BIOMED SCI 552: STATISTICAL THINKING

LECTURE 7: EXPLORATORY DATA ANALYSIS

QUESTIONS FROM THURSDAY?

THE PROBLEM SET

- People found this one a little more challenging
- That's a good thing
    - There is, on occasion, virtue in being a little stuck
- Remember: If you show your work and I can follow your logic, partial credit can be had

PROBLEM 1

- This is just an extension of the Bayes' Theorem problem we covered in class

$$P(Disease|Test = Positive) = \frac{P(Test = Positive|Disease)P(Disease)}{P(Test = Positive)}$$

$$P(Disease|Test = Positive) = \frac{0.98 * 0.0133}{\big((1 - 0.0133) * 0.05\big) + (0.0133 * 0.98)}$$

$$P(Disease|Test = Positive) = 0.209$$

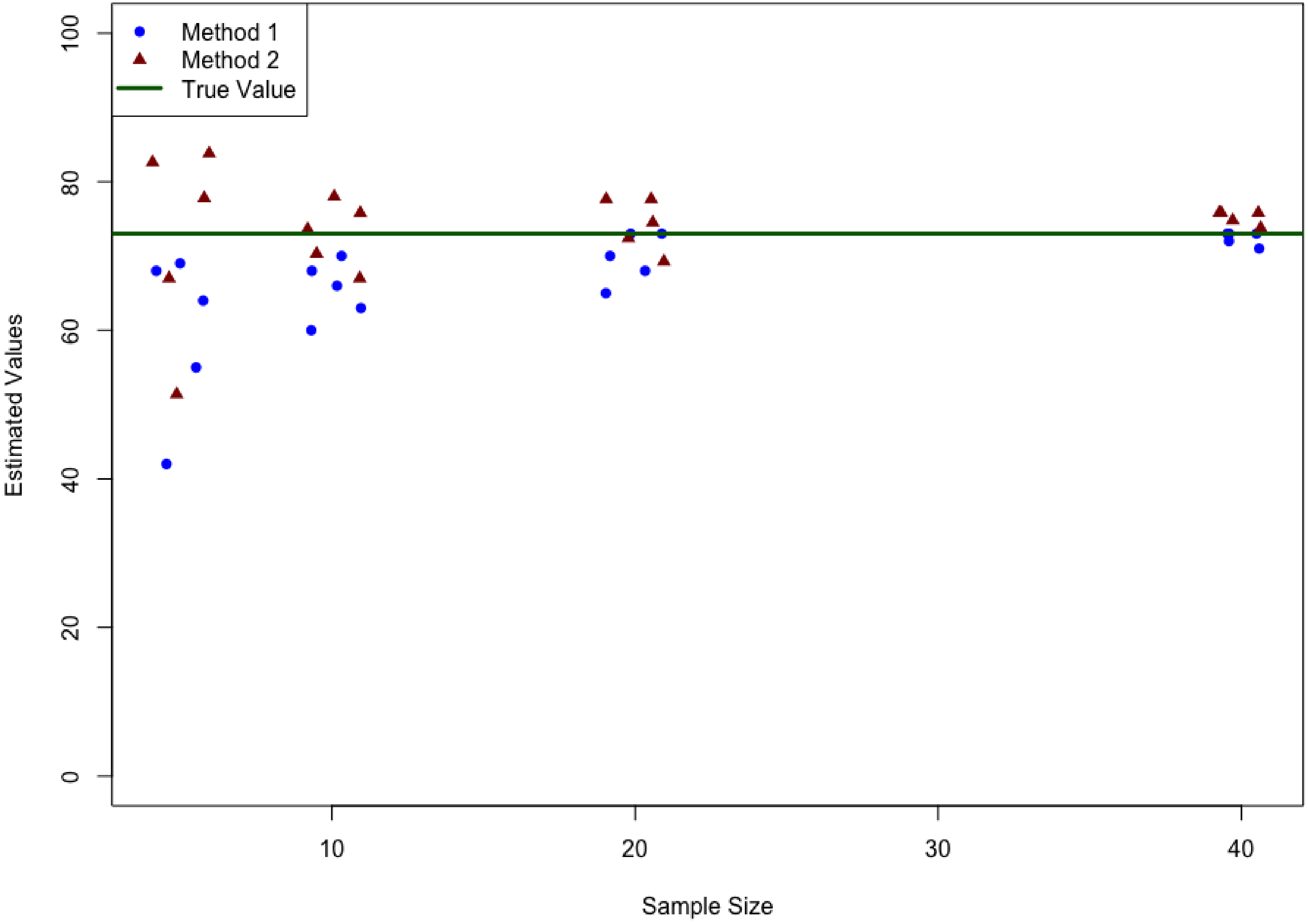
QUESTION 2

MY TAKE ON QUESTION 2

- What is the cost of the test?
- What is the consequence of not knowing vs. knowing?
- What interventions are the result of the test, and are these themselves harmful or expensive?

QUESTION 3

- This was mostly an opportunity to explore this
- There’s not necessarily *a* way to answer this question



● Method 1
▲ Method 2
True Value
Estimated Values
0
20
40
60
80
100
10
20
30
40
Sample Size

QUESTION 4

- I’ve been given a problem with a probability of success (1/6) and a number of trials (20) – this feels like a binomial distribution problem to me
- It’s 6 *or more* times, so I don’t just need the probability of 6, but of 6 to 20
- This seems laborious to me
- The probability of 0, 1, 2, 3, 4 and 5 is one minus the probability I want, and involves way less calculation

CALCULATIONS

$$f(s) = \frac{n!}{s!\,(n-s)!}p^s(1-p)^{n-s}$$

Or…

```
dbinom(s,size,prob,log=FALSE)
```

S	P(S)	Cumulative Probability
0	0.02606	0.02606
1	0.10428	0.13034
2	0.19818	0.32852
3	0.23787	0.56638
4	0.20224	0.76862
5	0.12946	0.89808

TAKE THE BET?

- P(Winning) = 1-0.89808 = 0.10192
- That’s an unlikely victory. But the payout is decent. On average, we expect to make $100*0.10192 = $10.192 on the bet. Let’s round that to $10.19
- On the other hand, on average, we expect to *lose* 0.89808*$20 or $17.96 on the bet
- Maybe find some other activity to occupy your time

IS BUYING A SINGLE LOTTERY TICKET RATIONAL?

