# *Statistics Unlocking the Power of Data 3rd edition Lock Solutions Manual*

SKU: 9781119682165-SOLUTIONS

**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 children of illegal immigrants should be able to attend public school is $\hat{p}=0.45$.

**2.7** The report describes the results of a sample, so the correct notation is $\hat{p}$. The proportion of US teens who say they have made a new friend online is $\hat{p}=605/1060=0.571$.

**2.8** Information is provided for an entire population so we use the notation $p$ for the proportion. The proportion is $p=554,665/2,220,087=0.250$.

**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** (a) We see that there are 200 cases total and 80 had Outcome A, so the proportion with Outcome A is $80/200 = 0.40$.

(b) We see that there are 200 cases total and 100 of them are in Group 1, so the proportion in Group 1 is $100/200 = 0.5$.

(c) There are 100 cases in Group 1, and 80 of these had Outcome B, so the proportion is $80/100 = 0.80$.

(d) We see that 80 of the cases had Outcome A and 60 of these were in Group 2, so the proportion is $60/80 = 0.75$.

**2.12** (a) We see that there are 100 cases total and 70 had Outcome A, so the proportion with Outcome A is $70/100 = 0.70$.

(b) We see that there are 100 cases total and 50 of them are in Group 1, so the proportion in Group 1 is $50/100 = 0.5$.

(c) There are 50 cases in Group 1, and 10 of these had Outcome B, so the proportion is $10/50 = 0.20$.

(d) We see that 70 of the cases had Outcome A and 30 of these were in Group 2, so the proportion is $30/70 = 0.429$.

**2.13** The Canadian census includes all Canadian adults, so this is a proportion for a population and the correct notation is $p$. We know that 81.7% is the same as 0.817 so we have $p = 0.817$.

**2.14** The Canadian census includes all Canadian adults, so this is a proportion for a population and the correct notation is $p$. We know that 45.7% is the same as 0.457 so we have $p = 0.457$.

**2.15** 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 $p = 0.624$.

**2.16** (a) The proportion not working is $2649/5204 = 0.509$.

(b) The number working is $1436 + 1119 = 2555$ so the proportion working is $2555/5204 = 0.491$. (We could have also arrived at this answer by taking one minus the proportion not working: $1 - 0.509 = 0.491$.)

(c) The proportion working on campus is $1436/5204 = 0.276$ and the proportion working off campus is $1119/5204 = 0.215$. We already found the proportion not working in part (a). The relative frequency table is shown.

| Paying job? | Relative frequency |
|---|---|
| Works on campus | 0.276 |
| Works off campus | 0.215 |
| Does not work | 0.509 |
| Total | 1.00 |

**2.17** (a) Since the proportions are given (instead of the counts) for the different options for this single categorical variable, this is a relative frequency table.

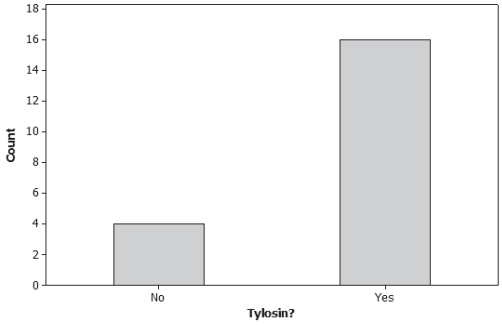
(b) Since the relative frequencies add up to 1, the proportion selecting an "Other" food delivery app is $1 - (0.276 + 0.267 + 0.252 + 0.121) = 0.084$.

**2.18** (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 |

**2.19** (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.

**2.20** (a) This is a population, since we are looking at all sports-related concussions over an entire year.

(b) The proportion who received their concussion playing football is $20293/100951 = 0.201$.

(c) The proportion who received their concussion riding bicycles is $23405/100951 = 0.232$.

(d) We cannot conclude that riding bicycles is more dangerous, because there are probably many more children riding bicycles than there are children playing football.

**2.21** (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.

**2.22** (a) There are 38 people in the control group and 26 of them developed the disease, so the proportion is $\hat{p}_C = 26/38 = 0.684$.

(b) There are 38 people in the teplizumab group and 16 of them developed the disease, so the proportion is $\hat{p}_T = 16/38 = 0.421$.

(c) We see that the differences in proportions is $\hat{p}_C - \hat{p}_T = 0.684 - 0.421 = 0.263$.

(d) Cases were randomized to treatment groups, so this is an experiment.

(e) Yes. Because this is an experiment, we can conclude that the drug causes a reduction in the disease.

**2.23** (a) We see that $\hat{p}_1 = 16/165 = 0.097$.

(b) We see that $\hat{p}_2 = 23/495 = 0.046$.

(c) The group with high social media use was more than twice as likely to develop symptoms of ADHD.

(d) The difference in proportions is $\hat{p}_1 - \hat{p}_2 = 0.097 - 0.046 = 0.051$.

(e) There were 660 teens in the study and 39 of them developed ADHD symptoms, so the proportion is $39/660 = 0.059$.

(f) The teens were not randomly assigned to high or low social media use, so this is an observational study.

(g) No, we cannot conclude that there is a causation effect since these results are from an observational study. There are many possible confounding variables.

**2.24** (a) When hiding, the proportion of the time the rats went into opaque boxes is $38/53 = 0.717$.

(b) When seeking, the proportion of the time the rats went into opaque boxes is $14/31 = 0.452$.

(c) The notation and value for the difference in sample proportions is

$$\hat{p}_1 - \hat{p}_2 = 0.717 - 0.452 = 0.265$$

**2.25** (a) The proportion of children who were given antibiotics is $438/616 = 0.711$.

(b) The proportion of children who were classified as overweight at age 9 is $181/616 = 0.294$.

(c) The proportion of those receiving antibiotics who were classified as overweight at age 9 is $144/438 = 0.329$.

(d) The proportion of those not receiving antibiotics who were classified as overweight at age 9 is $37/178 = 0.208$.

(e) Since $\hat{p}_A = 0.329$ and $\hat{p}_N = 0.208$, the difference in proportions is $\hat{p}_A - \hat{p}_N = 0.329 - 0.208 = 0.121$.

(f) Out of all children classified as overweight, the proportion who were given antibiotics is $144/181 = 0.796$.

**2.26** (a) The proportion feeling that the voices are mostly negative is $20/60 = 0.333$.

(b) The proportion of US participants feeling that the voices are mostly negative is $14/20 = 0.70$.

(c) There are 40 non-US participants in the study, and $4 + 2 = 6$ of them feel that the voices are mostly negative, so the proportion is $6/40 = 0.15$.

(d) The number of participants hearing positive voices is 29, and none of them is from the US, so the proportion is $0/29 = 0$.

(e) Yes, there appears to be a strong association between culture and how the voices are perceived.

**2.27** 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 $p_H - p_C = 0.097 - 0.052 = 0.045$.

**2.28** (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.

**2.29** (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.

**2.30** (a) There are two options for the group: therapy and no therapy. There are two options for the outcome: improvement or no improvement. The two-way table is shown, with totals included.

| | Improvement | No improvement | Total |
|---|---|---|---|
| Therapy | 14 | 6 | 20 |
| No therapy | 3 | 17 | 20 |
| Total | 17 | 23 | 40 |

(b) Seventeen people reported sleep improvement out of 40 people in the study, so the proportion is $17/40 = 0.425$.

(c) Twenty people received therapy and 14 reported improvement, so the proportion is $14/20 = 0.70$.

(d) Twenty people did not receive therapy and only 3 reported improvement, so the proportion is $3/20 = 0.15$.

(e) The difference in proportions is $\hat{p}_T - \hat{p}_N = 0.70 - 0.15 = 0.55$.

**2.31** (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.

**2.32** (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.

**2.33** (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.

**2.34** (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.

**2.35** (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.

**2.36** A two-way table to compare the smoking status of the *Reward* and *Deposit* groups is shown below.

| Group | Quit | Not quit | Total |
|---|---|---|---|
| Reward | 156 | 758 | 914 |
| Deposit | 78 | 68 | 146 |
| Total | 234 | 826 | 1060 |

To compare the success rates between the two treatments, we find the proportion in each group who quit smoking for the six months.

$$\text{Reward: } 156/914 = 0.171 \text{ vs Deposit: } 78/146 = 0.534$$

We see that the percentage of the reward only group who quit (17.1%) is quite a bit smaller than those who deposit some of their own money (53.4%).

**2.37** A two-way table to compare the participation rate of the *Reward* and *Deposit* groups is shown below.

| Group | Accepted | Declined | Total |
|---|---|---|---|
| Reward | 914 | 103 | 1017 |
| Deposit | 146 | 907 | 1053 |
| Total | 1060 | 1010 | 2070 |

To compare the participation rates between the two treatments, we find the proportion in each group who agreed to participate.

$$\text{Reward: } 914/1017 = 0.899 \text{ vs Deposit: } 146/1053 = 0.139$$

Not surprisingly, we see that the percentage in the *Reward* group who accepted the offer to participate in the program (89.9%) is much higher than in the *Deposit* group (13.9%) who were asked to risk some of their own money.

**2.38** The *Reward* group originally had 1017 subjects and $156 + 3 = 159$ eventually quit smoking, so the success rate in that group was $159/1017 = 0.156$ (15.6%). In the *Deposit* group a total of $78 + 30 = 108$ of the original 1053 subjects quit smoking to give a success rate of $108/1053 = 0.103$ (10.3%). So, overall, the reward only was the best financial incentive, but both groups did better than the subjects with no incentive.

**2.39** Recall that two variables are associated if values of one variable tend to be related to values of the other variable. In this case, it means that knowing whether a case is in Group 1 or Group 2 gives us information about the likely value of the Yes/No variable for that case.

(a) Dataset *B* shows a clear association between the two variables since we see that cases in Group 1 are more likely to give Yes values while cases in Group 2 are more likely to give No values. The results on the Yes/No variable are quite different between Group 1 and Group 2.

(b) Dataset *A* shows no association between the two variables, since values of the Yes/No variable are virtually the same between the two groups.

**2.40** (a) If there is a clear association, then there is an obvious difference in the outcomes based on which treatment is used. There are many possible answers, but the most extreme difference (in which A is always successful and B never is) is shown below.

| | Successful | Not successful | Total |
|---|---|---|---|
| Treatment A | 20 | 0 | 20 |
| Treatment B | 0 | 20 | 20 |
| Total | 20 | 20 | 40 |

(b) If there is no association, then there is no difference in the outcomes between Treatments A and B. There are many possible answers, but in every case the Treatment A and Treatment B rows would be the same or very similar. One possibility is shown in table below.

| | Successful | Not successful | Total |
|---|---|---|---|
| Treatment A | 15 | 5 | 20 |
| Treatment B | 15 | 5 | 20 |
| Total | 30 | 10 | 40 |

**2.41** (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 |

**2.42** Using technology we find the table below with counts and percentages for the superpower choices in this sample. It shows the most frequently chosen superpower is to freeze time (28.48%), while the least popular is super strength (4.19%).

| Superpower | Count | Percent |
|---|---|---|
| Fly | 120 | 26.49 |
| Freeze time | 129 | 28.48 |
| Invisibility | 80 | 17.66 |
| Super strength | 19 | 4.19 |
| Telepathy | 105 | 23.18 |
| N= | 453 | |

**2.43** (a) The *Year* variable in **StudentSurvey** has two missing values. Tallying the 360 nonmissing values gives the table below.

| FirstYear | Sophomore | Junior | Senior |
|---|---|---|---|
| 94 | 195 | 35 | 36 |

(b) The largest count is 195 sophomores. The relative frequency is $195/360 = 0.542$ or 54.2%.

**2.44** Tallying the 157 values in the *Server* variable of **RestaurantTips** gives the frequency table below.

| A | B | C |
|---|---|---|
| 60 | 65 | 32 |

Server B had the most bills for a relative frequency of $65/157 = 0.414$ or 41.4%.

**2.45** Here is a two-way table showing the distribution of *Year* and *Gender*, with column percentages to show the gender breakdown in each class year.

| | FirstYear | Junior | Senior | Sophomore | All |
|---|---|---|---|---|---|
| F | 43 | 18 | 10 | 96 | 167 |
| | 45.74 | 51.43 | 27.78 | 49.23 | 46.39 |
| M | 51 | 17 | 26 | 99 | 193 |
| | 54.26 | 48.57 | 72.22 | 50.77 | 53.61 |
| All | 94 | 35 | 36 | 195 | 360 |
| | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 |

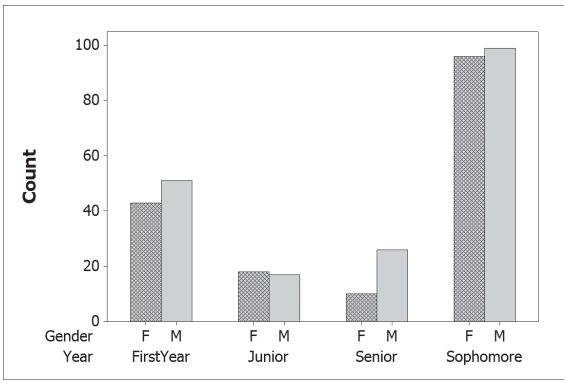
We see that the male/female split is close to 50/50 for most of the years, except for the senior year which appears to have a much higher proportion of males.

