DATA2002

Collecting data

Garth Tarr

THE UNIVERSITY OF SYDNEY

Sample surveys

Controlled experiments

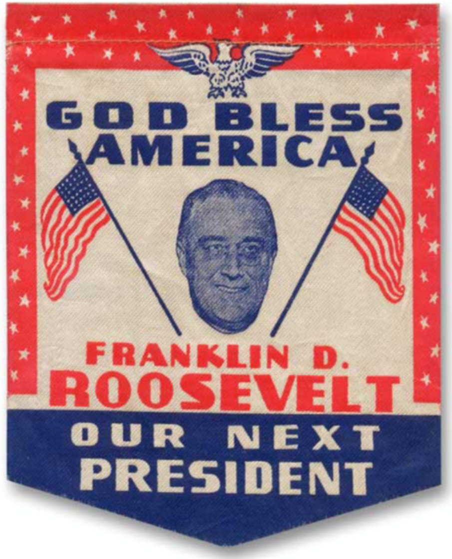
Observational studies

Simpsons Paradox

Sample surveys

The Literary Digest poll (1936)

- 1936 Franklin D. Roosevelt was completing his term of office
- America was struggling with high unemployment (9 million), just after the Great Depression
- 1936 US Presidential election
- Literary Digest polled **10 million** people (mail survey)
- 24% response rate (**2.4 million** people reply)
- They had predicted the winner since 1916
- They predicted victory for **Landon**



Results

- **Roosevelt** won by 62% to 38%.
- **Roosevelt** won 46 of 48 states.

Gallup poll

- George Gallup was setting up his survey organisation.
- He drew 3000 people and predicted the Digest results.
- He also drew 50,000 people and **correctly predicted** Roosevelt victory (the actual prediction was off by a bit); 56% predicted instead of 62%.

- Digest mailed questionnaires to **10 million people** with **2.4 million replies** and still failed to predict the winner. What went wrong?!?

Some data to play with if interested http://www.math.uah.edu/stat/data/LiteraryDigest.csv

Revision

- A **sample** is part of a **population**
- A **parameter** is a numerical fact about a population.
- Usually a parameter cannot be determined exactly, but can only be **estimated**
- A **statistic** can be computed from a sample, and used to estimate a parameter.
- A **statistic** summarises what the researcher knows. A **parameter** is what the researcher wants to know.
- When estimating a parameter, one major issue is **accuracy**: how close is the estimated statistic to the (unknown) true parameter?

Why not observe the whole population?

Typical limitations

- Hard to observe the population
- Not enough time
- Not enough money
- Not enough resources

Sampling

The solution is for us to draw samples and hope or expect to make general statement about the entire population.

Why sample?

- Reduce the number of measurements
- Save time, money and resources
- Might be essential in destructive testing

Notice this is in contrast with **census** and **experiment**.

Survey sampling

Survey design

- What survey design is appropriate for my study?
- How survey will be conducted/implemented?

Sampling is the process of selecting a subset of observations from an entire population of interest so that characteristics from the subset (sample) can be used to draw conclusion or making inference about the entire population.

Sampling procedure

- What sample size is needed for my study?
- How the design will affect the sample size?
- Appropriate survey design provides the best estimation with high reliability at the lowest cost with the available resources.

Selection bias

When a selection procedure is biased, taking a larger sample DOES NOT help. This just repeats the basic mistake at a larger scale.

Getting an opinion

- Should a woman have control over her own body, including her reproductive system ?
- Should a doctor be allowed to murder unborn children who can't defend themselves ?

Measurement bias

- Recall bias
- Sensitive questions
- Misinterpret the questions
- Wording of question
- Other attributes of the interview as a source of bias...

Schuman and Converse (1971) performed a study to check whether or not the race of the interviewer influenced responses after major racial riots in 1968 in Detroit. A sample of 495 African American were asked:

> "Do you personally feel that you can trust most white people, some white people, or none at all"

- White interviewer: 35% responded "most" ($n = 165$)
- African American interviewer: 7% responded "most" ($n = 330$)

Back to the 1936 US election

- Political split followed economic lines more closely in 1936 than previous elections.
- They later found that the 2.4 million responses didn't even represent the 10 million people who were sent the surveys.
- Lower-income and upper-income people tend not to respond so the survey was over-represented by middle-class people.

Which mode of survey administration is best?

