5.Description

PS 0700

Admin

- *Note:* Some material in these slides comes from Data Analysis for Social Science (DSS) by Llaudet and Imai

This Week

- We spent the last two weeks on *causality*
  - Randomized experiments + simple difference-in-means $\rightarrow$ estimate of the average causal effect
- We're going to spend the next two weeks on **description** (DSS Ch 3)
- This Week:
  - Concepts (not in DSS) but very important!
  - Summarizing a Single Variable
- Next Week:
  - Summarizing **Two** Variables!
  - Survey Sampling

Why Description?

- A key part of political science is understanding what patterns exist in the population
- If we do not know what exists in society, we cannot even begin to think about causal relationships!
- Many questions fall into this:
  - Which groups are more likely to turnout?
  - Why are certain members of Congress more effective at passing laws than others?
  - Which societies are more likely to have civil conflict?
- We often approach these questions with a causal *suspicion*, so I am going to be careful to describe answers to those questions as **descriptions** or "associations"
  - You have to have a good reason to believe that there is no spurious correlation to claim a causal relationship!

A Running Example

- We're going to use an important dataset on **legislative effectiveness** (Volden and Wiseman 2014)
- They want to measure which members of Congress are most effective at getting their legislation passed
- They explore who is more effective (e.g. men vs. women, new vs. experienced members, R vs. D, etc.)

Center for
Effective Lawmaking

University of Virginia
+
Vanderbilt University

- Okay, can't you just do a difference in means between men and women and be done?

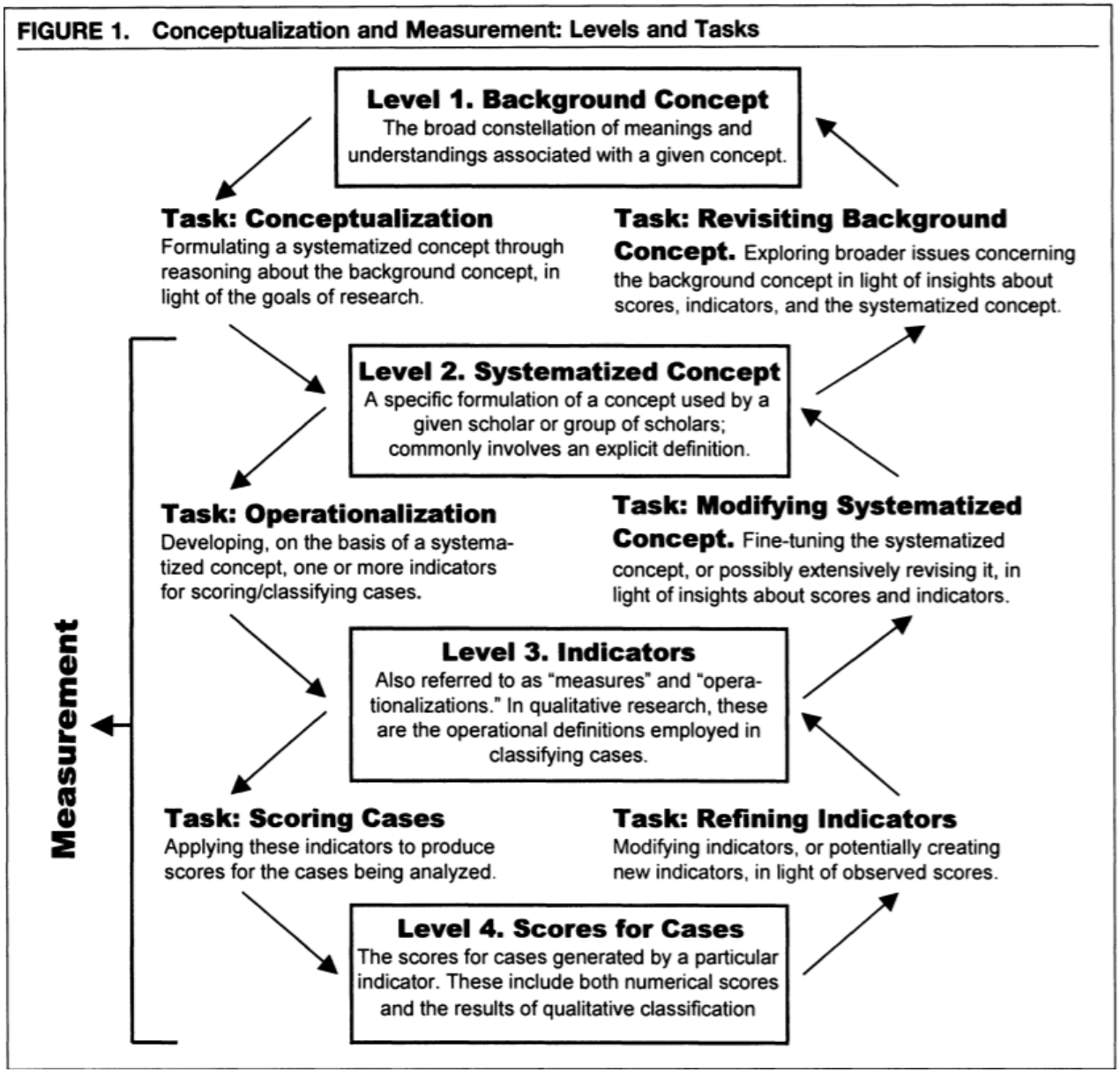

NOT SO FAST
MY FRIEND!!!!!
imgflip.com

Description Has Two Equally Important Parts

- Conceptualization and Measurement
  - How do we measure "effectiveness"?
  - What are we trying to capture? Where do those numerical scores come from?
- Statistical Analysis
  - Taking those scores as **given**, how do we summarize them?
- Often times the first step is the hardest!
- Always think critically and carefully about how researchers get their measures!

A Four Level Process

- We'll use a helpful diagram from Adcock and Collier (2001)



In Brief:

- Adcock and Collier go from “broad” to “specific” in four levels
- Background Concept
  - "The broad constellation of meanings and understandings associated with a given concept"
  - **In Brief**: The "layperson" understanding of the idea
- Systematized Concept
  - "A specific formulation of a concept used by a given scholar or group of scholars; commonly involves an explicit definition"
  - **In Brief**: Focus on one specific **aspect** of the broader idea that captures a clear, definable, part of the issue
- Indicators (Measures/Variables):
  - **In Brief**: The actual variables we want to collect. A **blueprint** or set of **rules** before one looks at data
- Scores:
  - **In Brief**: The actual **numbers** for each observation.

Background Concept (Legisative Effectiveness)

- Background Concept:
  - We are interested in studying **effectiveness**
  - What do you think effective means?
- Potentially many different meanings!
- Task 1: "Formulating a Systematized Concept"
  - How do we go from a lay understanding of "effective" to something useful?
- We think about what our **goals** are
- We want to compare members so our measure should
  - Be comparable across members
  - Be able to change over time
  - Capture different parts of the lawmaking process
  - Be measurable on different types of bills (who is more effective on agriculture policy vs. trade vs. foreign affairs)

Systematized Concept:

- Systematized concepts usually have a **broad but clear definition**
- **They do not have specific numbers, variables, or measures**
- From Volden and Wiseman's website

  We define legislative effectiveness to be the "proven ability to advance a member's agenda items through the legislative process and into law." In defining legislative effectiveness in this way, it is important to note that our definition consists of four separate components: proven ability, advancing legislation, members' agenda items, and progression through the legislative process into law...

