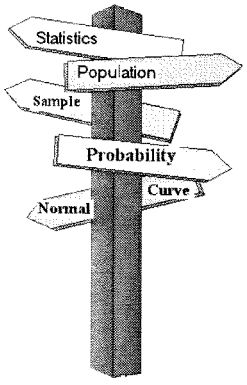
CHS Statistics

Summer Work

Chapter 1 Introduction
Variables and Process in Statistics



Zegeer (2013-2014)

CHAPTER 1

Introduction: Variables and Processes in Statistics

What can you accomplish with this book, and how?

Broker/Bigstockphoto.com

What's the key to solving statistics problems?

Before the semester starts, a statistics teacher wants to organize a box of hundreds of newspaper clippings and Internet reports collected in the past couple of years:

- "Dark Chocolate Might Reduce Blood Pressure"
- "Almost Half of U.S. Internet Users 'Google' Themselves"
- "Vampire Bat Saliva Researched for Stroke"
- "Environmental Mercury, Autism Linked by New Research"

There are several reports on smoking and on obesity, but for most of the topics—such as bat saliva—there is only one article. How can the teacher sort all of those articles in a way that will make them easy to access for future reference?

At the end of the semester, a group of statistics students are studying together, trying to solve practice final exam problems such as these:

- Suppose systolic blood pressures for 7 patients who ate dark chocolate daily for two weeks dropped an average of 5 points, whereas those of a control group of 6 patients who ate white chocolate remained unchanged. If the standardized difference between blood pressure decreases was 2.1, do we have convincing evidence that dark chocolate is beneficial?
- According to a 2007 report, 47% of 1,623 U.S. Internet users surveyed by the Pew Internet & American Life Project had searched for information about themselves online. Give a 95% confidence interval for the percentage of all U.S. Internet users who searched online for information about themselves.

- Researchers found that 9 out of 15 stroke patients receiving vampire bat saliva had an excellent recovery, compared with 4 out of 17 who were untreated. Does this provide evidence that bat saliva is effective in treating stroke patients?
- Research in a large sample of Texas school districts found that for every 1,000 pounds of environmentally released mercury, there was a 17% increase in autism rates. If one district has 300 additional pounds of environmental mercury compared to another, how much higher do we predict its autism rate to be?

The students may feel overwhelmed in trying to find the right approach to each of the problems, after having learned a whole semester's worth of various statistical procedures. How can the students figure out which procedure is the right one for each problem?

The answer for both teacher and students is a simple one, and it will also be the key for you to understand what this book is all about, from beginning to end. *The way we handle statistical problems depends on the number and types of variables involved.*

A *variable*, as the name suggests, is something that varies for different individuals: Blood pressure is a variable because it takes different values for different people; recovery from a stroke is a variable because some patients have an excellent recovery and others do not. The individuals with variable traits in many cases are people, but individuals can be anything that we are interested in—from penguins to school districts to planets.

The Big Picture: A Closer Look

Categorical variables are sometimes referred to as "qualitative" and quantitative variables are sometimes called "numerical."

Types of Variables: Categorical or Quantitative

Virtually all of the situations encountered in this book will involve either a *single variable* or the *relationship between two variables*. A variable's type is *categorical* if it takes qualitative values such as sex, race, or the response to a yes-or-no question. The type is *quantitative* if the variable takes number values for which arithmetic makes sense, such as age, number of siblings, or rating something on a scale of 1 to 10.

Definitions A **variable** is a characteristic that differs for different individuals. A **categorical variable** takes qualitative values that are not subject to the laws of arithmetic. A **quantitative variable** takes number values for which arithmetic makes sense. A **relationship** (also known as an **association**) exists between two variables if certain values of one tend to occur with certain values of the other.

The Big Picture: A Closer Look

Some number-valued variables, like ZIP codes, are categorical if the numbers are labels, not signifying an amount that can be quantified. For example, if half of a group of students have a ZIP code 15217 and the other half 15213, we can't say that the typical ZIP code is the average of these, 15215.

The statistics teacher can divide the clippings into just five piles:

1. One categorical variable
2. One quantitative variable
3. One categorical variable and one quantitative variable
4. Two categorical variables
5. Two quantitative variables

Likewise, the statistics students just need to identify the number and type of variables involved in each problem, and this will suggest what statistical procedure should be applied.

This book features "Students Talk Stats" examples and exercises that are discussions by four prototypical students, highlighting many of the most important

ideas in statistics. As you gradually rise to higher levels of understanding of statistical concepts and procedures, you may find you can relate to their struggles and discoveries. Our first such discussion will help you begin to develop the skill of identifying what types of variables are involved when you are presented with any report containing statistical information.

