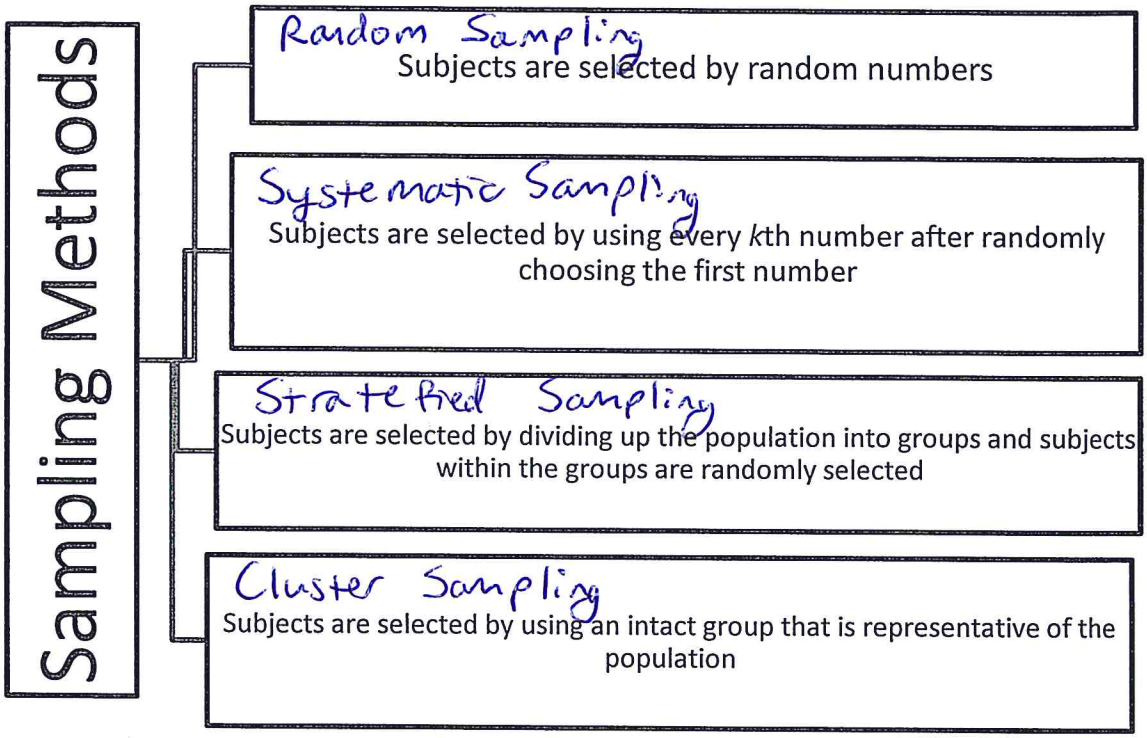

Surveys pros and cons

Telephone Surveys Pros not face to face so could be more honest, inexpensive Cons non-response, undercoverage only people with strong feelings would call in, don't know who is responding	**Personal Interviews** Pros can see body language more likely to respond Cons face to face may not be honest expensive
Mailed Questionnaire Pros very anonymous so more likely to be honest Cons non-response, undercoverage only people w/ strong feelings would call in, don't really know who is responding	**Direct Observation** Pros 1st hand experience likely to get information Cons expensive, possible legal issues, could have cultural bias
Surveying Records Pros factual data includes a variety of info. Cons legal issues to access must make sure source is reliable	**Internet Surveys** Pros-inexpensive, can reach people all over the globe in seconds Cons- do not know who is responding or if info is accurate

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: All adult Americans

Sample: Those that called in opinion

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.

No, only strong opinions would pay a fee and call to voice themselves. One person could have called multiple times. People without a radio could not call.

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: All voting Americans

Sample: 2.4 million that responded

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 1936 only wealthy people owned phones and cars. The wealthy tend to vote Republican (undercoverage)

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

  Statistic

- The 57% of voters who indicated they would vote for Alf Landon

  Statistic

- The 63% of voters who actually voted for Franklin Roosevelt

  Parameter

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

  Statistic

- The proportion of students at your school who use instant-messaging or text-messaging on a daily basis

  Parameter

- The average number of hours students at your school spent watching television last week

  Parameter

- The average number of hours students in your class slept last night

  Statistic

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

  Parameter

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

  Statistic

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

Statistic

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

Parameter

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: Categorical Average: Continuous + quantitative

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: 8300 undergrad + graduate students

Variable: purposefully injured self yes or no Type: Categorical Binary

b. Identify the population and sample.

Population: All College Students Sample: 8300 from Cornell + Princeton

c. What is the sample size in this study? 8300

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

Statistic

17% was based on a survey of our sample

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.

NO, NO, Prinaton + Cornell are high Stress very demanding Schools which could cause Students to cut themselves

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.

NO, it only looks at men from Harvard. These men tend to be wealthy. So there stress about money (a well known issue) is low.

Researchers found that 3312 of the respondents said that they almost never consumed candy, whereas 4529 did consume candy. total 7841

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

$\frac{4529}{7841}$ 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: $\frac{247}{3312} \approx 0.075$

Consumers: $\frac{267}{4529} \approx 0.059$

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: Harvard men

Explanatory: Candy eater or not Type: Categorical binary

Response: Died by 1993 Type: Categorical binary

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.

They may be more care free or have a healthy family history.

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: $\frac{1201}{3312} \approx 0.36$

Consumers: $\frac{1852}{4529} \approx 0.41$

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.

A larger proportion of non candy consumers smoked which could have affected their health

---

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: Statistical Students

Explanatory variable: Using Sports examples Type: Categorical binary

Response variable: Success in class Type: Categorical binary

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

Professor set up the class then observed the success of the students in each class.

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.

The earlier hour could have caused the lower performance

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: Amount of sleep Type: Quantitative continuous

Response: Whether they are obese Type: Categorical binary

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.

Confounding variable: activity level
Children who are active will need more rest then children who sit around all day

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

Yes, but with the same confounding variable.

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.