ETC5512: Wild Caught Data

Week 8

Synthetic and Simulated Data

Lecturer: *Emi Tanaka*

Department of Econometrics and Business Statistics

ETC5512.Clayton-x@monash.edu

13th May 2020

What is PISA?

- The Programme for International Student Assessment (PISA) is a triennial survey conducted by the Organization for Economic Cooperation and Development (OECD) on assessment measuring 15-year-old student performances in reading, mathematics and science.
- The goal of the PISA survey is to assess the workforce readiness of 15-year old students and used as a global metric for quality, equity and efficiency in school education.
- In 2018, PISA involved 79 countries and economies with assessment of about 600,000 students worldwide as a sample of 32 million 15-year olds in school.
- One domain is tested in detail for every PISA. In 2018, this was *reading* with mathematics and science as minor areas of assessment.

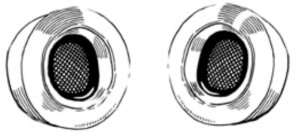
PISA 2018 Assessment and Analytical Framework

- **Reading literacy** is defined as students' capacity to understand, use, evaluate, reflect on and engage with texts in order to achieve one's goals, develop one's knowledge and potential, and participate in society.
- **Mathematics literacy** is defined as students' capacity to formulate, employ and interpret mathematics in a variety of contexts. It includes reasoning mathematically and using mathematical concepts, procedures, facts and tools to describe, explain and predict phenomena.
- **Science literacy** is defined as the ability to engage with science-related issues, and with the ideas of science, as a reflective citizen. A scientifically literate person is willing to engage in reasoned discourse about science and technology, which requires the competencies to explain phenomena scientifically, evaluate and design scientific enquiry, and interpret data and evidence scientifically.

PISA 2018 Assessment

- Assessments are mostly computer-based that lasts a total of 2 hours.
- The questions comprise a mixture of multiple choice and free entry.
- Different students may have different set of questions.
- Reading was tested for 1 hour and other topics for the remaining 1 hour.
- You can find an example of the test questions here.

MP3 PLAYERS

Music City MP3 Specialists		
MP3 player	Headphones	Speakers
		
155 zeds	86 zeds	79 zeds

Download the data from

http://www.oecd.org/pisa/data/2018database/

SPSS (TM) Data Files (compressed)
Student questionnaire data file

- The file is 494 MB so it will take a while to download.
- Keep a local copy for later use.

Data in proprietary formats

- The PISA data are provided in proprietary formats (SAS and SPSS).
- This means that the data are stored in a particular encoding scheme, designed so that decoding and reading the data is accomplished by particular software or hardware.
- In R, you can use the haven package to import the PISA data.

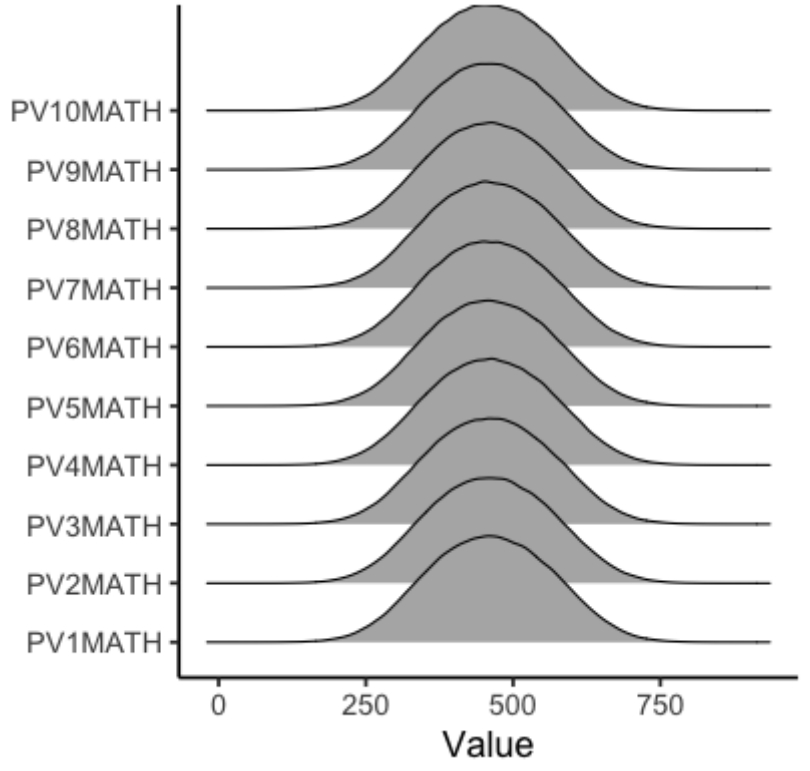
```
library(tidyverse)
library(haven)
pisa2018 <- read_sav(here::here("data", "CY07_MSU_STU_QQQ.sav")) %>%
                as_factor()  # swap code and labels for labelled factors
dim(pisa2018)

## [1] 612004   1118
```

- Since the data is big, it will take a while to read the data in.
- Every row corresponds to a student.

