Lab 6 - Smokers in Whickham

Simpson's paradox

Musab Isah (adopted from datasciencebox.org)

11 May 2022

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.

Learning goals

- Visualising relationships between variables

- Discovering Simpson's paradox via visualisations

Simpson's paradox, which also goes by several other names, is a phenomenon in probability and statistics, in which a trend appears in several different groups of data but disappears or reverses when these groups are combined.

Getting started

As usual, clone the lab files available here https://github.com/musabu/uhb-212-lab6.git to you Github account. Now, import Lab6 from you github account (*NOT INSTRUCTOR'S GITHUB ACCOUNT*) to rstudio.cloud. Open the R Markdown document, lab-07-simpsons.Rmd, and Knit it to make sure the document compiles without errors.

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.

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)
```

```
## Warning: package 'tidyverse' was built under R version 4.1.3
```

```
## Warning: package 'tidyr' was built under R version 4.1.3
```

```
## Warning: package 'readr' was built under R version 4.1.3
```

```
## Warning: package 'dplyr' was built under R version 4.1.3
```

```
library(mosaicData)
```

```
## Warning: package 'mosaicData' was built under R version 4.1.3
```

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`.

Exercises

1. What type of study do you think these data come from: observational or experiment? Why?

2. How many observations are in this dataset? What does each observation represent?

3. How many variables are in this dataset? What type of variable is each? Display each variable using an appropriate visualization.

4. What would you expect the relationship between smoking status and health outcome to be?

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.

```
Whickham %>%
  count(smoker, outcome)
```

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+"`

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?

```
Whickham %>%
  count(smoker, age_cat, outcome)
```

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.*