## Chapter 2: Organizing Data

Section 2.1

1. Class limits are possible data values, and they specify the span of data values that fall within a class. Class boundaries are not possible data values; they are values halfway between the upper class limit of one class and the lower class limit of the next class.

2. Each data value must fall into one class. Data values above 50 do not have a class.

3. The classes overlap. A data value such as 20 falls into two classes.

4. These class widths are 11.

5. $\text{Width} = \dfrac{82-20}{7} \approx 8.86$, so round up to 9. The class limits are $20-28, 29-37, 38-46, 47-55, 56-64, 65-73, 74-82$.

6. $\text{Width} = \dfrac{120-10}{5} = 22$, so round up to 23. The class limits are $10-32, 33-55, 56-78, 79-101, 102-124$.

7. **(a)** The distribution is most likely skewed right, with many short times and only a few long wait times.
   (b) A bimodal distribution might exist if there are different wait times during busy versus slow periods. During the morning rush, many long wait times might occur, but during the slow afternoon, most wait times will be very short.

8. The data set consists of the numbers 1 up through 100, with each value occurring once. The histogram will be uniform.

9. **(a)** Yes.
   (b)

10. (a)

Histogram of Salary

(b) Yes. Yes.

(c)

Histogram of Salary
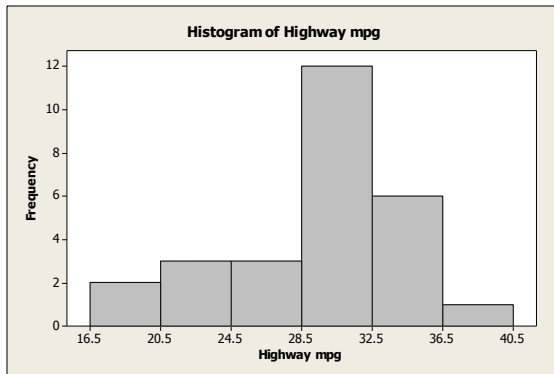
11. (a) The range of data seem to fall between 7 and 13 with the bulk of the data between 8 and 12.

(b) All three histograms are somewhat mound-shaped with the top of the mound between 9.5 and 10.5. In all three histograms the bulk of the data fall between 8 and 12.

12. (a) Graph (i) midpoint: 5; graph (ii) midpoint: 4; graph (iii) midpoint: 2.

(b) Graph (i) 0-17; Graph (ii) 1-16; Graph (iii) 0-28.

(c) Graph (iii) is most clearly skewed right; Graph (ii) is somewhat skewed right; Graph (i) is barely skewed right.

(d) No, each random sample of same size froma population is equally likely to be drawn. Sample (iii) most clearly reflects the properties of the population. Sample (ii) reflects the properties fairly well, but sample(i) seems to differ more from the described population.

13. (a) Because there are 50 data values, divide each cumulative frequency by 50 and convert to a percent.

(b) 35 states.

(c) 6 states.

(d) 2%

14. (a) Graph (i).

(b) Graph (iii)

(c) Graph (iii)

15. (a) Class width = 25

(b)

Class Limits	Class Boundaries	Midpoints	Frequency	Relative Frequency	Cumulative Frequency
236–260	235.5–260.5	248	4	0.07	4
261–285	260.5–285.5	273	9	0.16	13
286–310	285.5–310.5	298	25	0.44	38
311–335	310.5–335.5	323	16	0.28	54
336–360	335.5–360.5	348	3	0.05	57

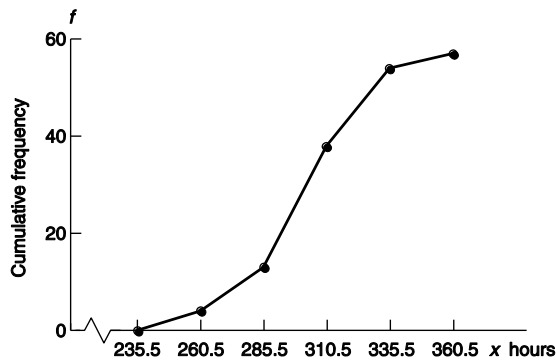**(c)**

(d)

Histogram of Finish Times
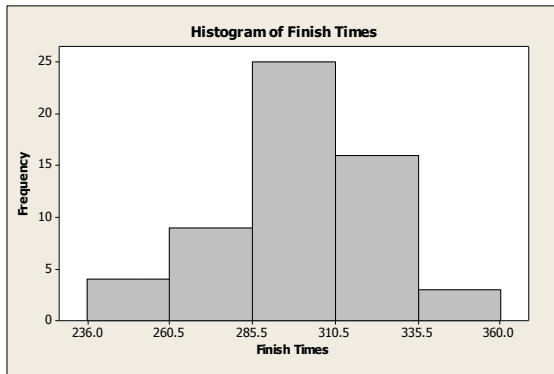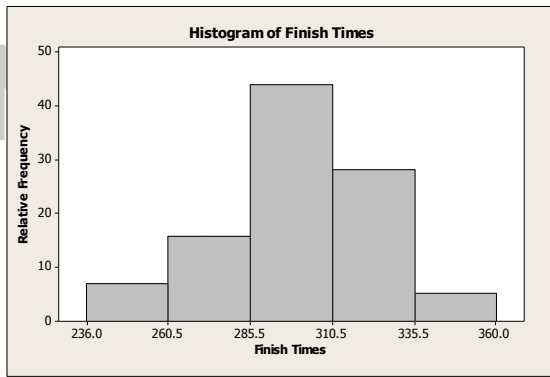
(e) This distribution is slightly skewed to the left but fairly mound-shaped, symmetric.
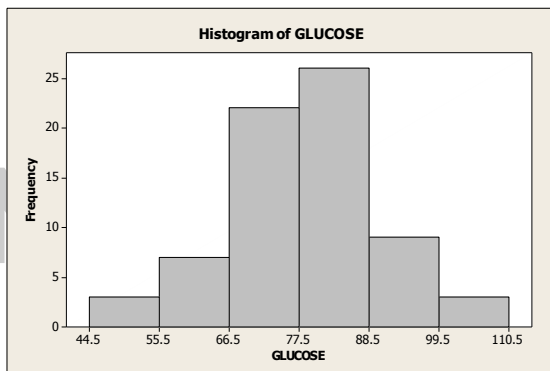
(f)

