[COGS 9] Discussion
Reading 4, EDA, Inferential Analysis

Reading Quiz 4 due on May 17th (Fri)

Assignment 2 due on May 20th (Mon)

Graphical Inference for Infovis

Is what we see really there?

What is inference and why do we need it?

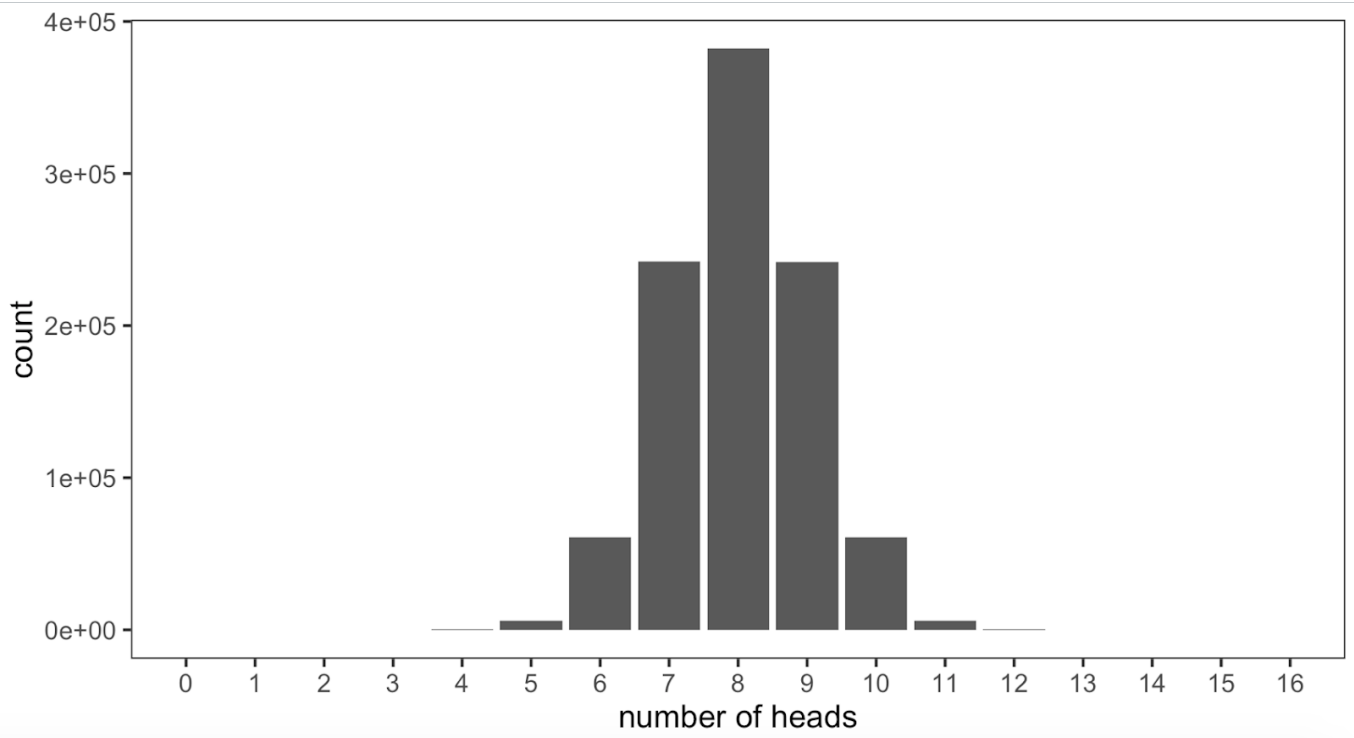
- Statistical inference is the process of drawing conclusions about a population based on a sample of data.
- Estimation involves using sample data to estimate the value of a population parameter, such as the mean or proportion.
- Hypothesis testing involves making decisions about whether a particular hypothesis is supported by the data, based on a set of statistical criteria and a chosen level of significance.

P-value

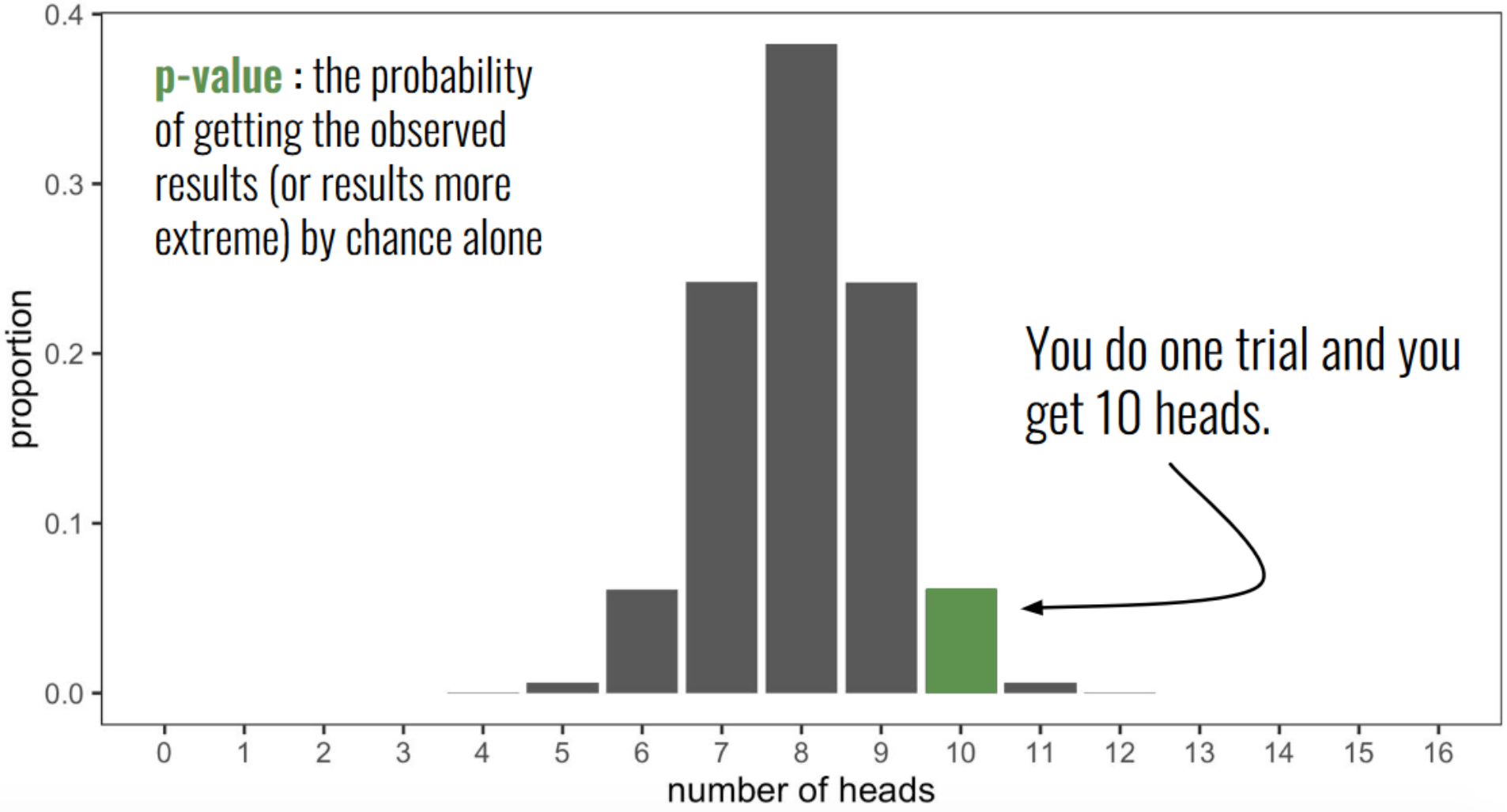
P-value = Probability of obtaining results at least as extreme as the observed results of a statistical hypothesis test, assuming that the null hypothesis is correct

