Lec 7: Relationships between two categorical variables (Two-way tables)

Corinne Riddell

September 11, 2022

Learning objectives for today

- How to visualize and quantify relationships between two categorical variables
- Two-way tables: marginal vs. conditional distributions
- Bar graphs: side by side vs. stacked
- Simpson's paradox

Readings

- Chapter 5 of Baldi & Moore
- Relationships in categorical data

Two-way tables

- Two-way stands for 2X2, as in a table with two columns and two rows
- Used to examine the relationship between 2 categorical variables, originally those with two levels
- Foundational to epidemiology, because of the types of variables we are often interested in

Classic 2X2 table format

Exposure group	Disease	No disease	Row total
Exposed	A	B	A+B
Not Exposed	C	D	C+D
Column total	A+C	B+D	A+B+C+D

Example: Lung cancer and smoking

Group	Lung Cancer	No Lung Cancer	Row total
Smoker	12	238	250
Non-smoker	7	743	750
Column total	19	981	1000

Marginal distribution

- The **marginal distribution** of a variable is the one that is **in the margin** of the table (i.e., the Row total or the Column total are the two margins of a two-way table).
- The marginal distribution is the distribution for a single categorical variable
- We learned in Ch.1 how to plot marginal distributions of categorical variables using `geom_bar()`

Marginal distribution

Group	Lung Cancer	No Lung Cancer	Row total
Smoker	12	238	250
Non-smoker	7	743	750
Column total	19	981	1000

- Overall, what % of the population has lung cancer?
  - Answer:
- Overall, what % of the population are smokers?
  - Answer:

Marginal distribution

Group	Lung Cancer	No Lung Cancer	Row total
Smoker	12	238	250
Non-smoker	7	743	750
Column total	19	981	1000

- Overall, what % of the population has lung cancer?
  - Answer: 19/1000 = 1.9%
- Overall, what % of the population are smokers?
  - Answer: 250/1000 25% smoking
- The **marginal** distribution of lung cancer is 1.9% lung cancer, 98.1% no lung cancer.

Conditional distribution

Group	Lung Cancer	No Lung Cancer	Row total
Smoker	12	238	250
Non-smoker	7	743	750
Column total	19	981	1000

- The **conditional distribution** is the distribution of one variable **within** or **conditional on** the level of a second variable
- What is the conditional distribution of lung cancer **given** smoking?
  - Answer:
- What is the conditional distribution of lung cancer **given** non-smoking?
  - Answer:

Conditional distribution

Group	Lung Cancer	No Lung Cancer	Row total
Smoker	12	238	250
Non-smoker	7	743	750
Column total	19	981	1000

- The **conditional distribution** is the distribution of one variable **within** or **conditional on** the level of a second variable
- What is the conditional distribution of lung cancer **given** smoking?
  - Answer: $12/250 = 4.8\%$ lung cancer and $238/250 = 95.2\%$ no lung cancer
- What is the conditional distribution of lung cancer **given** non-smoking?
  - Answer: $7/750 = 0.9\%$ lung cancer and $743/750 = 99.1\%$ no lung cancer

Visualization of conditional distributions

Marginal and conditional distributions in R

- We learned in Ch.1 how to plot marginal distributions of categorical variables using `geom_bar()`
- Can we generalize the use of `geom_bar()` to plot multiple conditional distributions? I.e., can we show the conditional distribution of lung cancer for smokers and non-smokers on the same plot?

First, we encode the data to read into R:

