Unit 2: Stemplots

PREREQUISITES

Stemplots require familiarity with place value in the number system. There are no statistical prerequisites.

ADDITIONAL TOPIC COVERAGE

Additional coverage of stemplots can be found in *The Basic Practice of Statistics*, Chapter 1, Picturing Distributions with Graphs.

ACTIVITY DESCRIPTION

Students are often more invested if they have an opportunity to analyze data that they have collected. Use the questions suggested in the Unit 2 Activity survey (these questions are listed below for your reference) or create alternative questions of your own.

MATERIALS

Survey questionnaire, one copy per student.

Prior to handing out the survey, ask students to wait a moment while you get things ready. Take your time – so that students have to wait a few moments. (This wait time is related to question 1.) Then hand out the survey.

Once students have answered the survey, they will need to turn in their responses. Combine student responses to each question into a single table (See Table T2.1 in the activity solutions for an example.) Make sure that the same units are used by all students – for example, height in inches. These data will be revisited in Unit 5's activity (boxplots), so save these data. (If you decide not to collect data from your class, use the sample data from Table T2.1 instead.)

As students make stemplots of the data on each variable, encourage them to experiment with using different stems. Sometimes it is helpful to expand the stems and other times it is helpful to truncate the data values and collapse the stems. The idea is to get a stemplot that reveals information about the data. Below is a copy of the suggested survey questionnaire. Feel free to adapt or revise the questions.

Survey Questionnaire

1. How long (in seconds) did you wait while your instructor was getting ready for this activity?

2. How much money in coins are you carrying with you right now?

3. To the nearest inch, how tall are you?

4. How long (in minutes) do you study, on average, for an exam?

5. On a typical day, how many minutes do you exercise?

6. Circle your gender: Male Female

Return your answers to your instructor.

THE VIDEO SOLUTIONS

1. Sample answer: head circumference, upper arm circumference, foot length, foot width, height.

2. The foot length data was fairly symmetric with a single peak. The center was around 26.8 inches.

3. City miles per gallon (mpg).

4. There were outliers at the upper end of fuel efficiency. A few cars got great gas mileage.

5. The data for the 2012 models exhibited more spread. There were vehicles that were more fuel efficient (for example, the Prius) in 2012 compared to 1984, but there were vehicles that were less fuel efficient (for example, SUVs) in 2012 compared to 1984.

UNIT ACTIVITY SOLUTIONS

Sample solutions are based on the data in Table T2.1.

Question 1 Wait (sec)	Question 2 Coins (cents)	Question 3 Height (in)	Question 4 Study (min)	Question 5 Exercise (min)	Question 6 Gender
40	77	68	30	75	Male
40	62	67	60	20	Female
75	175	73	30	0	Male
40	189	72	20	45	Male
50	120	71	15	90	Male
40	54	68	45	75	Female
45	26	66	60	30	Female
45	145	75	30	30	Male
40	0	69	120	0	Female
35	0	71	45	60	Male
45	35	72	30	30	Male
45	47	64	15	45	Female
45	125	72	20	90	Male
45	55	71	30	0	Male
40	35	69	60	45	Male
45	78	63	45	90	Female
55	157	65	45	20	Male
40	225	62	75	40	Female
40	92	64	30	60	Female
50	85	62	60	0	Female
35	35	64	45	30	Female
45	59	66	45	90	Female
50	145	60	30	45	Female
30	137	70	30	30	Male
50	142	69	20	45	Male
45	62	69	30	60	Male

Table T2.1. Sample data from unit activity survey.

1. Sample answers based on sample data in Table T2.1.

Time (sec)

```
3 | 0
3 | 55
4 | 00000000
4 | 555555555
5 | 0000
5 | 5
6 |
6 |
7 |
7 | 5
```

The stemplot for Time is single-peaked and fairly symmetric. The middle is somewhere in the 40s. There is one outlier at 75.

Money (cents)   Leaf unit = 10

```
0   00
0   2333
0   4555
0   6677
0   89
1
1   223
1   4445
1   7
1   8
2
2   2
```

For the stemplot above of the money (cents), we truncated the pennies place. These data are not symmetric. There are two clumps of data and one single high value in the 220s.

Height (in)

6	0
6	223
6	4445
6	667
6	889999
7	0111
7	2223
7	5

The height data are somewhat mound-shaped and roughly symmetric. The middle is around 67. There are no outliers.

Study Time (min)

1	55
2	000
3	000000000
4	555555
5	
6	0000
7	5
8	
9	
10	
11	
12	0

These data do not appear symmetric. The study times are mostly under 45 minutes. There is at least one outlier, the largest of which is 120 minutes.

Exercise (min)

0	0000
1	
2	00
3	00000
4	055555
5	
6	000
7	55
8	
9	0000

There are some gaps in the data. There are some people who don't exercise and the same number who exercise, on average, 90 minutes per day. The middle appears around 45.

2. The time estimates from male students were more spread out than for female students. A middle value for the female students looks to be at around 40 seconds and for male students at around 45 seconds. One male student was an outlier, at 75 seconds.

Time

Female		Male
	3	0
5	3	5
00000	4	000
5555	4	55555
00	5	00
	5	5
	6	
	6	
	7	
	7	5

3. The change in male students’ pockets split into two groups. One group was at the low end of the change carried by female students. However, the other group tended to have more change than the female students. One female was an outlier. She carried $2.25 in change, more than anyone else in the class.

Money

Female		Male
0	0	0
32	0	33
554	0	5
76	0	67
98	0	
4	1	
	1	223
	1	445
	1	7
	1	8
	2	
2	2	

EXERCISE SOLUTIONS

1. Sample answer: time to run one mile; time to complete an obstacle course; number of pull-ups completed without stopping, number of sit-ups completed without stopping; resting pulse rate (a low rate means more fit).

2. a.

2	25
3	45
4	1166679
5	449
6	0

b. It is roughly symmetric with the center at stem 4. It is unimodal.

c. The center is around 46. (There are 15 observations – the 7th, 8th, and 9th value of the ordered data are all 46.)

d. No. Although 60 is the largest data value it is not outside the pattern that includes two 54s and a 59. There is no gap between 60 and the other data values.

3. a.

46	99
47	99
48	25579
49	33345599
50	9
51	234557
52	033
53	99
54	268
55	5
56	348
57	01256
58	03456
59	02369

b. Sample answer: The distribution appears to have three peaks, one around the 490's, another around the 510's and a third around the 580's (or around the 570's – 590's). The distribution doesn't look very symmetric. At the lower end, there are few scores in the 460's and 470's and then the number of scores increases for the 480's and 490's. On the higher end, the pattern is reverse. There are a larger number of scores in the 590's, 580's and 570's and then the number of scores decreases in the 560's and 550's. The middle number is 523, which can serve as the center of the distribution. (Or students might suggest the middle is in the 520's because 24 scores are below this stem and 24 scores are above this stem.)

c.

Writing		Math
93	45	7
954	46	9
96654310	47	
97	48	79
9997751	49	00369
985	50	01112289
7631	51	13568
92	52	13568
6	53	79
765	54	1135
6431	55	09
973321	56	589
97553	57	023
	58	
1	59	1113
	60	2468
	61	27

d. Sample answer: The average Math SATs are more spread out than the average Writing SATs. The lowest Math SAT was 457 and the highest was 617, for a spread of 160 points. The lowest Writing SAT was 453 and the highest was 591, for a spread of 138 points. The center of the Writing SATs is in the 510's and the center of the Math SATs is in the 520's. (The actual middle number of the ordered data is 511 for the Writing SATs and 527 for the Math SATs.) There are gaps in the Math SAT data in the 470's and 580's; the average scores of 457 and 469 might be considered outliers. There is a gap in the 580's for the Writing SATs and 591 is a potential outlier. The Writing SATs appear multimodal, with a peak around the 470's and another around the 560's. Neither of the distributions appears to be roughly symmetric.

4. The distribution is not symmetric. The data is concentrated toward the lower numbers and then trails off as the numbers get larger. In other words, this show attracts a mostly young audience. The center is around 19 (the 22nd observation in the ordered data of 44 values). There is an outlier at 120, which is considerably larger than the second largest data value of only 65. Most likely this is a typo – maybe the person was 12 or 20 but certainly not 120. The other possibility is that someone was being funny and responded that he/she was 120.

0	566899
1	0012223445667789
2	001233467
3	0135
4	28
5	0025
6	05
7	
8	
9	
10	
11	
12	0

REVIEW QUESTIONS SOLUTIONS

1. a.

0	3344
0	5555556667788
1	02
1	79
2	01
2	68
3	
3	
4	
4	7
5	23
5	678
6	44
6	7788
7	01344
7	789
8	0
8	799
9	3

b. The states divide into two clusters, with students from one group of states participating in SATs at a very low rate and students in the second group participating at a much higher rate. The lower cluster varies from 3% to 28% and the upper cluster from 47% to 93%. The lower cluster is concentrated around the single digits and then trails in the teens and twenties. The upper cluster is unimodal and roughly symmetric with its center at around 70%. (In some states, most college-bound students take the SATs. In other states, the rival American College Testing, or ACT, exams dominate and only students applying to selective colleges take the SATs. This explains, in part, the two clusters.)

c. Similar to the 2010/2011 data, the 1990 data breaks into two clusters. The lower cluster of the 1990 percentages has a similar spread to the 2010/2011 percentages and has roughly the same number of data values in the lower cluster. The gap between lower and upper clusters begins at the same stem for both years, but is slightly wider for the 2010/2011 data than for the 1990 data. The upper cluster is reasonably symmetric in both years. However, the spread of the upper cluster is wider for the 2010/2011 data (47% to 93% for a difference of 46%) than for the 1990 data (42% to 74% for a difference of 32%). The center of the upper cluster for the 2010/2011 data is around 70%; the center for the 1990 upper cluster is only around 58%.

1990		2010-11
4	0	3344
99866555	0	5555555666
4322221000	1	02
765	1	79
420	2	01
85	2	68
	3	
	3	
4422	4	
95	4	7
442	5	23
98875	5	678
4220	6	44
987	6	7788
420	7	01344
	7	789
	8	0
	8	799
	9	3

2. Sample answer: The Army should stock boots that fit foot widths from 90 millimeters to 113 millimeters. There was one outlier, a soldier with a foot width of 119 millimeters – 6 millimeters larger than the second largest foot width. The boot for that soldier should be specially ordered

(See stemplot on next page...).

9	011
9	2223
9	455
9	77
9	888
10	0001
10	222333
10	445
10	77
10	88
11	000011
11	3
11	
11	
11	9

3. a.

Leaf unit = 0.1

Girls		Boys
4	13	257
97744210	14	0458
9988531	15	45577
53	16	03
	17	356
	18	
	19	
	20	6
	21	
	22	8
	23	
	24	5
2	25	
9	26	

b. Sample answer: The overall pattern for boys' BMI is a flat mount shape that is roughly symmetric. There is a gap in the 18s and 19s and then three possible outliers: 20.6, 22.8, and 24.5.

c. Sample answer: The overall pattern for girls' BMI is mound-shaped and roughly symmetric. However, there appear to be two outliers: 25.2 and 26.9.

d. Sample answer: Ignoring the outliers identified in (b) and (c), the girls' data is less spread out than the boys' data. Just by eyeballing the data, the girls' data is centered around 15.1 and the boys' data is centered a little higher at about 15.5. The outliers in the girls' data are more extreme than the potential outliers in the boys' data.

Unit 3: Histograms

PREREQUISITES

Histograms require the ability to group numbers by size into categories. Students will need to be able to compute proportions (relative frequencies) and percentages. This unit continues the discussion of describing distributions that began in Unit 2, Stemplots.

ADDITIONAL TOPIC COVERAGE

Additional information on histograms and other graphic displays can be found in *The Basic Practice of Statistics*, Chapter 1, Picturing Distributions with Graphs.

ACTIVITY DESCRIPTION

The Unit 3 activity focuses on quality control in the production of polished wafers used in the manufacture of microchips. The Wafer Thickness tool found in the Interactive Tools menu is required for this activity. Using this interactive, students can set three controls at three different levels. These controls affect the thickness distribution of polished wafers. The final task asks students to make a recommendation for the control settings so that the product is consistently close to the target thickness of 0.5 mm.

MATERIALS

Students will need access to the Wafer Thickness interactive from the online Interactive Tools menu.

The activity introduces students to histograms and the concept of variability. Students learn how histograms are constructed by watching a histogram being made in real time as the data are generated by the Wafer Thickness interactive. In addition, students should discover (at an informal level) that there are different sources of variability – this understanding will be useful preparation for future units. Here are some sources of variability that students should observe:

- Under the same control settings, thickness varies from wafer to wafer.
- Under the same control settings, the histograms from two samples of wafers will differ (variability due to sampling).
- Changing the control settings changes the distribution of wafer thickness (variability due to control settings).

Questions 4 and 5 are ideal for group work. In question 4a students must design a strategy for determining the effect that changes in the control settings have on the sample data. There are three control settings, each having three levels. Hence, there are 3 x 3 x 3 = 27 distinct sets of possible control levels. A carefully designed plan may reduce the number of settings used in the investigation. The variability due to sampling makes 4b somewhat difficult to answer.

Students may have to view more than one sample from a given set of control settings. If students get frustrated, have them start with Control 3. Control 3's effect on the spread of the data is probably the easiest to spot. Control 1's shift in location is also not difficult to observe, particularly if Control 3 is set at level 3. Control 2's effect on shape as well as location is the most difficult to ascertain and students may not be able to figure out how Control 2 affects the distribution of wafer thickness. (That's OK – this happens in the real world.)

In question 5, students need to make a recommendation on the best choice of settings for the three controls and to support that recommendation based on the histograms they have constructed. There is not a single correct answer to this question. Some settings clearly give better results than others – but a "best" choice of control settings is a point open to argument. Students should make a decision and defend it against other possibilities.

It should be noted that data from a single sample can be saved in a CSV file. Data from CSV files can be imported into statistics packages or worked with in Excel.

Students may be interested in seeing how real data on microchip thickness are gathered. The video clip at the following site shows a technician taking measurements from wafers on which microchips have been embedded:

http://www.youtube.com/watch?v=jG84UjCZboo

THE VIDEO SOLUTIONS

1. Time of first lightning flash.

2. Horizontal scale: Time of day in hours.
Vertical Scale: Percent of days with first lightning flash within that hour.

3. Roughly symmetric.

4. These were values that were separated from the overall pattern by a gap in the data.

5. The classes need to have equal width.

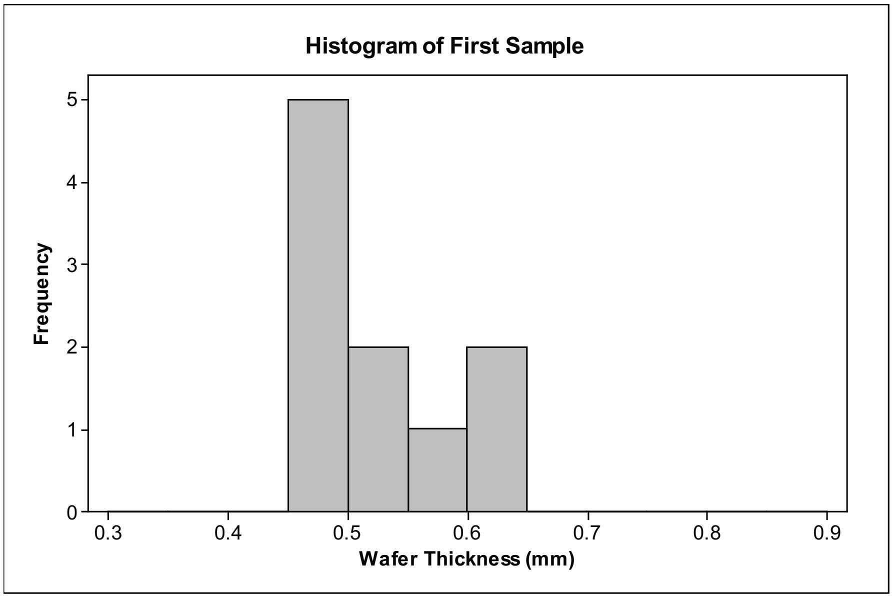
6. Using too many classes can make it difficult to summarize patterns connected with specific values on the horizontal axis. (In other words, you can't see the forest for the trees.) Too few classes can mask important patterns.

UNIT ACTIVITY SOLUTIONS

1. a. Sample answer based on the following sample data (in mm): 0.591, 0.483, 0.489, 0.452, 0.639, 0.523, 0.601, 0.511, 0.498, 0.467.

After the first wafer was measured, a rectangle was drawn above the interval 0.550 to 0.600 since the thickness 0.591 fell between these values. The second, third, and fourth rectangles were stacked on top of each other over the interval 0.450 to 0.500, since 0.483, 0.489, and 0.452 all fell in that interval. The process continued until a rectangle was drawn for each of the 10 measurements.

b. Sample answer: The histogram for the sample data from 1a appears below.



The histogram does not appear to be symmetric.

The interval 0.450 mm to 0.500 mm has the tallest bar and hence more wafers had thicknesses that fell in this interval than any other interval.

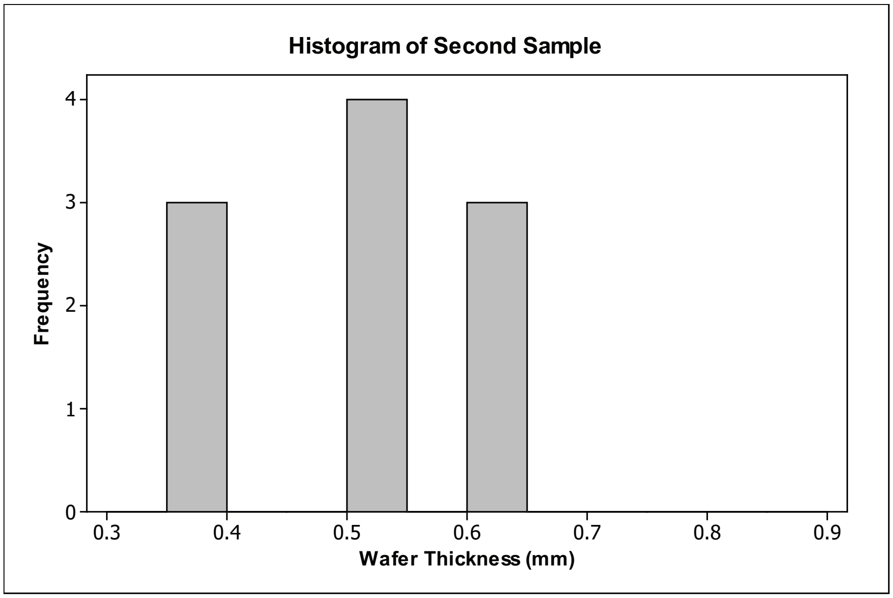
There are no gaps between the bars. However, none of the wafers had thicknesses that fell in the intervals 0.300 to 0.450 and 0.650 to 0.900.

The smallest data value fell in the interval from 0.450 to 0.500 and the largest data value fell in the interval from 0.600 to 0.650.

The thickness 0.5 mm does not appear to be a good choice for summarizing the location of these data. One bar falls to the left of 0.500 mm (the tallest bar) and three bars fall to the right of this value; perhaps 0.525 mm would be a better choice. The controls do not appear to be properly set to produce wafers of consistent 0.5 mm thickness.

2. a. Sample answer data for second sample: 0.389, 0.541, 0.525, 0.621, 0.543, 0.500, 0.638, 0.392, 0.382, 0.602.

b. Sample answer:



The histogram does not appear to be symmetric. (It would be symmetric if the second bar had been closer to the first bar.)

The interval 0.500 mm to 0.550 mm has the tallest bar and hence, more wafers had thicknesses that fell in this interval than any other interval.

There are gaps between each of the bars. None of the wafers had thicknesses that fell in the intervals 0.300 mm to 0.350 mm, 0.400 mm to 0.500 mm, 0.550 mm to 0.600 mm and 0.650 mm to 0.900 mm.

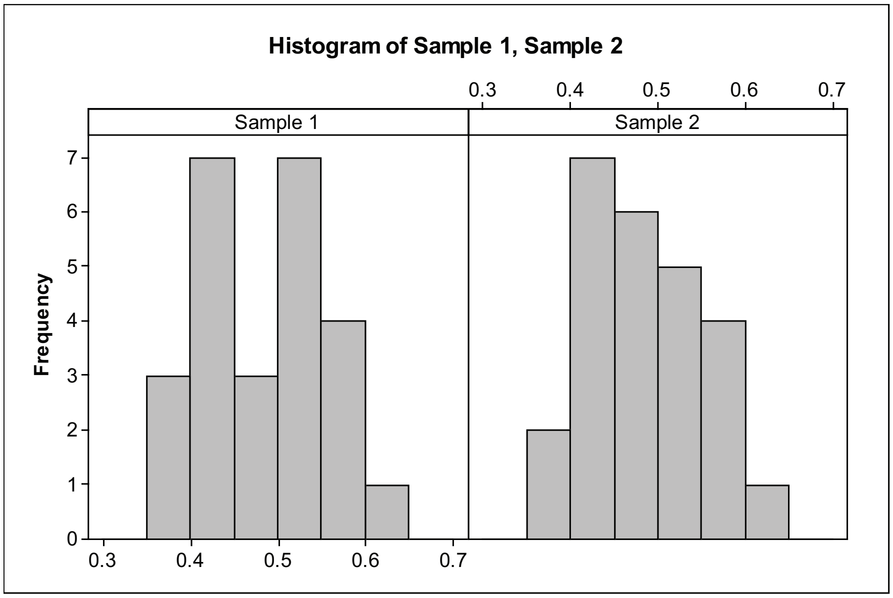
The smallest data value fell in the interval from 0.350 to 0.400 and the largest data value fell in the interval from 0.600 to 0.650.

The thickness 0.5 mm might be a good choice for summarizing the location of these data. It’s the lower endpoint of the interval corresponding to the tallest bar. The outside bars are the same height. Given there is a larger gap between the first and the second bar than there is between the second and the third bar, using the lower value of the middle bar's interval seems reasonable. So, it is somewhat reasonable to assume that the controls are properly set to produce wafers that are fairly consistently close to 0.5 mm in thickness.

3. Sample answer from sample data shown in histogram that follows descriptions of common features and differences.

Common features: Neither histogram is symmetric. In both samples, the data values are spread from 0.35 to 0.65. The highest bar occurs over the interval 0.400 to 0.450 in both histograms.

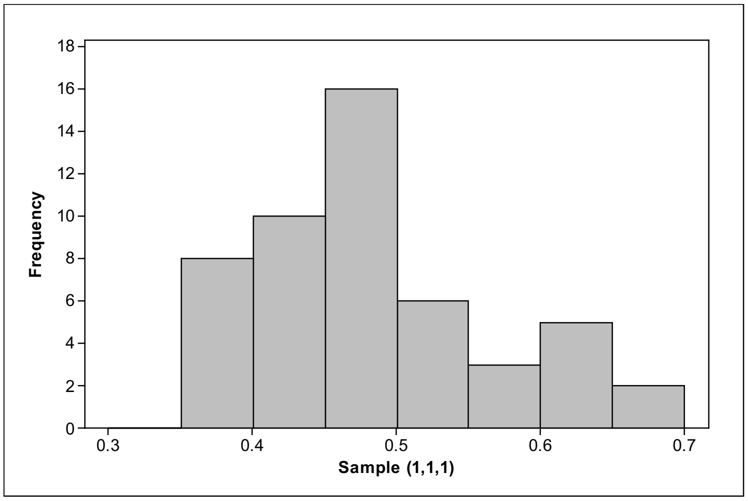
Differences: The two histograms appear different in shape. In the left histogram, the heights of the bars are irregular – down, up, down, up, down. However, in the right histogram, from the second bar to the last bar, the heights of the bars decrease.

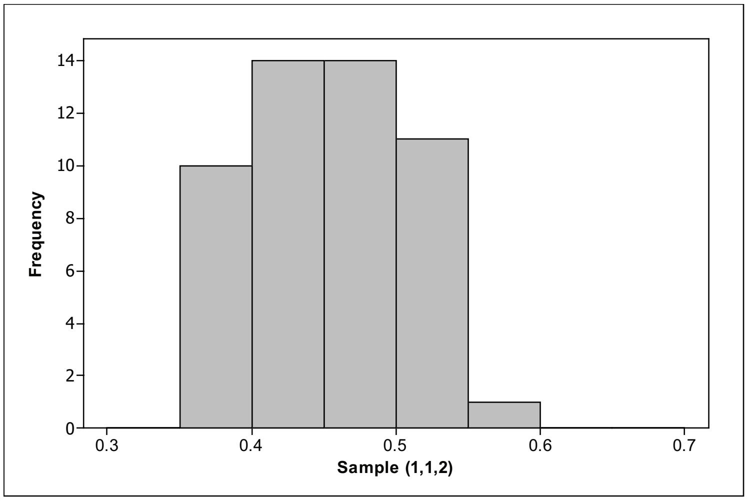


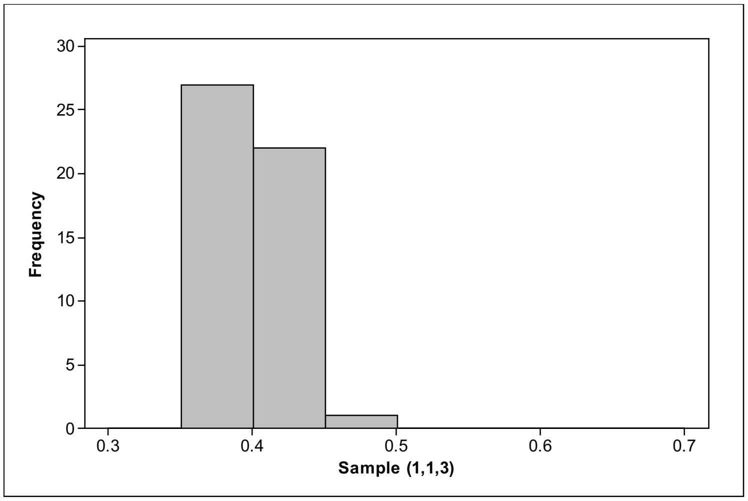
4. a. Sample: Change one control setting at a time and compare histograms to see what has changed. For example, start with the following settings: Control 1 = 1, Control 2 = 1 and Control 3 = 1. Change Control 3 from 1 to 2 to 3 and describe the change. Then choose different settings for Controls 1 and 2 and repeat the process described above. See if the observed pattern remains the same. If so, describe how the settings of Control 3 affect wafer thickness in sample data.

Adapt the strategy described above to determine how Controls 1 and 2 affect the thickness of wafers.

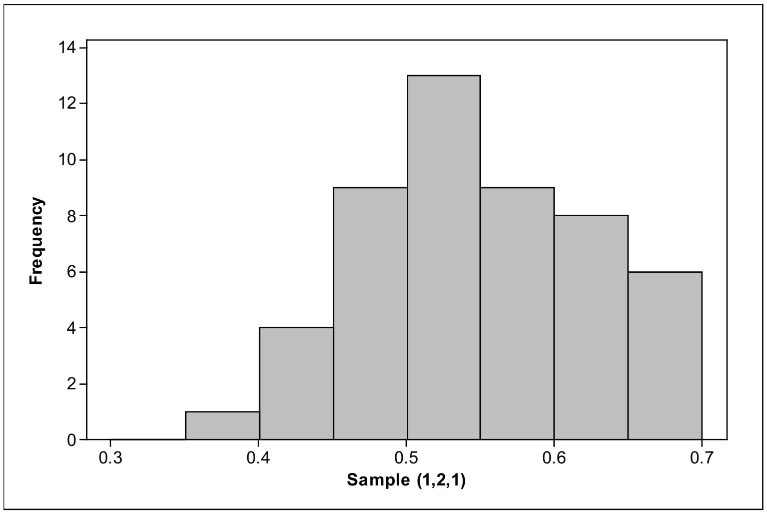
b. Sample answer: Samples were collected with Control 1 = 1 and Control 2 = 1 and then changing Control 3 from 1 to 2 to 3. The notation (1, 1, 1), (1, 1, 2), and (1, 1, 3) is used to identify the three control settings. In the histograms below, the most apparent change appears to be to in the spread of the data. The data are least spread out (thicknesses are most consistent) when Control 3 = 3. (It is almost as if the right tail shrinks as the level of Control 3 is increased.)

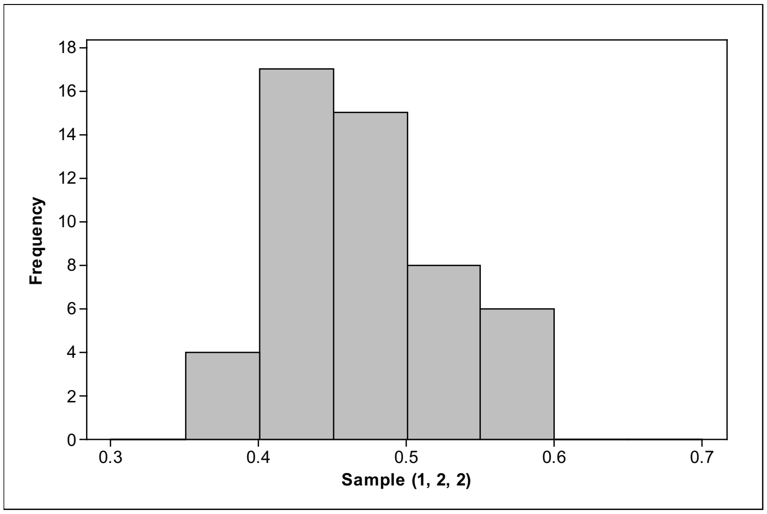


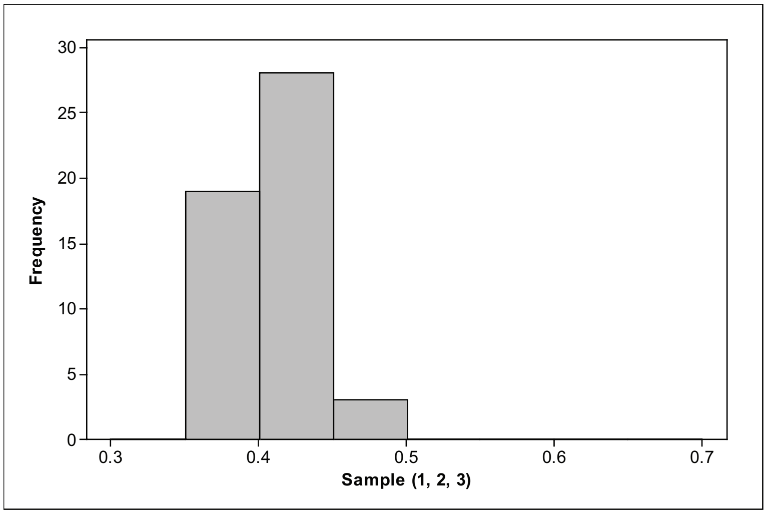




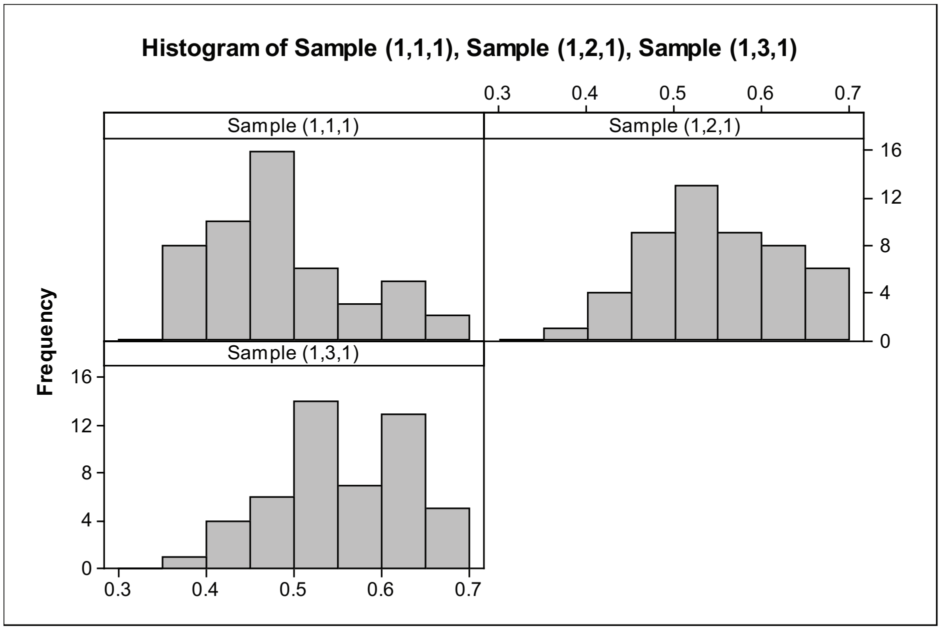
Next, we kept Control 1 = 1, set Control 2 = 2, and then changed Control 3 from 1 to 2 to 3. The histograms appear below. The same pattern of reduced spread occurred. The data are more consistent (less spread out) when Control 3 = 3.



