



1 Define Statistics and Statistical Thinking

What is statistics? Many people say that statistics is numbers. After all, we are bombarded by numbers that supposedly represent how we feel and who we are. For example, we hear on the radio that 50% of first marriages, 67% of second marriages, and 74% of third marriages end in divorce (Forest Institute of Professional Psychology, Springfield, MO).

Another interesting consideration about the “facts” we hear or read is that two different sources can report two different results. For example, a November 10, 2015 poll by CBS News and the *New York Times* indicated that 62% of Americans believed the country was on the wrong track. However, a November 17, 2015 poll by *Bloomberg* indicated that 69% of Americans believed the country was on the wrong track. Is it possible that the percent of Americans who believe the country is on the wrong track could increase by 7% in one week, or is something else going on? Statistics helps to provide the answer.

Certainly, statistics has a lot to do with numbers, but this definition is only partially correct. Statistics is also about where the numbers come from (that is, how they were obtained) and how closely the numbers reflect reality.

Definition **Statistics** is the science of collecting, organizing, summarizing, and analyzing information to draw conclusions or answer questions. In addition, statistics is about providing a measure of confidence in any conclusions.

Let's break this definition into four parts. The first part states that statistics involves the collection of information. The second refers to the organization and summarization of information. The third states that the information is analyzed to draw conclusions or answer specific questions. The fourth part states that results should be reported using some measure that represents how convinced we are that our conclusions reflect reality.

What is the information referred to in the definition? The information is **data**, which the *American Heritage Dictionary* defines as “a fact or proposition used to draw a conclusion or make a decision.” Data can be numerical, as in height, or nonnumerical, as in gender. In either case, data describe characteristics of an individual.

Other Words

Anecdotal means that the information being conveyed is based on casual observation, not scientific research.

Analysis of data can lead to powerful results. Data can be used to offset anecdotal claims, such as the suggestion that cellular telephones cause brain cancer. After carefully collecting, summarizing, and analyzing data regarding this phenomenon, it was determined that there is no link between cell phone usage and brain cancer. See Examples 1 and 2 in Section 1.2.

Because data are powerful, they can be dangerous when misused. The misuse of data usually occurs when data are incorrectly obtained or analyzed. For example, radio or television talk shows regularly ask poll questions for which respondents must call in or use the Internet to supply their vote. Most likely, the individuals who are going to call in are those who have a strong opinion about the topic. This group is not likely to be representative of people in general, so the results of the poll are not meaningful. Whenever we look at data, we should be mindful of where the data come from.

Even when data tell us that a relation exists, we need to investigate. For example, a study showed that breast-fed children have higher IQs than those who were not breast-fed. Does this study mean that a mother who breast-feeds her child will increase the child's IQ? Not necessarily. It may be that some factor other than breast-feeding contributes to the IQ of the children. In this case, it turns out that mothers who breast-feed generally have higher IQs than those who do not. Therefore, it may be genetics that leads to the higher IQ, not breast-feeding.* This illustrates an idea in statistics known as the *lurking variable*. A good statistical study will have a way of dealing with lurking variables.

A key aspect of data is that they vary. Consider the students in your classroom. Is everyone the same height? No. Does everyone have the same color hair? No. So, within groups there is variation. Now consider yourself. Do you eat the same amount of food each day? No. Do you sleep the same number of hours each day? No. So even considering an individual there is variation. Data vary. One goal of statistics is to describe and understand the sources of variation. Variability in data may help to explain the different results obtained by the CBS News/*New York Times* and *Bloomberg* polls described at the beginning of this section.

Because of this variability, the results that we obtain using data can vary. In a mathematics class, if Bob and Jane are asked to solve $3x + 5 = 11$, they will both obtain $x = 2$ as the solution when they use the correct procedures. In a statistics class, if Bob and Jane are asked to estimate the average commute time for workers in Dallas, Texas, they will likely get different answers, even though both use the correct procedure. The different answers occur because they likely surveyed different individuals, and these individuals have different commute times. Bob and Jane would get the same result if they both asked *all* commuters or the same commuters about their commutes, but how likely is this?

So, in mathematics when a problem is solved correctly, the results can be reported with 100% certainty. In statistics, when a problem is solved, the results do not have 100% certainty. In statistics, we might say that we are 95% confident that the average commute time in Dallas, Texas, is between 20 and 23 minutes. Uncertain results may seem disturbing now but will feel more comfortable as we proceed through the course.

Without certainty, how can statistics be useful? Statistics can provide an understanding of the world around us because recognizing where variability in data comes from can help us to control it. Understanding the techniques presented in this text will provide you with powerful tools that will give you the ability to analyze and critique media reports, make investment decisions, or conduct research on major purchases. This will help to make you an informed citizen, consumer of information, and critical and statistical thinker.

will help to make you an informed citizen, consumer of information, and critical and statistical thinker.

Definitions The entire group to be studied is called the **population**. An **individual** is a person or object that is a member of the population being studied. A **sample** is a subset of the population that is being studied. See Figure 1.