Identifying Types of Variables

Four students who have recently enrolled in a statistics class are browsing through news articles on the Internet, thinking about what kind of variables are involved.

Adam: *"I'm in the mood for chocolate, so I'm looking at this article that says* 'Dark Chocolate Might Reduce Blood Pressure'. *I'm pretty sure blood pressure is quantitative but couldn't chocolate go either way?"*

Brittany: *"Realistically, I think they'd just compare people who do and don't eat dark chocolate, which would make it categorical. Here's one that says* 'Almost Half of U.S. Internet Users "Google" Themselves'. *Half is a number so it's quantitative."*

Carlos: *"Half is talking about the overall fraction, but for each person, they just recorded whether or not they Googled themselves, so it's categorical. What about* 'Vampire Bat Saliva Researched for Stroke'? *I picture they handled bat saliva like Brittany said they'd handle chocolate—some people get it and others don't. I don't think it would be easy to put a number on recovery from a stroke, so that variable's probably categorical, too."*

Dominique: *"I'm confused about this one:* 'Environmental Mercury, Autism Linked by New Research'. *Mercury would be quantitative, and I think of autism as being categorical, but the report says they looked at autism* rates *in different school districts depending on how much mercury was in the area. Would that make autism quantitative?"*

Adam is correct that blood pressure is quantitative, and Brittany rightly guesses that chocolate consumption in this case would be categorical. Carlos has correctly identified Googling one's self as a categorical variable in the second article, and is on the right track that both bat saliva and stroke recovery would be categorical. Finally, although autism for individual people would be categorical, if a study considers autism *rates* for a sample of school districts, then the variable is quantitative. Dominique is right about both mercury and autism rate being quantitative variables in this study.

Practice: *Try Exercise 1.2 on page 11.*

Although variable type is usually fairly straightforward to identify, some "crossover" from one type to the other may take place, such as in the autism/mercury study discussed above by the four students, as well as in the following pair of examples.

EXAMPLE 1.1 WHEN A CATEGORICAL VARIABLE GIVES RISE TO A QUANTITATIVE VARIABLE

Background: Individual teenagers were surveyed as to whether they have used marijuana, and whether they have used harder drugs.

Teenager	Marijuana?	Harder Drugs?
#1	Yes	Yes
#2	No	No
#3	No	No
#4	Yes	No
...	...	...

Researchers then looked at the percentage of teenagers using marijuana and the percentage using harder drugs in various countries around the world to see if those two variables are related.

Country	% Marijuana	% Harder Drugs
#1	22	4
#2	37	16
#3	7	3
#4	23	14
...	...	...

Questions: What kinds of variables are involved in the first situation? What kinds of variables are involved in the second situation?

Responses: The first situation explores the relationship between two categorical variables. The second explores the relationship between two quantitative variables.

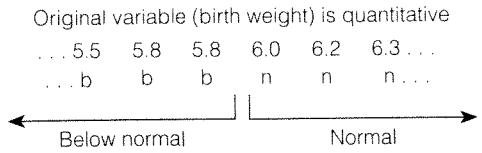
Practice: ***Try Exercise 1.6(a,b) on page 12.***

EXAMPLE 1.2 WHEN A QUANTITATIVE VARIABLE GIVES RISE TO A CATEGORICAL VARIABLE

Researchers studied the effects of having dental X-rays during pregnancy. First they recorded birth weights of babies, along with information as to whether the mothers had been X-rayed while pregnant. When it came time to analyze the data, they simply classified the babies as being below 6 pounds (considered below normal) or not, along with information about whether the mothers had been X-rayed while pregnant.

Questions: What kind of variables were involved in the first situation? In the second situation?

Responses: The first situation involves one categorical variable (mother had dental X-rays or not) and a quantitative variable (baby's weight). The second situation involves two categorical variables because babies' weights are now categorized into two groups.



Practice: *Try Exercise 1.8 on page 12.*

LOOKING AHEAD

Many real-life studies, including many discussed in this book, convert quantitative variables to categorical in order to simplify matters.

Handling Data for Two Types of Variables

We refer to recorded values of categorical or quantitative variables as *data*. The science of *statistics* is all about handling data.

Data are pieces of information about the values taken by variables for a set of individuals.

The science of **statistics** concerns itself with gathering data about a group of individuals, displaying and summarizing the data, and using the information provided by the data to draw conclusions about a larger group of individuals.

Before we go into detail about the process of gathering data, it helps to have an idea of how we will handle the data when the time comes. Categorical variables are summarized by telling *count* or *proportion* or *percentage* in the category of interest. The most common way of summarizing quantitative variables is with their *mean* (same thing as *average*), although we will discuss other useful summaries a bit later in this book.

