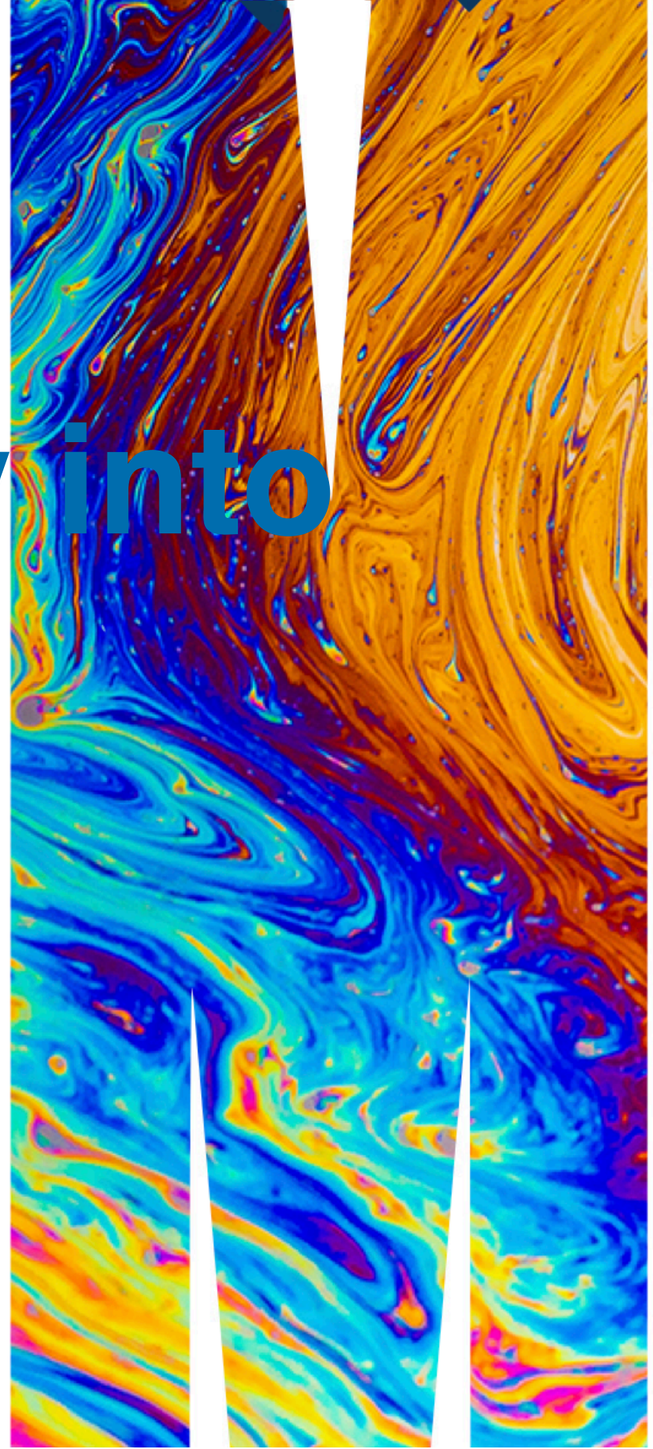
MONASH University

ETC5521: Diving Deeply into Data Exploration

Making comparisons between groups and strata

Professor Di Cook

Department of Econometrics and Business Statistics



At the heart of quantitative reasoning is a single question: Compared to what?

-Edward Tufte

Making comparisons

- Groups defined by strata labelled in categorical variables
- Observations in strata, same or different?
- Is there a baseline, or normal value?
- What are the dependencies in the way the data was collected?
- Are multiple samples recorded for the same individual, or recorded on different individuals?

How would you answer these questions?

- Are housing prices increasing more in Sydney or Melbourne?
- Is the quoted price of the unit/apartment I might buy reasonable, or is it too high?
- Are you more at risk of MPox in Australia or Germany?
- Is the Alfred or Epworth hospitals better for having a baby?
- It’s hot and dry today, is the risk of bushfires too high to go hiking?

Comparing strata

Case study: Melbourne's daily maximum temperature (1/2)

data R

Melbourne's daily maximum temperature from 1970 to 2020.

What are the strata in temporal data?

- How are the temperatures different across months?
- What about the temperature within a month?

Case study: Melbourne's daily maximum temperature (2/2)

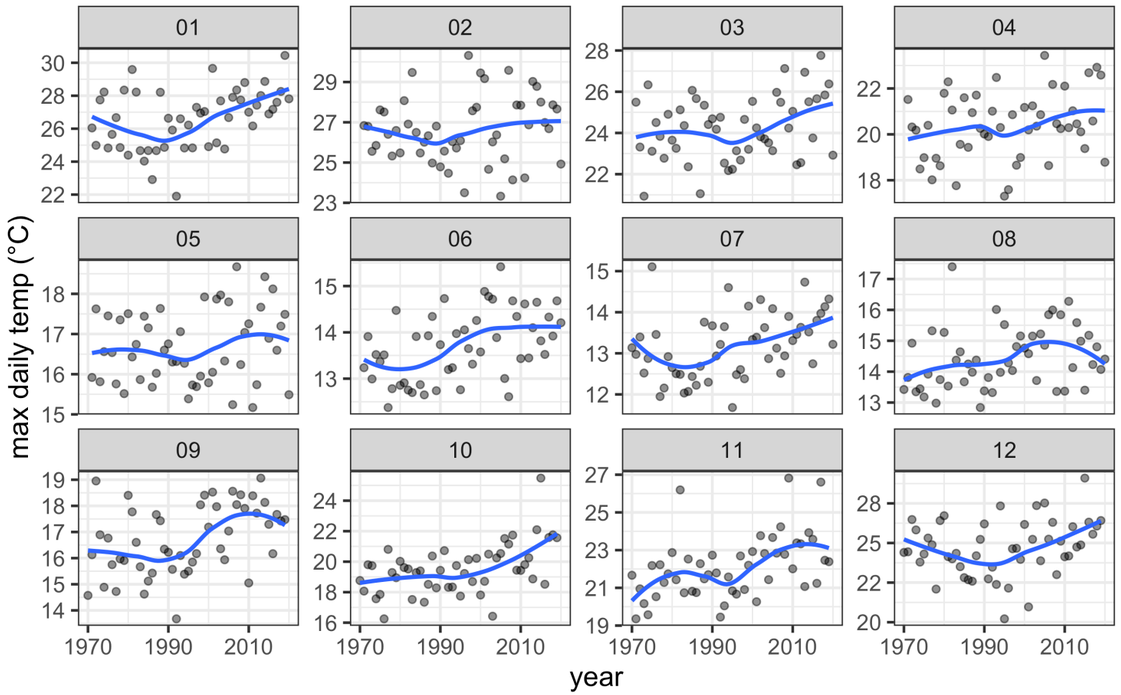
Why can we make the comparison across months?

Because it is the same location, and same years, for each month subset.

Is some variation in temperature each month due to changing climate?

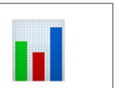
How would you check this?