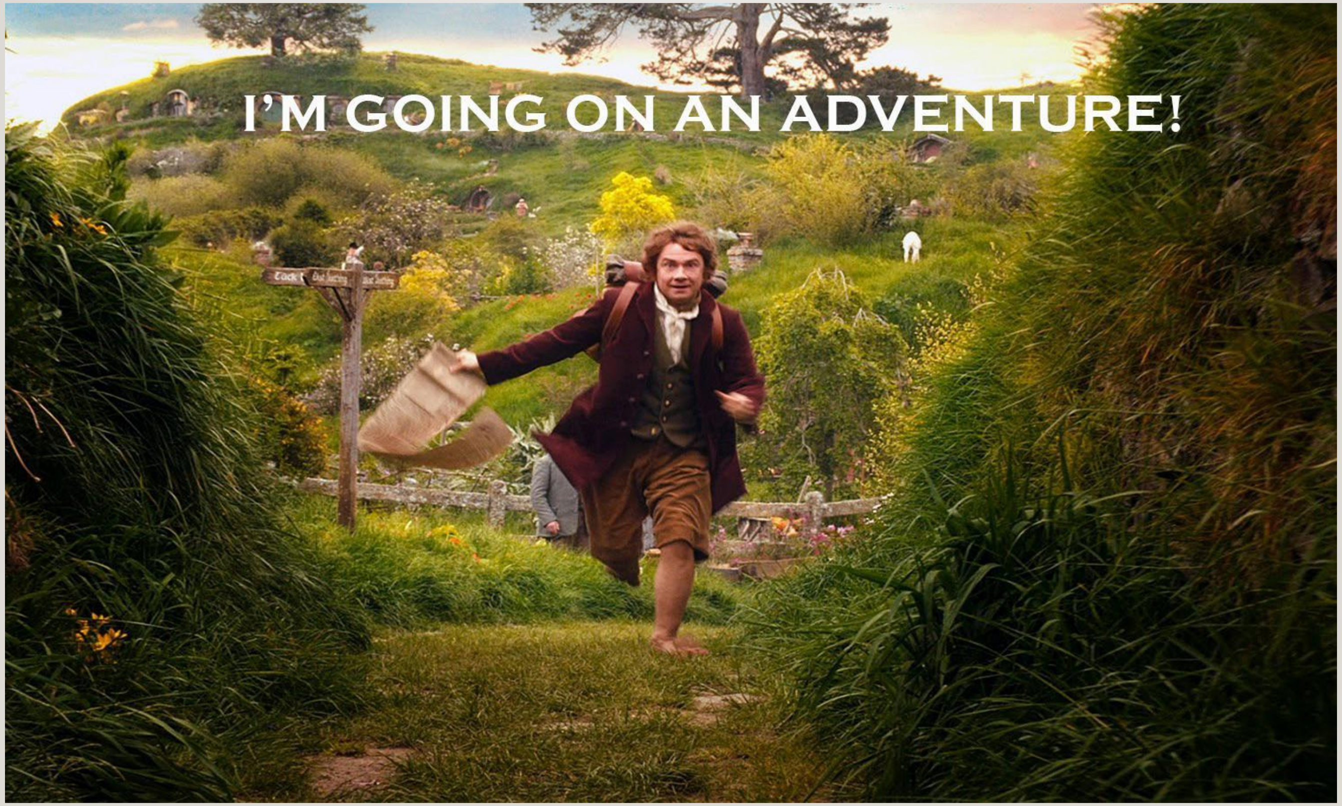
MAYBE…

- One lottery ticket is a rational purchase *if* the entertainment value of the fantasy of winning exceeds the cost of the ticket
- Expected Value of a Lottery Ticket:
- Cost = Ticket Price
- Value = Fantasy + A Potentially Large Number * Basically Zero = Fantasy

- **Two** lottery tickets doubles the price, for no increased value (you haven't increased the value of the fantasy, and the expected monetary reward is still, essentially, zero)

QUESTIONS?

EXPLORATORY DATA ANALYSIS



WHAT IS EXPLORATORY DATA ANALYSIS

- First, we're going to call it EDA for now to keep things short
- EDA is the process of…well…exploring your data
- EDA is an exploration of the characteristics of your data, often using data visualization, to help you identify patterns, spot errors, and suggest directions to take an analysis in
- It's very helpful for assessing your assumptions
- Essentially invented by John Tukey in 1970

ELEMENTS OF EDA

- Univariate, non-graphical techniques
- Univariate graphical techniques
- Multivariate, non-graphical techniques
- Multivariate, graphical techniques

UNIVARIATE TECHNIQUES

- These techniques are primarily about simply describing a single variable, to help you understand it
- What sort of distribution might it follow?
- Are there outliers or extreme values?
  - Are these in error, or are they simply extreme values?
  - I once found a 2000-year-old man in what was theoretically a clean data set
  - I also found some test data accidentally left in, because it looked funny

FIVE NUMBER SUMMARY OF A VARIABLE

- Two Extremes:
  - Minimum and Maximum
- The Median
- The Quartiles (aka the 25th and 75th percentiles)
- Why these and not the mean and standard deviation?

AN EXPERIMENT

- I give you two samples with the following information:

Sample 1 Mean	Sample 1 SD	Sample 2 Mean	Sample 2 SD
30.095	4.99	30.1202	8.99

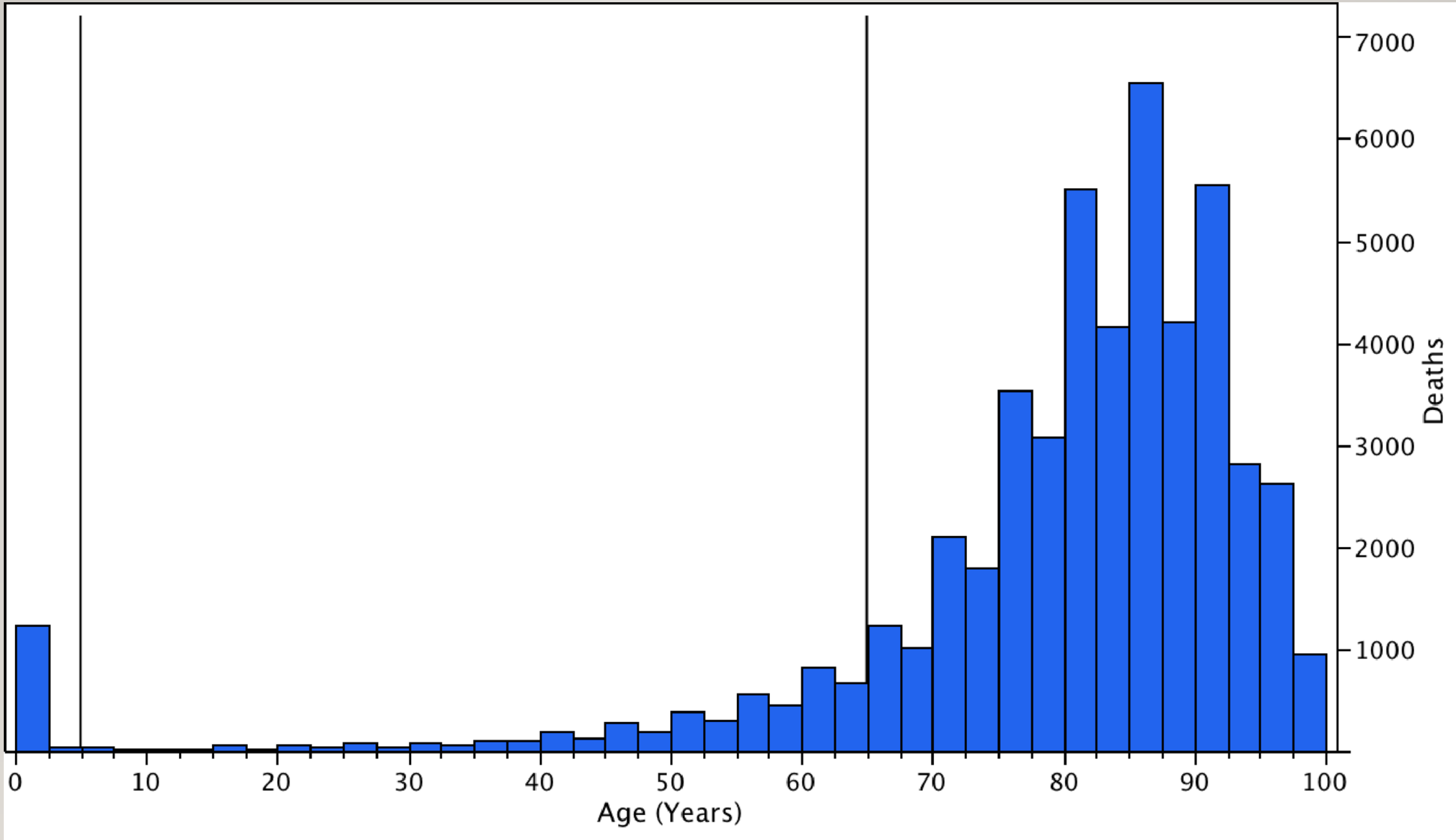
- Are these different samples? What’s wrong? It’s fairly hard to tell…
- Lets look at a five number summary, using the fivenum() function in R…