The **count** in a category of interest is simply the number of individuals in that category.

The **proportion** in a category of interest is the number of individuals in that category, divided by the total number of individuals considered.

The **percentage** in a category of interest is the proportion (as a decimal) multiplied by 100%.

The **mean** of a set of values is their sum divided by the total number of values.

Students may be misled to think that the variable of interest in a situation is quantitative because they see a number attached to it. In fact, that number may be a count or a proportion or a percentage summarizing values of a categorical variable. It may help to think about how data values are being recorded *for each individual* in a sample in order to decide whether the variable of interest is categorical or quantitative, as Carlos did in the four students' discussion on page 3.

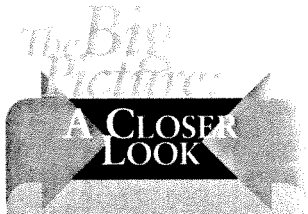
EXAMPLE 1.3 SUMMARIZING CATEGORICAL VARIABLES

Background: An article entitled "New Test-Taking Skill: Working the System" reports: "Indeed, although only a tiny fraction—1.9%—of students nationwide got special accommodations for the SAT, the percentage jumps fivefold for students at New England prep schools. At 20 prominent Northeastern private schools, nearly one in 10 students received special treatment."[1]

Question: What type of variable is featured here, and how is it summarized?

Response: For each student in the entire nation or in the private schools examined, it is recorded whether or not the student was granted special accommodations in taking the SAT test. This is a categorical variable, summarized by telling the percentage or proportion in the category of interest (receiving special accommodations).

Practice: ***Try Exercise 1.10 on page 12.***



In cases like this, where values of a quantitative variable are being compared for two or more categorical groups, a summary occasionally quantifies the differences by reporting what percent higher or lower another mean is from the original mean.

EXAMPLE 1.4 SUMMARIZING QUANTITATIVE VARIABLES

Background: An article entitled "Racial Gaps in Education Cause Income Tiers" reports: "On average, a white man with a college diploma earned about $65,000 in 2001. Similarly educated white women made about 40% less, while black and Hispanic men earned 30% less. . ."[2]

Question: How would earnings for each group (such as white women or Hispanic men) be summarized—with a mean or with a proportion?

Response: Earnings are a quantitative variable and could be summarized for each group with a mean, namely $65,000 for college-educated white men, and a mean that is less by 40% of $65,000 for college-educated white women—that is, $39,000.

Practice: ***Try Exercise 1.11 on page 12.***

Most of the data that statisticians handle, and most of the data that we encounter in our everyday lives, come from some subgroup, called a *sample*, as opposed to the entire group of interest, called the *population*. Occasionally, we have access to information about the entire population, gathered via a *census*. This was the case in Example 1.4 about earnings of various demographic groups in the United States.

Definitions A sample is a subset taken from a larger group, and the larger group of interest is the population.

A census, according to Webster's dictionary, is a "usually complete enumeration of the population," and we think of a census in general as a survey intended to include all citizens in a given area. When we talk about "*the* Census," we are referring to the U.S. Census, conducted regularly since 1790, and designed to gather more and more detailed information about America's population.

Once census results are summarized, as in Example 1.4, there are no further statistical procedures needed to draw conclusions about the "larger population."

EXAMPLE 1.5 WHEN INFORMATION IS PROVIDED FOR AN ENTIRE POPULATION

Background: "Are Feeding Tubes Over-Prescribed?" describes a Harvard Medical School study that "involved 1999 data from all 15,135 licensed U.S. nursing homes at the time."[3] The study found that "one-third of U.S. nursing home patients in the final stages of Alzheimer's and other forms of dementia are given feeding tubes, despite evidence that the practice serves no benefit and may even cause harm." The variable of interest here is whether or not nursing home patients in the final stages of Alzheimer's or other forms of dementia are given feeding tubes, a categorical variable that is summarized with the proportion 1/3.

Question: Why would it not be appropriate to generalize the study's results to a larger population?

Response: It is not possible to generalize this result to a larger group because it already refers to patients in *all* nursing homes at the time, rather than to a sample comprising a subset of those patients.