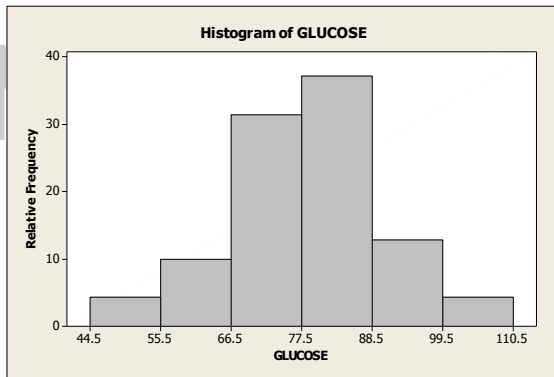Hours to Complete Iditarod—Ogive

16. (a) Class width = 11

(b)

Class Limits	Class Boundaries	Midpoint	Frequency	Relative Frequency	Cumulative Frequency
45–55	44.5–55.5	50	3	0.0429	3
56–66	55.5–66.5	61	7	0.8714	10
67–77	66.5–77.5	72	22	0.3143	32
78–88	77.5–88.5	83	26	0.3714	58
89–99	88.5–99.5	94	9	0.1286	67
100–110	99.5–110.5	105	3	0.0429	70

(c)

Histogram of GLUCOSE

11

(d)

Histogram of GLUCOSE
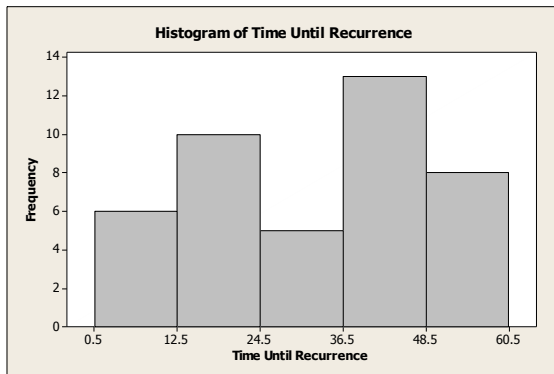
(e) Approximately mound-shaped, symmetric.

(f) To create the ogive, place a dot on the *x* axis at the lower class boundary of the first class, and then, for each class, place a dot above the upper class boundary value at the height of the cumulative frequency for the class. Connect the dots with line segments.

17. (a) Class width = 12

(b)

Class Limits	Class Boundaries	Midpoint	Frequency	Relative Frequency	Cumulative Frequency
1–12	0.5–12.5	6.5	6	0.14	6
13–24	12.5–24.5	18.5	10	0.24	16
25–36	24.5–36.5	30.5	5	0.12	21
37–48	36.5–48.5	42.5	13	0.31	34
49–60	48.5–60.5	54.5	8	0.19	42

(c)

Histogram of Time Until Recurrence

(d)

Histogram of Time Until Recurrence

(Relative Frequency vs. Time Until Recurrence; x-axis values: 0.5, 12.5, 24.5, 36.5, 48.5, 60.5; y-axis 0 to 35)

(e) The distribution is bimodal.

(f) To create the ogive, place a dot on the *x* axis at the lower class boundary of the first class, and then, for each class, place a dot above the upper class boundary value at the height of the cumulative frequency for the class. Connect the dots with line segments.
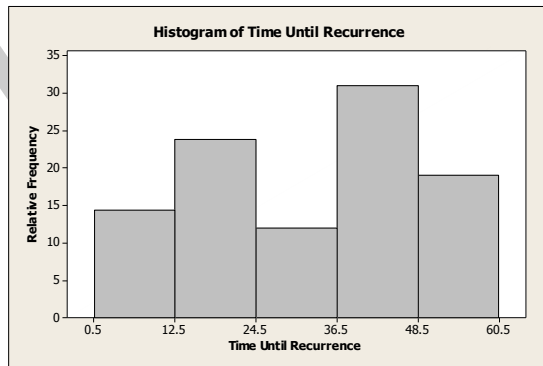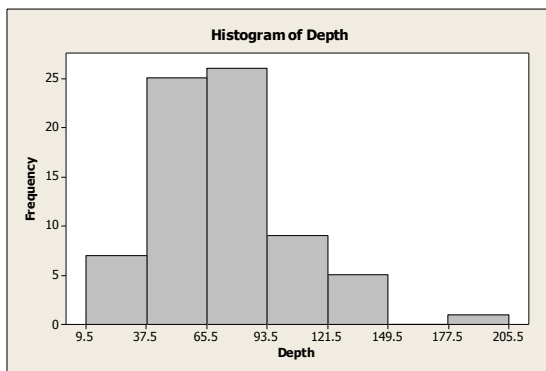
18. (a) Class width = 28.

(b)

Class Limits	Class Boundaries	Midpoint	Frequency	Relative Frequency
10–37	9.5–37.5	23.5	7	7
38–65	37.5–65.5	51.5	25	32
66–93	65.5–93.5	79.5	26	58
94–121	93.5–121.5	107.5	9	67
122–149	121.5–149.5	135.5	5	72
150–177	149.5–177.5	163.5	0	72
178–205	177.5–205.5	191.5	1	73

(c)

Histogram of Depth

(Frequency vs. Depth; x-axis values: 9.5, 37.5, 65.5, 93.5, 121.5, 149.5, 177.5, 205.5; y-axis 0 to 25)

(d)

(e) This distribution is skewed right with a possible outlier.

