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 two identification columns, ID *Country* and the abbreviated *Code*, there are 12 variables. We see that *Developed* is a categorical variable, while the other eleven 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 (a) The categorical variables are *Smoke*, *Vitamin*, *Gender*, *VitaminUse*, and *PriorSmoke*. The other eleven variables are quantitative.

(b) There are many possible answers. For example, one possible relationship of interest between two categorical variables is the relationship between smoking status and vitamin use. A possibly interesting relationship between two quantitative variables is between the amount of beta-carotene consumed in the diet (*BetaDiet*) and the concentration of beta-carotene in the blood (*BetaPlasma*). One possible relationship of interest between a categorical variable and a quantitative variable might be gender and the number of alcoholic drinks per week.

1.14 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.

1.15 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.

1.16 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.

1.17 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.

1.18 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.

1.19 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.

1.20 (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

1.21 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.

1.22 (a) The cases are the 41 participants.

(b) There are many variables in this study. The only categorical variable is whether or not the person participated in the meditation program. All other variables are quantitative variables. These variables include (at minimum):

- Brain wave activity before
- Brain wave activity after
- Brain wave activity 4 months later
- Immune response after 1 month
- Immune response after 2 months
- Negative survey before
- Negative survey after
- Positive survey before
- Positive survey after

(c) The explanatory variable is whether or not the person participated in the meditation program.

(d) The data set will have 41 rows (one for each participant) and at least 10 columns (one for each variable).

1.23 (a) The "most appealing" question would require just one categorical variable with four possible categories corresponding to the four flavors.

(b) Data for "which are appealing" would need four categorical variables, one for each flavor, with values of *yes* or *no*.

(c) The "rank the flavors" item would need four variables recording the rank given to each flavor. These could be considered categorical (first, second, ...) or quantitative (numerical value of the rank).

(d) The "rate the flavors" item would need four quantitative variables, each with a value between 1 and 10 for the rating assigned to that flavor.

1.24 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.

1.25 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.

1.26 Answers will vary.

Section 1.2 Solutions

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

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

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

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

1.31 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.

1.32 The sample is the fish in that one day's catch. The population is all fish in that area.

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 consists of the cookies that were in the package; the population is all Chips Ahoy! cookies.

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

1.37 (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.

1.38 (a) The sample is the 100 dogs selected.

(b) The population we are interested in is all dogs in New York City (licensed or not).

(c) A population we can generalize to, given our sample, is all licensed dogs in the New York City registry.

1.39 (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.

1.40 (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.

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

1.42 Yes, this is random sample from the population.

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

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

1.45 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.

1.46 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.

1.47 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.

1.48 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.

1.49 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.

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

1.51 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.

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

1.53 (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.

1.54 (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.

1.55 No, we cannot conclude that about 79% of all people think physical beauty matters, since this was a volunteer sample in which only people who decided to vote were included in the sample, and only people looking at *cnn.com* even had the opportunity to vote. The sample is the 38,485 people who voted. The population if we made such an incorrect conclusion would be all people. There is potential for sampling bias in every volunteer sample.

1.56 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.

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

1.58 (a) The sample is the 300 salons that were contacted.

(b) Yes, the sample was collected in a way that should be representative of all tanning salons.

(c) The salons are more interested in marketing what they offer than in giving facts. The responses are dishonest, and a completely inaccurate portrayal of the facts, because they are trying to get more business.

(d) Yes, the sample is well taken so the study is probably accurate in how salons market to teenage girls.

1.59 (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.

1.60 (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!

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

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

1.63 (a) Number the rows from 1 to 100 and the plants within each row from 1 to 300. Use a computer random number generator to pick a number between 1 and 100 to select a row and a second number between 1 and 300 to identify the plant within that row. Repeat until 30 different plants have been selected. Other options are possible: for example, we could number the plants from 1 to 3000 and randomly select 30 numbers between 1 and 3000.

(b) Here is the start of one sample.

```
Row     Plant
#94     #180
#83     # 81
#10     #222
```

Other options are possible.

Section 1.3 Solutions

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

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

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

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

1.68 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).

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

1.70 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.

1.71 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.

1.72 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.

1.73 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.

1.74 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.

1.75 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.

1.76 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.

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

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

1.79 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.

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

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

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

1.83 Snow falls when it is cold out and the heating plant will be used more on cold days than on warm days. Also, when snow falls, people have to shovel the snow and that can lead to back pain. Notice that the confounding variable has an association with *both* the variables of interest.

1.84 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.

1.85 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.

1.86 No. The researchers only measured the walking habits at the beginning of the study and did not actively control the amounts walked. This is an observational study, not an experiment, so the causal conclusion (that hiking reduces risk) is not justified. This doesn't mean that walking might not be helpful, just that this study does not establish a cause and effect relationship.

1.87 (a) The cases are university students. One variable is whether the student lives in a single-sex or co-ed dorm. This is a categorical variable. The other variable is how often the student reports hooking up for casual sex, which is quantitative.

(b) The type of dorm is the explanatory variable and the number of hook-ups is the response variable.

(c) Yes, apparently the studies show that students in same sex dorms hook-up for casual sex more often, so there is an association.

(d) Yes, the president is assuming that there is a causal relationship, since he states that "single sex dorms *reduce* the number of student hook-ups".

(e) There is no indication that any variable was manipulated, so the studies are probably observational studies.

(f) The type of student who requests a single sex dorm might be different from the type of student who requests a co-ed dorm. There are other possible confounding variables.

(g) No! We should not assume causation from an observational study.

(h) He is assuming causation when there may really only be association.

1.88 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.

1.89 (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.

1.90 (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.

1.91 (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.

1.92 (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.

1.93 (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!

1.94 (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.

1.95 (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.

1.96 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.