Assuming a fair coin

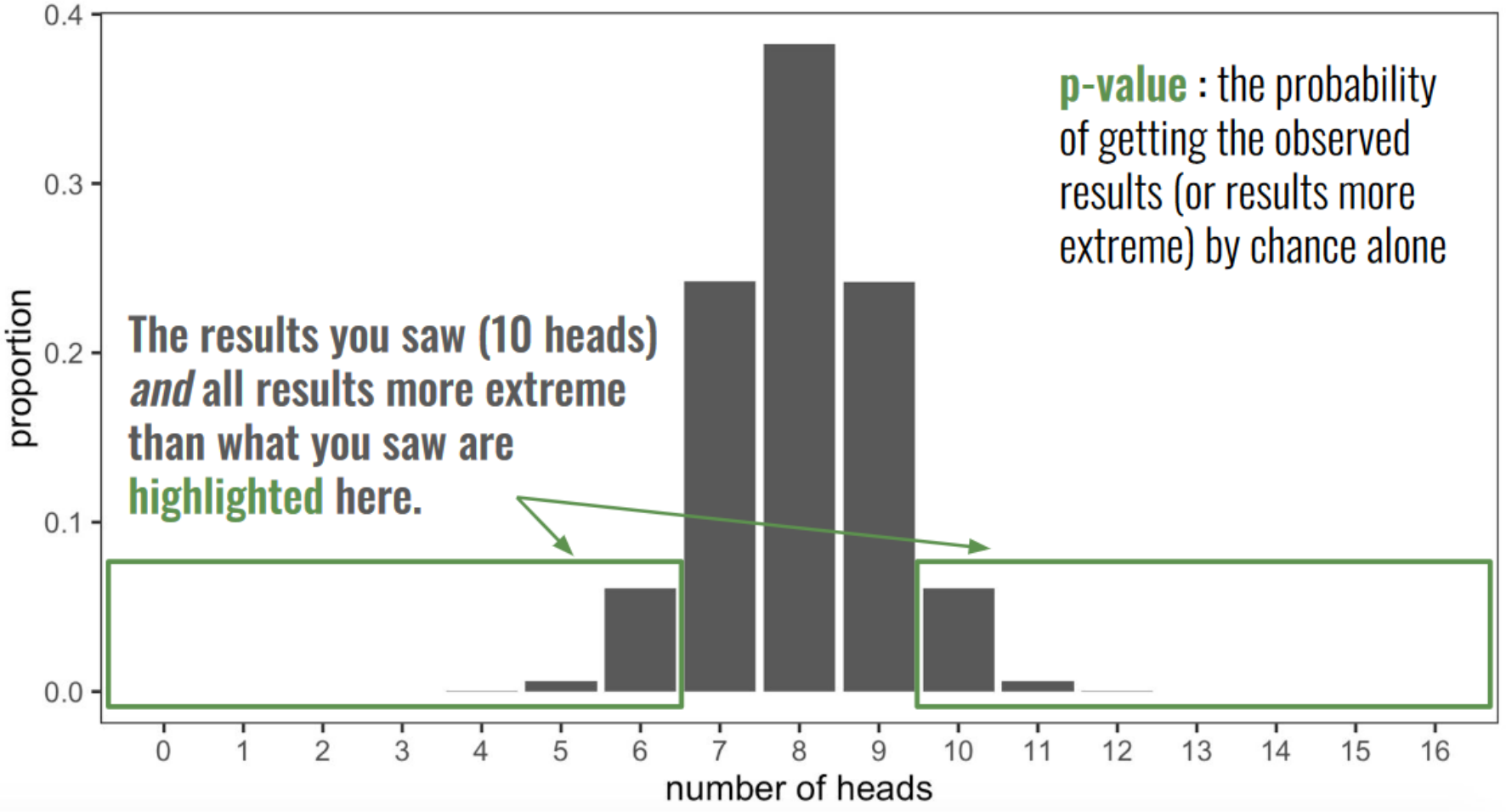
If we flip a coin 16 times and record the number of heads and then repeat this 16 flip trial 1 million times



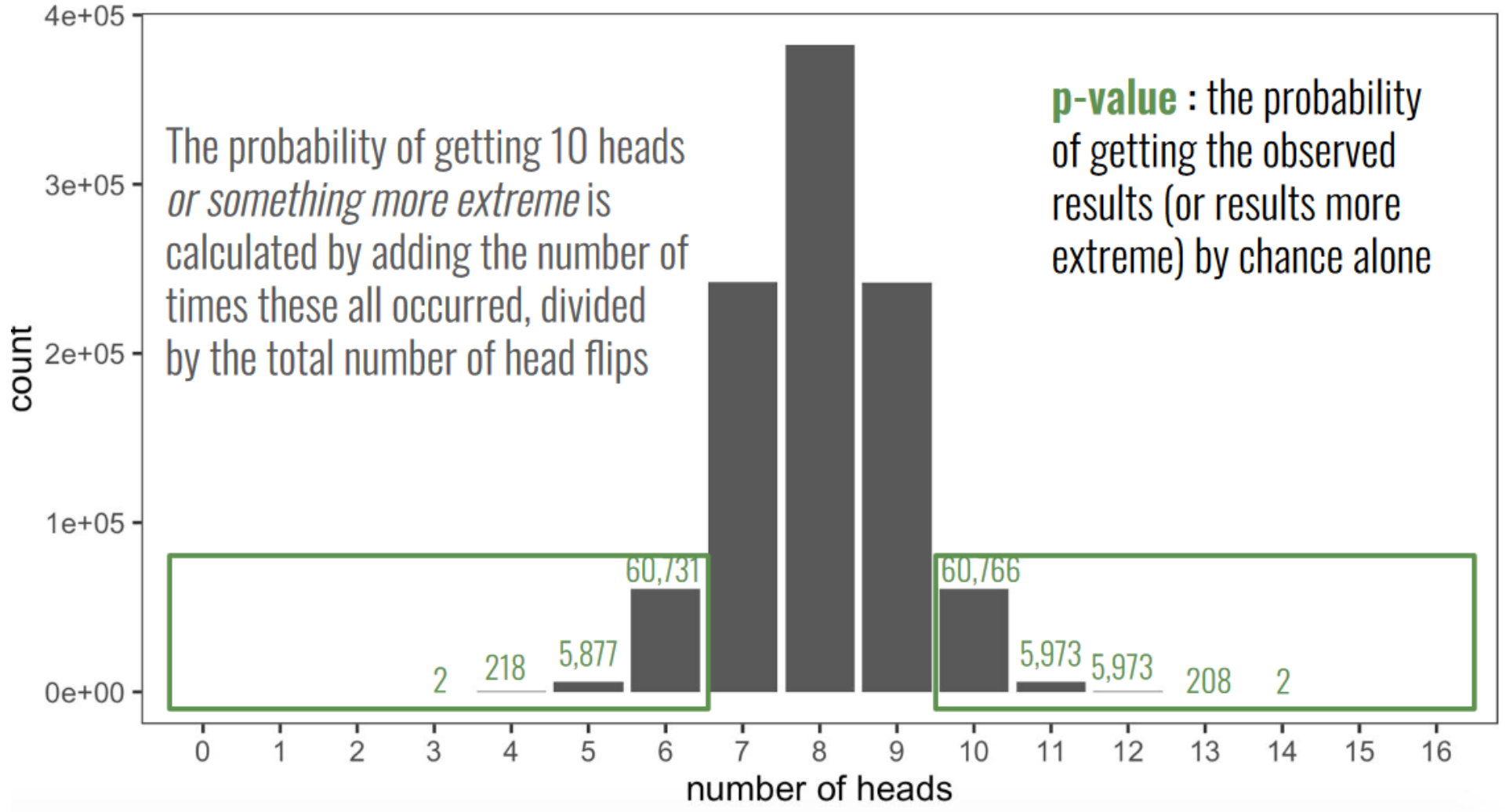
P-value


p-value : the probability of getting the observed results (or results more extreme) by chance alone
You do one trial and you get 10 heads.
proportion
0.4
0.3
0.2
0.1
0.0
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
number of heads

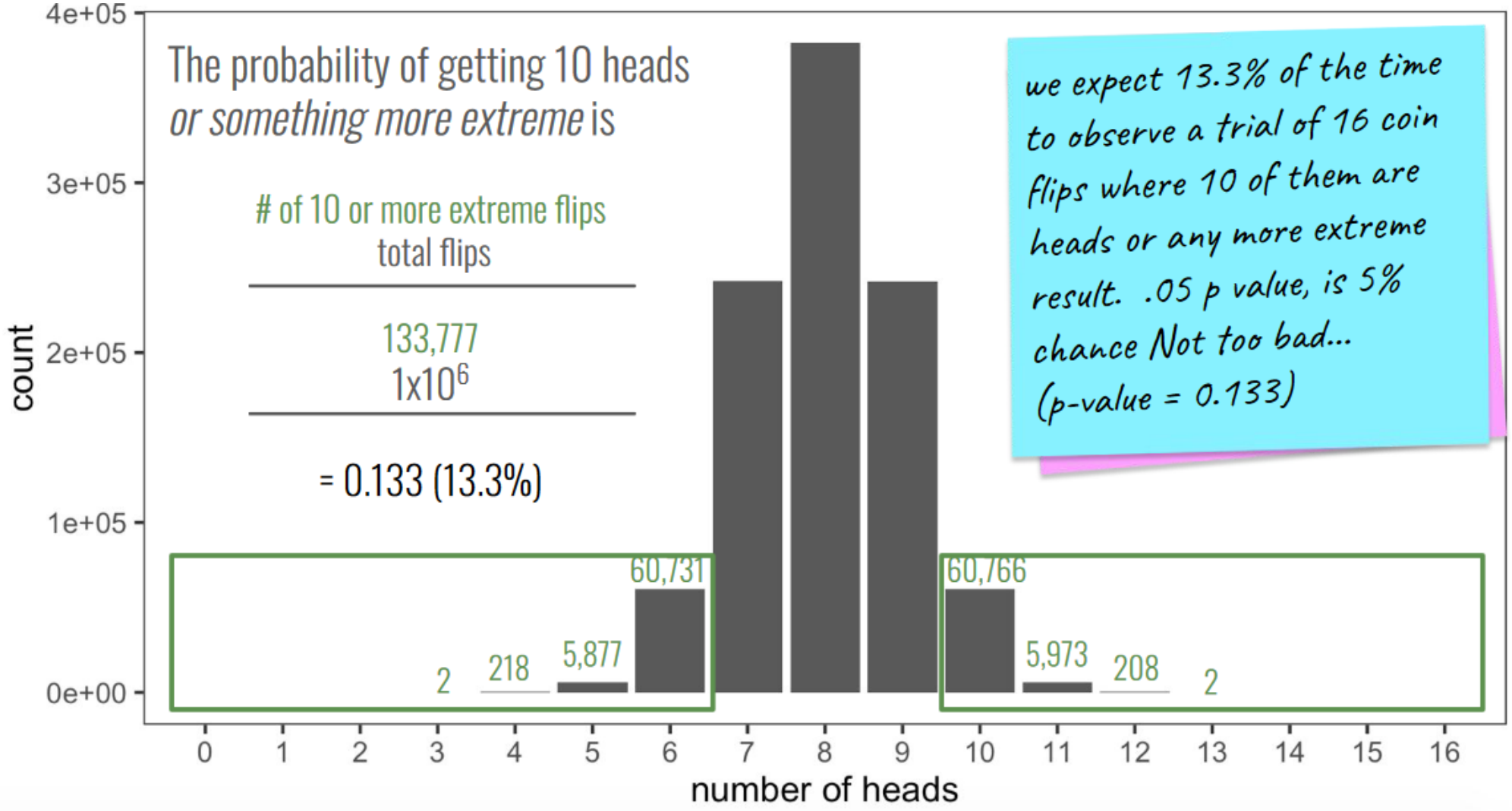
P-value


0.4
0.3
0.2
0.1
0.0
proportion
p-value : the probability of getting the observed results (or results more extreme) by chance alone
The results you saw (10 heads) *and* all results more extreme than what you saw are highlighted here.
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
number of heads

