Section 1.1 Solutions

1.1 (a) The cases are the people who are asked the question.

(b) The variable is whether each person supports the law or not. It is categorical.

1.2 (a) The cases are the 100 stocks.

(b) The variable is the percentage change, which is a numerical quantity, for each of the stocks. It is quantitative.

1.3 (a) The cases are the teenagers in the sample.

(b) The variable is the result (yes or no) indicating whether each teenager eats at least five servings a day of fruits and vegetables. It is categorical.

1.4 (a) The cases are the bunches of bananas in the sample.

(b) The variable is the number of days until the bananas go bad. It is quantitative.

1.5 (a) The 10 beams that were tested.

(b) The force at which each beam broke. It is quantitative.

1.6 (a) The cases are countries of the world.

(b) The variable is whether or not the literacy rate is over 75%. It is categorical.

1.7 Since we expect the number of years smoking cigarettes to impact lung capacity, we think of the number of years smoking as the explanatory variable and the lung capacity as the response variable.

1.8 Since we expect the amount of fertilizer used to impact the yield (and not the other way around), we think of the amount of fertilizer as the explanatory variable and the yield of the crop as the response variable.

1.9 Ingesting more alcoholic drinks will cause the level of alcohol in the blood to increase, so the number of drinks is the explanatory variable and blood alcohol content is the response.

1.10 The world record time will continue to decrease as the years go by so we expect the year to impact marathon record time. We think of the year as the explanatory variable and the record time as the response variable.

1.11 (a) *Year* and *HigherSAT* are categorical. The other six variables are all quantitative, although *Siblings* might be classified as either categorical or quantitative.

(b) There are many possible answers, such as "What proportion of the students are first year students?" or "What is the average weight of these students?"

(c) There are many possible answers, such as "Do seniors seem to weigh more than first year students?" or "Do students with high Verbal SAT scores seem to also have high Math SAT scores?"

1.12 (a) In addition to the identification column, *Country*, there are 24 variables. We see that *Developed* is a categorical variable, while the other 23 variables are all quantitative.

(b) There are many possible answers, such as "What is the average life expectancy for all countries of the world?" or "What proportion of countries are developed?"

(c) There are many possible answers, such as "Do countries with a greater land area have a larger percent rural?" or "Do countries that spend a relatively large amount on the military spend a relatively small amount on health care?" or "Do developed countries have a longer life expectancy than developing countries?"

1.13 There are at least two variables. One variable is whether or not the spider engaged in mock-sex. This variable is categorical and the explanatory variable. Another variable is length of time to reach the point of real mating once the spider is fully mature. This variable is quantitative and the response variable.

1.14 The individual cases are the lakes from which water samples were taken. For each lake in the sample, we record the concentration of estrogen in the water and the fertility level of fish. Both are quantitative variables.

1.15 There are two variables. One variable indicates the presence or absence of the gene variant and the second variable indicates which of the three ethnic groups the individual belongs to. Both variables are categorical.

1.16 (a) There are 8 cases, corresponding to the 8 rowers. The two variables are number of days to cross the Atlantic and gender. Number of days to cross the Atlantic is quantitative and gender is categorical.

(b) We need two columns, one for each variable. The columns can be in either order. See the table.

Time	Gender
40	Male
87	Male
78	Male
106	Male
67	Male
70	Female
153	Female
81	Female

1.17 (a) There are 10 cases, corresponding to the 10 cities. The two variables are population, which is quantitative, and the hemisphere the city is in, which is categorical.

(b) We need two columns, one for each variable. The columns can be in either order. See the table.

Population	Hemisphere
37	Eastern
26	Eastern
23	Eastern
22	Eastern
21	Eastern
21	Eastern
21	Eastern
21	Western
20	Western
19	Western

1.18 One variable is whether each male was fed a high-fat diet or a normal diet. This is the explanatory variable and it is categorical. The response variable is whether or not the daughters developed metabolic syndrome, which is also categorical.

1.19 One variable is whether the young female mice lived in an enriched environment or not. This is the explanatory variable and it is categorical. The response variable is how fast the offspring learned to navigate mazes and is quantitative.

