# Introduction to Data Science

## Getting Data and Simpson's Paradox

Joanna Bieri DATA101

# Important Information

- Email: joanna_bieri@redlands.edu
- Office Hours: Duke 209 Click Here for Joanna's Schedule

# Announcements

**NEXT WEEK - Data Ethics** This week you should be reading your book or articles.

# Data

# Where does data come from?

1. Scientific Studies
2. User Data
3. Geographic or Spatial Data

Today we will focus on Scientific Studies

# Two types of Scientific Studies

- **Observational Data** Data collected so that collection does not interfere with how the data arise. The goal is to establish associations. Cannot find causation.
- **Experimental Data** Randomly assign treatment groups, gather data from those groups. The goal is to establish causal connections.

# Variables in a Study

When we are trying to say that A causes B we use two variable types:

**Explanatory Variable**

**Response Variable**

# When interacting with Data

| | Random assignment | No random assignment | |
|---|---|---|---|
| Random sampling | Causal and generalizable | Not causal, but generalizable | Generalizable |
| No random sampling | Causal, but not generalizable | Neither causal nor generalizable | Not generalizable |
| | Causal | Not causal | |

# Climate Change Example

# Example - Climate Change

A July 2019 YouGov survey asked 1633 GB and 1333 USA randomly selected adults which of the following statements about the global environment best describes their view:

They had the choices:

- The climate is changing and human activity is mainly responsible
- The climate is changing and human activity is partly responsible, together with other factors
- The climate is changing but human activity is not responsible at all
- The climate is not changing

Source: YouGov - International Climate Change Survey

**What kind of study is this?**

**What kind of results can we find?**

# Climate data

```
file_name = 'data/yougov-climate.csv'
DF = pd.read_csv(file_name,index_col=0)
```

# First question:

What percent of **all respondents** think the climate is changing and human activity is mainly responsible?

# Get totals across rows and columns:

The **.sum()** command can sum the rows (axis=0) or the columns (axis=1) of a data frame.

```
DF.loc['total']= DF.sum(axis=0)
DF['total'] = DF.sum(axis=1)
```

# Calculating Percent

**Percent** is just the part divided by the total. So we need to:

- Find the total number of survey respondents.
- Find the total number of survey respondents who thought human activity is responsible.
- Divide those two numbers to get a decimal

```
np.float64(0.45178691840863117)
```

This means that overall 45.2% of people think that climate is changing and human activity is mainly responsible.

# Second question:

What percent of **GB respondents** think the climate is changing and human activity is mainly responsible?

`np.float64(0.5101041028781383)`

Of the people would responded from Great Britain, 51% think that climate is changing and human activity is mainly responsible.

# Third Question

What percent of **US respondents** think the climate is changing and human activity is mainly responsible?

```
np.float64(0.3803450862715679)
```

When we consider US respondents we see that 38% (lower than GB) believe that climate is changing and human activity is mainly responsible.

# What can we say?

- There is a clear difference between Great Britain and the USA.
- So there is an association between country and your answer to this question.
- Could there be other confounding variables? YES!

We cannot say that being from GB causes you to be more likely to believe climate changes is caused by human activity. We can only say that there is a relationship between country and your belief.

# Probability

# Conditional Probability

What do these percentages represent? We can say:

- If you are from the US there is a 38% probability that you believe climate change is caused by human activity.
- If you are from the GB there is a 51% probability that you believe climate change is caused by human activity.

This is called **Conditional Probability**

$$P(A|B):$$

Probability of event A given event B

# Conditional Probability

Example: two different questions:

1. What is the probability that it will be unseasonably warm tomorrow?
2. What is the probability that it will be unseasonably warm tomorrow **given** that it was unseasonably warm today?

# Independence

If knowing event A happened tells you something about event B happening - then these two things are linked - and we say A and B are not independent.

Otherwise events A and B are said to be independent and we know $P(A|B) = P(A)$ - because knowing B doesn't tell you anything about A.

# Simpson's Paradox

# Simpson's Paradox

- When we see a trend when we look at the data grouping one way, but the trend disappears or reverses when we look at the data another way.
- Not considering an important variable when studying a relationship can result in **Simpson's paradox**
- Simpson's paradox illustrates the effect that omission of an explanatory variable can have on the measure of association between another explanatory variable and a response variable
- The inclusion of a third variable in the analysis can change the apparent relationship between the other two variables

# Berkeley admission data example

- Study carried out by the Graduate Division of the University of California, Berkeley in the early 70's to evaluate whether there was a gender bias in graduate admissions.
- The data come from six departments. For confidentiality we'll call them A-F.
- We have information on whether the applicant was male or female and whether they were admitted or rejected. This is an old study so only two binary classifications were used.