Domain assessment scores

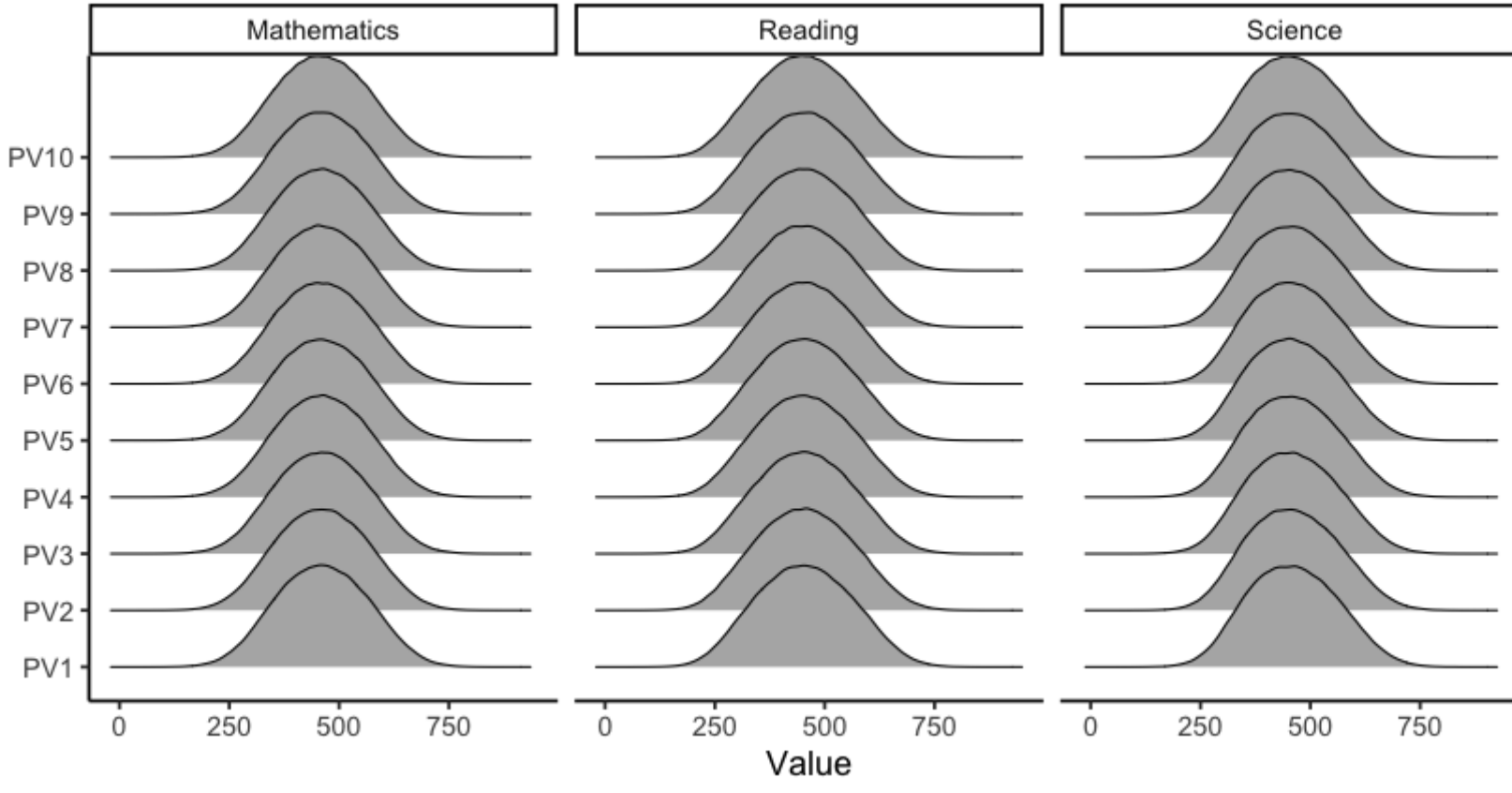
- PV1MATH = Plausible Value 1 in Mathematics
- PV1READ = Plausible Value 1 in Reading
- PV1SCIE = Plausible Value 1 in Science

```
pisa2018 %>%
  select(PV1MATH:PV10MATH) %>%
  pivot_longer(PV1MATH:PV10MATH,
               names_to = "Number",
               values_to = "Value") %>%
  # reorder factor so it is PV1MATH, ..., PV10MATH
  mutate(Number = fct_reorder(Number, Number,
         function(x) unique(parse_number(x)))) %>%
  ggplot(aes(x = Value, y = Number)) +
  labs(y = "") +
  ggridges::geom_density_ridges() +
    theme_classic(base_size = 18)
```



Domain score distribution by plausible value number

- Wait... is it too perfect?
- There are no outliers or unusual characteristics for the values.
- Also why are there 10 values?



What are "plausible values"?

- School assessments are typically concerned with accurately assessing **individual performance** for the purpose of *diagnosis*, *selection* or *ranking*.
- The goal of PISA is to compare the skills and knowledge of 15-year-old students across countries and economies.
- PISA supplies data for individual students but the assessment values are *not* raw data.
- The raw data are first quality checked and then used for **scaling** and **population modelling**.
- In brief, the **plausible values are generated from a model** that *capture sub-population or population characteristics*.
- Hence why the PISA data do not display individual characteristics.
- Thus PISA data should not be used to make precise inferences about individuals' domain performances.

Brief technical explanation of "plausible values"

1. *Item response theory scaling* of the cognitive responses estimates the item parameters that provide comparable latent scales across countries and cycles for each domain
2. *Multivariate latent regression* is fitted using item parameters estimates from 1.
3. For each student and each domain, *10 plausible values are drawn from posterior distribution* using the estimated model parameters in 2.
   - This is the gist of how the values are generated but the *technical details are beyond the scope of this course*.
   - For those interested, you can find detailed technical explanation from PISA 2018 Technical Report Chapter 9 Scaling PISA Data.

