COGS9-Intro to Data Science

Spring24 - Prof. Kyle Shannon

Discussion Section A01
Week 7

Teaching Assistant (TA): Abdullah
Instructional Assistant (IA): Kyra

Where to find all material

COGS 9

Search COGS 9 UCSD Podcast Gradescope

- Home
- Syllabus
- Readings
- Assignments
- Exam
- Final Project
- Office Hours
- Contact Us

This site uses Just the Docs, a

Introduction to Data Science

COGS 9 - UC San Diego - Prof. Kyle Shannon

Spring 2024 | SOLIS 107 | TU & TH 5:00-6:20PM

Welcome 👋

We are all very excited that you decided to join us on this whirlwind tour of data science. All relevant info, e.g. due dates, assignment links, etc. are found on this website. We look forward to teaching and working with all of you and hope to meet you in office hours. Check out the **Getting Started** section so you can hit the ground running when class starts!

NOTE
Week one I try to take as many students from the **waitlist** as I can, please email cogsadvising@ucsd.edu with further questions.

Discussion Sections

	Day	Time	Location	Staff	Materials
A01	Wed	12:00-12:50PM	CENTR 222	TA: Abdullah IAs: Kyra	View
A02	Wed	1:00-1:50PM	CENTR 222	TA: Kaushik IAs: Seshu, Vicky	View
A03	Wed	2:00-2:50PM	CENTR 222	TA: Matthew IAs: Jessica, Wenhua	View
A04	Wed	3:00-3:50PM	CENTR 222	TA: Vineeth IAs: Jiesen	View
A05	Wed	4:00-4:50PM	CENTR 222	TA: Vineeth IAs: Harshita	View

cogs9_TA Public

main | 1 Branch | 0 Tags | Go to file

AbdullahAshfaq Added week3 material

week2	Added week2 material
week3	Added week3 material
README.md	Update README.md

README

Cogs 9 Discussions-Intro to Data Science

Abdullah's discussion section material for COGS9 course

Upcoming Deadlines

Week 7		
Mon, May 13	PROJ	Final Project Part 1 due
Tue, May 14	EXTR	[Extra Credit] Team evaluation due
Tue, May 14	LECT	Machine Learning I
Thu, May 16	LECT	Machine Learning II
Fri, May 17	QUIZ	Reading Quiz 4 due
Fri, May 17	READ	Begin reading 5

Week 8		
Mon, May 20	ASSG	Assignment 2 due
Tue, May 21	LECT	Text & Geospatial Analysis
Thu, May 23	LECT	Giving a Data Science Talk

Week 9		
Mon, May 27	ASSG	Assignment 3 due
Tue, May 28	GLCT	Guest Lec 1. (prerecorded not in-class)

Discussion Sections Outline: Mostly Hands-on

- Week 2: Introductions, Making teams, Reading 1 (Part 1)
- Week 3: Reading 1 (Part 2), Python Basics with Jupyter Notebook
- Week 4: Reading 2, Getting data and wrangling it using Pandas
- Week 5: Reading 3, Assignment 1, Basics of SQL and Visualizations
- ~~Week 6: Reading 4, Final Project Part 1 reviews/discussions~~
- Week 7: Reading 4, Assignment 2, Data Visualization and EDA demo
- Week 8: Assignment 3, Machine Learning demo
- Week 9: Reading 5, Closing thoughts
- Week 10: Final Project Part 2 reviews/discussions

Today’s Outline

- Reading 4 Summary
- Assignment 2 Topics
- EDA

Participation = Extra Credit 😀

Reading 4
Attitudes and Perceptions of Data Visualization in Rural Pennsylvania

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

Which visualizations do people understand?

