Chapter 1

Introduction to data[1]

Department of Mathematics & Statistics
North Carolina A&T State University

[1] These notes use content from OpenIntro Statistics Slides by Mine Cetinkaya-Rundel.

Data Basics

Classroom survey

A survey was conducted on students in an Intro to Stat 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_extro*: Do you consider yourself introverted or extroverted?
- *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?

Data Matrix

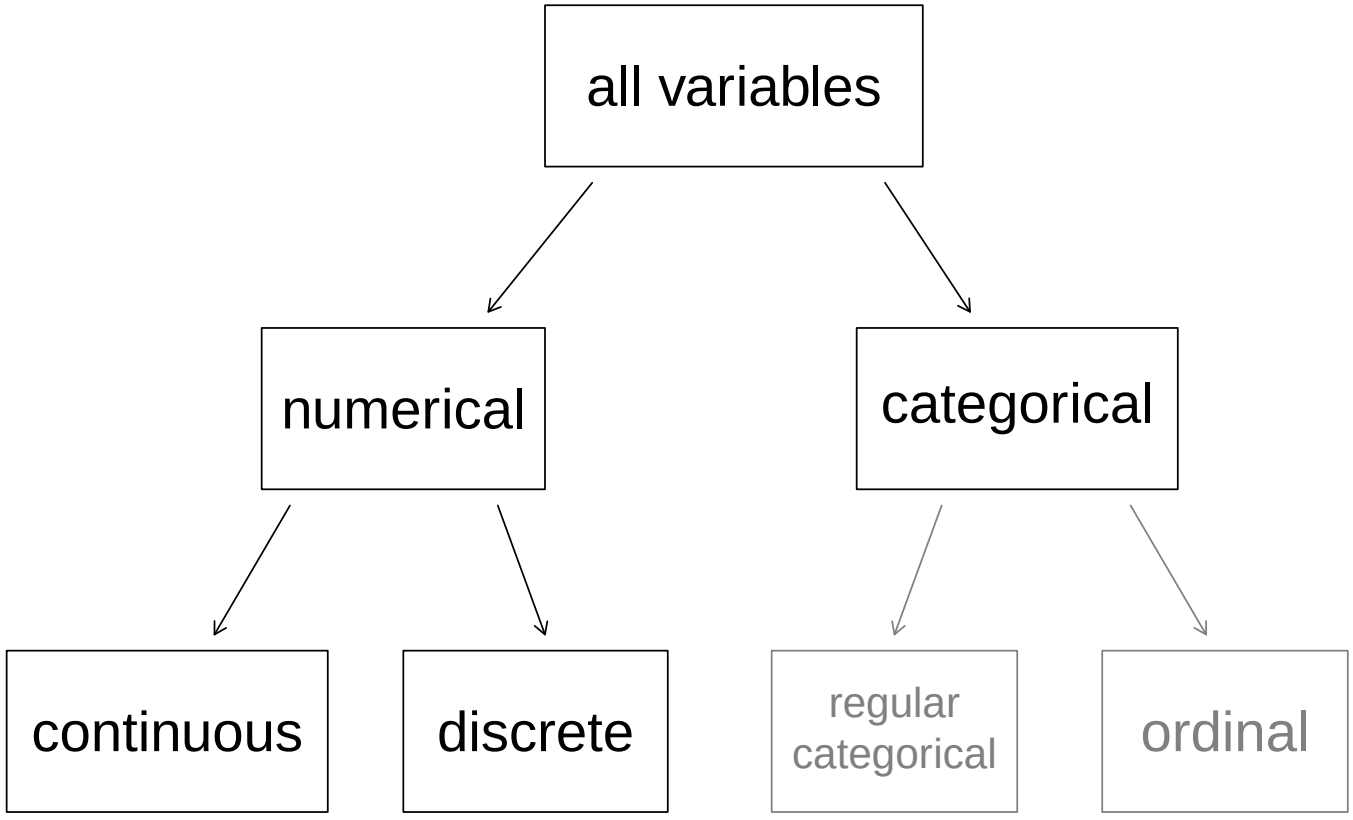
Data collected on students in a statistics class on a variety of variables:

variable ↓

Student	gender	intro_extro	...	dread
1	male	extrovert	...	3
2	female	extrovert	...	2
3	female	introvert	...	4
4	female	extrovert	...	2
⋮	⋮	⋮	⋮	⋮
86	male	extrovert	...	3

← observation

Types of variables



Types of variables

Student	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

- gender: Type?