PISA Data Management

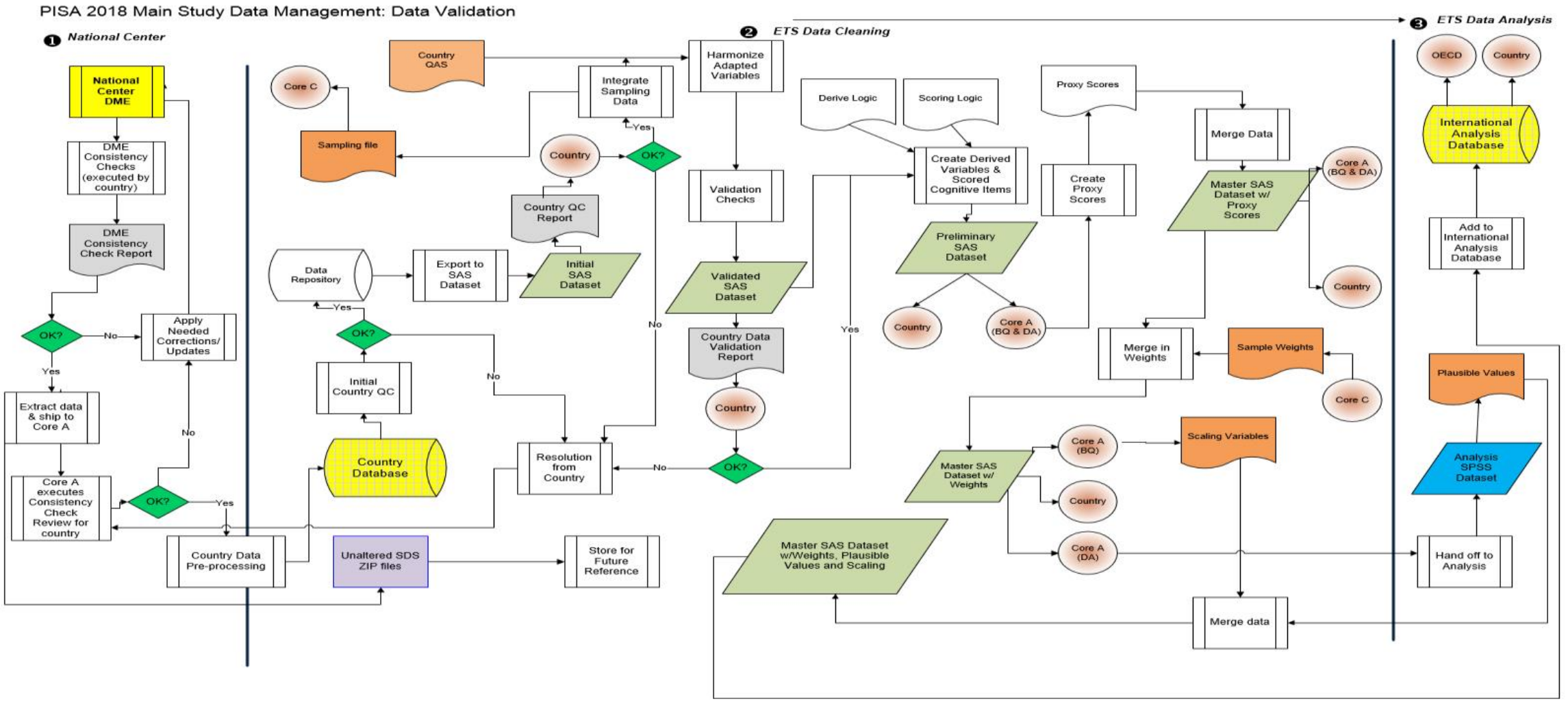


Diagram from PISA 2018 Technical Report Chapter 10 Data Management

Examining the gender gap across countries

This section is based on upcoming book by Hofmann, Cook, Vanderplas and Wang.

Are girls worse in maths than boys?

- The gender gap in mathematics is a common discussion, with the concern being that girls tend to score lower than boys on average in standardized math tests.
- The PISA data provides an opportunity to explore the gender gap across numerous countries.
- In the `pisa2018` data, the sex of the student is in variable `ST004D01T` and the country/region is in variable `CNT`.
- Let's rename these to sensible names, e.g. `sex` and `country`.
- We'll also modify some country names so that it can be joined with the map data later.
- We will focus on using `PV1MATH` and will not cover any analysis that require us to use all 10 plausible values in this course.

Code to clean PISA data

```
pisa2018c <- pisa2018 %>%
  rename(sex = ST004D01T, country = CNT) %>%
  filter(!is.na(sex)) %>% # filter two Canadian students where sex is missi
  filter(!is.na(PV1MATH)) %>%  # Vietnam is missing scores
  mutate(country = case_when(
    country == "Brunei Darussalam" ~ "Brunei",
    country == "United Kingdom" ~ "UK",
    country %in% c("Hong Kong", "B-S-J-Z (China)") ~ "China",
    country == "Korea" ~ "South Korea",
    country == "North Macedonia" ~ "Macedonia",
    country == "Baku (Azerbaijan)" ~ "Baku",
    country %in% c("Moscow Region (RUS)", "Tatarstan (RUS)",
                   "Russian Federation") ~ "Russia",
    country == "Slovak Republic" ~ "Slovakia",
    country == "Chinese Taipei" ~ "Taiwan",
    country == "United States" ~ "USA",
    TRUE ~ as.character(country)))
```

⚠ Plot 1: Gender difference in math scores by country

```
pisa2018c %>%
  group_by(sex, country) %>%
  summarise(avg = mean(PV1MATH)) %>%
  ungroup() %>%
  pivot_wider(country, names_from = sex,
              values_from = avg) %>%
  mutate(diff = Female - Male,
         country = fct_reorder(country, diff)) %>%
  ggplot(aes(x = diff, y = country)) +
  geom_point() +
  geom_vline(xintercept = 0, color = "red") +
  labs(y = "Country",
       x = "Difference in mean PV1 (girl - boy)") +
  theme_bw(base_size = 14)
```

👾 But wait how is the data collected?

Assessment Design

Sourced from PISA 2018 Integratated Design. Scroll down to see more information.

GROUP 1 – CBA Trend

FUO (Forms 01-18, 67-78)

Forms	Cluster 1	Cluster 2	Cluster 3	Cluster 4
01	S1	S4	M1	M2
02	S3	S6	M3	M4
03	S5	S2	M5	M6ab
04	M2	M3	S2	S5
05	M4	M5	S4	S1
06	M6ab	M1	S6	S3
07	M1	M4	R1	R2
08	M3	M6ab	R3	R4
09	M5	M2	R5	R6ab
10	R2	R3	M2	M5
11	R4	R5	M4	M1
12	R6ab	R1	M6ab	M3
13	R1	R4	S1	S2
14	R3	R6ab	S3	S4
15	R5	R2	S5	S6
16	S2	S3	R2	R5
17	S4	S5	R4	R1
18	S6	S1	R6ab	R3
67	M1	R1	FL1	FL2
68	R2	M2	FL1	FL3
69	M3	R3	FL2	FL1
70	R4	M4	FL2	FL3
71	M5	R5	FL3	FL1

GROUP 2 – CBA Trend/New R

VUO (Forms 19-42)

Forms	Cluster 1	Cluster 2	Cluster 3	Cluster 4
19	R1	R14	R12	R7
20	R2	R16	R14	R9
21	R3	R18	R16	R11
22	R4	R8	R18	R13
23	R5	R10	R8	R15
24	R6ab	R12	R10	R17
25	R13	R1	R10	R9
26	R15	R2	R12	R11
27	R17	R3	R14	R13
28	R7	R4	R16	R15
29	R9	R5	R18	R17
30	R11	R6ab	R8	R7
31	R11	R18	R1	R8
32	R13	R8	R2	R10
33	R15	R10	R3	R12
34	R17	R12	R4	R14
35	R7	R14	R5	R16
36	R9	R16	R6ab	R18
37	R16	R17	R15	R1
38	R18	R7	R17	R2
39	R8	R9	R7	R3
40	R10	R11	R9	R4
41	R12	R13	R11	R5

GROUP 3 – CBA New R/GC

FUO (Forms 43-66)

Forms	Cluster 1	Cluster 2	Cluster 3	Cluster 4
43	R7	R8	R10	R14
44	R8	R9	R11	R15
45	R9	R10	R12	R16
46	R10	R11	R13	R17
47	R11	R12	R14	R18
48	R12	R13	R15	R7
49	R13	R14	R16	R8
50	R14	R15	R17	R9
51	R15	R16	R18	R10
52	R16	R17	R7	R11
53	R17	R18	R8	R12
54	R18	R7	R9	R13
55	R7	R13	GC1	GC2
56	R8	R14	GC2	GC3
57	R9	R15	GC3	GC4
58	R10	R16	GC4	GC1
59	R11	R17	GC1	GC3
60	R12	R18	GC2	GC4
61	GC1	GC2	R13	R8
62	GC2	GC3	R14	R9
63	GC3	GC4	R15	R10
64	GC4	GC1	R16	R11
65	GC3	GC1	R17	R12

