Chapter 5 – Describing Distributions Numerically

1. In the news. Answers will vary.

2. In the news II. Answers will vary.

3. Summaries.
   a) The mean price of the electric smoothtop ranges is $1001.50.
   b) In order to find the median and the quartiles, the list must be ordered.
      
      565   750   850   900   1000   1050   1200   1250   1400
      
      The median price of the electric ranges is $1025.
      Quartile 1 = $850 and Quartile 3 = $1200.
   c) The range of the distribution of prices is Max – Min = $1400 – $565 = $835.
      The IQR = Q3 – Q1 = $1200 - $850 = $350.

   a) The mean annual number of deaths from tornadoes is 56.27.
   b) In order to find the median and the quartiles, the list must be ordered.
      
      25 30 33 39 39 40 53 67 69 94 130
      
      The median annual number of deaths from tornadoes is 40.
      Quartile 1 = 33 deaths, and Quartile 3 = 69 deaths.
      (Some statisticians consider the median to be part of the both the lower and upper halves of the ordered list. This changes the position of the quartiles slightly. If median is included, Q1 = 36, Q3 = 68. In practice, it rarely matters, since these measures of position are best for large data sets.)
   c) The range of the distribution of deaths is Max – Min = 130 – 25 = 105 deaths.
      The IQR = Q3 – Q1 = 69 – 33 = 36 deaths. (Or, the IQR = 68 – 38 = 32 deaths, if the median is included in both halves of the ordered list.)

5. 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.

6. 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.
7. **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.

8. **Singers.**

   a) 5-number summary: 60, 65, 66, 70, 76

   b) The boxplot of heights of the choir members is at the right.

   c) The mean height of the singers is 67.12 inches, and the standard deviation of the heights is 3.79 inches.

   d) The histogram of heights of the choir members is below the boxplot.

   e) 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.

9. **Standard deviation.**

   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.

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

   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.

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

   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.

   \[ \text{SD}(\text{Set 1}) = 4.24 \quad \text{SD}(\text{Set 2}) = 4.24 \]
10. Standard deviation.
   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.
   \[ SD(\text{Set 1}) = 2.12 \quad SD(\text{Set 2}) = 2.24 \]
   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.
   \[ SD(\text{Set 1}) = 36.06 \quad SD(\text{Set 2}) = 36.06 \]
   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. Range(\text{Set 1}) = 18 and 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.
   \[ SD(\text{Set 1}) = 6.03 \quad SD(\text{Set 2}) = 7.04 \]

11. 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.

   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.
13. Home runs.

5-number summary: 8, 14, 24.5, 33, 61  
IQR = Q3 – Q1 = 33 – 14 = 19 home runs.

Using the Outlier Rule (more than 1.5 IQRs beyond the quartiles) to find the fences:

Upper Fence:  
Q3 + 1.5(IQR) = 33 + 1.5(19)  
= 33 + 28.5  
= 61.5

Lower Fence:  
Q1 - 1.5(IQR) = 14 - 1.5(19)  
= 14 - 28.5  
= -14.5

Any year with more than 61.5 home runs or less than -14.5 home runs (a meaningless fence) is considered an outlier. Technically, the year in which Maris hit 61 home runs is not an outlier, but according to a less formal definition (any point that departs from the overall pattern), it seems strange. For this reason, 61 homeruns is displayed as an outlier in the boxplot.


a) The range is the distance between the minimum and maximum. 82 - 45 = 37 games.

b) IQR = Q3 – Q1 = 80 – 73.5 = 6.5 games.

c) Lower Fence:  
Q1 - 1.5(IQR) = 73.5 - 1.5(6.5)  
= 73.5 - 9.75  
= 63.75

Both the 45 game season and the 48 game season are well below the fence, so they could be considered outliers.

d) The 64 game season is very close to the lower fence of 63.75. It may (or may not) be considered an outlier. Personally, I think that the 64 game season is unusually low for Gretzky.