1.20 In the first study, the cases are the students. The only variable is whether or not the student has smoked a hookah. This is a categorical variable.
In the second study, the cases are the people in a hookah bar. The variables are the length of the session, the frequency of puffing, and the depth of inhalation. All are quantitative.
In the third study, the cases are the smoke samples, and the variables are the amount of tar, nicotine, and heavy metals. All three variables are quantitative.

1.21 (a) This description of the study mentions six variables: age, nose volume, nose surface area, nose height, nose width, and gender.

(b) One of the variables (gender) is categorical, and the other five are quantitative.

(c) There are six variables so the dataset will have six columns. The 859 participants are the cases, so the dataset will have 859 rows.

1.22 (a) The cases are the 47 participants.

(b) The description of the study includes three different variables: the score on the no-distractions test, the score on the test while texting, and whether or not the student considered him or herself to be good at multitasking. The two test score variables are quantitative and the multitasking variable is categorical.

(c) The dataset would have 47 rows (one for each participant) and three columns (one for each of the three variables.)

1.23 (a) The cases are the 40 people with insomnia who were included in the study.

(b) There are two variables. One is which group the person is assigned to, either therapy or not, and the other is whether or not the person reported sleep improvements. Both are categorical.

(c) The dataset would have two columns, one for each of the two variables, and 40 rows, one for each of the people in the study.

1.24 If we simply record age in years and income in dollars, the variables are quantitative. Often, however, in a survey, we don't ask for the exact age but rather what age category the participant falls in (20 - 29, 30 - 39, etc). Similarly, we often don't ask for exact income but for an income category (less than $10,000, between $10,000 and $25,000, etc.) If we ask participants what category they are in for each variable, then the variables are categorical.

1.25 We could sample people eligible to vote and ask them each their political party and whether they voted in the last election. The cases would be people eligible to vote that we collect data from. The variables would be political party and whether or not the person voted in the last election. Alternatively, we could ask whether each person plans to vote in an upcoming election.

1.26 We could survey a sample of people and ask their household income and measure happiness in some way, such as asking how happy they are on a scale of 1-10. The cases would be the people we collect data from. The variables in this case would be household income and happiness rating, although any two variables measuring wealth and happiness are possible.

1.27 Answers will vary.

Section 1.2 Solutions

1.28 This is a sample, because only a subset of fish are measured.

1.29 This is a population, because all customers are accounted for.

1.30 This is a population, because all registered vehicles are accounted for.

1.31 This is a sample, because only a subset of college students were sent the questionnaire.

1.32 The sample is the 120 people interviewed. The population might be all people in that town or all people that go to the mall in that town or a variety of other groups larger than and containing the 120 people in the sample.

1.33 The sample is the five hundred Canadian adults that were asked the question; the population is all Canadian adults.

1.34 The sample is the 100 customers surveyed; the population is all customers of the cell phone carrier.

1.35 The sample is the 1000 households which have databoxes attached to the televisions. The population is all US households with televisions.

1.36 (a) The sample is the 100 college students who were asked the question.

(b) The population we are interested in is all Americans.

(c) A population we can generalize to, given our sample, is college students.

1.37 (a) The sample is the 10 selected twitter accounts.

(b) The target population is all twitter accounts.

(c) The population we can generalize to, given the sample, is only twitter accounts of this author's followers, since this is the population from which the sample was drawn.

1.38 (a) The sample is the 1500 people who were contacted.

(b) The population we are interested in is all residents of the US.

(c) A population we can generalize to, given our sample, is residents of Minnesota.

1.39 (a) The sample is the girls who are on the selected basketball teams.

(b) The population we are interested in is all female high school students.

(c) A population we can generalize to, given our sample, is female high school students who are on a basketball team.

1.40 Yes, this is a random sample from the population.

1.41 Yes, this is random sample from the population.

1.42 No, this is not a random sample, because some employees may be more likely than others to actually complete the survey.

1.43 No, this is not a random sample, because certain segments of the population (e.g. those not attending college) cannot be selected.

1.44 No, this is not a random sample. We might think we can pick out a "representative sample", but we probably can't. We need to let a random number generator do it for us.

1.45 No, this is not a random sample, this is a volunteer sample, since the only people in the sample are those that self-select to respond to the online poll.