▶ Code

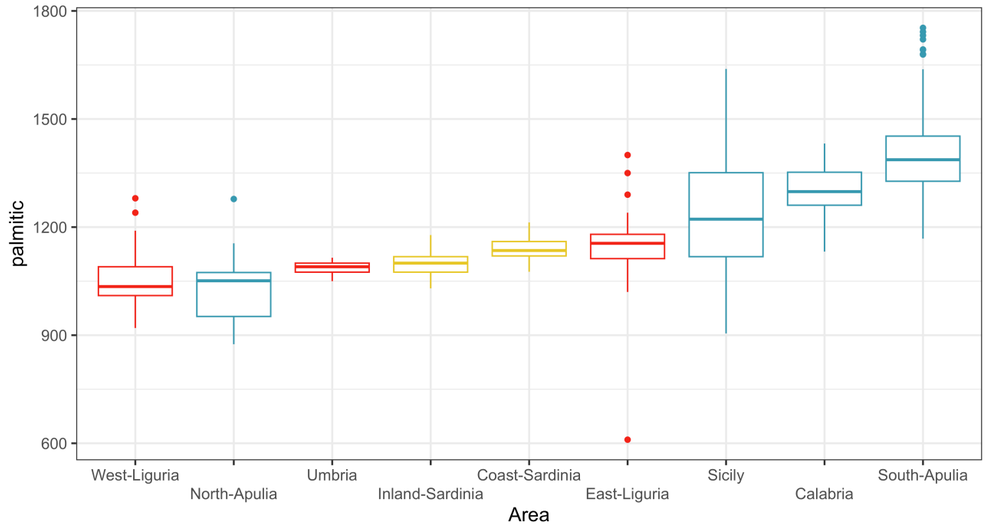


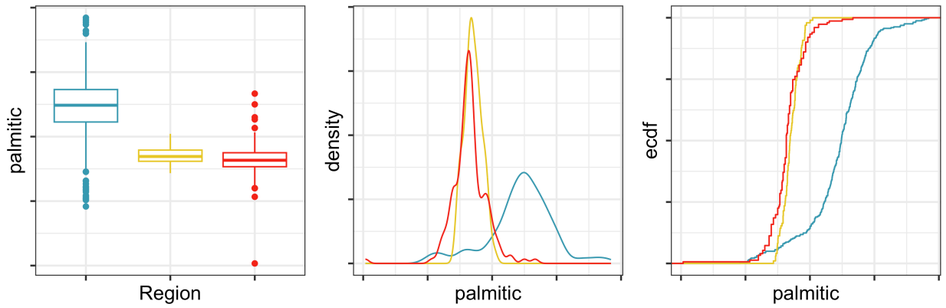
What is `scales="free_y"` for?

Case study: olive oils (1/4)



data R

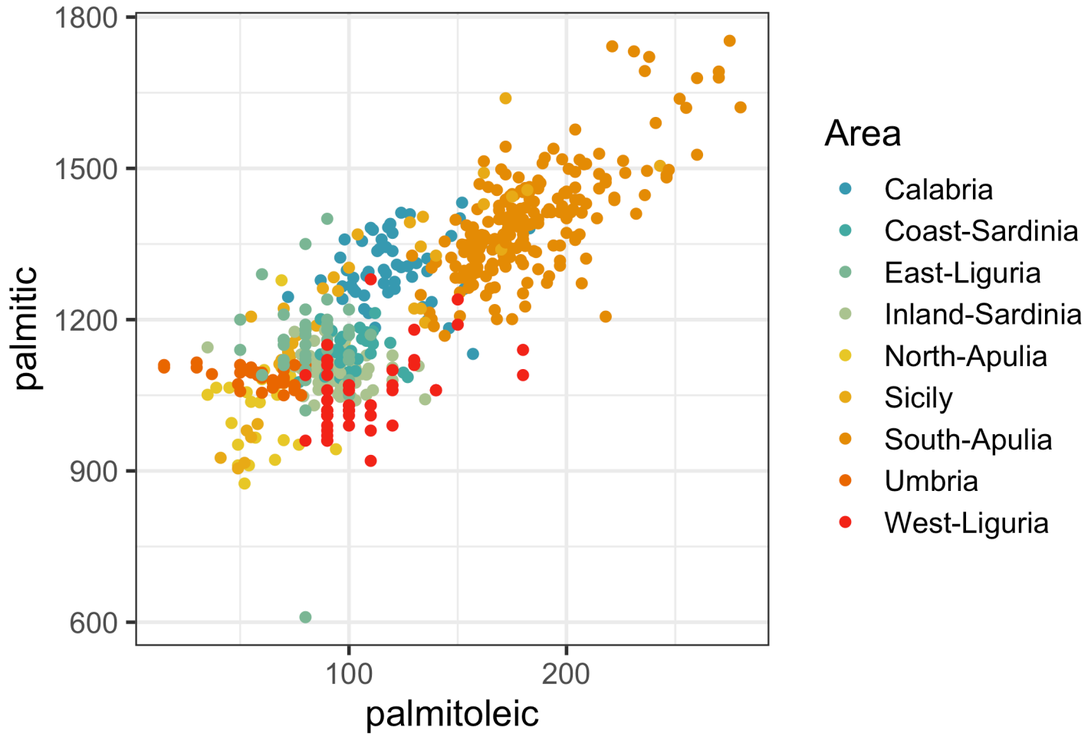




- The olive oil data consists of the percentage composition of 8 fatty acids (palmitic, palmitoleic, stearic, oleic, linoleic, linolenic, arachidic, eicosenoic) found in the lipid fraction of 572 Italian olive oils.
- There are 9 collection areas, 4 from southern Italy (North and South Apulia, Calabria, Sicily), two from Sardinia (Inland and Coastal) and 3 from northern Italy (Umbria, East and West Liguria).

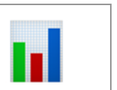
Case study: olive oils (2/4)

R

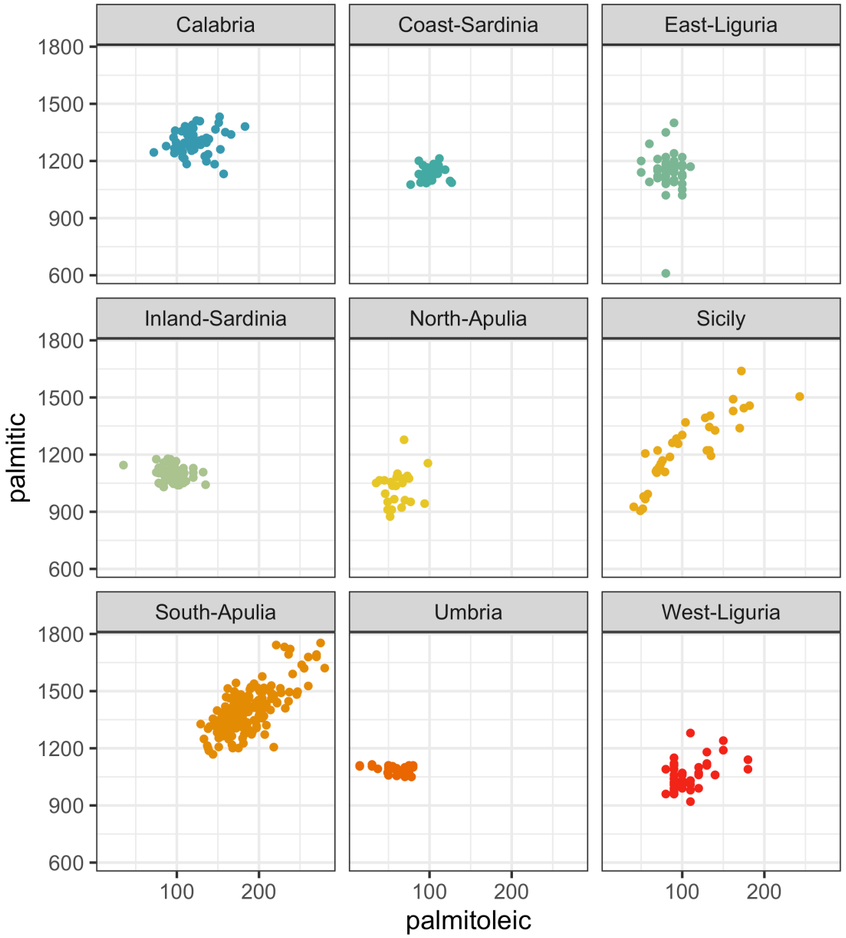


Colour is generally good to differentiate strata but if there are too many categories then it becomes hard to compare.

