PSY 503: Foundations of Statistical Methods in Psychological Science

Why Statistics?

Suyog Chandramouli

Zoom & 311 PSH (Princeton University)

8th September, 2024

What is statistics all about?

- Many different answers
- But all boils down to

  - “Statistics is about understanding variation”

- Discussion:
  - In your research studies/experiences, what types of variation have you come across?

What is variation?

- We can see variation in this class here.

What is variation?

- We can see variation in this class here.
  - You’re all different on many variables:
    - Hair color
    - Experience with statistics
    - What brand of computer you use
    - Where you went to high school
    - Height etc.

What is variation?

- We can see variation in this class here.
  - You're all different on many variables:
    - Hair color
    - Experience with statistics
    - What brand of computer you use
    - Where you went to high school
    - Height etc.
- Variation doesn't mean that everyone has to be different (e.g., some of you may have gone to the same university for undergrad).

What is variation?

- We can see variation in this class here.
  - You're all different on many variables:
    - Hair color
    - Experience with statistics
    - What brand of computer you use
    - Where you went to high school
    - Height etc.
- Variation doesn't mean that everyone has to be different (e.g., some of you may have gone to the same university for undergrad).
- Lack of variation is when everyone is the same on some variable.

Making sense of variation

- It is not easy to always explain the cause of some particular variation. The world is full of variation

Making sense of variation

- It is not easy to always explain the cause of some particular variation. The world is full of variation
- Smoking & Lung cancer

Making sense of variation

- It is not easy to always explain the cause of some particular variation. The world is full of variation
- Smoking & Lung cancer
  - **Late 19th century**: Increased cigarette use (due to mass production, marketing)

Making sense of variation

- It is not easy to always explain the cause of some particular variation. The world is full of variation
- Smoking & Lung cancer
  - **Late 19th century**: Increased cigarette use (due to mass production, marketing)
  - **1900s** : Lung cancer was considered rare. Smoking wasn't hypothesized to cause it.

Making sense of variation

- It is not easy to always explain the cause of some particular variation. The world is full of variation
- Smoking & Lung cancer
  - **Late 19th century**: Increased cigarette use (due to mass production, marketing)
  - **1900s** : Lung cancer was considered rare. Smoking wasn't hypothesized to cause it.
  - **1920s:** Several explanations including air pollution, Spanish flu, etc.

Making sense of variation

- It is not easy to always explain the cause of some particular variation. The world is full of variation
- Smoking & Lung cancer
  - **Late 19th century**: Increased cigarette use (due to mass production, marketing)
  - **1900s** : Lung cancer was considered rare. Smoking wasn't hypothesized to cause it.
  - **1920s:** Several explanations including air pollution, Spanish flu, etc.
  - **1940s**: Recognition of cigarette use as a very likely factor, based on epidemiology, controlled animal studies, etc.

Making sense of variation

- It is not easy to always explain the cause of some particular variation. The world is full of variation
- Smoking & Lung cancer
  - **Late 19th century**: Increased cigarette use (due to mass production, marketing)
  - **1900s** : Lung cancer was considered rare. Smoking wasn't hypothesized to cause it.
  - **1920s:** Several explanations including air pollution, Spanish flu, etc.
  - **1940s**: Recognition of cigarette use as a very likely factor, based on epidemiology, controlled animal studies, etc.
  - **1960s:** Propaganda by tobacco industry….

Making sense of variation

- It is not easy to always explain the cause of some particular variation. The world is full of variation
- Smoking & Lung cancer
  - **Late 19th century**: Increased cigarette use (due to mass production, marketing)
  - **1900s** : Lung cancer was considered rare. Smoking wasn't hypothesized to cause it.
  - **1920s:** Several explanations including air pollution, Spanish flu, etc.
  - **1940s**: Recognition of cigarette use as a very likely factor, based on epidemiology, controlled animal studies, etc.
  - **1960s:** Propaganda by tobacco industry….
- Controlled experiments help reduce/control variation in data (due to otherwise uncontrolled variables)

Making sense of variation

- It is not easy to always explain the cause of some particular variation. The world is full of variation
- Smoking & Lung cancer
  - **Late 19th century**: Increased cigarette use (due to mass production, marketing)
  - **1900s** : Lung cancer was considered rare. Smoking wasn't hypothesized to cause it.
  - **1920s:** Several explanations including air pollution, Spanish flu, etc.
  - **1940s**: Recognition of cigarette use as a very likely factor, based on epidemiology, controlled animal studies, etc.
  - **1960s:** Propaganda by tobacco industry….
- Controlled experiments help reduce/control variation in data (due to otherwise uncontrolled variables)
- Converging evidence, replications, etc. are helpful for establishing a claim.

Making sense of variation

- It is not easy to always explain the cause of some particular variation. The world is full of variation
- Smoking & Lung cancer
  - **Late 19th century**: Increased cigarette use (due to mass production, marketing)
  - **1900s** : Lung cancer was considered rare. Smoking wasn't hypothesized to cause it.
  - **1920s:** Several explanations including air pollution, Spanish flu, etc.
  - **1940s**: Recognition of cigarette use as a very likely factor, based on epidemiology, controlled animal studies, etc.
  - **1960s:** Propaganda by tobacco industry….
- Controlled experiments help reduce variation in data (due to otherwise uncontrolled variables)
- Converging evidence, replications, etc. are helpful for establishing a claim.
- Statistics looks to establish ***"general claims from limited (but sufficient) data"***

Statistics and Generalization

- Statistics can help establish “general claims from limited data”

Statistics and Generalization

- Statistics can help establish “general claims from limited data”
- Generalization is when we take what we learn from the data we have and extend it to situations we haven't directly observed

Statistics and Generalization

- Statistics can help establish “general claims from limited data”
- Generalization is when we take what we learn from the data we have and extend it to situations we haven't directly observed.
- Discussion:
  - What are some ways in which Statistics helps to make generalizations beyond immediate evidence?
    - i.e.
      - *we start with data about _____ and generalize to _____*

Statistics & Generalization (ROS)

- *we start with data about a sample and generalize to a population*

Statistics & Generalization (ROS)

- *we start with data about a sample and generalize to a population*

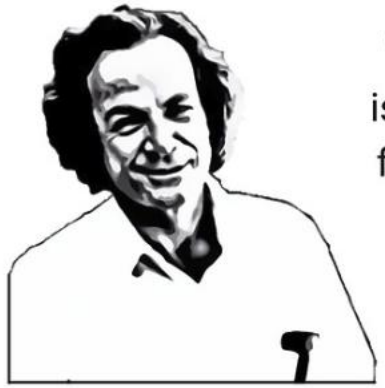
- *we start with data about a treatment and generalize to a control group OR to other possible treatment groups*

Statistics & Generalization (ROS)

- *we start with data about a sample and generalize to a population*
- *we start with data about a treatment and generalize to a control group OR to other possible treatment groups*
- *we start with data about a measurement and generalize to the underlying construct*

“Applying Common sense” and “gut instincts” to interpret data?

- Common sense is important. We care about expertise.
- But it is also easy to fool ourselves.

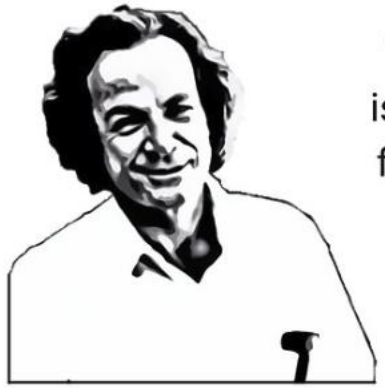


"The first principle is that you must not fool yourself — and you're the easiest person to fool."

Richard Feynman

“Applying Common sense” and “gut instincts” to interpret data?

- Common sense is important. We care about expertise.
- But it is also easy to fool ourselves.
- Belief bias:
  - when deciding whether an argument is valid, we’re influenced by believability

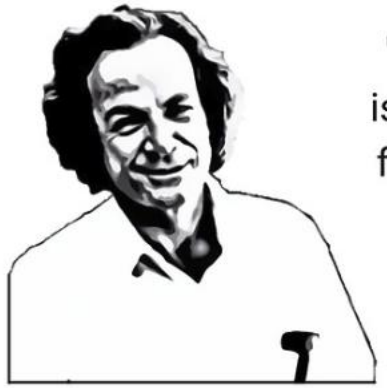


"The first principle is that you must not fool yourself — and you're the easiest person to fool."

Richard Feynman

“Applying Common sense” and “gut instincts” to interpret data?

- Common sense is important. We care about expertise.
- But it is also easy to fool ourselves.
- Belief bias:
  - when deciding whether an argument is valid, we’re influenced by believability



"The first principle is that you must not fool yourself — and you're the easiest person to fool."

