Lecture 6: Sampling Distributions

Intro Stats with Nathan Favero

February 23, 2023

AMERICAN UNIVERSITY
WASHINGTON, DC

Study guide: key points from this lecture

- We create an *estimate* (from a sample) when we want to know about a *population parameter* but can't observe it directly
- The *standard error* tells us how precise our estimate is (smaller standard errors indicate greater precision)
- Larger samples produce smaller standard errors
- *Degrees of freedom* describe the sample size
- A *sampling distribution* describes all the possible estimates we could have gotten, depending on what sample we drew
- Confidence intervals tell us a likely range for a parameter in the *population*
- When we have a small sample, it's important to use the *t distribution* (rather than the normal distribution) to create confidence intervals

Brief note on notation

- N: number of observations in population
- n: number of observations in sample
- μ: population mean $\mu = \frac{\sum x_i}{N}$
- $\bar{x}$: sample mean $\bar{x} = \frac{\sum x_i}{n}$
- σ: population standard deviation $\sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{N}}$
- s: sample standard deviation $s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$

Statistical inference

- Parameter (or estimand): a value that describes a population of interest; the (unknown) value we are trying to estimate
- Estimate: a value (or set of values) derived from a sample that is used as an approximation of a population parameter
- Estimator: a process (statistical formula) for calculating an estimate
- The sample mean is often used as an *estimate* of the population mean (where the population mean is an unobserved *parameter*).

Statistical inference

- Estimates based on random sampling (or a random treatment) are random variables, so they have probability distributions
- See our example, where we randomly select clients to tell us how satisfied they are, and then use the sample mean as our estimate of the population mean
- If we had selected different clients, we probably would have gotten a different estimate of the population mean
- As a thought experiment, imagine if we repeated this process of randomly selecting a sample over and over again, creating a new mean estimate each time
- If we were to make a histogram of all the mean estimates we obtained, we would see what the probability distribution for our mean estimate looks like

Standard errors

- Q: When we use a sample to estimate a population parameter, can we say anything about how accurate we believe that estimate is?
- Q: All else equal, a bigger sample (more observations) should give us a more precise estimate than a smaller sample. How can we describe this greater level of precision?
- Q: What if I want to report an estimate as a range of likely values rather than a single point estimate (e.g., the population mean probably lies between 4.2 and 4.8)?

- A: *Standard errors* help us to address all of these concerns

Standard errors

- Estimates based on random sampling (or other random processes) have probability distributions (since the value of the estimate depends on who gets selected into the sample)
- We call the probability distribution for an estimate the *sampling distribution*
- Standard error: the standard deviation of an estimate's sampling distribution (probability distribution)
  - You can think of the standard error as the typical distance between an estimate and the true parameter value (if you were to repeat the process of sampling and producing an estimate over and over again)
  - The standard error is a measure of how precise our estimation process is (taking into account sample size, how much variance is in the population, etc.)

Standard errors

- Again, part of what inferential statistics allows us to do is draw conclusions about how estimators behave based on assumptions we've made about the data generation process
  - Check the appendices for an example of the kinds of assumptions we often make
  - There is also a discussion of what makes a good estimator, and how we can choose an estimator if there are multiple options
- Once we've made a set of assumptions about how the data is generated, we can use math or simulations to determine the standard error for an estimation procedure

Standard errors

- If we use a sample mean to estimate the population mean, the standard error of the estimate will be the following (assuming *x* is normally distributed, a simple random sample is used to create the estimate, and the population size is infinite):

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

  - Notice the notation: σ is the standard deviation (of *x*); $\sigma_{\bar{x}}$ is the standard deviation of $\bar{x}$ (our mean estimate), a.k.a. the standard error of the sample mean
  - When the standard deviation of *x* is larger, the sample mean (of *x*) is a less precise estimate of the population mean ($\sigma_{\bar{x}}$ is bigger)
  - When the sample size (*n*) is larger, the sample mean (of *x*) offers a more precise estimate of the population mean ($\sigma_{\bar{x}}$ is smaller)

