Section 2.1 Solutions

2.1 The total number is $169 + 193 = 362$, so we have $\hat{p} = 169/362 = 0.4669$. We see that 46.69% are female.

2.2 Since the total number is $43 + 319 = 362$, we have $\hat{p} = 43/362 = 0.1188$. We see that 11.88% percent of the students in the sample are smokers.

2.3 The total number is $94 + 195 + 35 + 36 = 360$ and the number who are juniors or seniors is $35 + 36 = 71$. We have $\hat{p} = 71/360 = 0.1972$. We see that 19.72% percent of the students who identified their class year are juniors or seniors.

2.4 The total number of students who reported SAT scores is 355, so we have $\hat{p} = 205/355 = 0.5775$. We see that 57.75% have higher math SAT scores.

2.5 Since this describes a proportion for all residents of the US, the proportion is for a population and the correct notation is p. We see that the proportion of US residents who are foreign born is $p = 0.124$.

2.6 The report describes the results of a sample, so the correct notation is $\hat{p}$. We see that the proportion of likely voters in the sample who believe a woman president is likely in the next 10 years is $\hat{p} = 0.73$.

2.7 The report describes the results of a sample, so the correct notation is $\hat{p}$. The proportion of US adults who believe the government does not provide enough support for soldiers returning from Iraq or Afghanistan is $\hat{p} = 931/1502 = 0.62$.

2.8 Information is provided for an entire population so we use the notation p for the proportion. The proportion is $p = 1,114,273/1,547,990 = 0.72$.

2.9 A relative frequency table is a table showing the proportion in each category. We see that the proportion preferring an Academy award is $31/362 = 0.086$, the proportion preferring a Nobel prize is $149/362 = 0.412$, and the proportion preferring an Olympic gold medal is $182/362 = 0.503$. These are summarized in the relative frequency table below. In this case, the relative frequencies actually add to 1.001 due to round-off error.

Response	Relative Frequency
Academy award	0.086
Nobel prize	0.412
Olympic gold medal	0.503
Total	1.00

2.10 A relative frequency table is a table showing the proportion in each category. In this case, the categories we are given are "No piercings", "One or two piercings", and "More than two piercings". The relative frequency with no piercings is $188/361 = 0.521$, the relative frequency for one or two piercings is $82/361 = 0.227$. The total has to add to 361, so there are $361 - 188 - 82 = 91$ students with more than two piercings, and the relative frequency is $91/361 = 0.252$. These are summarized in the relative frequency table below.

Response	Relative Frequency
No piercings	0.521
One or two piercings	0.227
More than two piercings	0.252
Total	1.00

2.11 One possible table is shown below. You might also choose to include the totals both down and across. It is also perfectly correct to switch the rows and columns.

	1	2	3
A	3	1	8
B	4	3	1

2.12 One possible table is shown below. You might also choose to include the totals both down and across. It is also perfectly correct to switch the rows and columns.

	1	2	3
A	2	6	4
B	5	2	11

2.13 (a) The sample is the 119 players who were observed. The population is all people who play rock-paper-scissors. The variable records which of the three options each player plays. This is a categorical variable.

(b) A relative frequency table is shown below. We see that rock is selected much more frequently than the others, and then paper, with scissors selected least often.

Option selected	Relative frequency
Rock	0.555
Paper	0.328
Scissors	0.118
Total	1.0

(c) Since rock is selected most often, your best bet is to play paper.

(d) Your opponent is likely to play paper again, so you should play scissors.

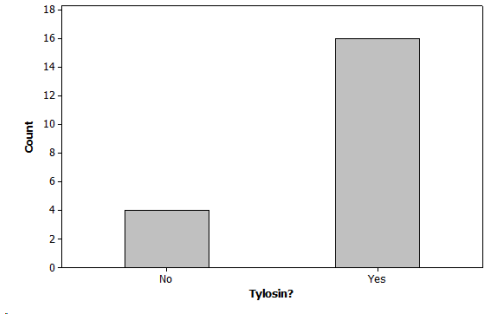
2.14 Since the dataset includes all professional soccer games, this is a population. The cases are soccer games and there are approximately 66,000 of them. The variable is whether or not the home team won the game; it is categorical. The relevant statistic is $p = 0.624$.

2.15 (a) The variable records whether or not tylosin appears in the dust samples. The individual cases in the study are the 20 dust samples.

(b) Here is a frequency table for the presence or absence of tylosin in the dust samples.

Category	Frequency
Tylosin	16
No tylosin	4
Total	20

(c) A bar chart for the frequencies is shown below.



(d) The table below shows the relative frequencies for cases with and without tylosin.

Category	Relative frequency
Tylosin	0.80
No tylosin	0.20
Total	1.00

2.16 (a) The total is $n = 2253$ so we divide each of the frequencies by the total. See the table. Notice that the relative frequencies add to 1.0, as we expect.

Cell phone owned	Relative Frequency
Android Smartphone	0.203
iPhone Smartphone	0.194
Blackberry Smartphone	0.063
Cell phone but not a smartphone	0.410
No cell phone	0.130
Total	1.0

(b) We see that 13% do not own a cell phone, 41% own a cell phone but not a smartphone, and 46% own a smartphone.

2.17 (a) There are two variables, both categorical. One is whether or not the dog selected the cancer sample and the other is whether or not the test was a breath test or a stool test.

(b) We need to include all possible outcomes for each variable when we make a two way table. The result variable has two options (dog is correct or dog is not correct) and the type of test variable has two options (breath or stool). The two-way table below summarizes these data.

	Breath test	Stool test	Total
Dog selects cancer	33	37	70
Dog does not select cancer	3	1	4
Total	36	38	74

(c) The dog got $33/36 = 0.917$ or 91.7% of the breath samples correct and $37/38 = 0.974$ or 97.4% of the stool samples correct.

(d) The dog got 70 tests correct and 37 of those were stool samples, so $37/70 = 0.529$ of the tests the dog got correct were stool samples.

2.18 (a) The table is given.

	HS or less	Some college	College grad	Total
Agree	363	176	196	735
Disagree	557	466	789	1812
Don't know	20	26	32	78
Total	940	668	1017	2625

(b) For the survey participants with a high school degree or less, we see that $363/940 = 0.386$ or 38.6% agree. For those with some college, the proportion is $176/668 = 0.263$, or 26.3% agree, and for those with a college degree, the proportion is $196/1017 = 0.193$, or 19.3% agree. There appears to be an association, and it seems that as education level goes up, the proportion who agree that every person has one true love goes down.

(c) We see that $1017/2625 = 0.387$, or 38.7% of the survey responders have a college degree or higher.

(d) A total of 1812 people disagreed and 557 of those have a high school degree or less, so we have $557/1812 = 0.307$, or 30.7% of the people who disagree have a high school degree or less.

2.19 (a) We compute the percentage of smokers in the female column and in the male column. For females, we see that $16/169 = 0.095$, so 9.5% of the females in the sample classify themselves as smokers. For males, we see that $27/193 = 0.140$, so 14% of the males in the sample classify themselves as smokers. In this sample, a larger percentage of males are smokers.

(b) For the entire sample, the proportion of smokers is $43/362 = 0.119$, or 11.9%.

(c) There are 43 smokers in the sample and 16 of them are female, so the proportion of smokers who are female is $16/43 = 0.372$, or 37.2%.

2.20 (a) This is an observational study since the researchers are observing the results after the fact and are not manipulating the gene directly to force a disruption. There are two variables: whether or not the person has dyslexia and whether or not the person has the DYXC1 break.

(b) Since $109 + 195 = 304$ people participated in the study, there will be 304 rows. Since there are two variables, there will be 2 columns: one for dyslexia or not and one for gene break or not.

(c) A two-way table showing the two groups and gene status is shown.

	Gene break	No break	Total
Dyslexia group	10	99	109
Control group	5	190	195
Total	15	289	304

(d) We look at each row (Dyslexia and Control) individually. For the dyslexia group, the proportion with the gene break is $10/109 = 0.092$. For the control group, the proportion with the gene break is $5/195 = 0.026$.

(e) There is a very substantial difference between the two proportions in part (d), so there appears to be an association between this particular genetic marker and dyslexia for the people in this sample. (As mentioned, we see in Chapter 4 how to determine whether we can generalize this result to the entire population.)

(f) We cannot assume a cause-and-effect relationship because this data comes from an observational study, not an experiment. There may be many confounding variables.

2.21 The two-way table with row and column totals is shown.

	Near-death experience	No such experience	Total
Cardiac arrest	11	105	116
Other cardiac problem	16	1463	1479
Total	27	1568	1595

To compare the two groups, we compute the percent of each group that had a near-death experience. For the cardiac arrest patients, the percent is $11/116 = 0.095 = 9.5\%$. For the patients with other cardiac problems, the percent is $16/1479 = 0.011 = 1.1\%$. We see that approximately 9.5% of the cardiac arrest patients reported a near-death experience, which appears to be much higher than the 1.1% of the other patients reporting this.

2.22 (a) The percent of pregnancies ending in miscarriage is $145/1009 = 14.4\%$.

(b) For each category, we compute the percent ending in miscarriage:

$$\begin{aligned} \text{Aspirin: Percent} &= \frac{5}{22} = 22.7\% \\ \text{Ibuprofen: Percent} &= \frac{13}{53} = 24.5\% \\ \text{Acetaminophen: Percent} &= \frac{24}{172} = 14.0\% \\ \text{No painkiller: Percent} &= \frac{103}{762} = 13.5\% \end{aligned}$$

The percent ending in miscarriage seems to be higher for those women who used aspirin or ibuprofen. Acetaminophen does not seem to pose a greater risk of miscarrying.

(c) This is an observational study. There are many possible confounding variables, including the fact that women who take painkillers may have other characteristics that are different from women who are not taking painkillers. It might be any of these other characteristics that is related to the increased proportion of miscarriages.

(d) We see that $22 + 53 = 75$ of the women took NSAIDs and $5 + 13 = 18$ of them miscarried, so we have

$$\text{NSAIDs: Percent} = \frac{5+13}{22+53} = \frac{18}{75} = 24.0\%$$

The percent of miscarriages is higher for women taking NSAIDs than it is for women who did not use painkillers. The use of acetominophen does not appear to significantly increase the risk. Pregnant women who do not want to miscarry might want to avoid taking NSAIDs.

(e) The original table in the exercise is not a two-way table since it does not list all outcomes for each of the variables. A two-way table (showing both "Miscarriage" and "No miscarriage") is given below.

	Miscarriage	No miscarriage	Total
Aspirin	5	17	22
Ibuprofen	13	40	53
Acetaminophen	24	148	172
No painkiller	103	659	762
Total	145	864	1009

(f) We have

$$\hat{p} = \frac{103}{145} = 71.0\%.$$

Notice that although certain painkillers appear to increase the risk of a miscarriage, it is still true that within this sample 71% of all miscarriages happened to women who did not use any painkiller.

2.23 (a) This is an experiment. Participants were actively assigned to receive either electrical stimulation or sham stimulation.

(b) The study appears to be single-blind, since it explicitly states that participants did not know which group they were in. It is not clear from the description whether the study was double-blind.

(c) There are two variables. One is whether or not the participants solved the problem and the other is which treatment (electrical stimulation or sham stimulation) the participants received. Both are categorical.

(d) Since the groups are equally split, there are 20 participants in each group. We know that 20% of the control group solved the problem, and 20% of 20 is $0.20(20) = 4$ so 4 solved the problem and 16 did not. Similarly, in the electrical stimulation group, $0.6(20) = 12$ solved the problem and 8 did not. See the table.

Treatment	Solved	Not solved
Sham	4	16
Electrical	12	8

(e) We see that $4 + 12 = 16$ people correctly solved the problem, and 12 of the 16 were in the electrical stimulation group, so the answer is $12/16 = 0.75$. We see that 75% of the people who correctly solved the problem had the electrical stimulation.

(f) We have $\hat{p}_E = 0.60$ and $\hat{p}_S = 0.20$ so the difference in proportions is $\hat{p}_E - \hat{p}_S = 0.60 - 0.20 = 0.40$.

(g) The proportions who correctly solved the problem are quite different between the two groups, so electrical stimulation does seem to help people gain insight on a new problem type.

2.24 Since these are population proportions, we use the notation p. We use p_H to represent the proportion of high school graduates unemployed and p_C to represent the proportion of college graduates (with a bachelor's degree) unemployed. (You might choose to use different subscripts, which is fine.) The difference in proportions is $p_H - p_C = 0.097 - 0.052 = 0.045$.

2.25 (a) Since no one assigned smoking or not to the participants, this is an observational study. Because this is an observational study, we can not use this data to determine whether smoking influences one's ability to get pregnant. We can only determine whether there is an association between smoking and ability to get pregnant.

(b) The sample collected is on women who went off birth control in order to become pregnant, so the population of interest is women who have gone off birth control in an attempt to become pregnant.