Interrogating the data

- So students who have BOOKID as Form 1-12 or 67-78 would have had mathematics component in their test.

```
pisa2018 %>%
  filter(BOOKID == "Form 13") %>%
  select(CNT, ST004D01T, BOOKID, PV1MATH)
```

```
## # A tibble: 20,511 x 4
##    CNT     ST004D01T BOOKID  PV1MATH
##    <fct>   <fct>     <fct>     <dbl>
##  1 Albania Female    Form 13    417.
##  2 Albania Male      Form 13    585.
##  3 Albania Female    Form 13    354.
##  4 Albania Male      Form 13    424.
##  5 Albania Female    Form 13    451.
##  6 Albania Female    Form 13    351.
##  7 Albania Female    Form 13    257.
```

- But there is a mathematics score for students who did not even sit a test with mathematics component!
- You will compare the math gender gap with all students vs. the subset of students who did sit the mathematics component during the tutorial 🔧

Sample survey data

- PISA data is collected from a complex multi-stage design which results in *different* inclusion probabilities of certain student/school characteristics.
- For example, in Australia, all indigeneous students (the minority group) are asked to participate.
- PISA data comes with two sets of weights:
  - Final student weights (`W_FSTUWT`). These scale the sample up to the size of the population within each country. If the unit of interest is the population of students within subset of countries, use this.
  - Senate weights (`SENWT`). These weights sum up to the same constant value, therefore each country will contribute equally to the analysis. If the unit of interest is the countries then use this.
- Without applying weights, students or schools with particular characteristics may be either under/over represented within the analysis.

Jerrim et al. (2017) "What Happens When Econometrics and Psychometrics Collide? An Example Using the PISA Data." Economics of Education Review 61 (December): 51–58.

Why accounting for sampling weights is important

- Suppose we have a class of 24 students with 20 boys and 4 girls.

7 2 2 6 2 5 4 9 2 7 5 1 7 0 3 2 4 1 4 5 8 7 5 1

The population average of this class is 4.12 with 3.9 for boys and 5.25 for girls.

- If we randomly select 6 students to participate in the survey, we expect 5 boys and 1 girl on average (Selected boys are 2, 3, 5, 7, 16 and girl is 2). The sample average score of selected boys is 2.4 and girls is 7, and total average is 3.17.
- But having equal number of boys and girls in the survey is important then the inclusion probability for a boy is 3/20 while for a girl is 3/4. (Now say selected boys are 5, 11, 20 and girls are 1, 2, 4). The sample average score of selected boys is 4 and girls is 5.33, and total sample average is 4.67.
- The sample average score (4.67) is higher than it should be due to over-representation of the girls in the sample.

Taking into weights into account

- In this case, the sampling weights are the inverse of the inclusion probability (20/3 for boys and 4/3 for girls).
- A weighted mean, $\hat{\mu}$, for values $x_1, \ldots, x_n$ with corresponding weights $w_1, \ldots, w_n$ is computed as

$$\hat{\mu} = \frac{1}{\sum_{i=1}^{n} w_i} \sum_{i=1}^{n} w_i x_i.$$

- So the class population mean can be estimated as

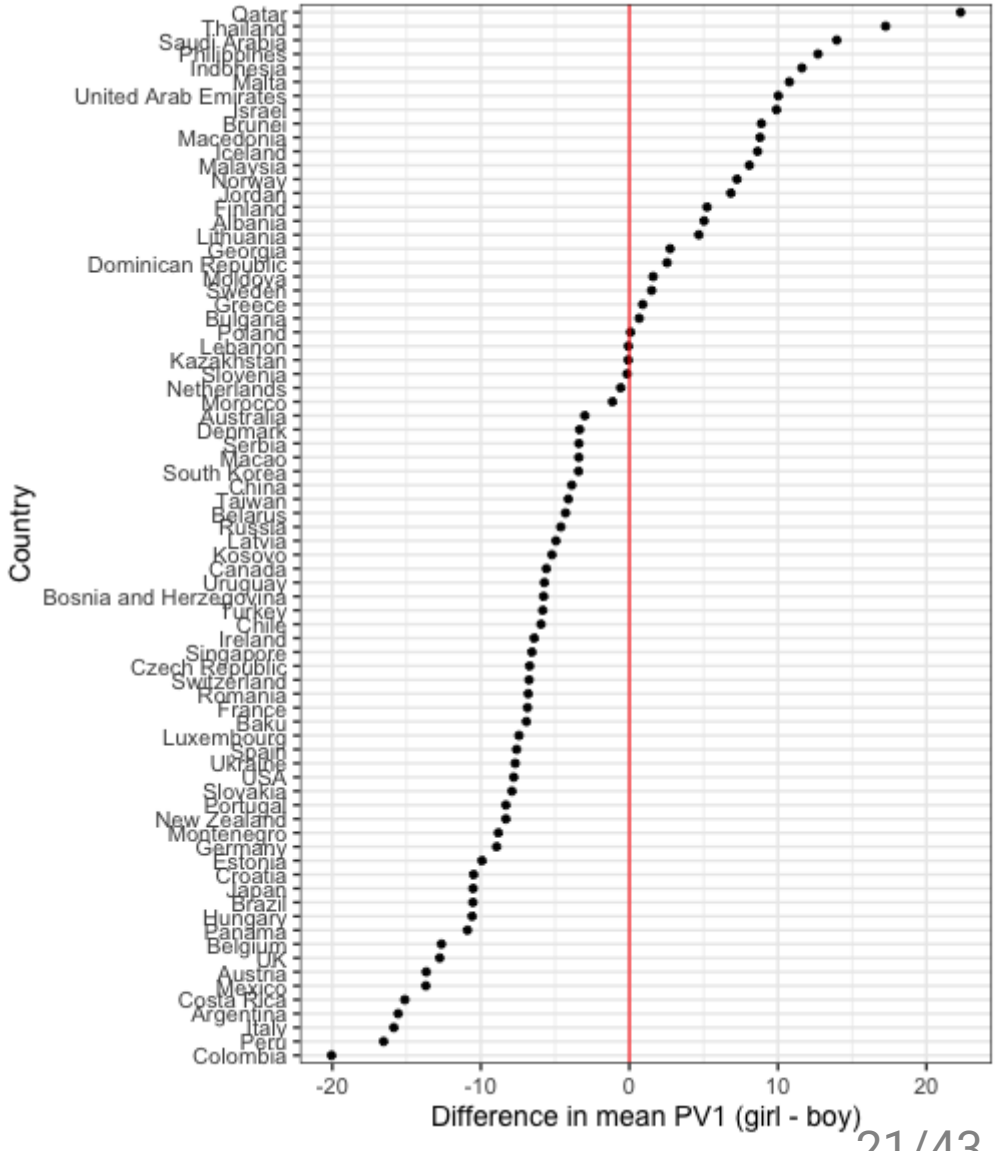
$$\frac{20/3 \times 4 + 4/3 \times 5.3333333}{20/3 + 4/3} = 4.2222222.$$

Notice that the estimate is closer to the class population mean.

