

Introduction to Data Science

Getting Data and Simpson's Paradox

Joanna Bieri DATA101

Mini Review - Two Types of Scientific Studies

REVIEW

- **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.

Mini Review - What Does This Mean About My Data?

REVIEW

	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	

Random Assignment: participants are assigned to treatment groups randomly. **Random Sampling:** randomly sample the population at large. **Generalizable:** the study should apply to the general population.

Mini Review - Climate Change Survey

REVIEW

A July 2019 YouGov survey asked 1633 GB and 1333 USA randomly selected adults which statement best describes their view. The first column is “The climate is changing and human activity is mainly responsible”.

DF

	mainly	partly	not at all	not changing	dont know
GB	833	604	49	33	114
US	507	493	120	80	133

What percent of all respondents think human activity is mainly responsible? What percent of GB respondents? Of US respondents?

Mini Review - Totals and Percents

REVIEW

```
DF.loc['total'] = DF.sum(axis=0)    # add down each column  
DF['total'] = DF.sum(axis=1)       # add across each row
```

Percent is just the part divided by the total.

```
DF['mainly']['total'] / DF['total'].loc['total']    # 0.452  
DF['mainly']['GB'] / DF['total'].loc['GB']         # 0.51  
DF['mainly']['US'] / DF['total'].loc['US']         # 0.38
```

Mini Review - Conditional Probability and Independence

REVIEW

- If you are from GB there is a 51% probability that you believe climate change is mainly caused by human activity. If you are from the US it is 38%.
- **Conditional Probability** $P(A | B)$: the probability of event A given event B.
- If knowing B happened tells you something about A, then A and B are **not independent**.
- If A and B are **independent** then $P(A | B) = P(A)$, because knowing B does not tell you anything about A.

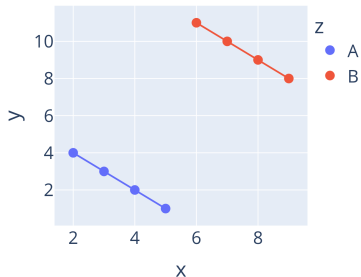
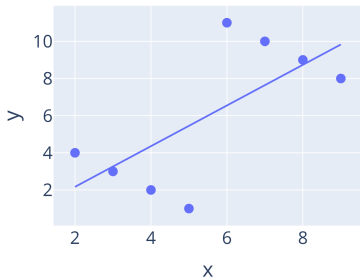
Mini Review - Simpson's Paradox

REVIEW

- We see a trend when we look at the data grouped 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**.
- Including a third variable in the analysis can change the apparent relationship between the other two variables.

Mini Review - Simpson's Paradox

REVIEW



Left: x and y with a line of best fit. Right: the same data when we include the variable z .

Types of Studies

ASK

A gym asks its members how much they exercise
and how well they sleep.

Members who exercise more sleep better.

Does exercise cause better sleep?

Types of Studies

ANSWER

- This is an observational study. Nobody was assigned to exercise.
- We can say: in this data, people who exercise more sleep better.
- We cannot say: exercise causes better sleep.
- There could be another variable (age, stress, job) that explains both.

Percent and Conditional Probability

PREDICT

df

	yes	no	total
GB	6	2	8
US	3	9	12
total	9	11	20

```
df['yes'].loc['GB'] / df['total'].loc['GB']
```

What number comes out? What does it mean in words?

Percent and Conditional Probability

ANSWER

```
df['yes'].loc['GB'] / df['total'].loc['GB']  
  
0.75
```

If a person is from GB there is a 75% probability that they said yes. We write this $P(\text{yes} \mid \text{GB}) = 0.75$. For the US it is $3/12 = 0.25$.

Percent and Conditional Probability

REVERSE

df

	yes	no	total
GB	6	2	8
US	3	9	12
total	9	11	20

We want the percent of ALL the people who said yes. What is the command?

Percent and Conditional Probability

ANSWER

```
df['yes'].loc['total'] / df['total'].loc['total']
```

```
0.45
```

Percent is the part divided by the total. Here the part is the yes column total (9) and the total is everyone (20).

Simpson's Paradox

PREDICT

df

	dept	gender	applied	admitted
0	easy	men	80	48
1	easy	women	20	14
2	hard	men	20	2
3	hard	women	80	16

```
df.groupby('gender', as_index=False)[  
    ['applied', 'admitted']].sum()
```

What comes out? Who was admitted at a higher rate?

Simpson's Paradox

ANSWER

```
df.groupby('gender', as_index=False)[  
    ['applied', 'admitted']].sum()
```

	gender	applied	admitted
0	men	100	50
1	women	100	30

Overall 50% of the men and 30% of the women were admitted.

Simpson's Paradox

ASK

df

	dept	gender	applied	admitted	rate
0	easy	men	80	48	0.6
1	easy	women	20	14	0.7
2	hard	men	20	2	0.1
3	hard	women	80	16	0.2

Same data, with $\text{rate} = \text{admitted}/\text{applied}$. Who has the higher rate in each department? Overall it was the men, 50% to 30%.
How can both be true?

Simpson's Paradox

ANSWER

- Women have the higher rate in BOTH departments.
- 80 of the 100 women applied to the hard department. 80 of the 100 men applied to the easy one.
- **Simpson's Paradox**: a trend reverses (or disappears) when we group the data by another variable.

Questions?

Questions?