TOPIC 1

Data and Variables

How can statistics be used to help decide the guilt or innocence of a nurse accused of murdering some of her patients? If chief executive officers tend to be taller than average, would this convince you that being tall provides advantages in the business world? Do some students do worse on standardized tests when they are first asked to indicate their race than when they are not, perhaps due to negative stereotypes of their academic ability? In this topic, you will begin your exploration of statistics and start gathering the building blocks and tools needed to address such questions.

Overview

Statistics is the science of reasoning from data, so a natural place to begin your study is by examining what is meant by the term *data*. You will find that data *vary*, and variability abounds in everyday life and in academic study. Indeed, the most fundamental principle in statistics is that of variability. If the world were perfectly predictable and showed no variability, you would not need to study statistics. Thus, you will learn about variables and consider their different classifications. You will also begin to experience the interesting research questions that you can investigate by collecting data and conducting statistical analyses.

Preliminaries

1. Kristen Gilbert, a nurse for a veteran's hospital, was charged with murdering some of her patients by administering fatal doses of a heart stimulant. Some of the evidence presented at her trial was statistical. Researchers combed through the records of all eight-hour shifts at the hospital in the previous eighteen months. What information do you think they recorded about each shift?

2. Telephone companies constantly collect data on cell phone calls in an effort to detect anomalies that might indicate fraud (in other words, that someone other than the rightful owner has been using the phone). What information should you record about each cell phone call, in order to develop a profile for the cell phone owner so that you could then monitor calls for anomalies?

In-Class Activities

Activity 1-1: Student Data

1-1, 1-5, 2-7, 2-8, 3-7, 7-8

You will encounter some fundamental ideas in statistics by first considering the word *statistics* itself.

a. Write two complete and grammatically correct sentences, explaining your primary reason for taking this course and then describing what the term *statistics* means to you.

b. For each word in your response to part a, record the number of letters in the word:

The numbers that you have recorded are data. Not all numbers are data, however. Data are numbers collected in a particular context. For example, the numbers 10, 3, and 7 do not constitute data in and of themselves. They are data, however, if they refer to the number of letters in the first three words of your response to part a.

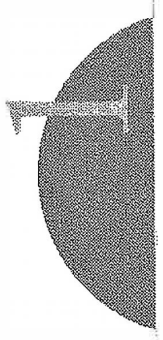
c. Did every word in your two sentences contain the same number of letters?

Although it might be obvious that different words contain different numbers of letters, the fact that the same measurement (e.g., "number of letters") can produce different responses illustrates one of the most fundamental ideas in statistics: variability.

A **variable** is any characteristic of a person or thing that can be assigned a number or a category. The person or thing to which the number or category is assigned, such as a student in your class, is called the **observational unit.** **Data** consist of the numbers or categories recorded for the observational units in a study. **Variability** refers to the phenomenon of a variable taking on different values or categories from observational unit to observational unit.

d. For the data that you recorded in part b, the observational units are the words that you wrote. What is the variable?

A **quantitative** variable measures a numerical characteristic such as height, whereas a **categorical** variable records a group designation such as gender. **Binary** variables are categorical variables with only two possible categories, for example, male and female.



e. Now consider the students in your class as observational units. Classify each of the following variables as categorical or quantitative. If it is categorical, also indicate whether or not it is binary.

- How many hours you have slept in the past 24 hours
- Whether you have slept for at least 7 hours in the past 24 hours
- How many states you have visited
- Handedness (which hand you write with)
- Day of the week on which you were born
- Gender
- Average study time per week
- Score on the first exam in this course

We will continue to focus your attention on observational units and variables in virtually all of the studies and data that you encounter throughout this book. Along the way, keep the following points in mind:

- This distinction between categorical and quantitative variables is quite important because determining which statistical tools to use for analyzing a given set of data often depends on the type of variables involved.
- Notice how the variable of *sleep time* can be measured either quantitatively (first bullet) or categorically (second bullet). This is true of many variables: The classification of the variable often depends on how the quantity is measured more than on any intrinsic property.
- A variable that takes on numerical values that are really just category labels, such as a zip code, is categorical.