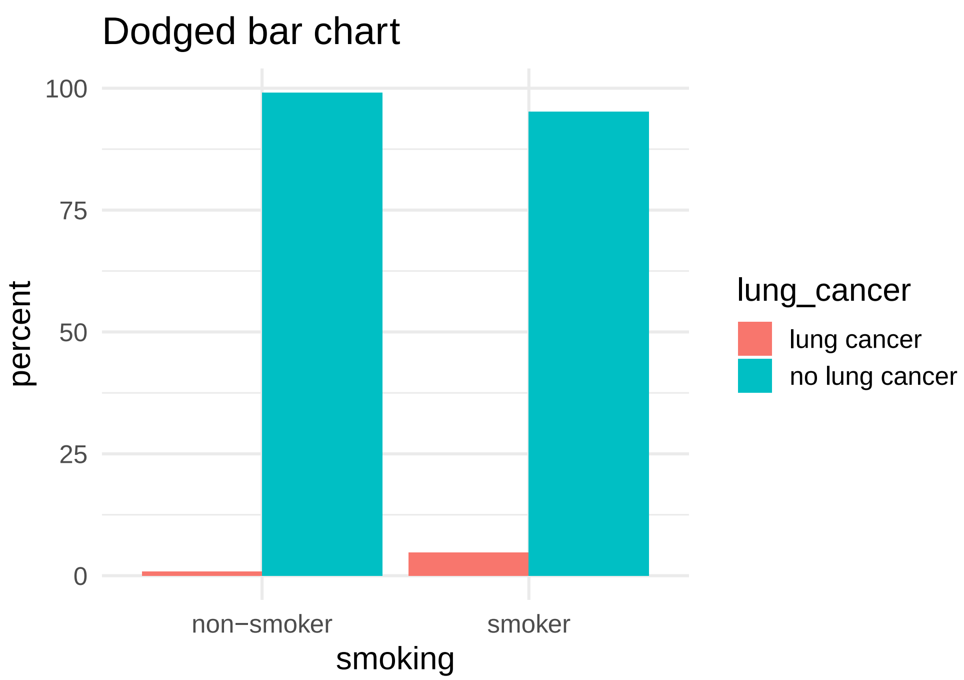
```
# students, you don't need to know how to do this
two_way <- tribble(~ smoking,      ~ lung_cancer,    ~ percent, ~number,
                     "smoker",       "lung cancer",    4.8,       12,
                     "smoker",       "no lung cancer", 95.2,      238,
                     "non-smoker",   "lung cancer",    0.9,       7,
                     "non-smoker",   "no lung cancer", 99.1,      743
                   )
```

Visualization of conditional distributions

> If there is an explanatory-response relationship, compare the conditional distribution of the response variable for the separate values of the explanatory variable.

Dodged bar chart for the visualization of conditional distributions

```
ggplot(two_way, aes(x = smoking, y = percent)) +
  geom_bar(aes(fill = lung_cancer), stat = "identity", position = "dodge")  +
  labs(title = "Dodged bar chart") + theme_minimal(base_size = 15)
```