Figure 1



Definitions A **statistic** is a numerical summary of a sample. **Descriptive statistics** consist of organizing and summarizing data. Descriptive statistics describe data through numerical summaries, tables, and graphs.

So 78% is a statistic because it is a numerical summary based on a sample. Descriptive statistics make it easier to get an overview of what the data are telling us.

If we extend the results of our sample to the population, we are performing **inferential statistics**.

Definition **Inferential statistics** uses methods that take a result from a sample, extend it to the population, and measure the **reliability** of the result.

The generalization contains uncertainty because a sample cannot tell us everything about a population. Therefore, **inferential statistics** includes a level of confidence in the results. So rather than saying that 78% of all students would return the money, we might say that we are 95% confident that between 74% and 82% of all students would return the money. Notice how this inferential statement includes a **level of confidence** (measure of reliability) in our results. It also includes a **range of values** to account for the variability in our results.

One goal of inferential statistics is to use statistics to estimate **parameters**.

Definition A **parameter** is a numerical summary of a population.



EXAMPLE 1 Parameter versus Statistic

Suppose 48.2% of all students on your campus own a car. This value represents a parameter because it is a numerical summary of a population. Suppose a sample of 100 students is obtained, and from this sample we find that 46% own a car. This value represents a statistic because it is a numerical summary of a sample.

• Now Work Problem 7

3 Distinguish between Qualitative and Quantitative Variables

Once a research objective is stated, a list of the information we want to learn about the individuals must be created. **Variables** are the characteristics of the individuals within the population. For example, recently my son and I planted a tomato plant in our backyard. We collected information about the tomatoes harvested from the plant. The individuals we studied were the tomatoes. The variable that interested us was the weight of a tomato. My son noted that the tomatoes had different weights even though they came from the same plant. He discovered that variables such as weight may vary.

If variables did not vary, they would be constants, and statistical inference would not be necessary. Think about it this way: If each tomato had the same weight, then knowing the weight of one tomato would allow us to determine the weights of all tomatoes. However, the weights of the tomatoes vary. One goal of research is to learn the causes of the variability so that we can learn to grow plants that yield the best tomatoes.

Variables can be classified into two groups: *qualitative* or *quantitative*.

Definitions

Qualitative, or categorical, variables allow for classification of individuals based on some attribute or characteristic.

Quantitative variables provide numerical measures of individuals. The values of a quantitative variable can be added or subtracted and provide meaningful results.

Many examples in this text will include a suggested **approach**, or a way to look at and organize a problem so that it can be solved. The approach will be a suggested method of *attack* toward solving the problem. This does not mean that the approach given is the only way to solve the problem, because many problems have more than one approach leading to a correct solution.

Variables can be classified into two groups: *qualitative* or *quantitative*.

Definitions **Qualitative, or categorical, variables** allow for classification of individuals based on some attribute or characteristic.

Quantitative variables provide numerical measures of individuals. The values of a quantitative variable can be added or subtracted and provide meaningful results.

Many examples in this text will include a suggested **approach**, or a way to look at and organize a problem so that it can be solved. The approach will be a suggested method of *attack* toward solving the problem. This does not mean that the approach given is the only way to solve the problem, because many problems have more than one approach leading to a correct solution.



EXAMPLE 3 Distinguishing between Qualitative and Quantitative Variables

Problem Determine whether the following variables are qualitative or quantitative.

- (a) Gender
- (b) Temperature
- (c) Number of days during the past week that a college student studied
- (d) Zip code

Approach Quantitative variables are numerical measures such that meaningful arithmetic operations can be performed on the values of the variable. Qualitative variables describe an attribute or characteristic of the individual that allows researchers to categorize the individual.

Solution

- (a) Gender is a qualitative variable because it allows a researcher to categorize the individual as male or female. Notice that arithmetic operations cannot be performed on these attributes.
- (b) Temperature is a quantitative variable because it is numeric, and operations such as addition and subtraction provide meaningful results. For example, 70°F is 10°F warmer than 60°F .
- (c) Number of days during the past week that a college student studied is a quantitative variable because it is numeric, and operations such as addition and subtraction provide meaningful results.
- (d) Zip code is a qualitative variable because it categorizes a location. Notice that, even though zip codes are numeric, adding or subtracting zip codes does not provide meaningful results.

Example 3(d) shows us that a variable may be qualitative while having numeric values. Just because the value of a variable is numeric does not mean that the variable is quantitative.

4 Distinguish between Discrete and Continuous Variables

We can further classify quantitative variables into two types: *discrete* or *continuous*.

Definitions

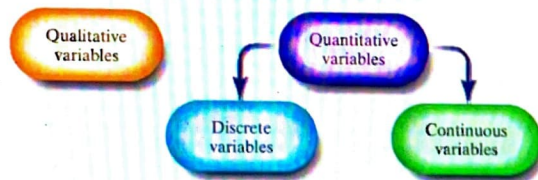
get the value
e variable, it
ou measure to
of a quantitative
continuous.

