





COGS 9 – Discussion Section A01 and A02

Kunal Rustagi (TA): Mon 9AM ([Zoom](#))
Connor McManigal (IA)
Yupei Sun(IA)



COGS9: Reading 4a

Data is Personal: Attitudes and
Perceptions of Data Visualization in
Rural Pennsylvania



Overview

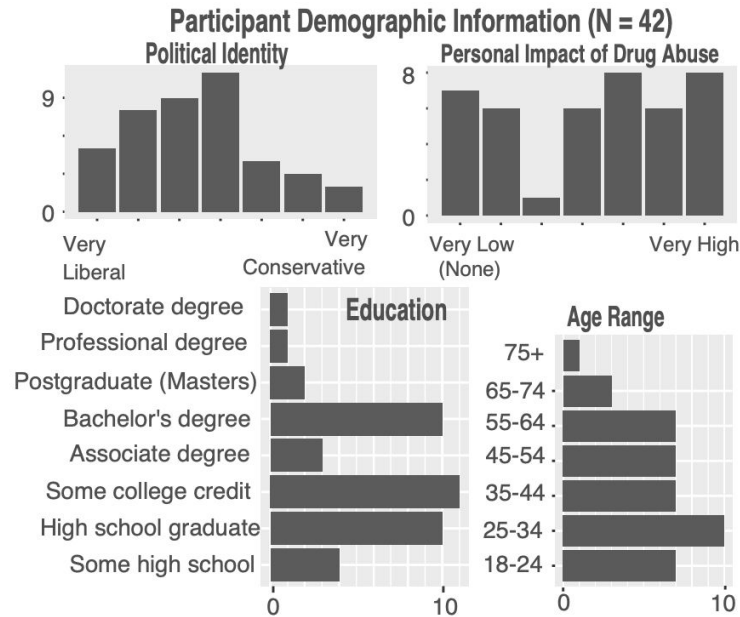
Overall, the paper emphasizes the importance of considering the unique challenges and profiles of underrepresented populations, particularly rural communities, when studying attitudes, biases, and literacy in data visualization. By addressing these issues, data visualization can become a powerful tool for empowering individuals in underrepresented communities to understand and engage with data effectively

Background

Encounters with data can be manipulated by several factors

- Experience or education
- Biases
- Attention
- Focus on people in rural settings is motivated by
 - The population's absence in the visualization literature
 - Gaps in education, income
 - Literacy may impact perceptions of data visualizations

Interviews in rural PA

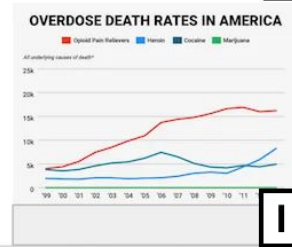
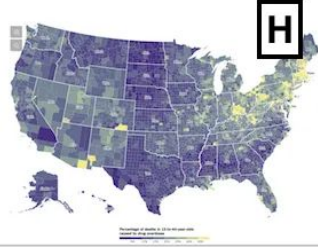
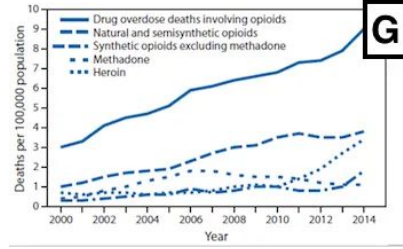
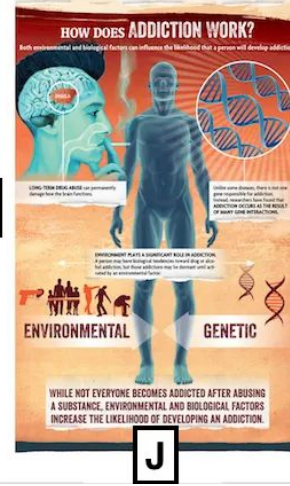
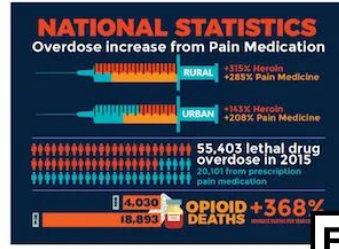
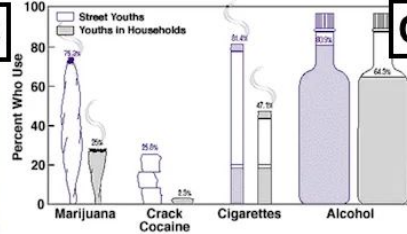
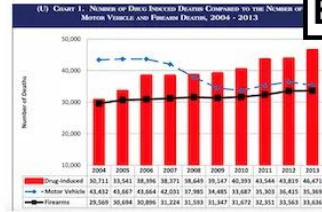
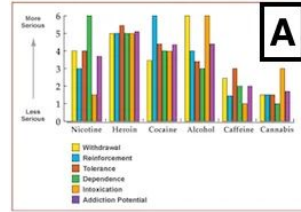