Practice: ***Try Exercise 1.14 on page 13.***

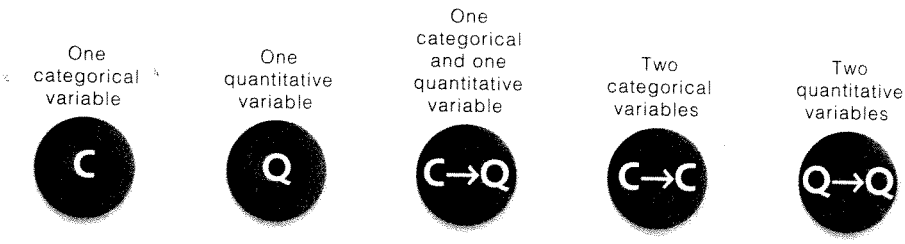
Roles of Variables: Explanatory or Response

By far the most interesting and useful statistical studies involve *relationships* between variables. How we approach the data will depend on what roles the variables play in their relationship. There are occasionally situations where two variables have "equal footing" in the relationship, such as in a study of the relationship between football teams' rankings in offense and in defense. For the most part, however, one variable is thought to cause changes in, or at least to explain, values of the other: It is called the *explanatory variable*. The other variable is impacted by, or responds to, the first: It is called the *response variable*. A more complicated relationship can involve more than one explanatory or response variable.

Causation exists between two variables if changes in values of the first are actually responsible for changes in values of the second.

The **explanatory variable** in a relationship between two variables is the one that is presumed to impact the other variable, called the **response variable.**

In the following diagram of the five possible situations introduced on page 2, the last three involve a relationship. The direction of the arrow goes from explanatory to response variable. Because relatively few actual situations of interest involve a quantitative explanatory and categorical response variable, and because the analysis is fairly advanced compared to the others, we will not analyze such situations in this book.



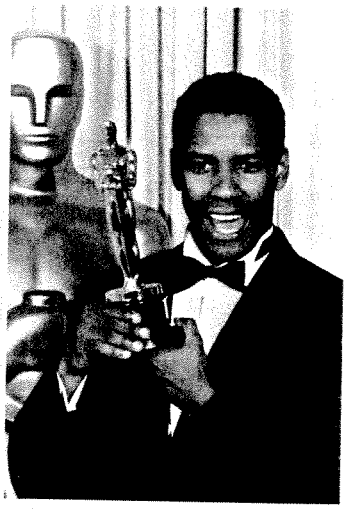
Example 1.6 illustrates the five situations in a variety of contexts.

EXAMPLE 1.6 IDENTIFYING VARIABLE TYPES AND ROLES

Background: Consider these headlines:

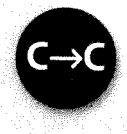
- "Men Are Twice as Likely as Women to Be Hit by Lightning"
- "35% of Returning Troops Seek Mental Health Aid"
- "Do Oscar Winners Live longer Than Less Successful Peers?"
- "Smaller, Hungrier Mice"
- "County's Average Weekly Wages at $811, Better Than U.S. Average"

Questions: What type of variables are involved in each of these situations? If the relationship between two variables is of interest, which plays the role of explanatory variable and which is the response?



Broker/BigstockPhoto.com

Thanks for helping me live longer?



Responses: "Men Are Twice as Likely as Women to Be Hit by Lightning": We consider two categorical variables—gender and whether or not a person is hit by lightning. Gender would be the explanatory variable and being hit by lightning or not is the response. The other way around wouldn't make sense because being hit by lightning could not have an impact on a person's gender.

"35% of Returning Troops Seek Mental Health Aid": Whether or not a returning soldier seeks mental health aid is a single categorical variable.

"Do Oscar Winners Live Longer Than Less Successful Peers?": This involves a categorical explanatory variable—being an Oscar winner or not—and a quantitative response variable—length of life.

"Smaller, Hungrier Mice": This brief headline suggests a relationship between two quantitative variables: the size of a mouse and its appetite. Size apparently plays the role of explanatory variable, so that as size goes down, the amount of food desired goes up.

"County's Average Weekly Wages at $811, Better Than U.S. Average": This involves just one quantitative variable—weekly wages. If wages for one county had been compared to those of another county, then there would have been an additional categorical explanatory variable. Comparing this county's wages to those of the United States in general is a different kind of comparison, where the county residents may be thought of as a single sample, coming from the larger population of U.S. residents.

Practice: *Try Exercise 1.17 on page 13.*

Statistics as a Four-Stage Process

Before we begin to learn about the first stage in the process of statistical analysis, we should consider how all the stages fit together to accomplish our overall goal. On page 5, we stated that, as a science, statistics is used to produce information from a sample, summarize it, and then draw conclusions about the larger population from which the sample came. Those conclusions, known as *statistical inference*, can be reached only if we have some knowledge of the workings of *random* behavior, which comes under the realm of the science of *probability*.

A **random** occurrence is one that happens by chance alone, and not according to a preference or an attempted influence.

