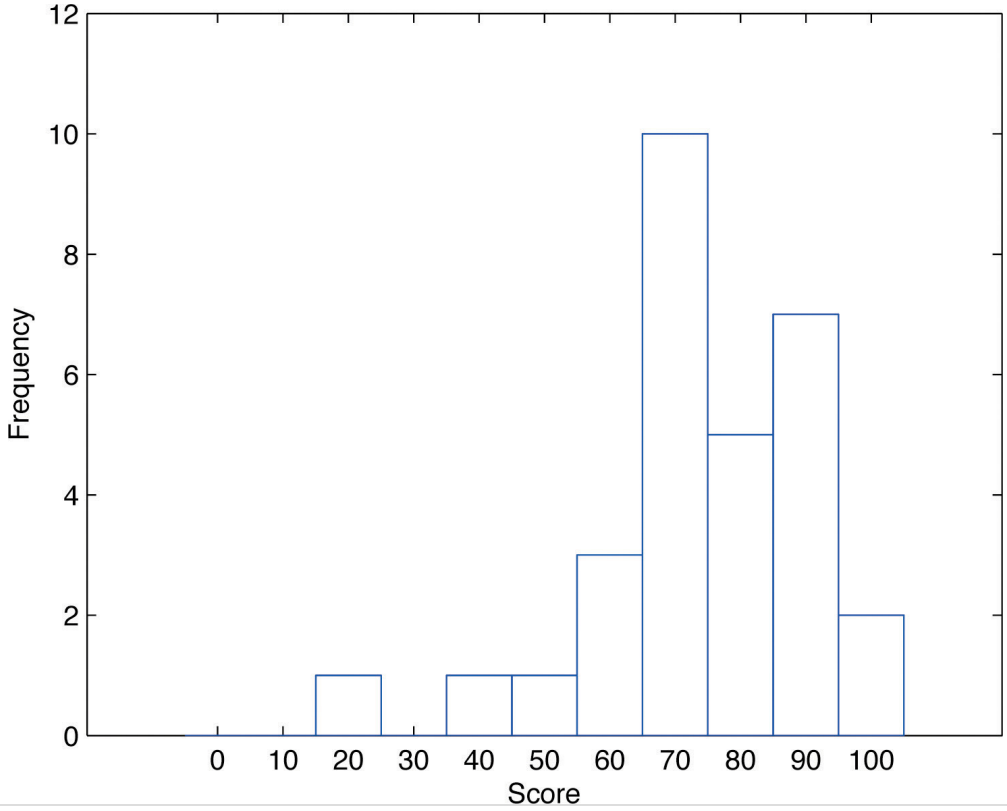
Note that all of the original data can be recovered from the stem and leaf diagram. This will not be true in the next two types of graphical displays.

Frequency Histograms

The stem and leaf diagram is not practical for large data sets, so we need a different, purely graphical way to represent data. A **frequency histogram**[1] is such a device. We will illustrate it using the same data set from the previous subsection. For the 30 scores on the exam, it is natural to group the scores on the standard ten-point scale, and count the number of scores in each group. Thus there are two 100s, seven scores in the 90s, six in the 80s, and so on. We then construct the diagram shown in Figure 2.3 "Frequency Histogram" by drawing for each group, or class, a vertical bar whose length is the number of observations in that group. In our example, the bar labeled 100 is 2 units long, the bar labeled 90 is 7 units long, and so on. While the individual data values are lost, we know the number in each class. This number is called the **frequency**[2] of the class, hence the name frequency histogram.

1. A graphical device showing how data are distributed across the range of their values by collecting them into classes and indicating the number of measurements in each class.
2. Of a class of measurements, the number of measurements in the data set that are in the class.

Figure 2.3 *Frequency Histogram*



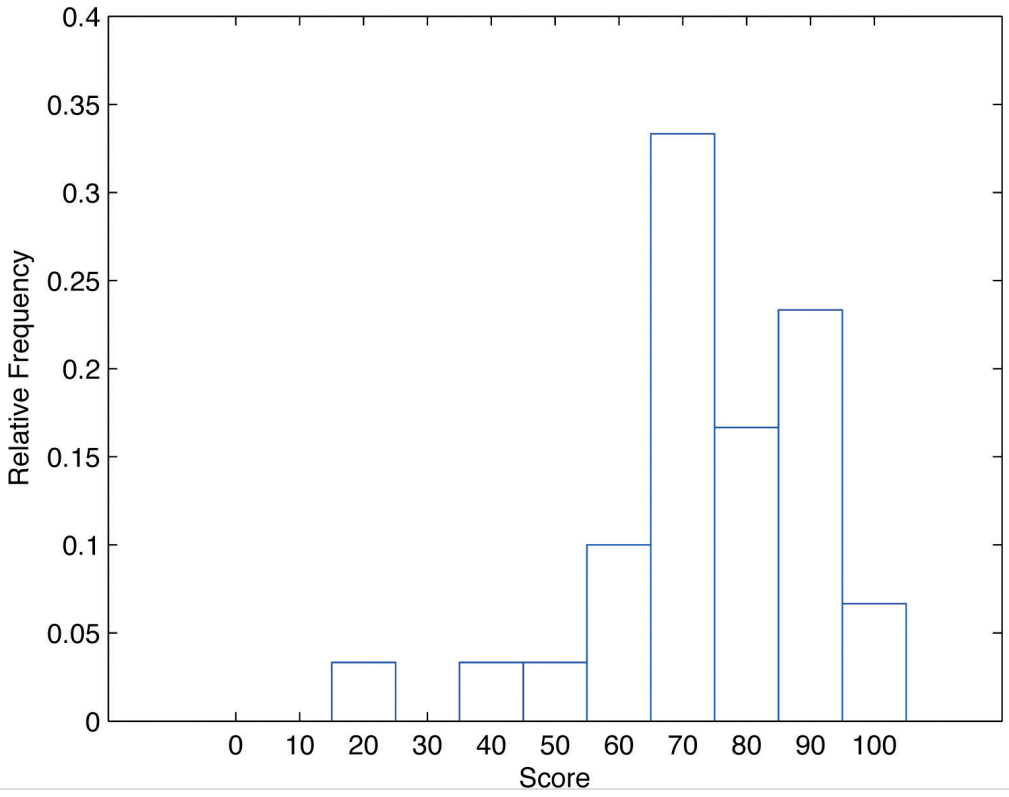
The same procedure can be applied to any collection of numerical data. Observations are grouped into several classes and the frequency (the number of observations) of each class is noted. These classes are arranged and indicated in order on the horizontal axis (called the *x*-axis), and for each group a vertical bar, whose length is the number of observations in that group, is drawn. The resulting display is a frequency histogram for the data. The similarity in Figure 2.1 "Stem and Leaf Diagram" and Figure 2.3 "Frequency Histogram" is apparent, particularly if you imagine turning the stem and leaf diagram on its side by rotating it a quarter turn counterclockwise.

In general, the definition of the classes in the frequency histogram is flexible. The general purpose of a frequency histogram is very much the same as that of a stem and leaf diagram, to provide a graphical display that gives a sense of data distribution across the range of values that appear. We will not discuss the process of constructing a histogram from data since in actual practice it is done automatically with statistical software or even handheld calculators.

Relative Frequency Histograms

In our example of the exam scores in a statistics class, five students scored in the 80s. The number 5 is the *frequency* of the group labeled "80s." Since there are 30 students in the entire statistics class, the proportion who scored in the 80s is 5/30. The number 5/30, which could also be expressed as $0.1\overline{6} \approx .1667$, or as 16.67%, is the **relative frequency**[3] of the group labeled "80s." Every group (the 70s, the 80s, and so on) has a relative frequency. We can thus construct a diagram by drawing for each group, or class, a vertical bar whose length is the relative frequency of that group. For example, the bar for the 80s will have length 5/30 unit, not 5 units. The diagram is a **relative frequency histogram**[4] for the data, and is shown in Figure 2.4 "Relative Frequency Histogram". It is exactly the same as the frequency histogram except that the vertical axis in the relative frequency histogram is not frequency but relative frequency.

Figure 2.4 *Relative Frequency Histogram*



3. Of a class of measurements, the proportion of all measurements in the data set that are in the class.

4. A graphical device showing how data are distributed across the range of their values by collecting them into classes and indicating the proportion of measurements in each class.

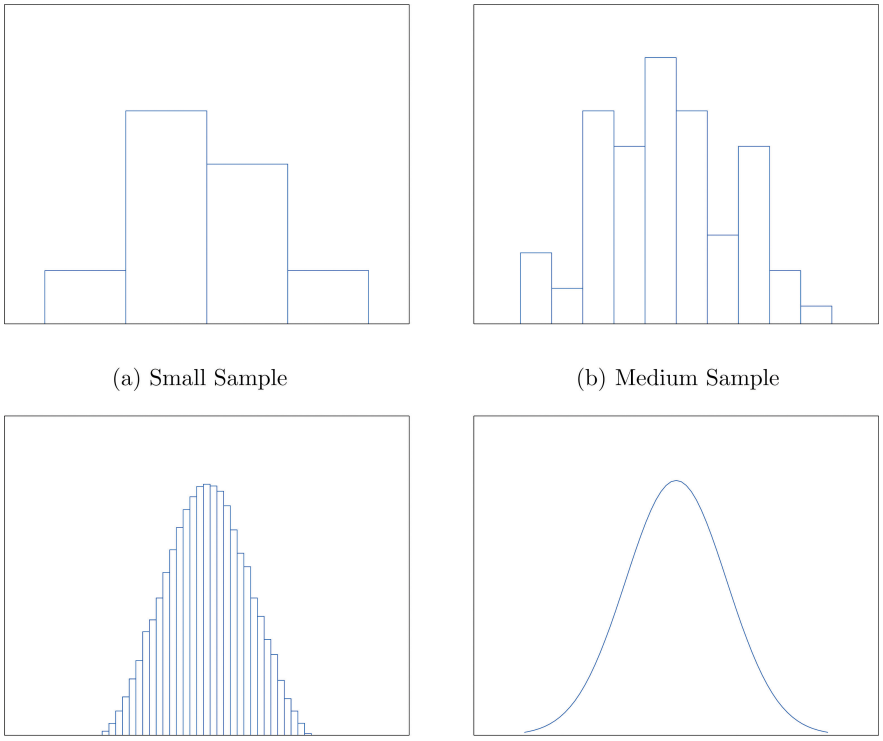
The same procedure can be applied to any collection of numerical data. Classes are selected, the relative frequency of each class is noted, the classes are arranged and indicated in order on the horizontal axis, and for each class a vertical bar, whose length is the relative frequency of the class, is drawn. The resulting display is a

relative frequency histogram for the data. A key point is that now if each vertical bar has width 1 unit, then the total area of all the bars is 1 or 100%.

Although the histograms in Figure 2.3 "Frequency Histogram" and Figure 2.4 "Relative Frequency Histogram" have the same appearance, the relative frequency histogram is more important for us, and it will be relative frequency histograms that will be used repeatedly to represent data in this text. To see why this is so, reflect on what it is that you are actually seeing in the diagrams that quickly and effectively communicates information to you about the data. It is the *relative sizes* of the bars. The bar labeled “70s” in either figure takes up 1/3 of the total area of all the bars, and although we may not think of this consciously, we perceive the proportion 1/3 in the figures, indicating that a third of the grades were in the 70s. The relative frequency histogram is important because the labeling on the vertical axis reflects what is important visually: the relative sizes of the bars.

When the size n of a sample is small only a few classes can be used in constructing a relative frequency histogram. Such a histogram might look something like the one in panel (a) of Figure 2.5 "Sample Size and Relative Frequency Histograms". If the sample size n were increased, then more classes could be used in constructing a relative frequency histogram and the vertical bars of the resulting histogram would be finer, as indicated in panel (b) of Figure 2.5 "Sample Size and Relative Frequency Histograms". For a very large sample the relative frequency histogram would look very fine, like the one in (c) of Figure 2.5 "Sample Size and Relative Frequency Histograms". If the sample size were to increase indefinitely then the corresponding relative frequency histogram would be so fine that it would look like a smooth curve, such as the one in panel (d) of Figure 2.5 "Sample Size and Relative Frequency Histograms".

Figure 2.5 *Sample Size and Relative Frequency Histograms*



(a) Small Sample

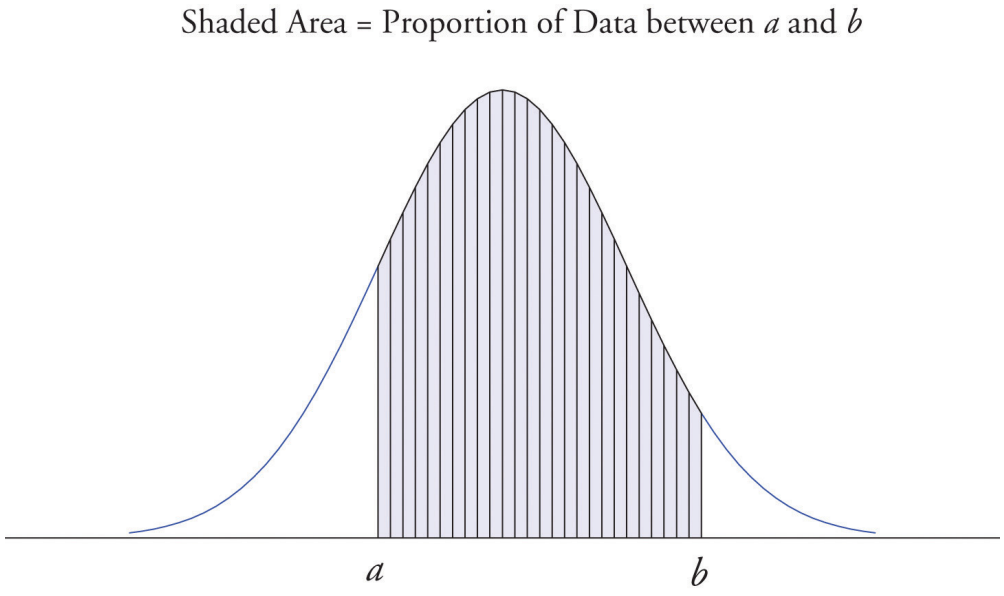
(b) Medium Sample

(c) Large Sample

(d) Very Large Sample

It is common in statistics to represent a population or a very large data set by a smooth curve. It is good to keep in mind that such a curve is actually just a very fine relative frequency histogram in which the exceedingly narrow vertical bars have disappeared. Because the area of each such vertical bar is the proportion of the data that lies in the interval of numbers over which that bar stands, this means that for any two numbers a and b, the proportion of the data that lies between the two numbers a and b is the area under the curve that is above the interval (a,b) in the horizontal axis. This is the area shown in Figure 2.6 "A Very Fine Relative Frequency Histogram". In particular the total area under the curve is 1, or 100%.

Figure 2.6 *A Very Fine Relative Frequency Histogram*



KEY TAKEAWAYS

- Graphical representations of large data sets provide a quick overview of the nature of the data.
- A population or a very large data set may be represented by a smooth curve. This curve is a very fine relative frequency histogram in which the exceedingly narrow vertical bars have been omitted.
- When a curve derived from a relative frequency histogram is used to describe a data set, the proportion of data with values between two numbers a and b is the area under the curve between a and b, as illustrated in Figure 2.6 "A Very Fine Relative Frequency Histogram".

EXERCISES

BASIC

1. Describe one difference between a frequency histogram and a relative frequency histogram.

2. Describe one advantage of a stem and leaf diagram over a frequency histogram.

3. Construct a stem and leaf diagram, a frequency histogram, and a relative frequency histogram for the following data set. For the histograms use classes 51–60, 61–70, and so on.

69 92 68 77 80
70 85 88 85 96

93 75 76 82 100
53 70 70 82 85

4. Construct a stem and leaf diagram, a frequency histogram, and a relative frequency histogram for the following data set. For the histograms use classes 6.0–6.9, 7.0–7.9, and so on.