Richard Feynman

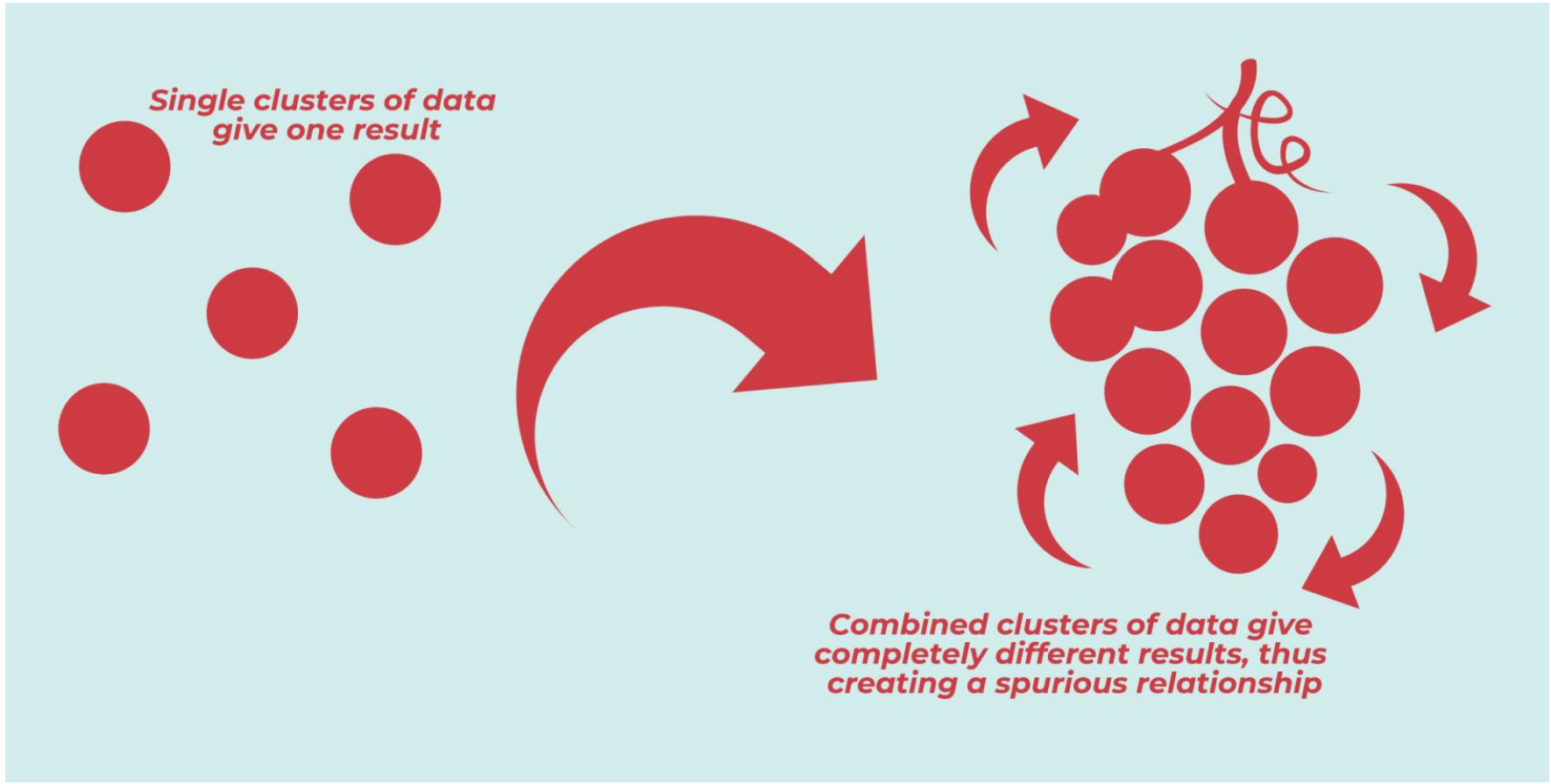
- Exercise: evaluate these statements (Which is valid?)

1. All cigarettes are expensive
Some addictive things are inexpensive
Therefore, some addictive things are not cigarettes

2. All addictive things are expensive
Some cigarettes are inexpensive
Therefore, some cigarettes are not addictive

Statistics / explaining variation requires careful thought too!

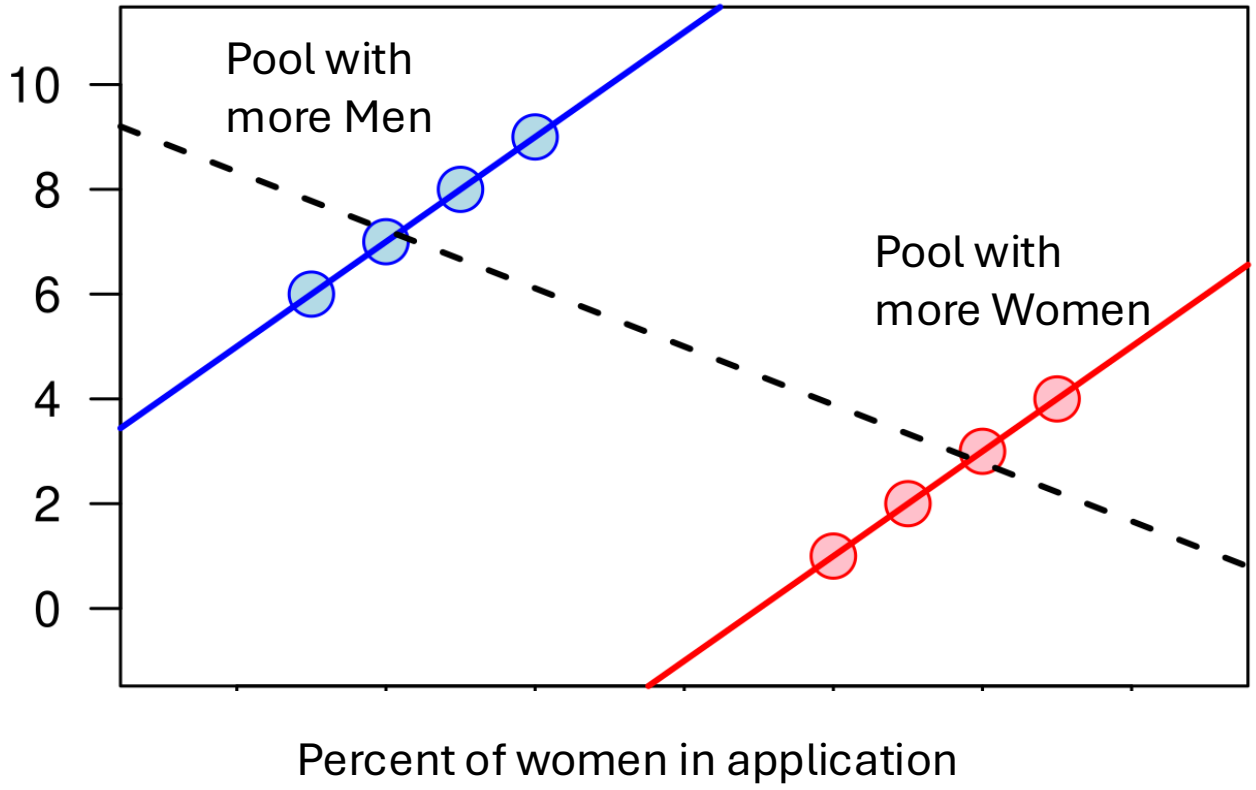
Statistics / explaining variation requires careful thought too!



- Simpson's paradox

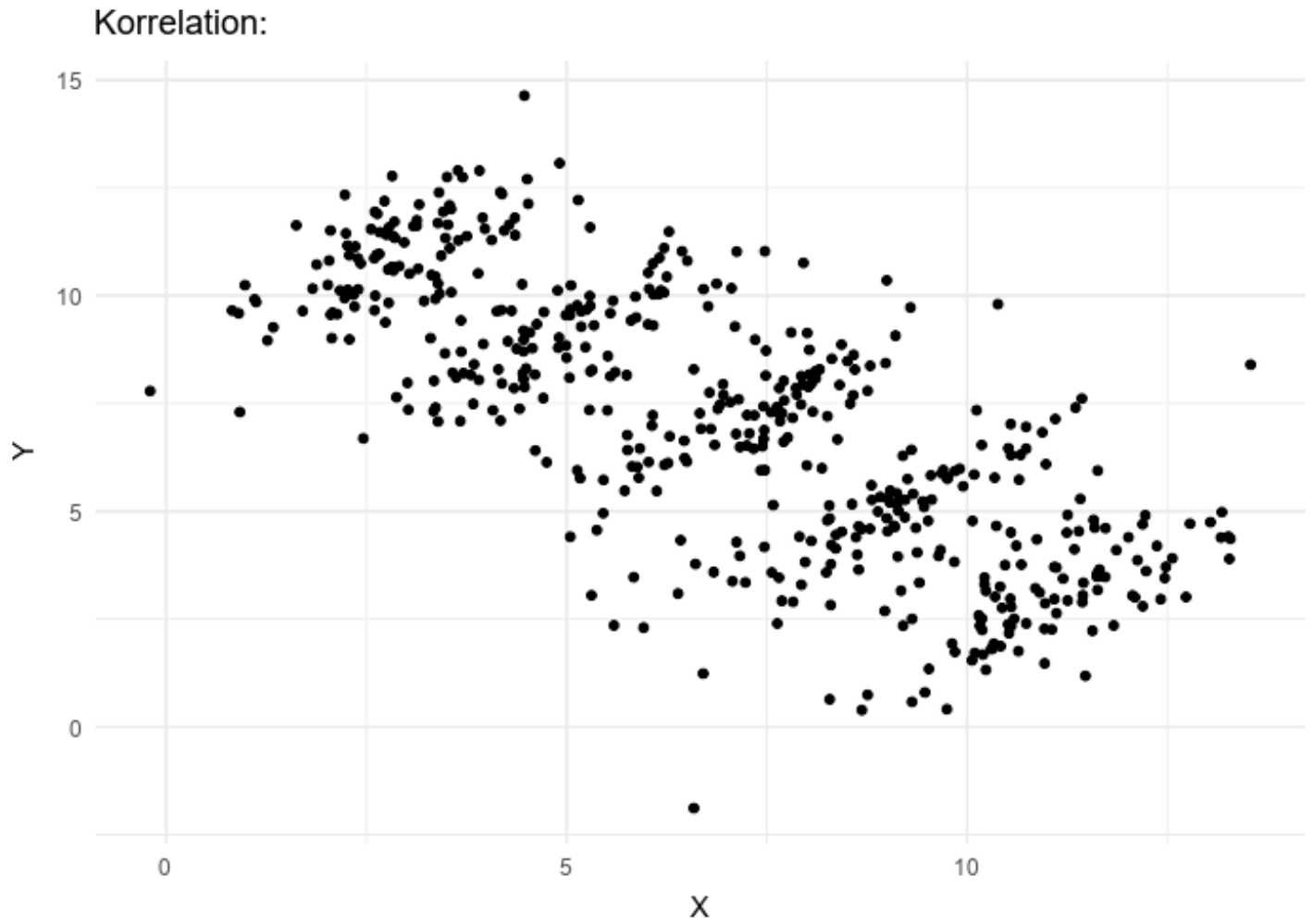
Example (Biased admissions in a university)

