STOR 155.010 Lecture 02
Data Basics

Aniruddhan (Ani) Ganesaraman

Fall 2026
UNC Chapel Hill

20 August 2026

Fun Facts: Guess Who?

- Trying to grow a 1,000+ lb pumpkin back home in Idaho
- Traveled to 6 of 7 continents before turning 12
- Runs a small home bakery, going for 6 years
- Has 5 cats and 2 dogs
- Can juggle
- Learned Dutch before visiting a friend in the Netherlands
- Building multiple startups simultaneously
- D1 football player + 4 state powerlifting records + first responder
- On the UNC women's lacrosse team
- Won a state championship in soccer
- Has been on swim team since age 4
- Exchange student from Spain
- Has 6 dogs

Class statistics

From the 45 responses, here's what we learned:

- **Class year**: 60% freshmen, 24% sophomores, and a handful of juniors and seniors.
- **Majors**: 14 Biology (Neuro, public health, pre-med), 12 Business, 9 Nursing. Also Applied Math, Economics, and Psychology.
- **Favorite Triangle restaurant**: Purple Bowl (3), Que Chula (2) and Carolina Coffee Shop (2).
- **Interests**: Health/medical topics (23), sports (13), politics (5), economics/business (5). Also MotoGP, organ transplantation.

Back to the CFS study: variables

Recall the chronic fatigue syndrome (CFS) trial. Below are some of the variables recorded for each patient in the study:

- `group`: Was the patient assigned to the treatment (cognitive behavior therapy) or control (relaxation) group?
- `age`: How old is the patient (in years)?
- `sex`: What is the patient's sex?
- `baseline_fatigue`: How severe was the patient's fatigue at the start of the study, on a scale from 0 (none) to 100 (extreme)?
- `completed`: Did the patient complete the study, or drop out?
- `outcome`: Did the patient have a good or poor outcome at the 6-month follow-up?

Data matrix

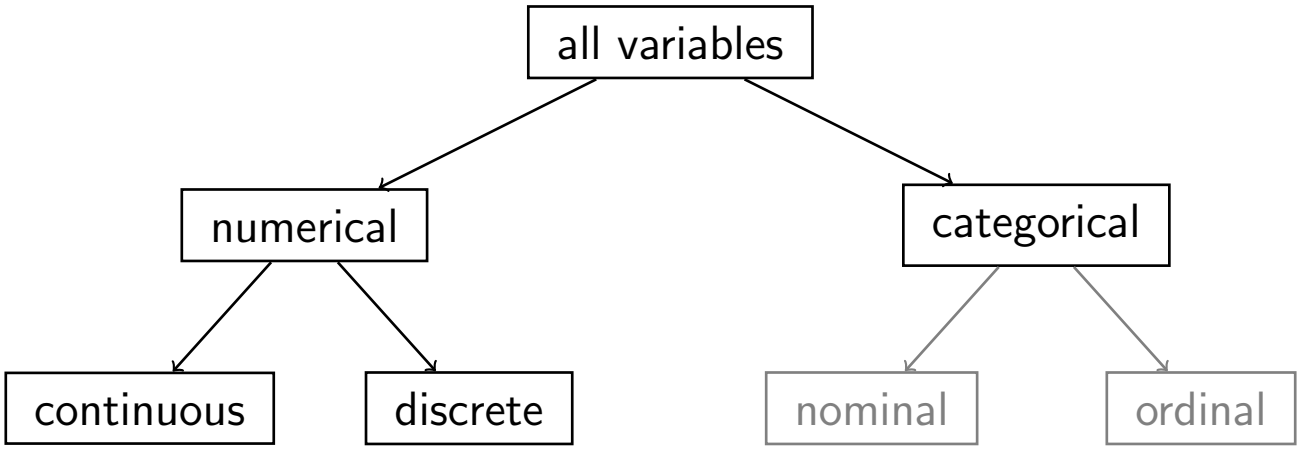
Data collected on patients in the CFS study on a variety of variables:

variable ↓

Pt.	group	age	⋯	outcome
1	treatment	34	⋯	good
2	control	41	⋯	poor
3	treatment	29	⋯	good
4	control	52	⋯	poor
⋮	⋮	⋮	⋮	⋮
53	control	37	⋯	good