- Visual literacy
  - Capability of a person “to read, comprehend, and interpret” graphs

- What can cause problems?
    - New graphic representation without training
    - Lack of familiarity

Procedure

- 10 different data visualizations that broadly involve the impact of drugs in the US
- Charts were chosen to represent a diverse set of features, including form, visual appeal, and source
- Each chart was presented to participants in color on individual sheets of paper

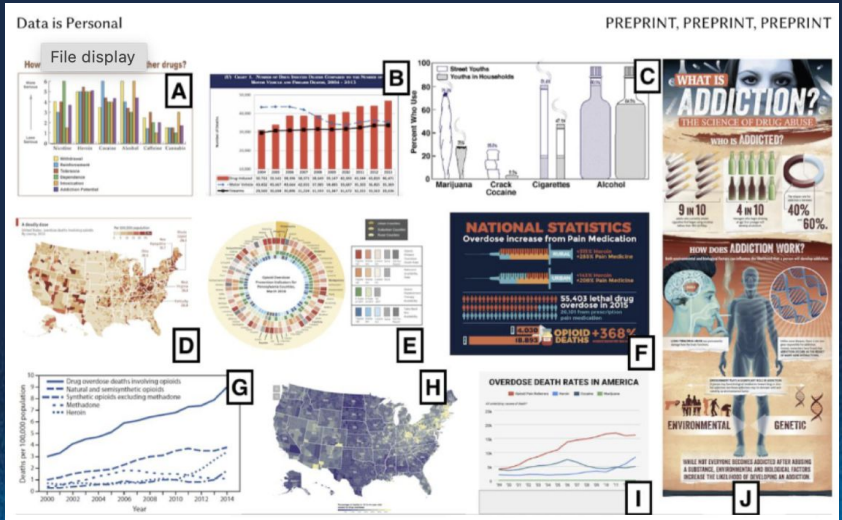


Figure 4: The graphs shown to participants. Each graph was presented on an independent sheet of paper

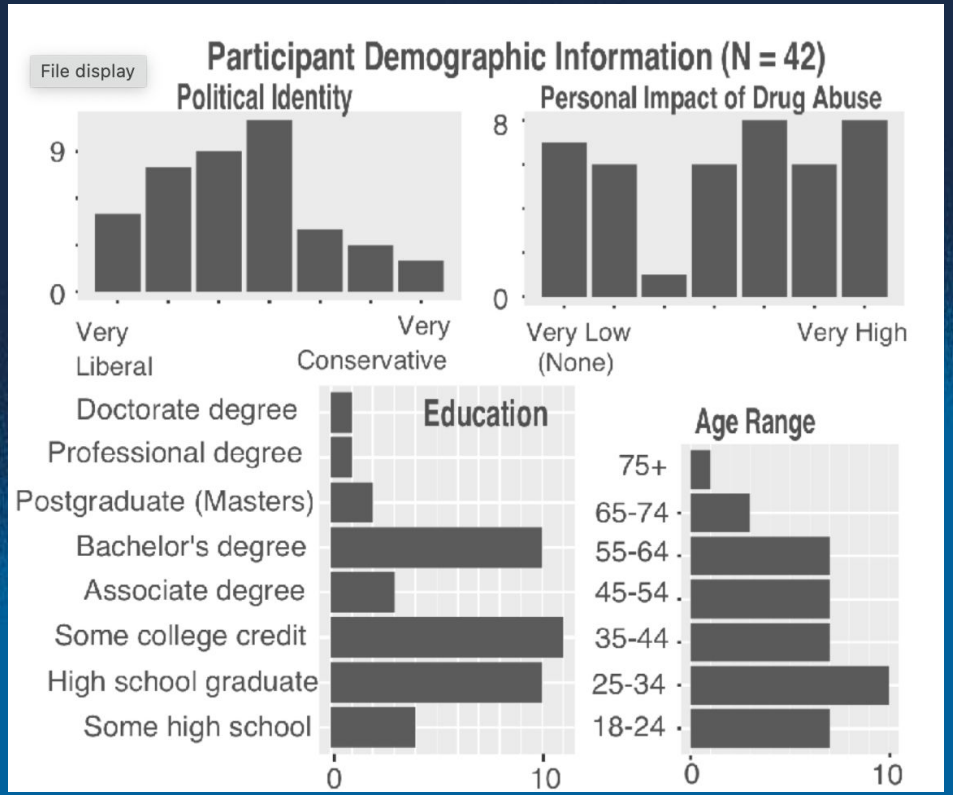
#	Topic	Type	Found on (Source)	Perceptions (Code Frequency)
A	File display cannabis vs. other drugs	Bar	National Institute on Drug Abuse (NIDA)	*Relatable*(4), *Informative*(2)
B	Comparison of drug, vehicle, and firearm deaths over time	Bar / Line	BreitBart	*Confusing*(2), *Informative*(2)
C	Drug use in 'street' youths vs. youths in households	Isotype	National Institute on Drug Abuse (NIDA)	*Simple*(3), *Not trusted*(3), *Clear*(2), *Relatable*(2)
D	Overdose deaths involving opioids by county	Map	The Economist	*Clear*(4), *Attractive*(3), *Confusing*(3), *Cluttered*(3), *Simple*(3), *Relatable*(3)
E	Opioid overdose prevention indicators for PA counties	Heat map	Drexel University	*Cluttered*(8), *Confusing*(8), *Clear*(4), *Colorful*(4), *Informative*(4)
F	Overdose increase from pain medication	Infographic	AgriMed (Medical Cannabis)	*Attractive*(5), *Confusing*(5), *Simple*(4)
G	Drug overdoses over time	Line	National Vital Statistics System (NVSS) - CDC	*Confusing*(6), *Simple*(3), *Cluttered*(2), *Intriguing*(2)
H	Overdose deaths by country (15-to-44-year olds)	Map	The New York Times	*Clear*(4), *Colorful*(3), *Relatable*(3), *Simple*(3)
I	Overdose death rates over time	Line	Business Insider	*Colorful*(16), *Attractive*(6), *Clear*(6), *Simple*(5)
J	The science of drug abuse	Infographic	Alternatives in Treatment (Rehab Center)	*Informative*(4), *Attractive*(3), *Relatable*(3)