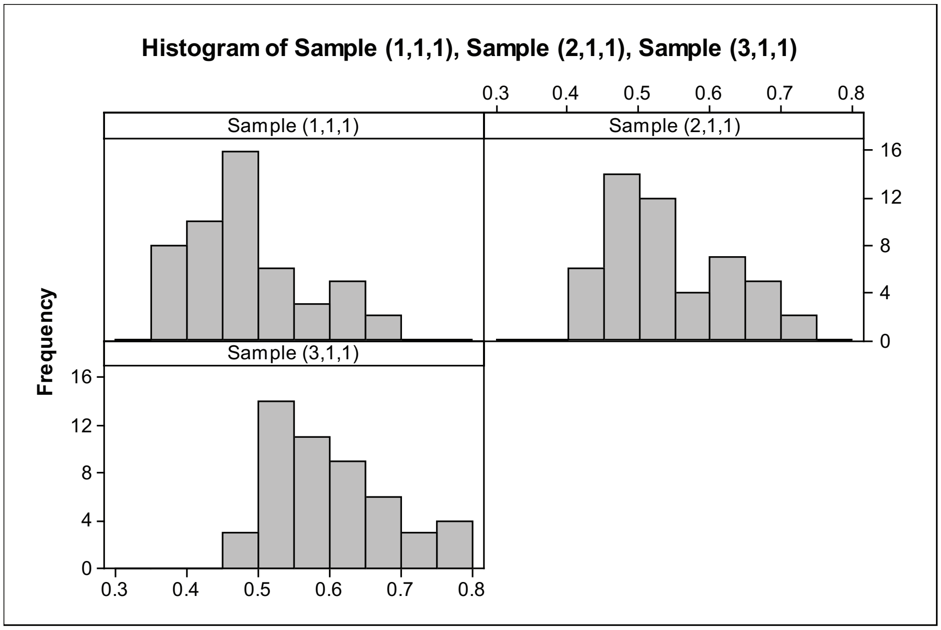




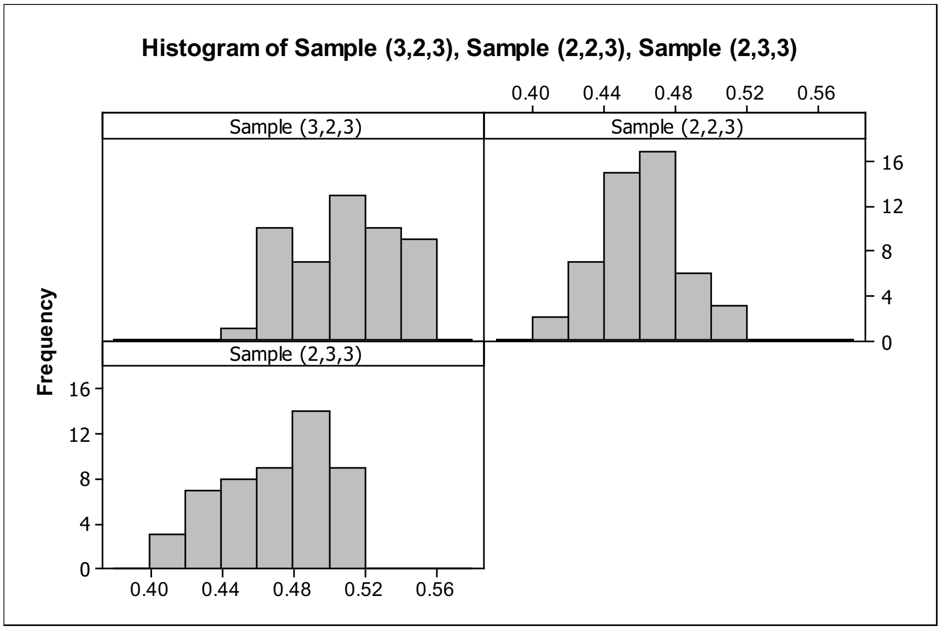
Next, we focus on the effect of Control 2. The histograms below compare settings (1,1,1), (1,2,1) and (1,3,1). When Control 2 = 1, the data appear more concentrated to the left. When Control 2 = 2, the data appear more symmetrical, and when Control 2 = 3, the data appear more concentrated to the right. So, we conclude that Control 2 affects the shape of the data.



Last, we change the settings for Control 1, leaving settings for Control 2 and Control 3 fixed. Below are histograms for samples from settings (1, 1, 1), (2, 1, 1) and (3, 1, 1). Changing the settings on Control 1 from 1 to 2 to 3 appeared to shift the bars in the histogram to the right – hence, increasing the thicknesses.

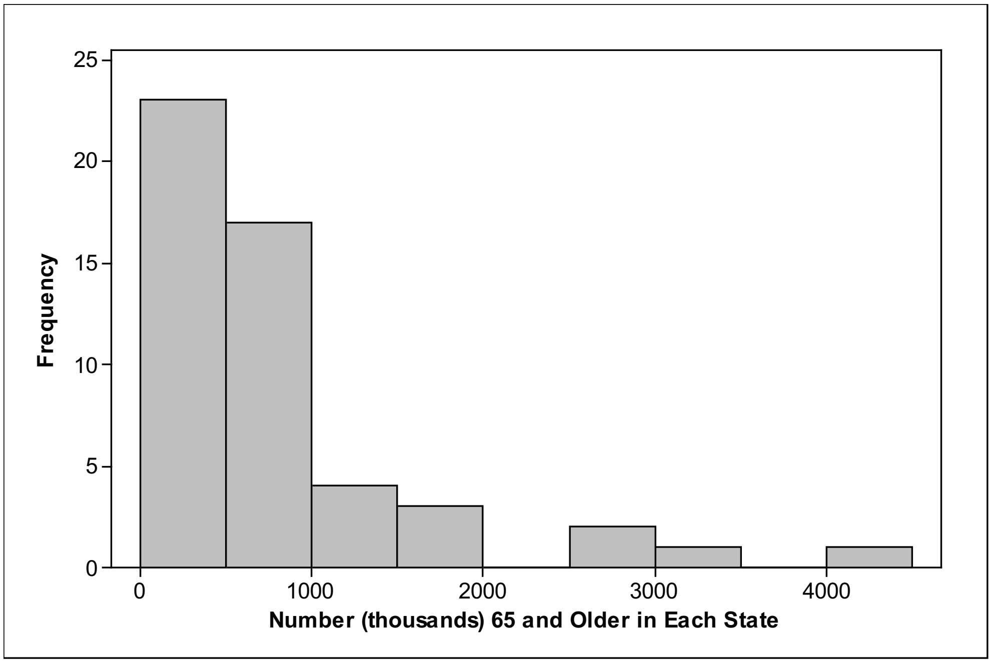


5. Sample answer (student answers will vary): We recommend settings (3,2,3). We chose Control 3 = 3 to reduce variability. We chose Control 2 = 2 so that we had balance between high and low values. Finally, we chose Control 1 = 3 to increase the thickness. We compare this choice of settings with (2,2,3) and (2,3,3) in the histograms below.

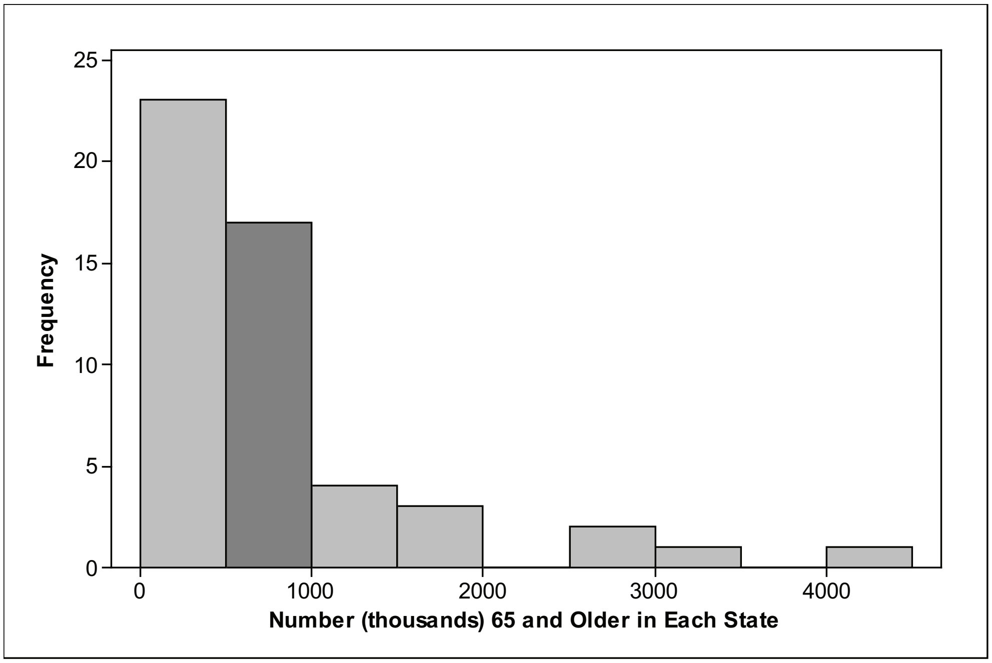


EXERCISE SOLUTIONS

1. a.

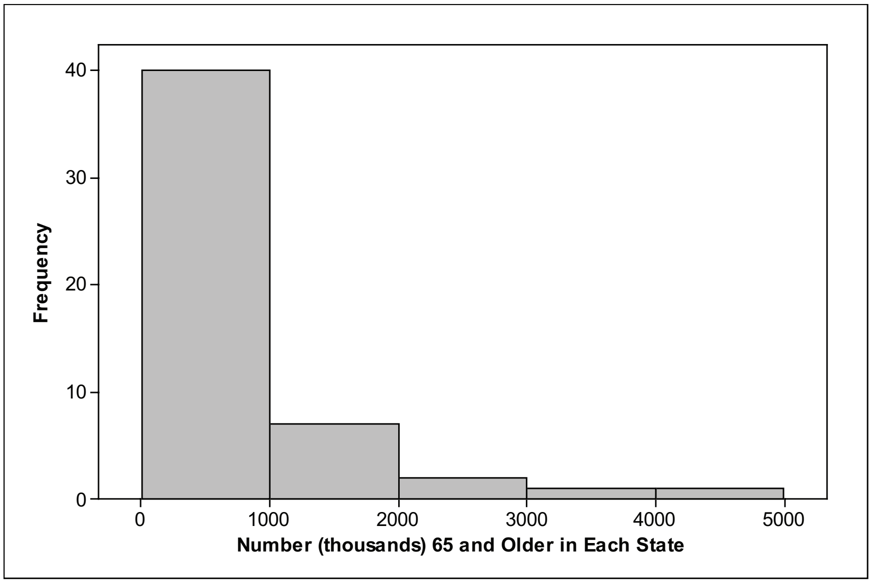


b. Sample answer (assuming the student's home state is Massachusetts): For Massachusetts, there were 903 thousand people 65 or older in 2010. Massachusetts' population of 65 and over appears to be fairly typical.

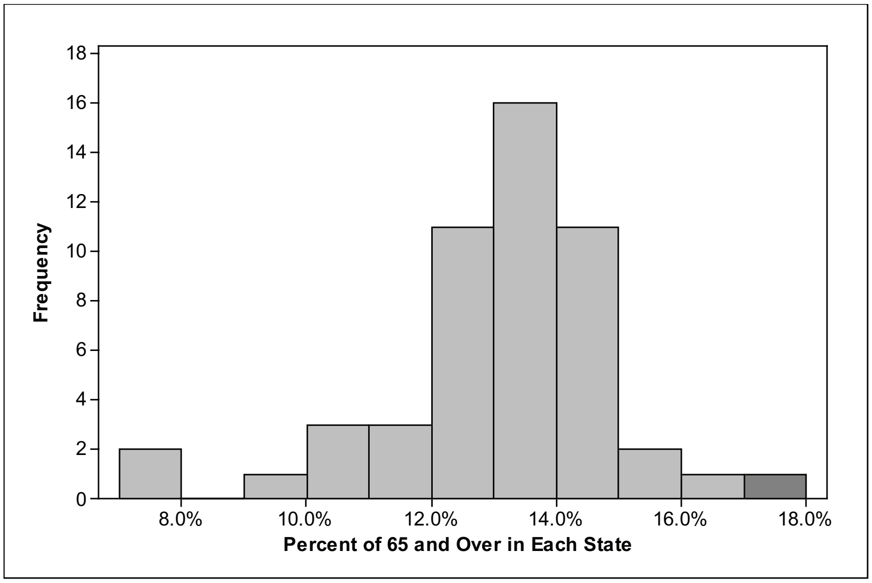


c. Sample answer: The distribution is skewed to the right. There are two gaps – one between 2,000 thousand and 2,500 thousand and the other between 3,500 thousand and 4,000 thousand. California with 4,247 thousand people 65 or over could be an outlier. Florida with 3,260 thousand, New York with 2,618 thousand, and Texas with 2,602 thousand might also be outliers (or they could simply be the tail of the overall pattern in the distribution).

d. In the histogram below, the gaps in the data are hidden. However, you still can observe an overall pattern that is skewed to the right.

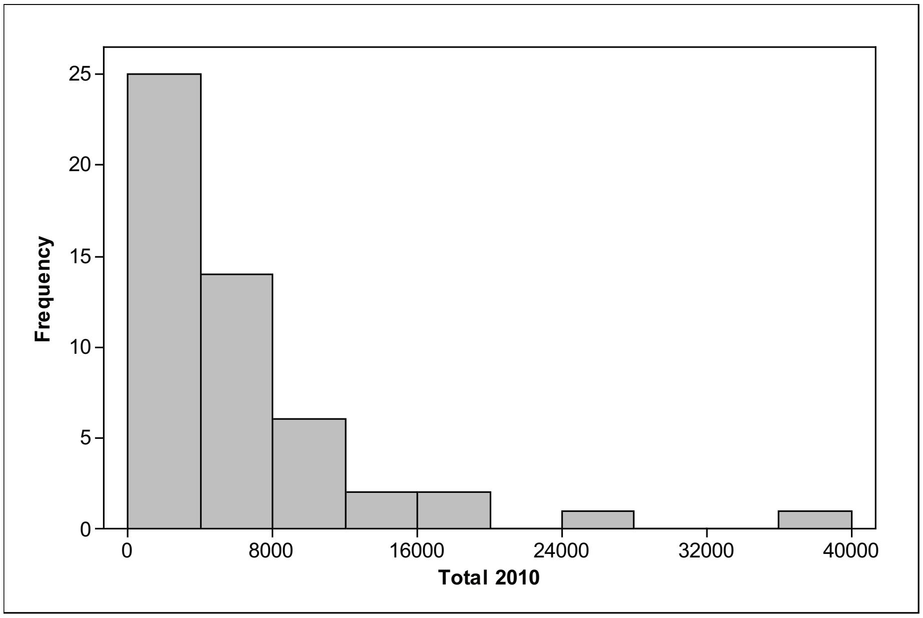


2. a. Sample answer (this time assuming the student's home state is Florida): For Florida, 17.3% of the people were 65 or older. Florida has a higher percentage of people 65 or older than all other states and the District of Columbia.



b. Sample answer: The overall pattern is roughly symmetric. There is a small gap – there are no percentages between 8% and 9%. South Carolina (7.8%) and Alaska (7.7%) might be outliers. However, they really don't appear to be unusual values – the gap is small and these values are at the upper end of the class interval from 7% to 8%.

3. a. Sample answer (students could have made other choices for the class sizes):

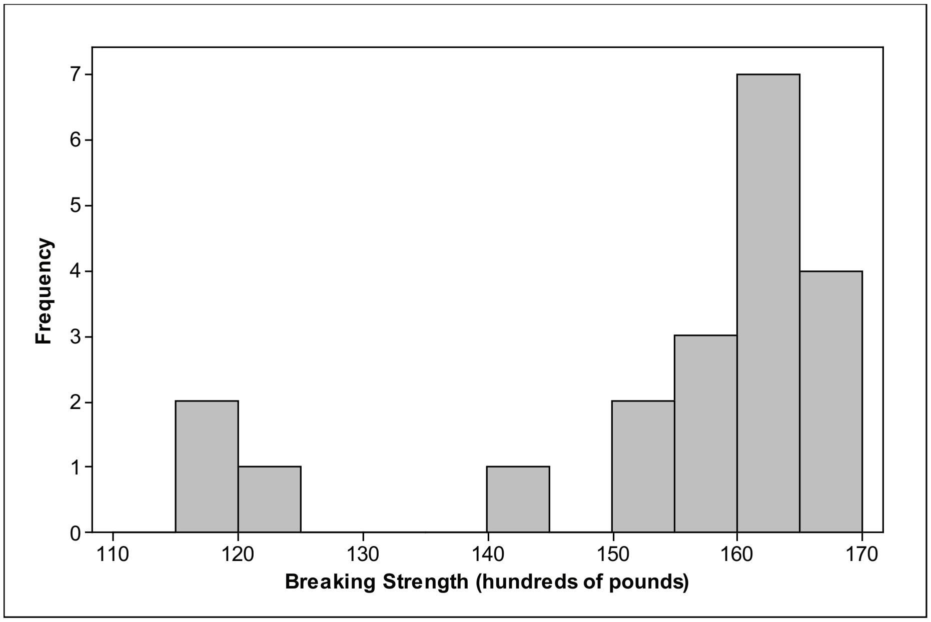


b. Sample answer: The overall pattern of the distribution of states' population sizes is skewed to the right. There are two gaps in the data, one between 20,000 thousand and 24,000 thousand and the other between 28,000 thousand and 36,000 thousand. California (37,254 thousand) is definitely an outlier. In addition, Texas (25,146 thousand) is a potential outlier.

4. a. Yes.

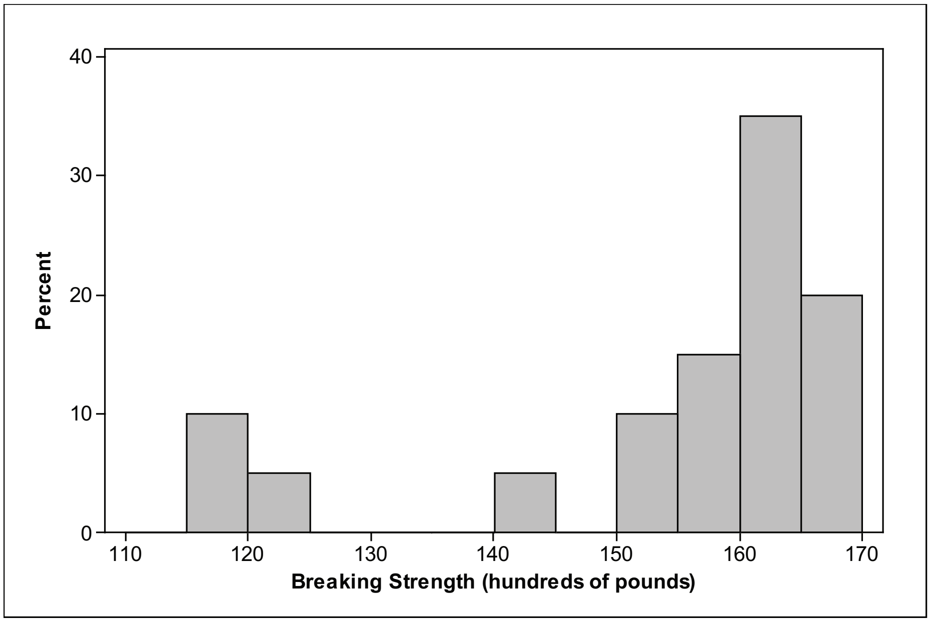
Sample explanation: If you look at the breaking strengths recorded in the first column, all the entries are different. In fact, all but four of the breaking strengths are distinct. So, breaking strength varied from stake to stake even though the stakes were nearly identical.

b.



c. The interval from 160 to 165 contained the most data.

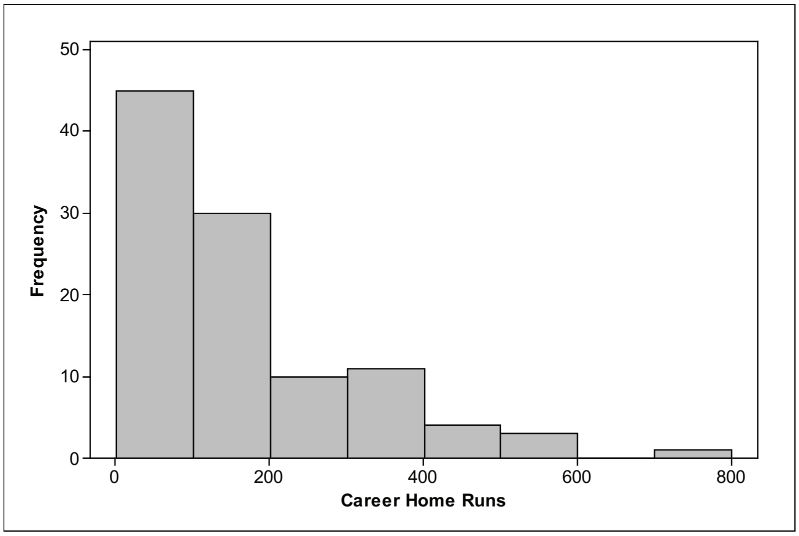
d. The histogram below looks exactly the same as the histogram in (b) – it has the same shape, the same gaps, and the same potential outliers. The only thing that changed was the scaling on the vertical axis.

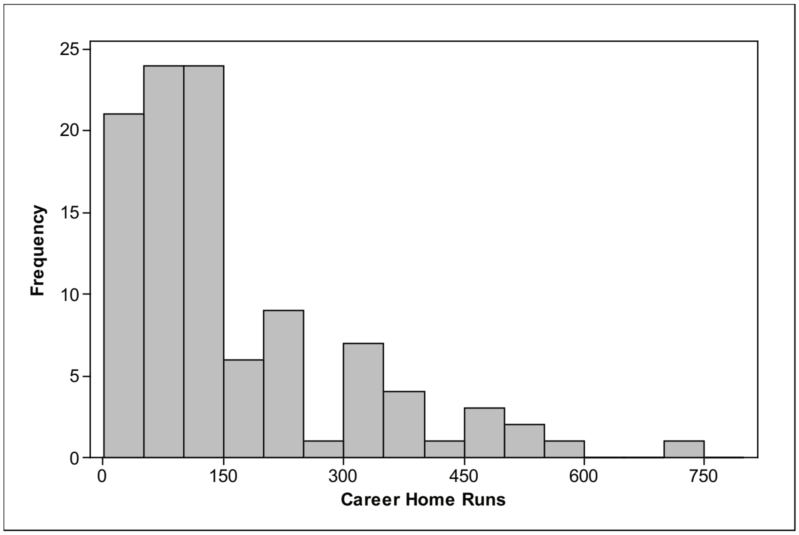


e. Sample answer. The overall pattern in the data is skewed to the left. The three data values between 115 and 125 represent a departure from the overall pattern and are sufficiently far from the rest of the data that they may be considered outliers. There are three class intervals containing no data that separate these potential outliers and the rest of the data.

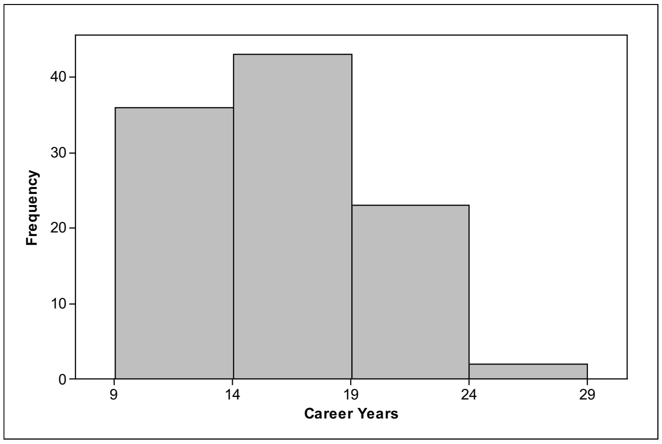
REVIEW QUESTIONS SOLUTIONS

1. Sample answer: The overall pattern in the first histogram is skewed to the right. There is a gap between 600 and 700 and one outlier between 700 and 800. The outlier is Babe Ruth's record 714 career home runs. Although the pattern in the second histogram could still be described as skewed to the right (because the tail of the data on the right is stretched out), the pattern is more jagged compared to the first histogram. There are a few secondary peaks and valleys apparent in the second histogram, which are not visible in the first histogram. Also interesting is the fact that the data values in the second class interval (100 to 200) of the first histogram are not evenly distributed when that class interval is divided in half, 100 to 150 and 150 to 200. There are 24 data values in the class interval 100 to 150 but only one-quarter as many from 150 to 200. A similar pattern holds when the class interval from 200 to 300 is divided into two class intervals.

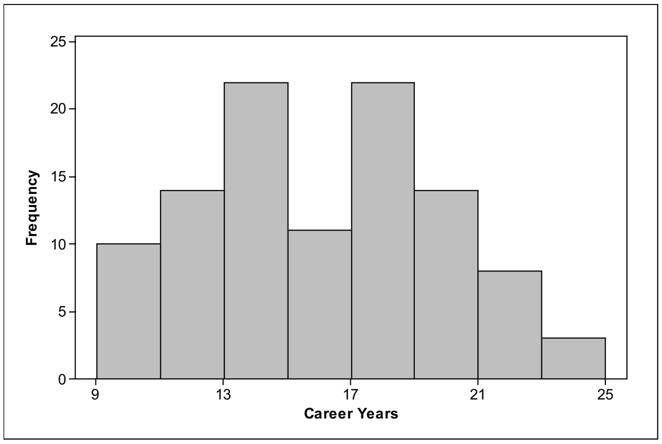




2. a.



Histogram 1



Histogram 2

b. Sample answer: Yes, for example Rod Carew had 19 career years. In the first histogram, his data value was classified in the class 19 – 24 and in the second histogram it was classified in the class 19 – 21.

c. Histogram 1: The shape appears unimodal and skewed right. Histogram 2: The shape appears bimodal and roughly symmetric. Changing the class intervals had a big effect on the overall shape.

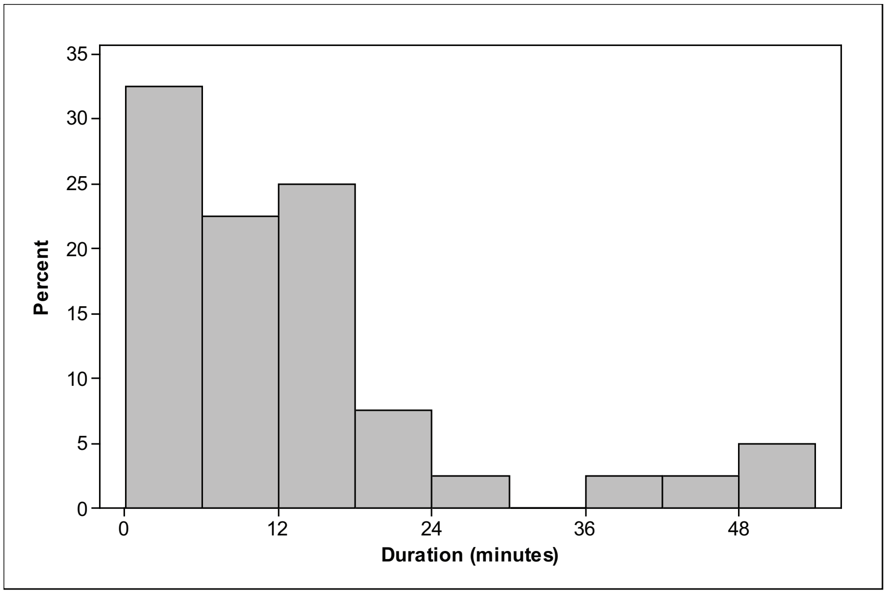
3. a.

Duration (minutes)	Frequency	Percent
0 – 6	13	32.5
6 – 12	9	22.5
12 – 18	10	25
18 – 24	3	7.5
24 – 30	1	2.5
30 – 36	1	2.5
36 – 42	0	0
42 – 48	1	2.5
48 – 54	2	5

b. 55%

c. 10%

d. *(See histogram on next page...)*



e. Sample answer: The distribution is skewed to the right. There is a gap between 30 and 36 minutes. There are two distinct groups of phone calls, those lasting under 30 minutes and a few lasting 36 or more minutes.

Unit 4: Measures of Center

PREREQUISITES

Students should be able to identify whether distributions are roughly symmetric or skewed given a histogram (Unit 3). The only mathematics prerequisite is knowledge of basic arithmetic operations (ordering, addition, division) needed to calculate the mean and median. Briefly introduce summation notation, $\sum x$, if that notation is new to students.

ADDITIONAL TOPIC COVERAGE

Additional coverage of measures of center can be found in *The Basic Practice of Statistics*, Chapter 2, Describing Distributions with Numbers.

ACTIVITY DESCRIPTION

The purpose of this activity is to help students learn how to assess the relationship between the mean and median based on the shape of the distribution. Students work with the Stemplots interactive from the Interactive Tools menu. The Stemplots interactive generates data and then organizes it into stemplots. Students use information from the graphic display to guess which is larger, the mean or the median. Then they calculate the mean and median. The interactive allows them to check their answers.

MATERIALS

Students will need access to the Stemplots tool from the Interactive Tool’s menu online.

THE VIDEO SOLUTIONS

1. The variable is the weekly wages for Americans, separated by gender.

2. The men's distribution is skewed to the right.

3. The medians of the two distributions differ. Median measures the 50-50 point. The median for men's wages was larger than the median for women's wages.

4. A few very large incomes inflate the mean of a group of incomes. Hence, these very large incomes would pull the mean up.

5. A few very large incomes have no effect on the median.

UNIT ACTIVITY SOLUTIONS

1. a. Sample answer (answers differ since the data are randomly generated): For the stemplot below, the mean should be smaller than the median. The few really small test scores should pull the mean down.

1	6
2	
3	56
4	
5	
6	12
7	4
8	235
9	5

b. There were 10 test grades: (10 + 1)/2 = 5.5. So, the median = (62 + 74)/2 = 68. To find the mean, the sum of the test scores is 629. So, the mean = (629)/10 = 62.9. The median is larger than the mean.

2. a. Sample answer: The mean and median should be fairly close. However, the graph is somewhat skewed to the left, so the mean should be lower than the median.

3	26
4	447
5	1258
6	6

b. median = (47 + 51)/2 = 49; mean = 485/10 = 48.5. The median is larger than mean.

3. a. Sample answer:

1	389
2	3
3	
4	2
5	268
6	1
7	59
8	88
9	
10	
11	
12	
13	256
14	3

b. The mean should be larger because there are 4 extremely large test scores in comparison to the other test scores. Those large scores will pull the mean up but not affect the median.

c. median = 61; mean = 71.6. The mean was larger than the median.

4. a. Sample answer:

5	19
6	116
7	256
8	33445
9	57
10	38

b. The mean should be somewhat smaller than the median because the distribution is skewed to the left.

c. median = 83 and mean = 79. The median was larger than the mean.

5. a. The plot has two peaks, one in the 30s and the other in the 60s.

3	234555555566666666667788999
4	0369
5	08
6	01233444555667779999999999
7	2233444789
8	0

b. median = 61.5; mean ≈ 53.3; there are two modes, one at 36 and the other at 69.

c. The median locates the upper peak, but does nothing to summarize the location of the lower peak. The mean is located where there is little data, and is not close to identifying the location of either of the peaks where there is a lot of data. Using the two modes gives the locations of the two peaks. So, in this case, the modes would be the best choice to describe the location of these data.

6. a. The plot appears roughly symmetric, with a single peak.

6	055
7	0022244445555566668888
8	0055
9	0

b. median = 75; mean = 75.1; mode = 75.

c. In this case, it doesn’t really matter which of the three numeric descriptors for the center you choose. They are all about the same.

EXERCISE SOLUTIONS

1. a. Either a histogram or a stemplot would be a good choice of graphic display.
Sample answer based on stemplot below: The distribution is skewed to the right. There is a gap in the 80 thousands, which makes the $90,000 salary appear to be an outlier.

4	89
5	00134569
6	345689
7	678
8	
9	0

b. (59 + 63)/2 = 61; median starting salary is $61,000.

c. (1241/20) = 62.05; mean starting salary is $62,050.

d. The mode is 50 thousand.

e. The mean starting salary is higher than the median. That's largely due to the $90,000 outlier but also due to the shape of the data, which is skewed to the right. The mode does not do a good job in measuring the center or location of these data.