Example (Biased admissions in a university)

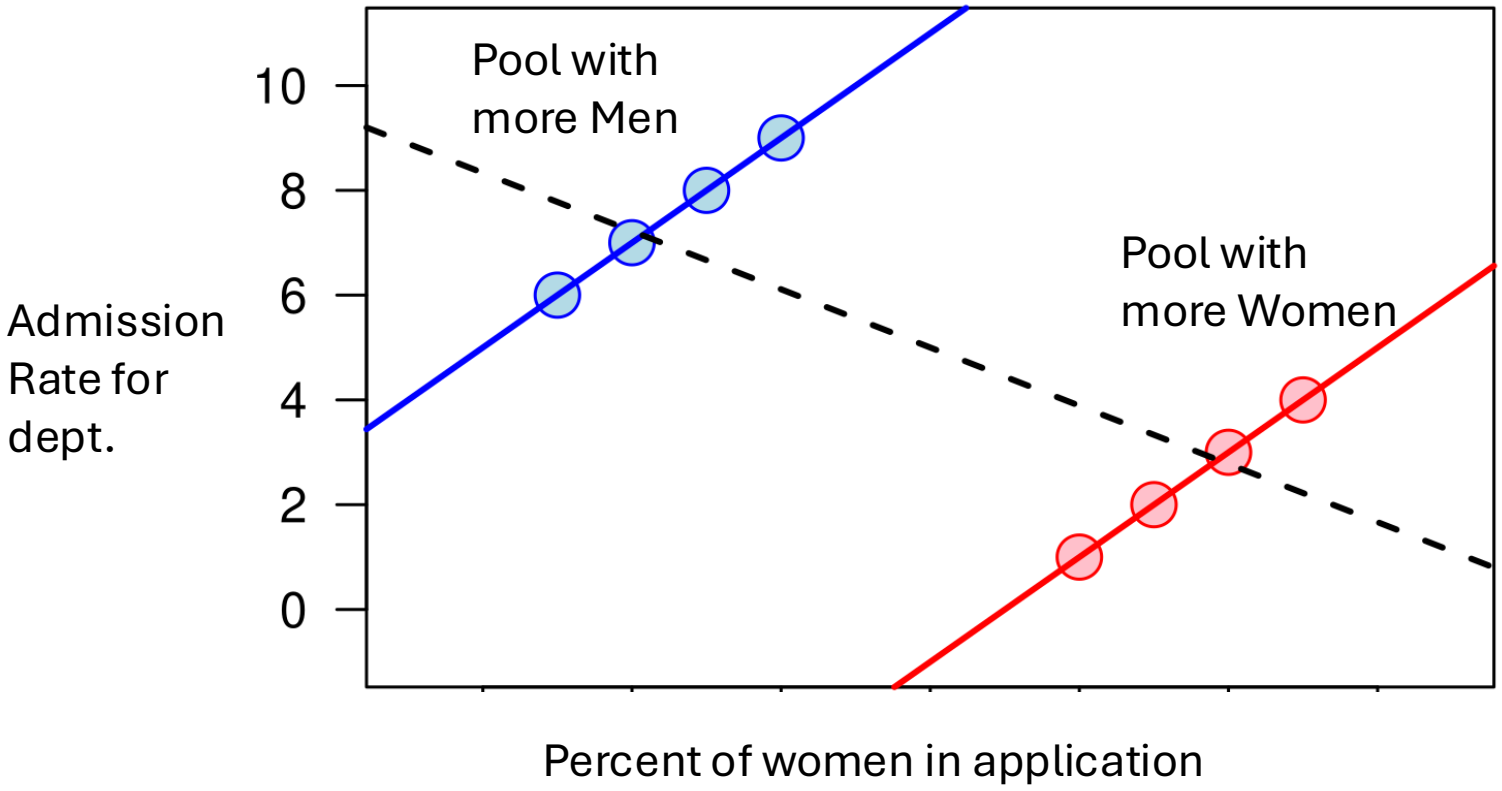


- Grouped data:
  - Admission rate favors women
- Pooled data
  - Admission rate disfavors women

More illustrative data pattern



Reason for the paradox



- Some departments have a larger intake
  - And polling results across the university ignores this aspect.

Instances of Simpson’s paradox

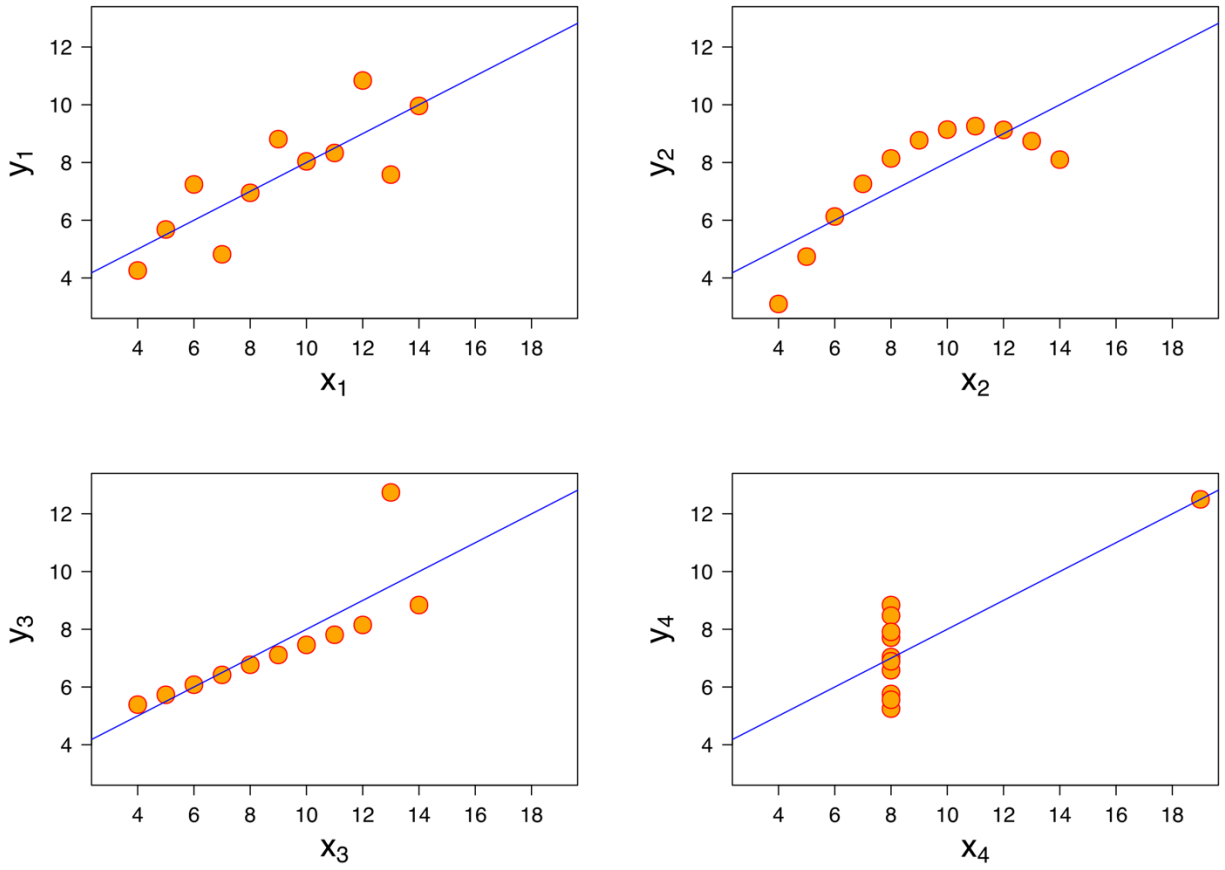
- Link

Instances of Simpson's paradox

- Link

- Looking at data / considering how we look at data is important

Anscombe's quartet



- All four sets have identical statistical parameters

Property	Value
Mean of x	9
Sample variance of x: s_x^2	11
Mean of y	7.50
Sample variance of y: s_y^2	4.125
Correlation between x and y	0.816
Linear regression line	$y = 3.00 + 0.500x$
Coefficient of determination of the linear regression: R^2	0.67

- Again, It's important to look at the data (not just summary statistics)

Exploratory Data Analysis

Exploratory Data Analysis

The goal during EDA is to develop an understanding of your data.

