Scientific studies, confounding, and Simpson's paradox

Dr. Maria Tackett

09.12.19

Click for PDF of slides

Announcements

- HW 01 - due **Monday, September 16 at 12p (noon)**
- Writing Exercise 1
  - Draft Response: 9/12 - 9/14
  - Peer Review: 9/15 - 9/17
  - Final Revision: 9/18 - 9/19

`mutate`

(from September 5 lecture)

`mutate` to add new variables

```
ncbikecrash %>%
  mutate(driver_alcohol_drugs_simplified = case_when(
    is.na(driver_alcohol_drugs)              ~ driver_alcohol_drugs,
    driver_alcohol_drugs == "Missing"        ~ NA_character_,
    str_detect(driver_alcohol_drugs, "Yes") ~ "Yes",
    TRUE                                     ~ "No"
  ))
```

"Save" when you `mutate`

Most often when you define a new variable with **`mutate`** you'll also want to save the resulting data frame, often by writing over the original data frame.

```
ncbikecrash <- ncbikecrash %>%
  mutate(driver_alcohol_drugs_simplified = case_when(
    is.na(driver_alcohol_drugs)              ~ driver_alcohol_drugs,
    driver_alcohol_drugs == "Missing"        ~ NA_character_,
    str_detect(driver_alcohol_drugs, "Yes") ~ "Yes",
    TRUE                                     ~ "No"
  ))
```

Check before you move on

```
ncbikecrash %>%
  count(driver_alcohol_drugs, driver_alcohol_drugs_simplified)
```

```
## # A tibble: 6 x 3
##   driver_alcohol_drugs                 driver_alcohol_drugs_simplified     n
##   <chr>                                <chr>                           <int>
## 1 <NA>                                 <NA>                             6654
## 2 Missing                              <NA>                               99
## 3 No                                   No                                695
## 4 Yes-Alcohol,  impairment suspected   Yes                                12
## 5 Yes-Alcohol, no impairment detected  Yes                                 3
## 6 Yes-Drugs, impairment suspected      Yes                                 4
```

```
ncbikecrash %>%
  count(driver_alcohol_drugs_simplified)
```

```
## # A tibble: 3 x 2
##   driver_alcohol_drugs_simplified     n
##   <chr>                           <int>
## 1 <NA>                             6753
## 2 No                                695
## 3 Yes                                19
```

Scientific studies

Scientific studies

- Observational
  - Collect data in a way that does not interfere with how the data arise ("observe")
  - Only establish an association
- Experimental
  - Randomly assign subjects to treatments
  - Establish causal connections

Design a study comparing average energy levels of people who do and do not exercise -- both as an observational study and as an experiment.

Study: "Cereal Keeps Girls Slim"

Girls who ate breakfast of any type had a lower average body mass index, a common obesity gauge, than those who said they didn't. The index was even lower for girls who said they ate cereal for breakfast, according to findings of the study conducted by the Maryland Medical Research Institute with funding from the National Institutes of Health (NIH) and cereal-maker General Mills.

[...]

The results were gleaned from a larger NIH survey of 2,379 girls in California, Ohio, and Maryland who were tracked between the ages of 9 and 19.

[...]

As part of the survey, the girls were asked once a year what they had eaten during the previous three days.

[...]

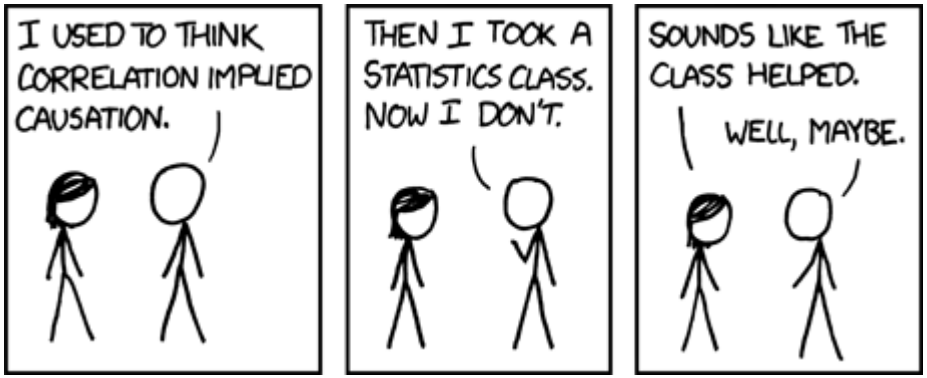
Source: Study: Cereal Keeps Girls Slim.

