Welcome!

SDS 220

- Instructor: Dr. Rebecca Kurtz-Garcia (she/her)
- Course Site/Syllabus
- Moodle/Gradescope
- Slack

Who are you?

Please take five minutes to introduce yourself to one or two of your neighbors:

1) Your name
2) Your pronouns, if you feel comfortable sharing
3) One fun or meaningful thing you did over winter break

What is one non-obvious thing that you all have in common?

What is statistics?

- Definition

Some common descriptions

- "Statistics is using data and knowledge about randomness to condense, communicate, and contextualize information and provide insight into the setting from which the data came." Jo Hardin
- "Statistics: The Art and Science of Learning from Data." Alan Agresti and Christine A. Franklin
- "The process of collecting, describing, summarizing, visualizing, analyzing, or inferring information from data." Rebecca Kurtz-Garcia

Recall the Three Phases of the Course

- Introduction and Descriptive Statistics
- Foundations of Statistical Inference
- Inferential Statistics

An Example Question

I recently heard a claim that:

The average height of a student at Smith is 62.5 inches.

An Example Question

I recently heard a claim that:

The average height of a student at Smith is 62.5 inches.

How can we check this claim?

Let's do it!

Lets take a few moments and record as many heights as we can. When asking your fellow students remember:

- (Re)introduce yourself
- Explain it is for a class
- Be polite

Record responses: https://forms.gle/vWU9NnkcUeixJGFG6

Results

- What did we observe?

Results

- What did we observe?
- Does this claim seem reasonable?

Results

- What did we observe?
- Does this claim seem reasonable?
- Should we accept/reject the claim?

Results

- What did we observe?
- Does this claim seem reasonable?
- Should we accept/reject the claim?
- What range of values would we say is reasonably close to the claim?

Future Problems

How would we find a range of values that are reasonably close to a claim for other scenarios?

- Different populations (heights of flowers, heights of fifth graders, ...)
- Different sample sizes
- Different claims (life expectancy, salary, proportion of SDS majors, proportion of credit card owners)

Future Problems

How would we find a range of values that are reasonably close to a claim for other scenarios?

- Different populations (heights of flowers, heights of fifth graders, ...)
- Different sample sizes
- Different claims (life expectancy, salary, proportion of SDS majors, proportion of credit card owners)

For any given claim about a mean/proportion/etc, we can use mathematics to determine a reasonable range of values! This is statistical inference.

Section 1.2

Observations, Variables, Data Frames

- **Data frame**: convenient and common way to organize data, especially if collecting data on a spreadsheet. Often called **tidy data**.
- **Observational unit/case**: the person on which measurements are taken. Usually a row of a data frame.
- **Variable**: characteristic being measured.