- Mail
- Personal interview
- Telephone
- Online poll at the end of an article
- Twitter poll

Bias

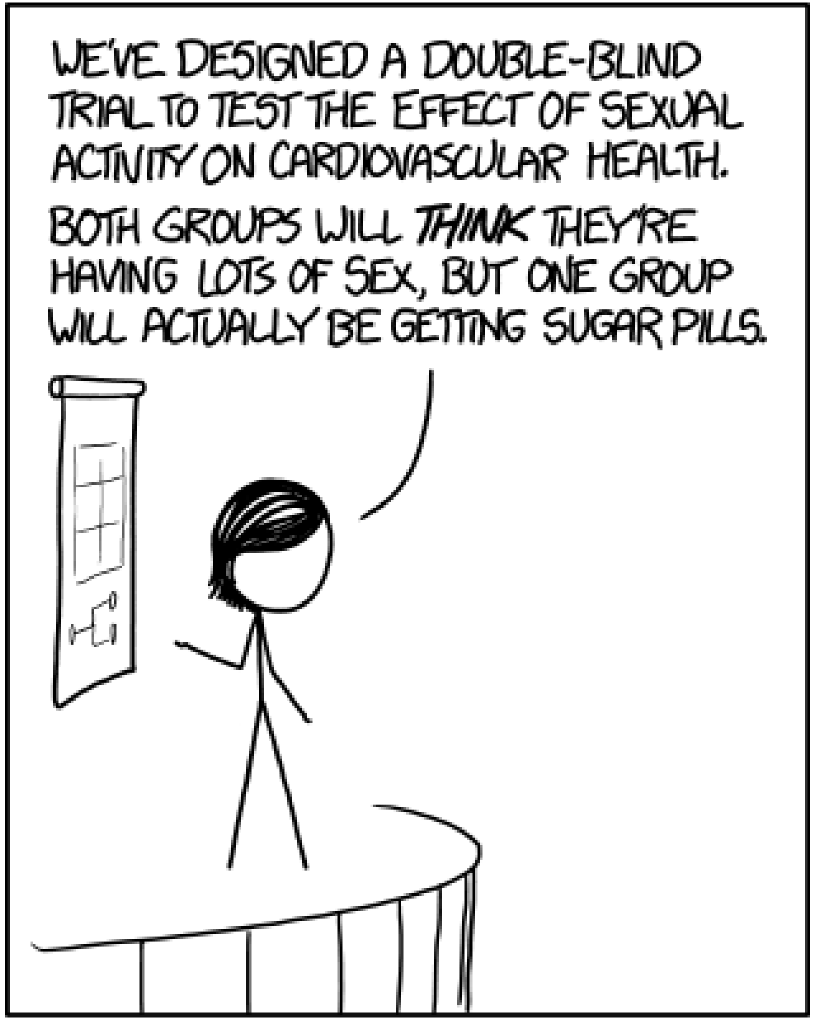
Bias is any factor that favours certain outcomes or responses, or influences an individual's responses. Bias may be unintentional (accidental), or intentional (to achieve certain results).

When looking at data from a survey think about

- **Selection bias / sampling bias**: the sample does not accurately represent the population. Example: Attendees at a Star Trek convention may report that their favorite genre is science fiction.
- **Non-response bias**: Certain groups are under-represented because they elect not to participate. Example: a restaurant may give each table a "customer satisfaction" survey with their bill.
- **Measurement or designed bias**: Bias factors in the sampling method influence the data obtained. Example: a respondent may answer questions in the way she thinks the questioner wants her to answer.

Controlled experiments

Randomised controlled double-blind trials



THE LIMITATIONS OF BLIND TRIALS

What is a randomised controlled double-blind study? Why is it good but rare?

1. Investigators obtain a representative sample of subjects.
2. Investigators randomly allocate the subjects into a **treatment group** and a **control group**.
3. The **control group** is given a placebo, but neither the subjects nor the investigators know the identity of the 2 groups (double-blind).
4. Investigators compare the responses of the 2 groups.
5. The design is good because we expect the 2 groups to be similar, hence any difference in the responses is likely to be caused by the treatment.

Observational studies

Does smoking cause cancer?

> "Tobacco smoking is the largest preventable cause of cancer, responsible for more cancer deaths in Australia than any other single factor. It is also directly responsible for many heart and lung diseases."
>
> -- *Australian Cancer Council*

The Sydney Morning Herald | National

NEWS SITE OF THE YEAR

SUBSCRIBE LOGIN / SIGN-UP

News Sport Business World Politics Comment Property Entertainment Lifestyle Travel Cars ...