3 possible explanations

- Eating breakfast causes girls to be slimmer
- Being slim causes girls to eat breakfast
- A third variable is responsible for both -- a confounding variable

A **confounding** variable is an an extraneous variable that affects both the explanatory and the response variable, and that make it seem like there is a relationship between them

Correlation != causation



Randall Munroe CC BY-NC 2.5 http://xkcd.com/552/

Studies and conclusions

	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	

Non-random samples: a cautionary tale

In 2016, the Natural Environment Research Council in England started an online competition in an effort to name a polar research ship. People were invited to submit suggestions and/or cast a vote for their favorite choice.

What type of sampling design is this?

What happened?

Conditional probability

Conditional probability

Notation: $P(A|B)$: Probability of event A given event B

A: it will be unseasonably warm tomorrow

B: it is unseasonably warm today

- What is the probability that it will be unseasonably warm tomorrow?
  - What is $P(A)$?
- What is the probability that it will be unseasonably warm tomorrow, given that is unseasonably warm today?
  - What is $P(A|B)$?

A January 2018 SurveyUSA poll asked 500 randomly selected Californians whether they are familiar with the DREAM act. The distribution of the responses by age category are shown below.

What proportion of **all respondents** are very familiar with the DREAM act?

	18 - 49	50+	Total
Very	90	32	122
Somewhat	125	86	211
Not very	56	33	89
Not at all	36	24	60
Not sure	9	9	18
Total	316	184	500

$$P(\text{Very}) = \frac{122}{500} = 0.244$$

Source: SurveyUSA News Poll 23754

A January 2018 SurveyUSA poll asked 500 randomly selected Californians whether they are familiar with the DREAM act. The distribution of the responses by age category are shown below.

What proportion of **respondents who are 18 - 49 years old** are very familiar with the DREAM act?

	18 - 49	50+	Total
Very	90	32	122
Somewhat	125	86	211
Not very	56	33	89
Not at all	36	24	60
Not sure	9	9	18
Total	316	184	500

$$P(\text{Very} \mid 18-49) = \frac{90}{316} = 0.285$$

A January 2018 SurveyUSA poll asked 500 randomly selected Californians whether they are familiar with the DREAM act. The distribution of the responses by age category are shown below.

What proportion of **respondents who are 50+ years old** are very familiar with the DREAM act?

	18 - 49	50+	Total
Very	90	32	122
Somewhat	125	86	211
Not very	56	33	89
Not at all	36	24	60
Not sure	9	9	18
Total	316	184	500

$$P(\text{Very} \mid 50+) = \frac{32}{184} = 0.173$$

Given that

- $P(\text{Very}) = \frac{122}{500} = 0.244$
- $P(\text{Very} \mid 18 - 49) = \frac{90}{316} = 0.285$
- $P(\text{Very} \mid 50+) = \frac{32}{184} = 0.173$

does there appear to be a relationship between age and familiarity with the DREAM act? Explain your reasoning.

Could there be another variable that explains this relationship?

Independence

Inspired by the previous example and how we used the conditional probabilities to make conclusions, come up with a definition of independent events. If easier, you can keep the context limited to the example (independence/dependence of familiarity with the DREAM act and age), but try to push yourself to make a more general statement.