Activity 1-4: Studies from *Blink*

1-4, 5-14

The following studies are all described in the popular book *Blink: The Power of Thinking Without Thinking*, by Malcolm Gladwell (2005). For each study, identify the observational units and variables. Also, classify each variable as quantitative or categorical.

a. An economist suspects that chief executive officers (CEOs) of American companies tend to be taller than the national average height of 69 inches, so she takes a random sample of 100 CEOs and records their heights.

Observational units:

Variable: Type:

b. A psychologist shows a videotaped interview of a married couple to a sample of 150 marriage counselors. Each counselor is asked to predict whether the couple will still be married five years later. The psychologist wants to test whether marriage counselors make the correct prediction more than half the time.

Observational units:

Variable: Type:

c. A psychologist gives an SAT-like exam to 200 African-American college students. Half of the students are randomly assigned to use a version of the exam that asks them ro indicate their race, and the other half are randomly assigned to use a version of the exam that does not ask them to indicate their race. The psychologist suspects that those students who are not asked to indicate their race will score significantly higher on the exam than those who are asked to indicate their race.

Observational units:

Variable 1: Type:

Variable 2: Type:

d. An economist randomly assigns four actors to go to ten different car dealerships each and negotiate the best price they can for a particular model of car. The four people are all the same age, dressed similarly, and tell the car salespeople that they have the same occupation and neighborhood of residence. One of the actors is a white male, one is a black male, one is a white female, and one is a black female.

The economist wants to test whether or not the average prices differ significantly among these four types of customers.

Observational units:

Variable 1: Type:

Variable 2: Type:

Variable 3: Type:

Activity 1-6: A Nurse Accused

1-6, 3-20, 6-10, 25-23

Statistical evidence played an important role in the murder trial of Kristen Gilbert, a nurse who was accused of murdering hospital patients by giving them fatal doses of a heart stimulant (Cobb and Gerlach, 2006). Hospital records for an eighteen-month period indicated that of the 257 eight-hour shifts that Gilbert worked, a patient died on 40 of those shifts (15.6%). But during the 1384 eight-hour shifts that Gilbert did not work, a patient died on only 34 of those shifts (2.5%). (You will learn how to analyze such data in Topics 6 and 21.)

a. Identify the observational units in this study. *Hint:* The correct answer here is more subtle than most students expect.

b. Identify the two variables mentioned in the preceding paragraph. Classify each as categorical (possibly binary) or quantitative.

Variable 1: Type:

Variable 2: Type:

Solution

a. The observational units are the eight-hour shifts.

b. One variable is whether or not Gilbert worked on the shift. This variable is categorical and binary. The other variable is whether or not a patient died on the shift. This variable is also categorical and binary.

Watch Out

It's tempting to call the patients the observational units, but that is not consistent with the data reported. The data indicate what happened on each *shift,* not what happened to each patient. The variables, therefore, need to refer to something that can be recorded about each shift, namely whether Gilbert worked that shift or not and whether a patient died on that shift or not. Notice that we are asking these variables as questions to be posed to each shift. Another way to spot the observational units is to focus on how many data values are in the study; in this case, there are 257 + 1384, or 1641, shifts, not 1641 patients.

Some common errors in reporting variables include

- Providing a summary, such as "the total number of patient deaths" or "the percentage who died on Gilbert's shifts"
- Giving an ambiguous answer, such as "patient deaths"
- Stating the research question rather than a variable, such as "Did patients die at a higher rate on Gilbert's shifts?"
- Describing a subset of the observational units, such as "the patients who died on Gilbert's shifts"

TOPIC 3

Drawing Conclusions from Studies

There is considerable concern about the issue of young people injuring themselves intentionally. Can you use statistics to better understand the seriousness of this problem, by estimating the proportion of college students who have attempted to injure themselves? On an issue of less societal importance, can you estimate what proportion of people believe that Elvis Presley faked his widely reported death, and would it matter which people you asked? Or consider a different kind of question, which may appear whimsical but may prove important: Do candy lovers live longer than other people? If so, is candy a secret to long life? In this topic, you will begin to study issues related to these questions, focusing on concerns that limit the scope of conclusions you can draw from some statistical studies.