Types of variables

Student	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

- gender: **Categorical**

Types of variables

Student	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

- gender: **Categorical**
- sleep: Type?

Types of variables

Student	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

- gender: **Categorical**
- sleep: **Numerical, Continuous**

Types of variables

Student	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

- gender: **Categorical**
- sleep: **Numerical, Continuous**
- bedtime: Type?

Types of variables

Student	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

- gender: **Categorical**
- sleep: **Numerical, Continuous**
- bedtime: **Categorical, Ordinal**

Types of variables

Student	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

- gender: **Categorical**
- sleep: **Numerical, Continuous**
- bedtime: **Categorical, Ordinal**
- countries: Type?

Types of variables

Student	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

- gender: **Categorical**
- sleep: **Numerical, Continuous**
- bedtime: **Categorical, Ordinal**
- countries: **Numerical, Discrete**

Types of variables

Student	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

- gender: **Categorical**
- sleep: **Numerical, Continuous**
- bedtime: **Categorical, Ordinal**
- countries: **Numerical, Discrete**
- dread: Type?

Types of variables

Student	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

- gender: **Categorical**
- sleep: **Numerical, Continuous**
- bedtime: **Categorical, Ordinal**
- countries: **Numerical, Discrete**
- dread: **Categorical, Ordinal**
  - Could also be used as **Numerical**

Practice

What type of variable is a telephone area code?

A) Numerical, Continuous
B) Numerical, Discrete
C) Categorical
D) Categorical, Ordinal

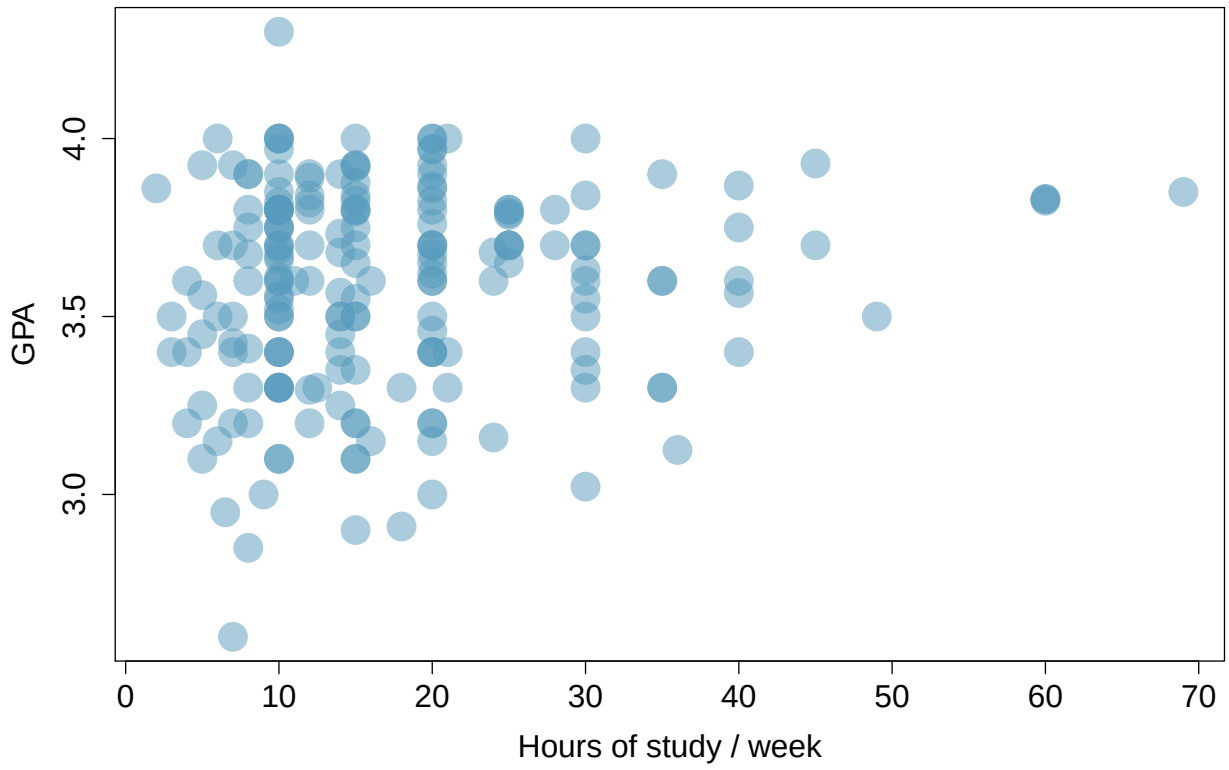
Practice

What type of variable is a telephone area code?

