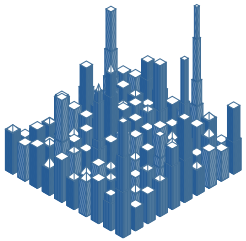
Chapter 6 Notes

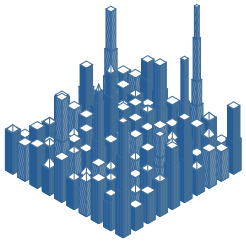
Samples and Observational Studies
STP-231

Arizona State University



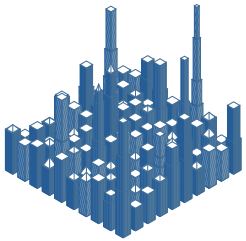
Objectives

- What is a population versus a sample?
- What is an observational study? An experiment?
- Randomness, bias, simple random samples (SRS)
- Other probability samples
- Sample surveys
- Comparitive observational studies



Noise versus Signal

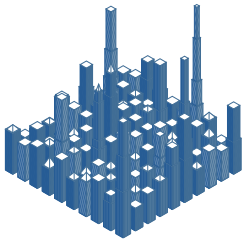
- Variability in data is inevitable, but it is important to the difference between noise and signal
- **Noise**: Is the variability we would expect by chance
- **Signal**: This is the variability due to a certain characteristic. In other words, this is a "real effect", not a statistical anamoly



Example

Below is a random sample of US lunch restaurant goers. Is there a difference between what men and women eat for lunch?

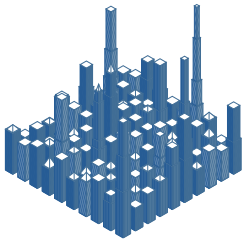
Diet	Men	Women	Total
Vegetarian	184	225	409
Non-Vegetarian	316	275	591
Total	500	500	1000



Observational Studies vs Experimental Studies

- **Observational study**: Record data on individuals without attempting to influence the responses

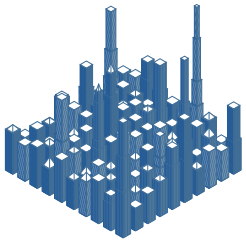
- For example, analyzing data from people who attended a certain university and their outcomes after the fact



Observational Studies vs Experimental Studies

- **Experimental study**: Deliberately impose a treatment on individuals and record their responses. Influential factors can be controlled.

- For example, a study over whether or not a new toothpaste helps peoples tooth health places half of the participants (randomly) into a control group who receive a placebo toothpaste, and the treatment group who receives the real experimental toothpaste



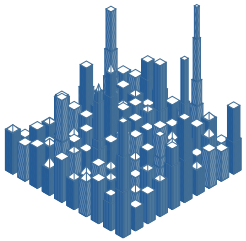
Observation vs Experimental

Observational Study

- We ONLY observe the subject
- Conclusions can be drawn about an association between two variable

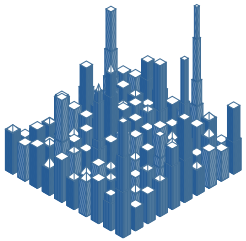
Experimental Design/Study

- We impose conditions on the subjects, i.e. we ask them to do something
- Conclusions can be drawn about cause & effect relationship between two variables



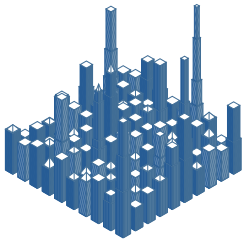
When to use each

- Use an experimental study if we wanna establish cause and effect
- Use observational study when:
  - Sometimes experimental studies are not ethical
  - Experimental studies take too long to complete, are too expensive, hard to properly design experiment
  - Sometimes causality is not that important, an association is enough
  - If we want to look at a historical study



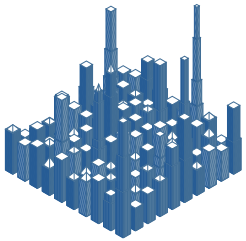
Observational Studies or Experiment?

- A 2013 Gallup study investigated how phrasing affects opinions of Americans regarding physician-assisted suicide. Telephone interviews were conducted with a random sample of 1,535 national adults. Using random assignment, 719 heard the question in form A and 816 the question in form B.
- The different forms worded the question different and 70% in form A said "should be allowed" versus 51% in form B.