(f) To create the ogive, place a dot on the *x* axis at the lower class boundary of the first class, and then, for each class, place a dot above the upper class boundary value at the height of the cumulative frequency for the class. Connect the dots with line segments.
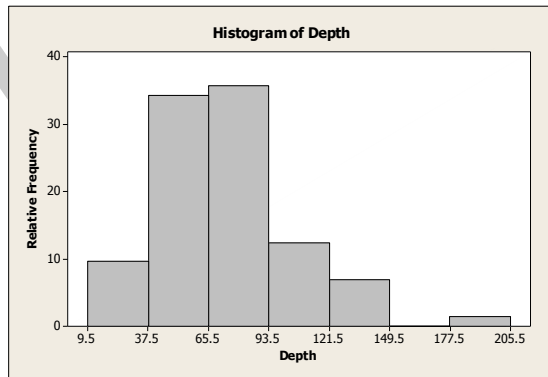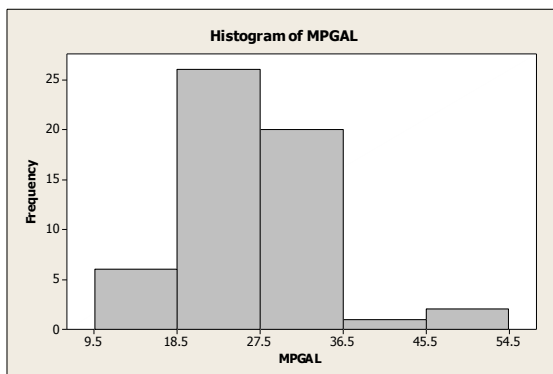
19. (a) Class width = 9

(b)

Class Limits	Class Boundaries	Midpoint	Frequency	Relative Frequency	Cumulative Frequency
10–18	9.5–18.5	14	6	0.11	6
19–27	18.5–27.5	23	26	0.47	32
28–36	27.5–36.5	32	20	0.36	52
37–45	36.5–45.5	41	1	0.02	53
46–54	45.5–54.5	50	2	0.04	55

(c)

(d)

Histogram of MPGAL

(e) This distribution is skewed right.

(f) To create the ogive, place a dot on the *x* axis at the lower class boundary of the first class, and then, for each class, place a dot above the upper class boundary value at the height of the cumulative frequency for the class. Connect the dots with line segments.
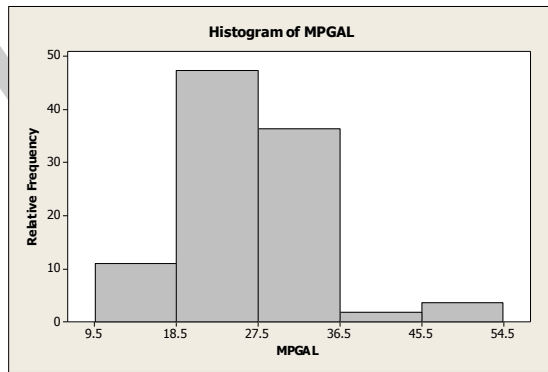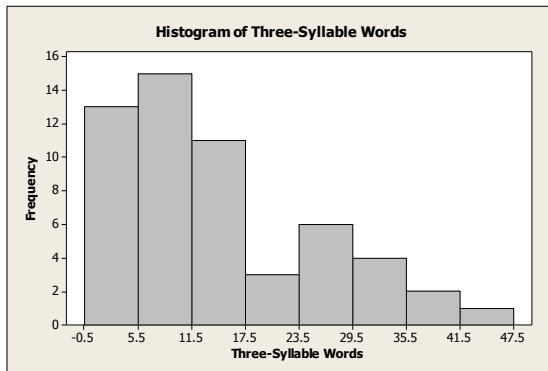
20. (a) Class width = 6

(b)

Class Limits	Class Boundaries	Midpoint	Frequency	Relative Frequency	Cumulative Frequency
0–5	0.5–5.5	2.5	13	0.24	13
6–11	5.5–11.5	8.5	15	0.27	28
12–17	11.5–17.5	14.5	11	0.20	39
18–23	17.5–23.5	20.5	3	0.05	42
24–29	23.5–29.5	26.5	6	0.11	48
30–35	29.5–35.5	32.5	4	0.07	52
36–41	35.5–41.5	38.5	2	0.04	54
42–47	41.5–47.5	44.5	1	0.02	55

(c)

Histogram of Three-Syllable Words

15

(d)

Histogram of Three-Syllable Words

(chart: Relative Frequency vs Three-Syllable Words, x-axis labels -0.5, 5.5, 11.5, 17.5, 23.5, 29.5, 35.5, 41.5, 47.5)
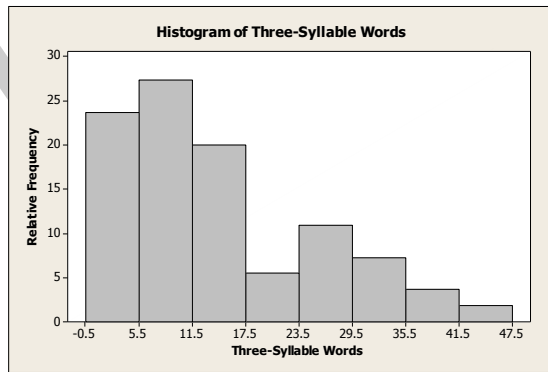
(e) The distribution is skewed right.

(f) To create the ogive, place a dot on the *x* axis at the lower class boundary of the first class, and then, for each class, place a dot above the upper class boundary value at the height of the cumulative frequency for the class. Connect the dots with line segments.

Words of Three Syllables or More—Ogive

(chart: Cumulative frequency *f* vs Number of Words *x*, x-axis labels −0.5, 5.5, 11.5, 17.5, 23.5, 29.5, 35.5, 41.5, 47.5)
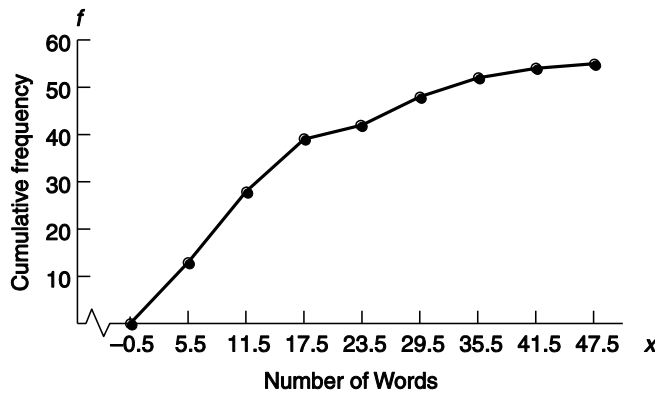
21. **(a)** Multiply each value by 100.

