

EDP308: STATISTICAL LITERACY

The University of Texas at Austin, Fall 2020

RAZ: Rebecca A. Zárate, MA

Overview

- Measurement
 - ▣ Conceptualize
 - ▣ Operationalize
- Qualitative vs. Quantitative Data
- Variables and Data Types
 - ▣ Qualitative: Categorical
 - Nominal, Ordinal
 - ▣ Quantitative: Discrete or Continuous
 - Interval, Ratio
- Data Type Examples
- Types of Data in R

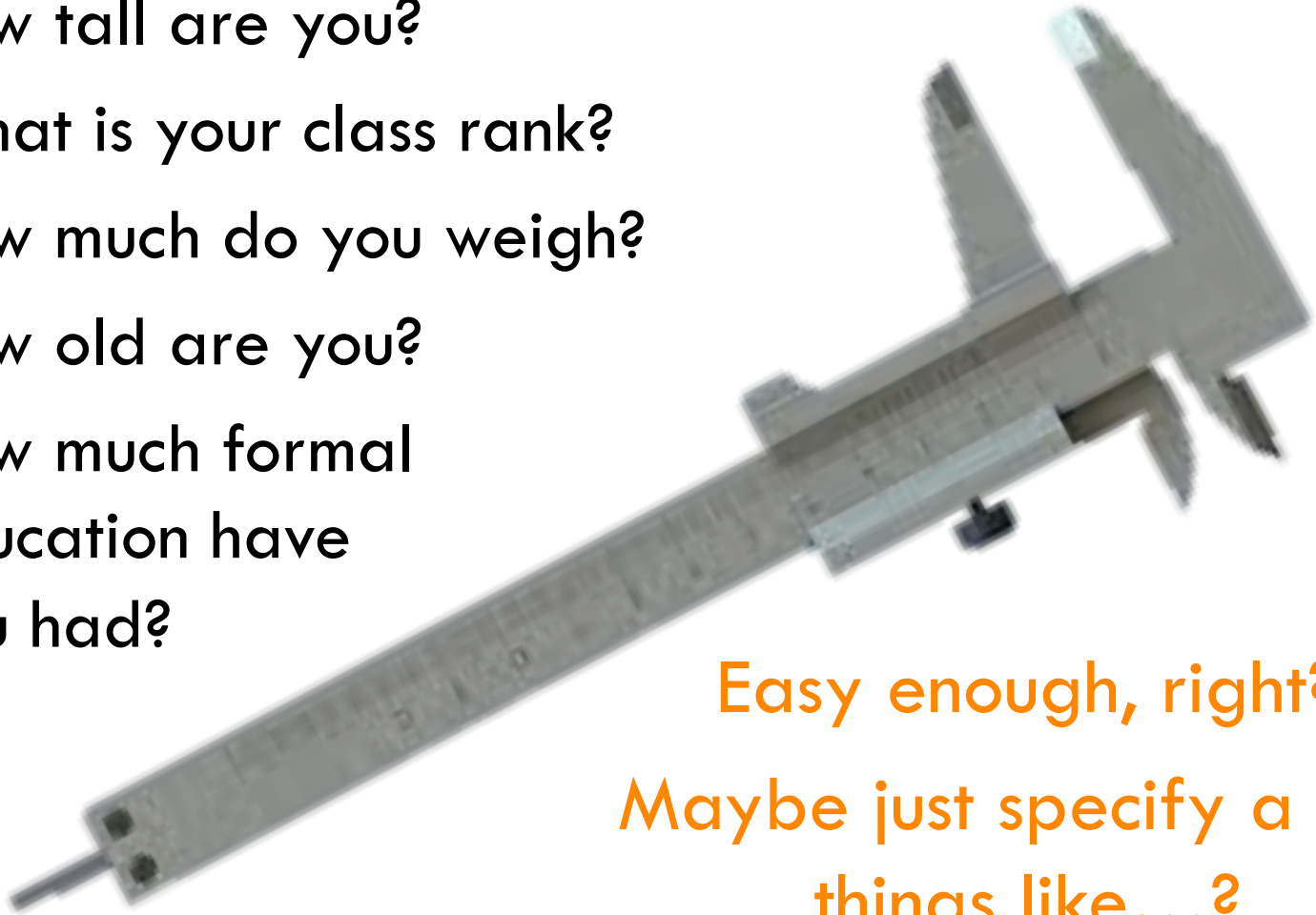
Measurement

Varying Variables

- What is a variable (noun)?
 - ▣ “A **variable** is any characteristics, number, or quantity that can be measured or counted. A **variable** may also be called a data item (or element). Age, sex, business income and expenses, country of birth, capital expenditure, class grades, eye color and vehicle type are examples of **variables**”
 - ▣ They are very variable (adjective)...

Give Me a Measurement of Some Variables

- How tall are you?
- What is your class rank?
- How much do you weigh?
- How old are you?
- How much formal education have you had?

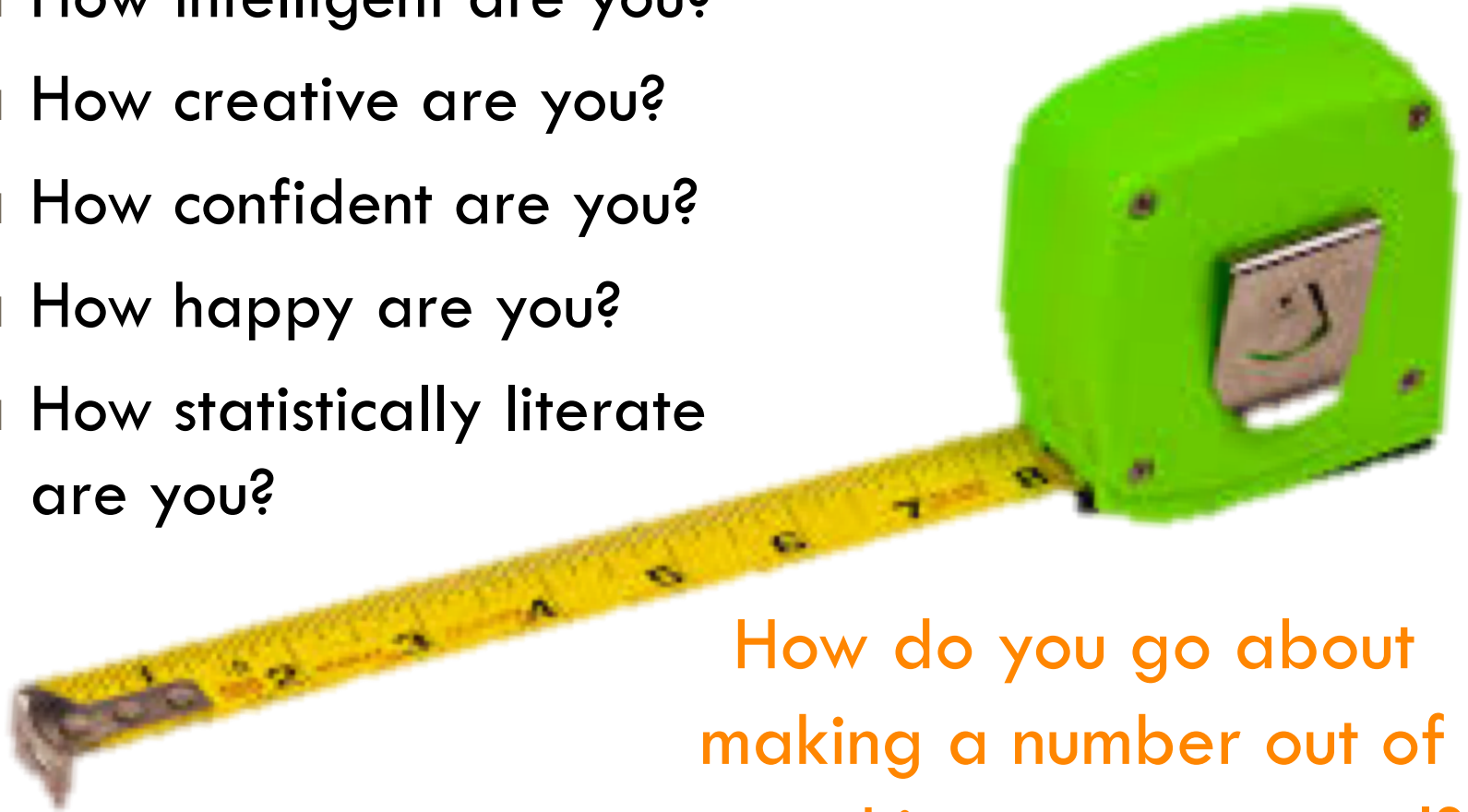


Easy enough, right?

Maybe just specify a few things like...?

Now, Give Me a Measurement.

- How intelligent are you?
- How creative are you?
- How confident are you?
- How happy are you?
- How statistically literate are you?



How do you go about making a number out of something so conceptual?

What is it the concept of...

Intelligence?

Creativity?

Confidence?

Happiness?

Statistical Literacy?

Conceptualization

- Conceptualization is the defining of your variables of interest and ideas into concrete definitions and constructs, thus enabling you to convey your ideas of what a particular concept means to you.
 - For example,
 - “Intelligence is the ability to acquire and apply knowledge”
 - “Creativeness is the ability to think and create new ideas”
 - “Confidence is a feeling of self-assurance arising from one’s appreciation of one’s own abilities or qualities”
 - “Happy is feelings feeling or showing pleasure or contentment”
 - “Statistical literacy is the ability to understand the gathering and analyzing data, as well as interpreting analysis results in order to critically evaluate findings reported in the media and in social science research.”

What's next?

Alright, you've conceptualized your concept. Now what?

Concepts to Operations

How do you go from a
concept to a measurable thing?

What steps would you need to take?

What questions would you need to ask?