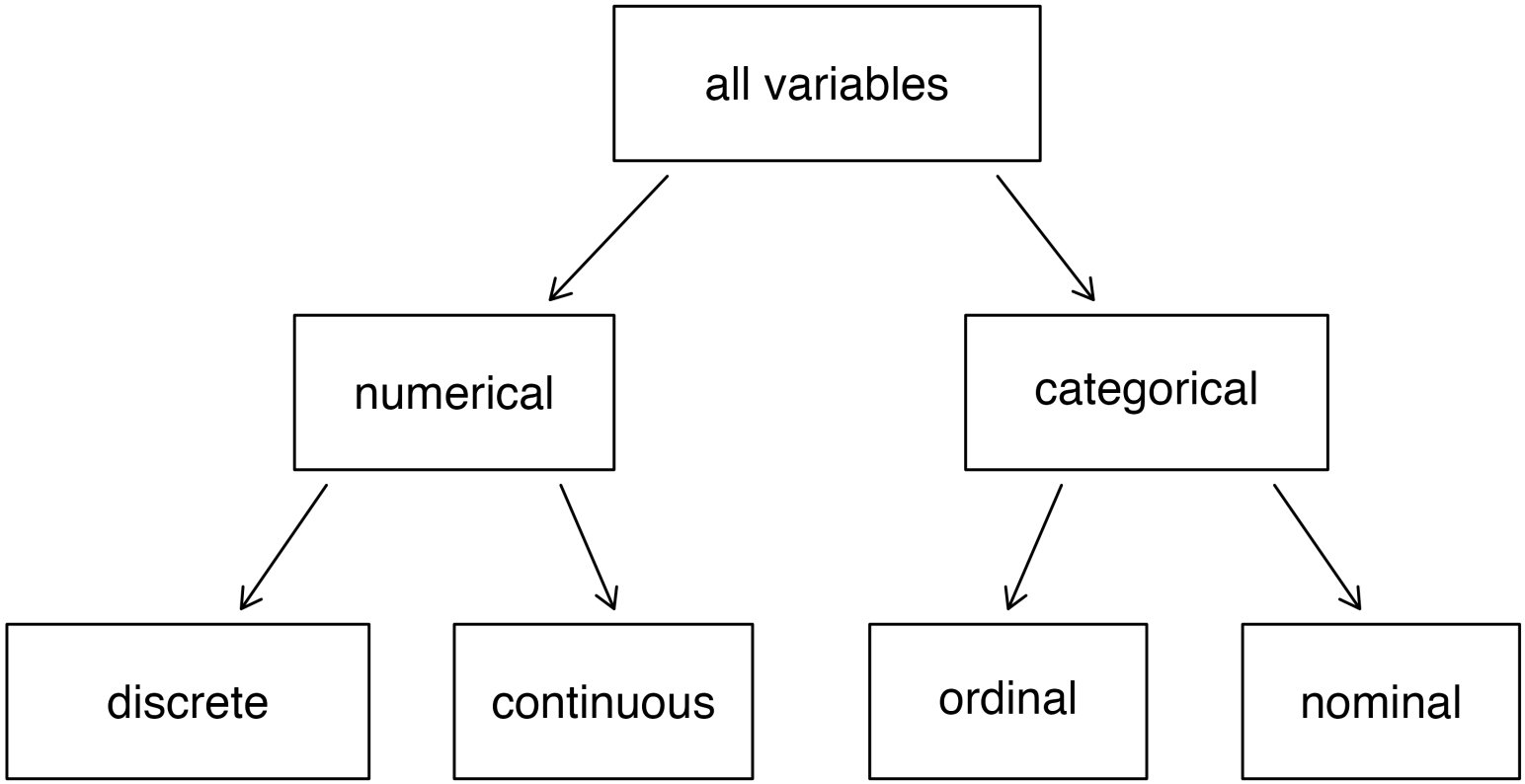
Observations, Variables, Data Frames

loan_amount	interest_rate	term	grade	state	total_income	homeownership
22,000	10.90	60	B	NJ	59,000	rent
6,000	9.92	36	B	CA	60,000	rent
25,000	26.30	36	E	SC	75,000	mortgage
6,000	9.92	36	B	CA	75,000	rent
25,000	9.43	60	B	OH	254,000	mortgage
6,400	9.92	36	B	IN	67,000	mortgage

