Module 4: Data Collection

4.1. Experimental and Observational Studies

Observational study

In an **observational study**, the researcher observes values of the response variable for the sampled subjects, but nothing is done to the subjects themselves (i.e., no treatment is imposed)

Experimental study

In an **experimental study**, or an experiment,a researcher assigns subjects to experimental conditions (or **treatments**). Each treatment corresponds to an assigned value of an explanatory variable. The researcher *manipulates* this variable in order to determine if this *causes* a change in the response.

Observational studies vs. experiments

It is not posible to establish cause-and-effect relationships with observational studies because of *extraneous variables*.

An **extraneous variable** is a variable that potentially influences the response variable being studied.

A **confounding** variable is a variable associated with the treatment or explanatory variable.

Well-designed experiments reduce (and can statistically eliminate) confounding variable.

Because treatments are randomly assigned, groups of subjects receiving different treatments will be balanced with respect to these extraneous variables. For example, groups will have similar distributions with respect to age, lifestyle, gender, and so on.

Does smoking cause cancer?

How do we know that smoking causes cancer? Although lung cancer rates are higher for smokers than non-smokers, we must consider and be able to dismiss several important confounding variables.

- Drinking: Drinking and smoking habits are correlated. How do we know that it is smoking, rather than drinking, that causes lung cancer?
- Exercise: Non-smokers are more likely than smokers to exercise regularly. Perhaps exercise can prevent cancer and is negatively correlated with smoking.
- Eating habits: It is likely that smokers do not eat as many vegetables as non-smokers. If eating vegetables prevents cancer, this might explain the association seen between smokers and lung cancer.
- Genetics: Is there an aspect of somebody's genetic makeup that makes them more likely to both smoke and to develop lung cancer?

Now consider the following experiment to determine whether smoking (tobacco) might cause cancer:

One hundred genetically identical mice are divided into two groups of 50 mice each. The first group is exposed to a fixed amount of tobacco smoke 3 times a day for a year; the second group was exposed only to air in the room. All mice were put on the same diet.

There are still extraneous variables that influence cancer rates. However, because this is an *experiment* these variables will not differ (statistically) between the treatment groups.

1. Genetics and diet are **controlled**.
2. All other factors (including exercise) are **randomized**.
3. The **treatment** is exposure to tobacco smoke

Any difference in cancer rates between the two groups is likely due to exposure to tobacco smoke.

link: http://tinyurl.com/nbaldae

Does using Facebook result in lower student grades?

4.2: Observational studies and sampling

When observational studies are appropriate

Not all studies seek to identify causal relationships between two variables.

For example, observational studies are appropriate for

- estimating the average height of all adult females in the United States
- gauging public opinion on a topic, such as capital punishment
- marketing studies to assess how people rate a new product
- estimating the efficiency of a sorting algorithm

The goal of an observational study (and an experiment) is to draw conclusions about a larger population of interest. This means that

1. the sample must be *representative* of the population
2. the subjects being surveyed must answer honestly and accurately (variable reflects question of interest)

Example

Suppose I want to know what percent of Eastern students prefer Google to Bing. One possible sampling design would be to sample the students in this class. Is this class representative of the population of all Eastern students?

No! Our sample should be representative of all Eastern students with respect to gender, age, class status, demographic data, proportion of student-athletes, etc. This is an example of a sampling design chosen out of convenience but one that does not represent the population

The key to selecting a sample that is representative of the population is to sample in such a way that each individual has the same chance of being in the sample. This is known as a **simple random sample**.

A **convenience sample** is a sample chosen out of convenience and is almost never representative of the population.

How survey questions are worded is extremely important

Consider these survey questions?

- What are your feelings toward Obamacare? Do you feel very positive, somewhat positive, neutral, somewhat negative, very negative, or do you not know enough to say?
- What are your feelings toward the Affordable Care Act (ACA)? Do you feel very positive, somewhat positive, neutral, somewhat negative, very negative, or do you not know enough to say?

Despite the fact that Obamacare and the ACA are the *same law*, 46% oppose Obamacare compared to 37% who oppose the Affordable Care Act (Oct. 2010).

Also see: https://www.youtube.com/watch?v=sx2scvIFGjE

The order and context of questions is important

1. On a scale of 1-10, with 1 being the least happy and 10 being the most happy, how happy are you with your life right now?
2. How many dates have you been on in the past two months?

vs.

1. How many dates have you been on in the past two months?
2. On a scale of 1-10, with 1 being the least happy and 10 being the most happy, how happy are you with your life right now?

4.3: Summary

Summary

- One can only conclude that there is a cause-and-effect relationship based on data collected from a well-designed experiment.
- The sample used to conduct the study must be representative of the population of interest.
- The wording of survey questions (and the specific variables measured) must be carefully defined and set a context through which the results should interpreted.

All of these should be taken into account when considering limitations or weaknesses of a research project.

4.4: Examples

Student Drug Testing Not Effective in Reducing Drug Use

A drug use questionnaire was filled out by over 76,000 students in 497 high schools and 225 middle schools nationwide. The study found that drug use was similar in schools that tested for drugs and in schools that did not test for drugs.

	Drug Use		
Drug Tests?	**Yes**	**No**	**n**
Yes	0.37	0.63	5,653
No	0.36	0.64	17,437

1. Is this an observational study or an experiment? An observational study since no variables are being manipulated. An experiment would have randomly selected the schools that perform drug testing.

2. Do you agree with the headline 'Student Drug Testing Not Effective in Reducing Drug Use'? No! It is possible that drug testing was performed primarily in schools with high drug use rates to begin with. Even if this led to a decrease in drug use, then overall you might not see a difference in drug use between schools that do and do not administer drug tests.

Example

Which diet is better? A group of 100 individuals who want to lose weight are randomly assigned to either Diet #1 or Diet #2. Those on Diet #1 will meet weekly with Dr. Jones while those on Diet #2 meet weekly with Dr. Smith. All patients are put on an exercise program which involves walking/jogging 20 minutes a day for 3 times a week.

After 3 months of dieting, subjects on Diet #1 have lost 18 pounds, on average, while those on Diet #2 have lost 6 pounds, on average. The researchers conclude that Diet #1 is more successful than Diet #2.

Is this the correct conclusion?

No! The doctor is completely confounded with the the treatment! It is possible that the personality of the doctor, rather than the diet itself, could explain the difference between the two groups. One solution is to use the same doctor for both groups, ideally in a way that the doctor does not know which patient is receiving which treatment.