Case study: olive oils (3/4)



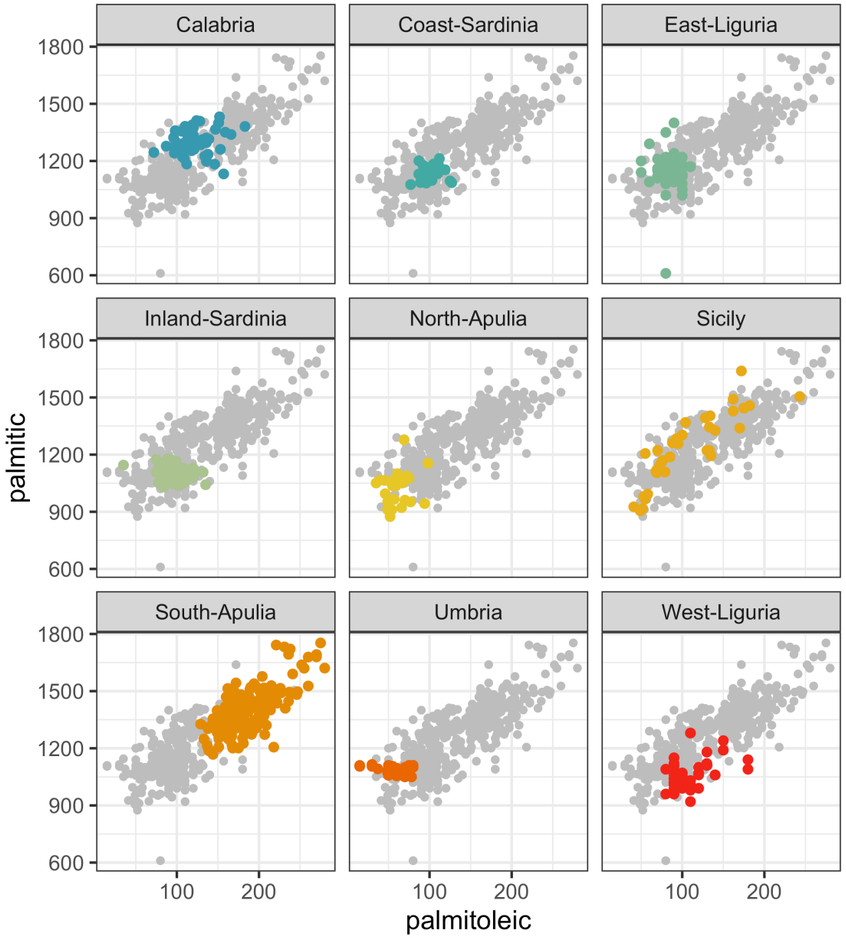
R



It can be hard to compare across plots, because we need to remember what the previous pattern was when focusing on the new cell.

Case study: olive oils (4/4)

R



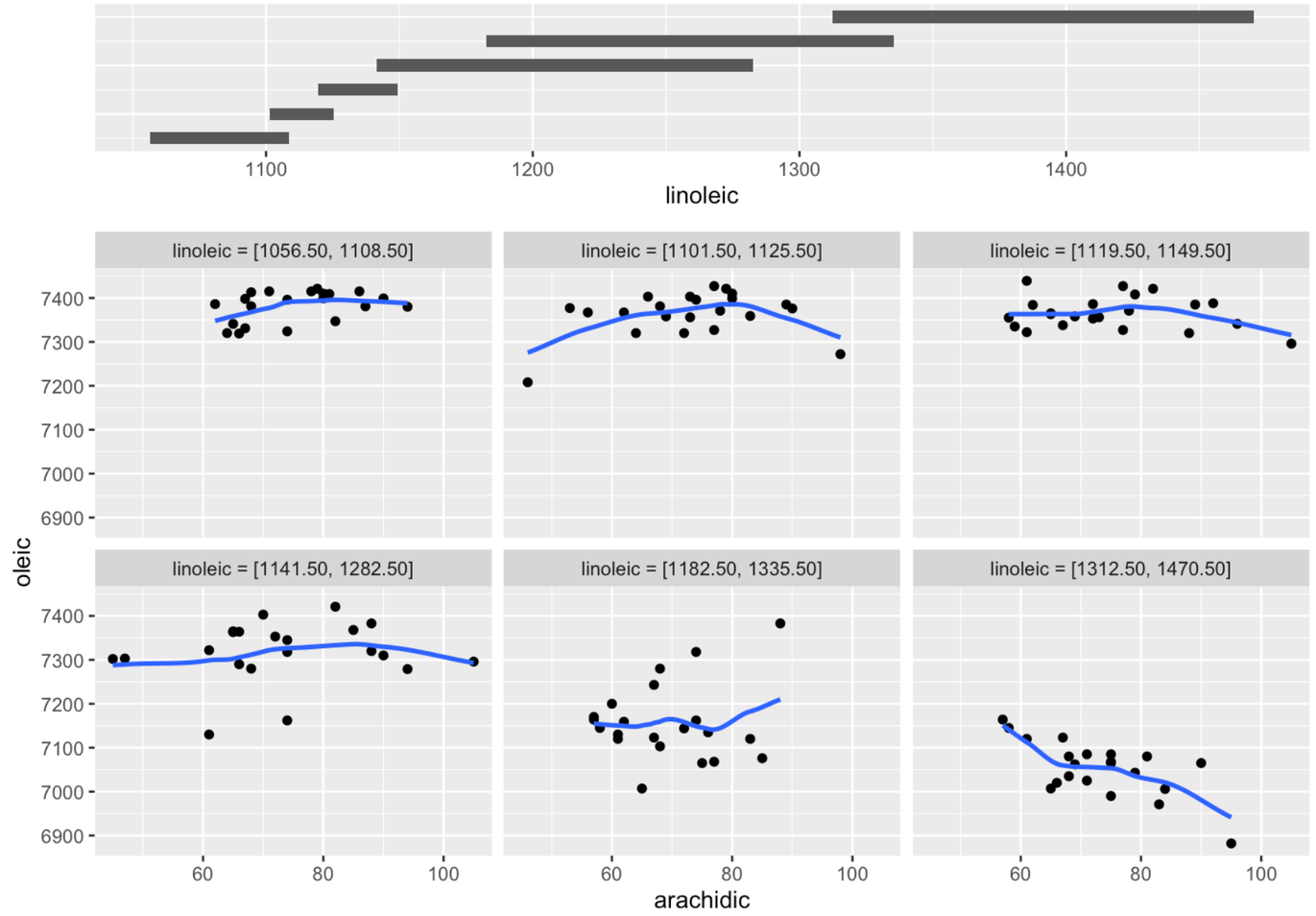
Comparison to all, by putting a shadow of all the data underneath the subset in each cell.

Strata from quantitative variable

The coplot divides the numerical variable into chunks, and facets by these. The chunks traditionally we overlapping.