Exploratory Data Analysis

The goal during EDA is to develop an understanding of your data.

The easiest way to do this is to use questions as tools to guide your investigation. When you ask a question, the question focuses your attention on a specific part of your dataset and helps you decide which graphs, models, or transformations to make.

Exploratory Data Analysis

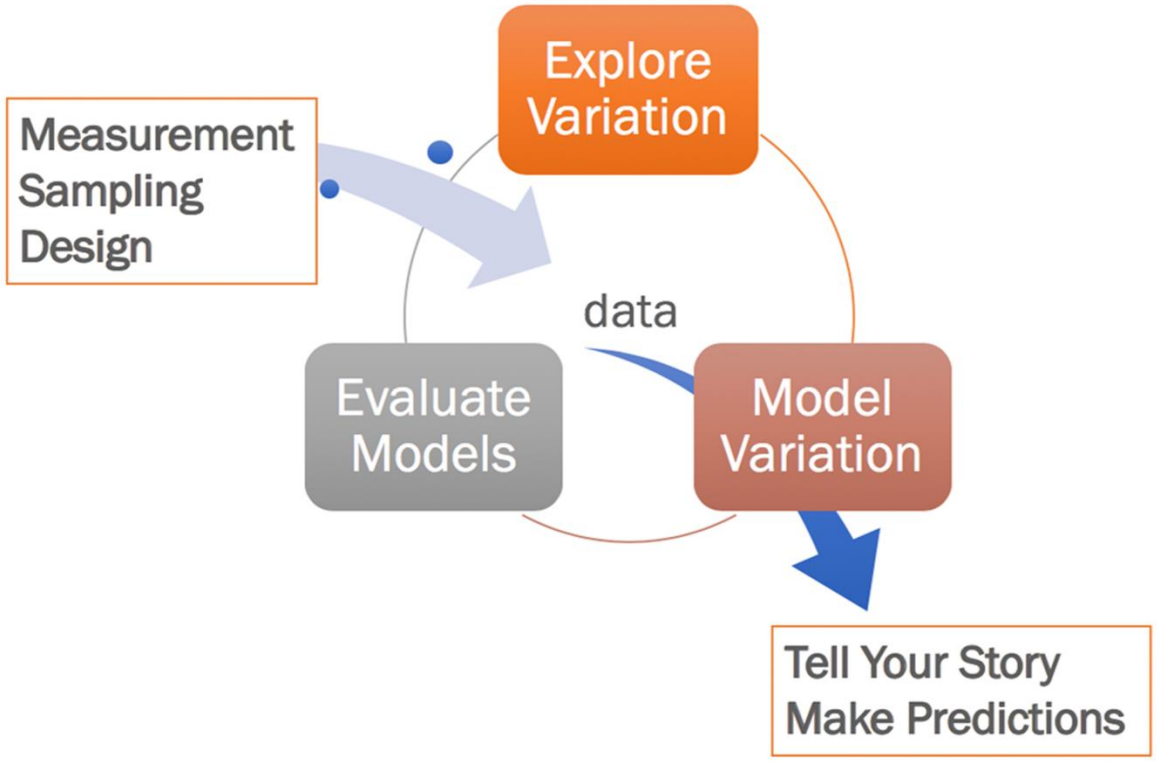
The goal during EDA is to develop an understanding of your data.

The easiest way to do this is to use questions as tools to guide your investigation. When you ask a question, the question focuses your attention on a specific part of your dataset and helps you decide which graphs, models, or transformations to make.

Examples:

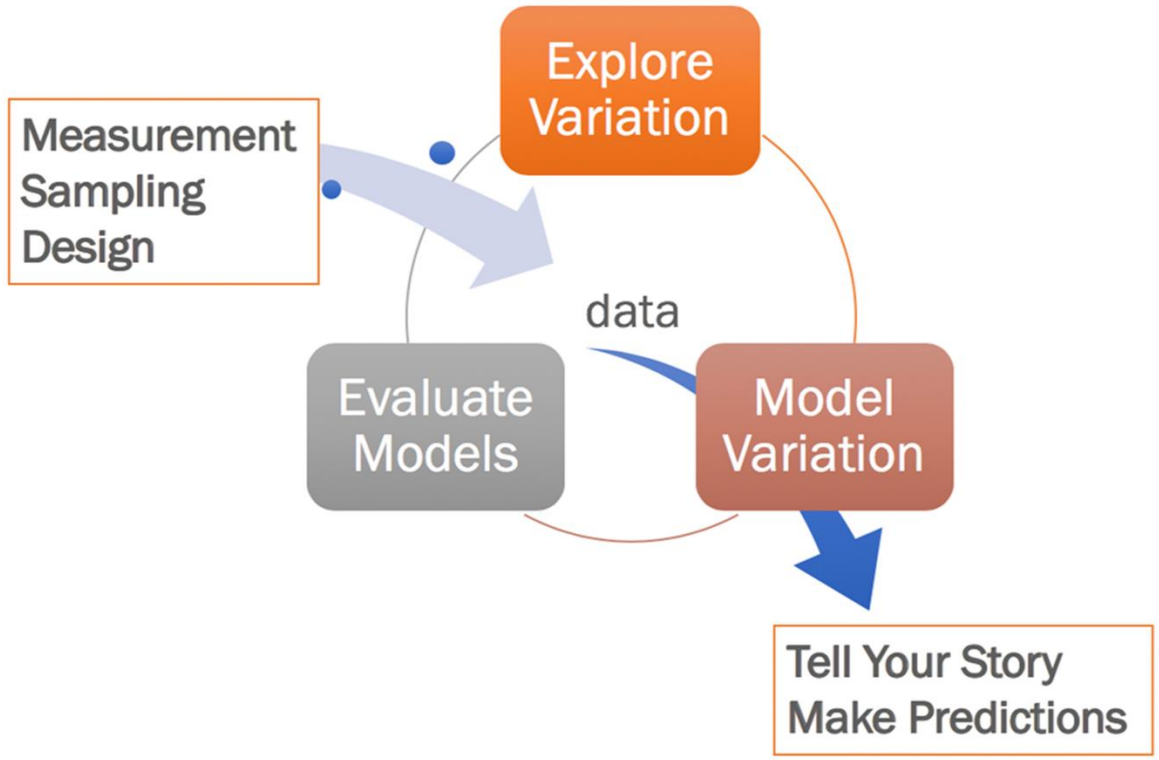
- What type of variation occurs within the variables?
- How do variables co-vary?
- Are there any unusual values in the data? What are salient / unexpected data patterns visible?

Statistics Cycle



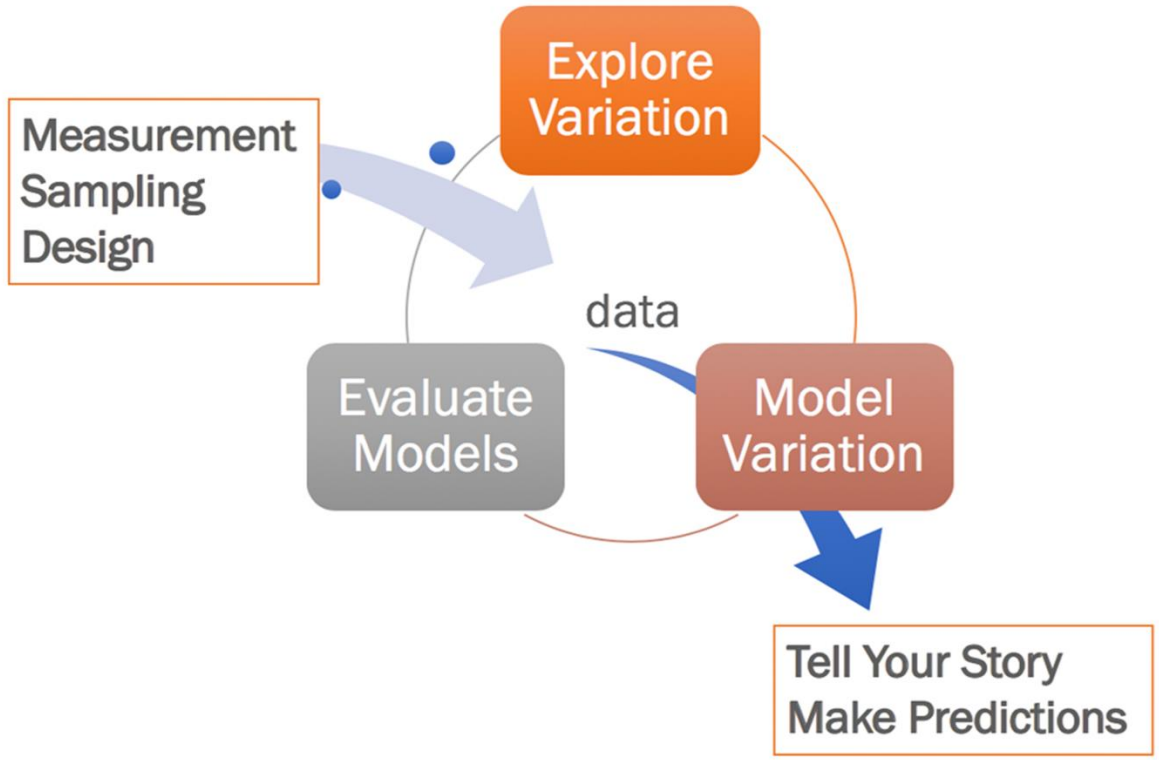
3 main parts to the course/ to statistics.