- Or you can use the `weighted.mean` function in R.

⚠ Plot 2: Gender difference in math scores by country

```
mathdiff_df <- pisa2018c %>%
  group_by(sex, country) %>%
  summarise(math = weighted.mean(PV1MATH,
                                 w = SENWT)) %>%
  ungroup() %>%
  pivot_wider(country, names_from = sex,
              values_from = math) %>%
  mutate(diff = Female - Male,
         country = fct_reorder(country, diff))

ggplot(mathdiff_df, aes(x = diff, y = country)) +
  geom_point() +
  geom_vline(xintercept = 0, color = "red") +
  labs(y = "Country",
       x = "Difference in mean PV1 (girl - boy)") +
  theme_bw(base_size = 14)
```

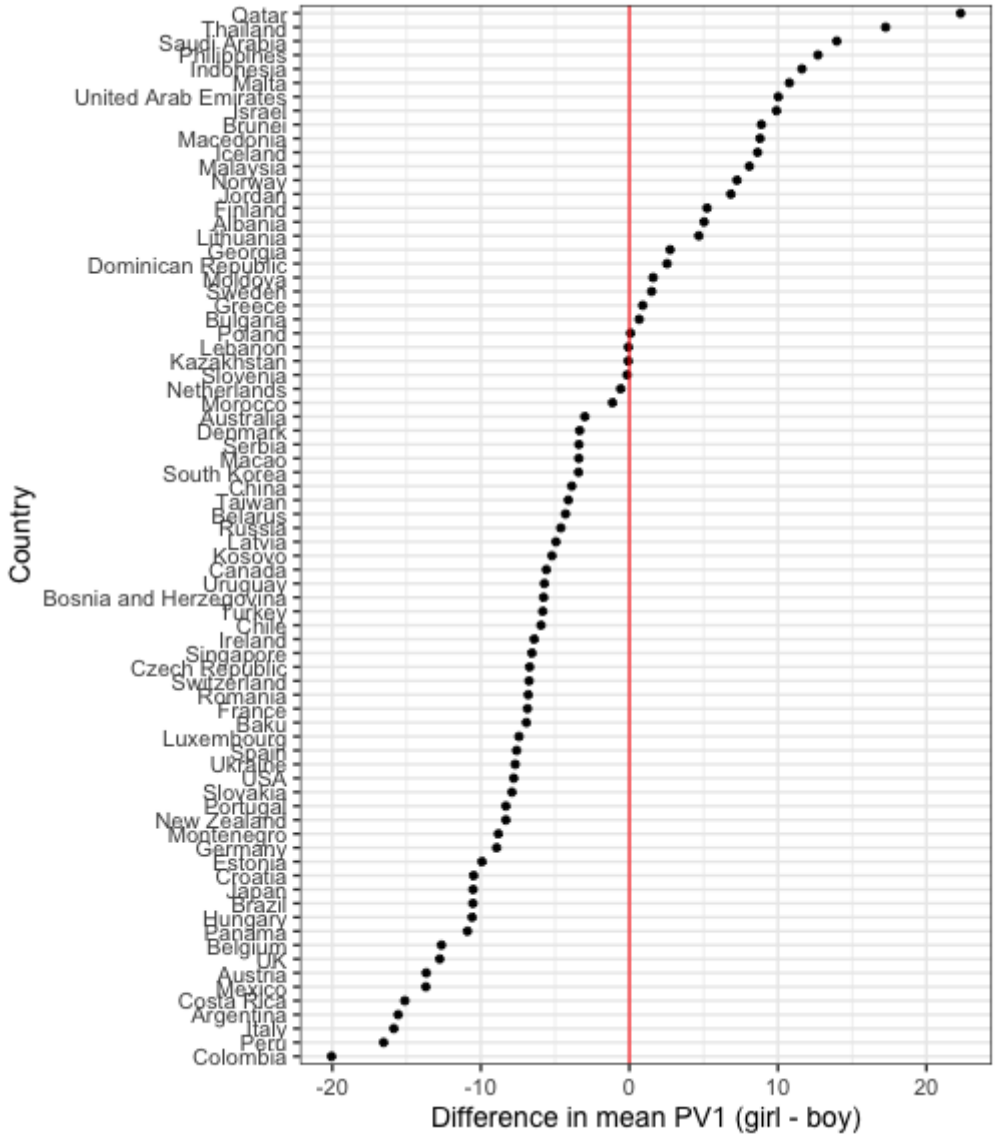


Mapping the math score differences by gender

```
map_data("world") %>% # function from ggplot2
  left_join(mathdiff_df, by = c("region" = "country")) %>%
  ggplot(aes(long, lat, group = group, fill = diff)) +
   geom_polygon(color = "black") + theme_void(base_size = 18) +
      scale_fill_gradient2("Math Gap", na.value="grey90",
                           low="#1B9E77", high="#D95F02", mid="white")
```

Bootstrap for estimating variance

Caution point estimates



- Non-zero point estimate does not mean that there is a *significant* difference in performance for mathematics by gender!
- There is uncertainty for every estimate (and prediction).
- The plot we saw before will be more useful if we plot the error bar, that represents the uncertainty, for each point estimate.
- But how do we calculate this uncertainty?

Bootstrap to measure uncertainty

- Bootstrap is a technique that uses **random sampling with replacement** of data to obtain properties of an estimator.
- Bootstrap is relatively **simple to apply**.
- Bootstrap can be **computationally expensive** as it requires a large number of times for the process to be applied (we will do 100 times but at least 200 times is recommended).
- In bootstrapping, it is assumed that the **observations are independent (or independent within blocks)**.
- For assessing the gender gap in mathematics scores across countries, we will be *resampling within country and gender*.
- The resampling process should generate data with the same number of observations as the original data.