Procedures:

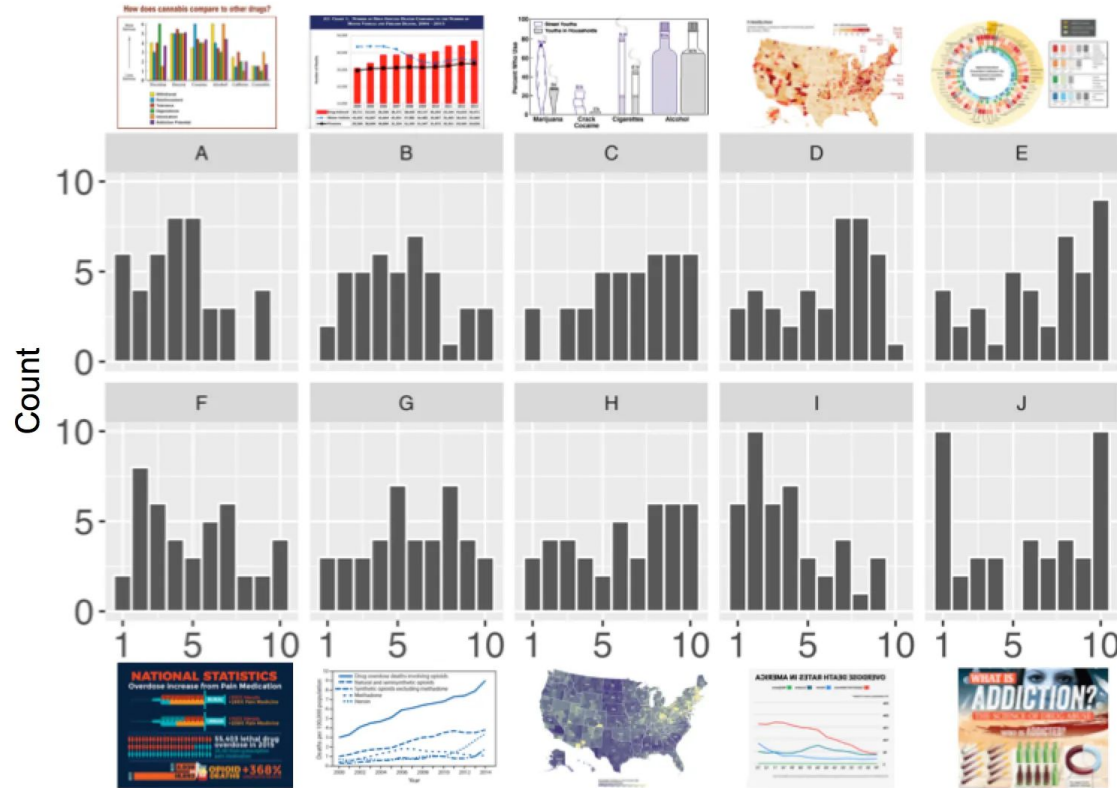
1. Introduction and consent
2. Graphs presentation and ranking
 - Ranking 10 graphs according to usefulness
3. Sources are revealed
 - Participants have the option to re rank the graphs
4. Demographics questions

Graphs used in the interview

How does cannabis compare to other drugs?