8.5 8.2 7.0 7.0 4.9
6.5 8.2 7.6 1.5 9.3

9.6 8.5 8.8 8.5 8.7
8.0 7.7 2.9 9.2 6.9

5. A data set contains $n = 10$ observations. The values x and their frequencies f are summarized in the following data frequency table.

x	−1	0	1	2
f	3	4	2	1

Construct a frequency histogram and a relative frequency histogram for the data set.

6. A data set contains the $n = 20$ observations The values x and their frequencies f are summarized in the following data frequency table.

x	-1	0	1	2
f	3	a	2	1

The frequency of the value 0 is missing. Find a and then sketch a frequency histogram and a relative frequency histogram for the data set.

7. A data set has the following frequency distribution table:

x	1	2	3	4
f	3	a	2	1

The number a is unknown. Can you construct a frequency histogram? If so, construct it. If not, say why not.

8. A table of some of the relative frequencies computed from a data set is

x	1	2	3	4
f / n	0.3	p	0.2	0.1

The number p is yet to be computed. Finish the table and construct the relative frequency histogram for the data set.

APPLICATIONS

9. The IQ scores of ten students randomly selected from an elementary school are given.

108 100 99 125 87
105 107 105 119 118

Grouping the measures in the 80s, the 90s, and so on, construct a stem and leaf diagram, a frequency histogram, and a relative frequency histogram.

10. The IQ scores of ten students randomly selected from an elementary school for academically gifted students are given.

133 140 152 142 137
145 160 138 139 138

Grouping the measures by their common hundreds and tens digits, construct a stem and leaf diagram, a frequency histogram, and a relative frequency histogram.

11. During a one-day blood drive 300 people donated blood at a mobile donation center. The blood types of these 300 donors are summarized in the table.

Blood Type	*O*	*A*	*B*	*AB*
Frequency	136	120	32	12

Construct a relative frequency histogram for the data set.

12. In a particular kitchen appliance store an electric automatic rice cooker is a popular item. The weekly sales for the last 20 weeks are shown.

20 15 14 14 18
15 17 16 16 18

15 19 12 13 9
19 15 15 16 15

Construct a relative frequency histogram with classes 6–10, 11–15, and 16–20.

ADDITIONAL EXERCISES

13. Random samples, each of size $n = 10$, were taken of the lengths in centimeters of three kinds of commercial fish, with the following results:

Sample 1:	108	100	99	125	87
	105	107	105	119	118
Sample 2:	133	140	152	142	137
	145	160	138	139	138
Sample 3:	82	60	83	82	82
	74	79	82	80	80

Grouping the measures by their common hundreds and tens digits, construct a stem and leaf diagram, a frequency histogram, and a relative frequency histogram for each of the samples. Compare the histograms and describe any patterns they exhibit.

14. During a one-day blood drive 300 people donated blood at a mobile donation center. The blood types of these 300 donors are summarized below.

Blood Type	*O*	*A*	*B*	*AB*
Frequency	136	120	32	12

Identify the blood type that has the highest relative frequency for these 300 people. Can you conclude that the blood type you identified is also most common for all people in the population at large? Explain.

15. In a particular kitchen appliance store, the weekly sales of an electric automatic rice cooker for the last 20 weeks are as follows.

20 15 14 14 18
15 17 16 16 18

15 19 12 13 9
19 15 15 16 15

In retail sales, too large an inventory ties up capital, while too small an inventory costs lost sales and customer satisfaction. Using the relative frequency histogram for these data, find approximately how many rice cookers must be in stock at the beginning of each week if

a. the store is not to run out of stock by the end of a week for more than 15% of the weeks; and
b. the store is not to run out of stock by the end of a week for more than 5% of the weeks.

ANSWERS

1. The vertical scale on one is the frequencies and on the other is the relative frequencies.

3.

5	3
6	8 9
7	0 0 0 5 6 7
8	0 2 3 5 5 5 8
9	2 3 6
10	0

Frequency and relative frequency histograms are similarly generated.

5. Noting that $n = 10$ the relative frequency table is:

x	-1	0	1	2
f / n	0.3	0.4	0.2	0.1

7. Since n is unknown, a is unknown, so the histogram cannot be constructed.

9.

8	7
9	9
10	0 5 5 7 8
11	8 9
12	5

Frequency and relative frequency histograms are similarly generated.

11. Noting $n = 300$, the relative frequency table is therefore:

Blood Type	O	A	B	AB
f / n	0.4533	0.4	0.1067	0.04

A relative frequency histogram is then generated.

13. The stem and leaf diagrams listed for Samples 1, 2, and 3 in that order.

6	
7	
8	7
9	9
10	0 5 5 7 8
11	8 9
12	5
13	
14	
15	
16	

6	
7	
8	
9	
10	
11	
12	
13	3 7 8 8 9
14	0 2 5
15	2
16	0

6	0
7	4 9
8	0 0 2 2 2 2 3
9	
10	
11	
12	
13	
14	
15	
16	

The frequency tables are given below in the same order.

Length	80 ~ 89	90 ~ 99	100 ~ 109
f	1	1	5

Length	110 ~ 119	120 ~ 129
f	2	1

Length	130 ~ 139	140 ~ 149	150 ~ 159
f	5	3	1

Length	160 ~ 169
f	1

Length	60 ~ 69	70 ~ 79	80 ~ 89
f	1	2	7

The relative frequency tables are given below in the same order.

Length	80 ~ 89	90 ~ 99	100 ~ 109
$f\ /\ n$	0.1	0.1	0.5

Length	110 ~ 119	120 ~ 129
$f\ /\ n$	0.2	0.1

Length	130 ~ 139	140 ~ 149	150 ~ 159
f / n	0.5	0.3	0.1

Length	160 ~ 169
f / n	0.1

Length	60 ~ 69	70 ~ 79	80 ~ 89
f / n	0.1	0.2	0.7

15. a. 19.
    b. 20.

2.2 Measures of Central Location

LEARNING OBJECTIVES

1. To learn the concept of the "center" of a data set.
2. To learn the meaning of each of three measures of the center of a data set—the mean, the median, and the mode—and how to compute each one.

This section could be titled "three kinds of averages of a data set." Any kind of "average" is meant to be an answer to the question "Where do the data center?" It is thus a measure of the central location of the data set. We will see that the nature of the data set, as indicated by a relative frequency histogram, will determine what constitutes a good answer. Different shapes of the histogram call for different measures of central location.

The Mean

The first measure of central location is the usual "average" that is familiar to everyone. In the formula in the following definition we introduce the standard summation notation Σ, where Σ is the capital Greek letter sigma. In general, the notation Σ followed by a second mathematical symbol means to add up all the values that the second symbol can take in the context of the problem. Here is an example to illustrate this.

EXAMPLE 1

Find Σx, Σx^2, and $\Sigma(x-1)^2$ for the data set

$$1 \quad 3 \quad 4$$

Solution:

$$\begin{aligned} \Sigma x &= 1+3+4=8 \\ \Sigma x^2 &= 1^2+3^2+4^2=1+9+16=26 \\ \Sigma(x-1)^2 &= (1-1)^2+(3-1)^2+(4-1)^2=0^2+2^2+3^2=13 \end{aligned}$$

In the definition we follow the convention of using lowercase n to denote the number of measurements in a sample, which is called the **sample size**.

Definition

The **sample mean**[5] *of a set of n sample data is the number* $\overline{x}$ *defined by the formula*

$$\overline{x} = \frac{\Sigma x}{n}$$

EXAMPLE 2

Find the mean of the sample data

$$2 \quad -1 \quad 0 \quad 2$$

Solution:

$$\overline{x} = \frac{\Sigma x}{n} = \frac{2+(-1)+0+2}{4} = \frac{3}{4} = 0.75$$

5. The familiar average of a sample data set.

EXAMPLE 3

A random sample of ten students is taken from the student body of a college and their GPAs are recorded as follows.

1.90 3.00 2.53 3.71 2.12 1.76 2.71 1.39 4.00 3.33

Find the sample mean.

Solution:

$$\overline{x} = \frac{\Sigma x}{n} = \frac{1.90 + 3.00 + 2.53 + 3.71 + 2.12 + 1.76 + 2.71 + 1.39 + 4.}{10}$$

$$= \frac{26.45}{10} = 2.645$$

EXAMPLE 4

A random sample of 19 women beyond child-bearing age gave the following data, where x is the number of children and f is the *frequency* of that value, the number of times it occurred in the data set.

x	0	1	2	3	4
f	3	6	6	3	1

Find the sample mean.

Solution:

In this example the data are presented by means of a data frequency table, introduced in Chapter 1 "Introduction". Each number in the first line of the table is a number that appears in the data set; the number below it is how many times it occurs. Thus the value 0 is observed three times, that is, three of the measurements in the data set are 0, the value 1 is observed six times, and so on. In the context of the problem this means that three women in the sample have had no children, six have had exactly one child, and so on. The explicit list of all the observations in this data set is therefore

$$0 \quad 0 \quad 0 \quad 1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 2 \quad 2 \quad 2 \quad 2 \quad 2 \quad 2 \quad 3 \quad 3 \quad 3 \quad 4$$

The sample size can be read directly from the table, without first listing the entire data set, as the sum of the frequencies: $n = 3 + 6 + 6 + 3 + 1 = 19$. The sample mean can be computed directly from the table as well:

$$\overline{x} = \frac{\Sigma x}{n} = \frac{0 \times 3 + 1 \times 6 + 2 \times 6 + 3 \times 3 + 4 \times 1}{19} = \frac{31}{19} = 1.6316$$

In the examples above the data sets were described as samples. Therefore the means were sample means, denoted by $\overline{x}$. If the data come from a census, so that there is a measurement for every element of the population, then the mean is calculated by exactly the same process of summing all the measurements and dividing by how many of them there are, but it is now the *population mean* and is denoted by μ, the lower case Greek letter mu.

Definition

The **population mean**[6] *of a set of N population data is the number μ defined by the formula*

$$\mu = \frac{\Sigma x}{N}$$

The mean of two numbers is the number that is halfway between them. For example, the average of the numbers 5 and 17 is $(5 + 17)/2 = 11$, which is 6 units above 5 and 6 units below 17. In this sense the average 11 is the "center" of the data set {5,17}. For larger data sets the mean can similarly be regarded as the "center" of the data.

The Median

To see why another concept of average is needed, consider the following situation. Suppose we are interested in the average yearly income of employees at a large corporation. We take a random sample of seven employees, obtaining the sample data (rounded to the nearest hundred dollars, and expressed in thousands of dollars).

$$24.8 \quad 22.8 \quad 24.6 \quad 192.5 \quad 25.2 \quad 18.5 \quad 23.7$$

The mean (rounded to one decimal place) is $\overline{x} = 47.4$, but the statement "the average income of employees at this corporation is \$47,400" is surely misleading. It is approximately twice what six of the seven employees in the sample make and is nowhere near what any of them makes. It is easy to see what went wrong: the presence of the one executive in the sample, whose salary is so large compared to everyone else's, caused the numerator in the formula for the sample mean to be far too large, pulling the mean far to the right of where we think that the average "ought" to be, namely around \$24,000 or \$25,000. The number 192.5 in our data set is called an **outlier**, a number that is far removed from most or all of the remaining measurements. Many times an outlier is the result of some sort of error, but not always, as is the case here. We would get a better measure of the "center" of the data if we were to arrange the data in numerical order,

$$18.5 \quad 22.8 \quad 23.7 \quad 24.6 \quad 24.8 \quad 25.2 \quad 192.5$$

6. The familiar average of a population data set.

then select the middle number in the list, in this case 24.6. The result is called the *median* of the data set, and has the property that roughly half of the measurements are larger than it is, and roughly half are smaller. In this sense it locates the center of the data. If there are an even number of measurements in the data set, then there will be two middle elements when all are lined up in order, so we take the mean of the middle two as the median. Thus we have the following definition.

Definition

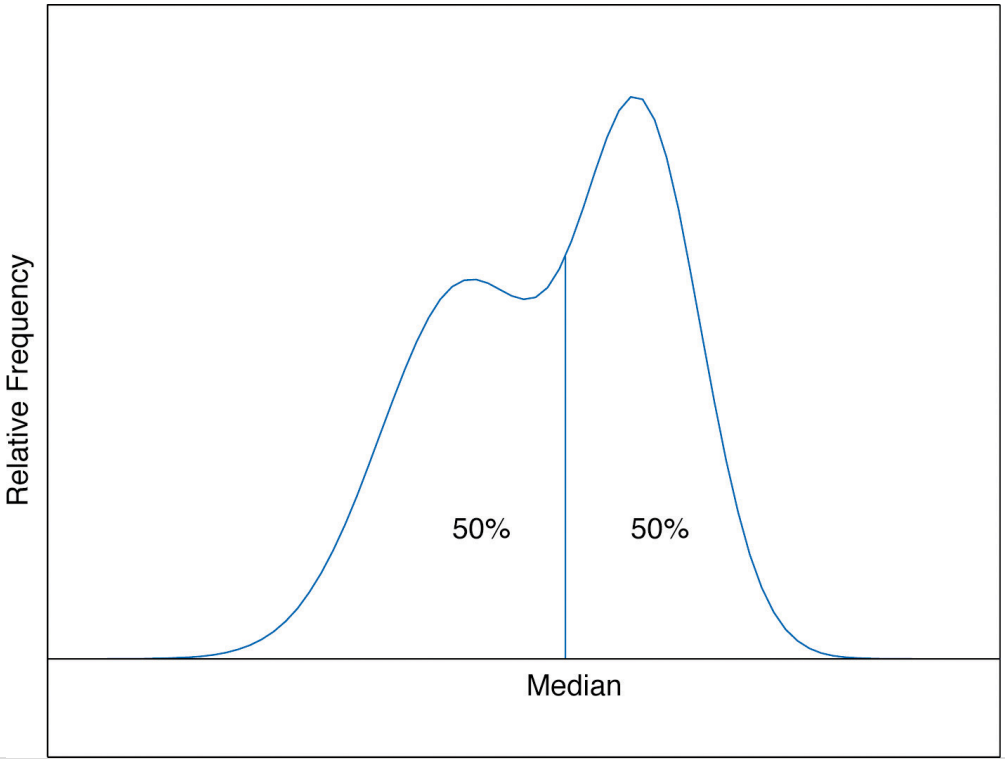
The **sample median**[7] $\tilde{x}$ *of a set of sample data for which there are an odd number of measurements is the middle measurement when the data are arranged in numerical order. The sample median* $\tilde{x}$ *of a set of sample data for which there are an even number of measurements is the mean of the two middle measurements when the data are arranged in numerical order.*

The population median is defined in a similar way, but we will not have occasion to refer to it again in this text.

The median is a value that divides the observations in a data set so that 50% of the data are on its left and the other 50% on its right. In accordance with Figure 2.6 "A Very Fine Relative Frequency Histogram", therefore, in the curve that represents the distribution of the data, a vertical line drawn at the median divides the area in two, area 0.5 (50% of the total area 1) to the left and area 0.5 (50% of the total area 1) to the right, as shown in Figure 2.7 "The Median". In our income example the median, $24,600, clearly gave a much better measure of the middle of the data set than did the mean $47,400. This is typical for situations in which the distribution is skewed. (Skewness and symmetry of distributions are discussed at the end of this subsection.)

7. The middle value when data are listed in numerical order.

Figure 2.7 The Median



EXAMPLE 5

Compute the sample median for the data of Note 2.11 "Example 2".

Solution:

The data in numerical order are -1, 0, 2, 2. The two middle measurements are 0 and 2, so $\widetilde{x} = (0 + 2) / 2 = 1$.

EXAMPLE 6

Compute the sample median for the data of Note 2.12 "Example 3".

Solution:

The data in numerical order are

$$1.39 \quad 1.76 \quad 1.90 \quad 2.12 \quad 2.53 \quad 2.71 \quad 3.00 \quad 3.33 \quad 3.71 \quad 4.00$$

The number of observations is ten, which is even, so there are two middle measurements, the fifth and sixth, which are 2.53 and 2.71. Therefore the median of these data is $\widetilde{x} = (2.53 + 2.71) \big/ 2 = 2.62.$

EXAMPLE 7

Compute the sample median for the data of Note 2.13 "Example 4".