A) Numerical, Continuous
B) Numerical, Discrete
C) Categorical
D) Categorical, Ordinal

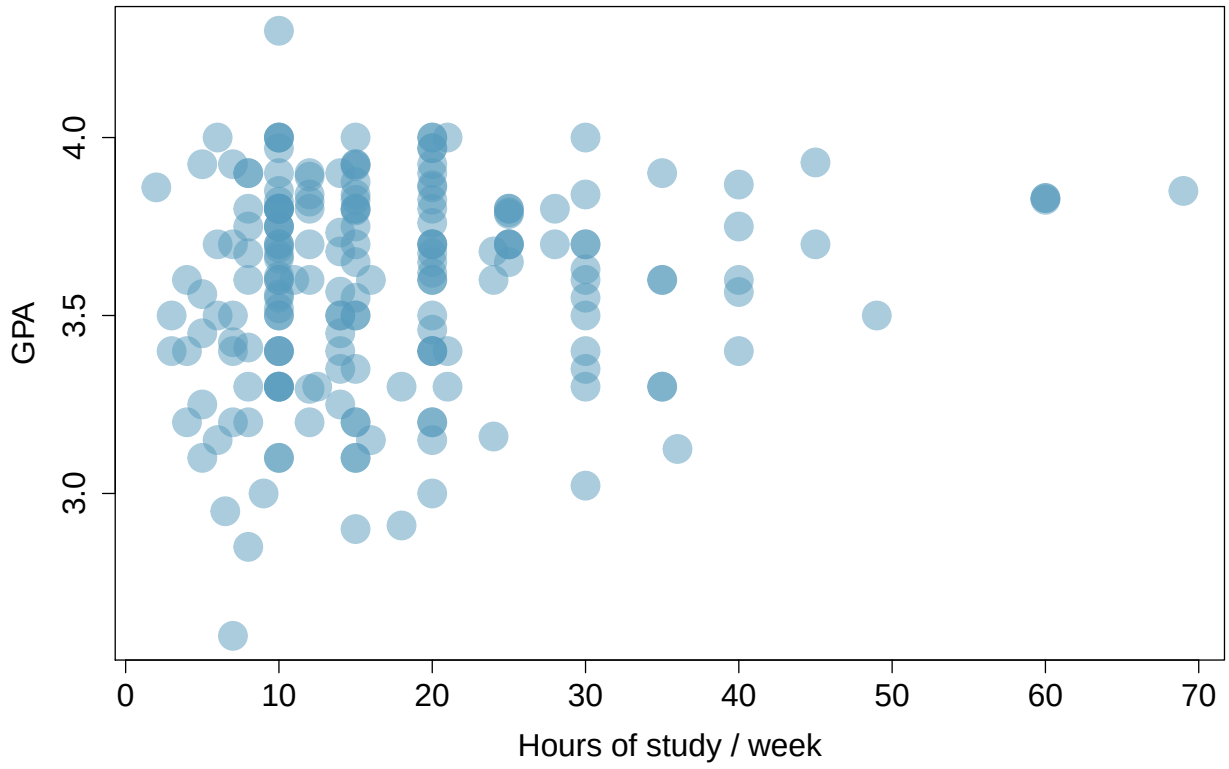
Relationships among variables

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



Relationships among variables

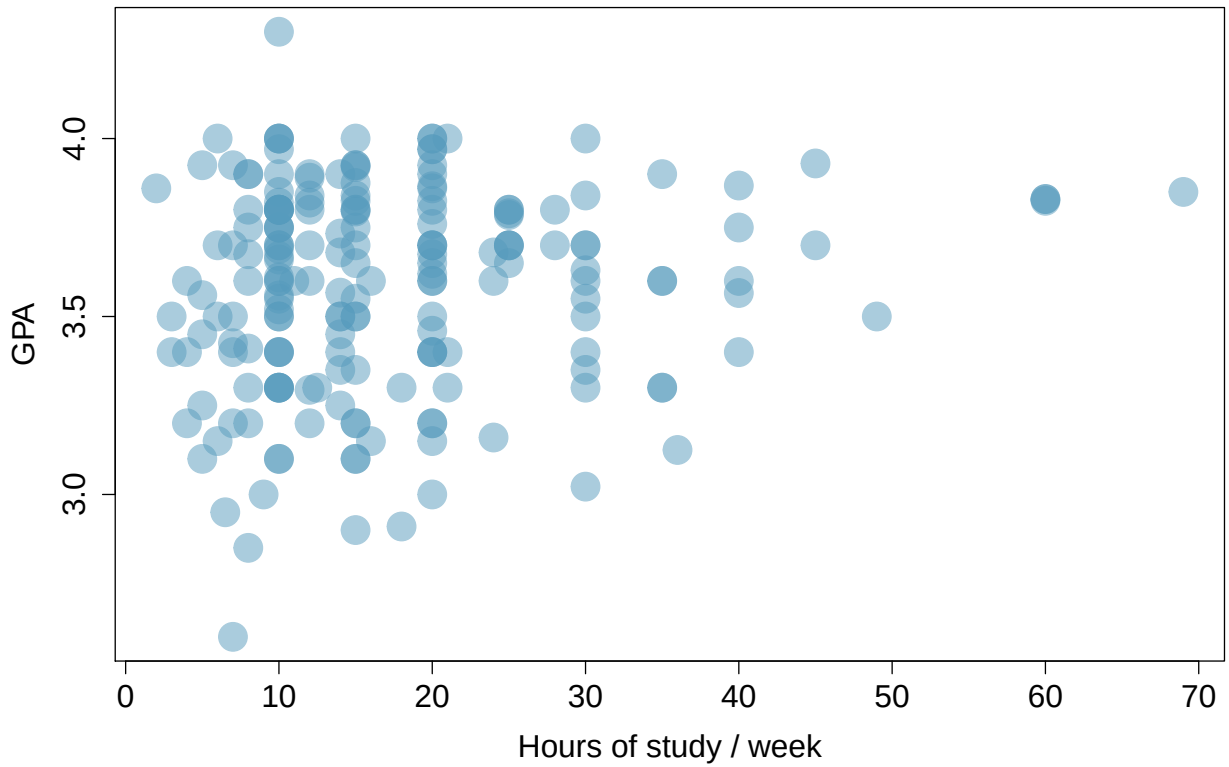
Does there appear to be a relationship between GPA and number of hours students study per week?



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

Relationships among variables

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 with GPA > 4.0, this is likely a data error.