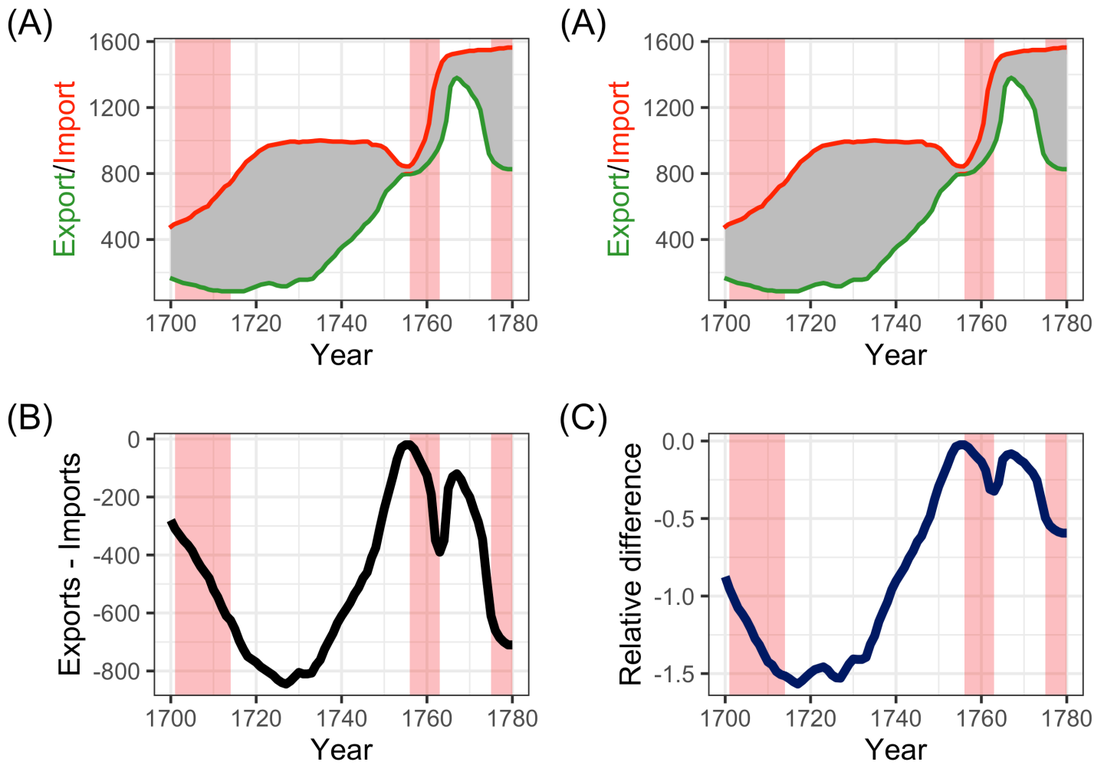
Becker, Cleveland and Shyu, (1996); Cleveland (1993)

▶ Code



Famous example: trade

data R



- The export from England to the East Indies and the import to England from the East Indies in millions of pounds (A).
- Import and export figures are easier to compare by plotting the difference like in (B).
- Relative difference may be more of an interest: (C) plots the relative difference with respect to the average of export and import values.
- The red area correspond to War of the Spanish Succession (1701-14), Seven Years' War (1756-63) and the American Revolutionary War (1775-83).

Paired samples

Pairing adjusts for individual differences

If we were wanting to measure the effect of incorporating a data analytics competition on student learning which is the best design?

METHOD A

- Divide students into two groups. Make sure that each group has similar types of students, so both groups are as similar as possible.
- One group gets an extra traditional assignment, and the other group participates in a data competition.
- Each student takes an exam on the content being taught.
- The scores are compared using side-by-side boxplots and a two-sample permutation test

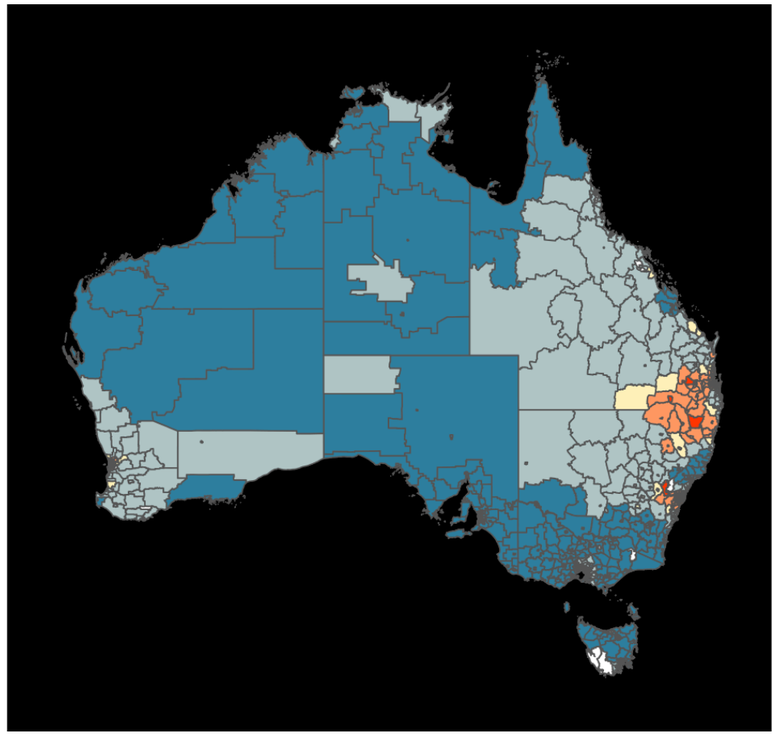
METHOD B

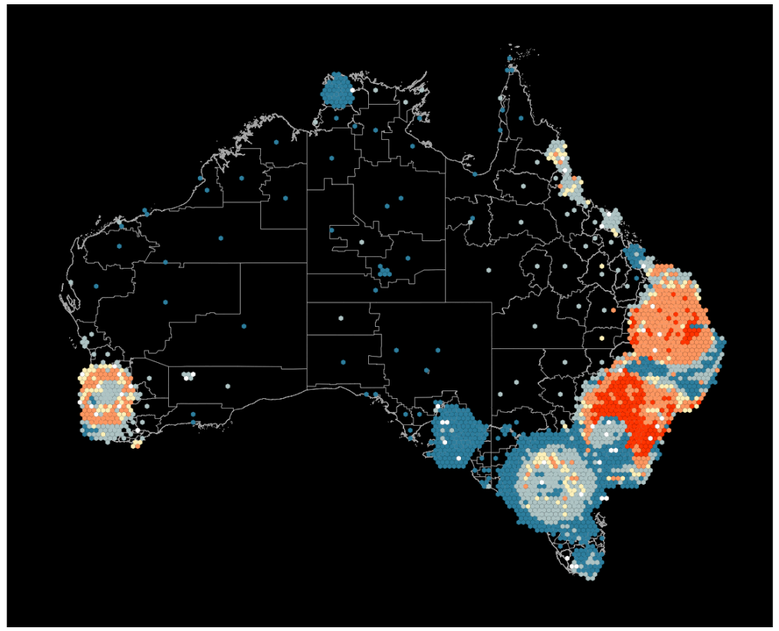
- Each student takes an exam on the content being taught. We'll call this their BEFORE score.
- Divide students into two groups. Make sure that each group has similar types of students, so both groups are as similar as possible.
- One group gets an extra traditional assignment, and the other group participates in a data competition.
- Each student takes an exam on the content being taught. We'll call this their AFTER score.
- The difference between the before and after scores are compared using side-by-side boxplots and a two-sample permutation test

What other modifications to the design can you think of?

Case study: choropleth vs hexagon tile (1/3)

The goal is to demonstrate that the *hexagon tile map is better than the choropleth* for communicating disease incidence across Australia.





The choropleth fills geographic regions (LGAs, SA2s, …) with colour corresponding to the thyroid cancer relative difference from the overall mean. The hexagons, are also filled this way.

Kobakian et al

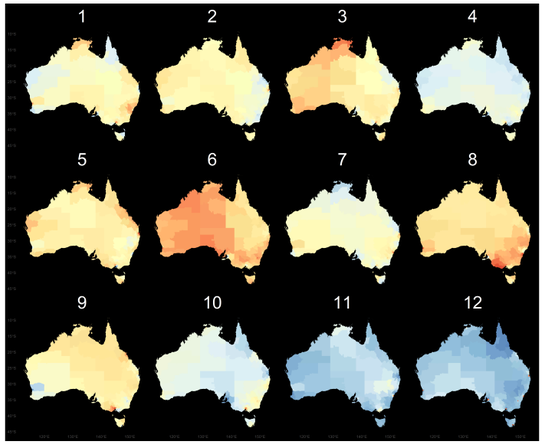
Case study: choropleth vs hexagon tile (2/3)

- Each participant can only see the (same) data once.
- Need to test for different types of spatial patterns.
- Need to repeat measure each type of pattern, and each participant.