Overview

You have begun to understand that data can be useful for gaining insights into interesting questions. But to what extent can statistics provide answers to these questions? This topic begins your introduction to key concepts that determine the scope of conclusions you can draw from a study. For example, when can you *generalize* the results of a study to a larger group than those used in the study itself? Also, why can't you always conclude that one variable *affects* another when a study shows a relationship between the variables?

As you consider those questions, you will encounter some more fundamental terms, such as population and sample, parameter and statistic, and explanatory and response variables. You will also study the important concepts of bias and confounding, and you will begin to understand why those concepts sometimes limit the scope of conclusions you can draw.

Preliminaries

1. Do you believe that Elvis Presley faked his death on August 16, 1977?

2. Take a guess for the percentage of adult Americans who believe that Elvis Presley faked his death.

3. Guess the percentage of American college students who have ever injured themselves intentionally.

4. Would you say that you consume candy rarely, sometimes, or often?

In-Class Activities

Activity 3-1: Elvis Presley and Alf Landon

3-1, 3-6, 16-5

Elvis Presley is reported to have died in his Graceland mansion on August 16, 1977. On the 12th anniversary of this event, a Dallas record company wanted to learn the opinions of all adult Americans on the issue of whether Elvis was really dead.

But of course they could not ask every adult American this question, so they sponsored a national call-in survey. Listeners of more than 100 radio stations were asked to call a 1-900 number (at a charge of $2.50) to voice an opinion concerning whether Elvis was really dead. It turned out that 56% of the callers thought that Elvis was alive.

This scenario is very common in statistics: wanting to learn about a large group based on data from a smaller group.

The **population** in a study refers to the *entire* group of people or objects (observational units) of interest. A **sample** is a (typically small) *part* of the population from whom or about which data are gathered to learn about the population as a whole. If the sample is selected carefully, so it is **representative** of (has similar characteristics to) the population, you can learn useful information. The number of observational units (people or objects) studied in a sample is the **sample size.**

a. Identify the population and sample in this study.

Population:

Sample:

b. Do you think that 56% accurately reflects the opinions of all Americans on this issue? If not, identify some of the flaws in the sampling method.

In 1936, *Literary Digest* magazine conducted the most extensive (to that date) public opinion poll in history. They mailed out questionnaires to over 10 million people whose names and addresses they had obtained from telephone books and vehicle registration lists. More than 2.4 million people responded, with 57% indicating they would vote for Republican Alf Landon in the upcoming presidential election. (Incumbent Democrat Franklin Roosevelt won the actual election, carrying 63% of the popular vote.)

c. Identify the population of interest and the sample actually used to study that population in this poll.

Population:

Sample:

d. Explain how *Literary Digest*'s prediction could have been so much in error. In particular, comment on why its sampling method made it vulnerable to overestimating support for the Republican candidate.

In both the Elvis study and the *Literary Digest* presidential election poll, the goal was to learn something about a very large population (all American adults and all American registered voters, respectively) by studying a sample. However, both studies used a poor method of selecting the sample from the population. In neither case was the sample representative of the population, so you could not accurately infer anything about the population of interest from the sample results. This is because the sampling methods used were *biased*.

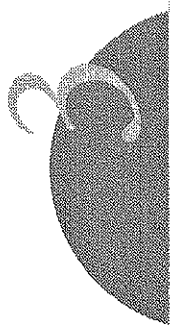
A sampling procedure is said to have **sampling bias** if it tends systematically to overrepresent certain segments of the population and to underrepresent others.