**2.46** Here is a two-way table showing the distribution of *Credit* and *Server*, with column percentages to show the credit card breakdown for each server.

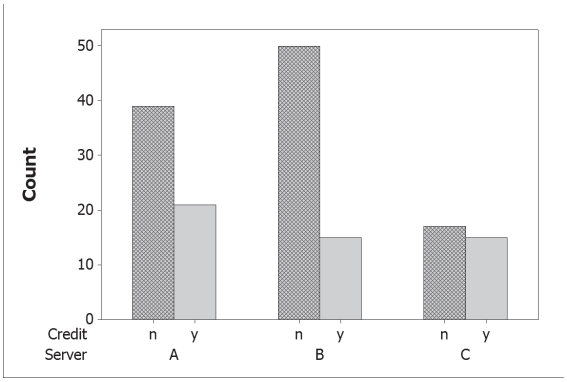
| | A | B | C | All |
|---|---|---|---|---|
| n | 39 | 50 | 17 | 106 |
| | 65.00 | 76.92 | 53.13 | 67.52 |
| y | 21 | 15 | 15 | 51 |
| | 35.00 | 23.08 | 46.88 | 32.48 |
| All | 60 | 65 | 32 | 157 |
| | 100.00 | 100.00 | 100.00 | 100.00 |

We see that cash ($Credit = n$) is used more frequently for all servers, but more often for Server B and closer to 50/50 for Server C.

**2.47** Here is a side-by-side bar chart showing the relationship between class year and gender for the **StudentSurvey** data. You might also choose a stacked bar chart or ribbon plot to show the relationship. Note that the categories are ordered alphabetically, rather than in year sequence.




**2.48** Here is a side-by-side bar chart showing the relationship between server and whether or not the bill was paid with a credit card for the **RestaurantTips** data. You might also choose a stacked bar chart or ribbon plot to show the relationship.




**2.49** Using technology and the data in **CollegeScores** we find the distribution below for the three types of *Control*.

| Control | Count | Percent |
|---|---|---|
| Private | 1832 | 29.83 |
| Profit | 2333 | 37.99 |
| Public | 1976 | 32.18 |
| N= | 6141 | |

The most frequent type of control is for profit with 2333 of the 6141 schools (37.99%).

**2.50** Using technology and the data in **CollegeScores** we find the distribution below for the five types of *MainDegree*.

| MainDegree | Count | Percent |
|---|---|---|
| 0 | 4 | 0.07 |
| 1 | 2689 | 43.79 |
| 2 | 1141 | 18.58 |
| 3 | 2012 | 32.76 |
| 4 | 295 | 4.80 |
| N= | 6141 | |

The most frequent type of primary degree is the certificate (1) found at 2689 of the 6141 schools (43.79%).

**2.51** Using technology and the data in **CollegeScores** we find the two-way table below for *MainDegree* and *Control*. Since we want to see the distribution within each type of *Control* (row), the table also shows the row percentages below each count.

```
 Rows: Control   Columns: MainDegree
               0      1      2      3      4     All
Private        4    175    167   1243    243    1832
            0.22   9.55   9.12  67.85  13.26  100.00
Profit         0   1906    225    170     32    2333
            0.00  81.70   9.64   7.29   1.37  100.00
Public         0    608    749    599     20    1976
            0.00  30.77  37.90  30.31   1.01  100.00
All            4   2689   1141   2012    295    6141
            0.07  43.79  18.58  32.76   4.80  100.00
```

The most common main degree for Private schools is the bachelor's degree (3) with 67.85%.
The most common main degree for Profit schools is the certificate (1) with 81.70%. Public schools are the most evenly spread, with the highest percentage being 37.90% for associate's degrees (2).

**2.52** Using technology and the data in **CollegeScores** we find the two-way table below for *MainDegree* and *Control*. Since we want to see the distribution within each type of *MainDegree* (column), the table also shows the column percentages below each count.

```
Rows: Control   Columns: MainDegree
                0       1       2       3       4     All
Private         4     175     167    1243     243    1832
           100.00    6.51   14.64   61.78   82.37   29.83
Profit          0    1906     225     170      32    2333
             0.00   70.88   19.72    8.45   10.85   37.99
Public          0     608     749     599      20    1976
             0.00   22.61   65.64   29.77    6.78   32.18
All             4    2689    1141    2012     295    6141
           100.00  100.00  100.00  100.00  100.00  100.00
```

The most common type of control for schools with mainly
No degrees (0) is Private (100.00%)
Certificates (1) is Profit (70.88%)
Associates (2) is Public (65.64%)
Bachelors (3) is Private (61.78%)
Graduates (4) is Private (82.37%)

**2.53** 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**

**2.54** Histograms A and H are both skewed to the left.

**2.55** Only histogram F is skewed to the right.

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

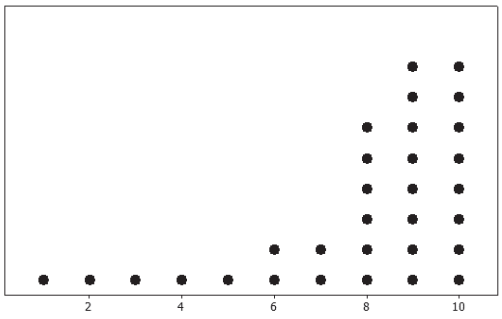
**2.57** While all of B, C, D, E, and G are approximately symmetric, only B, C, and E are also bell shaped.

**2.58** 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.

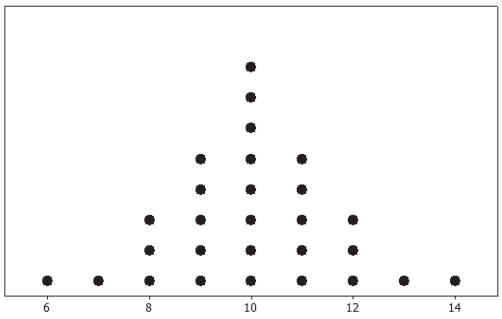
**2.59** Histograms E and G are both approximately symmetric, so the mean and median will be approximately equal. Histogram F is skewed right, so the mean should be larger then the median; while histogram H is skewed left, so the mean should be smaller then the median.

**2.60** 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.

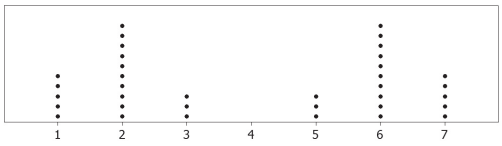
**2.61** There are many possible dotplots we could draw that would be clearly skewed to the left. One is shown.



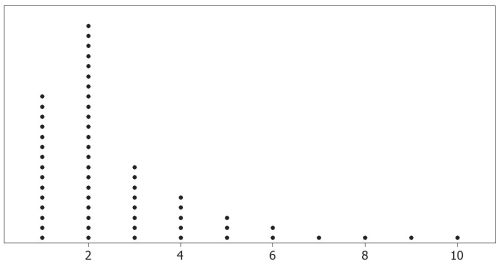

**2.62** There are many possible dotplots we could draw that are approximately symmetric and bell-shaped. One is shown.




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




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




**2.65** (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.

**2.66** (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.

**2.67** (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.

**2.68** (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.

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

**2.70** This mean is from a sample, so the notation is $\overline{x}$.

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

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

**2.73** (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:

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

**2.74** (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.

**2.75** (a) This is a mean. Since number of cats owned is always a whole number, a median of 2.39 is impossible.

(b) Since this is a right-skewed distribution, we expect the mean to be greater than the median.

**2.76** 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.

**2.77** (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 at least one outlier at just under 40 million. (The outlier represents the state of California.) The value between 28 and 30 million might also considered be an outlier.

(c) The median splits the data in half. The first two bars represent 23 of the 50 states, so the median will be somewhere in the third group, perhaps just under 5 million. (In fact, the median is 4.57 million.)

(d) The mean is the balance point for the histogram and is harder to estimate. It should be larger than the median because of the right skew. A bit over 6 million might be a good guess. (In fact, it is 6.5 million.)

**2.78** (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.

**2.79** (a) The distribution is skewed to the left since there are many values between about 74 and 84 and then a long tail going down to the low values to the left.

(b) Since half the values are above 74, the median is about 74. (The actual median is 74.3.)

(c) Since the data is skewed to the left, the mean will be less than the median so the mean will be less than 74. (The actual mean is 72.47.)

**2.80** (a) Since there are lots of small numbers and a few very large numbers, the distribution is skewed to the right.

(b) Since the few large numbers pull the mean up, the mean is 5.3 and the median is 1.

**2.81** The speed values tend to be larger than the distance values and the speed distribution is right-skewed, so the speed mean should be larger than the speed median. This implies that the speed mean is 632 ypm and the median is 575 ypm. The distance distribution is left-skewed, so its mean should be smaller than its median. Thus the distance mean is 409 miles and the median is 429 miles.

**2.82** (a) Using $N$ and $V$ for the subscripts to indicate noun or verb, we have $\overline{x}_N = 3.21$ and $\overline{x}_V = 3.67$.

(b) We have $\overline{x}_N - \overline{x}_V = 3.21 - 3.67 = -0.46$.

(c) Because the participants were randomly assigned to one of the two conditions, this is a randomized experiment.

**2.83** (a) The mean number of minutes on the treadmill for the mice receiving young blood is $\overline{x}_Y = 56.76$ minutes.

(b) The mean number of minutes on the treadmill for the mice receiving old blood is $\overline{x}_O = 34.69$ minutes.

(c) We see that $\overline{x}_Y - \overline{x}_O = 56.76 - 34.69 = 22.07$. The mice receiving young blood were able to run on the treadmill for 22 minutes longer, on average, than the mice receiving old blood.

(d) This is a randomized comparative experiment, as the mice were randomly assigned to the two groups.

(e) Yes, we can conclude causation since the data come from an experiment.

**2.84** (a) Since this is a sample, we use the notation $\overline{x}$ for the means. We use subscripts 1 and 2 for smartphone and desktop, respectively (or we could use $S$ and $D$ to be more clear). Using 1 and 2, we have $\overline{x}_1 = 230$ and $\overline{x}_2 = 120$.

(b) The difference in means is $230-120=110$ and the notation is $\overline{x}_1-\overline{x}_2$ so we have $\overline{x}_1-\overline{x}_2=230-120=110$.

**2.85** 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.

**2.86** The Canadian census includes all Canadian adults, so this is a mean for a population and the correct notation is $\mu$. We have $\mu=2.4$ people.

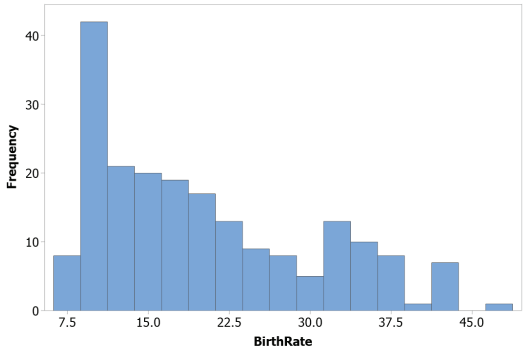
**2.87** (a) There are many possible answers. One way to force the outcome is to have a very small outlier, such as
2, 51, 52, 53, 54.
The median of these 5 numbers is 52 while the mean is 42.4.

(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.

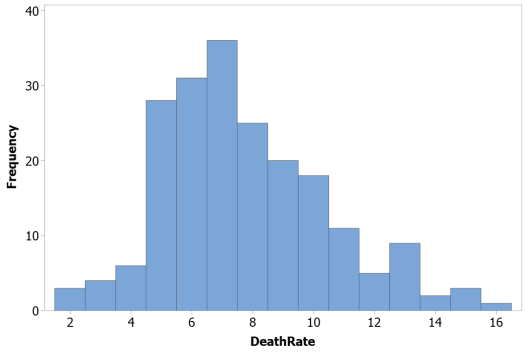
**2.88** There are many possible answers. Any variable that measures something with large outliers will work.

**2.89** The histogram is shown below.



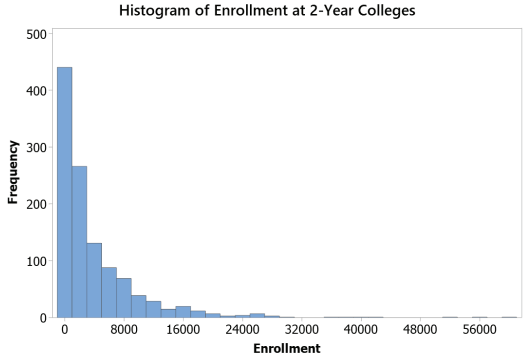

We see that it is strongly skewed to the right.

**2.90** The histogram is shown below.




We see that it is mildly right skewed.

