

Chapter 3 – Displaying and Summarizing Quantitative Data

1. **Histogram.** Answers will vary.
2. **Not a histogram.** Answers will vary.
3. **In the news.** Answers will vary.
4. **In the news II.** Answers will vary.
5. **Thinking about shape.**
 - a) The distribution of the number of speeding tickets each student in the senior class of a college has ever had is likely to be unimodal and skewed to the right. Most students will have very few speeding tickets (maybe 0 or 1), but a small percentage of students will likely have comparatively many (3 or more?) tickets.
 - b) The distribution of player's scores at the U.S. Open Golf Tournament would most likely be unimodal and slightly skewed to the right. The best golf players in the game will likely have around the same average score, but some golfers might be off their game and score 15 strokes above the mean. (Remember that high scores are undesirable in the game of golf!)
 - c) The weights of female babies in a particular hospital over the course of a year will likely have a distribution that is unimodal and symmetric. Most newborns have about the same weight, with some babies weighing more and less than this average. There may be slight skew to the left, since there seems to be a greater likelihood of premature birth (and low birth weight) than post-term birth (and high birth weight).
 - d) The distribution of the length of the average hair on the heads of students in a large class would likely be bimodal and skewed to the right. The average hair length of the males would be at one mode, and the average hair length of the females would be at the other mode, since women typically have longer hair than men. The distribution would be skewed to the right, since it is not possible to have hair length less than zero, but it is possible to have a variety of lengths of longer hair.
6. **More shapes.**
 - a) The distribution of the ages of people at a Little League game would likely be bimodal and skewed to the right. The average age of the players would be at one mode and the average age of the spectators (probably mostly parents) would be at the other mode. The distribution would be skewed to the right, since it is possible to have a greater variety of ages among the older people, while there is a natural left endpoint to the distribution at zero years of age.

- b) The distribution of the number of siblings of people in your class is likely to be unimodal and skewed to the right. Most people would have 0, 1, or 2 siblings, with some people having more siblings.
- c) The distribution of pulse rate of college-age males would likely be unimodal and symmetric. Most males' pulse rates would be around the average pulse rate for college-age males, with some males having lower and higher pulse rates.
- d) The distribution of the number of times each face of a die shows in 100 tosses would likely be uniform, with around 16 or 17 occurrences of each face (assuming the die had six sides).

7. Sugar in cereals.

- a) The distribution of the sugar content of breakfast cereals is bimodal, with a cluster of cereals with sugar content around 10% sugar and another cluster of cereals around 48% sugar. The lower cluster shows a bit of skew to the right. Most cereals in the lower cluster have between 0% and 10% sugar. The upper cluster is symmetric, with center around 45% sugar.
- b) There are two different types of breakfast cereals, those for children and those for adults. The children's cereals are likely to have higher sugar contents, to make them taste better (to kids, anyway!). Adult cereals often advertise low sugar content.

8. Singers.

- a) The distribution of the heights of singers in the chorus is bimodal, with a mode at around 65 inches and another mode around 71 inches. No chorus member has height below 60 inches or above 76 inches.
- b) The two modes probably represent the mean heights of the male and female members of the chorus.

9. Vineyards.

- a) There is information displayed about 36 vineyards and it appears that 28 of the vineyards are smaller than 60 acres. That's around 78% of the vineyards. (75% would be a good estimate!)
- b) The distribution of the size of 36 Finger Lakes vineyards is skewed to the right. Most vineyards are smaller than 75 acres, with a few larger ones, from 90 to 160 acres. One vineyard was larger than all the rest, over 240 acres. The mode of the distribution is between 0 and 30 acres.

10. Run times.

The distribution of runtimes is skewed to the right. The shortest runtime was around 28.5 minutes and the longest runtime was around 35.5 minutes. A typical run time was between 30 and 31 minutes, and the majority of runtimes were between 29 and 32 minutes. It is easier to run slightly slower than usual

24 *Part I Exploring and Understanding Data*

and end up with a longer runtime than it is to run slightly faster than usual and end up with a shorter runtime. This could account for the skew to the right seen in the distribution.

11. Heart attack stays.

- a) The distribution of length of stays is skewed to the right, so the mean is larger than the median.
- b) The distribution of the length of hospital stays of female heart attack patients is skewed to the right, with stays ranging from 1 day to 36 days. The distribution is centered around 8 days, with the majority of the hospital stays lasting between 1 and 15 days. There are a relatively few hospital stays longer than 27 days. Many patients have a stay of only one day, possibly because the patient died.
- c) The median and IQR would be used to summarize the distribution of hospital stays, since the distribution is strongly skewed.

12. Emails.

- a) The distribution of the number of emails sent is skewed to the right, so the mean is larger than the median.
- b) The distribution of the number of emails received from each student by a professor in a large introductory statistics class during an entire term is skewed to the right, with the number of emails ranging from 1 to 21 emails. The distribution is centered at about 2 emails, with many students only sending 1 email. There is one outlier in the distribution, a student who sent 21 emails. The next highest number of emails sent was only 8.
- c) The median and IQR would be used to summarize the distribution of the number of emails received, since the distribution is strongly skewed.