Solution:

The data in numerical order are

$$0 \quad 0 \quad 0 \quad 1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 2 \quad 2 \quad 2 \quad 2 \quad 2 \quad 2 \quad 3 \quad 3 \quad 3 \quad 4$$

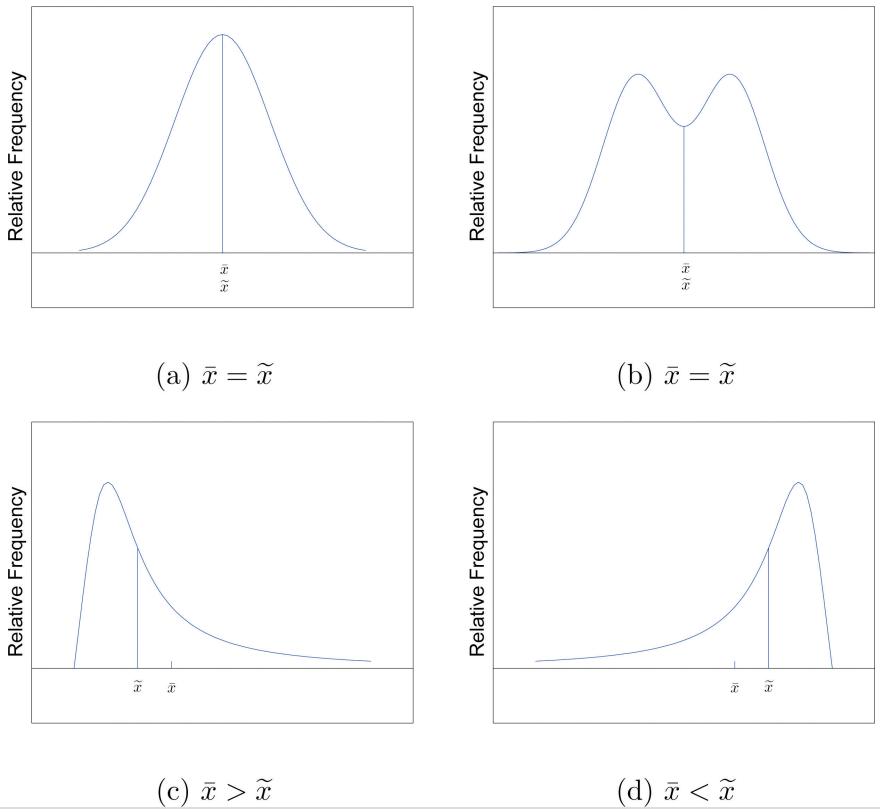
The number of observations is 19, which is odd, so there is one middle measurement, the tenth. Since the tenth measurement is 2, the median is $\widetilde{x} = 2.$

It is important to note that we could have computed the median without first explicitly listing all the observations in the data set. We already saw in Note 2.13 "Example 4" how to find the number of observations directly from the frequencies listed in the table: $n = 3 + 6 + 6 + 3 + 1 = 19.$ As just above we figure out that the median is the tenth observation. The second line of the table in Note 2.13 "Example 4" shows that when the data are listed in order there will be three 0s followed by six 1s, so the tenth observation is a 2. The median is therefore 2.

The relationship between the mean and the median for several common shapes of distributions is shown in Figure 2.8 "Skewness of Relative Frequency Histograms". The distributions in panels (a) and (b) are said to be *symmetric* because of the symmetry that they exhibit. The distributions in the remaining two panels are said to be *skewed*. In each distribution we have drawn a vertical line that divides the area under the curve in half, which in accordance with Figure 2.7 "The Median" is located at the median. The following facts are true in general:

a. When the distribution is symmetric, as in panels (a) and (b) of Figure 2.8 "Skewness of Relative Frequency Histograms", the mean and the median are equal.
b. When the distribution is as shown in panel (c) of Figure 2.8 "Skewness of Relative Frequency Histograms", it is said to be *skewed right*. The mean has been pulled to the right of the median by the long "right tail" of the distribution, the few relatively large data values.
c. When the distribution is as shown in panel (d) of Figure 2.8 "Skewness of Relative Frequency Histograms", it is said to be *skewed left*. The mean has been pulled to the left of the median by the long "left tail" of the distribution, the few relatively small data values.

Figure 2.8 *Skewness of Relative Frequency Histograms*



(a) $\bar{x} = \widetilde{x}$ (b) $\bar{x} = \widetilde{x}$

(c) $\bar{x} > \widetilde{x}$ (d) $\bar{x} < \widetilde{x}$

The Mode

Perhaps you have heard a statement like “The average number of automobiles owned by households in the United States is 1.37,” and have been amused at the thought of a fraction of an automobile sitting in a driveway. In such a context the following measure for central location might make more sense.

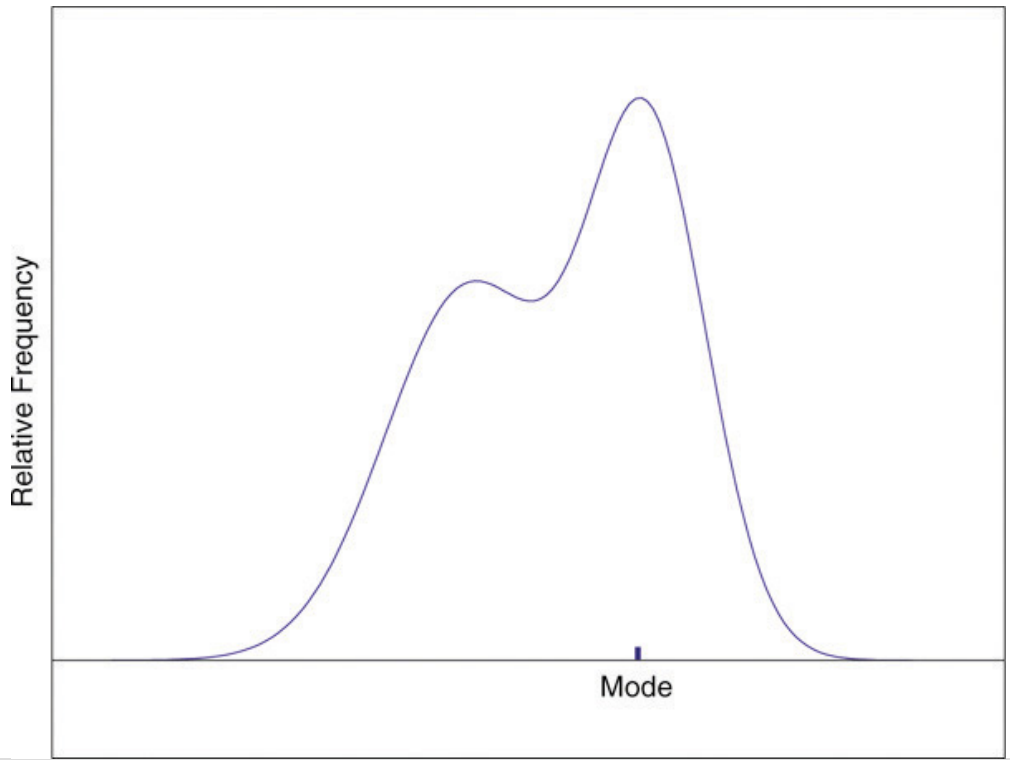
Definition

The **sample mode**[8] *of a set of sample data is the most frequently occurring value.*

The population mode is defined in a similar way, but we will not have occasion to refer to it again in this text.

On a relative frequency histogram, the highest point of the histogram corresponds to the mode of the data set. Figure 2.9 "Mode" illustrates the mode.

Figure 2.9 *Mode*



8. The most frequent value in a data set.

For any data set there is always exactly one mean and exactly one median. This need not be true of the mode; several different values could occur with the highest frequency, as we will see. It could even happen that every value occurs with the same frequency, in which case the concept of the mode does not make much sense.

EXAMPLE 8

Find the mode of the following data set.

$$-1 \quad 0 \quad 2 \quad 0$$

Solution:

The value 0 is most frequently observed and therefore the mode is 0.

EXAMPLE 9

Compute the sample mode for the data of Note 2.13 "Example 4".

Solution:

The two most frequently observed values in the data set are 1 and 2. Therefore mode is a set of two values: {1,2}.

The mode is a measure of central location since most real-life data sets have more observations near the center of the data range and fewer observations on the lower and upper ends. The value with the highest frequency is often in the middle of the data range.

KEY TAKEAWAY

The mean, the median, and the mode each answer the question “Where is the center of the data set?” The nature of the data set, as indicated by a relative frequency histogram, determines which one gives the best answer.

EXERCISES

BASIC

1. For the sample data set {1,2,6} find
   a. Σx
   b. Σx^2
   c. $\Sigma (x-3)$
   d. $\Sigma (x-3)^2$
2. For the sample data set $\{-1,0,1,4\}$ find
   a. Σx
   b. Σx^2
   c. $\Sigma (x-1)$
   d. $\Sigma (x-1)^2$
3. Find the mean, the median, and the mode for the sample
   $$1 \quad 2 \quad 3 \quad 4$$
4. Find the mean, the median, and the mode for the sample
   $$3 \quad 3 \quad 4 \quad 4$$
5. Find the mean, the median, and the mode for the sample
   $$2 \quad 1 \quad 2 \quad 7$$
6. Find the mean, the median, and the mode for the sample
   $$-1 \quad 0 \quad 1 \quad 4 \quad 1 \quad 1$$
7. Find the mean, the median, and the mode for the sample data represented by the table

x	1	2	7
f	1	2	1

8. Find the mean, the median, and the mode for the sample data represented by the table

x	−1	0	1	4
f	1	1	3	1

9. Create a sample data set of size $n = 3$ for which the mean $\overline{x}$ is greater than the median $\widetilde{x}$.

10. Create a sample data set of size $n = 3$ for which the mean $\bar{x}$ is less than the median $\tilde{x}$.

11. Create a sample data set of size $n = 4$ for which the mean $\bar{x}$, the median $\tilde{x}$, and the mode are all identical.

12. Create a data set of size $n = 4$ for which the median $\tilde{x}$ and the mode are identical but the mean $\bar{x}$ is different.

APPLICATIONS

13. Find the mean and the median for the LDL cholesterol level in a sample of ten heart patients.

132 162 133 145 148
139 147 160 150 153

14. Find the mean and the median, for the LDL cholesterol level in a sample of ten heart patients on a special diet.

127 152 138 110 152
113 131 148 135 158

15. Find the mean, the median, and the mode for the number of vehicles owned in a survey of 52 households.

x	0	1	2	3	4	5	6	7
f	2	12	15	11	6	3	1	2

16. The number of passengers in each of 120 randomly observed vehicles during morning rush hour was recorded, with the following results.

x	1	2	3	4	5
f	84	29	3	3	1

Find the mean, the median, and the mode of this data set.

17. Twenty-five 1-lb boxes of 16d nails were randomly selected and the number of nails in each box was counted, with the following results.

x	47	48	49	50	51
f	1	3	18	2	1

Find the mean, the median, and the mode of this data set.

ADDITIONAL EXERCISES

18. Five laboratory mice with thymus leukemia are observed for a predetermined period of 500 days. After 500 days, four mice have died but the fifth one survives. The recorded survival times for the five mice are

$$493 \quad 421 \quad 222 \quad 378 \quad 500^*$$

where 500^* indicates that the fifth mouse survived for at least 500 days but the survival time (i.e., the exact value of the observation) is unknown.

   a. Can you find the sample mean for the data set? If so, find it. If not, why not?
   b. Can you find the sample median for the data set? If so, find it. If not, why not?

19. Five laboratory mice with thymus leukemia are observed for a predetermined period of 500 days. After 450 days, three mice have died, and one of the remaining mice is sacrificed for analysis. By the end of the observational period, the last remaining mouse still survives. The recorded survival times for the five mice are

$$222 \quad 421 \quad 378 \quad 450^* \quad 500^*$$

where * indicates that the mouse survived for at least the given number of days but the exact value of the observation is unknown.

   a. Can you find the sample mean for the data set? If so, find it. If not, explain why not.
   b. Can you find the sample median for the data set? If so, find it. If not, explain why not.

20. A player keeps track of all the rolls of a pair of dice when playing a board game and obtains the following data.

x	2	3	4	5	6	7
f	10	29	40	56	68	77

x	8	9	10	11	12
f	67	55	39	28	11

Find the mean, the median, and the mode.

21. Cordelia records her daily commute time to work each day, to the nearest minute, for two months, and obtains the following data.

x	26	27	28	29	30	31	32
f	3	4	16	12	6	2	1

a. Based on the frequencies, do you expect the mean and the median to be about the same or markedly different, and why?
b. Compute the mean, the median, and the mode.

22. An ordered stem and leaf diagram gives the scores of 71 students on an exam.

10	0 0
9	1 1 1 1 2 3
8	0 1 1 2 2 3 4 5 7 8 8 9
7	0 0 0 1 1 2 4 4 5 6 6 6 7 7 7 8 8 9
6	0 1 2 2 2 3 4 4 5 7 7 7 7 8 8
5	0 2 3 3 4 4 6 7 7 8 9
4	2 5 6 8 8
3	9 9

a. Based on the shape of the display, do you expect the mean and the median to be about the same or markedly different, and why?
b. Compute the mean, the median, and the mode.

23. A man tosses a coin repeatedly until it lands heads and records the number of tosses required. (For example, if it lands heads on the first toss he records a 1; if it lands tails on the first two tosses and heads on the third he records a 3.) The data are shown.

x	1	2	3	4	5	6	7	8	9	10
f	384	208	98	56	28	12	8	2	3	1

a. Find the mean of the data.
b. Find the median of the data.

24. a. Construct a data set consisting of ten numbers, all but one of which is above average, where the average is the mean.
b. Is it possible to construct a data set as in part (a) when the average is the median? Explain.

25. Show that no matter what kind of average is used (mean, median, or mode) it is impossible for all members of a data set to be above average.

26. a. Twenty sacks of grain weigh a total of 1,003 lb. What is the mean weight per sack?
    b. Can the median weight per sack be calculated based on the information given? If not, construct two data sets with the same total but different medians.

27. Begin with the following set of data, call it Data Set I.

    5 −2 6 14 −3 0 1 4 3 2 5

    a. Compute the mean, median, and mode.
    b. Form a new data set, Data Set II, by adding 3 to each number in Data Set I. Calculate the mean, median, and mode of Data Set II.
    c. Form a new data set, Data Set III, by subtracting 6 from each number in Data Set I. Calculate the mean, median, and mode of Data Set III.
    d. Comparing the answers to parts (a), (b), and (c), can you guess the pattern? State the general principle that you expect to be true.

LARGE DATA SET EXERCISES

28. Large Data Set 1 lists the SAT scores and GPAs of 1,000 students.

    http://www.gone.2012books.lardbucket.org/sites/all/files/data1.xls

    a. Compute the mean and median of the 1,000 SAT scores.
    b. Compute the mean and median of the 1,000 GPAs.

29. Large Data Set 1 lists the SAT scores of 1,000 students.

    http://www.gone.2012books.lardbucket.org/sites/all/files/data1.xls

    a. Regard the data as arising from a census of all students at a high school, in which the SAT score of every student was measured. Compute the population mean μ.
    b. Regard the first 25 observations as a random sample drawn from this population. Compute the sample mean $\overline{x}$ and compare it to μ.
    c. Regard the next 25 observations as a random sample drawn from this population. Compute the sample mean $\overline{x}$ and compare it to μ.

30. Large Data Set 1 lists the GPAs of 1,000 students.

    http://www.gone.2012books.lardbucket.org/sites/all/files/data1.xls

    a. Regard the data as arising from a census of all freshman at a small college at the end of their first academic year of college study, in which the GPA of every such person was measured. Compute the population mean μ.

b. Regard the first 25 observations as a random sample drawn from this population. Compute the sample mean $\overline{x}$ and compare it to μ.
c. Regard the next 25 observations as a random sample drawn from this population. Compute the sample mean $\overline{x}$ and compare it to μ.

