DATA2002

Collecting data

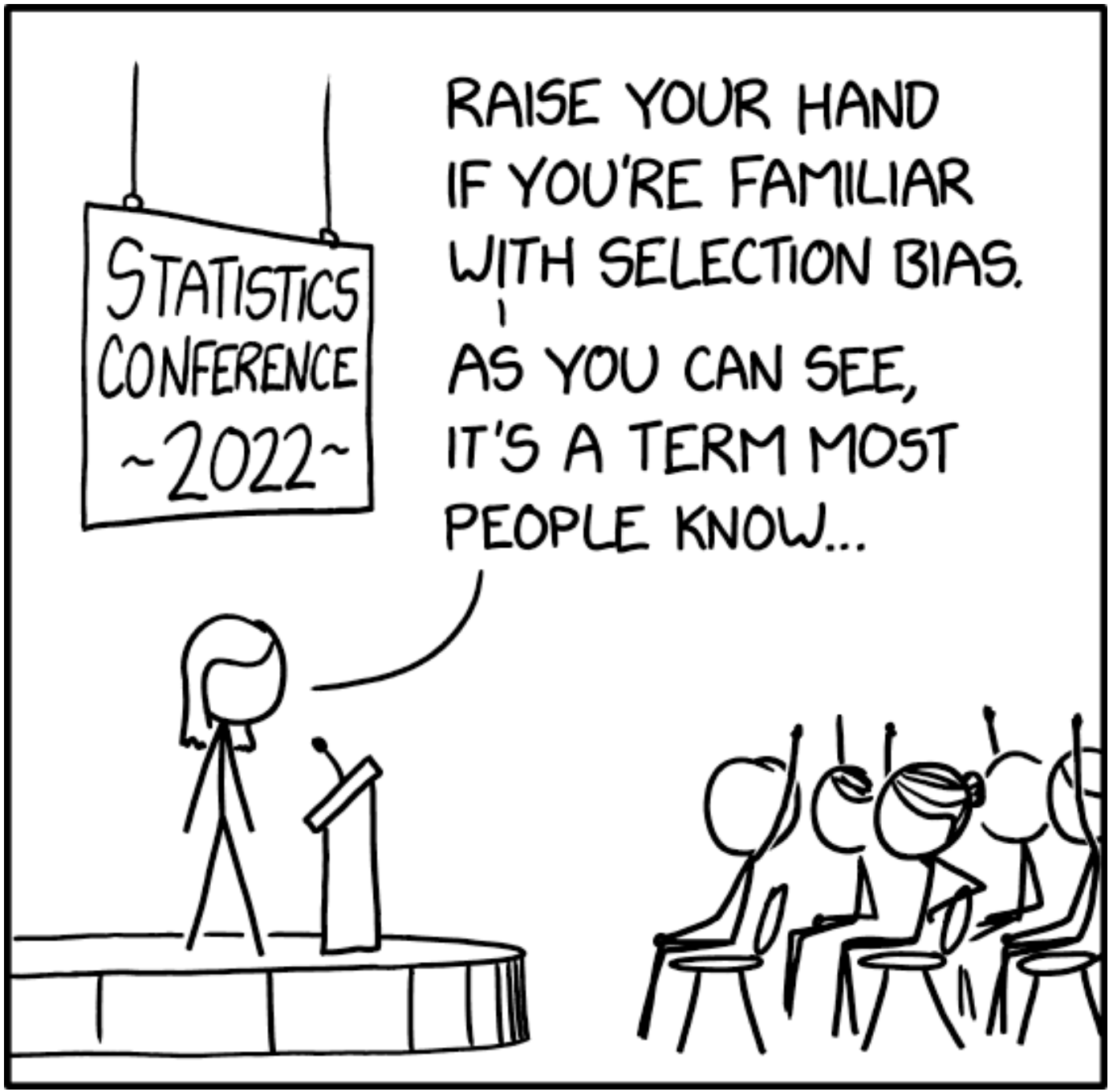
Garth Tarr
The University of Sydney

THE UNIVERSITY OF SYDNEY

In this lecture

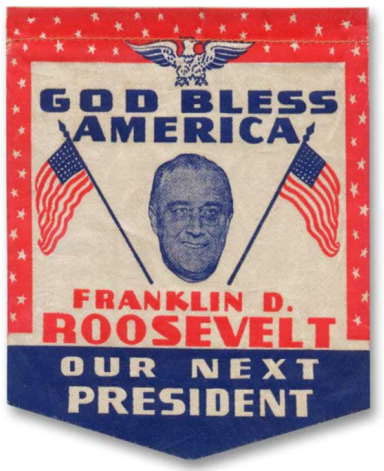
- Sample surveys
- Controlled experiments
- Observational studies
- Simpson’s paradox

Sample surveys


STATISTICS CONFERENCE ~2022~
RAISE YOUR HAND IF YOU'RE FAMILIAR WITH SELECTION BIAS.
AS YOU CAN SEE, IT'S A TERM MOST PEOPLE KNOW...

Polling fail

- 1936 US Presidential election
- Franklin D. Roosevelt was completing his term of office
- America was struggling with high unemployment (16%) following the Great Depression
- Literary Digest polled **10 million** people (mail survey)
- 24% response rate (**2.4 million** people reply)
- They had correctly predicted the winner at every election since 1916
- Predicted victory for Landon



Election 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?!?**

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

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

Sampling

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

Why sample?

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

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

Types of biases

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

Further reading: Catalogue of Bias

Getting an opinion

Phrasing 1

Should a woman have control over her own body, including her reproductive system?

Phrasing 2

Should a doctor be allowed to murder unborn children who can’t defend themselves?

Measurement bias

Schuman & 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

- The 2.4 million responses didn't even represent the 10 million people who were sent the surveys let alone the general voting population.
- **Non-response bias**: the people who didn't respond were different to those that did respond.
- **Selection bias**: addresses sourced from car registration and phone books (skewed towards wealthy Americans).

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

Quick quiz

Which mode of survey administration is best?

1. Mail
2. Personal interview (e.g. door knocking)
3. Telephone
4. Online poll in the middle of a news article
5. Twitter poll

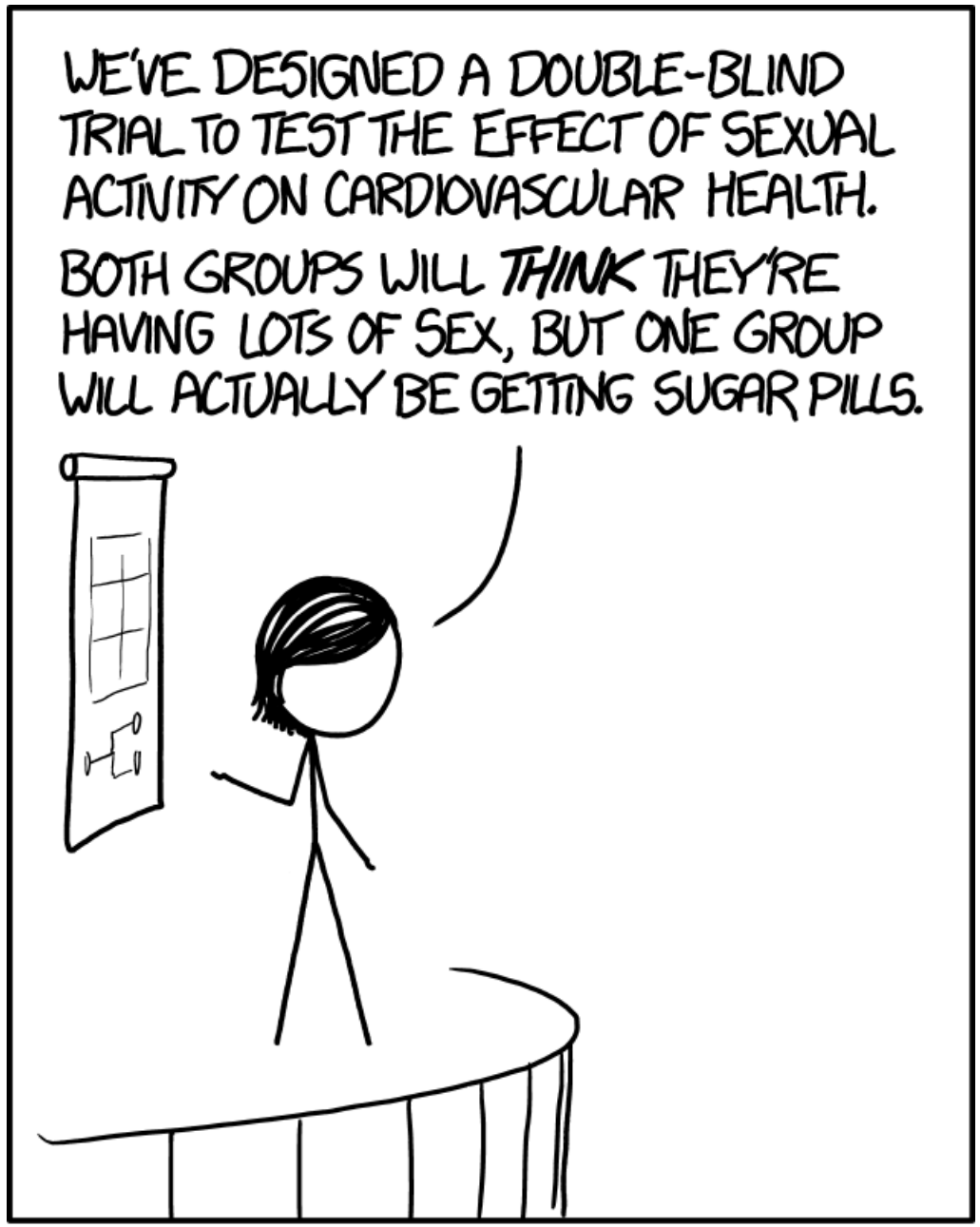
Bias

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



THE LIMITATIONS OF BLIND TRIALS

Randomised controlled double-blind trials

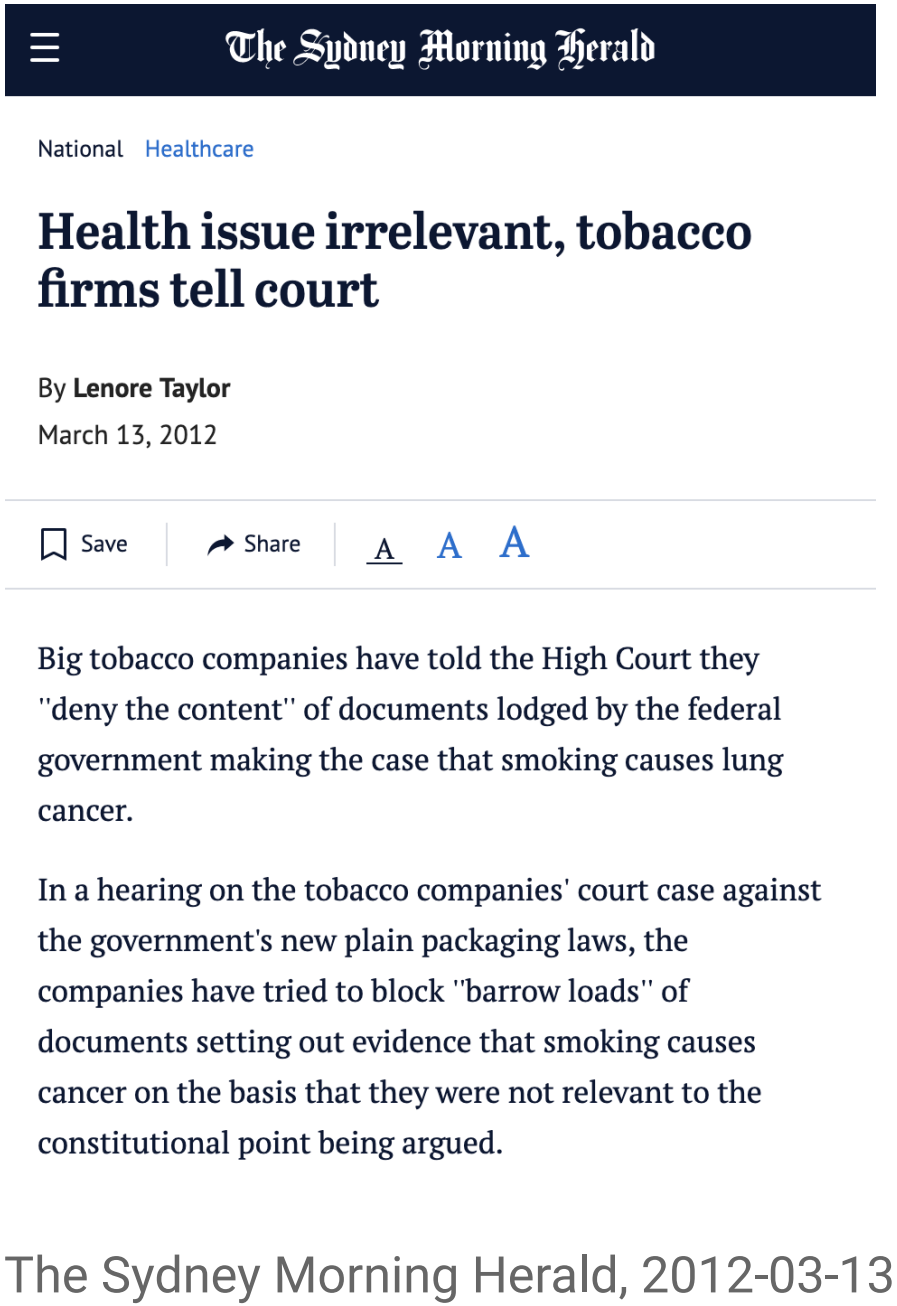
> What is a randomised controlled double-blind study? Why is it good but rare?

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 Healthcare

Health issue irrelevant, tobacco firms tell court

By **Lenore Taylor**
March 13, 2012

Save | Share | A A A

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.

In a hearing on the tobacco companies' court case against the government's new plain packaging laws, the companies have tried to block "barrow loads" of documents setting out evidence that smoking causes cancer on the basis that they were not relevant to the constitutional point being argued.

The Sydney Morning Herald, 2012-03-13

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...**"

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**.

Explore further: 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?

Is smoking good for your longevity?

A famous study by Appleton et al. (1996) considered data on female subjects 20 years apart. Two studies:

- initial data from a 1 in 6 survey from an electoral roll in a mixed urban and rural area near Newcastle upon Tyne UK Tunbridge et al. (1977).
- follow-up data 20 years later Vanderpump 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%

```
library(tidyverse)
x = read_csv("data/appleton1996.csv")
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
```

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

```
[1] 1314    2
```

```
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
```

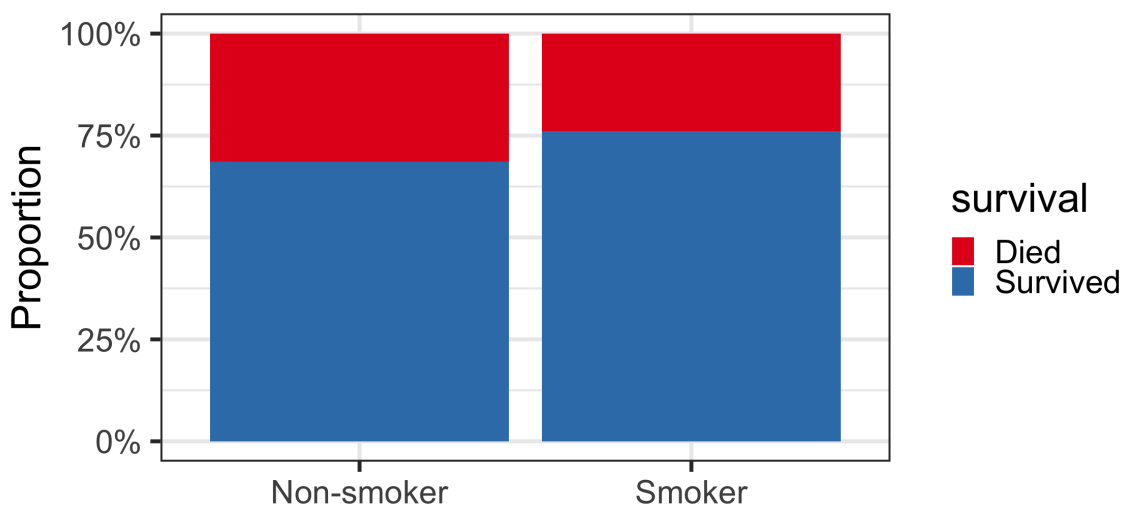
```
x_long %>% # without group_by()
  dplyr::summarise(
    rate = sum(survival == "Died")/n()
  )
```

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

The data file, `appleton1996.csv`, can be downloaded from here.

Initial results (survival over a 20 year period)

```
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")
```

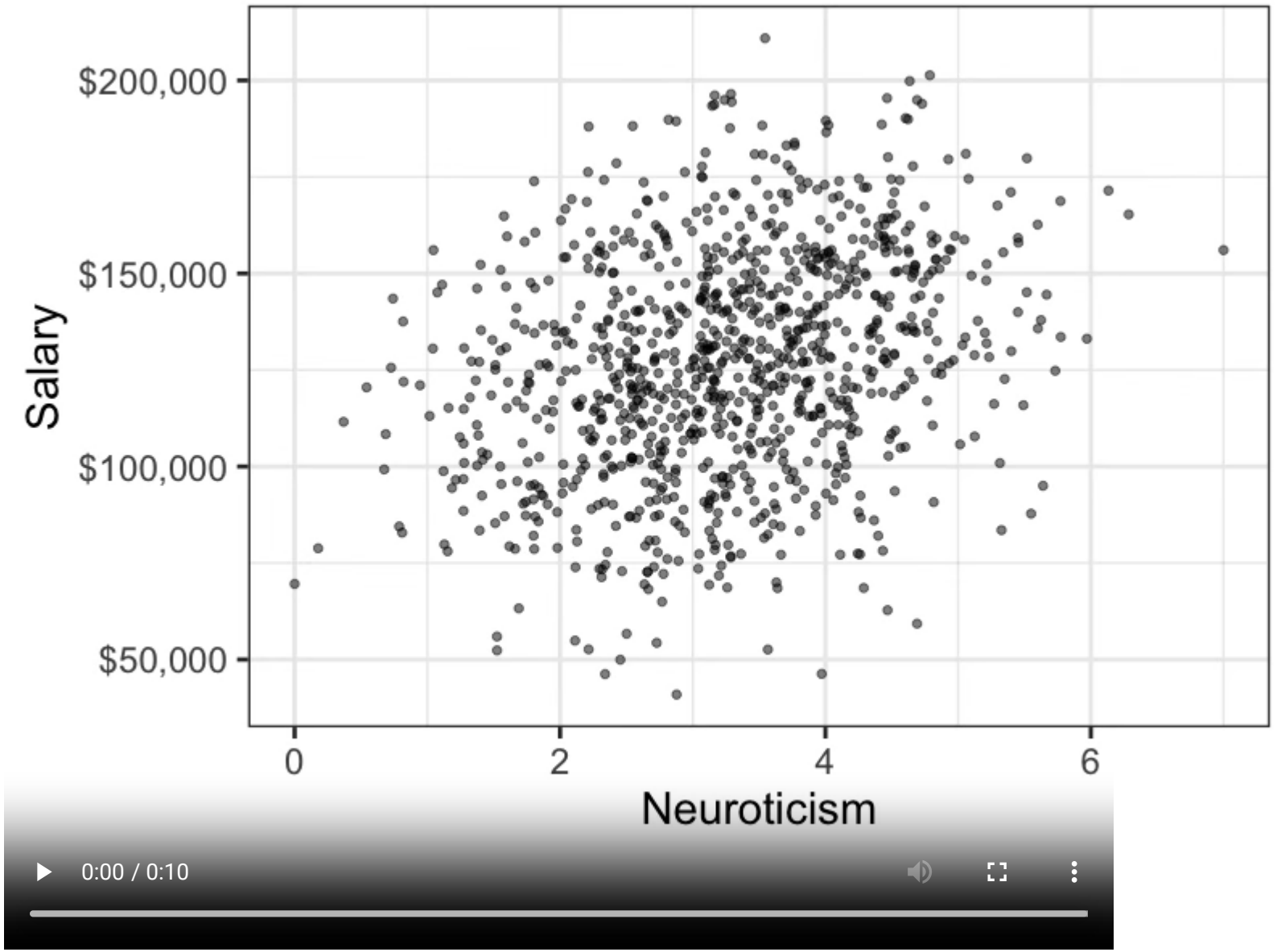


What does this data seem to say?

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

Simpson’s paradox

Simpson's paradox



What is Simpson's paradox?

- **Simpson's Paradox** was first mentioned by British statistician Udny Yule in 1903.
- It was named after Edward H. Simpson (Simpson, 1951)
- 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.

> **Important**
>
> Observational studies with a confounding variable can lead to **Simpson's paradox**

Mortality by age group

```
y = read_csv("data/appleton1996_age.csv")
dplyr::glimpse(y, width = 40)
```

```
Rows: 28
Columns: 4
$ status    <chr> "Smoker", "Smoker", …
$ survival  <chr> "Died", "Died", "Die…
$ age_group <chr> "18-24", "25-34", "3…
$ count     <dbl> 2, 3, 14, 27, 51, 29…
```

We can “spread” this long form data into a “wide” table that is easier for a human to read.

```
ytab = y %>%
  tidyr::pivot_wider(
    id_cols = age_group,
    names_from = c(status, survival),
    values_from = count,
    names_sep = " ")
```

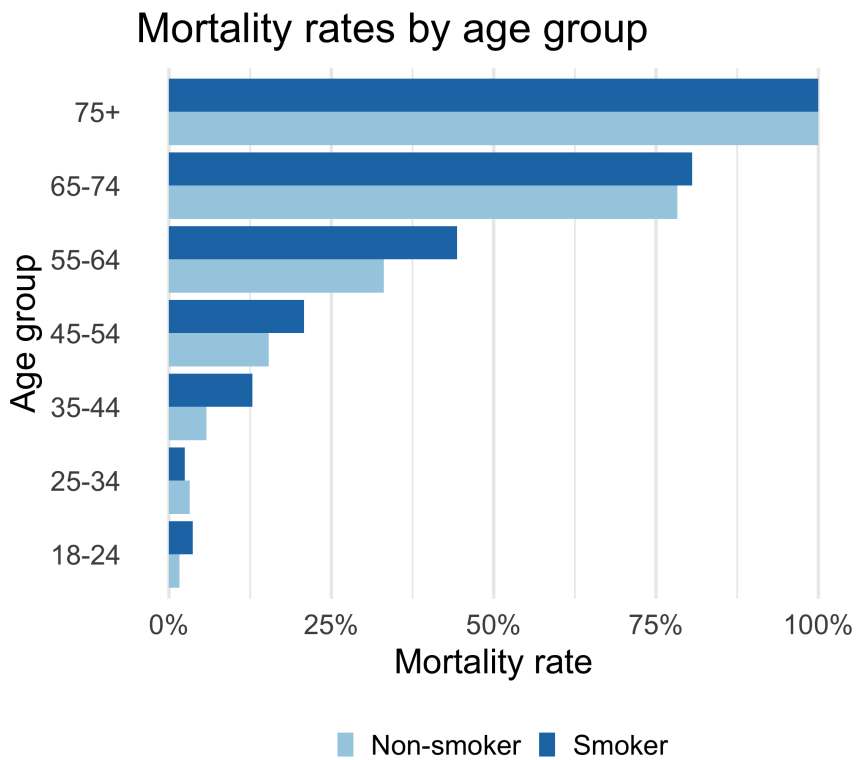
Age group	Smoker Died	Smoker Survived	Non-smoker Died	Non-smoker Survived
18-24	2	53	1	61
25-34	3	121	5	152
35-44	14	95	7	114
45-54	27	103	12	66
55-64	51	64	40	81
65-74	29	7	101	28
75+	13	0	64	0

The data file `appleton1996_age.csv` can be downloaded from here.

Mortality by age group

```
mortality = y %>%
  uncount(weights = count) %>%
  group_by(status, age_group) %>%
  summarise(rate = mean(survival=="Died"))
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()
```

```
p
```

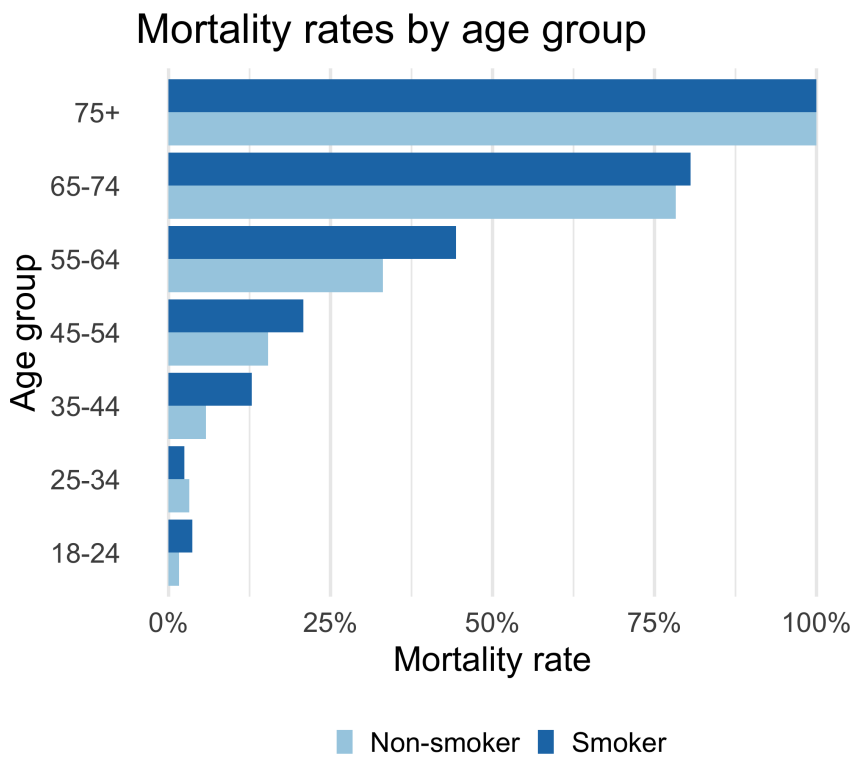


Mortality by age group

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
```

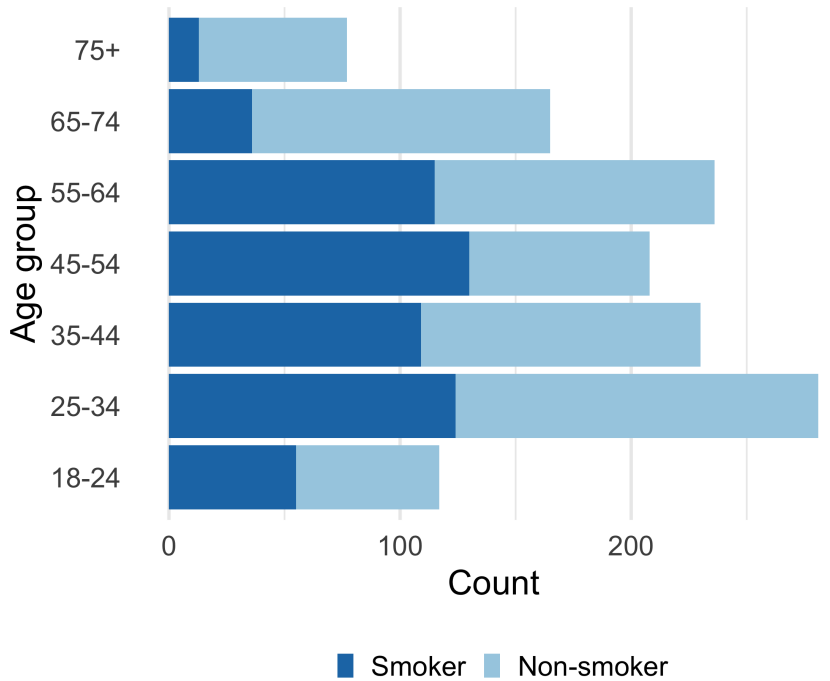


How did we got the wrong overall conclusion?

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
```

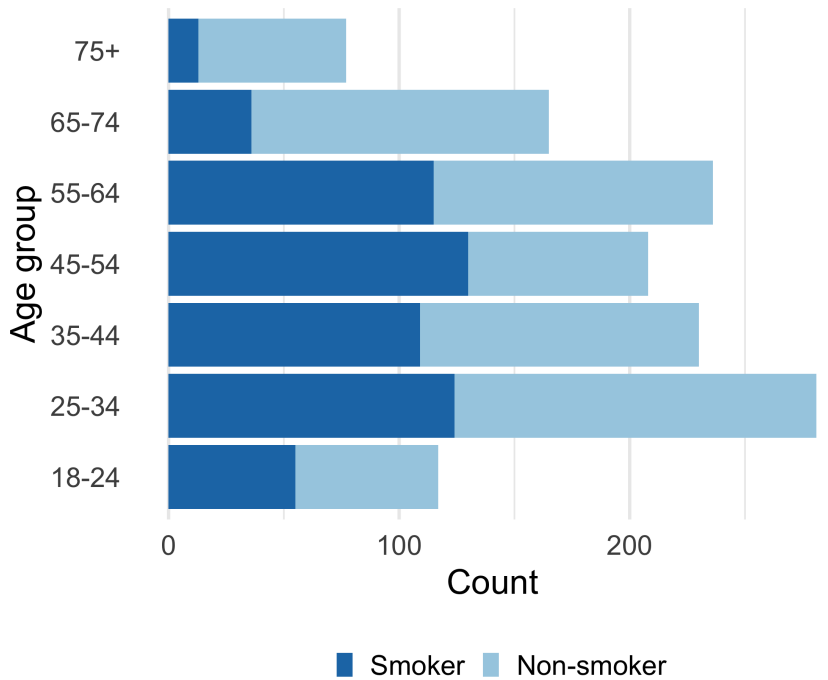


How did we got the wrong overall conclusion?

> **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**: a trend present within multiple groups can reverse when the groups are combined.

```
p2
```



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::mean(survival == "Died")` proportion 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())`

Further reading

We've only just scratched the surface of sampling and bias. Here are some resources to find out more.

- Harford, Tim (2014). Big data: are we making a big mistake? FT Magazine.
- Catalogue of Bias. Centre for Evidence-Based Medicine, University of Oxford.
- Berman et al. (2012) Simpson's Paradox: a cautionary tale in advanced analytics.

References

Appleton, D.R., French, J.M., & Vanderpump, M.P.J. (1996). Ignoring a covariate: An example of Simpson's paradox. *The American Statistician*, 50(4), 340–341. DOI: 10.1080/00031305.1996.10473563

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

Schuman, H., & Converse, J.M. (1971). The effects of black and white interviewers on black responses in 1968. *Public Opinion Quarterly*, 35(1), 44–68. DOI: 10.1086/267866

Simpson, E.H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 13(2), 238–241. http://www.jstor.org/stable/2984065

Tunbridge, W.M.G., Evered, D.C., Hall, R., Appleton, D., Brewis, M., Clark, F., … Smith, P.A. (1977). The spectrum of thyroid disease in a community: The Whickham survey. *Clinical Endocrinology*, 7(6), 481–493. DOI: 10.1111/j.1365-2265.1977.tb01340.x

Vanderpump, M.P.J., Tunbrldge, W.M.G., French, J.M., Appleton, D., Bates, D., Clark, F., … Young, E.T. (1995). The incidence of thyroid disorders in the community: A twenty-year follow-up of the whickham survey. *Clinical Endocrinology*, 43(1), 55–68. DOI: 10.1111/j.1365-2265.1995.tb01894.x