A **discrete variable** is a quantitative variable that has either a finite number of possible values or a countable number of possible values. The term *countable* means that the values result from counting, such as 0, 1, 2, 3, and so on. A discrete variable cannot take on every possible value between any two possible values.

A **continuous variable** is a quantitative variable that has an infinite number of possible values that are not countable. A continuous variable may take on every possible value between any two values.

continuous variables.

Figure 2



EXAMPLE 4 Distinguishing between Discrete and Continuous Variables

Problem Determine whether the quantitative variables are discrete or continuous.

- The number of heads obtained after flipping a coin five times.
- The number of cars that arrive at a McDonald's drive-thru between 12:00 P.M. and 1:00 P.M.
- The distance a 2015 Toyota Prius can travel in city driving conditions with a full tank of gas.

Approach A variable is discrete if its value results from counting. A variable is continuous if its value is measured.

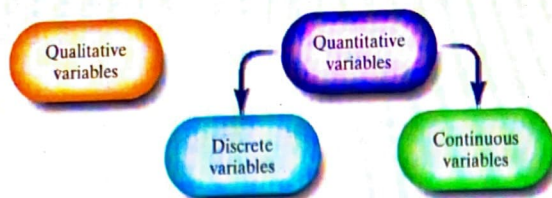
Solution

- The number of heads obtained by flipping a coin five times is a discrete variable



continuous variables.

Figure 2



EXAMPLE 4 Distinguishing between Discrete and Continuous Variables

Problem Determine whether the quantitative variables are discrete or continuous.

- (a) The number of heads obtained after flipping a coin five times.
- (b) The number of cars that arrive at a McDonald's drive-thru between 12:00 P.M. and 1:00 P.M.
- (c) The distance a 2015 Toyota Prius can travel in city driving conditions with a full tank of gas.

Approach A variable is discrete if its value results from counting. A variable is continuous if its value is measured.

Solution

- (a) The number of heads obtained by flipping a coin five times is a discrete variable because we can count the number of heads obtained. The possible values of this discrete variable are 0, 1, 2, 3, 4, 5.
- (b) The number of cars that arrive at a McDonald's drive-thru between 12:00 P.M. and 1:00 P.M. is a discrete variable because we find its value by counting the cars. The possible values of this discrete variable are 0, 1, 2, 3, 4, and so on. Notice that this number has no upper limit.
- (c) The distance traveled is a continuous variable because we measure the distance (miles, feet, inches, and so on).

• Now Work Problem 23



Continuous variables are often rounded. For example, if a certain make of car gets 24 miles per gallon (mpg) of gasoline, its miles per gallon must be greater than or equal to 23.5 and less than 24.5, or $23.5 \leq \text{mpg} < 24.5$.

The type of variable (qualitative, discrete, or continuous) dictates the methods that can be used to analyze the data.

The list of observed values for a variable is **data**. Gender is a variable; the observations male and female are data. **Qualitative data** are observations corresponding to a qualitative variable. **Quantitative data** are observations corresponding to a quantitative variable. **Discrete data** are observations corresponding to a discrete variable. **Continuous data** are observations corresponding to a continuous variable.



Approach A variable is discrete if its value results from counting. A variable is continuous if its value is measured.

Solution

- (a) The number of heads obtained by flipping a coin five times is a discrete variable because we can count the number of heads obtained. The possible values of this discrete variable are 0, 1, 2, 3, 4, 5.
- (b) The number of cars that arrive at a McDonald's drive-thru between 12:00 P.M. and 1:00 P.M. is a discrete variable because we find its value by counting the cars. The possible values of this discrete variable are 0, 1, 2, 3, 4, and so on. Notice that this number has no upper limit.
- (c) The distance traveled is a continuous variable because we measure the distance (miles, feet, inches, and so on).

• Now Work Problem 23

Continuous variables are often rounded. For example, if a certain make of car gets 24 miles per gallon (mpg) of gasoline, its miles per gallon must be greater than or equal to 23.5 and less than 24.5, or $23.5 \leq \text{mpg} < 24.5$.

The type of variable (qualitative, discrete, or continuous) dictates the methods that can be used to analyze the data.

The list of observed values for a variable is **data**. Gender is a variable; the observations male and female are data. **Qualitative data** are observations corresponding to a qualitative variable. **Quantitative data** are observations corresponding to a quantitative variable. **Discrete data** are observations corresponding to a discrete variable. **Continuous data** are observations corresponding to a continuous variable.



EXAMPLE 5 Distinguishing between Variables and Data

Problem Table 1 presents a group of selected countries and information regarding these countries as of July, 2015. Identify the individuals, variables, and data in Table 1.

Table 1

Country	Government Type	Life Expectancy	Population
---------	-----------------	-----------------	------------

Table 1

Country	Government Type	Life Expectancy (years)	Population (in millions)
Australia	Federal parliamentary democracy	82.15	22.8
Canada	Constitutional monarchy	81.76	35.1
France	Republic	81.75	66.6
Morocco	Constitutional monarchy	76.71	33.3
Poland	Republic	77.40	38.6
Sri Lanka	Republic	76.56	22.1
United States	Federal republic	79.68	321.4

Source: CIA World Factbook

Approach An individual is an object or person for whom we wish to obtain data. The variables are the characteristics of the individuals, and the data are the specific values of the variables.