13. Super Bowl points.

- a) The median number of points scored in the first 46 Super Bowl games is 45.5 points.
- b) The first quartile of the number of points scored in the first 46 Super Bowl games is 37 points. The third quartile is 55 points.
- c) $IQR = Q3 - Q1 = 55 - 37 = 18$.

Using the Outlier Rule (1.5 IQRs beyond quartiles):

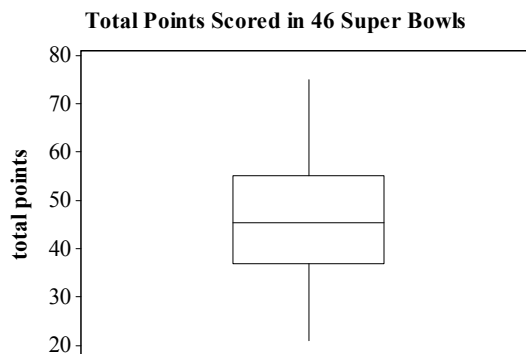
$$\begin{aligned}\text{Upper Fence: } Q3 + 1.5(IQR) &= 55 + 1.5(18) \\ &= 55 + 27 \\ &= 82\end{aligned}$$

The highest number of points scored in a Super Bowl is 75, which is lower than the upper fence of 82. There are no high outliers.

$$\begin{aligned} \text{Lower Fence: } Q1 - 1.5(\text{IQR}) &= 37 - 1.5(18) \\ &= 37 - 27 \\ &= 10 \end{aligned}$$

The lowest number of points scored in a Super Bowl is 21, which is higher than the lower fence of 10. There are no low outliers.

- d) A boxplot of the number of points scored in the first 46 Super Bowl games is at the right.
- e) In the first 46 Super Bowl games, the lowest number of points scored was 21, and the highest number of points scored was 75. The median number of points scored was 45.5, and the middle 50% of Super Bowls has between 37 and 55 points scored.



14. Super Bowl wins.

- a) The median winning margin in the first 46 Super Bowl games is 12.5 points.
- b) The first quartile of the winning margin in the first 46 Super Bowl games is 5 points. The third quartile is 19 points.
- c) $\text{IQR} = Q3 - Q1 = 19 - 5 = 14$.

Using the Outlier Rule (1.5 IQRs beyond quartiles):

$$\begin{aligned} \text{Upper Fence: } Q3 + 1.5(\text{IQR}) &= 19 + 1.5(14) \\ &= 19 + 21 \\ &= 40 \end{aligned}$$

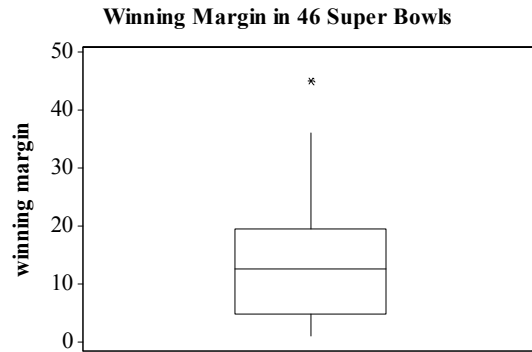
The highest winning margin in a Super Bowl is 45 points, which is higher than the upper fence of 40. This winning margin is an outlier.

$$\begin{aligned} \text{Lower Fence: } Q1 - 1.5(\text{IQR}) &= 5 - 1.5(16) \\ &= 5 - 21 \\ &= -16 \end{aligned}$$

The lowest winning margin in a Super Bowl is 1, which is higher than the lower fence of -16. There are no low outliers.

26 Part I Exploring and Understanding Data

- d) A boxplot of the number of points scored in the first 46 Super Bowl games is at the right.
- e) In the first 46 Super Bowl games the lowest winning margin was 1 point and the highest winning margin was 45 points, which was an outlier. The second highest winning margin was only 36 points. The median winning margin was 12.5 points, with the middle 50% of winning margins between 5 and 19 points.



15 Details.

- a) The boxplots do not show clusters and gaps, nor locations of multiple modes.
- b) Boxplots can give only general ideas about overall shape, and should not be used when more detail is needed.

16. Opposites.

- a) The tallest bars are in the narrow regions of the boxplot.
- b) The histogram shows taller bars clustered on the left and smaller bars more spread out to the right. The boxplot shows narrower regions on the left, and a longer right side of the box and a long whisker stretching to the right, as well as some potential outliers on the right.
- c) The histogram shows more precisely the locations of peaks and gaps.
- d) The boxplot shows the locations of points flagged as outliers by the $1.5 \times \text{IQR}$ outlier rule.