P-value


4e+05
3e+05
2e+05
1e+05
0e+00
count
The probability of getting 10 heads *or something more extreme* is calculated by adding the number of times these all occurred, divided by the total number of head flips
p-value : the probability of getting the observed results (or results more extreme) by chance alone
2
218
5,877
60,731
60,766
5,973
5,973
208
2
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
number of heads

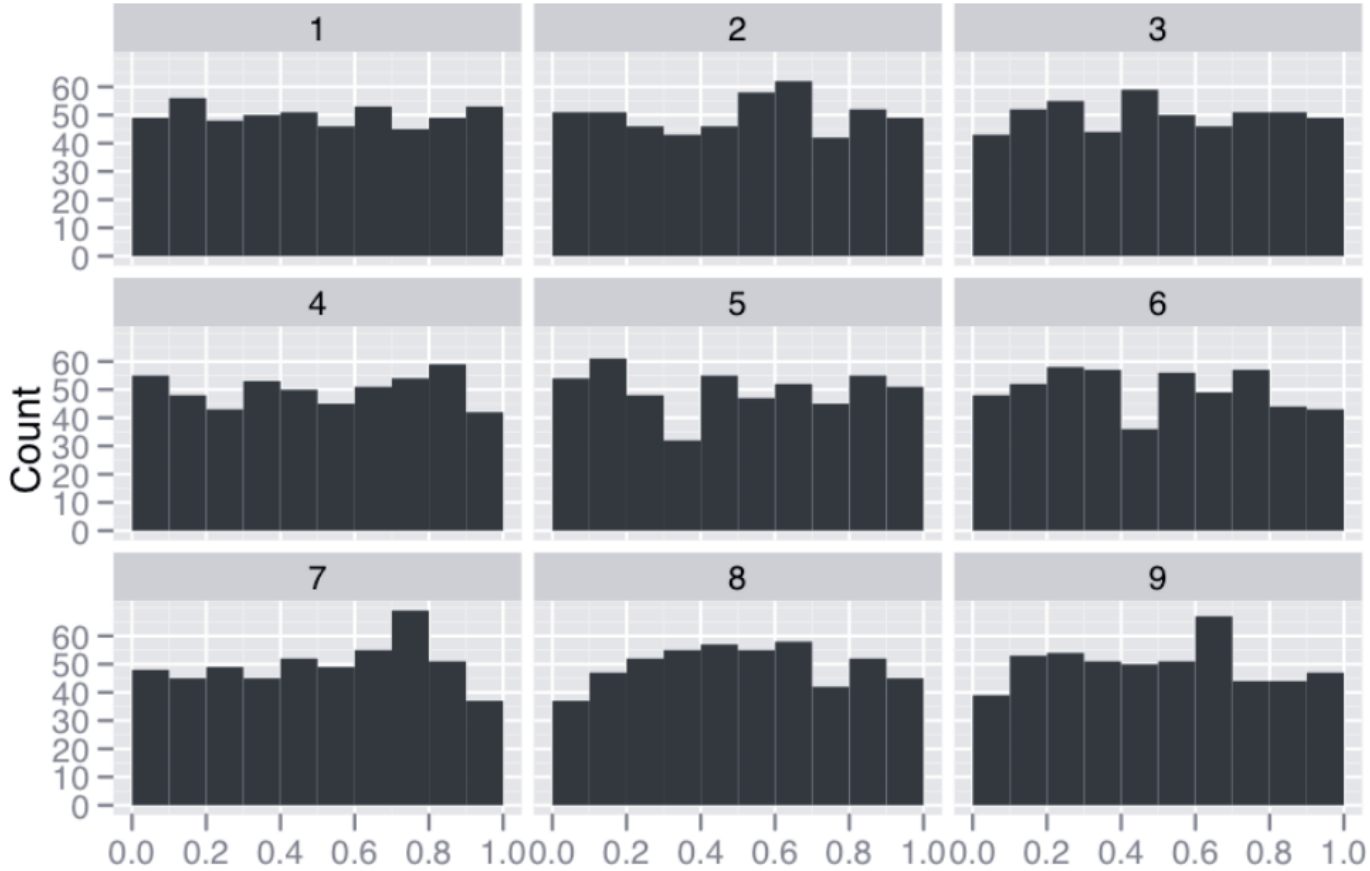
P-value


The probability of getting 10 heads
or something more extreme is
# of 10 or more extreme flips
total flips
133,777
1x10^6
= 0.133 (13.3%)
we expect 13.3% of the time to observe a trial of 16 coin flips where 10 of them are heads or any more extreme result. .05 p value, is 5% chance Not too bad... (p-value = 0.133)
count
4e+05
3e+05
2e+05
1e+05
0e+00
2
218
5,877
60,731
60,766
5,973
208
2
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
number of heads

Protocols of Graphical Inference

1) Rorschach: a calibrator, helping the analyst become accustomed to the vagaries of random.
2) Line-up (works like a police line-up): the suspect (test statistic plot) is hidden in a set of decoys. If the observer, who has not seen the suspect, can pick it out as being noticeably different, there is evidence that it is not innocent.

Misleading due to patterns in random noise



Example of Rorschach Protocol

To use the line-up protocol:

1) Identify the question the plot is trying to answer.
2) Characterize the null-hypothesis.
3) Figure out how to generate null datasets.

Selected visualizations in terms of their purpose and associated null distributions

1) Tag clouds: a visual representation of text data. It typically consists of a collection of tags, or keywords, that are displayed in different sizes or colors based on their frequency or importance.
2) Scatterplot: displays the relationship between two continuous variables, and answers the question: are x and y related in some way? The scatterplot can reveal many different types of relationships, e.g., linear trends, non-linear relationships and clustering.

Example of tag clouds

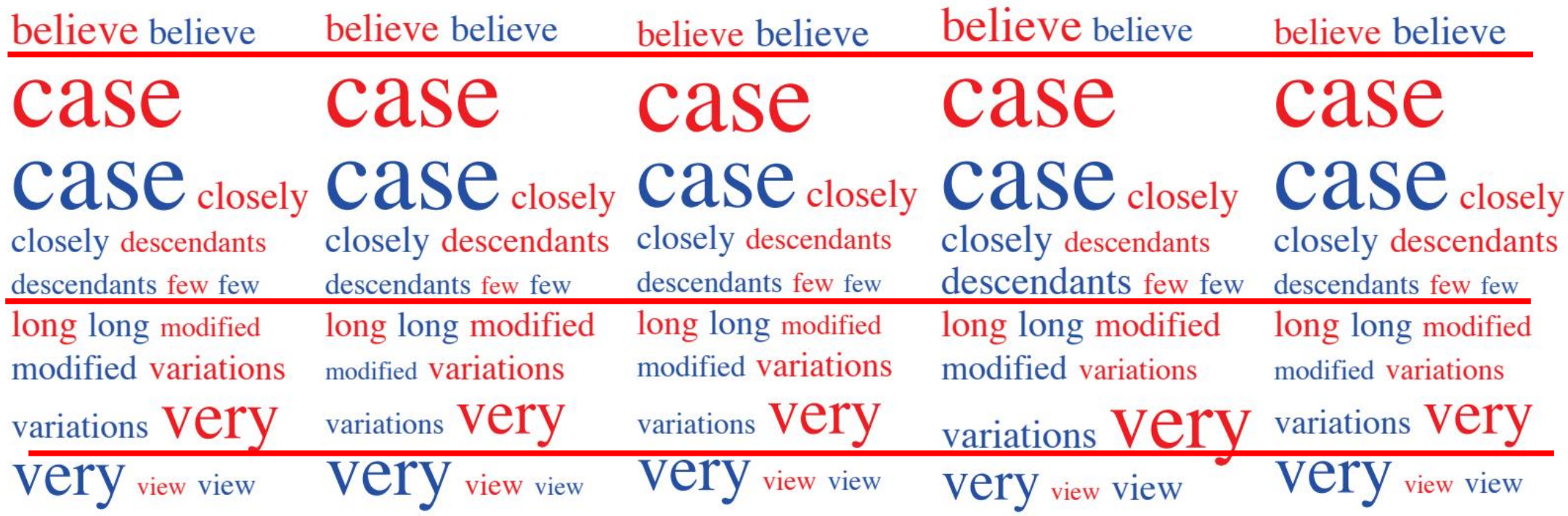


Fig. 5. Five tag clouds of selected words from the 1st (red) and 6th (blue) editions of Darwin's "Origin of Species". Four of the tag clouds were generated under the null hypothesis of no difference between editions, and one is the true data. Can you spot it?