Participants ranking



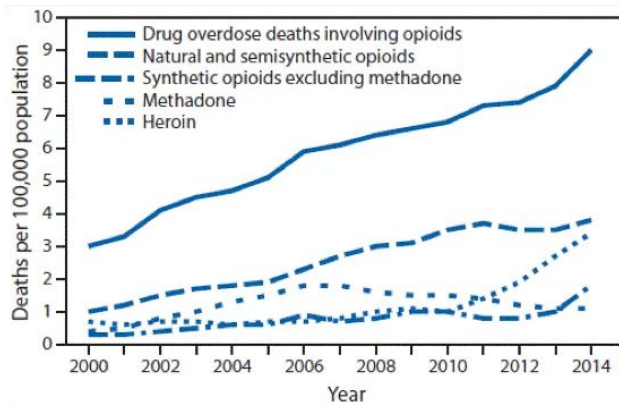
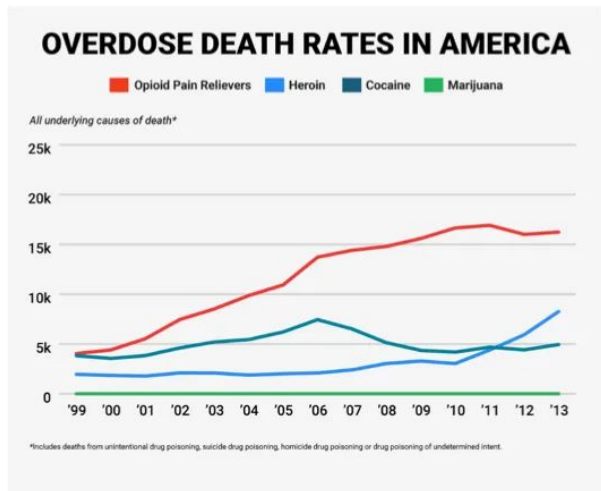
Participant Rankings (1 is high, 10 is low)

Observations:

- Data is messy
- Infographic like J has the most “1” votes and “10” votes

Analysis

Data can be intimate and personal. If someone found a personal connection to any graph, it didn't matter the color, the style or the technique

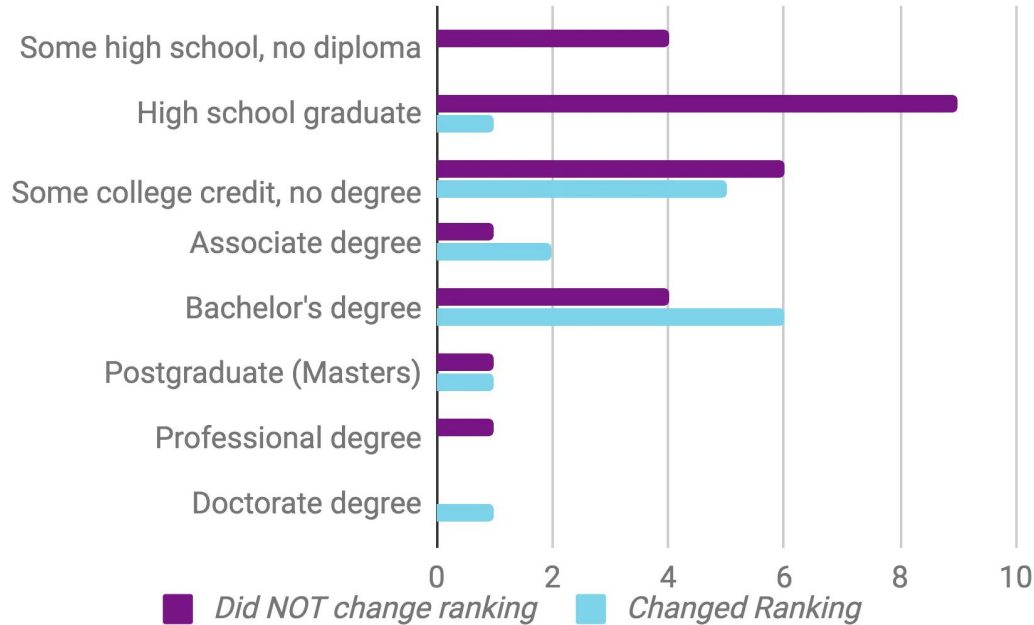


*I ranked it higher just for the simple fact that **I live in America** so I thought it was pretty relevant... more than the other one.*