(b)

Class Limits	Class Boundaries	Midpoint	Frequency
46–85	45.5–85.5	65.5	4
86–125	85.5–125.5	105.5	5
126–165	125.5–165.5	145.5	10
166–205	165.5–205.5	185.5	5
206–245	205.5–245.5	225.5	5
246–285	245.5–285.5	265.5	3

Histogram of Tonnes

(c)
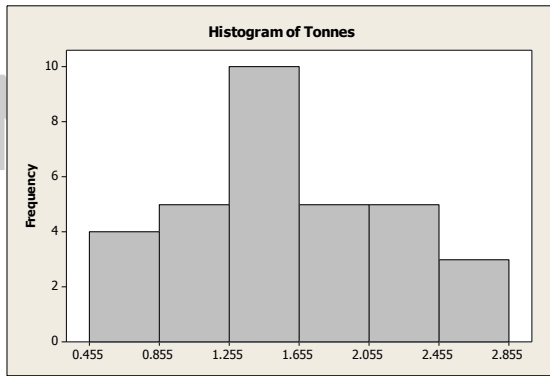
Class Limits	Class Boundaries	Midpoint	Frequency
0.46–0.85	0.455–0.855	0. 655	4
0.86–1.25	0.855–1.255	1.055	5
1.26–1.65	1.255–1.655	1.455	10
1.66–2.05	1.655–2.055	1.855	5
2.06–2.45	2.055–2.455	2.255	5
2.46–2.85	2.455–2.855	2.655	3

22. (a) Multiply each value by 1000.

(b)

Class Limits	Class Boundaries	Midpoint	Frequency
107–149	106.5–149.5	128	3
150–192	149.5–192.5	171	4
193–235	192.5–235.5	214	3
236–278	235.5–278.5	257	10
279–321	278.5–321.5	300	6

Histogram of Average

(c)
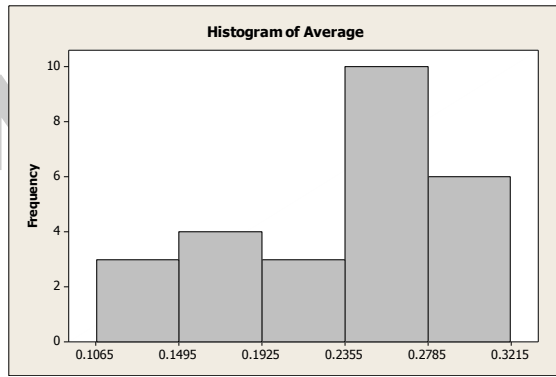
Class Limits	Class Boundaries	Midpoint	Frequency
0.107–0.149	0.1065–0.1495	0.128	3
0.150–0.192	0.1495–0.1925	0.171	4
0.193–0.235	0.1925–0.2355	0.214	3
0.236–0.278	0.2355–0.2785	0.257	10
0.279–0.321	0.2785–0.3215	0.300	6

23. (a) 1

(b) About $5/51 = 0.098 = 9.8\%$

(c) 650 to 750

24.

Dotplot of Finish Times
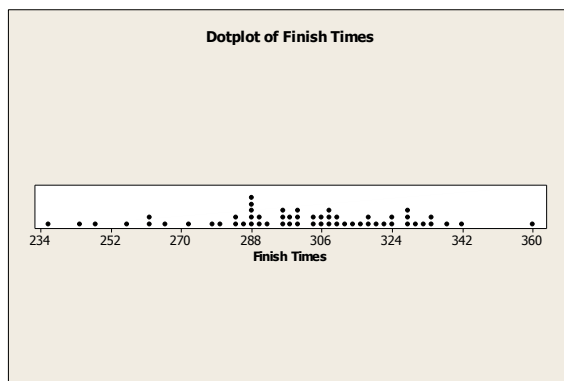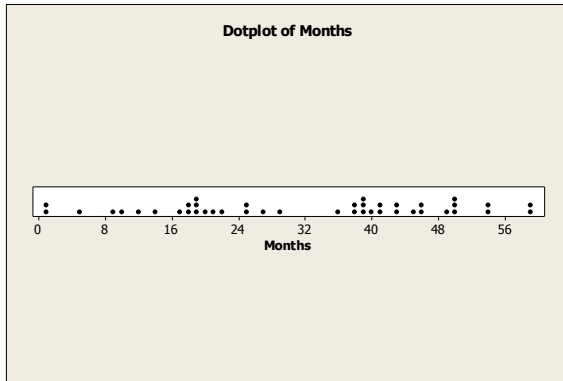
The dotplot shows some of the characteristics of the histogram, such as more dot density from 280 to 340, for instance, that corresponds roughly to the histogram bars of heights 25 and 16. However, the dotplot and histogram are somewhat difficult to compare because the dotplot can be thought of as a histogram with one value, the class mark (i.e., the data value), per class. Because the definitions of the classes (and therefore the class widths) differ, it is difficult to compare the two figures.

25.

Dotplot of Months

The dotplot shows some of the characteristics of the histogram, such as the concentration of most of the data in two peaks, one from 13 to 24 and another from 37 to 48. However, the dotplot and histogram are somewhat difficult to compare because the dotplot can be thought of as a histogram with one value, the class mark (i.e., the data value), per class. Because the definitions of the classes (and therefore the class widths) differ, it is difficult to compare the two figures.
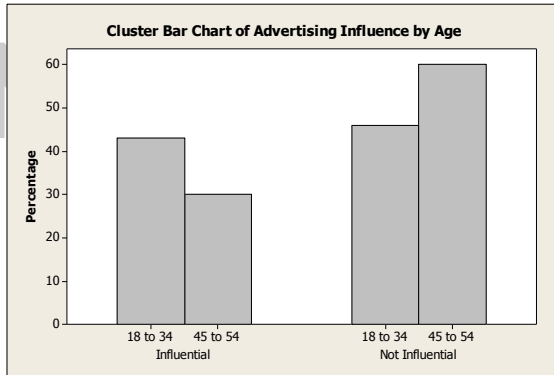
Section 2.2

1. **(a)** Yes, since the percentages total more than 100%.
    (b) No. In a circle graph, the percentages must total 100%.
    (c) Yes. The graph is organized from most frequently selected to least frequently selected.

2. This is not proper because the bars differ in both length and width.

3. A Pareto chart because it shows the five conditions in their order of importance to employees.

4. A time-series graph because the pattern of stock prices over time is more relevant than just the frequency of a specific range of closing prices.

