Lab 07 - Smokers in Whickham

Simpson's paradox

Tina Huynh

Contents

A study of conducted in Whickham, England recorded participants' age, smoking status at baseline, and then 20 years later recorded their health outcome. In this lab we analyse the relationships between these variables, first two at a time, and then controlling for the third.

1 Learning goals

- Visualising relationships between variables

- Discovering Simpson's paradox via visualisations

2 Getting started

Go to the course GitHub organization and locate your homework repo, clone it in RStudio and open the R Markdown document. Knit the document to make sure it compiles without errors.

2.1 Warm up

Before we introduce the data, let's warm up with some simple exercises. Update the YAML of your R Markdown file with your information, knit, commit, and push your changes. Make sure to commit with a meaningful commit message. Then, go to your repo on GitHub and confirm that your changes are visible in your Rmd **and** md files. If anything is missing, commit and push again.

2.2 Packages

We'll use the **tidyverse** package for much of the data wrangling and visualisation and the data lives in the **mosaicData** package. These packages are already installed for you. You can load them by running the following in your Console:

```
library(tidyverse)
library(mosaicData)
```

2.3 Data

The dataset we'll use is called Whickham from the **mosaicData** package. You can find out more about the dataset by inspecting their documentation, which you can access by running `?Whickham` in the Console or using the Help menu in RStudio to search for `Whickham`.

3 Exercises

1. What type of study do you think these data come from: observational or experiment? Why? I think these data come from an observational study because the researchers did not manipulate any variables, they simply observed and recorded the smoking status and health outcomes of the participants over a 20-year period.
2. How many observations are in this dataset? What does each observation represent? There are 5000 observations in this dataset. Each observation represents an individual participant in the study, with their corresponding age, smoking status at baseline, and health outcome after 20 years.
3. How many variables are in this dataset? What type of variable is each? Display each variable using an appropriate visualization. There are 3 variables in this dataset:

- Age: Continuous variable (numeric)
- Smoker: Categorical variable (factor with levels "yes" and "no")
- Outcome: Categorical variable (factor with levels "alive" and "dead")

```
# Visualizing Age
ggplot(Whickham, aes(x = age)) +
  geom_histogram(binwidth = 5, fill = "blue", color = "black") +
  labs(title = "Distribution of Age", x = "Age", y = "Count")
```

```
# Visualizing Smoker
ggplot(Whickham, aes(x = smoker)) +
  geom_bar(fill = "green", color = "black") +
  labs(title = "Smoking Status", x = "Smoker", y = "Count")
```

```r
# Visualizing Outcome
ggplot(Whickham, aes(x = outcome)) +
  geom_bar(fill = "red", color = "black") +
  labs(title = "Health Outcome", x = "Outcome", y = "Count")
```

4. What would you expect the relationship between smoking status and health outcome to be? I would expect that smokers are more likely to have a negative health outcome (i.e., be dead) compared to non-smokers. This expectation is based on the well-established link between smoking and various health issues, including respiratory diseases, cardiovascular diseases, and cancer, which can lead to higher mortality rates among smokers.

Knit, *commit, and push your changes to GitHub with an appropriate commit message. Make sure to commit and push all changed files so that your Git pane is cleared up afterwards.*

5. Create a visualization depicting the relationship between smoking status and health outcome. Briefly describe the relationship, and evaluate whether this meets your expectations. Additionally, calculate the relevant conditional probabilities to help your narrative. Here is some code to get you started:

```r
Whickham %>%
  count(smoker, outcome) %>%
  ggplot(aes(x = smoker, y = n, fill = outcome)) +
  geom_bar(stat = "identity", position = "fill") +
  labs(y = "Proportion", title = "Health Outcome by Smoking Status") +
  scale_y_continuous(labels = scales::percent)
```

```r
# Calculate conditional probabilities
Whickham %>%
  count(smoker, outcome) %>%
  group_by(smoker) %>%
  mutate(prop = n / sum(n)) %>%
  select(smoker, outcome, prop) %>%
    pivot_wider(names_from = outcome, values_from = prop)
```

Surprisingly, the visualization shows that smokers have a lower proportion of deaths (around 24%) compared to non-smokers (around 31%). This does NOT meet my expectations, as I would have expected smokers to have higher mortality rates. This counterintuitive result suggests there may be a confounding variable at play.

The conditional probabilities show that among smokers, 76.3% are alive and 23.7% are dead, while among non-smokers, 68.6% are alive and 31.4% are dead. This means smokers have a lower death rate (23.7%) compared to non-smokers (31.4%). This counterintuitive result suggests there may be a confounding variable at play.

6. Create a new variable called `age_cat` using the following scheme:

- `age <= 44 ~ "18-44"`
- `age > 44 & age <= 64 ~ "45-64"`
- `age > 64 ~ "65+"`

```r
Whickham <- Whickham %>%
  mutate(age_cat = case_when(
    age <= 44 ~ "18-44",
    age > 44 & age <= 64 ~ "45-64",
    age > 64 ~ "65+"
  )) %>%
  mutate(age_cat = factor(age_cat, levels = c("18-44", "45-64", "65+")))

table(Whickham$age_cat)
```

7. Re-create the visualization depicting the relationship between smoking status and health outcome, faceted by `age_cat`. What changed? What might explain this change? Extend the contingency table from earlier by breaking it down by age category and use it to help your narrative.

```
Whickham %>%
  count(smoker, age_cat, outcome) %>%
  ggplot(aes(x = smoker, y = n, fill = outcome)) +
  geom_bar(stat = "identity", position = "fill") +
  labs(y = "Proportion", title = "Health Outcome by Smoking Status and Age Category") +
  scale_y_continuous(labels = scales::percent) +
  facet_wrap(~age_cat)
```

```
# Extended contingency table with conditional probabilities by age group
Whickham %>%
  count(smoker, age_cat, outcome) %>%
  group_by(smoker, age_cat) %>%
  mutate(prop = n / sum(n)) %>%
  filter(outcome == "dead") %>%
  select(smoker, age_cat, prop) %>%
  pivot_wider(names_from = smoker, values_from = prop)
```

What changed: When we control for age, the relationship reverses! Within each age group, smokers now show higher mortality rates than non-smokers, which aligns with our expectations.

What explains this change: This is Simpson's Paradox. The original relationship was confounded by age because: 1. Younger people are more likely to smoke 2. Younger people are also much less likely to die during the 20-year study period 3. When we don't control for age, the protective effect of being young masks the harmful effect of smoking

The age-stratified analysis reveals the true relationship: smoking increases mortality risk within each age group.

Knit, *commit, and push your changes to GitHub with an appropriate commit message. Make sure to commit and push all changed files so that your Git pane is cleared up afterwards and review the md document on GitHub to make sure you're happy with the final state of your work.*

Now go back through your write up to make sure you've answered all questions and all of your R chunks are properly labeled. Once you decide that you are done with the homework, choose the knit drop down and select `Knit to tufte_handout` to generate a pdf. Download and submit that pdf to Canvas.