— 45–54 year old, associate's degree

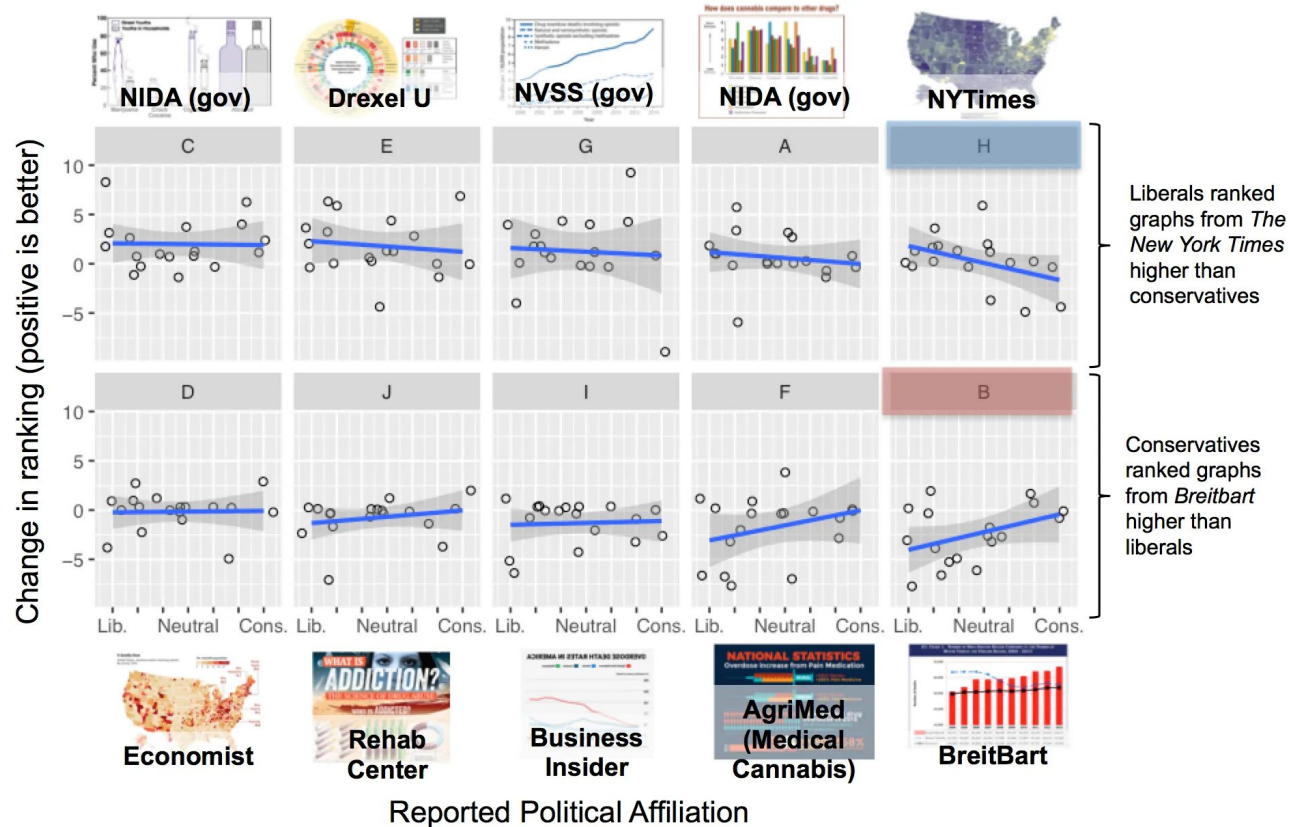
After revealing sources

Who changed their ranking?: Educational Background



- The decision not to change rankings aligned the educational background of our participants
- 12 of 22 participants that did not change their ranking either expressed beliefs that the source of the data visualization is irrelevant or that all sources were equal.