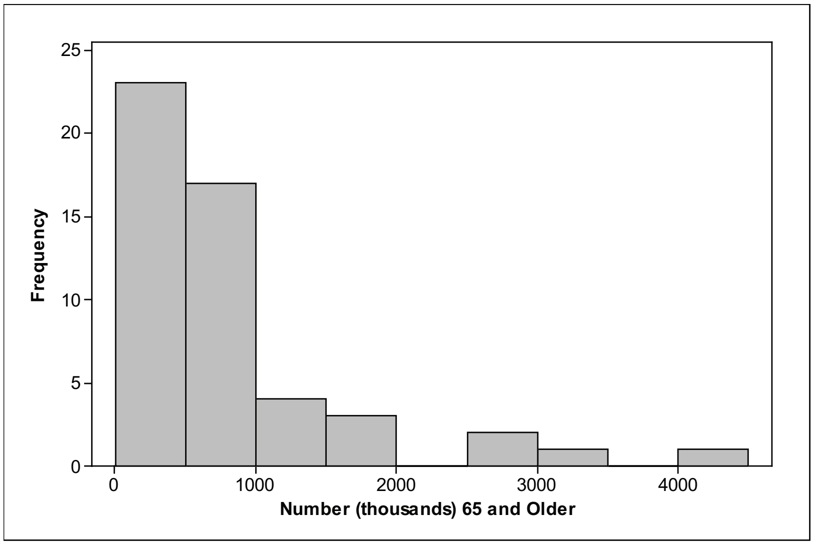
2. The median was $256,900 and the mean is $295,300. The mean is inflated because of a few extremely expensive houses, houses with prices in the millions.

3. a. Mean for all 10 years: (13 + 23 + 26 + 16 + 33 + 61 + 28 + 39 + 14 + 8)/10 = 261/10 = 26.1.
Mean (excluding 61): (13 + 23 + 26 + 16 + 33 + 28 + 39 + 14 + 8)/9 = 200/9 ≈22.22.
Omitting his record year lowers the mean by 3.88.

b. All 10 years: ordered data 8 13 14 16 23 26 28 33 39 61; median = (23 + 26)/2 = 24.5.
Excluding 61: ordered data 8 13 14 16 23 26 28 33 39; median = 23.
Omitting his record year lowers the median by 1.5. Hence, the median is less affected by the record number of home runs than the mean.

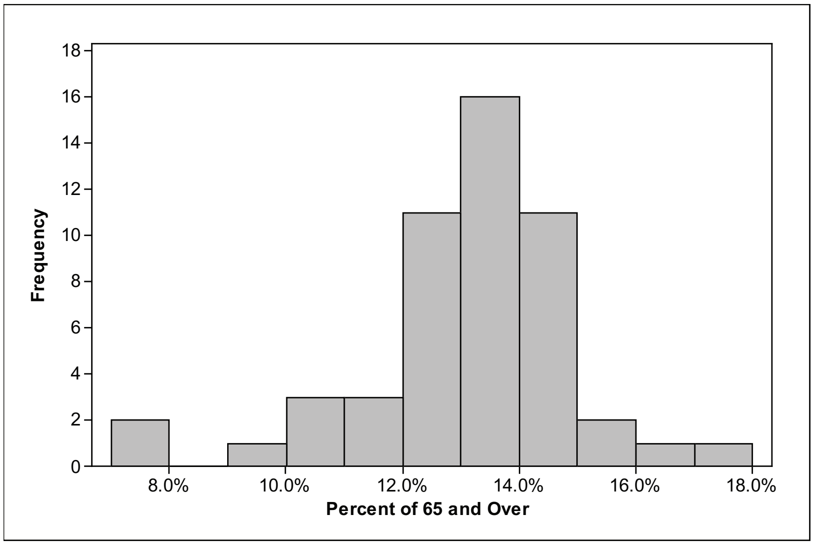
c. Sample answer: The mean overstates Maris's usual performance because of the influence of the outlier. But the median doesn't point to the great achievement of his career. Perhaps we should say “Maris hit 61 home runs in 1961, and averaged about 22 home runs a year in his other 9 years in the American League.”

4. a. The mean is 784,000 and the median is 534,000. Note from the histogram (shown below) that these data are skewed to the right with some possible outliers.



The skewed pattern and outliers inflate the mean. The better choice for describing the location of the 65 and older data is the median.

b. Because the histogram is roughly symmetric, the mean and median should be close in value. The mean is about 13.14% and the median is 13.51%. (The mean is slightly less because of the two percentages below 8%.) In this case, either the mean or median would be appropriate. However, from looking at the histogram below, the median looks slightly more central than the mean.



REVIEW QUESTIONS SOLUTIONS

1. a.

5	66
5	
6	0
6	2
6	4
6	
6	8888
7	000000
7	2222222
7	44444
7	66
7	88

b. Mean ≈ 70.1 beats/min; median = 72 beats/min; mode = 72 beats/min.

c. Sample answer: The median of 72 beats/min best describes a “typical” pulse rate for this man. In addition, the mode is also 72 beats/minute. (The mode is the man’s most frequent pulse rate.) There are a few days when the man’s pulse rate is very low. These low values tend to pull the mean down.

2. a. Approximatly 56 of the fish had mercury levels below 0.30 µg/g.

b. Approximately 27 of the fish from the sample had mercury levels at or above 0.30 µg/g. Hence, around 32.5% of the fish in the sample had levels of mercury concentration above the EPA guidelines.

c. Because the data are skewed to the right, the few high mercury concentration values in the tail will inflate the mean but not affect the median. Hence, the mean mercury concentration will be larger than the median mercury concentration.

3. a. To compute the mean, sum the data and divide by 25: $\bar{x}$ = 1103/25 = 44.12.

To compute the median, order the data from smallest to largest. Select the (25 + 1)/2, or 13th data value from the ordered list:

28	35	37	37	38	38	40	40	42	43	43	44	45
45	46	46	46	47	47	49	49	51	54	55	58	

The median is 45 fries in a bag.

b. A stemplot appears below. We've chosen one that divides the stem into increments of 5.

2	8
3	
3	57788
4	002334
4	556667799
5	14
5	58

c. In this case, the choice between the mean and median is a matter of judgment. The difference between the two is less than one fry per bag. Some students may prefer the median because of the potential outlier of 28, which drags the mean down slightly.

Unit 5: Boxplots

PREREQUISITES

Students should be able to use a stemplots (Unit 2) to aid in ordering data from smallest to largest. Given an ordered set of data, they should be able to compute the median, which was covered in Unit 4, Measures of Center.

ACTIVITY DESCRIPTION

The Unit 5 activity returns to the data collected in the survey for Unit 2's activity. In Unit 2's activity, the sample data, used for the sample solutions, were created rather than collected from a class, with the exception of the height data. If you chose not to collect the data from your class, let students use the sample data from Unit 2 to complete Unit 5's activity.

Students can compare the stemplots that they created for Unit 2's activity to the boxplots from this activity. If students are drawing the boxplots by hand, the stemplots will help them order the data from smallest to largest. Students are asked to complete around 11 boxplots for this activity. So, it may be best to use software to create the boxplots. Otherwise, students should work in small groups so that they can split up the work of constructing the boxplots among group members.

THE VIDEO SOLUTIONS

1. The different brands of hot dogs were compared by their calories.

2. The one-quarter point is called the first quartile.

3. The values in a five-number summary are the minimum, first quartile (Q_1), median, third quartile (Q_3), and maximum.

4. The interquartile range or IQR = $Q_3 - Q_1$.

5. The median of the poultry hot dogs is below the minimum for the beef hot dogs. So, half of the brands of poultry hot dogs have fewer calories than the lowest calorie brand of all-beef hot dogs.

UNIT ACTIVITY:

USING BOXPLOTS TO ANALYZE DATA SOLUTIONS

Unit 5 solutions are based on sample data from Unit 2's activity given in Table T5.1.

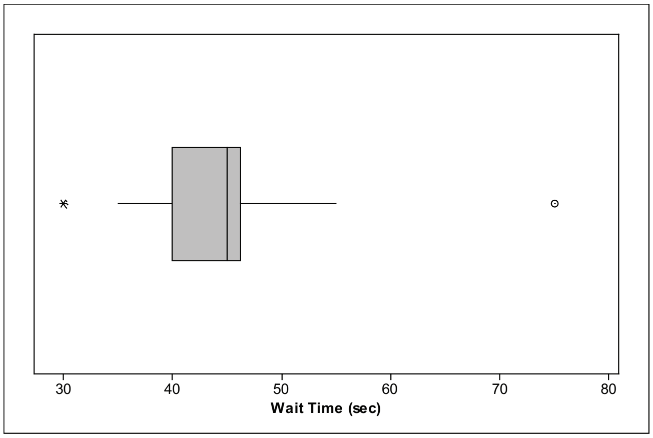
Question 1 Wait (sec)	Question 2 Coins (cents)	Question 3 Height (in)	Question 4 Study (min)	Question 5 Exercise (min)	Question 6 Gender
40	77	68	30	75	Male
40	62	67	60	20	Female
75	175	73	30	0	Male
40	189	72	20	45	Male
50	120	71	15	90	Male
40	54	68	45	75	Female
45	26	66	60	30	Female
45	145	75	30	30	Male
40	0	69	120	0	Female
35	0	71	45	60	Male
45	35	72	30	30	Male
45	47	64	15	45	Female
45	125	72	20	90	Male
45	55	71	30	0	Male
40	35	69	60	45	Male
45	78	63	45	90	Female
55	157	65	45	20	Male
40	225	62	75	40	Female
40	92	64	30	60	Female
50	85	62	60	0	Female
35	35	64	45	30	Female
45	59	66	45	90	Female
50	145	60	30	45	Female
30	137	70	30	30	Male
50	142	69	20	45	Male
45	62	69	30	60	Male

Table T5.1. Sample data for Unit 2 Activity survey questions.

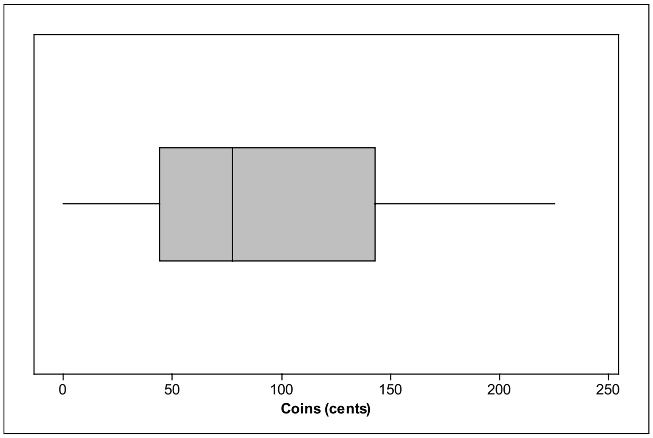
Minitab was used to create the boxplots for the sample answers. Hand drawn boxplots may differ slightly.

1. Sample answers:

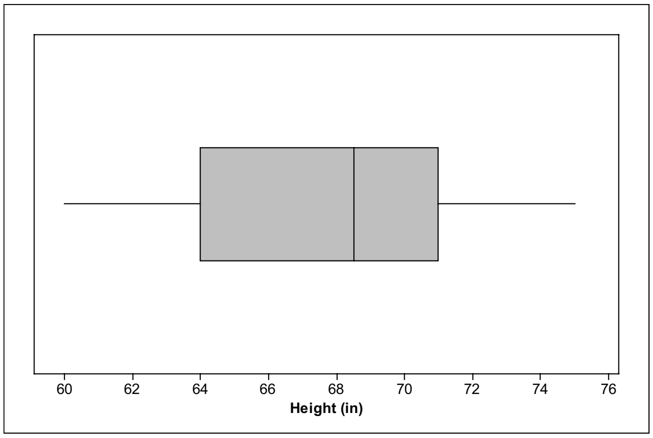
Question 1 – Wait Time: There were two outliers, a mild outlier on the low side and an extreme outlier on the high side. The times in the third quarter of the data are really concentrated compared to the times in the first, second and fourth quarter.



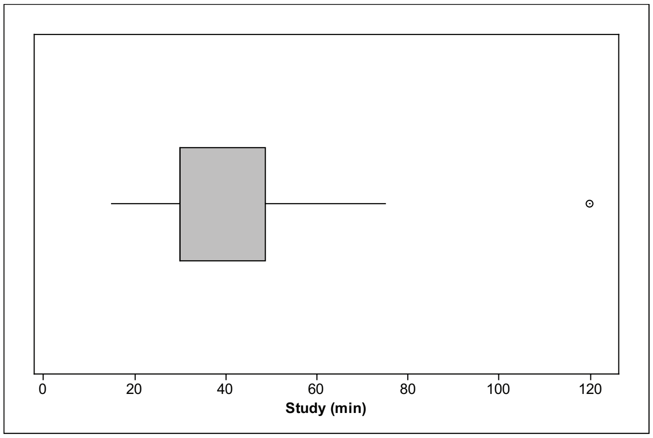
Question 2 – Coins: The amount of money in coins ranged from 0 cents to 225 cents. The data appear to be right skewed.



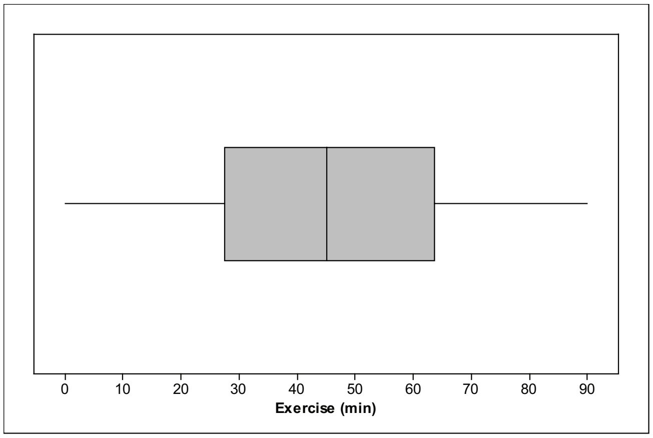
Question 3 – Height: The lower quarter of the heights and the upper quarter of the heights appear to have similar spread. However, the second quarter of the heights appear to be about twice as spread out as the third quarter of the heights.



Question 4 – Study Time: There was one person in class who claimed to study, on average, for 120 minutes per exam. That turned out to be an extreme outlier. However, there doesn't appear to be any dividing line in the box. That is because the first quartile and median were both 30. In fact, there were 9 students who claimed that they studied, on average, 30 minutes for an exam.

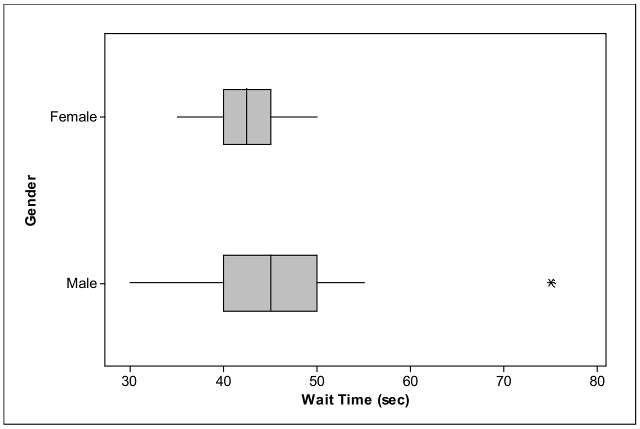


Question 5 – Exercise: The exercise times look pretty symmetric. Even though some people exercised on average for 90 minutes and others did not exercise, neither extreme turned out to be an outlier.

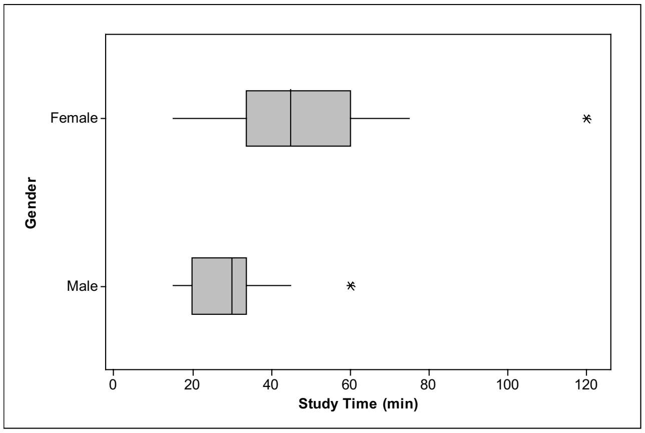


2. Comparative plots for how long female students felt they waited compared to male students are shown below. The outlier for the males turns out to be a mild outlier. The median time for the females was lower than for males. The data for the males were considerably more spread out than for the females, particularly if the range was used to describe the spread.

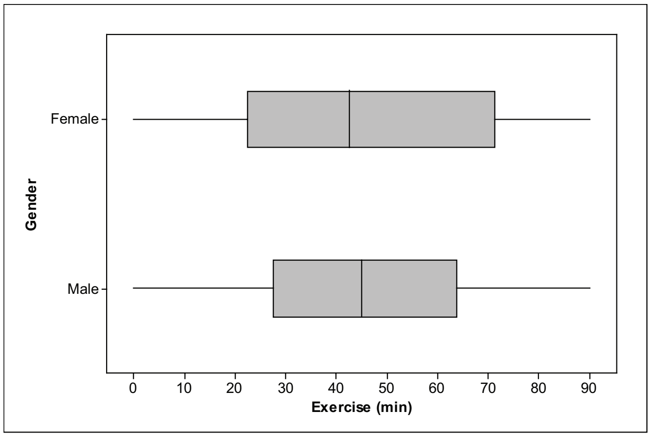
(See boxplot on next page...)



3. The study times for females were more spread out than for males. Using Minitab's algorithm for computing quartiles, the first quartile for the females equaled the third quartile for the males. Therefore, three-quarters of the females spent longer studying than three-quarters of the males. Both of the outliers are mild outliers, even though we had thought that the study time outlier for females would be an extreme outlier. That is most likely due to the larger IQR for the female study times.



4. The overall spread of the two data sets is the same, with a range of 90 minutes. The inner 50% of the data for the males is more concentrated than for females. The median for the males is very slightly higher than for females. Both distributions are roughly symmetric.



EXERCISE SOLUTIONS

1. a. First, the data need to be ordered from smallest to largest:

111	131	132	135	139	141	148	149	149	152
153	157	158	175	176	181	184	186	190	190

The median is computed by averaging the 10th and 11th data values:

median = (152 + 153)/2 = 152.5

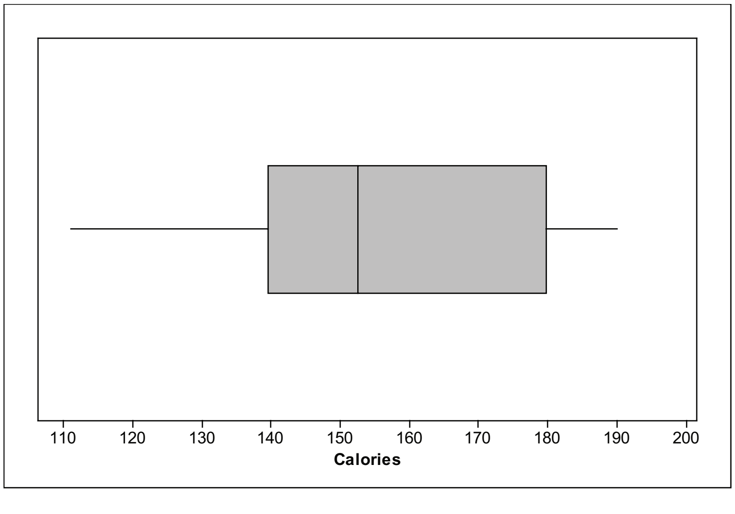
Q_1 is the median of the lower half of the data (top row): Q_1 = (139 + 141)/2 = 140
Q_3 is the median of the upper half of the data (bottom row): Q_3= (176 + 181)/2 = 178.5
5-number summary: 111, 140, 152.5, 178.5, 190

(Note: If quartiles are computed using Minitab, Q_1 = 139.50 and Q_3= 179.75. If Excel is used to compute the quartiles, Q_1= 140.5 and Q_3= 177.25. In all three cases, 5 data values or 25% of the data fall below Q_1. Similarly, 75% of the data fall below Q_3, regardless of whether Minitab's, Excel's or the hand-calculated value of is used.)

b. range = 190 – 111 = 79; IQR = 178.5 – 140 = 38.5. The range gives the spread between the minimum and maximum data value. The IQR tells how spread out the middle half of the data are.

c. No; 175 is below Q_3, the cutoff for the top quarter of the data.

2. a.



b. The second quarter (represented by left section of box) and the fourth quarter (represented by right whisker) appear close in length. Data values in the second quarter lie between 140 and 152.5, a distance of 12.5; data values in the fourth quarter lie between 178.5 and 190, a distance of 11.5. So, the data in the fourth quarter are the most concentrated.

c. The first quarter (represented by the left whisker) and third quarter (represented by right section of box) appear equally spread. Data values in the first quarter lie between 111 and 140, a spread of 29; data values in the third quarter lie between 152.5 and 178.5, a spread of 26. Hence, data in the first quarter exhibit the most spread.

d. 1.5 × 38.5 = 57.57; Q_1 – 57.75 = 82.25 and Q_3 + 57.57 = 236.25. None of the beef hot dogs in the sample had calories below 82.25 or above 236.25. Therefore, there are no outliers.

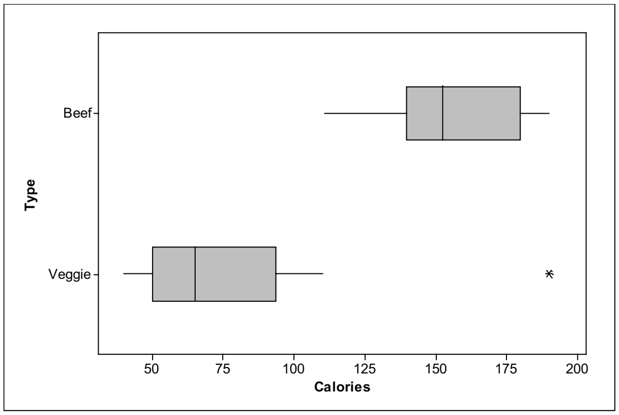
3. The stemplot appears below. Notice that there was one brand of beef hot dog that had low calories compared to the other brands. According to the 1.5 × IQR rule, this value was not sufficiently small compared to the rest of the data to be classified as an outlier. Ignoring that value, it appears that the beef hot dogs fall into two categories separated by a gap. The lower-calorie hot dogs have between 131 and 158 calories; the higher-calorie hot dogs have between 175 and 190 calories. This gap that appears to divide the beef hot dogs into two categories on the stemplot is not visible in the boxplot.

11	1
12	
13	1259
14	1899
15	2378
16	
17	56
18	146
19	00

4. a. Five-number summary: 40, 50, 65, 92.5, 190

b. For the veggie dogs: IQR = Q_3 - Q_1 = -42.5; step = 1.5 × 42.5 = 63.75
Inner upper fence: Q_3 + 1 step = 92.5 + 63.75 = 156.25
Outer upper fence: Q_3 + 2 steps = 220
Hence, 190 is a mild outlier.
See (c) for the modified boxplot. (The upper whisker ends at 110.)

c. Boxplots comparing calories of beef and veggie dogs.



d. As a group, the veggie dogs appear to have fewer calories than the beef dogs. The third quartile (right end of the box) for the veggie dogs is below the minimum calories for the beef dogs. Hence, at least three-quarters of the veggie dogs have fewer calories than the beef dogs. However, there is one veggie dog that has the same number of calories as the highest calorie beef hot dog. So, if you are trying to limit calories, you need to read the label to make sure you are getting a low-calorie veggie dog.

5. a. Ordered career home runs data is shown below. The 26th, 52th, 53rd, and 79th positions have been highlighted.

13	18	24	27	27	33	33	34	36	37
38	39	40	41	42	42	45	46	47	48
49	51	52	53	54	58	58	62	63	64
65	65	67	68	69	69	75	80	83	83
83	92	93	96	97	101	101	102	102	103
105	106	106	106	106	110	113	116	117	117
118	119	127	128	132	135	136	137	138	154
164	181	183	184	196	202	205	207	219	227
229	238	240	248	254	300	300	301	307	309
312	331	354	359	361	383	449	474	475	493
521	534	555	714						

b. Five-number summary: 13, 58, 106, 219, 714
(Keep in mind that statistical software may use a different algorithm for computing the first and third quartiles. For example, Minitab gives $Q_3 = 207 + 0.75(219 - 207) = 216$.)

c. Calculations of fences:

IQR = 219 – 58 = 161

step = (1.5)(161) = 241.5

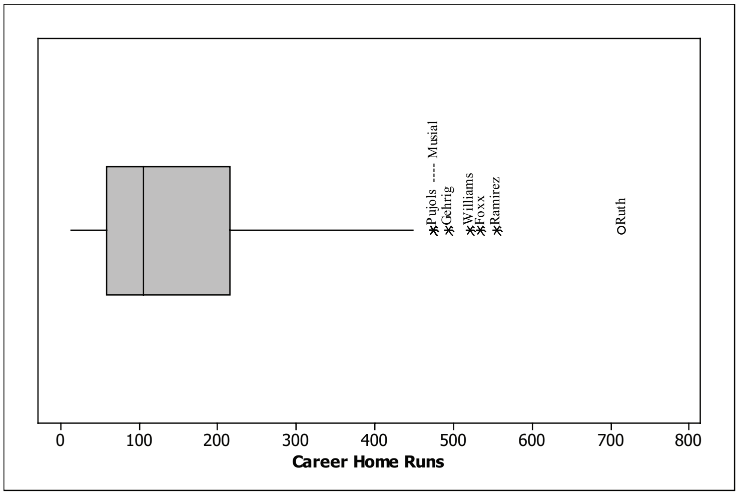
upper inner fence: 219 + 241.5 = 460.5

upper outer fence: 219 + 2(241.5) = 702

Mild outliers: 474 475 493 521 534 555

Extreme outlier: 714

d. The distribution of career home runs is skewed to the right. The right tail (fourth quarter) of the data is much more spread out than the left tail (first quarter) of the data.



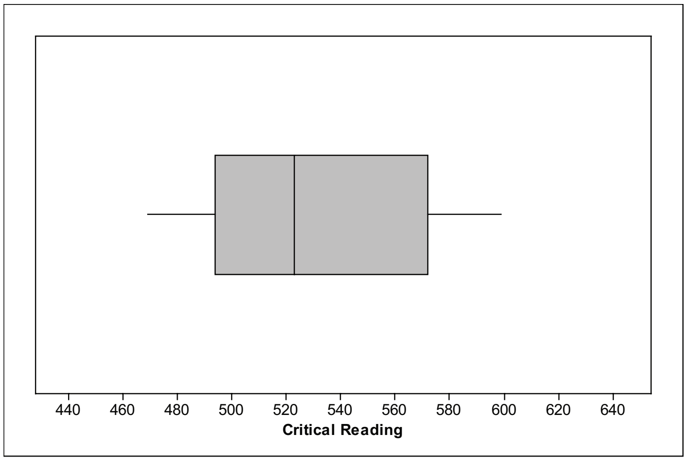
REVIEW QUESTIONS SOLUTIONS

1. a. 469, 494, 523, 572, 599

b. California's average SAT Critical Reading score does not fall in the top half of the states' average Critical Reading scores because 499 is below the median score of 523. It does fall above the bottom quarter because 499 is greater than the first quartile, which is 494.

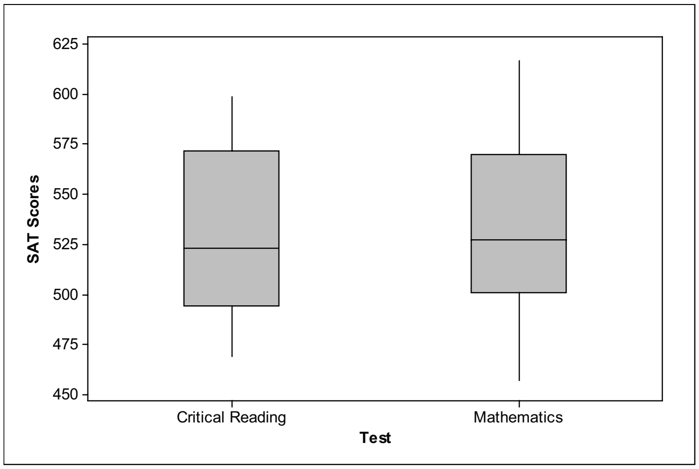
c. About 25% of the states; 12 states. That is because Wyoming's score is the third quartile.

d. In the boxplot below, the third quarter of the data appears to be more spread out than the other quarters of the data.



2. a. 457, 501, 527, 570, 617

b. Comparative boxplots appear below. (Note: Boxplots can be oriented either horizontally or vertically.)

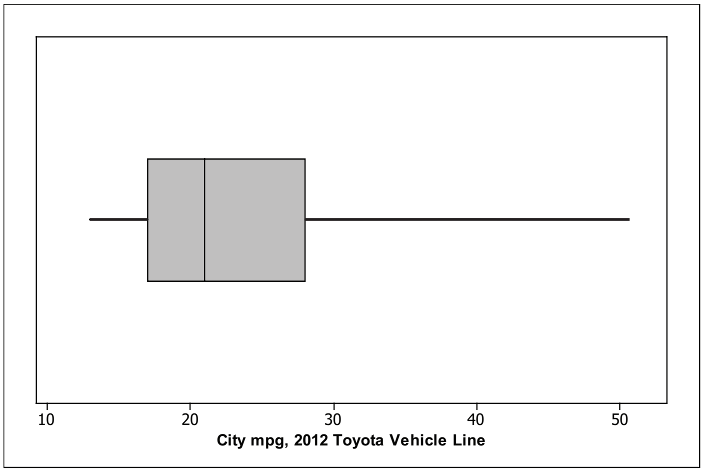


c. Sample answer: The states' average SAT Math scores are centered slightly higher than the SAT Critical Reading scores; the median for the Critical Reading scores is 523 and the median for the Math scores is 527. The middle half of the Critical Reading scores data is more spread out than the middle half of the Math scores; the IQR for Critical Reading scores is 78 compared to only 69 for the Math scores.

However, based on the length of the whiskers, the first and fourth quarters of the Math scores are more spread out than the first and fourth quarters of the Critical Reading scores. The length of the whiskers contributed to the range of the Math scores, which was 30 points higher than the range of the Critical Reading scores (range Math scores = 160 and range of Critical Reading scores = 130).

3. a. 13, 17 ,21, 28 , 51

b.

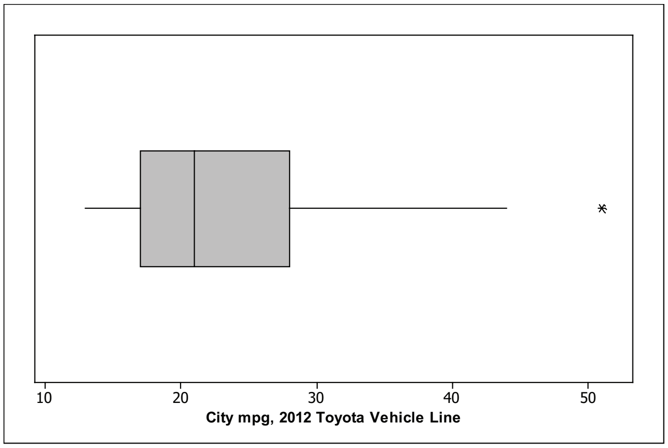


c. There appear to be three potential outliers: 43, 44, 51.

d. Calculations for fences:

IQR = 28 – 17 = 11; step = (1.5)(11) = 16.5
Lower inner fence = 17 – 16.5 = 0.5; no data values lie below this fence.
Upper inner fence = 28 + 16.5 = 44.5
Upper outer fence = 28 + 2(16.5) = 61

Only one data value, 51, is an outlier and it is a mild outlier because it falls between the two upper fences (see below).


10
20
30
40
50
City mpg, 2012 Toyota Vehicle Line

Unit 6: Standard Deviation

PREREQUISITES

Students should be familiar with graphic displays such as stemplots, histograms, and boxplots, which are covered in Units 2, 3, and 5, respectively. From these units, students should be familiar with the five-number summary and measures of spread related to the five-number summary, namely the interquartile range (IQR) and range. They should be able to identify whether or not the shape of a distribution is symmetric and to identify potential outliers from a graphic display. In addition, students need to be able to compute the mean of a set of data, which is covered in Unit 4, Measures of Center.

ADDITIONAL TOPIC COVERAGE

Additional coverage on standard deviation can be found in *The Basic Practice of Statistics*, Chapter 2, Describing Distributions with Numbers.

ACTIVITY DESCRIPTION

The purpose of this activity is to help students visualize the spread of the data by focusing on the concept of deviations from the mean. Students can work on this activity either individually or in groups.

In questions 1 and 2, students make dotplots of the data and then draw horizontal bars that represent deviations from the mean. Based on the lengths of the horizontal bars for the data sets in question 2, it should be obvious which data set has the larger standard deviation.

In questions 3 and 4, students are given histograms of five data sets, all of which have the same mean. In question 3, they compare two data sets at a time and determine which has the larger standard deviation. In one case, the two data sets have histograms that are mirror images of each other about a vertical line at the mean, and hence, these data sets have the same standard deviation. In question 4, students are given the actual standard deviations of the five data sets and are asked to match them to the histograms.