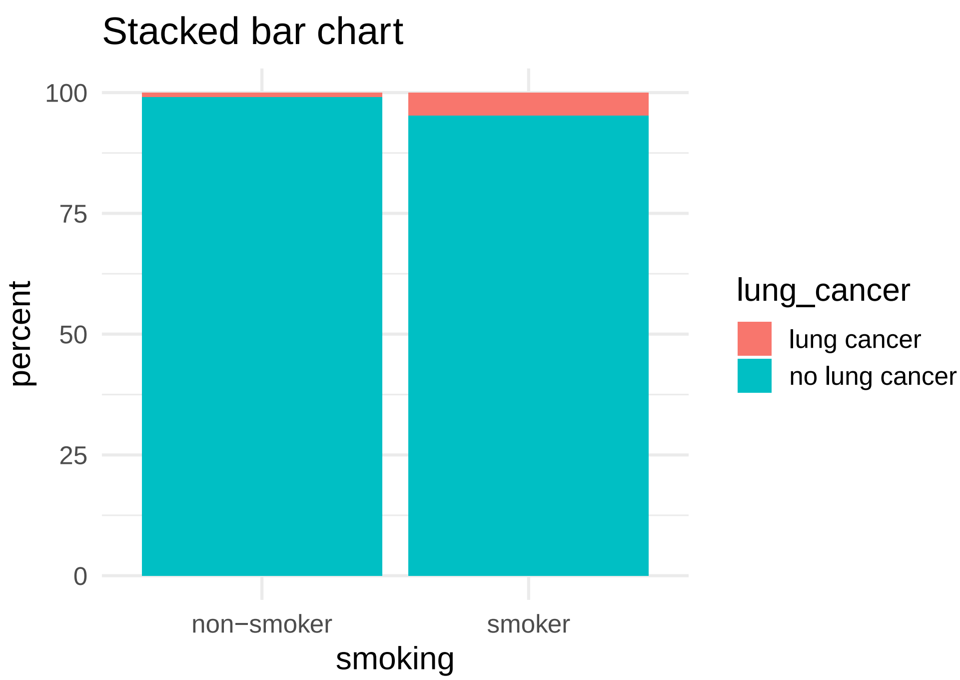


Syntax: Dodged bar chart for the visualization of conditional distributions

```
#students, remove eval=F if you copy this code chunk (or else the code won't compile)
ggplot(data, aes(x = exposure_variable, y = percent)) +
  geom_bar(aes(fill = outcome_variable), stat = "identity", position = "dodge")  +
  labs(title = "Dodged bar chart") +
  theme_minimal(base_size = 15)
```

Stacked bar chart for the visualization of conditional distributions

```
ggplot(two_way, aes(x = smoking, y = percent)) +
  geom_bar(aes(fill = lung_cancer), stat = "identity", position = "stack") +
  labs(title = "Stacked bar chart") + theme_minimal(base_size = 15)
```



Syntax: Stacked bar chart for the visualization of conditional distributions

```
#students, remove eval=F if you copy this code chunk (or else the code won't compile)
ggplot(data, aes(x = exposure_variable, y = percent)) +
  geom_bar(aes(fill = outcome_variable), stat = "identity", position = "stack")  +
  labs(title = "Stacked bar chart") +
  theme_minimal(base_size = 15)
```

Visualization of conditional distributions: three levels of response variable

- Stacked and dodged plots are less informative when there are only two levels of both variables.
- This is because once you know the percent of lung cancer among smokers, you also know the percent of non-lung cancer among smokers. This makes some of the information redundant.
- The plots are more informative if there are 3 or more levels for at least one of the variables

Visualization of conditional distributions: three levels of response variable

- Example 2: Shoe support by gender (Data from Baldi & Moore page 124 of Ed.4):

Group	Men	Women	**Row total**
Good support	94	137	231
Average support	1348	581	1929
Poor support	30	1182	1212
Column total	1472	1900	3372

Check your understanding!

Visualization of conditional distributions: three levels of response variable

- Example 2: Shoe support by gender (Data from Baldi & Moore page 124 of Ed.4):

Group	Men	Women	**Row total**
Good support	94	137	231
Average support	1348	581	1929
Poor support	30	1182	1212
Column total	1472	1900	3372

- The question: How does the distribution of support of shoes worn vary between men and women?

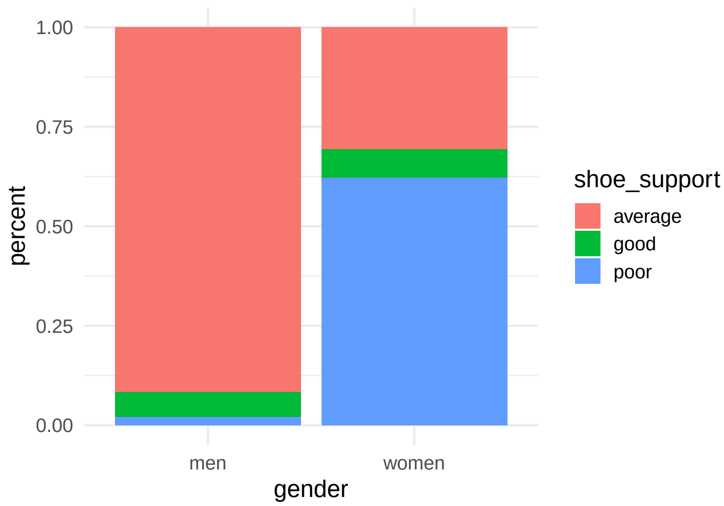
Visualization of conditional distributions: three levels of response variable

```
# students, you don't need to know how to do this
shoe_data <- tribble(~ shoe_support, ~ gender, ~ percent,
                       "good",         "men",    94/1472,
                       "average",      "men",    1348/1472,
                       "poor",         "men",    30/1472,
                       "good",         "women",  137/1900,
                       "average",      "women",  581/1900,
                       "poor",         "women",  1182/1900)
shoe_data
```

```
## # A tibble: 6 x 3
##   shoe_support gender percent
##   <chr>        <chr>    <dbl>
## 1 good         men     0.0639
## 2 average      men     0.916
## 3 poor         men     0.0204
## 4 good         women   0.0721
## 5 average      women   0.306
## 6 poor         women   0.622
```

