STAT 311: Gathering data: Observational Studies

Y. Samuel Wang

Summer 2016

Logistics

- Discussion of HW 1

Causal relationships

- Monday we focused primarily on surveys, a specific type of observational study where we typically are only interested in describing a population, but not necessarily drawing cause and effect relationships.
- Making statements about cause and effect relationships requires a bit more care

Evaluating an estimator

When we think about how well we are estimating a parameter, there are two main aspects we can think about

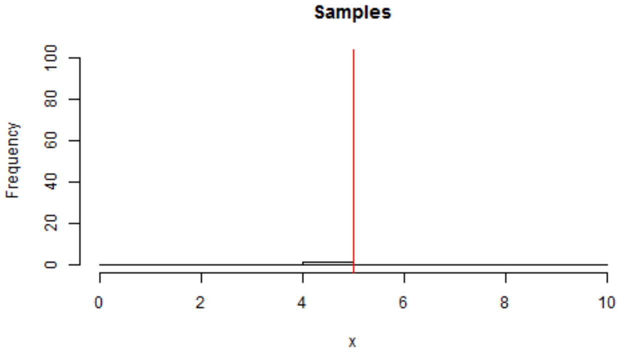
- Variability- How much does the statistic change from sample to sample?
- Bias- Are we systematically over/under estimating the parameter?

Sampling Variability

Each time we take a new sample, the statistic will be different. We can consider the distribution of statistics which we would get if we took many different samples.

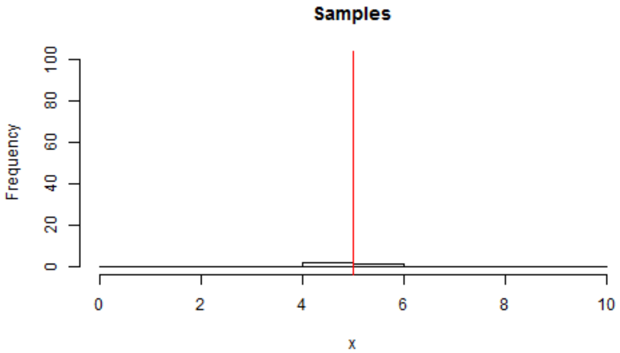
Sampling Variability

Each time we take a new sample, the statistic will be different. We can consider the distribution of statistics which we would get if we took many different samples.



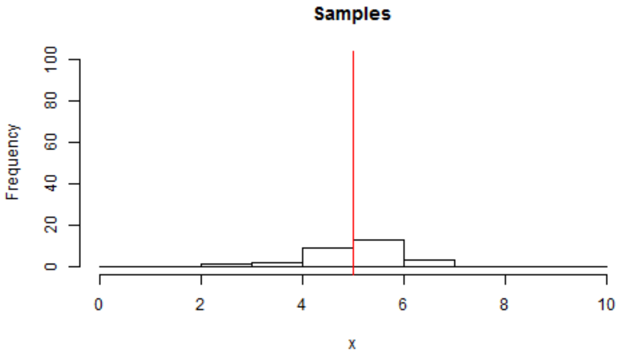
Sampling Variability

Each time we take a new sample, the statistic will be different. We can consider the distribution of statistics which we would get if we took many different samples.



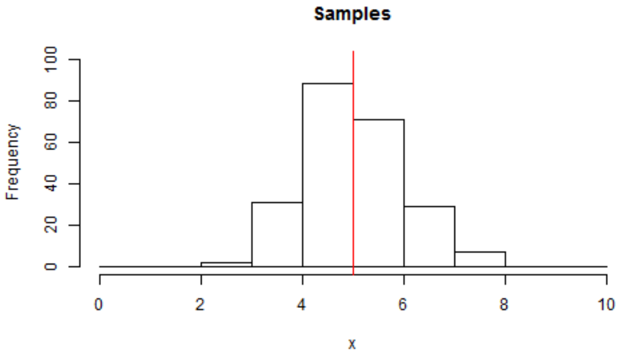
Sampling Variability

Each time we take a new sample, the statistic will be different. We can consider the distribution of statistics which we would get if we took many different samples.