THE VIDEO SOLUTIONS

1. The mean precipitation rates for the two cities were very close – Portland had a mean of 3.32 inches/month and Montreal had a mean of 3.4 inches/month. What differed between the two cities was the variability in the precipitation patterns. Montreal's precipitation rate was fairly constant from month to month; Portland's precipitation was heavy during winter months and light during summer months.

2. The sum of the deviations of individual data values from their mean, $\sum\left(x-\bar{x}\right)$, is always exactly 0.

3. The monthly sales data from the Manhattan Beach location is more variable than the sales data from the South Coast Plaza location.

4. No, standard deviation is always positive or 0. When you square deviations from the mean, they become positive or zero. Their sum is still positive or zero and the quotient after dividing the sum by *n* – 1 stays positive or zero. This final quantity is the variance. To get the standard deviation, take the square root of the variance, which gives a number greater than or equal to zero.

UNIT ACTIVITY SOLUTIONS

1. a. 1, 1, -3, 3, -2

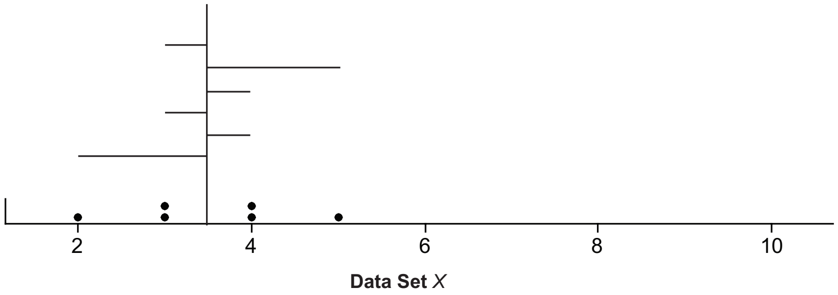
b. The sum is zero.

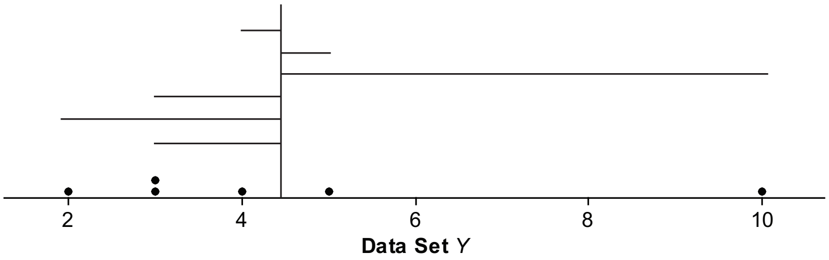
c. 1 + 1 + 9 + 9 + 4 = 24

d. $s^2 = 24/4 = 6$; $s = \sqrt{6} \approx 2.45$

2. a. Data Set *X*: $\bar{x} = 3.5$; Data Set *Y*: $\bar{y} = 4.5$

b.





c. The line segments representing the deviations from the mean tend to be longer for Data Set Y than for Data Set X. Since standard deviation is based on the deviations from the mean, Data Set Y will have the larger standard deviation.

d. For Data Set X: $s^2 = \frac{5.5}{5} = 1.1$; $s = \sqrt{1.1} \approx 1.05$

For Data Set Y: $s^2 = \frac{41.5}{5} = 8.3$; $s = \sqrt{8.3} \approx 2.88$

As was predicted in (c), the standard deviation for Data Set Y is larger than for Data Set X.

3. a. Data Set B has the larger standard deviation.

Sample answer: Data Set A is more concentrated around the mean of 2.5 and has its highest concentration of data between 2 and 3. Data Set B has data evenly spread from 0 to 5. So, there is a higher concentration of data in class intervals 0 to 1 and 4 to 5 for Data Set B than for Data Set A; these are the class intervals that are farthest from the mean. Therefore, Data Set B has the larger standard deviation.

b. Data Sets C and D have the same standard deviation.

Sample answer: The histograms are mirror images of each other about the vertical line at the mean, 2.5. So, the spread about the mean is the same for both data sets, just in opposite directions.

c. Data Set E has the larger standard deviation.

Sample answer: Data Set E has its highest concentration of data between class intervals 0 to 1 and 4 to 5, the class intervals that are farthest from the mean. A high proportion of the data from Data Set D is concentrated from 1 to 3, close to the mean of 2.5. Therefore, Data Set E is more spread out than Data Set D.

4. Standard deviations: $s_A = 1.124$, $s_B = 1.451$, $s_C = s_D = 1.026$, $s_E = 1.589$

EXERCISE SOLUTIONS

1. (100)2 = 10,000

2. a. Ninth-grade students:

$$\bar{x} = \frac{5+1+2+5+3+8}{6} = 4$$

$$s^2 = \frac{(5-4)^2+(1-4)^2+(2-4)^2+(5-4)^2+(3-4)^2+(8-4)^2}{6-1} = 6.4$$

$$s = \sqrt{6.4} \approx 2.53$$

12th-grade students:

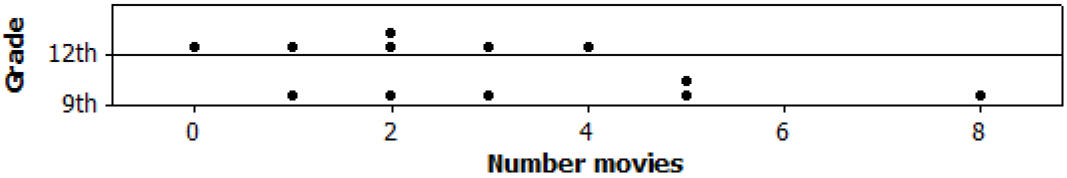
$$\bar{x} = \frac{4+2+0+2+3+1}{6} = 2$$

$$s^2 = \frac{(4-2)^2+(2-2)^2+(0-2)^2+(2-2)^2+(3-2)^2+(1-2)^2}{5} = 2$$

$$s = \sqrt{2} \approx 1.41$$

The data for the ninth-graders is more spread out.

b. The ninth-graders' data appears shifted to the right and more spread out compared to the 12th-graders' data.



3. a. $\bar{x} = \dfrac{7+3+4+7+5+10}{6} = 6$

$$s = \sqrt{\frac{(7-6)^2+(3-6)^2+(4-6)^2+(7-6)^2+(5-6)^2+(10-6)^2}{6-1}} = \sqrt{\frac{32}{5}} \approx 2.53$$

b. The mean increased by 2, the same amount that was added to the ninth-graders' data. The standard deviation stayed the same.

c. If you add 10 to each of the numbers in the ninth-grade data set, the mean will increase by 10. However, the standard deviation will still be 2.53.

4. a. Critical reading: s = 40.84; mathematics: s = 41.84; writing: s = 39.85

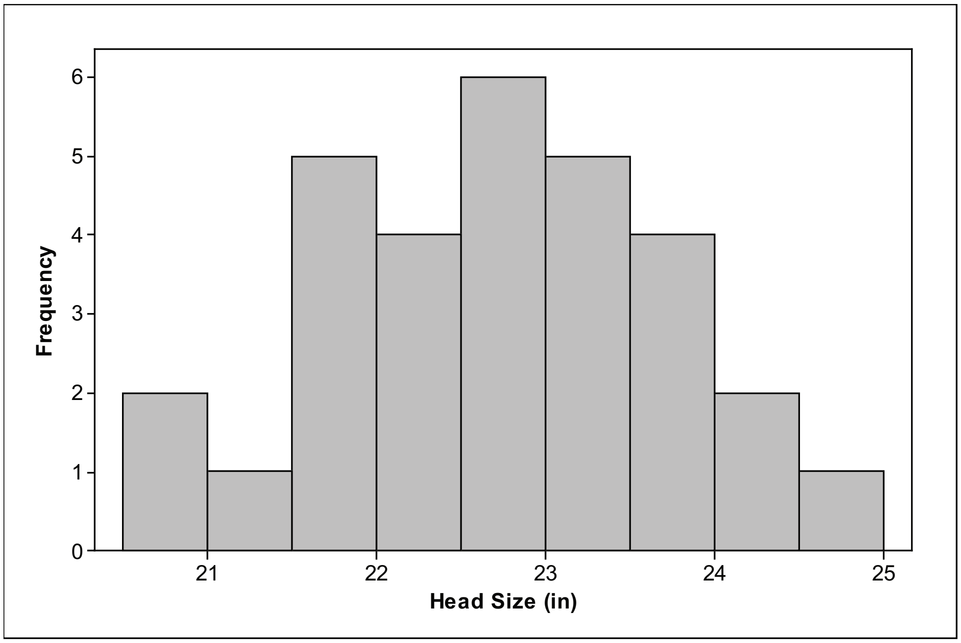
b. The mathematics SAT scores are the most spread out of the three exams.

REVIEW QUESTIONS SOLUTIONS

1. a. No. These numerical summaries help describe specific aspects of the distribution, especially center and spread. But they do not describe the exact shape of the distribution.

b. The answer is again No. In the case of the five-number summary, this is easier to see. The observations between, say, the third quartile and the maximum observation are free to move anywhere in that interval without changing the five-number summary.

2. a. From the histogram below, the shape of the histogram appears to be mound-shaped and roughly symmetric. Hence, this is a good distribution to summarize using $\bar{x}$ and s to measure center and spread, respectively.



b. Using technology, $\bar{x} \approx 22.647$ inches and $s \approx 1.026$ inches.

c. $\bar{x} - s \approx 22.647 - 1.026 = 21.621$; $\bar{x} + s \approx 22.647 + 1.026 = 23.673$

Using technology to sort the data from smallest to largest gives the following ordered list of data values. The data values that fall within one standard deviation of the mean have been shaded.

20.8 20.8 21.0 21.5 21.5 21.7 21.8 21.9 22.0 22.2 22.3
22.4 22.5 22.6 22.6 22.7 22.7 22.7 23.0 23.0 23.1 23.3
23.4 23.5 23.5 23.9 23.9 24.0 24.2 24.9

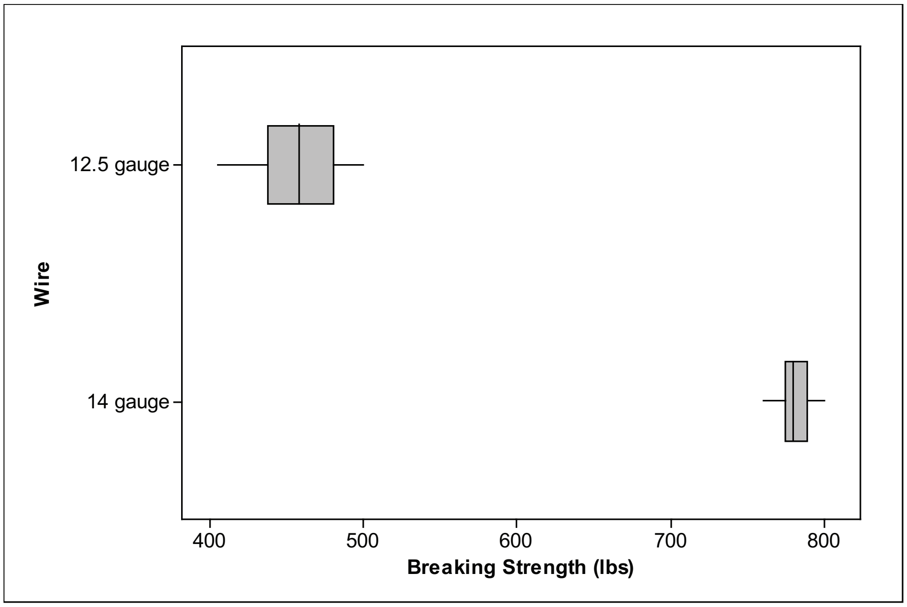
20/30 × 100% or around 66.7% of the data fall within one standard deviation of the mean.

3. a. 58.420 56.388 55.118 55.880 56.642 57.404
57.658 54.610 57.658 63.246 52.832 59.182
61.468 59.690 60.706 59.436 52.832 54.610
58.420 60.960 57.658 57.404 60.706 55.372
58.674 55.626 53.340 56.896 59.690 57.150

b. $s \approx 2.606$

c. The standard deviation in (b) is 2.54 times the standard deviation of the data in 2(b): (1.026)(2.54) ≈ 2.606

4. a. Sample answer: The two boxplots appear roughly symmetric.



b. The mean breaking strength for the 12.5-gauge, low-carbon wire is 458 lbs. The mean breaking strength for the 14-gauge, high-tensile wire is 780.75 lbs. The 14-gauge wire has the larger mean breaking strength.

c. 12.5-gauge wire: s = 26.48 lbs; 14-gauge wire: s = 9.90 lbs. The 12.5-gauge wire's data is more variable than the 14-gauge wire's data.

Unit 7: Normal Curves

PREREQUISITES

This unit requires an understanding of Unit 3, Histograms, Unit 4, Measures of Center, and Unit 6, Standard Deviation.

ADDITIONAL TOPIC COVERAGE

Additional coverage of normal curves can be found in *The Basic Practice of Statistics*, Chapter 3, The Normal Distributions.

ACTIVITY DESCRIPTION

The purpose of the activity is to familiarize students with the normal density curve. After completing the activity, they should be able to identify the mean and standard deviation of a normal distribution from its normal density curve. Students should also observe that most of the area under a normal density curve falls within three standard deviations of the mean. Therefore, the activity can serve as preparation for Unit 8, Normal Calculations, where students will learn the Empirical rule (or the 68-95-99.7% rule).

THE VIDEO SOLUTIONS

1. It is mound-shaped and symmetric; bell-shaped.

2. Since a normal curve is symmetric, the mean is at the line of symmetry.

3. The normal curve that is low and spread out has a larger standard deviation.

4. The mean arrival time for Year 33 appears to have decreased (shifted to an earlier date) compared to the mean in Year 1.

5. It has decreased from about 10% in Year 1 to 4% in Year 33.

UNIT ACTIVITY SOLUTIONS

1. a. Sample answer: Most of the data from a standard normal distribution should lie in the interval between -3 and 3. (Some students may pick a slightly wider interval. However, whatever interval they give, it should be centered at 0.)

b. Sample answer: Between -3 and 0. (Some students may give a value below -3.)

2. a. The dashed curve is flatter and more spread out than the standard normal density curve. Hence, it has a larger standard deviation than the standard normal distribution, which has a standard deviation of 1. So, the standard deviation will be larger than 1.

b. It looks like it should be around 1.5.

3. a. Normal density curves are symmetric about their mean. The line of symmetry for the dashed curve is at $\mu = 2$.

b. We would expect nearly all the data from this distribution to fall between -1 and 5. That looks to be about 3 standard deviations from the mean of 2.

4. Figure 7.12(a) is centered at $\mu = 15$. Going out three standard deviations on either side of 15 brings you to between 6 and 24, which would mean $\sigma = 3$.

So, Figure 7.12(a) goes with choice (ii).

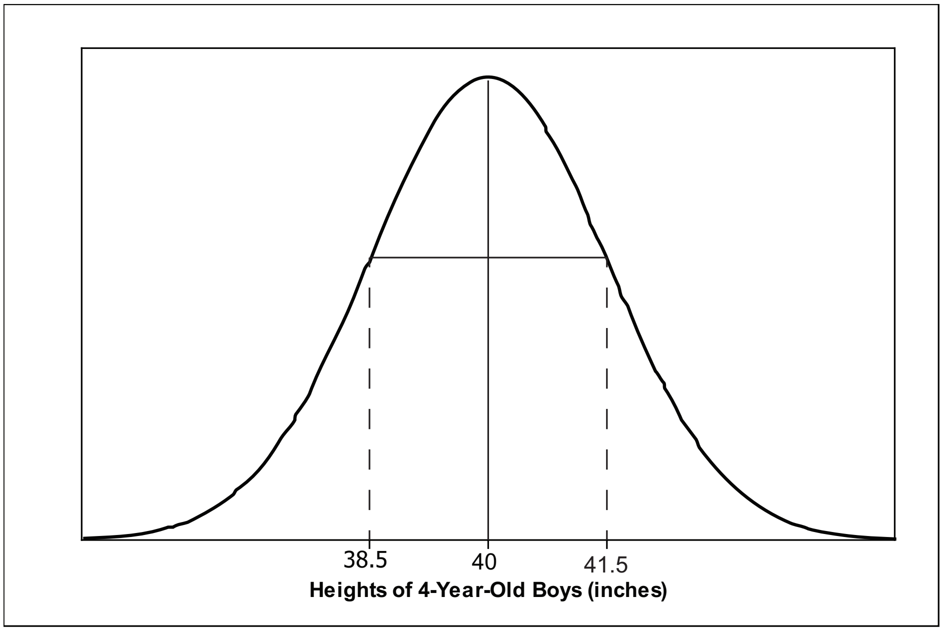
Figure 7.12(b) is centered at $\mu = 15$. Going out three standard deviations on either side of 5 brings you to between -1 and 11, which would mean that $\sigma = 2$. So, Figure 7.12(b) goes with choice (iv).

Figures 7.12(c) and (d) are centered at $\mu = 4$.

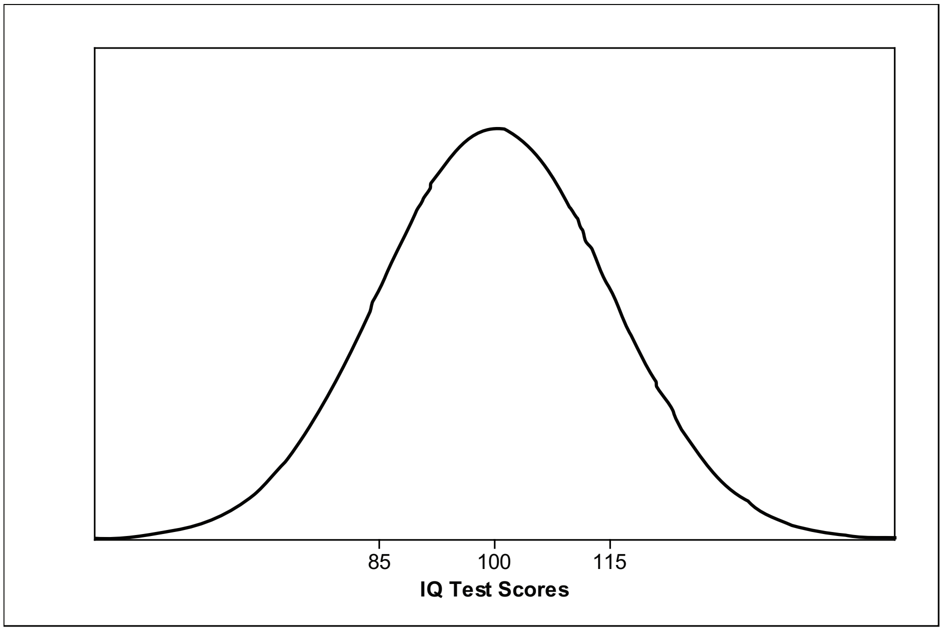
In each case, most of the data from the given distribution will fall within 3 standard deviations from the mean. Data from the distribution in Figure 7.12(d) is more spread out than data from the distribution in Figure 7.12(c). Hence, Figure 7.12(d) goes with choice (i), $\mu = 4$, $\sigma = 1$ and Figure 7.12(c) goes with choice (iii), $\mu = 4$, $\sigma = 0.5$.

EXERCISE SOLUTIONS

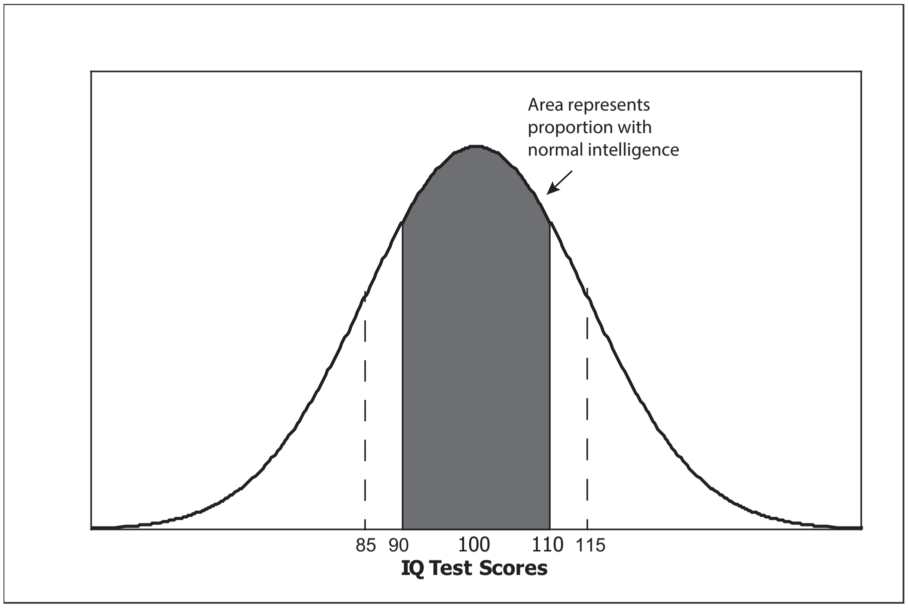
1.

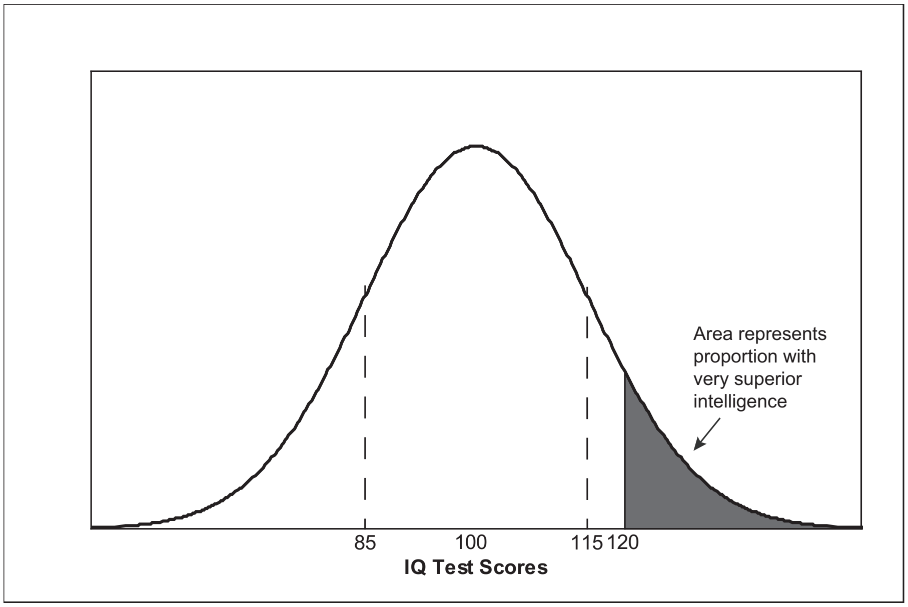


2. a.



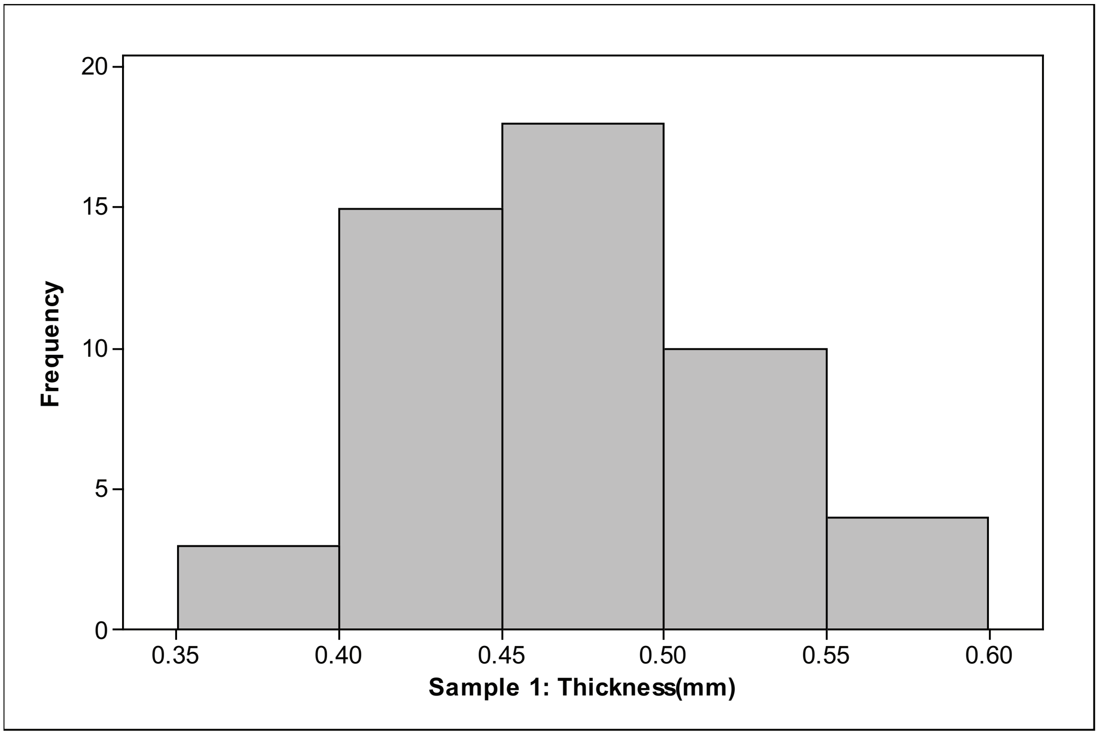
b.



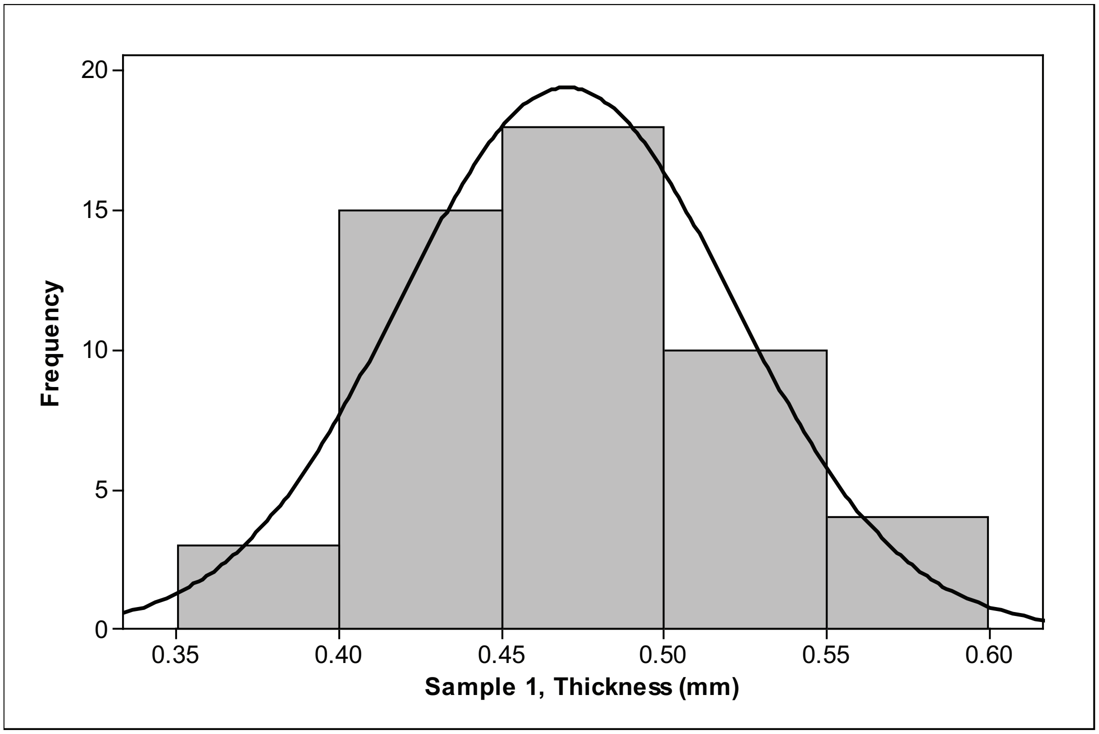


c. The shaded area under the curve over the interval from 90 to 110 appears larger than the shaded area under the curve to the right of 120. This indicates that there is a higher proportion of people with normal (or average) intelligence compared to people with very superior or genius intelligence.

3. a.

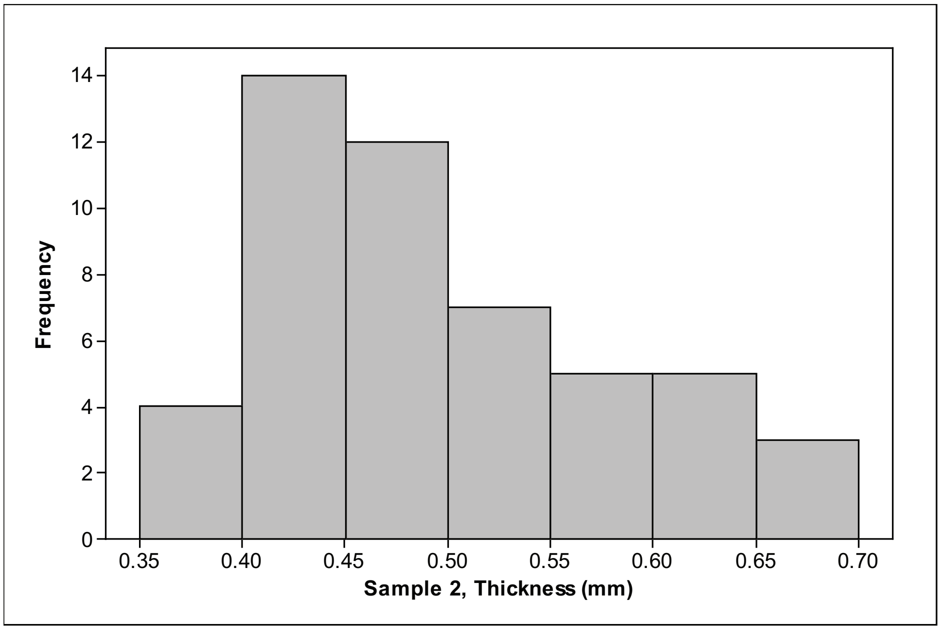


b.

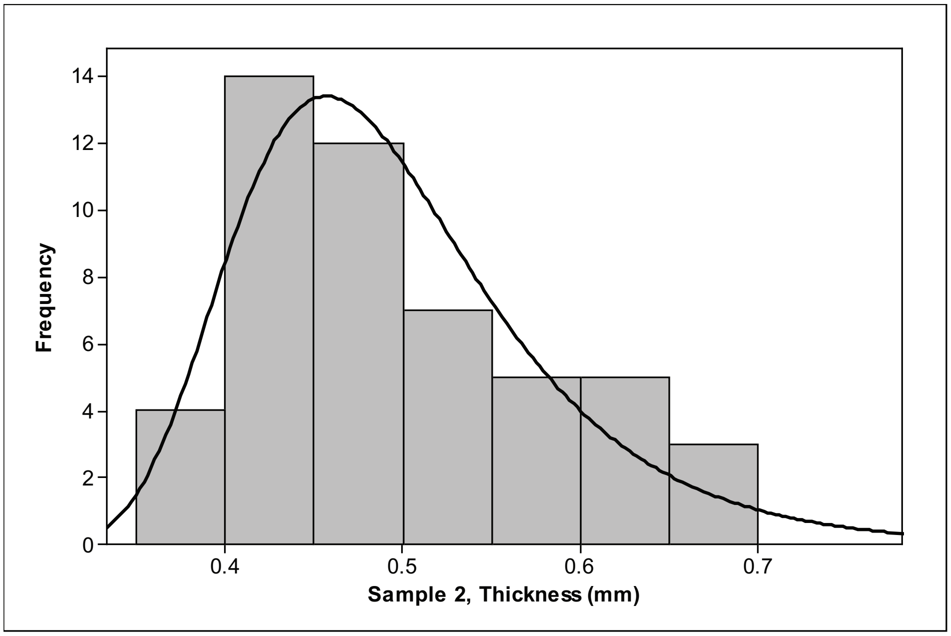


c. The balance point or mean appears to be less than 0.50.

4. a.



b. It does not appear that these data are from a normal distribution. The data appear skewed to the right and so the smoothed curve should also appear skewed to the right. Below is a sample smoothed curve that better captures the shape of this histogram than a normal curve.



c. Sample answer: The balance point appears closer to 0.5 mm with this distribution than for Sample 1 in question 3. The sample mean for Sample 2 is 0.497 mm, which is pretty close to 0.5.