Standard errors

- The prior formula helps us to understand what the probability distribution of the sample mean (of a simple random sample) looks like under different circumstances
- But if we're using a sample to study a population, we won't know what the population's standard deviation is, so we'll have to use an estimate instead
- The typical estimator for the standard error of a sample mean is:

$$s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

  - We use an estimate (s) of the standard deviation of x instead of the true population parameter (σ) since we don't know it

Confidence intervals

- So how do we interpret a standard error estimate?
- One of the easiest ways to understand a standard error is by using it to construct a confidence interval

Confidence intervals

- Example: estimating the population mean
  - We're about to learn about the t-distribution, which will help us create better confidence intervals when we have a small sample
  - For large samples ($n > 30$), the sampling distribution (probability distribution for our estimate) will follow a normal curve (because of the *central limit theorem*, as explained in the textbook)
  - For any normal curve, 95% of the data falls within 2 standard deviations of the mean, and about 70% of the data falls within 1 standard deviation of the mean (with software, we can easily find other confidence levels too)

$$\text{95\% C.I.: } \bar{x} \pm (2 \times s_{\bar{x}})$$

$$\text{70\% C.I.: } \bar{x} \pm (1 \times s_{\bar{x}})$$

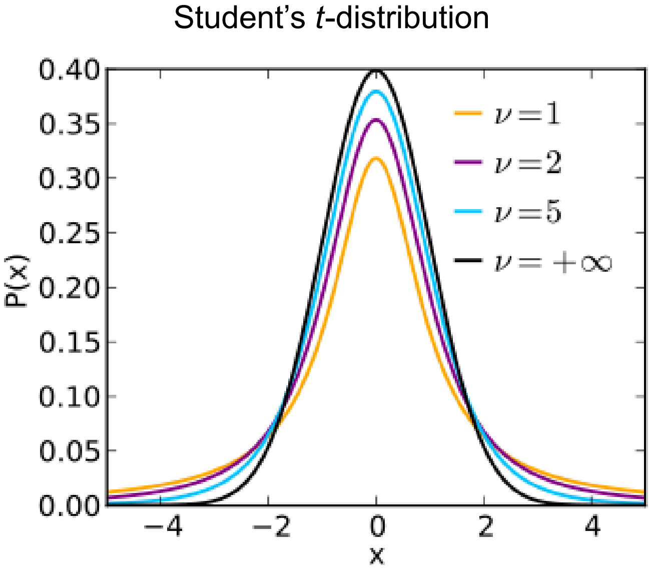
Student’s t distribution

- Problem: we don’t know the standard deviation of the population, so we can’t compute the exact standard error of our mean estimate
- We instead rely on an estimate of the standard error
- Since this estimate will be pretty far off sometimes, our confidence interval will be wider or narrower than it should be sometimes
- Thus, our confidence interval won’t actually contain the true population mean 95% of the time like it’s supposed to (if our sample is small)
- Solution: make a correction to account for the fact that we’re using an estimate of the standard error
    - Specifically, we’ll multiply the standard error estimate by a number larger than 1.96

Student’s t distribution

- So how do we figure out what number to multiply the standard error by?
- We use a distribution called the Student’s t distribution that was created (discovered) specifically for this purpose
- The Student’s t distribution is sort of like the standard normal distribution, except that its exact shape depends on the sample size
- 95% of the area under the standard normal distribution is within 1.96 units of the mean (0)
- With the t distribution, when the sample size is small, we’ll have to go more than 1.96 units from the mean to get to 95%

Student’s t distribution



Student’s t distribution

- With small samples, the standard error estimate is quite imprecise, so the tails of the t distribution are noticeably fatter than the standard normal distribution
- With large samples, the standard error estimate is very precise, so the t distribution is virtually indistinguishable from the standard normal distribution
- Using “critical values” from the t distribution makes it so that a 95% confidence interval actually will contain the true population mean 95% of the time, even with small samples (assuming other assumptions are met)
- Stata will automatically do this extra work for us; you just need to know that it uses a t distribution instead of a normal distribution to deal with small samples correctly