(c) We look in the total section of our two way table to find that out of the 678 women attempting to become pregnant, 244 succeeded in their first cycle, so $\hat{p} = 244/678 = 0.36$. For smokers we look only in the *Smoker* column of the two way table and observe 38 of 135 succeeded, so $\hat{p}_s = 38/135 = 0.28$. For non-smokers we look only in the *Non-smoker* column of the two way table and observe 206 of 543 succeeded, so $\hat{p}_{ns} = 206/543 = 0.38$.

(d) For the difference in proportions, we have $\hat{p}_{ns} - \hat{p}_s = 0.38 - 0.28 = 0.10$. This means that in this sample, the percent of non-smoking women successfully getting pregnant in the first cycle is 10 percentage points higher than the percent of smokers.

2.26 (a) The total number of respondents is 27,255 and the number in an abusive relationship is 2627, so the proportion is $2627/27255 = 0.096$. We see that about 9.6% of respondents have been in an emotionally abusive relationship in the last 12 months.

(b) We see that 2627 have been in an abusive relationship and 593 of these are male, so the proportion is $593/2627 = 0.226$. About 22.6% of those in abusive relationships are male.

(c) There are 8945 males in the survey and 593 of them have been in an abusive relationship, so the proportion is $593/8945 = 0.066$. About 6.6% of male college students have been in an abusive relationship in the last 12 months.

(d) There are 18310 females in the survey and 2034 of them have been in an abusive relationship, so the proportion is $2034/18310 = 0.111$. About 11.1% of female college students have been in an abusive relationship in the last 12 months.

2.27 (a) The total number of respondents is 27,268 and the number answering zero is 18,712, so the proportion is $18712/27268 = 0.686$. We see that about 68.6% of respondents have not had five or more drinks in a single sitting at any time during the last two weeks.

(b) We see that 853 students answer five or more times and 495 of these are male, so the proportion is $495/853 = 0.580$. About 58% of those reporting that they drank five or more alcoholic drinks at least five times in the last two weeks are male.

(c) There are 8,956 males in the survey and $912 + 495 = 1407$ of them report that they have had five or more alcoholic drinks at least three times, so the proportion is $1407/8956 = 0.157$. About 15.7% of male college students report having five or more alcoholic drinks at least three times in the last two weeks.

(d) There are 18,312 females in the survey and $966 + 358 = 1324$ of them report that they have had five or more alcoholic drinks at least three times, so the proportion is $1324/18312 = 0.072$. About 7.2% of female college students report having five or more alcoholic drinks at least three times in the last two weeks.

2.28 (a) We see in part (a) of the figure that both males and females are most likely to say that they had no drinks of alcohol the last time they socialized.

(b) We see in part (b) of the figure that both males and females are most likely to say that a typical student at their school would have 5 to 6 drinks the last time they socialized.

(c) No, perception does not match reality. Students believe that students at their school drink far more than they really do. Heavy drinkers tend to get noticed and skew student perceptions. When asked about a typical student and alcohol, students are much more likely to think of the heavy drinkers they know and not the non-drinkers.

2.29 (a) More females answered the survey since we see in graph (a) that the bar is much taller for females.

(b) It appears to be close to equal numbers saying they had no stress, since the height of the brown bars in graph (a) are similar. Graph (a) is the appropriate graph here since we are being asked about actual numbers not proportions.

(c) In this case, we are being asked about percents, so we use the relative frequencies in graph (b). We see in graph (b) that a greater percent of males said they had no stress.

(d) We are being asked about percents, so we use the relative frequencies in graph (b). We see in graph (b) that a greater percent of females said that stress had negatively affected their grades.

2.30 (a) Table where the vaccine has no effect (10% infected in both groups)

	Vaccine	No vaccine	Total
Malaria	20	30	50
No malaria	180	270	450
Total	200	300	500

(b) Table where the vaccine cuts the infection rate in half (from 10% to 5%).

	Vaccine	No vaccine	Total
Malaria	10	30	40
No malaria	190	270	460
Total	200	300	500

2.31 Graph (b) is the impostor. It shows more parochial students than private school students. The other three graphs have more private school students than parochial.

Section 2.2 Solutions

2.32 Histograms A and H are both skewed to the left.

2.33 Only histogram F is skewed to the right.

2.34 Histograms B, C, D, E, and G are all approximately symmetric.

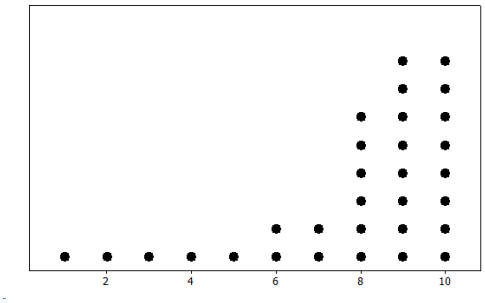
2.35 While all of B,C,D,E and G are approximately symmetric, only B,C and E are also bell shaped.

2.36 Histogram A is skewed left, so the mean should be smaller then the median. The other three histograms (B,C, and D) are approximately symmetric so the mean and median will be approximately equal.

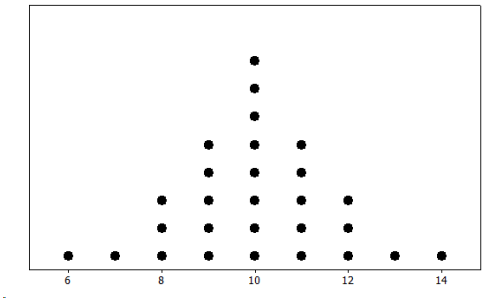
2.37 Histograms E and G are both approximately symmetric, so the mean and median will be approximately equal. Histogram F is skewed right, so the mean should be larger then the median; while histogram H is skewed left, so the mean should be smaller then the median.

2.38 Histogram C appears to have a mean close to 150, so it has the largest mean. Histogram H appears to have a mean around -2 or -3, so it has the smallest mean.

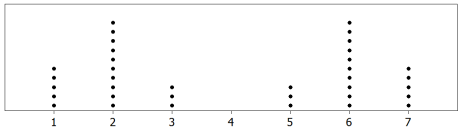
2.39 There are many possible dotplots we could draw that would be clearly skewed to the left. One is shown.



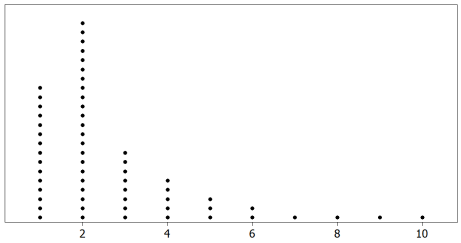
2.40 There are many possible dotplots we could draw that are approximately symmetric and bell-shaped. One is shown.



2.41 There are many possible dotplots we could draw that are approximately symmetric but not bell-shaped. One is shown.



2.42 There are many possible dotplots we could draw that are clearly skewed to the right. One is shown.



2.43 (a) We have $\overline{x} = (8 + 12 + 3 + 18 + 15)/5 = 11.2$.

(b) The median is the middle number when the numbers are put in order smallest to largest. In order, we have:

$$3 \quad 8 \quad 12 \quad 15 \quad 18.$$

The median is $m = 12$. Notice that there are two data values less than the median and two data values greater.

(c) There do not appear to be any particularly large or small data values relative to the rest, so there do not appear to be any outliers.

2.44 (a) We have $\overline{x} = (41 + 53 + 38 + 32 + 115 + 47 + 50)/7 = 53.714$.

(b) The median is the middle number when the numbers are put in order smallest to largest. In order, we have:

$$32 \quad 38 \quad 41 \quad 47 \quad 50 \quad 53 \quad 115.$$

The median is $m = 47$. Notice that there are three data values less than the median and three data values greater.

(c) The value 115 is significantly larger than all the other data values, so 115 is a likely outlier.

2.45 (a) We have $\overline{x} = (15 + 22 + 12 + 28 + 58 + 18 + 25 + 18)/8 = 24.5$.

(b) Since there are $n = 8$ values, the median is the average of the two middle numbers when the numbers are put in order smallest to largest. In order, we have:

$$12 \quad 15 \quad 18 \quad 18 \quad 22 \quad 25 \quad 28 \quad 58.$$

The median is the average of 18 and 22, so $m = 20$. Notice that there are four data values less than the median and four data values greater.

(c) The value 58 is significantly larger than all the other data values, so 58 is a likely outlier.

2.46 (a) We have $\overline{x} = (110 + 112 + 118 + 119 + 122 + 125 + 129 + 135 + 138 + 140)/10 = 124.8$.

(b) The data values are already in order smallest to largest. Since there are $n = 10$ values, the median is the average of the two middle numbers 122 and 125, so we have $m = (122 + 125)/2 = 123.5$. Notice that there are five data values less than the median and five data values greater.

(c) There do not appear to be any particularly large or small data values relative to the rest, so there do not appear to be any outliers.

2.47 This is a sample, so the correct notation is $\overline{x} = 2386$ calories per day.

2.48 This is a sample, so the correct notation is $\overline{x} = 60$ texts per day.

2.49 This is a population, so the correct notation is $\mu = 41.5$ yards per punt.

2.50 This is a population, so the correct notation is $\mu = 2.6$ television sets per household.

2.51 (a) We expect the mean to be larger since there appears to be a relatively large outlier (26.0) in the data values.

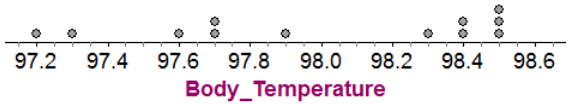
(b) There are eight numbers in the data set, so the mean is the sum of the values divided by 8. We have:

$$\text{Mean} = \frac{0.8 + 1.9 + 2.7 + 3.4 + 3.9 + 7.1 + 11.9 + 26.0}{8} = \frac{57.7}{8} = 7.2 \text{ mg/kg}.$$

The data values are already in order smallest to largest, and the median is the average of the two middle numbers. We have:

$$\text{Median} = \frac{3.4 + 3.9}{2} = 3.65.$$

2.52 (a) A dotplot of the body temperatures is shown below.



(b) We compute $\overline{x} = 98.0°\text{F}$. It is the balance point in the dotplot.

(c) There are $n = 12$ data values, so the median is the average of the two middle values. We have

$$m = \frac{97.9 + 98.3}{2} = 98.1°\text{F}.$$

This is a point in the dotplot that has six dots on either side.

2.53 (a) Since there are only 50 states and all of them are represented, this is the entire population.

(b) The distribution is skewed to the right. There appears to be an outlier at about 35 million. (The outlier represents the state of California.)

(c) The median splits the data in half and appears to be about 4 million. (In fact, it is exactly 4.170 million.)

(d) The mean is the balance point for the histogram and is harder to estimate. It appears to be about 6 million. (In fact, it is exactly 5.862 million.)

2.54 Since most insects have small weights, the frequency counts for small weights will be large and the frequency counts for larger weights will be quite small, so we expect the histogram to be skewed to the right. The mean will be larger since the outlier of 71 will pull the mean up.

2.55 (a) The distribution is skewed to the left.

(b) The median is the value with half the area to the left and half to the right. The value 5 has way more area on the right so it cannot be correct. If we draw a line at 7, there is more area to the left than the right. The answer must be between 5 and 7 and a line at 6.5 appears to split the area into approximately equal amounts. The median is about 6.5.

(c) Because the data is skewed to the left, the values in the longer tail on the left will pull the mean down. The mean will be smaller than the median.

2.56 (a) The distribution is skewed to the left since there are many values between about 72 and 82 and then a long tail going down to the outliers on the left.

(b) Since half the values are above 72, the median is about 72. (The actual median is 71.9.)

(c) Since the data is skewed to the left, the mean will be less than the median so the mean will be less than 72. (It is actually about 68.9.)

2.57 (a) We compute $\bar{x} = 26.6$. Since there are ten numbers, we average the two middle numbers to find the median. We have $m = (15 + 17)/2 = 16$.

(b) Without the outlier, we have $\bar{x} = 16.78$. Since $n = 9$, the median is the middle number. We have $m = 15$.

(c) The outlier has a very significant effect on the mean and very little effect on the median.

2.58 (a) The distribution has a right skew. There are a number of apparent outliers on the right side.

(b) The actual median is 140 ng/ml. Estimates between 120 and 160 are reasonable.

(c) The actual mean is 189.9 ng/ml. Estimates between 160 and 220 are reasonable. Note that the outliers and right skew should make the mean larger than the median.

2.59 (a) Many people send just a few text messages per day, so many of the data values will be relatively small. However, some people send and receive a very large number of text messages per day. We expect that the bulk of the data values will be small numbers, with a tail extending out to the right to some values that are very large. This describes a distribution that is skewed to the right.