Comparative boxplots are at the right. Generally, the southern and western states appear to have significantly higher growth between 1990 and 2000 than the northeastern and midwestern states. The growth rates in the southern and western states were also more variable.
16. 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) \[ \text{IQR} = Q_3 - Q_1 = 78 - 28 = 50. \]
      Using the Outlier Rule (1.5 IQRs beyond quartiles):
      \[ \text{Upper Fence: } Q_3 + 1.5(\text{IQR}) = 78 + 1.5(50) \]
      \[ = 78 + 75 \]
      \[ = 153 \]
      \[ \text{Lower Fence: Well below 0 campsites.} \]
      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.

17. Sip size.
   a) The distribution of height is most nearly symmetric and shows no outliers. That makes it the best candidate for summarizing with a mean. The distribution of weight is skewed to the right, which would inflate the value of the mean. The distribution of sip size has an outlier, which would inflate the value of the mean.
   b) The distribution of sip size shows a high outlier. The standard deviation is sensitive to outliers, so the IQR is a better choice for summarizing spread. It is resistant to outliers.

18. How tall?
   The histogram shows some low outliers in the distribution of height estimates. These are probably poor estimates and will pull the mean down. It is likely that an estimate of 56 inches was an attempt to guess 5’6”. The median is likely to give a better estimate of the professor’s true height.

19. Women’s basketball.
   a) Both girls have a median score of about 17 points per game, but Scyrine is much more consistent. Her IQR is about 2 points, while Alexandra’s is over 10.
   b) If the coach wants a consistent performer, she should take Scyrine. She’ll almost certainly deliver somewhere between 15 and 20 points. But, if she wants to take a chance and needs a “big game”, she should take Alexandra. Alex scores over 24 points about a quarter of the time. On the other hand, she scores under 11 points about as often.
20. Gas prices.
   a) Gas prices have been increasing on average over the three year period, and the spread has been increasing as well. The distribution of prices in 2002 was skewed to the left with several low outliers. Since then, the distribution has been increasingly skewed to the right. There is a high outlier in 2004, although it appears to be pretty close to the upper fence.
   b) The distribution of gas prices in 2004 shows the greatest range and the biggest IQR, so the prices varied a great deal.

   The distribution of marriage age of U.S. men is skewed right, with a typical man (as measured by the median) first marrying at around 24 years old. The middle 50% of male marriage ages is between about 23 and 26 years. For U.S. women, the distribution of marriage age is also skewed right, with median of around 21 years. The middle 50% of female marriage age is between about 20 and 23 years. When comparing the two distributions, the most striking feature is that the distributions are nearly identical in spread, but have different centers. Females typically seem to marry earlier than males. In fact, between 50% and 75% of the women marry at a younger age than any man.

22. Fuel economy.
   Cars with 4 cylinders generally get better gas mileage than cars with 6 cylinders, which generally get better gas mileage than cars with 8 cylinders. Additionally, the greater the number of cylinders, the more consistent the mileage becomes. 4 cylinder cars typically get between 27 – 33 mpg, 6 cylinder cars typically get between 18 – 22 mpg, and large cars typically get between 16 – 19 mpg.

23. Grapes.
   a) A Seneca Lake vineyard had the maximum case price of approximately $150.
   b) A Seneca Lake vineyard had the minimum case price of approximately $50.
   c) The Keuka Lake vineyards were consistently high.
   d) Cayuga Lake vineyards and Seneca Lake vineyards have approximately the same average case price, at about $100, while a typical Keuka Lake vineyard has a case price of about $130. Keuka Lake vineyards have consistently high case prices, between $120 and $140, with one low outlier at about $85 per case. Cayuga Lake vineyards have case prices from $70 to $135, and Seneca Lake vineyards have highly variable case prices, from $50 to $150.

24. Ozone.
   a) April had the highest recorded ozone level, approximately 440.
   b) February had the largest IQR of ozone level, approximately 50.
   c) August had the smallest range of ozone levels, approximately 50.
   d) January had a slightly lower median ozone level than June, 340 and 350, respectively, but June’s ozone levels were much more consistent.
e) Generally, ozone levels rose through the winter and were highest in the spring, then fell through the summer and were lowest in the fall. Additionally, ozone levels were very consistent in the summer, became more variable in the fall, were most variable in the winter, and became more consistent through the spring.