1.46 This sample is definitely biased because only students who are at the library on a Friday night can be selected. The random sample should be from all students.

1.47 This is biased because the way the question is worded is not at all objective. Although the sample is a random sample, the wording bias may distort the results.

1.48 This sample is biased because taking 10 apples off the top is not a random sample. The apples on the bottom of the truckload are probably more likely to be bruised.

1.49 From the description, it appears that this method of data collection is not biased.

1.50 This sample is biased because it is a volunteer survey in which people choose to participate or not. Most likely, the people taking the time to respond to the email will have stronger opinions than the rest of the student body.

1.51 Because this was a random sample of parents in Kansas City, the result can be generalized to all parents in Kansas City.

1.52 (a) No, the sample is almost certainly not representative, since it is a volunteer sample and only includes people who visit that website and who chose to participate in the poll.

(b) No, it is not appropriate to generalize since the sample is not representative.

1.53 (a) Yes, the sample is likely to be representative since it is a random sample.

(b) Yes, since the sample is a random sample, we can generalize to the population of all Canadian consumers.

1.54 (a) The sample is the 1,236 registered voters that were contacted. The intended population is all registered voters in Iowa.

(b) Yes, it is reasonable to generalize since the sample was selected randomly.

1.55 (a) The individual cases are the over 6000 restroom patrons who were observed. The description makes it clear that at least three variables are recorded. One is whether or not the person washed their hands, another is the gender of the individual, and a third is the location of the observation. All three are categorical.

(b) In a phone survey, people are likely to represent themselves in the best light and not always give completely honest answers. That is why it is important to also find other ways of collecting data, such as this method of observing people's actual habits in the restroom.

1.56 (a) The sample is the survey participants, the population is all professors at the University of Nebraska.

(b) No, we cannot conclude that the sample of survey responders is not representative of professors at the University of Nebraska since we are not given enough information to decide one way or the other.

(c) No, the 94% is based on self descriptions, which can be (and in this case, probably are) biased.

1.57 No. This is a volunteer sample, and there is reason to believe the participants are not representative of the population. For example, some may choose to participate because they LIKE alcohol and/or marijuana, and those in the sample may tend to have more experience with these substances than the overall population. In addition, the advertisements for the study were aired on rock radio stations in Sydney, so only those people who listen to rock radio stations in Sydney would hear about the option to participate.

1.58 Yes! The sample is a random sample so we can be quite confident that it is probably a representative sample.

1.59 (a) This is not a simple random sample from the population, since only those who saw and wanted to click and complete the survey were included.

(b) These results could also have been biased by how the survey was constructed. The wording of the questions might also introduce bias.

1.60 The study given found a relationship in a sample of rats. This relationship may not generalize to the human population.

1.61 The sample of planes that return from bombing missions was biased. More bullet holes were found in the wings and tail because planes that were shot in other regions were more likely to crash and not return.

1.62 (a) The population in the CPS is all US residents. (Also acceptable: US citizens, US households...)

(b) The population in the CES survey is all non-farm businesses and government agencies in the U.S.

(c) i. The CES survey would be more relevant, because the question pertains to companies.

ii. The CPS would be more relevant, because the question pertains to American people.

iii. The CPS would be more relevant, because the question pertains to people, not businesses.

1.63 (a) Since the NHANES sample is drawn from all people in the US, that is the population we can generalize to.

(b) Since the NHAMCS sample is drawn from patients in emergency rooms in the US, we can generalize the results to all emergency room patients in the US.

(c) i. NHANES: The question about an association between being overweight and developing diabetes applies to all people in the US, not just those who visit an emergency room.

ii. NHAMCS: This question asks specifically about the type of injury for people who go to an emergency room.

iii. NHAMCS: This question of average waiting time only applies to emergency room patients.

iv. NHANES: This question is asking about all US residents. Note that the proportion would be equal to one for the people sampled in NAMCS since they only get into the sample if they visit an emergency room!

1.64 Answers will vary. See the technology notes to see how to use specific technology to select a random sample.

1.65 Answers will vary. See the technology notes to see how to use specific technology to select a random sample.

Section 1.3 Solutions

1.66 The use of "improves" implies this is a causal association.

1.67 Since "no link is found" there is neither association nor causation.

1.68 The phrase "leads to deaths" indicates a causal association.

1.69 The phrase "more likely" indicates an association, but there is no claim that wealth *causes* people to lie, cheat or steal.

