Chapter 4 – Displaying Quantitative Data

1. **Statistics in print.** 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 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 42 Super Bowl games is 45 points.

b) The first quartile of the number of points scored in the first 42 Super Bowl games is 37 points. The third quartile is 55 points.

c) In the first 42 Super Bowl games, the lowest number of points scored was 17, and the highest number of points scored was 75. The median number of points scored was 45, 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 42 Super Bowl games is 13.5 points.

b) The first quartile of the winning margin in the first 42 Super Bowl games is 7 points. The third quartile is 21 points.

c) In the first 42 Super Bowl games the lowest winning margin was 1 point and the highest winning margin was 45 point, which was an outlier. The second highest winning margin was only 36 points. The median winning margin was 13.5 points, with the middle 50% of winning margins between 7 and 21 points.

15. 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(Set 1) = 2.24 SD(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(Set 1) = 3.61 SD(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(Set 1) = 4.24 SD(Set 2) = 4.24

16. 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(Set 1) = 2.12 SD(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(Set 1) = 36.06 SD(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. Range(Set 1) = 18 and Range(Set 2) = 22. Since the Range of Set 2 is greater than the Range of Set 1, the standard deviation is also larger.
SD(Set 1) = 6.03 SD(Set 2) = 7.04

17. Pizza prices.

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

18. Neck size.

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

19. 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 of $1.00.

20. 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.

21. Movie lengths.

a) A typical movie would be around 100 minutes long. This is near the center of the unimodal and slightly skewed histogram, with the outlier set aside.

b) You would be surprised to find that your movie ran for 150 minutes. Only 3 movies ran that long.

c) The mean run time would be higher, since the distribution of run times is skewed to the right, and also has a high outlier. The mean is be pulled towards this tail, while the median resistant.

22. 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 260 yards, and as long as about 320 yards.

b) Approximately 15% of professional male golfers drive less than 280 yards.

c) The actual mean drive is about 288.6 yards, so any estimate between 285 and 290 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 288.35.

23. Movie lengths II.

a) i) The distribution of movie running times is fairly consistent, with the middle 50% of running times between 97 and 119 minutes. The interquartile range is 22 minutes.

ii) The standard deviation of the distribution of movie running times is 19.6 minutes, which indicates that movies typically have running times fairly close to the mean running time.

b) Since the distribution of movie running times is skewed to the right and contains an outlier, the standard deviation is a poor choice of numerical summary for the spread. The interquartile range is better, since it is resistant to outliers.

24. Golf drives II.

a) i) The distribution of PGA golf drives is fairly consistent, with the middle 50% of the drives having distances between 282 and 292 yards. The interquartile range is 10 yards.

ii) The standard deviation of the distribution of PGA golf drives is 9.3 yards, which indicates that golf drives are typically within 9.3 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.

25. 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.

26. 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.

27. Movie budgets.

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

28. 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.

29. 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.

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.

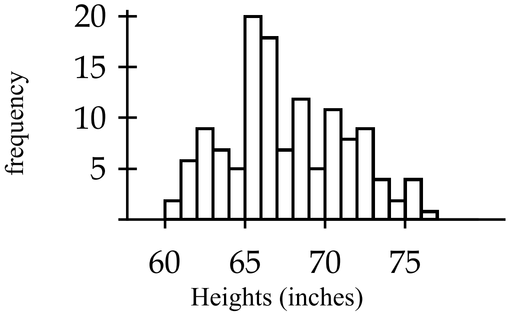
d) The IQR is the better measure of spread for the payroll distribution. The standard deviation and the range are both affected by the two higher salaries.

30. 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.

31. Gasoline.

a) Gasoline Prices

```
2.4 | 56
2.4 |
2.3 | 68
2.3 | 23
2.2 | 677789
2.2 | 1234
Key:
2.2 | 1 = $2.21/gal
```

b) The distribution of gas prices is unimodal and skewed to the right, centered around $2.27 per gallon, with most stations charging between $2.26 and $2.33 per gallon. The lowest and highest prices were $2.27 and $2.46 per gallon.

c) There is a gap in the distribution of gasoline prices. There were no stations that charged between $2.40 and $2.44.

32. The Great One.

a) Wayne Gretzsky –
Games played per season

```
8 | 000000122
7 | 8899
7 | 0344
6 |
6 | 4          Key:
5 |            7 | 8 = 78
5 |                    games
4 | 58
4 |
```

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.

33. 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. 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.

35. Home runs.

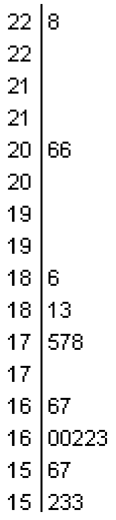
The distribution of the number of homeruns hit by Mark McGwire during the 1986 – 2000 seasons is skewed to the right, with a typical number of homeruns per season in the 30s. With the exception of 3 seasons in which McGwire hit fewer than 10 homeruns, his total number of homeruns per season was between 22 and the maximum of 70.