Simpson's paradox

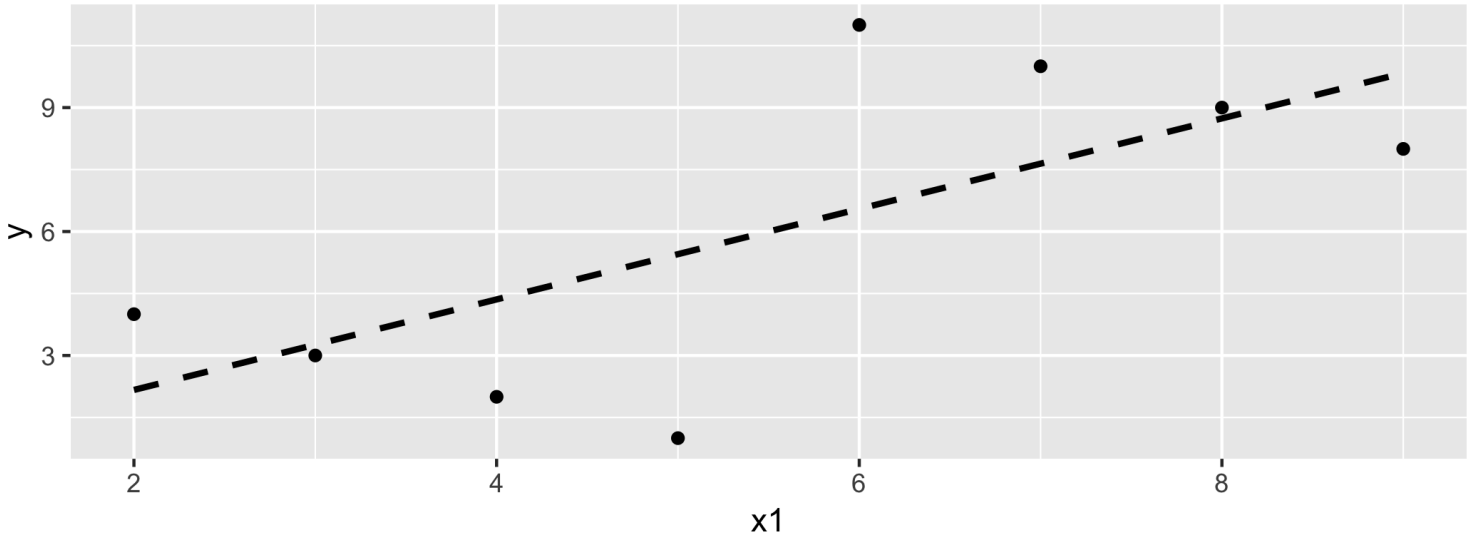
Relationships between variables

- **Bivariate relationship**: Fitness → Heart health
- **Multivariate relationship**: Calories + Age + Fitness → Heart health