31. Large Data Sets 7, 7A, and 7B list the survival times in days of 140 laboratory mice with thymic leukemia from onset to death.

http://www.gone.2012books.lardbucket.org/sites/all/files/data7.xls

http://www.gone.2012books.lardbucket.org/sites/all/files/data7A.xls

http://www.gone.2012books.lardbucket.org/sites/all/files/data7B.xls

a. Compute the mean and median survival time for all mice, without regard to gender.
b. Compute the mean and median survival time for the 65 male mice (separately recorded in Large Data Set 7A).
c. Compute the mean and median survival time for the 75 female mice (separately recorded in Large Data Set 7B).

ANSWERS

1. a. 9.
   b. 41.
   c. 0.
   d. 14.

3. $\overline{x} = 2.5, \widetilde{x} = 2.5, \text{mode} = \{1,2,3,4\}$.

5. $\overline{x} = 3, \widetilde{x} = 2, \text{mode} = 2.$

7. $\overline{x} = 3, \widetilde{x} = 2, \text{mode} = 2.$

9. {0,0,3}.

11. {0,1,1,2}.

13. $\overline{x} = 146.9, \widetilde{x} = 147.5$

15. $\overline{x} = 2.6, \widetilde{x} = 2, \text{mode} = 2$

17. $\overline{x} = 48.96, \widetilde{x} = 49, \text{mode} = 49$

19. a. No, the survival times of the fourth and fifth mice are unknown.
    b. Yes, $\widetilde{x} = 421.$

21. $\overline{x} = 28.55, \widetilde{x} = 28, \text{mode} = 28$

23. $\overline{x} = 2.05, \widetilde{x} = 2, \text{mode} = 1$

25. Mean: $nx_{\min} \leq \Sigma x$ so dividing by n yields $x_{\min} \leq \overline{x}$, so the minimum value is not above average. Median: the middle measurement, or average of the two middle measurements, $\widetilde{x}$, is at least as large as $x_{\min}$, so the minimum value is not above average. Mode: the mode is one of the measurements, and is not greater than itself.

27. a. $\overline{x} = 3.\overline{18}, \widetilde{x} = 3, \text{mode} = 5.$
    b. $\overline{x} = 6.\overline{18}, \widetilde{x} = 6, \text{mode} = 8.$
    c. $\overline{x} = -2.\overline{81}, \widetilde{x} = -3, \text{mode} = -1.$
    d. If a number is added to every measurement in a data set, then the mean, median, and mode all change by that number.

29. a. $\mu = 1528.74$
    b. $\overline{x} = 1502.8$
    c. $\overline{x} = 1535.2$

31. a. $\bar{x} = 553.4286$ and $\tilde{x} = 552.5$
b. $\bar{x} = 665.9692$ and $\tilde{x} = 667$
c. $\bar{x} = 455.8933$ and $\tilde{x} = 448$

2.3 Measures of Variability

LEARNING OBJECTIVES

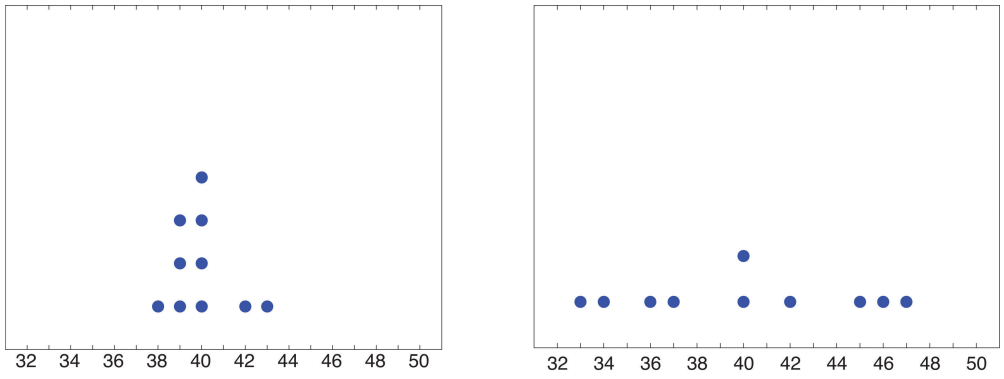
1. To learn the concept of the variability of a data set.
2. To learn how to compute three measures of the variability of a data set: the range, the variance, and the standard deviation.

Look at the two data sets in Table 2.1 "Two Data Sets" and the graphical representation of each, called a *dot plot*, in Figure 2.10 "Dot Plots of Data Sets".

Table 2.1 Two Data Sets

Data Set I:	40	38	42	40	39	39	43	40	39	40
Data Set II:	46	37	40	33	42	36	40	47	34	45

Figure 2.10 *Dot Plots of Data Sets*



(a) Set I (b) Set II

The two sets of ten measurements each center at the same value: they both have mean, median, and mode 40. Nevertheless a glance at the figure shows that they are markedly different. In Data Set I the measurements vary only slightly from the center, while for Data Set II the measurements vary greatly. Just as we have attached numbers to a data set to locate its center, we now wish to associate to each data set numbers that measure quantitatively how the data either scatter away

from the center or cluster close to it. These new quantities are called measures of variability, and we will discuss three of them.

The Range

The first measure of variability that we discuss is the simplest.

Definition

The **range**[9] *of a data set is the number R defined by the formula*

$$R = x_{\max} - x_{\min}$$

where $x_{\max}$ *is the largest measurement in the data set and* $x_{\min}$ *is the smallest.*

EXAMPLE 10

Find the range of each data set in Table 2.1 "Two Data Sets".

Solution:

For Data Set I the maximum is 43 and the minimum is 38, so the range is $R = 43 - 38 = 5$.

For Data Set II the maximum is 47 and the minimum is 33, so the range is $R = 47 - 33 = 14$.

The range is a measure of variability because it indicates the size of the interval over which the data points are distributed. A smaller range indicates less variability (less dispersion) among the data, whereas a larger range indicates the opposite.

The Variance and the Standard Deviation

The other two measures of variability that we will consider are more elaborate and also depend on whether the data set is just a sample drawn from a much larger population or is the whole population itself (that is, a census).

9. The variability of a data set as measured by the number $R = x_{\max} - x_{\min}$.

Definition

The **sample variance** *of a set of n sample data is the number s^2 defined by the formula*

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

which by algebra is equivalent to the formula

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

The **sample standard deviation**[10] *of a set of n sample data is the square root of the sample variance, hence is the number s given by the formulas*

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

Although the first formula in each case looks less complicated than the second, the latter is easier to use in hand computations, and is called a **shortcut formula**.

10. The variability of sample data as measured by the number $\sqrt{\frac{\Sigma(x-\overline{x})^2}{n-1}}$.

EXAMPLE 11

Find the sample variance and the sample standard deviation of Data Set II in Table 2.1 "Two Data Sets".

Solution:

To use the defining formula (the first formula) in the definition we first compute for each observation x its deviation $x - \overline{x}$ from the sample mean. Since the mean of the data is $\overline{x} = 40$, we obtain the ten numbers displayed in the second line of the supplied table.

x	46	37	40	33	42	36	40	47	34	45
$x - \overline{x}$	6	−3	0	−7	2	−4	0	7	−6	5

Then

$$\Sigma(x - \overline{x})^2 = 6^2 + (-3)^2 + 0^2 + (-7)^2 + 2^2 + (-4)^2 + 0^2 + 7^2 + (-6)^2 +$$

so

$$s^2 = \frac{\Sigma(x - \overline{x})^2}{n-1} = \frac{224}{9} = 24.\overline{8}$$

and

$$s = \sqrt{24.\overline{8}} \approx 4.99$$

The student is encouraged to compute the ten deviations for Data Set I and verify that their squares add up to 20, so that the sample variance and standard deviation of Data Set I are the much smaller numbers $s^2 = 20 / 9 = 2.\overline{2}$ and $s = \sqrt{20 / 9} \approx 1.49$.

EXAMPLE 12

Find the sample variance and the sample standard deviation of the ten GPAs in Note 2.12 "Example 3" in Section 2.2 "Measures of Central Location".

$$1.90 \quad 3.00 \quad 2.53 \quad 3.71 \quad 2.12 \quad 1.76 \quad 2.71 \quad 1.39 \quad 4.00 \quad 3.33$$

Solution:

Since

$$\Sigma x = 1.90 + 3.00 + 2.53 + 3.71 + 2.12 + 1.76 + 2.71 + 1.39 + 4.00 + 3.$$

and

$$\begin{aligned} \Sigma x^2 &= 1.90^2 + 3.00^2 + 2.53^2 + 3.71^2 + 2.12^2 + 1.76^2 \\ &\quad + 2.71^2 + 1.39^2 + 4.00^2 + 3.33^2 \\ &= 76.7321 \end{aligned}$$

the shortcut formula gives

$$s^2 = \frac{\Sigma x^2 - \frac{1}{n}(\Sigma x)^2}{n-1} = \frac{76.7321 - \frac{(26.45)^2}{10}}{10-1} = \frac{6.77185}{9} = .752427$$

and

$$s = \sqrt{.752427} \approx .867$$

The sample variance has different units from the data. For example, if the units in the data set were inches, the new units would be inches squared, or square inches. It is thus primarily of theoretical importance and will not be considered further in this text, except in passing.

If the data set comprises the whole population, then the *population* standard deviation, denoted σ (the lower case Greek letter sigma), and its square, the *population* variance σ^2, are defined as follows.

Definition

The **population variance** *and* **population standard deviation**[11] *of a set of N population data are the numbers σ^2 and σ defined by the formulas*

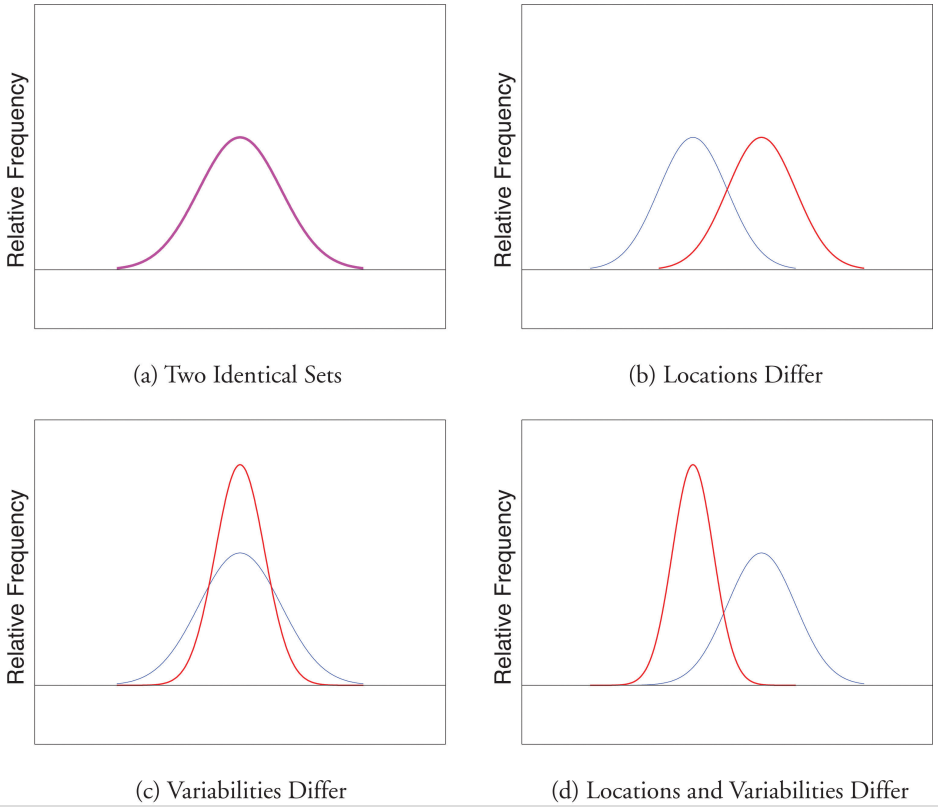
$$\sigma^2 = \frac{\Sigma(x-\mu)^2}{N} \quad \text{and} \quad \sigma = \sqrt{\frac{\Sigma(x-\mu)^2}{N}}$$

Note that the denominator in the fraction is the full number of observations, not that number reduced by one, as is the case with the sample standard deviation. Since most data sets are samples, we will always work with the sample standard deviation and variance.

Finally, in many real-life situations the most important statistical issues have to do with comparing the means and standard deviations of two data sets. Figure 2.11 "Difference between Two Data Sets" illustrates how a difference in one or both of the sample mean and the sample standard deviation are reflected in the appearance of the data set as shown by the curves derived from the relative frequency histograms built using the data.

11. The variability of population data as measured by the number $\sigma^2 = \frac{\Sigma(x-\mu)^2}{N}$.

Figure 2.11 *Difference between Two Data Sets*



(a) Two Identical Sets

(b) Locations Differ

(c) Variabilities Differ

(d) Locations and Variabilities Differ

KEY TAKEAWAY

The range, the standard deviation, and the variance each give a quantitative answer to the question "How variable are the data?"

EXERCISES

BASIC

1. Find the range, the variance, and the standard deviation for the following sample.

$$1 \quad 2 \quad 3 \quad 4$$

2. Find the range, the variance, and the standard deviation for the following sample.

$$2 \quad -3 \quad 6 \quad 0 \quad 3 \quad 1$$

3. Find the range, the variance, and the standard deviation for the following sample.

$$2 \quad 1 \quad 2 \quad 7$$

4. Find the range, the variance, and the standard deviation for the following sample.

$$-1 \quad 0 \quad 1 \quad 4 \quad 1 \quad 1$$

5. Find the range, the variance, and the standard deviation for the sample represented by the data frequency table.

x	1	2	7
f	1	2	1

6. Find the range, the variance, and the standard deviation for the sample represented by the data frequency table.

x	−1	0	1	4
f	1	1	3	1

APPLICATIONS

7. Find the range, the variance, and the standard deviation for the sample of ten IQ scores randomly selected from a school for academically gifted students.

$$132 \quad 162 \quad 133 \quad 145 \quad 148$$
$$139 \quad 147 \quad 160 \quad 150 \quad 153$$

8. Find the range, the variance and the standard deviation for the sample of ten IQ scores randomly selected from a school for academically gifted students.

142 152 138 145 148
139 147 155 150 153

ADDITIONAL EXERCISES

9. Consider the data set represented by the table

x	26	27	28	29	30	31	32
f	3	4	16	12	6	2	1

   a. Use the frequency table to find that $\Sigma x = 1256$ and $\Sigma x^2 = 35{,}926$.
   b. Use the information in part (a) to compute the sample mean and the sample standard deviation.

10. Find the sample standard deviation for the data

x	1	2	3	4	5
f	384	208	98	56	28

x	6	7	8	9	10
f	12	8	2	3	1

11. A random sample of 49 invoices for repairs at an automotive body shop is taken. The data are arrayed in the stem and leaf diagram shown. (Stems are thousands of dollars, leaves are hundreds, so that for example the largest observation is 3,800.)

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

For these data, $\Sigma x = 101{,}100$, $\Sigma x^2 = 244{,}830{,}000$.

   a. Compute the mean, median, and mode.
   b. Compute the range.

c. Compute the sample standard deviation.

12. What must be true of a data set if its standard deviation is 0?

13. A data set consisting of 25 measurements has standard deviation 0. One of the measurements has value 17. What are the other 24 measurements?

14. Create a sample data set of size $n = 3$ for which the range is 0 and the sample mean is 2.

15. Create a sample data set of size $n = 3$ for which the sample variance is 0 and the sample mean is 1.

16. The sample $\{-1,0,1\}$ has mean $\overline{x} = 0$ and standard deviation $s = 1$. Create a sample data set of size $n = 3$ for which $\overline{x} = 0$ and s is greater than 1.

17. The sample $\{-1,0,1\}$ has mean $\overline{x} = 0$ and standard deviation $s = 1$. Create a sample data set of size $n = 3$ for which $\overline{x} = 0$ and the standard deviation s is less than 1.