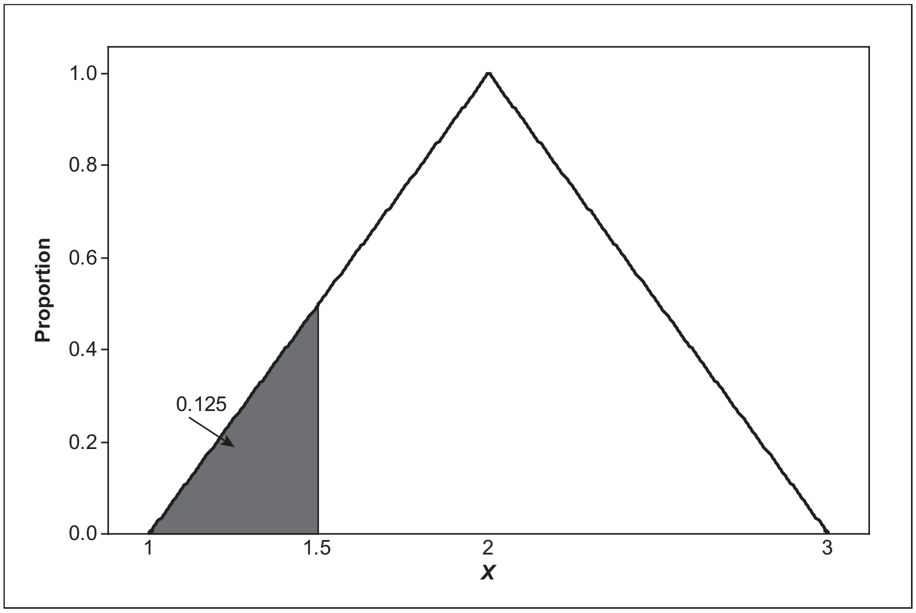
REVIEW QUESTIONS SOLUTIONS

1. a. The dashed curve represents a distribution with the larger mean. For normal curves, the mean is the line of symmetry. The lines of symmetry are around 25 and 30 for the solid curve and dashed curve, respectively.

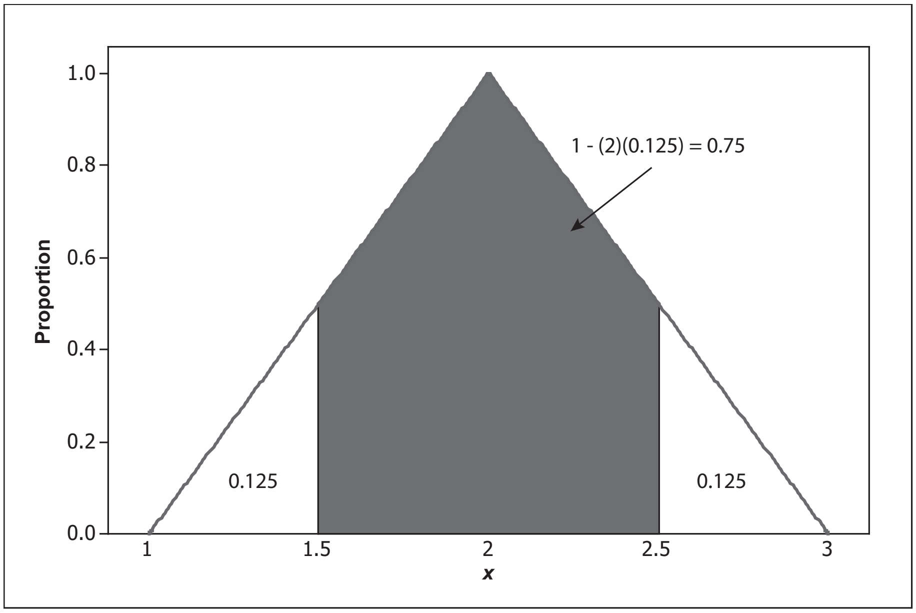
b. The solid curve represents a distribution with the larger standard deviation. It is flatter and its area under the curve is more spread out than for the dashed curve.

2. a. Because this distribution is symmetric, the mean is located at the line of symmetry for the curve. Therefore, μ = 2.0.

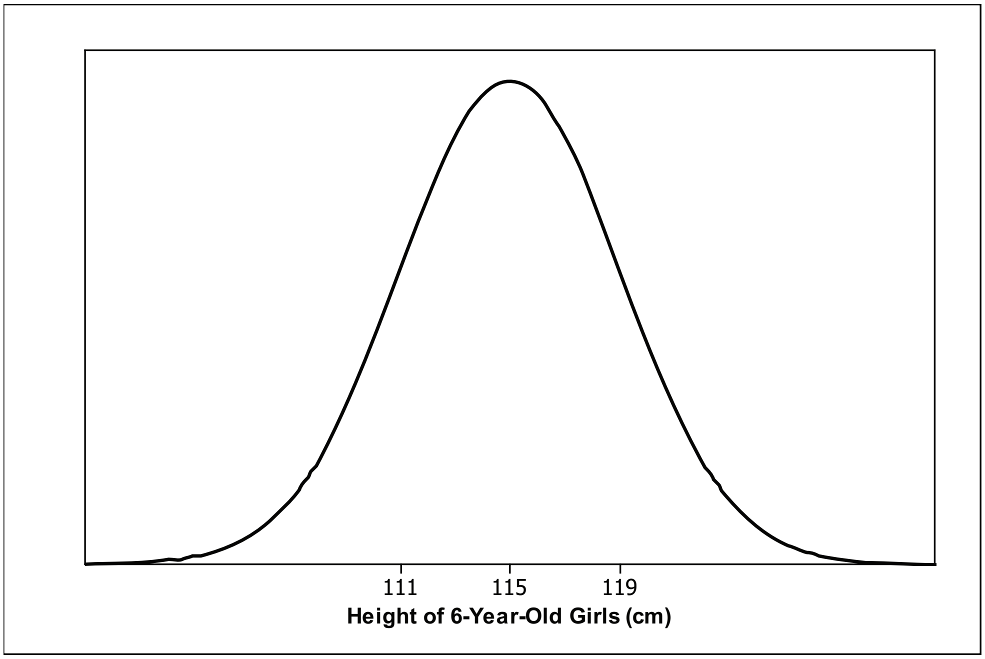
b. The area under the density curve to the left of 1.5 gives the proportion of data that fall below 1.5. This area forms the triangular region shown below. Area of triangle = ½(base)(height). So, in this case, proportion of data less than 1.5 is (1/2)(0.5)(0.5) = 1/8 = 0.125.



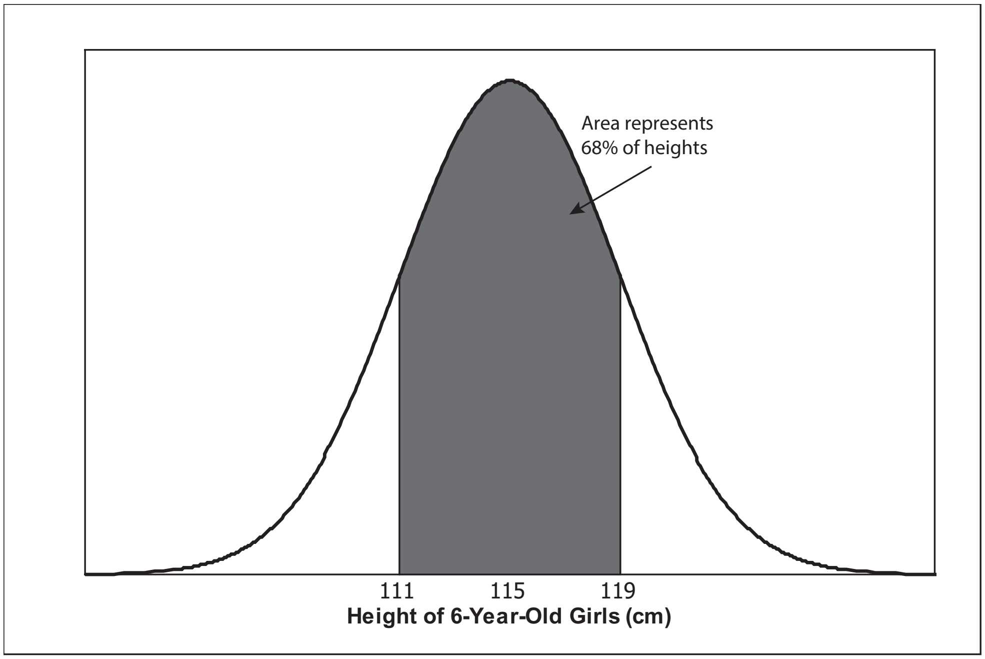
c. Since the density curve is symmetric, the proportion of data that is more than 2.5 is the same as the proportion of data that is less than 1.5. Because the area under a density curve is 1, the area of the region between 1.5 and 2.5 (shown below) is 1 - (2)(0.125) = 0.75.



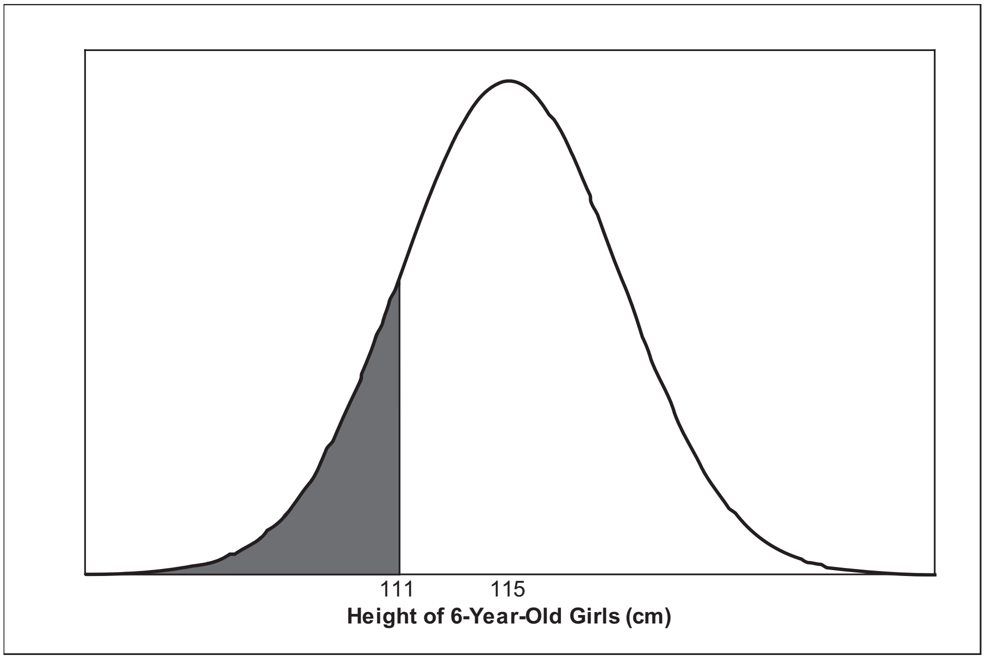
3. a.



b.

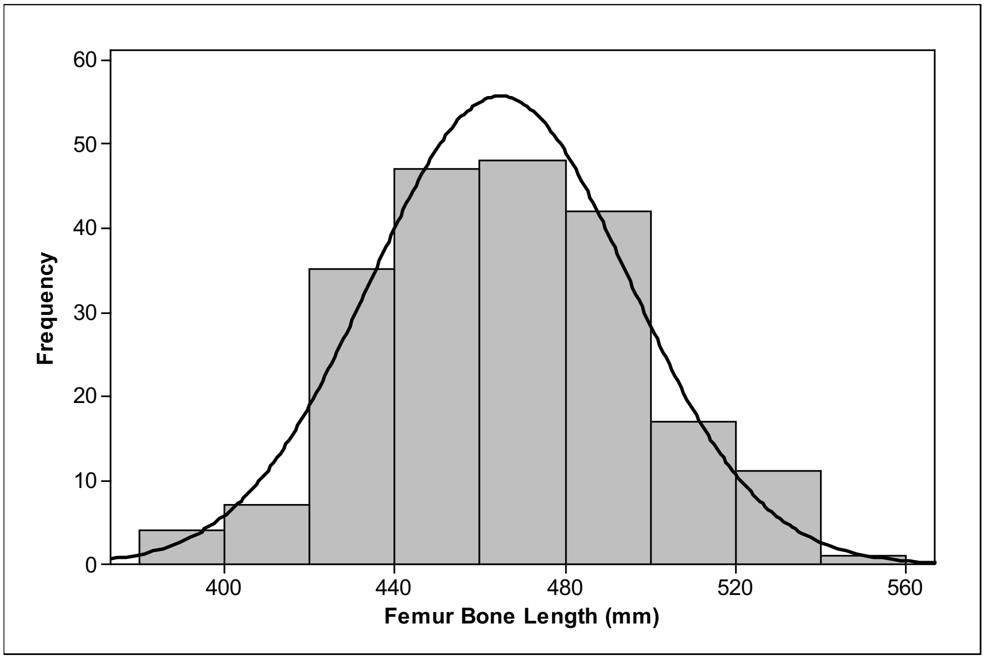


c. (100% - 68%)/2 = 32%/2 = 16%.



4. a. The data do not appear to be from a normal distribution. It looks as though there may be outliers to the left. There is a large peak between 70 and 72.5 and then a steep drop on either side. The data do not have the characteristic bell-shape of normal data.

b. The distribution of femur bone lengths looks like it could be from a normal distribution. The histogram has a symmetric mound shape. Below is a normal curve sketched over the histogram. (Software was used to fit a normal curve to these data.)



c. These data are strongly skewed to the right. Hence, they do not appear to be from a normal distribution.

Unit 8: Normal Calculations

PREREQUISITES

This unit requires familiarity with basic facts about normal distributions, which are covered in Unit 7, Normal Curves. In addition, students need some background on distributions, means, and standard deviations, which are covered in Units 3, 4, and 6, respectively.

ADDITIONAL TOPIC COVERAGE

Additional coverage of normal curves can be found in *The Basic Practice of Statistics*, Chapter 3, The Normal Distributions.

ACTIVITY DESCRIPTION

The purpose of this activity is to help students understand the connection between finding the proportion of data that fall in specific intervals and the areas under a density curve over those intervals. Students are given a graph of a standard normal density curve on which a rectangular grid has been superimposed. This allows them to determine areas by counting the number of rectangles under the density curve over specific intervals. Students estimate proportions by estimating areas. After estimating a collection of proportions, they use a *z*-table to see how close their estimates were to the actual proportions.

MATERIALS

Students will need either a hard copy of this activity or six copies of the standard normal curve from Figure 8.11. That curve has been reproduced on the next page so that it can be easily copied. In addition, they will need a copy of a *z*-table (or access to technology where they can find standard normal probabilities).

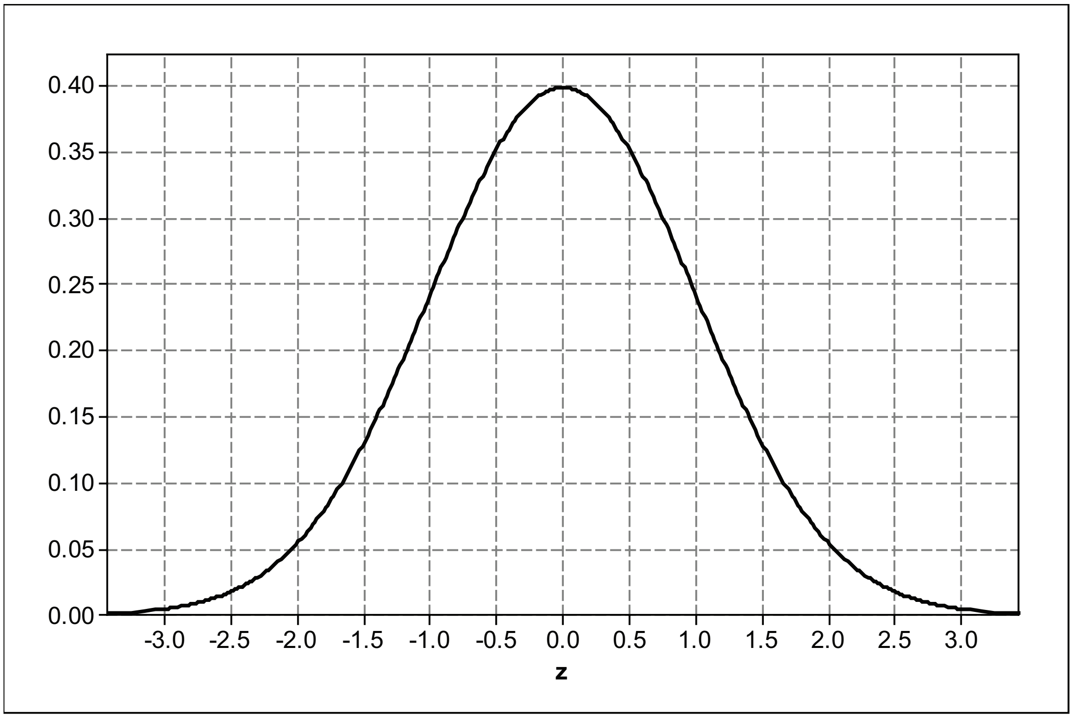


Figure 8.11. The standard normal density curve.

THE VIDEO SOLUTIONS

1. The 68-95-99.7% Rule.

2. At least 5 feet 10 inches tall.

3. Suppose *x* is an observation from a normal distribution with mean μ and standard deviation σ. To calculate the *z*-score, subtract μ from *x* and then divide the result by σ .

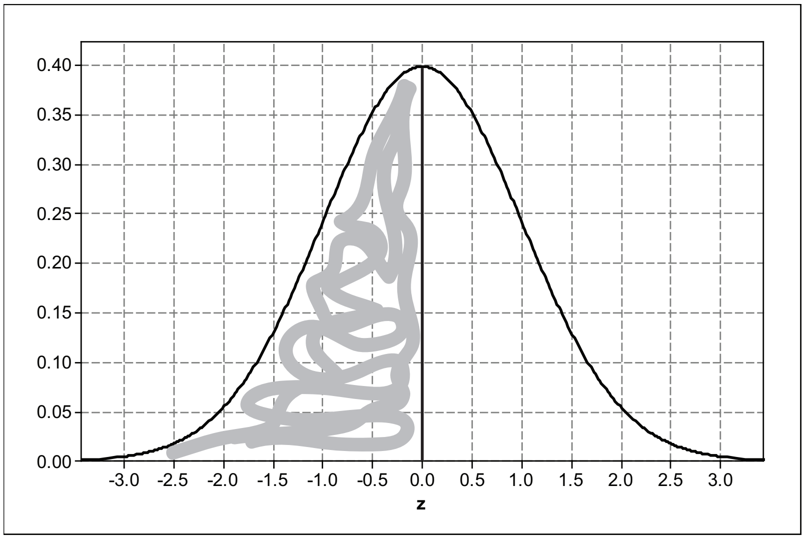
4. The eligibility *z*-score for women (1.48) is higher than for men (0.98). So, in order to join the Beanstalks, women's heights must be at least 1.48 standard deviations above the mean height for women while men's heights need only be at least 0.98 standard deviations above the mean height for men.

UNIT ACTIVITY:

USING AREA TO ESTIMATE STANDARD NORMAL PROPORTIONS SOLUTIONS

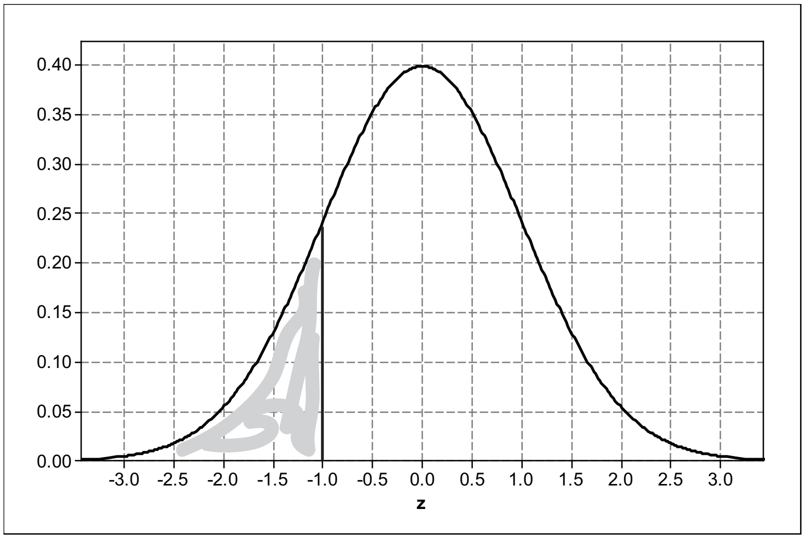
1. Sample answer: 40

2. a. Sample answer: There are 20 rectangles in the shaded region below.



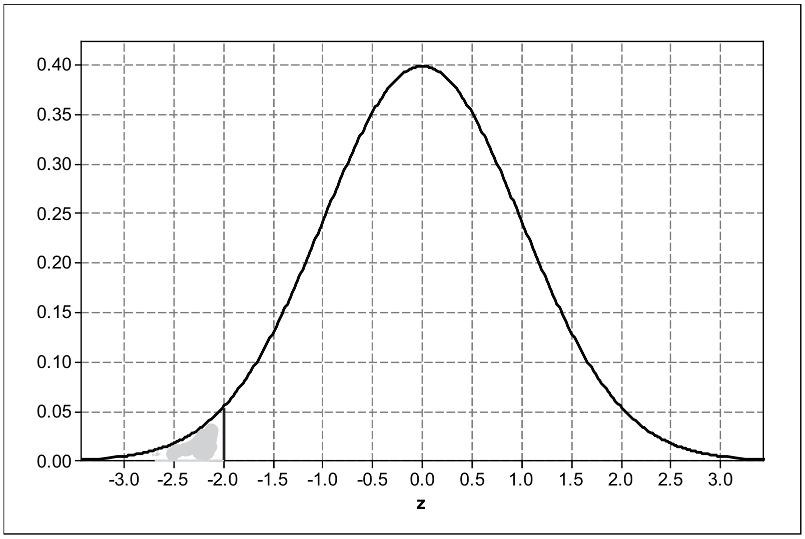
b. Proportion = 20/40 = 0.5

3. a. Sample answer: There are 6 ½ rectangles in the shaded region below.



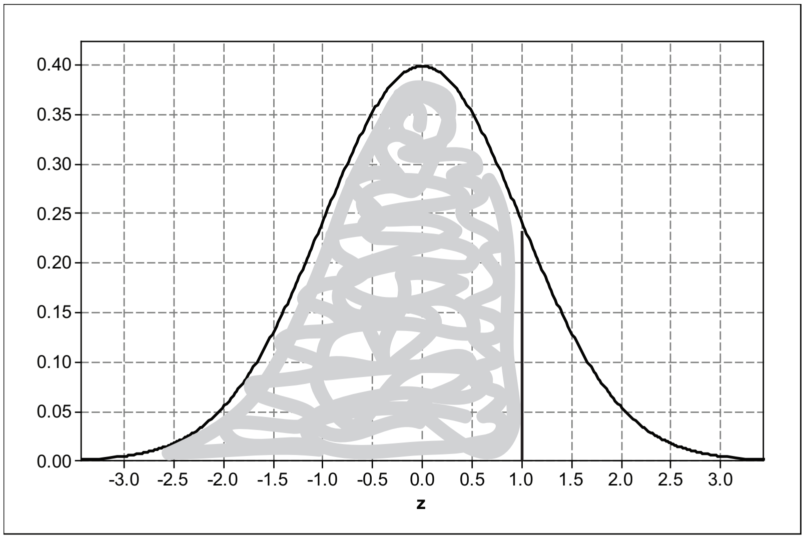
b. Proportion = 6.5/40 ≈ 0.1625

4. a. Sample answer: There is 1 rectangle in the shaded region below.



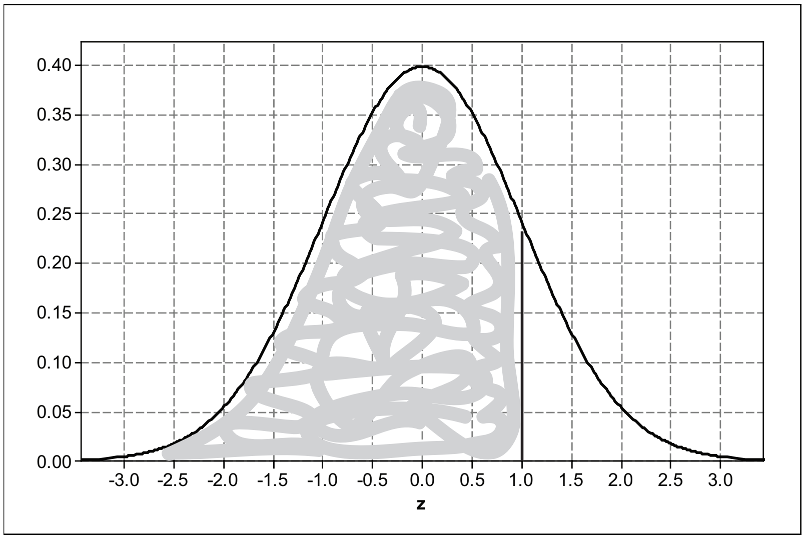
b. Proportion = 1/40 = 0.025

5. a. Sample answer: There are 33 ½ rectangles in the shaded region below.



b. Proportion = 33.5/40 = 0.8375

6. a. Sample answer: There are 39 rectangles in the shaded region below.



b. Proportion = 39/40 = 0.975

7. a. See solution to 7b.

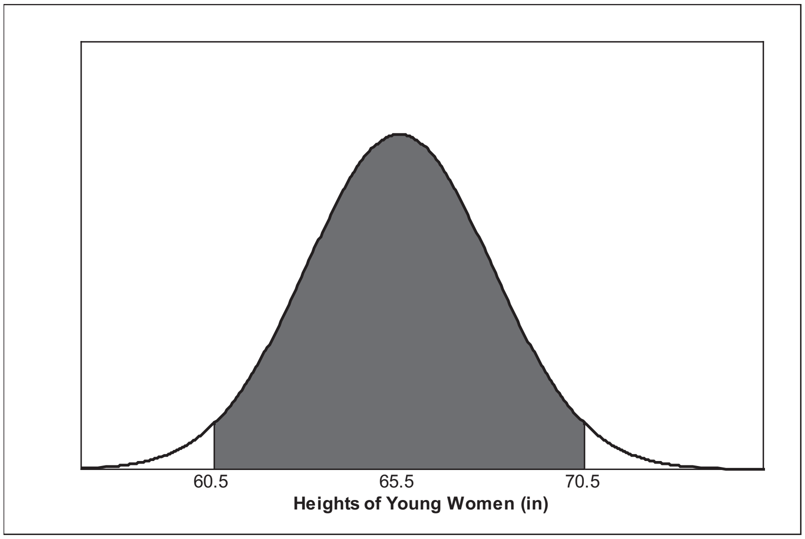
b.

z	Proportion from *z*-table	Estimated Proportion
-2.00	0.0228	0.0250
-1.00	0.1587	0.1625
0.00	0.5000	0.5000
1.00	0.8413	0.8375
2.00	0.9772	0.9750

c. Sample answer:
The estimated proportions from areas are very close to the values from the *z*-table.

EXERCISE SOLUTIONS

1. a. 65.5 – 2(2.5) = 60.5; 65.5 + 2(2.5) = 70.5



b. Around 95% of young women are between 60.5 and 70.5 inches tall (within two standard deviations of the mean). That means that 5% of young women are more than two standard deviations shorter or taller than the mean. Hence, 5%/2, or 2.5% of young women are more than 2 standard deviations taller than the mean.

c. 6 feet = 72 inches. Converting to a *z*-score gives *z* = (72 – 65.5)/2.5 = 2.6. This means that the 20-year-old woman is 2.6 standard deviations taller than the mean height of other young women.

2. a. We know that 68% of the scores are within 1 standard deviation of the mean – hence, between 400 and 600. That means that 32% are more than 1 standard deviation on either side of the mean. So, the percentage of scores above 600 is half of 32% or 16%.

b. Julie: *z* = (630 – 500)/100 = 1.3; John: *z* = (22 – 18)/6 ≈ 0.67. Julie did better than John because her score was 1.3 standard deviations above the mean while John’s score was only 0.67 standard deviations above the mean.

3. Sample answer: There are a few low priced homes, many moderately priced houses, some very expensive houses, and a few outrageously expensive mansions. So, the distribution of house prices is strongly skewed to the right and not normally distributed. The 68-95-99.7% rule should not be applied to house prices.

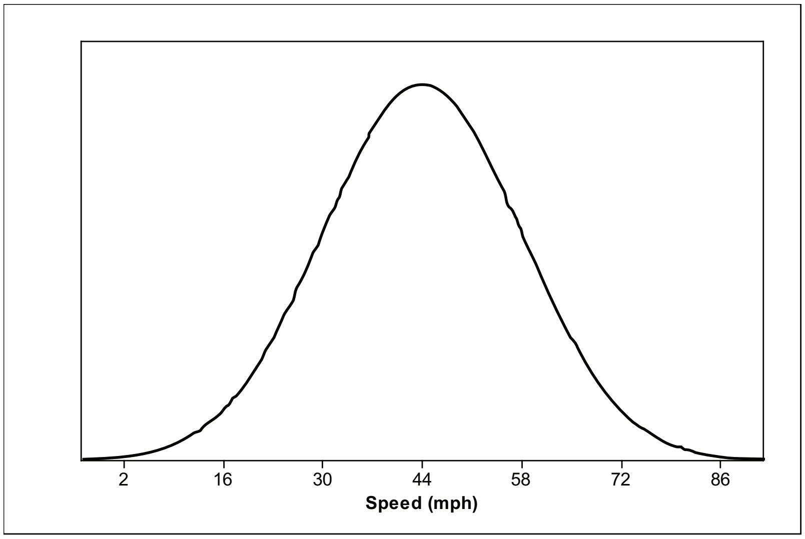
4. The table entry for *z* = -1 is 0.1587; so, 15.87% are less than -1.

The table entry for *z* = 2.25 is 0.9878; so, 98.78% are less than 2.25 and therefore, 1.22% are greater.

Because 98.78% are less than 2.25 and 15.87% are less than -1, the percentage lying between -1 and 2.25 is 98.78% - 15.87% = 82.91%, or about 83%.

REVIEW QUESTIONS SOLUTIONS

1. a.



b. This interval represents the data values that fall within one standard deviation of the mean. Using the 68-95-99.7 Rule, the percentage would be 68%. Thus, the proportion is 0.68.

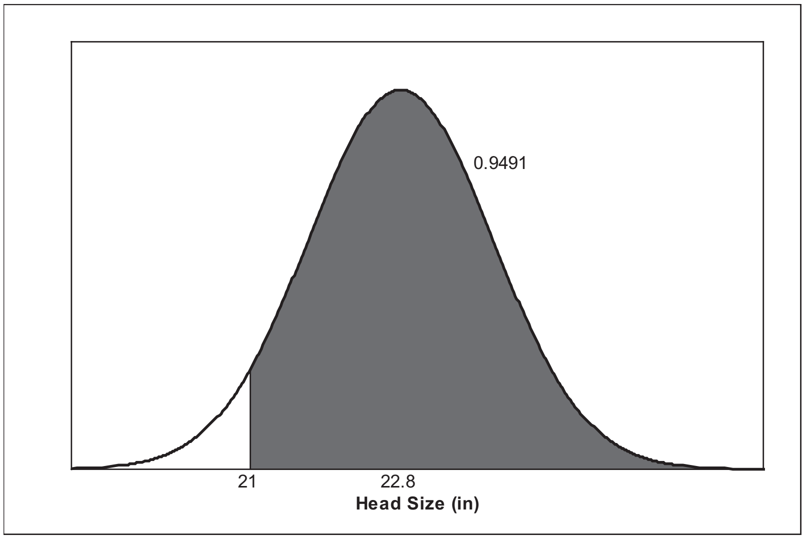
c. To find the proportion of speeds that are below 30 mph or above 58 mph, subtract 0.68 from 1: 1 – 0.68 = 0.32. The proportion of speeds that are below 30 mph is half this amount: 0.32/2 = 0.16.

d. The speed of 72 mph is two standard deviations from the mean of 44. We know that roughly 0.95 of the speeds fall within two standard deviations from the mean. Hence, 0.05 of the speeds fall beyond two standard deviations from the mean. Roughly 0.05/2 or 0.025 of the speeds exceeded 72 mph.

2. Carrie's standardized score on test B is $z = (79 - 65)/9 = 1.56$; Pat's standardized score on test A is $z = (85 - 78)/6 = 1.17$. Carrie has the higher standardized score. If both tests cover the same material and both were taken by similar groups of students, then Carrie did better than Pat because her score is higher relative to the overall distribution of scores.

