Introduction to Multivariate Regression & Econometrics

HED 612

Lecture 2

1. Prep
2. Review of Statistics
3. 10 Min Break
4. R Basics

Prep

Download Data and Open R Script

If you had trouble installing R/ R Studio, let me know before moving forward

Download Data and Open R Script

1. Create a new data folder called "ca"
   - hed612 »> data »> ca
2. Download the California Dataset from D2L (under Data)
   - Place the "caschool-v2.dta" dataset into the "ca" folder you created in the previous step
3. Download the Lecture 2 PDF and R script for this week
   - Place all files in hed612 »> lectures »> lecture2
4. Open the RProject you created last week (should be in your main HED612_S21 folder)
5. Once the RStudio window opens, open the Lecture 2 R script by clicking on:
   - file »> open file... »> [navigate to lecture 2 folder] »> lecture2.R

Review Homework

- Every week we will spend 10-15 min s a class or in groups reviewing the problem set you just submitted
- Problem Set #1 Common Questions [review as a class]
  - Why same Q4 and Q5?
    - I will often ask same questions multiple times seeking different solutions
    - Absolute and relative filepaths will techinically get you to the "right" answer; I want you to understand/get practice with the difference between these two approaches
  - Why a word document and an R script?
    - Problem sets will ask both "substantive" questions and "R syntax/code" questions
    - R scripts should only contain R syntax/code and short comments describing what you are doing via the R syntax/code
    - Word documents should be used for substantive answers; no need to copy in R syntax/code into the word document
- Questions or Concerns?
- Solutions to all problem sets with be posted on D2L after I finish grading
- If you need a bit more help with absolute/relative file paths
  - YouTube Video: https://www.youtube.com/watch?v=ephId3mYu9o&t=1s

Today and Next Week

Today

- Review of statistics [univariate]
- R Basics

HW & Reading

- HW#2 posted on D2L
- No reading for next week

Next Week

- Review of Statistics [bivariate]
- Introduction to bivariate regression

Review of Statistics

Statistical Inference

Goal of statistical inference is to infer something about the population from a sample

Population Parameter: a measure of the population

- e.g., mean household income for all U.S. households; proportion of all American voters approving president's performance;
- We don't know this 99.9999999% of the time because we don't have data on the entire population

Sample is part of, a subset, of the population

Estimator (or Statistic): a formula or procedure used to estimate the value of the population parameter using a sample of the population

- e.g., formula to calculate sample mean ($\bar{Y}$) or sample proportion ($\hat{p}$)

Point Estimate: numeric value generated from calculating an estimator from a specific sample of data

- e.g., sample mean household income is \$81,000, 2017 American Community Survey

Parameters and Point Estimates

Introductory statistics class

- Parameter: Population mean (μ_Y)
- Estimator: Sample mean ($\bar{Y}$)
- RQs: What is the mean GPA of UA students living in residence halls?

Multivariate regression class

- Parameter: Population regression coefficient (β)
- Estimator: Sample regression coefficient ($\hat{\beta}$)
- RQs: What is the effect of living in residence halls on GPA?

Notation for Parameters

Population Parameters

- Denoted by Greek setters, usually lowercase
  - μ : "mu" refers to population mean
  - σ: "sigma" refers to population standard deviation
  - β: "beta" refers to population regression coefficient
- Subscripts usually denote population parameters of certain variables
  - μ_Y : "mu Y" refers to population mean of the variable Y
  - σ_X: "sigma X" refers to population standard deviation of the variable X

Estimators of Population Parameters (two general approaches)

- Denoted using Greek letters with a "hat"
  - $\hat{\mu}_Y$: "mu Y hat" refers to the estimate of μ_Y
  - $\hat{\beta}_X$: "beta X hat" refers to the estimate of β_X
- Denoted using English/Arabic letters
  - $\bar{Y}$: "Y bar" refers to the estimate of μ_Y
  - s_X: "S of X" refers to the estimate of σ_X

Data Sources

Experimental Data: obtained from experiments designed to assess the causal effect of a "treatment" on an outcome

- randomized control trials (experiments) are the **gold standard** of program evaluation [will learn more about this]
- e.g., Tennessee STAR project

Observational Data: obtained from surveys, administrative records [focus of this course]

- researchers have/had no control on the "treatment"
- e.g., Beginning Postsecondary Students Longitudinal Study, High School Longitudinal Study, National Education Longitudinal Study

Data Types

Cross sectional data: data on different "observations" (e.g., students, classrooms, universities) for a single point in time [focus of this course]

- e.g., California Test Score data where "observations" are districts

Time-Series Data: data on a single "observation" collected at multiple time points

- e.g., US inflation and unemployment rate data 1959-2000

Longitudinal Data (or panel data): data on multiple "observations" at multiple time points

- e.g., Beginning Postsecondary Students Longitudinal Study, High School Longitudinal Study, National Education Longitudinal Study