Table 1: Graphs were chosen for representing diverse styles and sources. Codes are derived from interviews. When interpreting frequencies, recall that many participants chose to only comment on a select group of graphs

Procedure

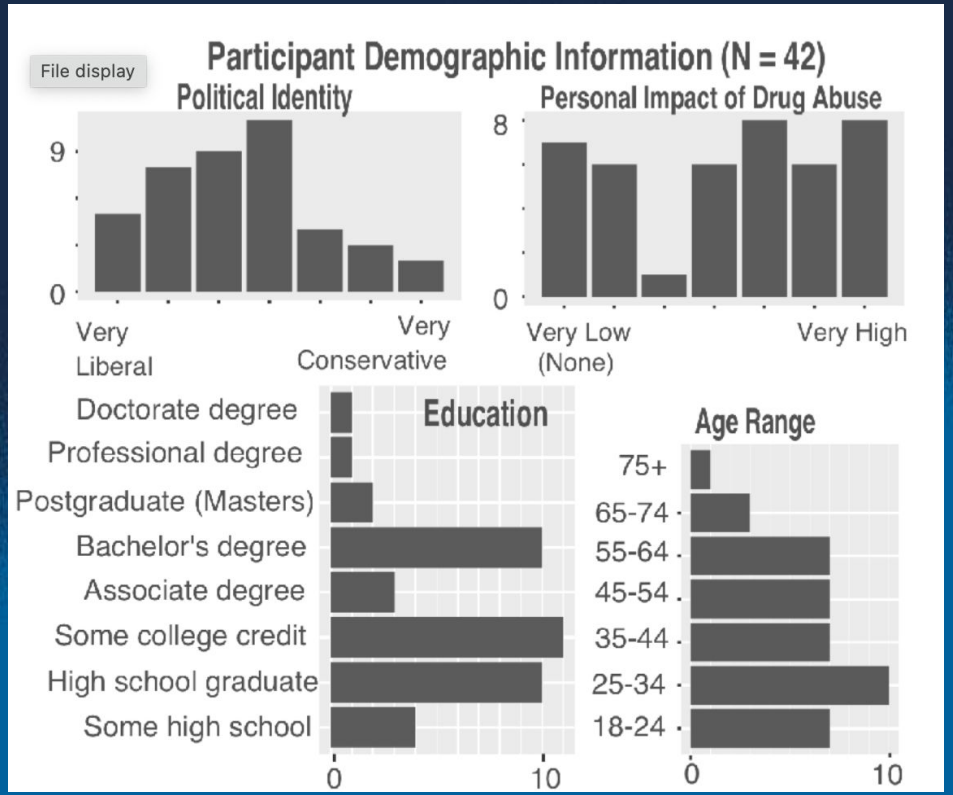
Participants

- Staff members at a local university. Participants largely identified as working in food services as cashier, line server, prep kitchen, or management
- Employees at a local construction site. Participants largely identified as working in demolition or labor
- Visitors of a local farmers market. Participates were diverse in their backgrounds and occupations