Example of scatterplot

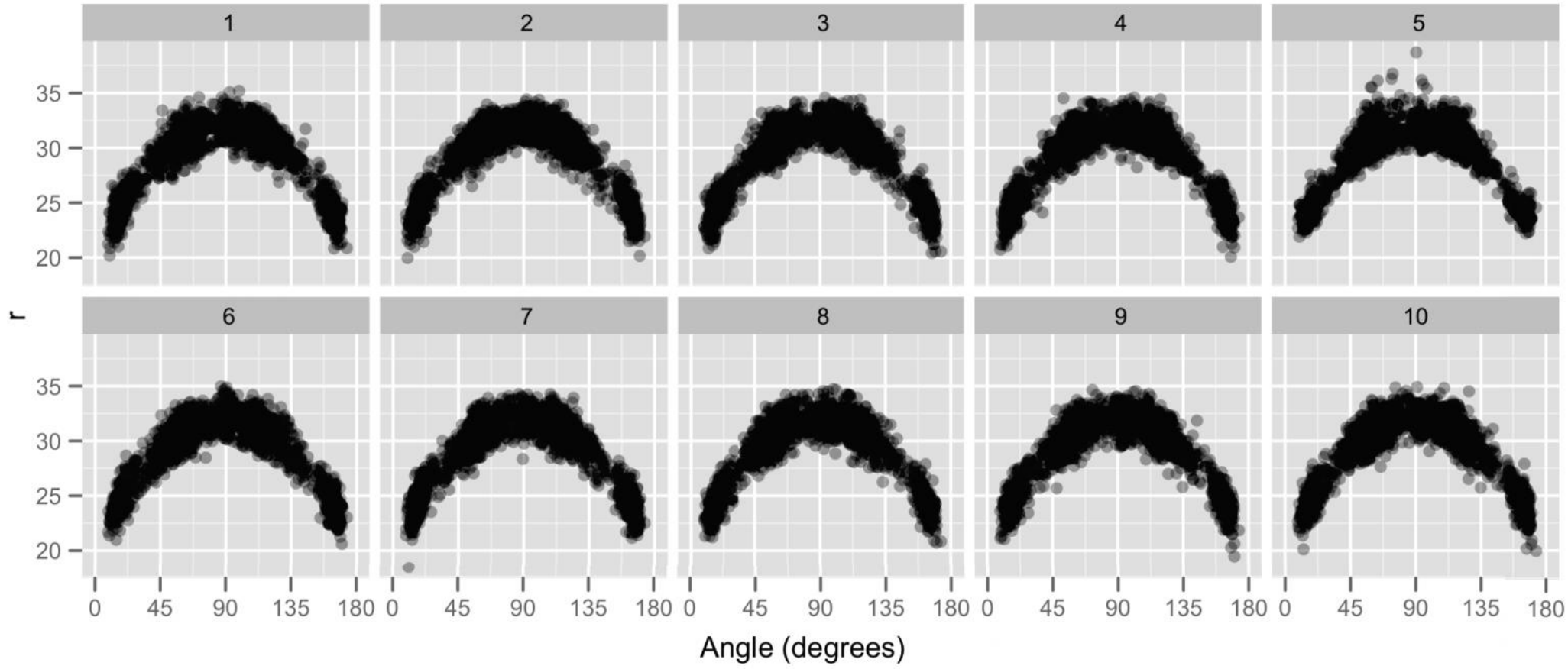


Fig. 6. Scatterplot of distance vs. angle for three pointers by the LA Lakers. True data is concealed in line-up of nine plots generated under the null hypothesis that there is a quadratic relationship between angle and distance.

The Power of Graphical Tests

- The probability of correctly convicting a guilty dataset. The capacity to detect specific structure in plots can depend on many things, including an appropriate choice of plot.
- The ability of graphical methods to detect patterns, trends, and differences in data that may not be apparent through traditional statistical tests.

Conclusion

- Rorschach and line-up protocols bring rigorous statistical inference to freeform data exploration.
- Both techniques center around identifying a null hypothesis, which then generates null datasets and null plots.
- The Rorschach provides a tool for calibrating our expectations of null data, while the line-up brings the techniques of formal statistical hypothesis testing to visualization.

Data is Personal

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 United States
- 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.

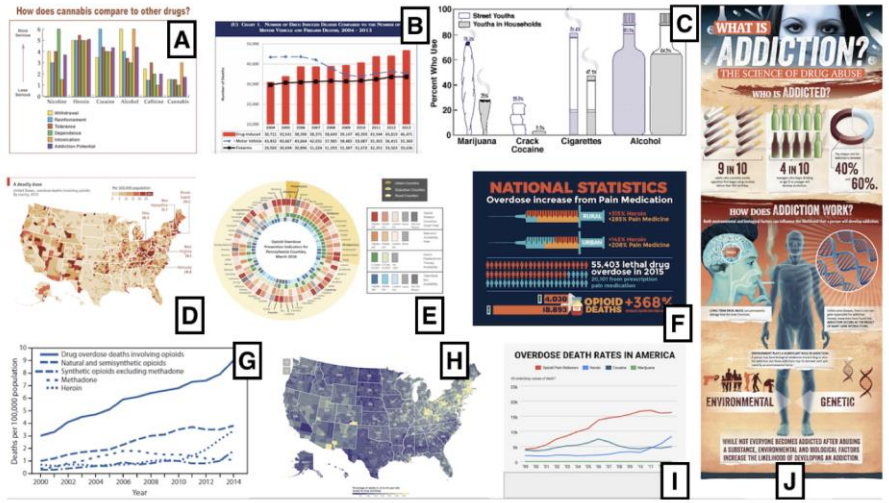


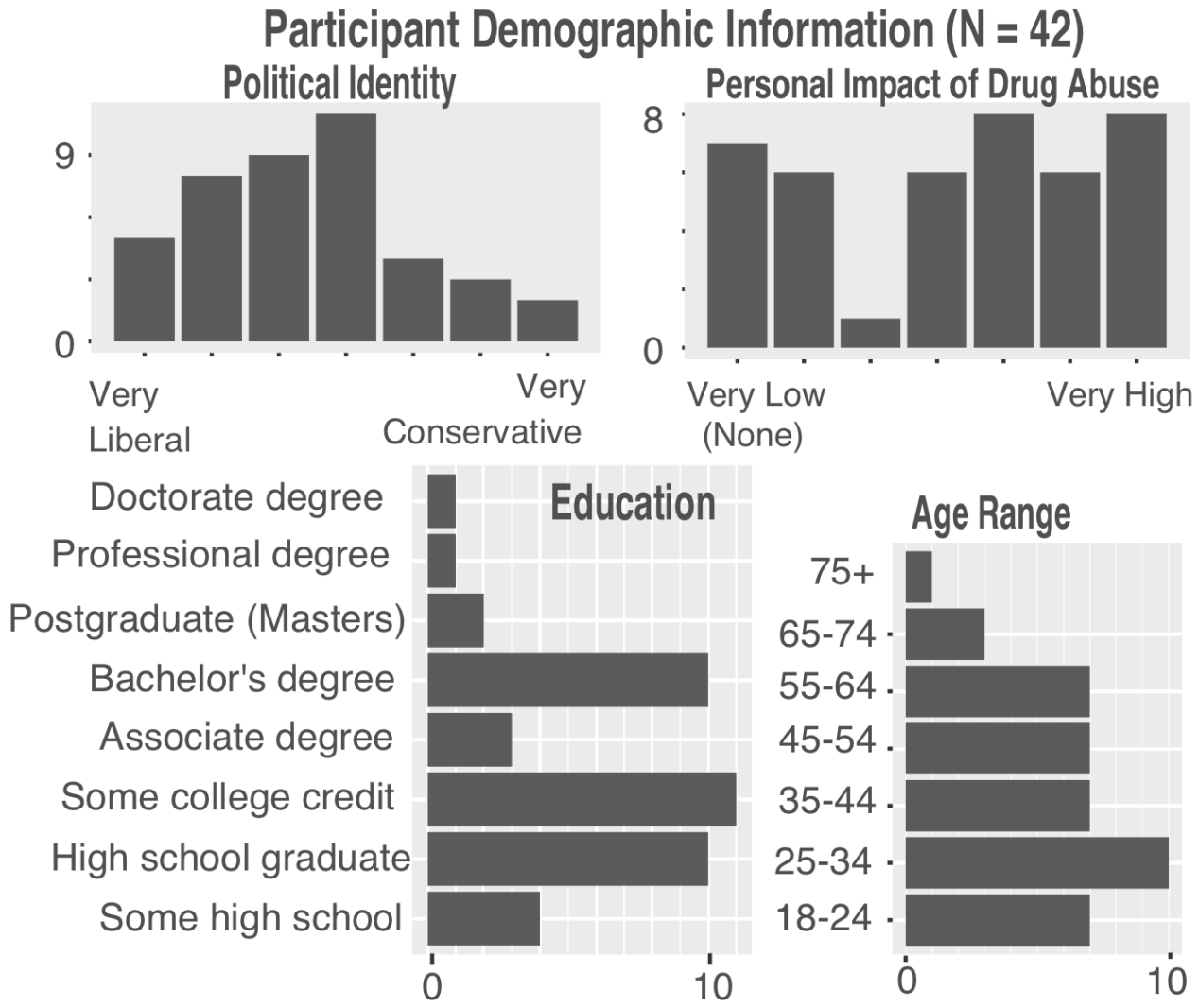
Figure 4: The graphs shown to participants. Each graph was presented on an independent sheet of paper