← *observation*

Types of Variables



- **Variable:** Anything we can measure, count, or record about our data points.
- **Numerical:** Numbers that you can do meaningful math with.
  - **Continuous:** Numbers that can be any exact decimal or fraction.
  - **Discrete:** Numbers you count whole one by one (no fractions).
- **Categorical:** Data that puts things into groups or labels instead of numbers.
  - **Nominal:** Categories with no natural order or ranking.
  - **Ordinal:** Categories that have a built-in ranking or logical order.

Practice

What type of variable represents a student's grade level (e.g., freshman, sophomore, junior, senior)?

(a) numerical, continuous

(b) numerical, discrete

(c) categorical, nominal

(d) categorical, ordinal

Classroom survey

A survey was conducted on students in an introductory statistics course. Below are a few of the questions on the survey, and the corresponding variables the data from the responses were stored in:

- `gender`: What is your gender?
- `intro_extra`: Do you consider yourself introverted or extraverted?
- `sleep`: How many hours do you sleep at night, on average?
- `bedtime`: What time do you usually go to bed?
- `countries`: How many countries have you visited?
- `dread`: On a scale of 1-5, how much do you dread being here?

Types of variables (cont.)

	gender	sleep	bedtime	countries	dread
1	male	5	12-2	13	3
2	female	7	10-12	7	2
3	female	5.5	12-2	1	4
4	female	7	12-2		2
5	female	3	12-2	1	3
6	female	3	12-2	9	4

Practice

What type of variable is the count of apples in a basket?