Procedure

- Age
- School district
- Political affiliation (“very liberal”(1) to “very conservative” (7))
- Familiarity with graphs and charts
- Educational background
- The extent to which they had been personally impacted by drugs and/or addiction

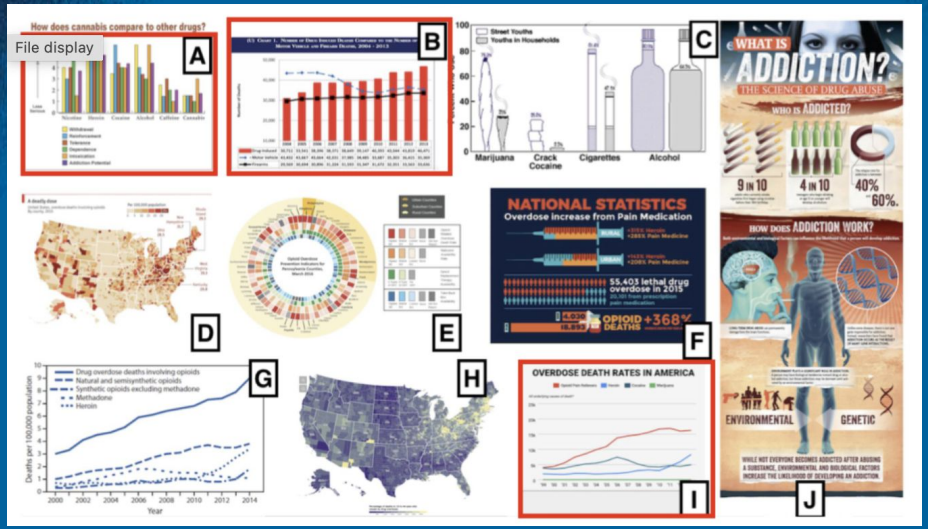


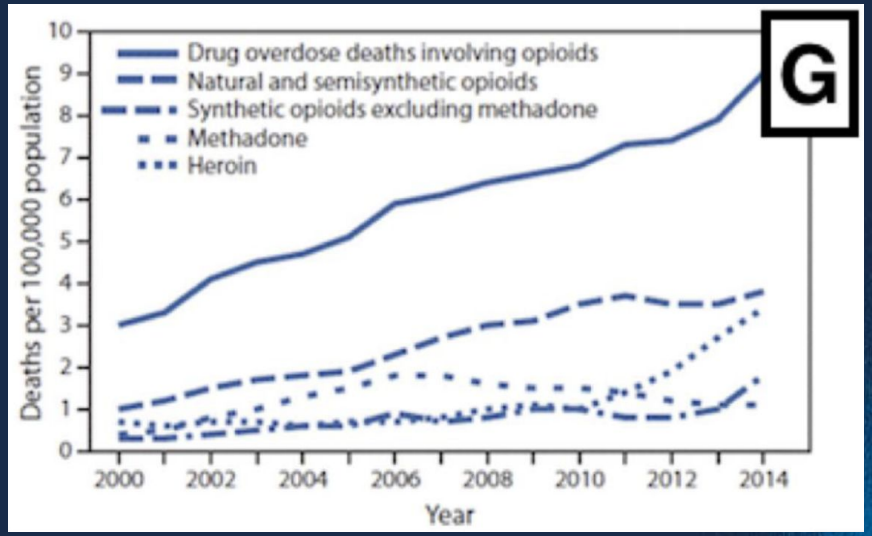
Procedure

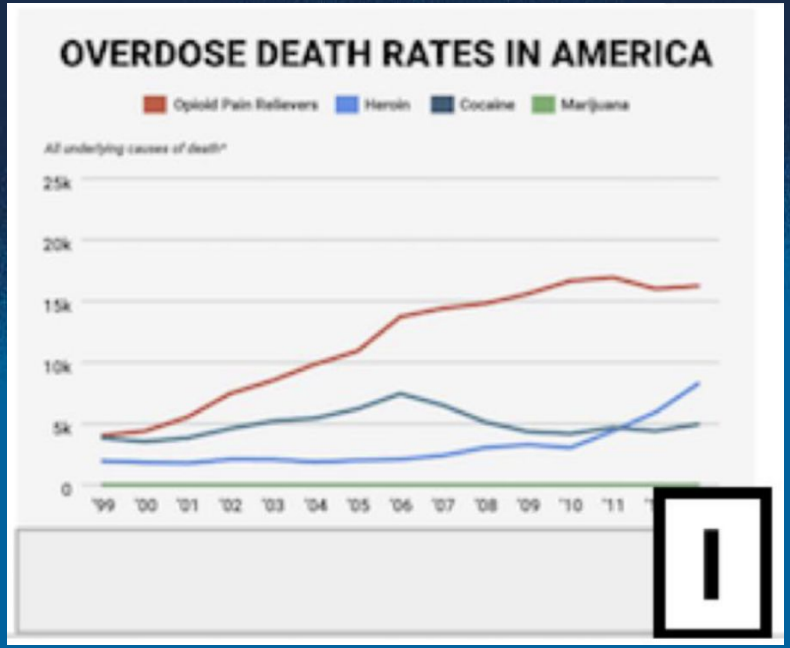
1. Introduction and consent
2. Graphs presentation and ranking
    a. “Based on how useful they are to you, arrange the graphs from most useful to least useful”
    b. “Useful” was successful in encouraging the participants to express opinions
3. Sources are revealed
4. Demographics questions (collected after the interview)