These scenarios also indicate some common problems that produce biased samples. Both are **convenience samples** to some extent because they reach those people most readily accessible (e.g., those listening to the radio station or listed in the phone book). Another problem is **voluntary response,** which refers to samples collected in such a way that members of the population decide for themselves whether or not to participate in the study. For instance, radio stations asked listeners to call in if they wanted to participate. The related problem of **nonresponse** can arise even if an unbiased sampling method is used (e.g., those who are not home when the survey is conducted may have longer working hours than those who participate). Furthermore, the **sampling frame** (the list used to select the subjects) in the *Literary Digest* poll was not representative of the population in 1936 because the wealthier segment of society

was more likely to have vehicles and telephones, which overrepresented those who would vote Republican.

A **parameter** is a number that describes a population, whereas a **statistic** is a number that describes a sample.

Note: To help keep this straight, notice that population and parameter start with the same letter, as do sample and statistic.



e. Identify each of the following as a parameter or a statistic:

- The 56% of callers who believed that Elvis was alive
- The 57% of voters who indicated they would vote for Alf Landon
- The 63% of voters who actually voted for Franklin Roosevelt

f. Consider the students in your class as a sample from the population of all students at your school. Identify each of the following as a parameter or a statistic:

- The proportion of students in your class who use instant-messaging or text-messaging on a daily basis
- The proportion of students at your school who use instant-messaging or text-messaging on a daily basis
- The average number of hours students at your school spent watching television last week
- The average number of hours students in your class slept last night

g. Identify each of the following as a parameter or a statistic. If you need to make an assumption about who or what the population of interest is in a given case, explain that.

- The proportion of voters who voted for President Bush in the 2004 election
- The proportion of voters surveyed by CNN who voted for John Kerry in the 2004 election

- The proportion of voters among your school's faculty members who voted for Ralph Nader in the 2004 election

- The average number of points scored in a Super Bowl game

h. What type of variable leads to a parameter or statistic that is a proportion? What type of variable leads to a parameter or statistic that is an average? *Hint:* Review your answers to the last few questions and look for a pattern.

Proportion: Average:

Watch Out

- A categorical variable leads to a parameter or statistic that is a proportion, whereas a quantitative variable usually leads to a parameter or statistic that is an average.
- Many students confuse a parameter with a population and a statistic with a sample. Remember that parameters and statistics are *numbers*, whereas populations and samples are groups of observational units (people or objects).
- If you believe that a sampling method is biased, suggest a likely direction for that bias. Do not simply say "there is bias" or "the results will be off" or "the results could be higher or lower." Remember, bias is a systematic tendency to err in a particular direction, not just to err.

Activity 3-2: Self-Injuries

An article published in the June 6, 2006, issue of the journal *Pediatrics* describes the results of a survey on the topic of college students injuring themselves intentionally (Whitlock, Eckenrode, and Silverman, 2006). Researchers invited 8300 undergraduate and graduate students at Cornell University and Princeton University to participate in the survey. A total of 2875 students responded, with 17% of them saying that they had purposefully injured themselves.

a. Identify the observational units and variable in this study. Also, classify the variable as categorical (also binary) or quantitative.

Observational units:

Variable: Type:

b. Identify the population and sample.

Population: Sample:

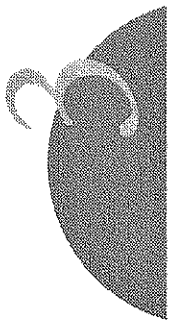
c. What is the sample size in this study?

d. Is 17% a parameter or a statistic? Explain.

e. Do you think it likely this sample is representative of the population of all college students in the world? What about of all college students in the U.S.? Explain.

Activity 3-3: Candy and Longevity

3-3, 21-27



Newspaper headlines proclaimed that chocolate lovers live longer, following the publication of a study titled "Life Is Sweet: Candy Consumption and Longevity" in the *British Medical Journal* (Lee and Paffenbarger, 1998). In 1988, researchers sent a health questionnaire to men who entered Harvard University as undergraduates between 1916 and 1950. They then obtained death certificates for those who died by the end of 1993.

a. With regard to learning about the health habits in the population of all adult Americans, do you consider this a representative sample? Explain.