Solution The individuals in the study are the countries: Australia, Canada, and so on. The variables measured for each country are *government type*, *life expectancy*, and *population*. The variable *government type* is qualitative because it categorizes the individual. The variables *life expectancy* and *population* are quantitative.

The quantitative variable *life expectancy* is continuous because it is measured. The quantitative variable *population* is discrete because we count people. The observations are the data. For example, the data corresponding to the variable *life expectancy* are 82.15, 81.76, 81.75, 76.71, 77.40, 76.56, and 79.68. The following data correspond to the individual Poland: a republic government with residents whose life expectancy is 77.40 years and population is 38.6 million people. Republic is an instance of qualitative data that results from observing the value of the qualitative variable *government type*. The life expectancy of 77.40 years is an instance of quantitative data that results from observing the value of the quantitative variable *life expectancy*.

5 Determine the Level of Measurement of a Variable

Rather than classify a variable as qualitative or quantitative, we can assign a level of measurement to the variable.

nominal level of measurement, however the naming scheme allows for the values of the variable to be arranged in a ranked or specific order.

A variable is at the **interval level of measurement** if it has the properties of the nominal level of measurement, however the naming scheme allows for the values of the variable to be arranged in a ranked or specific order. The observations are the data. For example, the data corresponding to the variable *life expectancy* are 2.15, 81.76, 81.75, 76.71, 77.40, 76.56, and 79.68. The following data correspond to the individual Poland: a republic government with residents whose life expectancy is 77.40 years and population is 38.6 million people. Republic is an instance of qualitative data that results from observing the value of the qualitative variable *government type*. The life expectancy of 77.40 years is an instance of quantitative data that results from observing the value of the quantitative variable *life expectancy*.

Determine the Level of Measurement of a Variable

Rather than classify a variable as qualitative or quantitative, we can assign a level of measurement to the variable.

A variable is at the **nominal level of measurement** if the values of the variable name, label, or categorize. In addition, the naming scheme does not allow for the values of the variable to be arranged in a ranked or specific order.

A variable is at the **ordinal level of measurement** if it has the properties of the nominal level of measurement, however the naming scheme allows for the values of the variable to be arranged in a ranked or specific order.

A variable is at the **interval level of measurement** if it has the properties of the ordinal level of measurement and the differences in the values of the variable have meaning. A value of zero does not mean the absence of the quantity. Arithmetic operations such as addition and subtraction can be performed on the values of the variable.

A variable is at the **ratio level of measurement** if it has the properties of the interval level of measurement and the ratios of the values of the variable have meaning. A value of zero means the absence of the quantity. Arithmetic operations such as multiplication and division can be performed on the values of the variable.



EXAMPLE 6 Determining the Level of Measurement of a Variable

Problem For each of the following variables, determine the level of measurement.

- (a) Gender
- (b) Temperature
- (c) Number of days during the past week that a college student studied
- (d) Letter grade earned in your statistics class

Approach For each variable, we ask the following: Does the variable simply categorize each individual? If so, the variable is nominal. Does the variable categorize *and* allow ranking of each value of the variable? If so, the variable is ordinal. Do differences in values of the variable have meaning, but a value of zero does not mean the absence of the quantity? If so, the variable is interval. Do ratios of values of the variable have meaning *and* there is a natural zero starting point? If so, the variable is ratio.

Solution

- (a) Gender is a variable measured at the nominal level because it only allows for categorization of male or female. Plus, it is not possible to rank gender classifications.
- (b) Temperature is a variable measured at the interval level because differences in the value of the variable make sense. For example, 70°F is 10°F warmer than 60°F . Notice that the ratio of temperatures does not represent a meaningful result. For example, 60°F is not twice as warm as 30°F . In addition, 0°F does not represent the absence of heat.
- (c) Number of days during the past week that a college student studied is measured at the ratio level, because the ratio of two values makes sense and a value of zero has meaning. For example, a student who studies four days studies twice as many days as a student who studies two days.
- (d) Letter grade is a variable measured at the ordinal level because the values of the variable can be ranked, but differences in values have no meaning. For example, an A is better than a B, but A - B has no meaning.

• Now Work Problem 31

When classifying variables according to their level of measurement, it is extremely important that we recognize what the variable is intended to measure. For example, suppose we want to know whether cars with 4-cylinder engines get better gas mileage than cars with 6-cylinder engines. Here, engine size represents a category of data and so the variable is nominal. On the other hand, if we want to know the average number of cylinders in cars in the United States, the variable is classified as ratio (an 8-cylinder engine has twice as many cylinders as a 4-cylinder engine).



1.1 Assess Your Understanding

and an Experiment

Once our research question is identified,

and an Experiment

Once our research question is developed, we must develop methods for obtaining the data that can be used to answer the questions posed in our research objective. There are two methods for collecting data, *observational studies* and *designed experiments*. To see the difference between these two methods, read the following two studies.