Statistics vs. Research Methods/ Design



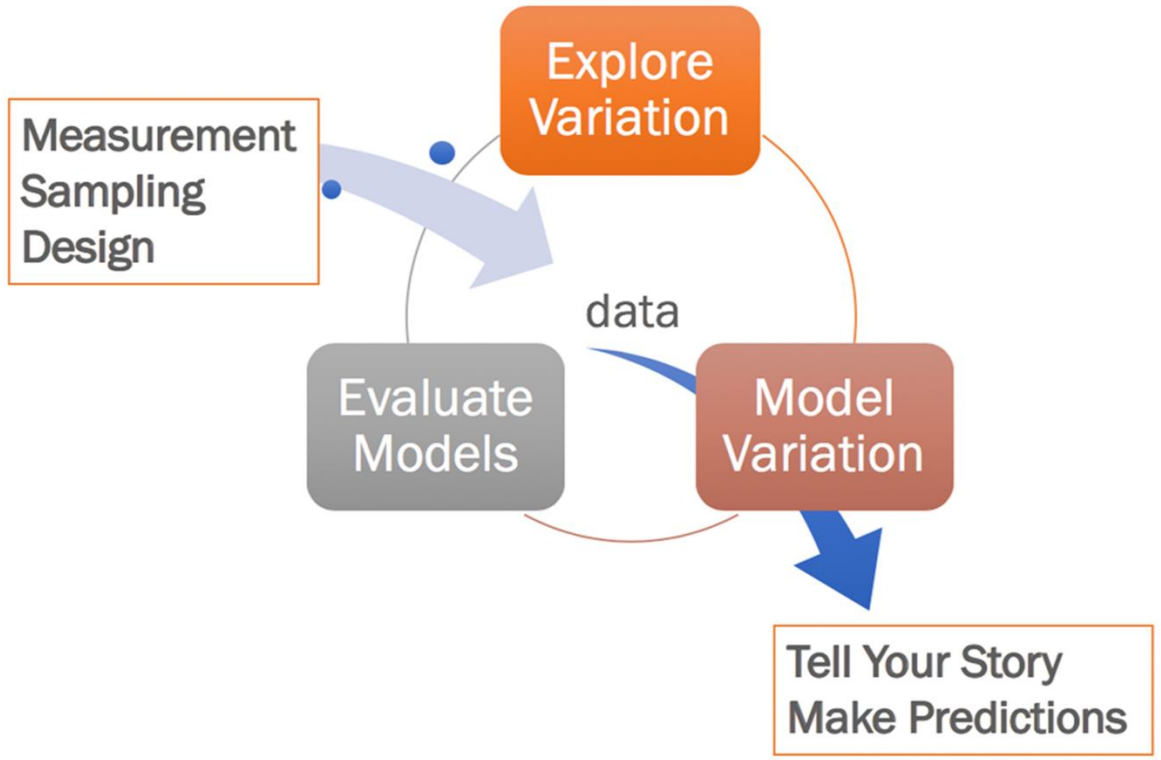
- Research design teaches you to:
  - how to plan and design experiments with good controls
  - how to collect data / good measurements, reduce bias in data collected

Statistics vs. Research Methods/ Design



- Research design teaches you to:
  - how to plan and design experiments with good controls
  - how to collect data / good measurements, reduce bias in data collected
- Statistics
  - what you might do with the data. That is, how to:
    - Process data,
    - Represent data, and
    - analyze data

Statistics vs. Research Methods/ Design



- Research design teaches you to:
  - how to plan and design experiments with good controls
  - how to collect data / good measurements, reduce bias in data collected
- Statistics
  - what you might do with the data. That is, how to:
    - Process data,
    - Represent data, and
    - analyze data
- Research design influenced the type of data collected influences the statistics we use
  - E.g. Likert Scale data vs. Response Time data

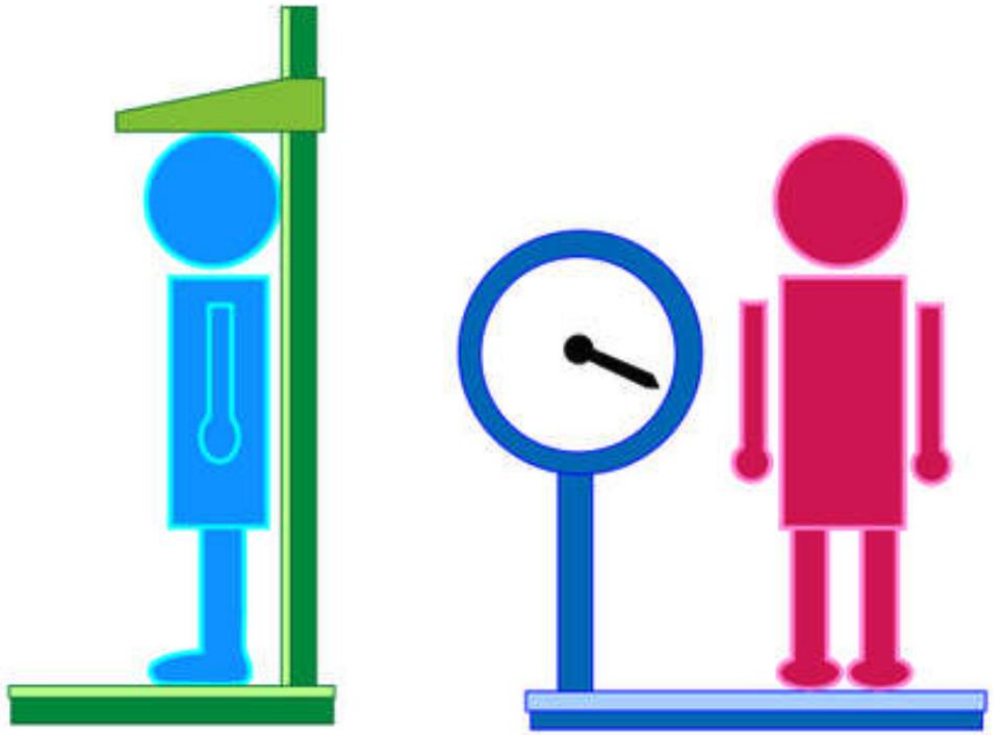
Measurements

Core assumption:

"Whatever exists at all exists in some amount. To know it thoroughly involves knowing its quantity as well as its quality." -Edward L. Thorndike

What is measurement?

The assignment of scores so that the scores represent some characteristic of the individuals



What may we want to measure?

- Depression
- Effort
- Intelligence
- Memory
- Social support
- Extroversion
- Eating behavior
- Parent child relationships
- Attention
- Burn out
- Hopelessness

What may we want to measure?

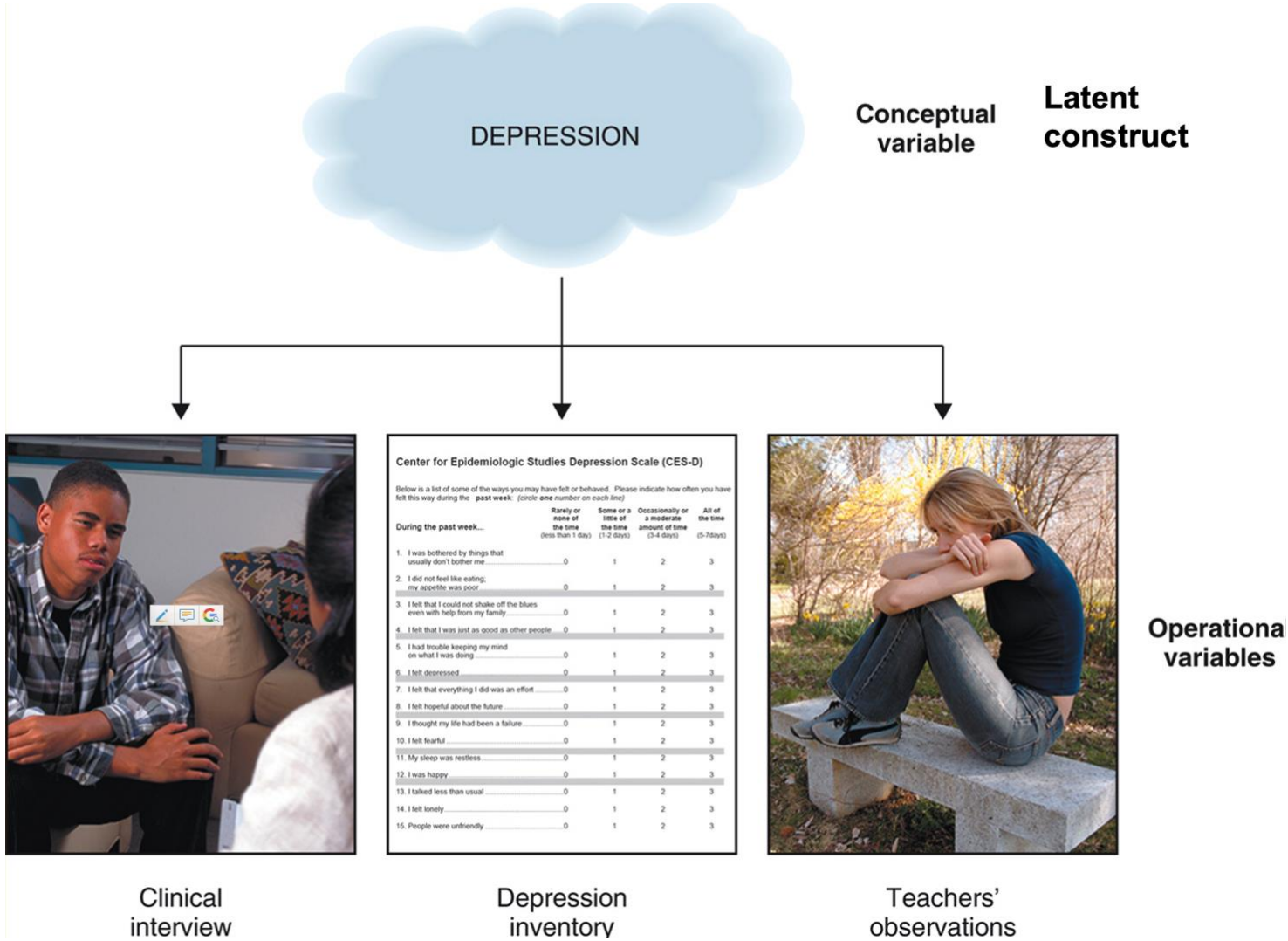
- Depression
- Effort
- Intelligence
- Memory
- Social support
- Extroversion
- Eating behavior
- Parent child relationships
- Attention
- Burn out
- Hopelessness