(a) numerical, continuous

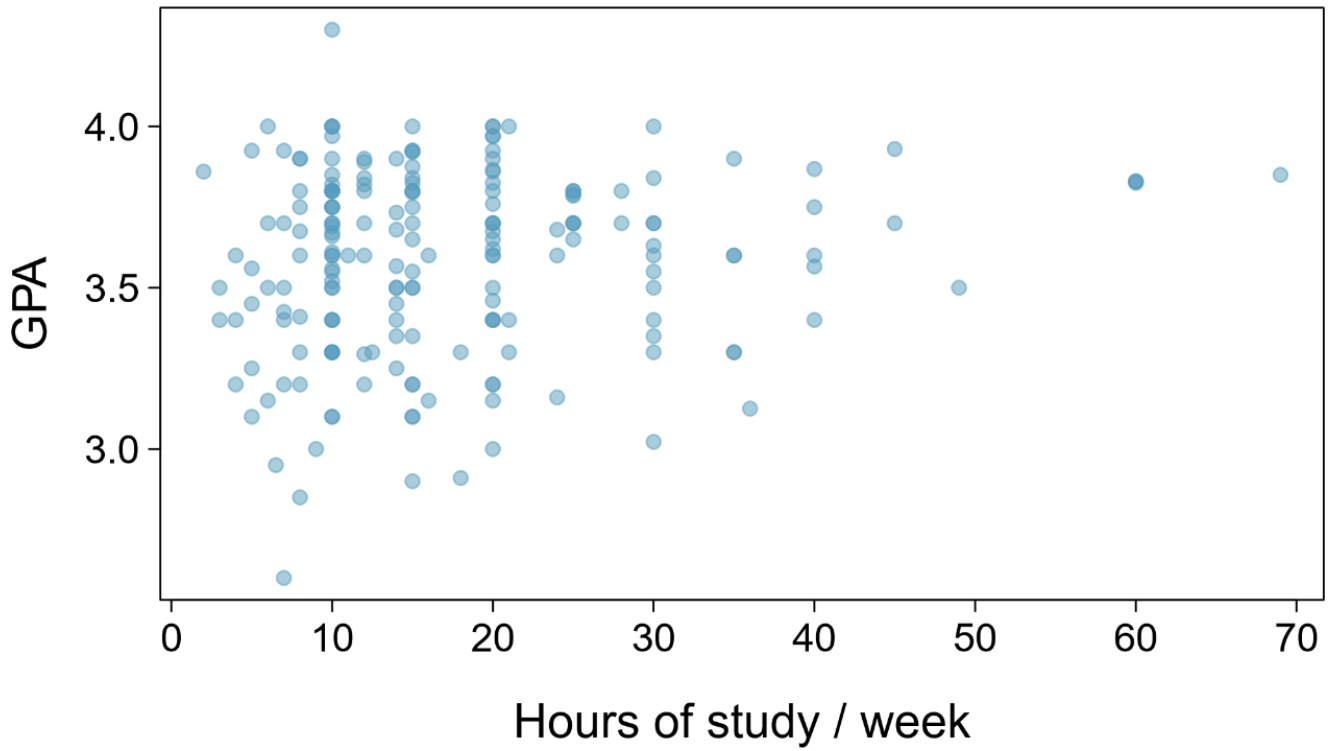
(b) numerical, discrete

(c) categorical, nominal

(d) categorical, ordinal

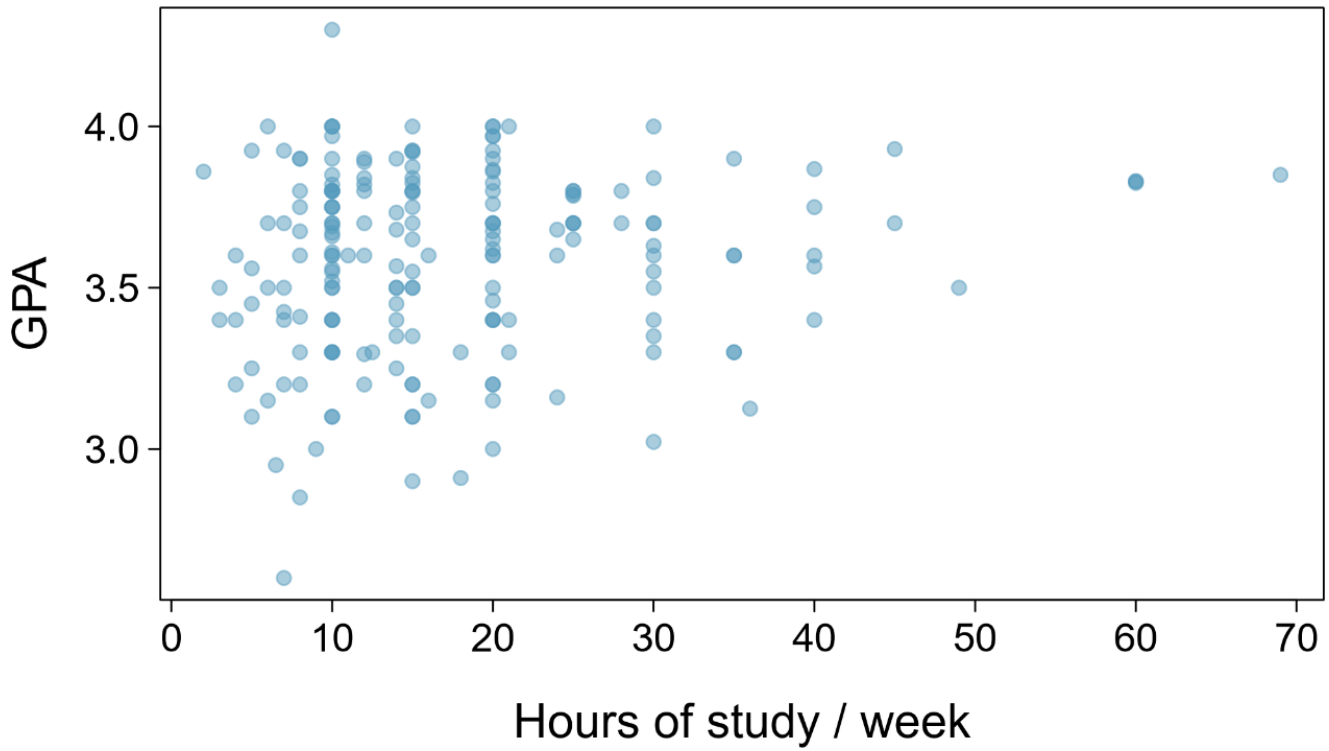
Scatterplot