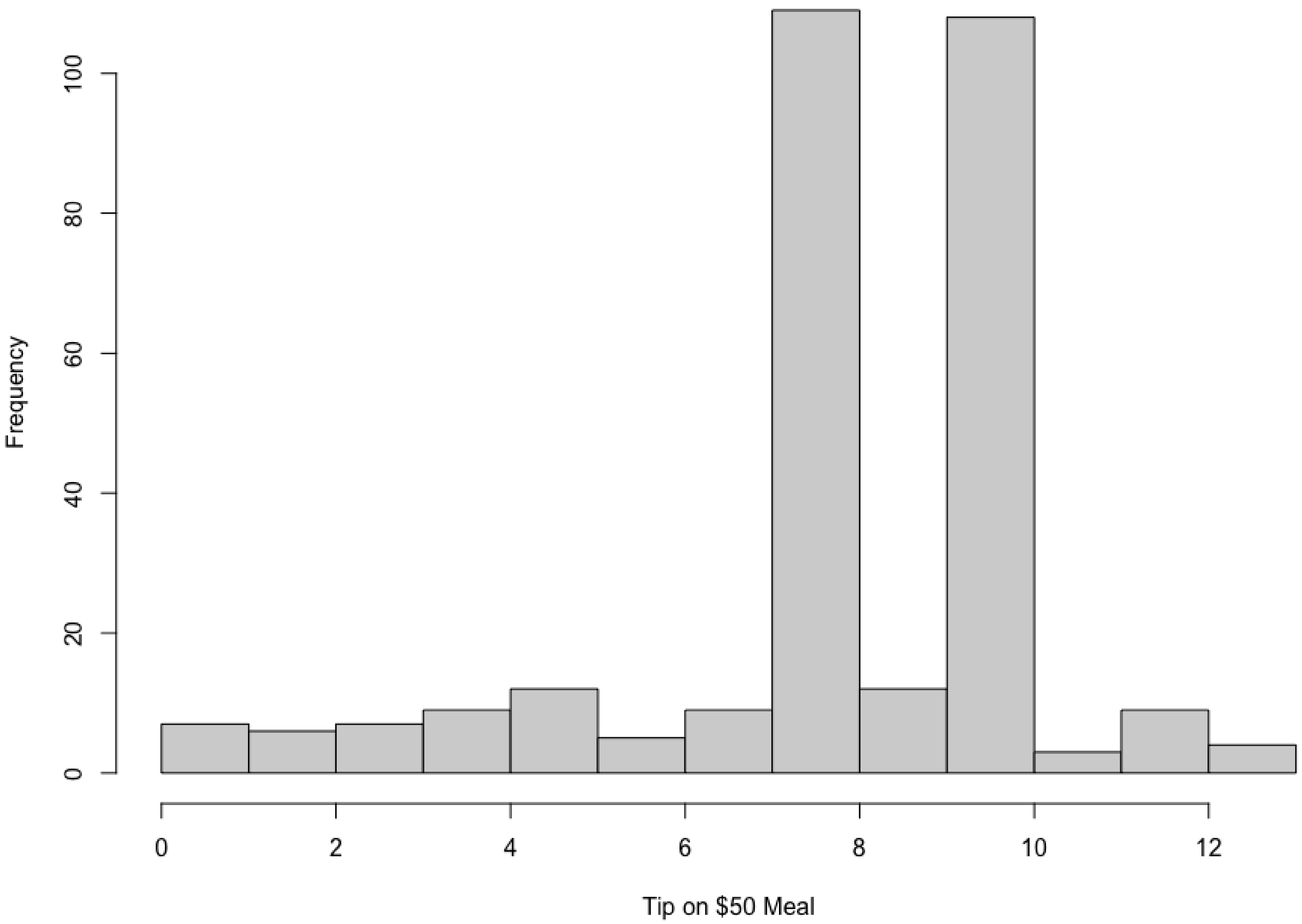
Sample	Min	25th Percentile	Median	75th Percentile	Max
1	10.87	26.68	30.11	33.42	49.22
2	11.60	26.65	30.03	33.45	778.0

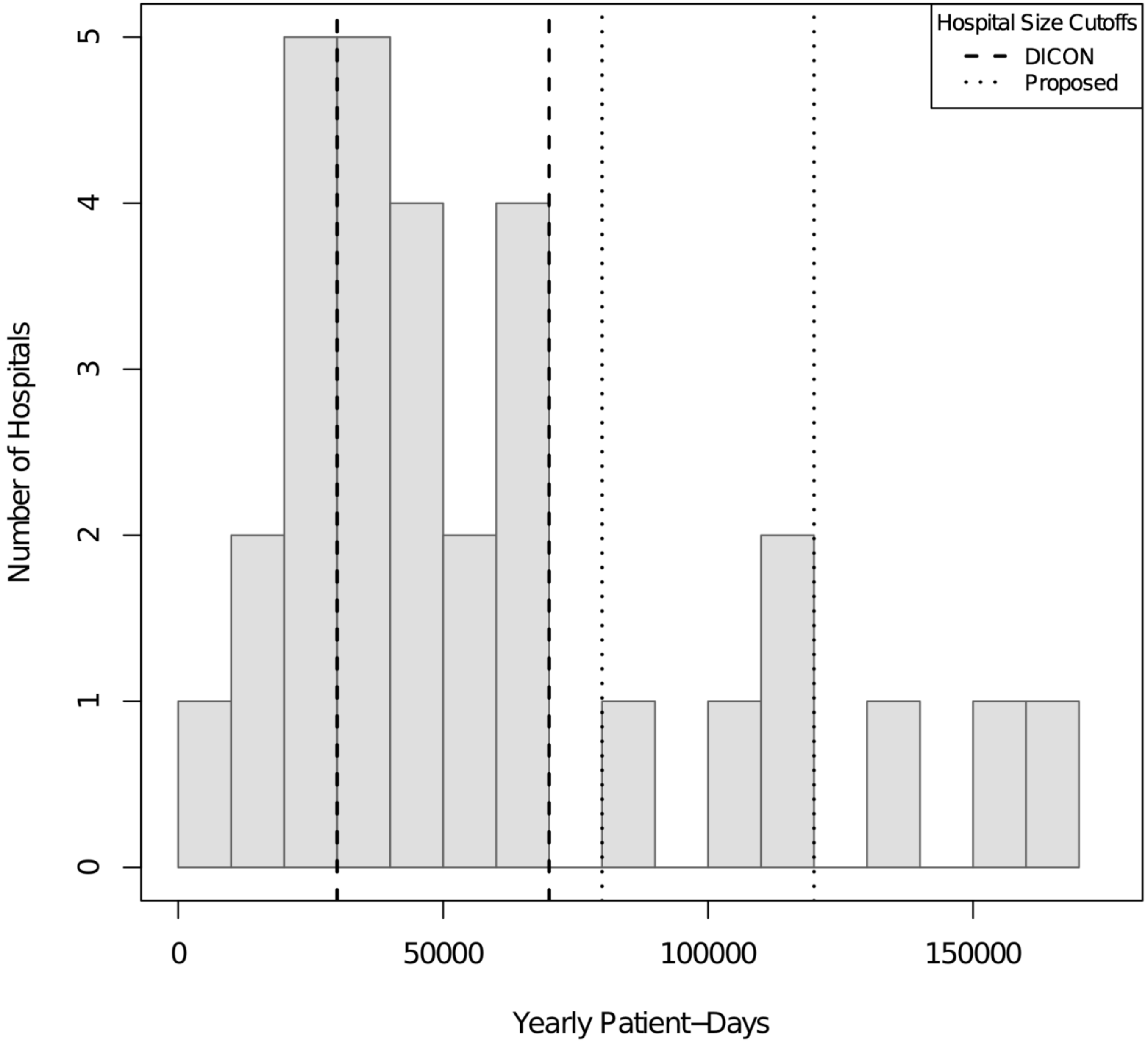
The problem is now obvious

VISUALIZING A SINGLE VARIABLE

- You can also visualize a variable using a histogram, density plot, or other methods
- This can show you things like break points in your data, natural cut offs, etc.
  - This is often very important if you want to categorize a continuous variable – it's often helpful to make sure you're not cutting a logical grouping in half
- This can also help you detect if there are thresholds in your data you were (or were not) expecting
  - Do values of a lab assay suddenly truncate?
  - Do you not enroll anyone of a particular value or lower (or higher) even though you know they exist?