These are constructs.

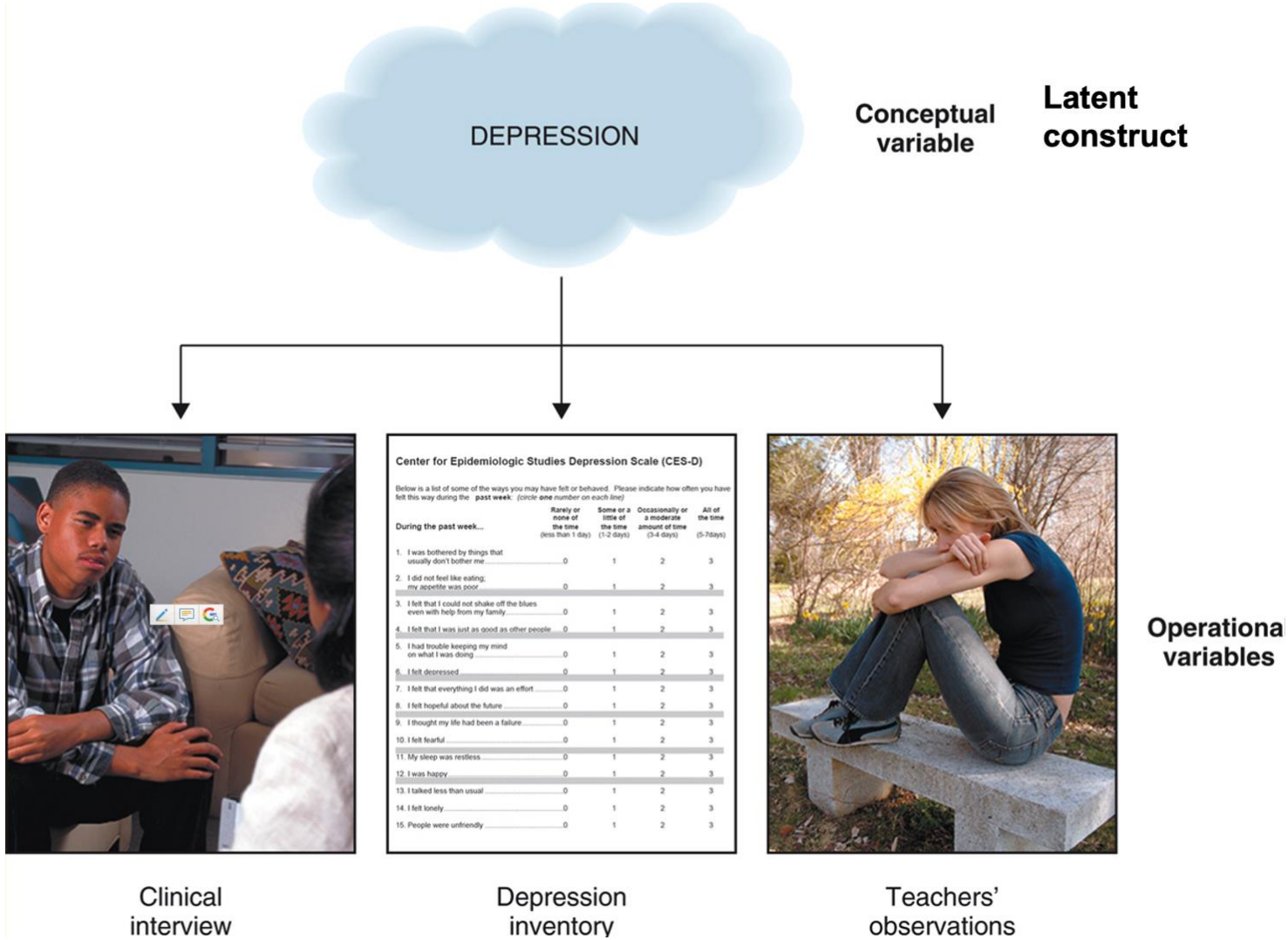
Constructs are “operationalized” by measurements

We are hoping that our measurements generalize well to the underlying construct.

Measurements & Constructs



Measurements & Constructs



What are some desirable properties of operationalized measurements?

Types of Measurements

Depression
Effort
Intelligence
Memory
Social support
Extroversion
Eating behavior
Parent child relationships
Attention
Burn out
Hopelessness

Types of Measurements

Depression
Effort
Intelligence
Memory
Social support
Extroversion
Eating behavior
Parent child relationships
Attention
Burn out
Hopelessness

You're all different on many variables:

- Hair color
- Experience with statistics
- What brand of computer you use
- Where you went to high school
- Height etc.

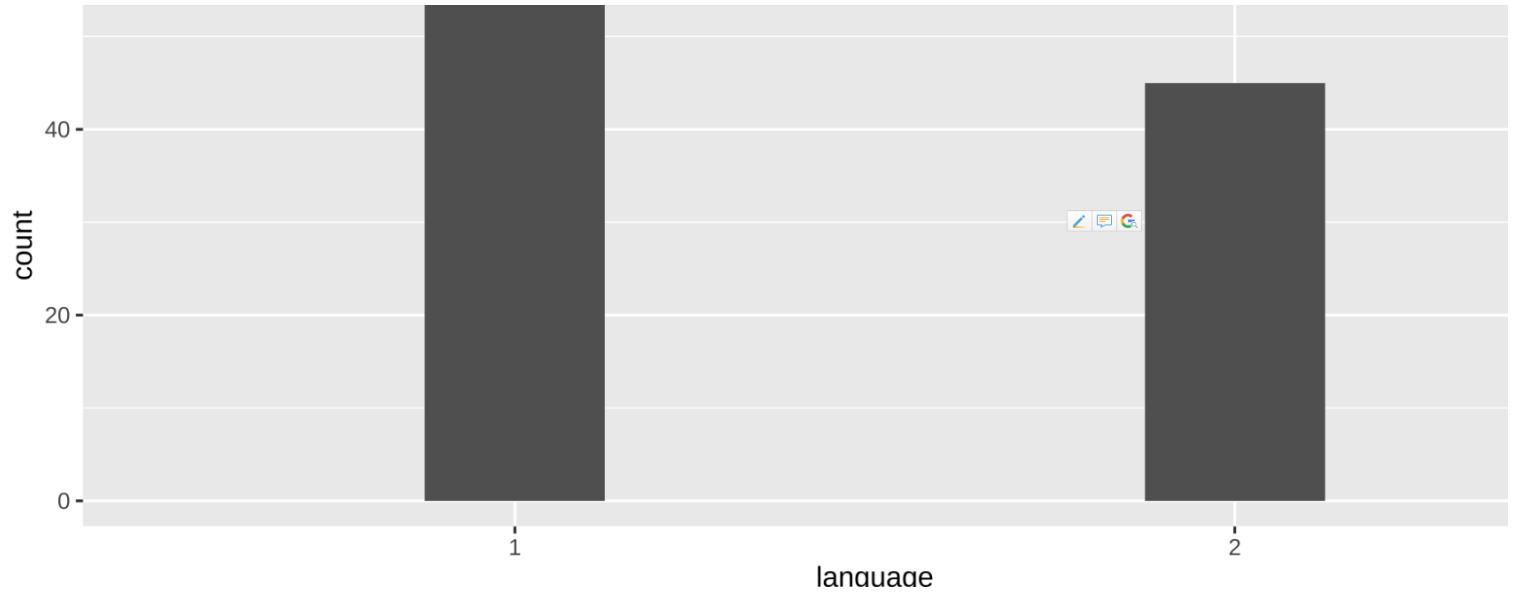
Types of Measurements -- Jargon

- Four ways of categorizing them:
  - **N**ominal
  - **O**rdinal
  - **I**nterval
  - **R**atio