17. Still rockin'.

- a) The histogram and boxplot of the distribution of "crowd crush" victims' ages both show that a typical crowd crush victim was approximately 18 - 20 years of age, that the range of ages is 36 years, that there are two outliers, one victim at age 36 - 38 and another victim at age 46 - 48.
- b) This histogram shows that there may have been two modes in the distribution of ages of "crowd crush" victims, one at 18 - 20 years of age and another at 22 - 24 years of age. Boxplots, in general, can show symmetry and skewness, but not features of shape like bimodality or uniformity.
- c) Median is the better measure of center, since the distribution of ages has outliers. Median is more resistant to outliers than the mean.
- d) IQR is a better measure of spread, since the distribution of ages has outliers. IQR is more resistant to outliers than the standard deviation.

18. Slalom times.

- a) The histogram and boxplot of the distribution of slalom times both show that a typical slalom time of around 92 seconds, that the range of slalom times is a little more than 20 seconds. Both distributions also show that the distribution of slalom times is skewed to the high end.
- b) The histogram shows that the distribution of slalom times is bimodal. In addition to a mode around 92 seconds, there was a small group of skiers with scores around 105 seconds. Boxplots, in general, can show symmetry and skewness, but not features of shape like bimodality or uniformity. However, the boxplot shows two possible outlying times that the histogram doesn't highlight.
- c) Since the distribution of slalom times is skewed, and contains possible outliers, the median is the better summary of center.
- d) In the presence of skewness and possible outliers, we'd prefer the IQR to the standard deviation as a measure of spread.

19. Camp sites.

- a) The distribution of the number of campsites in public parks in Vermont is skewed to the right, so median and IQR are appropriate measures of center and spread.

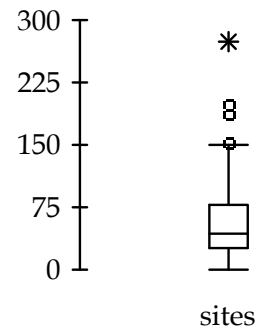
b) $IQR = Q3 - Q1 = 78 - 28 = 50$.

Using the Outlier Rule (1.5 IQRs beyond quartiles):

$$\begin{aligned} \text{Upper Fence: } Q3 + 1.5(IQR) &= 78 + 1.5(50) \\ &= 78 + 75 \\ &= 153 \end{aligned}$$

Lower Fence: Well below 0 campsites.

There are 3 parks with greater than 180 campsites. These are definitely outliers. There are 2 parks with between 150 and 160 campsites each. These may be outliers as well.



- c) A boxplot of the distribution of number of campsites is at the right.
- d) The distribution of the number of campsites at public parks in Vermont is unimodal and skewed to the right. The center of the distribution is approximately 44 campsites. The distribution of campsites is quite spread out, with several high outliers. These parks have in excess of 150 campsites each.

28 **Part I Exploring and Understanding Data**

20. **Outliers.**

a) $IQR = Q3 - Q1 = 116 - 98 = 18$.

Using the Outlier Rule (1.5 IQRs beyond quartiles):

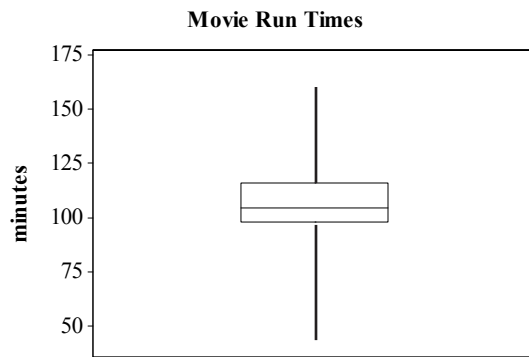
$$\begin{aligned} \text{Upper Fence: } Q3 + 1.5(IQR) &= 116 + 1.5(18) \\ &= 116 + 27 \\ &= 143 \end{aligned}$$

The longest run time is 160 minutes, which is higher than the upper fence of 143. This run time is an outlier. There may be others.

$$\begin{aligned} \text{Lower Fence: } Q1 - 1.5(IQR) &= 98 - 1.5(18) \\ &= 98 - 27 \\ &= 71 \end{aligned}$$

The shortest run time is 43 minutes, which is lower than the lower fence of 71. This run time is an outlier. There may be others.

b) We do not have enough information to know if there are other outliers, or where the last non-outlier point is on either side, so the boxplot can be drawn without outliers as shown. We can't say much about the shape of the distribution, since many different shapes could underlie the boxplot. The most we could say is that the distribution of movie run times is skewed slightly to the right.



21. **Standard deviation I.**

a) Set 2 has the greater standard deviation. Both sets have the same mean, 6, but set two has values that are generally farther away from the mean.

$$SD(\text{Set 1}) = 2.24 \quad SD(\text{Set 2}) = 3.16$$

b) Set 2 has the greater standard deviation. Both sets have the same mean (15), maximum (20), and minimum (10), but 11 and 19 are farther from the mean than 14 and 16.

$$SD(\text{Set 1}) = 3.61 \quad SD(\text{Set 2}) = 4.53$$

c) The standard deviations are the same. Set 2 is simply Set 1 + 80. Although the measures of center and position change, the spread is exactly the same.