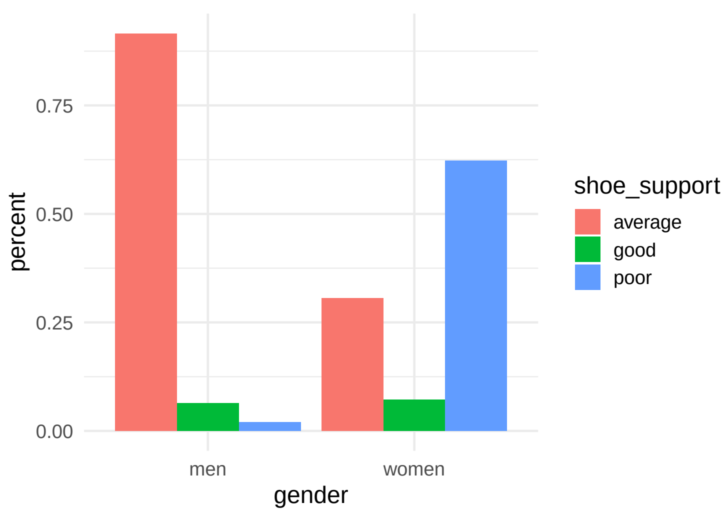
Stacked visualization when there are three levels of response

```
ggplot(shoe_data, aes(x = gender, y = percent)) +
  geom_bar(stat = "identity", aes(fill = shoe_support), position = "stack") +
  theme_minimal(base_size = 15)
```



Dodged visualization when there are three levels of response

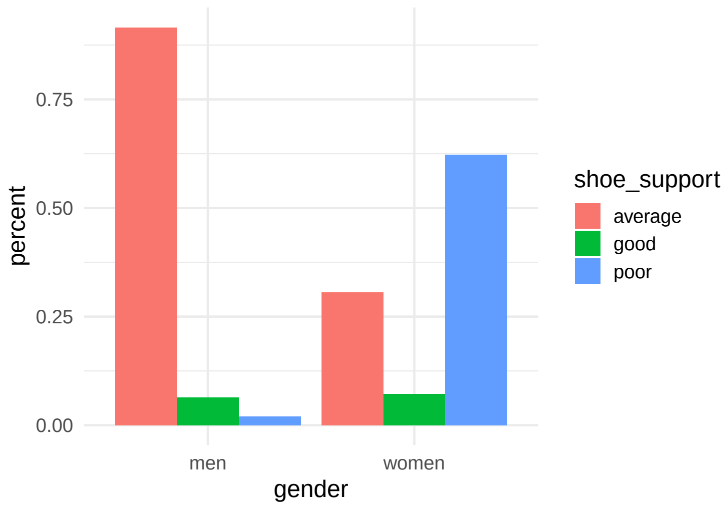
```
ggplot(shoe_data, aes(x = gender, y = percent)) +
  geom_bar(stat = "identity", aes(fill = shoe_support), position = "dodge") +
  theme_minimal(base_size = 15)
```



Dodged visualization when there are three levels of response

Question: what is misleading about the fill legend?

```
ggplot(shoe_data, aes(x = gender, y = percent)) +
  geom_bar(stat = "identity", aes(fill = shoe_support), position = "dodge") +
  theme_minimal(base_size = 15)
```



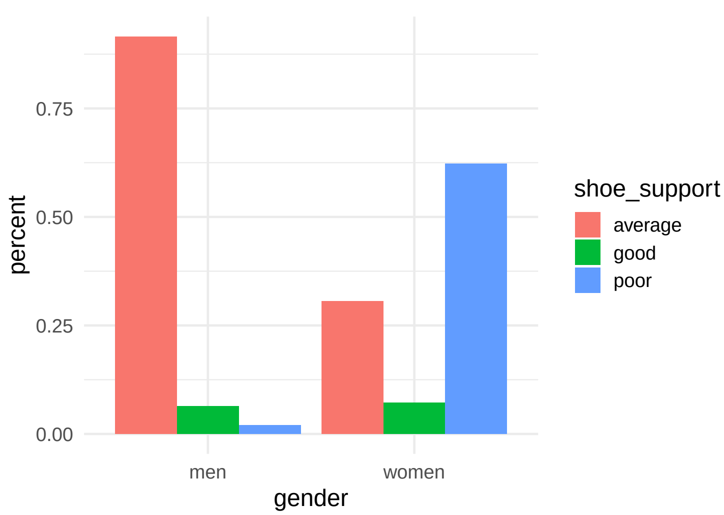
Dodged visualization when there are three levels of response

Question: what is misleading about the fill legend?