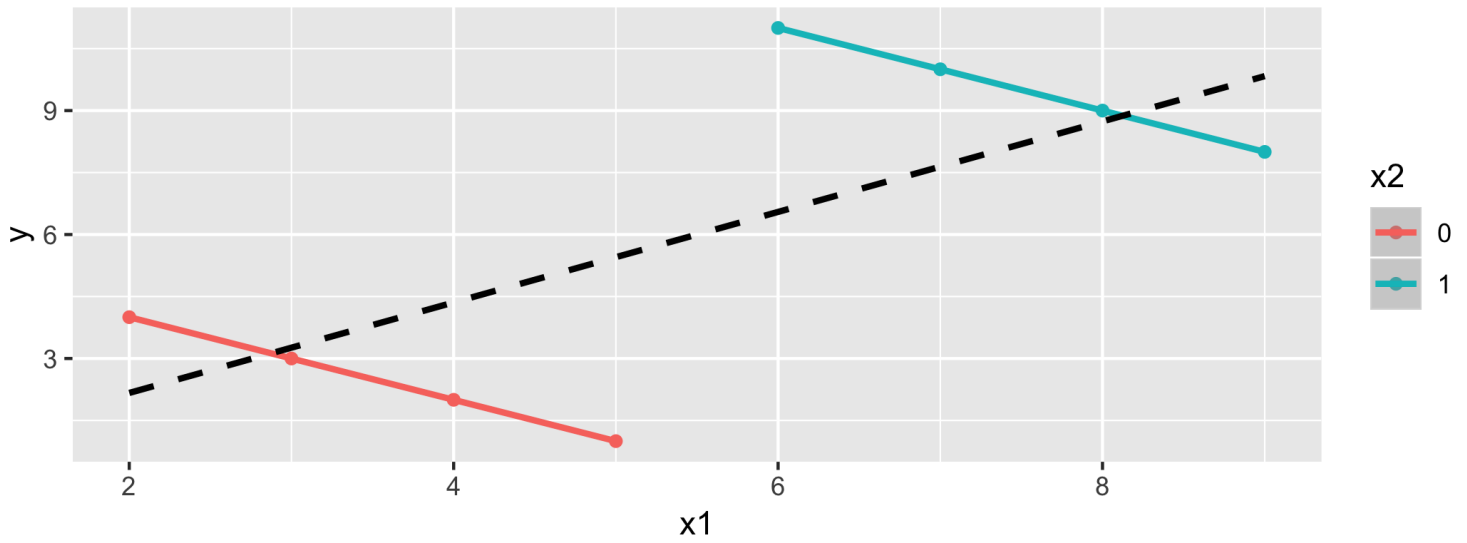
Simpson's paradox

- Not considering an important variable when studying a relationship can result in **Simpson's paradox**, a phenomenon in which the omission of one explanatory variable can affect the measure of association between another explanatory variable and a response variable.
- In other words, the inclusion of a third variable in the analysis can change the apparent relationship between the other two variables.

Simpson's paradox



Simpson's paradox



Data

- Is one hospital more effective than the other at treating a certain disease?
- To answer this question, we will analyze the treatment outcomes for a 100 patients at each hospital.

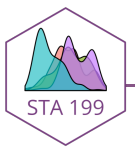
```
treatments %>% slice(1:20)
```

```
## # A tibble: 20 x 3
##    hospital disease     outcome
##    <fct>    <fct>       <fct>
##  1 A        Less Severe success
##  2 A        Less Severe success
##  3 A        Less Severe success
##  4 A        Less Severe success
##  5 A        Less Severe success
##  6 A        Less Severe success
##  7 A        Less Severe success
##  8 A        Less Severe success
##  9 A        Less Severe success
## 10 A        Less Severe success
## 11 A        Less Severe success
## 12 A        Less Severe success
## 13 A        Less Severe success
```

Glimpse at the data

```
glimpse(treatments)
```

```
## Observations: 200
## Variables: 3
## $ hospital <fct> A, A, A, A, A, A, A, A, A, A, A, A, A, A, A, A, A, A, A…
## $ disease  <fct> Less Severe, Less Severe, Less Severe, Less Severe, Les…
## $ outcome  <fct> success, success, success, success, success, success, s…
```



Overall distribution of treatment outcomes

What can you say about the overall distribution of outcomes by hospital? Hint: Calculate the following probabilities: $P(\text{Success}|\text{Hospital A})$ and $P(\text{Success}|\text{Hospital B})$.

```
treatments %>%
  count(hospital, outcome)
```

```
## # A tibble: 4 x 3
##   hospital outcome     n
##   <fct>    <fct>   <int>
## 1 A        failure    50
## 2 A        success    50
## 3 B        failure    32
## 4 B        success    68
```