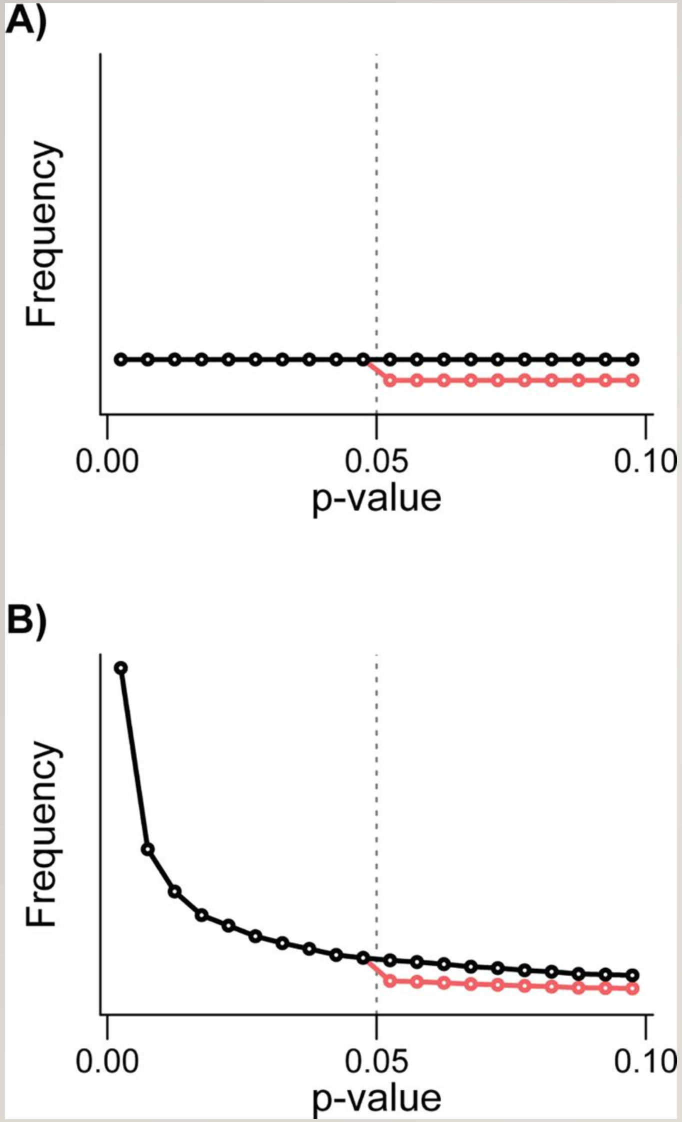

7000
6000
5000
4000
3000
2000
1000
Deaths
0
10
20
30
40
50
60
70
80
90
100
Age (Years)


Frequency
0
20
40
60
80
100
0
2
4
6
8
10
12
Tip on $50 Meal

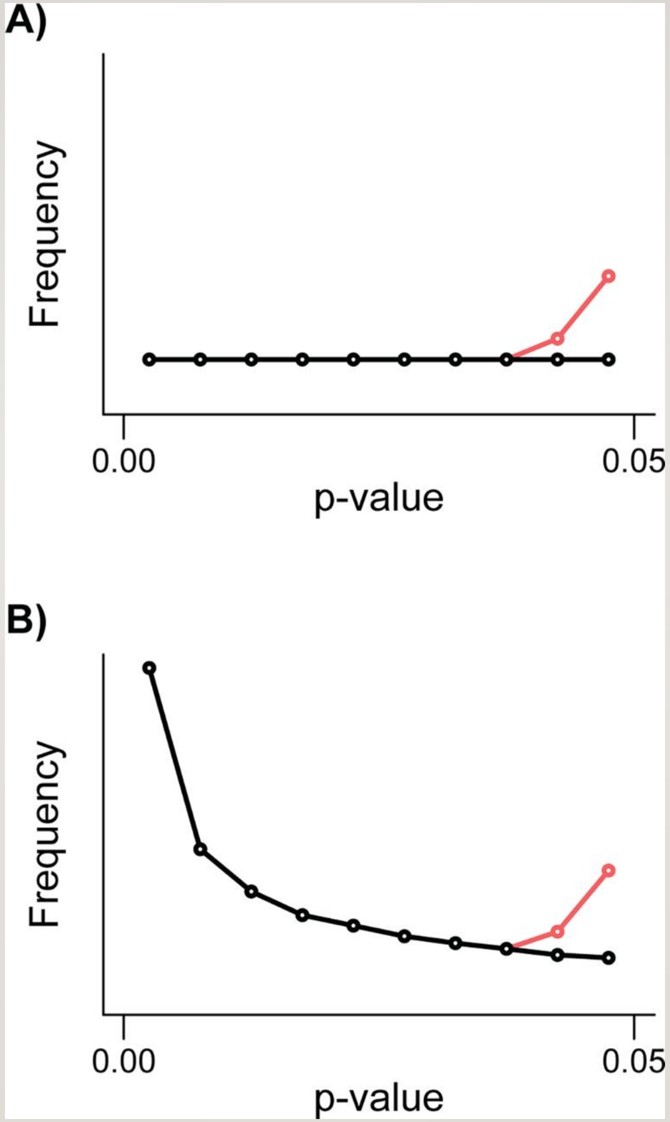

Hospital Size Cutoffs
DICON
Proposed
Number of Hospitals
0
1
2
3
4
5
0
50000
100000
150000
Yearly Patient–Days

P-HACKING AND PUBLICATION BIAS

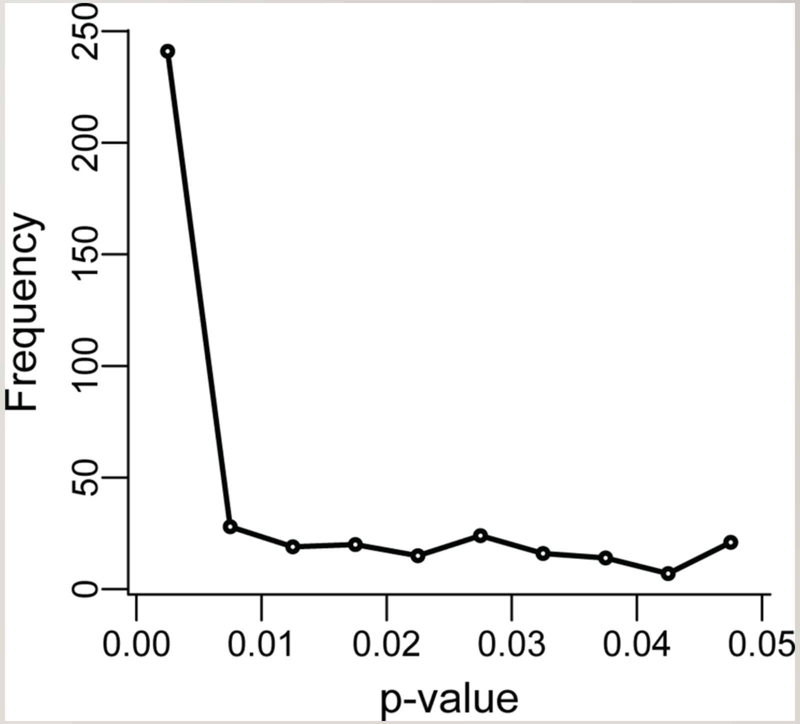
- One application of this type of visualization is detecting things like p-hacking and publication bias in a meta-analysis
- P-values should follow a theoretical distribution if they're not being influenced by researcher or publisher judgement
- But do they?
- Head ML, Holman L, Lanfear R, Kahn AT, Jennions MD (2015) The Extent and Consequences of P-Hacking in Science. *PLoS Biol* 13(3): e1002106. https://doi.org/10.1371/journal.pbio.1002106