Answer: It is in alphabetic order, which is different from the natural order of this variable.

Question 2: How can we change the order in the legend?

```
ggplot(shoe_data, aes(x = gender, y = percent)) +
  geom_bar(stat = "identity", aes(fill = shoe_support), position = "dodge") +
  theme_minimal(base_size = 15)
```

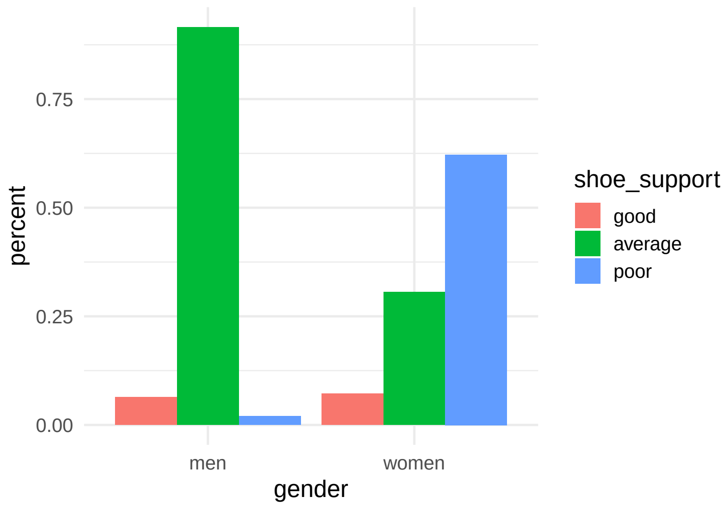


Dodged visualization when there are three levels of response

Question 2: How can we change the order in the legend?

Answer 2: Recall from the problem sets and lab how to reorder factor variables that affect the look of the plot:

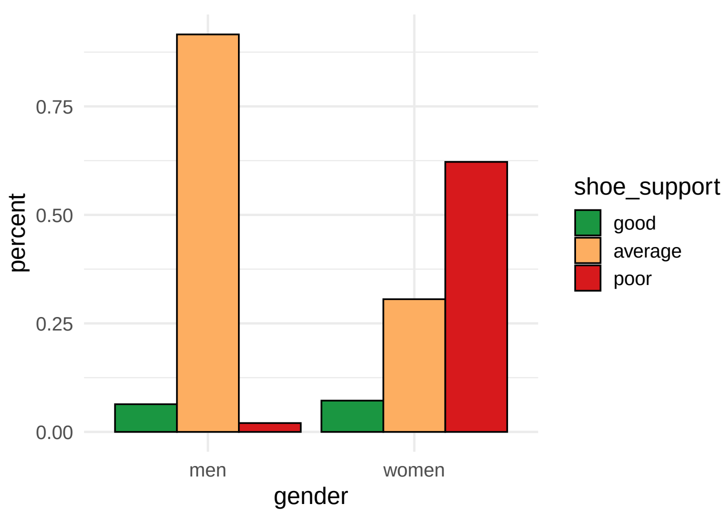
```r
shoe_data <- shoe_data %>%
  mutate(shoe_support = fct_relevel(shoe_support, "good", "average", "poor"))

ggplot(shoe_data, aes(x = gender, y = percent)) +
  geom_bar(stat = "identity", aes(fill = shoe_support), position = "dodge") +
  theme_minimal(base_size = 15)
```



Dodged visualization when there are three levels of response

You might also want to specify the colors used to communicate that poor shoe support is painful!

```r
ggplot(shoe_data, aes(x = gender, y = percent)) +
  geom_bar(stat = "identity", aes(fill = shoe_support), position = "dodge", col = "black") +
  theme_minimal(base_size = 15) +
  scale_fill_manual(values = c("#1a9641", "#fdae61", "#d7191c"))
```



Visualization of conditional distributions: three levels of response variable

In general, dodged plots are preferred over stacked plots. Why do you think that is?

