Notes and Exercises Chapter 2 – Study Design
Math 217
Saidi

Study Design

Populations and samples

Consider the following three research questions:

1. What is the average mercury content in swordfish in the Atlantic Ocean?
2. Over the last five years, what is the average time to complete a degree for Duke undergrads?
3. Does a new drug reduce the number of deaths in patients with severe heart disease?

Each research question refers to a target population. In the first question, the target population is all swordfish in the Atlantic Ocean, and each fish represents a case. Often times, it is not feasible to collect data for every case in a population. Collecting data for an entire population is called a census. A census is difficult because it is too expensive to collect data for the entire population, but it might also be because it is difficult or impossible to identify the entire population of interest!

Instead, a sample is taken. A sample is the data you have. Ideally, a sample is a small fraction of the population. For instance, 60 swordfish (or some other number) in the population might be selected, and this sample data may be used to provide an estimate of the population average and to answer the research question.

Populations and samples (parameters and statistics)

- **Statistic -** when a number is being calculated on a sample of data
- **Parameter -** when it is being calculated or considered for calculation on the entire population

The terms statistic and parameter are useful for communicating claims and models and will be used extensively in later chapters which delve into making inference on populations.

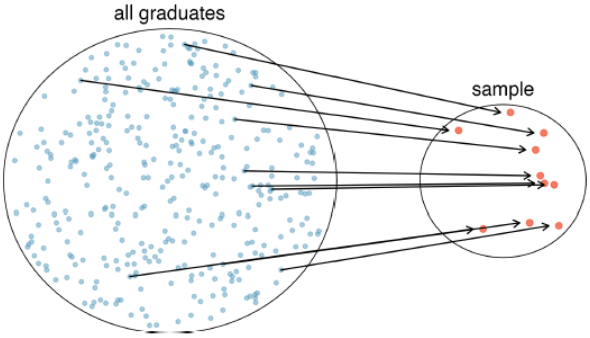
Anecdotal evidence

Consider the following possible responses to the three research questions:

1. A man on the news got mercury poisoning from eating swordfish, so the average mercury concentration in swordfish must be dangerously high.
2. I met two students who took more than 7 years to graduate from Duke, so it must take longer to graduate at Duke than at many other colleges.
3. My friend's dad had a heart attack and died after they gave him a new heart disease drug, so the drug must not work.

Each conclusion is based on data. However, there are two problems. First, the data only represent one or two cases. Second, and more importantly, it is unclear whether these cases are actually representative of the population. Data collected in this haphazard fashion are called anecdotal evidence.

Sampling from a population



Sampling bias - systematically over/underestimates a characteristic of the population; results from a biased sample are not to be trusted to make generalizations. Here are some examples:

- Voluntary
- Convenience
- Non-representative (selective)
- Non-response
- Information – research lacks integrity, fairness, reproducibility (See article Bats, parachutes and bridges – Alba 8/21 – Significance)

Sampling error – the chance error in the discrepancy between the sample and the population.

Examples of Sampling (with and without bias – you decide)

Suppose you want to estimate the average number of hours that students spend studying each week.
Each sample below has bias. Label each with the type of bias, and choose which of the following is the best method of sampling out of the 4?

(a) Go to the **library** and ask all the students there how much they study

(b) Email **all** students asking how much they study. Then randomly select a sample from all answers.

(c) Give a clicker question **in this class** and force every student to respond

(d) Stand **outside the student center** and ask everyone going in how much they study

In general, the best way to prevent bias is by using random sampling techniques.

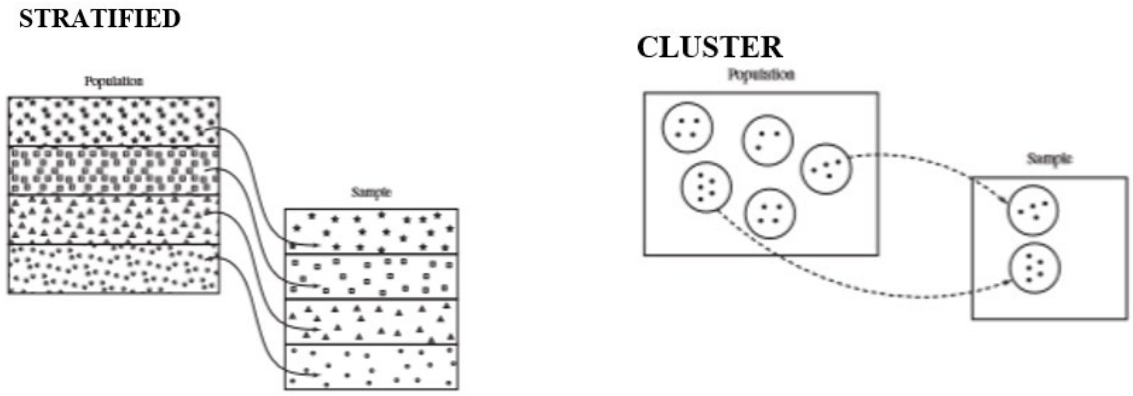
Sampling methods

- **Simple random sampling** - each case in the population has an equal chance of being included in the final sample and knowing that a case is included in a sample does not provide useful information about which other cases are included.