$$SD(\text{Set 1}) = 4.24 \quad SD(\text{Set 2}) = 4.24$$

22. Standard deviation II.

- a) Set 2 has the greater standard deviation. Both sets have the same mean (7), maximum (10), and minimum (4), but 6 and 8 are farther from the mean than 7.
 $SD(\text{Set 1}) = 2.12$ $SD(\text{Set 2}) = 2.24$
- b) The standard deviations are the same. Set 1 is simply Set 2 + 90. Although the measures of center and position are different, the spread is exactly the same.
 $SD(\text{Set 1}) = 36.06$ $SD(\text{Set 2}) = 36.06$
- c) Set 2 has the greater standard deviation. The central 4 values of Set 2 are simply the central 4 values of Set 1 +40, but the maximum and minimum of Set 2 are farther away from the mean than the maximum and minimum of Set 1.
 $\text{Range}(\text{Set 1}) = 18$ and $\text{Range}(\text{Set 2}) = 22$. Since the Range of Set 2 is greater than the Range of Set 1, the standard deviation is also larger.
 $SD(\text{Set 1}) = 6.03$ $SD(\text{Set 2}) = 7.04$

23. Pizza prices.

The mean and standard deviation would be used to summarize the distribution of pizza prices, since the distribution is unimodal and symmetric.

24. Neck size.

The mean and standard deviation would be used to summarize the distribution of neck sizes, since the distribution is unimodal and symmetric.

25. Pizza prices again.

- a) The mean pizza price is closest to \$2.60. That's the balancing point of the histogram.
- b) The standard deviation in pizza prices is closest to \$0.15, since that is the typical distance to the mean. There are no pizza prices as far as \$0.50 or \$1.00.

26. Neck sizes again.

- a) The mean neck size is closest to 15 inches. That's the balancing point of the histogram.
- b) The standard deviation in neck sizes is closest to 1 inch, because a typical value lies about 1 inch from the mean. There are a few points as far away as 3 inches from the mean, and none as far away as 5 inches. Those are too large to be the standard deviation.

27. Movie lengths.

- a) A typical movie would be around 105 minutes long. This is near the center of the unimodal and roughly symmetric histogram.
- b) You would be surprised to find that your movie ran for 150 minutes. Only 2 movies ran that long.

30 *Part I Exploring and Understanding Data*

- c) The mean run time would probably be roughly the same as the median run time, since the distribution is reasonably symmetric.

28. Golf drives.

- a) The distribution of golf drives is roughly unimodal and symmetric, with a typical drive of around 290 yards. Professional golfers on the men's PGA tour had drives that were as short as about 270 yards, and as long as about 320 yards.
- b) Approximately 10% of professional male golfers drive less than 280 yards.
- c) The actual mean drive is about 291.3 yards, so any estimate between 285 and 295 yards is reasonable.
- d) The distribution of golf drives is approximately symmetric, so the mean and the median should be relatively close. The actual median is 291.3 yards.

29. Movie lengths II.

- a)
 - i) The distribution of movie running times is fairly consistent, with the middle 50% of running times between 97 and 115 minutes. The interquartile range is 18 minutes.
 - ii) The standard deviation of the distribution of movie running times is 17.3 minutes, which indicates that movies typically have running times fairly close to the mean running time. On average, movie running times differed from the mean running time by only 17.3 minutes.
- b) Since the distribution of movie running times is reasonably symmetric, either measure of spread would be appropriate.

30. Golf drives II.

- a)
 - i) The distribution of PGA golf drives is fairly consistent, with the middle 50% of the drives having distances between 285.2 and 297.5 yards. The interquartile range is 12.3 yards.
 - ii) The standard deviation of the distribution of PGA golf drives is 8.4 yards, which indicates that golf drives are typically within 8.4 yards of the mean gold drive.
- b) Since the distribution of golf drives is reasonably symmetric, both the standard deviation and the interquartile range are reasonable measures of spread.

31. Fuel economy.

The standard deviation is highly sensitive to outliers. If the electric car's fuel economy is removed, the standard deviation would be much lower. The actual standard deviation would be 4.6 miles per gallon. The IQR would be affected very little, if at all.

32. Test scores correction.

- a) The median score will not be affected at all. The top score is still the top score. The mean will increase by $10 \div 25 = 0.4$ points.
- b) The IQR will not be affected at all, since the quartiles remain the same. The standard deviation will increase, since increasing the top score makes the distribution of test scores more spread out.

33. Mistake.

- a) As long as the boss's true salary of \$200,000 is still above the median, the median will be correct. The mean will be too large, since the total of all the salaries will decrease by $\$2,000,000 - \$200,000 = \$1,800,000$, once the mistake is corrected.
- b) The range will likely be too large. The boss's salary is probably the maximum, and a lower maximum would lead to a smaller range. The IQR will likely be unaffected, since the new maximum has no effect on the quartiles. The standard deviation will be too large, because the \$2,000,000 salary will have a large squared deviation from the mean.

34. Cold weather.

- a) The mean temperature will be lower. The median temperature will not change, since the incorrect temperature is still the lowest temperature, and the median is based only on position.
- b) The range and standard deviation in temperature will both increase, since the incorrect temperature is more extreme than the correct temperature. The IQR will not change, since the both the correct and incorrect scores are below the first quartile, and the IQR measures the distance between the first and third quartiles.