Bootstrap using R: Part 1

There are a number of ways of doing this in R but we will use `sample_n` function in `dplyr` 📦.

```
set.seed(2020) # for reproducibility
boot_sample1 <- pisa2018c %>%
  group_by(country, sex) %>%
  sample_n(size = n(), replace = TRUE)
```

We can then treat `boot_sample1` as we did before to obtain another set of estimates for the gender gap for mathematics score by country.

```
boot_sample1 %>%
  summarise(avg = weighted.mean(PV1MATH, SENWT)) %>%
  ungroup() %>%
  pivot_wider(country, names_from = sex, values_from = avg) %>%
  mutate(diff = Female - Male, country = fct_reorder(country, diff))
```

Bootstrap using R: Part 2

- We need to repeat this process a reasonable number of times.
- We will do 100 times.
- To make this process easier, we will use the `map_dfr` function from `purrr` 📦.

```
boot_ests <- map_dfr(1:100, ~{
    pisa2018c %>%
      group_by(country, sex) %>%
      sample_n(size = n(), replace = TRUE) %>%
      summarise(avg = weighted.mean(PV1MATH, SENWT)) %>%
      ungroup() %>%
      pivot_wider(country, names_from = sex, values_from = avg) %>%
      mutate(diff = Female - Male, country = fct_reorder(country, diff)) %>%
      mutate(boot_id = .x)
  })
```

- The `.x` is substituted from an element from the first argument in `map_dfr`.

Bootstrap using R: Part 3

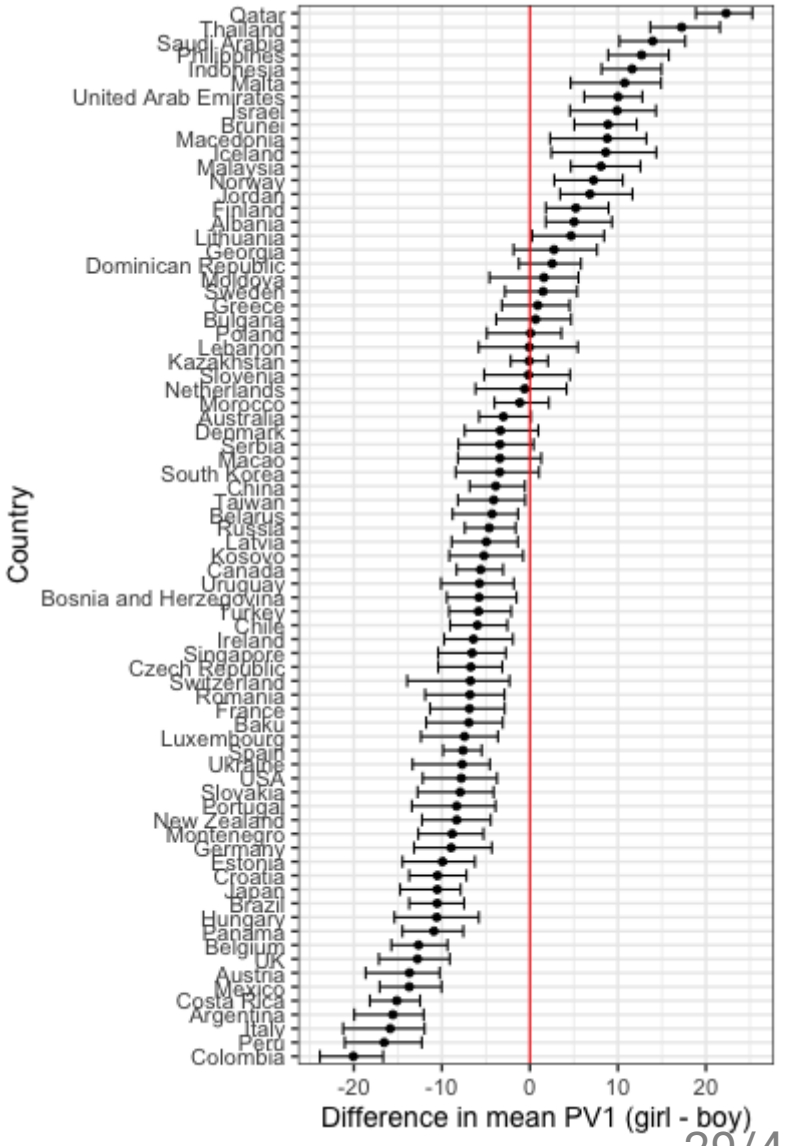
- We obtain a 90% confidence interval of mean differences by gender in mathematics score for each country by taking the 5% and 95% quantiles from the bootstrap estimates.
- There are many other ways to do this. Week 8 lab solution will show a different way.

```
mathdiff2_df <- boot_ests %>%
  group_by(country) %>%
  summarise(lower = sort(diff)[5],
            upper = sort(diff)[95]) %>%
  left_join(mathdiff_df, by = "country") %>%
  mutate(country = fct_reorder(country, diff))
```

Plot 3: Gender difference in math scores by country

- A better plot is then to draw this 90% confidence interval on the plot.

```
ggplot(mathdiff2_df, aes(diff, country)) +
  geom_point() +
  geom_errorbar(aes(xmin = lower,
                    xmax = upper)) +
  geom_vline(xintercept = 0, color = "red") +
  labs(y = "Country",
       x = "Difference in mean PV1 (girl - boy)") +
  theme_bw(base_size = 14)
```



Simulated data

"Real" and "Fake" Data

- **Real data** are data where observations are direct measurements from real-world phenomena.
- **Synthetic data** are data where observations are artificially generated with similar statistical properties to the real data.
- **Simulated data** are data where observations (and covariates) are simulated from a model.

- Data by default are assumed to be "real" so there is generally no need to explicitly refer data as real data.
- The terms synthetic and simulated data may be used interchangeably in some literature. Synthetic data may be simulated data and vice versa.
- The purpose of simulated data is to often study a statistical method and is commonly used in statistical literature (both in teaching and research).

Data Generating Process: Simple Linear Model

- Unlike typical data, the **data generating process is *known*** for simulated data.
- For example, we can generate a set of observation that is only linearly dependent on an independent continuous variable.
- Mathematically, the data generating process is a simple linear model:

$$y_i = \beta_0 + \beta_1 x_i + e_i,$$

where for $i = 1, \ldots, n$, y_i is the i-th response, x_i is the corresponding covariate, e_i is a random error, β_0 is the intercept and β_1 is the slope.

- Typically we assume that e_is are independent and $e_i \sim N(0, \sigma^2)$.
- In practice, we only observe those colored in blue and we estimate those colored in red (for simplicity assuming the data generating process is known but the model parameters are unknown).

Simulated Data in R: Simple Linear Model

Suppose that $n = 200, \beta_0 = 3, \beta_1 = -2$ and $\sigma^2 = 1$.

