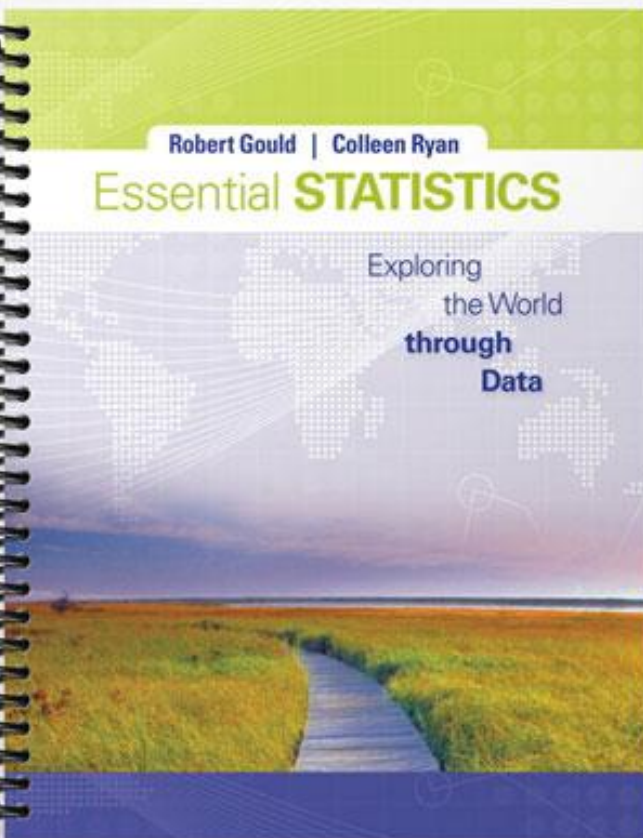# SOLUTIONS MANUAL

Robert Gould | Colleen Ryan

Essential **STATISTICS**

Exploring
the World
through
Data

Chapter 2: Picturing Variation with Graphs

SECTIONS 2.1 & 2.2

2.1   **a.**   There are 11 persons that would be classified as morbidly obese.

 b.   The percentage of persons in this sample who are morbidly obese is then $^{11}\!/_{134} \approx 0.082$, or approximately 8.2%. This sample has a substantially higher percentage of persons who are morbidly obese than the stated figure dating from 2005 of 3% who are morbidly obese.

2.2   **a.**   There are 21 adults whose cholesterol levels exceed 240 mg/dL.

 b.   The percentage of adults with unhealthy total cholesterol levels is then $^{21}\!/_{93} \approx 0.226$, or approximately 22.6%. This sample contains a higher percentage of adults who have unhealthy total cholesterol levels versus the stated figure of 18% from 2010.

2.3   Since there are a total of 25 CEOs, we simply divide each of the numeric labels on the vertical (frequency) axis. These are usually converted to decimals, rather than being listed as fractional values.

2.4   Since the right-most two bars of the histogram represent those colleges who charge $30,000 or more in tuition, we simply sum the relative frequencies for each of these bars. The total percentage is 6% and 6% of 85 is about 5.

2.5   **a.**   Only one home in the study does not have a television.

 b.   The largest number of televisions in a home is nine, as indicated by the right-most bar in the histogram.

 c.   27 homes have exactly three televisions.

 d.   The right-most two bars represent those homes with six or more televisions. There are a total of six homes that have six or more televisions.

 e.   The proportion of homes with six or more televisions is then $^{6}\!/_{90} \approx 0.0667$.

2.6   **a.**   The right-most bar could represent a value of 18 or 19 hours. Note that 20 hours, if it were the maximum value, would have appeared in the bar just to the right of the label "20," according to the right-hand rule.

 b.   Eight persons exercised 0 or 1 hour (as shown by the left-most bar).

 c.   Five persons exercised 10 or more hours (a total of three, one, and one persons respectively in the right-most three bars).

 d.   The proportion of persons who exercised 10 or more hours is $^{5}\!/_{50} = 0.10$.