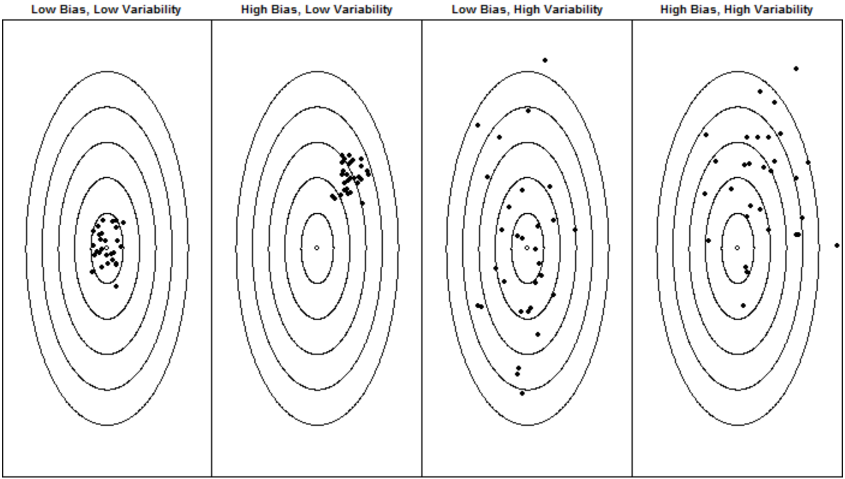


Sampling Variability

Each time we take a new sample, the statistic will be different. We can consider the distribution of statistics which we would get if we took many different samples.



Bias and Variability


Low Bias, Low Variability
High Bias, Low Variability
Low Bias, High Variability
High Bias, High Variability

Observational Studies vs Experiments

There are two main ways to gather data

- In an **experiment**, the researcher actively applies a treatment and then gathers data
- In an **observational study**, the researcher only observes and gathers data, but does not apply a treatment

Observational Studies vs Experiments

There are two main ways to gather data

- In an **experiment**, the researcher actively applies a treatment and then gathers data
- In an **observational study**, the researcher only observes and gathers data, but does not apply a treatment

- Experiments typically allow us to make stronger statements because we can rule out confounding variables more easily
- Experiments are not always possible

Observational Studies vs Experiments

There are two main ways to gather data

- In an **experiment**, the researcher actively applies a treatment and then gathers data
- In an **observational study**, the researcher only observes and gathers data, but does not apply a treatment

- Experiments typically allow us to make stronger statements because we can rule out confounding variables more easily
- Experiments are not always possible

We will mostly focus on observational studies and surveys today, and talk more about experiments on Wednesday

Why do we use observational studies

If carefully designed experiments allow us to make stronger statements, why do we still use observational studies?

- Cost
  - Observational studies are often much less expensive than designed experiments
- Logistics
  - Some treatments are simply impossible to carry out
  - Ex- What is the effect of global warming on wildlife diversity?
- Ethics
  - Cannot subject individuals to harmful treatments
  - Scientists have at times, have crossed ethical lines (ex. Tuskegee Syphilis Study)

Sampling Terminology

- The **Sampling frame** is the set of all observational units which had a chance of being included in the sample.
- Each observational unit in the sampling frame has a **Sampling probability**, which is the probability that they are included in the sample. This probability may be different for each observational unit.