Variables

Continuous Variables

- Variables that take on a continuum of possible values where the distance between one value and another is meaningful
- Examples: Age, income, GPA, test scores

Discrete Variables (or categorical variables)

- Variables that can only take on specific (discrete) integer values
- Can be nominal (school type), ordinal (likert scale), and quantitative (number of children)

Measures of Central Tendency: Sample Mean

Sample mean of Y or denoted as $\bar{Y}$

- $\bar{Y} = \frac{\sum_{i=1}^{n} Y_i}{n}$
  - where subscript *i* refers to the observation i

Example: Variable Y has the following six observations $(Y_1 ... Y_6)$

- Obs: $Y_1 = 5, Y_1 = 2, Y_3 = 13, Y_4 = 11, Y_5 = 18, Y_6 = 22$
- Calculate $\bar{Y}$

Other measures of central tendency:

- Median
- Mode

Measures of Dispersion: Standard Deviation

Sample standard deviation of Y or denoted as $\hat{\sigma}_Y$

- Standard deviation is, on average, how far away a random observation, Y_i, is from the sample mean, $\bar{Y}$
- $\hat{\sigma}_Y = \sqrt{\frac{\sum_{i=1}^{N}(Y_i - \overline{Y})^2}{N-1}}$
  - "square root of sum of squared deviations divided by n-1"

Example: Variable Y has the following six observations $(Y_1 ... Y_6)$

- Obs: $Y_1 = 5, Y_1 = 2, Y_3 = 13, Y_4 = 11, Y_5 = 18, Y_6 = 22$
- Calculate $\hat{\sigma}_Y$

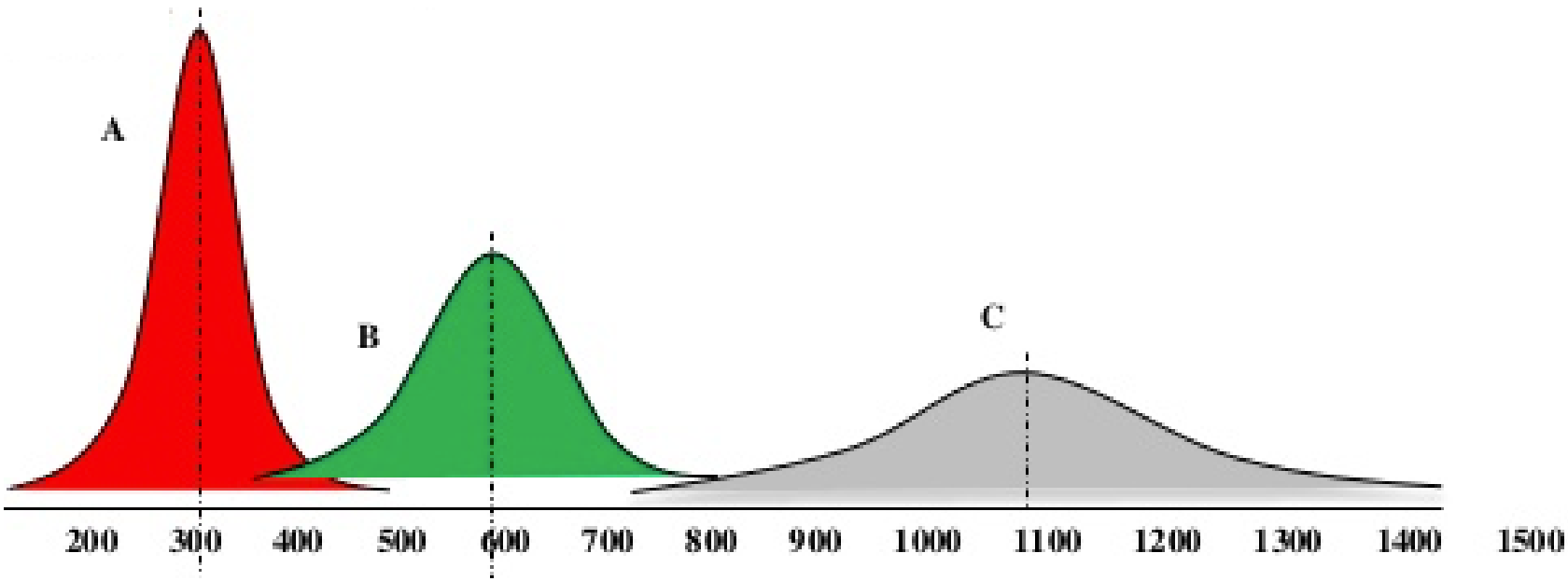
Other measures of dispersion:

- Variance (SD is the square root of the variance)
- Range

Sample Mean & Standard Deviation

Substantive interpretation(s):

- Which distribution has the greatest mean and lowest mean?
- Which distribution has the greatest standard deviation and lowest standard deviation?