35. Movie earnings.

The industry publication is using the median, while the watchdog group is using the mean. The mean is pulled higher by a few very expensive movies.

36. Sick days.

The company probably uses the mean, while the union uses the median number of sick days. The mean will likely be higher, since it is affected by probable right skew. Some employees may have many sick days, while most have relatively few.

37. Payroll.

- a) The mean salary is $\frac{(1200 + 700 + 6(400) + 4(500))}{12} = \525 .

The median salary is the middle of the ordered list:

400 400 400 400 400 400 500 500 500 500 700 1200

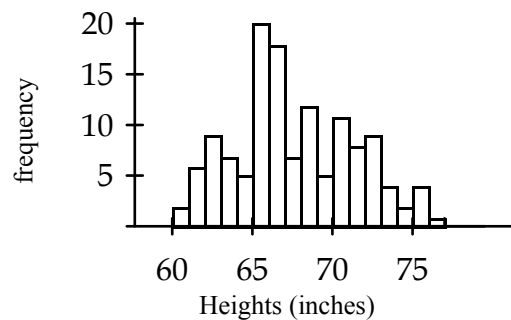
The median is \$450.

32 Part I Exploring and Understanding Data

- b) Only two employees, the supervisor and the inventory manager, earn more than the mean wage.
- c) The median better describes the wage of the typical worker. The mean is affected by the two higher salaries.
- d) The IQR is the better measure of spread for the payroll distribution. The standard deviation and the range are both affected by the two higher salaries.

38. Singers.

- a) 5-number summary: 60, 65, 66, 70, 76, so the median is 66 inches and the IQR is $70 - 65 = 5$ inches.
- b) The mean height of the singers is 67.12 inches, and the standard deviation of the heights is 3.79 inches.
- c) The histogram of heights of the choir members is at the right.
- d) The distribution of the heights of the choir members is bimodal (probably due to differences in height of men and women) and skewed slightly to the right. The median is 66 inches. The distribution is fairly spread out, with the middle 50% of the heights falling between 65 and 70 inches. There are no gaps or outliers in the distribution.

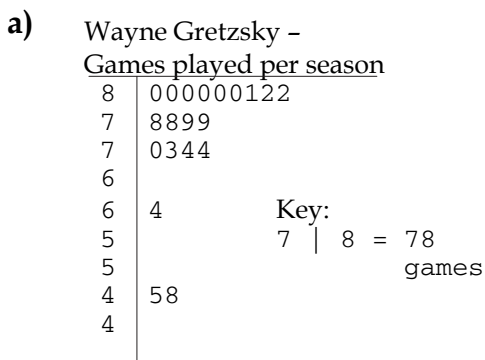


39. Gasoline 2011.

- a) A stem-and-leaf plot of the gasoline prices is at the right.
- b) The distribution of gas prices is bimodal, with two clusters, one centered around \$3.45 per gallon, and another centered around \$3.65 per gallon. The lowest and highest prices were \$3.31 and \$3.66 per gallon.
- c) There is a gap in the distribution of gasoline prices. There were no stations that charged between \$3.49 and \$3.58.

Gasoline Prices	
3.6	556
3.6	23
3.5	9
3.5	
3.4	6678
3.4	1233
3.3	5
3.3	1
Key : 3.3 1 = \$3.31/gal	

40. The Great One.



- b) The distribution of the number of games played by Wayne Gretzky is skewed to the left.
- c) Typically, Wayne Gretzky played about 80 games per season. The number of games played is tightly clustered in the upper 70s and low 80s.
- d) Two seasons are low outliers, when Gretzky played fewer than 50 games. He may have been injured during those seasons. Regardless of any possible reasons, these seasons were unusual compared to Gretzky's other seasons.

41. States.

- a) The distribution of state populations is skewed heavily to the right. Therefore, the median and IQR are the appropriate measures of center and spread.
- b) The mean population must be larger than the median population. The extreme values on the right affect the mean greatly and have no effect on the median.
- c) There are 51 entries in the stemplot, so the 26th entry must be the median. Counting in the ordered stemplot gives median = 4 million people. The middle of the lower 50% of the list (26 state populations) is between the 13th and 14th population, or 1.5 million people. The middle of the upper half of the list (26 state populations) is between the 13th and 14th population from the top, or 6 million people. The IQR = Q3 - Q1 = 6 - 1.5 = 4.5 million people.
- d) The distribution of population for the 50 U.S. States and Washington, D.C. is skewed heavily to the right. The median population is 4 million people, with 50% of states having populations between 1 and 6 million people. There is one outlier, a state with 34 million people. The next highest population is only 21 million.

34 Part I Exploring and Understanding Data

42. Wayne Gretzky.

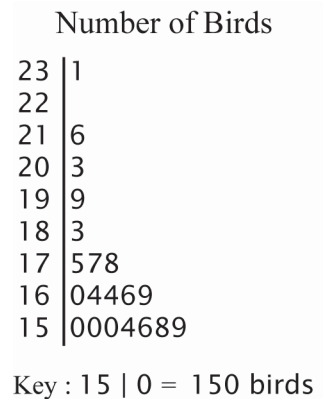
- a) The distribution of the number of games played per season by Wayne Gretzky is skewed to the left, and has low outliers. The median is more resistant to the skewness and outliers than the mean.
- b) The median, or middle of the ordered list, is 79 games. Both the 10th and 11th values are 79, so the median is the average of these two, also 79.
- c) The mean should be lower. There are two seasons when Gretzky played an unusually low number of games. Those seasons will pull the mean down.

43. A-Rod 2010.

The distribution of the number of homeruns hit by Alex Rodriguez during the 1994 - 2010 seasons is slightly skewed to the right, with a typical number of homeruns per season in the high 30s to low 40s, with two low outliers. With the exception of 2 seasons in which A-Rod hit fewer than 0 and 5 homeruns, his total number of homeruns per season was between 23 and the maximum of 57. The median is 36 homeruns.

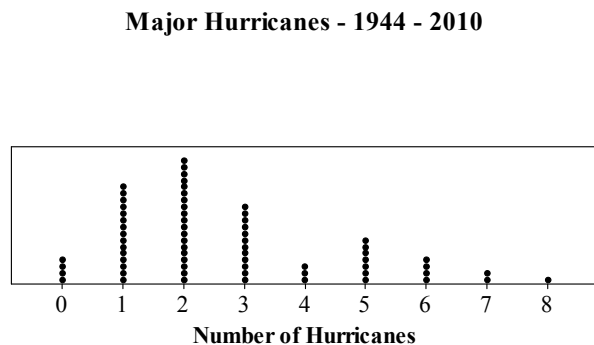
44. Bird species 2010.

- a) The results of the 2010 Laboratory of Ornithology Christmas Bird Count are displayed in the stem and leaf display at the right.
- b) The distribution of the number of birds spotted by participants in the 2010 Laboratory of Ornithology Christmas Bird Count is skewed right, with a median of 165 birds. There are two high outliers, with one participant spotting 216 birds and another spotting 231. With the exception of these outliers, most participants saw between 150 and 203 birds.



46. Hurricanes 2010.

- a) A dotplot of the number of hurricanes each year from 1944 through 2010 is displayed. Each dot represents a year in which there were that many hurricanes.



- b) The distribution of the number of hurricanes per year is unimodal and skewed to the right, with center around 2 hurricanes per year. The number of hurricanes per year ranges from 0 to 8. There are no outliers. There may be a second mode at 5 hurricanes per year, but since there were only 6 years in which 5 hurricanes occurred, this may simply be natural variability.

46. Horsepower.

The distribution of horsepower of cars reviewed by *Consumer Reports* is nearly uniform. The lowest horsepower was 65 and the highest was 155. The center of the distribution was around 105 horsepower.

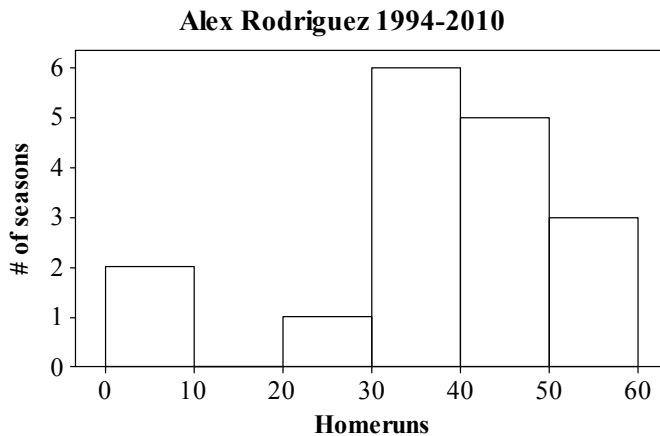
Horsepower

15		05
14		2
13		0358
12		0559
11		00555
10		359
9		00577
8		058
7		01158
6		55889

Key : 15 | 0 = 150 hp

47. A-Rod again 2010.

- a) This is not a histogram. The horizontal axis should the number of home runs per year, split into bins of a convenient width. The vertical axis should show the frequency; that is, the number of years in which A-Rod hit a number of home runs within the interval of each bin. The display shown is a bar chart/time plot hybrid that simply displays the data table visually. It is of no use in describing the shape, center, spread, or unusual features of the distribution of home runs hit per year by A-Rod.



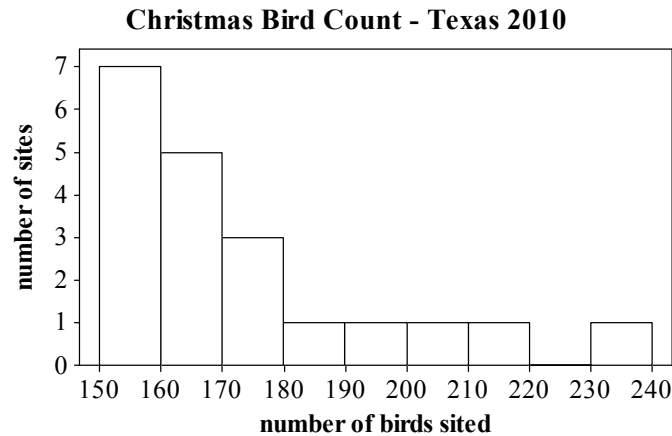
- b) The histogram is at the right.