5.

Bar Graph for Income vs Education
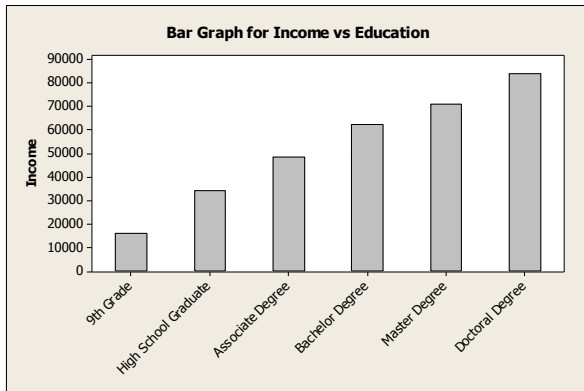
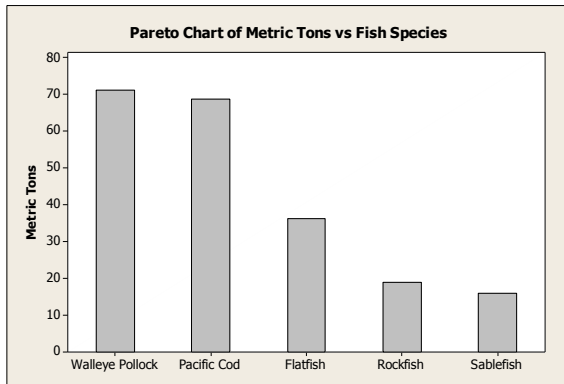**6.** **(a)** 45% of the 18 – 34 year olds and approximately 30% of the 45 – 54 year olds said "Influential". Perhaps the vertical scales should be labeled similarly.
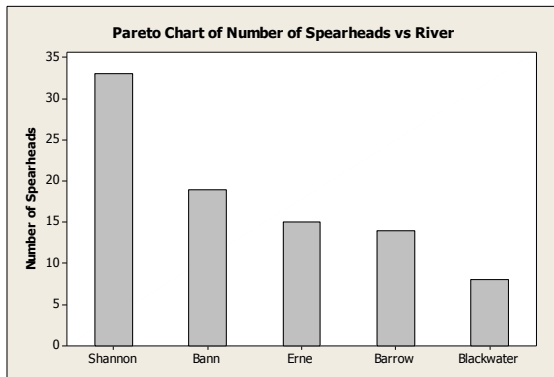
(b)

Cluster Bar Chart of Advertising Influence by Age

7.

Pareto Chart of Metric Tons vs Fish Species

8. **(a)**

Pareto Chart of Number of Spearheads vs River

(b)

Pie Chart of Number of Spearheads

9.

Pie Chart of Hiding Places

Under Bed 23.0%
Bathtub 6.0%
Freezer 3.0%
Closet 68.0%

10.

Pie Chart of Professor Activities

Consulting 6.0%
College Service 11.0%
Community Service 11.0%
Professional Growth 5.0%
Research 16.0%
Teaching 51.0%

11. (a)

Pareto Chart of Crime Rate vs. Crime Type

Crime Rate Per 100,000

House Burglary, Motor Vehicle Theft, Assault, Robbery, Rape, Murder

(b) Yes, but the graph would take into account only these particular crimes and would not indicate if multiple crimes occurred during the same incident.

12. (a)

Pareto Chart of Complaints

(b) Since the percentages do not add to 100%, a circle graph cannot be used. If we create an "other" category and assume that all other respondents fit this category, then a circle could be created.

13.
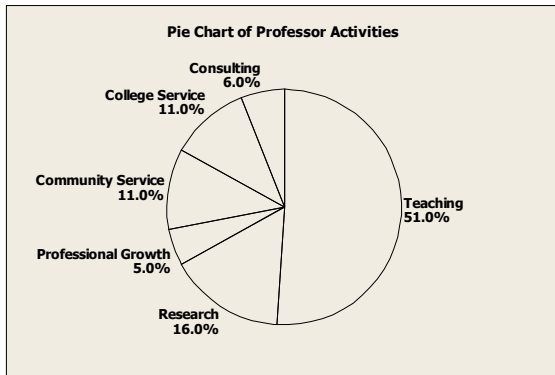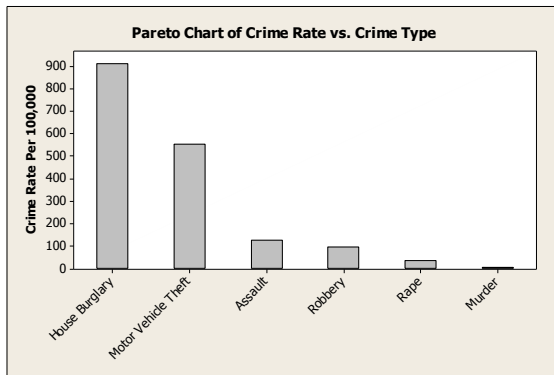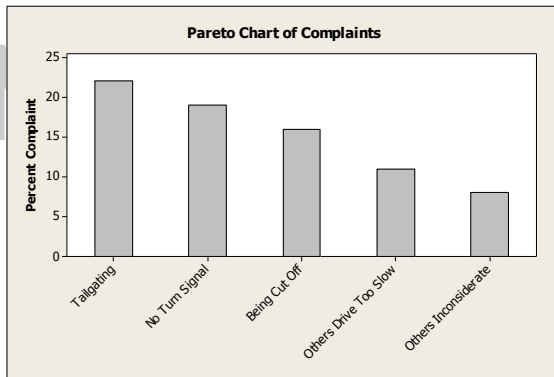

Time Series Plot of Elevation

14.