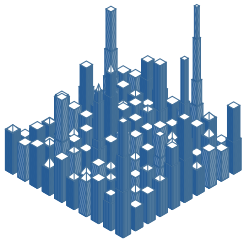
Observational Studies or Experiment?

- A 2013 Gallup study investigated how phrasing affects opinions of Americans regarding physician-assisted suicide. Telephone interviews were conducted with a random sample of 1,535 national adults. Using random assignment, 719 heard the question in form A and 816 the question in form B.
- The different forms worded the question different and 70% in form A said "should be allowed" versus 51% in form B.
- This is a randomized experiment because the groups were randomly assigned



Components in Experimental Study

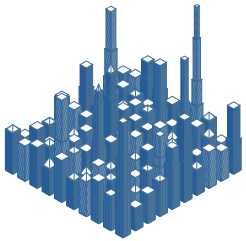
- **Response variable**: Outcome of interest
- **Treatment**: Conditions imposed on the subject. Treatment groups are the group of units in experiment who receive treatment, such as medication
- **Placebo**: The control group, which does not receive treatment. Sometimes placebo effect can lead to psuedo-treatment effect
- **Blind study**: Subjects do not know the treatment they are receiving
- **Double blind study**: Neither subject nor experimenter know which treatment subject receives
- **Panel bias**: A subject may behave differently if they participate in an experimental study



Confounding

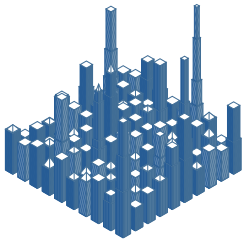
- Two variables are **confounded** when their effects on response variable cannot be distinguished. If the confounder affects both the treatment and outcome.
- Therefore, we cannot determine how one variable affects another if there is a third variable affecting both
- The "lurking" variable can affect the outcome, as long it is not associated with the "treatment". This can be accomplished if controlled for (we'll talk about this later)

Populations & Samples



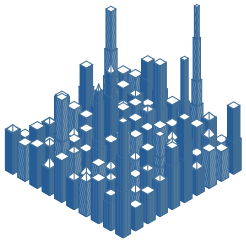
Population versus Sample

- **Population**: The entire group of individuals in which we are interested in but can't usually assess directly
- A **parameter** is a number summarizing a characteristic of the population.
- **Sample**: The part of the population we actually examine and for which we have data
- A **statistic** is a number summarizing a characteristic of a sample
- Parameters are denoted by Greek letters, i.e. μ for mean and lowercase English letters for statistics, such as $\overline{x}$ for sample mean



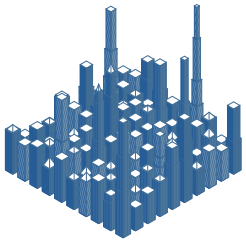
Role of Randomness in Sampling

- How do you select the individuals/units in a sample?
- **Probability sampling**: individuals or units are randomly selected; the sampling process is **unbiased**.
- **bias** is the systematic tendency for a study to favor certain outcomes.
- Think of bias as the accuracy of your study. How close is your estimate of a parameter (your statistic) to the true value of the parameter?



Bad Sampling

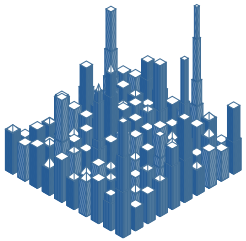
- Ann Landers summarizing 70% of responses of parents wrote in to say having kids was not worth it. But a random sample showed only 9% of parents believe this!
- This is because newsletters readers are not necessarily representative. This particular bunch was potentially disgruntled
- Another example. You are tasked with asking 200 people what they think about the legalization of marijuana and trying to predict how the state will vote. Is sampling the 200 from a college campus a representative sample? From a nursing home? From a mall?



Are these good samples?

Say you have some free time on a Tuesday night, and you want to estimate how much Netflix ASU students watch a week. You want a representative sample of the whole population, i.e. the student body. How good are these samples?