36. Bird species.

a) The results of the 1999 Laboratory of Ornithology Christmas Bird Count are displayed in the stem and leaf display at the right. This display uses split stems, to give the display a bit more definition. The lower stem contains leaves with digits 0,1,2,3,4 and the upper stem contains leaves with digits 5,6,7,8,9.

b) The distribution of the number of birds spotted by participants in the 1999 Laboratory of Ornithology Christmas Bird Count is skewed right, with a center at around 160 birds. There are several high outliers, with two participants spotting 206 birds and another spotting 228. With the exception of these outliers, most participants saw between 152 and 186 birds.

Christmas Bird
Count Totals
1999



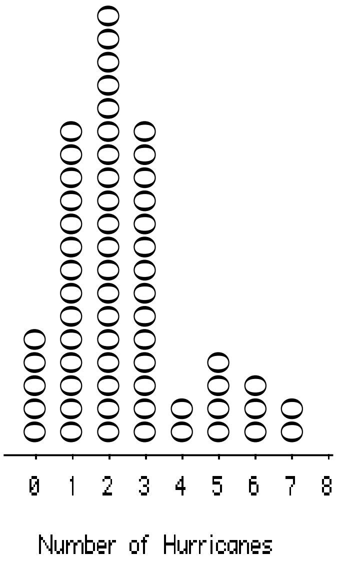
22	8
22	
21	
21	
20	66
20	
19	
19	
18	6
18	13
17	578
17	
16	67
16	00223
15	67
15	233

KEY:
18 | 6 = 186
species
spotted.

37. Hurricanes 2006.

a) A dotplot of the number of hurricanes each year from 1944 through 2006 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 7. There are no outliers. There may be a second mode at 5 hurricanes per year, but since there were only 4 years in which 5 hurricanes occurred, it is unlikely that this is anything other than natural variability.



38. 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.

Consumer Reports
Horsepower

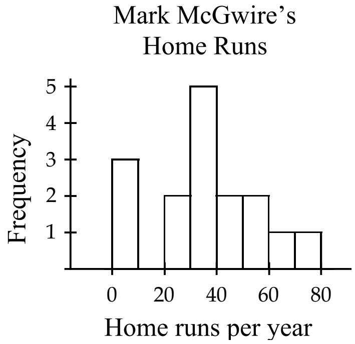
15	05
14	2
13	0358
12	0559
11	00555
10	359
9	00577
8	0058
7	01158
6	55889

KEY:
11 | 5 =
115 horsepower

39. Home runs, again.

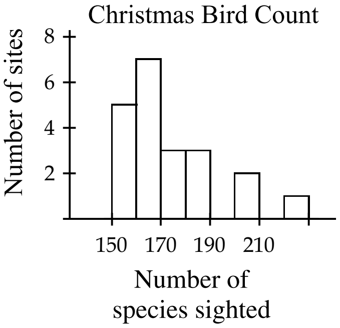
a) This is not a histogram. The horizontal axis should the number of home runs per year, split into bins of a convenient width. The vertical axis should show the frequency; that is, the number years in which McGwire hit a number of home runs within the interval of each bin. The display shown is a bar 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 McGwire.

b) The histogram is at the right.



40. Return of the birds.

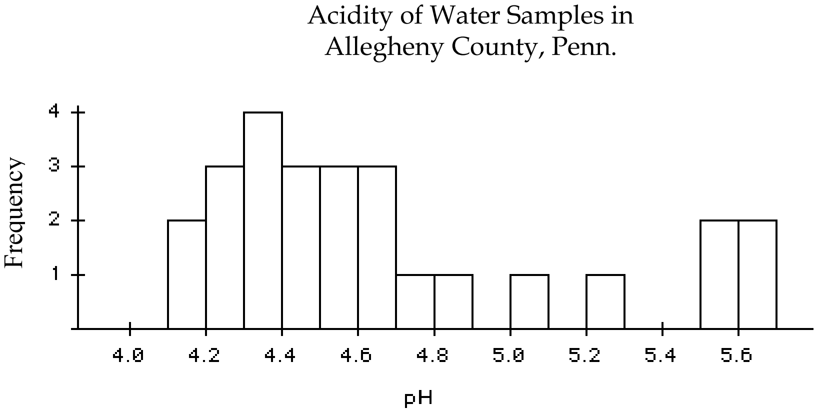
a) This is not a histogram. The horizontal axis should split the number of counts from each site into bins. The vertical axis should show the number of sites in each bin. The given graph is nothing more than a bar chart, showing the bird count from each site as its own bar. It is of absolutely no use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



b) The histogram is at the right.

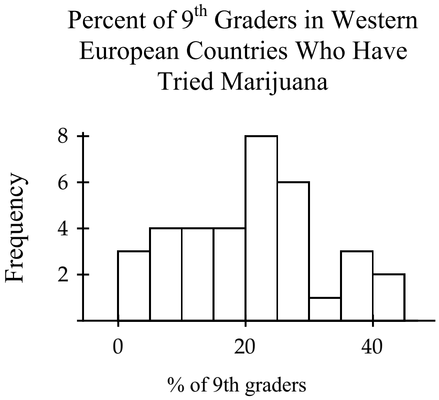
41. Acid rain.