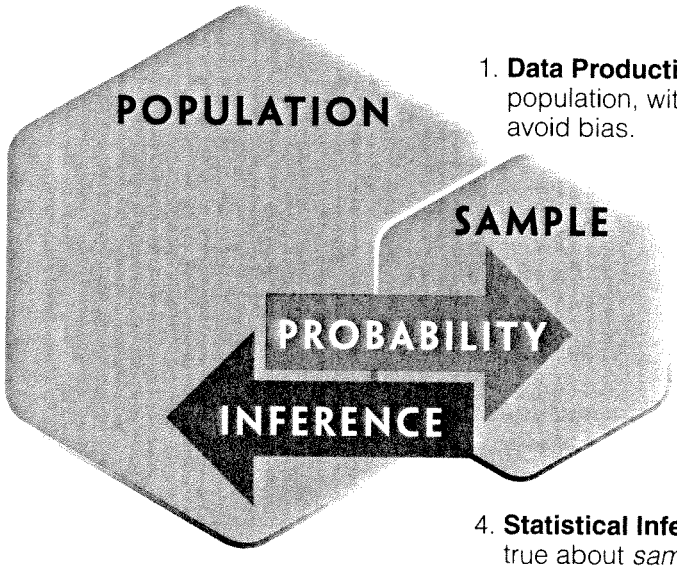
Probability is the formal study of the likelihood or chance of something occurring in a random situation. In the context of statistics, probability explores the behavior of random samples taken from a larger population.

Statistical inference is the scientific process of drawing conclusions about a population based on information from a sample.

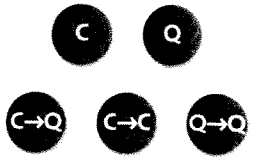
Thus, our goal can be reached in four stages, which will be addressed one at a time in the book's four parts.

1. **Data production:** How to select a representative sample, and how to properly assess values of variables for that sample.
2. **Displaying and summarizing data:** Depicting and describing single quantitative or categorical variables of interest, or relationships between variables if there are two variables involved.
3. **Probability:** The scientific process wherein we assume we actually know what is true for the entire population, and conclude what is likely to be true for a sample drawn at random from that population.
4. **Statistical inference:** Using what we have discovered about the variables of interest in a random sample to draw conclusions about those variables for the larger population.

It is easy for a student to lose sight of these long-term goals, as he or she concentrates on learning particular concepts and techniques. Throughout the book, the following diagram will help remind you of how each new topic fits into the "big picture." A reminder of variable types and roles is included because awareness of the variables involved is always an important part of the statistical picture.



1. **Data Production:** Take sample data from the population, with sampling and study designs that avoid bias.
2. **Displaying and Summarizing:** Use appropriate displays and summaries of the sample data, according to variable types and roles.



3. **Probability:** Assume we know what's true for the *population;* how should **random** *samples* behave?
4. **Statistical Inference:** Assume we only know what's true about *sampled* values of a single variable or relationship; what can we **infer** about the larger *population?*

EXAMPLE 1.7 IDENTIFYING THE FOUR PROCESSES

Background: Consider the following situations:

- A retail manager is asked to present some graphs and a brief report on her group's sales over the past several months, broken down into various types of merchandise.
- Before a bookstore's owners make plans for extensive renovations, they want to find out what customers already like about the store and what aspects are in need of change.
- A pharmaceutical company has carried out a study and determined proportions of patients experiencing nausea for those who take a certain medication and those who take a "dummy pill." The company wants to know what claims it can make about proportions of patients experiencing nausea in the general population for those who take the medication compared to those who don't.
- The proportion of all Americans who are of Hispanic origin is 0.13. We'd like to know how unlikely it would be to take a random sample of 1,000 Americans and find only 0.06 to be Hispanic.

Question: Which of the four processes is involved in each situation?

Response:

- The first is a task in displaying and summarizing data—namely, the information on sales of various types of merchandise.
- The second requires data production—namely, the design and implementation of a survey of the bookstore's customers.
- The third is a statistical inference problem, using information on tested patients to draw conclusions about side effects for any user of the medication.
- The final one is a probability problem because we seek the likelihood of obtaining a certain proportion in our sample who are Hispanic.

Practice: ***Try Exercise 1.23 on page 14.***

SUMMARY

Characteristics that can differ from one individual to another are called **variables.** Variables can be either categorical or quantitative. In statistics, we study single variables or relationships between variables. At times we merely focus on variables' values for a specific set of individuals, called a **sample.** More often, our goal is to generalize to a larger group, called the **population.**

Variables and Statistics

- Data are pieces of information about the values taken by variables for a set of individuals.
- The five variable situations to be covered in this book are:
  1. Single categorical variable
  2. Single quantitative variable
  3. Categorical explanatory and quantitative response variable
  4. Categorical explanatory and categorical response variable
  5. Quantitative explanatory and quantitative response variable

  Categorical variables can be summarized with **counts, proportions,** or **percentages.**

  Quantitative variables can be summarized with **means.**

  If individuals studied are entire groups, the percentage in a particular category for each group can be treated as a quantitative variable.

