Introduction to Statistics

Stat 120

March 28 2022

Something about me

- First year at Carleton
- Originally from Nepal
- PhD in Applied Statistics from UC-Riverside
- Diverse education background
- Avid learner and traveler

My webpage

COVID-19 related policies

- Stay home when sick. (Even if you don’t have COVID-19, you should stay home if you aren’t feeling well.)
- Follow CDC on testing, quarantine, and isolation.
- Follow the College mask-wearing policy

What will you learn?

- Analyzing data by doing exploratory data analysis
- Estimate some parameter of interest from the population
- Infer the population characteristics based in your estimation
- Quantify the uncertainty in the estimation

What will a typical day/week look like?

Before Class:

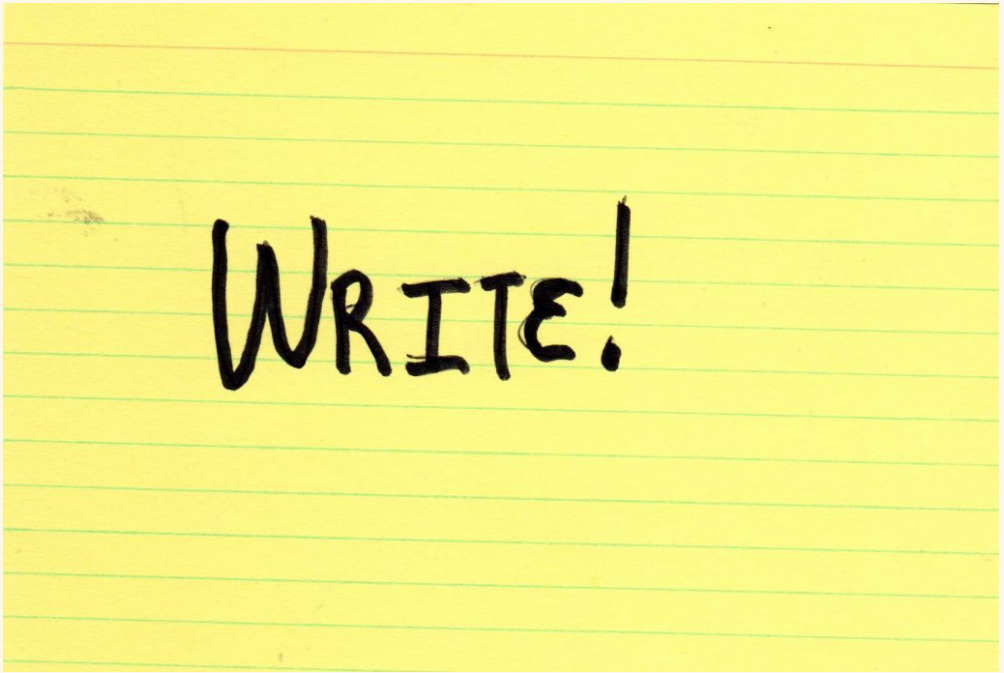
1. Some reading/video to introduce some topics
2. Will be updated in the weekly planner

During Class:

1. Mini-lectures
2. Hands-on group activities

Your introduction

- Your name?
- What gender pronouns do you use?
- Favourite Mathematician/Scientist/Person?
- Recent fun memories?



Please fill in!

Statistics is distinct from mathematics

Study of data and the uncertainties surrounding them

We will take a more conceptual route to statistics in this course

Software Component

- Statistical computing software called R
- RStudio gives nice user-friendly interface to R
- RMarkdown is platform in Rstudio to write your codes and results

We will gradually learn these things!

What and Why of Statistics.

- Science of collecting, describing, and analyzing data
  - Sampling
  - Exploratory Data Analysis
  - Inference
- Makes it easier to know the source of uncertainties
- Let's us take an unbiased viewpoint

Data: Cases and Variables

Data are a set of measurements taken on a set of individual units

- These are **cases** or units

Data is stored and presented in a dataset that comprises of variables measured on cases

- A **variable** is any characteristic that is recorded for each case

`EducationLiteracy` dataset from Lock5

Country	Code	Education	Literacy
Afghanistan	AFG	4.23	43.0
Albania	ALB	3.95	98.1
Algeria	DZA	NA	81.4
Andorra	AND	3.26	NA
Antigua and Barbuda	ATG	NA	99.0
Argentina	ARG	5.78	99.2
Armenia	ARM	2.81	99.7
Aruba	ABW	6.48	97.8
Australia	AUS	5.32	NA