# Here is the data

| | Department | Male Yes | Male No | Female Yes | Female No |
|---|---|---|---|---|---|
| 0 | A | 512 | 313 | 89 | 19 |
| 1 | B | 353 | 207 | 17 | 8 |
| 2 | C | 120 | 205 | 202 | 391 |
| 3 | D | 138 | 279 | 131 | 244 |
| 4 | E | 53 | 138 | 94 | 299 |
| 5 | F | 22 | 351 | 24 | 317 |

```
3    0.445188
Name: Number, dtype: float64
1    0.303542
Name: Number, dtype: float64
```

# Percent Males vs Females

$$P(Admit|Male) = 0.445$$

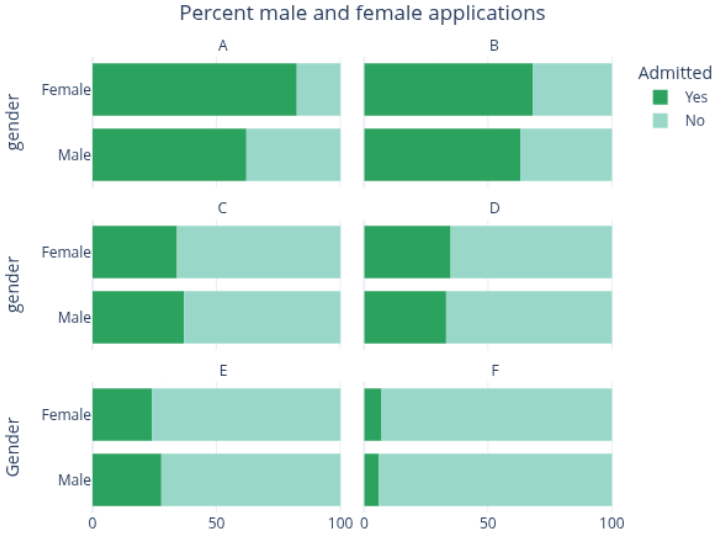
$$P(Admit|Female) = 0.304$$

So the answer is YES the admission rate is higher for males.

# Percent Males vs Females

```
Unable to display output for mime type(s): application/vnd.plo
```

# Gender distribution by department



# What do we see here?

Well actually in department A women were more likely to be admitted by quite a bit, and in the other departments they were either just barely less than men or more than men. This seems to contradict what we saw in our plot above!

Which picture is the real story?!?

# Take a closer look at the departments

How many women were applying in the first place?

What happens in departments that are balanced?

Are all departments the same?

| | Department | Male Yes | Male No | Female Yes | Female No |
|---|---|---|---|---|---|
| 0 | A | 512 | 313 | 89 | 19 |
| 1 | B | 353 | 207 | 17 | 8 |
| 2 | C | 120 | 205 | 202 | 391 |
| 3 | D | 138 | 279 | 131 | 244 |
| 4 | E | 53 | 138 | 94 | 299 |
| 5 | F | 22 | 351 | 24 | 317 |

# Example of Simposon's Paradox

Consider some data, put it in a scatter plot and draw a line of best fit.

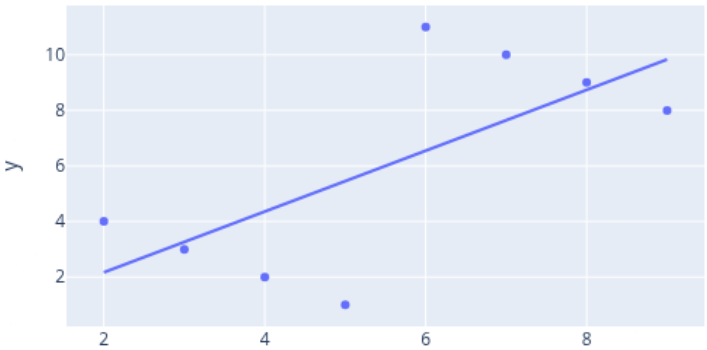
Here is the data

| | x | y | z |
|---|---|---|---|
| 0 | 2 | 4 | A |
| 1 | 3 | 3 | A |
| 2 | 4 | 2 | A |
| 3 | 5 | 1 | A |
| 4 | 6 | 11 | B |
| 5 | 7 | 10 | B |
| 6 | 8 | 9 | B |
| 7 | 9 | 8 | B |

# Example of Simposon's Paradox

```
fig = px.scatter(df,x='x',y='y',trendline="ols")

fig.show()
```



# Example of Simposon's Paradox

But we ignored one of the variables z. If we include z we get a completely different picture. On average the data looked very different than when considering the subgroups.

```
fig = px.scatter(df,x='x',y='y',color='z',trendline="ols")

fig.show()
```