#	Topic	Type	Found on (Source)	Perceptions (Code Frequency)
A	Severity of 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

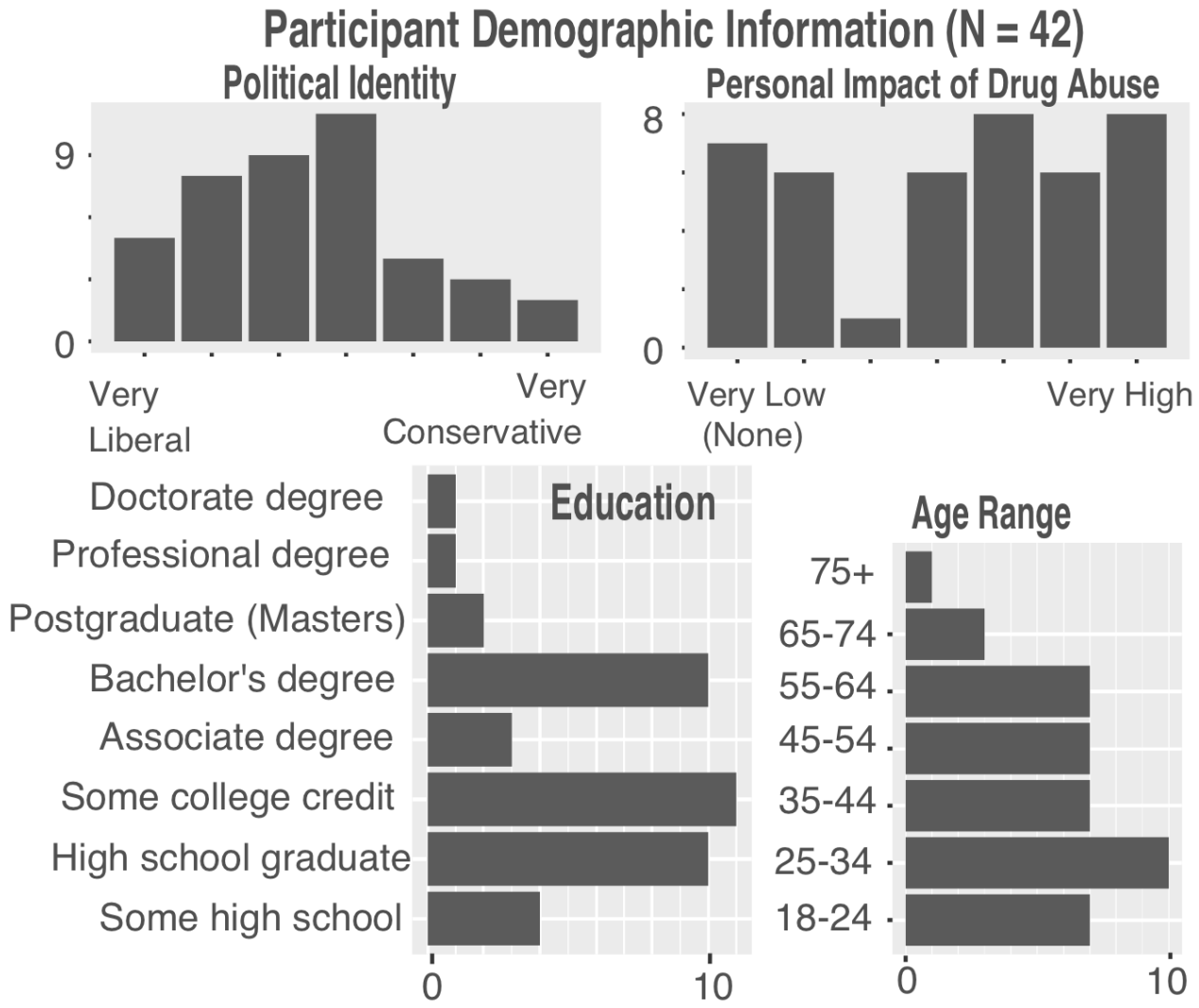
Procedure

- 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. Participants were diverse in their backgrounds and occupations.



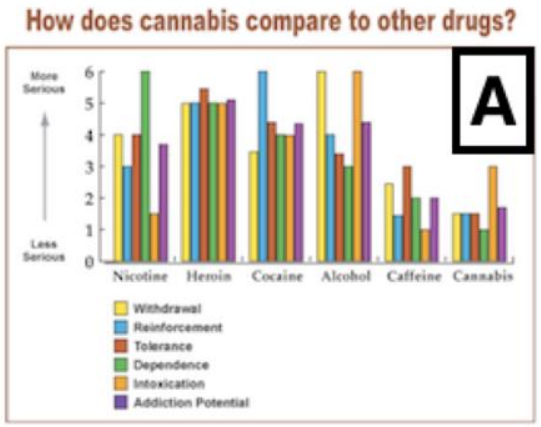
Procedure

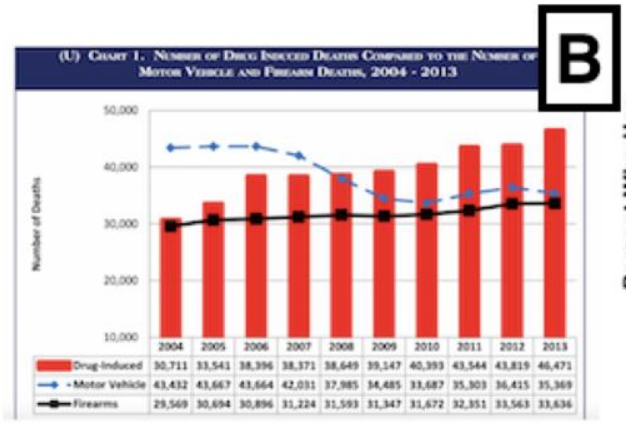
- Age
- School district,
- Political affiliation (“very liberal”(1) to “very conservative”(7))
- Familiarity with graphs and charts
- Educational background
- The extent to which they had been personally impacted by drugs and/or addiction

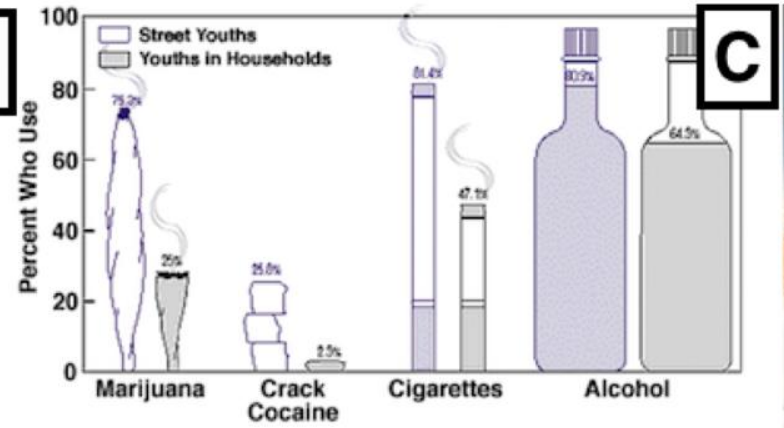


Procedure

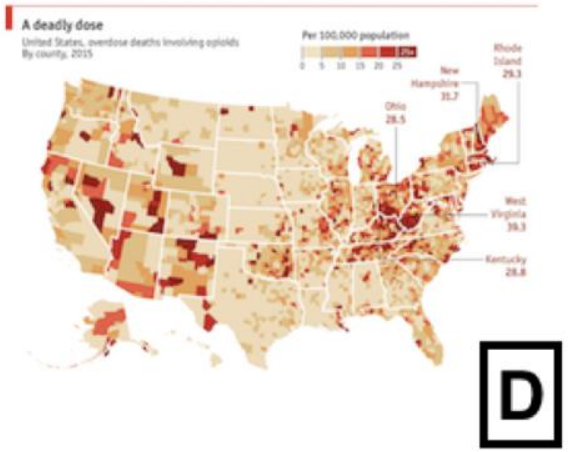
- Introduction and consent.
- Graphs presentation and ranking.
  - “Based on how useful they are to you, arrange the graphs from most useful to least useful”
  - ‘useful’ was successful in encouraging the participants to express opinions
- Sources are revealed
- Demographics questions (collected after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


How does cannabis compare to other drugs?
More Serious
Less Serious
Nicotine
Heroin
Cocaine
Alcohol
Caffeine
Cannabis
Withdrawal
Reinforcement
Tolerance
Dependence
Intoxication
Addiction Potential
A