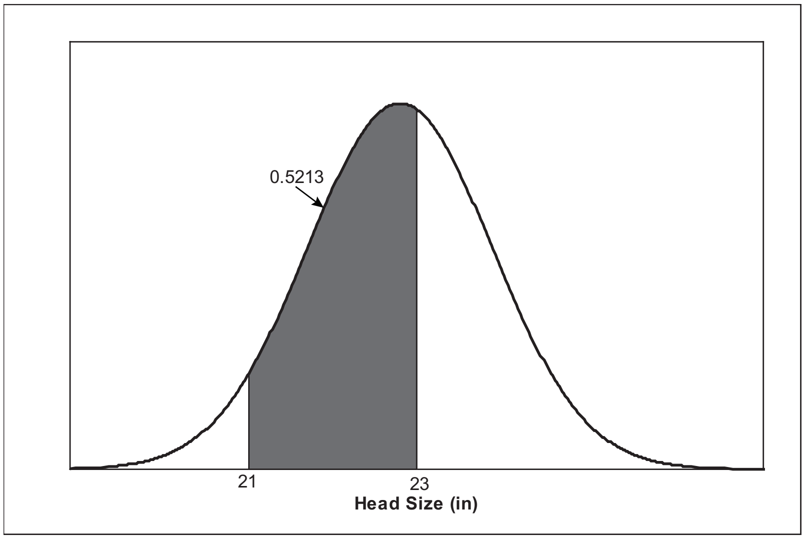
3. a. Convert 21 inches into a standardized value: $z = (21 - 22.8)/1.1 \approx - 1.64$. Using the standard normal table we get a proportion of 0.0505 soldiers with head sizes below the one observed. That means that 1 – 0.0505, or a proportion of 0.9495, or 94.95% of soldiers has head sizes above 21 inches.

Using Minitab, we did not have to first convert to a *z*-score. The result is slightly more accurate because we did not round a *z*-score to two decimals.

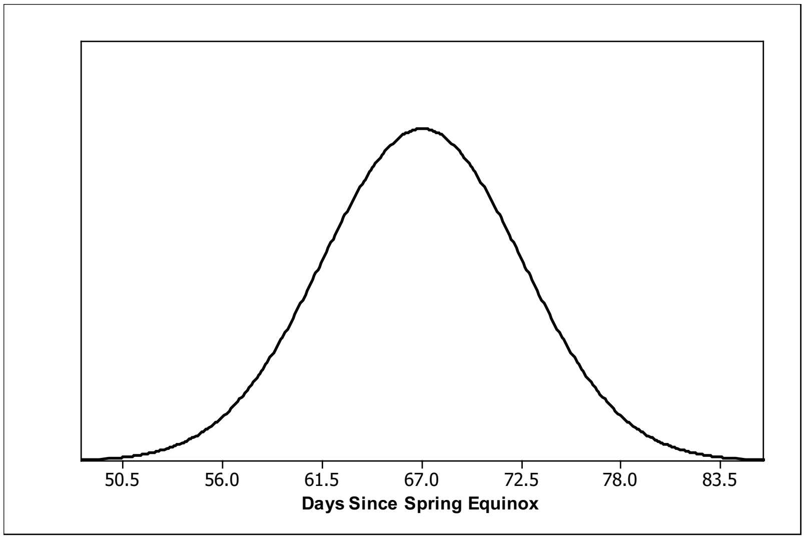


b. Converting 23 inches into a standardized value gives: $z = (23 - 22.8)/1.1 \approx 0.18$. Using the standard normal table, we get a proportion of 0.5714. The proportion of soldiers with head size between 21 inches and 23 inches is 0.5714 – 0.0505 = 0.5209, or around 52.09%.

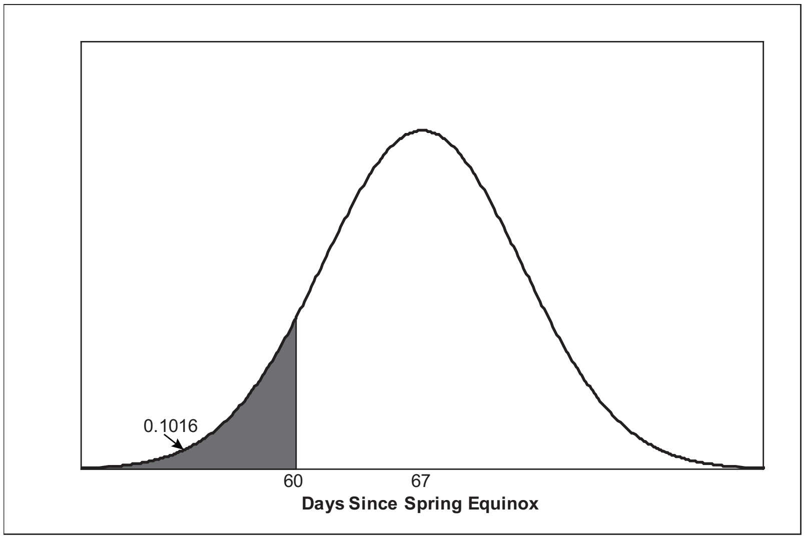
Using Minitab, we did not have to first convert to *z*-scores or subtract two proportions. Note there is a slight difference in the proportion below compared to the one above due to rounding *z*-scores to two decimals.



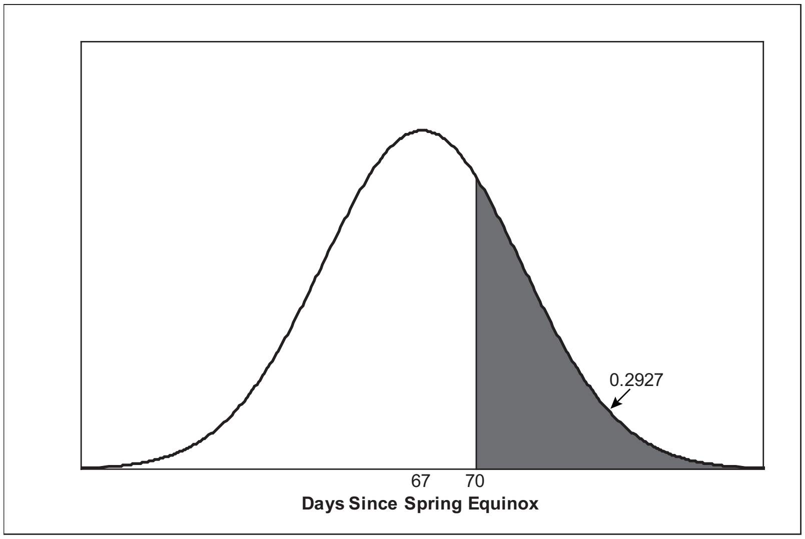
4. a.



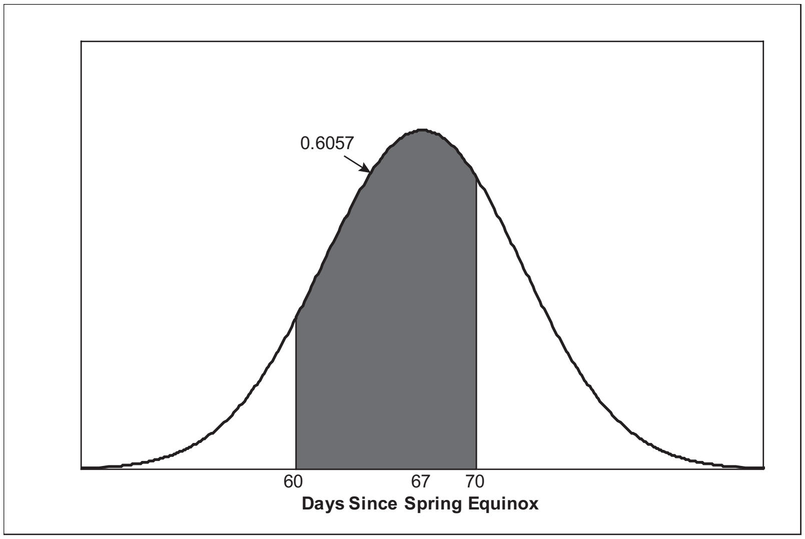
b. Around 10.1% of the Blackpoll Warblers arrived before day 60.



c. Around 29.3% of the Blackpoll Warblers arrived after day 70.



d. Around 60.6% of the Blackpoll Warblers arrived between days 60 and 70.



Unit 9: Checking Assumptions of Normality

PREREQUISITES

Students need to be familiar with histograms (Unit 3) and boxplots (Unit 5). The background on normal distributions covered in Unit 7, Normal Distributions, and Unit 8, Normal Calculations, are essential to this unit.

ACTIVITY DESCRIPTION

Students learn how to use normal quantile plots by examining a variety of shapes of data and the corresponding normal quantile plots. Students should work in small groups on this activity.

MATERIALS

Access to technology to construct normal quantile plots.

In questions 1 – 4, students are presented with pairs of histograms and normal quantile plots. In their groups, students should discuss the shapes of the histograms and how a given shape affects the pattern of dots in the normal quantile plots. In questions 5 and 6 students are given real data and asked to assess whether or not it is reasonable to assume the data are normally distributed.

For the final question, students are asked to collect their own data. Then they construct normal quantile plots to assess whether the distributions of the data they have collected are normally distributed. They can collect data on themselves – height, forearm length, head circumference, foot length, etc. They can also collect data from online sources; for example, data on sports – salaries of baseball players, lengths of tennis matches, a favorite basketball team’s scores for a season, etc. They could also collect data from their school – football scores for a season of games, exam scores from a class, and so forth.

In terms of technology to construct normal quantile plots (or normal probability plots), students can use statistical software, graphing calculators or spreadsheet software such as Excel.

Below are Excel instructions for constructing the simplified version of the normal quantile plot discussed in the Content Overview.

Step 1:
Label two columns: In cell A1, enter the variable name and in B1 enter Normal Quantile.

Step 2:
Enter your data in column A, starting in cell A2.

Step 3:
Use Sort to sort your data in column A from smallest to largest.

Step 4:
In cell B2 enter the following: =NORMINV((CELL("row",B2) -1)/(*n*+1),0,1). Replace *n* with the number of data values. Press Enter.

Step 5:
Click on cell B2. Then click on the lower right hand corner of cell B2 and drag down to create the quantiles.

Step 6:
Make a scatterplot of the data in column B versus the data in column A.

THE VIDEO SOLUTIONS

1. The curve is bell-shaped. It has one peak and is symmetric.

2. You expect a histogram that is roughly symmetric and mound-shaped.

3. You expect the boxplot to be roughly symmetric. The box should be concentrated in the middle of the display. The whiskers to the right and left of the box should be longer than the Q1 to median distance (or median to Q3 distance).

4. The pattern of the dots appears linear.

5. Skewed to the right.

UNIT ACTIVITY SOLUTIONS

1. a. The histogram is skewed to the right.

b. The pattern of the dots in the normal quantile is concave down, which you would expect when the histogram is skewed to the right.

c. No. The histogram is skewed to the right instead of being symmetric. The normal quantile plot is concave down instead of showing a straight-line pattern.

2. a. Sample answer: The histogram is mound-shaped and not quite symmetric. The first bar represents too many data values.

b. Sample answer: The normal quantile plot appears fairly linear. The dots all stay within the curved bands produced by Minitab. The three dots that are almost in a vertical line (corresponding to $v \approx 0$) are the result of the fact that too many data values fall in the first class interval.

c. Sample answer: Given that the pattern of dots is mostly linear, it seems reasonable to assume these data are from a normal distribution.

3. a. Sample answer: There appear to be three possible outliers. The major portion of the histogram is mound-shaped but not symmetric. However, the lack of symmetry may be due to choice of class interval size.

b. Sample answer: The overall pattern of the dots is fairly linear (there may be a slight concave up curvature). However, there are three dots on the plot that are separated from that pattern. These data values appear on the histogram to be outliers.

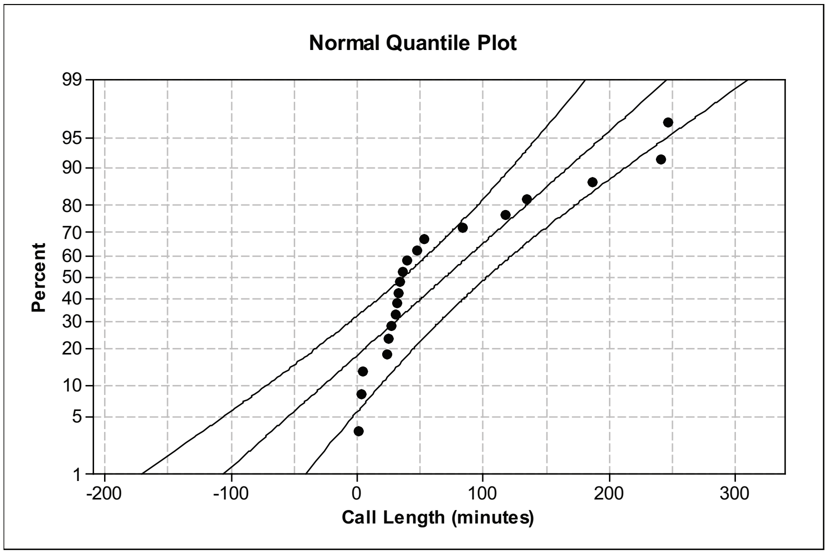
c. Sample answer: The three outliers probably indicate that these data are not normally distributed. (However, it is possible to observe outliers from normal distributions – but it would be unusual to observe three (in a data set of 25) that were this extreme.)

4. a. Sample answer: The histogram has two parts with a gap between them.

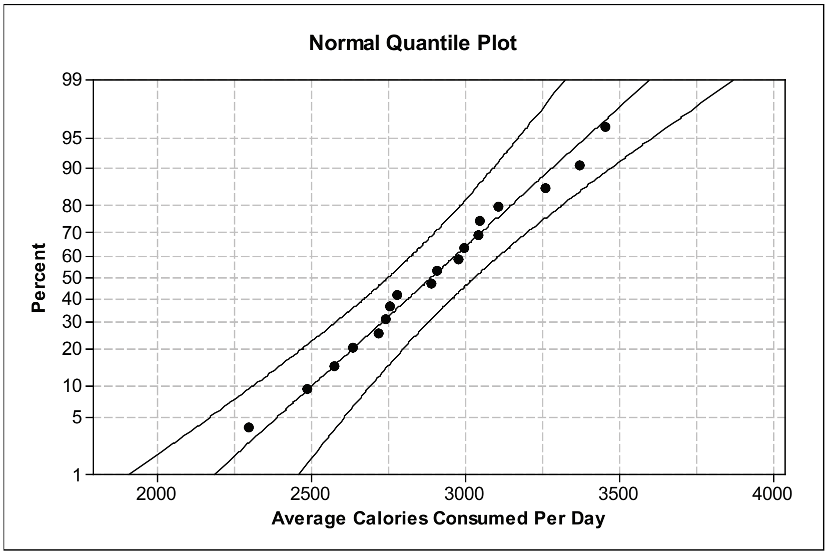
b. Sample answer: The plot appears to be in two pieces that are not aligned. So, the pattern does not appear mostly linear.

c. No. The normal quantile plot does not appear linear. The histogram is not unimodal and symmetric but instead has two peaks, one smaller than the other, with a gap in between.

5. The pattern of the dots in the normal quantile plot is severely curved concave down. Hence, the data are strongly skewed to the right. Hence, it is not reasonable to assume that call lengths are normally distributed.



6. The normal quantile plot shows a fairly straight-line pattern. Therefore, it is reasonable to assume that these data are from a normal distribution.

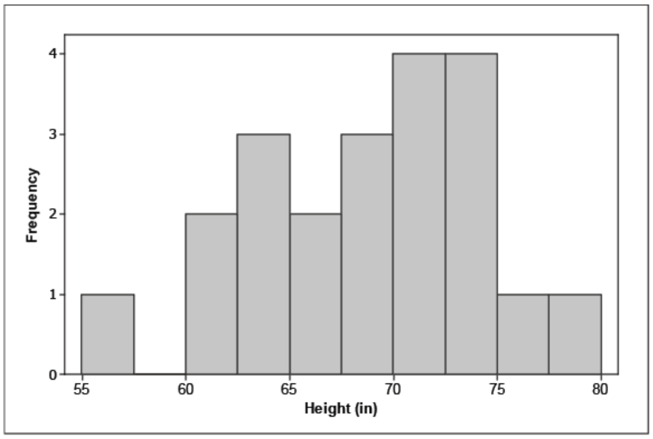


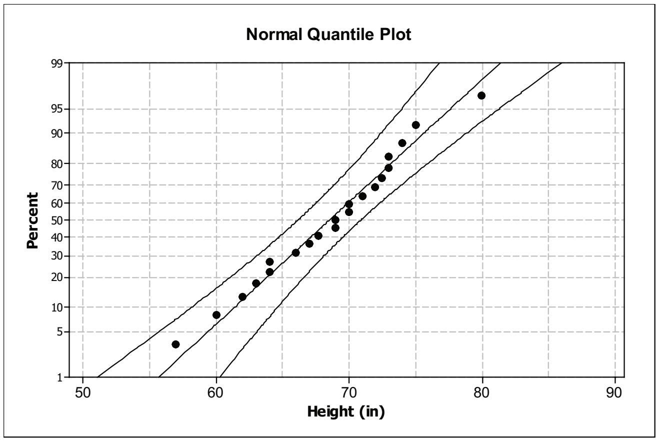
7. Sample answer:

Data on student heights collected from college students enrolled in an introductory statistics course:

62.0 60.0 64.0 69.0 66.0 63.0 64.0 57.0 67.0 72.5 70.0

73.0 69.0 74.0 70.0 71.0 80.0 73.0 75.0 72.0 67.7



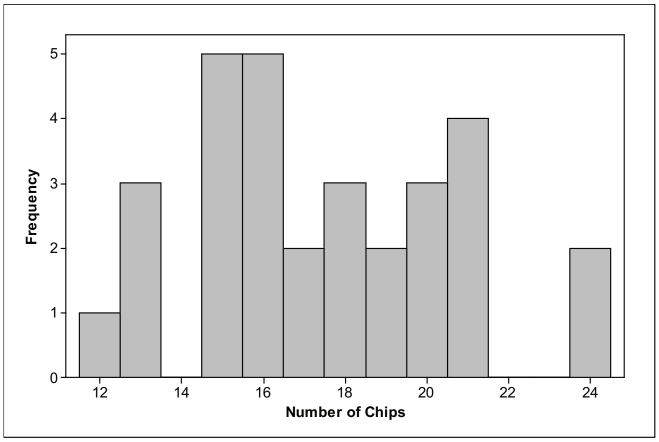


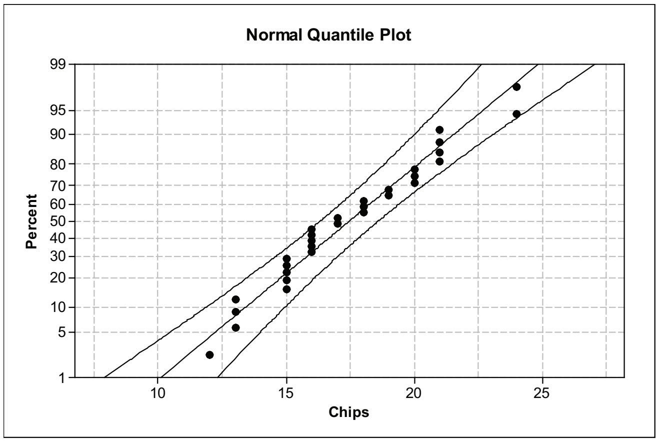
Based on the histogram, it appears that these data might not be normally distributed. However, the normal quantile plot looks pretty straight, so based on that plot, we conclude that it is reasonable to assume heights are normally distributed.

Data on number of chips in Chips Ahoy chocolate chip cookies collected by an introductory statistics class:

17 18 13 15 15 16 16 16 20 17 21 24 18 19 20

18 21 15 15 16 13 15 13 12 20 21 21 24 19 16





Based on the histogram, the data on number of chips do not look at all normal. However, the normal quantile plot tells a somewhat different story. Certainly these data are not perfectly normal. For one thing, the data are counts and as such take on only integer values. Data from a normal distribution can take on values in an interval.

However, the dots in the normal quantile plot stayed within the guide bands produced by Minitab. So, while the pattern in the normal quantile plot is not perfectly linear, they are close enough to being linear that we can conclude these data come from an approximately normal distribution.

EXERCISE SOLUTIONS

1. a. The 5th percentile is -1.645.

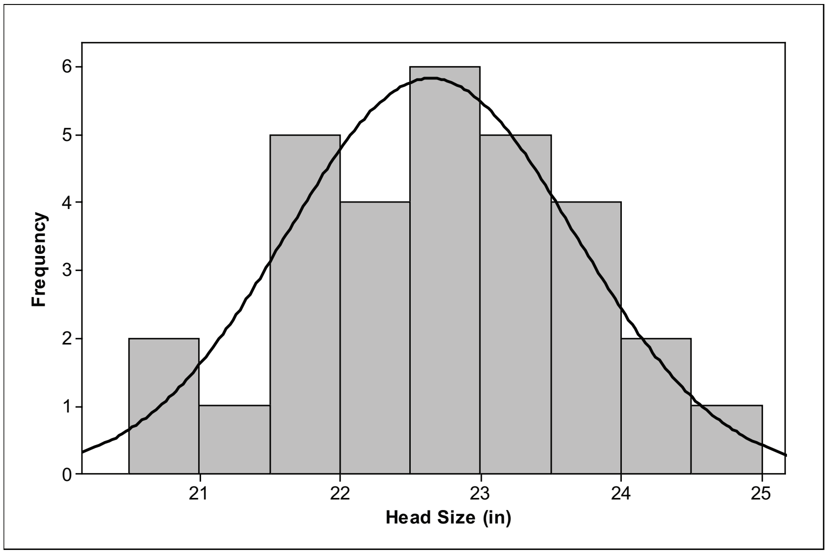
b. The 10th percentile is -1.282.

c. The 90th percentile is 1.282.

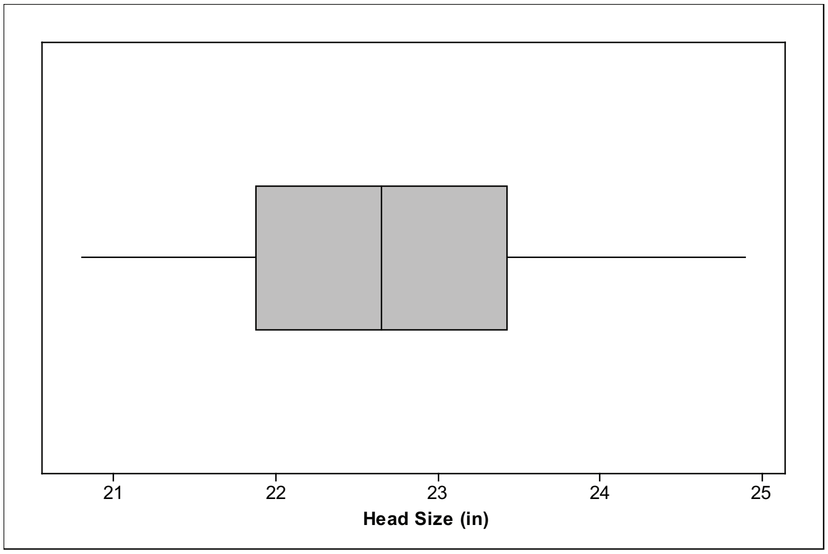
d. The 95th percentile is 1.645.

e. The 5th percentile is the opposite (or negative) of the 95th percentile. The 10th percentile is the opposite of the 90th percentile.

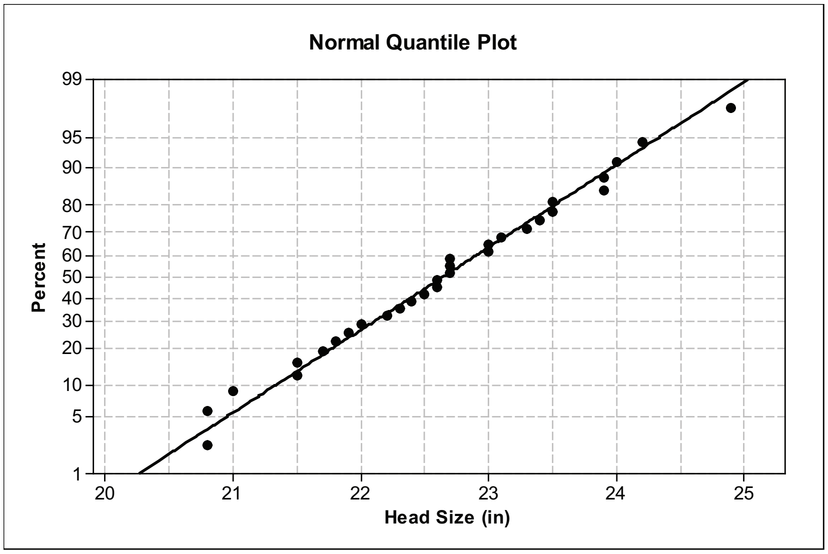
2. a. Sample answer: The histogram below is fairly mound-shaped and symmetric. However, the left side appears closer to the shape of the normal curve than the right side. Still, the fit of the normal curve appears to fit the shape of the histogram reasonably well.



b. The boxplot appears to reasonably represent normal data. Although the right whisker is slightly longer than the left whisker, the plot is fairly symmetric. The whiskers are each a bit longer than the half-width of the box. So, based on this boxplot it seems reasonable to conclude that these data are normally distributed.



c. We created the plot below using Minitab. The dots in the plot hug the line very closely. Therefore, based on this plot, it is reasonable to assume that soldiers' head sizes are normally distributed.



3. a. This histogram has a single peak and is roughly symmetric. The matching normal quantile plot should be fairly linear. So, it matches with Normal Quantile Plot #3 in Figure 9.29.

b. This histogram is skewed to the right. The matching normal quantile plot should show a concave down pattern. So, it matches with Normal Quantile Plot #1 in Figure 9.27.

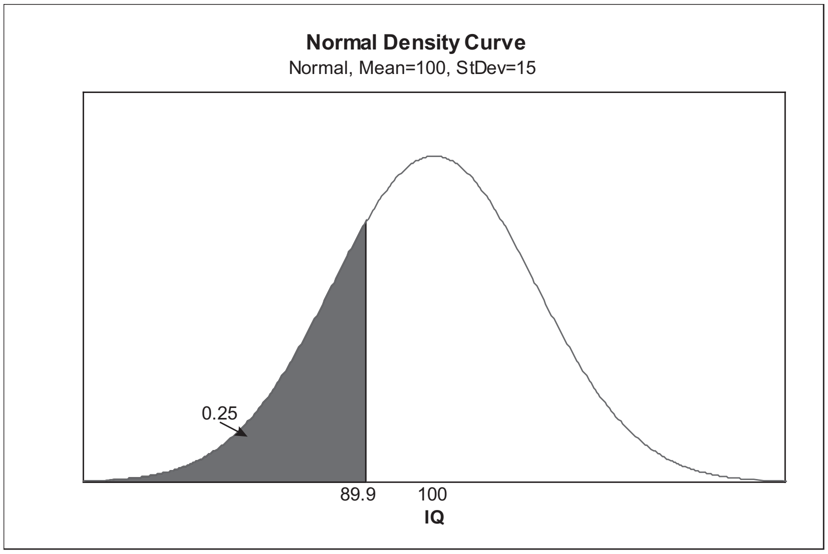
c. This histogram has a lot of data in the tails. It doesn't taper off in the tails the way a normal distribution should. So, its match is Normal Quantile Plot #2 in Figure 9.28.

REVIEW QUESTIONS SOLUTIONS

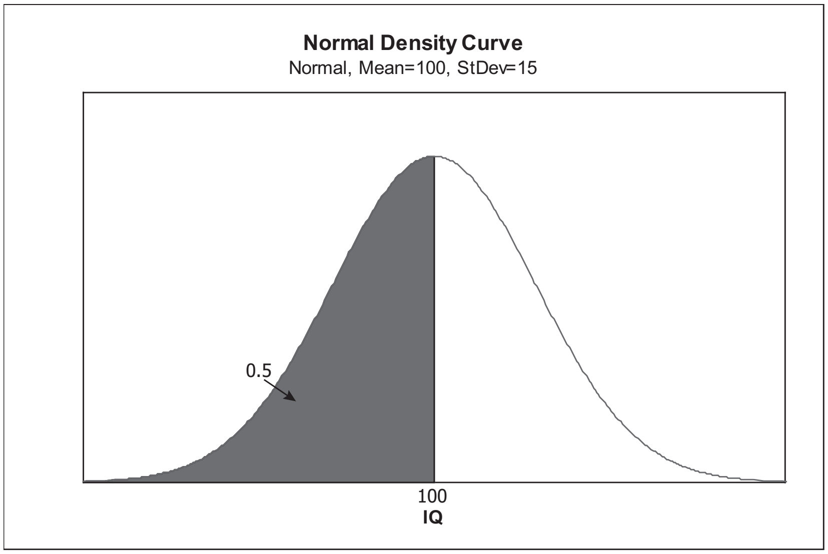
1. a. The quintiles divide the horizontal axis into five intervals. The area under the normal density curve over each of these intervals is 0.20. In other words, 20% of standard normal data will fall between two consecutive quintiles.

b. -0.8416, -0.2533, 0.2533, 0.8416

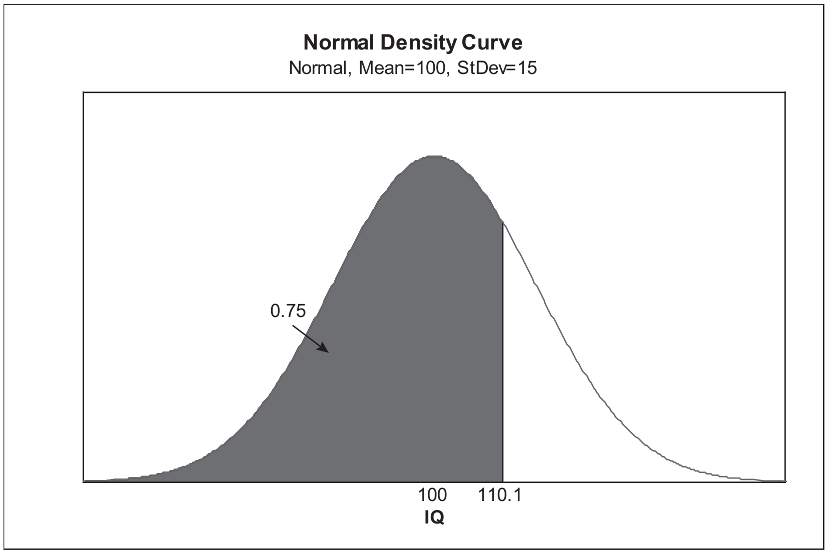
2. a. The 25th percentile for IQ scores is 89.9; 25% of IQ scores fall below this value.



b. Since the normal curve is symmetric about its mean, the 50th percentile is 100.

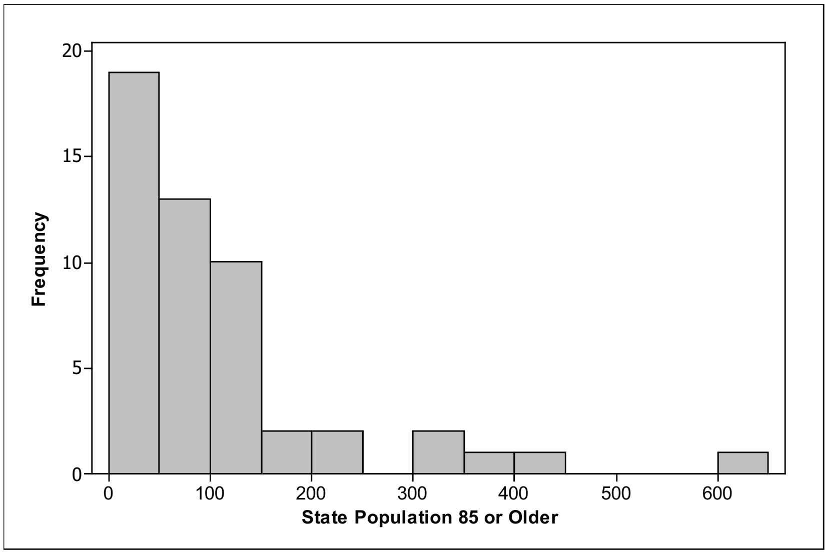


c. The 75th percentile for IQ scores is 110.1; 75% of IQ scores fall below this value.



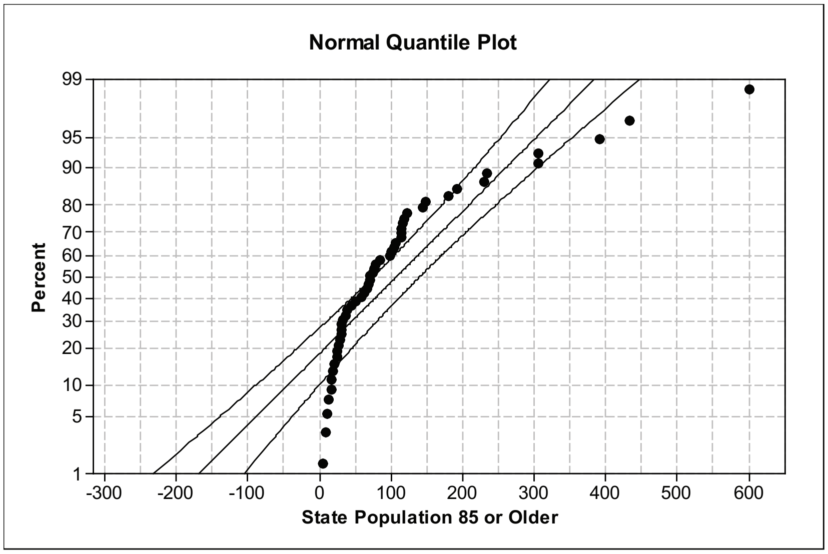
d. They are both approximately 10.1 units away from the mean of 100.