NSW News National Environment Health Education Good Weekend Investigations Galleries Clique Photos V ...

Home / News / Health

MARCH 13 2012

SAVE PRINT LICENSE ARTICLE

Health issue irrelevant, tobacco firms tell court

Lenore Taylor

SHARE TWEET MORE

Big tobacco companies have told the High Court they ''deny the content'' of documents lodged by the federal government making the case that smoking causes lung cancer.

MOST POPULAR

1 Crocodile attacks man in NT's Daly River region

In a hearing to the Australian High Court in 2012 disputing the introduction of cigarette plain packaging with health warnings, while British American Tobacco was prepared to accept that there are serious health consequences caused by smoking, Imperial Tobacco responded "**some people say that...**" [1]

1. Source: SMH

The need for observational studies

- By necessity, many research questions require an **observational study**, rather than a controlled experiment.
- For example, with a study on the effects of smoking, investigators cannot choose which subjects will be in the **treatment group** (smoking). Rather, they must **observe** medical results for the 2 groups.
- Similarly, most **educational research** is based on observational studies.
- The conclusions of observational studies require great care.

> Observational studies can not establish causation.

- A good **randomised controlled experiment** can establish *causation*, an **observational study** can only establish *association*.
- An observational study may *suggest* causation, but it can't *prove* causation.

Misleading hidden confounders

- Confounding occurs when the **treatment group** and **control group** differ by some third variable (other than the treatment) which influences the response that is studied.
- Confounders can be hard to find, and can mislead about a cause and effect relationship.
- Confounding (or lurking) variables can be introduced into a randomised study if any of the subjects drop out, causing **selection bias** or **survivor bias**. Similarly, if not all subjects keep taking the treatment or placebo, we get the confounding of **adherers** and **non-adherers**.

Examples: Spurious Correlations

Lung cancer associations

A study finds that having yellow fingertips is associated with lung cancer. Does having yellow fingertips cause lung cancer?

A study finds that smokers tend to have higher rates of lung cancer. Does smoking cause lung cancer?

Strategy for dealing with confounders

Sometimes we can make the groups more comparable by dividing them into subgroups with respect to the confounder.

For example, if alcohol consumption is a potential confounding factor for smoking's affect on liver cancer, we can divide our subjects into 3 groups:

- heavy drinkers
- medium drinkers
- light drinkers.

This is called **controlling** for alcohol consumption.

Controlling for confounding

We can control for confound by making 3 separate comparisons:

- heavy drinking: smokers vs non-smokers
- medium drinking: smokers vs non-smokers
- light drinking: smokers vs non-smokers

> What are the limitations of this strategy?

- This strategy is limited by our ability to identify all confounders and then divide the study by the confounders.
- This explains the long time to establish that smoking causes lung cancer. Researchers needed to control for factors such as health, fitness, diet, lifestyle, environment etc.

Is smoking good for your longevity?

A famous study by Appleton, French, and Vanderpump (1996) considered data on female subjects 20 years apart. The data came 2 studies:

- inital data from a 1 in 6 survey from an electoral roll in a mixed urban and rural area near Newcastle upon Tyne UK (Tunbridge, Evered, Hall, et al., 1977).
- follow-up data 20 years later (Vanderpump, Tunbrldge, French, et al., 1995).

The study concentrated on the 1314 women who were either smokers or non-smokers (in the full data, only 162 had stopped smoking and only 18 did not record their status).

Initial results (survival over a 20 year period)

Status	Died	Survived	Total	Mortality Rate
Smoker	139	443	582	23.9%
Non-smoker	230	502	732	31.4%
Total	369	945	1314	28.1%

```r
library(readr)
x = read_csv("https://git.io/fNGXk")
x
```

```
## # A tibble: 4 × 3
##   status     survival count
##   <chr>      <chr>    <dbl>
## 1 Smoker     Died       139
## 2 Smoker     Survived   443
## 3 Non-smoker Died       230
## 4 Non-smoker Survived   502
```

```r
library(tidyr)
# for uncount()
library(dplyr)
# for %>%, group_by() and summarise()
```

```r
x_long = tidyr::uncount(x, weights = count)
dim(x_long)
```

```
## [1] 1314    2
```

```r
x_long %>% dplyr::group_by(status) %>%
  dplyr::summarise(
    rate = sum(survival == "Died")/n()
  )
```

```
## # A tibble: 2 × 2
##   status      rate
##   <chr>      <dbl>
## 1 Non-smoker 0.314
## 2 Smoker     0.239
```

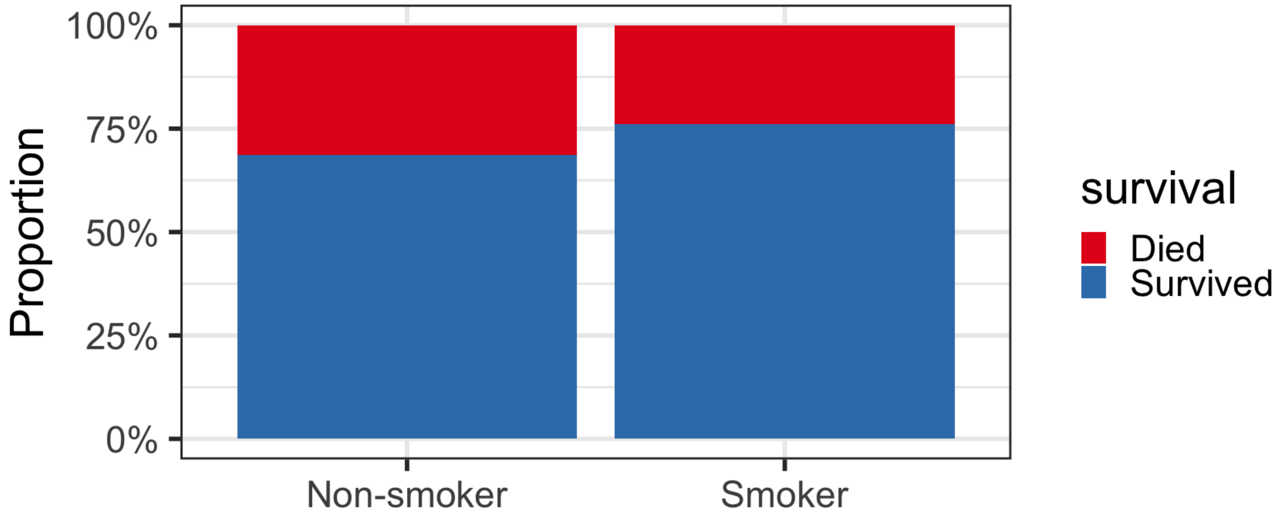
```r
x_long %>% # without group_by()
  dplyr::summarise(
    rate = sum(survival == "Died")/n()
  )
```

```
## # A tibble: 1 × 1
##    rate
##   <dbl>
## 1 0.281
```

The URL `https://git.io/fNGXk` is a shortened link to `https://raw.githubusercontent.com/DATA2002/data/master/appleton1996.csv`

Initial results (survival over a 20 year period)

```r
library(ggplot2)
library(scales)
ggplot(x, aes(x = status, y = count, fill = survival)) +
  geom_bar(stat = "identity", position="fill") +
  scale_y_continuous(labels = scales::percent_format()) +
  labs(x = "", y = "Proportion") +
  theme_bw(base_size = 30) +
  scale_fill_brewer(palette = "Set1")
```



What does this data seem to say?

- It seems to imply that smoking has a 'protective effect'.
- **Smokers live longer?**

Simpsons paradox

Observational studies with a confounding variable can lead to Simpson's Paradox

- **Simpson's Paradox** was first mentioned by British statistician Udny Yule in 1903.
- It was named after Edward H. Simpson[1]
- Sometimes there is a clear trend in individual groups of data that disappears when the groups are pooled together.
- It occurs when relationships between percentages in subgroups are reversed when the subgroups are combined, because of a confounding or lurking variable.
- The association between a pair of variables (X, Y) reverses sign upon conditioning of a third variable Z, regardless of the value taken by Z.

1. Original paper by Simpson (1951).

Mortality by age group

```r
y = readr::read_csv("https://git.io/fNGSi")
dplyr::glimpse(y)
```

```
## Rows: 28
## Columns: 4
## $ status    <chr> "Smoker", "Smoker", "Smoker", "Smoker", …
## $ survival  <chr> "Died", "Died", "Died", "Died", "Died", …
## $ age_group <chr> "18-24", "25-34", "35-44", "45-54", "55-…
## $ count     <dbl> 2, 3, 14, 27, 51, 29, 13, 53, 121, 95, 1…
```

We can "spread" this long form data into a "wide" table that is easier for a human to interpret.

```r
# might need dev version of tidyr for pivot_wider
# devtools::install_github("tidyverse/tidyr")
ytab = y %>%
  tidyr::pivot_wider(id_cols = age_group,
                     names_from = c(status, survival),
                     values_from = count,
                     names_sep = " ")
```

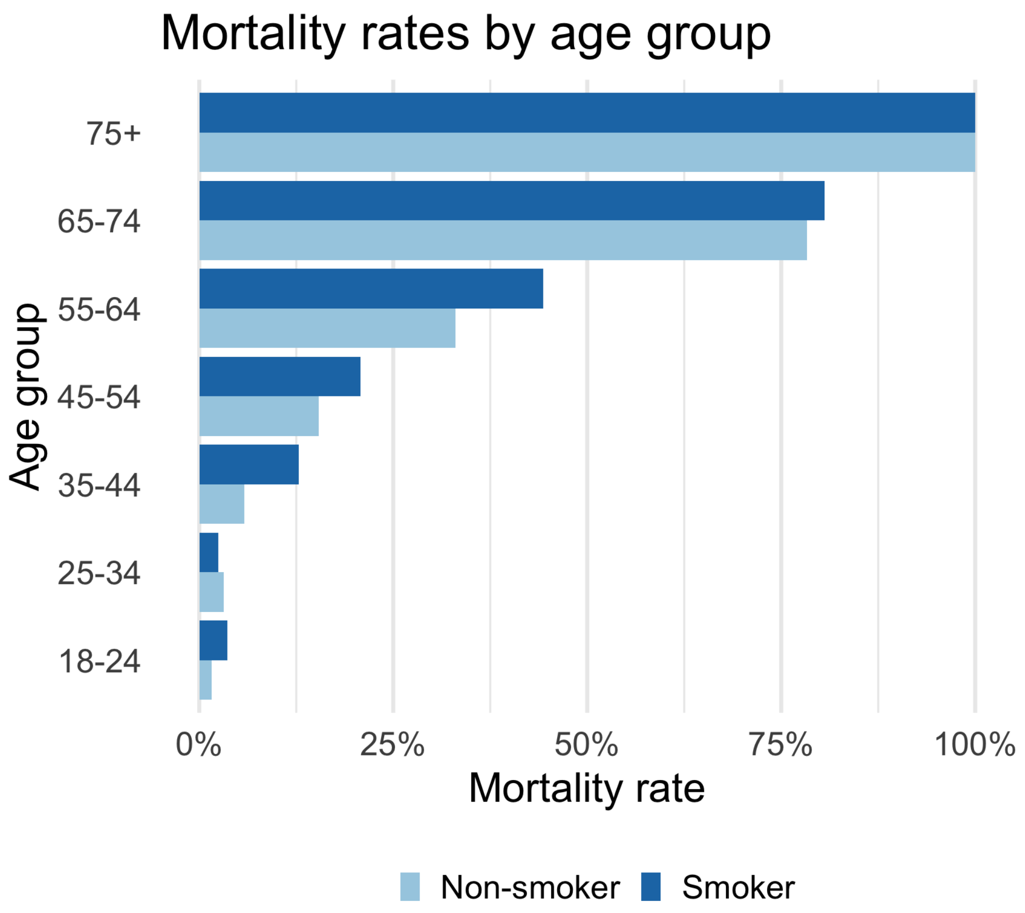
Age group	Smoker Died	Smoker Survived	Non-smoker Died	Non-smoker Survived
18-24	2	53	1	61
25-34	3	121	5	152
35-44	14	95	7	114
45-54	27	103	12	66
55-64	51	64	40	81
65-74	29	7	101	28
75+	13	0	64	0

The URL `https://git.io/fNGSi` is a shortened link to `https://raw.githubusercontent.com/DATA2002/data/master/appleton1996_age.csv`

Mortality by age group

```r
mortality = y %>%
  uncount(weights = count) %>%
  group_by(status, age_group) %>%
  summarise(rate = mean(survival=="Died"))
  # could also have used:
  # summarise(rate = sum(survival=="Died")/n())
  # note the n() function
p = mortality %>%
  ggplot() +
  aes(x = age_group, y = rate, fill = status) +
  geom_bar(stat = "identity", position = "dodge") +
  theme_minimal(base_size = 34) +
  scale_fill_brewer(palette = "Paired") +
  scale_y_continuous(labels = scales::percent_format()) +
  labs(title = "Mortality rates by age group",
       y = "Mortality rate",
       x = "Age group",
       fill = "") +
  theme(panel.grid.major.y = element_blank(),
        legend.position = "bottom") +
  coord_flip()
```

```r
p
```

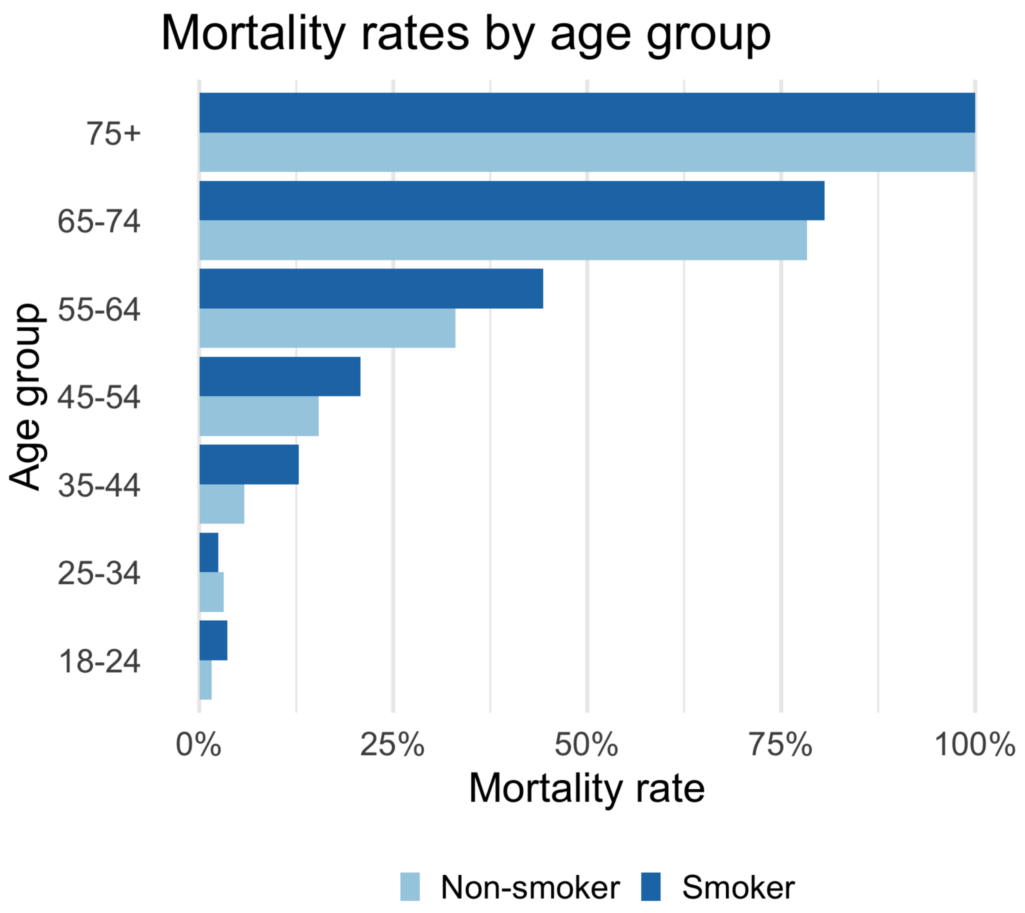


Mortality by age group

What does this summary of data reveal?

- Not many young people died.
- Most old people died.
- In the middle age groups, smokers tended to have higher mortality rates than non-smokers.

```
p
```

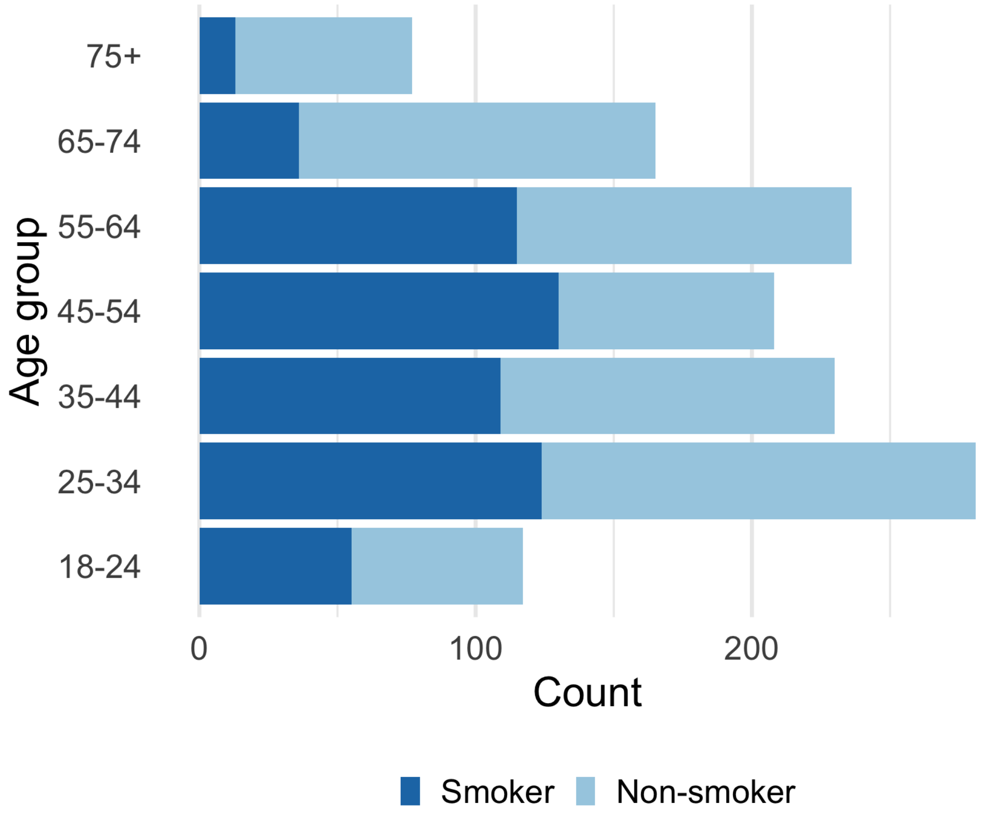


How did we got the wrong overall conclusion?

Consider the distribution of samples by smoking status across age groups.

```
p2 = y %>%
  uncount(weights = count) %>%
  ggplot() +
  aes(x = age_group, fill = status) +
  geom_bar() +
  theme_minimal(base_size = 34) +
  scale_fill_brewer(palette = "Paired") +
  labs(y = "Count",
       x = "Age group",
       fill = "") +
  theme(panel.grid.major.y = element_blank(),
        legend.position = "bottom") +
  guides(fill = guide_legend(reverse = TRUE)) +
  coord_flip()
```

```
p2
```

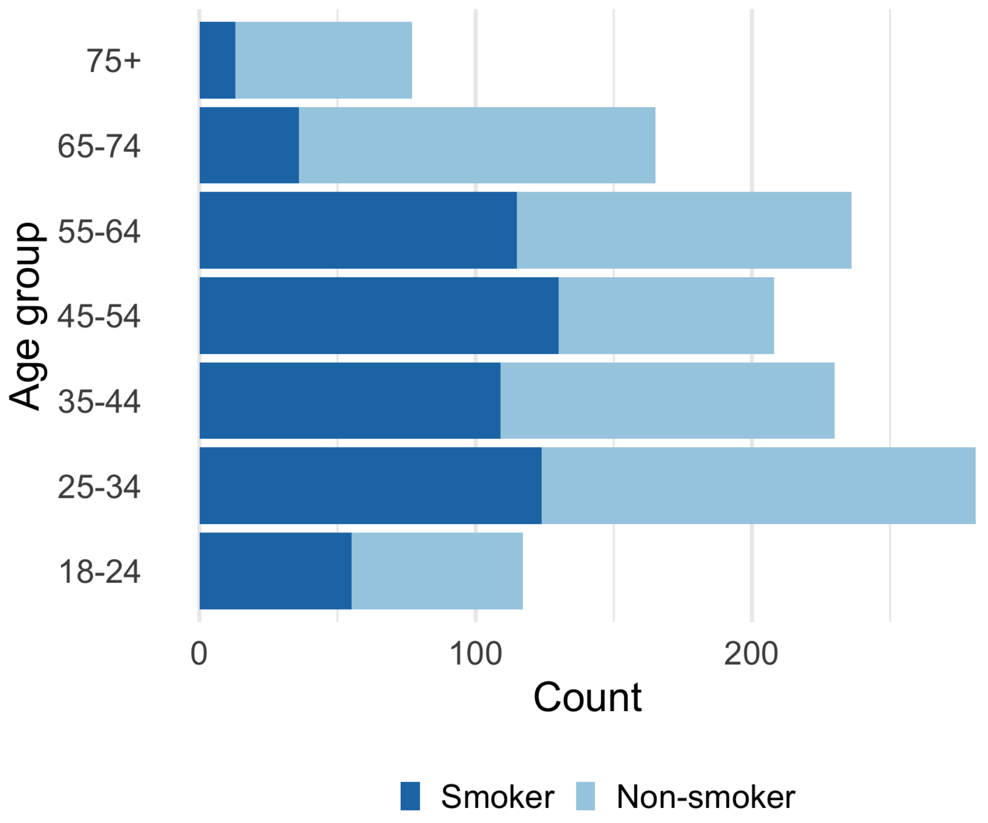


How did we got the wrong overall conclusion?

What does this summary of data reveal?

- As there are many more young women who smoked than older women, and as younger women are expected to live longer than older women, adding all the groups together makes smoking appear to be beneficial.
- This is a classic example of Simpson's Paradox phenomenon; it shows that a trend present within multiple groups can reverse when the groups are combined.

```
p2
```



Further reading: Berman, DalleMule, Greene, and Lucker (2012)

R packages and functions

- `readr::read_csv()` for reading in csv files
- `tidyr::uncount()` for converting tabulated data to observation level data
- `dplyr::glimpse()` for inspecting the structure of objects
- `dplyr::group_by()` for creating a grouping structure in your data
- `dplyr::summarise()` for extracting summary statistics from grouped data
- `dplyr::n()` for calculating the number of observations in a group
- `base::dim()` for finding the dimensions (`rows columns`) of a data frame
- `base::sum(survival == "Died")` counts the number of times the variable `survival` equals `"Died"`
- `ggplot2::ggplot()` and associated functions from the **ggplot2** package `aes()`, `geom_bar()`, `labs()`, `scale_fill_brewer()`, `theme_bw()`, `theme_minimal()`
- `scales::percent_format()` for nice formatting of **ggplot2** axes. E.g. to make the y axis nicely formatted, `scale_y_continuous(labels = scales::percent_format())`

References

Appleton, D. R., J. M. French, and M. P. J. Vanderpump (1996). "Ignoring a Covariate: An Example of Simpson's Paradox". In: *The American Statistician* 50.4, pp. 340-341. DOI: 10.1080/00031305.1996.10473563.

Berman, S., L. DalleMule, M. Greene, et al. (2012). "Simpson's Paradox: a cautionary tale in advanced analytics". In: *Significance*. URL: https://www.significancemagazine.com/14-the-statistics-dictionary/106-simpson-s-paradox-a-cautionary-tale-in-advanced-analytics.

Schuman, H. and J. M. Converse (1971). "The effects of black and white interviewers on black responses in 1968". In: *Public Opinion Quarterly* 35.1, pp. 44-68. ISSN: 0033-362X. DOI: 10.1086/267866.

Simpson, E. H. (1951). "The Interpretation of Interaction in Contingency Tables". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 13.2, pp. 238-241. URL: http://www.jstor.org/stable/2984065.

Tunbridge, W. M. G., D. C. Evered, R. Hall, et al. (1977). "The spectrum of thyroid disease in a community: the Whickham survey". In: *Clinical Endocrinology* 7.6, pp. 481-493. DOI: 10.1111/j.1365-2265.1977.tb01340.x.

Vanderpump, M. P. J., W. M. G. Tunbrldge, J. M. French, et al. (1995). "The incidence of thyroid disorders in the community: a twenty-year follow-up of the Whickham Survey". In: *Clinical Endocrinology* 43.1, pp. 55-68. DOI: 10.1111/j.1365-2265.1995.tb01894.x.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. New York, NY: Springer-Verlag. ISBN: 978-3-319-24277-4. URL: http://ggplot2.tidyverse.org/.

Wickham, H., R. François, L. Henry, et al. (2018). *dplyr: A Grammar of Data Manipulation*. R package version 0.7.5. URL: https://CRAN.R-project.org/package=dplyr.

Wickham, H. and L. Henry (2018). *tidyr: Easily Tidy Data with 'spread()' and 'gather()' Functions*. R package version 0.8.1. URL: https://CRAN.R-project.org/package=tidyr.

Wickham, H., J. Hester, and R. Francois (2017). *readr: Read Rectangular Text Data*. R package version 1.1.1. URL: https://CRAN.R-project.org/package=readr.