Crim 250: Statistics for the social sciences

Introduction

Outline

- Meet the toolkit: from the video you followed.
- Review Chapter 1 of DVB
- What is R?
- Exploratory data analysis using an example: Prisoner count in PA

The toolkit

- **R is like the engine —** Programming language for statistical computing and graphics.
- **RStudio is like the car —** Integrated Development Environment (IDE) for R (we'll use desktop version).
- **R Markdown is a way to produce nice-looking write-ups of R code (and more), on PDF, HTML, etc. —** Lightweight markup language for creating formatted text using a plain-text editor.
- **GitHub saves version of code online (and has version control) —** Software for tracking changes in a set of files. Desktop version allows you to work on your computer, but save changes online.

Breakout room activity

R

- **What is it? —** A programming language used for statistical computing and graphics that you can use to clean, analyze, and graph your data.

- **Who uses it? —** Most commonly used language in statistics and data science (which language is used depends on the discipline).

- **Who runs it? —** Free, open-source: can be edited in a collaborative and public manner on CRAN (The Comprehensive R Archive Network).

Why use it?

- **State-of-the-art —** You will always be able to perform the newest statistical analyses as soon as anyone thinks of them (over 15k packages in late 2020).

- **Quickly updated —** R will fix its bugs quickly and transparently.

- **Helpful community —** has brought together a community of programming and stats nerds (a.k.a., useRs) that you can turn to for help.

Visualization in R

- https://www.r-graph-gallery.com/

R data types

- R has 6 basic data types:
  - character: eg. “Maria”, “Cuellar 1987”, “1987”
  - numeric (real or decimal): eg. 1.5, 1987, 2e10
  - integer: eg. as.integer(3.14) gives 3
  - logical: true or false, eg. NA (missing value, either missing or not).
  - complex: 4+2i
  - raw: stores data as raw bytes

R data structures

- R has several data structures:
  - vector:
    - atomic vector: usually of same type, eg. x<- c(1,2,3)
    - list: acts as a container, eg. x<- list(1, "a", TRUE, 1 + (0+4i))
  - matrix: like a vector but with 2 dimensions, eg. m <- matrix(nrow = 2, ncol = 2)
  - data frame: VERY IMPORTANT in R, it's the data table.
  - factors: special vectors for categorical data, eg. factor(c("yes", "no", "no", "yes", "yes"))
  - tables: a frequency table (how many of a type in a vector or matrix)

An example:
Prisoner count in PA (1978-2016)

- Source: https://jacobdkaplan.com/
- How many prisoners are there in the US? (Note: not how many new admissions, just the count for that year.)
- See R code on Canvas.

Exploratory data analysis (EDA)

- The idea is to explore the data before you perform any analysis on it.
- EDA: Some of the most useful parts of statistics.
- Most of the statistics we see on the news are EDA.

Exploratory data analysis

- A method is either non-graphical (usually calculations of summary statistics) or graphical (summarizes data in a diagrammatic way).

- A method is either univariate (one variable) or multivariate (several, but usually two, variables).

  - Note: *It is almost always a good idea to perform univariate EDA on each of the components of a multivariate EDA before performing the multivariate EDA.*

- **-> The four types of EDA are univariate non-graphical, multivariate non-graphical, univariate graphical, and multivariate graphical.**

Categorical data: Univariate non-graphical

- The only useful univariate non-graphical techniques for categorical variables is some form of tabulation of the frequencies, usually along with calculation of the fraction (or percent) of data that falls in each category.

Statistic/Group	Green	Blue	Yellow	Total
Count	5	15	10	30
Proportion	0.167	0.5	0.33	1
Percent	16.7	50	33	100

Quantitative data EDA

- The characteristics of the population distribution of a quantitative variable are its **center, spread, modality** (number of peaks in the prob. density function), **shape** (including “heaviness of the tails”), and **outliers.**

- Our observed data represent just one sample out of an infinite number of possible samples.

Sample population standing in for population

- Univariate EDA for a quantitative variable is a way to make preliminary assessments about the population distribution of the variable using the data of the observed sample.

Central tendency

- The most common measure of central tendency is the mean. For skewed distribution or when there is concern about outliers, the median may be preferred.

DVB Chp 1: Define these

- Data:
- Data table:
- Cases/Records:
- Respondent:
- Subject or Participant:
- Experimental unit:
- Sample:
- Population:
- Variable:
- Categorical (or qualitative) variable:
- Quantitative variable:
- Units:
- Identifier variable:
- Ordinal variable:

Terms from DVB Chp 1

- Data: information about something.
- Data table: a way to organize information (think Excel table)
- Cases/Records: rows of data table
- Respondent: individuals who answer a survey
- Subject or Participant: person on whom we experiment.
- Experimental unit: inanimate subjects, eg. animals, plants, websites.
- Sample: what was collected…
- Population: from the group of interest
- Variable: eg. columns in a data table.
- Categorical (or qualitative) variable: eg. Green=1, Yellow=2, Blue=3.
- Quantitative variable: eg. Age, time of day, number of buses.
- Units: eg. meters, pounds, seconds.
- Identifier variable: eg. Student ID: 0001, 0002, etc.
- Ordinal variable: eg. How worried are you about global warming? Not very=1, Not sure=2, Very worried=3.

R Markdown cheat sheet

- https://www.rstudio.com/wp-content/uploads/2015/02/rmarkdown-cheatsheet.pdf