1.12 Countries of the World Information about the world’s countries is given in All Countries, introduced in Data 1.2 on page 7. You can find a description of the variables in Appendix B on page 691. For the full dataset:

(a) Indicate which of the variables are quantitative and which are categorical.
(b) List at least two questions we might ask about any one of these individual variables.
(c) List at least two questions we might ask about relationships between any two (or more) of these variables.

Solution
(a) Quantitative Variables: Land Area, Population, Energy, Rural, Military, Health, HIV, Internet, Birth Rate, Elderly Population, Life Expectancy, CO2, Cell, Electricity, GDP
Categorical Variables: Developed
 Identification Columns: Country, Code
(b) Any answers which contain two questions about individual variables are acceptable. For example, a question could be: Which country has the highest percentage of government expenditures directed toward the military?
(c) Any answers which contain two questions about relationships between any two (or more) variables is acceptable. For example, a question could be: How does the percentage of government expenditures directed towards healthcare in a country compare to the percentage of the population with HIV?

1.14 Spider Sex Play Spiders regularly engage in spider foreplay that does not culminate in mating. Male spiders mature faster than female spiders and often practice the mating routine on not-yet-mature females. Since male spiders run the wrist of getting eaten by female spiders, biologists wondered why spiders engage in this behavior. In one study, some spiders were allowed to participate in these nearmatings, while other maturing spiders were isolated. When the spiders were fully mature, the scientists observed real matings. They discovered that if either partner had participated at least once in mock sex, the pair reached the point of real mating significantly faster than inexperienced spiders did. (Mating faster is, apparently, a real advantage in the spider world.) Describe the variables, indicate whether each variable is quantitative or categorical, and indicate the explanatory and response variables.

Solution
Variables:
Isolated or not while immature - categorical - explanatory variable
Time until real mating occurred - quantitative - response variable

1.20 Rowing Solo Across the Atlantic Ocean On January 14, 2012, Andrew Brown of Great Britain set the world record time (40 days) for rowing solo across the northern Atlantic Ocean. On March 14, 2010, Katie Spotz of the United States became the youngest person to ever row solo across the Atlantic when she completed it in 70 days at the age of 22 years old. Table 1.3 shows times for males and females who rowed solo across the Atlantic Ocean in the last few years.
(a) How many cases are there in this dataset? How many variables are there and what are they? Is each categorical or quantitative?
(b) Display the information in Table 1.3 as a dataset with cases as rows and variables as columns.

Table 1: **Table 1.3 Number of days to row alone across the Atlantic Ocean**

| Male times: 40, 78, 78, 106, 67 |
| Female times: 70, 153, 81       |

Solution
(a) There are 8 cases. There are two variables - gender and number of days to cross the Atlantic. The variables are quantitative.

(b)

<table>
<thead>
<tr>
<th>Time</th>
<th>Gender</th>
</tr>
</thead>
<tbody>
<tr>
<td>40</td>
<td>Male</td>
</tr>
<tr>
<td>87</td>
<td>Male</td>
</tr>
<tr>
<td>78</td>
<td>Male</td>
</tr>
<tr>
<td>106</td>
<td>Male</td>
</tr>
<tr>
<td>67</td>
<td>Male</td>
</tr>
<tr>
<td>70</td>
<td>Female</td>
</tr>
<tr>
<td>153</td>
<td>Female</td>
</tr>
<tr>
<td>81</td>
<td>Female</td>
</tr>
</tbody>
</table>

**1.24 Political Party and Voter Turnout** Suppose that we want to investigate the question “Does voter turnout differ by political party?” How might we collect data to answer this question? What would the cases be? What would the variable(s) be?

Solution
Answers will vary. 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.

Note: For Exercises 1.47 to 1.51, indicate whether we should trust the results of the study. Is the method of data collection bias? If it is, explain why.

**1.47** Ask a random sample of students at the library on a Friday night “How many hours a week do you study?” to collect data to estimate the average number of hours a week that all college students study.

Solution
This sample is biased since students who are in the library on a Friday night may tend to study more than the average college student.

**1.48** Ask a random sample of people in a given school district “Excellent teachers are essential to the well-being of children in this community, and teachers truly deserve a salary raise this year.
Do you agree? Use the results to estimate the proportion of all people in the school district who support giving teachers a raise.

Solution
This sample is biased because of the way the question is asked. Since the question mentions that teachers are essential to the well being of the children, the responders may be persuaded to say that they agree. However, if the question was simply “Should teachers receive a raise?”, the responses may have been different.

1.49 Take 10 apples off the top of a truckload of apples and measure the amount of bruising on those apples to estimate how much bruising there is, on average, in the whole truckload.

Solution
This sample is biased since apples on the top of the truckload may tend to have less bruises than apples at the bottom of the truckload.