2.7   **a.**   NYC typically has the highest rents since most of the dots correspond to higher monthly rents than for the other cities.

 b.   Las Vegas has the smallest variation since the data are clustered tightly around the center (average) of the data.

 c.   Las Vegas' distribution is least skewed since there are no tails in the distribution as there are for the other cities' distributions.

2.8   **a.**   Los Angeles typically has the highest rental prices since most of the dots correspond to higher monthly rents than for the other cities.

 b.   LA has the highest variation in its rental prices since the data are more spread out than the other cities.

      c. We could see the sample data as bimodal (with modes at $400 and a bit above $450). However, it would be unwise to conclude that the Houston rental market is characterized by this bimodality without additional information on the method by which rental units were surveyed.

2.9   Because of the outlier, we would expect the distribution to be right-skewed with most CEO salaries clustered near the typical value of $50 million.

2.10   The distribution is most likely right-skewed since there are likely to be just a few persons who smoke a large number of cigarettes per day.

2.11   Since men and women have different average heights (men are generally taller), the distribution of armspans will be bimodal with a cluster of longer armspans (the men) and a cluster of shorter armspans (the women).

2.12   Differences between costs for private versus public institutions could account for the bimodality in the distribution.

2.13   The mode represents the typical age for the CEOs in the data set: since these are ages between 56 and 60, 58 would be a reasonable estimate.

2.14   The typical costs for tuition correspond to costs between $16,000 and $20,000. We average these two values to obtain $18,000 as a reasonable estimate.

2.15   The average commuting time by car appears to be approximately 16 minutes. The average commuting time by bus appears to be nearly twice this value. It is clear from the two histograms that the distribution representing commuting by bus is more variable, since the data are more spread out; the data for commuting by car is tightly clustered together.

2.16   **a.** The histograms are fairly similar: both are bimodal with modes at about $100 and $200 spent per month. The distribution of spending for women contains two respondents whose spending is significantly above average.
      b. Women tend to spend more. The effect of the two respondents who spend a significant amount per month is to raise the mean monthly spending on clothing for women.
      c. The distribution of spending for women would have slightly higher variation since more values are spread out farther away from the mean.

2.17   **a.** The data are multimodal with modes at 12 years, 14 years, 16 years and 18 years of formal education completed. These correspond with traditional lengths of time to graduation for various degrees.
      b. The bars corresponding to 16 or more years completed have estimated frequency counts of 300 (or so), 50, 125, 40, and 50 respectively. This gives a total of upwards of 600 persons who have completed a bachelor's degree or higher.
      c. We can approximate the percentage of persons with a bachelor's degree by $^{600}\!/_{2018} \approx 0.297$, or 29.7%. This figure is fairly similar to the figure from Wikipedia.

2.18   **a.** The distribution is right-skewed since the tail extends to the right.
      b. The mode of the distribution represents the typical number of siblings, which are two.
      c. Approximately 80 persons from the survey have no siblings.
      d. We have approximately $^{80}\!/_{2000} = 0.04$ or 4% with no siblings.

2.19   Both distributions show a peak corresponding to those graduating with a high school degree as well as smaller peaks corresponding to those graduating with a two-year degree, a four-year (college) degree, or a master's degree (usually two years after a bachelor's degree). However, in the respondents' distribution these peaks are much taller than the ones in their mothers' distribution. At the same time, the left tail of the respondents' distribution is lighter than the distribution of their mothers, meaning that comparatively fewer respondents than their mothers completed only a few years of formal education.

2.20   Both distributions are fairly symmetric and each show a peak at approximately 40 hours worked since it's common for men or women to work 40 hours per week. However, the distribution for hours worked by men has a "heavier" right-hand tail than that of the women, meaning that more men work more than 40 hours per week than women do.

2.21   Situation 1 matches distribution (b). The distribution shows few deaths for younger women and a progressively higher number of deaths for older women up to a certain age, then a slightly lower number of deaths for even older women (since fewer women are alive then).