18. Begin with the following set of data, call it Data Set I.

$$5 \quad -2 \quad 6 \quad 14 \quad -3 \quad 0 \quad 1 \quad 4 \quad 3 \quad 2 \quad 5$$

a. Compute the sample standard deviation of Data Set I.
b. Form a new data set, Data Set II, by adding 3 to each number in Data Set I. Calculate the sample standard deviation of Data Set II.
c. Form a new data set, Data Set III, by subtracting 6 from each number in Data Set I. Calculate the sample standard deviation of Data Set III.
d. Comparing the answers to parts (a), (b), and (c), can you guess the pattern? State the general principle that you expect to be true.

LARGE DATA SET EXERCISES

19. Large Data Set 1 lists the SAT scores and GPAs of 1,000 students.

http://www.gone.2012books.lardbucket.org/sites/all/files/data1.xls

a. Compute the range and sample standard deviation of the 1,000 SAT scores.
b. Compute the range and sample standard deviation of the 1,000 GPAs.

20. Large Data Set 1 lists the SAT scores of 1,000 students.

http://www.gone.2012books.lardbucket.org/sites/all/files/data1.xls

a. Regard the data as arising from a census of all students at a high school, in which the SAT score of every student was measured. Compute the population range and population standard deviation σ.

b. Regard the first 25 observations as a random sample drawn from this population. Compute the sample range and sample standard deviation s and compare them to the population range and σ.
c. Regard the next 25 observations as a random sample drawn from this population. Compute the sample range and sample standard deviation s and compare them to the population range and σ.

21. Large Data Set 1 lists the GPAs of 1,000 students.

http://www.gone.2012books.lardbucket.org/sites/all/files/data1.xls

a. Regard the data as arising from a census of all freshman at a small college at the end of their first academic year of college study, in which the GPA of every such person was measured. Compute the population range and population standard deviation σ.
b. Regard the first 25 observations as a random sample drawn from this population. Compute the sample range and sample standard deviation s and compare them to the population range and σ.
c. Regard the next 25 observations as a random sample drawn from this population. Compute the sample range and sample standard deviation s and compare them to the population range and σ.

22. Large Data Sets 7, 7A, and 7B list the survival times in days of 140 laboratory mice with thymic leukemia from onset to death.

http://www.gone.2012books.lardbucket.org/sites/all/files/data7.xls

http://www.gone.2012books.lardbucket.org/sites/all/files/data7A.xls

http://www.gone.2012books.lardbucket.org/sites/all/files/data7B.xls

a. Compute the range and sample standard deviation of survival time for all mice, without regard to gender.
b. Compute the range and sample standard deviation of survival time for the 65 male mice (separately recorded in Large Data Set 7A).
c. Compute the range and sample standard deviation of survival time for the 75 female mice (separately recorded in Large Data Set 7B). Do you see a difference in the results for male and female mice? Does it appear to be significant?

ANSWERS

1. $R = 3$, $s^2 = 1.7$, $s = 1.3$.
3. $R = 6$, $s^2 = 7.\overline{3}$, $s = 2.7$.
5. $R = 6$, $s^2 = 7.3$, $s = 2.7$.
7. $R = 30$, $s^2 = 103.2$, $s = 10.2$.
9. $\overline{x} = 28.55$, $s = 1.3$.
11. a. $\overline{x} = 2063, \widetilde{x} = 2000, \text{mode} = 2000.$
    b. $R = 3400$.
    c. $s = 869$.
13. All are 17.
15. {1,1,1}
17. One example is $\{-.5, 0, .5\}$.
19. a. $R = 1350$ and $s = 212.5455$
    b. $R = 4.00$ and $s = 0.7407$
21. a. $R = 4.00$ and $\sigma = 0.740375$
    b. $R = 3.04$ and $s = 0.808045$
    c. $R = 2.49$ and $s = 0.657843$

2.4 Relative Position of Data

LEARNING OBJECTIVES

1. To learn the concept of the relative position of an element of a data set.
2. To learn the meaning of each of two measures, the percentile rank and the z-score, of the relative position of a measurement and how to compute each one.
3. To learn the meaning of the three quartiles associated to a data set and how to compute them.
4. To learn the meaning of the five-number summary of a data set, how to construct the box plot associated to it, and how to interpret the box plot.

When you take an exam, what is often as important as your actual score on the exam is the way your score compares to other students' performance. If you made a 70 but the average score (whether the mean, median, or mode) was 85, you did relatively poorly. If you made a 70 but the average score was only 55 then you did relatively well. In general, the significance of one observed value in a data set strongly depends on how that value compares to the other observed values in a data set. Therefore we wish to attach to each observed value a number that measures its relative position.

Percentiles and Quartiles

Anyone who has taken a national standardized test is familiar with the idea of being given both a score on the exam and a "percentile ranking" of that score. You may be told that your score was 625 and that it is the 85th percentile. The first number tells how you actually did on the exam; the second says that 85% of the scores on the exam were less than or equal to your score, 625.

Definition

Given an observed value x in a data set, x is the **Pth percentile**[12] *of the data if the percentage of the data that are less than or equal to x is P. The number P is the* **percentile rank**[13] *of x.*

12. The measurement x, if it exists, such that P percent of the data are less than or equal to x.
13. Of a measurement x, the percentage of the data that are less than or equal to x.

EXAMPLE 13

What percentile is the value 1.39 in the data set of ten GPAs considered in Note 2.12 "Example 3" in Section 2.2 "Measures of Central Location"? What percentile is the value 3.33?

Solution:

The data written in increasing order are

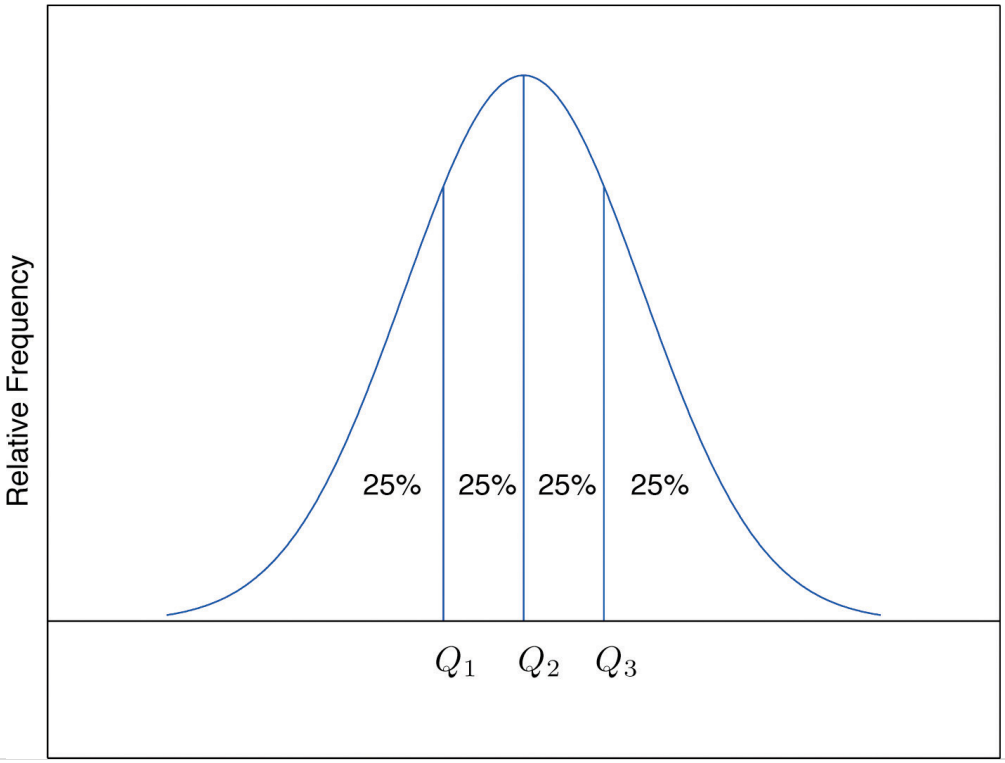
1.39 1.76 1.90 2.12 2.53 2.71 3.00 3.33 3.71 4.00

The only data value that is less than or equal to 1.39 is 1.39 itself. Since 1 is 1⁄10 = .10 or 10% of 10, the value 1.39 is the 10th percentile. Eight data values are less than or equal to 3.33. Since 8 is 8⁄10 = .80 or 80% of 10, the value 3.33 is the 80th percentile.

The Pth percentile cuts the data set in two so that approximately P% of the data lie below it and $(100 - P)$% of the data lie above it. In particular, the three percentiles that cut the data into fourths, as shown in Figure 2.12 "Data Division by Quartiles", are called the **quartiles**[14]. The following simple computational definition of the three quartiles works well in practice.

14. Of a data set, the three numbers Q_1, Q_2, Q_3 that divide the data approximately into fourths.

Figure 2.12 *Data Division by Quartiles*



Definition

For any data set:

1. *The* **second quartile** Q_2 *of the data set is its median.*
2. *Define two subsets:*
    1. *the lower set: all observations that are strictly less than* Q_2;
    2. *the upper set: all observations that are strictly greater than* Q_2.
3. *The* **first quartile** Q_1 *of the data set is the median of the lower set.*
4. *The* **third quartile** Q_3 *of the data set is the median of the upper set.*

EXAMPLE 14

Find the quartiles of the data set of GPAs of Note 2.12 "Example 3" in Section 2.2 "Measures of Central Location".

Solution:

As in the previous example we first list the data in numerical order:

$$1.39 \quad 1.76 \quad 1.90 \quad 2.12 \quad 2.53 \quad 2.71 \quad 3.00 \quad 3.33 \quad 3.71 \quad 4.00$$

This data set has $n = 10$ observations. Since 10 is an even number, the median is the mean of the two middle observations: $\widetilde{x} = (2.53 + 2.71) \Big/ 2 = 2.62$. Thus the second quartile is $Q_2 = 2.62$. The lower and upper subsets are

$$\text{Lower: } L = \{1.39, 1.76, 1.90, 2.12, 2.53\}$$
$$\text{Upper: } U = \{2.71, 3.00, 3.33, 3.71, 4.00\}$$

Each has an odd number of elements, so the median of each is its middle observation. Thus the first quartile is $Q_1 = 1.90$, the median of L, and the third quartile is $Q_3 = 3.33$, the median of U.

EXAMPLE 15

Adjoin the observation 3.88 to the data set of the previous example and find the quartiles of the new set of data.

Solution:

As in the previous example we first list the data in numerical order:

1.39 1.76 1.90 2.12 2.53 2.71 3.00 3.33 3.71 3.88 4.00

This data set has 11 observations. The second quartile is its median, the middle value 2.71. Thus $Q_2 = 2.71$. The lower and upper subsets are now

$$\text{Lower: } L = \{1.39, 1.76, 1.90, 2.12, 2.53\}$$
$$\text{Upper: } U = \{3.00, 3.33, 3.71, 3.88, 4.00\}$$

The lower set L has median the middle value 1.90, so $Q_1 = 1.90$. The upper set has median the middle value 3.71, so $Q_3 = 3.71$.

In addition to the three quartiles, the two extreme values, the minimum $x_{\min}$ and the maximum $x_{\max}$ are also useful in describing the entire data set. Together these five numbers are called the **five-number summary**[15] of the data set:

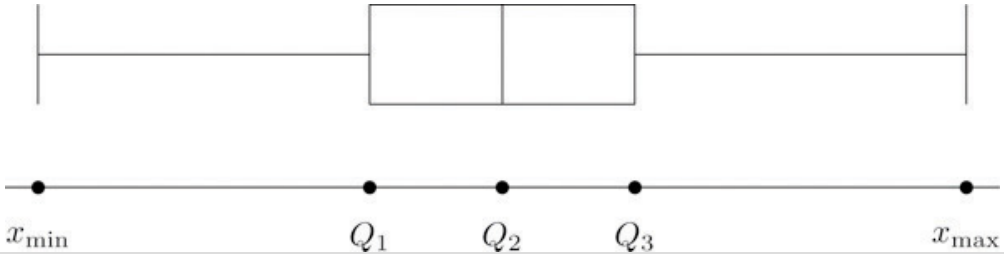
$$\{x_{\min}, Q_1, Q_2, Q_3, x_{\max}\}$$

The five-number summary is used to construct a **box plot**[16] as in Figure 2.13 "The Box Plot". Each of the five numbers is represented by a vertical line segment, a box is formed using the line segments at Q_1 and Q_3 as its two vertical sides, and two horizontal line segments are extended from the vertical segments marking Q_1 and Q_3 to the adjacent extreme values. (The two horizontal line segments are referred to as "whiskers," and the diagram is sometimes called a "box and whisker plot.") We caution the reader that there are other types of box plots that differ somewhat from the ones we are constructing, although all are based on the three quartiles.

15. Of a data set, the list $\{x_{min}, Q_1, Q_2, Q_3, x_{max}\}$.

16. For a data set, a diagram constructed using the five-number summary, as in Figure 2.13 "The Box Plot", which graphically summarizes the distribution of the data.

Figure 2.13 *The Box Plot*



Note that the distance from Q_1 to Q_3 is the length of the interval over which the middle half of the data range. Thus it has the following special name.

Definition

The **interquartile range (IQR)**[17] *is the quantity*

$$IQR = Q_3 - Q_1$$

17. Of a data set, the difference between the first and third quartiles.

EXAMPLE 16

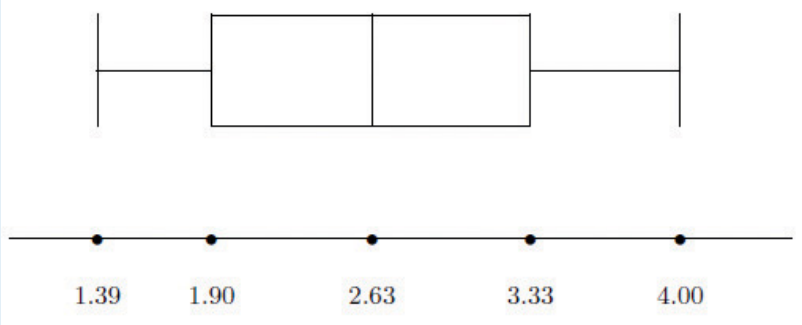
Construct a box plot and find the IQR for the data in Note 2.44 "Example 14".

Solution:

From our work in Note 2.44 "Example 14" we know that the five-number summary is

$$x_{\min} = 1.39 \quad Q_1 = 1.90 \quad Q_2 = 2.62 \quad Q_3 = 3.33 \quad x_{\max} = 4.00$$

The box plot is



The interquartile range is $\text{IQR} = 3.33 - 1.90 = 1.43.$

z-scores

Another way to locate a particular observation x in a data set is to compute its distance from the mean in units of standard deviation.

Definition

The ***z*-score**[18] *of an observation x is the number z given by the computational formula*

$$z = \frac{x - \overline{x}}{s} \quad or \quad z = \frac{x - \mu}{\sigma}$$

according to whether the data set is a sample or is the entire population.