Publication Bias



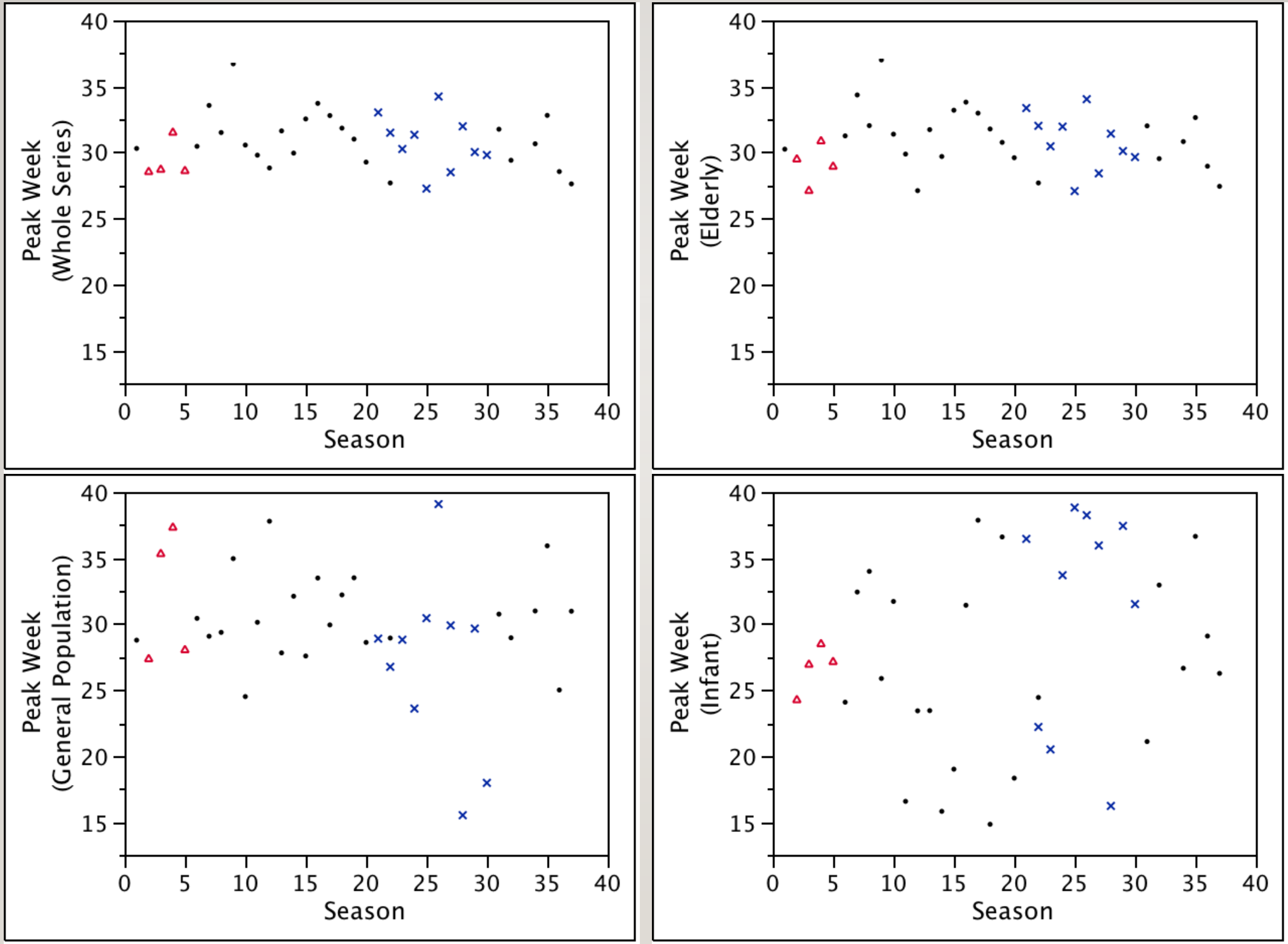
P-Hacking

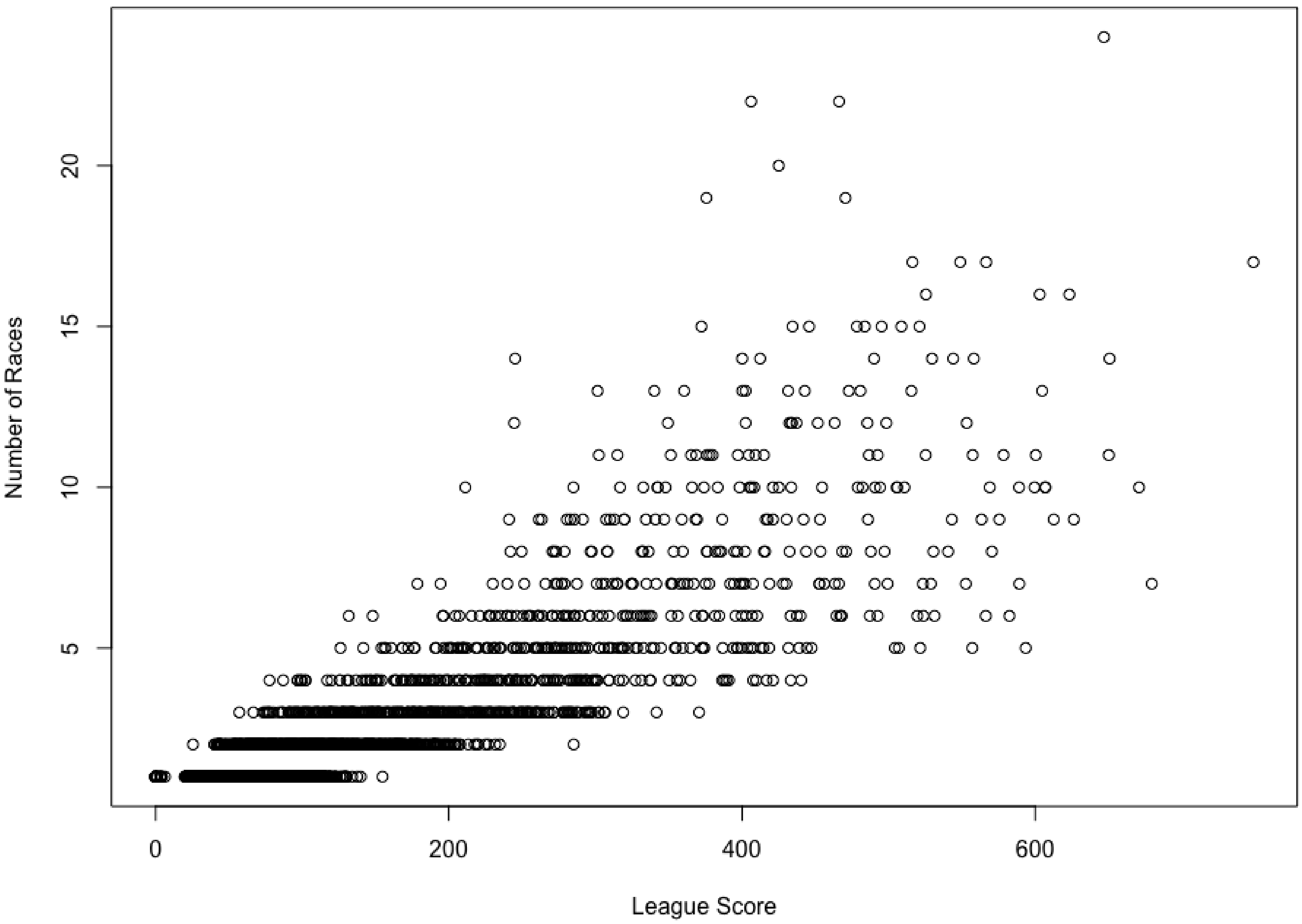


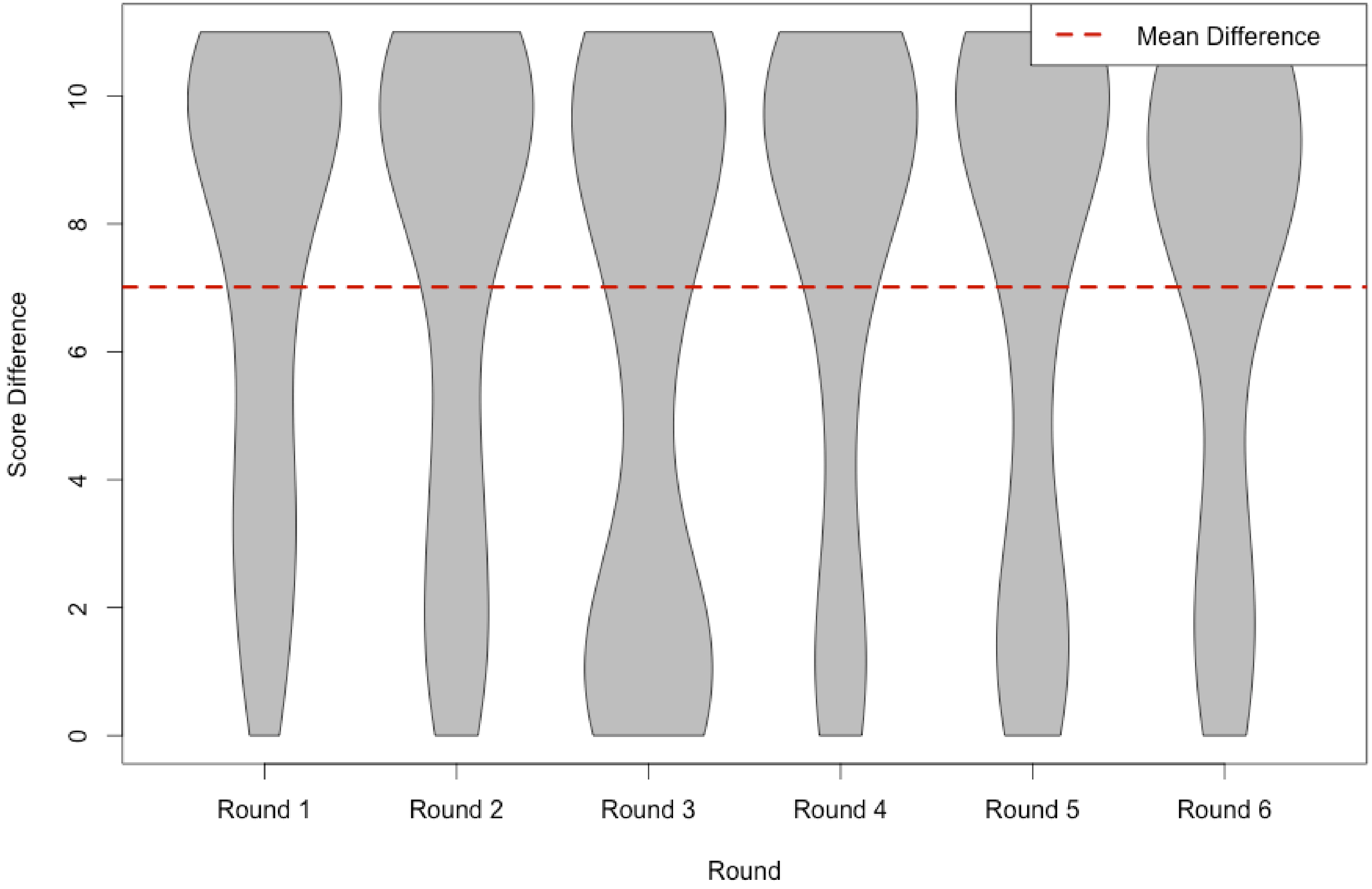
A Meta-Analysis

MULTIVARIATE ANALYSIS

- Much of EDA is concerned with understanding the relationship between two or more variables, for a number of reasons
- These can be visualized with scatter plots, box or violin plots, and a number of other methods