1.50 Take a random sample of one type of printer and test each printer to see how many pages of text each will print before the ink runs out. Use the average from the sample to estimate how many pages, on average, all printers of this type will last before the ink runs out.

Solution
This sample is not biased.

1.51 Send an email to a random sample of students at a university asking them to reply to the question: “Do you think this university should fund an ultimate frisbee team?” A small number of students reply. Use the replies to estimate the proportion of all students at the university who support this use of funds.

Solution
This sample is biased since the students involved volunteered to answer the survey.

1.53 How Many People Wash Their Hands After Using the Washroom? In Example 1.10 on page 16, we introduce a study by researchers from Harris Interactive who were interested in determining what percent of people wash their hands after using the washroom. They collected data by standing in public restrooms and pretending to comb their hair or put on make-up as they observed patrons’ behavior. Public restrooms were observed at Turner’s Field in Atlanta, Penn Station and Grand Central Station in New York, the Museum of Science and Industry and the Shedd Aquarium in Chicago, and the Ferry Terminal Farmers Market in San Francisco. Of the over 6000 people whose behavior was observed, 85% washed their hands. Women were more likely to wash their hands: 93% of women washed, while only 77% of men did. The Museum of Science and Industry in Chicago had the highest hand-washing rate, while men at Turner’s Field in Atlanta had the lowest.

(a) What are the cases? What are the variables? Classify each variable as quantitative or categorical?

(b) In a separate telephone survey of more than 1000 adults, more than 96% said they always wash their hands after using a public restroom. Why do you think there is such a discrepancy in the percent from the telephone survey compared to the percent observed?
Solution
(a) The cases are the 6000 restroom patrons observed. The variables are wash, gender, and location. They are categorical variables.

(b) The discrepancy may be caused, because people are not always honest in self-reporting.

1.56 Effects of Alcohol and Marijuana In 1986 the Federal Office of Road Safety in Australia conducted an experiment to assess the effects of alcohol and marijuana on mood and performance. Participants were volunteers who responded to advertisements for the study on two rock radio stations in Sydney. Each volunteer was given a randomly determined combination of the two drugs, then tested and observed. Is the sample likely representative of all Australians? Why or why not?

Solution
The sample is probably not representative of all Australians, because first of all the people who participated were volunteers. Additionally, the advertisements were only made to people who list to rock radio stations. Thus, not every Australian would have heard the advertisements.

1.63 Strawberry Fields A strawberry farmer has planted 100 rows of plants, each 12 inches apart, and there are about 300 plants in each row. He would like to select a random sample of 30 plants to estimate the average number and weight of the berries per plant.

(a) Explain how he might choose the specific plants to include in the sample.

(b) Carry out your procedure from (a) to identify the first three plants selected for the sample.

Solution
(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 30000 and randomly select 30 numbers between 1 and 30000.

(b) Answers will vary for this question, but the procedure should be explained and the three numbers which were obtained should be listed.

Here is the start of one sample.
Row Plant
#94 #180
#83 # 81
#10 #222

1.84 Salt on Roads and Accidents Three situations are described at the start of this section, on page 29. In the third bullet, we describe an association between the amount of salt spread on the roads and the number of accidents. Describe a possible confounding variable and explain how it fits the definition of a confounding variable.

Solution
The weather condition could be a confounding variable. It is possible that the times when more salt are applied to the roads correlate with bad weather conditions such as snowfall and strong winds. Thus, the accidents could occur, because of the weather conditions, but it also happens that these
are the times when more salt has been applied to the road. Notice that the confounding variable has an association with both the variables of interest.

**1.87 Single-Sex Dorms and Hooking Up** The president of a large university recently announced that the school would be switching to dorms that are all single-sex, because, he says, research shows that single-sex dorms reduce the number of student hook-ups for casual sex. He cites studies showing that, in universities that offer both same-sex and coed housing, students in coed dorms report hooking up for casual sex more often.

(a) What are the cases in the studies cited by the university president? What are the two variables being discussed? Identify each as categorical or quantitative.

(b) Which is the explanatory variable and which is the response variable?

(c) According to the second sentence, does there appear to be an association between the variables?

(d) Use the first sentence to determine whether the university president is assuming a causal relationship between the variables.

(e) Use the second sentence to determine whether the cited studies appear to be observational studies or experiments.

(f) Name a confounding variable that might be influencing the association. (Hint: Students usually request one type of dorm or the other.

(g) Can we conclude from the information in the studies that single-sex dorms reduce the number of student hook-ups?

(h) What common mistake is the university president making?

**Solution**

(a) The cases being studied are the students in the dorms. The two variables are the type of dorm the student lives in and the number of hook ups. The type of dorm is a categorical variable, and the number of hook ups is quantitative.