1.70 The phrase "tend to be more educated" indicates an association, but there is no claim that owning a cat causes more education (or that better education causes people to prefer cats).

1.71 The statements imply that eating more fiber will cause people to lose wait, so this is a causal association.

1.72 One possible confounding variable is temperature (or season). More people eat ice cream, and go swimming, in warm weather. Other answers are possible. Remember that a confounding variable should be associated with both of the variables of interest.

1.73 One possible confounding variable is population. Increasing population in the world over time may mean more beef and more pork is consumed. Other answers are possible. Remember that a confounding variable should be associated with both of the variables of interest.

1.74 One possible confounding variable is wealth. People who own a yacht are likely wealthy and can afford a sports car. Other answers are possible. Remember that a confounding variable should be associated with both of the variables of interest.

1.75 One possible confounding variable is snow in the winter. When there is more snow, sales of both toboggans and mittens will be higher. Remember that a confounding variable should be associated with both of the variables of interest.

1.76 One possible confounding variable is number of cars (and also number of people). If there are lots of cars, there will be more pavement and more air pollution. Remember that a confounding variable should be associated with both of the variables of interest.

1.77 One possible confounding variable is gender. Males usually have shorter hair and are taller. Other answers are possible. Remember that a confounding variable should be associated with both of the variables of interest.

1.78 We are not manipulating any variables in this study, we are only collecting information (rice preference and metabolism) as they exist. This is an observational study.

1.79 We are actively manipulating the explanatory variable (playing music or not), so this is an experiment.

1.80 We are actively manipulating the explanatory variable (planting trees or not), so this is an experiment.

1.81 We are not manipulating any variables in this study, we are only measuring things (omega-3 oils and water acidity) as they exist. This is an observational study.

1.82 Data were collected after the fact from sprinters, marathon runners, and non-athletes. No genes were manipulated. These data came from an observational study.

1.83 The penguins in this study were randomly assigned to get either a metal or an electronic tag so this is an experiment.

1.84 All three studies are experiments since the scientists actively control the treatment (tears or salt solution).

1.85 A possible confounding variable is amount of snow and ice on the roads. When more snow and ice has fallen, more salt will be needed *and* more people will have accidents. Notice that the confounding variable has an association with *both* the variables of interest.

1.86 Age or grade level! Certainly, students in sixth grade can read substantially better than students in first grade and they are also substantially taller. Grade level is influencing both of the variables and it is a confounding variable. If we look at individual grades one at a time, the association could easily disappear.

1.87 Yes, this study provides evidence that louder music causes people to drink more beer, because the explanatory variable (volume of music) was randomly determined by the researchers and an association was found.

1.88 Study 1 is a randomized, controlled experiment, while Study 2 is observational. Therefore, Study 1 provides better evidence of causation, while study 2 is more prone to confounding variables (for example, people who eat more nuts may also be more healthy in other ways).

1.89 (a) Yes. Because the study in mice was a randomized experiment, we can conclude causation.

(b) No. Since it appears that the study in humans was an observational study, it is not appropriate to conclude causation. Although the headline may still be true for humans, we cannot make this conclusion based on the study described.

1.90 No, this was an observational study and allowed for many confounding variables.

1.91 (a) The cases are the 2,623 schoolchildren.

(b) The explanatory variable is the amount of greenery around the schools.

(c) The response variable is the score on the memory and attention tests.

(d) Yes, the headline implies that more green space causes kids to be smarter.

(e) No variables were manipulated so this is an observational study.

(f) No! Since this is not an experiment, we cannot conclude causation.

(g) The socioeconomic status of the children is a possible confounding variable, since it is likely to effect both the amount of green space and also the test scores. There are other possible answers.

1.92 (a) The cases are the Danish men who were included in the study.

(b) The explanatory variable is whether or not the person has been in the hospital for an infection. Since the answer is either yes or no, this is a categorical variable.

(c) The response variable is the IQ score for the person. We can find an average so this is a quantitative variable.

(d) Yes, the headline implies that infections lower IQ, which is causation.

(e) The explanatory variable was not manipulated, so this is an observational study.