Simpson's Paradox

Simpson's Paradox: Example from Baldi and Moore

- Let's load these data that examines mortality rates by community and age group across two communities

```
#this is the data from page 131 of edition 4 of baldi and moore
simp_data <- tribble(~ age_grp, ~ community, ~ deaths, ~ pop,
                       "0-34",    "A",         20,       1000,
                       "35-64",   "A",         120,      3000,
                       "65+",     "A",         360,      6000,
                       "all",     "A",         500,      10000,
                       "0-34",    "B",         180,      6000,
                       "35-64",   "B",         150,      3000,
                       "65+",     "B",         70,       1000,
                       "all",     "B",         400,      10000)

simp_data <- simp_data %>%
  mutate(death_per_1000 = (deaths/pop) * 1000)

simp_data_no_all <- simp_data %>% filter(age_grp != "all")
```

Simpson's Paradox: Example from Baldi and Moore

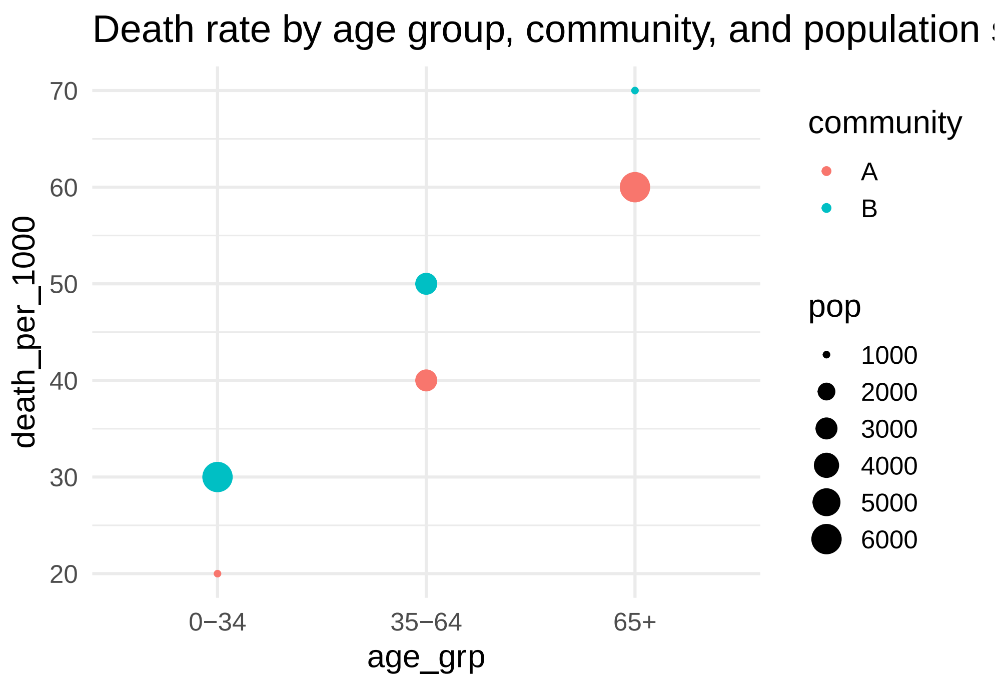
```
simp_data
```

```
## # A tibble: 8 x 5
##   age_grp community deaths   pop death_per_1000
##   <chr>   <chr>      <dbl> <dbl>          <dbl>
## 1 0-34    A             20  1000             20
## 2 35-64   A            120  3000             40
## 3 65+     A            360  6000             60
## 4 all     A            500 10000             50
## 5 0-34    B            180  6000             30
## 6 35-64   B            150  3000             50
## 7 65+     B             70  1000             70
## 8 all     B            400 10000             40
```

Simpson's Paradox Example: Plot only the conditional data

- Plot the mortality rates according to age group and community and link the point size to population size

```r
ggplot(simp_data_no_all, aes(x = age_grp, y = death_per_1000)) +
  geom_point(aes(col = community, size = pop)) +
  labs(title = "Death rate by age group, community, and population size") +
  theme_minimal(base_size = 15)
```