(b) The people with a very large number of text messages will pull up the mean but not the median, so we expect the mean to be 39.1 text messages and the median to be 10 messages. The fact that the mean and the median are so different indicates that the data is significantly skewed to the right, and almost certainly has some high outliers. In addition, the median of 10 messages implies that half the people averaged less than 10 text messages a day and half averaged more.

2.60 (a) We have $\overline{x}_f = 6.40$.

(b) We have $\overline{x}_m = 6.81$.

(c) We see that $\overline{x}_m - \overline{x}_f = 6.81 - 6.40 = 0.41$. In this sample, the males, on average, spent 0.41 more hours per week exercising than the females.

2.61 The notation for a median is m. We use m_H to represent the median earnings for high school graduates and m_C to represent the median earnings for college graduates. (You might choose to use different subscripts, which is fine.) The difference in medians is $m_H - m_C = 626 - 1025 = -399$. College graduates earn about \$400 more per week than high school graduates.

2.62 (a) This is an experiment since the treatment was randomly assigned and imposed.

(b) The cases are the 24 fruit flies. There are two variables. The explanatory variable is which of the two groups the fly is in. The response variable is percent of time the alcoholic mixture is selected.

(c) Using $\overline{x}_R$ for the mean of the rejected group and $\overline{x}_M$ for the mean for the mated group, we have $\overline{x}_R - \overline{x}_M = 0.73 - 0.47 = 0.26$.

(d) Yes, since this was a randomized experiment.

2.63 (a) There are many possible answers. One way to force the outcome is to have a very small outlier, such as
2, 51, 52, 53, 54.
The median of these 5 numbers is 52 while the mean is 42.4.

(b) There are many possible answers. One way to force the outcome is to have a very large outlier, such as
2, 3, 4, 5, 200.
The median of these 5 numbers is 4 while the mean is 42.8.

(c) There are many possible answers. One option is the following:
2, 3, 4, 5, 6.
Both the mean and the median are 4.

2.64 There are many possible answers. Any variable that measures something with large outliers will work.

2.65 The values are in order smallest to largest, and since more than half the values are 1, the median is 1. We calculate the mean to be $\overline{x} = 3.2$. In this case, the mean is probably a better value (despite the fact that 12 might be an outlier) since it allows us to see that some of the data values are above 1.

2.66 (a) It appears that the mean of the married women is higher than the mean of the never married women. We expect that the mean and the median will be the most different for the never married women, since that data is quite skewed while the married data is more symmetric.

(b) We have $n = 1000$ in each case. For the married women, we see that 162 women had 0 children, 190 had 1 child, and 290 had 2 children, so $162 + 190 + 290 = 642$ had 0, 1, or 2 children. Less than half the women had 0 or 1 child and more than half the women had 0, 1, or 2 children so the median is 2. For the never married women, more than half the women had 0 children, so the median is 0.

Section 2.3 Solutions

2.67 (a) Using technology, we see that the mean is $\overline{x} = 17.36$ with a standard deviation of $s = 5.73$.

(b) Using technology, we see that the five number summary is $(10, 13, 17, 21, 28)$. Notice that these five numbers divide the data into fourths.

2.68 (a) Using technology, we see that the mean is $\overline{x} = 15.09$ with a standard deviation of $s = 13.30$.

(b) Using technology, we see that the five number summary is $(1, 4, 10, 25, 42)$. Notice that these five numbers divide the data into fourths.

2.69 (a) Using technology, we see that the mean is $\overline{x} = 10.4$ with a standard deviation of $s = 5.32$.

(b) Using technology, we see that the five number summary is $(4, 5, 11, 14, 22)$. Notice that these five numbers divide the data into fourths.

2.70 (a) Using technology, we see that the mean is $\overline{x} = 59.73$ with a standard deviation of $s = 17.89$.

(b) Using technology, we see that the five number summary is $(25, 43, 64, 75, 80)$. Notice that these five numbers divide the data into fourths.

2.71 (a) Using technology, we see that the mean is $\overline{x} = 9.05$ hours per week with a standard deviation of $s = 5.74$.

(b) Using technology, we see that the five number summary is $(0, 5, 8, 12, 40)$. Notice that these five numbers divide the data into fourths.

2.72 (a) Using technology, we see that the mean is $\overline{x} = 6.50$ hour per week with a standard deviation of $s = 5.58$.

(b) Using technology, we see that the five number summary is $(0, 3, 5, 9.5, 40)$. Notice that these five numbers divide the data into fourths.

2.73 We know that the standard deviation is a measure of how spread out the data are, so larger standard deviations go with more spread out data. All of these histograms are centered at 10 and have the same horizontal scale, so we need only look at the spread. We see that $s = 1$ goes with Histogram B and $s = 3$ goes with Histogram C and $s = 5$ goes with Histogram A.

2.74 Remember that a standard deviation is an approximate measure of the average distance of the data from the mean. Be sure to pay close attention to the scale on the horizontal axis for each histogram.

(a) V

(b) III

(c) IV

(d) I

(e) VI

(f) II

2.75 Remember that the five number summary divides the data (and hence the area in the histogram) into fourths.

(a) II

(b) V

(c) IV

(d) I

(e) III

(f) VI

2.76 Remember that the five number summary divides the data (and hence the area in the histogram) into fourths.

(a) This shows a distribution pretty evenly spread out across the numbers 1 through 9, so this five number summary matches histogram W.

(b) This shows a distribution that is more closely bunched in the center, since 50% of the data is between 4 and 6. This five number summary matches histogram X.

(c) Since the top 50% of the data is between 7 and 9, this data is left skewed and matches histogram Y.

(d) Since both the minimum and the first quartile are 1, there is at least 25% of the data at 1, so this five number summary matches histogram Z.

2.77 The mean appears to be at about $\overline{x} \approx 500$. Since 95% of the data appear to be between about 460 and 540, we see that two standard deviations is about 40 so one standard deviation is about 20. We estimate the standard deviation to be between 20 and 25.

2.78 The 10^{th}-percentile is the value with 10% of the data values below it, so a reasonable estimate would be between 460 and 470. The 90^{th}-percentile is the value with about 10% of the values above it, so a reasonable estimate would be between 530 and 540.

2.79 The minimum appears to be at 440, the median at 500, and the maximum at 560. The quartiles are a bit harder to estimate accurately. It appears that the lower quartile is about 485 and the upper quartile is about 515, so the five number summary is approximately $(440, 485, 500, 515, 560)$.

2.80 The mean appears to be about 68. Since the data is relatively bell-shaped, we can estimate the standard deviation using the 95% rule. Since there are 100 dots in the dotplot, we want to find the boundaries with 2 or 3 dots more extreme on either side. This gives boundaries from 59 to 76, which is 8 or 9 units above and below the mean. We estimate the standard deviation to be about 4.5.

2.81 Since there are exactly $n = 100$ data points, the 10^{th}-percentile is the value with 10 dots to the left of it. We see that this is at the value 62. Similarly, the 90^{th}-percentile is the value with 10 dots to the right of it. This is the value 73.

2.82 We see that the minimum value is 58 and the maximum is 77. We can count the dots to find the value at the 25^{th}-percentile, the 50^{th}-percentile, and the 75^{th}-percentile to find the quartiles and the median. We see that $Q_1 = 65$, the median is at 68, and $Q_3 = 70$. The five number summary is (58, 65, 68, 70, 77).

2.83 This dataset is very symmetric.

2.84 For this dataset, half of the values are clustered between 100 and 115, and the other half are very spread out to the right between 115 and 220. This distribution is skewed to the right.

2.85 For this dataset, half of the values are clustered between 22 and 27, and the other half are spread out to the left all the way down to 0. This distribution is skewed to the left.

2.86 This data appears to be quite symmetric about the median of 36.3.

2.87 We have

$$Z\text{-score} = \frac{\text{Value} - \text{Mean}}{\text{Standard deviation}} = \frac{243 - 200}{25} = 1.72.$$

This value is 1.72 standard deviations above the mean.

2.88 We have

$$Z\text{-score} = \frac{\text{Value} - \text{Mean}}{\text{Standard deviation}} = \frac{88 - 96}{10} = -0.8.$$

This value is 0.80 standard deviations below the mean, which is likely to be relatively near the center of the distribution.

2.89 We have

$$Z\text{-score} = \frac{\text{Value} - \text{Mean}}{\text{Standard deviation}} = \frac{5.2 - 12}{2.3} = -2.96.$$

This value is 2.96 standard deviations below the mean, which is quite extreme in the lower tail.

2.90 We have

$$Z\text{-score} = \frac{\text{Value} - \text{Mean}}{\text{Standard deviation}} = \frac{8.1 - 5}{2} = 1.55.$$

This value is 1.55 standard deviations above the mean.

2.91 The 95% rule says that 95% of the data should be within two standard deviations of the mean, so the interval is:

$$\begin{array}{rcl} \text{Mean} & \pm & 2 \cdot \text{StDev} \\ 200 & \pm & 2 \cdot (25) \\ 200 & \pm & 50 \\ 150 & \text{to} & 250. \end{array}$$

We expect 95% of the data to be between 150 and 250.

2.92 The 95% rule says that 95% of the data should be within two standard deviations of the mean, so the interval is:

$$\begin{array}{rcl} \text{Mean} & \pm & 2 \cdot \text{StDev} \\ 10 & \pm & 2 \cdot (3) \\ 10 & \pm & 6 \\ 4 & \text{to} & 16. \end{array}$$

We expect 95% of the data to be between 4 and 16.

2.93 The 95% rule says that 95% of the data should be within two standard deviations of the mean, so the interval is:

$$\begin{array}{rcl} \text{Mean} & \pm & 2 \cdot \text{StDev} \\ 1000 & \pm & 2 \cdot (10) \\ 1000 & \pm & 20 \\ 980 & \text{to} & 1020. \end{array}$$

We expect 95% of the data to be between 980 and 1020.

2.94 The 95% rule says that 95% of the data should be within two standard deviations of the mean, so the interval is:

$$\begin{array}{rcl} \text{Mean} & \pm & 2 \cdot \text{StDev} \\ 1500 & \pm & 2 \cdot (300) \\ 1500 & \pm & 600 \\ 900 & \text{to} & 2100. \end{array}$$

We expect 95% of the data to be between 900 and 2100.

2.95 (a) The numbers range from 46 to 61 and seem to be grouped around 53. We estimate that $\overline{x} \approx 53$.

(b) The standard deviation is roughly the typical distance of a data value from the mean. All of the data values are within 8 units of the estimated mean of 53, so the standard deviation is definitely not 52, 10, or 55. A typical distance from the mean is clearly greater than 1, so we estimate that $s \approx 5$.

(c) Using a calculator or computer, we see $\overline{x} = 52.875$ and $s = 5.07$.

2.96 (a) We see in the computer output that the mean obesity rate is $\mu = 24.552\%$ and the standard deviation is $\sigma = 3.044\%$.

(b) We see that the largest value is 30.9, so we compute the z-score as:

$$z\text{-score} = \frac{x - \mu}{\sigma} = \frac{30.9 - 24.552}{3.044} = 2.085.$$

The maximum of 30.9% obese (which comes from the state of Mississippi) is slightly more than two standard deviations above the mean.
We compute the z-score for the smallest percent obese, 17.8%, similarly:

$$z\text{-score} = \frac{x - \mu}{\sigma} = \frac{17.8 - 24.552}{3.044} = -2.218.$$

The minimum of 17.8% obese, from the state of Colorado, is about 2.2 standard deviations below the mean. Both the maximum and the minimum might be considered mild outliers.

(c) Since the distribution is relatively symmetric and bell-shaped, we expect that about 95% of the data will lie within two standard deviations of the mean. We have:

$$\mu - 2\sigma = 24.552 - 2(3.044) = 18.464 \qquad \text{and} \qquad \mu + 2\sigma = 24.552 + 2(3.044) = 30.640.$$

We expect about 95% of the data to lie between 18.464% and 30.640%. In fact, this general rule is very accurate in this case, since the percent of the population that is obese lies within this range for 47 of the 50 states, which is 94%. (The only states outside the range are Colorado, Mississippi, and Louisiana.)

2.97 (a) We see in the computer output that the five number summary is (17.800, 22.175, 24.400, 26.825, 30.900).

(b) The range is the difference between the maximum and the minimum, so we have Range $= 30.9 - 17.8 = 13.1$. The interquartile range is the difference between the first and third quartiles, so we have $IQR = 26.825 - 22.175 = 4.65$.

(c) The 15^{th}-percentile is between the minimum and the first quartile, so it will be between 17.8 and 22.175. The 60^{th}-percentile is between the median and the third quartile, so it will be between 24.4 and 26.825.

2.98 (a) We use technology to see that the mean is 56.10 and the standard deviation is 7.50.

(b) We see that 4 of the 10 values are larger than the mean. None of these four are in the early five years and all 4 of them are in the later five years. People seem to be getting better at eating hot dogs!