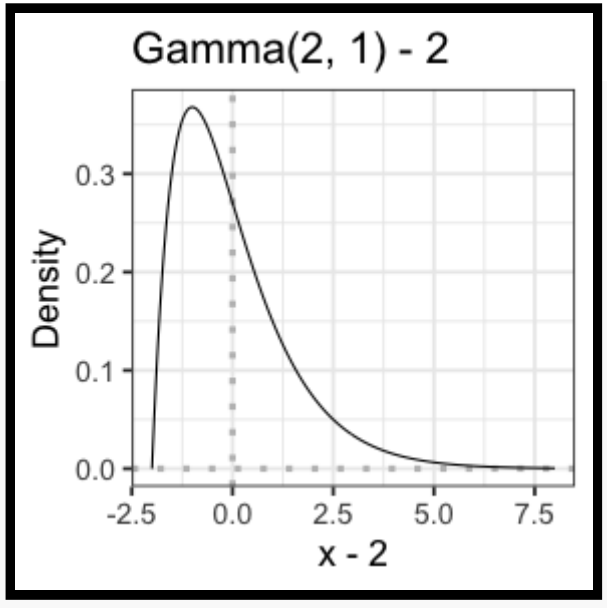
```
set.seed(2020) # for reproducibility
n <- 200 # sample size
b0 <- 3  # intercept
b1 <- -2 # slope
sim_df <- tibble(id = 1:n) %>%  # initialise data set
    mutate(x = runif(n(), 0, 10), # draw x from Uniform[0,10]
           y = b0 + b1 * x + rnorm(n(), 0, 1))
```

Obtain least squares estimates (or maximum likelihood estimate) for β_0 and β_1:

```
fit1 <- lm(y ~ x, data = sim_df)
coef(fit1)

## (Intercept)           x
##    2.975177   -2.020419
```

Simulated Data in R: Simple Linear Model

- Suppose now that $e_i \sim \text{Gamma}(2, 1) - 2$.
- How good are the estimates under least squares when the error is not normally distributed?




```
sim2_df <- tibble(id = 1:n) %>%
    mutate(x = runif(n(), 0, 10),
           y = b0 + b1 * x + rgamma(n(), 2, 1) - 2)
fit2 <- lm(y ~ x, data = sim2_df)
coef(fit2)

## (Intercept)           x
##    2.818092   -1.997902
```

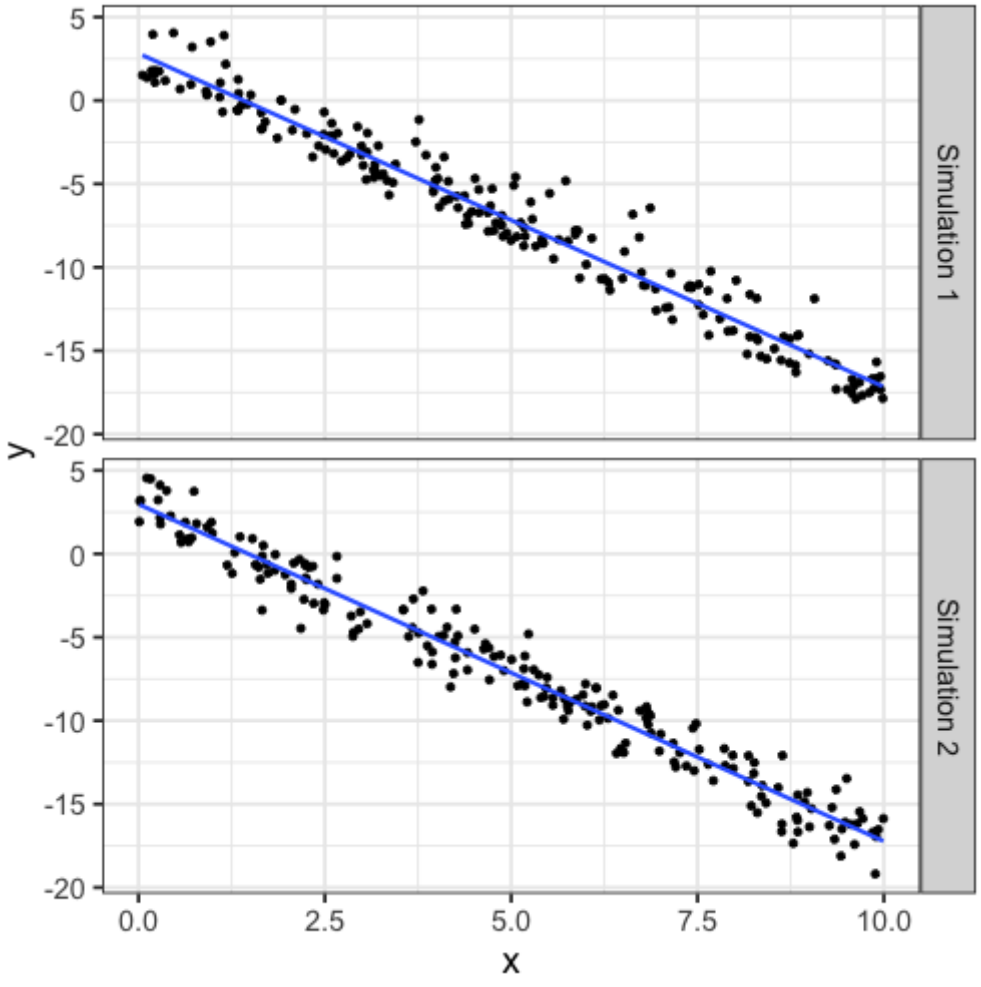
- The estimate of the slope is still good but the estimate of the intercept is not as good as before.

Visual Inference

Assessing estimates based on a scatter plot

- We could also superimpose the least squares fit onto the scatter plot.

```
combined_df <- sim2_df %>%
  mutate(sim = "Simulation 1") %>%
  rbind(mutate(sim_df,
               sim = "Simulation 2"))
ggplot(combined_df, aes(x, y)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  facet_grid(sim ~ .) +
  theme_bw(base_size = 18)
```