- There's also quantitative methods such as correlation coefficients, or linear models of the relationships
    - Once again, to be covered in a future module


Peak Week (Whole Series)
Peak Week (Elderly)
Peak Week (General Population)
Peak Week (Infant)
Season
40
35
30
25
20
15
0
5
10
15
20
25
30
35
40


Number of Races
5
10
15
20
0
200
400
600
League Score

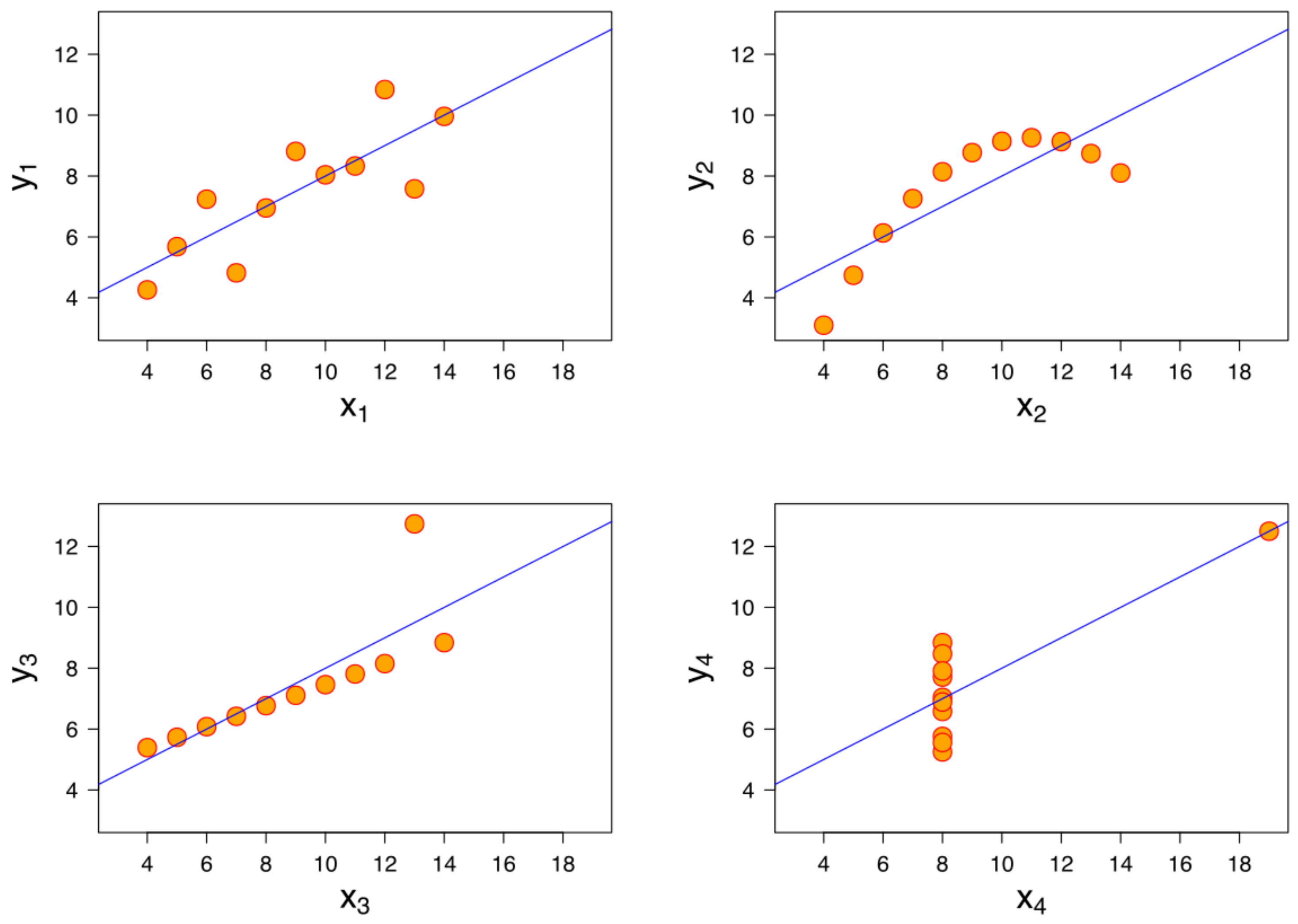

Mean Difference
Score Difference
0
2
4
6
8
10
Round 1
Round 2
Round 3
Round 4
Round 5
Round 6
Round