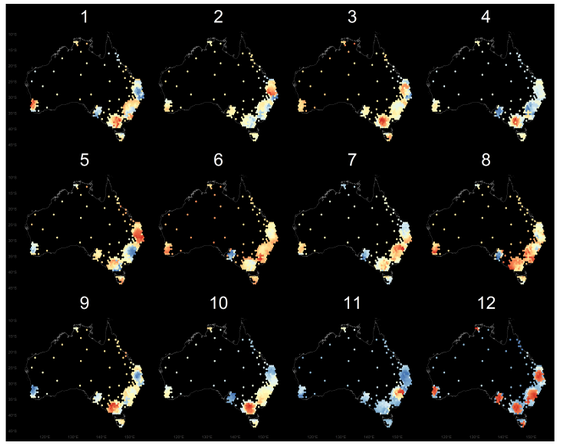
Pairing is done on the data set. Four different data sets used for each pattern.

Trial 1

Participant 1

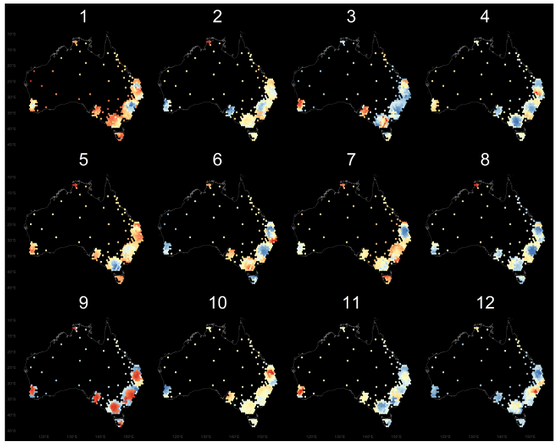


Participant 2

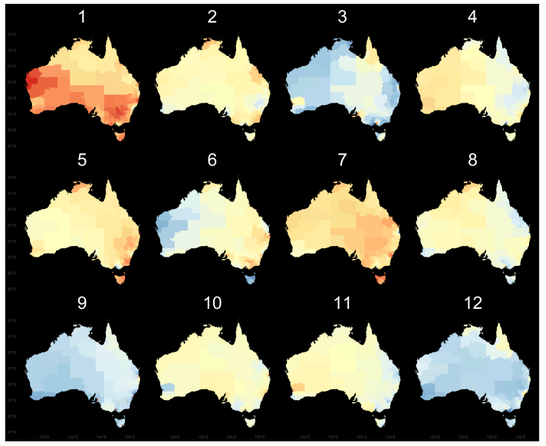


Trial 2

Participant 1



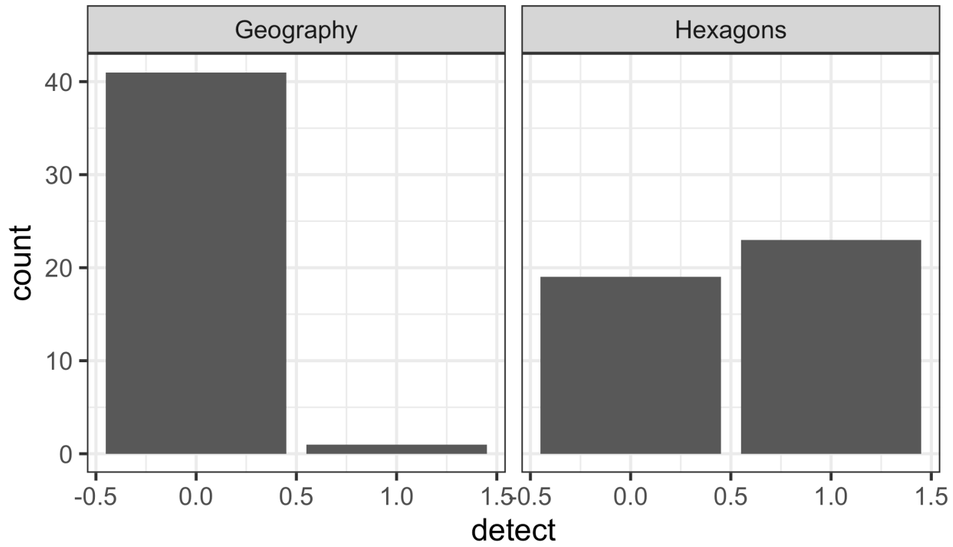
Participant 2



Case study: choropleth vs hexagon tile (3/3)

Ignore the pairing

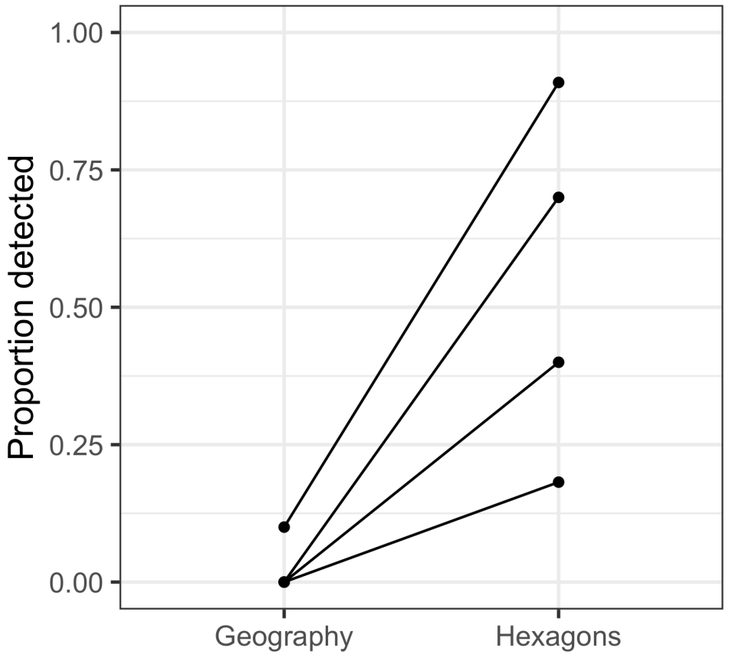
▶ Code



Looks like detection rate about 50-50 for hexagon tile map, which is better than almost zero for choropleth map.

Account for the pairing

▶ Code



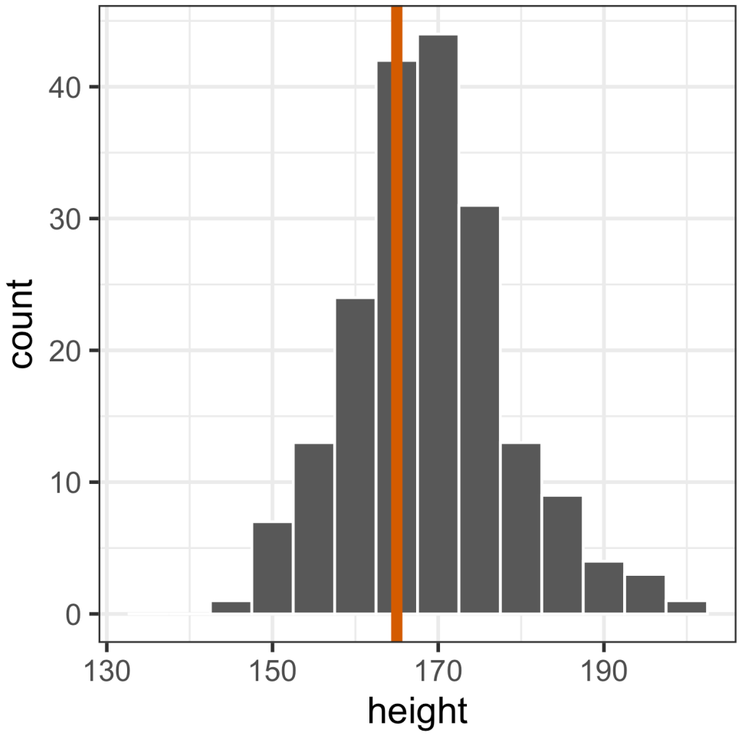
For each data set, the hexagon tile map performed better.