Student’s t distribution

- In the context of the t distribution, the sample size is expressed using something called “degrees of freedom”
- Degrees of freedom: a way to measure the sample size
    - Whenever you hear “degrees of freedom,” think “sample size”
    - The degrees of freedom won’t be exactly equal to your number of observations because it’s a measure of sample size that adjusts for how many things you are estimating; the details aren’t important for this class, but you can read more in the textbook if you’re curious

Student’s t distribution

- Example:
  - We want to estimate the average gas mileage for a certain car type based on a sample of 12 cars that have been driven
  - Type into Stata:

    ```
    use http://www.stata-press.com/data/r14/fuel
    sum mpg1
    ```

  - We can compute a standard error estimate: $\frac{s}{\sqrt{n}} = \frac{2.73}{\sqrt{12}} = .79$
  - If we were to simply multiple the standard error estimate by 1.96 to create our confidence interval, we’d get [19.5, 22.5]
  - But we need to use a number larger than 1.96 to account for the fact that we use an estimate of the standard error (and have a small sample)

Student's t distribution

- Example:
  - Stata will automatically calculate the confidence interval for us using the t-distribution if we type:

    ```
    mean mpg1
    ```
  - The correct confidence interval is [19.3, 22.7], which is slightly wider than the [19.5, 22.5] we got when we used a normal distribution

Appendix 1: Estimating satisfaction

Return to main slides

- Let's pretend that we (as managers of an organization/program) want to figure out what the typical satisfaction level is among the clients of a particular program
- Rather than contact everyone, we decide to draw a random sample
- We start with a list of all clients and then use a random number generator to select a subset of them to contact
- Satisfaction question: "Tell us how satisfied you are with our services on a scale from 1-10, where a 10 means you're perfectly satisfied"
- For simplicity, we'll assume that everyone we contact responds to our question

A1: Estimating satisfaction (example)

- Q: In this example, what is our population?
- A: Every client of the program

- Q: What is our sample?
- A: Just the clients who respond to our question

- Q: What parameter do we want to learn about?
- A: Mean (average) satisfaction in the population

A1: Estimating satisfaction (example)

- Once we've drawn our sample, we can use the following formulas to help us make our estimates (*n* is the number of observations in the sample):

$$\text{Sample mean: } \bar{x} = \frac{\sum x_i}{n}$$

$$\text{Sample standard deviation: } s = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}}$$

$$\text{Mean standard error: } s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

$$\text{95\% confidence interval for mean estimate: } \bar{x} \pm (2 \times s_{\bar{x}})$$

A1: Estimating satisfaction (example)

- Let's say we get a sample of 5 clients with satisfaction levels of: 9, 1, 9, 5, 6

$$\text{Mean: } \bar{x} = \frac{9+1+9+5+6}{5} = 6$$

$$\text{S.D.: } s = \sqrt{\frac{(3)^2+(-5)^2+(3)^2+(-1)^2+(0)^2}{5-1}} = \sqrt{11} \approx 3.3$$

$$\text{Standard error: } s_{\bar{x}} = \frac{\sqrt{11}}{\sqrt{5}} \approx 1.5$$

$$\text{95\% C.I. for mean estimate: } 6 \pm (2 \times 1.5) = [3, 9]$$

- Our estimate of the population mean (average satisfaction among all clients) is 6 (the sample mean)
- We're 95% confident the population mean lies between 3 and 9

A1: Estimating satisfaction (example)