**2.91** (a) Here is a histogram for the *Enrollment* variable in **CollegeScores2yr**.

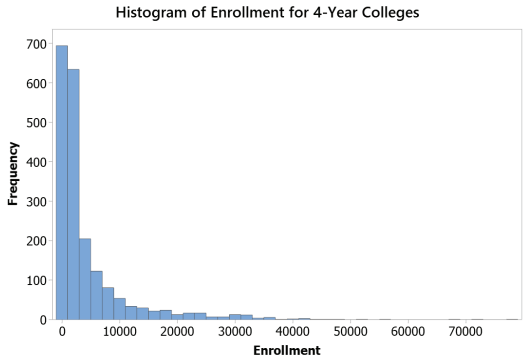



The distribution shows a very strong skew to the right.

(b) Using technology, we find the mean enrollment at two-year colleges is 4,050 students and the median is 1,748 students.

(c) As expected for the strong right skew in the graph, the mean enrollment is much larger than the median.

**2.92** (a) Here is a histogram for the *Enrollment* variable in **CollegeScores4yr**.



The distribution shows a very strong skew to the right.

(b) Using technology, we find the mean enrollment at four-year colleges is 4,485 students and the median is 1,722 students.

(c) As expected for the strong right skew in the graph, the mean enrollment is much larger than the median.

**2.93** (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, 213 had 1 child, and 344 had 2 children, so $162 + 213 + 344 = 719$ 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**

**2.94** (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.

**2.95** (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.

**2.96** (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.

**2.97** (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.

**2.98** (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.

**2.99** (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.

**2.100** 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.

**2.101** 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

**2.102** 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

**2.103** 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.

**2.104** 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.

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

**2.106** 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 $(440, 485, 500, 515, 560)$.

**2.107** 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.

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

**2.109** We see that the minimum value is 58 and the maximum is 77. We can count the dots to find the value at the 25th-percentile, the 50th-percentile, and the 75th-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).

**2.110** This dataset is very symmetric.

**2.111** 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.

**2.112** 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.

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

**2.114** 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.

**2.115** 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.

**2.116** 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.

**2.117** 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.

**2.118** 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.

**2.119** 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.

**2.120** 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.

**2.121** 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.

**2.122** (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 technology, we see $\overline{x} = 52.875$ and $s = 5.07$.

**2.123** (a) We see in the computer output that the five number summary is (119, 162, 171, 180, 210).

(b) The range is the difference between the maximum and the minimum, so we have Range $= 210 - 119 = 91$. The interquartile range is the difference between the first and third quartiles, so we have $IQR = 180 - 162 = 18$.

(c) The 30th-percentile is between the first quartile and the median, so it will be between 162 and 171. The 80th-percentile is between the third quartile and the maximum, so it will be between 180 and 210.

**2.124** (a) We see in the computer output that the mean armspan is $\overline{x} = 170.76$ cm and the standard deviation is $s = 13.46$ cm.

(b) We see that the largest value is 210, so we compute the $z$-score as:

$$z\text{-score} = \frac{x - \overline{x}}{s} = \frac{210 - 170.76}{13.46} = 2.915$$

The maximum armspan in this sample is 2.915 (or almost 3) standard deviations above the mean. We compute the $z$-score for the smallest armspan, 119, similarly:

$$z\text{-score} = \frac{x - \overline{x}}{s} = \frac{119 - 170.76}{13.46} = -3.845$$

The minimum armspan in this sample is 3.845 standard deviations below the mean. The minimum would probably be considered an outlier.

(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:

$$\overline{x} - 2 \cdot s = 170.76 - 2(13.46) = 143.84 \qquad \text{and} \qquad \overline{x} + 2 \cdot s = 170.76 + 2(13.46) = 197.68$$

We expect about 95% of the data to lie between 143.84 and 197.68. (In fact, the 95% rule is very accurate in this case, since 397 of the 415 armspan values lie between these two numbers, which is 95.7%.)

**2.125** (a) We see in the computer output that the mean obesity rate is $\mu = 31.43\%$ and the standard deviation is $\sigma = 3.82\%$.

(b) We see that the largest value is 39.5, so we compute the $z$-score as:

$$z\text{-score} = \frac{x-\mu}{\sigma} = \frac{39.5-31.43}{3.82} = 2.11$$

The maximum of 39.5% obese (which occurs for both Mississippi and West Virginia) is slightly more than two standard deviations above the mean.
We compute the $z$-score for the smallest percent obese, 23%, similarly:

$$z\text{-score} = \frac{x-\mu}{\sigma} = \frac{23.0-31.43}{3.82} = -2.21$$

The minimum of 23.0% obese, from the state of Colorado, is about 2.2 standard deviations below the mean. The minimum might be considered a mild outlier.

(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:

$$\mu - 2\sigma = 31.43 - 2(3.82) = 23.79 \qquad \text{and} \qquad \mu + 2\sigma = 31.43 + 2(3.82) = 39.07$$

We expect about 95% of the data to lie between 23.79% and 39.07%. 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 on the low side and West Virginia and Mississippi on the high side.)

**2.126** (a) We see in the computer output that the five number summary is (23.0, 28.6, 30.9, 34.4, 39.5).

(b) The range is the difference between the maximum and the minimum, so we have Range $= 39.5 - 23.0 = 16.5$. The interquartile range is the difference between the first and third quartiles, so we have $IQR = 34.4 - 28.6 = 5.8$.

(c) The 15th-percentile is between the minimum and the first quartile, so it will be between 23.0 and 28.6. The 60th-percentile is between the median and the third quartile, so it will be between 30.9 and 34.4.

**2.127** (a) The z-score for Brazil will be positive because the value for Brazil is higher than the mean.

(b) $z = (6.24 - 4.8)/1.66 = 0.87$

(c) The range is $max - min = 12.46 - 1.46 = 11.0$.

(d) The IQR is $Q3 - Q1 = 5.61 - 3.76 = 1.85$.

**2.128** (a) We use technology to see that the mean is 61.56 hot dogs and the standard deviation is 8.83.

(b) Over the 18 year period, 10 of the years had counts above the mean of 61.56 hot dogs. Only two of these were in the first nine years (2007 and 2009); however eight of the nine later years (all but 2014) were above the mean. People seem to be getting better at eating hot dogs!

**2.129** (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.

**2.130** (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.

**2.131** (a) The 10th 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 75th 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 $4.0 - 2.0 = 2.0$.

**2.132** The histogram is relatively symmetric and bell-shaped. The mean appears to be approximately 440 billion dollars. To estimate the standard deviation, we estimate an interval centered at 440 that contains approximately 95% of the data. The interval from about 320 to 560 appears to contain almost all the data. Since 320 is 120 units below the mean around 440 and 560 is 120 units above the mean, by the 95% rule we estimate that 2 times the standard deviation is about 120, so the standard deviation appears to be approximately 60 billion dollars. Note that we can only get a rough approximation from the histogram. To find the exact values of the mean and standard deviation, we would use technology and all the values in the dataset.

**2.133** Using technology, we see that the mean is 427.8 billion dollars and the standard deviation is 61.0 billion dollars. The 95% rule says that 95% of the data should be within two standard deviations of the mean, so the interval is:

$$\begin{array}{rcl} \text{Mean} & \pm & 2 \cdot \text{StDev} \\ 427.8 & \pm & 2 \cdot (61.0) \\ 427.8 & \pm & 122.0 \\ 305.8 & \text{to} & 549.8 \end{array}$$

We expect 95% of US monthly retail sales over this time period to be between 305.8 billion dollars and 549.8 billion dollars.

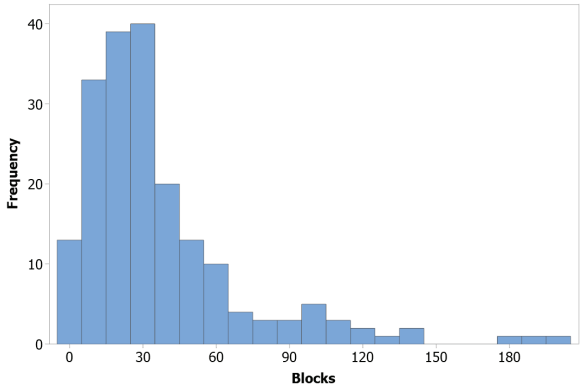
**2.134** The data is not at all bell-shaped, so it is not appropriate to use the 95% rule with this data.

**2.135** (a) Using technology, we see that the mean is 37.0 blocks in a season and the standard deviation is 34.2 blocks.

(b) Using technology, we see that the five number summary is $(0, 16, 28, 45, 199)$.

(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, 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.

**2.136** 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

$$z\text{-score for } FGPct = \frac{0.442 - 0.463}{0.057} = -0.37 \qquad z\text{-score for } Points = \frac{2818 - 981}{478} = 3.84$$

$$z\text{-score for } Assists = \frac{586 - 219}{148} = 2.48 \qquad z\text{-score for } Steals = \frac{158 - 63.5}{31.2} = 3.03$$

The most impressive statistic is his total points, which is about 3.84 standard deviations above the mean. The least impressive is his field goal percentage, which is 0.37 standard deviations below the mean. Although just a bit below average in field goal percentage, he is well above the mean for the other three variables.

**2.137** (a) We calculate z-scores using the summary statistics for each: Reading and Writing $= \frac{600-536}{102} = 0.627$, Math $= \frac{600-531}{114} = 0.605$.

(b) Stanley's more unusual score was in the Reading and Writing component, since he has the higher z-score in this section.

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

**2.138** (a) Using technology, we find that the mean internet speed is $\overline{x} = 24.86$ (Mb/sec) a and the standard deviation is $s = 10.92$.

(b) Using technology, we find that the mean hours online is $\overline{x} = 5.69$ and the standard deviation is $s = 1.54$.

(c) There may be some relationship between internet speeds and hours online. Brazil has a much slower internet speed and the highest time online by considerable margin, but the pattern does not seem so obvious with the other countries.

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

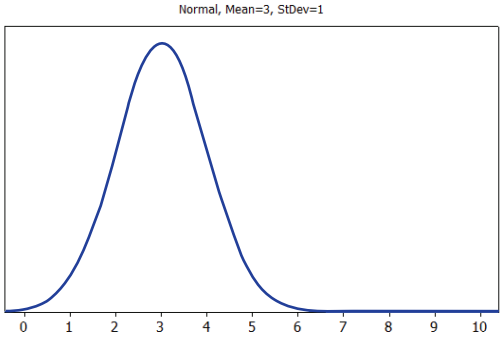
(b) The maximum of 6662 is clearly an outlier and we expect it to pull the mean above the median. Since the median is 1667, the mean should be larger than 1667, but not too much larger. The mean of this data set is 1796.

(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.

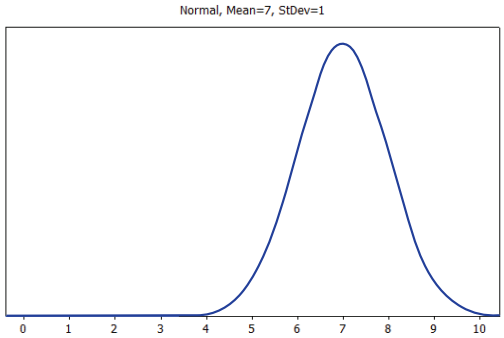
**2.140** (a) The smallest possible standard deviation is zero, which is the case if the numbers don't deviate at all from the mean. The dataset is:
5, 5, 5, 5, 5, 5.

(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.

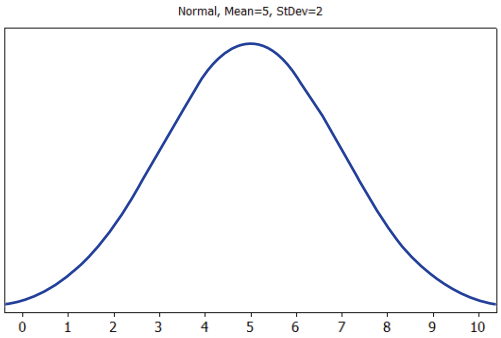
**2.141** A bell-shaped distribution with mean 3 and standard deviation 1.



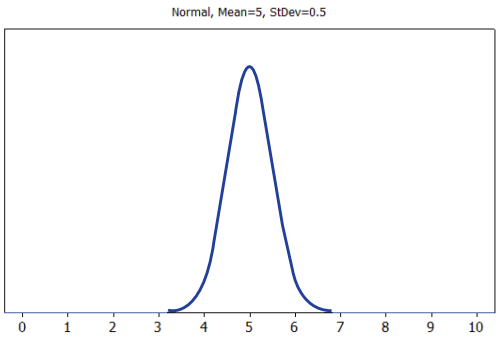

**2.142** A bell-shaped distribution with mean 7 and standard deviation 1.




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




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




**2.145** Using technology, we see that the mean rating is 57.58, the standard deviation is 28.06, and the five number summary is (0, 33, 61, 84, 99).

**2.146** Using technology, we see that the mean audience rating is 62.18, the standard deviation is 18.21, and the five number summary is (10, 49, 64, 77, 99).

**2.147** Using technology, we see that the sample size is $n = 1412$, the mean distance is $\overline{x} = 408.7$, the standard deviation is $s = 78.1$, and the five number summary is (137, 375, 429, 448, 577). (Note that different stat packages may give slightly different values for the quartiles.)

**2.148** (a) Using technology the mean braking distance is $\overline{x} = 130.8$ feet and the standard deviation is $s = 6.92$ feet.