Time Series Plot of Height
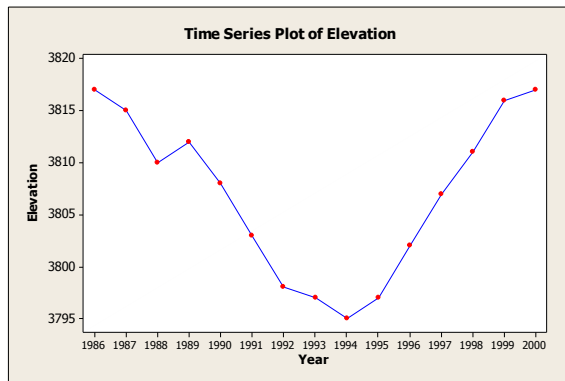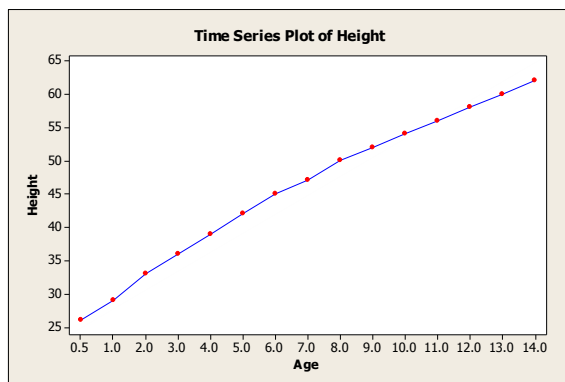
15. (a) The size of the donut hole is smaller on the rightmost graph. To correct the situation, simply make all donuts exactly the same size, with the radii of the respective holes the same size as well. Data labels showing the percentages for each response would also be useful.

(b) The college graduates answered "no" more frequently than those with a high school education or less.

16. (a) Worry that the technologies cause too much distraction and are dangerous.

(b) The Matures.

(c) Gen X.

Section 2.3

1. **(a)** The smallest value is 47 and the largest value is 97, so we need stems 4, 5, 6, 7, 8, and 9. Use the tens digit as the stem and the ones digit as the leaf.

Longevity of Cowboys

4	7 = 47 years
4	7
5	2 7 8 8
6	1 6 6 8 8
7	0 2 2 3 3 5 6 7
8	4 4 4 5 6 6 7 9
9	0 1 1 2 3 7

(b) Yes, these cowboys certainly lived long lives, as evidenced by the high frequency of leaves for stems 7, 8, and 9 (i.e., 70-, 80-, and 90-year-olds).

2. The largest value is 91 (percent of wetlands lost) and the smallest value is 9 (percent), which is coded as 09. We need stems 0 to 9. Use the tens digit as the stem and the ones digit as the leaf. The percentages are concentrated from 20% to 50%. These data are fairly symmetric, perhaps slightly skewed right. There is a gap showing that none of the lower 48 states has lost from 10% to 19% of its wetlands.

Percent of Wetlands Lost

4	0 = 40%
0	9
1	
2	0 3 4 7 7 8
3	0 1 3 5 5 5 6 7 8 8 9
4	2 2 6 6 6 8 9 9
5	0 0 0 2 2 4 6 6 9 9
6	0 7
7	2 3 4
8	1 5 7 7 9
9	0 1

3. The longest average length of stay is 11.1 days in North Dakota, and the shortest is 5.2 days in Utah. We need stems from 5 to 11. Use the digit(s) to the left of the decimal point as the stem and the digit to the right as the leaf.

Average Length of Hospital Stay

5	2 = 5.2 days
5	2 3 5 5 6 7
6	0 2 4 6 6 7 7 8 8 8 8 9 9
7	0 0 0 0 0 1 1 1 2 2 2 3 3 3 3 4 4 5 5 6 6 8
8	4 5 7
9	4 6 9
10	0 3
11	1

The distribution is skewed right.

4.

Number of Hospitals per State

0	8 = 8 hospitals		
0	8	15	
1	1 2 5 6 9	16	2
2	1 7 7	17	5
3	5 7 8	18	
4	1 2 7	19	3
5	1 2 3 9	20	9
6	1 6 8	21	
7	1	22	7
8	8	23	1 6
9	0 2 6 8		
10	1 2 7	42	1
11	3 3 7 9	43	
12	2 3 9	44	0
13	3 3 6		
14	8		

Texas and California have the highest number of hospitals, 421 and 440, respectively. Both states have large populations and large areas. The four largest states by area are Alaska, Texas, California, and Montana.

5. **(a)** The longest time during 1961−1980 is 23 minutes (i.e., 2:23), and the shortest time is 9 minutes (2:09). We need stems 0, 1, and 2. We'll use the tens digit as the stem and the ones digit as the leaf, placing leaves 0, 1, 2, 3, and 4 on the first stem and leaves 5, 6, 7, 8, and 9 on the second stem.

Minutes Beyond 2 Hours (1961−1980)

0	9 = 9 minutes past 2 hours
0	9 9
1	0 0 2 3 3
1	5 5 6 6 7 8 8 9
2	0 2 3 3

(b) The longest time during the period 1981−2000 was 14 (2:14) and the shortest was 7 (2:07), so we'll need stems 0 and 1 only.

Minutes Beyond 2 Hours (1981−2000)

	7 = 7 minutes past 2 hours
0	7 7 7 8 8 8 8 9 9 9 9 9 9 9 9
1	0 0 1 1 4

(c) There were seven times under 2:15 during 1961−1980, and there were 20 times under 2:15 during 1981−2000.

6. **(a)** The largest (worst) score in the first round was 75; the smallest (best) score was 65. We need stems 6 and 7. Leaves 0–4 go on the first stem, and leaves 5–9 belong on the second stem.

First-Round Scores

6	5 = score of 65
6	5 6 7 7
7	0 1 1 1 1 1 1 1 1 1 1 1 2 2 2 3 3 3 3 4 4 4
7	5 5 5 5 5 5 5

(b) The largest score in the fourth round was 74, and the smallest was 68. Here we need stems 6 and 7.

Fourth-Round Scores

6	8 = score of 68
6	8 9 9 9 9 9
7	0 0 0 0 1 1 1 1 1 1 1 1 2 2 2 2 2 2 3 3 3 3 3 4 4 4

(c) Scores are lower in the fourth round. In the first round, both the low and high scores were more extreme than in the fourth round.

7. The largest value in the data is 29.8 mg of tar per cigarette smoked, and the smallest value is 1.0. We will need stems from 1 to 29, and we will use the numbers to the right of the decimal point as the leaves.