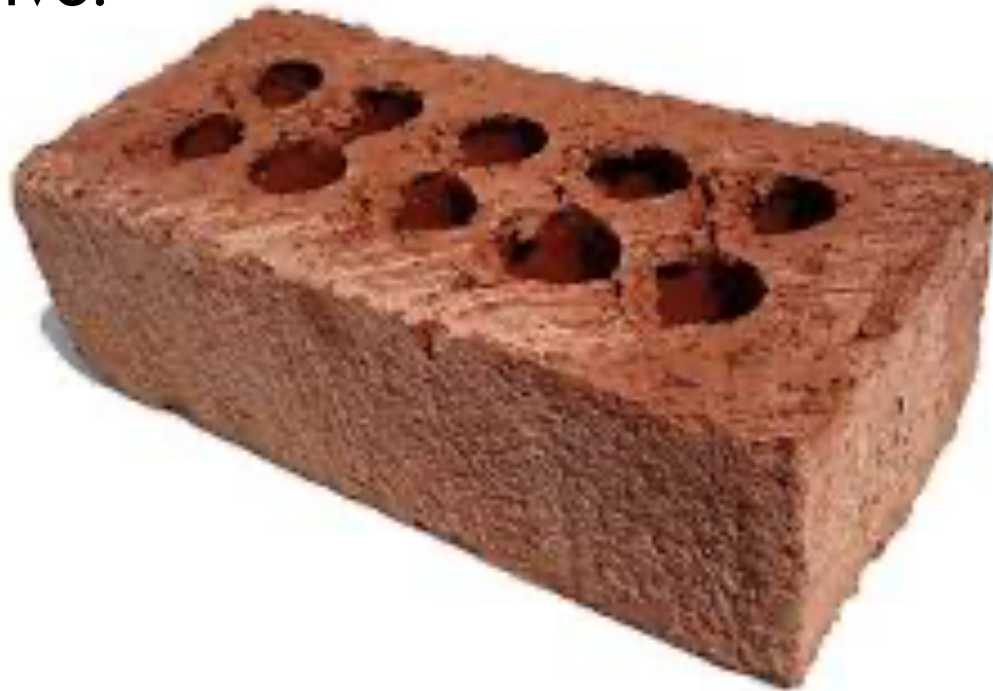
Operationalization

- “A process of defining the measurement of a phenomenon that is not directly measurable... Defining what is and is not a part of the concept.”
- Operationalization is the refining of your concept to a level of clarity that you can measure it.

How could we measure creativity?

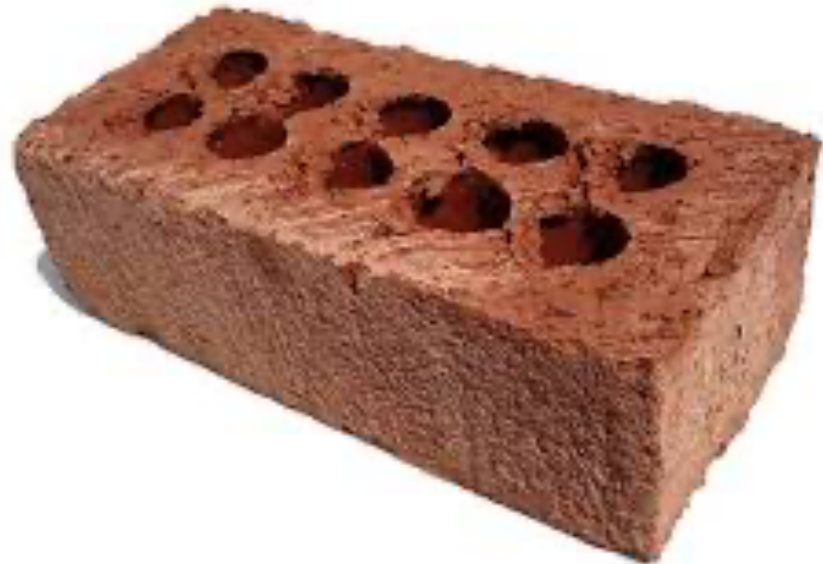
An Activity: Here's a brick...

- What can you do with this brick?
- List as many things as you can in one minute!
- Be creative!



An Activity: Here's a brick...

- How would you “score” this task?
 - ▣ Number of things they wrote down?
 - ▣ The degree of commonness of their response?
 - ▣ What about things that are really far out?



Time to Operate!

- Health
- Anxiety
- Extroversion
- Personal Space
- Type A Personality
- Statistical Literacy

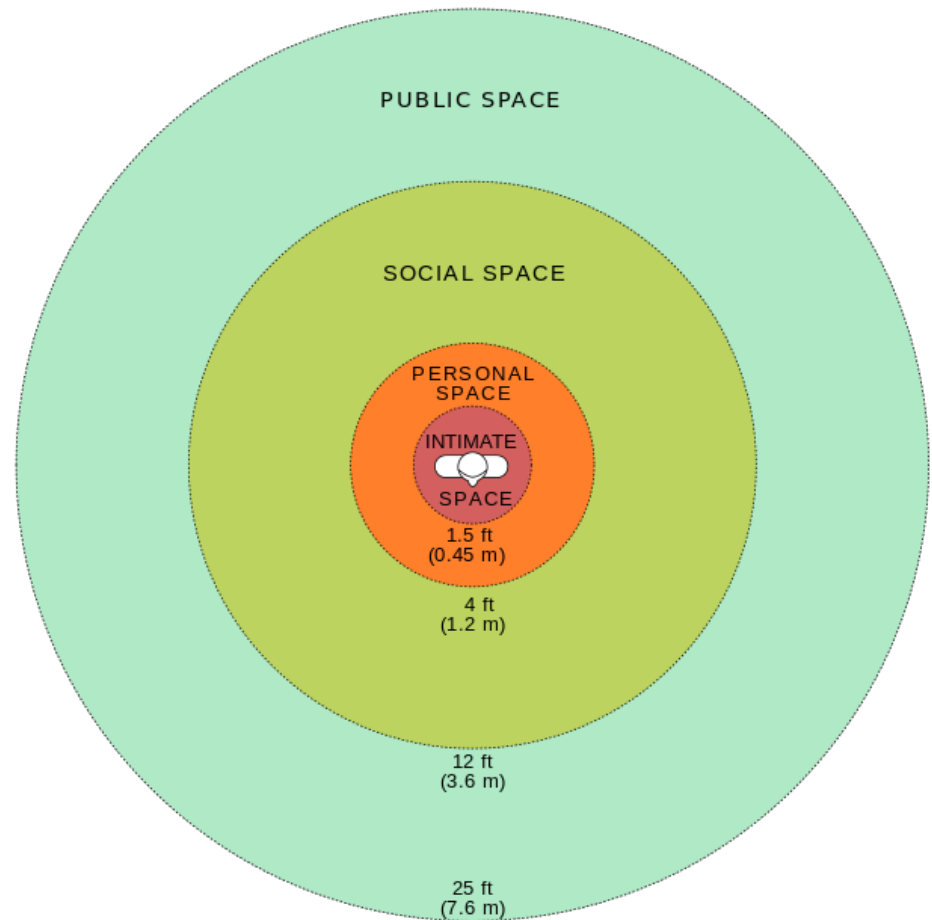


How would you conceptualize and operationalize some of the concepts above.

Personal Space Example

Nice job tuning an abstract concept into something measurable.

(But this is quite culturally dependent...)



What to measure and how?

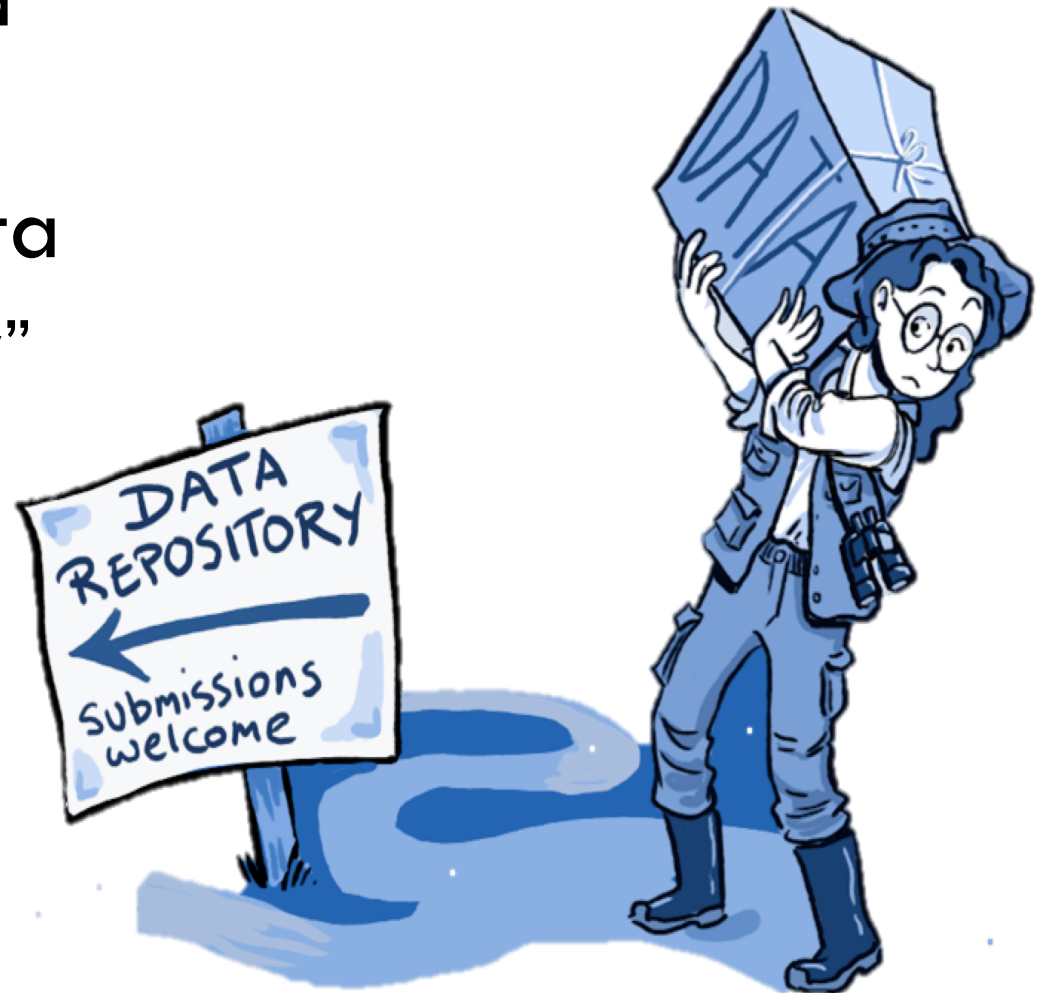
You could ask people to *describe* their own idea of personal space or you could ask them *where* (in inches and feet) their personal space begin?

What is the difference between those two questions?

Quality of Quantity?

Show Me the Data!

- Qualitative Data
 - ▣ “qualit” = “quality”
- Quantitative Data
 - ▣ “quant” = “quantity”



Data & Descriptions

What is
this?



Describe and
measure it.

Data & Descriptions

Qualitative

- ❑ A portrait of a woman with a thorn necklace and animals.
- ❑ Thorns piercing her neck and dripping blood
- ❑ Solemn looking
- ❑ Feels intense



What is this?
Describe it.