(b) An interval for $\overline{x} \pm 2 \cdot s = 130.8 \pm 2 \cdot 6.92 = 130.8/pm13.8$ goes from 117.0 to 144.6 feet. There are five cars with braking distance less than 117.0 feet: Corvette (107), Porsche 911 (108), BMW Z4 (111), Camaro (112), and Audi TT (113). Only one car, the Toyota Sequoia (146), has a braking distance more than two standard deviations above the mean.

**2.149** (a) Using technology we find the following summary statistics for the two variables, which include standard deviation (StDev), range, and interquartile range (IQR) for each variable.

```
Variable   N  Mean  StDev Minimum   Q1 Median  Q3 Maximum Range  IQR
CityMPG  110 16.19   3.74   10.0  13.0   15.5 19.0    28.0   18.0  6.0
HwyMPG   110 34.00   7.19   21.0  28.0   32.5 39.0    54.0   33.0 11.0
```

(b) The range (18), IQR (6) and standard deviation (3.74) for *CityMPG* are all smaller than the respective range (33), IQR (11), and standard deviation (7.193) for *HwyMPG*.

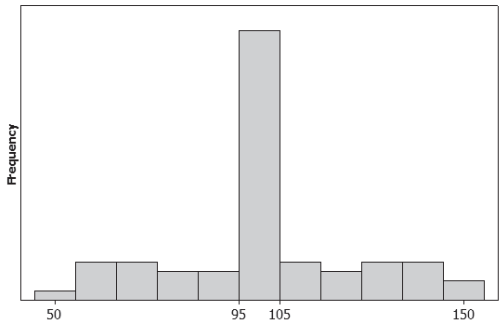
**2.150** (a) Using technology we find that $\overline{x} = 10.29$ hours and $s = 10.19$ hours.

(b) The computed $z$-score for a homework hours value of $x = 80$ is

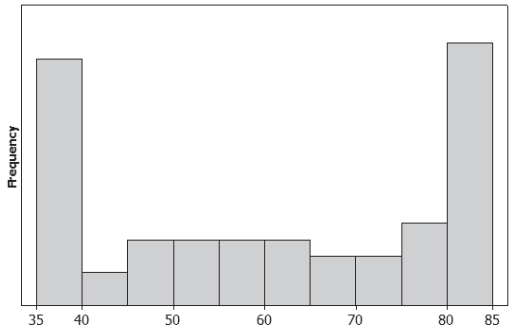
$$z = \frac{x - \overline{x}}{s} = \frac{80 - 10.29}{10.19} = 6.84$$

This amount of homework claimed by this student is 6.84 standard deviations above the mean for this sample. While this is not impossible, it would be very surprising to find a student working 80 hours per week on homework!

**2.151** (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.




**2.152** (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**

**2.153** 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.

**2.154** 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.

**2.155** (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.

**2.156** (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.

**2.157** (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.

**2.158** (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.

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

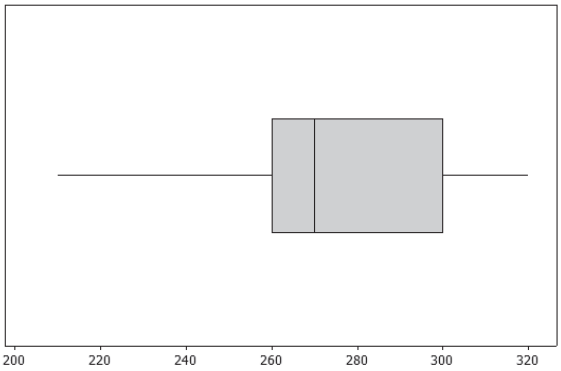
$$Q_1 - 1.5(IQR) = 260 - 1.5(40) = 200,$$

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:




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

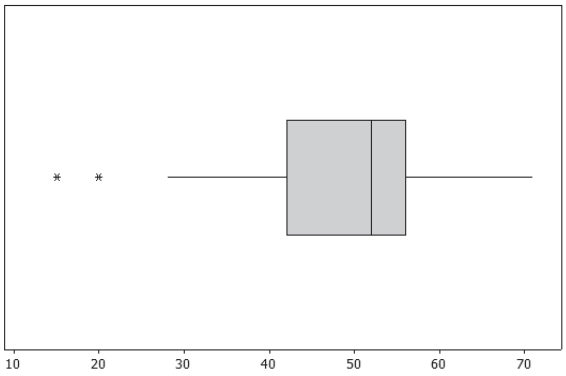
$$Q_1 - 1.5(IQR) = 42 - 1.5(14) = 21,$$

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.




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

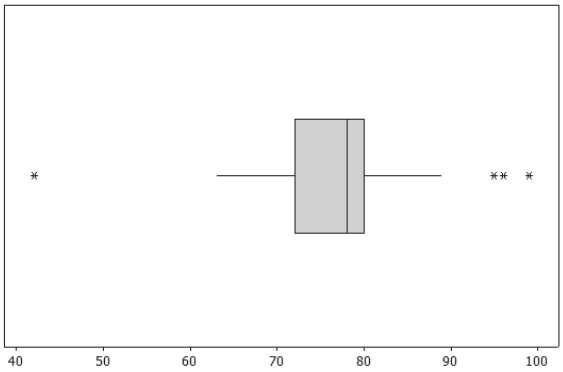
$$Q_1 - 1.5(IQR) = 72 - 1.5(8) = 60,$$

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.




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

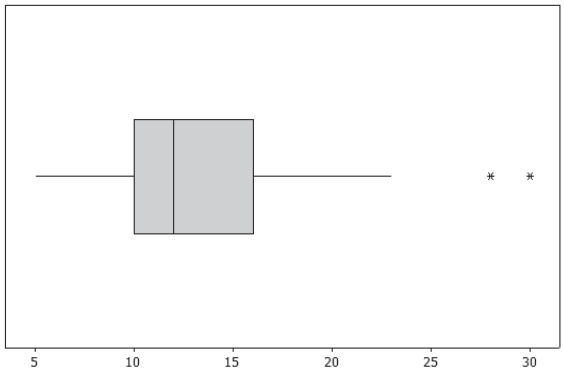
$$Q_1 - 1.5(IQR) = 10 - 1.5(6) = 1,$$

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.




**2.163** (a) The distribution is left-skewed.

(b) Half of all literacy rates are between about 73% and 98%.

(c) The five number summary is about (15, 73, 92, 98, 100).

(d) The true median is most likely lower than that shown here, because data is more likely to be available for developed countries, which have higher literacy rates.

**2.164** (a) The median runtime for the mice receiving old blood appears to be about 29 minutes.

(b) This data appear to be skewed to right, so we expect the mean to be larger than the median.

(c) The mice receiving young blood appear to be able to run for much longer.

(d) Yes, there appears to be an association between these variables since the runtimes are generally much larger for those mice receiving young blood.

(e) No, there are no outliers in either group.

**2.165** (a) The explanatory variable is the group the person is in, and it is categorical. The response variable is hippocampus volume, and it is quantitative.

(b) The control group of people who never played football appears to have the largest hippocampal volume, while the football players with a history of concussion appears to have the smallest.

(c) Yes, there are two outliers (one high and one low) in the group of football players with a history of concussion.

(d) The third quartile appears to be at about 7000 $\mu$L.

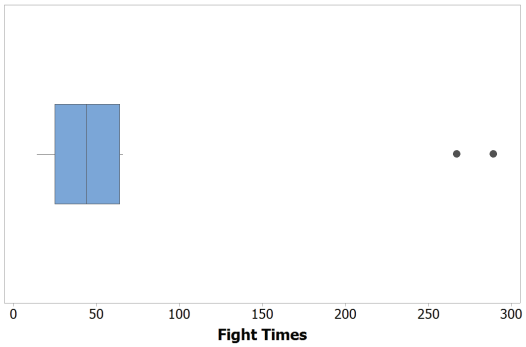
(e) Yes, there is a quite obvious association, with those playing football having smaller brain hippocampus volume and those playing football with concussions having even smaller volume.

(f) No, we cannot conclude causation as the data come from an observational study and not an experiment. There are many possible confounding variables.

**2.166** (a) We see that $IQR = 64 - 25 = 39$ so $Q_3 + 1.5(IQR) = 64 + 1.5(39) = 122.5$. We see that there are two high outliers, at 289 seconds and 267 seconds.

(b) We see that $Q_1 - 1.5(IQR) = 25 - 1.5(39) = -33.5$. There are definitely no low outliers!

(c) See the figure.



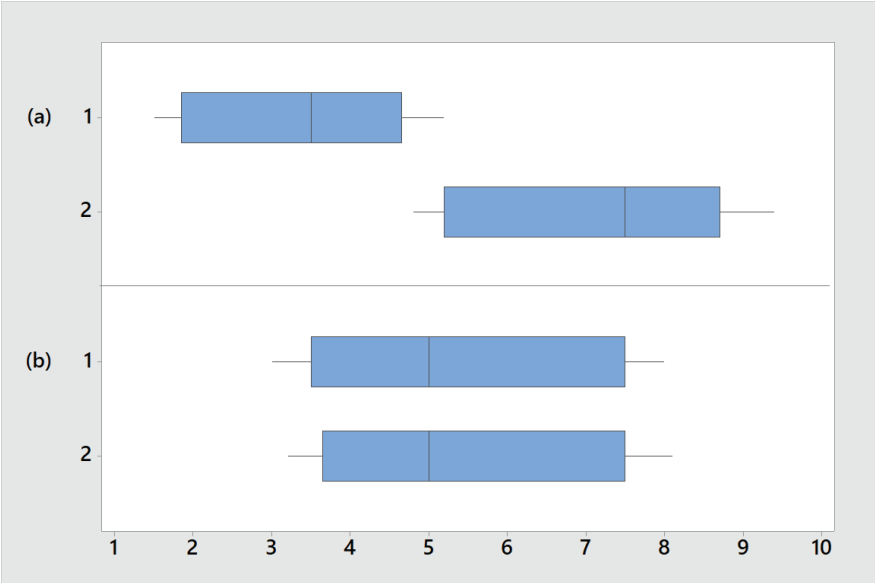

(d) Due to the right skew of the distribution, and large positive outliers we would expect the mean to be greater than the median.

**2.167** Recall that two variables are associated if values of one variable tend to be related to values of the other variable. In this case, it means that knowing whether a case is in group 1 or group 2 gives us information about the likely value of the quantitative variable for that case.

(a) Dataset $C$ shows the strongest association between the two variables since the values of the quantitative variable are quite different between group 1 and group 2.

(b) Dataset $A$ shows no evidence of an association between the two variables, since values of the quantitative variable are virtually the same between the two groups.

**2.168** Recall that two variables are associated if values of one variable tend to be related to values of the other variable. In this case, it means that knowing what category a case is in, in the categorical variable, gives useful information about what the likely values are for the quantitative variable. Possible answers are shown below. Answers may vary.




**2.169** (a) Most of the data is between 0 and 500, and then the data stretches way out to the right to some very large outliers. This data is very much skewed to the right.

(b) The data appear to range from about 0 to about 10,000, so the range is about $10000 - 0 = 10000$.

(c) The median appears to be about 250 (although this is difficult to estimate precisely due to the scale of the graph). About half of all movies recover less than 250% of their budget, and half recover more than 250% 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 268.9 while the mean is 435.7.)

**2.170** 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 10, there are no low outliers.

**2.171** (a) Action movies appear to have the largest budgets, while horror movies appear to have the smallest budgets.

(b) Action movies have by far the biggest spread in the budgets, with horror movies and comedies appearing to have the smallest spread (but not much smaller than dramas).

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

**2.172** (a) The highest mean is in Drama (69.52), while the lowest mean is in Horror movies (43.56).

(b) The highest median is in Drama (75), while the lowest median is in Horror movies (38.5).

(c) The lowest score is 18 and it is for a Horror movie. The highest score is 97, and it is obtained by a Drama.

(d) The genre with the largest number of movies is Drama, with $n = 52$.

(e) We see that $\overline{x}_C = 55.64$ while $\overline{x}_H = 43.56$, so the difference in mean score between the two types of movies is $\overline{x}_C - \overline{x}_H = 55.64 - 43.56 = 12.08$.

**2.173** (a) The lowest level of physical activity appears to be in the South, and the highest, in general, appears to be in the West (although the highest individual state is in the Northeast).

(b) There are high outliers in the Midwest (Wisconsin) and Northeast (Vermont) and a low outlier in the West (Nevada).

(c) Yes, the boxplots are very different between the different regions. All of the states except the outlier in the West are larger than all the values in the South. The Midwest and Northeast tend to be between these two extremes.

**2.174** (a) The median corresponds to the middle line in each box. These are at about the same place, around 1380 hits, for both leagues. In fact, the actual medians from the data are 1379 (AL) and 1378 (NL). These are remarkably close.

(b) Although the medians are similar, the other values of the five number summary (minimum, maximum, $Q_1$, and $Q_3$) are all smaller for the National League. The boxplot is more symmetric for the National League and appears to have a bit less variability.