  ... The LES is constructed to measure how successful a given Representative or Senator is at moving his or her own legislative agenda items (meaning, the bills that he/she sponsors) through different stages of the legislative process...

- What might be **added** to their concept?
- This definition uses an **individual-centric** measure of lawmaking
  - You must be the one who introduces the bill to get credit
  - But what about if a colleague puts your bill into theirs? That is probably effective too!
- You might think about **revising** your systematized concept if you were interested in the "team-work" function of effectiveness
- From the authors:

  That said, other efforts that may be commonly considered "legislative effectiveness," such as working behind the scenes to help others' bills pass, having one's legislative proposals incorporated into other legislators' bills (which then advance further in the legislative process), serving as Speaker of the House or party leader, or blocking proposals of opponents, are not included in calculating the LES.

Operationalization

- With this systematized concept, next is **variables**/measures
- What is our blueprint or rules for how we should capture the systematized concept?
- The authors care about two dimensions:
  - How **important** are the bills that a member introduces?
  - How **far** do they go in the legislative process?
- Quality vs. Quantity:
  - You introduced the COVID relief bill and it got passed $>$ many bills to rename the post office in your district
- Actual Progression in Congress:
  - If you introduced an important bill (Medicare for All) that went **nowhere** vs. your bill to provide greater transparency on how government contracts that **got enacted** into law
- For each bill a member introduces, let's record
  - How important it is?
  - How far did it go?

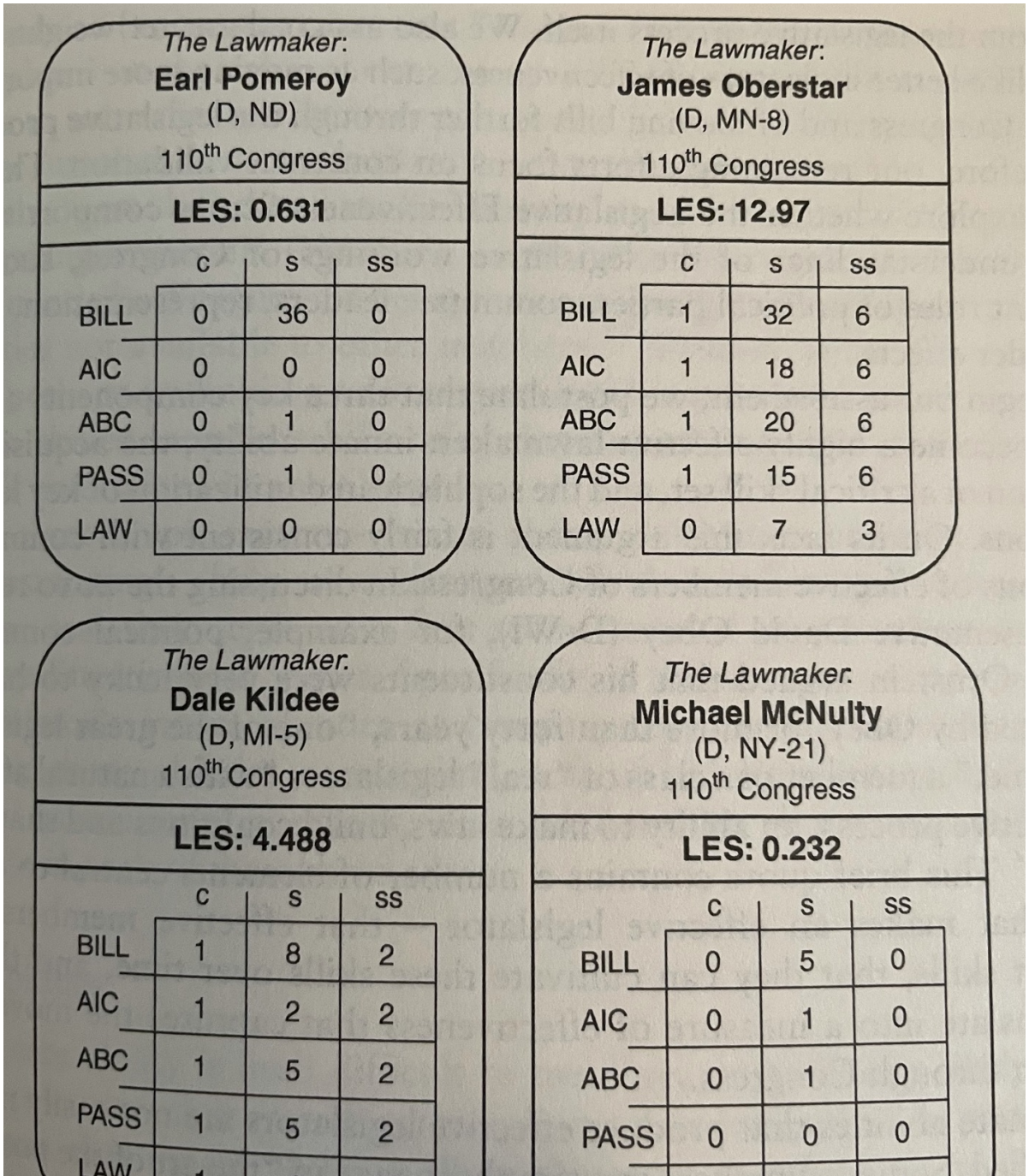
Variables Used in the Score

- Five “Stages”
- Three Levels of “Importance”

	Commemorative	Substantive	Significant
Sponsored			
"Action in Committee"			
"Action beyond Committee"			
Passed Chamber			
Became Law			

- Could try to think about better ways to do this!

Some Score Cards



The Lawmaker:
Earl Pomeroy
(D, ND)
110th Congress

LES: 0.631

	C	S	SS
BILL	0	36	0
AIC	0	0	0
ABC	0	1	0
PASS	0	1	0
LAW	0	0	0

The Lawmaker:
James Oberstar
(D, MN-8)
110th Congress

LES: 12.97

	C	S	SS
BILL	1	32	6
AIC	1	18	6
ABC	1	20	6
PASS	1	15	6
LAW	0	7	3

The Lawmaker:
Dale Kildee
(D, MI-5)
110th Congress

LES: 4.488

	C	S	SS
BILL	1	8	2
AIC	1	2	2
ABC	1	5	2
PASS	1	5	2
LAW			

The Lawmaker:
Michael McNulty
(D, NY-21)
110th Congress

LES: 0.232

	C	S	SS
BILL	0	5	0
AIC	0	1	0
ABC	0	1	0
PASS	0	0	0

Scoring

- Given our indicators/variables, what are the actual scores we assign to each observation?
- For each member/Congress, how many bills fall into each of the categories?
- Sometimes easy
  - (e.g. “introduced”, “passed”)
- Sometimes hard
  - What count as “any action in committee”?
  - How do we define “significant”?
- You might modify or change your variables if you discover hard cases or unusual variation when scoring