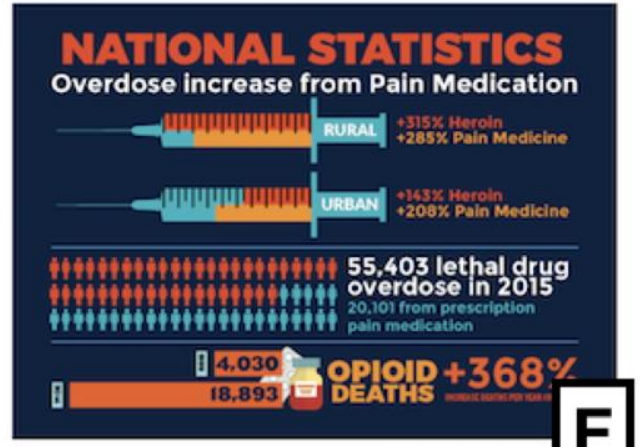

(U) Chart 1. Number of Drug Induced Deaths Compared to the Number of Motor Vehicle and Firearm Deaths, 2004 - 2013
Number of Deaths
Drug-Induced
Motor Vehicle
Firearms
B

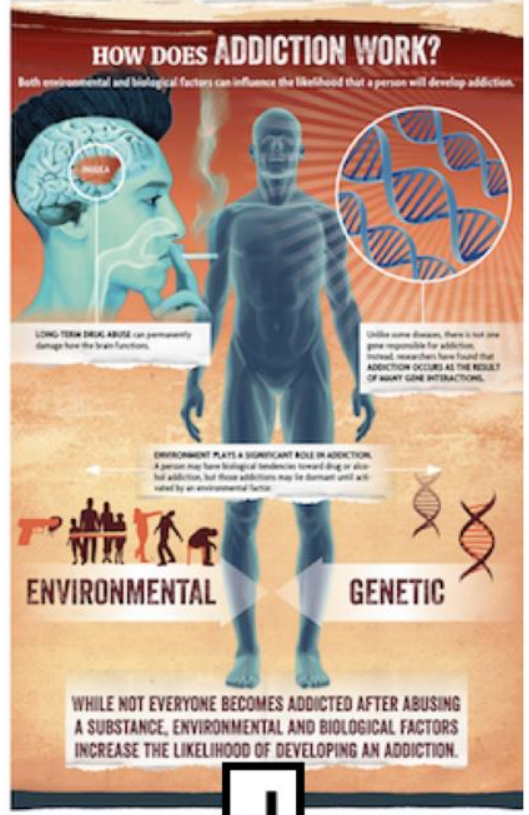

Street Youths
Youths in Households
Percent Who Use
Marijuana
Crack Cocaine
Cigarettes
Alcohol
C

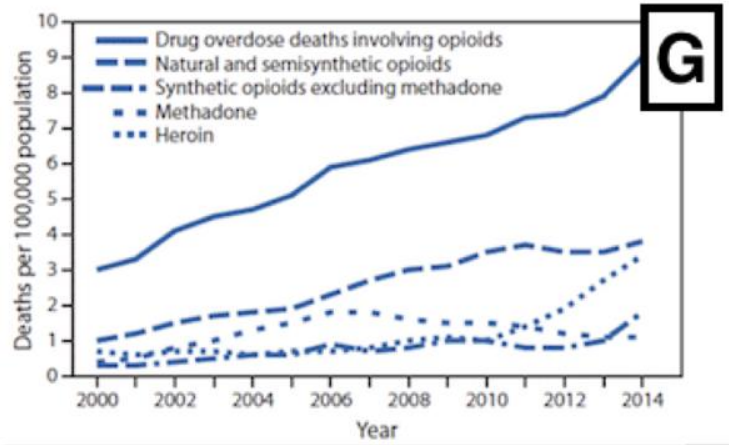

WHAT IS
ADDICTION?
THE SCIENCE OF DRUG ABUSE
WHO IS ADDICTED?
9 IN 10
4 IN 10
40% and 60%.

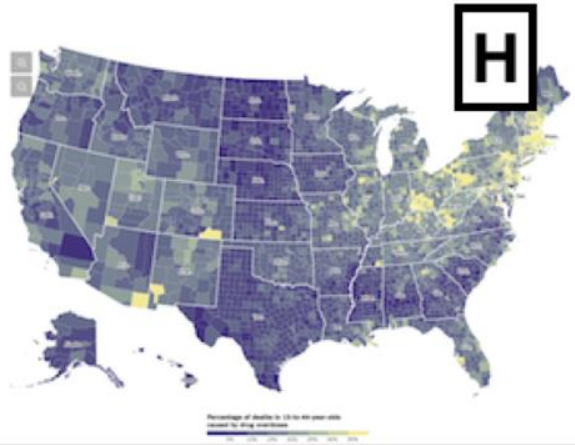

A deadly dose
Per 100,000 population
D

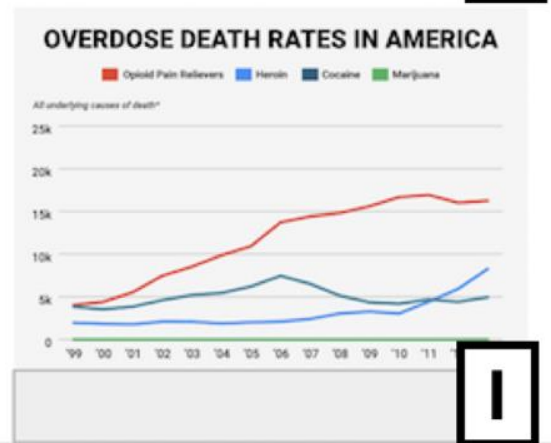

Opioid Overdose Prevention Indicators for Pennsylvania Counties, March 2016
Urban Counties
Suburban Counties
Rural Counties
E


NATIONAL STATISTICS
Overdose increase from Pain Medication
RURAL
+315% Heroin
+285% Pain Medicine
URBAN
+143% Heroin
+208% Pain Medicine
55,403 lethal drug overdose in 2015
20,101 from prescription pain medication
4,030
18,893
OPIOID DEATHS
+368%
F


HOW DOES ADDICTION WORK?
Both environmental and biological factors can influence the likelihood that a person will develop addiction.
ENVIRONMENTAL
GENETIC
WHILE NOT EVERYONE BECOMES ADDICTED AFTER ABUSING A SUBSTANCE, ENVIRONMENTAL AND BIOLOGICAL FACTORS INCREASE THE LIKELIHOOD OF DEVELOPING AN ADDICTION.
J

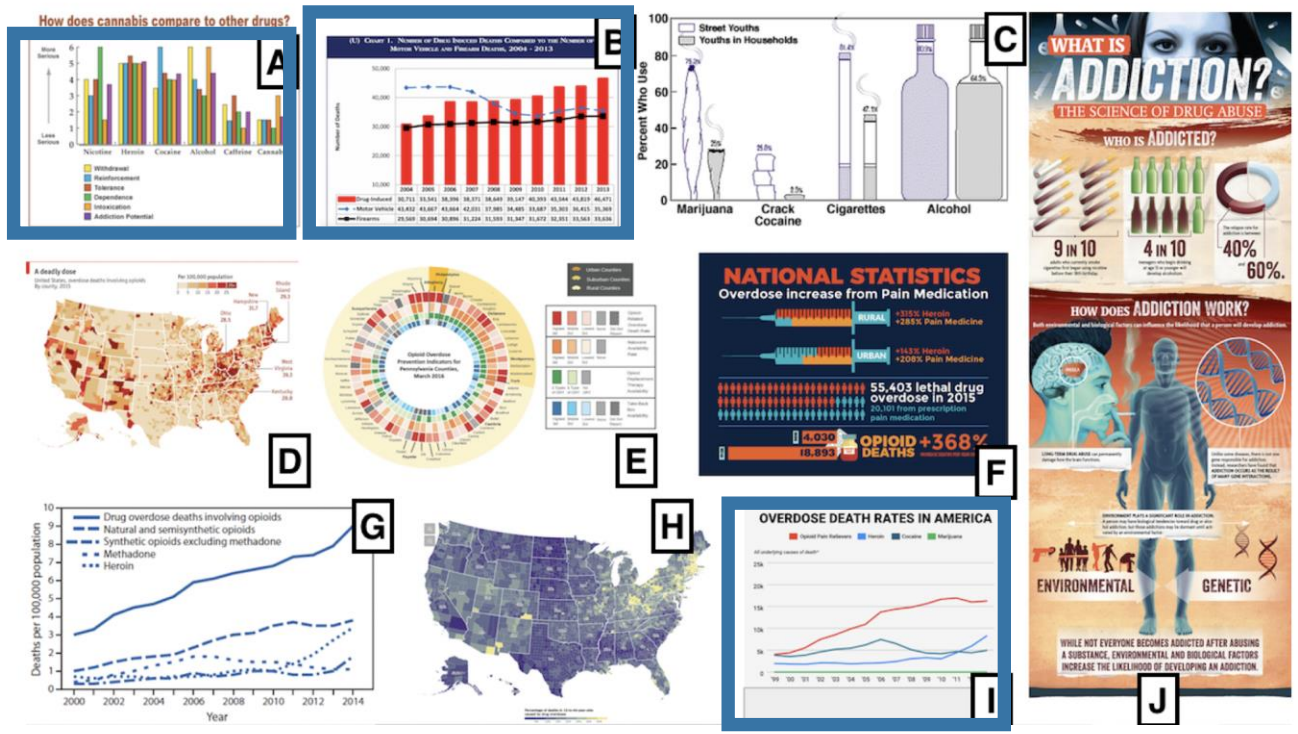

Drug overdose deaths involving opioids
Natural and semisynthetic opioids
Synthetic opioids excluding methadone
Methadone
Heroin
Deaths per 100,000 population
Year
G