THE IMPORTANCE OF VISUALIZATION

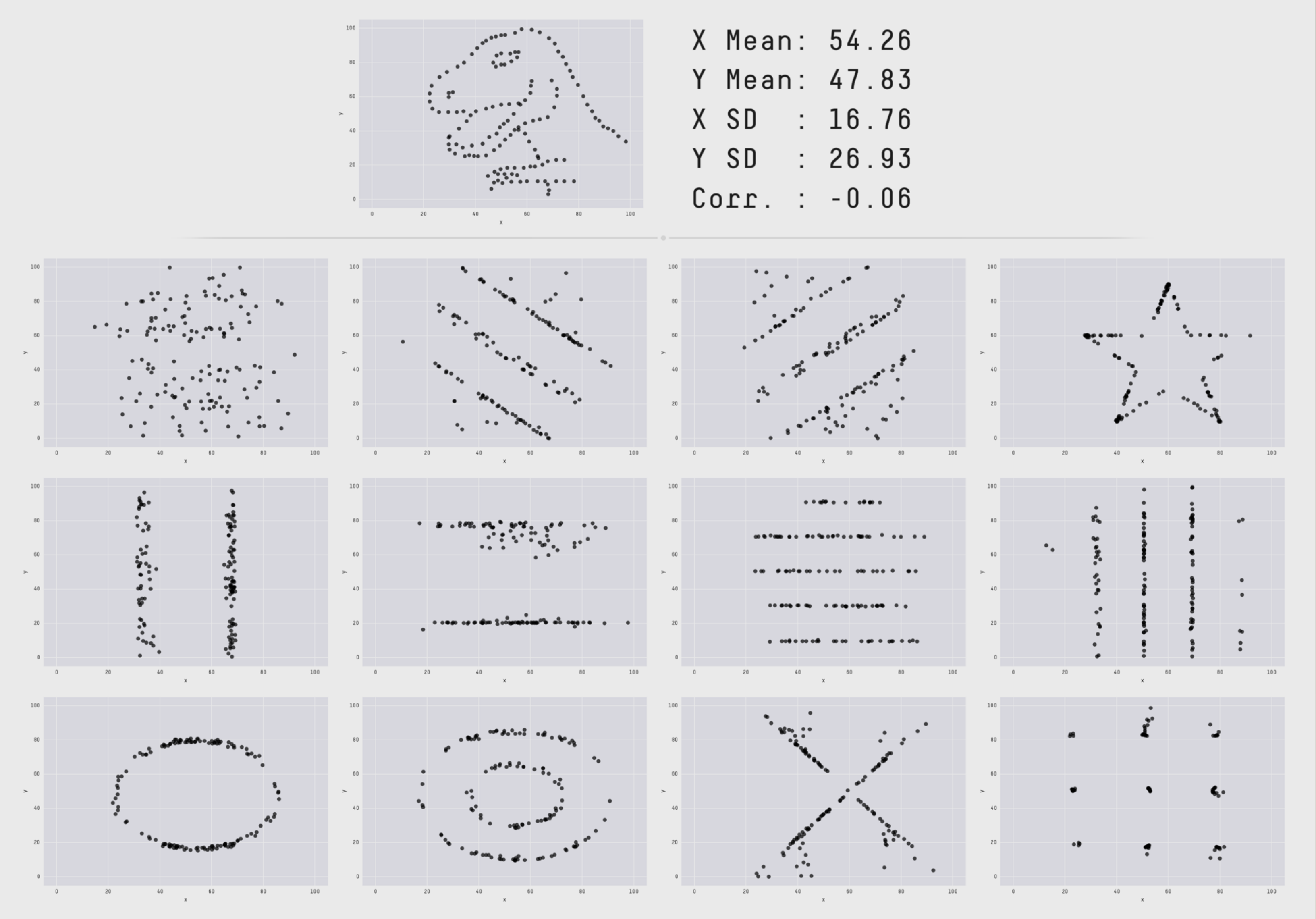
- There is often a perception that numerical measures – means, medians, correlation coefficients, etc. are cold and precise and preferable to the fuzzy business of looking at a graph and going "That's odd…"
- Human pattern recognition is a double edged sword – it's very good, but it's also overtuned
- Francis Anscombe designed four datasets, each with 11 points, in 1973 to explore this phenomenon

ANSCOMBE'S QUARTET

- Each dataset has:
  - An exact mean of X (9)
  - An exact variance of X (11)
  - A mean of Y to two decimal places (7.50)
  - A variance of Y of 4.125±0.003
  - A correlation between X and Y of 0.816 to three decimal places
  - A linear regression line of y = 3.00 + 0.500x to two and three decimal places respectively
  - A coefficient of determination (R^2) of 0.67 to two decimal places



y1
x1
y2
x2
y3
x3
y4
x4
4
6
8
10
12
14
16
18



X Mean: 54.26

Y Mean: 47.83

X SD : 16.76

Y SD : 26.93

Corr. : -0.06

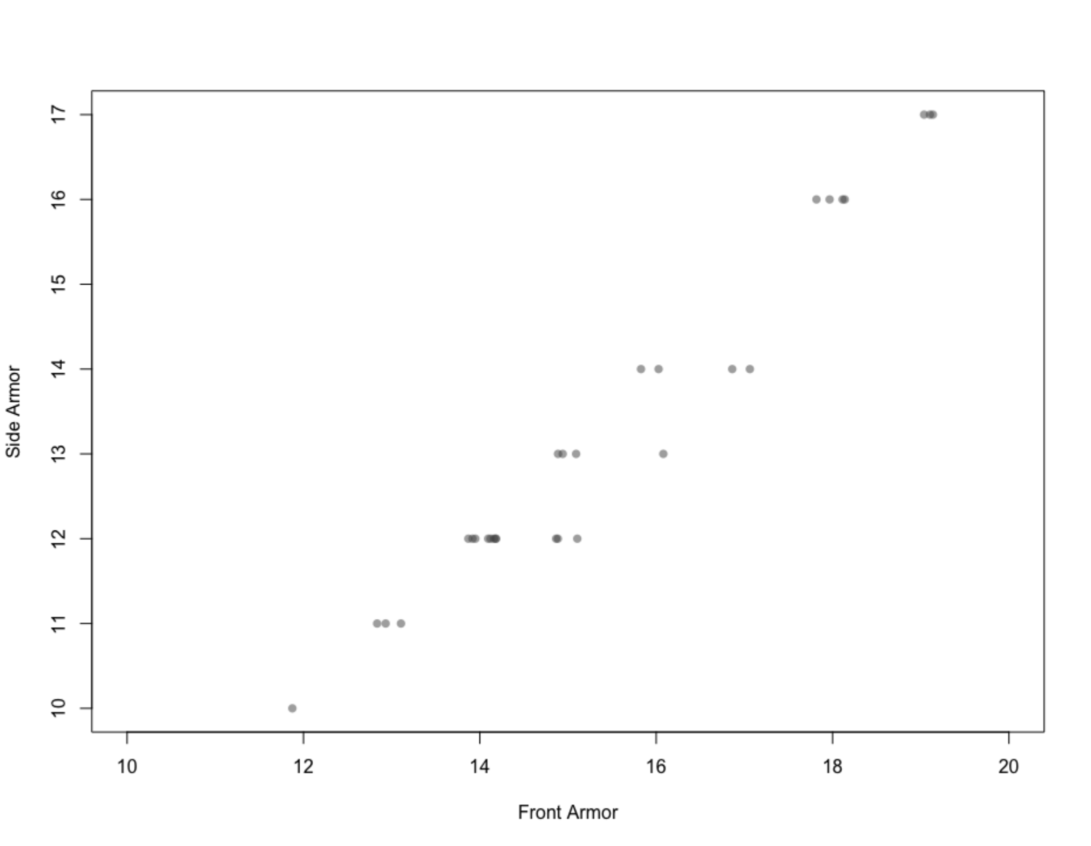
Justin Matejka and George Fitzmaurice. 2017. Same Stats, Different Graphs: Generating Datasets with Varied Appearance and Identical Statistics through Simulated Annealing. In Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17). Association for Computing Machinery, New York, NY, USA, 1290–1294. https://doi.org/10.1145/3025453.3025912