- Why bother about this jargon?
  - It influences how you do stats on this measure
  - It influences how you store these variables in your (R) code

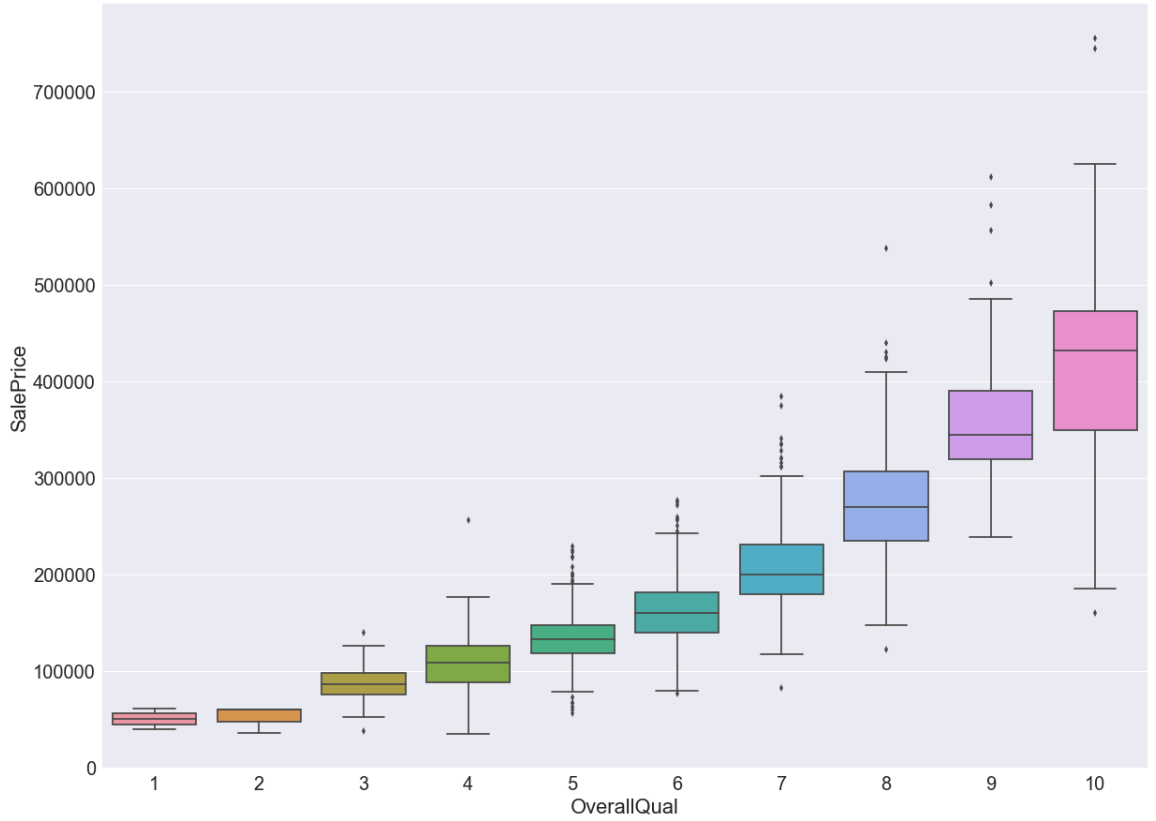
Nominal

- Nominal ≈ ***name*** ≈ categorical
  - No rank order
  - Often Presented as frequency counts
    - Stats are often on frequency counts
- Example:
  - Gender/sex, eye color



Ordinal

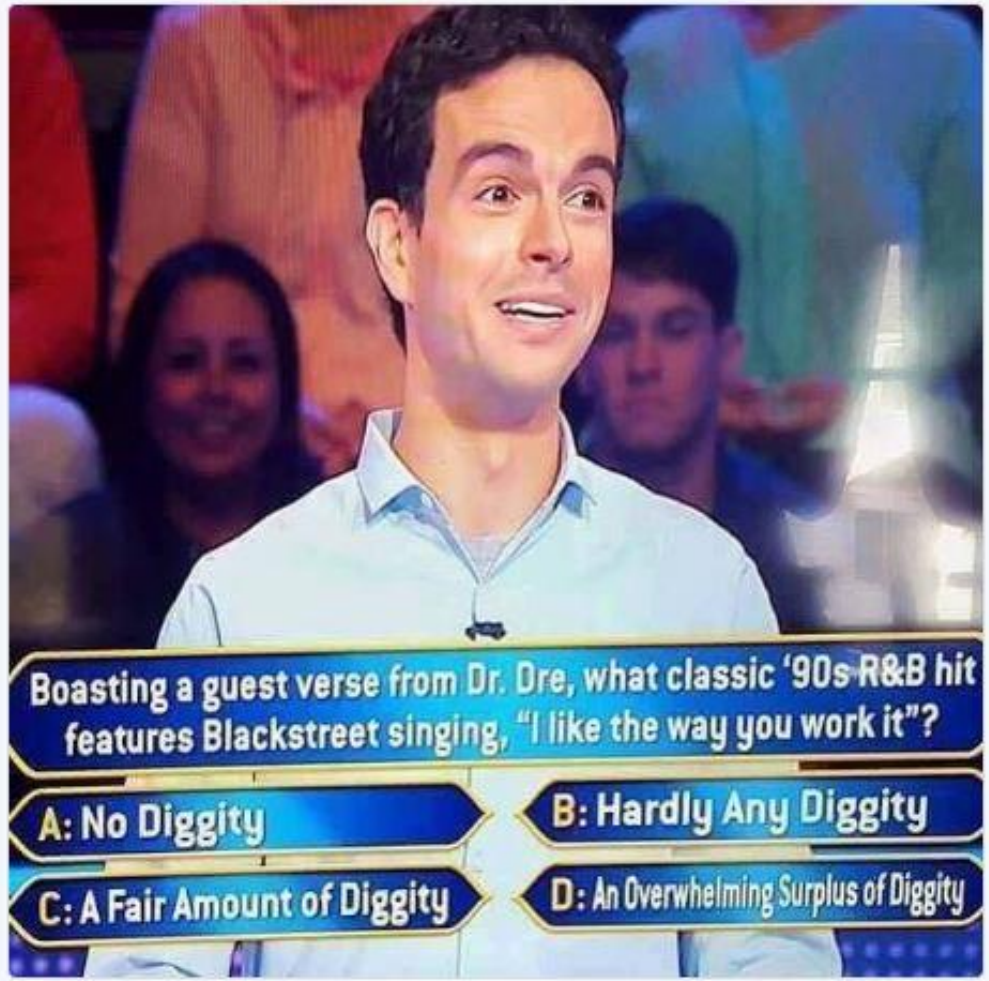
- Nominal+
  - Scores indicating rank
  - Increase/decrease (**i.e. order**) is meaningful
    - Comparative information (1 > 2 > 3)
    - No magnitude information (1 > 2 != 2 > 3)
- Most statistical techniques inappropriate



PsychBrief Retweeted

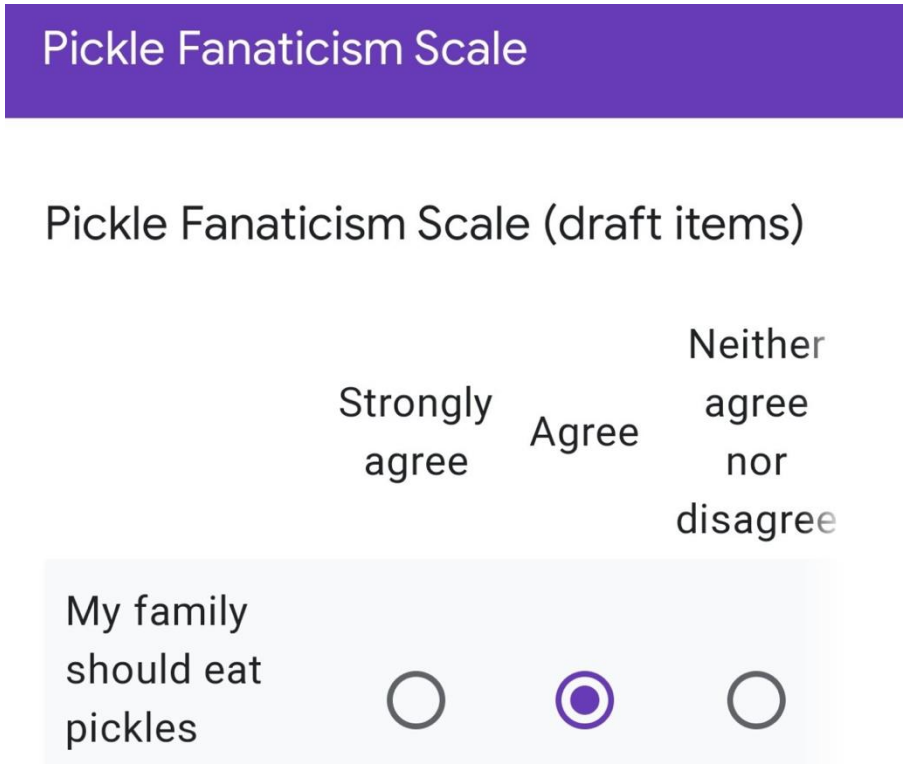
[title] ♥ .(.Zannah.). [surname] @zannah · 9h

This is the scale that I will be using for everything from now on.