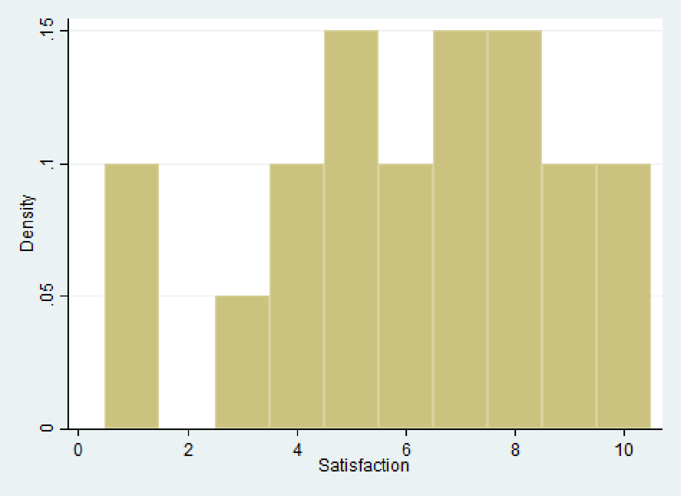
- The formulas I've used here technically don't work so well with small samples ($n < 30$), so this is just a rough approximation
- For example, when forming confindence intervals, we should be multiplying the standard error by a number bigger than 2, as explained in the main slides on the t-distribution

A1: Estimating satisfaction (example)

- In the real world, we use samples when we don’t have access to data for the whole population
- But to learn about the properties of random samples, it’s helpful to play around with a dataset where we do have data for the whole population
- I’m going to pretend that there are only 20 clients in our program and that I have data for all the clients
- In the dataset I’m using, the true population mean is 6.15

A1: Estimating satisfaction (example)

- Here's a histogram of the population data (satisfaction levels for all 20 clients):

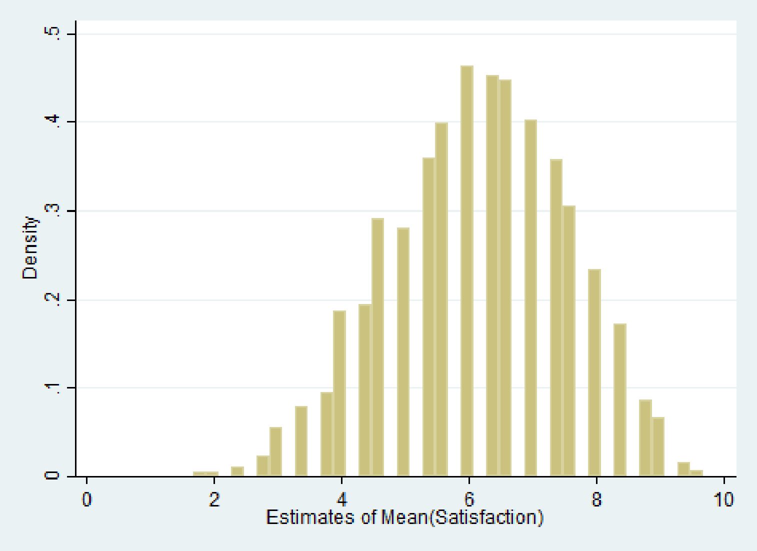


A1: Estimating satisfaction (example)

- To start with, I’ll take a random sample of 3 clients
- I’ll then repeat this process of taking a random sample and generating an estimate 10,000 times to see what kinds of estimates I get

A1: Estimating satisfaction (example)

- Here's a histogram of my 10,000 estimates:



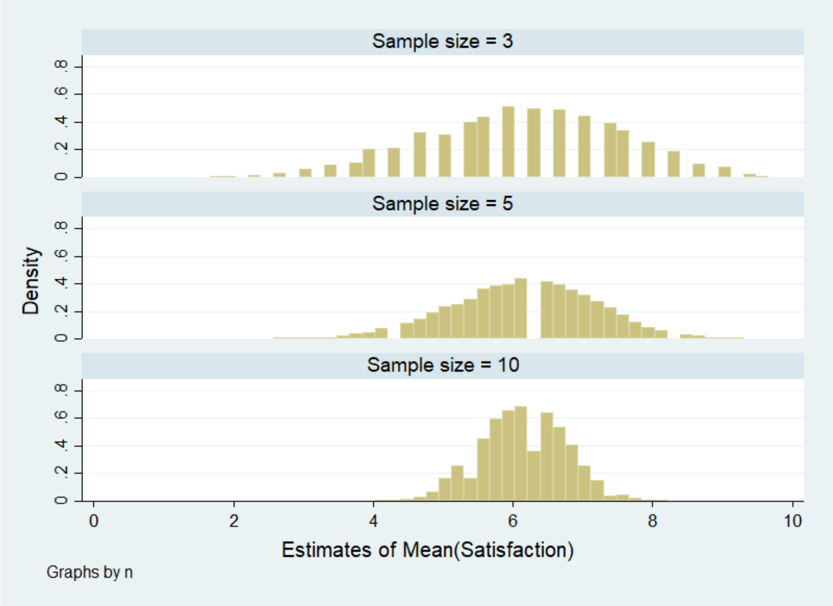
A1: Estimating satisfaction (example)

- Notice that many different estimates are possible, depending on who ends up in my random sample
- I can describe the relative likelihood of getting different estimates using a probability distribution, which we'll call a *sampling distribution*
- The estimates do center around the true value of the population mean (6.15), which means this estimation approach (estimator) is *unbiased*
- I can call the standard deviation of this sampling distribution (the distribution of possible estimates) the *standard error*

A1: Estimating satisfaction (example)

- The distribution of estimates looks sort of like a normal distribution (although there are gaps because of our small sample size)
  - This is because of the *central limit theorem*, which we'll learn more about later
- I can approximate using the general rule for normal distributions that we learned last week–that 95% of the data falls within two standard deviations of the mean (here, 95% of the estimates fall within two *standard errors* of the true *population mean*)
- Now, I'll try collecting many different samples using sample sizes of 5 and 10

A1: Estimating satisfaction (example)



A1: Estimating satisfaction (example)

- As the sample size increases, the estimates are more tightly clustered around the true population mean (the variance of the estimates decreases)
- In other words, larger sample sizes yield more precise estimates

Return to main slides

Appendix 2: Estimating treatment effects

Return to main slides

- We can also use statistical inference to estimate a treatment effect
- Say we have some independent variable (a treatment) coded as either a 1 (unit was treated) or 0 (unit was not treated)
- We can create a very simple model of what causes the values of our dependent variable Y:

$$Y = bX + \varepsilon$$

  - b is the treatment effect (would be equal to zero if the treatment has no effect)
  - ε represents the cumulative effect of all factors besides the treatment that affect the dependent variable

A2: Estimating treatment effects

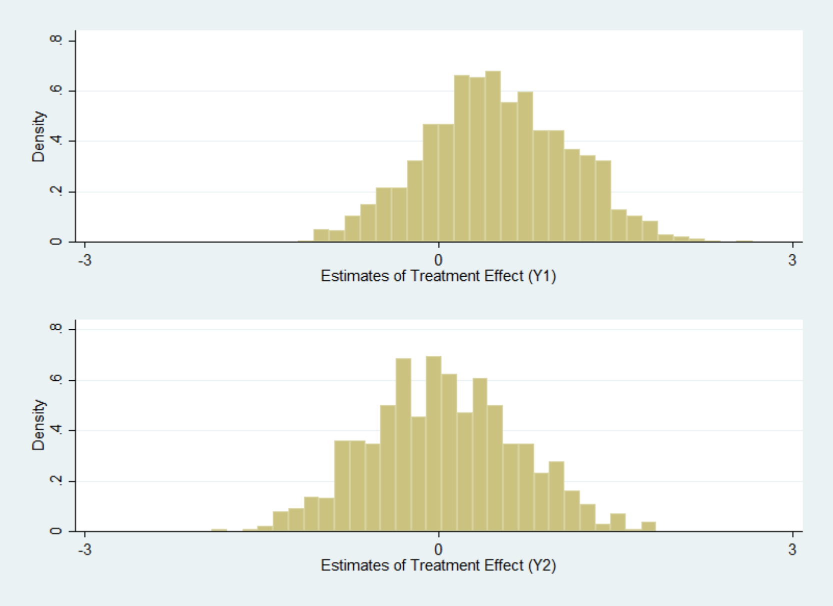
$$Y = bX + \varepsilon$$

- We only observe X and Y, but we'll try to estimate *b*
- Key assumption: ε (which we don't observe) is uncorrelated with *X*
- We can run a regression to estimate *b*
- I'll run two sets of simulations: one where $b = .5$ and one where $b = 0$:

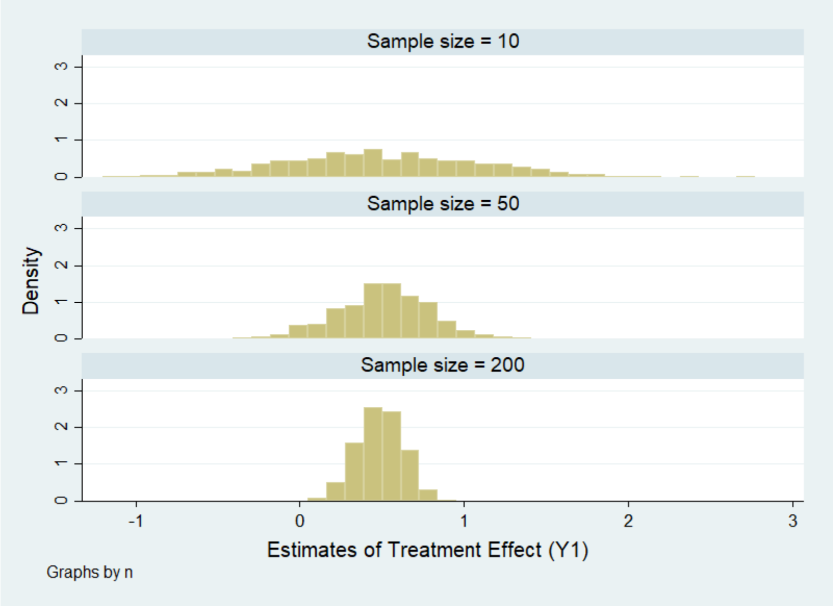
$$Y_1 = .5 \times X + \varepsilon_1$$

$$Y_2 = 0 \times X + \varepsilon_2$$

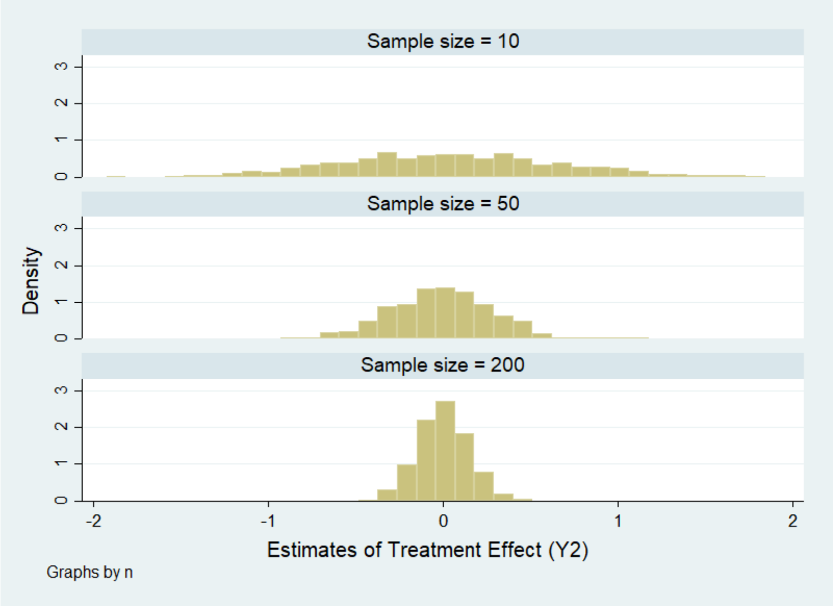
A2: Estimating treatment effects



A2: Estimating treatment effects



A2: Estimating treatment effects



Appendix 3: Using assumptions to estimate the standard error

Return to main slides

- Statisticians have shown that our formula for estimating the standard error is accurate (unbiased) under certain strict conditions:
  1. The population mean is estimated using the sample mean from a *simple random sample*
  2. x is normally distributed
  3. The population is infinite in size
- Assumption 1 is the most important
- Violating assumption 2 is usually okay as long as the sample is reasonably large (greater than about 30)
- Violating assumption 3 is usually okay as long as the population is a lot bigger than the sample (if not, there are other formulas that can correct for this)

Appendix 4: Choosing estimators

Return to main slides

- Expected value: we often use the term "expected value" when discussing the mean or center of a probability distribution
- Bias: the difference between a population parameter and the expected value of an estimator of that parameter
- Unbiased: the expected value equals the true parameter
- Ideally, we want to use estimators that are *unbiased* (all else equal)

A4: Choosing estimators

- How do we know if an estimator is unbiased?
- In real research, we can't measure the true parameter, and we can't know for sure whether our estimates are unbiased
- But inferential statistics allows us to determine whether an estimator is unbiased *under a particular set of assumptions*
- In other words, we can use inferential statistics to see how different estimators behave under different assumed parameters, data generation processes, etc.

A4: Choosing estimators

- To see whether an estimator is unbiased under a given set of assumptions, we need to find the expected value of the estimator and then see if it is equal to the (assumed) population parameter.
    - Analytical approach: use math to figure out the expected value
    - Numerical/simulated approach: simulate many different samples and then use the average estimate from these samples as an approximation of the expected value (fits with definition of expected value)

A4: Choosing estimators

- Example: estimating the population mean
  - Assuming we have a simple random sample...