25. **Wild card Summer Olympics.**

a) Since the mean time is larger than the median time, the distribution of qualifying swim times should be skewed to the right.

b) Even if the distribution was generally symmetric, some swimmers with very high times might have been outliers in the distribution.

c) The distribution of qualifying times in the men’s 100 meter swim at the 2000 Olympics in Sydney was unimodal and skewed slightly to the right, with all but one swimmer having times between 47 and 63 seconds. One swimmer, Eric Moussambani, had a qualifying time of over 110 seconds.

26. **Unemployment.**

The distribution of unemployment rates reported by the U.S. Bureau of Labor Statistics must have had about 30\% of the rates above 4.2\% and the rest below. In other words, 4.2\% should be slightly below the third quartile (25\% of rates above). Other than that, the distribution cannot be accurately determined. An example of one possible distribution is shown in the boxplot at the right.

27. **Test scores.**

a) Class 3 had the highest mean score, probably somewhere in the 70s. The other two classes had mean scores in the 60s.

b) Class 3 had the highest median score, probably somewhere in the 80s. The other two classes had median scores in the 60s.

c) Class 3 has median higher than the mean, since the distribution of scores is skewed to the left. The mean is pulled toward the tail. The other two classes have roughly symmetric distributions, with mean and median approximately equal.

d) Class 1 has the smallest standard deviation. Most scores are clustered close to the mean.

e) Class 1 probably has the smallest IQR. However, without the actual scores, it is impossible to calculate the exact IQR of these classes. We can estimate them. Class 1 has 24 students, so Q1 is between the 6th and 7th scores, somewhere between 50 and 60. Q3 is between the 18th and 19th scores, somewhere between 70 and 80. The IQR is at least 10 and at most 30. Class 3 also has 24 scores. Q1 is between 60 and 70. Q3 is between the 18th score (80-90) and the 19th score (90-100), meaning it could be between 80 and 100. The IQR is at least 10 and at most 40. Class 2 seems to have the largest IQR, of at least 30 and at most 50.
28. Test scores.
   a) Class 3 did better overall. Class 3 has Q1 at about the same score as the median of either of
      the other two classes. This means that 75% of the students in Class 3 scored higher than the
      median of the other two classes.
   b) Classes 1 and 2 had distributions of test scores that were roughly symmetric, while Class 3
      had a distribution of scores that was skewed to the left. All classes were fairly spread out,
      with scores ranging from approximately 30 to 100.
   c) Class A is Class 1  Class B is Class 2  Class C is Class 3

29. 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.

30. Golf courses.
   a) The range of the lengths of the golf courses in Max – Min = 6796 – 5185 = 1611 yards.
   b) In any distribution, 50% of scores lie between Quartile 1 and Quartile 3. In this case,
      Quartile 1 = 5585.75 yards and Quartile 3 = 6131 yards.
   c) The distribution of golf course lengths appears roughly symmetric, so the mean is the
      appropriate measure of center. In this case, the mean is 5892.91 yards.
   d) The distribution of the lengths of all the golf courses in Vermont is roughly unimodal and
      symmetric. The mean length of the golf courses is approximately 5900 yards. Vermont has
      golf courses anywhere from 5185 yards to 6796 yards long. There are no outliers in the
      distribution.

31. Graduation?
   a) The distribution of the percent of incoming college freshman who graduate on time is
      roughly symmetric. The mean and the median are reasonably close to one another and the
      quartiles are approximately the same distance from the mean.
48  **Part I Exploring and Understanding Data**

b) Upper Fence:  \[ Q_3 + 1.5(IQR) = 74.75 + 1.5(74.75 - 59.15) \]
\[ = 74.75 + 23.4 \]
\[ = 98.15 \]

Lower Fence:  \[ Q_1 - 1.5(IQR) = 59.15 - 1.5(74.75 - 59.15) \]
\[ = 59.15 - 23.4 \]
\[ = 35.75 \]

Since the maximum value of the distribution of the percent of incoming freshmen who graduate on time is 87.4% and the upper fence is 98.15%, there are no high outliers. Likewise, since the minimum is 43.2% and the lower fence is 35.75%, there are no low outliers. Since the minimum and maximum percentages are within the fences, all percentages must be within the fences.

c) A boxplot of the distribution of the percent of incoming freshmen who graduate on time is at the right.

d) The distribution of the percent of incoming freshmen who graduate on time is roughly symmetric, with mean of approximately 68% of freshmen graduating on time. Universities surveyed had between 43.2% and 87.4% of students graduating on time, with the middle 50% of universities reporting between 59.15% and 74.75% graduating on time.