Situation 2 matches distribution (c). The distribution is right-tailed since there are few institutions with substantially high tuition costs (likely private colleges) and more institutions with lower tuition costs.

Situation 3 matches distribution (a). The distribution shows the six possible outcomes (the numbers 1 through 6) and we would expect that a fair die would show each of the numbers about an equal number of times.

2.22   Situation 1 matches distribution (a). Larger numbers of students will do well on an easy test and fewer will do badly.

Situation 2 matches distribution (b). We would expect more people to watch a smaller number of hours of TV.

Situation 3 matches distribution (c). The distribution of heights will be symmetric with few adults who are either very short or very tall.

2.23   We follow the steps described in the solution to this Guided Exercise.
(Step 1)   First, create a pair of dotplots on the same axis.
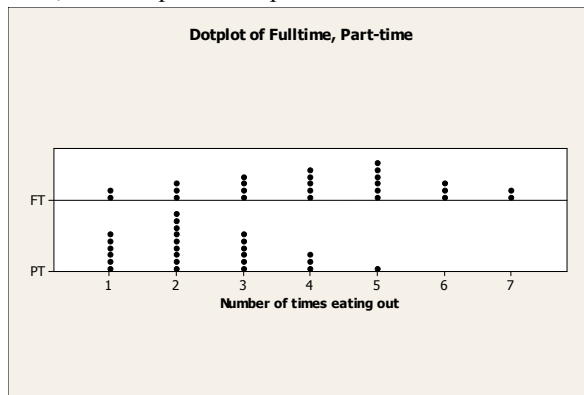
(Step 2)   The data for part-time workers is somewhat right-skewed, while the data for fulltime workers is a bit left-skewed.

(Step 3)   It is fairly clear by comparing the centers of the data sets that those with fulltime jobs eat out more often than those with only part-time employment.

(Step 4)   The data for part-time workers is less spread out since it ranges from 1 to 5, while the data for fulltime workers is more spread out since it ranges from 1 to 7.

(Step 5)   Since there are no gaps in either data set, neither data set contains outliers.

(Step 6)   Those with fulltime employment typically eat out more often in a month than do part-time workers. The data for fulltime workers is slightly left-skewed and the data for part-time workers is slightly right skewed. Also, the data for fulltime workers is more spread out than for part-time workers. There are no outliers in either data set.

2.24   By comparing the dotplots for the weights of baseball players with those of soccer players, we observe that the distribution of weights of baseball players is right-skewed and the distribution of weights of soccer players is left-skewed. It appears that there are outliers (at 225 lbs and 230 lbs) for the baseball players but no outliers for the soccer players' weights.
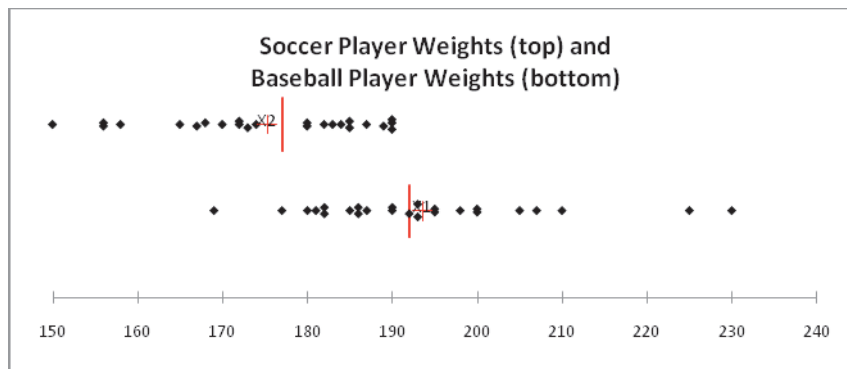