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

**2.176** (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:

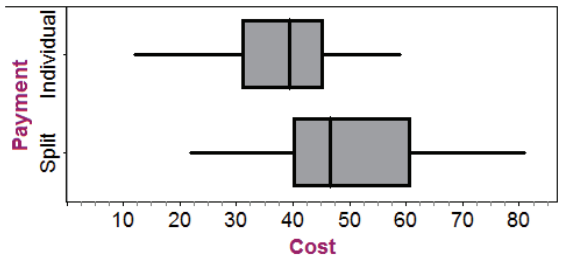
$$z\text{-score for } 185 = \frac{x - \text{Mean}}{\text{Standard deviation}} = \frac{185 - 30}{49.4} = 3.138$$

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.

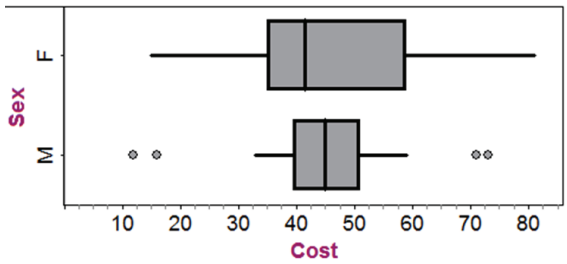
**2.177** (a) The five number summaries for Individual (12, 31, 39.5, 45.5, 59) and Split (22, 40, 46.5, 61, 81) show that costs tend to be higher when subjects are splitting the bill. This is also true for the means ($\overline{x}_I = 37.29$ vs $\overline{x}_S = 50.92$), while the standard deviations and IQRs show slightly more variability for those splitting the bill ($s_I = 12.54$ and $IQR_I = 14.5$ vs $s_S = 14.33$ and $IQR_S = 21$).

(b) Side-by-side plot show the distributions of costs tend to be relatively symmetric for both groups, but generally higher and slightly more variable for those splitting the bill.

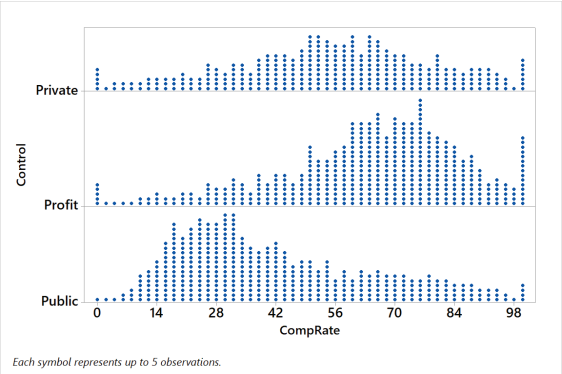



**2.178** (a) The five number summaries for females (15, 35, 41.5, 59, 81) and males (12, 39.5, 45, 51, 73) are fairly similar with a somewhat larger interquartile range for the females (24 vs 11.5). The means ($\overline{x}_f = 44.46$ vs $\overline{x}_m = 43.75$) and standard deviations ($s_f = 15.48$ vs $s_m = 14.81$) are similar for both groups.

(b) Side-by-side boxplots show the distributions of costs tend to be relatively symmetric for both females and males, although the smaller IQR for males produces 2 outlier values at each end of the distribution. Both distributions are centered at roughly the same point (a bit over 40 shekels).



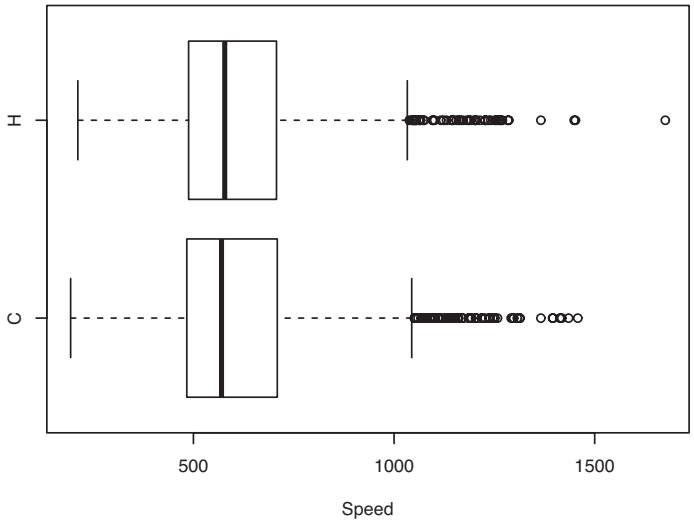

**2.179** Here are side-by-side dotplots to compare the completion rates between the three types of control.




We see that all three distributions include values for 0% to 100% (and frequently have a cluster of values at those extremes). The private schools have a relatively symmetric distribution with a center that is below that of profit schools, but well above the public schools. The for profit schools show a left skew, while the public schools are right skewed. The ranges are the same (100) and standard deviations are likely to be similar.

Note, you might also compare side-by-side boxplots or histograms.

**2.180** Using technology, we create the side-by-side boxplots as in the figure below.




(a) There are high outliers, but no low outliers for the hens.

(b) There are high outliers, but no low outliers for the cocks.

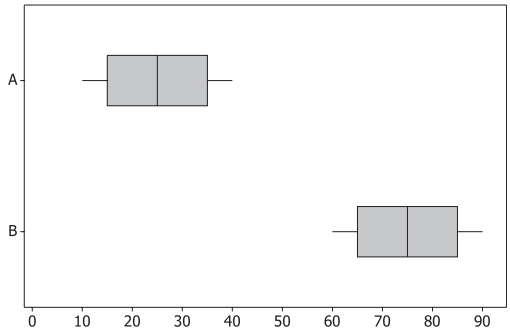
(c) The speed boxplots are very similar for hens and cocks. Both are right skewed with lots of high outliers, and very similar five number summaries (expect for the hen with the fastest speed). There does not appear to be an association between sex and speed for homing pigeons.

**2.181** Using technology we find some summary statistics for the *CompRate* variable in **CollegeScores** as shown below.

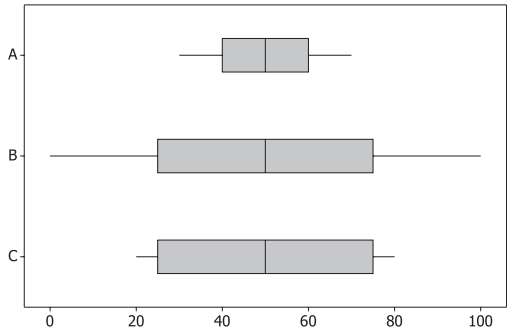
```
Variable:CompRate
Control      N   N*    Mean   StDev  Minimum      Q1  Median      Q3  Maximum
Private   1432  400  55.370  22.849    0.000  40.657  56.560  70.815  100.000
Profit    2096  237  64.229  20.928    0.000  52.940  66.670  78.570  100.000
Public    1908   68  41.826  22.657    0.000  24.243  35.960  56.813  100.000
```

Comparing means (or medians) between the types of control we see that for profit schools tend to have higher completion rates (mean=64.2, median=66.7), while public schools are lower (mean=41.8, median=36.0) and private schools tend to be in between (mean=55.4, median=56.6). The ranges are all equal to 100 and the standard deviations are similar (22.8, 20.9, and 22.7). Note, however, that a sizable number of schools (especially private ones) have missing values (as shown in the column labeled N*).

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




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




**2.184** Answers will vary.

**2.185** Answers will vary.

**Section 2.5 Solutions**

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

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

**2.188** 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).

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

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

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

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

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

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

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

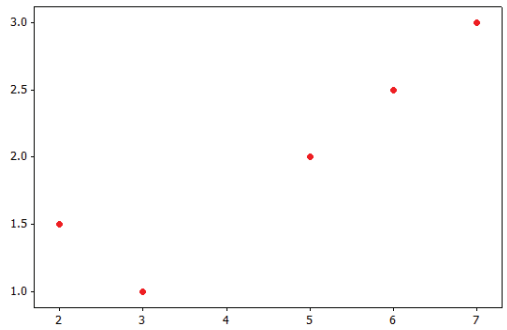
**2.196** We wear more clothes when it is cold outside, so as the temperature goes down, the amount of clothes worn goes up. This describes a negative association.

**2.197** 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.

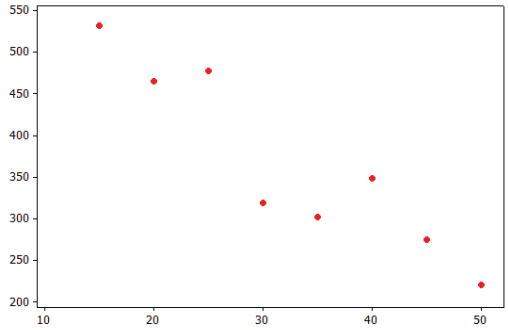
**2.198** Usually there are not many people in a heavily wooded area and there are not many trees in a heavily populated area such as the middle of a city. This would be a negative relationship.

**2.199** 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.

**2.200** See the figure below.



**2.201** See the figure below.




**2.202** The correlation is $r = 0.915$.

**2.203** The correlation is $r = -0.932$.

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

**2.205** (a) The two variables are amount of iron in the soil and amount of potassium in spinach grown in that soil. Amount of iron in the soil is the explanatory variable since it appears to impact the amount of potassium in the spinach.

(b) Since higher amounts of iron imply lower levels of potassium, this is a negative association.

(c) We know that 840 mg is the average amount of potassium. If soil has high levels of iron, we expect the amount of potassium to be less than 840 mg.

(d) If soil has low levels of iron, we expect the amount of potassium to be greater than 840 mg.

**2.206** (a) (i) This is a positive association. (ii) This is a negative association.

(b) No, we cannot conclude a cause-and-effect relationship since the results do not come from an experiment. (However, additional experiments do seem to indicate causation in this case.)

**2.207** (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.

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

**2.209** (a) The association appears to be positive. This means that higher levels of self-reported depression are associated with higher levels of clinical levels of depression. This makes sense since both scales are meant to measure approximately the same thing.

(b) i. A case in the upper left would represent a person who has low levels of self-reported depression symptoms but who scores high on the clinical depression assessment.

ii. A case in the upper right would represent a person who scores high on both the self-reported scale and the clinical scale.

iii. A case in the lower left would represent a person who scores low on both the self-reported scale and the clinical scale.

iv. A case in the lower right would represent a person who has high levels of self-reported depression symptoms but who has a low score on the clinical depression assessment.

(c) For the case farthest to the right, the self-reported score is approximately 54 and the clinical score is approximately 27.

**2.210** All the cases are male-female married couples.

(a) A case in the top left corner of the scatterplot represents a couple with an old husband and a young wife.

(b) A case in the top right corner of the scatterplot represents an old couple with an old husband and an old wife.

(c) A case in the bottom left corner of the scatterplot represents a young couple with a young husband and a young wife.

(d) A case in the bottom right corner of the scatterplot represents a couple with an young husband and an old wife.

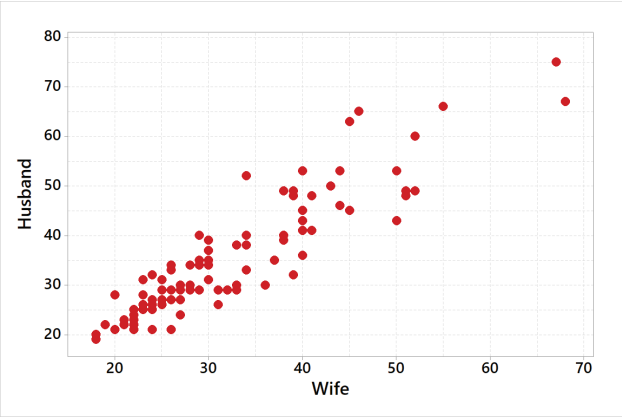
**2.211** (a) BMI, cortisol level, depression score, and heart rate are all positively correlated with social jetlag, while weekday hours of sleep and physical activity are negatively correlated.

(b) No, we cannot conclude causation since this is an observational study and not an experiment.

**2.212** (a) We see that years of football appears to have a stronger association with brain hippocampal volume than with the cognitive percentile score, so the correlation $-0.465$ goes with Graph (b) while the correlation $-0.366$ goes with Graph (a).

(b) More years playing football tends to be associated with smaller brain size and a lower cognitive percentile score.

**2.213** The scatterplot is shown.



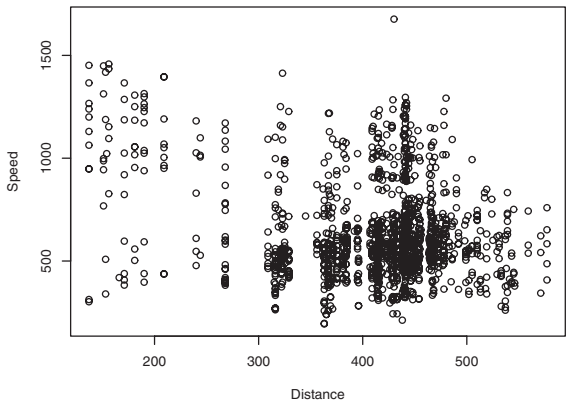

(a) We see on the scatterplot that one other 50-year-old woman got married during this time period. Her husband was about 43 at the time of the marriage.

(b) The oldest man to get married during this time period was 75 years old. The youngest was about 19 years old.