Explanatory and reponse variables

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

$$\text{explanatory variable} \xrightarrow{\text{might affect}} \text{response variable} \tag{1}$$

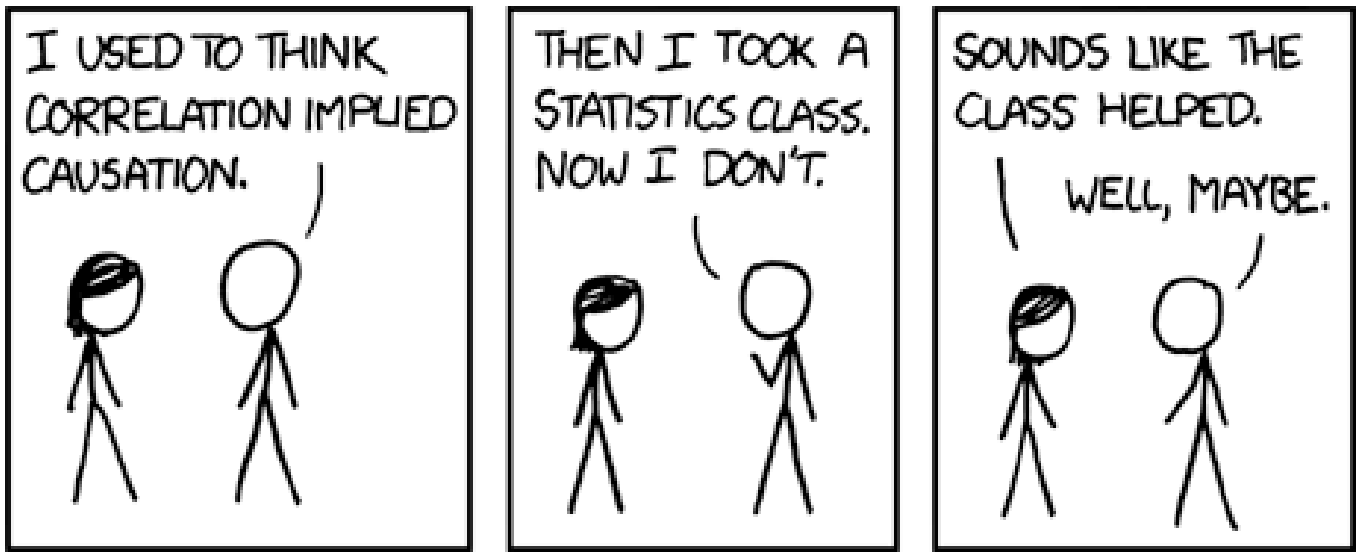
- Labeling variables as explanatory and response does not guarantee the relationship between the two is actually causal, even if there is an association identified between the two variables. We use these labels only to keep track of which variable we suspect affects the other.