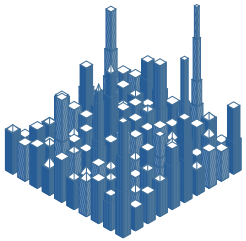
- You sample 100 students at Pete's trivia night on Mill avenue
- You sample 100 students at the library
- You ask 100 students at the gym
- You ask 100 people at the MU



The Simple Random Sample

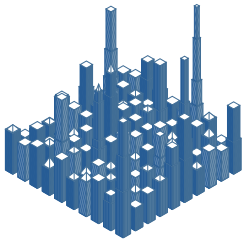
- A **Simple random sample** (**SRS**) is made of randomly selected individuals. Each individual in the population has the same probability of being in the sample. $\binom{N}{n}$ possible samples with n/N probability of being the sample drawn. All the possible samples of size n have the same chance of being the sample drawn
- How do we draw an SRS? Usually we assign random numbers to every unit and draw till we get the desired size. Without a computer, we could draw from a hat, or use a table.
- Example in *R*:

```
set.seed(12296)
names<-randomNames::randomNames(35)
our_sample<-names[sample(35,10)];our_sample
```



Other Probability Samples

- A **stratified random sample**: make sure your sample has known percentages of individuals of certain types (strata)
- America's State of Mind report was based on a probability sample of Medco's de-identified database of members with 24 months of continuous insurance enrollment. Sampling was stratified by age group and sex to match the demographics of the whole customer base.
- A **multistage sample**: select your final sample in stages, by sampling successively within a sample within a sample

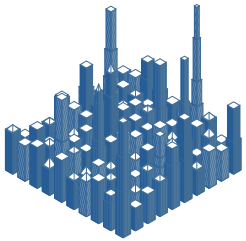


Example of 2-stage Sample

The National Youth Tobacco Survey administered in schools use a sampling procedure to generate a nationally representative sample of students in grades 6-12. Sampling is probabilistic and consists of selecting:

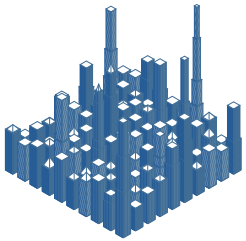
- Counties as primary sampling units (PSU)
- Schools within each selected PSU
- Classes within each selected school

These studies are cheaper to conduct (for example have to sample less counties, meaning less distance between schools), but estimates of population totals and means vary more (bad)!



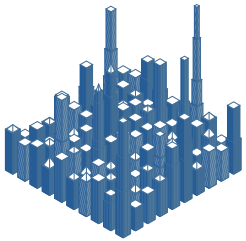
Sample Surveys

- A **sample survey** is an observational study that relies on a random sample drawn from the entire population
- Opinion polls are sample surveys that typically use voter registries or telephone numbers to select their samples
- In epidemiology, sample survey are used to establish the **incidence** (rate of new cases per year) and the **prevalence** (rate of all cases at one point in time) of various medical conditions, diseases, and lifestyles. These are typically stratified or multistage samples.



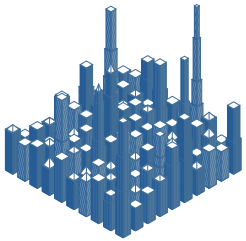
Some Challenges

- **Undercoverage**: Parts of the population are systematically left out
- **Nonresponse**: Some people choose not to answer/participate
- **Wording effects**: Biased or leading questions, and complicated/confusing statements can influence survey results
- **Response bias**: If people lie, forget, or misanswer
- **Endogeneity**: Outcome of survey affects estimand survey is trying to estimate!



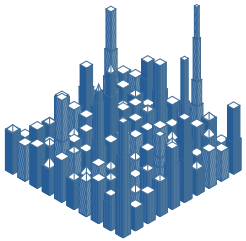
How are we biasing

- Try to think about how your sampling issues may bias your estimate of the population
- For example, if we over sample from a certain population how might these affect our estimate?
- However, trying to predict the direction of bias could be an issue. At some point we are just guessing and our predispositions could affect analysis. Almost impossible to predict confounding!