2.99 (a) See the table.

Year	Joey	Takeru	Difference
2009	68	64	4
2008	59	59	0
2007	66	63	3
2006	52	54	-2
2005	32	49	-17

(b) For the five differences, we use technology to see that the mean is -2.4 and the standard deviation is 8.5.

2.100 (a) We use technology to see that the mean is $\overline{x} = 85.25$ and the standard deviation is $s = 33.18$.

(b) The longest time is 153 days and the z-score for that is

$$Z\text{-score for 153 days } = \frac{x - \overline{x}}{s} = \frac{153 - 85.25}{33.18} = 2.04.$$

The shortest time is 40 days, with a z-score of

$$Z\text{-score for 40 days } = \frac{40 - \overline{x}}{s} = \frac{40 - 85.25}{33.18} = -1.36.$$

The longest time in the sample is slightly more than two standard deviations above the mean, while the shortest time is 1.36 standard deviations below the mean.

2.101 (a) We expect that 95% of the data will lie between $\overline{x} \pm 2s$. In this case, the mean is $\overline{x} = 2.31$ and the standard deviation is $s = 0.96$, so 95% of the data lies between $2.31 \pm 2(0.96)$. Since $2.31 - 2(0.96) = 0.39$ and $2.31 + 2(0.96) = 4.23$, we estimate that about 95% of the temperature increases will lie between 0.39° and 4.23°.

(b) Since $\overline{x} = 2.31$ and $s = 0.96$, the z-score for a temperature increase of 4.9° is

$$z\text{-score } = \frac{x - \overline{x}}{s} = \frac{4.9 - 2.31}{0.96} = 2.70.$$

The temperature increase for this man is 2.7 standard deviations above the mean.

2.102 (a) The 10^{th} percentile is the value with 10% of the area of the histogram to the left of it. This appears to be at about 2.5 or 2.6. A (self-reported) grade point average of about 2.6 has 10% of the reported values below it (and 90% above). The 75^{th} percentile appears to be at about 3.4. A grade point average of about 3.4 is greater than 75% of reported grade point averages.

(b) It appears that the highest GPA in the dataset is 4.0 and the lowest is 2.0, so the range is $4.0 - 2.0 = 2.0$.

2.103 (a) Using software, we see that $\overline{x} = 0.272$ and $s = 0.237$.

(b) The largest concentration is 0.851. The z-score is

$$z\text{-score } = \frac{x - \overline{x}}{s} = \frac{0.851 - 0.272}{0.237} = 2.44.$$

The largest value is almost two and a half standard deviations above the mean, and appears to be an outlier.

(c) Using software, we see that

$$\text{Five number summary } = (0.073, 0.118, 0.158, 0.358, 0.851).$$

(d) The range is $0.851 - 0.073 = 0.778$ and the interquartile range is $IQR = 0.358 - 0.118 = 0.240$.

2.104 (a) The data is heavily skewed and there appear to be some large outliers. It is most appropriate to use the five number summary.

(b) No, it is not appropriate to use that rule with this distribution. That rule is useful when data is symmetric and bell-shaped.

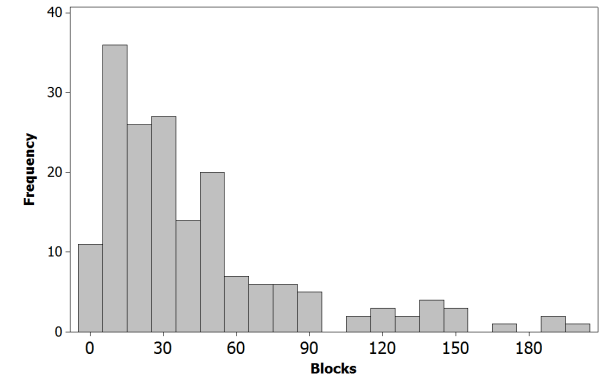
2.105 The data is not at all bell-shaped, so it is not appropriate to use the 95% rule with this data.

2.106 (a) Using technology, we see that the mean is 42.86 blocks in a season and the standard deviation is 41.35 blocks.

(b) Using technology, we see that the five number summary is $(3, 14, 30, 53.5, 198)$.

(c) The five number summary from part (b) is more resistant to outliers and is often more appropriate if the data is heavily skewed.

(d) We create either a histogram or dotplot or boxplot. A histogram of the data in *Blocks* is shown. We see that the distribution is heavily skewed to the right.



(e) This distribution is not at all bell-shaped, so it is not appropriate to use the 95% rule with this distribution.

2.107 We first calculate the z-scores for the four values. In each case, we use the fact that the z-score is the value minus the mean, divided by the standard deviation. We have

$$z\text{-score for } FGPct = \frac{0.510 - 0.464}{0.053} = 0.868 \qquad z\text{-score for } Points = \frac{2111 - 994}{414} = 2.698$$

$$z\text{-score for } Assists = \frac{554 - 220}{170} = 1.965 \qquad z\text{-score for } Steals = \frac{124 - 68.2}{31.5} = 1.771$$

The most impressive statistic is his total points, which is about 2.7 standard deviations above the mean. The least impressive is his field goal percentage, which is only 0.868 standard deviations above the mean. He is above the mean on all four, however, and substantially above the mean on three.

2.108 (a) We calculate z-scores using the summary statistics for each: Critical Reading $= \frac{600-497}{114} = 0.904$, Math $= \frac{600-514}{117} = 0.735$, Writing $= \frac{600-489}{113} = 0.982$.

(b) Stanley's most unusual score was in the Writing component, since he has the highest z-score in this section. His least unusual score was in Mathematics.

(c) Stanley performed best on Writing, since this is the highest z-score.

2.109 (a) Using technology, we find that the mean is $\overline{x} = 65.89$ percent with a fast connection and the standard deviation is $s = 18.29$.

(b) Using technology, we find that the mean is $\overline{x} = 26.18$ hours online and the standard deviation is $s = 3.41$.

(c) It is not clear if there is a relationship or not. The Swiss have the fastest connection times and the lowest time online and Brazil has the slowest connection time and the highest time online, but the pattern does not seem to be obvious with the other countries.

2.110 (a) The average for both joggers is 45, so they are the same.

(b) The averages are the same, but the set of times for jogger 1 has a much lower standard deviation.

2.111 (a) Using technology, we see that the mean is $\overline{x} = 13.15$ years with a standard deviation of $s = 7.24$ years.

(b) We have

$$\text{The } z\text{-score for the elephant } = \frac{\text{Elephant's value} - \text{Mean}}{\text{Standard deviation}} = \frac{40 - 13.15}{7.24} = 3.71.$$

The elephant is 3.71 standard deviations above the mean, which is way out in the upper tail of the distribution. The elephant is a strong outlier!

2.112 (a) The range is 6662-445 = 6217 and the interquartile range is $IQR = 2106 - 1334 = 772$.

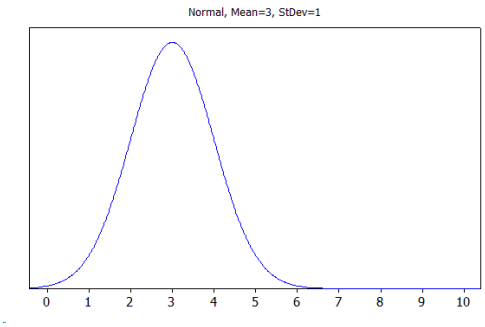
(b) The maximum of 6662 is clearly an outlier and we expect it to pull the mean above the median. Since the median is 1667, the mean should be larger than 1667, but not too much larger. The mean of this data set is 1796.

(c) The best estimate of the standard deviation is 680. We see from the five number summary that about 50% of the data is within roughly 400 of the median, so the standard deviation is definitely bigger than 200. The two values above 680 would be way too large to give an estimated distance of the data values from the mean, so the only reasonable answer is 680. The actual standard deviation is 680.3.

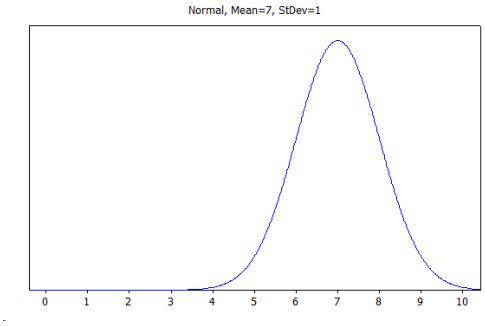
2.113 (a) The smallest possible standard deviation is zero, which is the case if the numbers don't deviate at all from the mean. The dataset is:
5, 5, 5, 5, 5, 5.

(b) The largest possible standard deviation occurs if all the numbers are as far as possible from the mean of 5. Since we are limited to numbers only between 1 and 9 (and since we have to keep the mean at 5), the best we can do is three values at 1 and three values at 9. The dataset is:
1, 1, 1, 9, 9, 9.

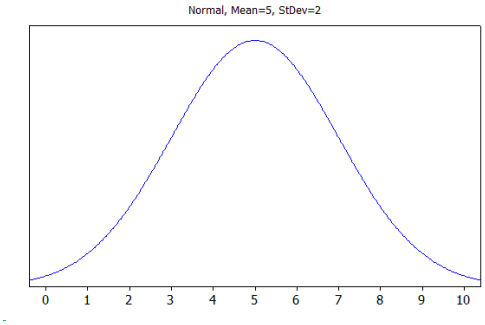
2.114 A bell-shaped distribution with mean 3 and standard deviation 1.



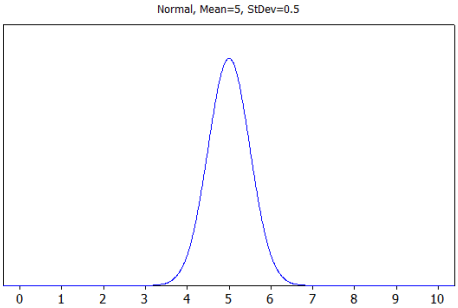
2.115 A bell-shaped distribution with mean 7 and standard deviation 1.



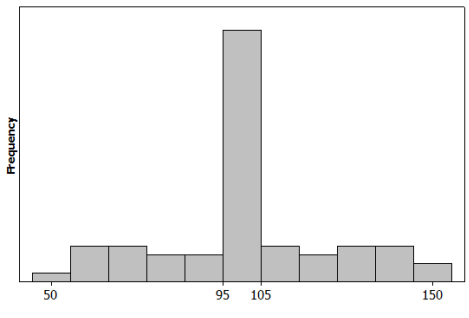
2.116 A bell-shaped distribution with mean 5 and standard deviation 2.



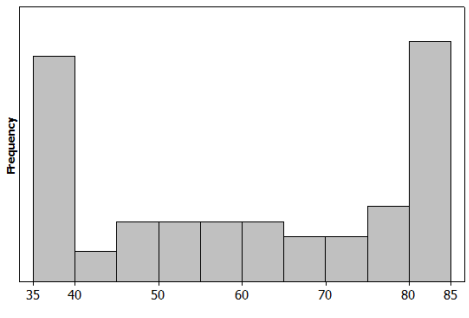
2.117 A bell-shaped distribution with mean 5 and standard deviation 0.5.



2.118 (a) One half of the data should have a range of 10 and all of the data should have a range of 100. The data is very bunched in the middle, with long tails on the sides. One possible histogram is shown.



(b) One half of the data should have a range of 40 and all of the data should have a range of 50. This is a bit tricky – it means the outside 50% of the data fits in only 10 units, so the data is actually clumped on the outside margins. One possible histogram is shown.



2.119 (a) The rough estimate is $(130 - 35)/5 = 19$ bpm compared to the actual standard deviation of s=12.2 bpm. The possible outliers pull the rough estimate up quite a bit. Without the two outliers, the rough estimate is $(96 - 35)/5 = 12.2$, exactly matching the actual standard deviation.

(b) The rough estimate is $(40-0)/5 = 8$. The rough estimate is a bit high compared to the actual standard deviation of $s = 5.741$.

(c) The rough estimate is $(40 - 1)/5 = 7.8$. The rough estimate is quite close to the actual standard deviation of $s = 7.24$.

Section 2.4 Solutions

2.120 We match the five number summary with the maximum, first quartile, median, third quartile, and maximum shown in the boxplot.

(a) This five number summary matches boxplot S.

(b) This five number summary matches boxplot R.

(c) This five number summary matches boxplot Q.

(d) This five number summary matches boxplot T. Notice that at least 25% of the data is exactly the number 12, since 12 is both the minimum and the first quartile.

2.121 We match the five number summary with the maximum, first quartile, median, third quartile, and maximum shown in the boxplot.

(a) This five number summary matches boxplot W.

(b) This five number summary matches boxplot X.

(c) This five number summary matches boxplot Y.

(d) This five number summary matches boxplot Z.

2.122 (a) Half of the data lies between 585 and 595, while the other half (the left tail) is stretched all the way from 585 down to about 50. This distribution is skewed to the left.