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.

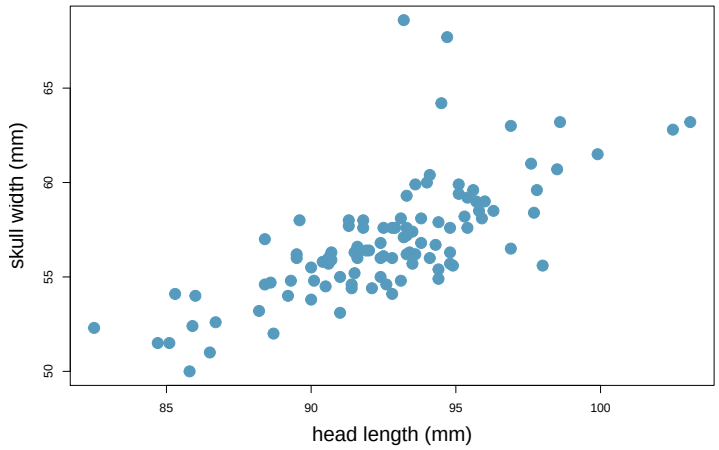
Association vs. Causation

- When two variables show some connection with one another, they are called **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, association does not imply causation, and causation can only be inferred from a randomized experiment.



Practice

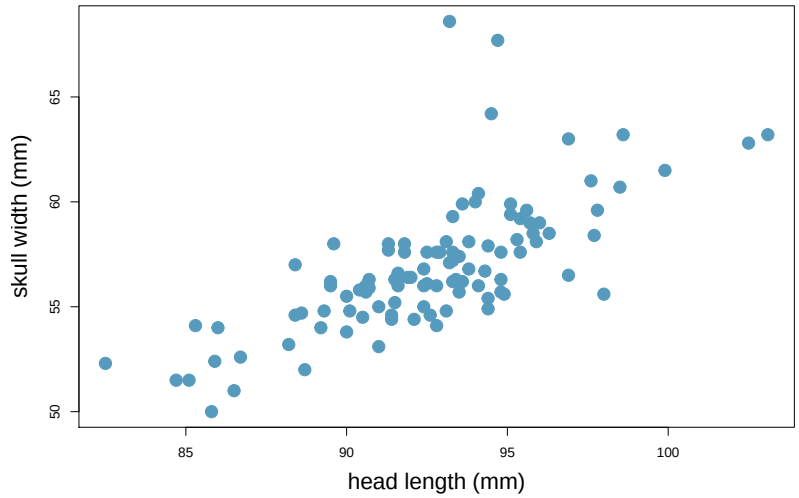
Based on the scatterplot on the right, which of the following statements is correct about the head and skull lengths of possums?



A) There is no relationships between head length and skull width, i.e. the variables are independent.
B) Head length and skull width are positively associated.
C) Skull width and head length are negatively associated.
D) A longer head causes the skull to be wider.
E) A wider skull causes the head to be longer.

Practice

Based on the scatterplot on the right, which of the following statements is correct about the head and skull lengths of possums?



A) There is no relationships between head length and skull width, i.e. the variables are independent.
B) Head length and skull width are positively associated.
C) Skull width and head length are negatively associated.
D) A longer head causes the skull to be wider.
E) A wider skull causes the head to be longer.