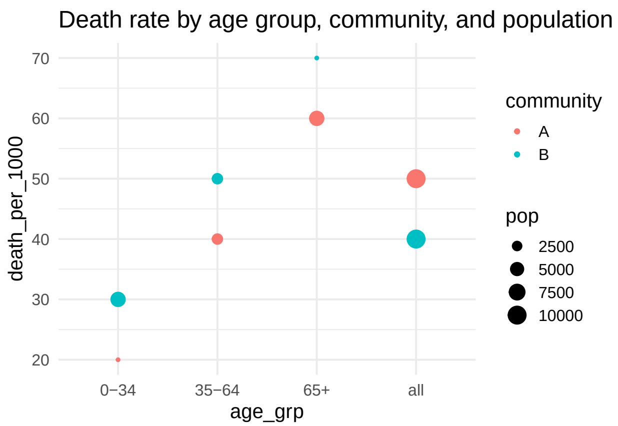
Observations from this visualization:

1.
2.
3.

If someone ask you which community has higher mortality, what would you say?

Simpson's Paradox Example: Add the marginal data

- Add in the **marginal** data (not conditional on age)
- Notice that the mortality rates for the communities overall show community A having a higher rate than community B. Why?

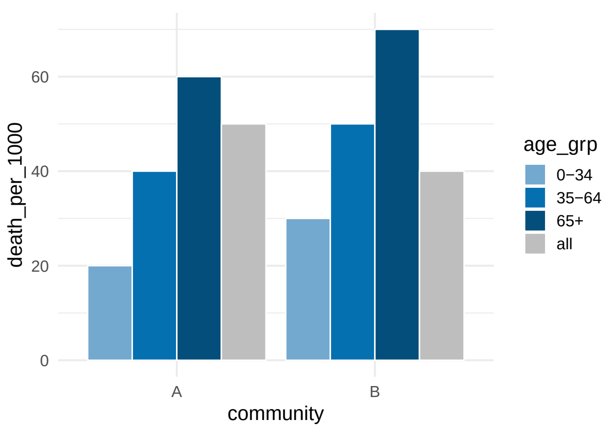


Simpson's Paradox

"An association or comparison that holds for all of several groups can **reverse direction** when the data are combined to form a single group. This reversal is called **Simpson's Paradox**"

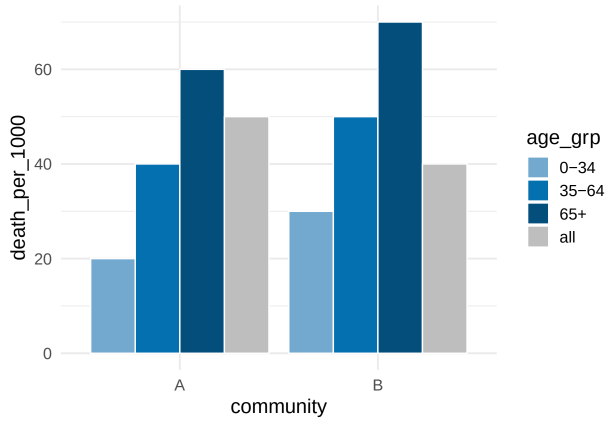
Simpson's Paradox

- Here are the same data shown using a bar chart
- Notice that the mortality rate for each of the blue-shaded bars in community B is higher than the correponding bar for community A, but the overall bar (shaded in gray) shows a reversal.



Simpson's Paradox

- With a bar chart we can't use `aes(size = pop)`, so it is harder to see why the paradox is occuring.
- It is because we are taking a weighted average of each age-specific bar with weights proportional to the number of people of each age group in each community



Simpson's Paradox in Berkeley Admissions

- There is a famous example of Simpson's paradox related to admissions to Berkeley by gender
- Watch it here!

Recap: What new code and statistical concepts did we learn?

1. `geom_bar(aes(col = var), stat = "identity", position = "dodge")`
2. `geom_bar(aes(col = var), stat = "identity", position = "stack")`
3. Marginal distribution vs. conditional distribution
4. Simpson's Paradox