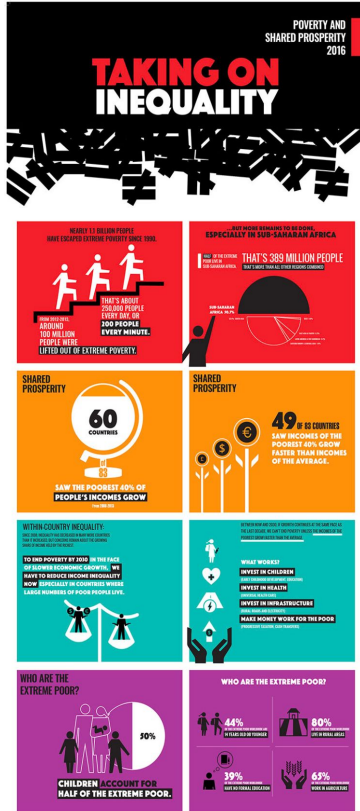
Ranking change V.S Political affiliation



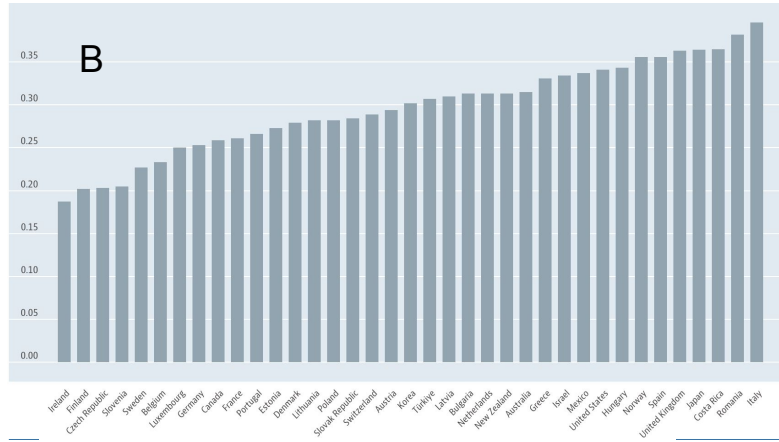
Discussion

- 42 interviews revealed a complex tapestry of motivations, preferences, and beliefs that impacted the way that participants prioritized data visualizations
- Participants valued clarity and simplicity, prefer simple bar graphs and line charts. Others might value more on color and visual appeal
- Highly educated people were more likely to value the source of a data visualization, and that trust can align with political identity

Exercise: Rank the following graphs according to usefulness

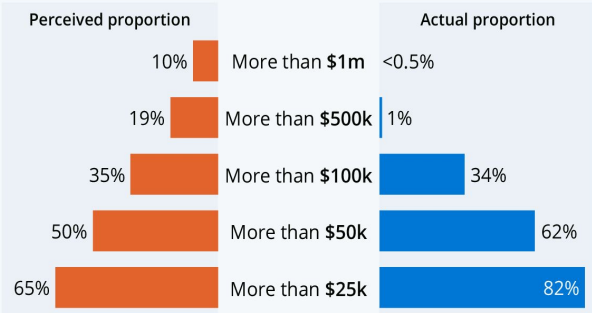


A



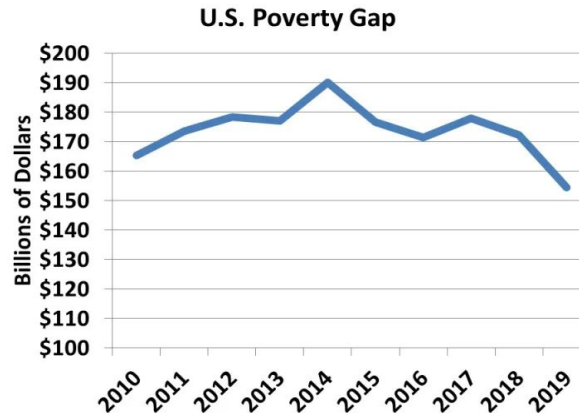
The United States' Real and Perceived Income Gap