Quantitative

- ❑ Painted in 1940
- ❑ Size: 61.25 cm x 47 cm
- ❑ It took XX number of hours to paint

“Autorretrato con Collar de Espinas” by Frida Kahlo (1940)

Pros and Cons: Qual and Quant

Qualitative

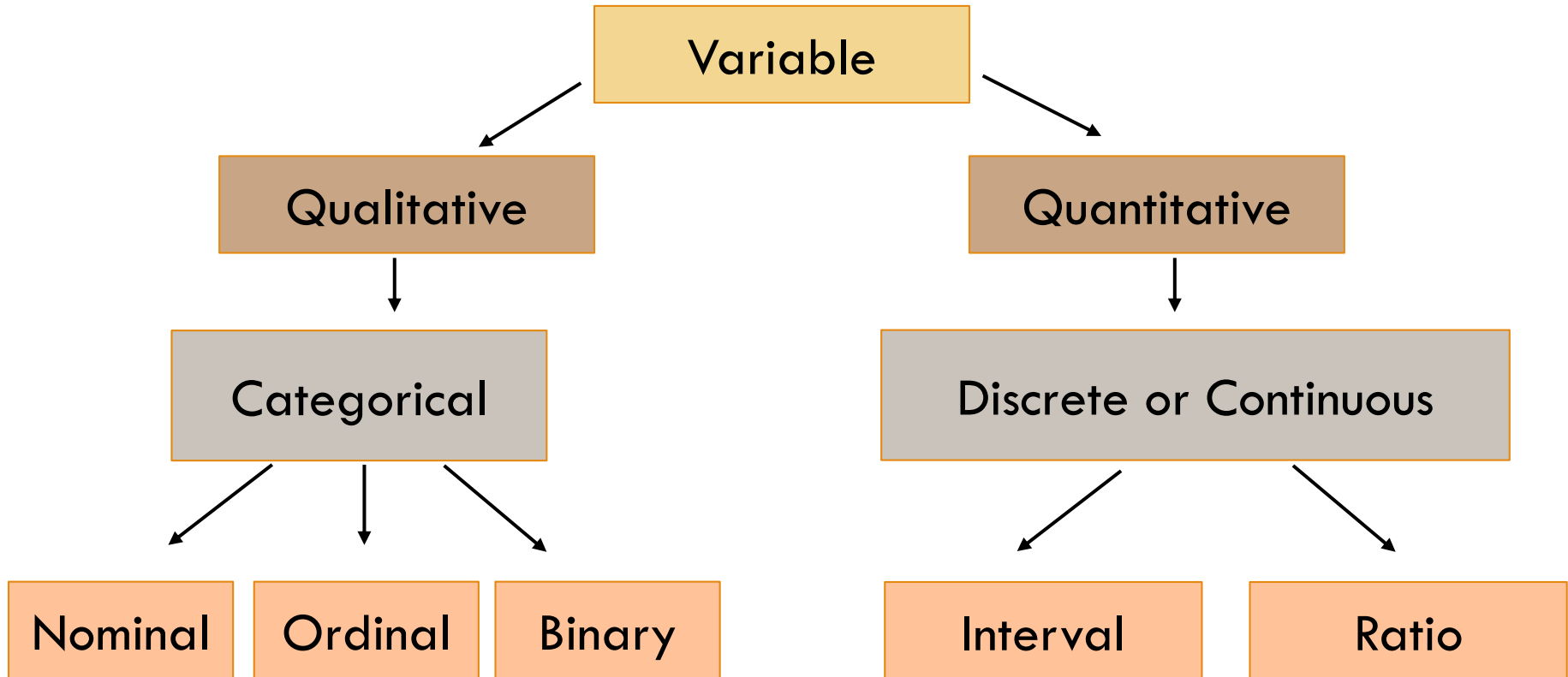
- ▣ Rich information about human behavior
 - Observations and interviews
- ▣ Analyze themes
- ▣ Can be hard to quantify and reproduce

Quantitative

- ▣ Concrete numbers
- ▣ Measurable and more easily reproduced
- ▣ Can run statistics and perform analyses on the numbers
- ▣ Can summarize and test

Variable and Data Types

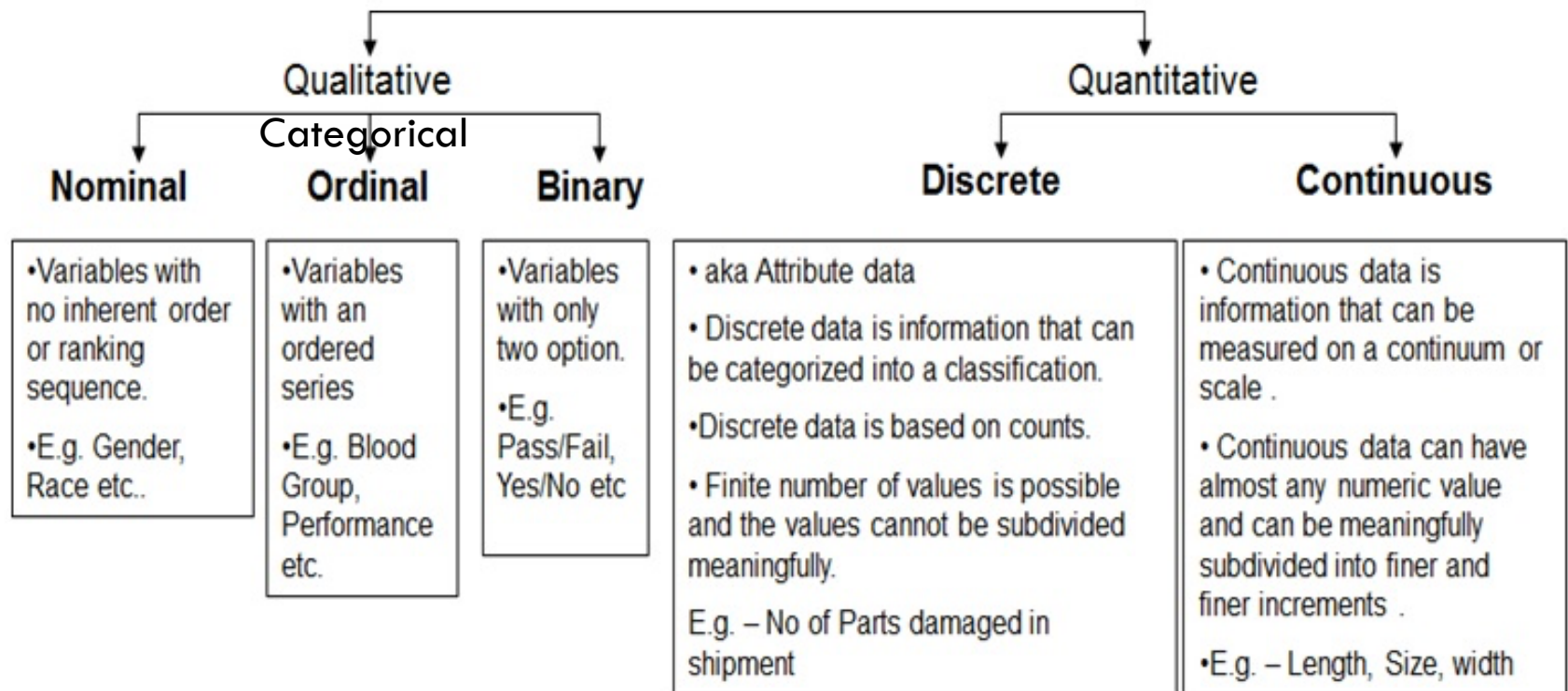
What can it be?



Types of Data

- Qualitative (Categorical)
 - ▣ Nominal
 - ▣ Ordinal
- Quantitative (Discrete or Continuous)
 - ▣ Interval
 - ▣ Ratio

Qualitative vs. Quantitative



Qualitative: Categorical

What's in a Name?



- Nominal Data:
 - Purely a name or category
 - Differentiates between items or subjects
 - Qualitative in nature
 - Limited math and computation ability
 - Grouping variable
 - You can assign numbers to a category, BUT they do not have a true numerical value or relationship
 - For example, you may code Males (0) and Females (1)
 - Males are not “worth” 0, and adding all the 0s would equal 0
 - You could count the number of 0s and you know how many males you have in a study → Frequency

Valueless Numbers

Subject ID#	Gender	Age
103	0	15
345	0	19
430	1	12
957	0	13

Which of the columns above are examples of Nominal data?

Nominal Data: Categories

- Gender
- Favorite color
- Ethnicity
- Political Affiliation
- Opinion on X
- Handedness
- Religious Affiliation
- Type of Fidgeting
- Major in college
- Car model
- Geographic location
- Favorite food
- Music interest
- Netflix preferences
- Language
- True/False

Frequency Distributions

- The number of times a certain categorical variable is observed, a count of the number of people in a certain category

Possible Values	Tally Marks	Frequency
Brown		2
Black		6
White		1
Blue		12
Purple		17
Orange		11
Yellow		7
Red		18

Favorite Color - Relative Frequencies

Possible Values	Tally Marks	Frequency	Relative Frequency
Red	IIII IIII IIII III	18	$18/74 = .24, 24\%$
Purple	IIII IIII IIII II	17	$17/74 = .23, 23\%$
Blue	IIII IIII II	12	$12/74 = .16, 16\%$
Orange	IIII IIII I	11	$11/74 = .15, 15\%$
Yellow	IIII II	7	$7/74 = .09, 9\%$
Black	IIII I	6	$6/74 = .08, 8\%$
Brown	II	2	$2/74 = .03, 3\%$
White	I	1	$1/74 = .01, 1\%$

Relative frequency is the frequency count (also called proportion) of a certain event relative to all the other events (or times someone picked a certain color).
Ex. 18 out of 74 people chose Red, 24% of all respondents chose Red.

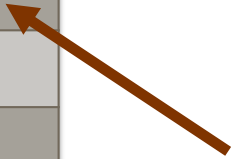
Favorite Color - Relative Frequencies

Possible Values	Tally Marks	Frequency	Relative Frequency
Red	IIII IIII IIII III	18	$18/74 = .24, 24\%$
Purple	IIII IIII IIII II	17	$17/74 = .23, 23\%$
Blue	IIII IIII II	12	$12/74 = .16, 16\%$
Orange	IIII IIII I	11	$11/74 = .15, 15\%$
Yellow	IIII II	7	$7/74 = .09, 9\%$
Black	IIII I	6	$6/74 = .08, 8\%$
Brown	II	2	$2/74 = .03, 3\%$
White	I	1	$1/74 = .01, 1\%$

Based on this sample, what is the probability I ask somewhat what their favorite color is and they say “Red”?