Final Effectiveness Scores

- Use this formula to create the final scores:

$$LES_{it} = \left[\begin{array}{l} \left(\frac{\alpha BILL_{it}^{C} + \beta BILL_{it}^{S} + \gamma BILL_{it}^{SS}}{\alpha \sum_{j=1}^{N} BILL_{jt}^{C} + \beta \sum_{j=1}^{N} BILL_{jt}^{S} + \gamma \sum_{j=1}^{N} BILL_{jt}^{SS}}\right) \\ + \left(\frac{\alpha AIC_{it}^{C} + \beta AIC_{it}^{S} + \gamma AIC_{it}^{SS}}{\alpha \sum_{j=1}^{N} AIC_{jt}^{C} + \beta \sum_{j=1}^{N} AIC_{jt}^{S} + \gamma \sum_{j=1}^{N} AIC_{jt}^{SS}}\right) \\ + \left(\frac{\alpha ABC_{it}^{C} + \beta ABC_{it}^{S} + \gamma ABC_{it}^{SS}}{\alpha \sum_{j=1}^{N} ABC_{jt}^{C} + \beta \sum_{j=1}^{N} ABC_{jt}^{S} + \gamma \sum_{j=1}^{N} ABC_{jt}^{SS}}\right) \\ + \left(\frac{\alpha PASS_{it}^{C} + \beta PASS_{it}^{S} + \gamma PASS_{it}^{SS}}{\alpha \sum_{j=1}^{N} PASS_{jt}^{C} + \beta \sum_{j=1}^{N} PASS_{jt}^{S} + \gamma \sum_{j=1}^{N} PASS_{jt}^{SS}}\right) \\ + \left(\frac{\alpha LAW_{it}^{C} + \beta LAW_{it}^{S} + \gamma LAW_{it}^{SS}}{\alpha \sum_{j=1}^{N} LAW_{jt}^{C} + \beta \sum_{j=1}^{N} LAW_{jt}^{S} + \gamma \sum_{j=1}^{N} LAW_{jt}^{SS}}\right) \end{array} \right] \left[\frac{N}{5}\right]$$

- Complicated but the gist is:
  - Weight "significant" x10 of commemorative, "substantive" x5 of commemorative and then take a weighted average at each stage and sum.
- Why those weights? Why a weighted average?
- The authors **validate** their measure
  - Check that it gives similar results with different weights
  - It matches certain prior expectations

Summing Up

- Creating our Data (Conceptualization/Measurement) has four steps:
  - Take the general idea you are interested in (Background Concept)
  - Create a specific definition relevant to your research (Systematized Concept)
  - Create a set of rules for making variables that capture the concept (Indicators/Variables)
  - Get the actual scores for each observation using the rules
- At each stage, think about whether we should adjust
  - Actually coding data $\rightarrow$ discover new things $\rightarrow$ adjust the concept

Breakout Room

- Imagine you wanted to revise the legislative effectiveness score to take account of "teamwork"

  That said, other efforts that may be commonly considered "legislative effectiveness," such as working behind the scenes to help others' bills pass, having one's legislative proposals incorporated into other legislators' bills (which then advance further in the legislative process), serving as Speaker of the House or party leader, or blocking proposals of opponents, are not included in calculating the LES.

- How might you modify their systematized concept to include some part of "team work"?

  We define legislative effectiveness to be the "proven ability to advance a member's agenda items through the legislative process and into law."

- What **variable** might you create to measure this? Try to think of something that can be scored "objectively".

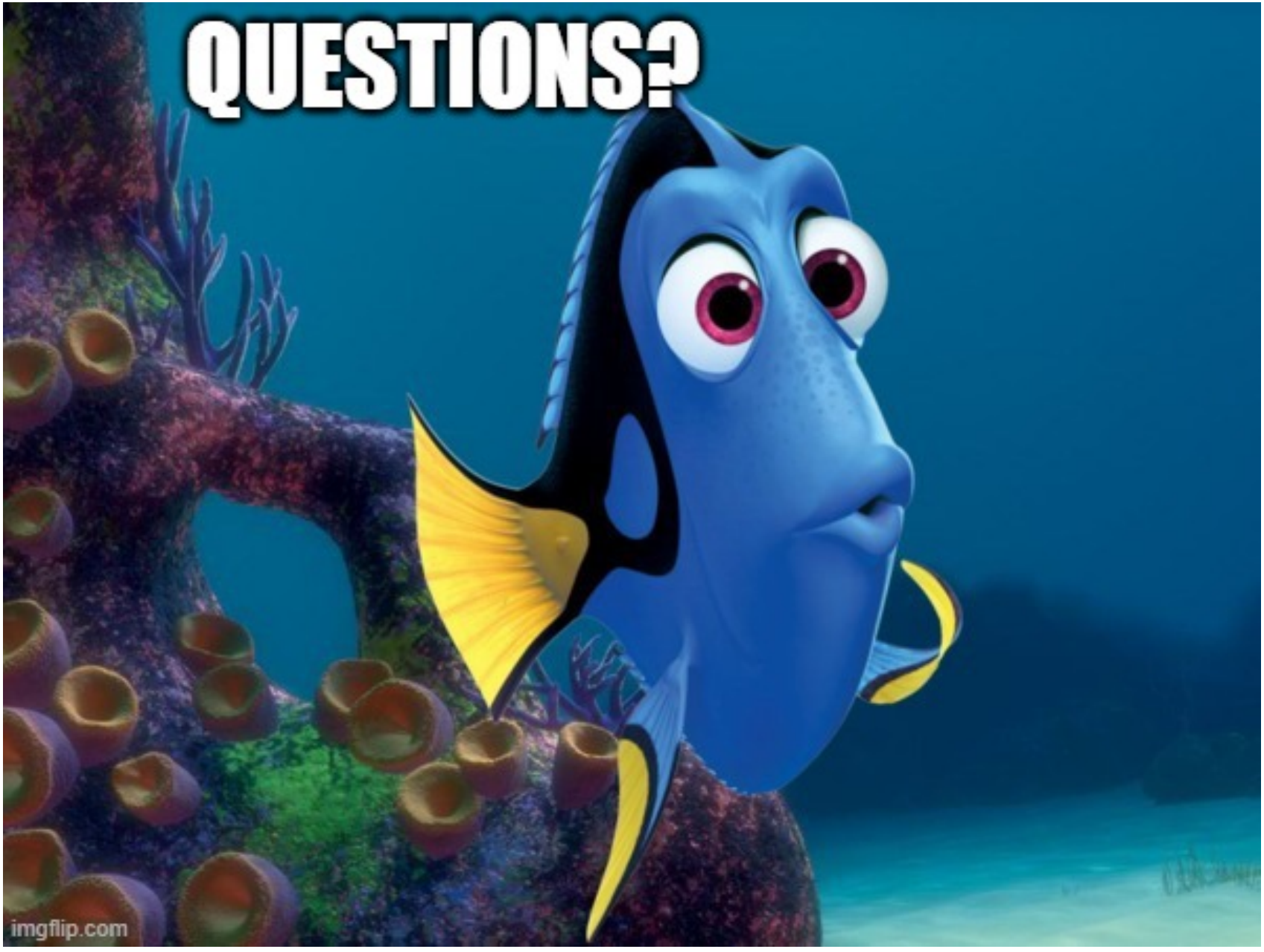
Answers

You may say something like

> *We define legislative effectiveness to be the "proven ability to advance a member's **or their colleagues'** agenda items through the legislative process and into law."*

One way to measure it would be to count all bills a member co-sponsored in their effectiveness score.