(b) There are 3 low outliers.

(c) We see that the median is at about 585 and the distribution is skewed left, so the mean is less than the median. A reasonable estimate for the mean is about 575 or 580.

2.123 (a) Half of the data appears to lie in the small area between 20 and 40, while the other half (the right tail) appears to extend all the way up from 40 to about 140. This distribution appears to be skewed to the right.

(b) Since there are no asterisks on the graph, there are no outliers.

(c) The median is approximately 40 and since the distribution is skewed to the right, we expect the values out in the right tail to pull the mean up above the median. A reasonable estimate for the mean is about 50.

2.124 (a) This distribution looks very symmetric.

(b) Since there are no asterisks on the graph, there are no outliers.

(c) We see that the median is at approximately 135. Since the distribution is symmetric, we expect the mean to be very close to the median, so we estimate the mean to be about 135.

2.125 (a) This distribution isn't perfectly symmetric but it is close. Despite the presence of outliers on both sides, there does not seem to be a distinct skew either way. This distribution is approximately symmetric.

(b) There appear to be 3 low outliers and 2 high outliers.

(c) Since the distribution is approximately symmetric, we expect the mean to be close to the median, which appears to be at about 1200. We estimate that the mean is about 1200.

2.126 (a) We see that $Q_1 = 260$ and $Q_3 = 300$ so the interquartile range is IQR $= 300 - 260 = 40$. We compute

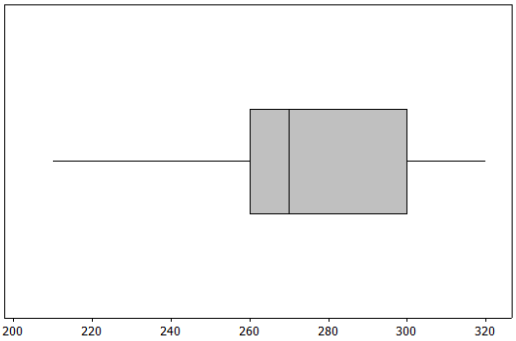
$$Q_1 - 1.5(IQR) = 260 - 1.5(40) = 200,$$

and

$$Q_3 + 1.5(IQR) = 300 + 1.5(40) = 360.$$

Since the minimum (210) and maximum (320) values lie inside these values, there are no outliers.

(b) Boxplot:



2.127 (a) We see that $Q_1 = 42$ and $Q_3 = 56$ so the interquartile range is IQR $= 56 - 42 = 14$. We compute

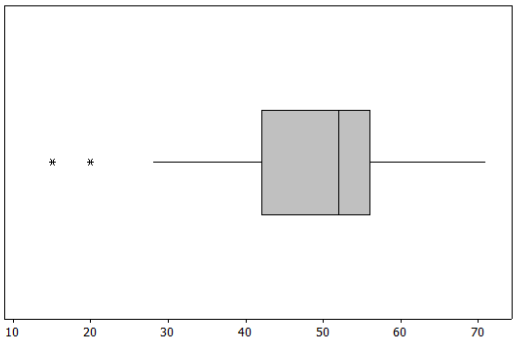
$$Q_1 - 1.5(IQR) = 42 - 1.5(14) = 21,$$

and

$$Q_3 + 1.5(IQR) = 56 + 1.5(14) = 77.$$

There are two data values that fall outside these values. We see that 15 and 20 are both small outliers.

(b) Notice that the line on the left of the boxplot extends down to 28, the smallest data value that is not an outlier.



2.128 (a) We see that $Q_1 = 72$ and $Q_3 = 80$ so the interquartile range is IQR $= 80 - 72 = 8$. We compute

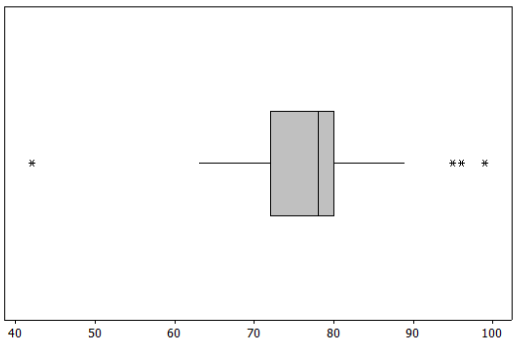
$$Q_1 - 1.5(IQR) = 72 - 1.5(8) = 60,$$

and

$$Q_3 + 1.5(IQR) = 80 + 1.5(8) = 92.$$

There are four data values that fall outside these values, one on the low side and three on the high side. We see that 42 is a low outlier and 95, 96, and 99 are all high outliers.

(b) Notice that the line on the left of the boxplot extends down to 63, the smallest data value that is not an outlier, while the line on the right extends up to 89, the largest data value that is not an outlier.



2.129 (a) We see that $Q_1 = 10$ and $Q_3 = 16$ so the interquartile range is IQR $= 16 - 10 = 6$. We compute

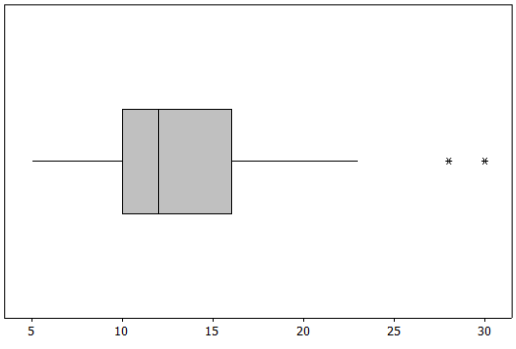
$$Q_1 - 1.5(IQR) = 10 - 1.5(6) = 1,$$

and

$$Q_3 + 1.5(IQR) = 16 + 1.5(6) = 25.$$

There are no small outliers and there are two large outliers, at 28 and 30.

(b) Notice that the line on the right extends up to 23, the largest data value that is not an outlier.



2.130 (a) Most of the data is between 0 and 5, and then the data stretches way out to the right to some very large outliers. This data is skewed to the right.

(b) The data appear to range from about 0 to about 65, so the range is about $65 - 0 = 65$.

(c) The median appears to be about 2. About half of all movies recover less than 200% of their budget, and half recover more than 200% of the budget.

(d) The very large outliers will pull the mean up, so we expect the mean to be larger than the median. (In fact, the median is 2.2 while the mean is 3.315.)

2.131 We see from the five number summary that the interquartile range is $IQR = Q_3 - Q_1 = 77 - 49 = 28$. Using the IQR, we compute:

$$\begin{aligned} Q_1 - 1.5 \cdot IQR &= 49 - 1.5(28) = 49 - 42 = 7. \\ Q_3 + 1.5 \cdot IQR &= 77 + 1.5(28) = 77 + 42 = 119. \end{aligned}$$

Scores greater than 119 are impossible since the scale only goes up to 100. An audience score less than 7 would qualify as a low outlier. (That would have to be a *very* bad movie!) Since the minimum rating (seen in the five number summary) is 24, there are no low outliers.

2.132 (a) Action movies appear to have the largest budgets, while horror and drama movies appear to have the smallest budgets.

(b) Action movies have by far the biggest spread in the budgets, with dramas appearing to have the smallest spread.

(c) Yes, there definitely appears to be an association between genre and budgets, with action movies having substantially larger budgets than the other three types.

2.133 (a) The highest mean is in Drama, while the lowest mean is in Horror movies.

(b) The highest median is in Drama, while the lowest median is in Action movies.

(c) The lowest score is 24 and it is for a Thriller. The highest score is 93, and it is obtained by both an Action and a Comedy.

(d) The genre with the largest number of movies is Action, with $n = 32$.

2.134 (a) The lowest level of physical activity appears to be in the South, and the highest appears to be in the West.

(b) The biggest range is in the Midwest.

(c) There are no outliers in any of the regions.

(d) Yes, the boxplots are very different between the different regions.

2.135 (a) We see that the interquartile range is $IQR = Q_3 - Q_1 = 149 - 15 = 134$. We compute:

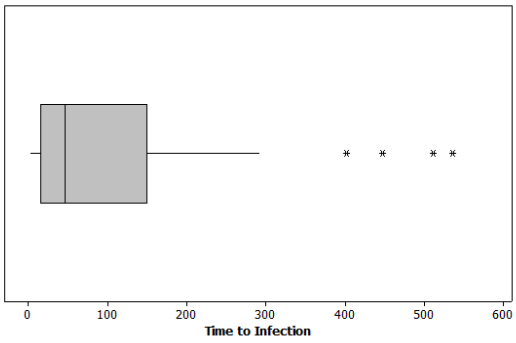
$$Q_1 - 1.5(IQR) = 15 - 1.5(134) = 15 - 201 = -186.$$

and

$$Q_3 + 1.5(IQR) = 149 + 1.5(134) = 149 + 201 = 350.$$

Outliers are any values outside these fences. In this case, there are four outliers that are larger than 350. The four outliers are 402, 447, 511, and 536.

(b) A boxplot of time to infection is shown:



2.136 (a) The median corresponds to the middle line in each box. An estimated median for the AL is about 1455 hits, and an estimated median for the NL is about 1410 hits (other answers are possible, but should be similar). So an estimated difference in median number of hits is $1455 - 1410 = 45$ hits. The American League has more hits.

(b) The other obvious difference is that the variability is greater for teams in the American League. Other correct ways to state this are that the standard deviation is greater, the IQR is greater, or the range is greater.

2.137 (a) We have IQR $= 2106 - 1334 = 772$, so the upper boundary for non-outlier data values is:

$$\begin{aligned} Q_3 + 1.5(IQR) &= 2106 + 1.5(772) \\ &= 2106 + 1158 \\ &= 3264. \end{aligned}$$

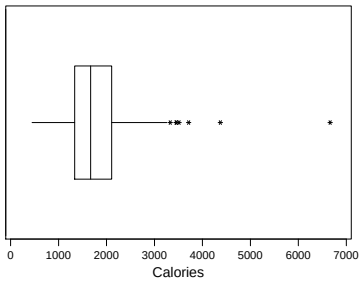
Any data value above 3264 is an outlier, so the seven largest calorie counts are all outliers.

(b) We have already seen that $IQR = 772$, so the lower boundary for non-outlier data values is

$$\begin{aligned} Q_1 - 1.5(IQR) &= 1334 - 1.5(772) \\ &= 1334 - 1158 \\ &= 176. \end{aligned}$$

We see in the five number summary that the minimum data value is 445, so there are no values below 176 and no low outliers.

(c) A boxplot of daily calorie consumption is shown:



2.138 (a) The median appears to be about 500 calories higher for the males than for the females. The largest outlier of 6662 calories in one day is a male, but the females have many more outliers.

(b) Yes, there does appear to be an association. Females appear to have significantly lower calorie consumption than males. We see that every number in the five number summary is higher for males than it is for females. The median for females is even lower than the first quartile for males.

2.139 The side-by-side boxplots are almost identical. Vitamin use appears to have no effect on the concentration of retinol in the blood.

2.140 The blood pressures have a relatively symmetric distribution ranging from a low of around 35 mm Hg to a high just over 250 mm Hg. The middle 50% of blood pressures are between 110 mm Hg and 150 mm Hg with a median value of 130 mm Hg. There are two unusually low blood pressures at around 35 and 46 and three unusually high blood pressures at 210, 220 and 255.
The five number summary appears to be about $(35, 110, 130, 150, 255)$.

2.141 Both distributions are relatively symmetric with one or two outliers. In general, the blood pressures of patients who lived appear to be slightly higher as a group than those of the patients who died. The middle 50% box for the surviving patients is shifted to the right of the box for patients who died and shows a smaller interquartile range. Both quartiles and the median are larger for the surviving group. Note that the boxplots give no information about how many patients are in each group. From the original data table, we can find that 40 of the 200 patients died and the rest survived.

2.142 (a) Yes, there does appear to be an association. Honeybees appear to dance many more circuits for a high quality option.

(b) It is obvious that there are no low outliers, so we look for high outliers in each case. For the high quality group, we have $IQR = 122.5 - 7.5 = 115$, so outliers would be those beyond

$$Q_3 + 1.5 \cdot IQR = 122.5 + 1.5(115) = 122.5 + 172.5 = 295.$$

There are two outliers in the high quality group, with one at the maximum of 440 and the other appearing on the dotplot to be at approximately 330. These two honeybee scouts must have been very enthusiastic about this possible new home!
For the low quality group, we have $IQR = 42.5 - 0 = 42.5$, so outliers would be those beyond

$$Q_3 + 1.5 \cdot IQR = 42.5 + 1.5(42.5) = 42.5 + 63.75 = 106.25.$$

There are three outliers in the low quality group, with one at the maximum of 185 and the other two appearing on the dotplot to be at approximately 140 and 175. Notice that none of these three outliers would be considered outliers in the high quality group.