Favorite Color - Relative Frequencies

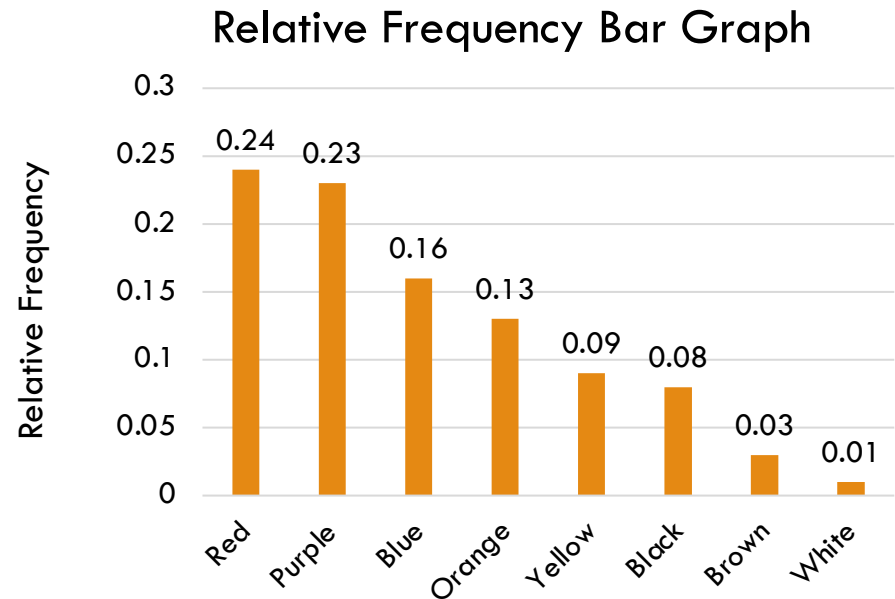
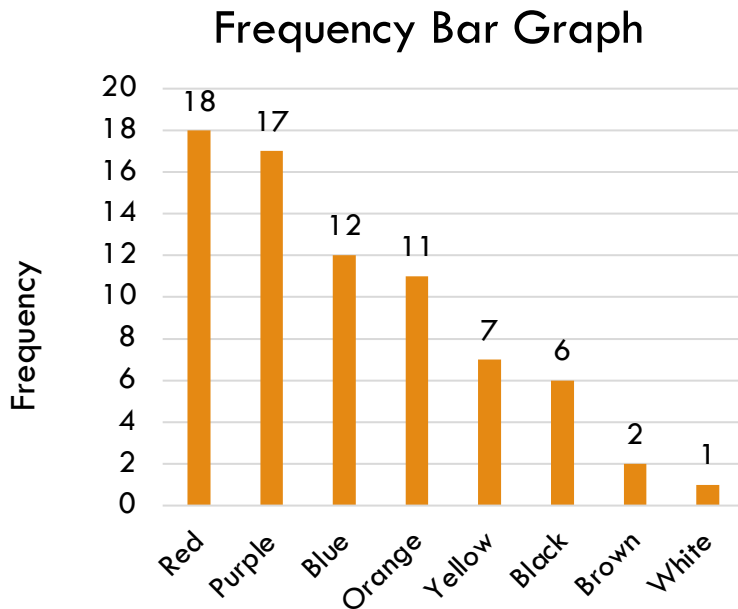
Possible Values	Tally Marks	Frequency	Relative Frequency
Red	IIII IIII IIII III	18	$18/74 = .24, 24\%$
Purple	IIII IIII IIII II	17	$17/74 = .23, 23\%$
Blue	IIII IIII II	12	$12/74 = .16, 16\%$
Orange	IIII IIII I	11	$11/74 = .15, 15\%$
Yellow	IIII II	7	$7/74 = .09, 9\%$
Black	IIII I	6	$6/74 = .08, 8\%$
Brown	II	2	$2/74 = .03, 3\%$
White	I	1	$1/74 = .01, 1\%$



Based on this sample, there is a 24% chance a person's favorite color is Red.

Based on this sample, what is the probability I ask somewhat what their favorite color is and they say “Red”?

Frequency Visualizations – Bar Graphs



Nominal variables are usually presented like this.

Notice the different Y axes.

Both are conveying the same data but in slightly different ways.

Get in Order!

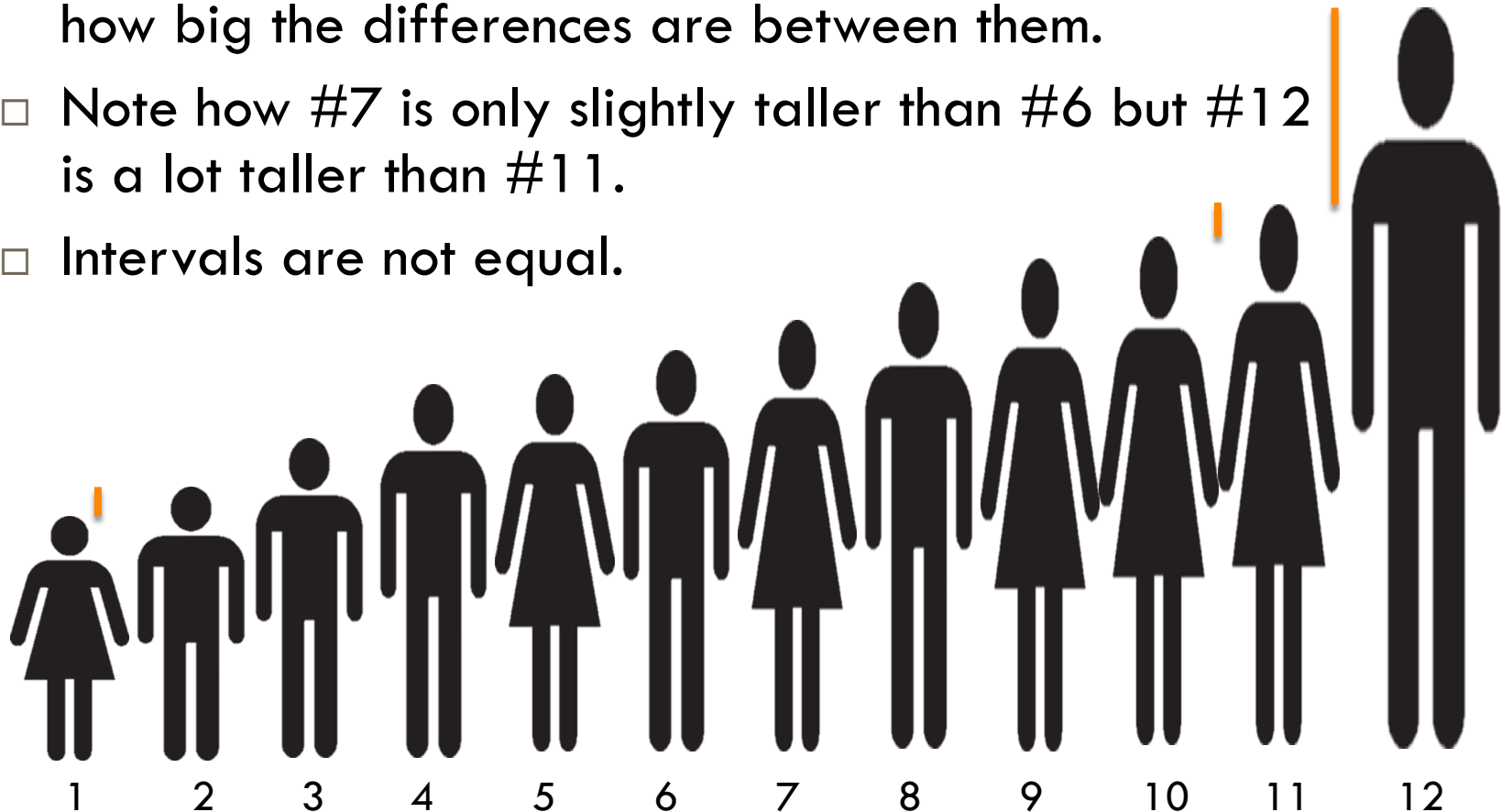
1st 2nd 3rd

□ Ordinal Data:

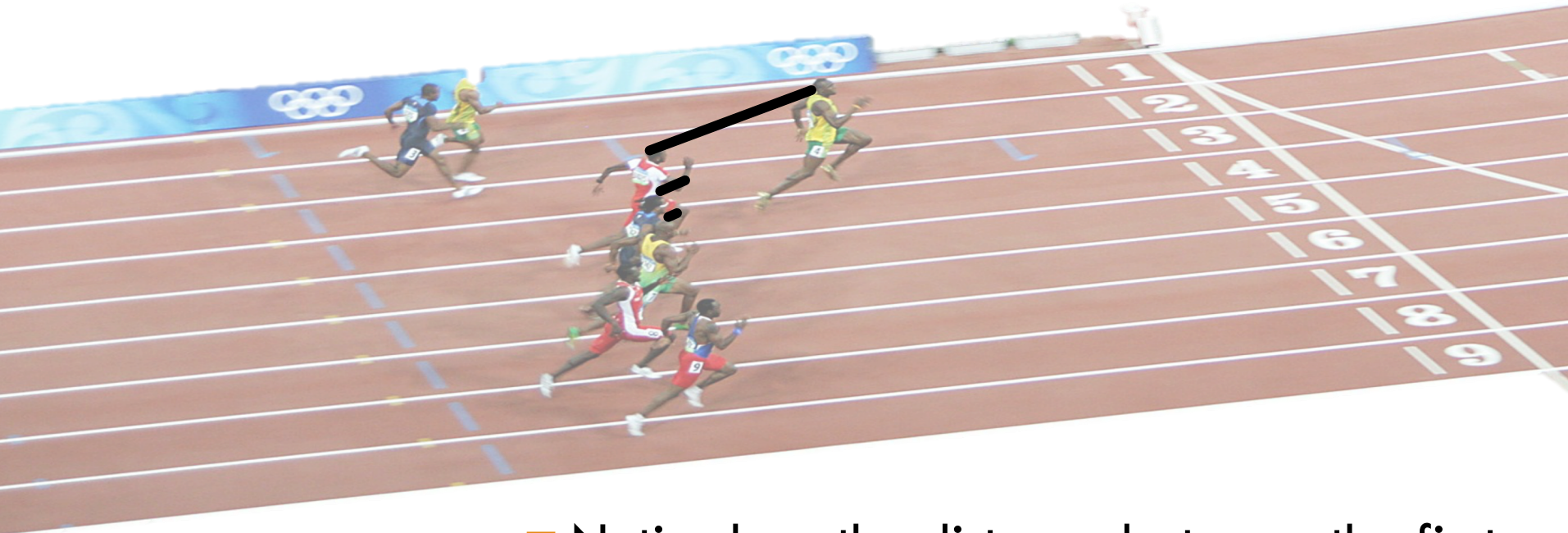
- Tells you about the rank order and can be sorted
- Does NOT tell you the difference between them
 - For example, we could put all of us in order of height
 - We know who is taller than who...
 - But we would NOT know by how much, we don't know the increments between us
 - How much taller is the tallest person than the next tallest person? A foot? An inch? Is the next person exactly an inch shorter? We don't know.
 - Cannot do meaningful additions and subtractions

Get in Order

- We know person #12 is the tallest, and that #11 is taller than #10 and below, but we do not know how big the differences are between them.
- Note how #7 is only slightly taller than #6 but #12 is a lot taller than #11.
- Intervals are not equal.