(b) Explanatory variable - type of dorm; response variable - number of hook ups

(c) Yes, there is an association between the variables.

(d) Yes, he is assuming a causal relationship.

(e) The cited studies appear to be observational studies.

(f) Since students tend to select where they want to live, it may be possible that those who would like to have hook ups would choose to live in the coed dorms, while those who are not as interested in hook ups may choose to live in a single-sex dorm.

(g) No.

(h) He is assuming causation when he shouldn’t be.

**1.95 Alcohol and Reaction Time** Does alcohol increase reaction time? Design a randomized experiment to address this question using the method described in each case. Assume the participants are 40 college seniors and the response variable is time to react to an image on a screen after drinking either alcohol or water. Be sure to explain how randomization is used in each
case.
(a) A randomized comparative experiment with two groups getting two separate treatments.
(b) A matched pairs experiment.

Solution
Answers will vary. These are examples of acceptable solutions.

(a) Use a random process to assign 20 of the students to drink water and the other 20 students to
drink alcohol. Then show them the image and record their reaction.

(b) Have each of the 40 students drink water and alcohol on two different days. Use a random
process to determine whether a student will drink water on the first day and alcohol on the second
day or alcohol on the first day and water on the second day.

Computer Exercises

Before completing these exercises, you will need to install R onto your computer. The site
[cran.r-project.org](http://cran.r-project.org) has installation information and directions for both Mac and Windows com-
puters. Also see the R Users Guide from the textbook authors on the course web page. You can
get easy access to all data from the textbook by installing the package Lock5Data in R. To do this,
after installing and starting R, install the package. You only need to do this step once.

```r
> install.packages("Lock5Data")
```

To actually load the this package into your active session, type the following command. You need
to do this each session.

```r
> library("Lock5Data")
```

To load a data set from the library, use the `data()` command. For example, to load the AllCountries
data set, do this.

```r
> data(AllCountries)
```

Alternatively, without the package you may load this data into R from the file. The following
command assumes that the file with the data is in your working directory.

```r
> AllCountries = read.csv("AllCountries.csv")
```

**R problem 1** Load the data set AllCountries from the textbook into R. Write a function that
will take a random sample of some of these countries and calculate the mean land area in square
kilometers. Use the function to take sample of ten of these countries. For the assignment you turn
in on paper, include the ten sampled countries and their mean land area. In addition, send an
email message to the instructor and the TA with an attached file named countries.R that includes
your code. We should be able to source the file into R and run your code successfully (assuming
we run R in a directory with the file that contains the data).

Solution

R Code:

```r
# Load the data AllCountries into R
data(AllCountries)

# Look at the structure of the data
```
str(AllCountries)

# Look at the variables in the data set
names(AllCountries)

# Attach the variable names so we can use them
attach(AllCountries)

# Write a function which will take a random sample of the countries
countries <- function(size){
  sample <- sample(1:213,size)
  data <- AllCountries[sample,c(1,3)]
  print("The countries sampled and their land area are:",quote=FALSE)
  print(data)
  mean <- mean(data[1:10,2])
  print("The mean is:",quote=FALSE)
  return(mean)
}

# Run the function
countries(10)

Example of R output after using function:

> countries(10)

[1] The countries sampled and their land area are:

        Country    LandArea
   62  Ethiopia  1000000
   56   Ecuador  248360
 197  Turks and Caicos Islands     950
   64     Fiji     18270
   11  Australia  7682300
 113  Lithuania   62670
   72    Ghana    227540
 135   Myanmar    653520
   80 Guinea-Bissau   28120
   99     Kenya   569140

[1] The mean is:
[1] 1049087

R problem 2 The game Dominion is a card game that uses cards unique to the game. In the basic game, there are 25 kingdom cards, but only 10 will be used in any given game. The names of the cards are listed below. Write an R function called dominion() that will randomly sample ten of the 25 card types and write their names to the screen. Send an email message to the TA and instructor with a file called dominion.R that contains your code. We should be able to source the file into R and run your code successfully.

Card names: Adventurer, Bureaucrat, Cellar, Chancellor, Chapel, Council Room, Feast, Festival, Laboratory, Library, Market, Militia, Mine, Moat, Money Lender, Remodel, Smithy, Spy, Thief,

Solution

R Code:

```r
# Write a function that will sample 10 of the kingdoms
dominion <- function(size=10)
{
  sample <- sample(kingdoms, size)
  return(sample)
}

# Use the function to sample 10 kingdoms
dominion(10)
```

Example of R output after using function:

```r
> dominion(10)
[1] "Mine"    "Laboratory" "Council Room" "Militia"    "Cellar"
[6] "Money Lender" "Gardens"  "Library"   "Chapel"   "Chancellor"
```