Normalising

Am I short?

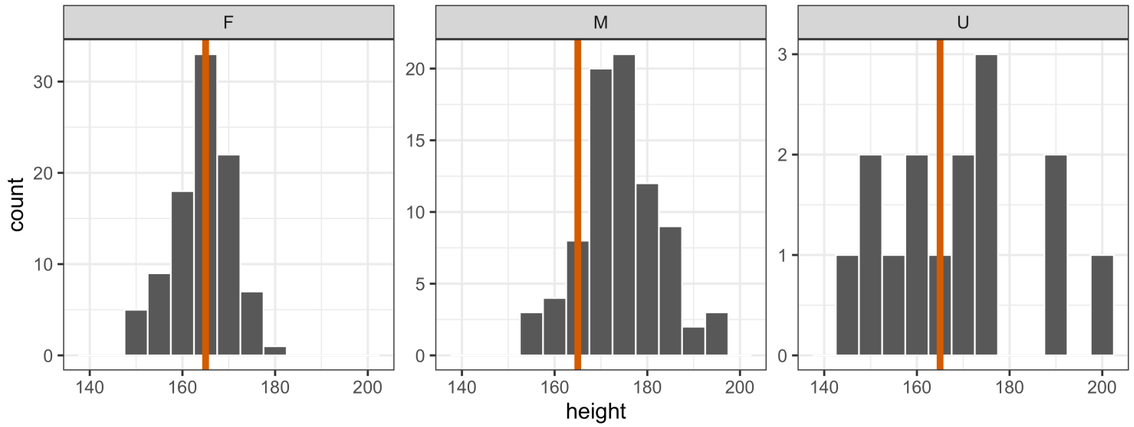
I am 165 cms tall.

▶ Code



But, there are strata in humans, so compared to what would be better?

▶ Code



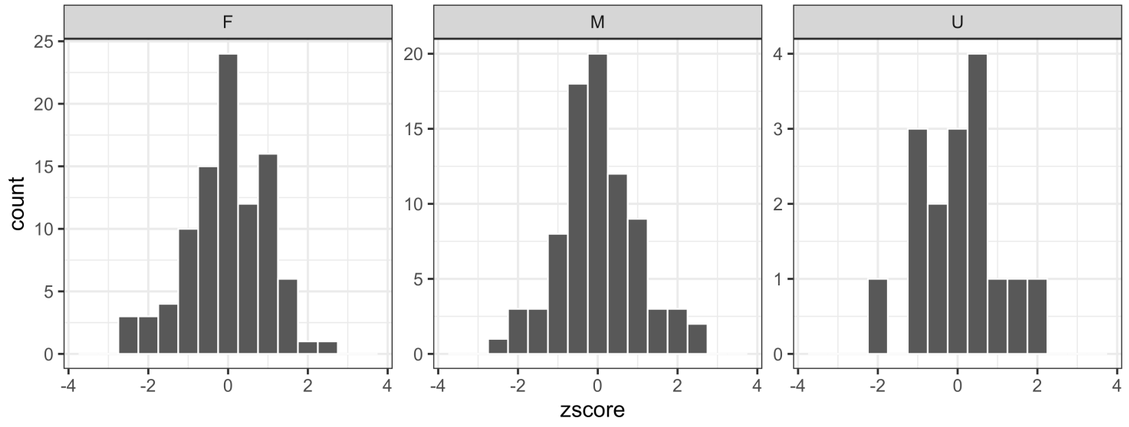
Nope, I'm average height.

Normalising

Within each strata convert values to a z-score.

$$z = \frac{x - \bar{x}}{s}$$

```
1 df22 <- df22 |>
2   group_by(sex) |>
3   mutate(zscore = (height -
4     mean(height))/sd(height))
```



- females: $\bar{x} = 164.43$, $s = 6.41$
- males: $\bar{x} = 174.23$, $s = 8.71$
- unknown: $\bar{x} = 165.74$, $s = 18.15$

My z-score is 0.09.

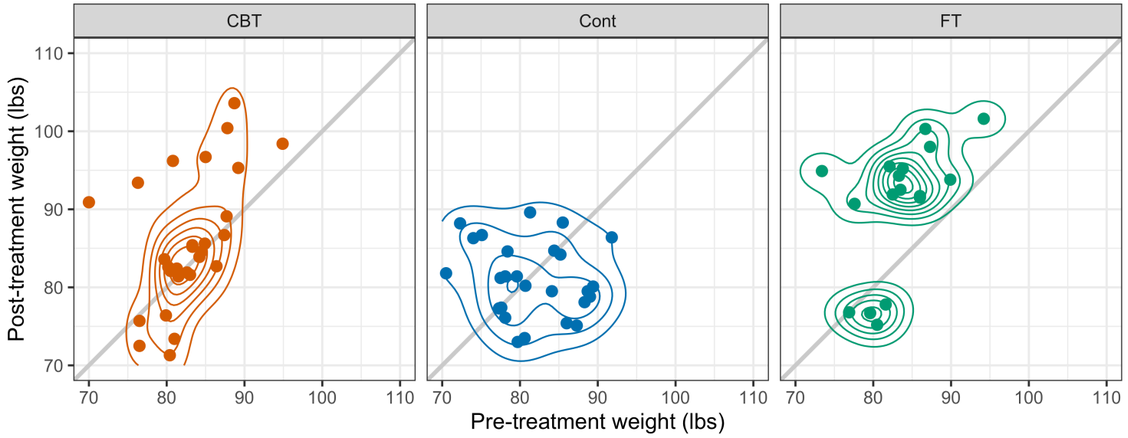
Rob's height is 170 cms. His z-score is -0.49.

I am relatively TALLER than Rob.

Baselines

Relative to a baseline

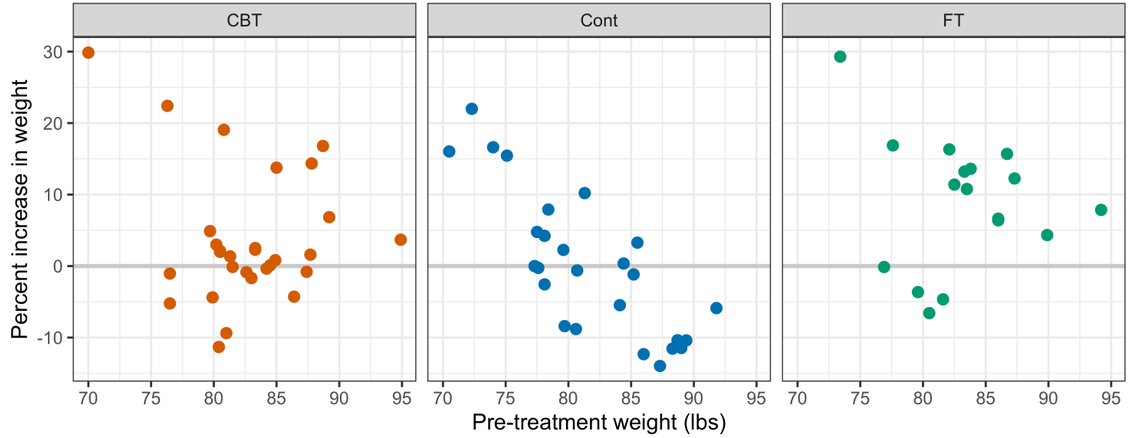
▶ Code



- Primary comparison is before treatment weight, and after treatment weight.
- Three different treatments.