Overall distribution of treatment outcomes

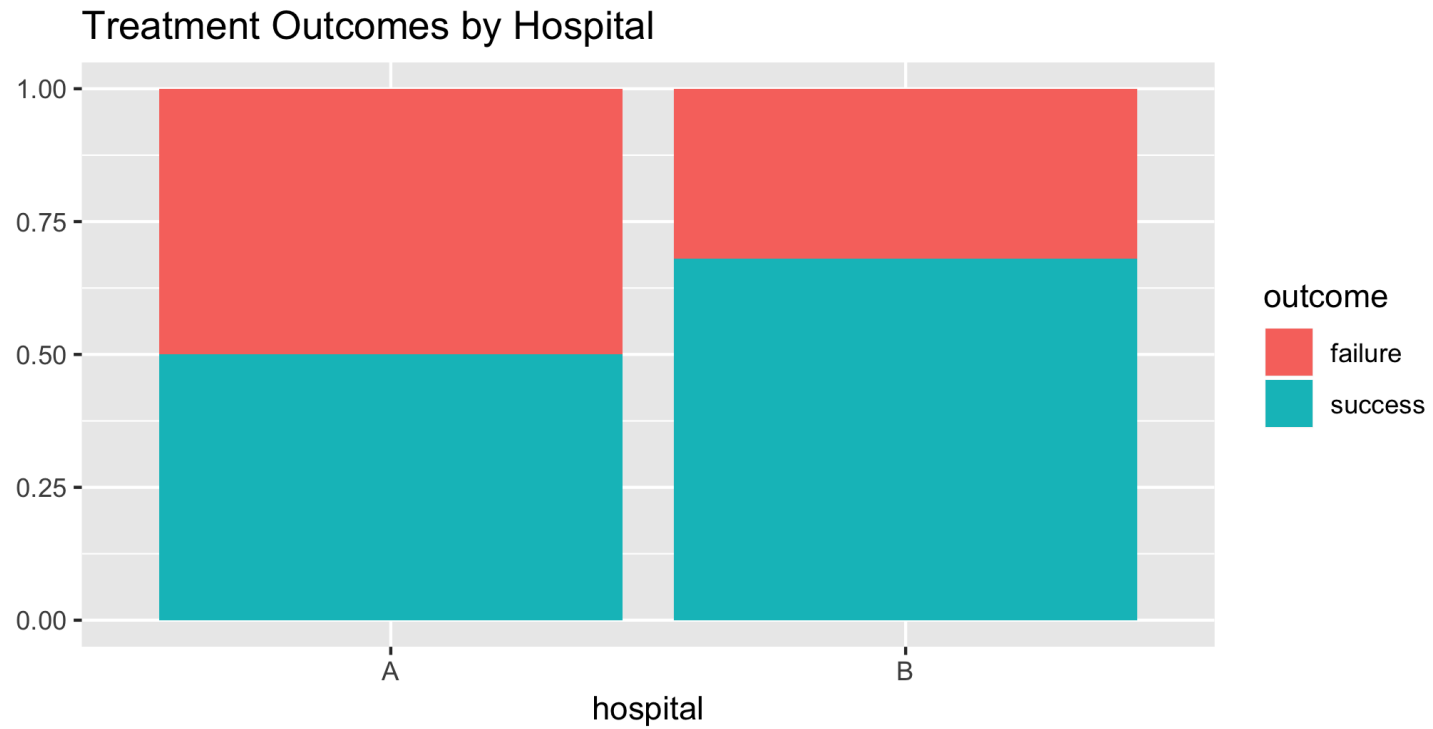
What type of visualization would be appropriate for representing this data?

```
treatments %>%
  count(hospital, outcome) %>%
  group_by(hospital) %>%
  mutate(prop_success = n / sum(n))
```

```
## # A tibble: 4 x 4
## # Groups:   hospital [2]
##   hospital outcome     n prop_success
##   <fct>    <fct>   <int>        <dbl>
## 1 A        failure    50         0.5
## 2 A        success    50         0.5
## 3 B        failure    32         0.32
## 4 B        success    68         0.68
```

Overall distribution of treatment outcomes

```
ggplot(treatments, mapping = aes(x = hospital, fill = outcome)) +
  geom_bar(position = "fill") +
  labs(y = "", title = "Treatment Outcomes by Hospital")
```