Questions



Single Variable Summary

- Let's focus on this Legislative Effectiveness Data
- Given that we have **made** this score, how do we use it?
- **Descriptive statistics** are numerical summaries of our variables
  - Condense the entire dataset into a small, interpretable, set of numbers.
- Four popular types of summary:
  - A table!
  - What is the **average** or **typical** value?
  - How "spread out" is it? How much **variability** is there?
  - Visual summaries

Summary with a Table

- If our variable has a **small number** of unique values, show the frequencies

```
library(readxl)
cel_data <- readxl::read_excel('cel_house_data.xlsx')
```

- What is the gender distribution? 1128 F + 9135 M

```
table(cel_data$female)
```

```
##
##    0    1
## 9135 1128
```

```
# 89% Male; 11% Female
prop.table(table(cel_data$female))
```

```
##
##         0         1
## 0.8900906 0.1099094
```

- Can also table with more values, although harder to read!
- Seniority (Number of Terms in Congress)

```
table(cel_data$seniority)
```

```
##
##    1    2    3    4    5    6    7    8    9   10   11   12   13   1
## 1716 1440 1271 1058  882  754  629  533  430  350  280  226  180  13
##   17   18   19   20   21   22   23   24   25   26   27   28   29   3
##   57   44   27   19   13    9    6    4    3    3    3    1    1
```

Don't Table a Continuous Variable!

```
table(cel_data$les_score)
```

```
##
##                   0 0.0010336817940697 0.0082694543525576  0.01137
##                   1                   1                   1
##   0.01550522632896  0.0196399539709091  0.0227409992367029  0.02584
##                   1                   1                   1
##   0.03101045265793  0.0382462255656719  0.0434146337211132  0.04754
##                   2                   1                   1
##  0.051684089004993  0.0568524971604347   0.057886179536581   0.0620
##                   1                   1                   1
##  0.082694537937641   0.087862953543663  0.0899303108453751  0.09923
##                   1                   2                   1
##   0.10440185666084   0.1064692214131364   0.1095702648162844   0.1147
##                   1                   2                   1
##   0.12404181063175   0.1250754892826084   0.1273961365222934   0.1302
##                   1                   1                   1
##   0.13705362379551   0.1380873173475274   0.1411883533000954   0.1463
##                   2                   1                   1
##     0.1550522595644   0.1602206677198414   0.1664227694272994   0.1715
##                   1                   1                   1
##   0.17633353173733   0.1767595857381824   0.1773672252893454   0.1788
```

"Typical Values"

- The most common typical value is the **mean**
- The sum divided by the number of observations

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N}$$

```
mean(cel_data$les_score)
```

```
## [1] 1
```

- You might also use the **median**:
- If you sorted all of the values from largest to smallest, what value is in the **middle**

$$\text{median} = \begin{cases} \text{middle value} & \text{if number of entries is odd} \\ \frac{\text{sum of two middle values}}{2} & \text{if number of entries is even} \end{cases}$$

An Example of the Median

- The median is less affected by **outliers** (extremely large or small values)
  - Data 1: [1, 2, 3, 5]
    - Mean is 2.75; Median is 2.5
  - Data 2: [1, 2, 3, 1000]
    - Mean is 251.5; Median is 2.5
- How does Elon Musk affect the **mean** of the income distribution in the US?
  - A lot!
- How does he affect the **median**? **Basically not at all!**

```r
# Much lower than the mean!
median(cel_data$les_score)
```

```
## [1] 0.4678169
```

- Consider the most effective legislator in the data:
  - Charles Rangel (110th Congress; D NY)
  - Score of 18.686!
  - Was chair of the powerful House Ways and Means Committee
- Many outliers could affect the mean

The Most Popular Measure of "Variability"

- Are all the data close to the center or spread out?
- The "variability" of the data or the "spread"
- **Standard deviation**: On average, how far away are data points from the mean? Higher SD $\rightarrow$ more variability

$$\text{standard deviation} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

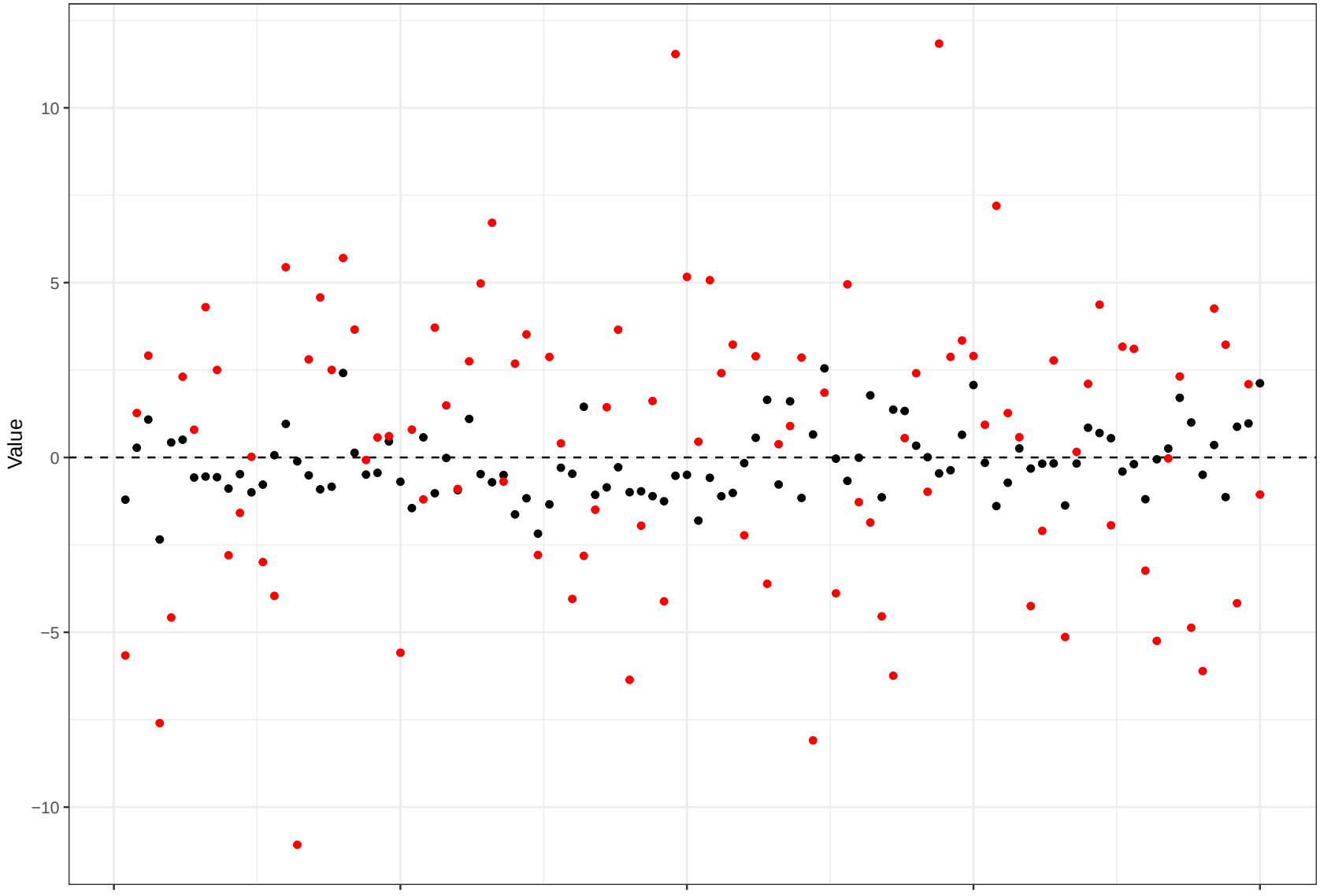
From DSS (Chapter 3):