  - ...the expected value of the sample mean is equal to the population mean (this can be proven by analytical or numerical means).
  - Therefore, the sample mean is an unbiased estimator of the population mean (if we have a simple random sample).

A4: Choosing estimators

- Another example: estimating the population variance
  - Assuming we have a simple random sample...
  - ...the expected value of the sample's variance (calculated using the formula we learned before; also shown below, with slightly different notation) is smaller than the true population variance. Therefore, it is a *biased* estimator of the population variance.

$$\frac{\sum(x_i - \bar{x})^2}{n}$$

  - Intuition:
    - Since variance is based on the (squared) distance between data points and the mean...
    - ...and since $\bar{x}$ is the sample mean rather than the true population mean (which is presumably unknown)...
    - ...we will tend to underestimate the variance because the sample's mean will be closer to the data than the population mean.

A4: Choosing estimators

- Another example: estimating the population variance
  - Alternative estimator:

$$s^2 = \frac{\sum(x_i - \bar{x})^2}{n - 1}$$

  - Assuming a simple random sample, this is an unbiased estimator of the population variance
  - This statistic is known as the “sample variance” (as opposed to the formula we learned in week 2, known as the “population variance”)
  - Notice that only difference is division by $n - 1$ instead of n
  - As the sample gets larger (n increases), the difference between the two formulas gets smaller
    - Intuition: As the sample size increases, the sample mean ($\bar{x}$) becomes a more precise estimate of the population mean (μ), making the prior formula a better estimator

A4: Choosing estimators

- All else equal, we prefer estimators with smaller standard errors
- Example: stratification can give you smaller standard errors
- For basic estimation setups, we can usually find an estimator that minimizes bias (is unbiased) and that also minimizes the standard error
- But sometimes (especially with more complicated data structures or statistical models), we face a tradeoff between (1) an unbiased estimator with a larger standard error and (2) a slightly biased estimator with a smaller standard error
  - To choose between these two options, we have to decide if we want to have a more precise estimate that tends to be slightly off the mark or a less precise estimate that is on the mark

Return to main slides