3. a. The histogram of the state populations of residents 85 and older is strongly skewed to the right. It is not reasonable to assume that these data are normally distributed.

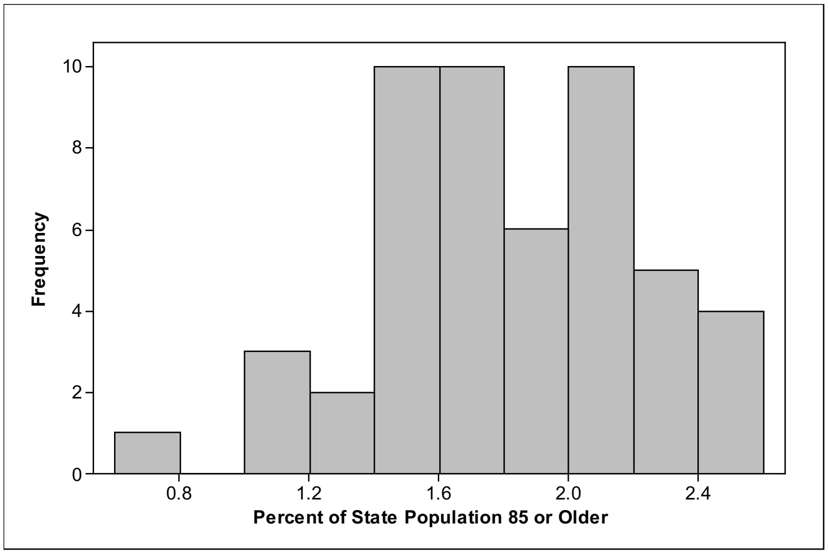


b. Because the shape of the histogram is skewed to the right, the normal quantile plot should be concave down.

c. Since the normal quantile plot below is concave down and not linear, it is not reasonable to assume that population sizes of residents 85 or older are normally distributed.

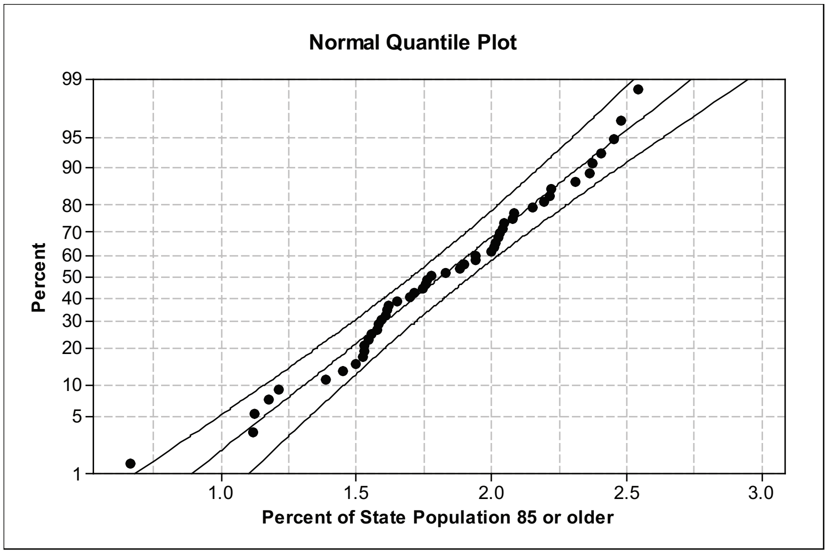


4. a. Sample answer: The histogram is roughly mound-shaped and symmetric. However, there is one valley with two peaks on either side, which may just be the messy nature of real world data. Whether or not it is reasonable to assume these data are approximately normally distributed is somewhat difficult to say. However, these data are closer to having a normal distribution than the population sizes dealt with in question 3.



b. Sample answer: Mostly linear. The data is mound-shaped (in a ragged sort of way) and very roughly symmetric. This could be an example of messy real-world data that is approximately normally distributed.

c. The normal quantile plot is mostly linear. All but one of the dots lie inside the curved bands of the normal probability plot. So, based on this plot, it is reasonable to assume that the percentages are approximately normally distributed.


Normal Quantile Plot
Percent
99
95
90
80
70
60
50
40
30
20
10
5
1
1.0
1.5
2.0
2.5
3.0
Percent of State Population 85 or older

Unit 10: Scatterplots

PREREQUISITES

Units 10, 11, and 12 form a natural cluster on describing relationships between two quantitative variables. This unit on scatterplots can be taught at a lower level than Units 11 and 12 on regression and correlation. For this unit, students need to be able to draw axes and plot ordered pairs.

ADDITIONAL TOPIC COVERAGE

Additional coverage on scatterplots can be found in *The Basic Practice of Statistics*, Chapter 4, Scatterplots and Correlation

ACTIVITY DESCRIPTION

For this activity, you will need to collect the following data from one or more classes.

- Have male students record their heights and the heights of their fathers.
- Have female students record their heights and the heights of their mothers.

Using the parent's height as the explanatory variable, students make scatterplots of student height versus parent height for females and the same for males. By combining the data and using different symbols (or colors) for males and females, students can work with multivariate data that consists of two quantitative variables and one categorical variable.

If you decide *not* to have students from your class collect the data, then use the sample data in Table T10.1, which follows this activity description. The sample answers are based on these data, which were collected from several sections of a college introductory statistics course. If your class data are from high school students, don't be surprised if you find a linear pattern for females and you do not find a linear pattern for males. Males finish growing later than females and that may influence the pattern that appears in the scatterplots.

Females		Males	
Student Height	Mother's Height (inches)	Student Height	Father's Height (inches)
62	61	75	74
66	66	72	70
68	64	70	68
60.25	62	72.5	74
68	66	67.7	71
64	61	73	70
62	65	67	61
67	64	67.7	68
66	70	65	67
63	66	71	71
68	66	70	68
64	63	75	76
60	62	73	70
69	65	70	67
66	65	71	74
67	62	69	68
66.5	64	71	72
65	64	72	70
66	65	74	72
62	62	73	68
64	69	69	72
69	70	72	70
64	66	80	70
57	62	69	68
64	60	72	78
63	65	70	68
60	63	70	68
62	59	68	68
66	67	66	68
		69	68
		72	68
		72	68

Table 10.1: Sample Class Data

THE VIDEO SOLUTIONS

1. A manatee is a large, slow-moving sea mammal. They can weigh more than half a ton and are mainly plant eaters.

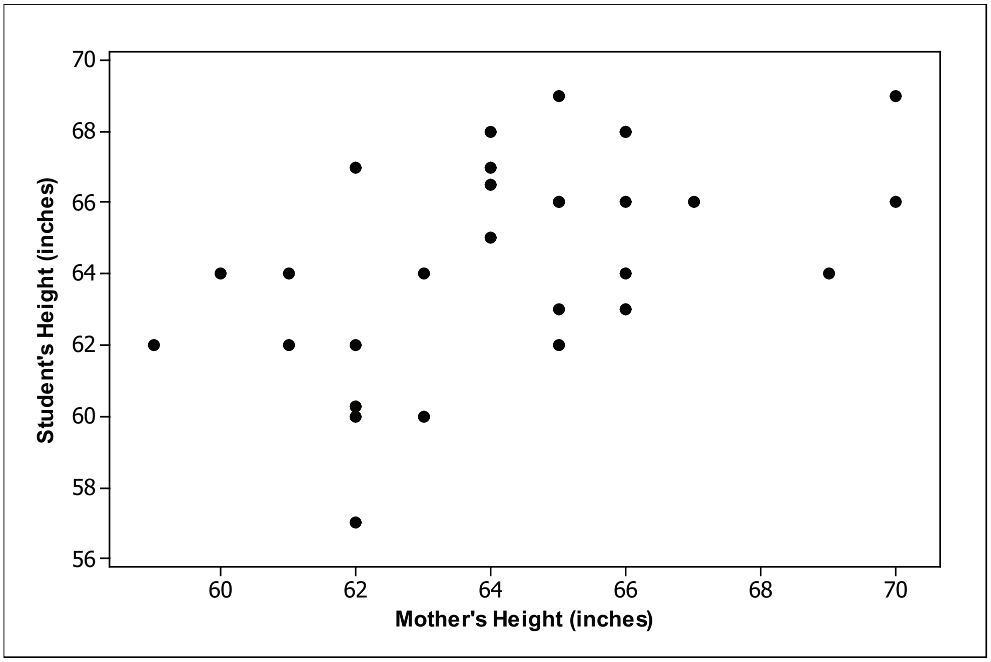
2. There is a positive association between manatees killed by powerboats and the number of powerboat registrations. In other words, as the number of powerboat registrations increases, the number of manatees killed also tends to increase.

3. The number of powerboat registrations is the explanatory variable.

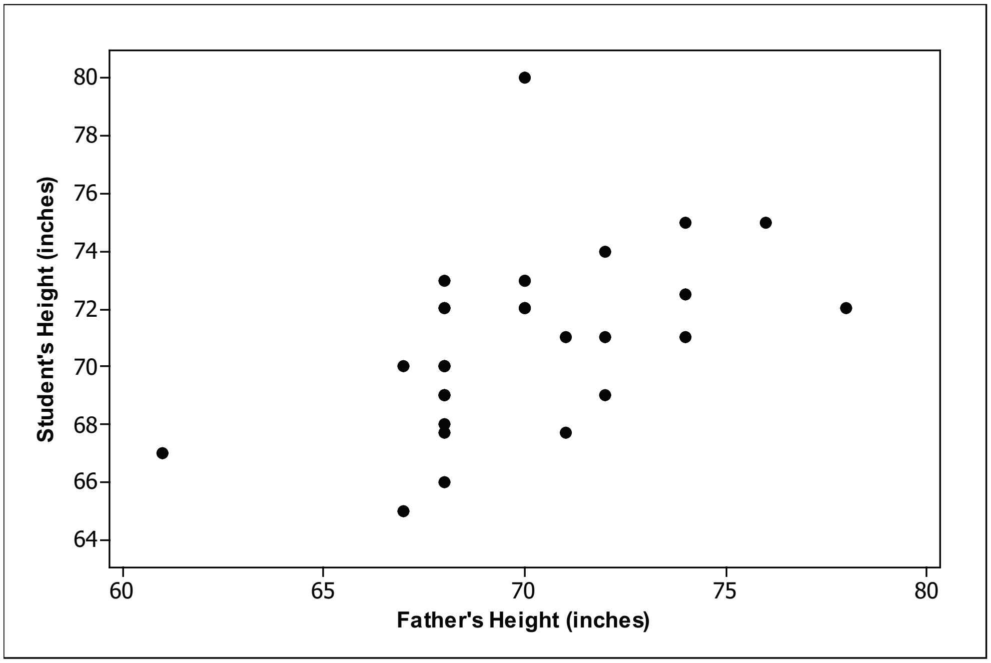
4. As one variable increases, the other tends to decrease. For example, in making pies the time it takes for you to make a particular type of pie decreases with the number of times you have made that type of pie. (The more pies you make the faster you get.)

UNIT ACTIVITY SOLUTIONS

1. Sample answer based on sample class data. (See Activity Description for sample class data.) Students could choose different scaling for axes.

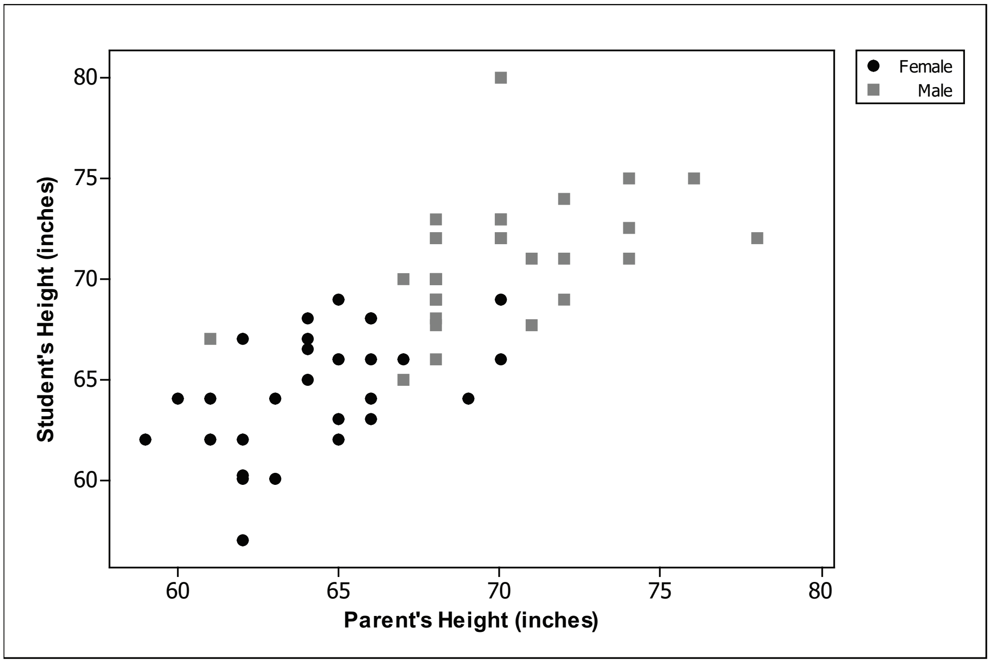


2. Sample answer:



3. Sample answer: Both scatterplots appear to have linear form with positive association. The linear trend for the males appears stronger than the linear trend for the females, but that could be just because different scaling has been used for the two graphs. For the males, there appear to be two outliers. One outlier seems to be consistent with the overall pattern, but that male is the shortest male student (and his father is also on the short side). However, there is one student whose father appears to be average in height, but that student is the tallest student in the class.

4. Sample answer:



5. Sample answer: The scatterplot of the combined data has linear form, which appears to be moderately strong. As might be expected, most of the data from the female students appears in the lower left of the scatterplot, which indicates that female students and their mothers tended to be shorter than male students and their fathers.

EXERCISE SOLUTIONS

1. a. The amount of time spent studying for a statistics exam is the explanatory variable and the grade on the exam is the response variable.

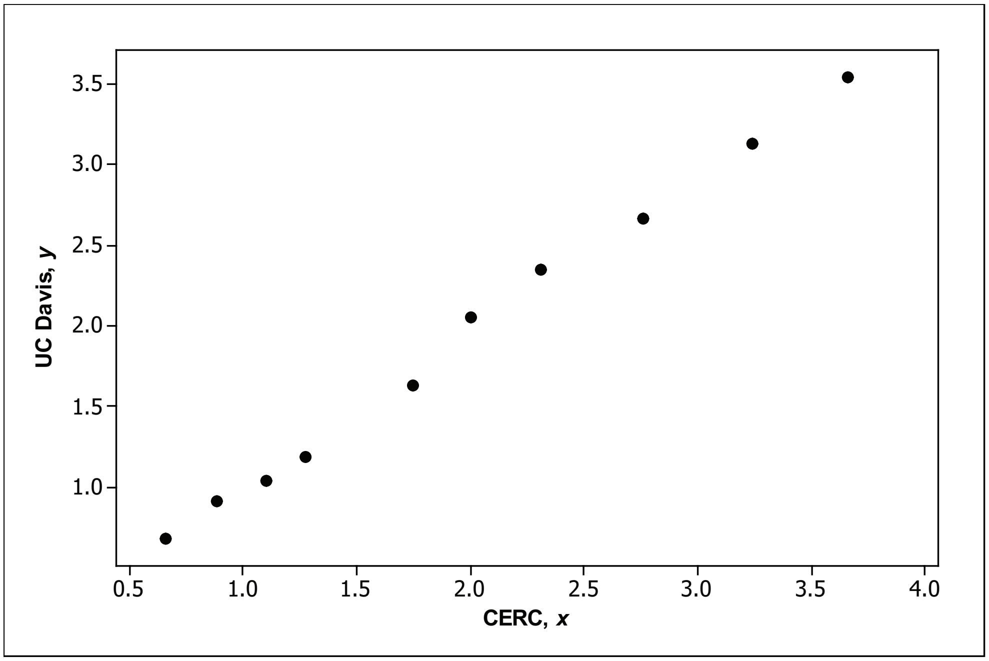
b. Sample answer #1: There is a relationship but neither explains the other.

Sample answer #2: Taller people tend to be heavier and hence height is the explanatory variable and weight is the response variable. You might want to try to predict a person's weight given how tall they are.

c. Rainfall helps explain crop yields, so rainfall is the explanatory variable and crop yields is the response variable.

d. There is a relationship between hand length and foot length. However, neither explains the other.

2. a.

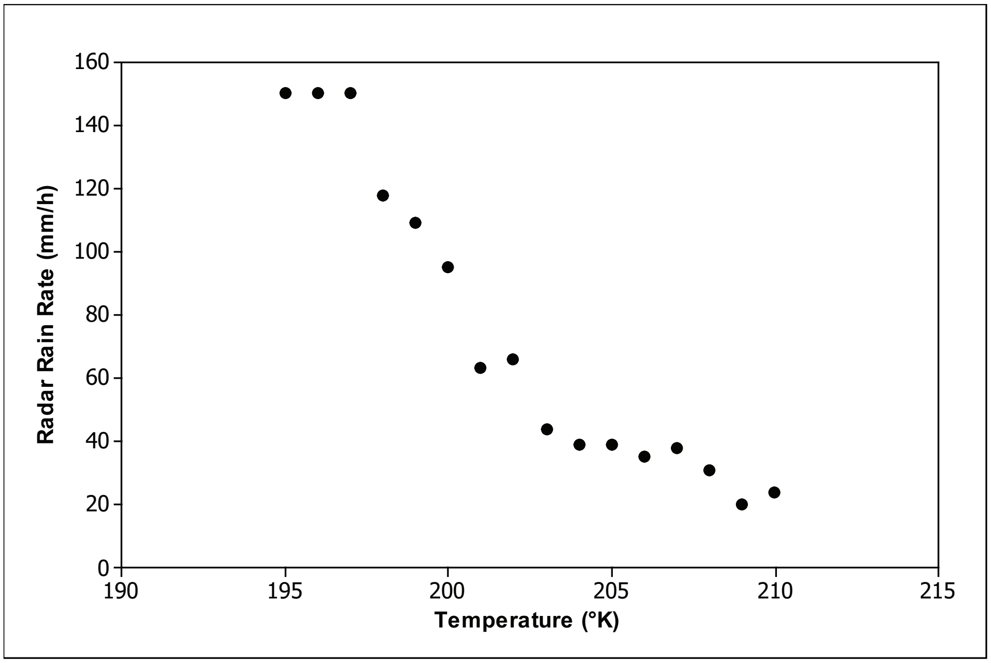


b. This is an example of positive association. Above-average mercury concentrations from CERC are associated with above-average mercury concentrations from UC Davis and below-average mercury concentrations from CERC are associated with below.

c. The pattern is linear because the dots fall close to a straight line.

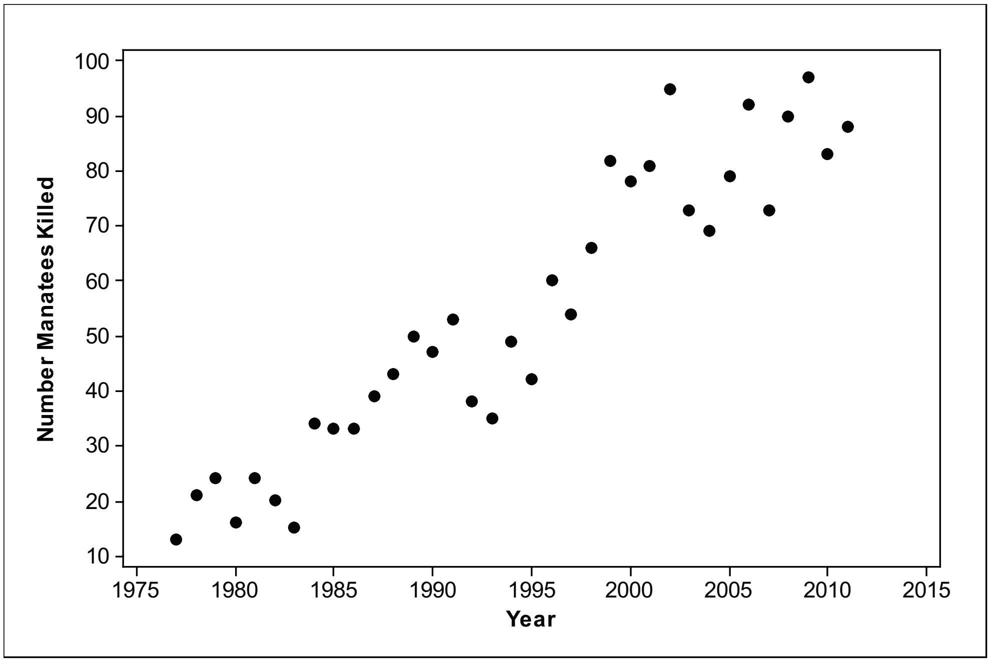
3. a. The temperature is the explanatory variable. We would like to use cloud top brightness temperatures to explain rain rate.

b.

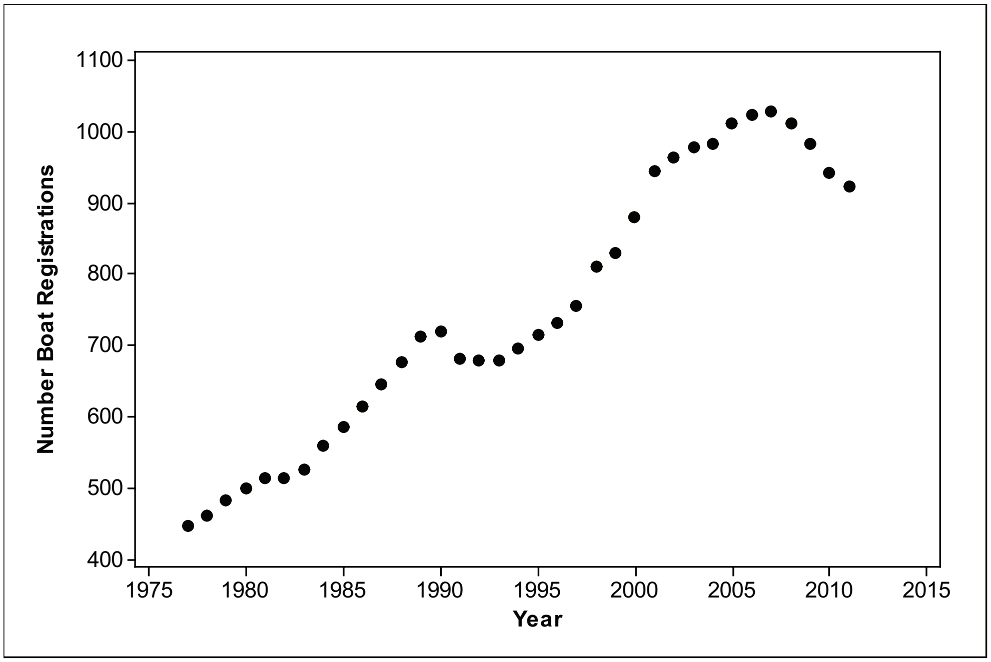


c. The association is negative – as temperature increases, rain rate tends to decrease. The pattern appears roughly linear. However, there does appear to be a slight upward bend to the pattern.

4. a. The pattern appears to be linear. As the years increase, the number of manatees killed by powerboats also tends to increase.



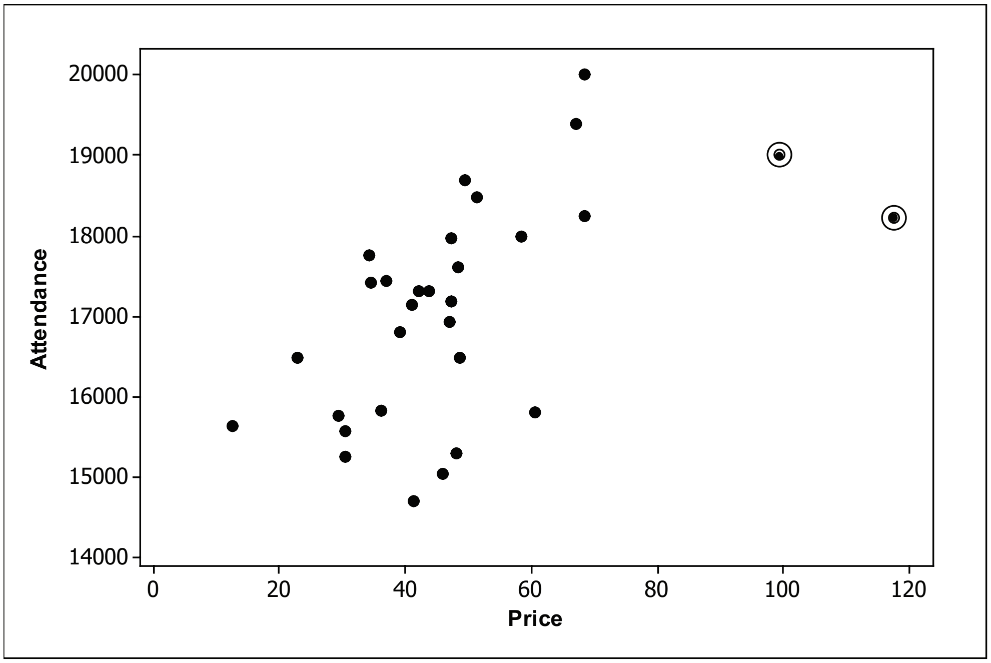
b. The pattern appears to be nonlinear. The association between the two variables appeared positive up until 2007, then it switched to negative.



REVIEW QUESTIONS SOLUTIONS

1. a. Better teams may be more popular. So, a good team can charge higher prices and also have higher attendance than a weak team. That's a *positive association*. However, high prices will drive some fans away, so we might expect higher prices to go with lower attendance. That's a negative association. We need to have the actual data before we know which of the situations described above is correct.

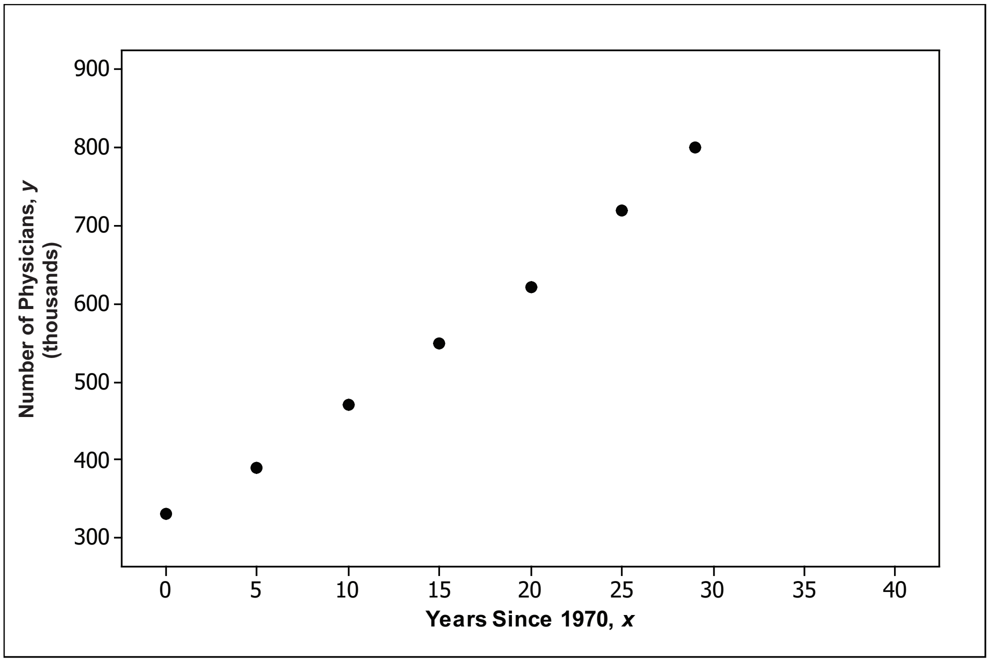
b.



Sample answer: The relationship between price and attendance appears to be a positive association. In this scatterplot, we are using ticket price to explain attendance. As price increases, attendance also tends to increase. There appear to be two outliers from what otherwise might be a linear pattern. The teams associated with these two outliers are the New York Knicks and the Los Angeles Lakers, the two teams with the highest ticket prices. These outliers could be an indication that after a certain point, high prices will begin to adversely affect attendance levels.

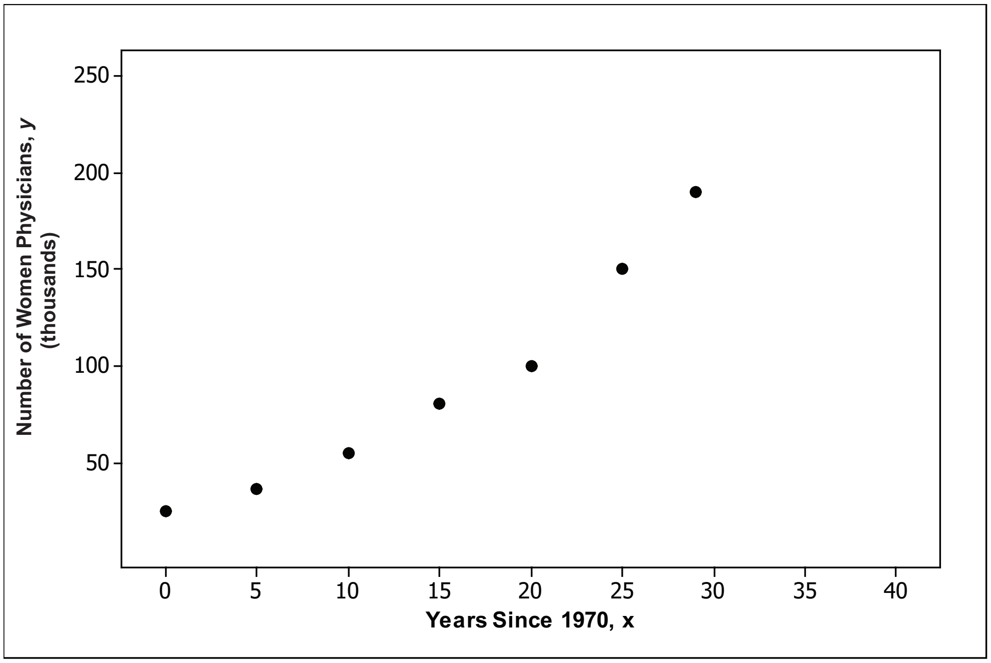
2. a) Values for *x* (second column): 0, 5, 10, 15, 20, 25, 29.

b) Note: Entries in Total Number of Physicians column in Table 10.7 have been rounded to nearest 10,000.



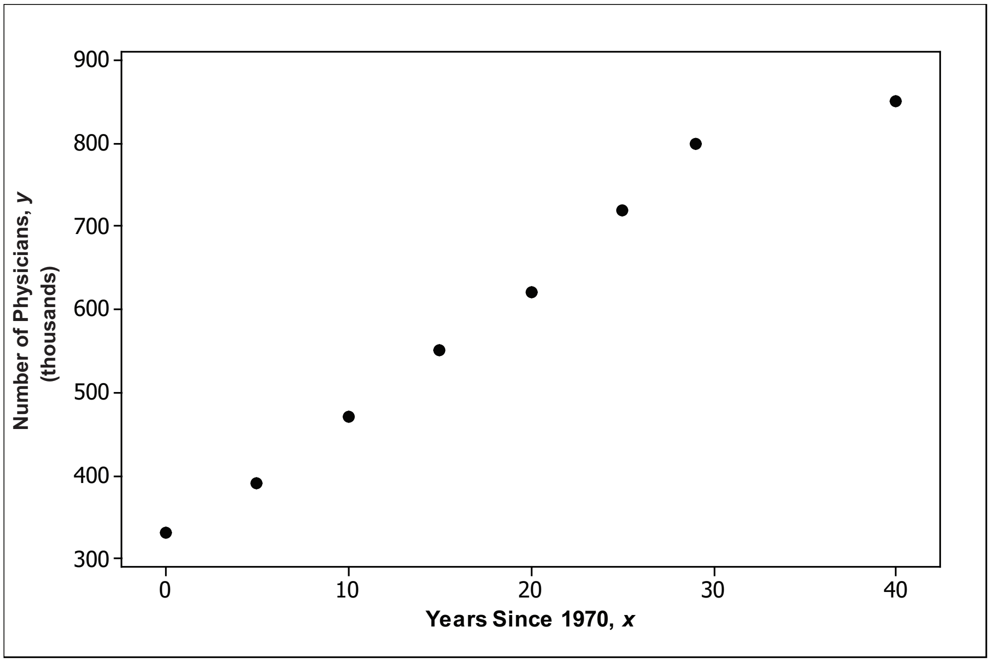
c. The pattern of the dots appears to have linear form.