(f) No, since this was not an experiment, it is not appropriate to conclude causation. There are many possible lurking variables that might impact both IQ and the likelihood of having an infection.

1.93 (a) The explanatoy variable is amount of leisure time spent sitting and the response variable is whether or not the person gets cancer.

(b) This is an observational study because the explanatory variable was not randomly assigned.

(c) No, we cannot conclude spending more leisure time sitting causes cancer in women because this is an observational study.

(d) No, we also cannot conclude that spending more leisure time sitting does not cause cancer in women; because this was an observational study we can make no conclusions about causality. Sitting may or may not cause cancer.

1.94 (a) The explanatory variable is whether or not the participants were given access to food and drink after 10pm, or just water. The response variables are reaction time and number of attention lapses.

(b) This is a randomized experiment because the explanatory variable was randomly assigned.

(c) Yes, we can conclude that eating late at night worsens reaction time and increases attention lapses for sleep deprived people.

(d) No, there are not likely to be confounding variables, because this was a randomized experiment.

1.95 (a) This is an observational study because the explanatory variable (time spent on affection after sex) was not randomly assigned.

(b) No, because this is an observational study. It's quite possible that people in stronger, more loving relationships simply tend to spend more time cuddling after sex, not that cuddling after sex causes relationship happiness.

(c) No, the phrase "boosts" implies a causal relationship, which cannot be supported by an observational study.

(d) No, the phrase "promotes" implies a causal relationship, which cannot be supported by an observational study.

1.96 (a) This is an observational study because we cannot randomly determine when a child learns to talk.

(b) No, we cannot conclude that early language skills reduce preschool tantrums because this is an observational study.

(c) There are many possible confounding variables, such as the intelligence of the child or parental involvement.

1.97 (a) The explanatory variable is whether the person just had a full night of sleep or 24 hours of being awake. The response variable is ability to recognize facial expressions.

(b) This is a randomized experiment, a matched pairs experiment because each person received both treatments.

(c) Yes, we can conclude that missing a night of sleep hinders the ability to recognize facial expressions, because the explanatory variable was randomly assigned.

(d) No, we cannot conclude that better quality of REM sleep improves ability to recognize facial expressions, because the explanatory variable in this case (quality of REM sleep) was not randomly assigned.

1.98 (a) No, we cannot conclude that drinking diet soda causes weight gain because the study was observational and prone to confounding variables. For example, it is possible that seniors who drink more diet soda do so because they know they are prone to gaining weight.

(b) Neither study is perfect. The study on senior citizens is observational and the study on rats may not generalize to humans. However, together these studies provide a more convincing case that diet soda can cause weight gain. Opinions can vary on how strong this evidence for causation is.

1.99 (a) This is an experiment since the background color was actively assigned by the researchers.

(b) The explanatory variable is the background color, which is categorical. The response variable is the attractiveness rating, which is quantitative.

(c) The men were randomly divided into the two groups. Blinding was used by not telling the participants or those working with them the purpose of the study.

(d) Yes. Since this was a well-designed randomized experiment, we can conclude that there is a causal relationship.

1.100 (a) It is an observational study since no one assigned some people to live in a city and some to live in the country.

(b) No, since we can never conclude from an observational study that there is a causal association.

(c) The 2011 study is also an observational study, since, again, no one assigned some people to live in a city and some to live in the country.

(d) The explanatory variable is whether or not the participant lives in the city or the country, which is a categorical variable. The response variable is level of activity in stress centers of the brain, which is quantitative.

(e) No! The results come from an observational study, so we cannot conclude a causal relationship.

1.101 The explanatory variables are the type of *payment* and *sex*. Only the type of payment can be randomly assigned. The number of *items* ordered and *cost* are response variables.

1.102 (a) The explanatory variable is whether or not the person had a good night's sleep or is sleep-deprived. The response variable is attractiveness rating.

(b) Since the explanatory variable was actively manipulated, this is an experiment. The two treatments are well-rested and sleep-deprived. Since all 23 subjects were photographed with both treatments, this is a matched pairs experiment.

(c) Yes, we can conclude that sleep-deprivation causes people to look less attractive, because this is an experiment.