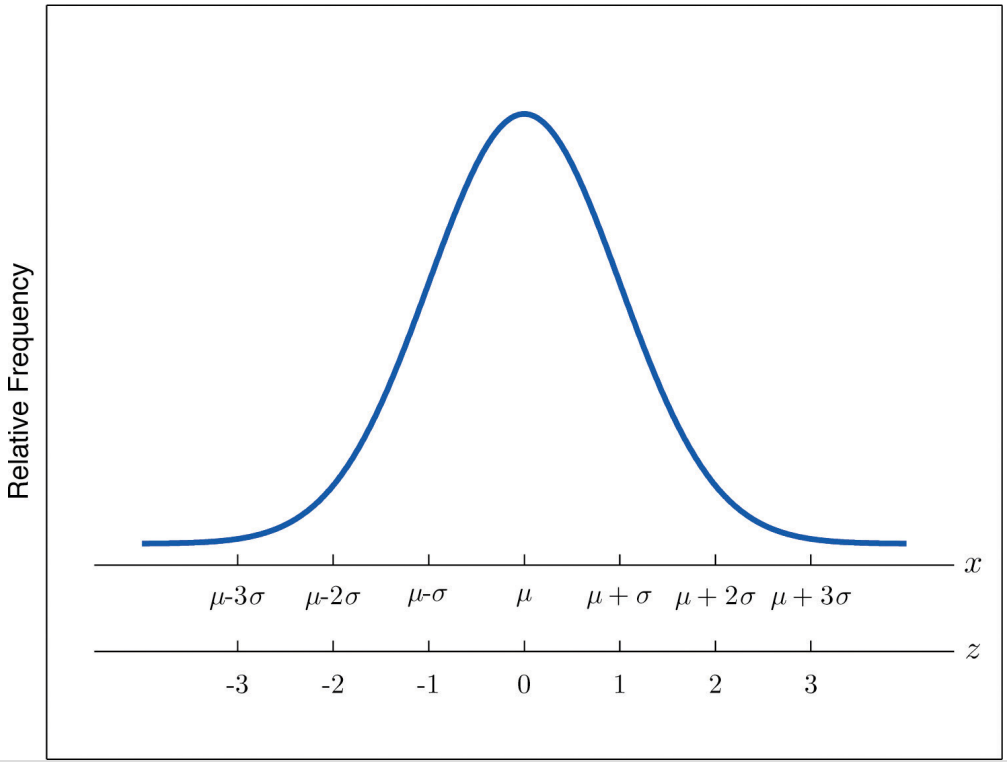
The formulas in the definition allow us to compute the *z*-score when *x* is known. If the *z*-score is known then *x* can be recovered using the corresponding inverse formulas

$$x = \overline{x} + sz \quad \text{or} \quad x = \mu + \sigma z$$

The *z*-score indicates how many standard deviations an individual observation *x* is from the center of the data set, its mean. If *z* is negative then *x* is below average. If *z* is 0 then *x* is equal to the average. If *z* is positive then *x* is above average. See Figure 2.14.

18. Of a measurement *x*, the distance of *x* from the mean in units of standard deviation.

Figure 2.14 x-Scale versus z-Score



EXAMPLE 17

Find the *z*-scores for all ten observations in the GPA sample data in Note 2.12 "Example 3" in Section 2.2 "Measures of Central Location".

1.90 3.00 2.53 3.71 2.12 1.76 2.71 1.39 4.00 3.33

Solution:

For these data $\overline{x} = 2.645$ and $s = 0.8674$. The first observation $x = 1.9$ in the data set has *z*-score

$$z = \frac{x - \overline{x}}{s} = \frac{1.9 - 2.645}{0.8674} = -0.8589$$

which means that $x = 1.90$ is 0.8589 standard deviations *below* the sample mean. The second observation $x = 3.00$ has *z*-score

$$z = \frac{x - \overline{x}}{s} = \frac{3.00 - 2.645}{0.8674} = 0.4093$$

which means that $x = 3.00$ is 0.4093 standard deviations *above* the sample mean. Repeating the process for the remaining observations gives the full set of *z*-scores

−0.86 0.41 −0.13 1.23 −0.61 −1.02 0.07 −1.45 1.56 0.79

EXAMPLE 18

Suppose the mean and standard deviation of the GPAs of all currently registered students at a college are $\mu = 2.70$ and $\sigma = 0.50$. The z-scores of the GPAs of two students, Antonio and Beatrice, are $z = -0.62$ and $z = 1.28$, respectively. What are their GPAs?

Solution:

Using the second formula right after the definition of z-scores we compute the GPAs as

$$\text{Antonio: } x = \mu + z\ \sigma = 2.70 + (-0.62)\ (0.50) = 2.39$$

$$\text{Beatrice: } x = \mu + z\ \sigma = 2.70 + (1.28)\ (0.50) = 3.34$$

KEY TAKEAWAYS

- The percentile rank and z-score of a measurement indicate its relative position with regard to the other measurements in a data set.
- The three quartiles divide a data set into fourths.
- The five-number summary and its associated box plot summarize the location and distribution of the data.

EXERCISES

BASIC

1. Consider the data set

69	92	68	77	80
93	75	76	82	100
70	85	88	85	96
53	70	70	82	85

   a. Find the percentile rank of 82.
   b. Find the percentile rank of 68.

2. Consider the data set

8.5	8.2	7.0	7.0	4.9
9.6	8.5	8.8	8.5	8.7
6.5	8.2	7.6	1.5	9.3
8.0	7.7	2.9	9.2	6.9

   a. Find the percentile rank of 6.5.
   b. Find the percentile rank of 7.7.

3. Consider the data set represented by the ordered stem and leaf diagram

10	0 0
9	1 1 1 1 2 3
8	0 1 1 2 2 3 4 5 7 8 8 9
7	0 0 0 1 1 2 4 4 5 6 6 6 7 7 7 8 8 9
6	0 1 2 2 2 3 4 4 5 7 7 7 7 8 8
5	0 2 3 3 4 4 6 7 7 8 9
4	2 5 6 8 8
3	9 9

   a. Find the percentile rank of the grade 75.
   b. Find the percentile rank of the grade 57.

4. Is the 90th percentile of a data set always equal to 90%? Why or why not?

5. The 29th percentile in a large data set is 5.
   a. Approximately what percentage of the observations are less than 5?
   b. Approximately what percentage of the observations are greater than 5?
6. The 54th percentile in a large data set is 98.6.
   a. Approximately what percentage of the observations are less than 98.6?
   b. Approximately what percentage of the observations are greater than 98.6?
7. In a large data set the 29th percentile is 5 and the 79th percentile is 10. Approximately what percentage of observations lie between 5 and 10?
8. In a large data set the 40th percentile is 125 and the 82nd percentile is 158. Approximately what percentage of observations lie between 125 and 158?
9. Find the five-number summary and the IQR and sketch the box plot for the sample represented by the stem and leaf diagram in Figure 2.2 "Ordered Stem and Leaf Diagram".
10. Find the five-number summary and the IQR and sketch the box plot for the sample explicitly displayed in Note 2.20 "Example 7" in Section 2.2 "Measures of Central Location".
11. Find the five-number summary and the IQR and sketch the box plot for the sample represented by the data frequency table

x	1	2	5	8	9
f	5	2	3	6	4

12. Find the five-number summary and the IQR and sketch the box plot for the sample represented by the data frequency table

x	−5	−3	−2	−1	0	1	3	4	5
f	2	1	3	2	4	1	1	2	1

13. Find the z-score of each measurement in the following sample data set.

$$-5 \quad 6 \quad 2 \quad -1 \quad 0$$

14. Find the z-score of each measurement in the following sample data set.

$$1.6 \quad 5.2 \quad 2.8 \quad 3.7 \quad 4.0$$

15. The sample with data frequency table

x	1	2	7
f	1	2	1

has mean $\overline{x} = 3$ and standard deviation $s \approx 2.71$. Find the z-score for every value in the sample.

16. The sample with data frequency table

x	-1	0	1	4
f	1	1	3	1

has mean $\overline{x} = 1$ and standard deviation $s \approx 1.67$. Find the z-score for every value in the sample.

17. For the population

$$0 \quad 0 \quad 2 \quad 2$$

compute each of the following.

a. The population mean μ.
b. The population variance σ^2.
c. The population standard deviation σ.
d. The z-score for every value in the population data set.

18. For the population

$$0.5 \quad 2.1 \quad 4.4 \quad 1.0$$

compute each of the following.

a. The population mean μ.
b. The population variance σ^2.
c. The population standard deviation σ.
d. The z-score for every value in the population data set.

19. A measurement x in a sample with mean $\overline{x} = 10$ and standard deviation $s = 3$ has z-score $z = 2$. Find x.

20. A measurement x in a sample with mean $\overline{x} = 10$ and standard deviation $s = 3$ has z-score $z = -1$. Find x.

21. A measurement x in a population with mean $\mu = 2.3$ and standard deviation $\sigma = 1.3$ has z-score $z = 2$. Find x.

22. A measurement x in a sample with mean $\mu = 2.3$ and standard deviation $\sigma = 1.3$ has z-score $z = -1.2$. Find x.

APPLICATIONS

23. The weekly sales for the last 20 weeks in a kitchen appliance store for an electric automatic rice cooker are

20	15	14	14	18
15	19	12	13	9
15	17	16	16	18
19	15	15	16	15

a. Find the percentile rank of 15.
b. If the sample accurately reflects the population, then what percentage of weeks would an inventory of 15 rice cookers be adequate?

24. The table shows the number of vehicles owned in a survey of 52 households.

x	0	1	2	3	4	5	6	7
f	2	12	15	11	6	3	1	2

a. Find the percentile rank of 2.
b. If the sample accurately reflects the population, then what percentage of households have at most two vehicles?

25. For two months Cordelia records her daily commute time to work each day to the nearest minute and obtains the following data:

x	26	27	28	29	30	31	32
f	3	4	16	12	6	2	1

Cordelia is supposed to be at work at 8:00 a.m. but refuses to leave her house before 7:30 a.m.

a. Find the percentile rank of 30, the time she has to get to work.
b. Assuming that the sample accurately reflects the population of *all* of Cordelia's commute times, use your answer to part (a) to predict the proportion of the work days she is late for work.

26. The mean score on a standardized grammar exam is 49.6; the standard deviation is 1.35. Dromio is told that the z-score of his exam score is –1.19.

a. Is Dromio's score above average or below average?
b. What was Dromio's actual score on the exam?

27. A random sample of 49 invoices for repairs at an automotive body shop is taken. The data are arrayed in the stem and leaf diagram shown. (Stems are

thousands of dollars, leaves are hundreds, so that for example the largest observation is 3,800.)

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

For these data, $\Sigma x = 101{,}100$, $\Sigma x^2 = 244{,}830{,}000$.

a. Find the z-score of the repair that cost \$1,100.
b. Find the z-score of the repairs that cost \$2,700.

28. The stem and leaf diagram shows the time in seconds that callers to a telephone-order center were on hold before their call was taken.

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

a. Find the quartiles.
b. Give the five-number summary of the data.
c. Find the range and the IQR.

ADDITIONAL EXERCISES

29. Consider the data set represented by the ordered stem and leaf diagram

10	0 0
9	1 1 1 1 2 3
8	0 1 1 2 2 3 4 5 7 8 8 9
7	0 0 0 1 1 2 4 4 5 6 6 6 7 7 7 8 8 9
6	0 1 2 2 2 3 4 4 5 7 7 7 7 8 8
5	0 2 3 3 4 4 6 7 7 8 9
4	2 5 6 8 8
3	9 9

a. Find the three quartiles.
b. Give the five-number summary of the data.
c. Find the range and the IQR.

30. For the following stem and leaf diagram the units on the stems are thousands and the units on the leaves are hundreds, so that for example the largest observation is 3,800.

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

a. Find the percentile rank of 800.
b. Find the percentile rank of 3,200.

31. Find the five-number summary for the following sample data.

x	26	27	28	29	30	31	32
f	3	4	16	12	6	2	1

32. Find the five-number summary for the following sample data.

x	1	2	3	4	5	6	7	8	9	10
f	384	208	98	56	28	12	8	2	3	1

33. For the following stem and leaf diagram the units on the stems are thousands and the units on the leaves are hundreds, so that for example the largest observation is 3,800.

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

    a. Find the three quartiles.
    b. Find the IQR.
    c. Give the five-number summary of the data.

34. Determine whether the following statement is true. "In any data set, if an observation x_1 is greater than another observation x_2, then the z-score of x_1 is greater than the z-score of x_2."

35. Emilia and Ferdinand took the same freshman chemistry course, Emilia in the fall, Ferdinand in the spring. Emilia made an 83 on the common final exam that she took, on which the mean was 76 and the standard deviation 8. Ferdinand made a 79 on the common final exam that he took, which was more difficult, since the mean was 65 and the standard deviation 12. The one who has a higher z-score did relatively better. Was it Emilia or Ferdinand?

36. Refer to the previous exercise. On the final exam in the same course the following semester, the mean is 68 and the standard deviation is 9. What grade on the exam matches Emilia's performance? Ferdinand's?

37. Rosencrantz and Guildenstern are on a weight-reducing diet. Rosencrantz, who weighs 178 lb, belongs to an age and body-type group for which the mean weight is 145 lb and the standard deviation is 15 lb. Guildenstern, who weighs 204 lb, belongs to an age and body-type group for which the mean weight is 165 lb and the standard deviation is 20 lb. Assuming z-scores are good measures for comparison in this context, who is more overweight for his age and body type?

LARGE DATA SET EXERCISES

38. Large Data Set 1 lists the SAT scores and GPAs of 1,000 students.

    http://www.gone.2012books.lardbucket.org/sites/all/files/data1.xls

    a. Compute the three quartiles and the interquartile range of the 1,000 SAT scores.
    b. Compute the three quartiles and the interquartile range of the 1,000 GPAs.

39. Large Data Set 10 records the scores of 72 students on a statistics exam.

    http://www.gone.2012books.lardbucket.org/sites/all/files/data10.xls

    a. Compute the five-number summary of the data.
    b. Describe in words the performance of the class on the exam in the light of the result in part (a).

40. Large Data Sets 3 and 3A list the heights of 174 customers entering a shoe store.

    http://www.gone.2012books.lardbucket.org/sites/all/files/data3.xls

    http://www.gone.2012books.lardbucket.org/sites/all/files/data3A.xls

    a. Compute the five-number summary of the heights, without regard to gender.
    b. Compute the five-number summary of the heights of the men in the sample.
    c. Compute the five-number summary of the heights of the women in the sample.

41. Large Data Sets 7, 7A, and 7B list the survival times in days of 140 laboratory mice with thymic leukemia from onset to death.

    http://www.gone.2012books.lardbucket.org/sites/all/files/data7.xls

    http://www.gone.2012books.lardbucket.org/sites/all/files/data7A.xls

    http://www.gone.2012books.lardbucket.org/sites/all/files/data7B.xls

    a. Compute the three quartiles and the interquartile range of the survival times for all mice, without regard to gender.
    b. Compute the three quartiles and the interquartile range of the survival times for the 65 male mice (separately recorded in Large Data Set 7A).
    c. Compute the three quartiles and the interquartile range of the survival times for the 75 female mice (separately recorded in Large Data Set 7B).

ANSWERS

1. a. 60.
   b. 10.

3. a. 59.
   b. 23.

5. a. 29.
   b. 71.

7. 50%.

9. $x_{\min} = 25, Q_1 = 70, Q_2 = 77.5, Q_3 = 90, x_{\max} = 100,$ $\text{IQR} = 20$

11. $x_{\min} = 1, Q_1 = 1.5, Q_2 = 6.5, Q_3 = 8, x_{\max} = 9, \text{IQR} = 6.5$

13. –1.3, 1.39, 0.4, –0.35, –0.11.

15. $z = -0.74$ for $x = 1$, $z = -0.37$ for $x = 2$, $z = 1.48$ for $x = 7$.

17. a. 1.
    b. 1.
    c. 1.
    d. $z = -1$ for $x = 0$, $z = 1$ for $x = 2$.

19. 16.

21. 4.9.

23. a. 55.
    b. 55.

25. a. 93.
    b. 0.07.

27. a. –1.11.
    b. 0.73.

29. a. $Q_1 = 59, Q_2 = 70, Q_3 = 81.$
    b. $x_{\min} = 39, Q_1 = 59, Q_2 = 70, Q_3 = 81, x_{\max} = 100.$
    c. $R = 61, \text{IQR} = 22.$

31. $x_{\min} = 26, Q_1 = 28, Q_2 = 28, Q_3 = 29, x_{\max} = 32.$

33. a. $Q_1 = 1450, Q_2 = 2000, Q_3 = 2800.$
    b. $\text{IQR} = 1350.$

c. $x_{\min} = 400, Q_1 = 1450, Q_2 = 2000, Q_3 = 2800,$
$x_{\max} = 3800.$

35. Emilia: $z =. 875$, Ferdinand: $z = 1.1\overline{6}.$

37. Rosencrantz: $z = 2.2$, Guildenstern: $z = 1.95$. Rosencrantz is more overweight for his age and body type.

39. a. $x_{min} = 15, Q_1 = 51, Q_2 = 67, Q_3 = 82,$ and $x_{max} = 97.$
b. The data set appears to be skewed to the left.