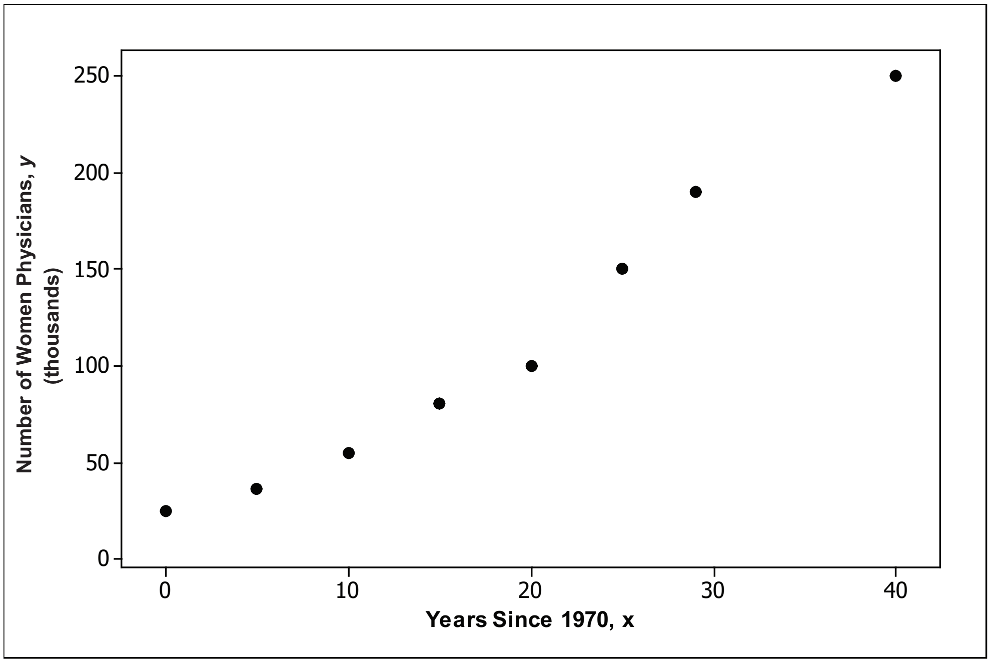
d.



e. Sample: The scatterplot of number of women physicians versus *x* is nonlinear. The pattern of dots appears to curve upward as the values of *x* increase. No matter what line you try to draw, the pattern of points will not appear randomly scattered above and below your line.

f.

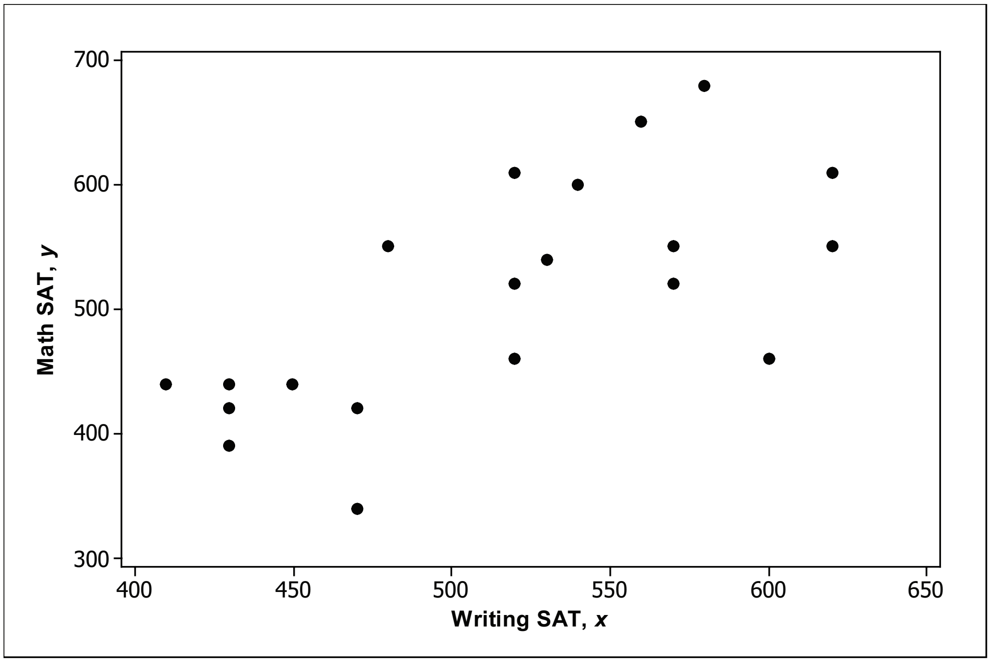




The added data value corresponding to the year 2010 deviates from the overall pattern of the original data. It looks as if the growth pattern slowed down between 1999 and 2010 compared to the growth pattern from 1979 to 1999.

Sample answer: The added data value corresponding to the year 2010 appears to fit fairly well with the overall pattern of the original data. However, it appears slightly lower than what might be expected – this could indicate that the rate of increase is slowing or it could just be normal variability of data about the overall curved pattern.

3. a.



b. There is a positive association between Math and Writing SAT scores. Students with above average Writing SATs also tend to have above average Math SATs and students with below average Writing SATs tend to have below average Math SATs. The relationship appears to have linear form.

Unit 11: Fitting Lines to Data

PREREQUISITES

Students must be familiar with material from Unit 10, Scatterplots. They must have some mathematical background on linear functions and know that the graph of a linear function is a line. Students will need to be prepared for a change in notation from what they are used to seeing in mathematics textbooks, $y = mx + b$. In this unit, a generic linear model is written as $y = a + bx$, which is consistent with the notation used in many introductory statistics textbooks. In addition, students should be familiar with the meaning of summation notation.

ACTIVITY DESCRIPTION

The unit activity provides forearm length and foot length data collected from 26 college students enrolled in an introductory statistics course. Students will need to use technology (statistical software, spreadsheet software, or graphing calculators) for computing the equation of the least-squares regression line.

MATERIALS

Access to technology with regression capabilities; graph paper (optional).

For question 2 students are asked to make a scatterplot of the data and to graph the least-squares line. They can use technology and then make a rough sketch of the results. If you want them to make a scatterplot by hand and then graph the least-squares line, they will need graph paper.

Question 7 asks students to compare the SSE for the least-squares line to the SSE of another line. From theory they should know that the least-squares line has the smaller SSE. However, as an extension to question 7, students are asked to calculate the 26 residuals associated with each of the two lines and then to calculate the SSEs. Calculator lists or Excel work really well for this computation.

Fitting a line to data that are related to the students themselves is often more interesting than working with data provided from the outside. Consider a second extension to this activity. In addition to the data provided in the activity, substitute data collected from your class and have students repeat the activity. Alternatively, students could gather data on other variables, such as height and armspan, or height and forearm length. If you want an example of a weaker relationship, have students collect data on height and how high they can jump. To gather the jump data, tape a yard stick on a wall (or use chalk to mark off a scale in inches). As a student jumps, an observer can record his or her height. Another alternative is to gather data on height and stride length. Forensic scientists might be able to determine the stride length of a criminal from foot prints and then use that information to predict the person's height.

THE VIDEO SOLUTIONS

1. Researchers push long tubes that have scales along the side into the snowpack.

2. It is the difference between the actual *y*-value and the *y*-value predicted by the least-squares line.

3. It finds the line for which the sum of the squares of the residuals is smallest.

4. Substitute the value for snowpack into the equation of the least-squares line to get the predicted value of water runoff.

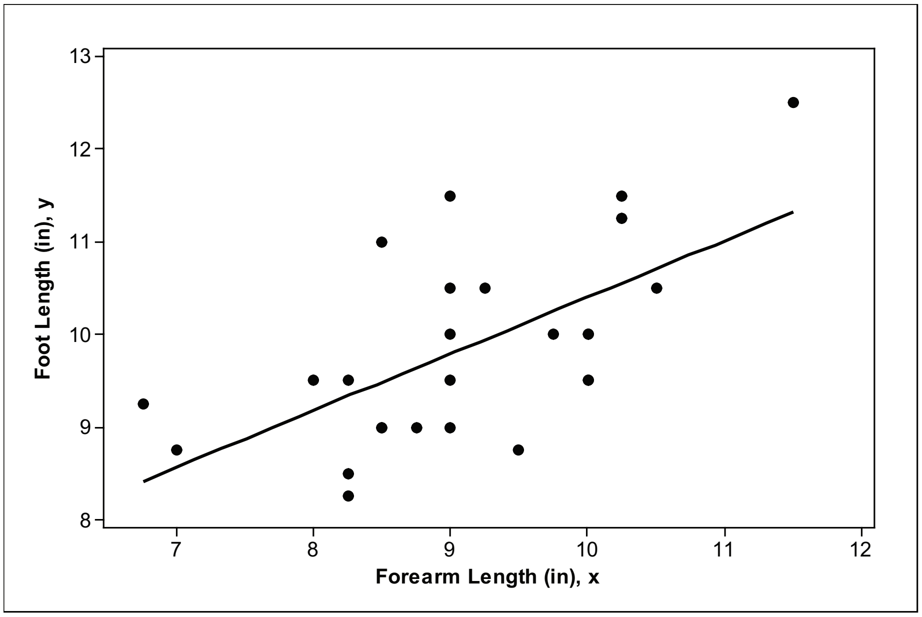
5. If the dots appear randomly scattered with no strong pattern, then the regression line is adequate to describe the pattern in the data. If the dots in a residual plot appear to have a strong curved pattern, then the linear model is not adequate to describe the pattern in the data and you need to look for a new model.

UNIT ACTIVITY SOLUTIONS

1. a. See solution to question 2.

b. Yes. The pattern of the dots in the scatterplot go from the lower left to the upper right.

2. The equation of the least-squares line is y = 4.291 + 0.6112x. Yes, the line provides a reasonable summary of the forearm-foot length data.



3. $\hat{y} = 4.291 + 0.6112(10.5) \approx 10.7 \text{ inches}$.

4. First, calculate the predicted y-value: $\hat{y} = 4.291 + 0.6112(10) = 10.403$.

Residual = 9.50 – 10.403 = -0.903 inch.

5. The dots in the residual plot appear randomly scattered with no strong pattern. Therefore, the least-squares line is adequate to describe the pattern in the data.

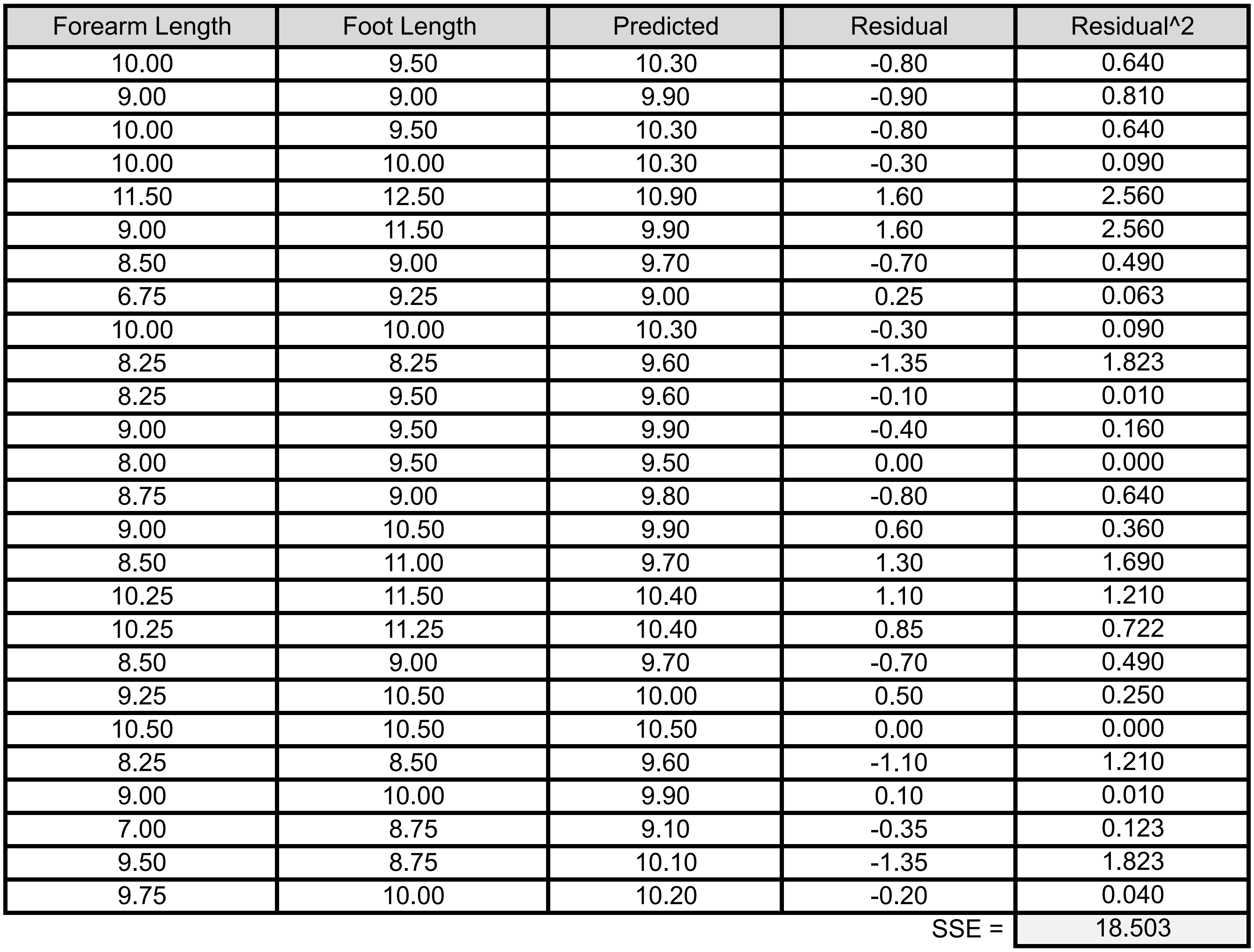
6. y = -4.525 + 1.579x; It didn't even come out close to being the same. Clearly Sarah's strategy for fixing Danny's error was faulty.

7. The SSE for her line will be larger. The least-squares line is the line with the smallest SSE of all possible lines.

Extension to question 7: As expected, the least-squares line has the smaller SSE.

SSE for least-squares line = 17.025.

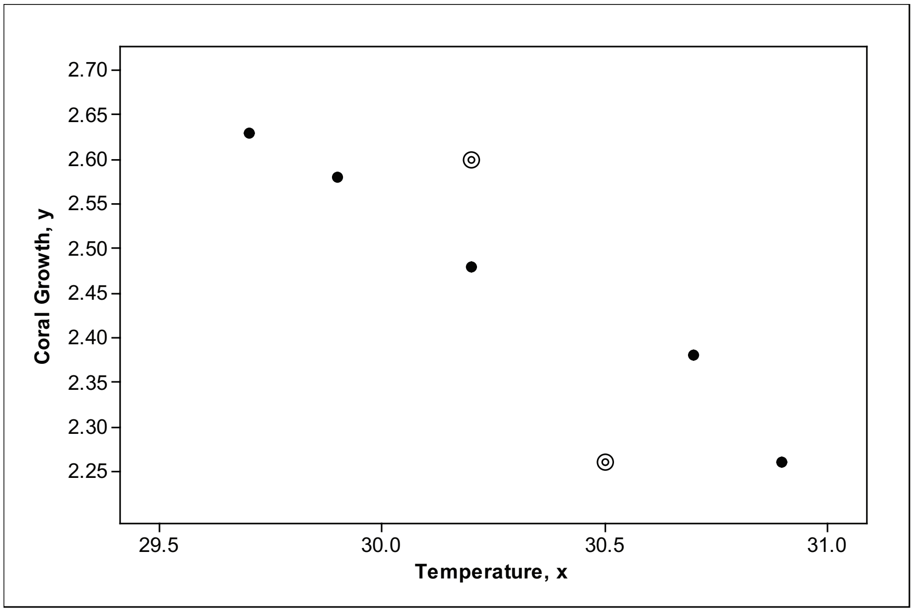
SSE for Linda’s line = 18.503. (See spreadsheet table below for calculation of the residuals for Linda’s line.)



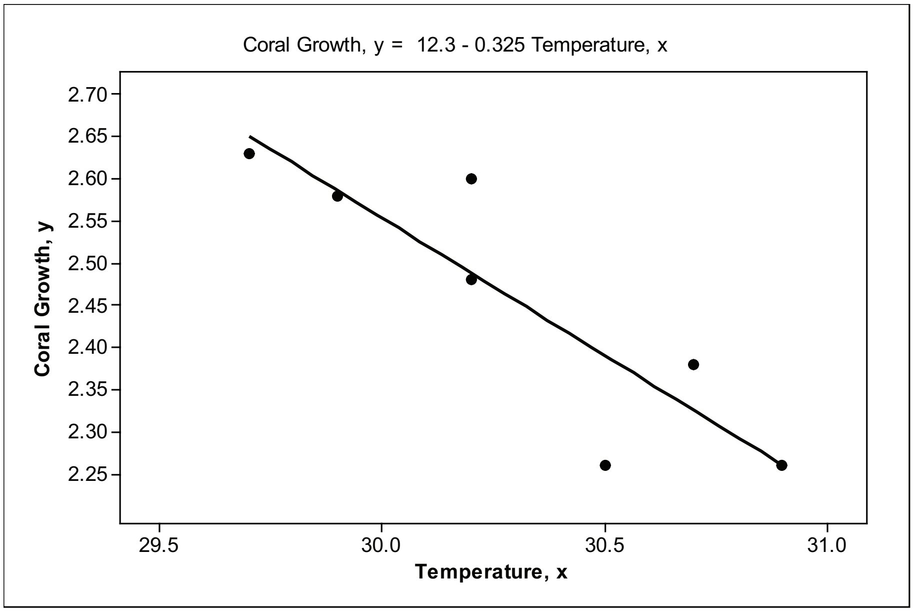
Forearm Length	Foot Length	Predicted	Residual	Residual^2
10.00	9.50	10.30	-0.80	0.640
9.00	9.00	9.90	-0.90	0.810
10.00	9.50	10.30	-0.80	0.640
10.00	10.00	10.30	-0.30	0.090
11.50	12.50	10.90	1.60	2.560
9.00	11.50	9.90	1.60	2.560
8.50	9.00	9.70	-0.70	0.490
6.75	9.25	9.00	0.25	0.063
10.00	10.00	10.30	-0.30	0.090
8.25	8.25	9.60	-1.35	1.823
8.25	9.50	9.60	-0.10	0.010
9.00	9.50	9.90	-0.40	0.160
8.00	9.50	9.50	0.00	0.000
8.75	9.00	9.80	-0.80	0.640
9.00	10.50	9.90	0.60	0.360
8.50	11.00	9.70	1.30	1.690
10.25	11.50	10.40	1.10	1.210
10.25	11.25	10.40	0.85	0.722
8.50	9.00	9.70	-0.70	0.490
9.25	10.50	10.00	0.50	0.250
10.50	10.50	10.50	0.00	0.000
8.25	8.50	9.60	-1.10	1.210
9.00	10.00	9.90	0.10	0.010
7.00	8.75	9.10	-0.35	0.123
9.50	8.75	10.10	-1.35	1.823
9.75	10.00	10.20	-0.20	0.040
			SSE =	18.503

EXERCISE SOLUTIONS

1. a. Sample answer: There do not appear to be any real outliers. However, the two circled data points depart somewhat from what would otherwise be a strong linear pattern.



b.



2. a. $\bar{x} = 30.3$; $\bar{y} = 2.46$

b.

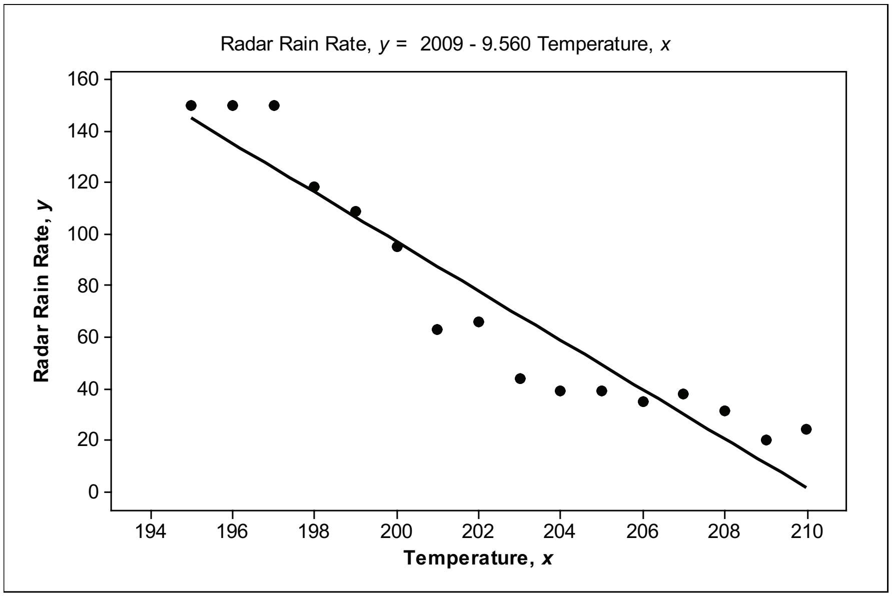
x	y	$(x-30.3)$	$(y-2.46)$	$(x-30.3)(y-2.46)$	$(x-30.3)^2$
29.7	2.63	-0.6	0.17	-0.102	0.36
29.9	2.58	-0.4	0.12	-0.048	0.16
30.2	2.6	-0.1	0.14	-0.014	0.01
30.2	2.48	-0.1	0.02	-0.002	0.01
30.5	2.26	0.2	-0.2	-0.04	0.04
30.7	2.38	0.4	-0.08	-0.032	0.16
30.9	2.26	0.6	-0.2	-0.12	0.36
			Sum =	-0.358	1.1

c. The slope *b* = -0.358/1.10 ≈ -0.325; the *y*-intercept *a* = 2.46 – (-0.325)(30.3) ≈ 12.3

3. a. Sample answer: The dots appear randomly scattered (although it is hard to tell if there is a strong pattern with so few points). Four dots are below the *x*-axis and 3 are on or above the *x*-axis. So, it appears that the line has taken out all the pattern in the data leaving only random noise. We can conclude that the least-squares regression line is adequate to describe the pattern in the data.

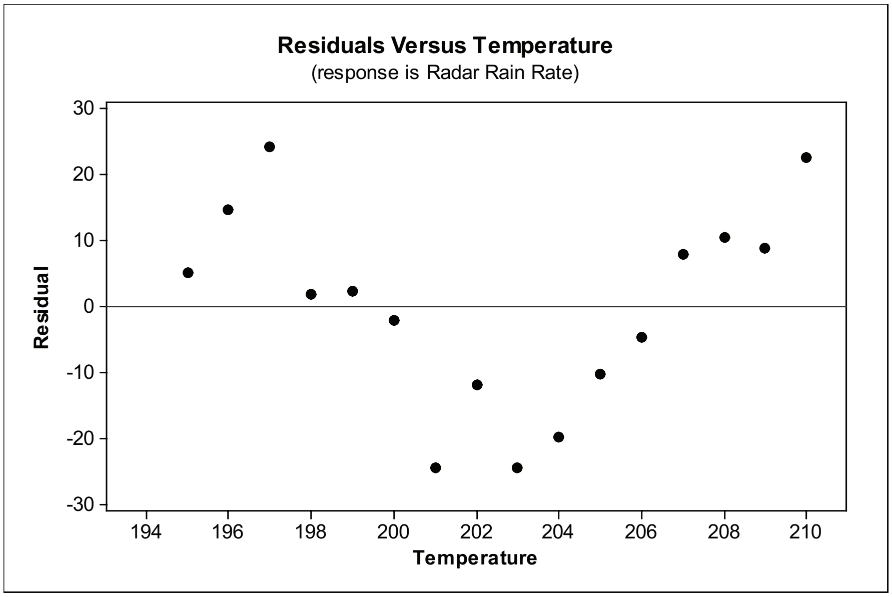
b. $\hat{y} = 12.3 - 0.325(40) \approx -0.7$

4. a. *y* = 2,009 - 9.56*x*, where *x* is temperature and *y* is radar rain rate.



b. 2,009 – 9.56(220) = -94.2 mm/h. No, you can't have a negative rain rate. This is an example of what can happen when you extrapolate beyond the observed data values.

c. There is a strong V-shaped pattern to the dots in the residual plot. A straight-line model is not adequate to describe the pattern in these data.

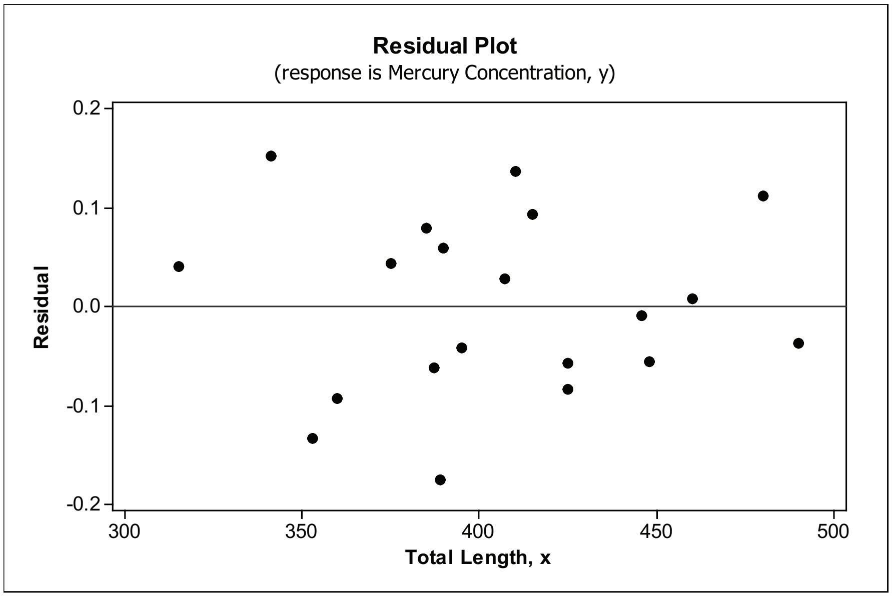

Residuals Versus Temperature
(response is Radar Rain Rate)
Residual
30
20
10
0
-10
-20
-30
194
196
198
200
202
204
206
208
210
Temperature

REVIEW QUESTIONS SOLUTIONS

1. a. Femur bone length is the explanatory variable since we wish to use it to explain a person's height. Therefore, height is the response variable.

b. The data appear to form a linear pattern. Men with longer femurs tend to be taller than men with shorter femurs. Thus, the association is positive.

c. The equation of the least-squares line is *y* = 51.01 + 0.2637*x*. The scatterplot with a graph of the least-squares line appears below.



d. Sample answer: There appears to be one outlier – data point (508, 198). This male is taller than we would expect given the pattern in the data. (See graph in (c). This point has been plotted with an open double circle.) The man is 198 cm tall or around 6' 6" tall. There are men that are this tall, even though it is quite tall for a man. So, it doesn't appear to be an error.

2. a. The slope is 0.2637; for each one mm increase in femur bone length we expect about a 0.26 cm increase in height. This makes sense in context.

The *y*-intercept is 51.01; this is the predicted height in centimeters of a person whose femur length is 0 mm. A femur length of 0 mm is far outside the range of observed femur lengths – the person would be missing his/her thigh. This does not make sense in context.

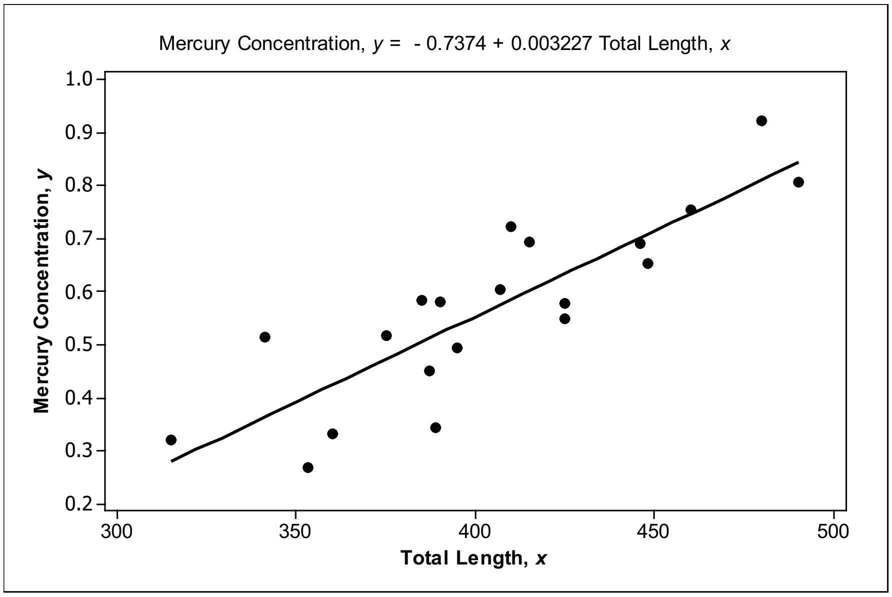
b. The slope gives the predicted change in height (cm) for each one millimeter increase in femur length. Hence, we would predict a (5)(0.264) or around 1.3 cm difference in height in response to a 5 mm difference in femur length.

c. $\hat{y} = 51.0 + 0.264(475) \approx 176.4 \text{ cm}$; or a little less than 5' 9½" tall. This is a reasonable height for a man.

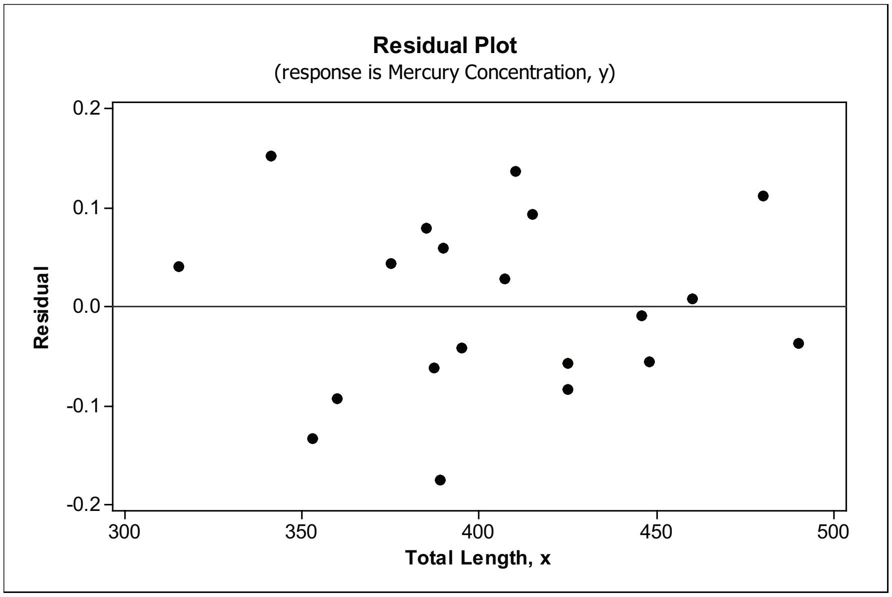
d. $\hat{y} = 51.0 + 0.264(250) \approx 117 \text{ cm}$; or around 3' 10" tall. This is an example of extrapolation. The smallest femur length in the data is 422 mm. All of the data is for adult males. We have no idea if the relationship between height and femur length is the same for children as it is for adults.

3. a. Total length is the explanatory variable, *x*, and mercury concentration is the response variable, *y*.

b. *y* = 0.0032*x* - 0.7374, where *x* is fish length and *y* is mercury concentration.



c. The least-squares line is adequate to describe the overall pattern in the data. The dots in the residual plot appear to be randomly scattered. Also, there is a good split between dots above the *x*-axis and below the *x*-axis.



d. For each additional 1 mm in fish length, we expect mercury concentration to increase by 0.0032 μg/g. This makes sense in the given context.

e. The *y*-intercept of the least squares line indicates that (0, - 0.7374) lies on the least-squares line. This means that a fish that is 0 mm in length will have a mercury concentration level of -0.7374 μg/g. It does not make sense for a fish to have zero length or a negative level of mercury concentration.

4. a. We used the equation of the least-squares line with constants rounded to four decimals: *y* = 0.0032*x* - 0.7374.

The prediction of mercury concentration is 0.0032(430) - 0.7374 = 0.6386 μg/g.

This prediction is an example of interpolation since *x* = 430 mm is between 315 mm and 490 mm, the range of the fish lengths in the observed data.

b. The prediction is 0.0032(90) - 0.7374 = -0.4494 μg/g.

This is an example of extrapolation. The length of the fish is far below the length of the smallest fish represented in the data. Furthermore, the sample of fish in the observed data were all of legal/edible size and this fish is not of legal/edible size.

5. a. SSE = $(0)^2 + (3)^2 + (0)^2 + (-2)^2 = 13$

b. SSE = $(-1.6)^2 + (2.3)^2 + (0.2)^2 + (-0.9)^2 = 8.7$

c. The least-squares line is the line that has the smallest SSE of all lines.