2.25   Since there are a variety of prices, it makes sense to represent the data with a histogram. The distribution of textbook prices appears to be bimodal with modes at approximately $30 and $90.
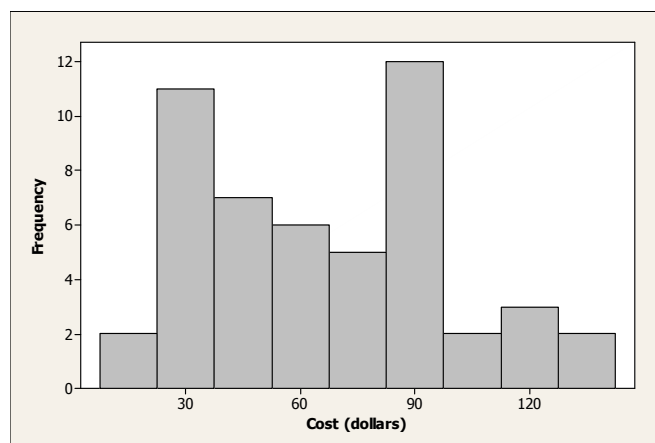

2.26   Since there are a variety of prices, it makes sense to represent the data with a histogram. The distribution appears to be fairly symmetric.
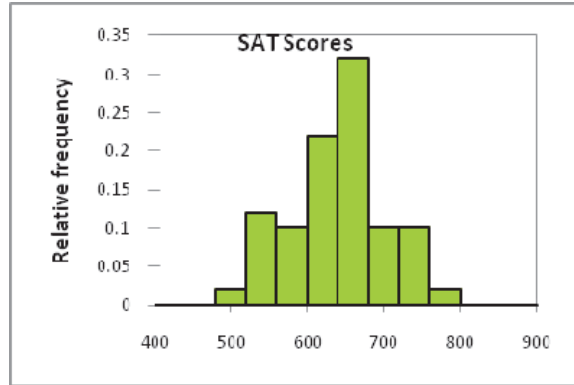
2.27 Either a dotplot or histogram would be appropriate to graph the data. Below is a histogram that shows the distribution of lifespans of the animals in the data set. Observe the distribution is right-skewed. There are three outliers in the data set: the animals with average lifespans of 35years or more. These are the hippo and the two kinds of elephants. If we were to include a data value for humans, it would appear far to the right as another outlier.
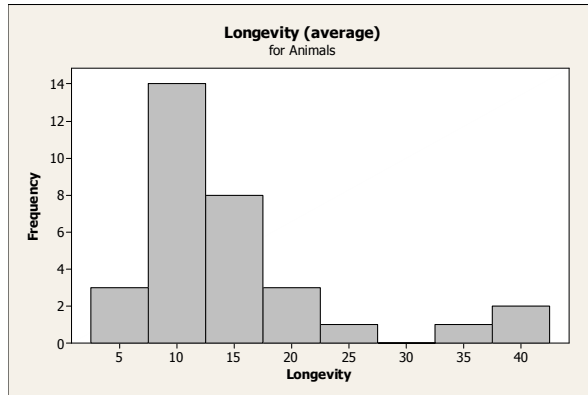
2.28 Either a dotplot or histogram would be appropriate to graph the data. Below is a dotplot that shows the distribution of gestational periods of the animals in the data set. The distribution is right-skewed. It appears there are two clusters of data, one centered near 90 days and another centered near 250 days. There are two outliers: the two kinds of elephants each with a gestational period of more than 630 days. Humans (266 days) would be near the middle of the dotplot.
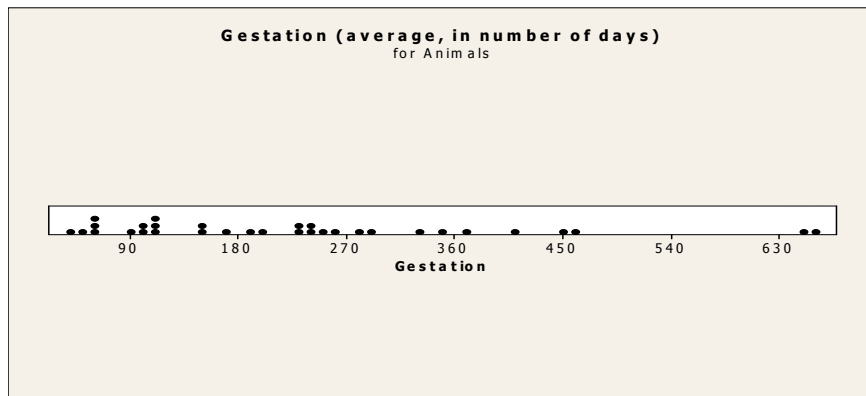
2.29   Situation 1 matches distribution (b). The distribution is bimodal, with one mode for females (who are generally shorter than males) and one mode for males.
Situation 2 matches distribution (a). The distribution is fairly symmetric, with an outlier representing one (or a small number of) student(s) who slept very little the previous night.
Situation 3 matches distribution (c). The majority of students will have zero or a few number of accidents.