Ways to select a sample

Two main ways of selecting a sample:

- **Convenience**: The sampling probability for each individual is unknown
- **Probability**: The sampling probability for each individual is known

Probability Sample: Simple Random Sample

- Basic idea: Every observational unit in the sampling frame has an equal probability of being selected
- Why do we use this: It is easy to implement
- Example: I take the entire list of Seattle Public School and randomly select (uniformly) 1500 students from the list
- Cautions: Can have large variability in the estimate

Probability Sample: Stratified

- Basic idea: We first split our sampling frame into subgroups based on observable characteristics, then we perform a simple random sample within each subgroup
- Why do we use this: It ensures that we properly represent important subgroups in the population. This can be especially helpful if certain subgroups are rare
- Example: I split the entire list of Seattle Public School students into 3 lists: elementary students, middle school students, and high school students. Then I randomly sample 500 students (uniformly) from each subgroup

Probability Sample: Cluster

- Basic idea: We first split our sampling frame into clusters (or subgroups) based on observable characteristics, then we randomly select a set number of clusters
- Why do we use this: It can be more convenient to sample an entire cluster rather than specific individuals
- Example: I randomly select 10 schools, and then give a survey to everyone in each of those schools
- Caution: Can also have high variability in the sample

Probability Sample: Cluster

- Basic idea: We first split our sampling frame into clusters (or subgroups) based on observable characteristics, then we randomly select a set number of clusters
- Why do we use this: It can be more convenient to sample an entire cluster rather than specific individuals
- Example: I randomly select 10 schools, and then give a survey to everyone in each of those schools
- Caution: Can also have high variability in the sample

Note how this is different from stratified sampling. Here, I select entire subgroups, rather than selecting individuals in a subgroup

Probability Sample: Systematic Sampling

- Basic idea: Select individuals from our list in a given interval
- Why do we use this: Easy to select the individuals
- Example: I go down the Seattle Public School Registrar list and select every 100th student
- Caution: Can still be biased if there is some sort of systemic pattern in the underlying process. For example, if I want to check for defects on a manufacturing line, and I check the first item in every batch. If the manufacturing process starts out okay, but gets worse as each batch goes along, I will not catch that in my data.

Probability Sample: Multi-stage

- Basic idea: combining several different type of sampling techniques into one larger process
- Why do we use this: Can get the benefits of each of the other sampling techniques
- Example: I split all the schools into elementary schools, middle schools and high schools. Then I randomly select 3 schools from each subgroup. (stratified + cluster sampling)

Convenience Sample: Voluntary

- Basic idea: Allow individuals to self-select whether or not they are in the sample
- Why do we use this: Can be easier to get a sample
- Example: Let students decide whether or not to take the survey
- Caution: This often results in a sample which is not representative of the population. Typically the people who take the time to respond feel most passionately about the issues at stake and do not represent the average individual. Think about yelp reviews.

What can go wrong?

- What if sampling frame is not representative of the population?
- What if people who have been selected to be in the sample do not respond?
- What if the survey itself causes bias?

Non-representative Sampling frame

If my initial sampling frame is not representative of my population, simple random samples will not give me unbiased estimates. In particular, **undercoverage** may occur if certain individuals are systematically not included.

- Random digit dialing
- Likely voters
- Internet access
- 1936 Presidential election

Non-response

If my sampling frame represents the population well, but certain individuals are less likely to respond this can still create a bias in my estimates

We have three different categories to describe how the data might be missing

Missing Data

- Missing completely at random: non-response is completely unrelated to the parameter of interest. Your estimate is still unbiased

  Ex- Some questions are only asked to randomly selected individuals

- Missing at random: the relationship between non-response and the parameter of interest can be full accounted for by observable variables. You can adjust your estimate based on observed variables to make it unbiased.

  Ex- We measure blood pressure of all individuals, but only measure heart rate of individuals with high blood pressure

- Missing not at random: the non-response is directly dependent on your parameter of interest. Not a lot you can do to fix your estimate.

  Ex- When asking about HIV status, individuals who are HIV+ may be less likely to respond

Response bias

The surveys themselves can often induce biases in the response

- Deliberately leading wording in questions
  Example- Do you think Vladamir Putin should be prosecuted for the unjust invasion of Ukraine?
- Embarrassing responses
  Example- What is your GPA?
- Confusingly worded or imprecise questions
  Example- Shouldn't convicted felons be allowed to work in schools after their release?
- Ordering of questions or possible answers
  Example- Individuals are more likely to select the first answer in a multiple choice question