  A quantitative variable can be converted into a categorical variable by grouping into ranges of values.
- The science of statistics is concerned with gathering data, summarizing it, and using that information to draw conclusions about a larger population. The latter process is known as **statistical inference.**
- A **census** gathers information about an entire population rather than just a sample.
- When the relationship between two variables is of interest, it should be determined which (if any) plays the role of **explanatory variable** and which is the **response variable.**
- A **random** occurrence is one that happens by chance alone, and **probability** is the formal study of randomness.
- The four stages in the "big picture" of statistics are
  1. Data production
  2. Displaying and summarizing data
  3. Probability
  4. Statistical inference

EXERCISES

Note: Asterisked numbers indicate exercises whose answers are provided in the Solutions to Selected Exercises section, on page 689.

*1.1 Students were asked to rate their instructor's preparation for class as being excellent, good, or needs improvement. Response to this question is what types of variable—quantitative or categorical?

*1.2 Suppose researchers want to investigate how weight can affect blood pressure. Tell what types of variables each of these situations involves.

a. Individuals' weights and blood pressures are recorded.

b. Individuals are classified as being normal or overweight, and their blood pressures are recorded.

c. Individuals are classified as having high or low blood pressure, and their weights are recorded in kilograms.

d. Individuals are assessed as having high or low blood pressure, and as being normal or overweight.

1.3 Prospective subjects for a study had their blood pressures recorded.

a. Is the variable of interest quantitative or categorical?

b. Would results best be summarized with a mean or with a proportion?

1.4 Before the 2004 presidential election in the United States, there was a great deal of interest concerning public opinion of the war in Iraq. For each of the following situations, tell what individuals are being studied, what variable is of interest, and whether the variable is categorical or quantitative.

a. People around the world were surveyed as to whether they approved or disapproved of the Iraq war.

b. People in various countries were surveyed as to whether they approved or disapproved of the Iraq war. For each country, it was determined what percentage of its people disapproved of the war.

c. The *Guardian*—a British newspaper—reported that 8 of 10 countries surveyed by leading newspapers (such as the *Guardian*, Canada's *La Presse*, and Japan's *Asahi Shimbun*) disapproved of the Iraq war.

1.5 Based on a survey of a few thousand people, a newspaper reporter wants to draw conclusions about how a country's citizens in general feel about the war in Iraq. At this point, is the reporter mainly concerned with data production, displaying and summarizing data, probability, or performing statistical inference?

*1.6 For parts (a) and (b), tell who or what individuals are being studied, identify the variable of interest, and tell whether it is categorical or quantitative; then answer the question in part (c).

a. Adults were surveyed as to whether they were married, single, or divorced.

b. The *New York Times* reported, state by state, the divorce rate per 1,000 married adults in 2003. The lowest rate was in Massachusetts, with 5.7 divorces per 1,000 married people, and the highest was in Nevada, with 14.6 per 1,000.

c. Assume we have Census data on marital status of people in the United States. Are those people considered to be a sample or a population?

1.7 A *New York Times* reporter decides to convey information about American divorce rates by including a map of the United States. Each state is shaded from light to dark depending on how high its divorce rate is. At this point, is the reporter mainly concerned with data production, displaying and summarizing data, probability, or performing statistical inference?

*1.8 "Can Mom's Drinking Lower Kids' IQ?" examined the relationship between mothers' consumption of alcohol during pregnancy and their children's IQs. The mothers were classified as being abstainers (0 alcoholic drinks per day), light drinkers (up to 0.5 per day), moderate drinkers (0.5 to 1 per day), or heavy drinkers (more than 1 per day). Is alcohol consumption being treated as a categorical or a quantitative variable?

1.9 An article reported costs of ski-lift tickets in various resorts in a region as being less than \$20, \$20 to \$40, \$40 to \$50, or more than \$50. Is ticket price being treated as a categorical or a quantitative variable?

*1.10 A British survey reported in 2006 states: "Nearly 40 percent of 106 students who answered questionnaires about their attitudes said they couldn't cope without their cell phone."[4]

a. What type of variable is being considered?

b. How is the variable summarized?