32. Vineyards.

a) The distribution of size of Finger Lakes vineyards is skewed heavily to the right. The mean size is a great deal higher than the median size.

b) Upper Fence:  \[ Q_3 + 1.5(IQR) = 55 + 1.5(55 - 18.5) \]
\[ = 55 + 54.75 \]
\[ = 109.75 \]

Lower Fence:  \[ Q_1 - 1.5(IQR) = 18.5 - 1.5(55 - 18.5) \]
\[ = 18.5 - 54.75 \]
\[ = -36.25 \]

The maximum of 250 acres is well above the upper fence of 109.75 acres. Therefore, there is at least one high outlier, 250 acres. Since the lower fence is negative, there are no low outliers, since it is certainly impossible to have a vineyard with negative size.
c) The boxplot of the distribution of sizes of Finger Lakes vineyards is at the right. There may be additional outliers, but we are sure that there is at least one, the maximum.

d) The distribution of sizes of Finger Lakes vineyards is skewed to the right. Many vineyards have moderate sizes, with the middle 50% of vineyards consisting of 18.5 to 55 acres. The smallest vineyard is 6 acres. At least one vineyard is comparatively bigger, at 250 acres.

33. Caffeine.

a) Who – 45 volunteers. What – Level of caffeine consumption and memory test score. When – Not specified. Where – Not specified. Why – The student researchers want to see the possible effects of caffeine on memory. How – It appears that the researchers imposed the treatment of level of caffeine consumption in an experiment. However, this point is not clear. Perhaps they allowed the subjects to choose their own level of caffeine.

b) Variables – Caffeine level is a categorical variable with three levels: no caffeine, low caffeine, and high caffeine. Test score is a quantitative variable, measured in number of items recalled correctly.

c) The groups consuming no caffeine and low caffeine had comparable memory test scores. A typical score from these groups was around 21. However, the scores of the group consuming no caffeine were more consistent, with a smaller range and smaller interquartile range than the scores of the group consuming low caffeine. The group consuming high caffeine had lower memory scores in general, with a median score of about 19. No one in the high caffeine group scored above 24, but 25% of each of the other groups scored above 24.
34. SAT scores.

a) Parallel boxplots comparing the scores of boys and girls SAT scores are at the right.

b) Females in this graduating class scored slightly higher on the Verbal SAT, with a median of 625, compared to the median of 600 for the males. Additionally, the females had higher first and third quartiles. The IQR of the males’ scores was slightly smaller, than the IQR for the females’ scores, indicating a bit more consistency in male scores. However, the overall spread of male scores was greater than that of female scores, with males having both the minimum and maximum score. Both distributions of scores were slightly skewed to the left.

35. Derby speeds.

a) The median speed is the speed at which 50% of the winning horses ran slower. Find 50% on the left, move straight over to the graph and down to a speed of about 36 mph.

b) Quartile 1 is at 25% on the left, and Quartile 3 is at 75% on the left. Matching these to the ogive, Q1 = 34.5 mph and Q3 = 36.5 mph.

c) Range = Max – Min = 38 – 31 = 7 mph
   IQR = Q3 – Q1 = 36.5 – 34.5 = 2 mph

d) A boxplot of winning Kentucky Derby Speeds is at the right.

e) The distribution of winning speeds in the Kentucky Derby is skewed to the left. The lowest winning speed is just under 31 mph and the fastest speed is about 37.5 mph. The median speed is approximately 36 mph, and 75% of winning speeds are above 34.5 mph. Only a few percent of winners have had speeds below 33 mph.

36. Cholesterol.

A boxplot for the distribution of cholesterol levels of 1400 men is at the right. The five number summary is estimated from the ogive to be: 100, 200, 230, 260, 425.