Analysis

- The most common codes associated with graphs across our interviews are: colorful(29), confusing(29), clear(26), simple(26), relatable(21), attractive(20), informative(19), cluttered(17)
- Gravitated towards straightforward visual encodings
- Simple bar graphs (graphs A, B) and line graphs (Graph I) emerged as among our more highly ranked charts

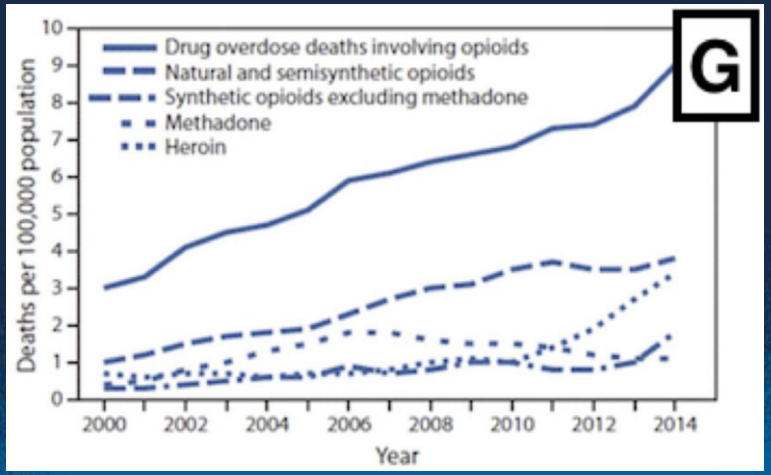


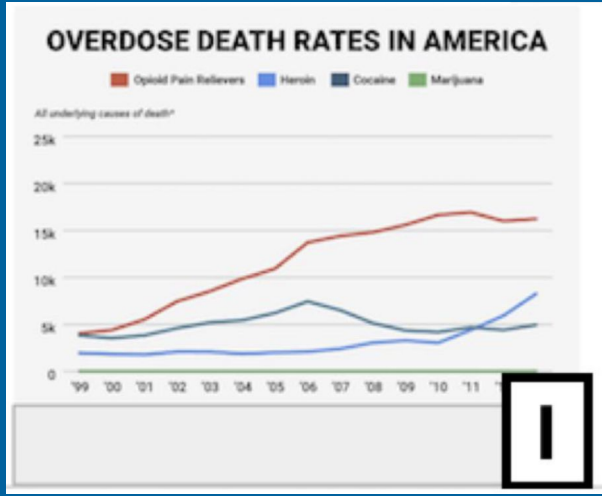

G
Drug overdose deaths involving opioids
Natural and semisynthetic opioids
Synthetic opioids excluding methadone
Methadone
Heroin
Deaths per 100,000 population
Year
0
1
2
3
4
5
6
7
8
9
10
2000
2002
2004
2006
2008
2010
2012
2014