41. a. $Q_1 = 440, Q_2 = 552.5, Q_3 = 661,$ and $IQR = 221.$
b. $Q_1 = 641, Q_2 = 667, Q_3 = 700,$ and $IQR = 59.$
c. $Q_1 = 407, Q_2 = 448, Q_3 = 504,$ and $IQR = 97.$

2.5 The Empirical Rule and Chebyshev's Theorem

LEARNING OBJECTIVES

1. To learn what the value of the standard deviation of a data set implies about how the data scatter away from the mean as described by the Empirical Rule and Chebyshev's Theorem.
2. To use the Empirical Rule and Chebyshev's Theorem to draw conclusions about a data set.

You probably have a good intuitive grasp of what the average of a data set says about that data set. In this section we begin to learn what the standard deviation has to tell us about the nature of the data set.

The Empirical Rule

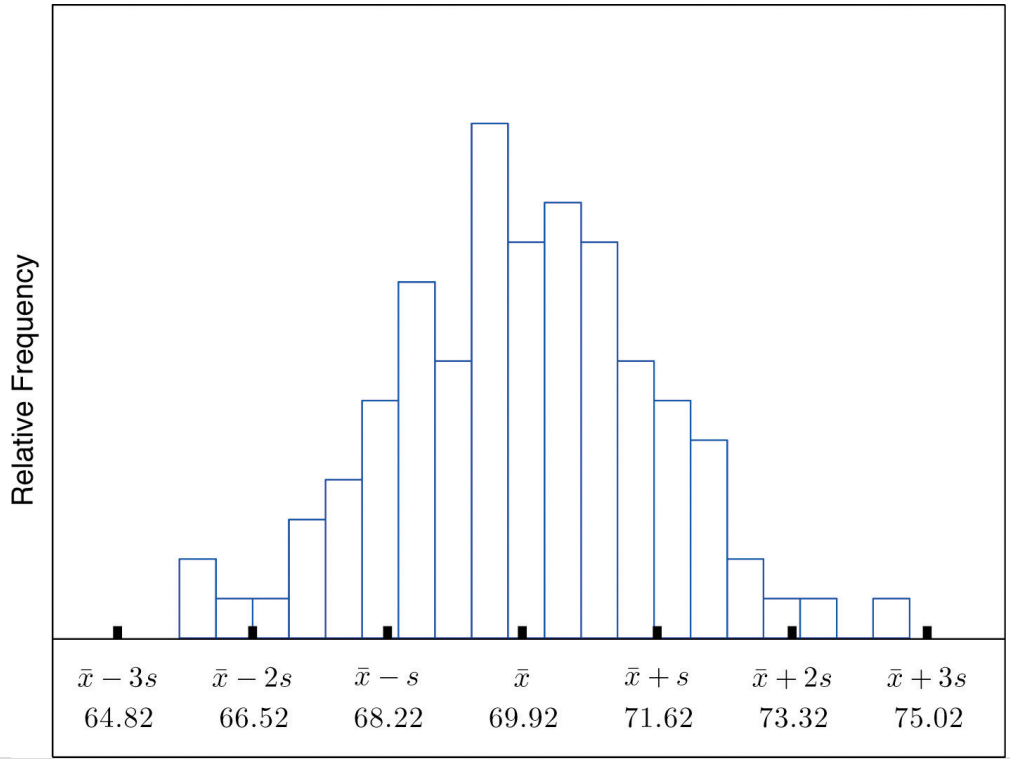
We start by examining a specific set of data. Table 2.2 "Heights of Men" shows the heights in inches of 100 randomly selected adult men. A relative frequency histogram for the data is shown in Figure 2.15 "Heights of Adult Men". The mean and standard deviation of the data are, rounded to two decimal places, $\overline{x} = 69.92$ and $s = 1.70$. If we go through the data and count the number of observations that are within one standard deviation of the mean, that is, that are between $69.92 - 1.70 = 68.22$ and $69.92 + 1.70 = 71.62$ inches, there are 69 of them. If we count the number of observations that are within two standard deviations of the mean, that is, that are between $69.92 - 2\,(1.70) = 66.52$ and $69.92 + 2\,(1.70) = 73.32$ inches, there are 95 of them. All of the measurements are within three standard deviations of the mean, that is, between $69.92 - 3\,(1.70) = 64.822$ and $69.92 + 3\,(1.70) = 75.02$ inches. These tallies are not coincidences, but are in agreement with the following result that has been found to be widely applicable.

Table 2.2 Heights of Men

68.7	72.3	71.3	72.5	70.6	68.2	70.1	68.4	68.6	70.6
73.7	70.5	71.0	70.9	69.3	69.4	69.7	69.1	71.5	68.6
70.9	70.0	70.4	68.9	69.4	69.4	69.2	70.7	70.5	69.9
69.8	69.8	68.6	69.5	71.6	66.2	72.4	70.7	67.7	69.1

68.8	69.3	68.9	74.8	68.0	71.2	68.3	70.2	71.9	70.4
71.9	72.2	70.0	68.7	67.9	71.1	69.0	70.8	67.3	71.8
70.3	68.8	67.2	73.0	70.4	67.8	70.0	69.5	70.1	72.0
72.2	67.6	67.0	70.3	71.2	65.6	68.1	70.8	71.4	70.2
70.1	67.5	71.3	71.5	71.0	69.1	69.5	71.1	66.8	71.8
69.6	72.7	72.8	69.6	65.9	68.0	69.7	68.7	69.8	69.7

Figure 2.15 *Heights of Adult Men*

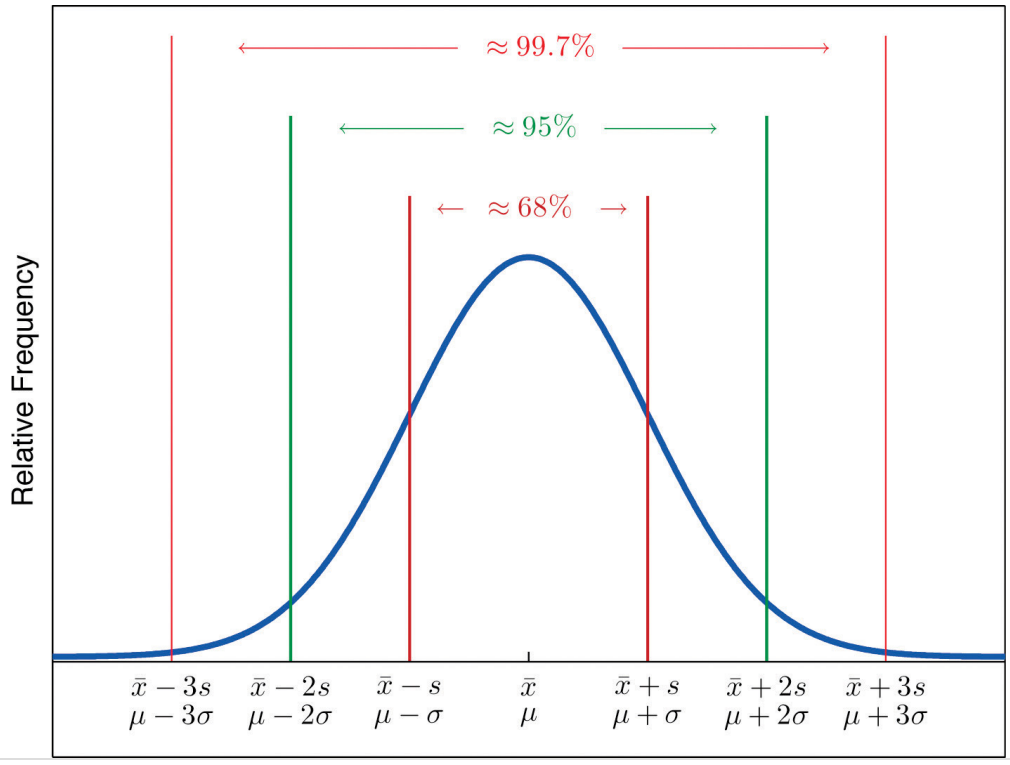


The Empirical Rule

If a data set has an approximately bell-shaped relative frequency histogram, then (see Figure 2.16 "The Empirical Rule")

1. approximately 68% of the data lie within one standard deviation of the mean, that is, in the interval with endpoints $\overline{x} \pm s$ for samples and with endpoints $\mu \pm \sigma$ for populations;
2. approximately 95% of the data lie within two standard deviations of the mean, that is, in the interval with endpoints $\overline{x} \pm 2s$ for samples and with endpoints $\mu \pm 2\sigma$ for populations; and
3. approximately 99.7% of the data lies within three standard deviations of the mean, that is, in the interval with endpoints $\overline{x} \pm 3s$ for samples and with endpoints $\mu \pm 3\sigma$ for populations.

Figure 2.16 The Empirical Rule



Two key points in regard to the Empirical Rule are that the data distribution must be approximately *bell-shaped* and that the percentages are only *approximately* true. The Empirical Rule does not apply to data sets with severely asymmetric distributions, and the actual percentage of observations in any of the intervals specified by the rule could be either greater or less than those given in the rule. We see this with the example of the heights of the men: the Empirical Rule suggested 68 observations between 68.22 and 71.62 inches but we counted 69.

EXAMPLE 19

Heights of 18-year-old males have a bell-shaped distribution with mean 69.6 inches and standard deviation 1.4 inches.

a. About what proportion of all such men are between 68.2 and 71 inches tall?
b. What interval centered on the mean should contain about 95% of all such men?

Solution:

A sketch of the distribution of heights is given in Figure 2.17 "Distribution of Heights".

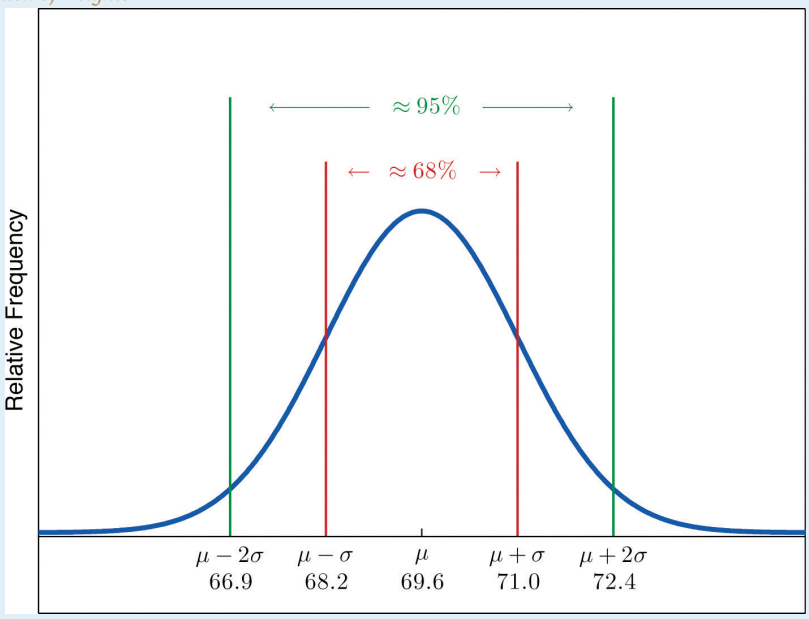
a. Since the interval from 68.2 to 71.0 has endpoints $\overline{x} - s$ and $\overline{x} + s$, by the Empirical Rule about 68% of all 18-year-old males should have heights in this range.

b. By the Empirical Rule the shortest such interval has endpoints $\overline{x} - 2s$ and $\overline{x} + 2s$. Since

$$\overline{x} - 2s = 69.6 - 2(1.4) = 66.8 \quad \text{and} \quad \overline{x} + 2s = 69.6 + 2(1.4) = 72.4$$

the interval in question is the interval from 66.8 inches to 72.4 inches.

Figure 2.17
Distribution of Heights



EXAMPLE 20

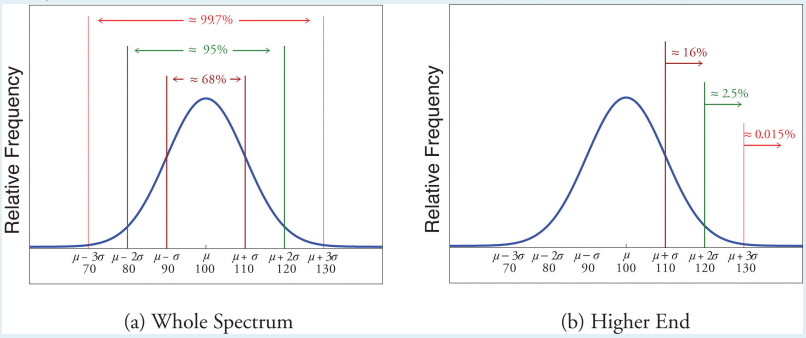
Scores on IQ tests have a bell-shaped distribution with mean $\mu = 100$ and standard deviation $\sigma = 10$. Discuss what the Empirical Rule implies concerning individuals with IQ scores of 110, 120, and 130.

Solution:

A sketch of the IQ distribution is given in Figure 2.18 "Distribution of IQ Scores". The Empirical Rule states that

1. approximately 68% of the IQ scores in the population lie between 90 and 110,
2. approximately 95% of the IQ scores in the population lie between 80 and 120, and
3. approximately 99.7% of the IQ scores in the population lie between 70 and 130.

Figure 2.18
Distribution of IQ Scores



(a) Whole Spectrum (b) Higher End

Since 68% of the IQ scores lie *within* the interval from 90 to 110, it must be the case that 32% lie *outside* that interval. By symmetry approximately half of that 32%, or 16% of all IQ scores, will lie above 110. If 16% lie above 110, then 84% lie below. We conclude that the IQ score 110 is the 84th percentile.

The same analysis applies to the score 120. Since approximately 95% of all IQ scores lie within the interval form 80 to 120, only 5% lie outside it, and half of them, or 2.5% of all scores, are above 120. The IQ score 120 is thus higher than 97.5% of all IQ scores, and is quite a high score.

By a similar argument, only 15/100 of 1% of all adults, or about one or two in every thousand, would have an IQ score above 130. This fact makes the score 130 extremely high.

Chebyshev's Theorem

The Empirical Rule does not apply to all data sets, only to those that are bell-shaped, and even then is stated in terms of approximations. A result that applies to every data set is known as Chebyshev's Theorem.

