Study Design

Learning Outcomes

- Explain the purpose of random sampling and its effect on scope of inference
- Explain the purpose of random assignment and its effect on scope of inference
- Identify whether a study is observational or an experiment
- Identify confounding variables in observational studies and explain why they are confounding

Background

Statistics is the study of how best to collect, analyze, and draw conclusions from data. Statistical inference will allow us to make a statement about a population parameter based on a sample statistic.

- Population: all the units (people, stores, animals, ...) of interest.
- Sample: a subset of the population which gets measured or observed in our study.
- Parameter: a number which describes a characteristic of the population. These values are never completely known except for small populations which can be enumerated. We will use:
  - μ pronounced mew) to represent the population mean.
  - σ (pronounced sigma) to represent the population's standard deviation (spread).
  - π (pronounced pie) to represent a population proportion.
  - ρ (the Greek letter ρ which sounds just like row) for correlation between two quantitative variables in a population.
  - β_1 (read it as beta-one) slope of a true linear relationship between two quantitative variables in a population.
- Statistic: a number which describes a characteristic of the sample and can be computed from the sample. We will use:
  - $\bar{x}$ (read it as ex-bar) to represent the sample mean (or average value).
  - s to represent the sample's standard deviation (spread).
  - $\hat{p}$ (read it as pea-hat) to represent a sample proportion. (We often use a hat to represent a statistic.)
  - r for correlation between two quantitative variables in a sample.
  - b_1 for slope of the "best fitting" line between two quantitative variables in a sample.

There are two parts to study design: how was the sample selected and how was the study conducted. First we will look at sampling.

Sampling from a population

In the text, four unbiased sampling methods are discussed:

- Simple random sampling: each case in the population has an equal chance of being included in the sample
- Stratified sampling: the population is divided into groups of similar observational units called strata and then a second sampling method (usually simple random sampling) is used with each stratum
- Cluster sampling: observational units are group into clusters, then a random sample of clusters is taken and **all** observational units within the selected clusters is measured
- Multistage sampling: similar to a cluster sampling but only a random selection of observational units within the cluster is measured

These above methods all represent an unbiased method of sampling. Another popular sampling method is convenience sampling. A convenience sample is where individuals who are easily accessible are more likely to be included in the sample such as surveying only Stat 216 students about university policies.

In these next questions, identify the target population, the sample, the variable, and the sampling method.

1. To determine if the proportion of out of state undergraduate students at Montana State University has increased in the last 10 years, a statistics instructor surveyed current undergraduate students. She randomly selected 50 students from each college at the University and asked them their state of residency.

Target population:

Sample:

Variable:

Sampling Method:

2. PEW Research surveys US adults about many different topics. Recently a survey was conducted to assess current presidential approval. A random sample of 6395 US adults was taken. Of those surveyed, 42% say they agree with President Trump on many or nearly all of the top issues facing the country today.

Target population:

Sample:

Variable:

Sampling Method:

3. A television station is interested in predicting whether or not voters in its listening area are opposed to legalizing marijuana for adult use. It asks its viewers to phone in and indicate whether they are in favor of this or opposed to this. Of the 2241 viewers who phoned in, forty-five percent were opposed to legalizing marijuana.

Target population:

Sample:

Variable:

Sampling Method:

4. To gauge the interest in a new swimming pool, a local organization divided Gallatin County into neighborhood groups of 100 households. Five neighborhoods were selected and surveys were sent to each household in those areas.

Target population:

Sample:

Variable:

Sampling Method:

5. Of the above scenarios, which would be a biased method of sampling? Explain your answer.

6. The Bozeman school district is interested in surveying parents of students about their opinions on returning to school this fall following the COVID-19 pandemic.

a. What is the target population?

b. Explain how would they select a simple random sample.

c. Explain how would they select a stratified sample.

d. Explain how would they select a cluster sample.

Study Design

Before discussing different study designs, we must first introduce two roles that variables may play in a study. The **explanatory variable** is the variable that may explain the change in the response variable. In an experiment this is the variable in the study that the researcher can manipulate and vary. The **response variable** is the variable that is impacted or changed by the explanatory variable. This is the outcome variable we observe (our data). A **confounding variable** is a variable related both to the explanatory variable and the response variable so that the effects on the response variable cannot be separated from the effects of the explanatory variable.

The two main study designs we will cover are observational studies and experiments. In an **observational study**, groups of cases are compared by observation. The data itself is not manipulated. When we randomly assign treatments to observational units, this study design is called an **experiment**.

Both the sampling method and the study design will help to determine the **scope of inference** for a study. Remember that only in a randomized experiment can we conclude a **causal** (cause and effect) relationship between the explanatory and response variable.

Scope of Inference: If evidence of an association is found in our sample, what can be concluded?

	Study Type	
Selection of cases	Randomized experiment	Observational study
Random sample (and no other sampling bias)	Causal relationship, and can generalize results to population.	Cannot conclude causal relationship, but can generalize results to population.
No random sample (or other sampling bias)	Causal relationship, but cannot generalize results to a population.	Cannot conclude causal relationship, and cannot generalize results to a population.

Inferences to population can be made

Can only generalize to those similar to the sample due to potential sampling bias

Can draw cause-and-effect conclusions

Can only discuss association due to potential confounding variables

For the next exercises, identify the explanatory variable, the response variable, a potential confounding variable, and the study design.

7. In a study on sleep habits, it was found that students who got at least 7 hours of sleep performed better on exams than students who got less than 7 hours of sleep.

Explanatory Variable:

Response Variable:

Confounding Variable:

Explain why this is a potential confounding variable:

Study design:

What is the scope of inference for this study?

8. A study in 2009 was designed to see the effect of drinking beverages with alcohol and or caffeine on neuropsychological status. Twenty seven women college student volunteers were randomly assigned to drink either a caffeinated energy drink, a caffeinated energy drink with alcohol, or a non-alcoholic, non-caffeinated control beverage.

Pre- and post-test assessments were conducted using alternate forms of the Repeatable Battery for the Assessment of Neuropsychological Status (RBANS). RBANS comes in two forms (A and B) which have been shown to give the same scores to people in terms of cognitive status. One form of the test was administered several days before the subjects were given the treatment, and another was given 30 minutes after they had drunk the assigned beverage. The response we will look at is the change in RBANS scores (post score minus pre score).

Explanatory Variable:

Response Variable:

Confounding Variable:

Explain why this is a potential confounding variable:

Study design:

What is the scope of inference for this study?

9. The pharmaceutical company, Moderna Therapeutics is working in conjunction with the National Institute of Health towards a vaccine for COVID-19 and has recently begun Phase 3 clinical trials. US Clinical research sites will enroll 30,000 volunteers without COVID-19 to participate. Participants will be randomly assigned to receive either the candidate vaccine or a saline placebo. They will then be followed to assess vaccine related symptoms and development of COVID-19. The trial is blinded, so the investigators and the participants will not know who is assigned to which group.

Explanatory Variable:

Response Variable:

Confounding Variable:

Explain why this is a potential confounding variable:

Study design:

What is the scope of inference for this study?

10. In which of the studies described above would the potential confounding variable identified not be an issue? Explain your answer.