2.30   Situation 1 matches distribution (c). We would expect the distribution of SAT scores for students accepted to be fairly symmetric.
Situation 2 matches distribution (a). Since the weights of both males and females are combined, we would expect the resulting distribution to be bimodal with one mode corresponding to the weights of females and another corresponding to the weights of males (who are in general heavier than females).
Situation 3 matches distribution (b). Since the class meets in the morning, we would expect primarily traditional-age college students to attend.

SECTIONS 2.3 & 2.4

2.31   More students changed their answers from wrong to right. Thus, the data do not support the claim that students should not change their initial choice.

2.32   **a.**   More students changed their answers from wrong to right. Thus, the data supports the claim that students should change their initial choice.
b.   Overall, changing your answers leads to higher grades since more students changed from wrong to right.

2.33   **a.**   The mode for both time periods is the Baidu search engine since it is most used. In 2010, the Baidu search engine is preferred over Google even more so than in 2009.
b.   There was more variation in the data in 2009 since more persons chose to use search engines other than the preferred one (Baidu).

2.34   Over the time period, Internet Explorer lost market share to Firefox and Chrome (which wasn't available in 2007). In 2007, Internet Explorer was the mode (most preferred) and in 2010, the mode was Firefox. There was more variability in 2010.

2.35   **a.**   The classification of Democrats (not strong) has the most men. It is usually easier to compare the heights of the bars, rather than the size of the wedges in a pie chart.
b.   The political affiliation with the second largest number of men is Republican (not strong).

2.36   **a.**   The political affiliation with the most women is Democrat (not strong). It is usually easier to compare the heights of the bars, rather than the size of the wedges in a pie chart.
b.   The political affiliation with the second largest number of women is Democrat (strong).
c.   Yes, there is evidence that women are more liberal because the first and second largest groups of women are both Democratic, while men are more evenly split between Democrat and Republican.

2.37   **a.**   In 2030 fewer adults will be between the ages of 25 and 64 and more will be over 65 years old than in 2010. The category of persons aged 24 and below is predicted to stay relatively the same size during this timeframe.
b.   With a greater number of retired persons and fewer working adults in 2030 than are currently retired or

working, less money will be available for a larger number of eligible retired persons. It will be difficult to continue operating the Social Security program as is; changes will be needed to address these constraints.

2.38   **a.** The mode in each of the given years is of Midsize cars.
       b. From 1985 to 2000, the percentage of small cars dropped slightly, but then increased from 2000 to 2007.
       c. From 1985 to 2000, the percentage of large cars went down, but then rebounded some from 2000 to 2007.