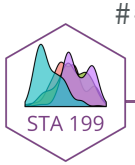


Distribution of treatment outcomes, by disease severity

What can you say about the distribution of treatment outcomes after accounting for the severity of the disease.

```
treatments %>%
  count(disease, hospital, outcome)
```

```
## # A tibble: 8 x 4
##   disease     hospital outcome     n
##   <fct>       <fct>    <fct>   <int>
## 1 Less Severe A        failure     2
## 2 Less Severe A        success    18
## 3 Less Severe B        failure    16
## 4 Less Severe B        success    64
## 5 More Severe A        failure    48
## 6 More Severe A        success    32
## 7 More Severe B        failure    16
## 8 More Severe B        success     4
```



Distribution of treatment outcomes, by disease severity

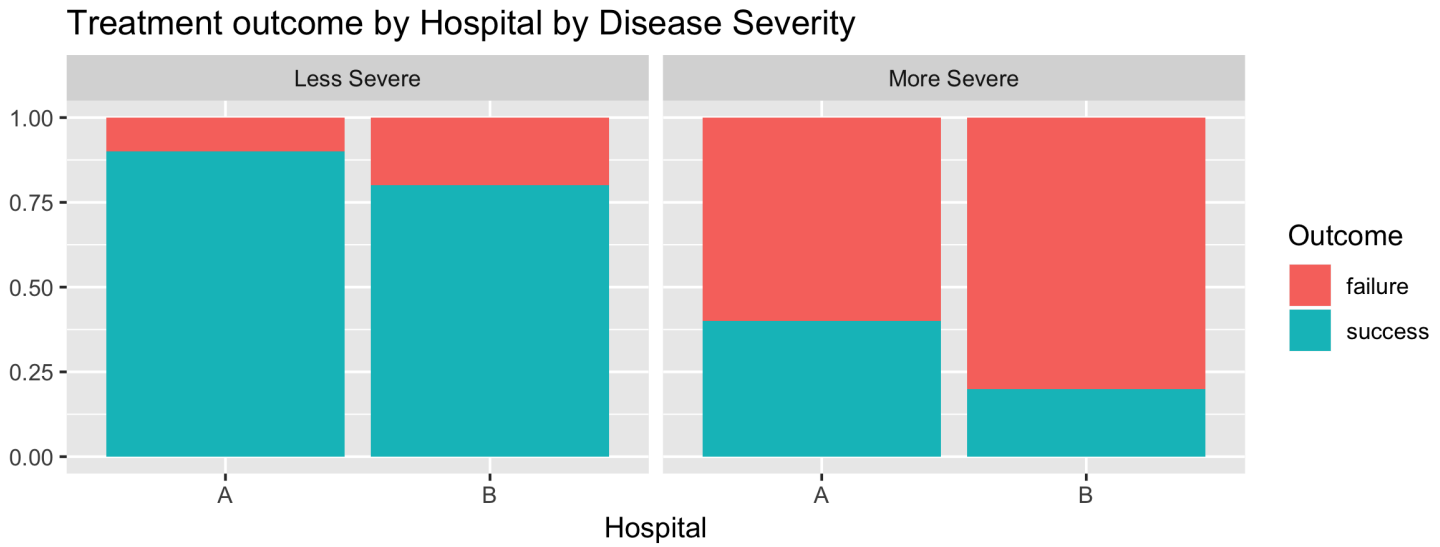
What type of visualization would be appropriate for representing these data?

```
treatments %>%
  count(disease, hospital, outcome) %>%
  group_by(disease, hospital) %>%
  mutate(perc_success = n / sum(n)) %>%
  filter(outcome == "success")
```