Actual and perceived share of U.S. households in the following income brackets*



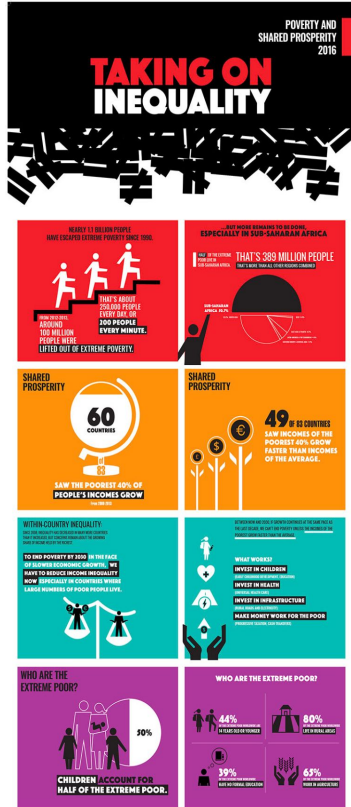
* Estimates based on a survey of 1,000 adults (aged 18+) in the U.S., median weighted responses, Jan 2022. Income data from 2020 U.S. Census. Sources: YouGov, U.S. Census Bureau

C

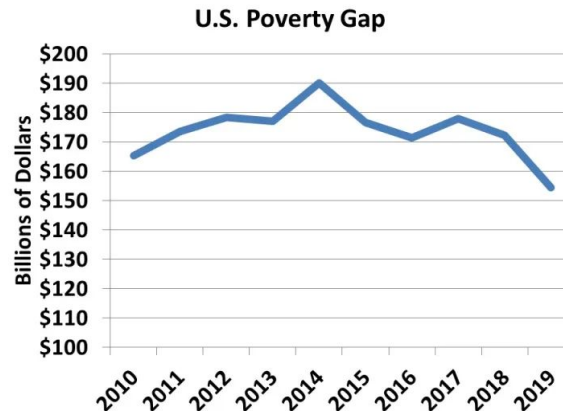
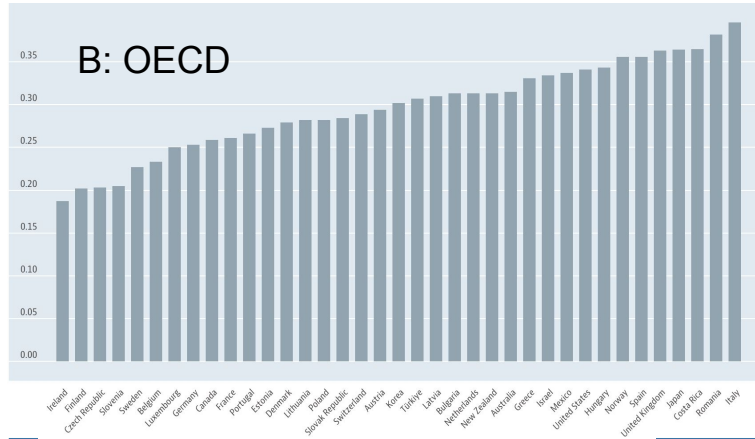


D

Re rank the graphs after revealing the sources



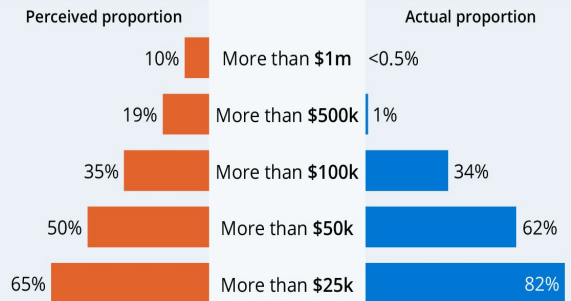
A: The World Bank



D: Federal Safety Net

The United States' Real and Perceived Income Gap

Actual and perceived share of U.S. households in the following income brackets*



* Estimates based on a survey of 1,000 adults (aged 18+) in the U.S., median weighted responses, Jan 2022. Income data from 2020 U.S. Census. Sources: YouGov, U.S. Census Bureau