(c) There are two dots to the right of 65 on the scatterplot, so 2 of the women who got married during this time period were older than 65. There are three dots to the left of 20, so 3 of the women who got married during this time period were less than 20 years old.

**2.214** Using the dataset **MarriageAges** and *StatKey* or other technology, we see that the correlation is 0.914. This is a sample correlation, so the notation is $r$ and we have $r = 0.914$.

**2.215** Using technology, we create the scatterplot as in the figure below.




(a) There appears to be a negative association between distance and speed. The pigeons flying shorter distances (under 300 miles) tend to have more speeds above 750 ypm, while the pigeons flying longer distances more often have speeds below 750 ypm.

(b) From the scatterplot, the highest value for speed appears to be a pigeon that flew for about 430 miles. This value can be confirmed by looking at the first case in the dataset, because the pigeons are ordered from highest to lowest speed.

(c) Using technology, we find that the correlation between speed and distance for these 1412 pigeons is $r = -0.231$. This is consistent with the negative association observed in the scatterplot.

**2.216** (a) A positive associate would indicate that teams that win more in the pre-season tend to win more in the regular season, while a negative association would indicate that teams that win more in the pre-season tend to lose more in the regular season.

(b) This is a small positive correlation, so it tells you there is a very weak positive linear relationship between these two variables.

**2.217** 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.

**2.218** Both variables in the study are categorical, not quantitative. Correlation is a statistical measure of the linear relationship between two quantitative variables.

**2.219** (a) There is a strong negative linear association between genetic diversity and distance from East Africa.

(b) The correlation is $r = -0.83$. The correlation is clearly negative, correlations must be between $-1$ and 1, and the association is relatively strong, indicating that the negative correlation is likely closer to $-1$ than 0.

(c) America has both the lowest genetic diversity and is the farthest from East Africa.

(d) Based on genetic diversity, populations closer to East Africa appear better suited to adapt to change because they have greater genetic diversity.

**2.220** (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.

**2.221** (a) A beer rated highly by both wife and husband.

(b) A beer rated low by both wife and husband.

(c) A beer rated high by the wife and low by the husband.

(d) A beer rated low by the wife and high by the husband.

**2.222** (a) The couple rated the most beers in 2011.

(b) The highest overall average was in 2019.

(c) The highest range of average ratings was in 2011.

(d) The highest average rated beer was tasted in 2011.

**2.223** (a) The correlation is 0.348.

(b) The wife rated the Pumpkin Ale by the Carolina Beer Company highest, in 2011.

(c) The husband rated Master of Pumpkins by Troegg and Trick or Treat: Chocolate Pumpkin Porter by Evil Genius Beer Company highest, in 2019.

**2.224** (a) The correlation is 0.088. The positive correlation means the ratings generally increase over time.

(b) The low outlier is Harvest Moon by Blue Moon.

(c) If the outlier were to be removed, the correlation would decrease (in fact, it turns from positive to negative).

**2.225** (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.

**2.226** (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.

**2.227** (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.

**2.228** (a) Women who are close to age 30 (specifically, ages 29, 30, 31) rate men who are the same age as they are the most attractive.

(b) Women in their 20s tend to rate older men as more attractive.

(c) Women past age 31 tend to rate younger men as more attractive.

(d) There appears to be a strong positive correlation between the variables, so the correct answer is 0.9. (The actual correlation is 0.982.)

**2.229** (a) Men who are age 20 rate women who are the same age as they are the most attractive. This is the only age for which this is true. Men of all other ages rate younger women as more attractive.

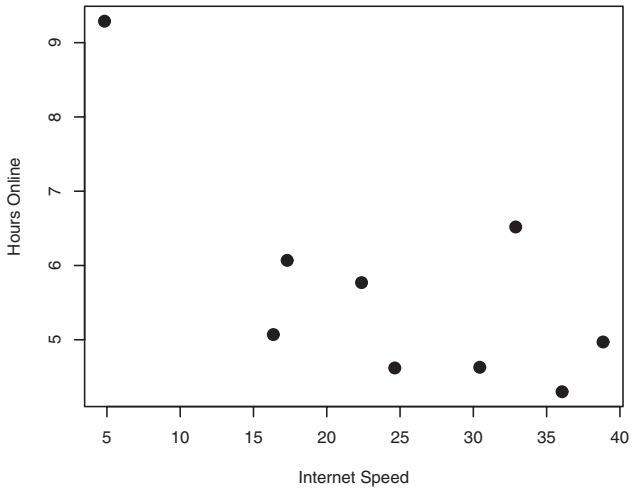
(b) Men of all ages rate women who are in their early 20s as the most attractive.

(c) There is very little association between these variables, so the best answer is 0. (The actual correlation is 0.287.)

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

(b) A negative relationship implies that as internet speed goes up, time online goes down. This might make sense because if internet 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 is one clear outlier. The point in the top left corner, corresponding to Brazil, has much lower internet speed and high number of hours online.




(d) If we ignore the outlier for Brazil, there is still a bit of a negative association, but the variables do not appear to have a strong relationship in either direction.

(e) For all nine countries, the correlation is $r = -0.704$. If we leave out Brazil, the correlation for the remaining eight countries changes to $r = -0.289$. The data point for Braazil appears to have strong influence on the correlation.

(f) No! Even with a negative correlation, these data comes from an observational study, so we cannot conclude that there is a causal association.

**2.231** (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.

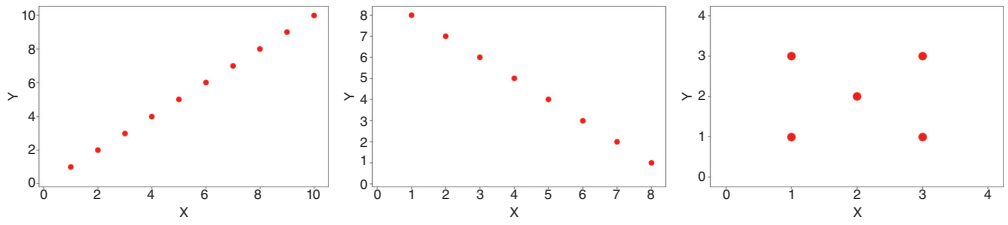
(c) The correlation is positive and close to, but not equal to, 1. A reasonable estimate is $r \approx 0.9$.

(d) There are no obvious outliers.

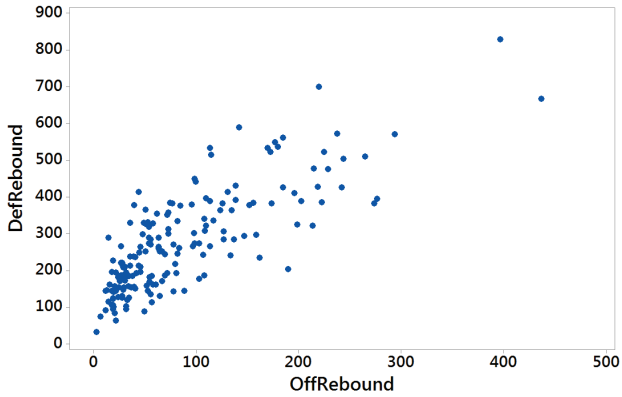
(e) The width of that iris appears to be about 11 mm.

(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).

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




**2.233** (a) See the figure.

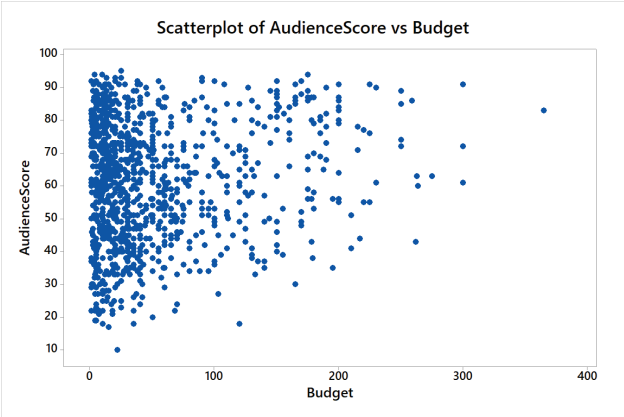



(b) There is a relatively strong positive relationship, which means players who have lots of defensive rebounds also tend to have lots of offensive rebounds. This makes sense since players who are very tall tend to get lots of rebounds at either end of the court.

(c) There are two outliers at the right with unusually high numbers of offensive rebounds. We see in the data file that these two players are Andre Drummond who has the most offensive (423) and defensive (809) rebounds and Steve Adams who has the second most offensive rebounds (391) but a more modest number (369) of defensive rebounds.

(d) We use technology to see that the correlation is 0.751. This correlation matches the strong positive linear relationship we see in the scatterplot.

**2.234** (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) The outlier has a budget of 365 million dollars. This movie is *Avengers: Age of Ultron* and the audience score is 83.

(c) The movie with the lowest audience rating (of 10) is *Just Getting Started* with a budget of 22.0 million dollars.

(d) We use technology to see that the correlation is 0.132.

**2.235** Answers will vary

**Section 2.6 Solutions**

**2.236** (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.

**2.237** (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!

**2.238** (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.

**2.239** (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.

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

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

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

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

**2.244** (a) The case with the largest negative residual appears to have a self-reported score of about 38, a clinical score of about 5, a predicted clinical score on the line of about 35. These values would give a residual of approximately $5 - 35 = -30$.

(b) For the case with a large positive residual and a self-reported score of 0, the clinical score appears to be approximately 31 and the predicted clinical score appears to be approximately 5. These values give a residual of approximately $31 - 5 = 26$.

**2.245** (a) The car with the largest positive residual has a *QtrMile* time of about 15 seconds and *Acc030* time of about 2.2 seconds. If we check the **Cars2020** data this turns out to be a Mazda Miata-MX5.

(b) The car with the most extreme negative residual has a *QtrMile* time of about 12.5 seconds and *Acc030* time of about 2.0 seconds. If we check the **Cars2020** data this turns out to be a Chevrolet Corvette.

**2.246** (a) The plot for all car models shows a stronger association between *QtrMile* times and *Acc030*. The fast accelerating (smaller times) sporty cars in the lower left show more scatter away from the line.

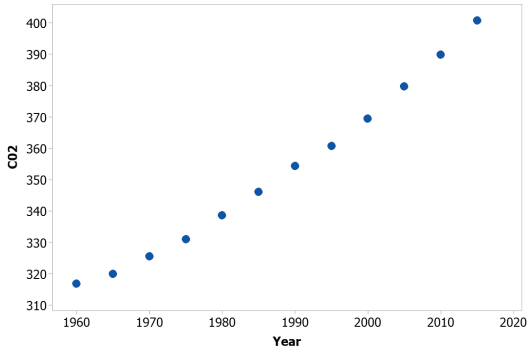
(b) The slope shown in the plot for sporty cars (2.487) is larger than the slope shown for all cars (2.219).

(c) Using technology, the correlation between *QtrMile* and *Acc030* for all the car models is $r = 0.947$. This is larger than the correlation (0.885) for just the sporty cars.

(d) The information from the correlations is more consistent with the stronger association shown in the plot of all car models, than the comparison of slopes.

**2.247** (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 $\widehat{CO_2} = -2701.2 + 1.5366(Year)$.

(e) The slope is 1.5366. Carbon dioxide concentrations in the atmosphere have been going up at a rate of about 1.5366 ppm each year.

(f) The intercept is $-2701.2$. 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 $\widehat{CO_2} = -2701.2 + 1.5366(2003) = 376.6$. This seems reasonable since the value lies between the data values for years 2000 and 2005.
In 2025, the predicted $CO_2$ concentration is $\widehat{CO_2} = -2701.2 + 1.5366(2025) = 410.4$. We have less confidence in this prediction since we can't be sure the linear trend will continue that far into the future beyond our data.

(h) In 2010, the predicted $CO_2$ concentration is $\widehat{CO_2} = -2701.2 + 1.5366(2010) = 387.37$. We see in the data that the actual concentration that year is 389.90, so the residual is $389.90 - 387.37 = 2.53$. The $CO_2$ concentration in 2010 was above the predicted value.

**2.248** (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.

**2.249** (a) The trend is positive, with a fairly 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 2014, 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 $r = 0.843$.

(d) We use technology to find the regression line: $\widehat{HotDogs} = -2743.6 + 1.3953 \cdot Year$.

(e) The slope indicates that the winning number is going up by about 1.4 more hot dogs each year. People better keep practicing!

(f) The predicted number in 2020 is $\widehat{HotDogs} = -2743.6 + 1.3953 \cdot (2020) = 74.9$, which would beat the all-time record.

(g) It is not appropriate to use this regression line to predict the winning number in 2030 because that is extrapolating too far away from the years that were used to create the dataset.

**2.250** (a) Fluoride exposure is not being randomly assigned, so this is an observational study.

(b) The explanatory variable is amount of fluoride exposure in pregnant women. The response variable is IQ scores of their children.

(c) As fluoride exposure goes up, the IQ score goes down, so this is a negative correlation.

(d) The sentence is describing the slope of the regression line.

**2.251** (a) This is a negative association since increases in elevation are associated with decreases in cancer incidence.

(b) The sentence is telling us the slope of the regression line.

(c) The explanatory variable is elevation and the response variable is lung cancer incidence.