*1.11 "In a study of 87 French and Swiss college students, researchers gave half of them sunscreen with a protection factor of 10 and the other half with a factor of 30. The students, who weren't told which lotion they received, went on summer vacations and recorded the amount of time they spent in the sun. Users of the stronger sunscreen spent 25% more time in the sun, mostly sunbathing, the study found . . . students in the study often waited until their skin turned red before rushing to the shade."[5]

a. Is time spent in the sun being treated as a quantitative or a categorical variable?

b. How would researchers summarize time spent in the sun for each group (those with the stronger and those with the weaker sunscreen)?

1.12 A newspaper article entitled "Teens Most Likely to Have Sex at Home" notes that of the sexually active teens surveyed in the year 2000, "56% said they first had sex at their family's home or at the home of their partner's family."[6]

a. What is the variable of interest?

b. Is the variable of interest quantitative or categorical?

c. How is the variable being summarized?

1.13 Based on results of a survey of sexually active teenagers, sociologists would like to be able to say whether or not a majority of all sexually active teenagers first had sex at their or their partner's home. At this point, are the sociologists mainly concerned with data production, displaying and summarizing data, probability, or performing statistical inference?

*1.14 The *New York Times* reports: "Three out of four workers drove to their jobs by themselves in 2006, according to another finding by the Census Bureau."[7] Should we consider the workers studied to be a sample or a population?

1.15 Mortality rates in the United States during the 1980s and 1990s were studied by county, race, gender, and income, with the following results: "Asian-Americans, average per-capita income of $21,566, have a life expectancy of 84.9 years . . . Western American Indians, $10,029, 72.7 years . . ."[8] Are these numbers referring to samples or populations?

1.16 The American Association of Retired People (AARP) conducted a survey in which it was discovered that 63% of adult Americans don't want to live to be at least 100. On average, those polled wanted to live to the age of 91.

a. Should we consider the Americans polled to be a sample or a population?

b. There is a categorical variable of interest in the survey; tell roughly how the survey question was phrased to obtain those responses.

c. There is a quantitative variable of interest in the survey; tell roughly how the survey question was phrased to obtain those responses.

*1.17 The *New York Times* reported on a study of gadgets and appliances in American homes. For each of the following results, tell which of the five variable situations is involved, choosing from the following:

- C: single categorical variable
- Q: single quantitative variable
- C → Q: categorical explanatory variable and quantitative response variable
- C → C: categorical explanatory variable and categorical response variable
- Q → Q: quantitative explanatory variable and quantitative response variable

a. For each of the 17 appliances studied, the *Times* reported the percentage of American homes in 2001 that had the appliance. For example, microwave ovens were in 96% of the homes and answering machines were in 78% of the homes. (1) C (2) Q (3) C → Q (4) C → C (5) Q → Q

b. The study made a comparison of percentage owning each appliance in 2001 to the percentage owning the appliance in 1987. For example, microwave ovens were in 66% of the homes in 1987 as opposed to 96% in 2001. Answering machines were in 10% of the homes in 1987 as opposed to 78% in 2001. (1) C (2) Q (3) C → Q (4) C → C (5) Q → Q

c. The study reported 2.5 television sets owned per household in 2001. (1) C (2) Q (3) C → Q (4) C → C (5) Q → Q

1.18 The *New York Times* reported on a study of gadgets and appliances in American homes. For each of the 17 appliances studied, it told the percentage of American homes in 2001 that had the appliance. For example, microwave ovens were in 96% of the homes and answering machines were in 78% of the homes.

a. Who or what are the individuals being studied?

b. What is the variable of interest?

c. Is the variable of interest quantitative or categorical?

1.19 The study that looked at prevalence of various appliances in homes in 2001, as described in Exercises 1.17 and 1.18, made a comparison to the percentages for each appliance in 1987. For example, microwave ovens were in 66% of the homes in 1987 as opposed to 96% in 2001. Answering machines were in 10% of the homes in 1987 as opposed to 78% in 2001.

a. There are two variables involved; what is the explanatory variable?

b. Tell whether the explanatory variable is quantitative or categorical.

c. What is the response variable?

d. Tell whether the response variable is quantitative or categorical.

e. In which year would you expect percentages to be higher overall—1987 or 2001, or both the same?

1.20 The *New York Times* study of appliances reported 2.5 television sets per household in 2001.

a. Is the variable of interest quantitative or categorical?

b. Is the reported summary a mean or a proportion?

