Introduction to applied statistical methods

Lecture 1

Xiaofei Shi

Course setup:

- Instructor: Xiaofei Shi (xs2427@columbia.edu)
  OH: Thursday 1:30 - 2:30 pm via zoom

- TA: Zhanhao Zhang (zz2760@columbia.edu)
  Yizi Zhang (yz4123@columbia.edu)
  OH: Monday 11:00 - 12:00 pm & 2:00 - 3:00 pm via zoom

- Zoom ID: 956 4866 1979 Passcode: 067214

- If you need more help, feel free to contact us through emails

Course setup:

- Course website: https://innerpeas.github.io/ASM_index.html
- We are using Piazza for discussions
- There are 5 homework and a course project, no exam
- We are using Coursework for submitting and grading
- Evaluation: 50% homework + 30% course project + 20% participation

Class participation

- Subject to change

- According to the current return to school procedure, all class will be in person. In order to protect everyone in the classroom while provide the education we promised you,
  - Please wear a facial mask that covers your nose and mouth
  - Please keep 6-feet social distance
  - Follow the gateway testing

- Discussion and Tutorial will be in person, please bring a laptop!

Teaching style

- To learn statistics well, you need to be able to translate human problems into mathematical terms then into working code.
- Learning is a constant application of the scientific method.
- Ask questions.
- Feedback and communications are important.

Objective

- What you should know after taking this class

- What and how you'll be learning in this class

Expectation changes for 3000+ level course

- You will need to work backwards from a desired outcome
  - Projects will be less spelled out
- If we just taught method A in class, you might need method B instead for the project
- Course is designed to “solve” new problems using both old and new methods
- If you don’t ask a question in this semester, you’re missing out
- You could be interviewing soon, so act like it.
  - No more linear rate learning

Your “expected” background so far

- Statistical Computing
  - Data cleaning
  - Data visualization
- Linear Regression
  - Properties of the regression line (unbiasedness, adding features can hurt)
  - Interpret results and identify appropriate use from regression
  - Diagnostics of regression
  - Modern regression
- Intro Stats
  - Hypothesis testing
  - Random variables
  - Summary statistics

Flow of a typical class

- Before class
  - Do the assigned reading (if any)
- During class
  - Try things out
  - Think about the message
  - Ask questions
- After class
  - Glance over all homework before doing them
  - Do the assignments on time

Applied Statistics is a very broad category

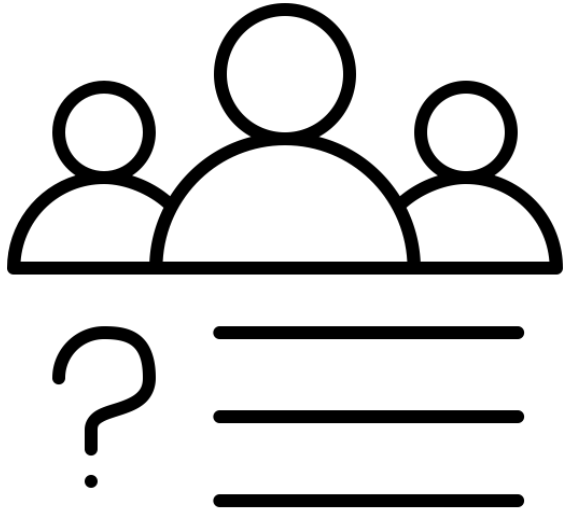
- You could argue any Statistics that uses real data is applied statistics
- “Applied” does **not** mean there is no math
- Your job will require more than interpreting results

What do you think applied statistics will enable you to do?

Topic 1: where does data come from

- Sampling
- Data quality
- Fairness, accountability, transparency

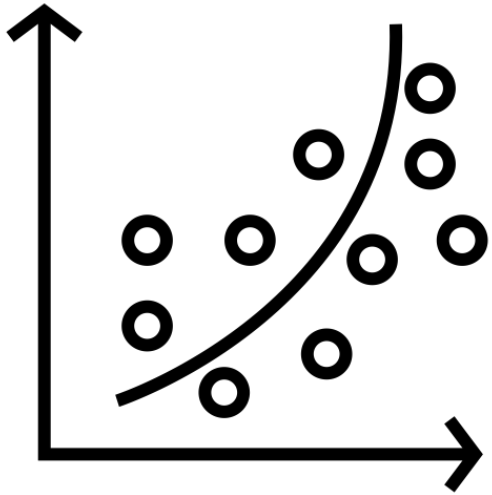
Sounds boring but is the most common weakness I see in students.



Created by sevgenjory
from Noun Project

Topic 2: Inference via regression

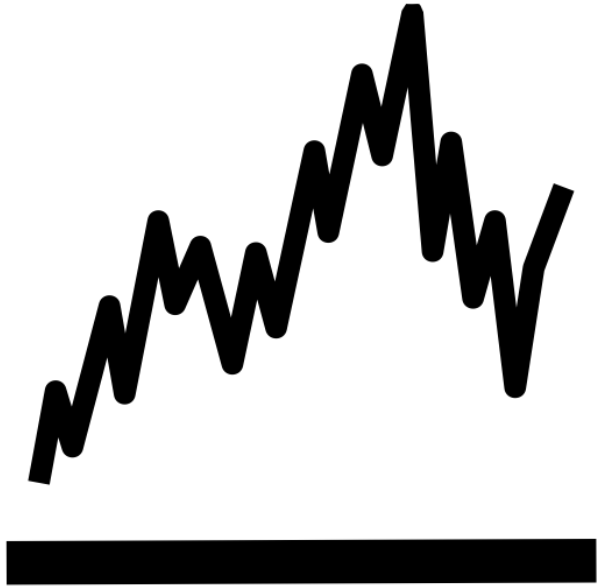
- Model comparison and improvement
- Bayesian statistics



Created by Brand Mania
from Noun Project

Topic 3: Data with dependencies

- Time series
- Spatial statistics



Created by tom
from Noun Project

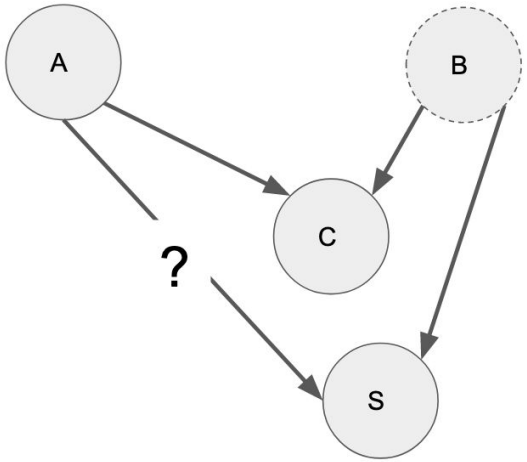
Topic 4: Censored data

- Survival analysis
- “Conditionally at random”

Created by Yugudesign
from Noun Project

Topic 5: Causal inference

- Randomized controlled studies and their limitations
- Methods beyond randomized controlled studies
  - Propensity score matching
  - Instrumental variables
  - Discontinuity
  - Difference in difference



Class Survey

- This is how participation will be measured!

Thoughts/question on remote learning and college education?