OVERDOSE DEATH RATES IN AMERICA
Opioid Pain Relievers
Heroin
Cocaine
Marijuana
All underlying causes of death*
25k
20k
15k
10k
5k
0
'99
'00
'01
'02
'03
'04
'05
'06
'07
'08
'09
'10
'11
I

Graph G and I

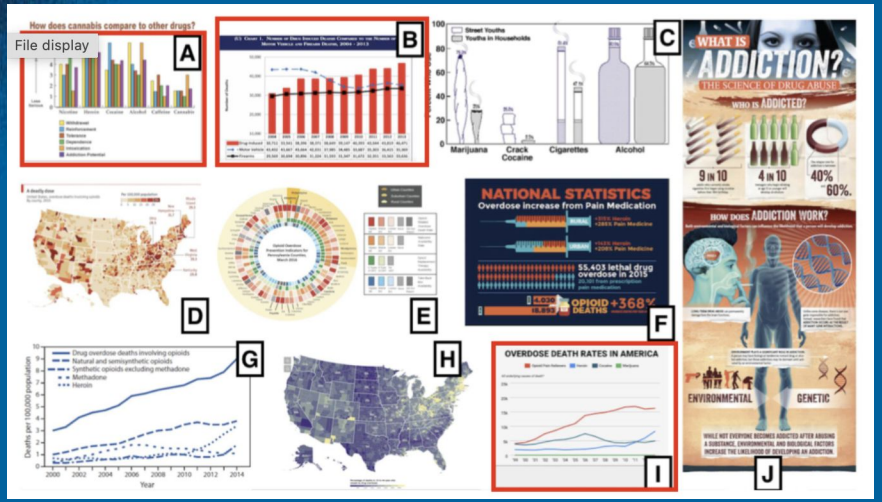
- Critiques of clarity and aesthetics often blurred together for our participants
- 16 participants identified color as a distinguishing factor
- Often ambiguous as to whether color referenced general appeal or an improved visual encoding





Infographics

- Graph J received the most polarizing rankings of any chart
- Participants who had positive feelings about infographics (Graphs F and J) found them to be clear(5), simple(5), and attractive(8)
- Infographics were often rated lower by older people



Unchanged Ranking

- Source is irrelevant(9): expressed that the source does not impact the data and/or presentation
- Ranked on other criteria(5): expressed that their initial ranking was based on other criteria(visuals, interest) and that criteria had not changed
- No reason(4) could not (or was not willing to_ articulate any reason for maintaining their rankings
- All sources are trusted(3): perceived that all sources were equally trustworthy

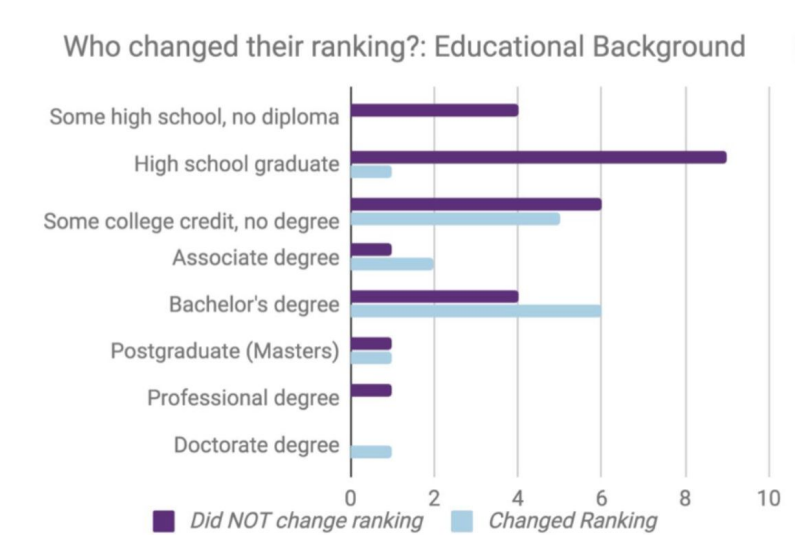


Figure 5: More educated participants were more likely to change their rankings after seeing the graph's source