H


OVERDOSE DEATH RATES IN AMERICA
Opioid Pain Relievers
Heroin
Cocaine
Marijuana
I

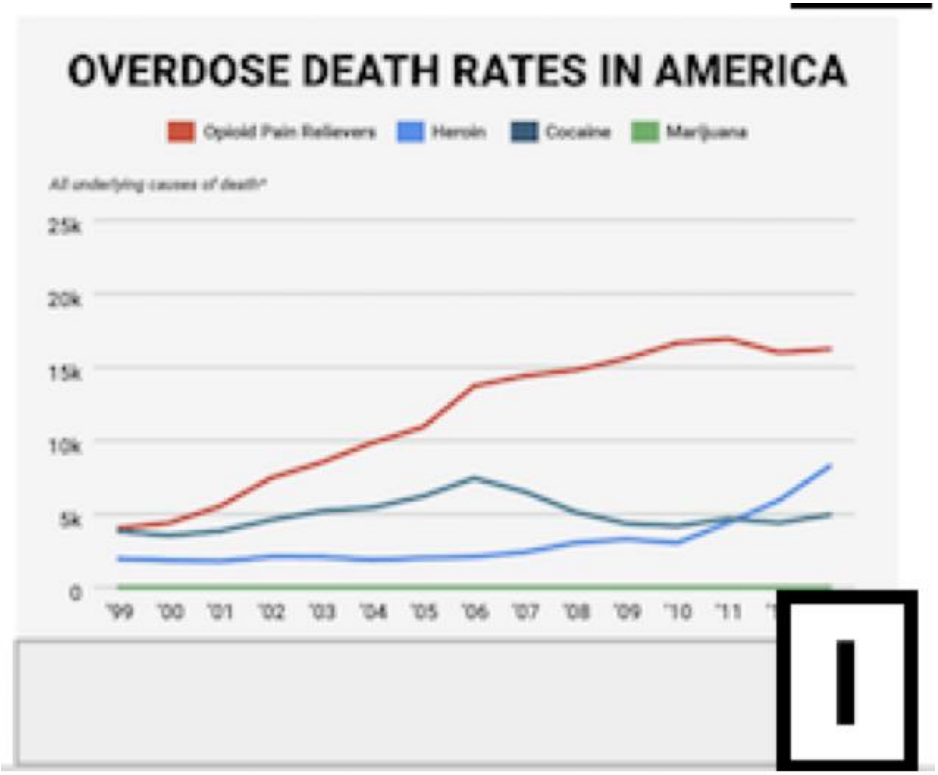
Analysis

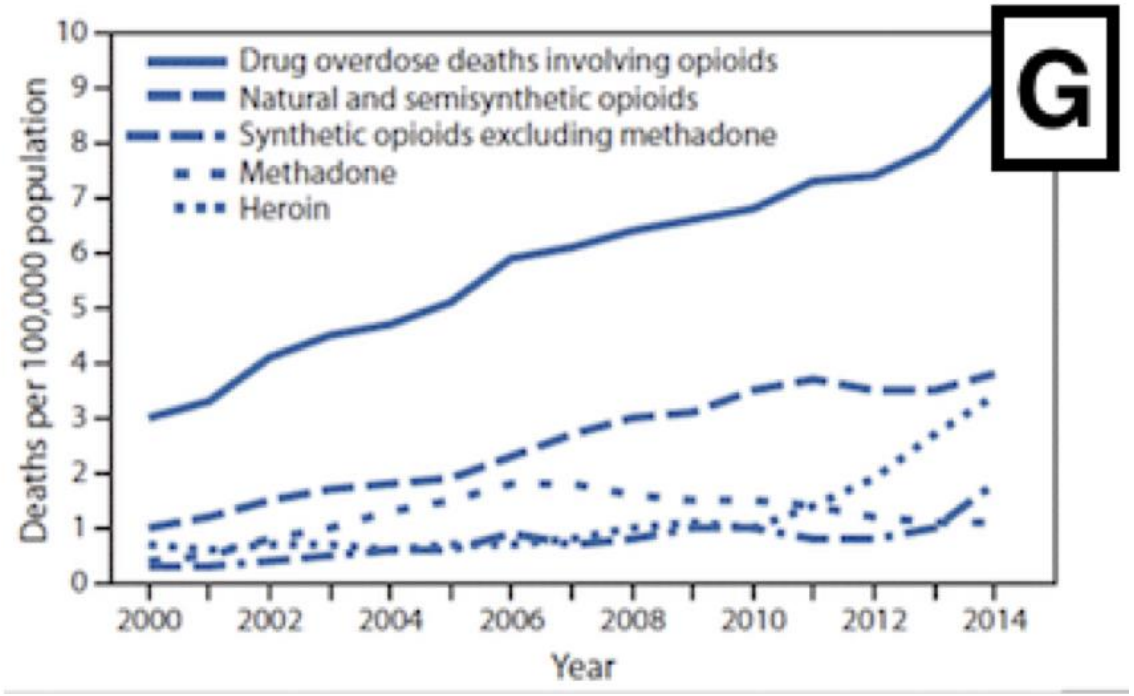
- The most common codes associated with graphs across our interviews are as follows: Colorful (29) , Confusing (29), Clear (26), Simple (26), Relatable (21), Attractive (20), Informative (19), Cluttered (17)
- gravitated towards straightforward visual encodings
- Simple bar graphs (Graphs A, B) and line graphs (Graph I) emerged as among our more highly ranked charts



Graph G and I

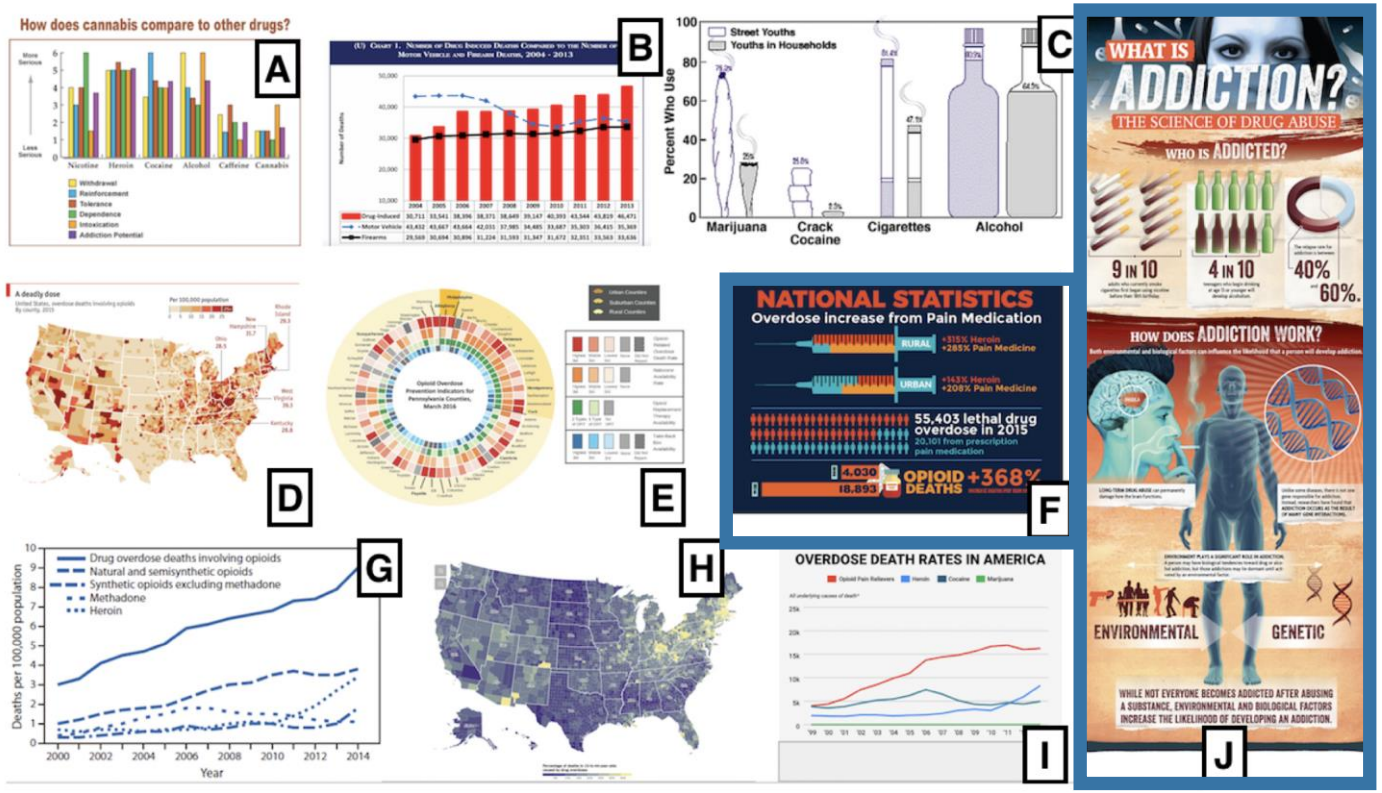
- critiques of clarity and aesthetics often blurred together for our participants
- 16 participants identified color as a distinguishing factor
- often ambiguous as to whether color referenced general appeal or an improved visual encoding





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



Changing of rankings

- 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.