Observations, Variables, Data Frames

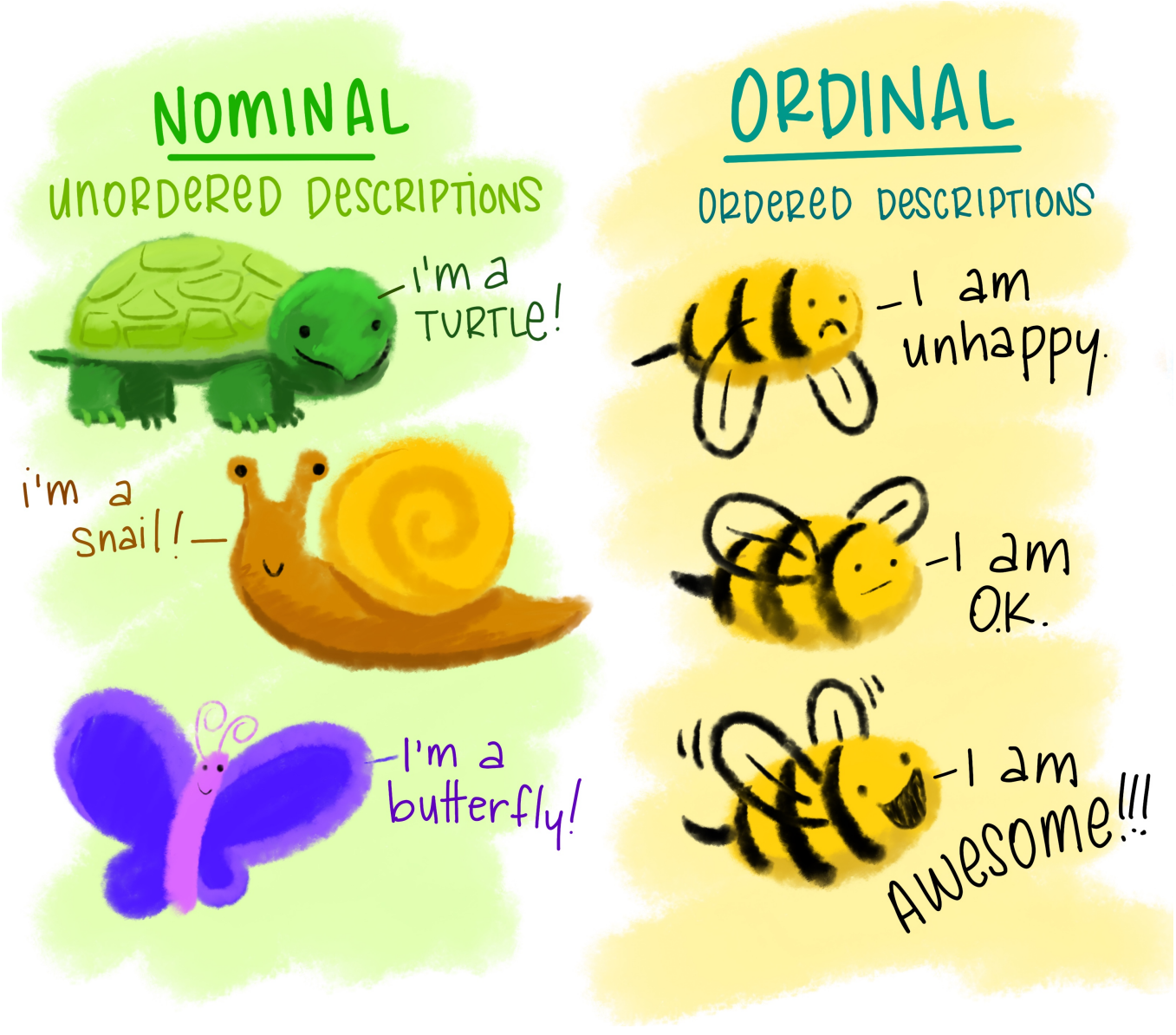
Observational Unit	Variable
US Voter	Prefered political candidate
Patient in clinical trial	COVID status
Mazda RX4	MPG
Loan Application	Interest Rate

Types of Variables



Types of Variables

Types of Variables



Types of Variables

- The distinction between types can be fuzzy, especially for discrete and ordinal. The subtle difference is *scale*.

Exercise

Return to your classmates and make a **data frame** by asking the following questions:

- What year are you?
- In which house do you live?
- What is your hometown?
- How many siblings do you have?
- What is the furthest away from Northampton (in miles) you were over the summer?

Exercise

Return to your classmates and make a **data frame** by asking the following questions:

- What year are you?
- In which house do you live?
- What is your hometown?
- How many siblings do you have?
- What is the furthest away from Northampton (in miles) you were over the summer?

For each variable, describe the type of variable that it is (e.g., categorical/numerical, discrete/continuous, ordinal, etc.)