1. Subtract each data point by the mean.
2. Square each resulting difference.
3. Take the sum of these values
4. Divide by $n - 1$
5. Take the square root.

- **Variance** = standard deviation^2
- Why not just take the average deviations without squaring? It will be **zero**!

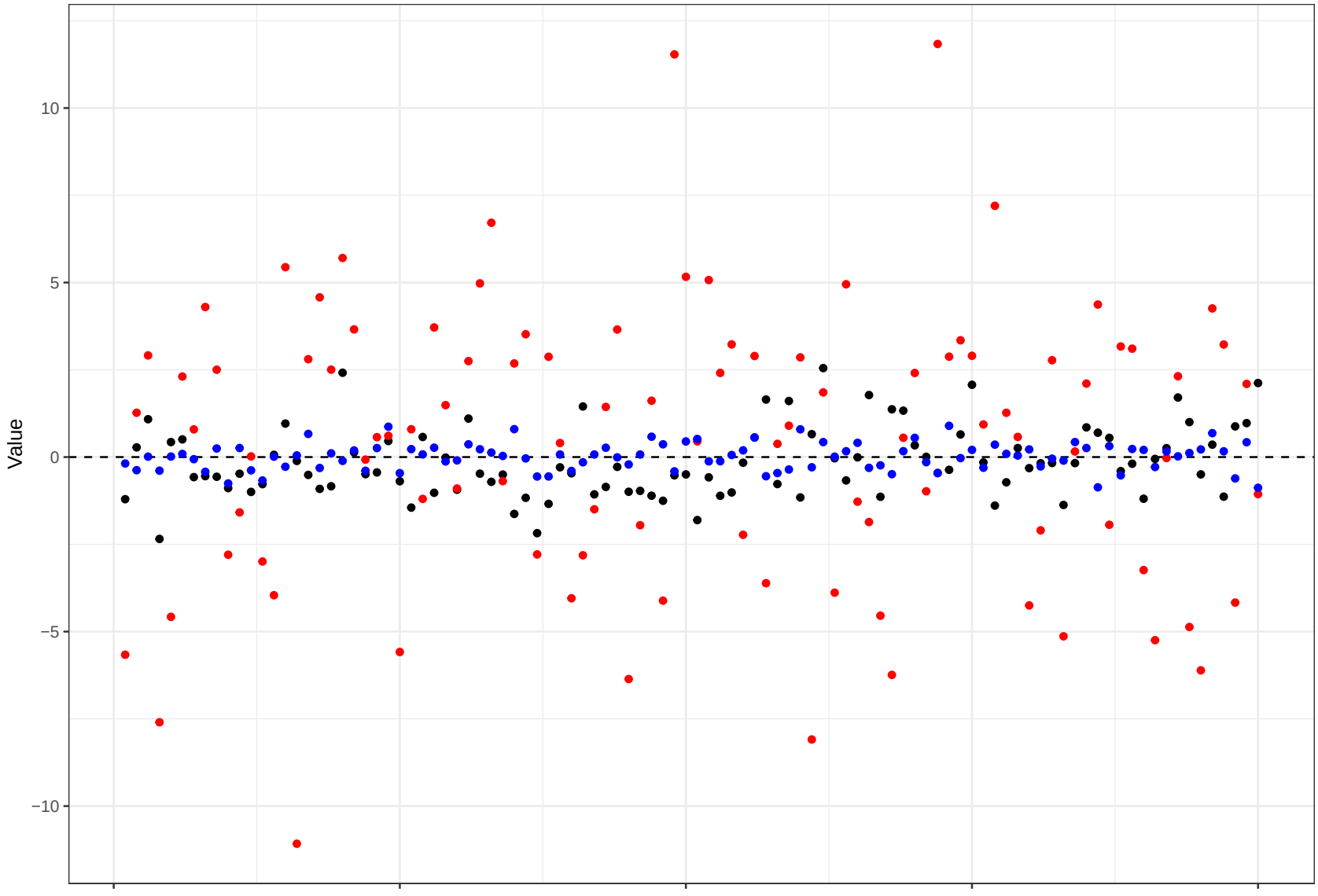
An Example (Part 1)

- Black has SD = 1; red has SD = 4; both have mean zero



An Example (Part 2)

- Compare blue that has SD of $1/3 \rightarrow$ much closer to mean



Other Measures of "Variability"

- **Range**: $[\min(X),\ \max(X)]$
- **Quantile** (a fancy name for **percentile**):
  - 25th percentile = lower quartile (25% of the data below this value)
  - 50th percentile = median (50% of the data below this value)
  - 75th percentile = upper quartile (75% of the data below this value)
- **Interquartile range** (IQR): a measure of variability
  - How spread out is the middle half of the data?
  - Is most of the data really close to the median or are the values spread out?
  - What is the difference between the 25th and 75th percentiles?