EXAMPLE 1 Cellular Phones and Brain Tumors

Researchers wanted to determine whether there is an association between mobile phone use and brain tumors. To do so, 791,710 middle-aged women in the United Kingdom were followed over a period of 7 years. During this time, there were 1261 incidences of brain tumors. The researchers compared the women who never used a mobile phone to those who used mobile phones and found no significant difference in the incidence rate of brain tumors between the two groups.

Source: Benson, V. S. et al. "Mobile Phone Use and Risk of Brain Neoplasms and Other Cancers: Prospective Study." *International Journal of Epidemiology* 2013 Jun; 42(3): 792–802



EXAMPLE 2 Cellular Phones and Brain Tumors

Researchers examined "whether chronic exposure to radio frequency (RF) radiation at two common cell phone signals—835.62 megahertz, a frequency used by analogue cell phones, and 847.74 megahertz, a frequency used by digital cell phones—caused brain tumors in rats." To do so, the researchers randomly divided 480 rats into three groups. The rats in group 1 were exposed to the analogue cell phone frequency; the rats in group 2 were exposed to the digital frequency; the rats in group 3 received no radiation. The exposure was done for 4 hours a day, 5 days a week for 2 years. The rats in all three groups were treated the same, except for the RF exposure.

After 505 days of exposure, the researchers "... found no statistically significant increases in any tumor type, including brain, liver, lung or kidney, compared to the control group."

Source: M. La Regina, E. Moros, W. Pickard, W. Straube, J. L. Roti Roti, "The Effect of Chronic Exposure to 835.62 MHz FMCW or 847.74 MHz CDMA on the Incidence of Spontaneous Tumors in Rats," *Bioelectromagnetic Society Conference*, June 25, 2002

In both studies, the goal was to determine if radio frequencies from cell phones increase the risk of contracting brain tumors. Whether or not brain cancer was contracted is the *response variable*. The level of cell phone usage is the *explanatory variable*. In research, we wish to determine how varying the amount of an *explanatory variable* affects the value of a *response variable*.

What are the differences between the studies in Examples 1 and 2? Obviously, in Example 1 the study was conducted on humans, while the study in Example 2 was

Researchers examined "whether chronic exposure to radio frequency (RF) radiation at two common cell phone signals—835.62 megahertz, a frequency used by analogue

at two common cell phone signals—835.62 megahertz, a frequency used by analogue cell phones, and 847.74 megahertz, a frequency used by digital cell phones—caused brain tumors in rats.” To do so, the researchers randomly divided 480 rats into three groups. The rats in group 1 were exposed to the analogue cell phone frequency; the rats in group 2 were exposed to the digital frequency; the rats in group 3 received no radiation. The exposure was done for 4 hours a day, 5 days a week for 2 years. The rats in all three groups were treated the same, except for the RF exposure.

After 505 days of exposure, the researchers “. . . found no statistically significant increases in any tumor type, including brain, liver, lung or kidney, compared to the control group.”

Source: M. La Regina, E. Moros, W. Pickard, W. Straube, J. L. Roti Roti, “The Effect of Chronic Exposure to 835.62 MHz FMCW or 847.74 MHz CDMA on the Incidence of Spontaneous Tumors in Rats,” *Bioelectromagnetic Society Conference*, June 25, 2002

In both studies, the goal was to determine if radio frequencies from cell phones increase the risk of contracting brain tumors. Whether or not brain cancer was contracted is the *response variable*. The level of cell phone usage is the *explanatory variable*. In research, we wish to determine how varying the amount of an **explanatory variable** affects the value of a **response variable**.

What are the differences between the studies in Examples 1 and 2? Obviously, in Example 1 the study was conducted on humans, while the study in Example 2 was conducted on rats. However, there is a bigger difference. In Example 1, no attempt was

SECTION 1.2 Observational Studies versus Designed Experiments

15

made to influence the individuals in the study. The researchers simply followed the women over time to determine their use of cell phones. In other words, no attempt was made to influence the value of the explanatory variable, radio-frequency exposure (cell phone use). Because the researchers simply recorded the behavior of the participants, the study in Example 1 is an *observational study*.

Definition An **observational study** measures the value of the response variable without attempting to influence the value of either the response or explanatory variables. That is, in an observational study, the researcher observes the behavior of the individuals without trying to influence the outcome of the study.

In the study in Example 2, the researchers obtained 480 rats and divided the rats into three groups. Each group was *intentionally* exposed to various levels of radiation.

made to influence the individuals in the study. The researchers simply followed the women over time to determine their use of cell phones. In other words, no attempt was made to influence the value of the explanatory variable, radio-frequency exposure (cell phone use). Because the researchers simply recorded the behavior of the participants, the study in Example 1 is an *observational study*.

Definition An **observational study** measures the value of the response variable without attempting to influence the value of either the response or explanatory variables. That is, in an **observational study**, the researcher observes the behavior of the individuals without trying to influence the outcome of the study.