**2.252** (a) For someone who has played 8 years of football, we have $\widehat{Cognition} = 102 - 3.34(8) = 75.28$. A person playing 8 years of football is predicted to be at the 75.28 percentile in cognitive ability. For someone who has played 14 years of football, we have $\widehat{Cognition} = 102 - 3.34(14) = 55.24$. A person playing 14 years of football is predicted to be at the 55.24 percentile in cognitive ability.

(b) The slope is $-3.34$. For every additional year playing football, cognitive percentile is predicted to go down 3.34.

(c) The intercept corresponds to 0 years of playing football, so it is not reasonable to interpret the intercept because we are extrapolating too far away from the original values. Also the intercept is 102, and it is impossible to be above the 100th percentile.

**2.253** (a) For this case, the number of years playing football is 18, predicted brain size appears to be about 2700, actual brain size appears to be about 2400, and the residual is $2400 - 2700 = -300$.

(b) The largest positive residual corresponds to the point farthest above the line. For this point, the number of years playing football appears to be 12, predicted brain size appears to be about 3000, actual brain size appears to be about 3900, so the residual is $3900 - 3000 = 900$.

(c) The largest negative residual corresponds to the point farthest below the line. For this point, the number of years playing football appears to be 13, predicted brain size appears to be about 2900, actual brain size appears to be about 2200, so the residual is $2200 - 2900 = -700$.

**2.254** The slope of the regression line indicates that the TV hours decrease by about 0.397 hours for each additional year, so over the decade the total decrease is about 3.97 hours.

**2.255** The slope of the regression line indicates that the computer hours increase by about 0.467 hours for each additional year, so over the decade the total increase is about 4.67 hours.

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

(b) The slope or the regression line is 0.1443. This means that we predict that obtaining 1 more run leads to about 0.1443 more wins.

(c) The predicted number of wins for Houston is $\widehat{Wins} = -31.94 + 0.1443(920) = 100.8$ games. The residual is $107 - 100.8 = 6.2$, meaning that the Giants won 6.2 more games than expected with 920 runs, so they were efficient at winning games.

**2.257** The slope is 0.839. 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 0.839 if the approval rating goes up by 1.

The $y$-intercept is $-36.76$. 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 $-36.76$ 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!

**2.258** (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.

**2.259** (a) The explanatory variable is pre-season wins, the response variable is regular season wins.

(b) For a team that won 2 games in the pre-season, the predicted number of wins is $7.27 + 0.35(2) = 7.97$ wins.

(c) The slope is 0.35, which implies that for every 1 pre-season win we predict 0.35 more regular season wins.

(d) The intercept is 7.27, which implies that we predict a team with 0 pre-season wins will win 7.27 regular season games. This is a reasonable assumption and prediction.

(e) The regression line predicts 42.27 wins for a team that wins 100 pre-season games, which is definitely not appropriate because we are extrapolating well away from the possible number of pre-season wins. (There are only four pre-season games.)

**2.260** (a) We predict that each year the number of bee colonies decreases by 8.358 thousand.

(b) The $y$-intercept in the context presented here refers to the predicted number of bee colonies in 0 AD, which is way too far outside the range of values used to create the line to give us anything meaningful. If we adjusted the year, the intercept would indicate the predicted number of bees colonies in 1995, which is reasonable.

(c) We see that $\widehat{Colonies} = 19,292 - 8.358(2100) = 1,740.2$ thousand colonies, but this is not appropriate because it extrapolates so far from the data.

**2.261** 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 $40 - 20 = 20$.

**2.262** (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%.

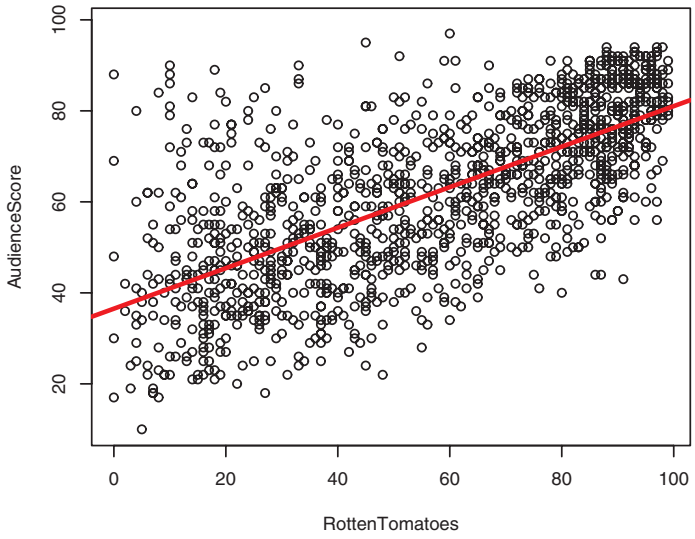
(c) This person has an abdomen circumference of approximately 113 cm. The predicted body fat percent for this abdomen circumference appears on the line to be about 32%, which gives a residual of about $40 - 32 = 8$.

**2.263** (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 $11.3 - 19.825 = -8.525$.

**2.264** (a) The scatterplot of *RottenTomatoes* vs *AudienceScore* with regression line (using *RottenTomatoes* as the predictor) is shown below. We see a positive association with audience scores tending to be higher when critic scores are higher.




(b) The fitted prediction equation is $\widehat{AudienceScore} = 36.52 + 0.4451 \cdot RottenTomatoes$. The slope indicates that for every increase of one point in the rotten tomatoes (critics) rating, we expect the audience rating to increase by about 0.45 points.

(c) Substituting a rotten tomatoes value of 78 into the prediction equation gives a predicted audience score for *Green Room* of 71.24.

$$\widehat{AudienceScore} = 36.52 + 0.4451 \cdot 78 = 71.24$$

The actual audience score is 91, so the residual is $91 - 71.24 = 19.76$.

**2.265** (a) Using technology to fit a regression we get the prediction equation

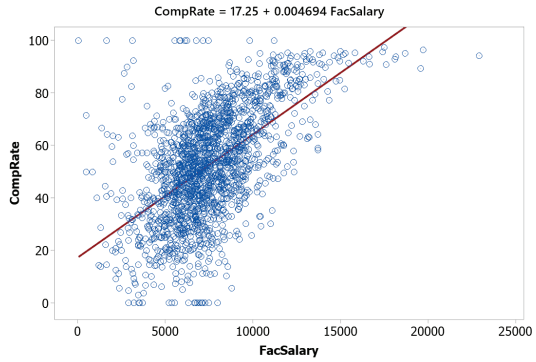
$$\widehat{CompRate} = 17.25 + 0.00469 \cdot FacSalary$$

(b) Substituting $FacSalary = 5000$ into the prediction equation for part (a) we have

$$\widehat{CompRate} = 17.25 + 0.004694 \cdot 5000 = 40.72$$

We predict a four-year school with average faculty salary of \$5,000 to have a completion rate of 40.72%.

(c) The slope of the fitted line (0.004694) is positive, indicating that schools with higher faculty salaries will be predicted to have higher completion rates. While the magnitude of the coefficient might appear to be small, the *FacSalary* values are quite large. An increase of 0.004694 percent for a \$1 in increase in faculty salary, would mean a 4.695% increase for an additional \$1,000 in salary. The scatterplot with the fitted line on it shown below indicates this can yield a substantial change in completion percentage over the range of faculty salaries.



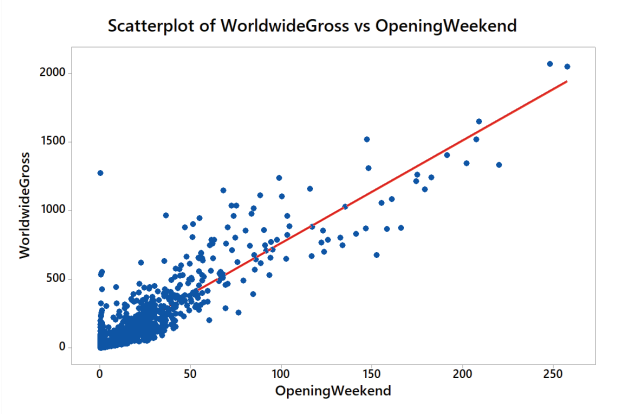

**2.266** (a) Using technology, the fitted prediction equation is $\widehat{HwyMPG} = 6.69 + 1.687 \cdot CityMPG$. The slope indicates that, for every increase of one mpg in the city rating, we expect the highway rating to go up by about 1.687 mpg.

(b) Putting $CityMPG = 23$ into the prediction equation, we find that the predicted highway mpg for a Toyota Corolla is

$$\widehat{HwyPPG} = 6.69 + 1.687 \cdot 23 = 45.49$$

The residual ($Actual - Predicted$) is $40 - 45.59 = -5.59$ mpg.

**2.267** (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.



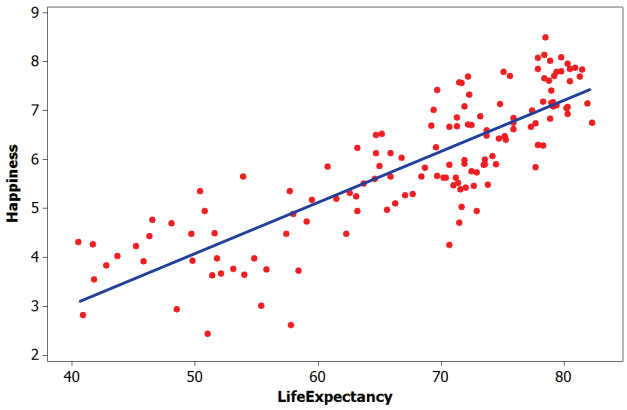

(b) The movie *Frozen* had a world gross of \$1272.45 million, but an opening weekend of only 0.24 million.

(c) We find that the correlation is $r = 0.907$.

(d) The regression line is $\widehat{WorldGross} = 15.4 + 7.50 \cdot OpeningWeekend$.

(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 $\widehat{WorldGross} = 15.4 + 7.5 \cdot (50) = 390.4$ million dollars.

**2.268** (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.

**2.269** Answers will vary.

**Section 2.7 Solutions**

**2.270** (a) Two variables are shown in the scatterplot, and both are quantitative. The range for Variable1 appears to be about $29 - 13 = 16$. The range for Variable2 appears to be about $160 - 70 = 90$.

(b) The association appears to be positive.

(c) Variable2 is on the vertical axis and is the response variable. The slope of the line is clearly positive, verifying our answer to part (b).

(d) The new variable is categorical, with four different groups, labeled A, B, C, D.

(e) The association between Variable1 and Variable2 appears to be negative in all four different groups.

(f) The regression line shows a negative slope in all four different groups.

(g) The association switches from appearing to be positive to being clearly negative, regardless of which group a case is in.

**2.271** (a) The variable *Happiness* is quantitative, the variable *Footprint* is quantitative, and the variable *Region* is categorical.

(b) Regions 1 (Latin America) and 2 (Western nations) seem to have the greatest happiness score. Region 2 (Western nations) seems to have the greatest ecological footprint.

(c) Region 4 (Sub-Saharan Africa) seems to have the lowest happiness scores. The ecological footprint also tends to be low in that region.

(d) Yes, overall, having a greater ecological footprint appears to be associated with having a higher happiness score.

(e) No. Considering only those countries in region 2 (Western nations), there does not seem to be a relationship between happiness and footprint. A greater ecological footprint does not appear to be associated with a higher happiness score.

(f) It is in the top left, with a relatively large happiness score and a relatively small ecological footprint.

(g) For countries in region 4, we should focus on trying to increase the happiness score. For countries in region 2, we should focus on decreasing the ecological footprint.

**2.272** (a) Three variables: height, weight, and body fat percentage. All three are quantitative.

(b) There appears to be a positive relationship. As height increases, weight tends to increase.

(c) The bubbles tend to be larger on the top half of the scatterplot. This makes sense since heavier people (with weight above 200 pounds) might be more likely to have a larger body fat percentage.

(d) The one weighing 125 pounds has the bigger bubble and hence the larger body fat percentage (despite weighing less.)

(e) The 125 pound man with a height of 66 inches appears to have a much larger body fat percentage than the one at 67 inches.

(f) The person with the largest weight (over 260 pounds) appears to have one of the largest bubbles on the graph, so this person's body fat percentage is relatively large.

(g) The person with the largest height (about 78 inches) has a bubble that is pretty average in size, so that person's body fat percentage is pretty average. The person who is third largest in height has quite a small bubble, so that person's body fat percentage is relatively small.

(h) One way to incorporate a fourth variable of gender would be to use different color bubbles for the two genders.

**2.273** (a) There are three variables displayed in the figure. Hippocampus size is quantitative and number of years playing football is quantitative, while the group (control, football with concussion, football without concussion) is categorical.

(b) The blue dots represent the study participants who are in the control group and have never played football, so the "years of football" for all these participants is zero.

(c) The general trend in the scatterplot shows a negative association between years playing football and hippocampus size.

(d) The reddish brown line (for the football with concussion group) is lower, which tells us that football players with a history of concussions appear to have smaller hippocampus sizes.

(e) The green line (for football players without a history of concussions) has the steeper slope.

**2.274** (a) The $CO_2$ concentration in 1960 appears to be about 317 ppm. In 2015, it appears to be about 400 ppm.

(b) It is increasing throughout this period.

(c) During this long history, the concentration is primarily oscillating.

(d) It looks very much like a vertical line.

(e) It appears that, before 1950, the concentration of $CO_2$ had never been above 300 ppm.

**2.275** (a) While it varies each year, the general trend for both positions is running the 40 yard dash in less time, so faster.

(b) Cornerbacks were slightly faster.

(c) No, from year to year it varies which position runs faster on average.

**2.276** (a) No, the point differential was never below 0, so the Warriors never trailed in this game.

(b) The Warriors had their largest lead in the 2nd quarter, at right around +30.

**2.277** (a) There are 7 purple dots at the 24-minute mark, so they were losing in 7 games at halftime.

(b) There are 2 purple dots at the 0-minute mark, so they only lost two of these games (both times by more than 13 points.)

**2.278** (a) There is a negative association between the two variables that generally persists over all time points.

(b) The number of babies per Russian woman decreases dramatically from 1941 to 1943 (from approximately 4.5 to just under 2). The age at first marriage does not change substantially.

(c) Generally, number of babies decrease and age at first marriage increases for Libyan woman from 1973 to 2005.

**2.279** (a) In 1970 the majority of people living in extreme poverty were from Asia.

(b) In 2018 the majority of people living in extreme poverty were from Africa.

(c) In 1970, the distribution has two peaks (it is *bimodal*), with one peak below \$1/day and another between \$10/day and \$20/day. This clearly distinguishes the "third world" from the "first world". In 2018, the overall distribution is more symmetric and bell-shaped on the log-scale provided.

**2.280** (a) Most Americans are sleeping at 6am, as the majority of dots are in the "Sleeping" category.

(b) At 10am, more Americans are working, as there are more dots in the "Working" category than any other category.

(c) Slightly more Americans are eating & drinking at 7pm (approximately 14%) than at 6pm (approximately 11%).

(d) Answers will vary; any correct observation is acceptable.

**2.281** (a) There are 6 different categories shown, from "<10%" to "≥30%".

(b) In 1990, it appears that the highest category needed is "10%–14%". In 2000, the highest category needed is "20%–24%". In 2010, many states are in the top category "≥30%".

**2.282** (a) The first year the 15–19% category was needed was 1990 (the first year included in the sequence) with one state (Mississippi) in that category. The 20–24% category first appeared in 2000. The 25–29% category first appeared in 2003, the 30–34% in 2006 (again, with only one state: Mississippi), and the 35+ category appeared in 2013 (in two states: Mississippi and West Virginia.)

(b) Answers will vary.

**2.283** (a) We see in the spaghetti plot that for *every* state, the percent obese more than doubled during this time period.

(b) There is more variability in 2018.

(c) In 1990, the state with the largest percent obese is Mississippi, with 15.0% obese. In 2018, the state with the smallest percent obese is Colorado, with 23.0% obese.

**2.284** Answers will vary from person to person.

**2.285** (a) We see that "bro" is most commonly used in Texas and surrounding states.

(b) We see that "buddy" is most commonly used in the midwest.

**2.286** (a) In 1880, Rhode Island spent the most ($82.72) per person on cotton. Texas spent the least (only $0.01) per person.

(b) Cotton was most expensive (high price of $1.90 per pound) in 1864.

**2.287** (a) The eastern half of the country is much more heavily populated than the western half.

(b) Answers will vary.

**2.288** The top baby girl name in 2014 was Sophia, and the top name in 1880 was Mary.

**2.289** (a) We can see that there are two distinct clusters of dots, with one cluster having smaller petals (in length and width) than the other.

(b) Setosa has the smallest petals, while Virginica has the largest petals.

**2.290** (a) In a scatterplot showing an association between petal length and petal width, there is a pretty clear distinction between the green dots and the red dots. In a scatterplot showing sepal width and sepal length, however, the green and red dots are mostly mixed together. Thus, the petal length and width show a clearer distinction.

(b) Looking only at the black dots (for Setosa), we see that the association between sepal width and petal length appears to be neither positive nor negative. (The dots appear either mostly horizontal or mostly vertical depending on which variable you put on which axis. In either case, the association does not look positive or negative.)

**2.291** Sepal length is on the vertical axis, and the green dots are highest up.

**2.292** (a) In 1950 there were a lot of babies and toddlers (kids 0–4) and very few old people.

(b) In 1950 the distribution of age is right-skewed.

(c) In 2060 there are many more old people and many fewer babies than there were in 1950.

(d) In 2060, it is projected that there will be more females than males in the 85+ range (females live longer, on average).

**2.293** (a) Whites are decreasing the most in terms of percentage of the US population, from 85% in 1960 to (projected) 43% in 2060.

(b) Hispanics are increasing the most in terms of percentage of the US population, from 4% in 1960 to (projected) 31% in 2060.

(c) Blacks are staying the most constant in terms of percentage of the US population, from 10% in 1960 to (projected) 13% in 2060.

**2.294** Direction and speed of the wind at each station are needed to make this map.

**2.295** (a) The commutes on the carbon bike tend to be longer in distance than the commutes on the steel bike.

(b) The carbon bike is slightly faster, on average, but also tends to cover more distance on the commute, making the overall trip longer in minutes. Perhaps the devices measuring distance, time, or speed give slightly different readings between the two bikes.

(c) Distance is associated with both type of bike and commute time in minutes, confounding the relationship between the two. Distance is a confounding variable.

(d) To minimize commute time, Dr. Groves should ride the carbon bike (which has a slightly faster average speed), but take the shorter distance commutes that are typically taken with the steel bike in this dataset. He might also want to check the accuracy of the recording devices used with each bike.

**2.296** (a) The median Democrat value became more liberal from 1994 to 1998, then stayed relatively constant until 2011, and from 2011 to 2014 it became much more liberal.

(b) The median Republican value became more liberal from 1994 to 2003, and then shifted to becoming more conservative from 2004 to 2014.

(c) There is much more political polarization in 2014 than in 1994.

(d) The two parties started moving rapidly away from each other in 2011 (although the Republican party started moving away from the Democratic party in 2004).

(e) In 2014, the politically active people are much more polarized than the general public.

(f) Answers will vary.

**2.297** (a) In 2014, New England had the highest support for same-sex marriage.

(b) In 2014, South Central had the lowest support for same-sex marriage.

(c) The Mountain states displayed the smallest increase (about 9%), and the Midwest displayed the largest (about 24%, although New England and Pacific are close at 23%).

(d) This data could also have been visualized with a spaghetti plot, in which each time series would be shown on the same plot, or with a dynamic heat map, in which a map of the US could have been shown color-coded according to the level of support, with the colors changing over time.

**2.298** (a) People who bought organic food were richer than those who didn't.

(b) Richer people are more likely to report very good or excellent health.

(c) Income is a confounding variable because it is associated with both whether or not someone bought organic and with their self-reported health status.

(d) No, income is a confounding variable. We cannot determine whether the observed difference in reported health is due to buying organic food, or just due to the fact that people who buy organic are richer and richer people are healthier.

**2.299** (a) Yes, after breaking it down by income, it is still generally true that people who bought organic are more likely to report very good or excellent health, because the red points are above the blue points in all except one category.

(b) No. This is still an observational study, and there are many other potential confounding variables.

**2.300** (a) For all stones, percutaneous nephrolithotomy is more successful ($289/350 = 83\%$ success for percutaneous nephrolithotomy, as opposed to $273/350 = 78\%$ success for open surgery).

(b) For small stones, open surgery is more successful ($81/87 = 93\%$ success for open surgery, as opposed to $234/270 = 87\%$ for percutaneous nephrolithotomy).

(c) For large stones, open surgery is more successful ($192/263 = 73\%$ success for open surgery, as opposed to $55/80 = 69\%$ for percutaneous nephrolithotomy).

(d) Smaller stones result in higher success rates for both treatments. This is most easily seen visually (more light gray for small stones), but can also be calculated numerically: small stone treatment has success rates of $(81 + 234)/(87 + 270) = 315/357 = 88\%$ when combining both treatments, while large stone treatment has lower success rates of $(192 + 55)/(263 + 80) = 247/343 = 72\%$ when combining both treatments.

(e) Percutaneous nephrolithotomy is much more commonly used for small stones. On the small stones plot, ignoring success, the bar for percutaneous nephrolithotomy is much taller than the bar for open surgery.

(f) Open surgery is much more commonly used for large stones. On the large stones plot, ignoring success, the bar for open surgery is much taller than the bar for percutaneous nephrolithotomy.

(g) Open surgery is more commonly used for large stones, which are harder to treat (have a lower success rate in general than small stones). Therefore, even though open surgery is more successful for each stone size, it appears to be worse overall when the stone sizes are combines.

(h) This was probably not a randomized experiment. If it had been a randomized experiment, then the numbers of small and large stones should have been approximately the same between the two treatments.

**2.301** Answers will vary.

**2.302** According to this graphic, greenhouse gases are warming the world.

**2.303** There is a strong tendency for blacks to live near other blacks in one region of St. Louis (near the middle of the city), and for whites to live near other whites.

**2.304** (a) Mornings tend to have more cloud cover than evenings. This may be most visible in the circle plot where the lines extend further from the center (no clouds) in the morning hours.

(b) Summer shows the most variability in cloud cover throughout the day. This may be most visible in the spaghetti plot which shows more variability in cloud cover during days in July and August.

(c) Answers will vary.

**2.305** (a) Early morning (6 am to 7 am) has the highest percent cloud cover in August.

(b) Summer (around August) tends to be the least windy season for Chicago.

**2.306** Solar wind is likely responsible for particles leaving the atmosphere of Mars.

**2.307** (a) This video is conveying the endangered status and variety of gazelle species.

(b) Answers will vary.

**2.308** (a) Student gatherings at a university based on smart phone tracking.

(b) Answers will vary.

**2.309** Answers will vary.

**2.310** (a) Using technology and the data in **Collegescores4yr** we find the mean faculty salary is \$7091 at private schools and \$8520 at public schools, so public schools have higher average salary.

(b) The mean completion rate is higher at private schools (55.7%) than at public schools (50.2%).

(c) From parts (a) and (b) it looks like schools with higher faculty salaries (public) tend to have lower completion rates. This would suggest a negative association between the two variables.

(d) The blue circles appear more at lower faculty salaries (to the left), but have higher completion rates (towards the top). This suggests the blue points and line are for private colleges. Note also that the mean faculty salary should predict the mean completion rate in each group, so (7091, 55.7) should be on the line for private schools and (8520, 50.2) should be on the line for public schools.

(e) The regression lines in the scatterplot make it clear that we have a positive association between faculty salaries and completion rates for both private and public schools. This is not consistent with the answer based only on comparing the means for each group.

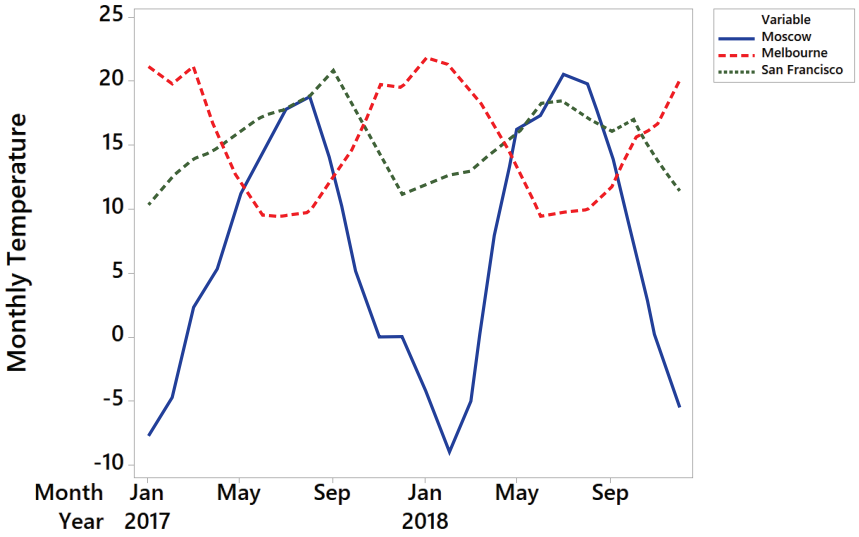
**2.311** Answer will vary.

**2.312** Answer will vary.

**2.313** Answer will vary.

**2.314** Answers will vary.

**2.315** Although we could use separate time series plots, the figure below puts all three series on the same scale (as a spaghetti plot) to ease comparisons.




(a) The most obvious difference between Moscow and San Francisco is that the variability is much higher in Moscow temperatures. The low temperatures are around the same time of the year (January), but are much lower in Moscow. The warmest months have more similar temperatures.

(b) The variability is similar between Melbourne and San Francisco, but the seasonal patterns are reversed. Since Melbourne is in the Southern Hemisphere, its warmest months are in January and February, when San Francisco, in the Northern Hemisphere, tends to be the coldest. Melbourne's chilliest period in July is around when San Francisco is enjoying warmer temperatures.

**2.316** Answers will vary.

**2.317** Answers will vary.

**2.318** Answers will vary.