Assignment 2

P-value

Credits:

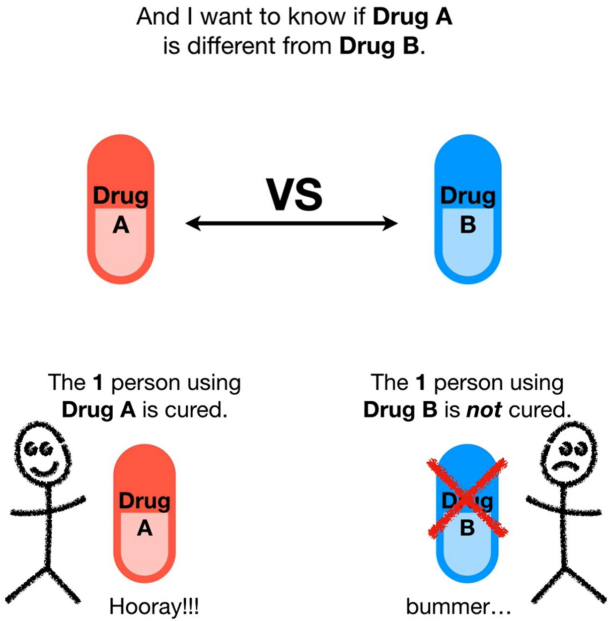
StatQuest YouTube Videos on P-Value

p-values: What they are and how to interpret them

How to calculate p-value

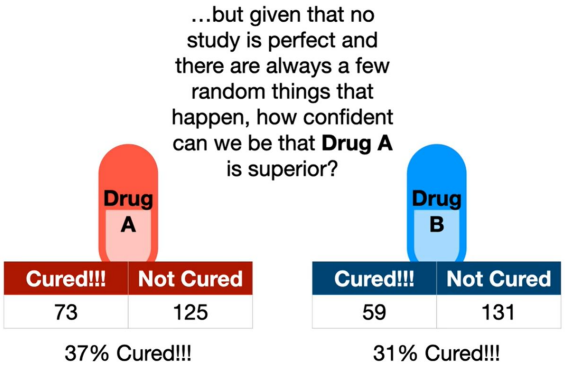
p-hacking: What it is and how to avoid it!

P-Value: Intuition



Can we conclude that A is better than B?

No because there can be many reasons why B didn't work on that 1 person

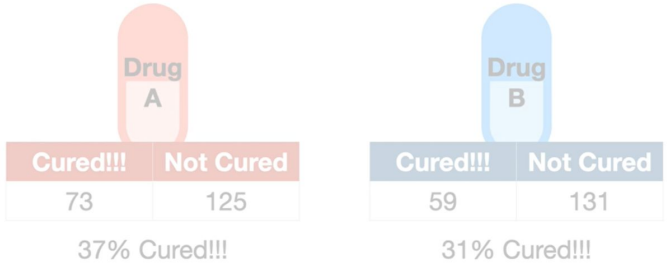


P-value is used for this purpose.

P-value is between 0 and 1. It quantifies how confident are we that A is different from B

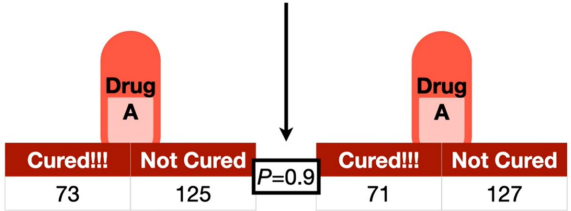
Close the P-value is to 0, the more confident we are that A and B are different

In practice, a commonly used threshold is **0.05**. It means that if there is no difference between **Drug A** and **Drug B**, and if we did this exact same experiment a bunch of times, then only **5%** of those experiments would result in the wrong decision.



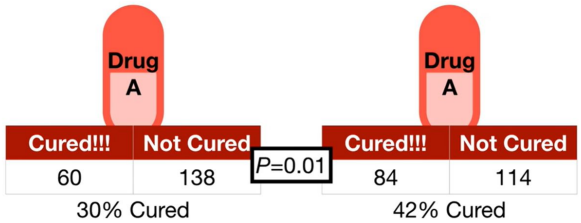
To see how p-value of 0.05 means 5% chance of error, let's give same drug to 2 groups

Thus, we would say that we fail to see a difference between the two groups.



psst!!! The p-value was calculated using **Fisher's Exact Test**. To learn more, the link is in the description below.