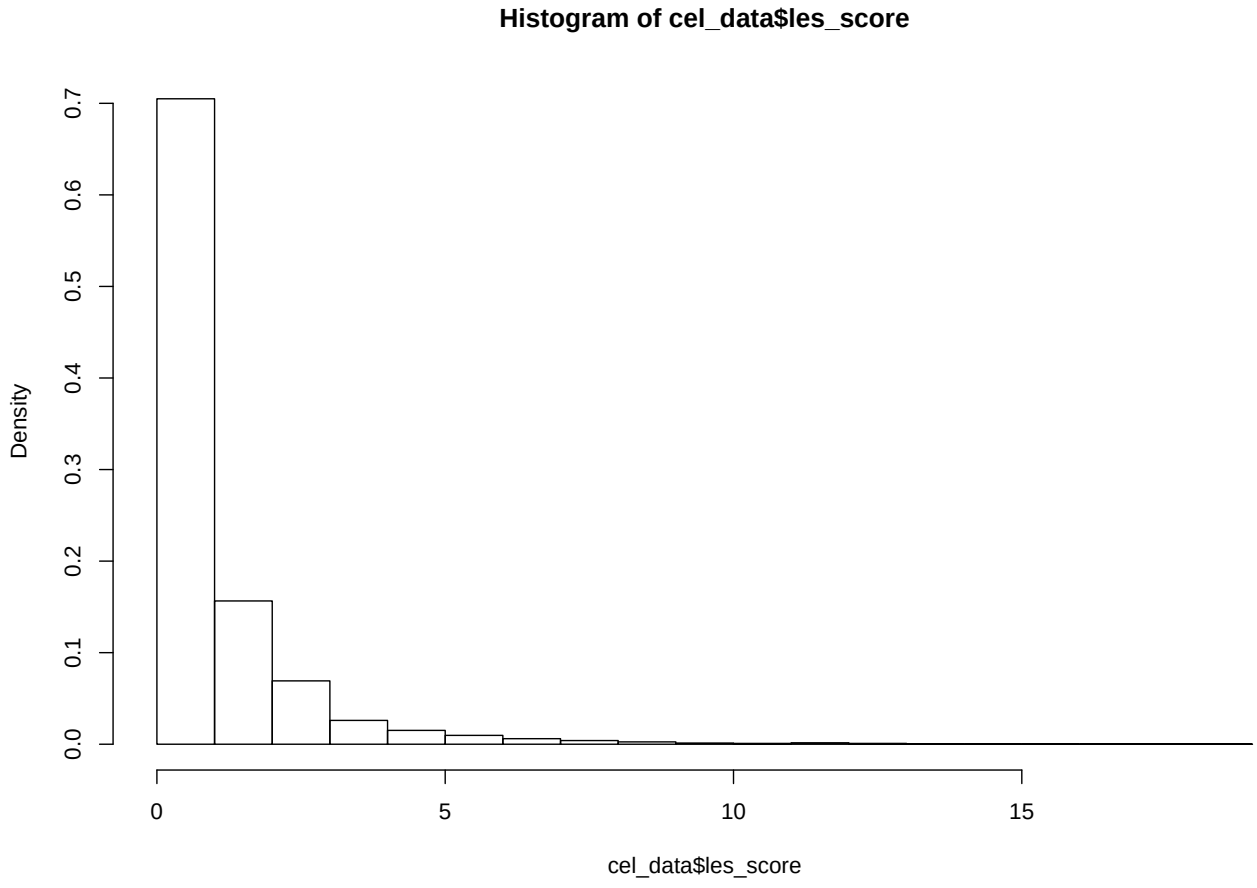
Numbers Are Nice But...

- Often hard to interpret what a standard deviation means
- Very popular to show the actual distribution of the variable
- We can use a **histogram**
- How to create a histogram? ("density histograms"; DSS, ch. 3)
  1. create bins along the variable of interest
  2. count number of observations in each bin
  3. divide by the **total** number of observations to get the **proportion of observations** (proportion * 100 = percent)
  4. **density** = bin height

$$\text{density} = \frac{\text{proportion of observations in bin}}{\text{bin width}}$$

- You can pick the number of bins or the width to make a smoother/coarser approximation

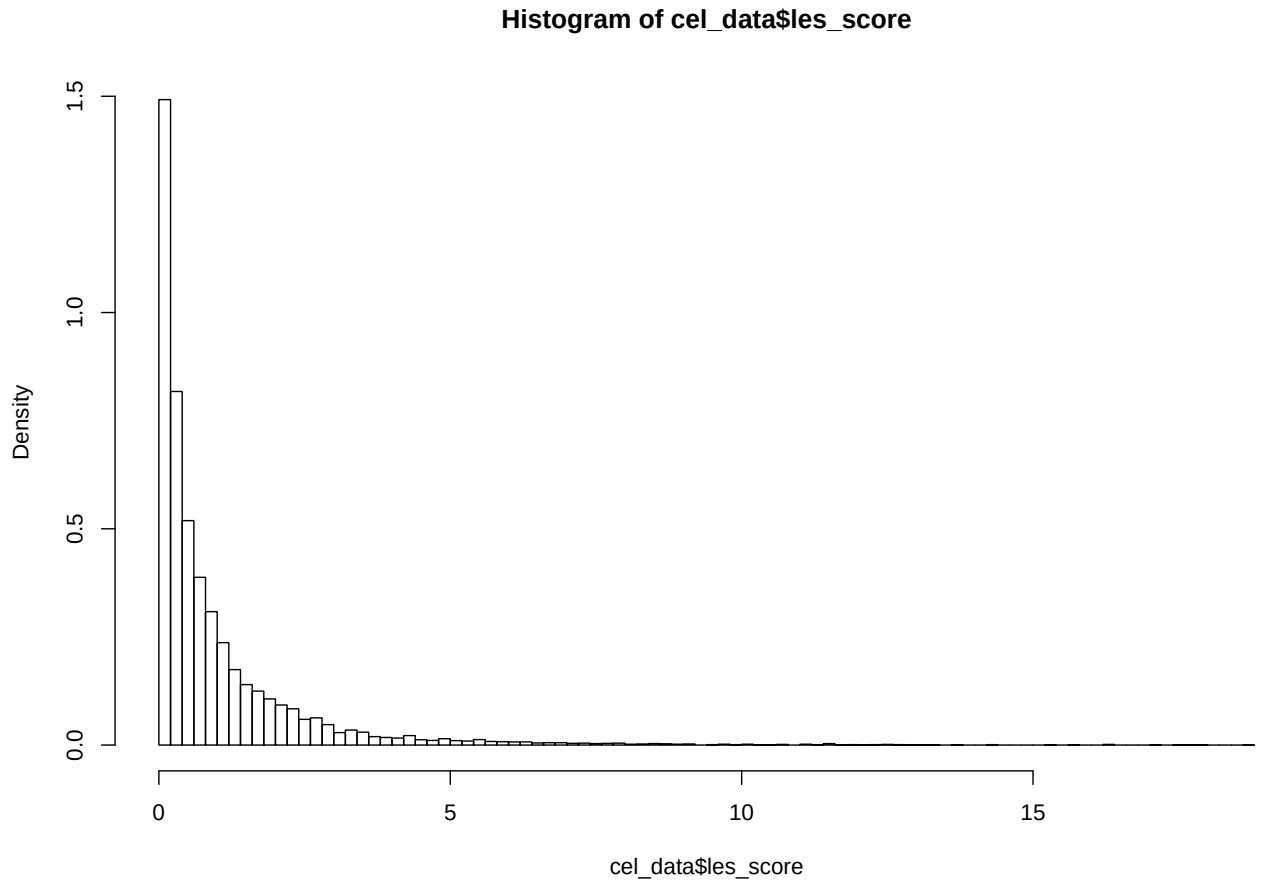
```
hist(cel_data$les_score, freq = FALSE)
```