(c) The difference in means is $\overline{x}_H - \overline{x}_L = 90.5 - 30.0 = 60.5$.

(d) We see in the five number summary that the largest value in the high quality group is 440. We find:

$$z\text{-score for } 440 \ = \frac{x - \text{Mean}}{\text{Standard deviation}} = \frac{440 - 90.5}{94.6} = 3.695.$$

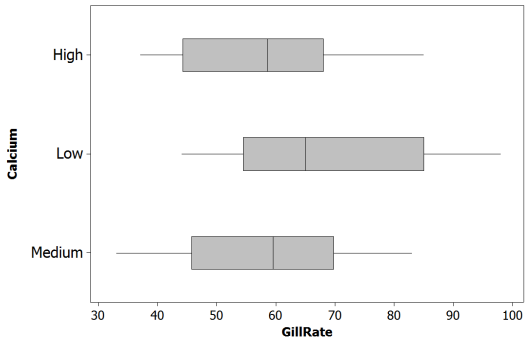
We see in the five number summary that the largest value in the low quality group is 185. We find:

$$z\text{-score for } 185 \ = \frac{x - \text{Mean}}{\text{Standard deviation}} = \frac{185 - 30}{49.4} = 3.138.$$

Both the largest values are more than 3 standard deviations above the mean. The one in the high quality group is a bit larger relative to its group.

(e) No, since the data are not bell-shaped.

2.143 (a) See the figure. It appears that respiration rate is higher when calcium levels are low, and that there is not much difference in respiration rate between medium and high levels of calcium.



(b) With a low calcium level, the mean is 68.50 beats per minute with a standard deviation of 16.23. With a medium level, the mean is 58.67 beats per minute with a standard deviation of 14.28. With a high level of calcium, the mean is 58.17 beats per minute with a standard deviation of 13.78. Again, we see that respiration is highest with low calcium.

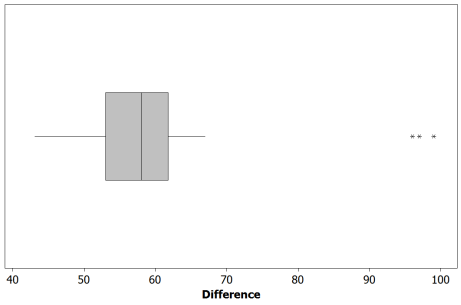
(c) This is an experiment since the calcium level was actively manipulated.

2.144 (a) The explanatory variable is whether the traffic lights are on a fixed or flexible system. This variable is categorical. The response variable is the delay time, in seconds, which is quantitative.

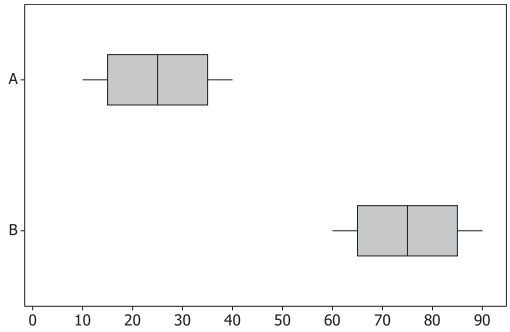
(b) Using technology we find the mean and standard deviation for each sample:
Timed: $\overline{x}_T = 105$ seconds and $s_T = 14.1$ seconds
Flexible: $\overline{x}_F = 44$ seconds and $s_F = 3.4$ seconds
This shows that the mean delay time is much less, 61 seconds or more than a full minute, with the flexible light system. We also see that the variability is much smaller with the flexible system.

(c) For the differences we have $\overline{x}_D = 61$ seconds and $s_D = 15.2$ seconds.

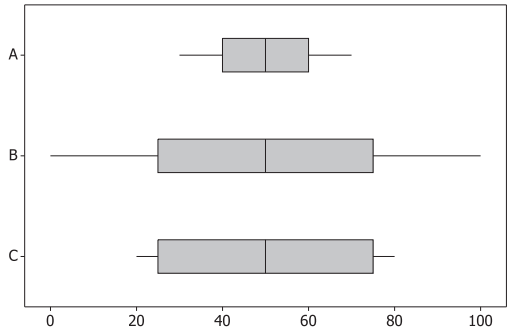
(d) The boxplot is shown. We see that there are 3 large outliers. Since this is a boxplot of the differences, this means there were three simulation runs where the flexible system *really* improved the time.



2.145 Here is one possible graph of the side-by-side boxplots:



2.146 Here is one possible graph of the side-by-side boxplots:



2.147 Answers will vary.

2.148 Answers will vary.

Section 2.5 Solutions

2.149 A correlation of -1 means the points all lie exactly on a line and there is a negative association. The matching scatterplot is (b).

2.150 A correlation of 0 means there appears to be no linear association in the scatterplot, so the matching scatterplot is (c).

2.151 A correlation of 0.8 means there is an obvious positive linear association in the scatterplot, but there is some deviation from a perfect straight line. The matching scatterplot is (d).

2.152 A correlation of 1 means the points all lie exactly on a line and there is a positive association. The matching scatterplot is (a).

2.153 The correlation represents almost no linear relationship, so the matching scatterplot is (c).

2.154 The correlation represents a mild negative association, so the matching scatterplot is (a)

2.155 The correlation shows a strong positive linear association, so the matching scatterplot is (d).

2.156 The correlation shows a strong negative association, so the matching scatterplot is (b).

2.157 We expect that larger houses will cost more to heat, so we expect a positive association.

2.158 Since the amount of gas goes down as distance driven goes up, we expect a negative association.

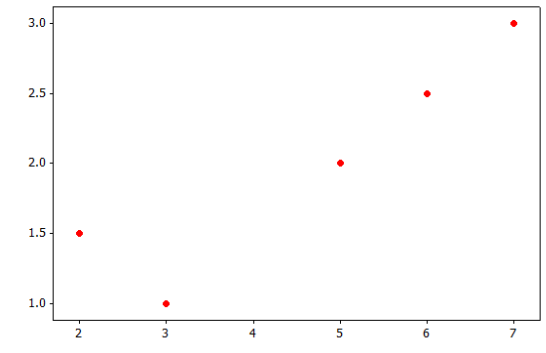
2.159 We wear more clothes when it is cold outside, so as the temperature goes down, the amount of clothes worn goes up. This describes a negative association.

2.160 Usually someone who sends lots of texts also gets lots of them in return, and someone who does not text very often does not get very many texts. This describes a positive relationship.

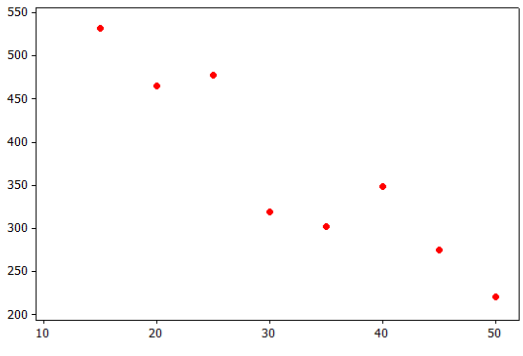
2.161 Usually there are not many people in a heavily wooded area and there are not many trees in a heavily populated area such as the middle of a city. This would be a negative relationship.

2.162 While it is certainly not a perfect relationship, we generally expect that more time spent studying will result in a higher exam grade. This describes a positive relationship.

2.163 See the figure below.



2.164 See the figure below.

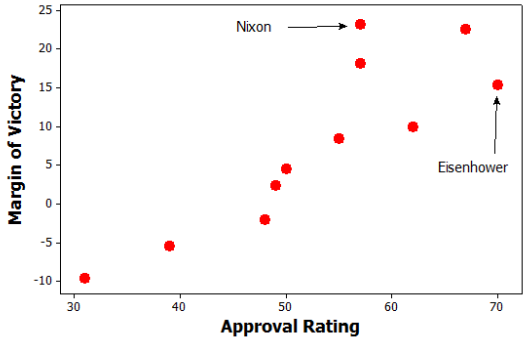


2.165 The correlation is $r = 0.915$.

2.166 The correlation is $r = -0.932$.

2.167 (a) The incumbent lost 3 times and won 8 times. We have $3/11 = 0.273$. Since 1940, the sitting president has lost his bid for re-election 27.3% of the time.

(b) President Eisenhower had the highest approval rating at 70%. President Nixon had the highest margin of victory at 23.2%.



2.168 (a) A positive association means that large values of one variable tend to be associated with large values of the other; in this case, that taller people generally weigh more. A negative association means that large values of one variable tend to be associated with small values of the other; in this case, that tall people generally weigh less than shorter people. Since we expect taller people to generally weigh more, we expect a positive relationship between these two variables.

(b) In the scatterplot, we see a positive upward relationship in the trend (as we expect) but it is not very strong. It appears to be approximately linear.

(c) The outlier in the lower right corner appears to have height about 83 inches (or 6 ft 11 inches) and weight about 135 pounds. This is a very tall thin person! (It is reasonable to suspect that this person may have entered the height incorrectly on the survey. Outliers can help us catch data-entry errors.)

2.169 The explanatory variable is roasting time and the response variable is the amount of caffeine. The two variables have a negative association.

2.170 (a) More nurturing is associated with larger hippocampus size, so this is a positive association.

(b) Larger hippocampus size is associated with more resiliency, so this is a positive association.

(c) An experiment would involve randomly assigning some children to get lots of nurturing while randomly assigning some other children to get less nurturing. After many years, we would measure the size of the hippocampus in their brains. It is clearly not ethical to assign some children to not get nurtured!

(d) We cannot conclude that there is a cause and effect relationship in humans. No experiment has been done and there are many possible confounding variables. We can, however, conclude that there is a cause and effect relationship in animals, since the animal results come from experiments. This causation in animals probably increases the likelihood that there is a causation effect in humans as well.

2.171 Since more cheating by the mother is associated with more cheating by the daughter, this is a positive association.

2.172 (a) The dots go up as we move left to right, so there appears to be a positive relationship. In context, that means that as a country's residents use more of the planet's resources, they tend to be happier and healthier.

(b) The bottom left is an area with low happiness and low ecological footprint, so they are countries whose residents are not very happy and don't use many of the planet's resources.

(c) Costa Rica is the highest dot – a black dot with an ecological footprint of about 2.0.

(d) For ecological footprints between 0 and 6, there is a strong positive relationship. For ecological footprints between 6 and 10, however, there does not seem to be any relationship. Using more resources appears to improve happiness up to a point but not beyond that.

(e) Countries in the top left are high on the happiness scale but are relatively low on resource use.

(f) There are many possible observations one could make, such as that countries in Sub-Saharan Africa are low on the happiness and well-being scale and also low on the use of resources, while Western Nations are high on happiness but also very high on the use of the planet's resources.

(g) For those in the bottom left (such as many countries in Sub-Saharan Africa), efforts should be devoted to improving the well-being of the people. For those in the top right (such as many Western nations), efforts should be devoted to reducing the use of the planet's resources.

2.173 (a) There appears to be a negative association, which means in this context that states with a larger proportion of the population eating lots of vegetables tend to have a lower obesity rate. This makes sense since people who eat lots of vegetables are less likely to be obese.

(b) A healthy state would include a large percentage of people eating lots of vegetables and a small percentage of people who are obese, which corresponds to the bottom right corner. An unhealthy state would include a small percentage of people eating lots of vegetables and a high obesity rate, which corresponds to the top left corner.

(c) There is one point furthest in the bottom right corner and it corresponds to the state of Vermont. There are three dots in the unhealthy top left corner, and they correspond to the states of Mississippi, Kentucky, and Oklahoma.

(d) Since all 50 US States are included, this is a population. The correct notation is ρ.

(e) The two variables appear to be negatively correlated with a moderately strong linear relationship. The correct correlation is $\rho = -0.605$.

(f) No, correlation does not imply cause and effect relationships.

(g) No, correlation does not imply cause and effect relationships.

(h) That point corresponds to the state of Colorado.

2.174 (a) A negative relationship would mean that old people were married to young people and vice versa. It would mean that an 80-year-old might more likely be married to a 20-year-old than to another 80-year-old.

(b) A positive relationship would mean that old people tended to be married to old people and young people tended to be married to young people.

(c) A positive relation is expected between these two variables.

(d) We expect a very strong linear relationship since it is quite common for people to be married to someone similar in age.

(e) Yes, a strong correlation implies an association (but not causation!).

2.175 (a) There are three variables mentioned: how closed a person's body language is, level of stress hormone in the body, and how powerful the person felt. Since results are recorded on numerical scales that represent a range for body language and powerful, all three variables are quantitative.

(b) People with a more closed posture (low values on the scale) tended to have higher levels of stress hormones, so there appears to be a negative relationship. If the scale for posture had been reversed, the answer would be the opposite. A positive or negative relationship can depend on how the data is recorded.