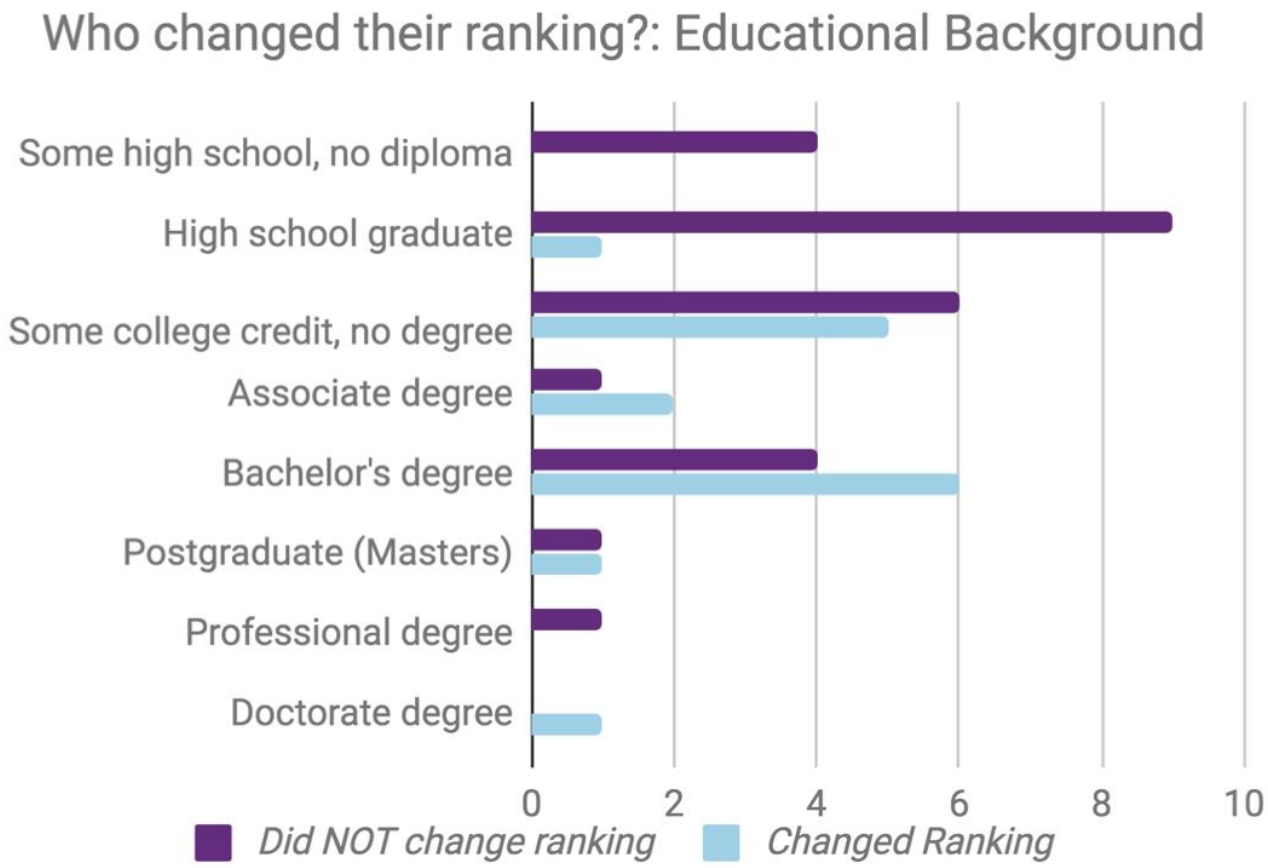


Figure 5: More educated participants were more likely to change their rankings after seeing the graph's source

Exploratory Data Analysis

Data -> Descriptive Analysis -> Exploratory Analysis -> Product

Exploratory: Inferential, Predictive, Causal, Mechanistic

- Inferential: Statistics, Frequentist, Bayesian, Text & Geospatial analysis
- Predictive: Statistical Learning/ML, Deep, Reinforcement Learning
- Causal: How variable X correlates to Y
- Mechanistic: How much does variable X affect Y

Univariate, Bivariate, Multivariate

Explanatory (Independent) vs Response (Dependent) variables

Source of data (Zipcode vs hometown), explore missing data

Don’t do EDA to give you the result you want

Exploratory Data Analysis

Checklist of things to do during EDA

- ❑ Investigate missing values
- ❑ Understand outliers
- ❑ Add filters, transform and scale data
- ❑ Calculate numerical summaries
- ❑ Generate plots to explore relationships
- ❑ Handle proportions correctly
- ❑ Use tables to scan data
- ❑ Search for patterns

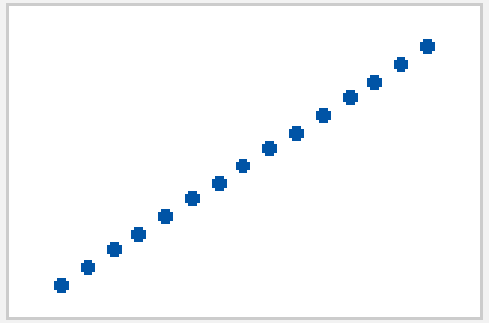
EDA
Demo

Correlation Coefficients

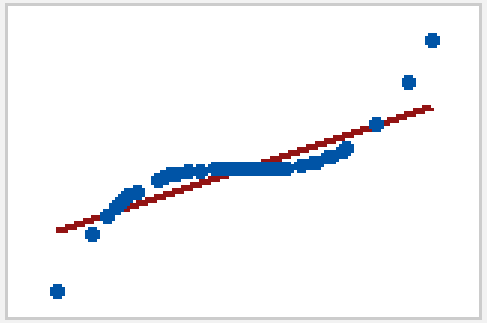
Pearson Correlation = Linear relationship between two variables

Spearman Correlation = Monotonic relationship between two sets

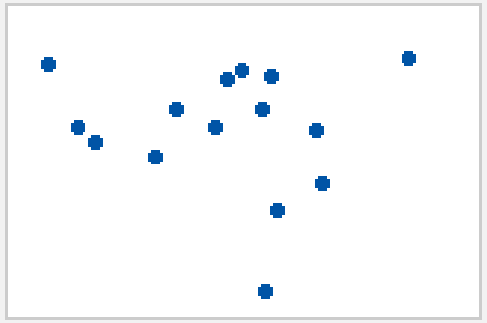
Pearson = +1, Spearman = +1



Pearson = +0.851, Spearman = +1



Pearson = −0.093, Spearman = −0.093



T-Test

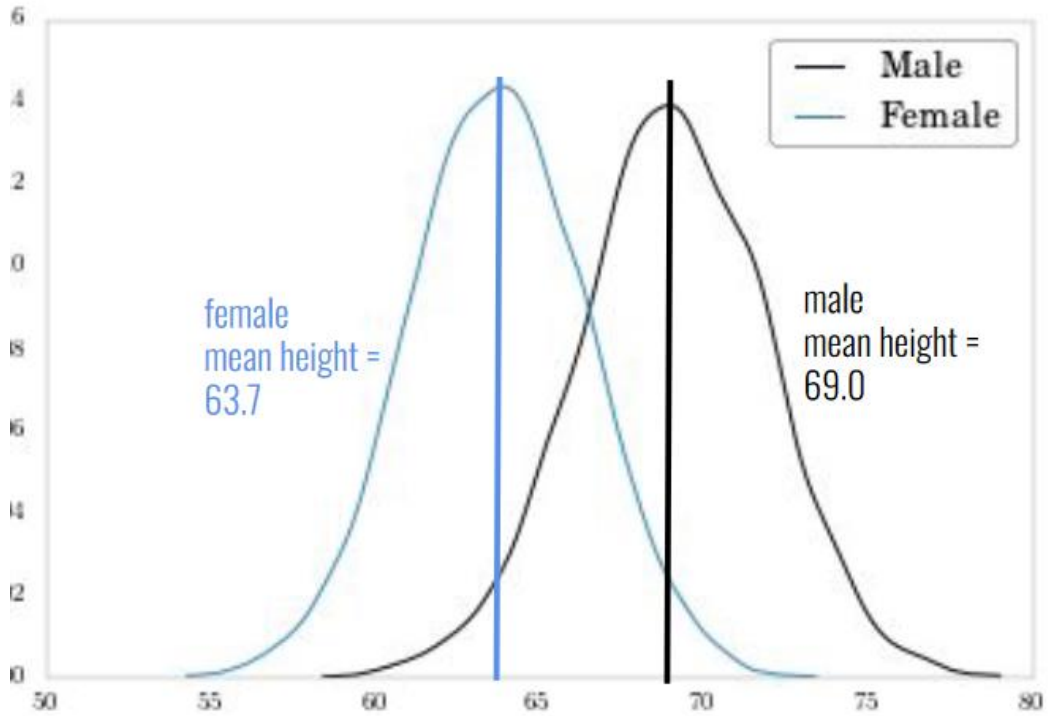
Test for differentiating between the means of 2 groups

- Data should be continuous
- Normally distributed
- Large enough sample size
- Equal variance between groups

Greater magnitude of T implies that there is a statistically significant difference b/w the 2 groups

t-statistic: -95.6

p-value < 0.001



Inferential Analysis Demo