STAT 305: Chapter 2

Data Collection

Amin Shirazi

ashirazist.github.io/stat305_s2020.github.io

Quick Recap: Populations and Samples

Recap

Making Generalizations

Recap

Making Generalizations

When performing an experiment or gathering data in an observational study, the (main?) goal is to take the information you learn and apply it *outside* of your experiment - i.e., to make *generalizations*. For instance, we may wish to

- describe a relationship between two groups when we do not have the time or ability to gather information from from each member of the two groups,
- use the results of our experiment to predict the outcome of a scenario that has not yet occured,
- explain what part of a process are making the largest contribution to inconsistent results, and so on.

Our ability to make *valid* generalizations heavily depends on the validity of two parts of the study's setup: our **population** and our **sample**.

Recap

Making Generalizations

Populations

Recap

Populations

def: A **population** is the entire group of objects about which one wishes to gather information in a statistical study.

Important: A study's population should be *clearly described* - there should be no question about which objects are in the population and which are not. If a study's population is *not clearly described*, then regardless of how well you execute the mechanics of the study, you will be left with the following conclusion:

> In conclusion, after performing this study we can safely say that our results can be applied to ???

Quick question:

If our goal is to make statements about a population, why don't we just study the population?

Recap

Making Generalizations

Populations

Recap

Populations

- Logistical issues
- Timeless
- Expensive
- Destructive to the objects under study
- ...

Recap

Making Generalizations

Populations

Samples

Recap

Samples

def: A **sample** is the group of objects on which one actually gathers data.

These should be members of the population about which one wishes to gather information in a statistical study.

Recap

Making Generalizations

Populations

Samples

Sampling

Recap

Getting Samples

The purpose of the sample is to be a representation of the population that can actually be studied in depth. Thus, the goal when gathering the sample is to make sure that there is no question that the sample actually does represent the population. A good sampling tecnique gives your study a indisputable connection between the sample and the population.

The gold standard of sampling methods is **Simple Random Sampling**. Using SRS, every possible sample of the same size has the same likelihood of being the sample used in the study.

Recap

Making Generalizations

Populations

Samples

Sampling

Recap

Getting Samples

However, real-world physical constraints may make simple random essentially impossible. In other words, there are "possible samples" from our population that are more likely to used in our study than others. The degree to which our study makes using some samples more likely than others is called **bias**.

In this case, we may have to make (or ask others to make) additional assumptions in order to minimize *the impact of the biased sampling* and still connect the sample we have with the population we are interested in.

Recap

Making Generalizations

Populations

Samples

Sampling

Recap

Getting Samples

Example: In a study of lifetime of lightbulbs, we took 100 consecutive lightbulbs off the factory line and measured their effective lifetime. We found that approximately 95% of lightbulbs survived 2,000 hours of use. We determine that 95% of the lightbulbs produced by our plant will survive 2,000 of use.

- population: all lightbulbs of the plant.
- sample: 100 consecutive lightbulb.
- hidden assumption connecting the sample and the population:
- highly biased?: yes.

Recap

Making Generalizations

Populations

Samples

Sampling

Recap

Getting Samples

Example: In a study of video games effects on emotions, 200 college students were asked how often they played video games and how often they felt angry. The researches found a strong positive correlation between the number of hours spent playing video games and the number of times the student felt anger. They concluded that video games led to increased anger.

- population:
- sample:
- hidden assumption connecting the sample and the population:
- highly biased?:

Recap

Making Generalizations

Populations

Samples

Sampling

Recap

Getting Samples

Example: As part of a study of the health of animals on campus, a field worker set traps and captured 200 squirrels. Once captured, a squirrel was measured and weighed, had it's age estimated, and blood was drawn to test for disease. After being held for a day, the squirrel was chipped and returned to the wild. The researchers reported that squirrels on campus were underweight.

- population:
- sample:
- hidden assumption connecting the sample and the population:
- highly biased?: