# L07: Relationships between two categorical variables

February 5, 2020

# Learning objectives for today



Today we will focus on how to visualize and quantify relationships between two categorical variables

- Two way tables
- marginal vs conditional distributions
- Bar graphs
- side by side
- stacked
- Simpson's paradox

# Reminder

Categorical variables are just that, categories.

These can be nominal (no underlying order)

or

ordinal (ordered)

# Two-way tables

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

Classic 2X2 looks like this:

| Exposure group | Disease | No disease | Row total |
|---|---|---|---|
| Exposed | A | B | A+B |
| Not Exposed | C | D | C+D |
| Column total | A+C | B+D | Total # observations |

# 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 distributions



- 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 distributions



| 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?
- Overall, what % of the population are smokers?

# Marginal distributions

| 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 distributions

| 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 distribution of lung cancer conditional on the individuals being smokers?
- What is the conditional distribution of lung cancer given individuals are non-smoking?

# Conditional distributions

| 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 of lung cancer given smoking is:
  - $12/250 = 4.8\%$ smokers and $238/250 = 95.2\%$
- The conditional distribution of lung cancer given non-smoking is:
  - $7/750 = 0.9\%$ smokers and $743/750 = 99.1\%$ non-smokers

# Visualizing categorical variables in R

# 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 our use of `geom_bar()` to allow us 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?

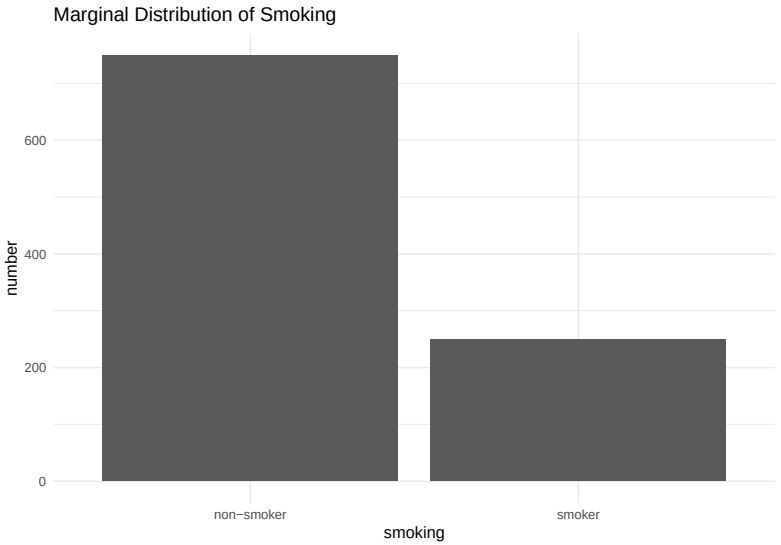
# Visualization in R



First, we encode the data to read into R:

```
library(dplyr)
# students, you don't need to know how to do this
two_way_data <- 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)
```

# Bar chart for the visualization of marginal distributions



# 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."

- This allows you to visualize the distribution of the response variable for varying levels of the exposure variable.

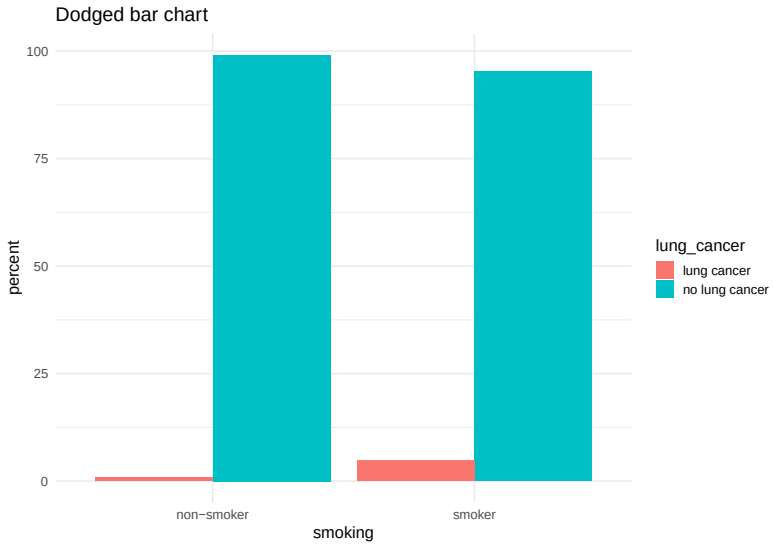
# Dodged bar chart for the visualization of conditional distributions



Syntax:

```
ggplot(two_way_data, 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)
```

# Dodged bar chart for the visualization of conditional distributions

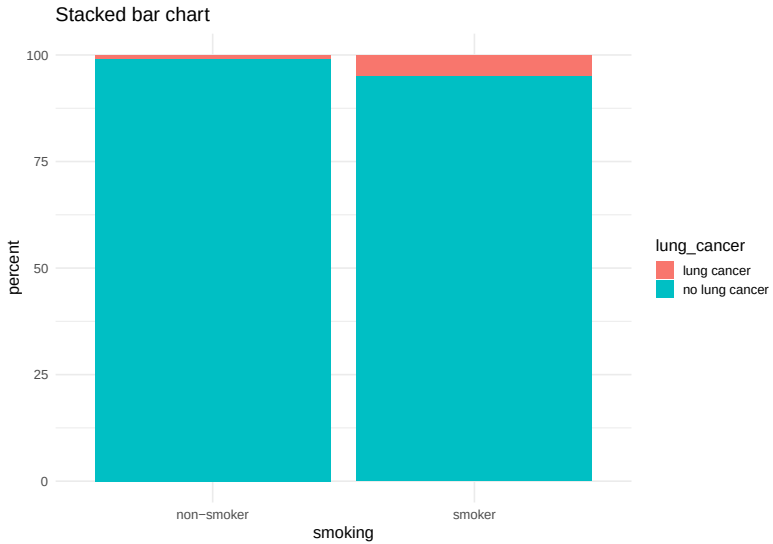


# Stacked bar chart for the visualization of conditional distributions



```
ggplot(two_way_data, 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)
```

# Stacked bar chart for the visualization of conditional distributions



# Visualization of conditional distributions: three levels of response variable

- Plots like the one above make less sense when there are only two levels of both of the 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.
- Here is another example with 3 levels: Shoe support by gender (from ch. 5):

| Group | Men | Women |
|---|---|---|
| Good support | 94 | 137 |
| Average support | 1348 | 581 |
| Poor support | 30 | 1182 |
| Column total | 1472 | 1900 |

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

# Participation

# Visualization of conditional distributions: three levels of response variable

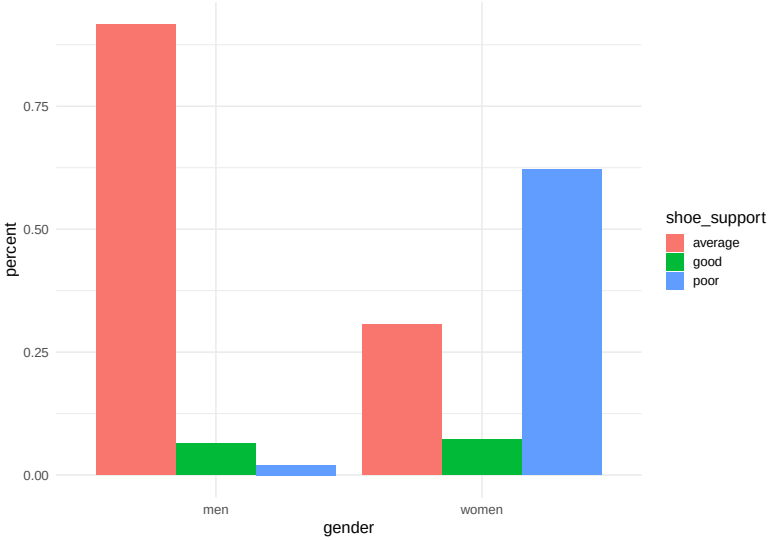
Example using shoe support data from Baldi & Moore page 124 of Ed.4

```
## # 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
```

# Visualization of conditional distributions: three levels of response variable

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

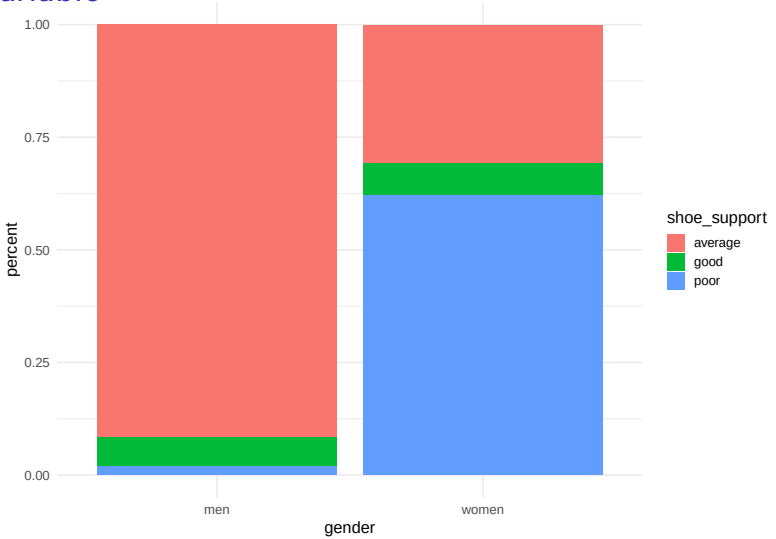
# Visualization of conditional distributions: three levels of response variable



# Visualization of conditional distributions: three levels of response variable

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

# Visualization of conditional distributions: three levels of response variable



# Visualization of conditional distributions: three levels of response variable

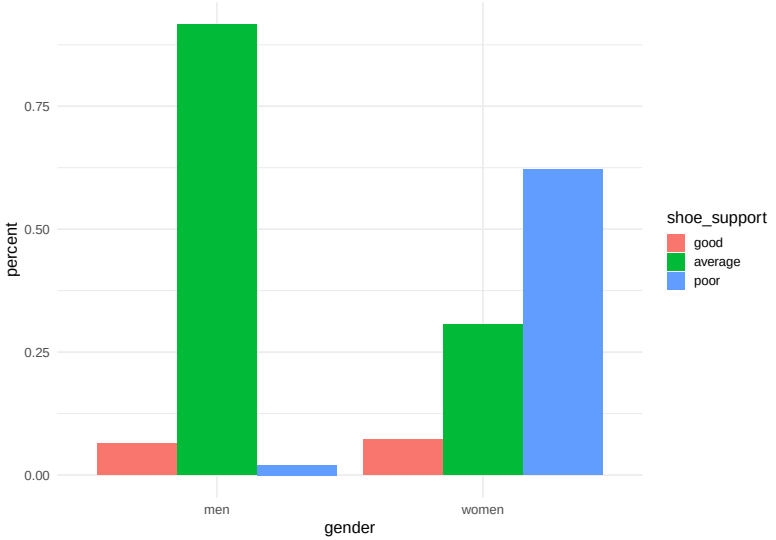


Recall from last class we learned how to reorder factor variables that affect the look of the plot:

```
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)
```

# Visualization of conditional distributions: three levels of response variable



# Visualization of conditional distributions: three levels of response variable

Why might we prefer dodged plots to stacked plots?

# Simpson's Paradox

# Simpson's Paradox: Example from Baldi and Moore

- Here is the data presented in your book to illustrate Simpson's paradox.
- It looks at mortality rates by community and age group for two communities

```
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")
```

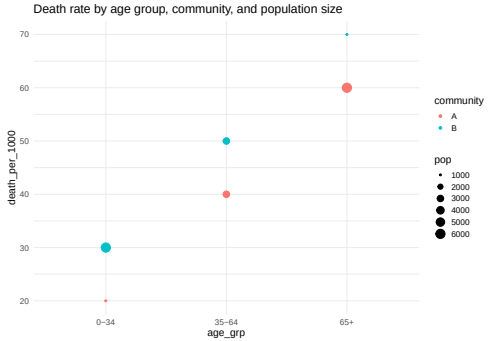
# Simpson's Paradox Example: Only Conditional data



Plot the mortality rates according to age group and community, linking size of dot to population size

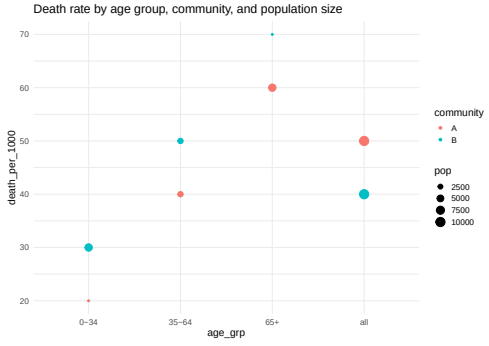
```
ggplot(simp_data_no_all, aes(x = age_grp, y = death_per_1000)) +
geom_point(aes(col = community, size = pop)) +
labs(title = "Death rate by age group, community, and population size") +
theme_minimal(base_size = 15)
```

# Simpson's Paradox Example: Only Conditional data




- What do we notice about mortality by age groups?
- Which community is larger?
- If someone ask you which community has higher mortality, what would you say?

# Simpson's Paradox Example: with marginal data



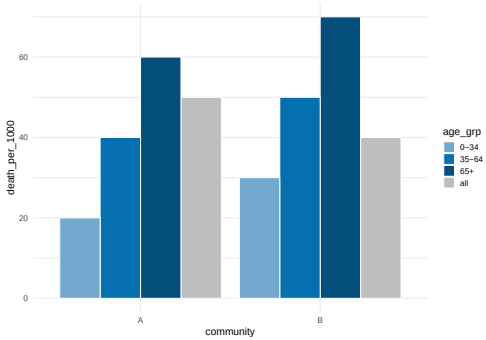

- Notice that the mortality rates for the communities overall show community A having a higher rate than community B. Why is that?

# 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




Which visualization gives you more information?

# Simpson's Paradox Berkeley example

A famous example of Simpson's paradox related to admissions to Berkeley by gender:

Watch: https://www.youtube.com/watch?v=E_ME4P9fQbo

# Recap: Code and concepts

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

# Comic Relief