Researchers found that 3312 of the respondents said that they almost never consumed candy, whereas 4529 did consume candy.

b. Determine the proportion of respondents who consumed candy. Is this a parameter or a statistic?

Of the 3312 nonconsumers of candy, 247 had died by the end of 1993, compared to 267 of the 4529 consumers of candy.

c. Calculate the proportion of deaths in each group.

Nonconsumers: Consumers:

The variable whose effect you want to study is the **explanatory variable.** The variable that you suspect is affected by the other variable is the **response variable.**

d. Identify the observational units in this study. Also identify the explanatory and response variables. Classify each variable as categorical (also binary) or quantitative.

Observational units:

Explanatory: Type:

Response: Type:

The researchers went on to show that this difference in proportions is too large to have reasonably occurred by chance. They also used more sophisticated analyses to estimate that candy consumers in this study would enjoy 0.92 added years of life, compared with nonconsumers.

e. Even if you focus on this group of males who attended Harvard and not some larger population, it is not reasonable to conclude that candy consumption caused the lower death rate and the higher longevity. Provide an alternative explanation for why candy consumers might live longer than nonconsumers.

In part e, you identified a possible second difference between the candy consumers and the nonconsumers, which often happens when the individuals self-select into the explanatory variable groups. Whenever a second variable changes between the explanatory variable groups, you cannot conclude that the explanatory variable causes an effect on the response variable. You have no way of knowing whether the explanatory variable or some other variable led to the different response variable outcomes in the two groups.

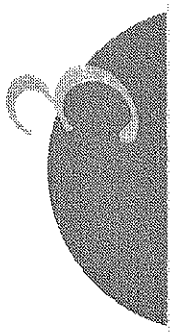
In an **observational study,** researchers passively observe and record information about observational units. An observational study may establish an association or relationship between the explanatory and response variables, but you cannot draw a cause-and-effect conclusion between the explanatory and response variables from an observational study.

The researchers provided data on more health-related variables in this study. Among the 3312 nonconsumers of candy, 1201 had never smoked, compared to 1852 who had never smoked among the 4529 consumers.

f. Among the nonconsumers, calculate the proportion who never smoked. Then do the same for the candy consumers.

Nonconsumers: Consumers:

g. Comment on what the calculations in part f reveal, and how they might help to explain why candy consumers in this study tended to live longer than nonconsumers.



An observational study does not control for the possible effects of variables that are not considered in the study but could affect the response variable. These unrecorded variables are called **lurking variables.** Lurking variables can have effects on the response variable that are confounded with those of the explanatory variable. A **confounding variable** is a lurking variable whose effects on the response variable are indistinguishable from the effects of the explanatory variable.

When confounding variables are present, even if you observe a difference in the response variable between treatment groups, you have no way of knowing which variable (explanatory or confounding), or some combination of the two, is responsible because the treatment groups differ in more ways than simply the explanatory variable. Thus, you cannot draw cause-and-effect conclusions from observational studies because *confounding* might provide an alternative explanation for any observed relationship.

In this study, the person's smoking status is confounded with his candy consumption because those who consumed candy were less likely to smoke than those who did not consume candy. Because smoking status is known to be associated with longevity, you cannot say whether it was the lack of smoking or the candy consumption, or both (or something else entirely) that led to a tendency to live longer in the candy-consuming group.

In order to conduct a study in which you can draw a cause-and-effect conclusion, you have to impose the explanatory variable on subjects (i.e., assign the subjects to "treatment groups") in such a way that the groups are nearly identical except for the explanatory variable (eating candy or not, for example). Then, if the groups are found to differ substantially on the response variable as well, you can attribute that difference to the explanatory variable. You will study strategies for designing such a study and assigning subjects to treatment groups in Topic 5.