- Each bin is **one** wide so the height is the **proportion**

Can vary the number of "bins" to make a smoother plot

```
hist(cel_data$les_score, breaks = 100, freq = FALSE)
```

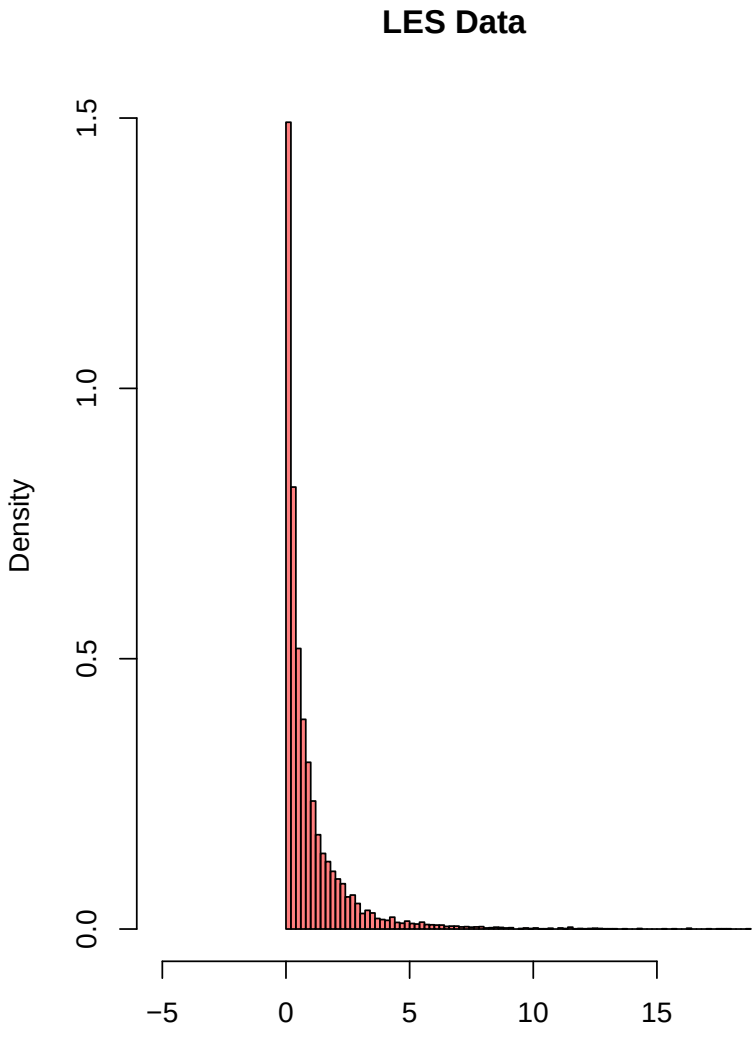


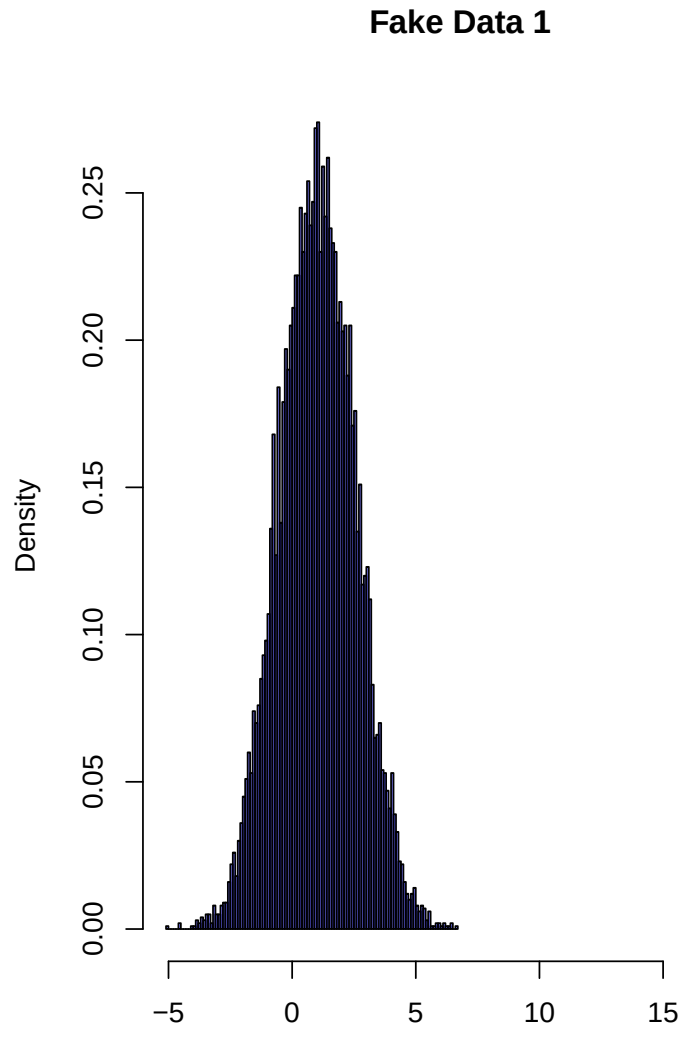
- Bins are 0.2 wide; proportion = "height" * 0.2
- First bin (0 to 0.2) is 1.492 tall; contains 0.3 proportion of data