(c) People with a more closed posture (low values on the scale) tended to feel less powerful (low values on that scale), so there appears to be a positive relationship. If both scales were reversed, the answer would not change. If only one of the scales was reversed, the answer would change.

2.176 (a) A positive relationship would imply that a student who is good at one of the tests is also likely to be good at the other – that students are generally either good at both or bad at both. A negative relationship implies that students tend to be good at either math or verbal but not both.

(b) A student in the top left is good at verbal and bad at math. A student in the top right is good at both. A student in the bottom left is bad at both, and a student in the bottom right is good at math and bad at verbal.

(c) There is not a strong linear relationship as the dots appear to be all over the place. This tells you that the scores students get on the math and verbal SAT exams are not very closely related.

(d) Since the linear relationship is not very strong, the correlation is likely to be one of the values closest to zero – either -0.235 or 0.445. Since there is more white space in the top left and bottom right corners than in the other two corners, the weak relationship appears to be a positive one. The correct correlation is 0.445.

2.177 (a) A positive relationship would imply that a student who exercises lots also watches lots of television, and a student who doesn't exercise also doesn't watch much TV. A negative relationship implies that students who exercise lots tend to not watch much TV and students who watch lots of TV tend to not exercise much.

(b) A student in the top left exercises lots and watches very little television. A student in the top right spends lots of hours exercising and also spends lots of hours watching television. (Notice that there are no students in this portion of the scatterplot.) A student in the bottom left does not spend much time either exercising or watching television. (Notice that there are lots of students in this corner.) A student in the bottom right watches lots of television and doesn't exercise very much.

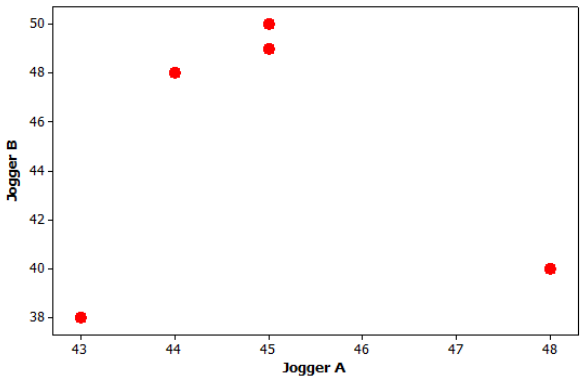
(c) The outlier on the right watches a great deal of television and spends very little time exercising. The outlier on the top spends a great deal of time exercising and watches almost no television.

(d) There is essentially no linear relationship between the number of hours spent exercising and the number of hours spent watching television.

2.178 (a) The data point (204,52) for patient #772 is the dot just above the tick mark for 200 on the systolic blood pressure axis.

(b) No. The rest of the data in this scatterplot are fairly randomly distributed and show no clear association in either direction between heart rate and blood pressure.

2.179 (a) Here is a scatterplot of Jogger A vs Jogger B.



(b) The correlation between the two joggers is -0.096.

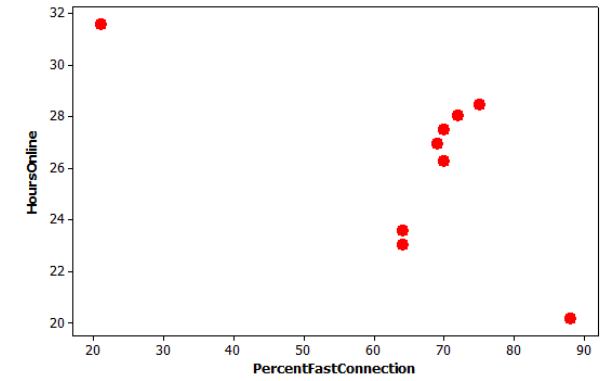
(c) The correlation between the two joggers with the windy race added is now 0.562.

(d) Adding the results from the windy day has a very strong effect on the relationship between the two joggers!

2.180 (a) A positive relationship implies that as connection speed goes up, time online goes up. This might make sense because being online is more enjoyable with a fast connection speed, so people may spend more time online.

(b) A negative relationship implies that as connection speed goes up, time online goes down. This might make sense because if connection speed is fast, people can accomplish what they need to accomplish online in a shorter amount of time so they spend less time online waiting.

(c) See the scatterplot below. These two variables have a negative association. There are two outliers. One, in the top left corner, corresponds to Brazil, which has a low percent with a fast connection and a high number of hours online. A second, in the bottom right, corresponds to Switzerland, which has a high percent fast connection and a low hours online.



(d) If we ignore the two outliers, the variables appear to have a positive relationship.

(e) The correlation for this sample of countries is $r = -0.649$. The correlation is pretty strong and negative, so it is being heavily influenced by the two outliers.

(f) No! This data comes from an observational study, so we cannot conclude that there is a causal association.

2.181 Type of liquor is a categorical variable, so correlation should not be computed and a positive relationship has no meaning. There cannot be a *linear* relationship involving a categorical variable.

2.182 (a) We see in the scatterplot that the relationship is positive. This makes sense for irises: petals which are long are generally wider also.

(b) There is a relatively strong linear relationship between these variables.

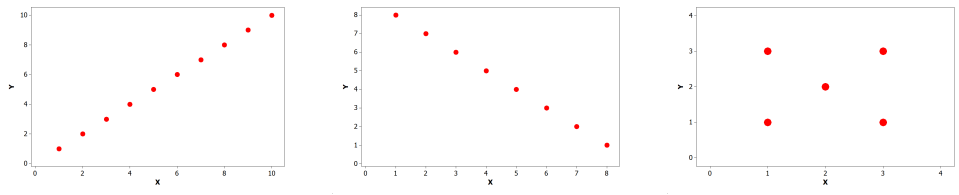
(c) The correlation is positive and close to, but not equal to, 1. A reasonable estimate is $r \approx 0.9$.

(d) There are no obvious outliers.

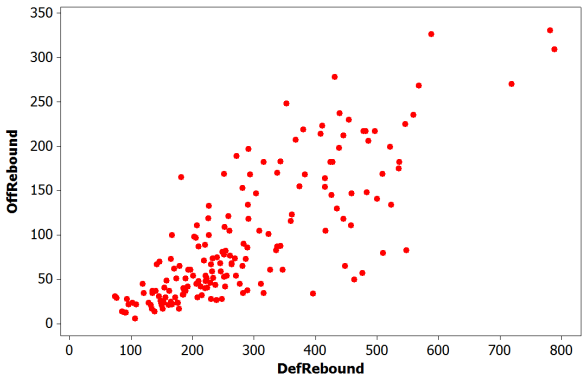
(e) The width of that iris appears to be about 11 mm.

(f) There are two obvious clumps in the scatterplot that probably correspond to different species of iris. One type has significantly smaller petals than the other(s).

2.183 There are many ways to draw the scatterplots. One possibility for each is shown in the figure.



2.184 (a) See the figure.

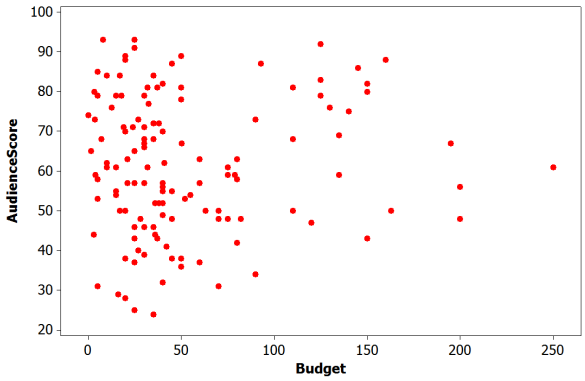


(b) There is a relatively strong positive relationship, which means players who have lots of defensive rebounds also tend to have lots of offensive rebounds. This makes sense since players who are very tall tend to get lots of rebounds in either end.

(c) There are three outliers with more than 700 defensive rebounds. We see in the data file that these three players are Dwight Howard with 789, Kevin Love with 782, and Blake Griffin with 719.

(d) We use technology to see that the correlation is 0.803. This correlation matches the strong positive linear relationship we see in the scatterplot.

2.185 (a) Since we are looking to see if budget affects audience score, we put the explanatory variable (budget) on the horizontal axis and the response variable (audience score) on the vertical axis. See the figure.



(b) There is not a strong linear relationship. In context, this means that having a larger budget for a movie does not tend to make for a higher (or lower) score from the audience.

(c) The outlier has a budget about 250 million dollars. This movie is *Pirates of the Caribbean* and the audience score is 61. The movie with a budget of 125 million dollars and an audience score over 90 is *Harry Potter and the Deathly Hallows, Part 2*.

(d) We use technology to see that the correlation is 0.084. This correlation near 0 matches the lack of an association we see in the scatterplot.

2.186 Answers will vary

Section 2.6 Solutions

2.187 (a) The predicted value for the data point is $\widehat{Hgt} = 24.3 + 2.74(12) = 57.18$ inches. The residual is $60 - 57.18 = 2.82$. This child is 2.82 inches taller than the predicted height.

(b) The slope 2.74 tells the expected change in Hgt given a one year increase in Age. We expect a child to grow about 2.74 inches per year.

(c) The intercept 24.3 tells the Hgt when the Age is 0, or the height (or length) of a newborn. This context does make sense, although the estimate is rather high.

2.188 (a) The predicted value for the data point is $\widehat{BAC} = -0.0127 + 0.018(3) = 0.0413$. The residual is $0.08 - 0.0413 = 0.0387$. This individual's BAC was 0.0387 higher than predicted.

(b) The slope of 0.018 tells us the expected change in BAC given a one drink increase in drinks. We expect one drink by this individual to increase BAC by 0.018.

(c) The intercept of -0.0127 tells us that the BAC of someone who has consumed no drinks is negative. The context makes sense, but a negative BAC is not possible!

2.189 (a) The predicted value for the data point is $\widehat{Weight} = 95 + 11.7(5) = 153.5$ lbs. The residual is $150 - 153.5 = -3.5$. This individual is capable of bench pressing 3.5 pounds less than predicted.

(b) The slope 11.7 tells the expected change in Weight given a one hour a week increase in Training. If an individual trains an hour more each week, the predicted weight the individual is capable of bench pressing would go up 11.7 pounds.

(c) The intercept 95 tells the Weight when the hours Training is 0, or the bench press capability of an individual who never lifts weights. This intercept does make sense in context.

2.190 (a) The predicted value for the data point is $\widehat{Grade} = 41.0 + 3.8(10) = 79$. The residual is $81 - 79 = 2$. This student did two points better than predicted.

(b) The slope 3.8 tells the expected change in Grade given a one hour increase in Study. We expect the grade to go up by 3.8 for every additional hour spent studying.

(c) The intercept 41.0 tells the Grade when Study is 0. The expected grade is 41 if the student does not study at all. This context makes sense.

2.191 The regression equation is $\hat{Y} = 0.395 + 0.349X$.

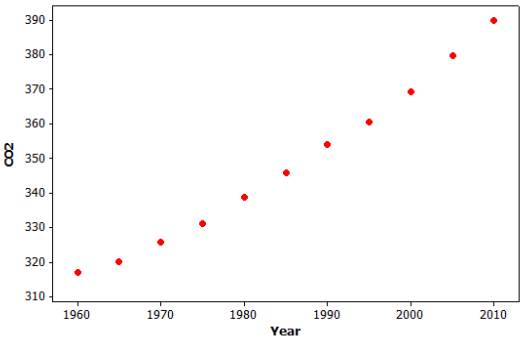
2.192 The regression equation is $\hat{Y} = 47.267 + 1.843X$.

2.193 The regression equation is $\hat{Y} = 111.7 - 0.84X$.

2.194 The regression equation is $\hat{Y} = 641.62 - 8.42X$.

2.195 (a) Year is the explanatory variable and CO_2 concentration is the response variable.

(b) A scatterplot of CO_2 vs $Year$ is shown. There is a very strong linear relationship in the data.



(c) We find that $r = 0.993$. This correlation is very close to 1 and matches the very strong linear relationship we see in the scatterplot.

(d) We see that $\widehat{CO_2} = -2571 + 1.47(Year)$.

(e) The slope is 1.47. Carbon dioxide concentrations in the atmosphere have been going up at a rate of about 1.47 ppm each year.

(f) The intercept is -2571. This is the expected CO_2 concentration in the year 0, but clearly doesn't make any sense since the concentration can't be negative. The linear trend clearly does not extend back that far and we can't extrapolate back all the way to the year 0.

(g) In 2003, the predicted CO_2 concentration is $\widehat{CO_2} = -2571+1.47(2003) = 373.41$. This seems reasonable since the value lies between the data values for years 2000 and 2005.
In 2020, the predicted CO_2 concentration is $\widehat{CO_2} = -2571 + 1.47(2020) = 398.4$. We have less confidence in this prediction since we can't be sure the linear trend will continue.