Note that the issue of generalizing results from a sample to a larger population is a completely different issue than drawing a cause-and-effect conclusion. In this study, it might be reasonable to generalize the finding about a relationship between candy and longevity to all males who attend Ivy League colleges. But it might not be safe to generalize to all males because those who attend Harvard probably have access to better healthcare and other advantages. It would certainly be risky to generalize this finding to women, as their bodies may respond differently to candy than men's bodies do. In Topic 4, you will learn how to select a sample from a population so that it is likely to be representative.

Activity 3-4: Sporting Examples

2-6, **3-4,** 8-14, 10-11, 22-26

Recall from Activity 2-6 that a statistics professor compared academic performance between two sections of students: one taught using sports examples exclusively and the other taught using a variety of examples. The sections were clearly advertised, and students signed up for whichever section they preferred. The sports section was offered at an earlier hour of the morning than the regular section. The professor found that the students taught using sports examples exclusively tended to perform more poorly than students taught with a variety of examples.

a. Identify the observational units and explanatory and response variables. Also classify the variables' type.

Observational units:

Explanatory variable: Type:

Response variable: Type:

b. Explain why this is an *observational* study.

c. Is it legitimate to conclude that the sports examples *caused* the lower academic performance from students? If so, explain. If not, identify a potential confounding variable and explain why it is confounded with the explanatory variable. *Hint:* Describe how the confounding variable provides an alternative explanation for the observed difference in academic performance between the two groups. Be sure to explain the connection of your proposed confounding variable to both the explanatory variable and the response variable.

Activity 3-5: Childhood Obesity and Sleep

A March 2006 article in the *International Journal of Obesity* described a study involving 422 children aged 5–10 from primary schools in the city of Trois-Rivieres, Quebec, (Chaput, Brunet, and Tremblay, 2006). The researchers found that children who reported sleeping more hours per night were less likely to be obese than children who reported sleeping fewer hours.

a. Identify the explanatory and response variables in this study. Also classify them.

Explanatory: Type:

Response: Type:

b. Is it legitimate to conclude from this study that less sleep caused the higher rate of obesity in Quebec children? If so, explain. If not, identify a confounding variable and explain why its effect on the response is confounded with that of the explanatory variable.

c. Do you think that the study's conclusion (of a relationship between sleep and obesity) applies to children outside of Quebec? Explain.

Solution

a. The explanatory variable is the amount of sleep that a child gets per night. This is a quantitative variable, although it would be categorical if the sleep information were reported only in intervals. The response variable is whether the child is obese, which is a binary categorical variable.

b. This is an observational study because the researchers passively recorded information about the child's sleeping habit. They did not impose a certain amount of sleep on children. Therefore, it is not appropriate to draw a cause-and-effect conclusion that less sleep causes a higher rate of obesity. Children who get less sleep may differ in some other way that could account for the increased rate of obesity. For example, amount of exercise could be a confounding variable. Perhaps children who exercise less have more trouble sleeping, in which case exercise would be confounded with sleep. You have no way of knowing whether the higher rate of obesity is due to less sleep or less exercise, or both, or some other variable that is also related to both sleep and obesity.

c. The population from which these children were selected is apparently all children aged 5–10 in primary schools in the city of Trois-Rivieres. These Quebec children might not be representative of all children in this age group worldwide, so you should be cautious about generalizing that a relationship between sleep and obesity exists for children around the world.

Watch Out

- Confounding is a tricky concept to grasp. Many students find it especially hard to express a confounding variable as a legitimate variable, and many also neglect to explain its connection both to the explanatory variable and to the response variable. For example, in the candy and longevity study, a confounding variable is a man's smoking habit/status. It is not enough to say that this variable is confounding because smokers tend not to live as long; you also need to note that candy consumers are less likely to smoke than nonconsumers.
- After learning that cause-and-effect conclusions cannot be drawn from observational studies, some students overreact and believe that observational studies are useless. On the contrary, an observational study can still be interesting and important by establishing a relationship between two variables, even though you cannot draw a cause-and-effect conclusion. For example, although the sleep and obesity study does not establish a cause-and-effect relationship between sleep and obesity, the connection is still interesting and important for children and parents to know about. Of course, there may be a cause-and-effect relationship between sleep and obesity; the point is simply that an observational study cannot establish this conclusion.

