ORCA 2500/4500

cjd2119@columbia.edu

Hw 25%
Midterm 25%
Final Exam 25%
Final Project 25%

extra credit
class participation

Course objectives

① teach basic python, (data stack)
SQL

② teach basic statistics and probability

③ regression

④ machine learning

What is data science?

→ multidisciplinary field
that combines the tools of comput,
statistics and machine learn to
solve problem!

→ approach-agnostic

→ what is statistics?
what is machine learn?

① descriptive statistics
↳ summariz and visualiz
data

② inferential statistics
↳ make conclusions in
a mathematically rigorous
way about data

③ learn patts in data an

applying it to making predictions

Populations and sampling

Typically we are interested in understanding the difference between two populations, or learning about some parameters of a population.

Instead of looking at every single member of a population, we're going to look at subset.

↳ how do we generate the subset

Causality

- Not enough to understand that two different populations are different → what to intervene on?

Simpson's predox

→ sick patients given thapton to by a new drg

→ 350 take it, 350 don't
(78.1) (83.1)

	Drug	No drg
M	81/87 (93.1)	234/270 (87.1)
F	192/263 (73.1)	55/80 (69.1)
Combined	78.1	83.1

gold standard for determing
causality →
randomized experiment

Observational study

→ a study in which scientists make conclusions based on data they have observed but had no hand in generating

confounding variable

variable that is correlated with both treatment and the outcome

Experiment

study in which scientists determine the assignment of the treatment