Unwin, Hofmann and Cook (2013)

▶ Code

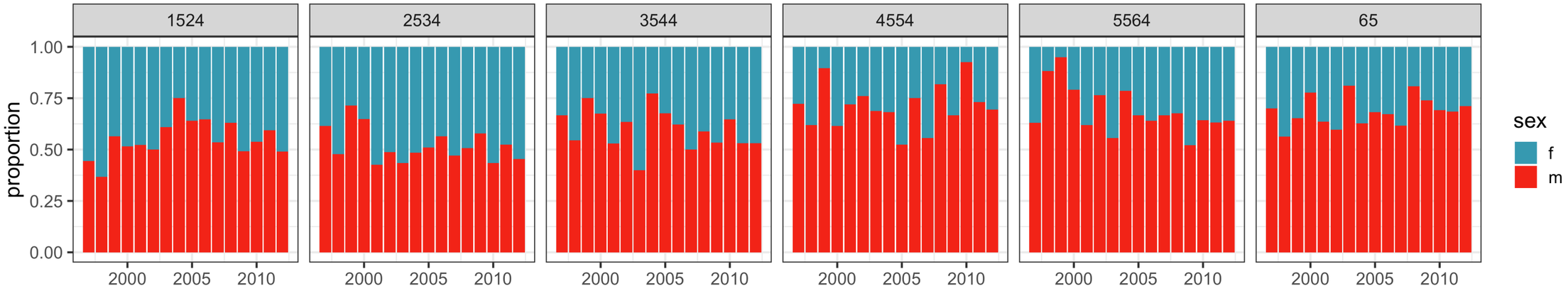


- Compute the difference
- Compare difference relative to before weight
- Before weight is used as the baseline

Proportions

Case study: tuberculosis (1/3)

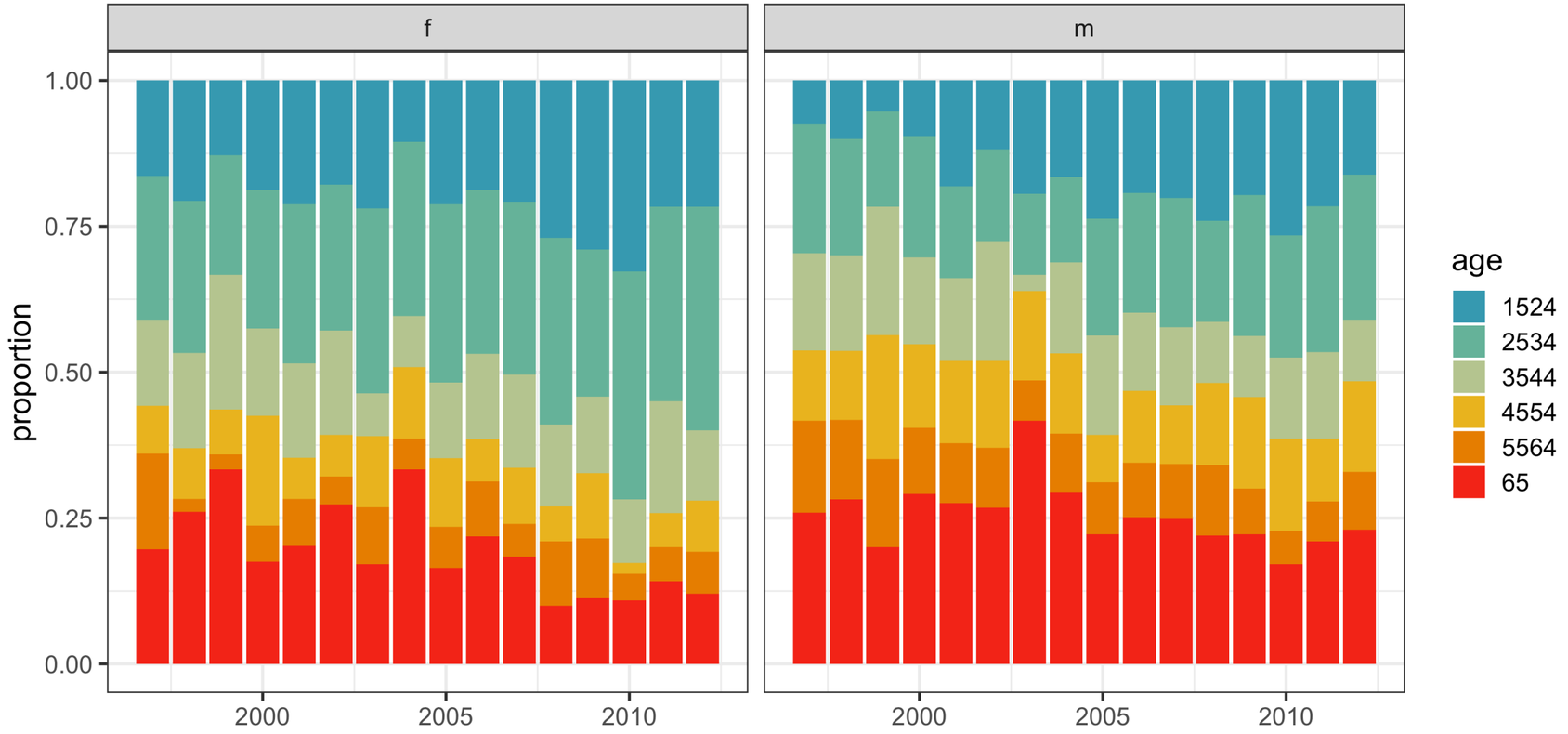
▶ Code



Primary comparison is `sex`, relative to yearly trend.

Case study: tuberculosis (2/3)

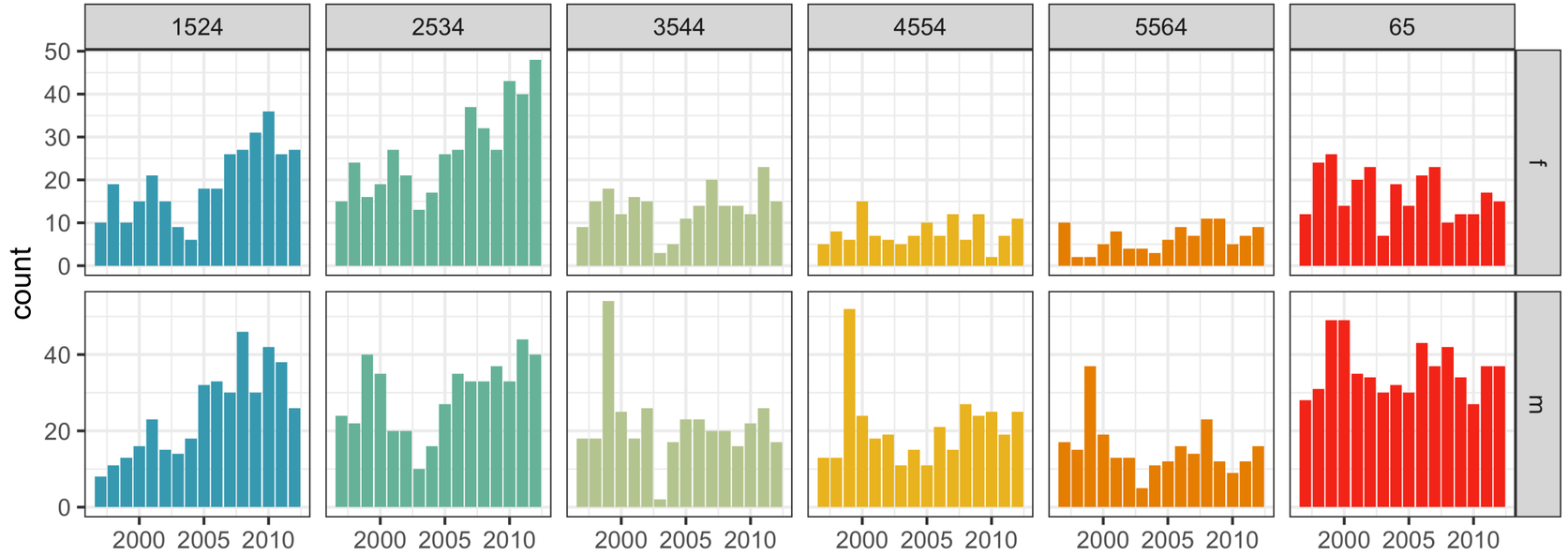
▶ Code



Primary comparison is age, relative to yearly trend.