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.



42. Marijuana 2003.

The distribution of the percentage of 9th graders in 20 Western European countries who have tried marijuana is unimodal, but a small cluster of countries have percentages over 35%. Romania, at 3%, has the lowest percentage of 9th graders who have tried marijuana. Czech Republic has the highest percentage, at 44%. A typical country might have a percentage of approximately 20%.



43. 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.

44. 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.

45. 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.

46. 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.

47. Math scores 2005.

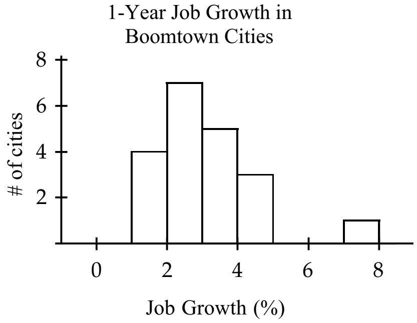
a) Median: 239
IQR: 9
Mean: 237.6
Standard deviation: 5.7

b) Since the distribution of Math scores is skewed to the left, it is probably better to report the median and IQR.

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 239. Scores range from 224 to 247, with the middle 50% of the scores falling between 233 and 242. Several low scores, namely New Mexico, Alabama, and Mississippi, pull down the mean, making the median a better measure of center.

48. Boomtowns.

a) A histogram of the job growth rates of *Inc.* magazine's top 20 boomtowns is at the right. A boxplot, stemplot, or dotplot would also have been an acceptable display.



b) The mean predicted growth rate is 3.07% and the median predicted growth rate is 2.85%. The mean is higher because the outlier, the predicted growth rate in Las Vegas, NV, pulls it up.

c) The median would be the appropriate measure of center of the distribution of predicted job growth rates, since the distribution contains an outlier.

d) The standard deviation of the distribution of predicted job growth rates is 1.37% and the IQR is 1.1%.

e) The IQR is the appropriate measure of spread, because the outlier influences the standard deviation.

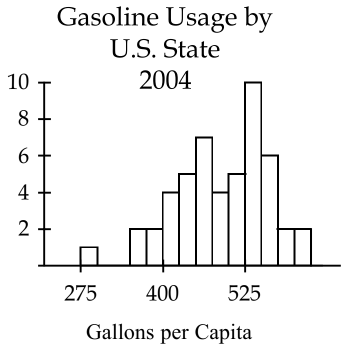
f) If 1.2% were subtracted from each of the predicted job growth rates, the mean and median would each decrease by 1.2%. The standard deviation and the IQR would not change.

g) If we were to set aside Las Vegas, an outlier, the mean would decrease. The outlier was pulling it up. The standard deviation would decrease, since the presence of the outlier gave the impression of more spread. The median and IQR would be relatively unaffected, since those measures are resistant to the presence of outliers, although they would change slightly, since they are each based upon relative position. With the outlier removed, there would only be 19 job predicted job growth rates, instead of 20. This would cause the median and the quartiles to shift down slightly.

h) The distribution of job growth rates is roughly unimodal and symmetric except for the one outlier, Las Vegas at 7.5%. The median growth rate for these cities is 2.85%. The middle 50% of the cities had growth rates between 2.25% and 3.35%, for an interquartile range of 1.1%. The median and IQR are the best measures of spread, unless the growth rate for Las Vegas is set aside. If we do this, the mean and standard deviation for the remaining cities are 2.84% and 0.91%, respectively.

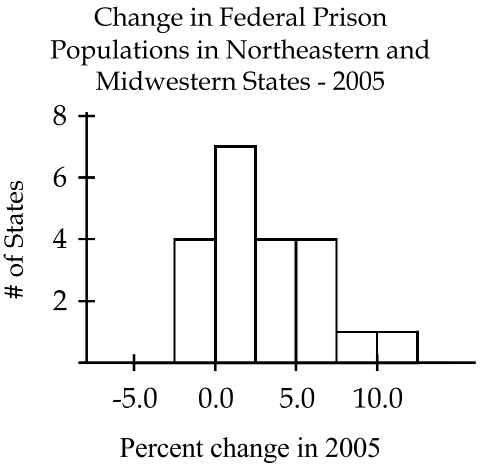
49. Gasoline usage 2004.

In 2004, per capita gasoline usage by state in the United States averaged approximately 500 gallons (mean 488.7, median 500.5). The distribution of gasoline usage was bimodal, and slightly skewed to the left, with one low outlier, New York. This state used much less gasoline per person than other states. The IQR of the distribution was 96.9 gallons per person, with the middle 50% of states having gasoline usage between 447.5 and 544.4 gallons per person.



50. Prisons 2005.

In the year 2005, the median increase in federal prison populations in 20 northeastern and midwestern states was 2.3%. Only 4 of the 20 states showed a decrease in federal prison population. The distribution is unimodal and skewed to the right. The large IQR of 4.7% indicates much variability from state to state, with 25% of these states experiencing prison population increases in excess of 5.5%.


Change in Federal Prison Populations in Northeastern and Midwestern States - 2005
# of States
Percent change in 2005