Sample Mean & Standard Deviation

Substantive interpretation(s):

Consider the SAT scores of students at a large, public high school and the SAT scores of the Harvard freshman class:

- Which do you think has a larger mean?
- Which do you think has a larger standard deviation?

Sampling Distribution & Central Limit Theorem

Many of the statistical tests we use (e.g., t-tests, confidence intervals) assume that populations we work with are normally distributed.

- A bit unrealistic (outliers, skewness, bimodal)

An appropriate sample size and *Central Limit Theorem* can help us get around this assumption!

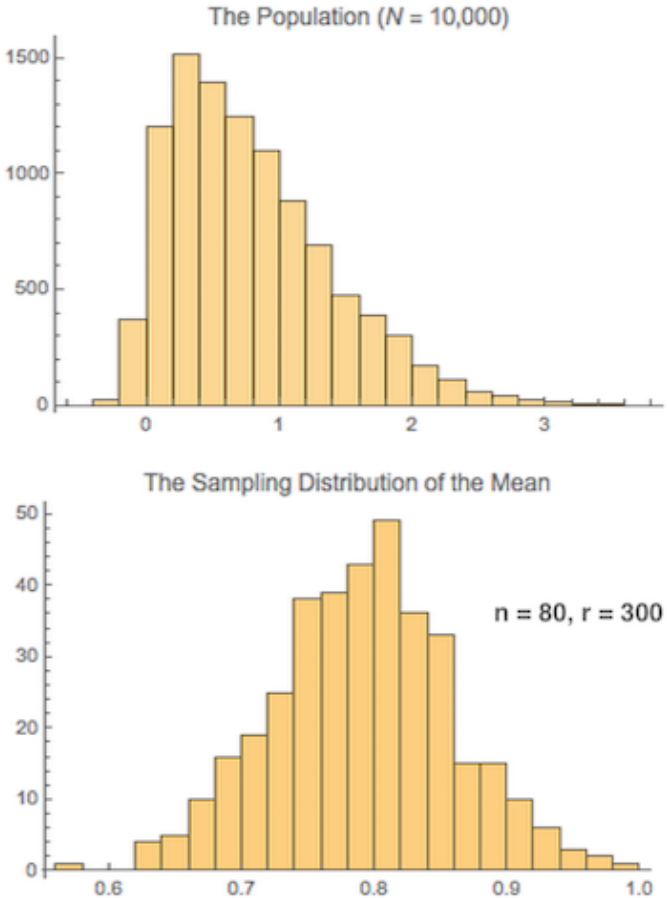
- We usually determine appropriate sample size via exploratory data analysis

Central Limit Theorem: No matter the distribution of our population parameter, given a suffficiently large sample size, the sampling distribution of the estimate (e.g., mean) for a variable will approximate a normal distribution.

Sampling Distribution & Central Limit Theorem

Sampling Distribution

- **Mean of Sample Means** $(\bar{Y}_Y)$ = population mean (μ_Y)
- **Standard Error** $SE(\bar{Y})$: the standard deviation of the sampling distribution. In other words it is the average distance between a random sample mean and the mean of sample means.



Interactive presentation link

10 Min Break

R Basics

R Script Basics

Go to the lecture2.R script I had you download before class

Within an R script you can run commands in different ways:

1. One command at a time by highlighting only that command and clicking the "Run" Button
2. Several commands at a time by highlighting several commands and clicking the "Run" Button
3. Run the entire R script by not highlighting any commands and clicking the "Run" Button

Comments

- `#` Will comment out the remainder of syntax on that line
- `" "` Will comment out anything between the quotation marks (multiple lines)

Tidyverse

- R sytanx for this class will all be given to you.
- For problem sets, you often will get an example of the code you need to run in the lecture or directly in the instructions; you'll often use that code as template for you the specific task you are working on
- While you have the template; you will often need to change a "couple of minor things" (e.g., the variable name; the "function" for the statistical test you are trying to run, etc.)
- So you will still need to develop some "intuition" as to what the code is doing
- All code (very few exceptions) will be written in the "tidyverse" way as opposed to base R
  - Base R: the "default" coding approach to R; it can be very inefficient
  - Tidyverse: most popular way to write R code that does simple tasks (everything we do in this class is considered simple)

Tidyverse

- Most distinctive part of Tidyverse approach to R syntax is using pipes or piping operator `%>%`
  - Allows for "sequential computations" or multiple commands/tasks in the same line of code
  - I usually think of the words "AND THEN" every time I see a pipe
- `caschool %>% select(enrl_tot, computer, read_scr) %>% var_label()`
  - `caschool` = open the california school dataset **AND THEN**
  - `select(enrl_tot, computer, read_scr)` = select the three variables enrol_tot, computer, read_scr **AND THEN**
  - `var_label()` = print the labels for the three variables selected