1.103 (a) We randomly divide the participants into two groups of 25 each. Half will be given fluoxetine and half will get a placebo.

(b) The placebo pills will look exactly like the fluoxetine pills and will be taken the same way, but they will not have any active ingredients.

(c) The patients won't know who is getting which type of pill (the fluoxetine or the placebo) and the people treating the patients and administering the questionnaire won't know who is in which group.

1.104 (a) The explanatory variable is amount of sleep and the response variable is growth in height.

(b) We would take a sample of children and randomly divide them into two groups. One group would get lots of sleep and the other would be deprived of sleep. Then after some time passed, we would compare the amount of height increase for the children in the two groups.

(c) An experiment is necessary in order to verify a cause and effect relationship, but it would definitely not be appropriate to randomly assign some of the kids to be sleep-deprived for long periods of time just for the purposes of the experiment!

1.105 (a) Randomly assign 25 people to carbo-load and 25 people to not carbo-load and then measure each person's athletic performance the following day.

(b) We would have each person carbo-load and not carbo-load, on different days (preferably different weeks). The order would be randomly determined, so some people would carbo-load first and other people would carbo-load second. In both cases athletic performance would be measured the following day and we would look at the difference in performance for each person between the two treatments.

(c) The matched pairs experiment is probably better because we are able to compare the different effects for the same person. It is more precise comparing one person's athletic performance under two different treatments, rather than different people's athletic performance under two different treatments.

1.106 (a) Randomly divide the students into two groups of 20 students each. One group gets alcohol and the other gets water. Measure reaction time for students in both groups.

(b) Measure reaction time for all 40 students both ways: after drinking alcohol and after drinking water. Do the tests on separate days and randomize the order in which the students are given the different treatments. Measure the difference in reaction time for each student.

1.107 Answers will vary. Example: The total amount of pizza consumed and the total amount of cheese consumed, per year, over the last century. Eating more pizza causes people to eat more cheese, but the overall rise in population is also a confounding variable.

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 $p=793,986/1,672,395=0.475$.

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 $30/70 = 0.429$.

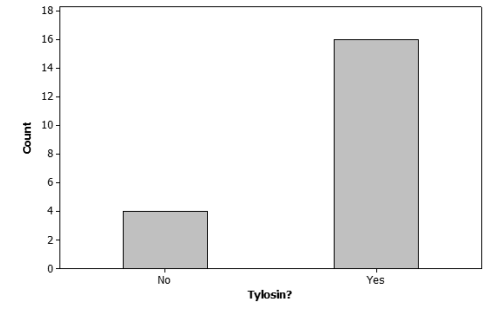
2.13 Since the dataset includes all professional soccer games, this is a population. The cases are soccer games and there are approximately 66,000 of them. The variable is whether or not the home team won the game; it is categorical. The relevant statistic is $p = 0.624$.

2.14 (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

2.15 (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.

2.16 (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.

2.17 (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.

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

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

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

2.19 (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 $144/181 = 0.796$.

2.20 (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.

2.21 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 $p_H - p_C = 0.097 - 0.052 = 0.045$.

2.22 (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.

2.23 (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.

2.24 (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 $\hat{p}_T - \hat{p}_N = 0.70 - 0.15 = 0.55$.

2.25 (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.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

2.30 A 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%).

2.31 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.

2.32 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.

2.33 (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

2.34 (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

2.35 (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%.

2.36 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%.

2.37 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

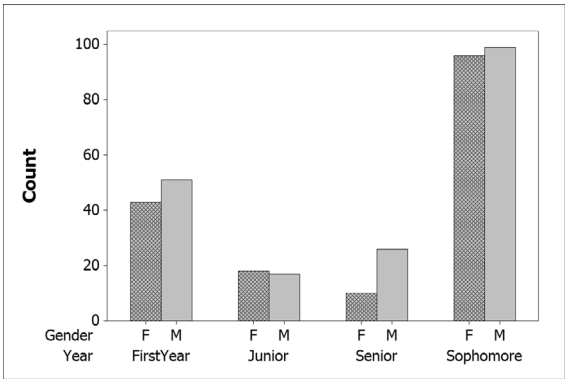
We see that the male/female split is close to 50/50 for most of the years, except for the senior year which appears to have a much higher proportion of males.