Case study: tuberculosis (3/3)

- Code



Primary comparison is `year` trend, separately for age and sex.

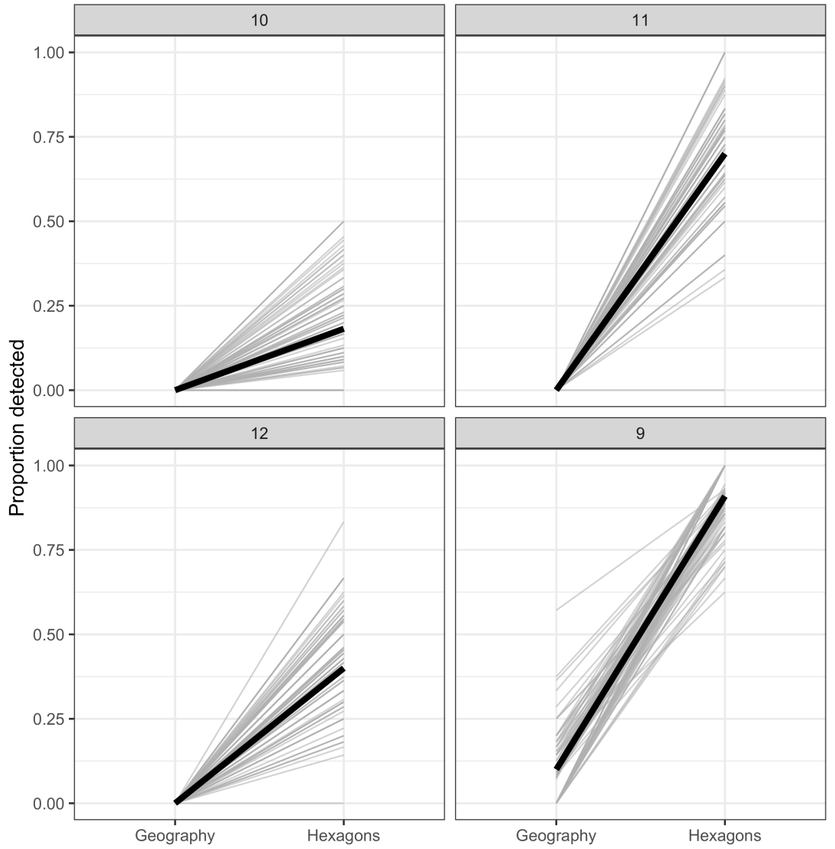
Inference

Bootstrap confidence intervals

data R

- Confidence intervals show what might happen to estimates with different samples,
- with the same dependence structure.
- Sample the current sample, but don't change anything else.
- Reason for sampling with replacement, is to keep sample size the same - we know that variance decreases with smaller sample size.

For choropleth vs hexagon tiles, sample participants with replacement.

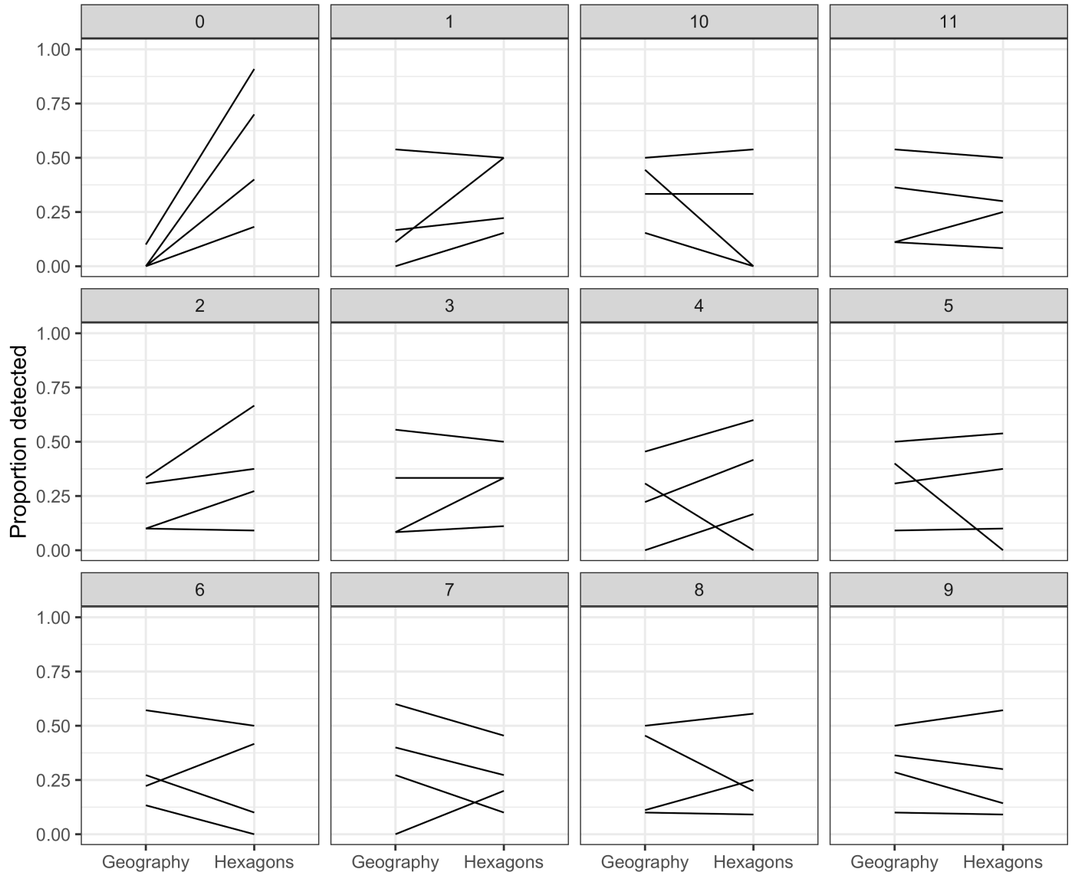


Lineups

- Lineups show what might happen to estimates with null samples, where (by construction) there is no relationship.
- Thus you need to break dependence structure.

For choropleth vs hexagon tiles, randomise the type of plot each participant received. This breaks any dependence between type and detection rate.

▶ Code



Take-aways

- In *comparison to what* is especially important for exploring observational data.
- Avoid reporting spurious associations by accounting for dependencies correctly.
- Adjust for individual variation, by
  - pairing (or multiple repeated measures)
  - relative to a baseline
  - on a standard scale
- Determining in *comparison to what* can be hard.

Resources

- Wilke (2019) Fundamentals of Data Visualization Chapters 9, 10, 11
- Unwin (2015) Graphical Data Analysis with R Chapter 10