2.39 A Pareto chart shows the percentage of students with majors in each respective area. The mode is Social Science majors but variation is high (since there are also many humanities majors).
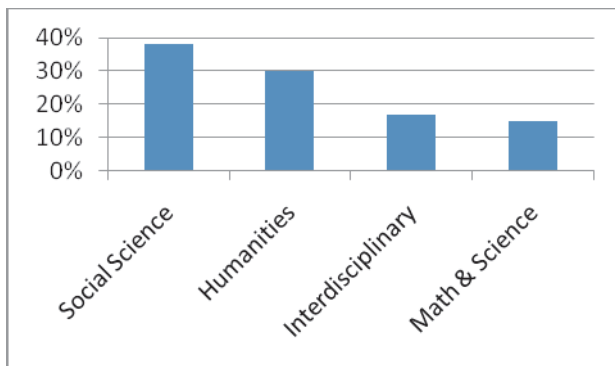
2.40 A Pareto chart shows the number of adoptions from the various countries. China is the mode but variation is high (since many adoptions are also from other countries). You could have instead used a bar or pie chart for your answers.
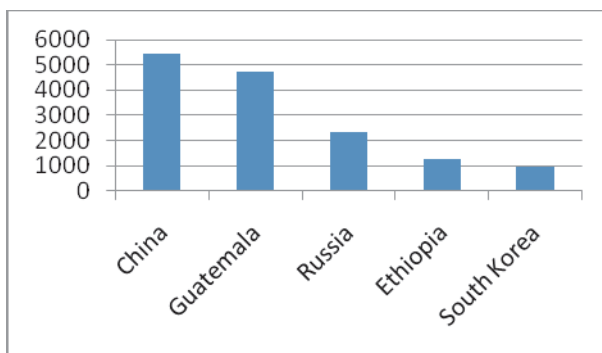

CHAPTER REVIEW EXERCISES

2.41 A histogram for each gender would be the most appropriate graph for the all the data since the data could be grouped into ranges (bins) based on the number of hours of TV watched. It would also be possible to use a dotplot or a stemplot for each gender.

2.42 A bar chart with bars representing the various job categories would be the best way to represent the data, since the data for both females and males could be given on the same graph (by differentiating the bars in some way). A pie chart could also be used to show the various job categories but you would need two pie charts, one for each gender.

2.43 Yes, because the study was randomized and double blind and involved treatment and control groups, it is possible to infer causation.
    a. Participants on HRT have higher rates of heart disease, stroke, pulmonary embolism, and breast cancer, but benefit from lower rates of hip fractures, as well as endometrial and colorectal cancers.
    b. Reporting the rates per 10,000 women allows comparisons to be made between the control and treatment groups, in case these groups have different numbers of women in them.

2.44   **a.** The data are likely reported in rates per 100 persons because there were different numbers of men and women in the comparison groups.
    b. Males aged 44 or below are substantially more likely to require treatment in the ER. Men and women between the ages of 45 and 74 seek treatment in the ER in approximately equal rates, while women aged 75 or older require more visits to the ER.

2.45 The graph is deceptive because the size of the bars makes it appear that women on HRT are more than twice as likely to get breast cancer as those who took a placebo. In reality, the difference between breast cancer rates for the two groups is only about 8 women out of 10,000. Scaling the vertical axis so that it starts at 0 would correct this problem.

2.46 The histogram should be changed so that the label "More" is replaced with the maximum data value appears of 90. Additionally, the bars in the histogram should be widened so that they touch.

2.47   **a.** To represent a single characteristic (gender) for a group (the elementary school class), a pie chart should be used rather than a histogram.
    b. A histogram should be used to display the data rather than a pie chart because the number of boys in each of the 45 different classrooms likely varies. Each of the histogram's bars would represent the number of classrooms that contain a particular range in quantity of boys (e.g. one bar might represent how many classrooms contain 7 to 9 boys).

2.48   **a.** A pie chart rather than a histogram would be most appropriate to represent responses to a single question (left- or right-handed) for this group.
    b. A histogram should be used to display the data rather than a pie chart, because the number of right-handed children in each of the 45 different classrooms likely varies. Each of the histogram's bars would represent the number of classrooms that contain a particular range in quantity of right-handed children (e.g., one bar might represent how many classrooms contain 7 to 9 children who are right-handed).

2.49 Both histograms are fairly symmetric (bell-shaped), with similar spread (variability). The primary difference between the two is the location of their centers: it appears that the average temperature for the 1880-1905 data are about 56.7 degrees F, while for the 1980-2005 it is about 57.9 degrees F. Therefore, average global temperatures have increased by a bit more than 1 degree F over this period of time; this may be evidence of global warming.