- Does there appear to be a relationship between GPA and number of hours students study per week?



Scatterplot

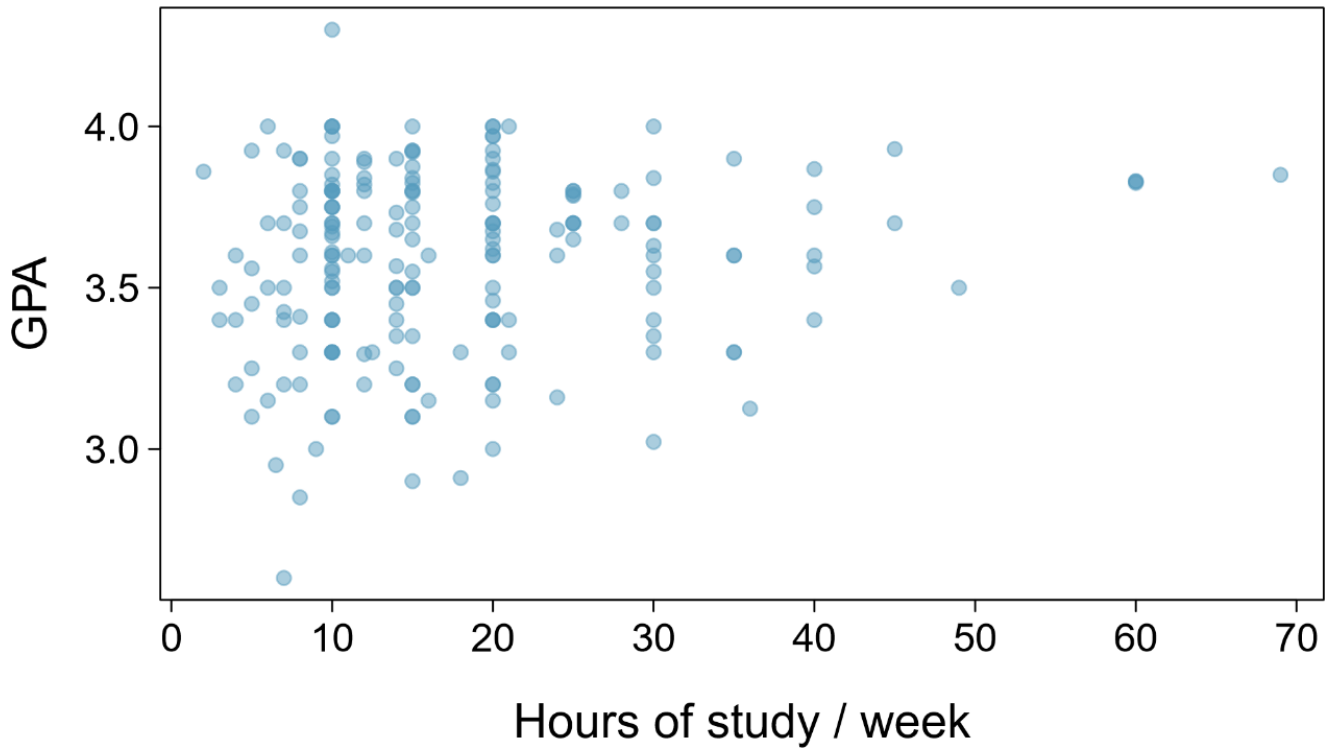
- Does there appear to be a relationship between GPA and number of hours students study per week?