The distribution of cholesterol levels is skewed to the right, and tightly clustered around the median. The median cholesterol level is approximately 230, with the middle 50% of cholesterol levels between 200 and 260. A very small percentage of men had cholesterol below 150 or above 325.
37. Reading scores.
   a) The highest score for boys was 6, which is higher than the highest score for girls, 5.9.
   b) The range of scores for boys is greater than the range of scores for girls.
      \[ \text{Range} = \text{Max} - \text{Min} \]
      \[ \text{Range(Boys)} = 4 \quad \text{Range(Girls)} = 3.1 \]
   c) The girls had the greater IQR.
      \[ \text{IQR} = Q_3 - Q_1 \]
      \[ \text{IQR(Boys)} = 4.9 - 3.9 = 1 \quad \text{IQR(Girls)} = 5.2 - 3.8 = 1.4 \]
   d) The distribution of boys’ scores is more skewed. The quartiles are not the same distance from the median. In the distribution of girls’ scores, Q1 is 0.7 units below the median, while Q3 is 0.7 units above the median.
   e) Overall, the girls did better on the reading test. The median, 4.5, was higher than the median for the boys, 4.3. Additionally, the upper quartile score was higher for girls than boys, 5.2 compared to 4.9. The girls’ lower quartile score was slightly lower than the boys’ lower quartile score, 3.8 compared to 3.9.
   f) The overall mean is calculated by weighting each mean by the number of students.
      \[ \frac{14(4.2) + 11(4.6)}{25} = 4.38 \]

38. Rainmakers?
   a) Median and IQR, as well as the quartiles are the appropriate summary statistics, since the distribution of amount of rain produced is either skewed to the right or has high outliers. Indication of skewness and outliers can be seen in the comparison of median and mean. The mean amount of rain produced is significantly higher than the median for both seeded and unseeded clouds. Skewness or outliers pulled up the sensitive mean.
   b) There is evidence that the seeded clouds produced more rain. The median and both quartiles are higher than the corresponding statistics for unseeded clouds. In fact, the median amount of rainfall for seeded clouds is 221.60 acre-feet, about 5 times the median amount for unseeded clouds.

39. Phone calls.
   a) A boxplot of the distribution of Net2Phone rates is at the right.
   b) The mean rate for Net2Phone is 13.7 cents per minute. The median rate is 9.9 cents per minute. The rates for India and Pakistan are outliers, so the median is the appropriate measure of center.
c) The IQR of the distribution of Net2Phone rates is 7.6 cents per minute and the standard deviation is 11.8 cents per minute. The rates for India and Pakistan are outliers, so the IQR is the appropriate measure of spread.

d) Upper Fence: \[ Q3 + 1.5\text{IQR} = 15.5 + 1.5(15.5 - 7.9) \]
\[ = 15.5 + 11.4 \]
\[ = 26.9 \]

Lower Fence: \[ Q1 - 1.5\text{IQR} = 7.9 - 1.5(15.5 - 7.9) \]
\[ = 7.9 - 11.4 \]
\[ = -3.5 \]

Any country that has a rate higher than 26.9 cents per minute is an outlier. Both India and Pakistan have rates of 49 cents per minute. No country could be a low outlier. It is impossible to have a rate lower than –3.5 cents per minute.

e) The distribution of the long distance rates for Net2Phone is unimodal and symmetric except for two outliers at 49 cents per minute. These are the rates for India and Pakistan. The median rate is 9.9 cents per minute. The middle 50% of rates are between 7.9 and 15.5 cents per minute for an IQR of 7.6 cents per minute.

f) The rates of these 24 countries may not be representative of the rates of all 250 countries that are serviced by Net2Phone. It would be unwise to assume that these rates could tell us anything about Net2Phone rates in general.

40. Job growth.

a) A histogram of the job growth rates predicted by Standard and Poor’s DRI is shown at the right. A boxplot, stemplot, or dotplot would also have been an acceptable display.

b) The mean predicted growth rate is 2.317% and the median predicted growth rate is 2.235%. 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 0.427% and the IQR is 0.515%.

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 omit 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 predicted by Standard and Poor’s DRI is roughly unimodal and symmetric except for the one outlier, Las Vegas at 3.72%. The median growth rate for these cities is 2.235%. The middle 50% of the cities had growth rates between 2.045% and 2.560%, for an interquartile range of 0.515%.