2.50 Both histograms are left-skewed. However, the data for the top law schools has less variability (ranging from 80% to 100%), while the data for the bottom group of law schools has more variability (ranging from

60% to 100%). It is also clear that the average percentage employed after nine months is much higher for those who attended a top law school. For the top group it appears the average percentage employed is about 96%, for the bottom group the average percentage employed is closer to 85%.

2.51   Answers will vary. For example, one could create a set of hypothetical quiz scores for a class that in general performed badly on the quiz, in particular where a large portion of the class has similar low scores clustered near one another but a few students performed well. Here are 15 such scores (percentages): 48, 51, 60, 66, 92, 78, 57, 70, 63, 58, 83, 61, 60, 75, 68.
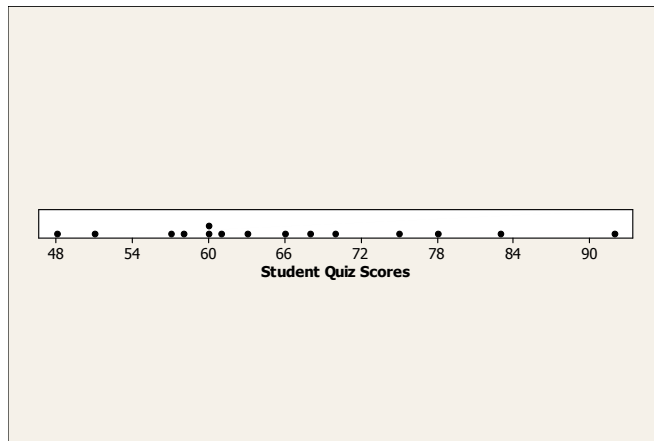

2.52   Answers will vary. For example, one could create a set of hypothetical quiz scores for a class that in general performed badly on the quiz, in particular where a large portion of the class has similar low scores clustered near one another but a few students performed even worse and a few others do well. Here are 15 such scores (percentages): 86, 71, 64, 66, 64, 78, 57, 70, 68, 78, 74, 61, 74, 72, 68.
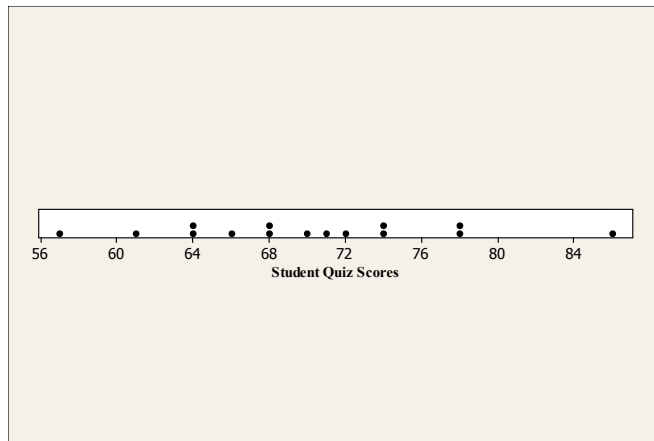