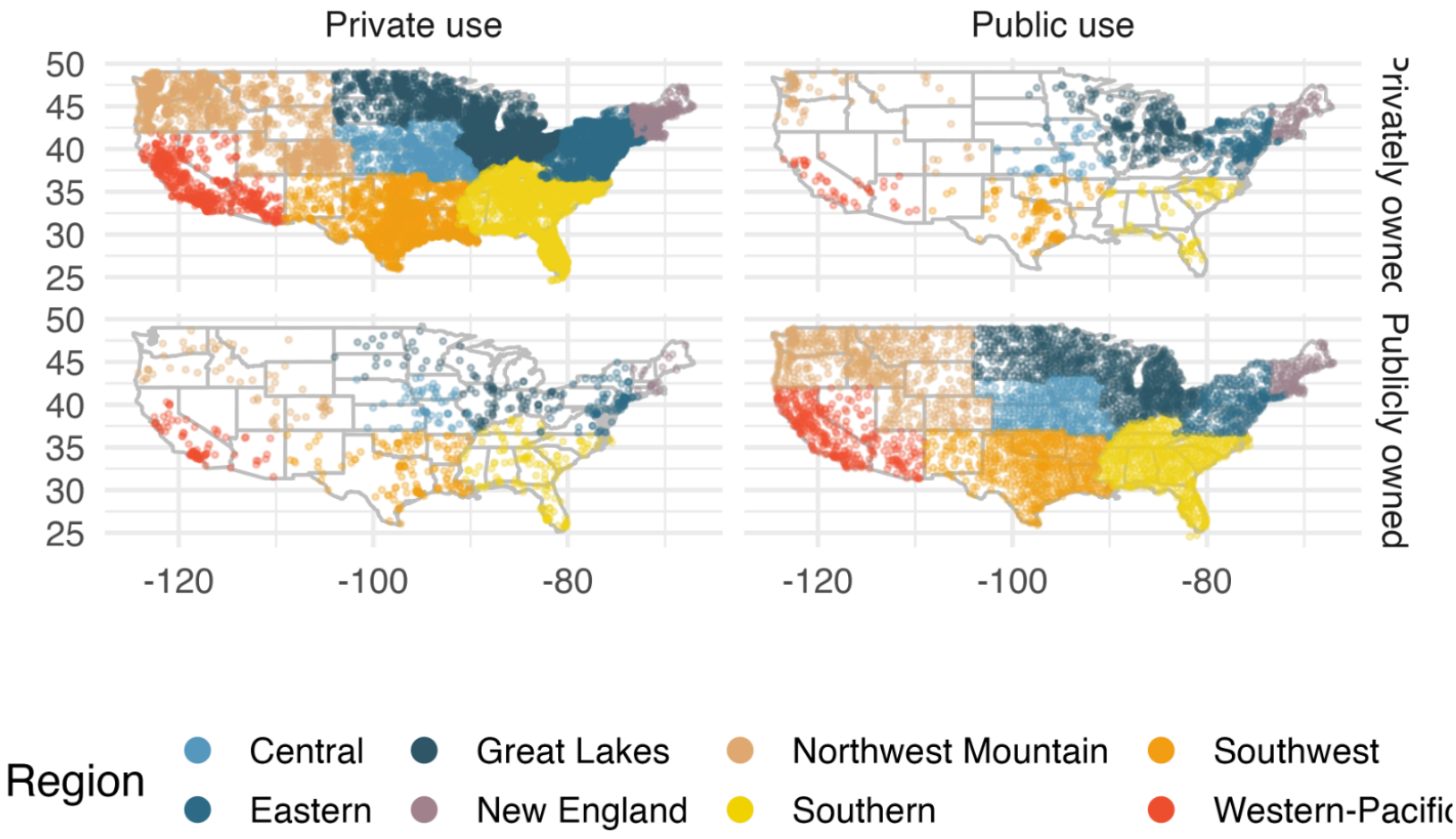
Exercise

[IMS 1.13] US Airports. The visualization below shows the geographical distribution of airports in the contiguous United States and Washington, DC. This visualization was constructed based on a dataset where each observation is an airport.

a) List the variables you believe were necessary to create this visualization.

b) Indicate whether each variable in the study is numerical or categorical. If numerical, identify as continuous or discrete. If categorical, indicate if the variable is ordinal.

Exercise


Private use
Public use
Privately owned
Publicly owned
50
45
40
35
30
25
-120
-100
-80
Region
Central
Great Lakes
Northwest Mountain
Southwest
Eastern
New England
Southern
Western-Pacific