Wrap Up

This topic introduced you to two sets of key terms that will recur throughout this course: population and sample, parameter and statistic. The population of interest in the self-injury study might well be all college students in the U.S., but the sample consisted only of those students at Cornell and Princeton who responded to a survey. The parameter was then the proportion of students who had tried to injure themselves among all college students, whereas the statistic was the proportion of students who admitted to having injured themselves among those Cornell and Princeton students who responded to the survey. Another set of important terms is explanatory and response variables. For example, whether or not a person consumed candy regularly is an explanatory variable, and whether or not the person survived the five-year study period is a response variable whose behavior you think might be *explained* by the candy-consumption habits. Although we have introduced many new terms in this topic, it is important to learn them quickly as they recur throughout the course.

You have also learned about two things that can prevent you from drawing certain conclusions, bias and confounding. Sampling bias occurs when the sampling method tends to produce samples that are not representative of the population, in which case you cannot *generalize* findings in the sample to the larger population. The *Literary Digest*'s sampling method was biased: It overrepresented support for the Republican challenger Alf Landon by sampling only those (wealthier) Americans with a telephone or vehicle, and also because the sample only consisted of those who voluntarily chose to respond. Confounding can always occur with observational studies, precluding you from drawing cause-and-effect conclusions because the groups

you are comparing might differ in more ways than just the explanatory variable. For example, smoking status was a confounding variable in the candy study, because smokers were less likely to eat candy than nonsmokers were. Thus, you cannot tell whether the increased longevity of candy lovers is due to the candy or to not smoking. This is an issue even if you are not trying to generalize the results to a larger population.

Most importantly, you have begun to consider two key questions to ask of statistical studies:

- To what population can you reasonably generalize the results of a study?
- Can you reasonably draw a cause-and-effect connection between the explanatory variable and the response variable?

The answer to the first question depends on how the sample was selected. The answer to the second question depends on whether or not the explanatory variable was assigned to the observational units.

Some useful definitions to remember and habits to develop from this topic include

- The **population** is the entire collection of observational units (people or objects) of interest; the **sample** is the part of the population on which you gather data.
- A **parameter** is a number that describes a population; a **statistic** is a number that describes a sample. Notice that these are numbers, not people or objects as population and sample are.
- An **explanatory variable** is one whose effect you study; a **response variable** is the outcome you record.
- **Sampling bias** is the systematic tendency of a sampling method to overrepresent some parts of the population and underrepresent others.
- An **observational study** is one in which researchers record information passively, without attempting to impose the explanatory variable on the observational units.
- A **confounding variable** is one whose effect on the response variable cannot be distinguished from the effect of the explanatory variable. To properly identify a confounding variable, discuss how it is related to both the explanatory and the response variables.
- You cannot draw cause-and-effect conclusions from observational studies because other factors (confounding variables) might differ between the groups.

In the next two topics, you will learn how to answer two key scope-of-conclusion questions, generalizability and causation. You will also learn how to design studies well in the first place, so that you can make generalizations and/or draw causal conclusions. More specifically, you will discover how to select a sample from the population of interest (Topic 4) and how to assign subjects to treatment groups (Topic 5). Both of these design considerations involve the concept of randomness, but always keep in mind that these are two separate issues.

● ● ● Homework Activities

Activity 3-6: Elvis Presley and Alf Landon

3-1, 3-6, 16-5

The question of whether Elvis Presley faked his death in 1977 has also been asked on an Internet site called `misterpoll.com`, where anyone can post a poll question and get responses from whoever sees the site and chooses to respond.

TOPIC 5

Designing Experiments

Do strength shoes (modified athletic shoes with a 4 cm platform attached to the front half of the sole) really help a person jump farther? How would you design a study to investigate this claim? What factors in a memory experiment affect how many letters a person can memorize correctly? Can a nicotine lozenge help smokers who want to quit? One common aspect of these questions is an interest in finding out whether or not one variable has an effect on another variable. The question is not simply whether those who take a nicotine lozenge tend to quit smoking, but whether you can say that their quitting is because of the lozenge and not some other factor. This topic teaches you how to design studies so they produce data that can answer such questions.