C: Statista



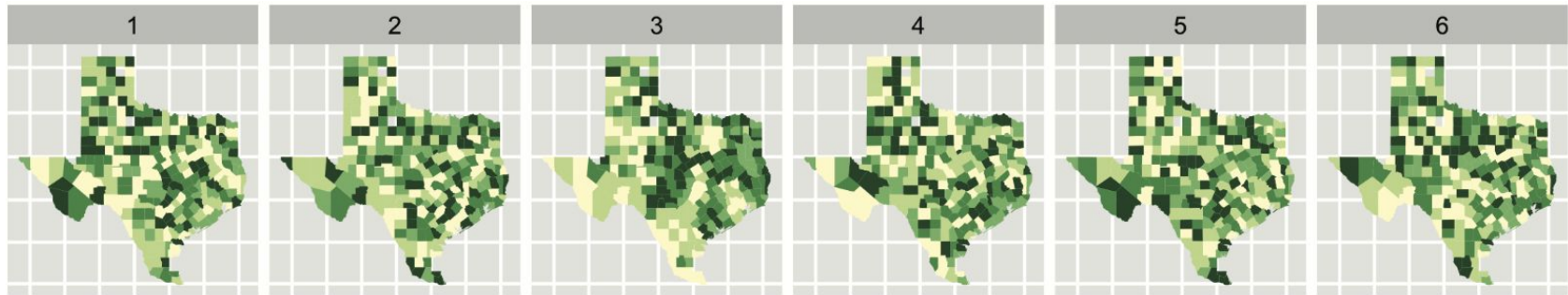
COGS9: Reading 4b

Graphical Inference for Infovis



Role of Statistics in Information Visualization(Infovis)

- Infovis combines curiosity and skepticism in data exploration
- Infovis provides tools of curiosity to uncover new relationships
- Statistical methods provide tools of skepticism to verify relationships



Graphical Inference

- A tool for skepticism that can be applied in a curiosity-driven context
- Graphical inference bridges the gap between curiosity and skepticism
- It allows us to control for **apophenia**(the innate human tendency to see patterns within noise)
- Helps us answer the question: “Is the relationship we are observing really there?”
- Helps us minimize false positives and avoids unfounded findings

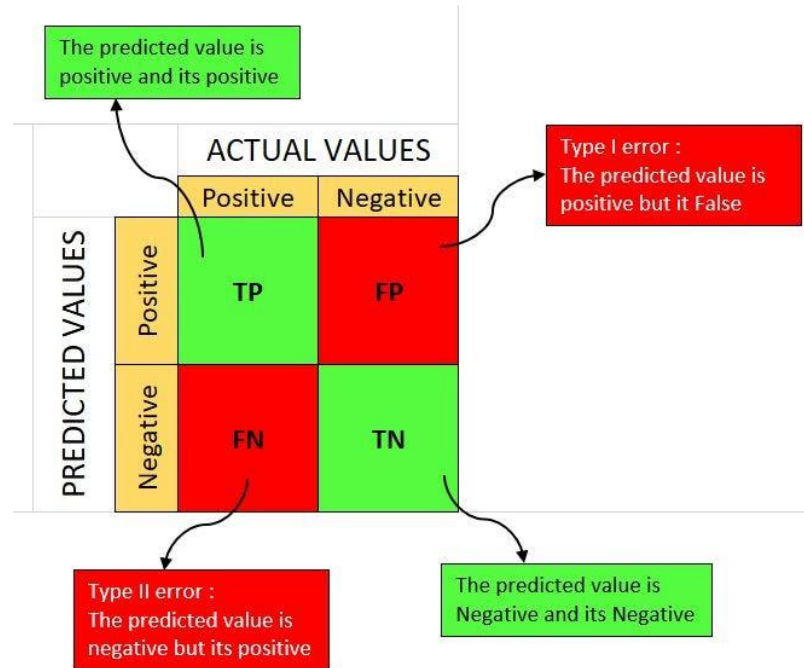
What is inference and why do we need it?