Each row = case & Each column = variable

Click here for Lock5 datasets for the second edition

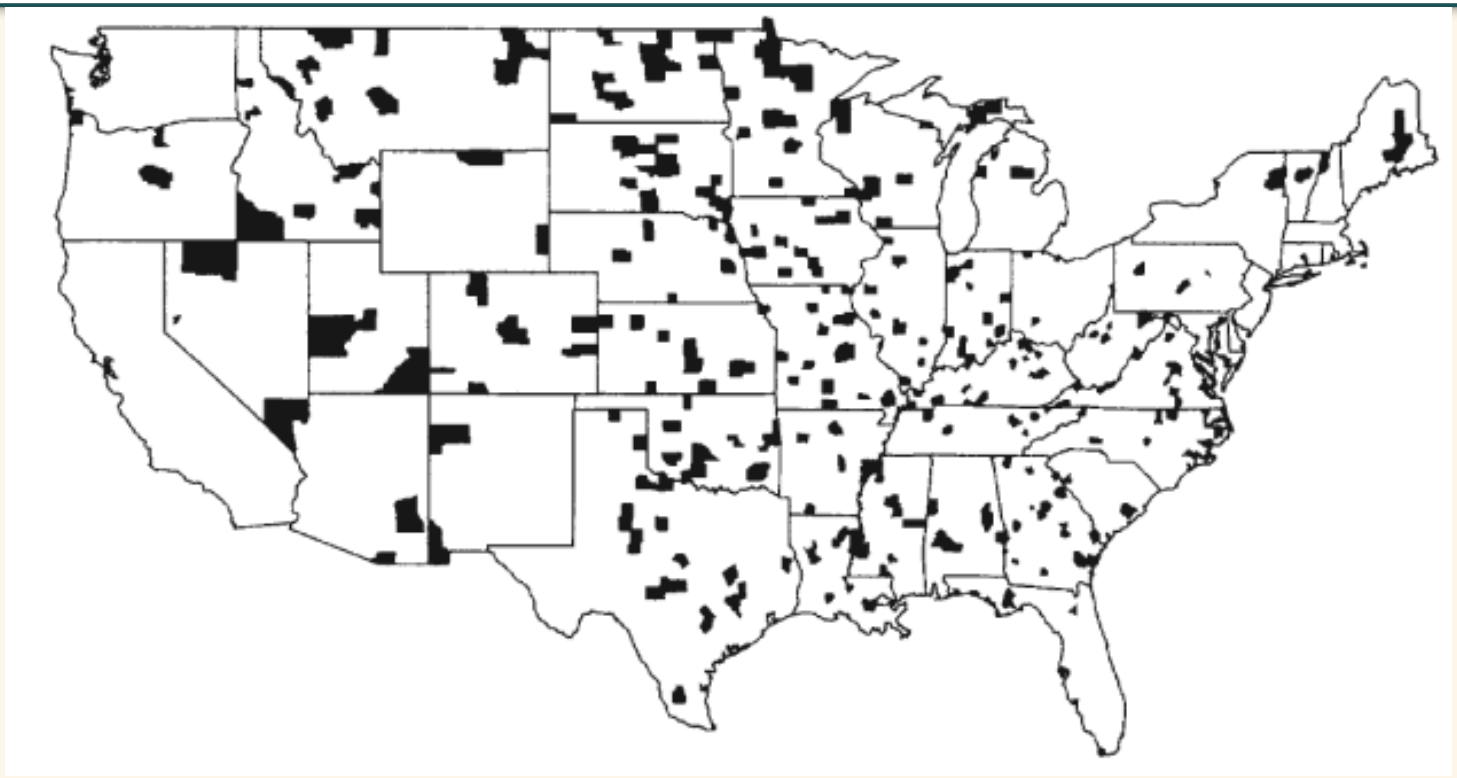
Categorical Versus Quantitative

Variables are classified as either categorical or quantitative:

- A categorical variable divides the cases into groups. e.g. gender, country, state etc.
- A quantitative variable measures a numerical quantity for each case, e.g. age, height, sleep hours, blood pressure etc

Kidney cancer

Counties with the highest kidney cancer rates



Source: Gelman et. al. Bayesian Data Anaylsis, CRC Press, 2004

Kidney cancer

If the cases in the kidney cancer dataset are people, then the measured variable is **categorical**

- We categorize each person as either having kidney cancer or not which is categorical.

Kidney cancer

If the cases in the kidney cancer dataset are counties, then the measured variable is **quantitative**

- Data collected at the county level is aggregated across all people living in the county. We then get rates of cancer which are numbers (quantitative).

Variable manipulations

Can use numbers to code categories of categorical variable

- e.g Gender (1 for male and 2 for female)

Can convert quantitative variable into categorical groups

- e.g. Income (0-50000 as Low, 50000+ as High)

Categorical variables are sometimes collapsed into fewer levels

- e.g. no HS degree, HS degree, some college, 2 year degree

Explanatory and Response Variable

When one variable helps us understand or predict values of another variable, we call the former the **explanatory variable** and the latter the **response variable**

Example 1: Does meditation help reduce stress?

- explanatory variable: **meditation**
- response variable: **stress level**

Example 2: Does sugar consumption increase hyperactivity?

- explanatory variable: **sugar consumption**
- response variable: **hyperactive behavior**

Your Turn 1

10:00

- Your turn features are regularly used in this class
- This is your time to gauge your learning
- Think of these as lab time under my supervision
- Feel free to communicate with your neighbors and form groups

Please download the in class activity file for Day 1 from moodle.

- Save this file to your course folder.
- We will go over this .Rmd file together.
- Once you are done, save the file for further reading.