(h) In 2010, the predicted CO_2 concentration is $\widehat{CO_2} = -2571 + 1.47(2010) = 383.7$. We see in the data that the actual concentration that year is 389.78, so the residual is $389.78 - 383.7 = 6.08$. The CO_2 concentration in 2010 was quite a bit above the predicted value.

2.196 (a) The explanatory variable is the duration of the waggle dance. We use it to predict the response variable which is the distance to the source.

(b) Yes, there is a very strong positive linear trend in the data.

(c) We use technology to see that the correlation is $r = 0.994$. These honeybees are very precise with their timing!

(d) We use technology to see that the regression line is $\widehat{Distance} = -399 + 1174 \cdot Duration$.

(e) The slope is 1174, and indicates that the distance to the source goes up by 1174 meters if the dance lasts one more second.

(f) If the dance lasts 1 second, we predict that the source is $\widehat{Distance} = -399 + 1174(1) = 775$ meters away. If the dance lasts 3 seconds, we predict that the source is $\widehat{Distance} = -399 + 1174(3) = 3123$ meters away.

2.197 (a) The trend is positive, with a mild linear relationship.

(b) The residual is the vertical distance from the point to the line, which is much larger in 2007 than it is in 2008. We see that in 2010, the point is below the line so the observed value is less than the predicted value. The residual is negative.

(c) We use technology to see that the correlation is $r = 0.692$.

(d) We use technology to find the regression line: $\widehat{HotDogs} = -3385 + 1.72 \cdot Year$.

(e) The slope indicates that the winning number is going up by about 1.72 more hot dogs each year. People better keep practicing!

(f) The predicted number in 2012 is $\widehat{HotDogs} = -3385 + 1.72 \cdot (2012) = 75.64$, which is a very large number of hot dogs!

(g) It is not appropriate to use this regression line to predict the winning number in 2020 because that is extrapolating too far away from the years that were used to create the dataset.

2.198 (a) We are using runs to predict wins, so the explanatory variable is runs and the response variable is wins.

(b) The intercept means that if a team got no runs in a season they would win 0.362 games. This makes sense, if a team doesn't get any runs they won't win any games - but that isn't realistic for actual baseball teams over an entire season. The slope means that we predict that obtaining 1 more run leads to about 0.114 more wins.

(c) The predicted number of wins for Oakland is $\widehat{Wins} = 0.362 + 0.114(663) = 75.9$ games. The residual is $81 - 75.9 = 5.1$, meaning that the A's won 5.1 more games than expected with 663 runs, so they were very efficient at winning games.

2.199 The slope is 0.836. The slope tells us the expected change in the response variable (Margin) given a one unit increase in the predictor variable (Approval). In this case, we expect the margin of victory to go up by 0.836 if the approval rating goes up by 1.
The y-intercept is -36.5. This intercept tell us the expected value of the response variable (Margin) when the predictor variable (Approval) is zero. In this case, we expect the margin of victory to be -36.5 if the approval rating is 0. In other words, if *no one* approves of the job the president is doing, the president will lose in a landslide. This is not surprising!

2.200 (a) For a height of 60, the predicted weight is $\widehat{Weight} = -170 + 4.82(60) = 119.2$. The predicted weight for a person who is 5 feet tall is 119.2 pounds. For a height of 72, the predicted weight is $\widehat{Weight} = -170 + 4.82(72) = 177.04$. The predicted weight for a person who is 6 feet tall is about 177 pounds.

(b) The slope is 4.82. For an additional inch in height, weight is predicted to go up by 4.82 pounds.

(c) The intercept is -170, and indicates that a person who is 0 inches tall will weigh -170 pounds. Clearly, it doesn't make any sense to predict the weight of a person who is 0 inches tall! It also doesn't make sense to have a negative weight.

(d) For a "height" of 20 inches, the line predicts a weight of $\widehat{Weight} = -170 + 4.82(20) = -73.6$ pounds. This is a ridiculous answer. We cannot use the regression line in this case because the line is based on data for adults (specifically, college students), and we should not extrapolate so far from the data used to create the line.

2.201 The man with the largest positive residual weighs about 190 pounds and has a body fat percentage about 40%. The predicted body fat percent for this man is about 20% so the residual is about $40 - 20 = 20$.

2.202 (a) There is a stronger positive linear trend, and therefore a larger correlation, in the one using abdomen.

(b) The actual body fat percent, from the dot, appears to be about 34% while the predicted value on the line appears to be about 40%.

(c) This person has an abdomen circumference of approximately 113 cm. The predicted body fat percent for this abdomen circumference appears on the line to be about 32%, which gives a residual of about $40 - 32 = 8$.

2.203 (a) For 35 cm, the predicted body fat percent is $\widehat{BodyFat} = -47.9 + 1.75 \cdot (35) = 13.35\%$. For a neck circumference of 40 cm, the predicted body fat percent is $\widehat{BodyFat} = -47.9 + 1.75 \cdot (40) = 22.1\%$.

(b) The slope of 1.75 indicates that as neck circumference goes up by 1 cm, body fat percent goes up by 1.75.

(c) The predicted body fat percent for this man is $\widehat{BodyFat} = -47.9 + 1.75 \cdot (38.7) = 19.825\%$, so the residual is $11.3 - 19.825 = -8.525$.

2.204 (a) There is a strong positive linear relationship.

(b) Using T to represent temperature and R to represent chirp rate, we see that the regression line for these data is $\hat{T} = 37.68 + 0.23R$.

(c) We use the regression line to find the predicted value for each data point, and then subtract to find the residuals. The results are given in the table. We see that the predicted values are all quite close to the actual values, so the residuals are all relatively small.

Chirp rate (R)	Temperature (T)	Predicted Temp ($\hat{T}$)	Residual
81	54.5	56.31	-1.81
97	59.5	59.99	-0.49
103	63.5	61.37	2.13
123	67.5	65.97	1.53
150	72.0	72.18	-0.18
182	78.5	79.54	-1.04
195	83.0	82.53	0.47

2.205 (a) We are attempting to predict rural population with land area, so land area is the explanatory variable, and percent rural is the response.

(b) There appears to be some positive correlation between these two variables, so the most likely correlation is 0.60.

(c) Using technology the regression line is: $\widehat{Rural} = 28.99 + 0.079(LandArea)$. The slope is 0.079, which means percent rural goes up by about 0.079 with each increase in 1,000 sq km of country size.

(d) The intercept does not make sense, since a country of size zero would have no population at all!

(e) The most influential country is the one in the far top right, which is Uzbekistan (UZB). This is due to the fact that Uzbekistan is much larger then any of the other countries sampled, so it appears to be an outlier for the explanatory variable.

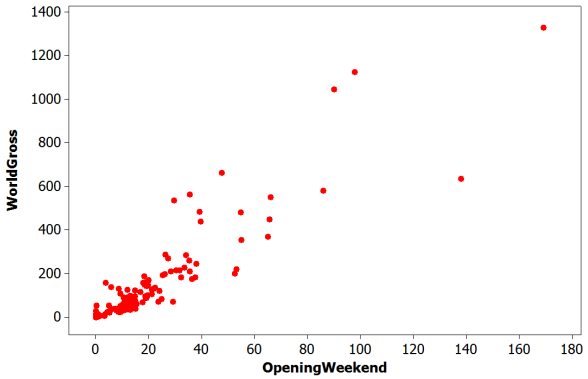
(f) Predicting the percent rural for USA with the prediction equation gives $\widehat{Rural} = 28.99 + 0.079(9147.4) = 751.63$. This implies that 752% of the United States population lives in rural areas, which doesn't make any sense at all. The regression line does not work at all for the US, because we are extrapolating so far outside of the original sample of 10 land area values. The US is much larger in area than any of the countries that happened to be picked for the random sample.

2.206 (a) Using technology the regression line is: $\widehat{Rural} = 37.85 - 0.002(LandArea)$.

(b) The slope with the USA is -0.002, the slope without is 0.08. These are two very different slopes. Adding the USA has a strong effect, because the United States is an extreme outlier in terms of size.

(c) Predicting the USA percent rural with the new regression line gives $\widehat{Rural} = 37.85 - 0.002(9147.4) = 19.5\%$, compared to $\widehat{Rural} = 28.99 + 0.079(9147.4) = 751.63\%$. The prediction is much better when the United States is included because we are no longer extrapolating so far outside the data.

2.207 (a) See the figure. We see that there is a relatively strong positive linear trend. It appears that the opening weekend is a reasonably good predictor of future total world earnings for a movie.



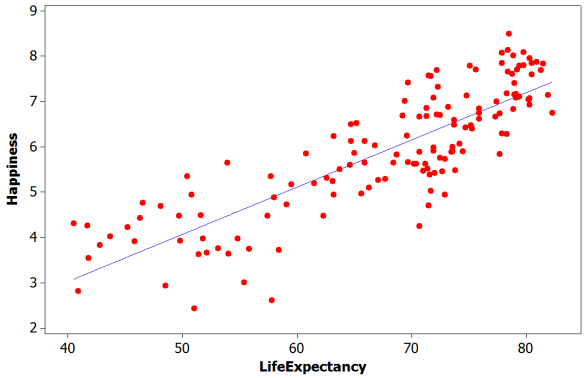
(b) The movie is *Harry Potter and the Deathly Hallows, Part 2*.

(c) We find that the correlation is $r = 0.904$.

(d) The regression line is $\widehat{WorldGross} = -8.7 + 7.70 \cdot OpeningWeekend$.

(e) If a movie makes 50 million dollars in its opening weekend, we predict that total world earnings for the year for the movie will be $\widehat{WorldGross} = -8.7 + 7.70 \cdot (50) = 376.3$ million dollars.

2.208 (a) See the figure. There is a clear linear trend, so it is reasonable to construct a regression line.



(b) Using technology, we see that the regression line is $\widehat{Happiness} = -1.09 + 0.103 \cdot LifeExpectancy$.

(c) The slope of 0.103 indicates that for an additional year of life expectancy, the happiness rating goes up by 0.103.

2.209 Answers will vary.

2.210 (a) The three variables are: average growth rate over the decade, predicted 2021 debt-to-GDP ratio, and predicted 2021 deficit.

(b) We have $\widehat{Ratio} = 129 - 19.1(Growth)$.

(c) The slope indicates that as the growth rate goes up by 1%, the predicted 2021 debt-to-GDP ratio goes down by 19.1%. This is a very impressive and good effect! The intercept of 129 indicates that if the growth rate is 0% (indicating no economic growth over the decade), the predicted 2021 debt-to-GDP ratio is 129%, which would be very bad for the country.

(d) Since $\widehat{Ratio} = 129 - 19.1(2) = 90.8$, we see that at a growth rate of 2%, the debt-to-GDP ratio is predicted to be 90.8% in 2021. Since $\widehat{Ratio} = 129 - 19.1(4) = 52.6$, we see that at a growth rate of 4%, the debt-to-GDP ratio is predicted to be 52.6% in 2021. This is a very big difference!

(e) We see from the answers to part (d) that the answer is likely to be slightly bigger than 2%. To find the actual value, we substitute a Ratio of 90% into the regression line and solve for the growth rate:

$$\begin{aligned} \widehat{Ratio} &= 129 - 19.1(Growth) \\ 90 &= 129 - 19.1(Growth) \\ -39 &= -19.1(Growth) \\ Growth &= \frac{-39}{-19.1} = 2.04. \end{aligned}$$

If the growth rate averages 2.04% over the decade 2011 to 2021, we expect the ratio to hit 90% in 2021.

(f) We have $\widehat{Deficit} = 2765 - 680(Growth)$.

(g) The slope indicates that as the growth rate goes up by 1%, the predicted deficit in 2021 goes down by 680 billion dollars. This is again a very impressive and good effect! The intercept of 2765 indicates that if the growth rate is 0% (indicating no economic growth), the predicted 2021 deficit is 2765 billion dollars, or 2.765 trillion dollars. This would not be a good outcome.

(h) Since $\widehat{Deficit} = 2765 - 680(2) = 1405$, we see that at a growth rate of 2%, the deficit is predicted to be 1405 billion dollars in 2021. Since $\widehat{Deficit} = 2765 - 680(4) = 45$, we see that at a growth rate of 4%, the deficit is predicted to be 45 billion dollars in 2021. Again, this is a very big difference!

(i) We see from the answers to part (h) that the answer is likely to be slightly bigger than 2%. To find the actual value, we substitute a Deficit of 1.4 trillion (which is 1400 billion) into the regression line and solve for the growth rate:

$$\begin{aligned} \widehat{Deficit} &= 2765 - 680(Growth) \\ 1400 &= 2765 - 680(Growth) \\ -1365 &= -680(Growth) \\ Growth &= \frac{-1365}{-680} = 2.007. \end{aligned}$$

If the growth rate averages 2.007% (or just over 2%) over the decade 2011 to 2021, we expect the deficit to be at the current level of $1.4 trillion in 2021.