- Two components of statistical inference:
 - Testing (is there a difference?)
 - Estimation (how big is the difference?)
- We don't want our conclusions to apply only to a small sample, but a large fraction of humanity
- For graphics, we want to address the question "Is what we see really there?"
 - In other words, "Is what we see in a plot of the sample an accurate reflection of the entire population?"

Criminal Justice Analogy

- The accused (data set) will be judged guilty or innocent based on the results of a trial (statistical test)
- Each trial has a defense (advocating for the null hypothesis) and a prosecution (advocating for the alternative hypothesis)
- On the basis of how evidence (the test statistic) compares to a standard (the p-value), the judge makes a decision to convict (reject the null) or acquit (fail to reject the null)

True Positives, True Negatives, False Positives, False Negatives & Confusion Matrices



Protocols of Graphical Inference

Rorschach Protocol:

- A calibrator that helps an analyst become accustomed to the peculiarities of random data
- We use this protocol to calibrate our vision to the natural variability in plots in which the data is generated from scenarios consistent with the null hypothesis
- Helps us learn which random features we might spuriously identify

Line-up Protocol:

- Provides a simple inferential process to produce a valid p-value for a data plot
 - Generate $n-1$ decoys (null data sets)
 - Make plots of the decoys and randomly position a plot of true data
 - Can the observer spot the real data?
- Works like a police lineup(the suspect is hidden in a set of decoys)

Common Examples:

1. If we are interested in the spatial trend in a data map, then the null hypothesis might be that location and value are independent
 - To generate null datasets we permute the value column
2. In a scatterplot, an initial hypothesis might be that there is no relationship between x and y
 - We can generate null hypotheses by permuting either the x or y variables
3. If we have clustered the data and are displaying the results with a colored scatter, we might be interested to know if the clusters are well separated
 - The null hypothesis is that cluster membership and position are independent, and we can generate null datasets by permuting the cluster id column

Power

- **Power**: the probability of correctly “convicting” a “guilty dataset”
- Perception psychology guides the choice of effective plots to detect specific structure
- Mapping variables to perceptual properties aids accurate interpretation
- **Aggregation**(summarizing or combining data points into a smaller set of representative values) improves structure detection in large datasets

Reading Practice Questions

Q1: Authors suggest that a null dataset for a scatterplot can be created by “permuting the x or y variable”. What does this phrase mean?

- A. Randomly change the order of values for both variables
- B. Swap the variable names
- C. Randomly change the order of one of the variables
- D. Swap a different variable in for either x or y

Q1: Authors suggest that a null dataset for a scatterplot can be created by “permuting the x or y variable”. What does this phrase mean?

- A. Randomly change the order of values for both variables
- B. Swap the variable names
- C. **Randomly change the order of one of the variables**
- D. Swap a different variable in for either x or y

Explanation: Permuting means to rearrange or alter. So when we permute either x or y, we change the order of one of these variables.

Q2: Traditional statistical tests work best when...

- A. The data suggests some sort of relationship
- B. The data is well behaved and follows a known distribution
- C. The data are complex, but can still be worked with
- D. The data follow an unknown distribution

Q2: Traditional statistical tests work best when...

- A. The data suggests some sort of relationship
- B. The data is well behaved and follows a known distribution**
- C. The data are complex, but can still be worked with
- D. The data follow an unknown distribution

Explanation: Traditional statistical tests are often developed on specific assumptions about the data, such as normality or independence. When these assumptions are met, the tests are more reliable.

Open Q/A + Public Office Hours

- If you weren't here for the attendance you can come up now
- No section next week!