2.53   Dotplots, histograms, or 2-sided stemplots are all suitable to represent the data. The paired dotplot shows that the distribution of prices for West LA is bimodal (with modes near $500,000 and $1,000,000), while the distribution of prices for Midtown LA is much more symmetric. Furthermore, home prices in West LA exhibit significantly higher variability than those for Midtown LA.
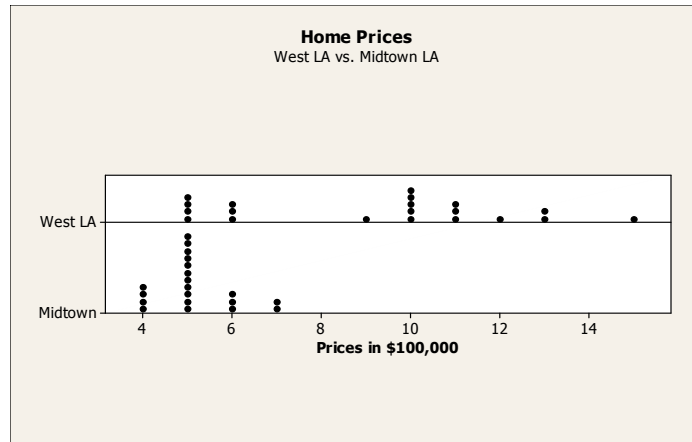
2.54   Either dotplots or histograms would be suitable to represent the data. A paired dotplot shows that the distribution of data obtained when a traffic camera is operating is skewed-right with a mode of 1, while the distribution of data obtained when no camera is present is symmetric with a higher mode. The "no camera" data exhibits higher variability.
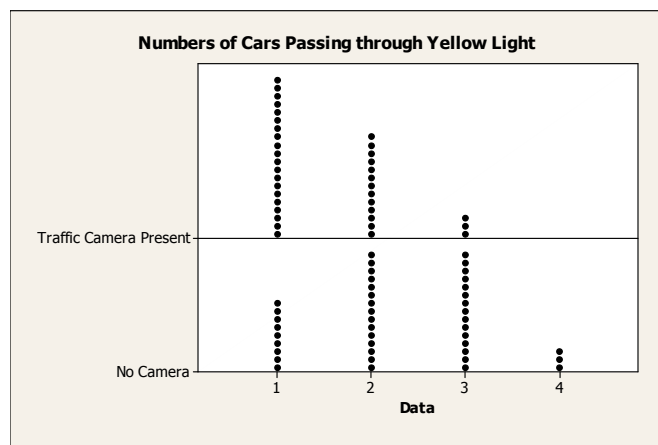
2.55   The distribution of data for the 100 companies with the highest net incomes will likely be right-skewed since there will be a few companies with extremely high net incomes. The histogram shows this to be true. Exxon Mobil is an outlier.
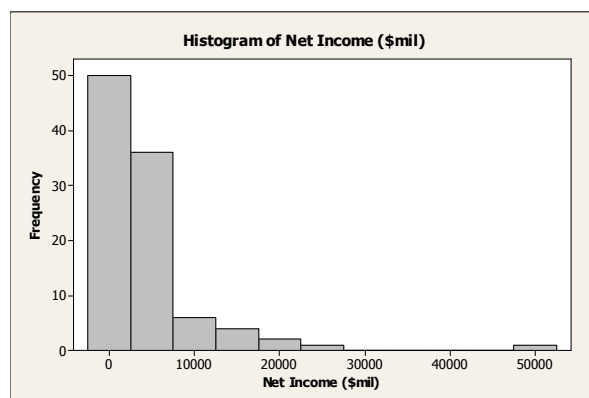
2.56 **a.** If both men's and women's weight data were combined, the resulting distribution would likely be bimodal since men's ideal weights should be higher than those of women.

   b. Histograms will vary depending on the software used. For the histogram shown, the bin width is 15 (pounds). The distribution is bimodal with modes corresponding to 120 lbs and 165 lbs.
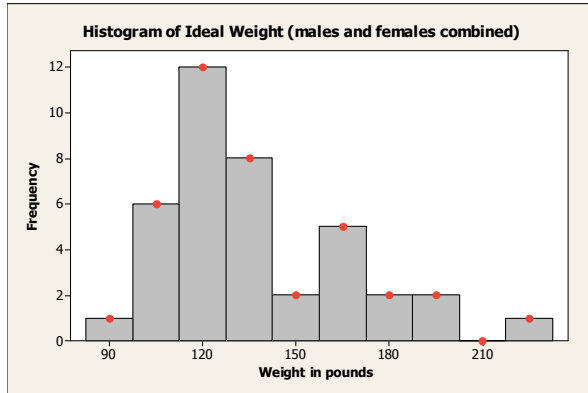
   c. Histograms will vary depending on the software used. For the histogram shown, the bin width is 20 (pounds). The distribution is right-skewed.