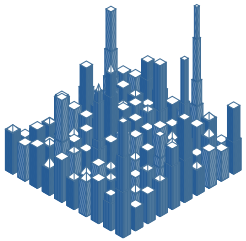
Undercoverage

- Polls and surveys may have a harder time reaching younger people who do not use traditional devices that are used as mediums for surveys, like landlines
- Certains groups of people that may be disenfranchised may be harder to reach or not a group that is reached out to
- Implication of undercoverage is other groups are **over – sampled**. Weighting is one way to correct this



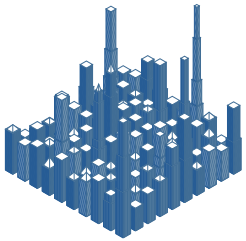
Nonresponse

- The Census Bureau's American Community Survey is about 97.5% via mail with reminders, with mandatory response
- University of Chicago's General Social Survey (GSS): Has about 70% response in person
- Pew Research Center methodology survey has about 10% response
- Private polling firms such as SurveyUSA has about 10% response in 2002. Even lower in 2020 with shift away from landlines. Online polls have low participation rates, but high amount of users (could affect how representative your sample is)



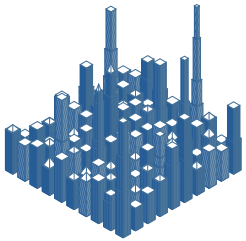
Wording Effects

- A Gallup 2013 study investigated how phrasing effects the opinions of Americans regarding physician-assisted suicide. Recall, we had form A and form B
- Form A: When a person has a disease that cannot be cured, do you think doctors should be allowed by law to end the patient's life by some painless means if the patient and his family request it?
- Form B: When a person has a disease that cannot be cured and is living in severe pain, do you think the doctors should or should not be allowed by law to assist the patients to commit suicide if the patient requests it?
- Even though this is an experiment that was randomly assigned, the wording makes the conclusion less robust



Comparitive Observational Studies

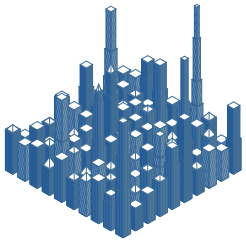
- **Case – control studies** start with 2 random samples of individuals with different outcomes, and look for exposure factors in the subjects' past ("retrospective")
- Individuals with the condition are cases, and those without are controls
- Good for studying rare conditions. Selecting controls is challenging
- **Cohort studies** enlist individuals of common demographic and keep of them over a long period of time ("prospective"). Individuals who later develop a condition are compared to those who don't develop the condition.
- Cohort studies examine the compounded effect of factors over time. Good for studying common conditions.



A Case-Control Study Example

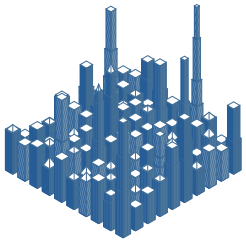
- Alfatoxins are secreted by a fungus found in damaged crops and can cause severe poisoning and death.

- The Kenya Ministry of Health investigated a 2004 outbreak of aflatoxicosis resulting in over 300 cases of liver failure. A sample of 40 case patients and 80 healthy controls were asked how they had stored and prepared their maize.



A Case-Control Study Example

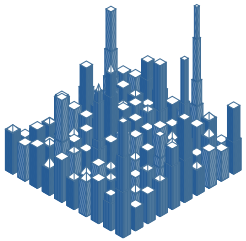
- The case patients were randomly selected from a list of individuals admitted to a hospital during the 2004 outbreak for for unexplained acute jaundice.
  Control individuals were selected to be as similar to the case patients as possible, yet randomly selected.
- Preliminary data suggested that soil, microclimate, and farming practices mights have played a role, but not age or gender
- For each case patient, two individuals from the patients' village with no history of jaundice symptoms were randomly selected.



Example of Cohort Study (1)

- The Nurses' Health Study is one of the largest prospective observational studies designed to examine factors that may affect major chronic diseases in women.

- Since 1976, the study has followed a cohort of over 100,000 registered nurses. Every two years, they receive a follow-up questionnaire about diseases and health-related topics, with 90% response rate each time.



Example of Cohort Study (2)

- 2007 Report on Age-Related Memory Loss: About 20,000 women ages 70+ had completed telephone interviews every two years to assess their memory with a set of cognitive tests. One of the findings: the more women walked during their late 50s and 60s, the better their memory score was at age 70 and older.

- However, because this is an observational study, we cannot conclude a causal effect of walking on protecting against memory loss.