1.21 Suppose television advertisers want to know if age plays a role in people's response to a rather unconventional ad that might be aired during the next Super Bowl. The ad is shown to a variety of viewers. Keeping in mind that the explanatory variable is not necessarily the first one mentioned, classify each of the following possible approaches as involving one of these relationships:

- C → C: categorical explanatory variable and categorical response variable
- C → Q: categorical explanatory variable and quantitative response variable
- Q → C: quantitative explanatory variable and categorical response variable
- Q → Q: quantitative explanatory variable and quantitative response variable

a. They ask whether or not a viewer likes the ad, and record his or her age.

b. They classify a viewer as being youth, young adult, middle-aged, or senior citizen, and whether or not he or she likes the ad.

c. Viewers' ages are recorded, along with their rating of the ad on a scale of 1 (most unfavorable) to 10 (most favorable).

d. Viewers' ratings of the ad on a scale of 1 to 10 are recorded, along with the viewers' age group as being youth, young adult, middle-aged, or senior citizen.

1.22 Television advertisers are trying to decide which of the approaches outlined in Exercise 1.21 to use in an upcoming study of age and response to an advertisement. At this point, are they mainly concerned with data production, displaying and summarizing data, probability, or performing statistical inference?

*1.23 A department head wants to investigate the quality of teaching of a professor who is coming up for tenure. Tell which of the four processes (data production, displaying and summarizing, probability, or statistical inference) is involved in each of these stages:

a. The department head considers whether to simply ask students to rate various aspects of the professor's performance on a 5-point scale, or whether to also ask them to write a paragraph describing their experience in that professor's class.

b. A sample of students is surveyed, and scores on a 5-point scale are averaged for each aspect of the professor's performance.

c. If all of the professor's students would give an average rating no higher than 4.0 on preparedness, it would be very unlikely to get a sample of 20 students' ratings averaging at least 4.3 on preparedness.

d. Based on the responses of sampled students, the department head concludes that the mean preparedness rating for *all* of the professor's students is higher than 4.0.

1.24 *Men's Health* magazine used data on body mass index, back-surgery rates, usage of gyms, etc. to grade the quality of men's "abs" (abdominal muscles) in 60 cities across the country. If each city was given a rating between 0 and 4, such as 2.75 for Pittsburgh, then how is the variable of interest being treated—as quantitative or categorical?

1.25 Suppose *Men's Health* magazine wants to present the results of the survey described in Exercise 1.24 in a way that is both appealing

and informative. Is the magazine mainly concerned with data production, displaying and summarizing data, probability, or performing statistical inference?

1.26 Anthropologists studied gender differences in public restroom graffiti, noting whether the graffiti occurred in a men's or women's room, and classifying writings as being competitive and derogatory or advisory and sympathetic.

a. There are two variables mentioned here; what is the explanatory variable?
b. Tell whether the explanatory variable is quantitative or categorical.
c. What is the response variable?
d. Tell whether the response variable is quantitative or categorical.
e. Would type of writings for each gender be summarized with means or proportions?

Typical graffiti for women's room?

1.27 If researchers report that alcoholics are three times as likely to smoke compared to nonalcoholics, do they consider smoking to be the explanatory variable or the response?

1.28 If researchers report that smokers are 10 times as likely to be alcoholics compared to nonsmokers, do they consider smoking to be the explanatory variable or the response?

1.29 The Centers for Disease Control and Prevention noted that "the price of a pack of cigarettes went up 90% between 1997 and 2003."[9] Suppose students in an introductory statistics course have been asked to identify the two variables of interest here, then tell which is explanatory and which is response, and whether each is quantitative or categorical. Which student has the correct answer?

Adam: The explanatory variable is price of cigarettes, and it's categorical because it was summarized with a percentage. The response is year, and it's quantitative because it takes number values.

Brittany: The roles are reversed: Year is the quantitative explanatory variable and price is the categorical response.

Carlos: Year is the explanatory variable, and because just two values are possible, it's categorical. Price is the response and it's quantitative—90% just tells how much the price has changed from the year 1997 to the year 2003.

Dominique: Both variables are quantitative because they both take number values; year is explanatory because it affects the price.

1.30 One-third of all nursing home patients with Alzheimer's and other forms of dementia are given feeding tubes. Researchers want to know how unlikely it would be to find more than half in a random sample of 100 such patients to have been given feeding tubes. Are the researchers mainly concerned with data production, displaying and summarizing data, probability, or performing statistical inference?

Discovering Research: VARIABLE TYPES AND ROLES

1.31 Hand in an article or report about a statistical study; tell what variable or variables are involved and whether they are quantitative or categorical. If there are two variables, tell which is explanatory and which is response. If summaries are mentioned, tell whether they are reporting means or proportions or something else.

Reporting on Research: VARIABLE TYPES AND ROLES

1.32 Use the results of Exercise 1.6 and relevant findings from the Internet to make a report on divorce in the United States that relies on statistical information.