In the study in Example 2, the researchers obtained 480 rats and divided the rats into three groups. Each group was *intentionally* exposed to various levels of radiation. The researchers then compared the number of rats that had brain tumors. Clearly, there was an attempt to influence the individuals in this study because the value of the explanatory variable (exposure to radio frequency) was influenced. Because the researchers controlled the value of the explanatory variable, we call the study in Example 2 a *designed experiment*.

Definition If a researcher assigns the individuals in a study to a certain group, intentionally changes the value of an explanatory variable, and then records the value of the response variable for each group, the study is a **designed experiment**.

• Now Work Problem 9

Which Is Better? A Designed Experiment or an Observational Study?

To answer this question, let's consider another study.



EXAMPLE 3 Do Flu Shots Benefit Seniors?

Researchers wanted to determine the long-term benefits of the influenza vaccine on seniors aged 65 years and older by looking at records of over 36,000 seniors for 10 years. The seniors were divided into two groups. Group 1 were seniors who chose to get a flu vaccination shot, and group 2 were seniors who chose not to get a flu vaccination shot. After observing the seniors for 10 years, it was determined that seniors who get flu shots are 27% less likely to be hospitalized for pneumonia or influenza and 48% less likely to die from pneumonia or influenza.

Source: Kristin L. Nichol, MD, MPH, MBA, James D. Nordin, MD, MPH, David B. Nelson, PhD, John P. Mullooly, PhD, Eelko Hak, PhD, "Effectiveness of Influenza Vaccine in the Community-Dwelling Elderly," *New England Journal of Medicine* 357:1373–1381, 2007

Wow! The results of this study sound great! All seniors should go out and get

Definition A **lurking variable** is an explanatory variable that was not considered in a study, but that affects the value of the response variable in the study. In addition, lurking variables are typically related to explanatory variables considered in the study.

Definition A **lurking variable** is an explanatory variable that was not considered in a study, but that affects the value of the response variable in the study. In addition, lurking variables are typically related to explanatory variables considered in the study.

In the influenza study, possible lurking variables might be age, health status, or mobility of the senior. How can we manage the effect of lurking variables? One possibility is to look at the individuals in the study to determine if they differ in any significant way. For example, it turns out in the influenza study that the seniors who elected to get a flu shot were actually *less* healthy than those who did not. The researchers also accounted for race and income. The authors identified another potential lurking variable, *functional status*, meaning the ability of the seniors to conduct day-to-day activities on their own. The authors were able to adjust their results for this variable as well.

Even after accounting for all the potential lurking variables in the study, the authors were still careful to conclude that getting an influenza shot is *associated* with a lower risk of being hospitalized or dying from influenza. The authors used the term *associated*, instead of saying the influenza shots *caused* a lower risk of death, because the study was observational.

Observational studies do not allow a researcher to claim causation, only association.

Designed experiments, on the other hand, are used whenever control of certain variables is possible and desirable. This type of research allows the researcher to identify certain cause and effect relationships among the variables in the study.

So why ever conduct an observational study if we can't claim causation? Often, it is unethical to conduct an experiment. Consider the link between smoking and lung cancer. In a designed experiment to determine if smoking causes lung cancer in humans, a researcher would divide a group of volunteers into group 1 who would smoke a pack of cigarettes every day for the next 10 years, and group 2 who would not smoke. In addition, eating habits, sleeping habits, and exercise would be controlled so that the only difference between the two groups was smoking. After 10 years the experiment's researcher would compare the proportion of participants in the study who contract lung cancer in the smoking group to the nonsmoking group. If the two proportions differ significantly, it could be said that smoking causes cancer. This designed experiment is able to control many of the factors that might affect whether one contracts lung cancer that would not be controlled in an observational study, however, it is a very unethical study.

Other reasons exist for conducting observational studies over designed experiments. An article in support of observational studies states, "observational studies have several advantages over designed experiments, including lower cost, greater timeliness, and a broader range of patients." (Source: Kjell Benson, BA, and Arthur J. Hartz, MD, PhD. "A Comparison of Observational Studies and Randomized Controlled Trials." *New*

Observational studies do not allow a researcher to claim causation, only association.

Observational studies do not allow a researcher to claim causation, only association.

Designed experiments, on the other hand, are used whenever control of certain variables is possible and desirable. This type of research allows the researcher to identify certain cause and effect relationships among the variables in the study.

So why ever conduct an observational study if we can't claim causation? Often, it is unethical to conduct an experiment. Consider the link between smoking and lung cancer. In a designed experiment to determine if smoking causes lung cancer in humans, a researcher would divide a group of volunteers into group 1 who would smoke a pack of cigarettes every day for the next 10 years, and group 2 who would not smoke. In addition, eating habits, sleeping habits, and exercise would be controlled so that the only difference between the two groups was smoking. After 10 years the experiment's researcher would compare the proportion of participants in the study who contract lung cancer in the smoking group to the nonsmoking group. If the two proportions differ significantly, it could be said that smoking causes cancer. This designed experiment is able to control many of the factors that might affect whether one contracts lung cancer that would not be controlled in an observational study, however, it is a very unethical study.