As a result, the **p-value** for this specific run of the experiment is **0.01**, since the results are pretty different.



By chance, Group B can get more people with placebo effect

Statistic Lingo

In fancy statistical lingo, the idea of trying to determine if these drugs are the same or not is called **Hypothesis Testing**.

Drug A

Cured!!!	Not Cured
73	125

37% Cured!!!

Drug B

Cured!!!	Not Cured
59	131

31% Cured!!!

The **Null Hypothesis** is that the drugs are the same…

…and the **p-value** helps us decide if we should reject the **Null Hypothesis** or not.

Drug A

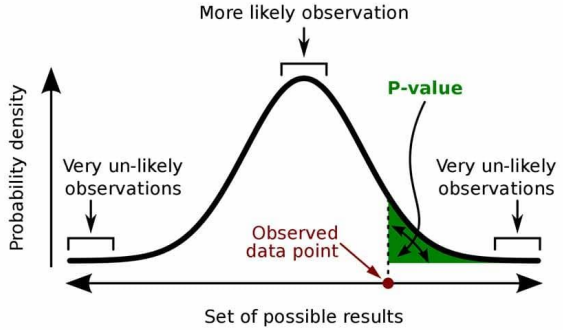
Cured!!!	Not Cured
73	125

37% Cured!!!

Drug B

Cured!!!	Not Cured
59	131

31% Cured!!!



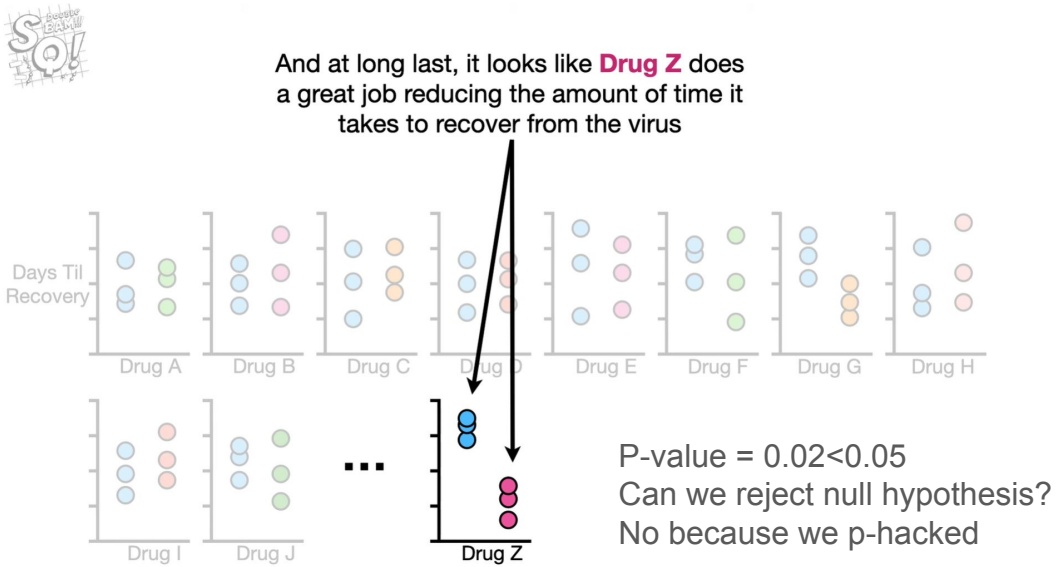
A **p-value** (shaded green area) is the probability of an observed (or more extreme) result assuming that the null hypothesis is true.

P-hacking

P-hacking refers to the misuse and abuse of analysis techniques and results in being fooled by false positives.

We developed drugs to reduce time to recovery for a virus. Now we want to test which of these is effective. We give people drugs and compare with group that did not get drugs and try to find one that works.

No Drug



P-value = 0.02<0.05
Can we reject null hypothesis?
No because we p-hacked

Hypothesis Testing and EDA

Let’s go to Python