More Ranks and Orders



- Notice how the distance between the first place runner to the second place runner is not the same distance of the second place runner to the third place runner.

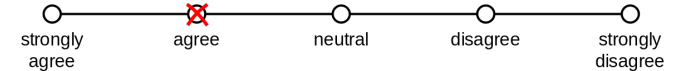


Likert Scales

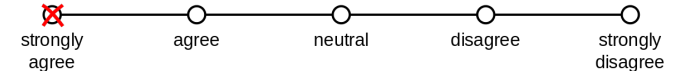
- Likert Scales are commonly used in social science research.
- Technically, they're ordinal because we do not know if the difference between strongly agree and agree is the same amount of difference from agree to neutral
 - ▣ Researchers to take some liberties and sometimes treat Likert scale data as interval data...

Website User Survey

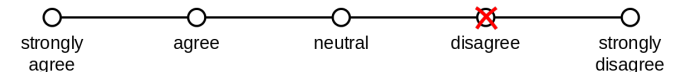
1. The website has a user friendly interface.



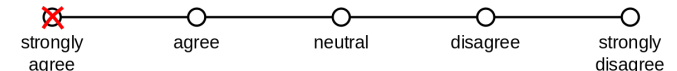
2. The website is easy to navigate.



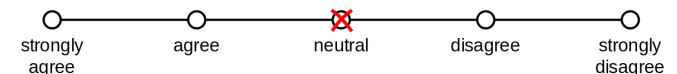
3. The website's pages generally have good images.



4. The website allows users to upload pictures easily.



5. The website has a pleasing color scheme.



Ordinal Data

- School rank
- Percentile rank
- Hotness rank
- Famousness rank
- Creativeness rank
- Likert scales*
- Socioeconomic status (high-medium-low)
- Level of agreement (disagree-agree)

Quantitative: Discrete and Continuous

Quantitative: Discrete vs. Continuous Data

Discrete

- Fixed set of options, finite number of choices
- Can't cut up into $\frac{1}{2}$'s
 - ▣ The number of laptops or cell phones you have
 - ▣ The number of cars you own
 - ▣ Number of significant others in the past

Continuous

- Infinite number of values
- Decimals and refinement
- If you can keep measuring/cutting it more and more it's continuous
 - ▣ Ex. Age* can be calculated to an infinite level,
 - 33 years, 1 months, 4 days, 7 hours, 12 minutes, 45 seconds, 120 ms, etc...

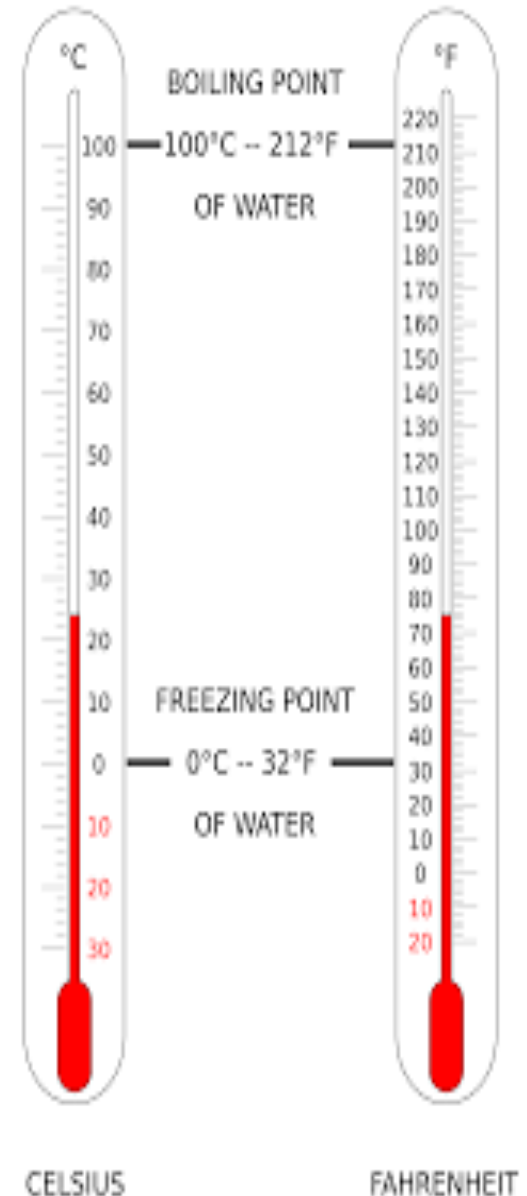
Continuous variables can sometimes be discretized at the discretion of the researcher.

Discretizing Continuous Variables

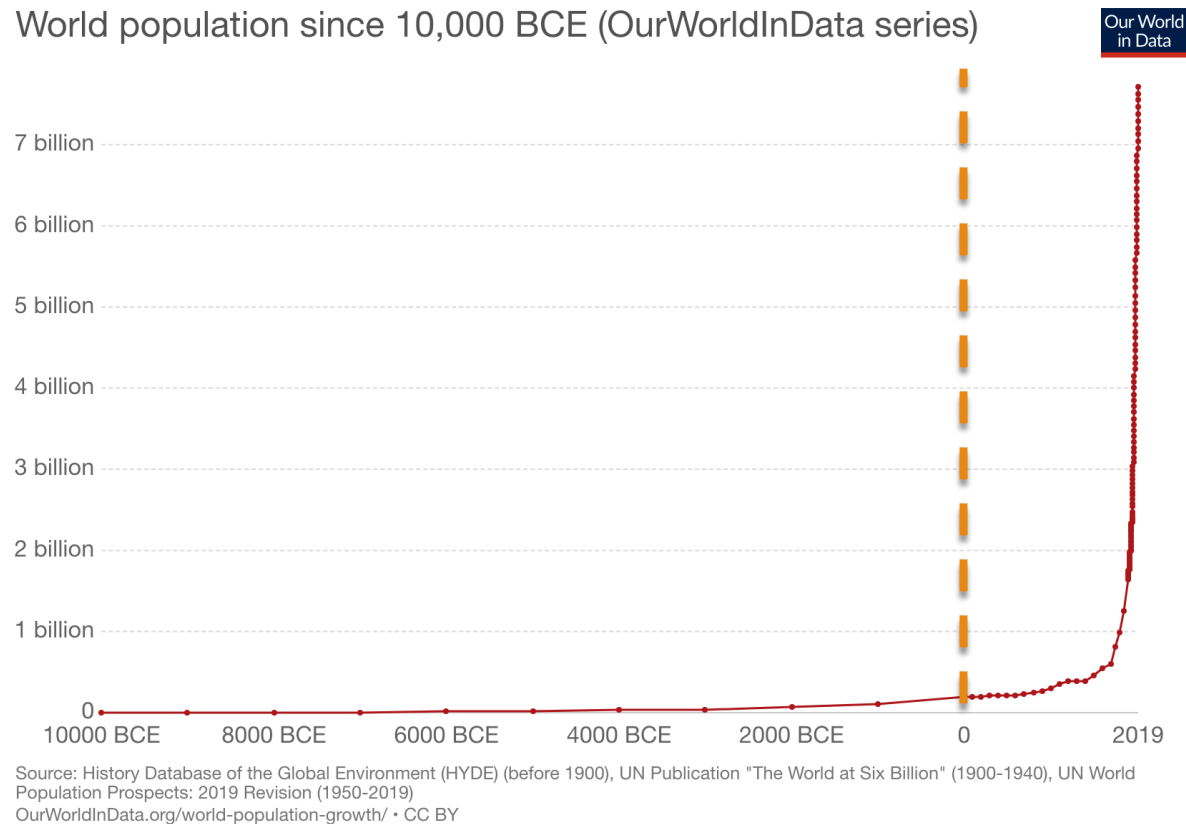
Age Group	Age
Infant	0-1 year
Toddler	1-3 years
Pre-Schooler Child	3-5 years
School Age Child	6-12 years
Adolescent	13-18 years
Young Adult	19-40 years
Middle Age Adult	41-65 years
Older Adult	65 + years

To what degree?

- Interval Data:
 - Similar to ordinal in that it is in order, but the increments, or distances between points, ARE equal.
 - Ex. 2° to 3° is the same as from 92° to 93°
 - But, the ratios (how many times a number is contained in a second number) between them are not meaningful
 - Ex. 80° is not twice as hot as 40°
 - Why not?
 - Because Celsius and Fahrenheit thermometers use an arbitrary 0° point



Arbitrary Chosen Zeros



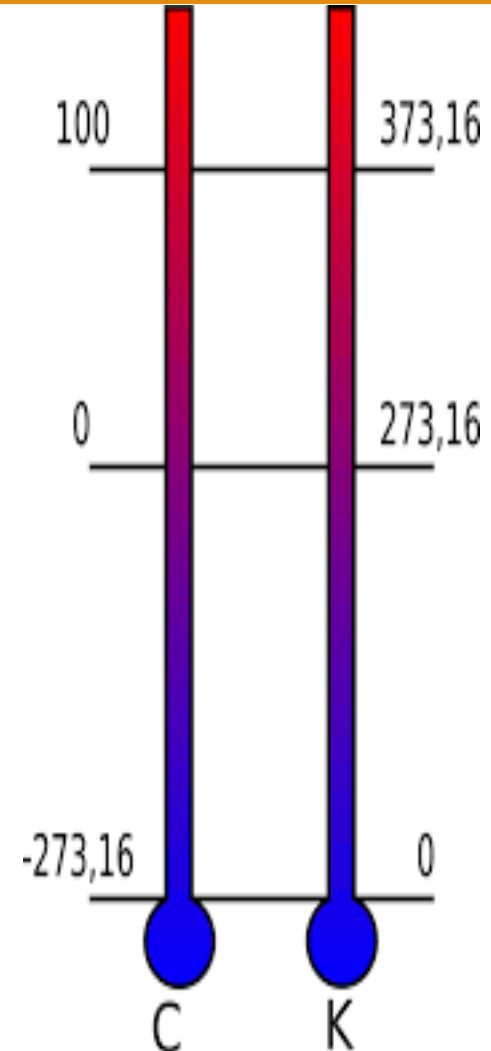
Time, as far as BC and AD, is an example of interval data because the zero start point is arbitrary.

