Week 7 lab

COGS 108, 9:00-9:50AM (B01)

https://github.com/alexavndra/cogs108-b01

Reminders!!

- Data Checkpoint due TODAY at 11:59PM
  - Submit on GitHub (commit + push)
- D6 is due Friday, November 17th at 11:59PM
- Office hours is still ongoing!
  - https://calendly.com/alexandrarh/office-hours
    - FYI, book early! It’s getting PACKED.

Data Checkpoint: Thee Rundown

Data section: What you do

You’ve talked about the ideal and real datasets already, so now what? We start getting prepped for the upcoming EDA!

What you need to do:

- ➢ Identify your datasets
  - ○ Describe what’s in them (briefly)
- ➢ Data wrangling
  - ○ Grab/create your dataset (Pandas?)
- ➢ Cleaning up
  - ○ Getting rid of any unnecessary parts from your data (e.g. null values)!

What about the Project Proposal improvements????

Earn those points back by making the changes on the Data Checkpoint notebook! No need to let us know, we will cross compare the improvements made with the original and determine the re-grade!

- ➢ Can do this with other checkpoints!

Inferential analysis

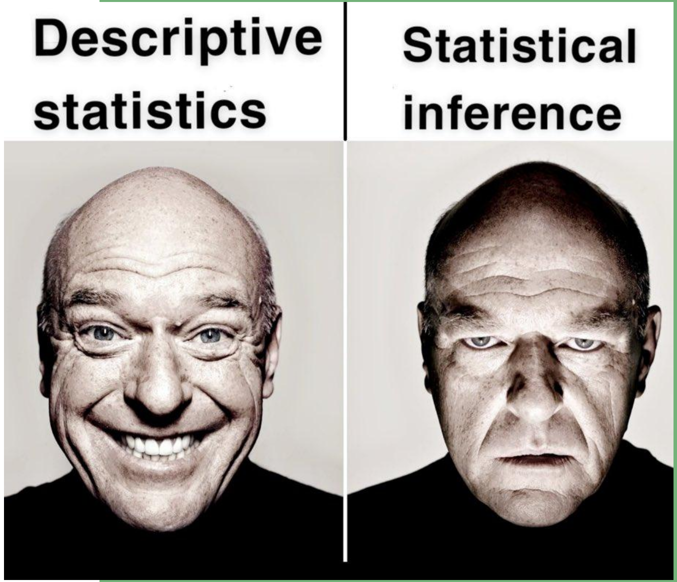
Inferential Analysis

AKA, me when I use my remaining brain cells to come up with a single thought in MATH 20E or 180A

Inferential analysis is the usage of data analysis techniques to infer “properties of a population, for example by testing hypotheses and deriving estimates” of certain groups within the given set (wikipedia.com)

Basically…

- ➢ Using data visualizations to look at sampled data
- ➢ Taking sampled data and using predictor techniques to determine the relationship between sample variables
  - ○ Interpolation vs. extrapolation
- ➢ Drawing conclusions about the data and determining whether hypothesis is right or wrong
  - ○ Null hypothesis?



Types of inferential analysis (most popular imo)

Linear Regression (linear and logistic): Analysis model that allows predictions based on the dependence of the "Y" variable on the "X" variable

- ➢ Usually represented by scatter plots + line of best fit
- ➢ Can use multiple dependent variables (linear)

Analysis of Variance (ANOVA): When multiple averages of a dataset are analyzed for significant differences

- ➢ Great for comparing numerous groups (obviously)

Statistical significance (T-test): Comparing *two* averages based on their common dependent variable (e.g. difference in spending between San Diegans and New Yorkians on local coffee)

- ➢ Best when wanting to examine effects on different groups, like medicine efficacy!

intellspot.com

D6: Inference, but make it a case study...

An overview of D6

Do Pulitzers help newspapers keep readers?, or more specifically *By looking at Pulitzer prizes awarded and changes in readership, can we determine what the effect of prestige is on the viewership at the 50 most popular newspapers between 2004 and 2013?*

- ➢ Using provided data to predict + analyze which factor has the highest influence on viewership
- ➢ Find the variables necessary for analyzation
- ➢ What analysis we will perform

What we'll use to answer D6

Bar graphs: Examining circulations/peaks of newspapers in 2004 and 2013

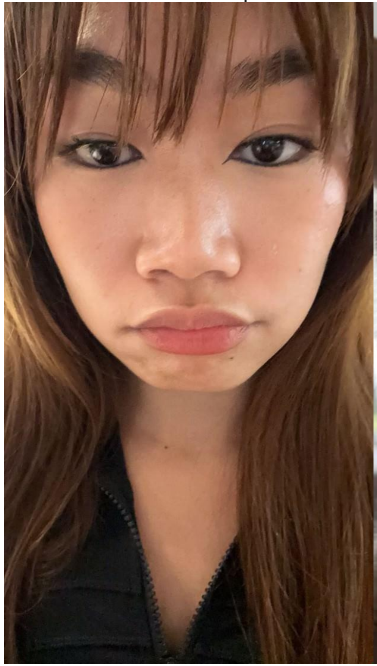
Distribution plots: Bar graphs, but with distribution curves; helps us visualize the gradual distribution of how many people have won *x* amount of Pulitzers, etc.

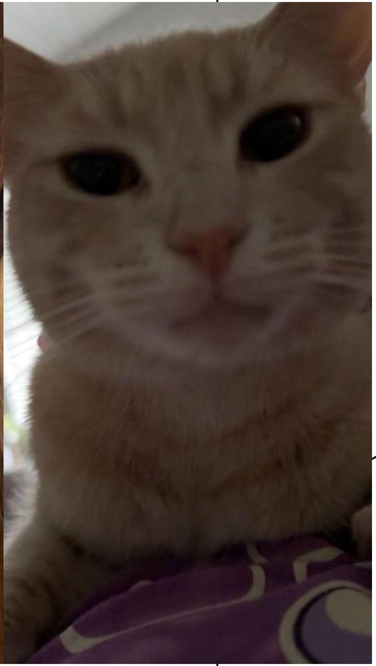
Scatter plots: Seeing how the winners and finalists of different years depend + compare to one another

- ➢ We'll be using a Line of Best Fit and Linear Regression (OLS) on this to further analyze this!

Next time!

D7, A3, nonparametrics...





D6 Demo