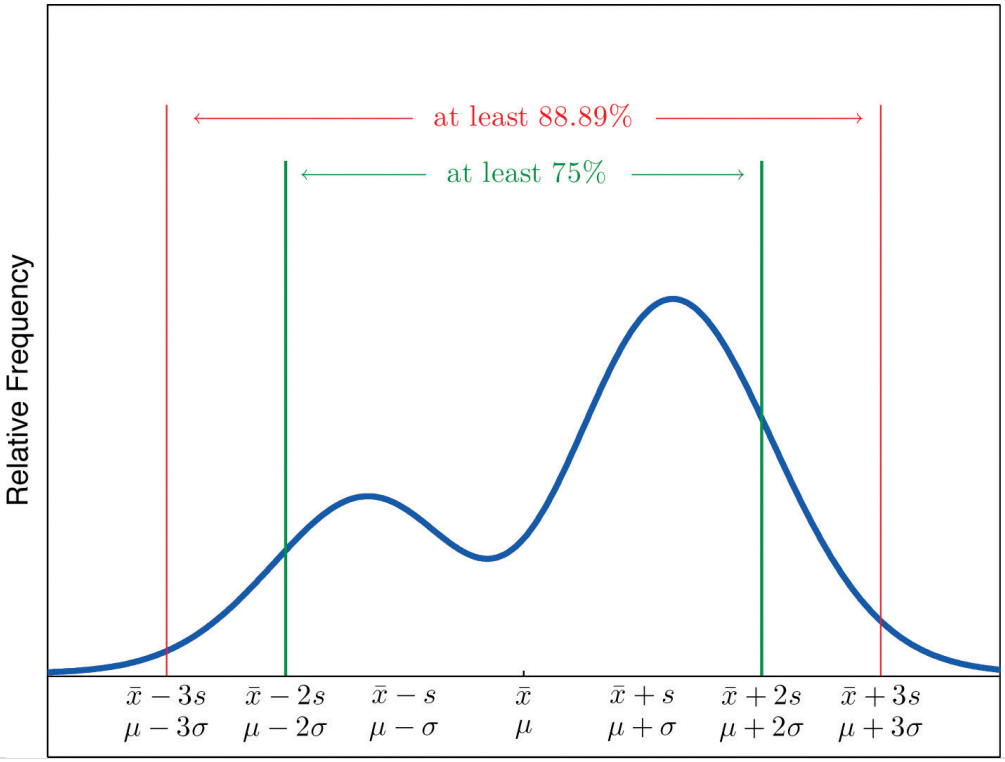
Chebyshev's Theorem

For any numerical data set,

1. at least 3/4 of the data lie within two standard deviations of the mean, that is, in the interval with endpoints $\overline{x} \pm 2s$ for samples and with endpoints $\mu \pm 2\sigma$ for populations;
2. at least 8/9 of the data lie within three standard deviations of the mean, that is, in the interval with endpoints $\overline{x} \pm 3s$ for samples and with endpoints $\mu \pm 3\sigma$ for populations;
3. at least $1 - 1 \Big/ k^2$ of the data lie within k standard deviations of the mean, that is, in the interval with endpoints $\overline{x} \pm ks$ for samples and with endpoints $\mu \pm k\sigma$ for populations, where k is any positive whole number that is greater than 1.

Figure 2.19 "Chebyshev's Theorem" gives a visual illustration of Chebyshev's Theorem.

Figure 2.19 *Chebyshev's Theorem*



It is important to pay careful attention to the words "at least" at the beginning of each of the three parts. The theorem gives the *minimum* proportion of the data which must lie within a given number of standard deviations of the mean; the true proportions found within the indicated regions could be greater than what the theorem guarantees.

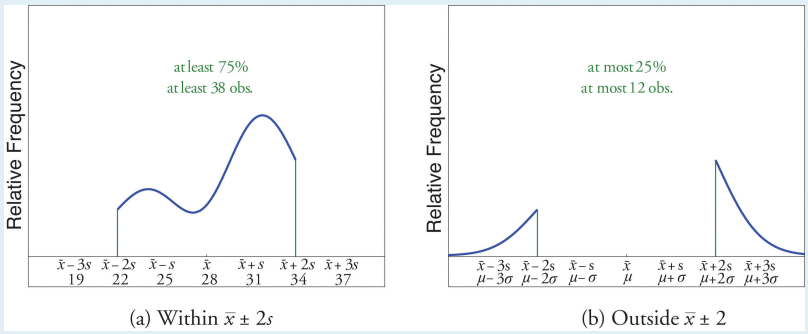
EXAMPLE 21

A sample of size $n = 50$ has mean $\bar{x} = 28$ and standard deviation $s = 3$. Without knowing anything else about the sample, what can be said about the number of observations that lie in the interval (22,34)? What can be said about the number of observations that lie outside that interval?

Solution:

The interval (22,34) is the one that is formed by adding and subtracting two standard deviations from the mean. By Chebyshev's Theorem, at least 3/4 of the data are within this interval. Since 3/4 of 50 is 37.5, this means that at least 37.5 observations are in the interval. But one cannot take a fractional observation, so we conclude that at least 38 observations must lie inside the interval (22,34).

If at least 3/4 of the observations are in the interval, then at most 1/4 of them are outside it. Since 1/4 of 50 is 12.5, at most 12.5 observations are outside the interval. Since again a fraction of an observation is impossible, x (22,34).



(a) Within $\bar{x} \pm 2s$ (b) Outside $\bar{x} \pm 2$

EXAMPLE 22

The number of vehicles passing through a busy intersection between 8:00 a.m. and 10:00 a.m. was observed and recorded on every weekday morning of the last year. The data set contains $n = 251$ numbers. The sample mean is $\overline{x} = 725$ and the sample standard deviation is $s = 25$. Identify which of the following statements *must* be true.

1. On approximately 95% of the weekday mornings last year the number of vehicles passing through the intersection from 8:00 a.m. to 10:00 a.m. was between 675 and 775.
2. On at least 75% of the weekday mornings last year the number of vehicles passing through the intersection from 8:00 a.m. to 10:00 a.m. was between 675 and 775.
3. On at least 189 weekday mornings last year the number of vehicles passing through the intersection from 8:00 a.m. to 10:00 a.m. was between 675 and 775.
4. On at most 25% of the weekday mornings last year the number of vehicles passing through the intersection from 8:00 a.m. to 10:00 a.m. was either less than 675 or greater than 775.
5. On at most 12.5% of the weekday mornings last year the number of vehicles passing through the intersection from 8:00 a.m. to 10:00 a.m. was less than 675.
6. On at most 25% of the weekday mornings last year the number of vehicles passing through the intersection from 8:00 a.m. to 10:00 a.m. was less than 675.

Solution:

1. Since it is not stated that the relative frequency histogram of the data is bell-shaped, the Empirical Rule does not apply. Statement (1) is based on the Empirical Rule and therefore it might not be correct.
2. Statement (2) is a direct application of part (1) of Chebyshev's Theorem because $(\overline{x}-2s, \overline{x}+2s) = (675,775)$. It must be correct.
3. Statement (3) says the same thing as statement (2) because 75% of 251 is 188.25, so the minimum whole number of observations in this interval is 189. Thus statement (3) is definitely correct.
4. Statement (4) says the same thing as statement (2) but in different words, and therefore is definitely correct.
5. Statement (4), which is definitely correct, states that at most 25% of the time either fewer than 675 or more than 775 vehicles passed through the intersection. Statement (5) says that half of that 25% corresponds to

days of light traffic. This would be correct if the relative frequency histogram of the data were known to be symmetric. But this is not stated; perhaps all of the observations outside the interval (675,775) are less than 75. Thus statement (5) might not be correct.

6. Statement (4) is definitely correct and statement (4) implies statement (6): even if every measurement that is outside the interval (675,775) is less than 675 (which is conceivable, since symmetry is not known to hold), even so at most 25% of all observations are less than 675. Thus statement (6) must definitely be correct.

KEY TAKEAWAYS

- The Empirical Rule is an approximation that applies only to data sets with a bell-shaped relative frequency histogram. It estimates the proportion of the measurements that lie within one, two, and three standard deviations of the mean.
- Chebyshev's Theorem is a fact that applies to all possible data sets. It describes the minimum proportion of the measurements that lie must within one, two, or more standard deviations of the mean.

EXERCISES

BASIC

1. State the Empirical Rule.
2. Describe the conditions under which the Empirical Rule may be applied.
3. State Chebyshev's Theorem.
4. Describe the conditions under which Chebyshev's Theorem may be applied.
5. A sample data set with a bell-shaped distribution has mean $\overline{x} = 6$ and standard deviation $s = 2$. Find the approximate proportion of observations in the data set that lie:
   a. between 4 and 8;
   b. between 2 and 10;
   c. between 0 and 12.
6. A population data set with a bell-shaped distribution has mean $\mu = 6$ and standard deviation $\sigma = 2$. Find the approximate proportion of observations in the data set that lie:
   a. between 4 and 8;
   b. between 2 and 10;
   c. between 0 and 12.
7. A population data set with a bell-shaped distribution has mean $\mu = 2$ and standard deviation $\sigma = 1.1$. Find the approximate proportion of observations in the data set that lie:
   a. above 2;
   b. above 3.1;
   c. between 2 and 3.1.
8. A sample data set with a bell-shaped distribution has mean $\overline{x} = 2$ and standard deviation $s = 1.1$. Find the approximate proportion of observations in the data set that lie:
   a. below −0.2;
   b. below 3.1;
   c. between −1.3 and 0.9.

9. A population data set with a bell-shaped distribution and size $N = 500$ has mean $\mu = 2$ and standard deviation $\sigma = 1.1$. Find the approximate number of observations in the data set that lie:

    a. above 2;
    b. above 3.1;
    c. between 2 and 3.1.

10. A sample data set with a bell-shaped distribution and size $n = 128$ has mean $\overline{x} = 2$ and standard deviation $s = 1.1$. Find the approximate number of observations in the data set that lie:

    a. below –0.2;
    b. below 3.1;
    c. between –1.3 and 0.9.

11. A sample data set has mean $\overline{x} = 6$ and standard deviation $s = 2$. Find the minimum proportion of observations in the data set that must lie:

    a. between 2 and 10;
    b. between 0 and 12;
    c. between 4 and 8.

12. A population data set has mean $\mu = 2$ and standard deviation $\sigma = 1.1$. Find the minimum proportion of observations in the data set that must lie:

    a. between –0.2 and 4.2;
    b. between –1.3 and 5.3.

13. A population data set of size $N = 500$ has mean $\mu = 5.2$ and standard deviation $\sigma = 1.1$. Find the minimum number of observations in the data set that must lie:

    a. between 3 and 7.4;
    b. between 1.9 and 8.5.

14. A sample data set of size $n = 128$ has mean $\overline{x} = 2$ and standard deviation $s = 2$. Find the minimum number of observations in the data set that must lie:

    a. between –2 and 6 (including –2 and 6);
    b. between –4 and 8 (including –4 and 8).

15. A sample data set of size $n = 30$ has mean $\overline{x} = 6$ and standard deviation $s = 2$.

    a. What is the maximum proportion of observations in the data set that can lie outside the interval (2,10)?
    b. What can be said about the proportion of observations in the data set that are below 2?

c. What can be said about the proportion of observations in the data set that are above 10?
d. What can be said about the number of observations in the data set that are above 10?

16. A population data set has mean $\mu = 2$ and standard deviation $\sigma = 1.1$.

a. What is the maximum proportion of observations in the data set that can lie outside the interval $(-1.3, 5.3)$?
b. What can be said about the proportion of observations in the data set that are below –1.3?
c. What can be said about the proportion of observations in the data set that are above 5.3?

APPLICATIONS

17. Scores on a final exam taken by 1,200 students have a bell-shaped distribution with mean 72 and standard deviation 9.

a. What is the median score on the exam?
b. About how many students scored between 63 and 81?
c. About how many students scored between 72 and 90?
d. About how many students scored below 54?

18. Lengths of fish caught by a commercial fishing boat have a bell-shaped distribution with mean 23 inches and standard deviation 1.5 inches.

a. About what proportion of all fish caught are between 20 inches and 26 inches long?
b. About what proportion of all fish caught are between 20 inches and 23 inches long?
c. About how long is the longest fish caught (only a small fraction of a percent are longer)?

19. Hockey pucks used in professional hockey games must weigh between 5.5 and 6 ounces. If the weight of pucks manufactured by a particular process is bell-shaped, has mean 5.75 ounces and standard deviation 0.125 ounce, what proportion of the pucks will be usable in professional games?

20. Hockey pucks used in professional hockey games must weigh between 5.5 and 6 ounces. If the weight of pucks manufactured by a particular process is bell-shaped and has mean 5.75 ounces, how large can the standard deviation be if 99.7% of the pucks are to be usable in professional games?

21. Speeds of vehicles on a section of highway have a bell-shaped distribution with mean 60 mph and standard deviation 2.5 mph.

    a. If the speed limit is 55 mph, about what proportion of vehicles are speeding?
    b. What is the median speed for vehicles on this highway?
    c. What is the percentile rank of the speed 65 mph?
    d. What speed corresponds to the 16th percentile?

22. Suppose that, as in the previous exercise, speeds of vehicles on a section of highway have mean 60 mph and standard deviation 2.5 mph, but now the distribution of speeds is unknown.

    a. If the speed limit is 55 mph, at least what proportion of vehicles must speeding?
    b. What can be said about the proportion of vehicles going 65 mph or faster?

23. An instructor announces to the class that the scores on a recent exam had a bell-shaped distribution with mean 75 and standard deviation 5.

    a. What is the median score?
    b. Approximately what proportion of students in the class scored between 70 and 80?
    c. Approximately what proportion of students in the class scored above 85?
    d. What is the percentile rank of the score 85?

24. The GPAs of all currently registered students at a large university have a bell-shaped distribution with mean 2.7 and standard deviation 0.6. Students with a GPA below 1.5 are placed on academic probation. Approximately what percentage of currently registered students at the university are on academic probation?

25. Thirty-six students took an exam on which the average was 80 and the standard deviation was 6. A rumor says that five students had scores 61 or below. Can the rumor be true? Why or why not?

ADDITIONAL EXERCISES

26. For the sample data

x	26	27	28	29	30	31	32
f	3	4	16	12	6	2	1

$\Sigma x = 1{,}256$ and $\Sigma x^2 = 35{,}926$.

a. Compute the mean and the standard deviation.
b. About how many of the measurements does the Empirical Rule predict will be in the interval $(\overline{x} - s, \overline{x} + s)$ the interval $(\overline{x} - 2s, \overline{x} + 2s)$, and the interval $(\overline{x} - 3s, \overline{x} + 3s)$?
c. Compute the number of measurements that are actually in each of the intervals listed in part (a), and compare to the predicted numbers.

27. A sample of size $n = 80$ has mean 139 and standard deviation 13, but nothing else is known about it.

a. What can be said about the number of observations that lie in the interval (126,152)?
b. What can be said about the number of observations that lie in the interval (113,165)?
c. What can be said about the number of observations that exceed 165?
d. What can be said about the number of observations that either exceed 165 or are less than 113?

28. For the sample data

x	1	2	3	4	5
f	84	29	3	3	1

$\Sigma x = 168$ and $\Sigma x^2 = 300$.

a. Compute the sample mean and the sample standard deviation.
b. Considering the shape of the data set, do you expect the Empirical Rule to apply? Count the number of measurements within one standard deviation of the mean and compare it to the number predicted by the Empirical Rule.
c. What does Chebyshev's Rule say about the number of measurements within one standard deviation of the mean?
d. Count the number of measurements within two standard deviations of the mean and compare it to the minimum number guaranteed by Chebyshev's Theorem to lie in that interval.

29. For the sample data set

x	47	48	49	50	51
f	1	3	18	2	1

$\Sigma x = 1224$ and $\Sigma x^2 = 59{,}940$.

a. Compute the sample mean and the sample standard deviation.
b. Considering the shape of the data set, do you expect the Empirical Rule to apply? Count the number of measurements within one standard deviation of the mean and compare it to the number predicted by the Empirical Rule.

c. What does Chebyshev's Rule say about the number of measurements within one standard deviation of the mean?
d. Count the number of measurements within two standard deviations of the mean and compare it to the minimum number guaranteed by Chebyshev's Theorem to lie in that interval.

ANSWERS

1. See the displayed statement in the text.

3. See the displayed statement in the text.

5.
   a. 0.68.
   b. 0.95.
   c. 0.997.

7.
   a. 0.5.
   b. 0.16.
   c. 0.34.

9.
   a. 250.
   b. 80.
   c. 170.

11.
   a. 3/4.
   b. 8/9.
   c. 0.

13.
   a. 375.
   b. 445.

15.
   a. At most 0.25.
   b. At most 0.25.
   c. At most 0.25.
   d. At most 7.

17.
   a. 72.
   b. 816.
   c. 570.
   d. 30.

19. 0.95.

21.
   a. 0.975.
   b. 60.
   c. 97.5.
   d. 57.5.

23.
   a. 75.
   b. 0.68.
   c. 0.025.
   d. 0.975.

25. By Chebyshev's Theorem at most 1⁄9 of the scores can be below 62, so the rumor is impossible.

27.
a. Nothing.
b. It is at least 60.
c. It is at most 20.
d. It is at most 20.

29.
a. $\overline{x} = 48.96$, $s = 0.7348$.
b. Roughly bell-shaped, the Empirical Rule should apply. True count: 18, predicted: 17.
c. Nothing.
d. True count: 23, guaranteed: at least 18.75, hence at least 19.