48. Return of the birds.

- a) This is not a histogram. The horizontal axis should split the number of counts from each site into bins. The vertical axis should show the number of sites in each bin. The given graph is nothing more than a bar chart, showing the bird count from each site as its own bar. It is of absolutely no use for describing the shape, center, spread, or unusual features of the distribution of bird counts.

36 **Part I Exploring and Understanding Data**

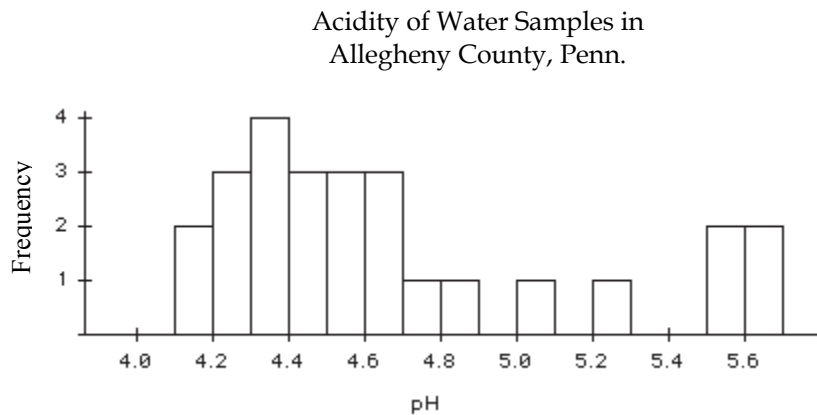
b) The histogram is below.



49. Acid rain.

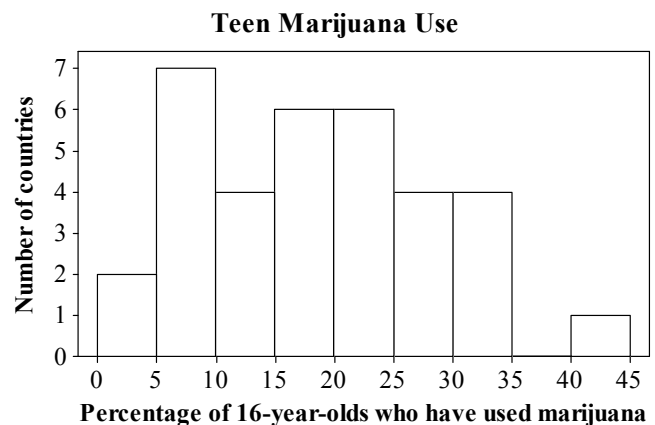
The distribution of the pH readings of water samples in Allegheny County, Penn. is bimodal. A roughly uniform cluster is centered around a pH of 4.4. This cluster ranges from pH of 4.1 to 4.9. Another

smaller, tightly packed cluster is centered around a pH of 5.6. Two readings in the middle seem to belong to neither cluster.



50. Marijuana 2007.

The distribution of the percentage of 16-year-olds in 34 countries who have used marijuana is somewhat bimodal, with 9 countries having between 3% and 10% of 16-year-olds having used marijuana. Another group of 12 countries has between 15% and 25% of teens who have used marijuana. Armenia, at 3%, had the lowest percentage of 16-year-olds who have tried marijuana. Czech Republic had the highest percentage, at 45%. A typical country might have a percentage of approximately 20%.



51. Final grades.

The width of the bars is much too wide to be of much use. The distribution of grades is skewed to the left, but not much more information can be gathered.

52. Final grades revisited.

- a) This display has a bar width that is much too narrow. As it is, the histogram is only slightly more useful than a list of scores. It does little to summarize the distribution of final exam scores.
- b) The distribution of test scores is skewed to the left, with center at approximately 170 points. There are several low outliers below 100 points, but other than that, the distribution of scores is fairly tightly clustered.

53. Zip codes.

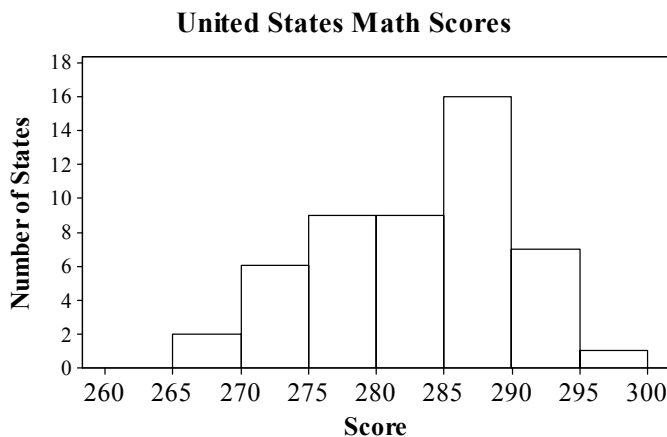
Even though zip codes are numbers, they are not quantitative in nature. Zip codes are categories. A histogram is not an appropriate display for categorical data. The histogram the Holes R Us staff member displayed doesn't take into account that some 5-digit numbers do not correspond to zip codes or that zip codes falling into the same classes may not even represent similar cities or towns. The employee could design a better display by constructing a bar chart that groups together zip codes representing areas with similar demographics and geographic locations.