- **Stratified sampling** - population is divided into groups called **strata.** The strata are chosen so that similar cases are grouped together, then a second sampling method, usually simple random sampling, is employed within each stratum. Stratified sampling is especially useful when the cases in each stratum are very similar with respect to the outcome of interest. The downside is that analyzing data from a stratified sample is a more complex task than analyzing data from a simple random sample.

- **Cluster sample** - break up the population into many groups, called clusters. Then sample a fixed number of clusters and include all observations from each of those clusters in the sample. A multistage sample is like a cluster sample, but rather than keeping all observations in each cluster, we would collect a random sample within each selected cluster. A downside of these methods is that more advanced techniques are typically required to analyze the data.

- Studies where the researchers assign treatments to cases are called **experiments.** When this assignment includes **randomization**, e.g., using a coin flip to decide which treatment a patient receives, it is called a **randomized experiment**. Randomized experiments are fundamentally important when trying to show a causal connection between two variables.



2.2.1 Principles of experimental design

1. **Controlling.** Researchers assign treatments to cases, and they do their best to control any other differences in the groups.

2. **Randomization.** Researchers randomize patients into treatment groups to account for variables that cannot be controlled. Randomizing patients into the treatment or control group helps even out such differences.

3. **Replication.** The more cases researchers observe, the more accurately they can estimate the effect of the explanatory variable on the response. In a single study, we replicate by collecting a sufficiently large sample. What is considered sufficiently large varies from experiment to experiment, but at a minimum we want to have multiple subjects (experimental units) per treatment group. Another way of achieving replication is replicating an entire study to verify an earlier finding.

4. **Blocking.** Researchers sometimes know or suspect that variables, other than the treatment, influence the response. Under these circumstances, they may first group individuals based on this variable into blocks and then randomize cases within each block to the treatment groups. This strategy is often referred to as blocking.

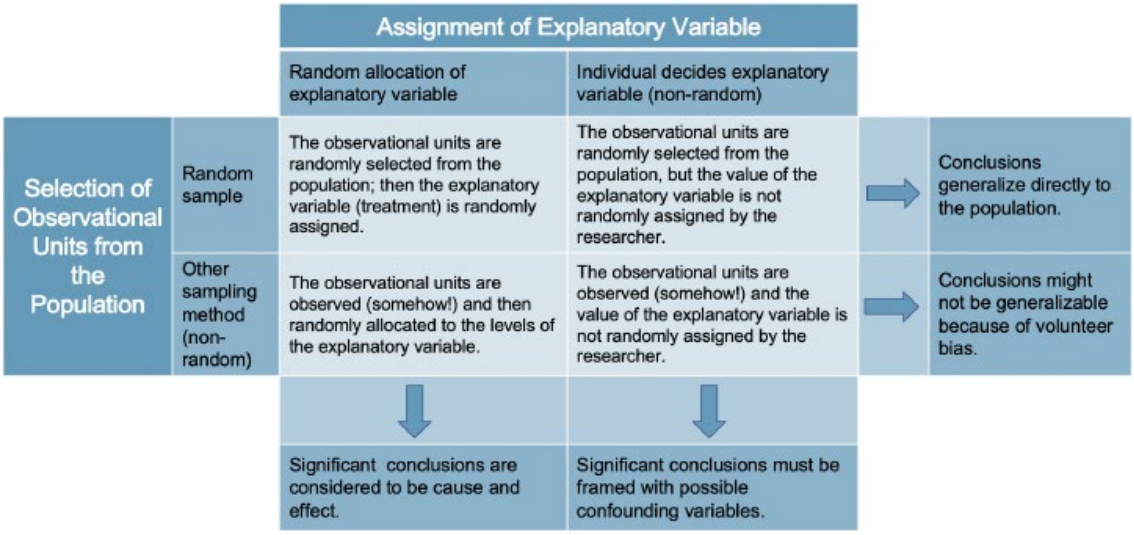
Types of Evidence

- **Observational –**
    - Prospective study
    - Retrospective study

- **Experimental –**
- **Placebo –**
- **Blind and double-blind sampling -**

**** Randomized experiments:** Considered the gold standard in scientific investigation;

**** Case-control studies** – studies match subjects with a disease to similar control subjects (age, race, gender, ….) then look back retrospectively to compare how frequently the exposure to a risk factor is present in each group



Homework Chapter 2

1. Review section 2.4 (the chapter review)
2. Suggested problems from textbook section 2.5 exercises: 3 – 27, odd only