2.38 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.

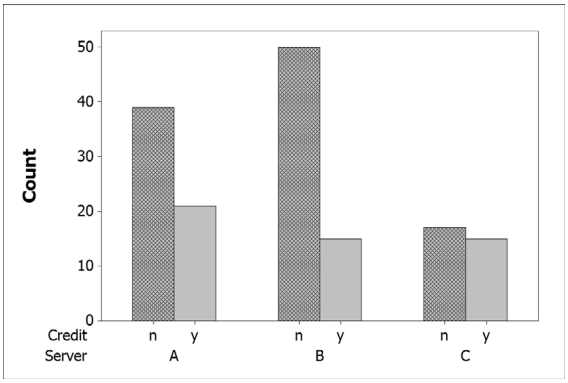
	A	B	C	All
n	39	50	17	106
	65.00	76.92	53.13	67.52
y	21	15	15	51
	35.00	23.08	46.88	32.48
All	60	65	32	157
	100.00	100.00	100.00	100.00

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.

2.39 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.



2.40 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.



2.41 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

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

2.43 Only histogram F is skewed to the right.

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

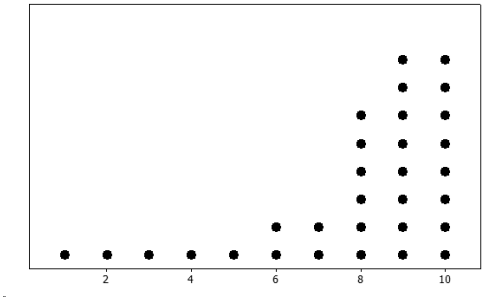
2.45 While all of B,C,D,E and G are approximately symmetric, only B,C and E are also bell shaped.

2.46 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.

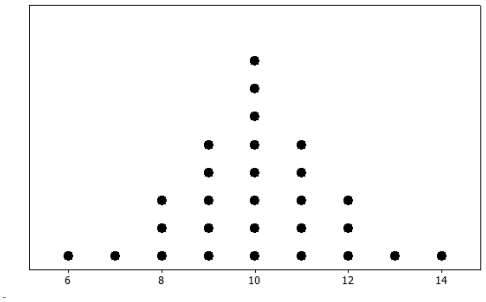
2.47 Histograms E and G are both approximately symmetric, so the mean and median will be approximately equal. Histogram F is skewed right, so the mean should be larger then the median; while histogram H is skewed left, so the mean should be smaller then the median.

2.48 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.

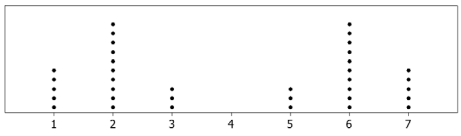
2.49 There are many possible dotplots we could draw that would be clearly skewed to the left. One is shown.



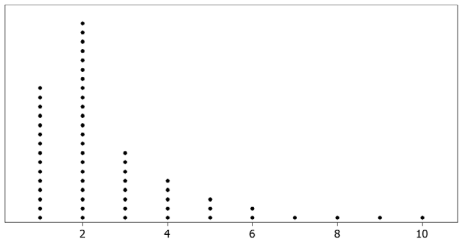
2.50 There are many possible dotplots we could draw that are approximately symmetric and bell-shaped. One is shown.



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



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



2.53 (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.

2.54 (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.

2.55 (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.

2.56 (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.

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

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

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

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

2.61 (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:

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

2.62 (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.

2.63 (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.

2.64 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.

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

(b) The distribution is skewed to the right. There appears to be an outlier at about 40 million. (The outlier represents the state of California.)

(c) The median splits the data in half and appears to be about 4 million. (In fact, it is 4.53 million.)

(d) The mean is the balance point for the histogram and is harder to estimate. It appears to be about 6 million. (In fact, it is 6.36 million.)

2.66 (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.

2.67 (a) The distribution is skewed to the left since there are many values between about 74 and 83 and then a long tail going down to the outliers on the left.

(b) Since half the values are above 74, the median is about 74. (The actual median is 73.8.)

(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 70.842.)

2.68 (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.

2.69 (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.

2.70 (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 $\overline{x}_1 - \overline{x}_2 = 230 - 120 = 110$.