Interval Data

- IQ scores
- ACT, SAT, GRE, MCAT, LSAT
- Calendar Year (i.e. 2020)
- Degrees of temperature (F and C)
 - ▣ 0 degrees does not mean absence of temperature
 - ▣ A temperature of 80F is not 2x as hot as 40F
- Likert scales*
- Ratio data converted to standardized metric (ex. z-score)

Ratio Scale

- Has equal intervals and an *absolute zero*.
 - A TRUE zero, not an arbitrary one
 - Time in regard to your age, does have a true 0, you can measure from now back to before you were born
 - Many hard sciences have ratio scale data
 - Kelvin temperature is a ratio scale because it has a true absolute zero, the zero degree temperature at which molecules stop moving
 - As opposed to 0° Celsius, which does NOT mean no heat
 - Countable quantities
 - With zero as a possibility
 - Ex. Can't score a zero on an IQ test



Ratio Data

- Age
- Weight
- Income
- Years of education
- Minutes from time X
- Hours of Study
- Amount spent on X
- Time since Big Bang
- Number of miles driven
- Time (with 0 start time)
- Distance (from 0 point)
- Family size

Sometimes Uncertainty

- Some scales of measurement are easier to distinguish and determine.
 - ▣ Year (since when? Birth, AD, Big Bang)
- Some scales are treated as a different type of data (when justified) for certain purposes.
 - ▣ Ordinal treated as interval data used for social sciences
- So, as with most things in life, sometimes it depends...
 - ▣ And sometimes you can change things to suit your needs...

Clarifying Questions

- Nominal questions typically start with “What” or “Which” such as,
 - ▣ “What (which) is your favorite color?”
- Interval and Ratio can seem quite similar, when trying to distinguish between the two, try to ask,
 - ▣ “Can I have ZERO of this particular thing?”
 - Ex. Can’t have a zero intelligence score – perhaps contrary to some evidence...
 - ▣ “How much/many of something?”

Data Type Examples

Siblings

□ How many siblings do you have?

Continuous, Discrete, or Categorical?



Siblings

□ How many siblings do you have?

Discrete and ratio because you can't have a $1/4^{\text{th}}$ of a sibling (let's not get too picky about $1/2$'s) and you can have a zero siblings



Scales

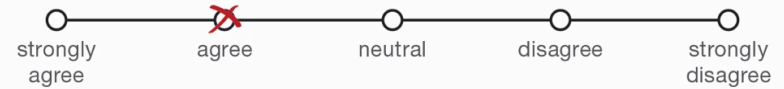
- Frequently used psychological measurements.

- ▣ Likert Scales

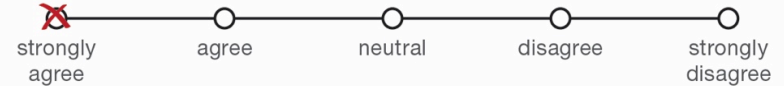
What level of measurement?

Nominal, Ordinal, Interval, or Ratio?

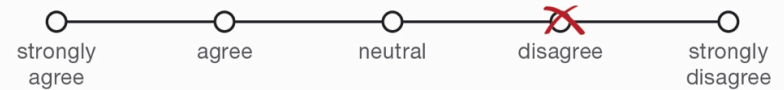
1. Wikipedia has a user friendly interface.



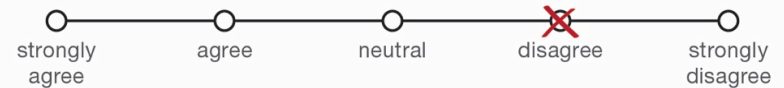
2. Wikipedia is usually my first resource for research.



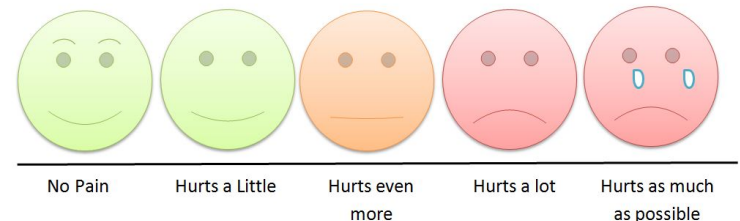
3. Wikipedia pages generally have good images.



4. Wikipedia allows users to upload pictures easily.



5. Wikipedia has a pleasing color scheme.



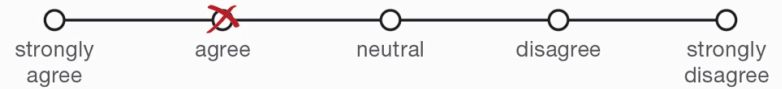
Scales

- Frequently used psychological measurements.

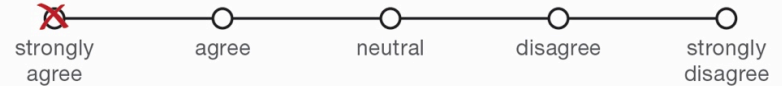
- Likert Scales

Technically Ordinal but usually treated as Interval, and discrete in this example.

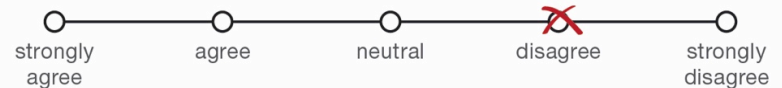
1. Wikipedia has a user friendly interface.



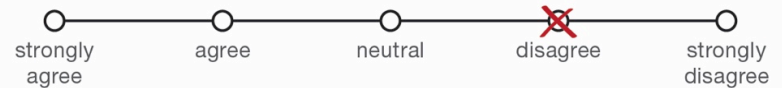
2. Wikipedia is usually my first resource for research.



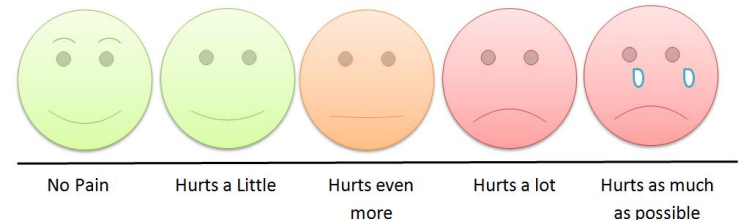
3. Wikipedia pages generally have good images.



4. Wikipedia allows users to upload pictures easily.



5. Wikipedia has a pleasing color scheme.

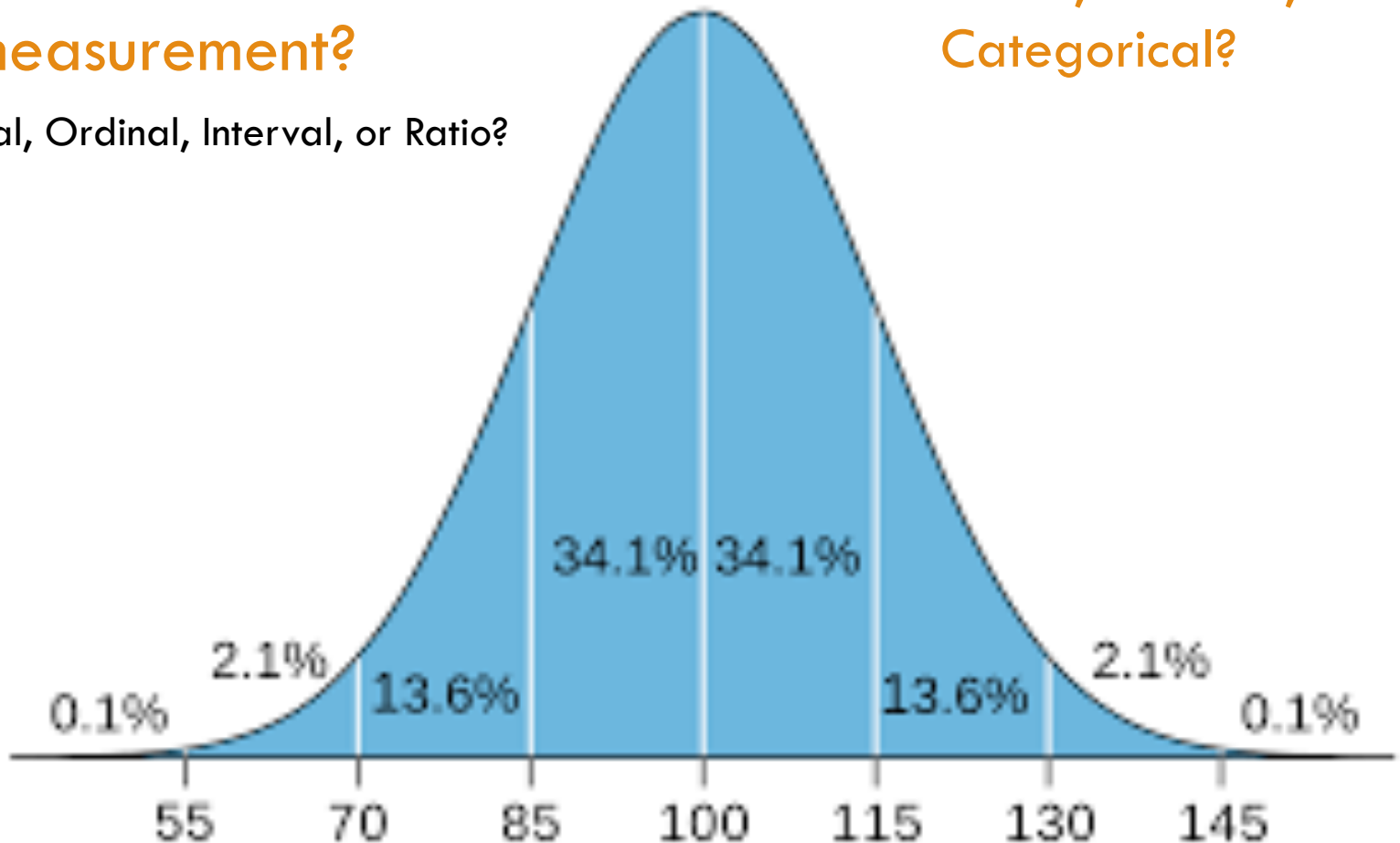


Intelligence (IQ) Tests

What level of measurement?

Nominal, Ordinal, Interval, or Ratio?

Continuous, Discrete, or Categorical?