Other reasons exist for conducting observational studies over designed experiments. An article in support of observational studies states, "observational studies have several advantages over designed experiments, including lower cost, greater timeliness, and a broader range of patients." (Source: Kjell Benson, BA, and Arthur J. Hartz, MD, PhD. "A Comparison of Observational Studies and Randomized, Controlled Trials," *New England Journal of Medicine* 342:1878-1886, 2000).

One final thought regarding confounding. In designed experiments, it is possible to have two explanatory variables in a study that are related to each other and related to the response variable. For example, suppose Professor Egner wanted to conduct an experiment in which she compared student success using online homework versus traditional textbook homework. To do the study, she taught her morning statistics class using the online homework and her afternoon class using traditional textbook homework. At the end of the semester, she compared the final exam scores for the online section to the textbook section. If the morning section had higher scores, could Professor Egner conclude that online homework is the cause of higher exam scores? Not necessarily. It is possible that the morning class had students who were more motivated. It is impossible to know whether the outcome was due to the online homework or to the time at which the class was taught. In this sense, we say that the time of day the class is taught is a *confounding variable*.

Definition A **confounding variable** is an explanatory variable that was considered in a study whose effect cannot be distinguished from a second explanatory variable in the study.

The big difference between lurking variables and confounding variables is that lurking variables are not considered in the study (for example, we did not consider lifestyle in the pneumonia study) whereas confounding variables are measured in the study (for example, we measured morning versus afternoon classes).

So lurking variables are related to both the explanatory and response variables, and this relation is what creates the apparent association between the explanatory and response variable in the study. For example, lifestyle (healthy or not) is associated with the likelihood of getting an influenza shot as well as the likelihood of contracting pneumonia or influenza.

A confounding variable is a variable in a study that does not necessarily have any association with the other explanatory variable, but does have an effect on the response variable. Perhaps morning students are more motivated, and this is what led to the higher final exam scores, not the homework delivery system.

The bottom line is that both lurking variables and confounding variables can confound the results of a study, so a researcher should be mindful of their potential existence.

We will continue to look at obtaining data through various types of observational studies until Section 1.6, when we will look at designed experiments.



2 Explain the Various Types of Observational Studies

There are three major categories of observational studies: (1) cross-sectional studies, (2) case-control studies, and (3) cohort studies.

Cross-sectional Studies These observational studies collect information about individuals at a specific point in time or over a very short period of time.

For example, a researcher might want to assess the risk associated with smoking by looking at a group of people, determining how many are smokers, and comparing the rate of lung cancer of the smokers to the nonsmokers.

An advantage of cross-sectional studies is that they are cheap and quick to do. However, they have limitations. For our lung cancer study, individuals might develop cancer after the data are collected, so our study will not give the full picture.

Case-control Studies These studies are **retrospective**, meaning that they require individuals to look back in time or require the researcher to look at existing records. In



Explain the various types of Observational Studies

There are three major categories of observational studies: (1) cross-sectional studies, (2) case-control studies, and (3) cohort studies.

Cross-sectional Studies These observational studies collect information about individuals at a specific point in time or over a very short period of time.

For example, a researcher might want to assess the risk associated with smoking by looking at a group of people, determining how many are smokers, and comparing the rate of lung cancer of the smokers to the nonsmokers.

An advantage of cross-sectional studies is that they are cheap and quick to do. However, they have limitations. For our lung cancer study, individuals might develop cancer after the data are collected, so our study will not give the full picture.

Case-control Studies These studies are **retrospective**, meaning that they require individuals to look back in time or require the researcher to look at existing records. In case-control studies, individuals who have a certain characteristic may be matched with those who do not.

For example, we might match individuals who smoke with those who do not. When we say "match" individuals, we mean that we would like the individuals in the study to be as similar (homogeneous) as possible in terms of demographics and other variables that may affect the response variable. Once homogeneous groups are established, we would ask the individuals in each group how much they smoked over the past 25 years. The rate of lung cancer between the two groups would then be compared.

A disadvantage to this type of study is that it requires individuals to recall information from the past. It also requires the individuals to be truthful in their responses. An advantage of case-control studies is that they can be done relatively quickly and inexpensively.

Cohort Studies A cohort study first identifies a group of individuals to participate in the study (the cohort). The cohort is then observed over a long period of time. During this period, characteristics about the individuals are recorded and some individuals will be exposed to certain factors (not intentionally) and others will not. At the end of the study the value of the response variable is recorded for the individuals.

Typically, cohort studies require many individuals to participate over long periods of time. Because the data are collected over time, cohort studies are **prospective**. Another problem with cohort studies is that individuals tend to drop out due to the long time frame. This could lead to misleading results. That said, cohort studies are the most powerful of the observational studies.

One of the largest cohort studies is the Framingham Heart Study. In this study, more than 10,000 individuals have been monitored since 1948. The study continues to this day, with the grandchildren of the original participants taking part in the study. This cohort study is responsible for many of the breakthroughs in understanding heart disease. Its cost is in excess of \$10 million.