Overview

In the previous topic, you studied the idea of random sampling as a fundamental principle by which to gather information about a population. Sometimes, however, your goal is not to describe a population but to investigate whether one variable has an effect on another variable. This topic introduces you to the design of controlled experiments for this purpose, contrasting them with the observational studies that you explored in earlier topics. You will discover principles for designing controlled experiments and learn how to avoid the problem of confounding variables. You will also explore properties of random assignment and see how this process differs from random sampling, which you studied in Topic 4.

Preliminaries

1. Your instructor will distribute a sequence of letters and give you twenty seconds to memorize as many as you can. Record how many you remember correctly, in the right order.

2. Record these memory scores for your classmates, where the score is the number of letters they remember in the correct order before the first mistake. Also keep track of which version of the letters each student had.

In-Class Activities

Activity 5-1: Testing Strength Shoes

5-1, 5-2, 5-3

The strength shoe is a modified athletic shoe with a 4 cm platform attached to the front half of the sole. Its manufacturer claims that this shoe can increase a person's jumping ability.

a. If your friend who wears strength shoes can jump much farther than another friend who wears ordinary shoes, would you consider that compelling evidence that strength shoes really do increase jumping ability? Explain.

Anecdotal evidence results from situations that come to mind easily and is of little value in scientific research. Much of the practice of statistics involves designing studies and collecting data so people do not have to rely on anecdotal evidence.

Now suppose that you take a random sample of individuals, identify who does and does not wear strength shoes, and then compare their jumping ability.

b. Identify the explanatory and response variables in this study. Also classify each as categorical (also binary) or quantitative.

Explanatory: Type:

Response: Type:

c. Even if the strength shoe group tends to jump much farther than the other group, can you conclude legitimately that strength shoes cause longer jumps? Explain.

The problem with this study, as with all observational studies, is that you do not know whether the two groups might differ in more ways than simply the explanatory variable. For example, subjects who choose to wear the strength shoes could be more athletic to begin with than those who opt to wear the ordinary shoes.

When investigating whether one variable causes an effect on another, researchers create a comparison group and assign subjects to the explanatory variable groups.

An **experiment** is a study in which the experimenter *actively imposes* the **treatment** (explanatory variable group) on the subjects. Ideally, the groups of subjects are identical in all respects other than the explanatory variable, so the researcher can then see the explanatory variable's direct effects on the response variable.

A 1993 study published in the *American Journal of Sports Medicine* investigated the strength shoe claim with a group of 12 intercollegiate track and field participants (Cook et al., 1993). Suppose you also want to investigate this claim, and you recruit 12 of your friends to serve as subjects. You plan to have 6 people wear strength shoes and the other 6 wear ordinary shoes and then measure their jumps.

d. How might you assign subjects to these two groups in an effort to balance out potentially confounding variables?

Random assignment is the preferred method of assigning subjects to treatments (explanatory variable groups) in an experiment: Each subject has an equal chance of being assigned to any of the treatment groups. Such a study is called a **randomized comparative experiment.**

e. Describe in detail how you might implement the process of randomly assigning subjects to treatments.

Activity 5-2: Testing Strength Shoes

5-1, **5-2,** 5-3

In this activity, you will explore properties of random assignment.

Reconsider the experiment described in the previous activity. Suppose that your 12 subjects are listed in the following table. You record their gender and height (in inches) because you suspect that these variables might be related to jumping ability:

Name	Gender	Height	Name	Gender	Height	Name	Gender	Height
Anna	female	61	Kyle	male	71	Patrick	male	70
Audrey	female	67	Mary	female	66	Peter	male	69
Barbie	female	63	Matt	male	73	Russ	male	68
Brad	male	70	Michael	male	71	Shawn	male	67