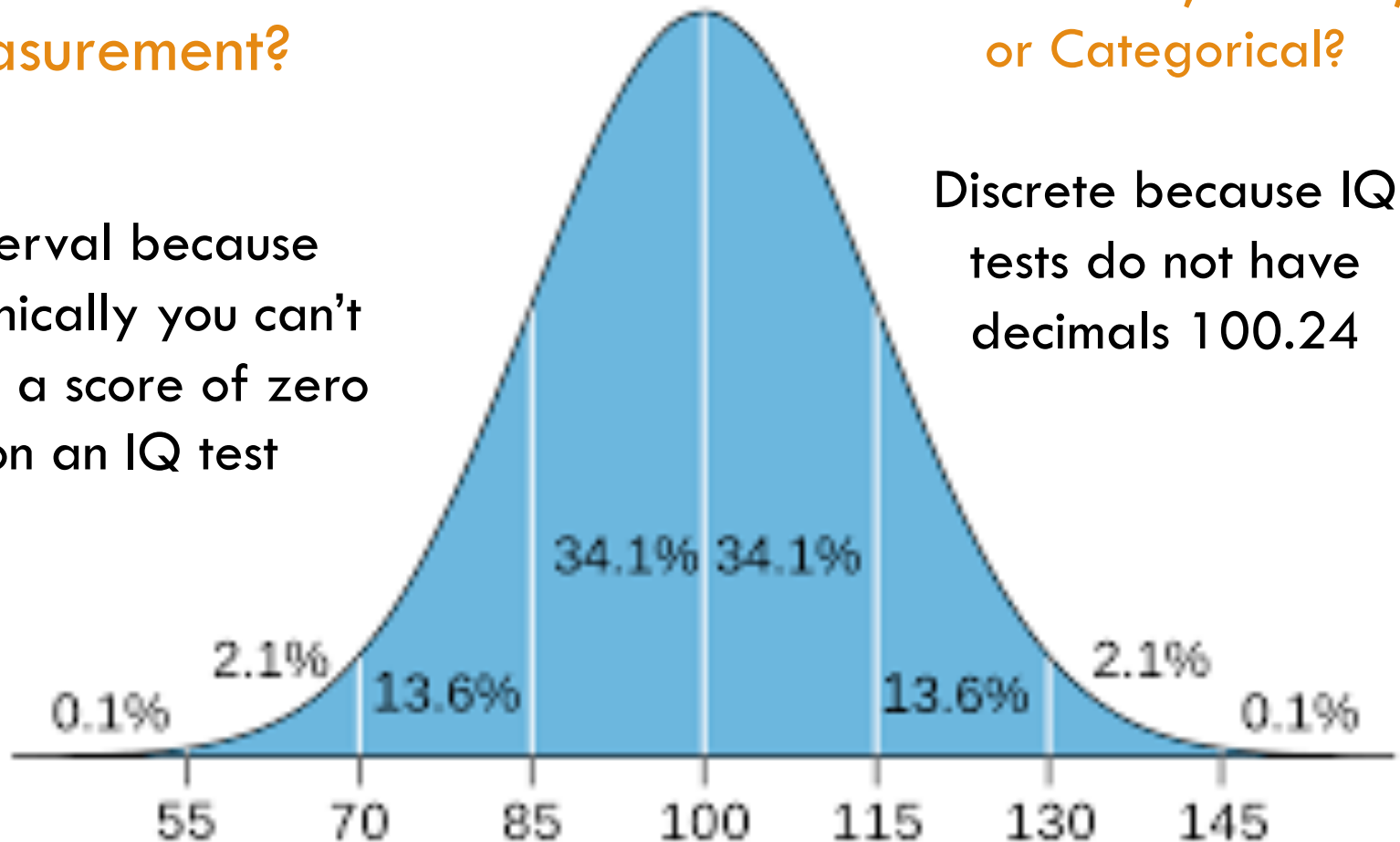
Intelligence (IQ) Tests

What level of measurement?

Interval because technically you can't have a score of zero on an IQ test

Continuous, Discrete, or Categorical?

Discrete because IQ tests do not have decimals 100.24



Event Related Potentials

□ Reaction Times

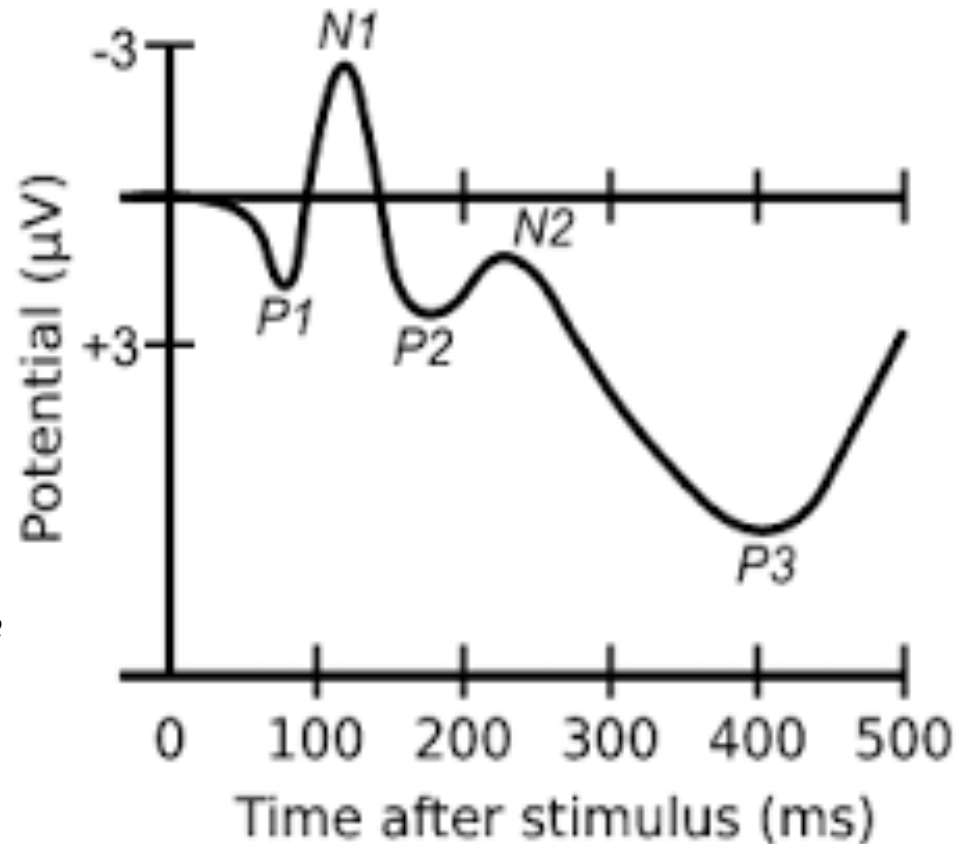
▣ A stimulus (like a tone) is delivered, a person responds

▣ Measure reaction time

What level of measurement?

Nominal, Ordinal, Interval, or Ratio?

Continuous, Discrete, or Categorical?



Event Related Potentials

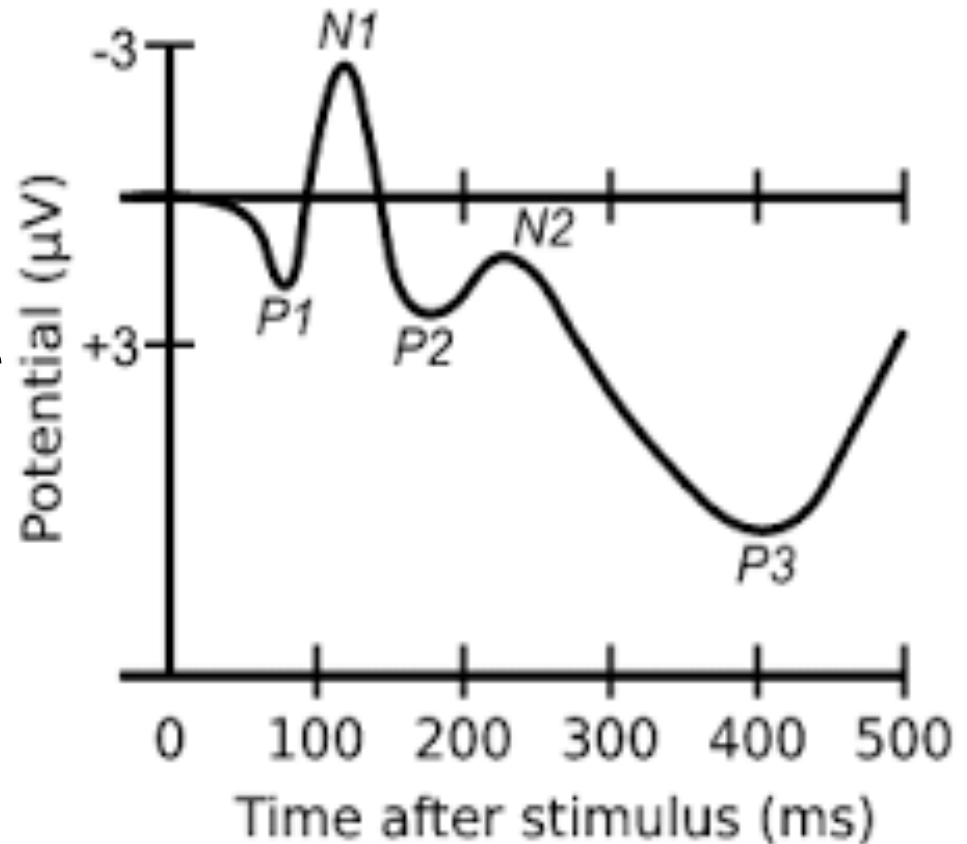
□ Reaction Times

- ▣ A stimulus (like a tone) is delivered, a person responds
- ▣ Measure reaction time

What level of measurement?

Ratio because we start from zero

Continuous because we can technically have 200.324 ms



Next Up...

- Now that we have a sense for the types of data we might be dealing with, let's start describing some of it with...

Describing Data

Types of Data in R

Types of Data in R

- R is very sensitive about the data you put into it. It knows that for certain data types you cannot calculate a mean
 - ▣ In the Males = 0 and Females = 1 example, R will assign this data as a “Factor”
 - ▣ A “Factor” is basically a Categorical (Nominal) variable that has been recoded into numbers, these numbers do not have a numeric value, they are just symbols in this context

Types of Data in R

- There are a few different “classes” of data in R, but the most important ones for us are:
 - Numeric
 - These can be ANY number, like 1.61803...
 - These are CONTINUOUS
 - Integer
 - These can only be WHOLE number, like 2
 - These are DISCRETE
 - Factor
 - These can be labelled with the Categorical name (ex. “Male”) or as their numeric label (ex. “0”), regardless of which one you see, in the background R knows these are labels
 - You cannot “math” these because this is nominal data
 - Character
 - There are purely text, like the word “Male”
 - You cannot “math” these because they are just words
- } You can do mathematical operations on these.