A Picture is Worth 1,000 Words

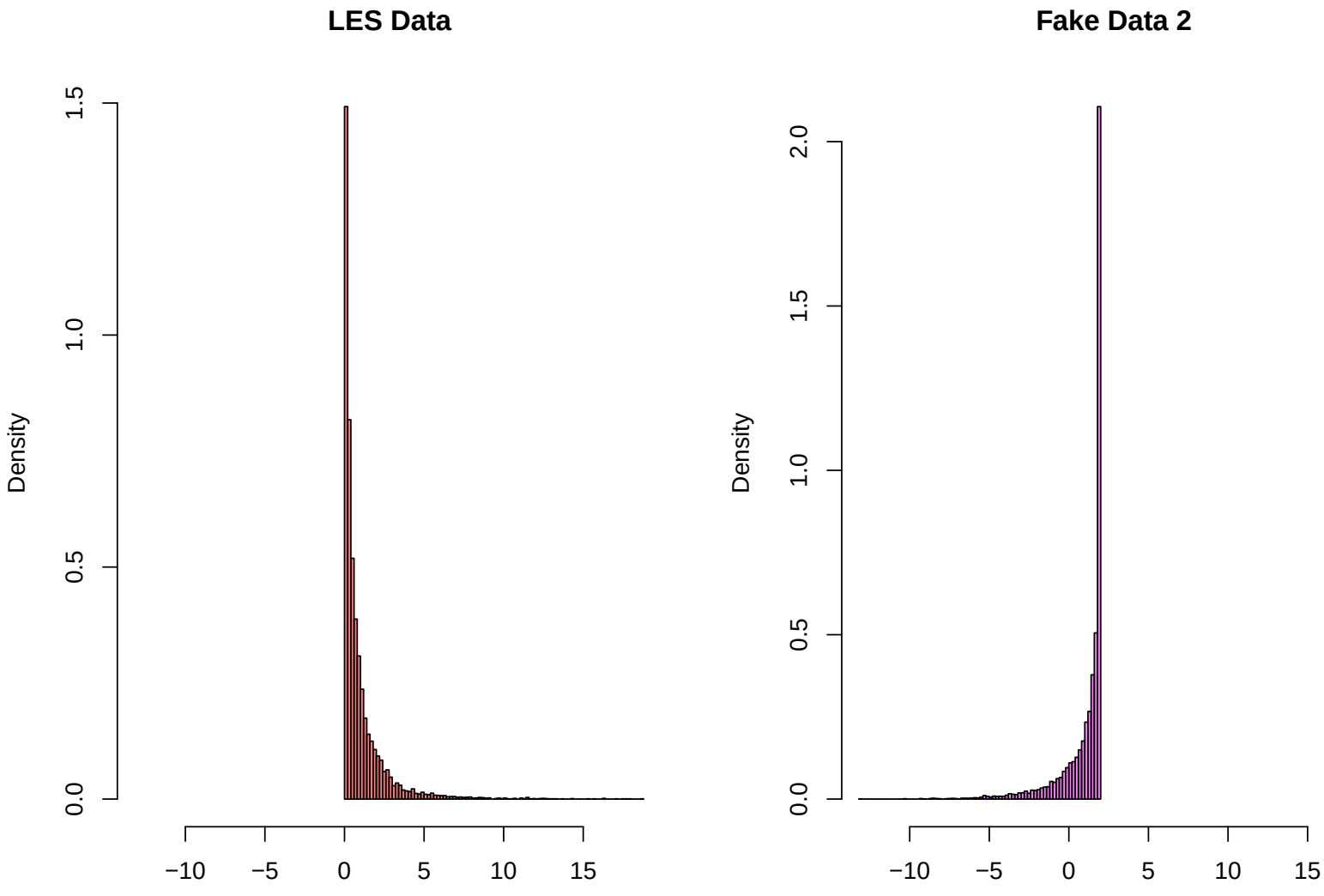
- We can see very clearly that the distribution is *skewed* or not symmetrical
- It is known as **right skewed** because it has a few very large values to the **right** of the main distribution
- The mean is **larger** than the median
- It suggests that we should use the **median** to summarize the distribution
  - The mean is sensitive to outliers
- The same "mean" and "standard deviation" can characterize very different distributions
- Always **look** at your data

Fake Data 1: Same Mean / Variance as LES





Fake Data 2: Same Mean / Variance as LES

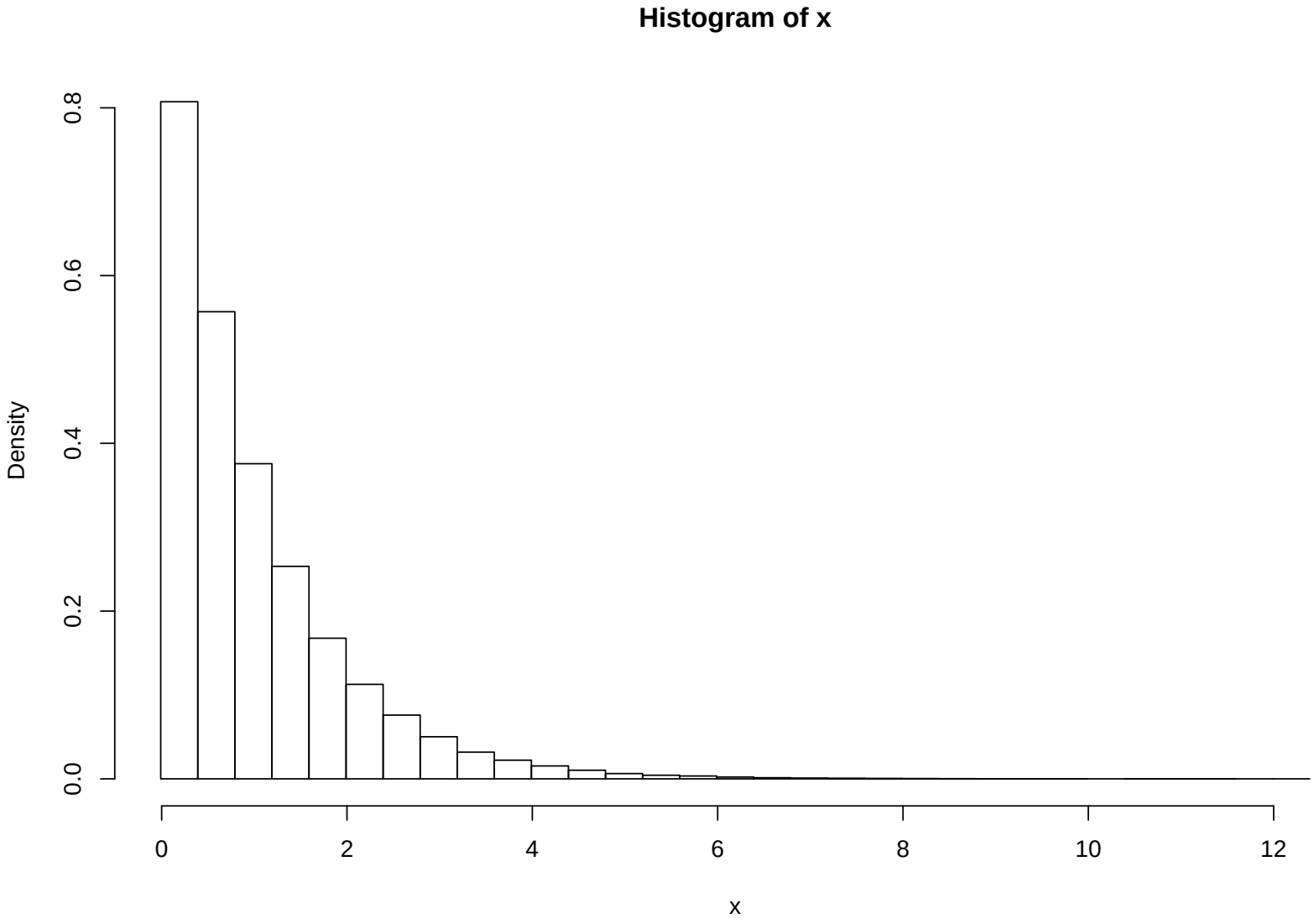


Questions

Practice Questions

- Consider the following data: [-10, 0, 10, 25, 100]
- What is the mean and median? Which do you think represents a better measure of the "typical value" and why?
- Look at the histogram on the next page: Noting that the bins are each 0.4 wide, approximately what proportion of the data is in the first bin?

Histogram



Answers

We can do the first part in R!

```
v <- c(-10, 0, 10, 25, 100)
mean(v)
```

```
## [1] 25
```

```
median(v)
```

```
## [1] 10
```

I think the median is better because 100 is a reasonably large outlier

The first bin is about 0.8 tall. Given the formula that height = proportion/width, we get that 0.8 = propotion / 0.4, so about 0.32 of the data is between 0 and 0.4.