Some Concluding Remarks about Observational Studies versus Designed Experiments

Is a designed experiment always superior to an observational study? Not necessarily. Plus, observational studies play a role in the research process. For example, because cross-sectional and case-control observational studies are relatively inexpensive, they allow researchers to explore possible associations prior to undertaking large cohort studies or designing experiments.

Also, it is not always possible to conduct an experiment. For example, we could not conduct an experiment to investigate the perceived link between high tension wires and leukemia (on humans). Do you see why?

Work Problem 19

Existing Sources of Data and Census Data

The saying "*There is no point in reinventing the wheel*" applies to spending energy obtaining data that already exist. If a researcher wishes to conduct a study and an appropriate data set exists, it would be silly to collect the data from scratch. For example, various federal agencies regularly collect data that are available to the public. Some of these agencies include the Centers for Disease Control and Prevention (www.cdc.gov), the Internal Revenue Service (www.irs.gov), and the Department of Justice (www.bjs.gov). A great website that lists virtually all the sources of federal data is fedstats.sites.usa.gov. Another great source of data is the General Social Survey (GSS) administered by the University of Chicago. This survey regularly asks "demographic and attitudinal questions" of individuals around the country. The website is <https://gssdataexplorer.norc.umd.edu>.

Another source of data is a *census*.

Definition A **census** is a list of all individuals in a population along with certain characteristics of each individual.

The United States conducts a census every 10 years to learn the demographic makeup of the United States. Everyone whose usual residence is within the borders of the United States must fill out a questionnaire packet. The cost of obtaining the census in 2010 was approximately \$5.4 billion, and about 635,000 American workers were hired to

The United States conducts a census every 10 years to learn the demo

The United States conducts a census every 10 years to learn the demographic makeup of the United States. Everyone whose usual residence is within the borders of the United States must fill out a questionnaire packet. The cost of obtaining the census in 2010 was approximately \$5.4 billion; about 635,000 temporary workers were hired to assist in collecting the data.

Why is the U.S. Census so important? The results of the census are used to determine the number of representatives in the House of Representatives in each state, congressional districts, distribution of funds for government programs (such as Medicaid), and planning for the construction of schools and roads. The first census of the United States was obtained in 1790 under the direction of Thomas Jefferson. It is a constitutional mandate that a census be conducted every 10 years.

Is the United States successful in obtaining a census? Not entirely. Some individuals go uncounted due to illiteracy, language issues, and homelessness. Given the political stakes that are based on the census, politicians often debate how to count these individuals. Statisticians have offered solutions to the counting problem. If you wish, go to www.census.gov and in the search box type *count homeless*. You will find many articles on the Census Bureau's attempt to count the homeless. The bottom line is that even census data can have flaws.

1.3 Simple Random Sampling



Objective 1 Obtain a simple random sample

Sampling

Besides the observational studies that we looked at in Section 1.2, observational studies can also be conducted by administering a survey. When administering a survey, the researcher must first identify the population that is to be targeted. For example, the Gallup Organization regularly surveys Americans about various pop-culture and political issues. Often, the population of interest is Americans aged 18 years or older. Of course, the Gallup Organization cannot survey *all* adult Americans (there are over 200 million), so instead the group typically surveys a *random sample* of about 1000 adult Americans.

Definition **Random sampling** is the process of using chance to select individuals from a population to be included in the sample.

1.3 Simple Random Sampling



Objective 1 Obtain a simple random sample

Sampling

Besides the observational studies that we looked at in Section 1.2, observational studies can also be conducted by administering a survey. When administering a survey, the researcher must first identify the population that is to be targeted. For example, the Gallup Organization regularly surveys Americans about various pop-culture and political issues. Often, the population of interest is Americans aged 18 years or older. Of course, the Gallup Organization cannot survey *all* adult Americans (there are over 200 million), so instead the group typically surveys a *random sample* of about 1000 adult Americans.

Definition **Random sampling** is the process of using chance to select individuals from a population to be included in the sample.

For the results of a survey to be reliable, the characteristics of the individuals in the sample must be representative of the characteristics of the individuals in the population. The key to obtaining a sample representative of a population is to let *chance* or *randomness* play a role in dictating which individuals are in the sample, rather than convenience. **If convenience is used to obtain a sample, the results of the survey are meaningless.**

Suppose that Gallup wants to know the proportion of adult Americans who consider themselves to be baseball fans. If Gallup obtained a sample by standing outside of Fenway Park (home of the Boston Red Sox professional baseball team), the survey results are not likely to be reliable. Why? Clearly, the individuals in the sample do not accurately reflect the makeup of the entire population. As another example, suppose you wanted to learn the proportion of students on your campus who work. It might be convenient to survey the students in your statistics class, but do these students represent the overall student body? Does the proportion of freshmen, sophomores, juniors, and seniors in your class mirror the proportion of freshmen, sophomores, juniors, and seniors on campus? Does the proportion of males and females in your class resemble the proportion of males and females across campus? Probably not. For this reason, the convenient sample is not representative of the population, which means any results reported from your survey are misleading.

We will discuss four basic sampling techniques: *simple random sampling*, *stratified*