41. Math scores.

   a) 5-number summary: 275, 448, 499, 531, 604
       IQR: 83
       Mean: 487.2
       Standard deviation: 72.3

   b) The distribution of average math achievement scores for eighth graders in 38 nations is unimodal, skewed slightly to the left, and centered at about 500. South Africa’s score of 275 is the lowest and is a departure from the overall pattern of the data. Not including this score, scores range from 337 to 604, with the middle 50% of the scores falling between 448 and 531. The United States score of 502, although well above the mean of 487.2, is actually near the middle of the distribution. Several low scores pull down the mean, making the median a better measure of center.
42. Prisons.

a) A histogram of the increase in federal prison populations for northeastern and midwestern states in 1999 is shown at the right.

b) Since the distribution is roughly unimodal and symmetric, mean and standard deviation are the appropriate measures of center and spread. The five number summary will also be useful.
   Mean = 5.08%
   Std. Dev. = 1.79%
   5-number summary:
   1.3%, 3.95%, 5.4%, 6.35%, 8%.

c) The distribution of percent increases in federal prison populations in 20 northeastern and midwestern states during 1999 is unimodal and roughly symmetric. There are no outliers.
   The mean percent increase was 5.08% with a standard deviation of 1.79%. The middle 50% of states had an increase of between 3.95% and 6.35%.

43. Gasoline usage.

In 2000, per capita gasoline usage by state in the United States averaged approximately 500 gallons. The distribution of gasoline usage was slightly skewed to the left, with two low outliers, North Carolina and Hawaii. These states used much less gasoline per person than other states. The IQR of the distribution was 82 gallons per person, with the middle 50% of states having gasoline usage between 453 and 535 gallons per person.

44. Industrial experiment.

First of all, there is an extreme outlier in the distribution of distances for the slow speed drilling. One hole was drilled almost an inch away from the center of the target! If that distance is correct, the engineers at the computer production plant should investigate the slow speed drilling process closely. It may be plagued by extreme, intermittent inaccuracy. The outlier in the slow speed drilling process is so extreme that no graphical display can display the distribution in a meaningful way while including that outlier. That distance should be removed before looking at a plot of the drilling distances.
With the outlier removed, we can determine that the slow drilling process is more accurate. The greatest distance from the target for the slow drilling process, 0.000098 inches, is still more accurate than the smallest distance for the fast drilling process, 0.000100 inches.

45. Customer Database.
   a) The mean of 54.41 is meaningless. The codes assigned to the titles are categories, even though the categories are represented by numbers. Averaging is only appropriate for quantitative data.
   b) Typically, the mean and standard deviation are influenced by outliers and skewness.
   c) There is no point in discussing the reason for the differences in the summary statistics, because the summary statistics are only appropriate for quantitative data. The title codes are categorical.

46. Zip codes revisited
   a) Neither mean nor median is appropriate for these data. Zip codes are categories, and mean and median are only appropriate for quantitative data.
   b) Neither standard deviation nor IQR is appropriate for these data. Zip codes are categories, and standard deviation and IQR are only appropriate for quantitative data.
   c) 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. Eye and hair color.
   The graph is not appropriate. Boxplots are for quantitative data, and these are categorical data, although coded as numbers. The numbers used for hair color and eye color are arbitrary, so the boxplot and any accompanying statistics for eye color make no sense.

48. Stereograms.
   a) The two variables discussed in the description are fusion time and treatment group.
   b) Fusion time is a quantitative variable, measured in seconds. Treatment group is a categorical variable, with subjects either receiving verbal clues only, or visual and verbal clues.
   c) Generally, the Visual/Verbal group had shorter fusion times than the No/Verbal group. The median for the Visual/Verbal group was approximately the same as the lower quartile for the No/Verbal group. The No/Verbal Group also had an extreme outlier, with at least one subject whose fusion time was approximately 50 seconds. There is evidence that visual information may reduce fusion time.
49. Stereograms, revisited.

The re-expression using logarithms has a distribution that is more symmetric than the original distribution of fusion times, and the re-expression has no outliers. This symmetry makes it easier to compare the two groups.

50. Stereograms, yet again.

The reciprocal re-expression of the fusion times has a distribution that shows skewness, as in the original distribution of fusion times, although it has no outliers. Of the three, the logarithmic re-expression seems to be the best.