Milligrams of Tar per Cigarette

1	0 = 1.0 mg tar
1	0
2	
3	
4	1 5
5	
6	
7	3 8
8	0 6 8
9	0
10	
11	4
12	0 4 8
13	7
14	1 5 9
15	0 1 2 8
16	0 6
17	0
.	
.	
.	
29	8

8. The largest value in the data set is 23.5 mg of carbon monoxide per cigarette smoked, and the smallest is 1.5. We need stems from 1 to 23, and we'll use the numbers to the right of the decimal point as leaves.

Milligrams of Carbon Monoxide

1	5 = 1.5 mg CO
1	5
2	
3	
4	9
5	4
6	
7	
8	5
9	0 5
10	0 2 2 6
11	
12	3 6
13	0 6 9
14	4 9
15	0 4 9
16	3 6
17	5
18	5
.	
.	
.	
23	5

9. The largest value in the data set is 2.03 mg of nicotine per cigarette smoked. The smallest value is 0.13. We will need stems 0, 1, and 2. We will use the number to the left of the decimal point as the stem and the first number to the right of the decimal point as the leaf. The number 2 placed to the right of the decimal point (the hundredths digit) will be truncated (not rounded).

Milligrams of Nicotine per Cigarette

0	1 = 0.1 milligram
0	1 4 4
0	5 6 6 6 7 7 7 8 8 9 9 9
1	0 0 0 0 0 0 0 1 2
1	
2	0

10. (a) For Site I, the least depth is 25 cm, and the greatest depth is 110 cm. For Site II, the least depth is 20 cm, and the greatest depth is 125 cm.
   (b) The Site I depth distribution is fairly symmetric, centered near 70 cm. Site II is fairly uniform in shape except that there is a huge gap with no artifacts from 70 to 100 cm.
   (c) It would appear that Site II probably was unoccupied during the time period associated with 70 to 100 cm.

Chapter Review Problems

1. (a) Bar graphs, Pareto charts, pie charts
   **(b)** All, but quantitative data must be categorized to use a bar graph, Pareto chart, or pie chart.

2. A time-series graph because a change over time is most relevant

3. Any large gaps between bars or stems might indicate potential outliers.

4. Dotplots and stem-and-leaf displays both show every data value. Stem-and-leaf plots group the data with the same stem, whereas dotplots only group the data with identical values.

5. **(a)** Figure 2-1(a) (in the text) is essentially a bar graph with a "horizontal" axis showing years and a "vertical" axis showing miles per gallon. However, in depicting the data as a highway and showing them in perspective, the ability to correctly compare bar heights visually has been lost. For example, determining what would appear to be the bar heights by measuring from the white line on the road to the edge of the road along a line drawn from the year to its mpg value, we get the bar height for 1983 to be approximately ⅞ inch and the bar height for 1985 to be approximately 1⅜ inches (i.e., 11/8 inches). Taking the ratio of the given bar heights, we see that the bar for 1985 should be $\frac{27.5}{26} \approx 1.06$ times the length of the 1983 bar. However, the measurements show a ratio of $\frac{\frac{11}{8}}{\frac{7}{8}} = \frac{11}{7} \approx 1.60$; i.e., the 1985 bar is (visually) 1.6 times the length of the 1983 bar. Also, the years are evenly spaced numerically, but the figure shows the more recent years to be more widely spaced owing to the use of perspective.

(b) Figure 2-1(b) is a time-series graph showing the years on the x axis and miles per gallon on the y axis. Everything is to scale and not distorted visually by the use of perspective. It is easy to see the mpg standards for each year, and you also can see how fuel economy standards for new cars have changed over the 8 years shown (i.e., a steep increase in the early years and a leveling off in the later years).

6. **(a)** We estimate the 1980 prison population at approximately 140 prisoners per 100,000 and the 1997 population at approximately 440 prisoners per 100,000 people.

(b) The number of inmates per 100,000 increased every year.

(c) The population 266,574,000 is 2,665.74 × 100,000, and 444 per 100,000 is $\frac{444}{100,000}$.

So $\frac{444}{100,000} \times (2,665.74 \times 100,000) \approx 1,183,589$ prisoners.

The projected 2020 population is 323,724,000, or 3,237.24 × 100,000.

So $\frac{444}{100,000} \times (3,237.24 \times 100,000) \approx 1,437,335$ prisoners.

7. Owing to rounding, the percentages are slightly different from those in the text.
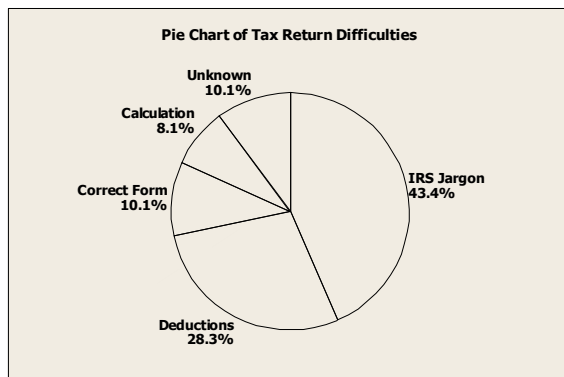
8. **(a)** Since the ages are two-digit numbers, use the ten's digit as the stem and the one's digit as the leaf.

Age of DUI Arrests

1	6 = 16 years
1	6 8
2	0 1 1 2 2 2 3 4 4 5 6 6 6 7 7 7 9
3	0 0 1 1 2 3 4 4 5 5 6 7 8 9
4	0 0 1 3 5 6 7 7 9 9
5	1 3 5 6 8
6	3 4

(b) The largest age is 64 and the smallest is 16, so the class width for seven classes is $\frac{64-16}{7} \approx 6.86$; use 7.
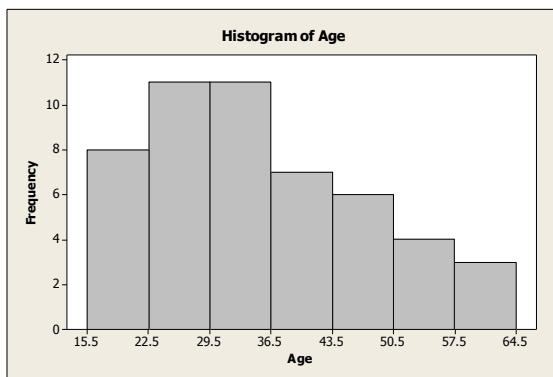
The lower class limit for the first class is 16; the lower class limit for the second class is $16 + 7 = 23$. The total number of data points is 50, so calculate the relative frequency by dividing the class frequency by 50.