- Can you spot anything unusual about any of the data points?

Scatterplot

- Does there appear to be a relationship between GPA and number of hours students study per week?



- Can you spot anything unusual about any of the data points?
- *There is one student with GPA > 4.0, which is not possible. This looks like a data error.*

Explanatory and Response Variables

- **Explanatory Variables (Predictors)**: These are the variables that are used to predict or explain variations in the response variable. Often referred to as independent variables.
- **Response Variable (Outcome)**: This is the variable that we want to predict or explain. It is also known as the dependent variable.
- **Relationship**: The relationship between explanatory and response variables may be causal or correlative and is often expressed in a mathematical or statistical model.

Explanatory and response variables

- To identify the explanatory variable in a pair of variables, identify which of the two is suspected of affecting the other:

$$\text{explanatory variable} \xrightarrow{\textit{might affect}} \text{response variable}$$

- Labeling variables as explanatory and response does not guarantee the relationship between the two is actually causal, even if there is an association identified between the two variables.

- We use these labels only to keep track of which variable we suspect affects the other.

- **Correlation DOES NOT imply causation**!

Two primary types of data collection

- **Observational studies:** Collect data in a way that does not directly interfere with how the data arise (e.g., surveys).
    - Can provide evidence of a naturally occurring association between variables, but they cannot by themselves show a causal connection.

Two primary types of data collection

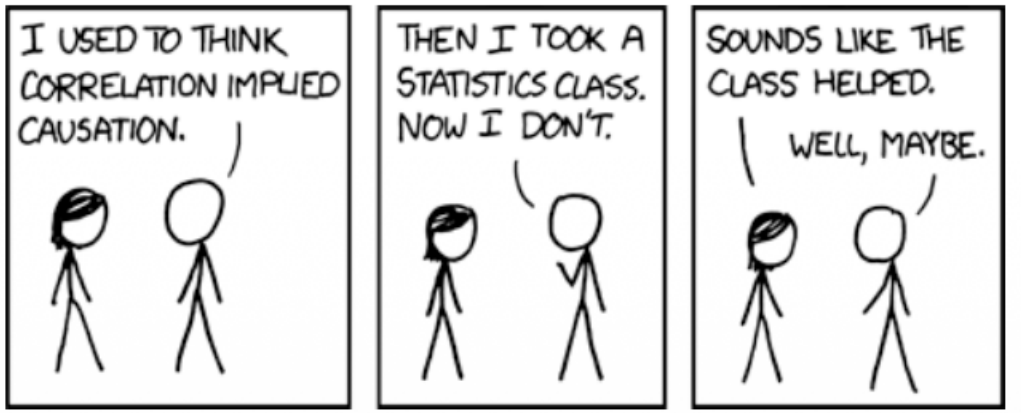
- **Observational studies:** Collect data in a way that does not directly interfere with how the data arise (e.g., surveys).
    - Can provide evidence of a naturally occurring association between variables, but they cannot by themselves show a causal connection.
- **Experiment:** Researchers randomly assign subjects to various treatments in order to establish causal connections between the explanatory and response variables.
    - In general, association does not imply causation, and causation can only be inferred from a randomized experiment.