54. Zip codes revisited

The statistics cannot tell us very much since zip codes are categorical. However, there is *some* information in the first digit of zip codes. They indicate a general East (0-1) to West (8-9) direction. So, the distribution shows that a large portion of their sales occurs in the West and another in the 32000 area. But a bar chart of the first digits would be the appropriate display to show this information.

55. Math scores 2009.

- a) Median: 284
IQR: 10
Mean: 282.98
Standard deviation: 7.60
- b) Since the distribution of Math scores is skewed to the left, it is probably better to report the median and IQR.

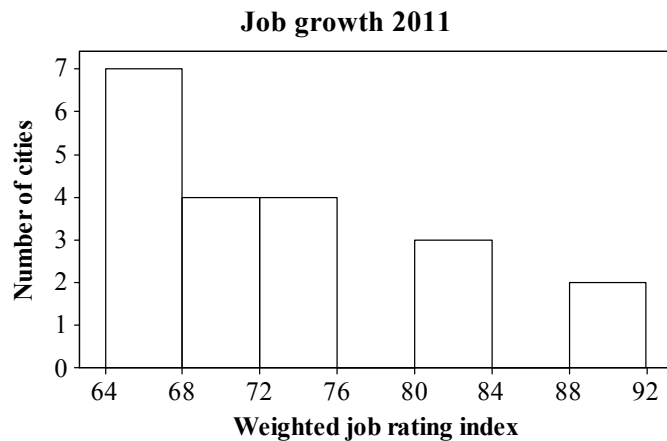


38 **Part I Exploring and Understanding Data**

- c) The distribution of average math achievement scores for eighth graders in the United States is skewed slightly to the left, and roughly unimodal. The distribution is centered at 283. Scores range from 265 to 299, with the middle 50% of the scores falling between 278 and 288. Several low scores, namely California, West Virginia, New Mexico, Alabama, and Mississippi, pull down the mean, making the median a better measure of center.

56. **Boomtowns 2011.**

- a) A histogram of the job growth rates of NewGeography.com's best cities for job growth is at the right. A boxplot, stemplot, or dotplot would also have been an acceptable display.



- b) The mean weighted job rating index is 73.03% and the median weighted job rating index is 71.80%. The mean is higher because distribution is skewed to the right.
- c) The median would be the appropriate measure of center of the distribution of weighted job rating indices, since the distribution is skewed to the right.
- d) The standard deviation of the distribution of weighted job rating indices is 7.61% and the IQR is 10.10%.
- e) The IQR is the appropriate measure of spread, because the skewness influences the standard deviation.
- f) If 49.23% were subtracted from each of the weighted job rating indices, the mean and median would each decrease by 49.23%. The standard deviation and the IQR would not change.
- g) If we were to set aside Austin-Round Rock-San Marcos, the highest weighted job rating index, the mean would decrease. The skewness was pulling it up. The standard deviation would decrease, since the skewness gave the impression of more spread. The median and IQR would be relatively unaffected, since those measures are resistant to the presence of skewness, although they would change slightly, since they are each based upon relative position. With the highest rating removed, there would only be 19 rating indices, instead of 20. This would cause the median and the quartiles to shift down slightly.

- h) The distribution of weighted job rating indices is roughly unimodal and skewed to the right. The median weighted job rating index for these cities is 71.80%. The middle 50% of the cities had weighted job rating indices between 67.25% and 77.35%, for an interquartile range of 10.10%. The median and IQR are the best measures of spread, since the distribution is skewed.

57. Population growth.

The distribution of population growth among the 50 United States and the District of Columbia is unimodal and skewed to the right. Most states experienced modest growth, as measured by percent change in population between 2000 and 2009. Every state experienced positive growth, with 0.3% being the lowest, in Michigan. The median population growth was 7.0%, with the middle 50% of states experiencing between 3.65% and 11.60% growth, for an IQR of 7.95. The distribution contains three high outliers. Utah experienced 24.7% population growth, Arizona 28.6% growth, and Nevada 32.3%.

58. Prisons 2006.

In the year 2006, the median increase in federal prison populations in 21 northeastern and midwestern states was 3.4%, with the middle 50% of the states experiencing between 0.45% to 5.3% growth. Only 4 of the 21 states showed a decrease in federal prison population. The distribution is unimodal and symmetric, with the exception of one low outlier, Maine, whose prison population fell by 7.7% in 2006. On the top end, 25% of these states experienced prison population increases in excess of 5.4%.