Age Distribution of DUI Arrests

Class Limits	Class Boundaries	Midpoint	Frequency	Relative Frequency	Cumulative Frequency
16–22	15.5–22.5	19	8	0.16	8
23–29	22.5–29.5	26	11	0.22	19
30–36	29.5–36.5	33	11	0.22	30
37–43	36.5–43.5	40	7	0.14	37
44–50	43.5–50.5	47	6	0.12	43
51–57	50.5–57.5	54	4	0.08	47
58–64	57.5–64.5	61	3	0.06	50

The class boundaries are the average of the upper class limit of the next class. The midpoint is the average of the class limits for that class.

(c)

(d) This distribution is skewed right.

9. **(a)** The largest value is 142 mm, and the smallest value is 69. For seven classes, we need a class width of $\frac{142-69}{7} \approx 10.4$; use 11. The lower class limit of the first class is 69, and the lower class limit of the second class is $69 + 11 = 80$.

The class boundaries are the average of the upper class limit of one class and the lower class limit of the next higher class. The midpoint is the average of the class limits for that class. There are 60 data values total, so the relative frequency is the class frequency divided by 60.

| Class Limits	Class Boundaries	Midpoint	Frequency	Relative Frequency	Cumulative Frequency
69–79	68.5–79.5	74	2	0.03	2
80–90	79.5–90.5	85	3	0.05	5
91–101	90.5–101.5	96	8	0.13	13
102–112	101.5–112.5	107	19	0.32	32
113–123	112.5–123.5	118	22	0.37	54
124–134	123.5–134.5	129	3	0.05	57
135–145	134.5–145.5	140	3	0.05	60

(b)

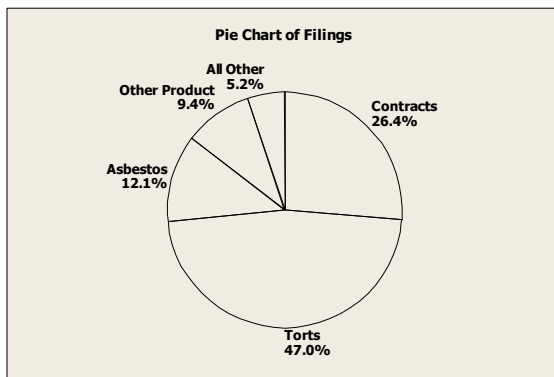**(c)**
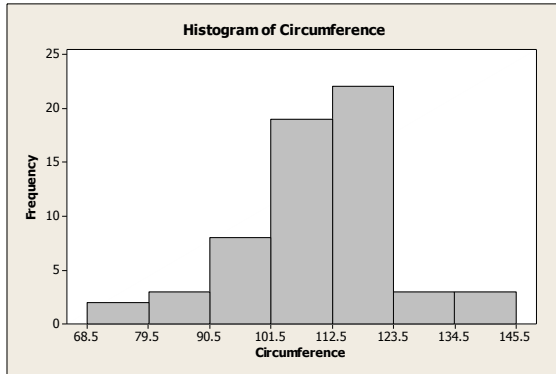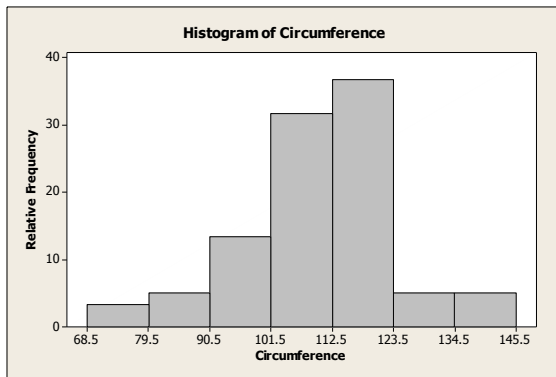
(d) This distribution is skewed left.
(e) The ogive begins on the *x* axis at the lower class boundary and connects dots placed at (*x*, *y*) coordinates (upper class boundary, cumulative frequency).

10. (a) General torts occur most frequently.

Pareto Chart of Filings vs Type

(b)

Pie Chart of Filings
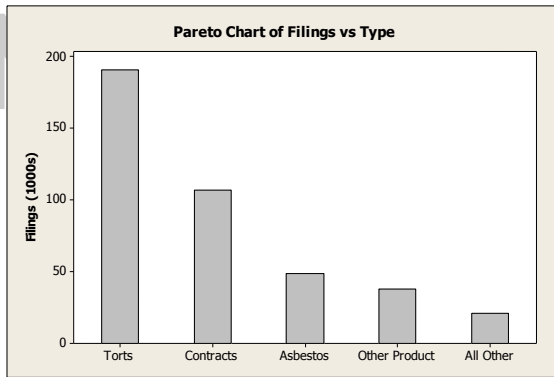
11. (a) To determine the decade that contained the most samples, count *both* rows (if shown) of leaves; recall that leaves 0–4 belong on the first line and leaves 5–9 belong on the second line when two lines per stem are used. The greatest number of leaves is found on stem 124, i.e., the 1240s (the 40s decade in the 1200s), with 40 samples.

(b) The number of samples with tree-ring dates 1200 to 1239 A.D. is $28 + 3 + 19 + 25 = 75$.

(c) The dates of the longest interval with no sample values are 1204 through 1211 A.D. This might mean that for these 8 years, the pueblo was unoccupied (thus no new or repaired structures), or that the population remained stable (no new structures needed), or that, say, weather conditions were favorable those years, so existing structures didn't need repair. If relatively few new structures were built or repaired during this period, their tree rings might have been missed during sample selection.

12. (a) It appears that the percentage of adults who consider using a cell phone while driving to be dangerous increases as age increases.

(b) In each age category except the oldest group, a smaller percentage of people say they never use a cell phone while driving than the percentage of people who say cell use while driving is very dangerous. If people do not use a cell phone or do not drive, they may perceive that using a cell phone while driving is dangerous than people who do use cell phones while driving. The differing survey populations make it difficult to draw conclusions about inconsistencies in behavior and opinion.