Correlation vs. Causation

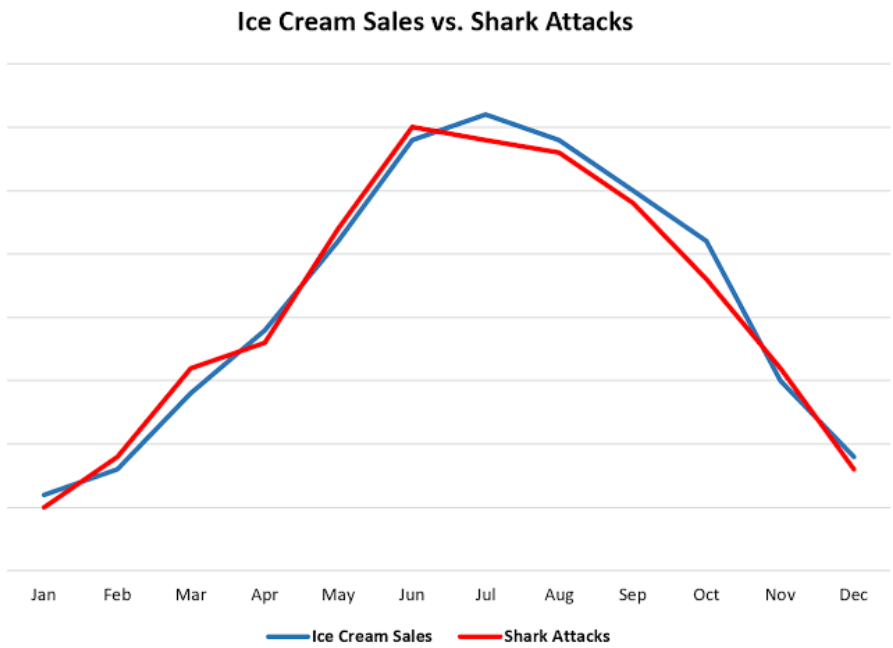
- When two variables show some connection with one another, they are called correlated/associated variables.
    - Associated variables can also be called dependent variables and vice-versa.
- If two variables are not associated, i.e., there is no evident connection between the two, then they are said to be independent.
- In general, correlation does not imply causation, and causation can only be inferred from a randomized experiment.

Correlation vs. Causation

- When two variables show some connection with one another, they are called correlated/associated variables.
  - Associated variables can also be called dependent variables and vice-versa.
- If two variables are not associated, i.e., there is no evident connection between the two, then they are said to be independent.
- In general, correlation does not imply causation, and causation can only be inferred from a randomized experiment.



Correlation vs. Causation (contd.)



Correlation

- A statistical measure that describes the extent to which two variables change together.
- If one variable tends to go up when the other goes up, there is a positive correlation.
- If one variable tends to go down when the other goes up, there is a negative correlation.
- *Example:* Imagine you're examining the relationship between the amount of time students spend studying for a test and their respective test scores.

Correlation

- A statistical measure that describes the extent to which two variables change together.
- If one variable tends to go up when the other goes up, there is a positive correlation.
- If one variable tends to go down when the other goes up, there is a negative correlation.
- *Example:* Imagine you're examining the relationship between the amount of time students spend studying for a test and their respective test scores. You might find that, generally, as the number of hours spent studying increases, test scores also increase. This would be a positive correlation.

Causation

- A change in one variable is responsible for a change in another. Three conditions must be met:

Causation

- A change in one variable is responsible for a change in another. Three conditions must be met:
  1. There's a correlation between the two variables.

Causation

- A change in one variable is responsible for a change in another. Three conditions must be met:
  1. There's a correlation between the two variables.
  2. The cause happens before the effect.

Causation

- A change in one variable is responsible for a change in another. Three conditions must be met:
  1. There's a correlation between the two variables.
  2. The cause happens before the effect.
  3. There aren't any other factors that could explain the relationship.

Causation

- A change in one variable is responsible for a change in another. Three conditions must be met:
  1. There's a correlation between the two variables.
  2. The cause happens before the effect.
  3. There aren't any other factors that could explain the relationship.
- *Example:* Taking a painkiller (like ibuprofen) to reduce a headache. If headaches consistently diminish or disappear after taking the pill, and no other factors are at play (e.g., you didn't also rest in a dark room), then we can infer a causal relationship between the painkiller and the reduction of headache pain.

Reading and Further References

- Today and next class: Sections 1.1 and 1.2 of OpenIntro Statistics
- More optional but fun readings:
  - *How to Lie with Statistics* by Darrell Huff
  - *The Signal and the Noise* by Nate Silver (fun for anyone into sports, elections, or weather forecasting)
  - *Naked Statistics: Stripping the Dread from the Data* by Charles Wheelan