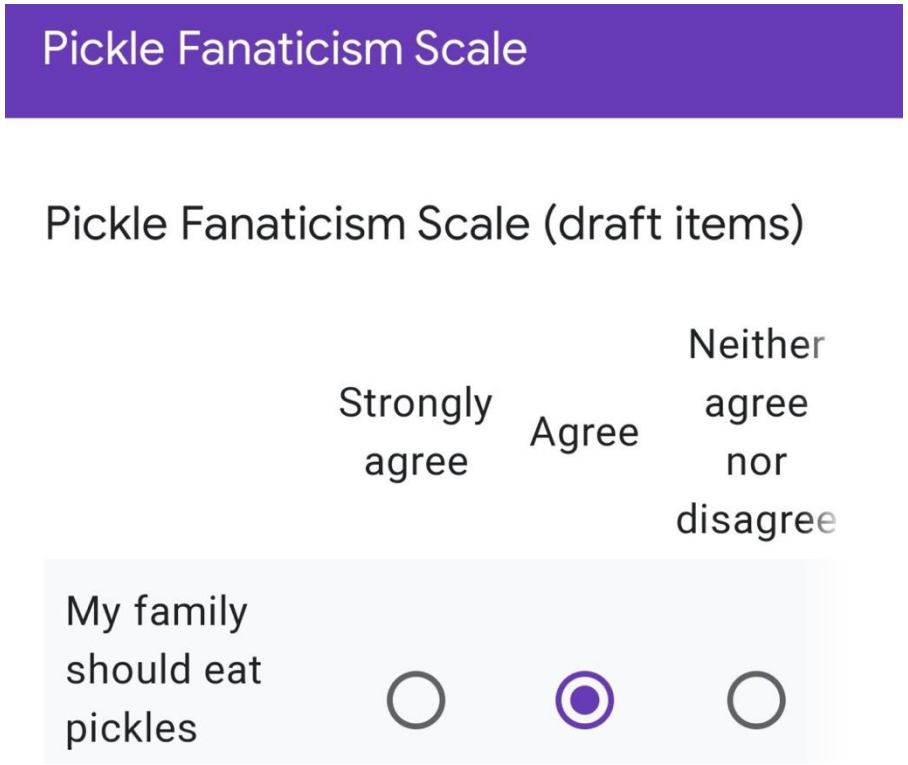


50 2.6K 6.1K

Likert Scale



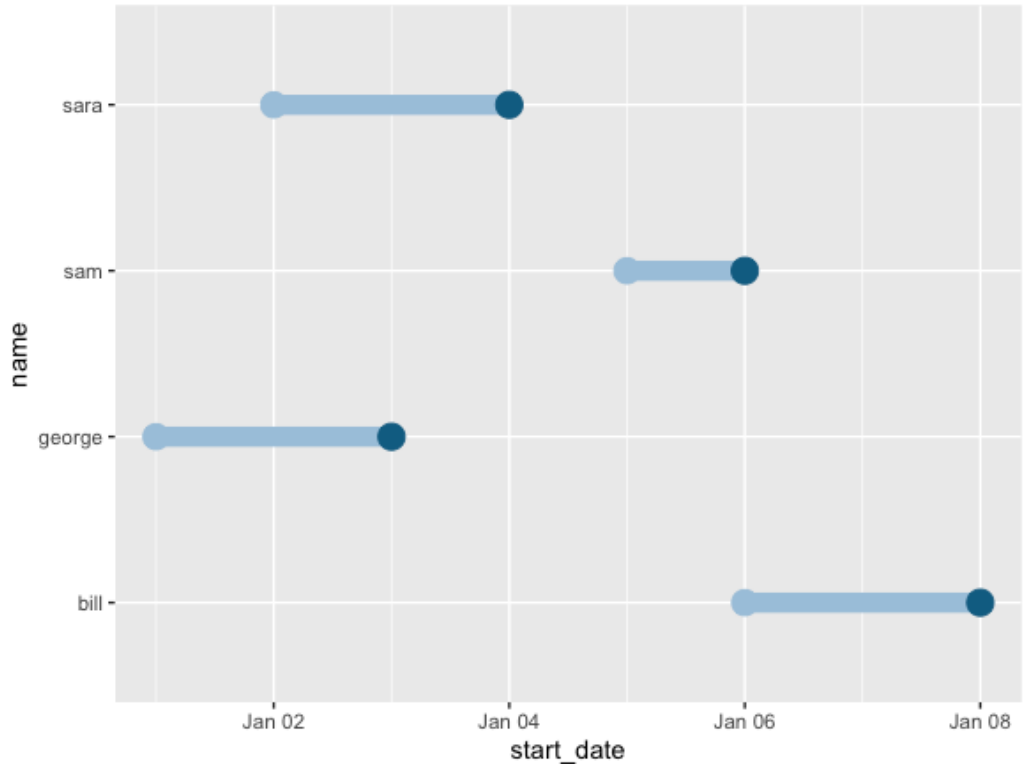
Likert Scale



Nominal or Ordinal ?

Interval

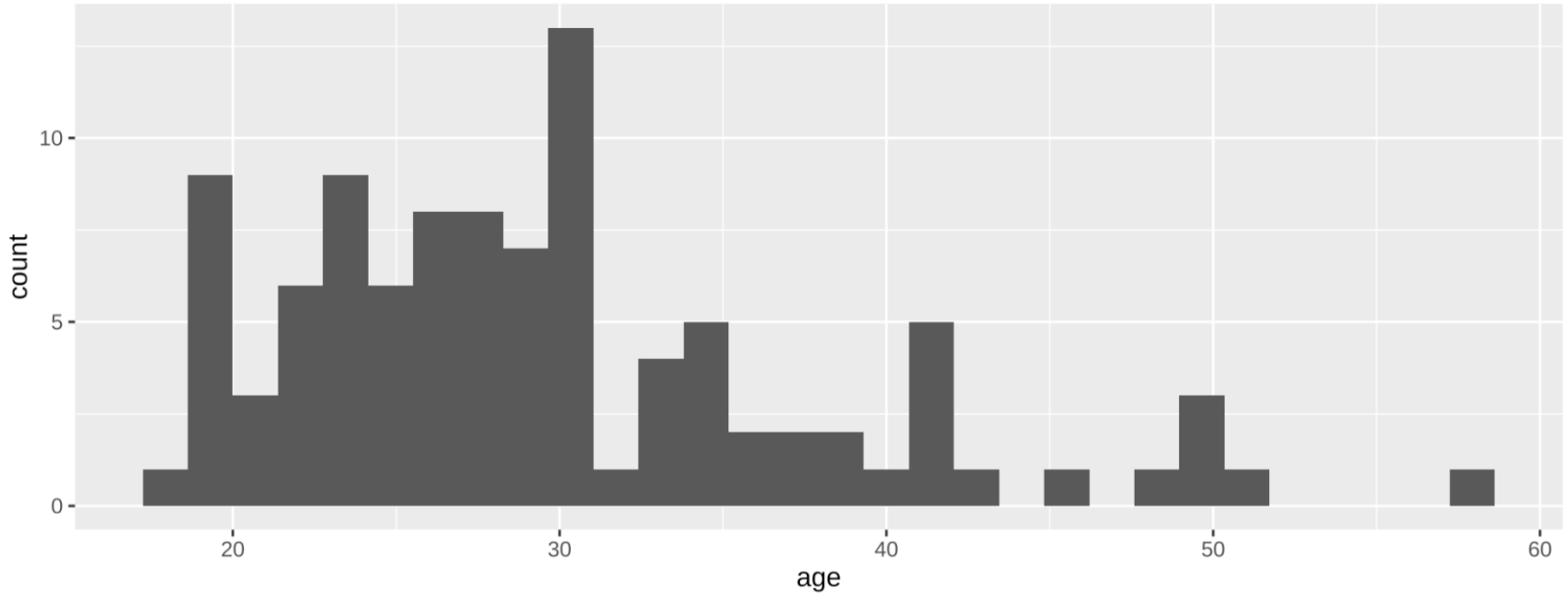
- Differences in adjacent ***intervals*** are equal
- Numeric, but lacks true 0;
  lacks multiplicative properties
- Examples:
  - Temperature (Celsius/ Fahreneit)
  - Time



Ratio

- All properties of interval plus:
  - True zero
  - Ratios are meaningful

- Examples:
  - Reaction times
  - Test Performance



Olympic example

Rank by Gold	Country	# of Athletes	# of Gold Medals
1	USA	1,200	46
2	People's Republic of China	600	38
3	Great Britain	1,500	29

- Different variables shown here:
  - Rank
  - Country
  - Number of athletes
  - Number of Gold medals

Levels of Measurement

Nominal	Ordinal	Interval	Ratio
"Eye color"	"Level of satisfaction"	"Temperature"	"Height"
Named	Named	Named	Named
	Natural order	Natural order	Natural order
		Equal interval between variables	Equal interval between variables
			Has a "true zero" value, thus ratio between values can be calculated

Information content in measurement

- Information content
  - Nominal -> Ordinal -> Interval -> Ratio
  - Can always go backwards but never forward
  - Higher level = higher sensitivity = higher power

Continuous or Discrete?

nominal & **ordinal** are always *discrete*

interval & **ratio** can be *discrete* OR *continuous*

If you know a variable is *discrete*, what does that tell you about its measurement level?

What if the variable is *continuous*?

Questions?

Finis