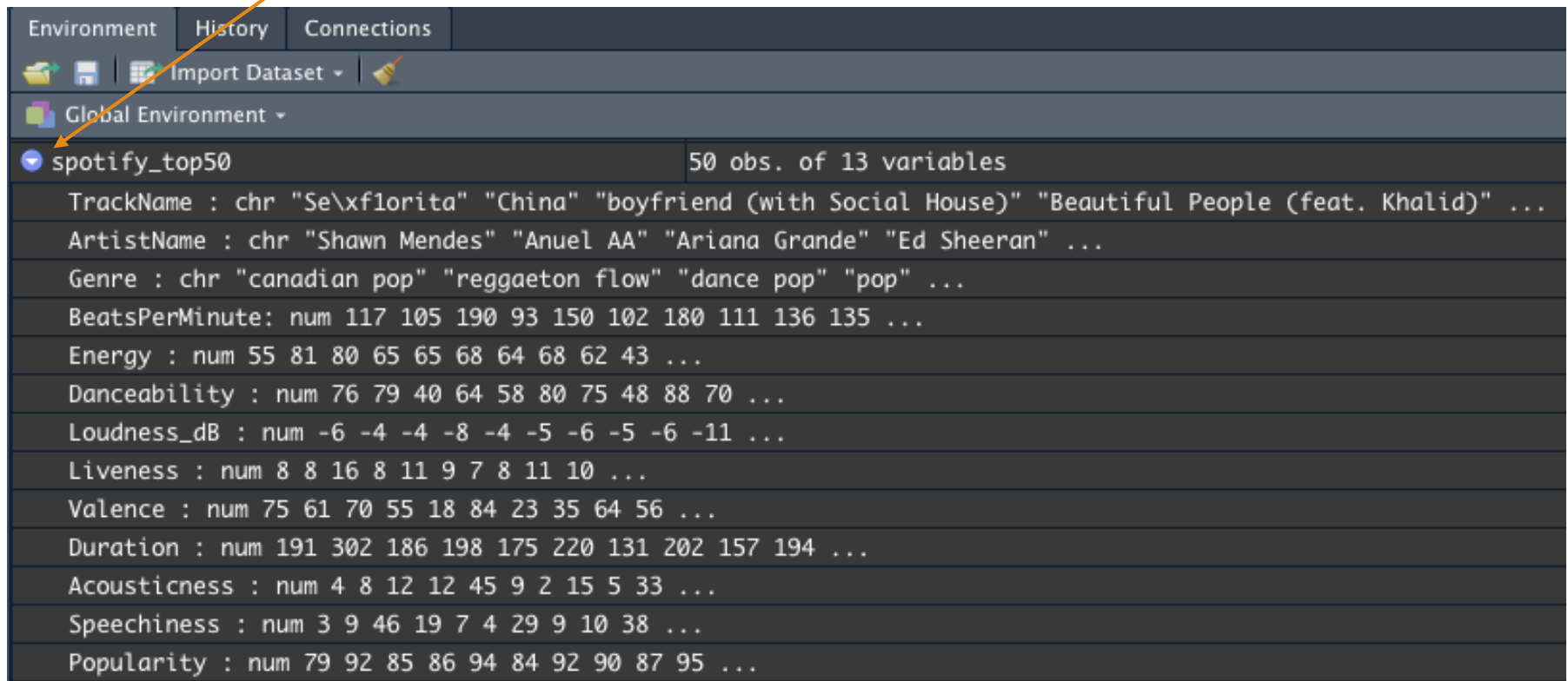
Types of Data in R

- You can check the type of data using “class()” and (in some cases) change the type of data in R

```
#####  
##### CLASS #####  
#####  
  
# Load in the data, here we are reading in a .csv file  
spotify_top50 <- read.csv("spotify_top50.csv")  
  
View(spotify_top50)  
  
# To check what type of data a particular column is you can use the "class()" function in R  
class(spotify_top50$Genre)  
  
# R is treating the Genre as a character... But we want this to be a factor,  
# because music Genre is a type of category  
# We can reassign the class by using the function "factor()"  
spotify_top50$Genre <- factor(spotify_top50$Genre)  
  
# We can check to make sure it worked with the class() function again,  
# now R is treating it as a Factor with 21 different categories  
class(spotify_top50$Genre)
```

Types of Data in R

- You can also look in the Global Environment and click the little blue drop down to see what columns and of what class you have in the data



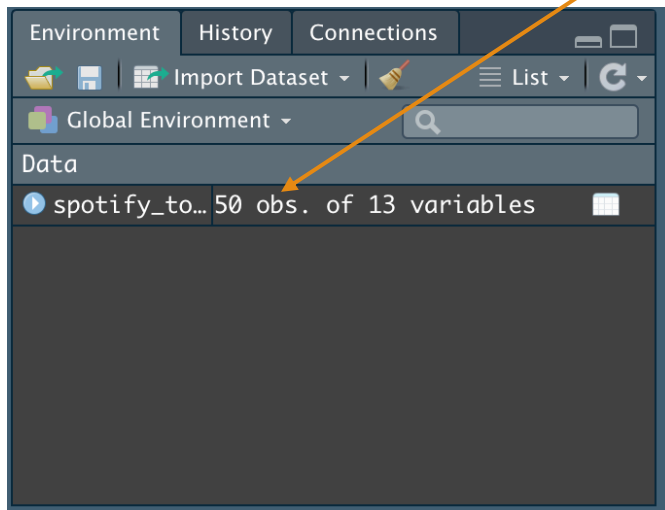
The screenshot shows the R Studio interface with the Global Environment pane open. The 'spotify_top50' dataset is selected, and its structure is displayed. A blue arrow points to the dropdown arrow next to the dataset name.

```
Environment | History | Connections | Import Dataset | Global Environment | spotify_top50 (50 obs. of 13 variables)
```

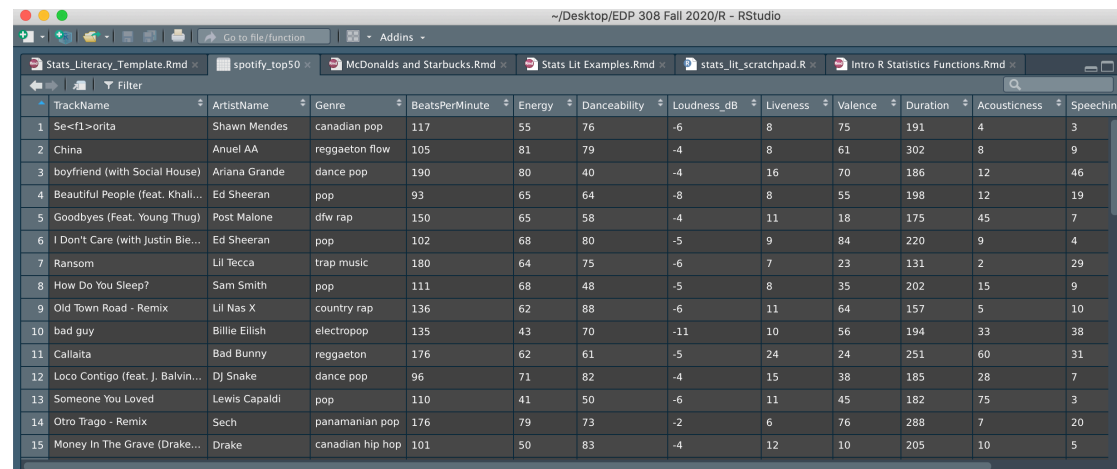
Variable	Class	Values
TrackName	chr	"Se\xflorita" "China" "boyfriend (with Social House)" "Beautiful People (feat. Khalid)" ...
ArtistName	chr	"Shawn Mendes" "Anuel AA" "Ariana Grande" "Ed Sheeran" ...
Genre	chr	"canadian pop" "reggaeton flow" "dance pop" "pop" ...
BeatsPerMinute	num	117 105 190 93 150 102 180 111 136 135 ...
Energy	num	55 81 80 65 65 68 64 68 62 43 ...
Danceability	num	76 79 40 64 58 80 75 48 88 70 ...
Loudness_dB	num	-6 -4 -4 -8 -4 -5 -6 -5 -6 -11 ...
Liveness	num	8 8 16 8 11 9 7 8 11 10 ...
Valence	num	75 61 70 55 18 84 23 35 64 56 ...
Duration	num	191 302 186 198 175 220 131 202 157 194 ...
Acousticness	num	4 8 12 12 45 9 2 15 5 33 ...
Speechiness	num	3 9 46 19 7 4 29 9 10 38 ...
Popularity	num	79 92 85 86 94 84 92 90 87 95 ...

Types of Data in R

- You can also look at the dataset like you would in a program like Excel by clicking the dataset or using the “View()” function



The screenshot shows the RStudio Environment pane. At the top, there are tabs for 'Environment', 'History', and 'Connections'. Below these are icons for 'Import Dataset', 'List', and a refresh button. The main area is labeled 'Global Environment' and contains a search bar. Under the 'Data' section, a dataset named 'spotify_to...' is listed with the description '50 obs. of 13 variables'. An orange arrow points from the text 'View()' in the slide to this dataset entry.



The screenshot shows a data table in RStudio. The table has 15 rows and 12 columns. The columns are: TrackName, ArtistName, Genre, BeatsPerMinute, Energy, Danceability, Loudness_dB, Liveness, Valence, Duration, Acousticness, and Speechiness. The data is as follows:

	TrackName	ArtistName	Genre	BeatsPerMinute	Energy	Danceability	Loudness_dB	Liveness	Valence	Duration	Acousticness	Speechiness
1	Se<fl>-orita	Shawn Mendes	canadian pop	117	55	76	-6	8	75	191	4	3
2	China	Anuel AA	reggaeton flow	105	81	79	-4	8	61	302	8	9
3	boyfriend (with Social House)	Ariana Grande	dance pop	190	80	40	-4	16	70	186	12	46
4	Beautiful People (feat. Khalid)	Ed Sheeran	pop	93	65	64	-8	8	55	198	12	19
5	Goodbyes (Feat. Young Thug)	Post Malone	dfw rap	150	65	58	-4	11	18	175	45	7
6	I Don't Care (with Justin Bieber)	Ed Sheeran	pop	102	68	80	-5	9	84	220	9	4
7	Ransom	Lil Tecca	trap music	180	64	75	-6	7	23	131	2	29
8	How Do You Sleep?	Sam Smith	pop	111	68	48	-5	8	35	202	15	9
9	Old Town Road - Remix	Lil Nas X	country rap	136	62	88	-6	11	64	157	5	10
10	bad guy	Billie Eilish	electropop	135	43	70	-11	10	56	194	33	38
11	Callaita	Bad Bunny	reggaeton	176	62	61	-5	24	24	251	60	31
12	Loco Contigo (feat. J. Balvin)	DJ Snake	dance pop	96	71	82	-4	15	38	185	28	7
13	Someone You Loved	Lewis Capaldi	pop	110	41	50	-6	11	45	182	75	3
14	Otro Trago - Remix	Sech	panamanian pop	176	79	73	-2	6	76	288	7	20
15	Money In The Grave (Drake)	Drake	canadian hip hop	101	50	83	-4	12	10	205	10	5