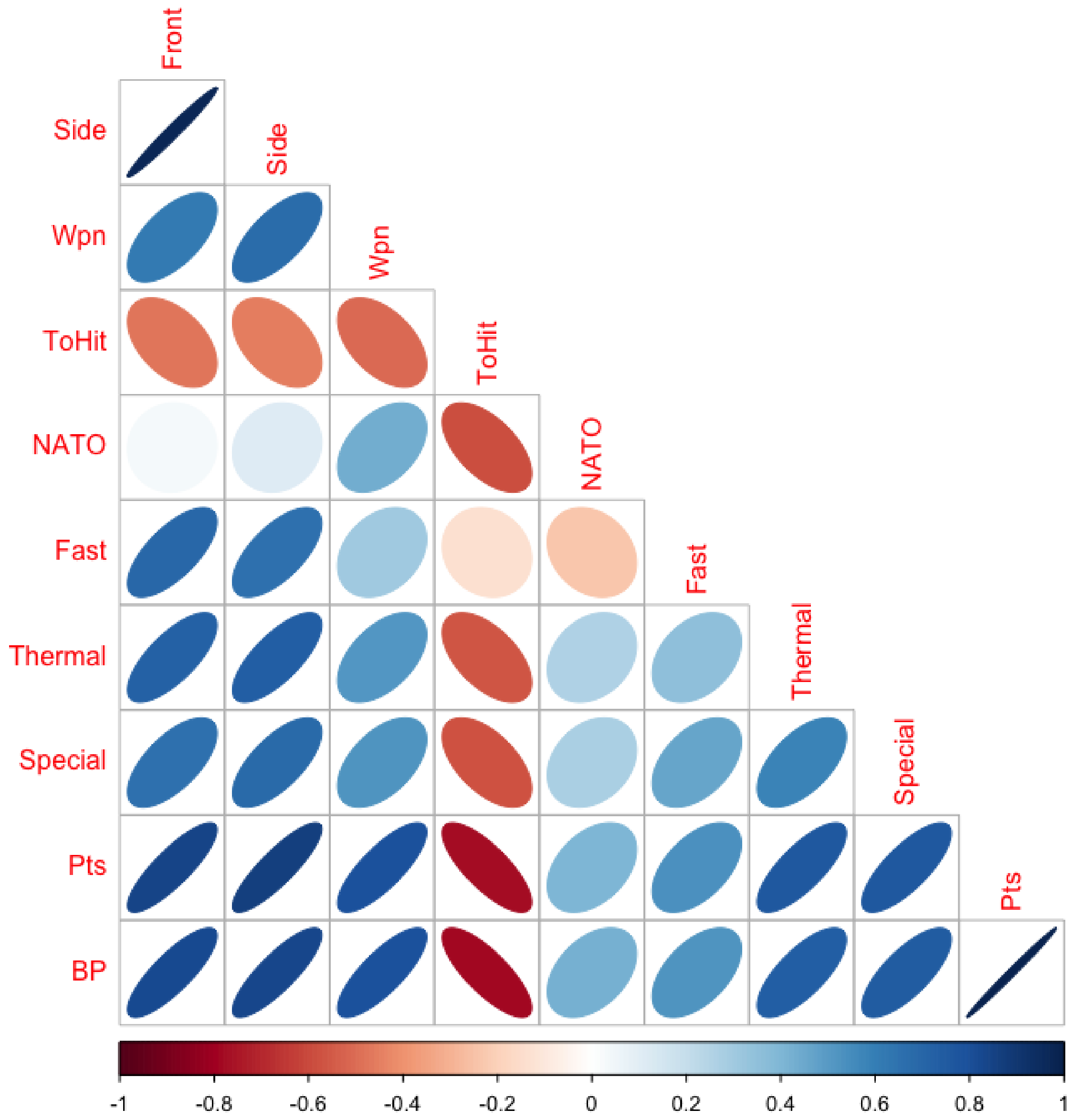
SPECIAL THINGS TO LOOK OUT FOR

- Very strong correlations
- Missing data patterns
  - Especially missing data patterns associated with another variable
- Areas of empty data
  - i.e.“empty cells”

VERY STRONG CORRELATIONS

- Very strong correlations suggest that one variable very well predicts another
- If this is what you *want*, good
- If you think you're going to include both of these variables in an analysis, this is bad, as you're going to add variance for no gain in information




Front
Side
Wpn
ToHit
NATO
Fast
Thermal
Special
Pts
Side
Wpn
ToHit
NATO
Fast
Thermal
Special
Pts
BP
-1
-0.8
-0.6
-0.4
-0.2
0
0.2
0.4
0.6
0.8
1

COMPLETE SEPARATION

- What can happen in this case is that one variable completely and perfectly predicts another
- This is called "complete separation"
  - There's also quasi-complete separation, which is you're dangerously close to this
- This is a problem, because you're likely not explaining something, but rather statistically representing a tautology

MISSING DATA PATTERNS

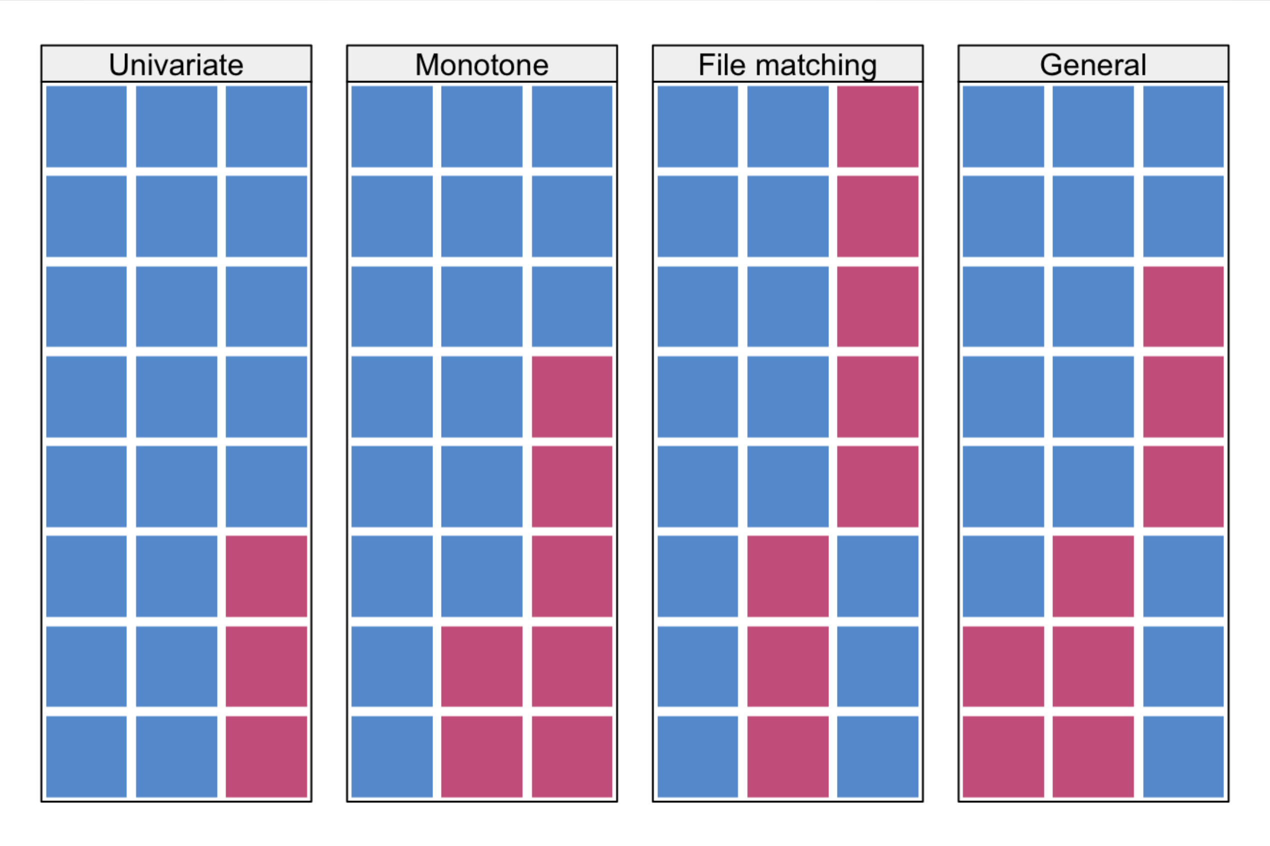


Figure by Stefan van Burren

EMPTY CELLS

- One of the underlying assumption of many of the methods we use is called "Positivity"
- In the observed data, there must be a non-zero probability of every individual or unit receiving any level of the treatment or exposure variable
- You take a sample of students in Pullman, and one of the variables you ask about is whether or not they drive a stick shift car
- Zero students say yes
- Is this a violation of the positivity assumption?

AN EXAMPLE

- You have been hired by the United States Navy to study non-combat related injuries in sailors
- You have data from 1967 to 2010 (the next 14 years remain classified)
- You have been given a number of variables: Age, what type of ship the sailor serves on, their rank and job, gender, physical fitness scores, height, weight, etc.
- Can anyone guess the problem?

POSITIVITY PROBLEMS

- Merely not having someone in a cell is not a problem
  - Again, think stochastically – how did someone end up in your sample, and what alternative samples might there be
  - Rare is not impossible
  - But impossible is bad
- These questions come up a lot in social epidemiology and related fields, and can get somewhat philosophical
  - Is a minoritized population that is not found in the highest income neighborhoods, while the majority population is not found in the lowest income neighborhoods (with overlap in the middle) a positivity violation or not?