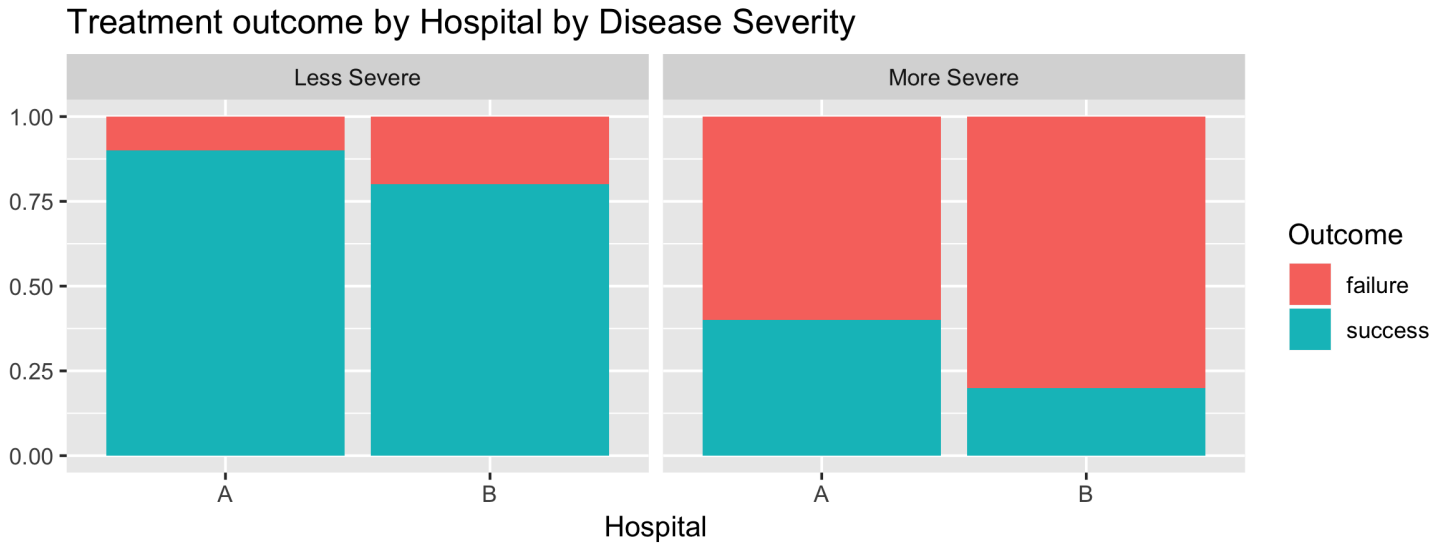
```
## # A tibble: 4 x 5
## # Groups:   disease, hospital [4]
##   disease     hospital outcome     n perc_success
##   <fct>       <fct>    <fct>   <int>        <dbl>
## 1 Less Severe A        success    18          0.9
## 2 Less Severe B        success    64          0.8
## 3 More Severe A        success    32          0.4
## 4 More Severe B        success     4          0.2
```

Distribution of treatment outcomes, by disease severity

```
ggplot(treatments, mapping = aes(x = hospital, fill = outcome)) +
  geom_bar(position = "fill") +
  facet_grid(. ~ disease) +
  labs(x = "Hospital", y = "", fill = "Outcome",
       title = "Treatment outcome by Hospital by Disease Severity")
```



Distribution of treatment outcomes, by disease severity



Why do you think Simpson's paradox occurred? In other words, why is the overall success rate much lower for Hospital A, even though the success rate for Hospital A is higher for each disease severity level?

Writing Assignments

Writing Assignments

- We will do 3 writing assignments over the course of the semester.
- The purpose of the writing assignments is to help you engage with complex statistical concept in a new way
- Each assignment will consist of 3 parts:
  - Initial draft (~ 15 min.)
  - Peer review (~ 15 min.)
  - Revision (~ 15 min.)
- You will be graded on
  - Timely submission for each component
  - Reasonable attempt at each component
  - Correctness of revised response

Writing

- If you haven't already registered for Eli, follow the instructions to register.
- Go to https://app.elireview.com/unit and log in.
- Click Prompt 01: Warm-up and submit your response.
- You will receive an email on Sunday for the next step - peer review.
- See the course schedule for the writing assignment schedule.