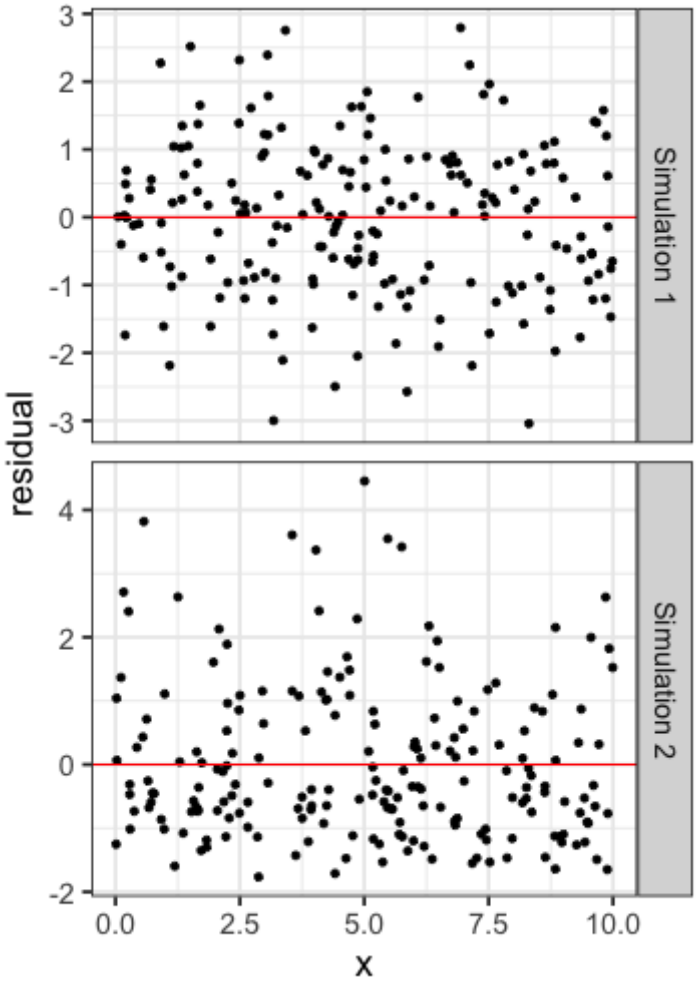


- We could also do a formal statistical test for the slope (and intercept) but we have to validate our assumption.

Assessing model assumption based on the residual plot

- We assume that $e_i \sim N(0, \sigma^2)$.
- We assess the residual plot to check this assumption.

```
combined2_df <- combined_df %>%
  mutate(residual = c(fit1$residuals, fit2$residuals))

ggplot(combined2_df, aes(x, residual)) +
  geom_point() +
  geom_hline(yintercept = 0, color = "red") +
  facet_grid(sim ~ ., scales = "free") +
  theme_bw(base_size = 18)
```



- Generally, we check that there is "no pattern" in the residual plot.
- What do we mean by "no pattern"?

Null data

$$H_0 : e_i \sim N(0, \hat{\sigma}^2) \qquad \text{vs.} \qquad H_1 : \text{not } H_0$$

- $\hat{\sigma}$ is the estimate of σ from the model fit.

```
c(summary(fit1)$sigma, summary(fit2)$sigma)
```

```
## [1] 1.140534 1.229640
```

- We simulate observations of size $n = 200$ from $N(0, \hat{\sigma}^2)$. These are called **null data** as it is generated under the null hypothesis.
- We draw a scatter plot with these observations against the original x-values.
- We repeat this process $K = 19$ times then produce a **lineup** of the plots with the original residual plot randomly included into it.

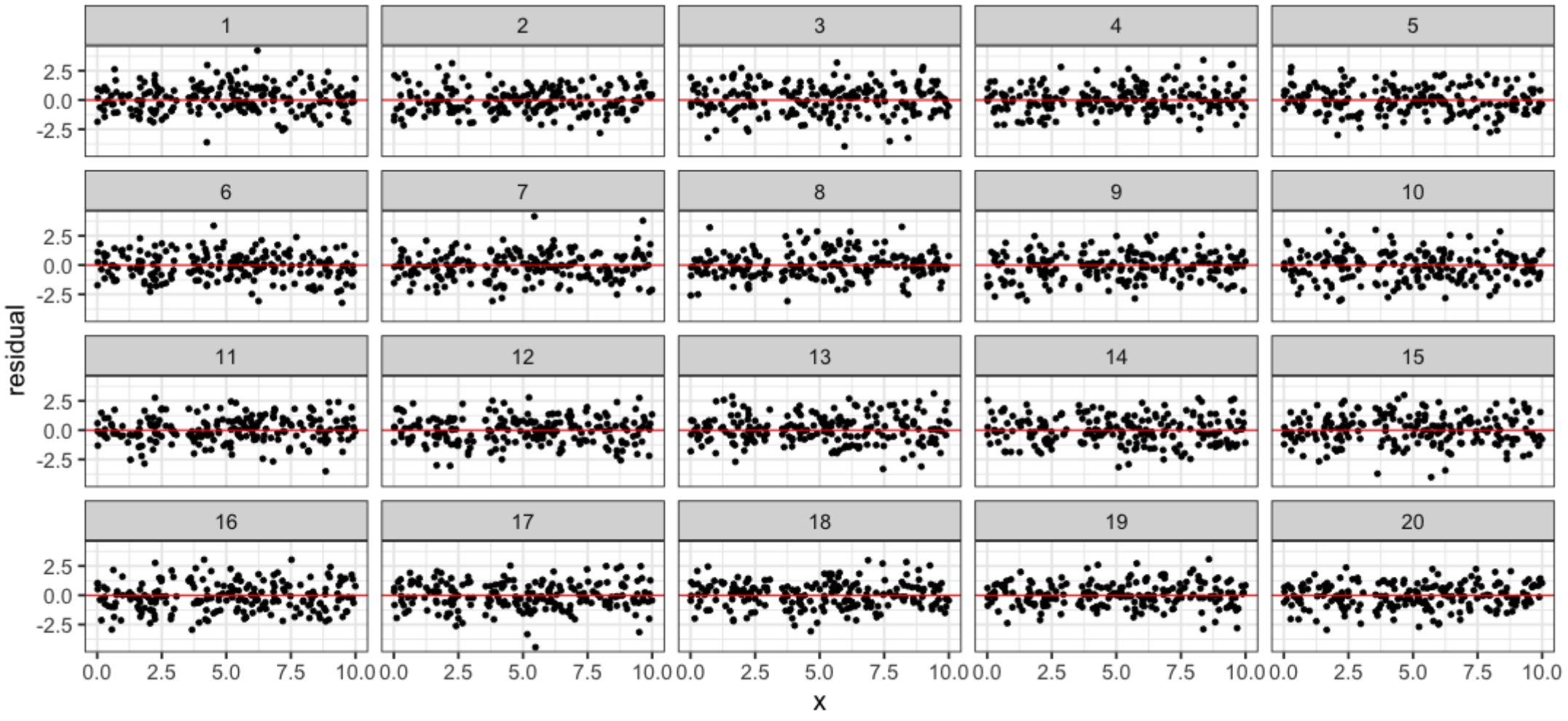
Null data using R

- The process of generating null data and lineups is facilitated by the `nullabor` 📦

```
library(nullabor)
method1 <- null_dist("residual", "norm",
                params =
                  list(mean = 0,
                       sd = summary(fit1)$sigma))
sim_df$residual <- fit1$residuals
null1_df <- lineup(method1, sim_df)
ggplot(null1_df, aes(x, residual)) +
  geom_point() +
  geom_hline(yintercept = 0, color = "red") +
  facet_wrap(~.sample) +
  theme_bw(base_size = 18)
```

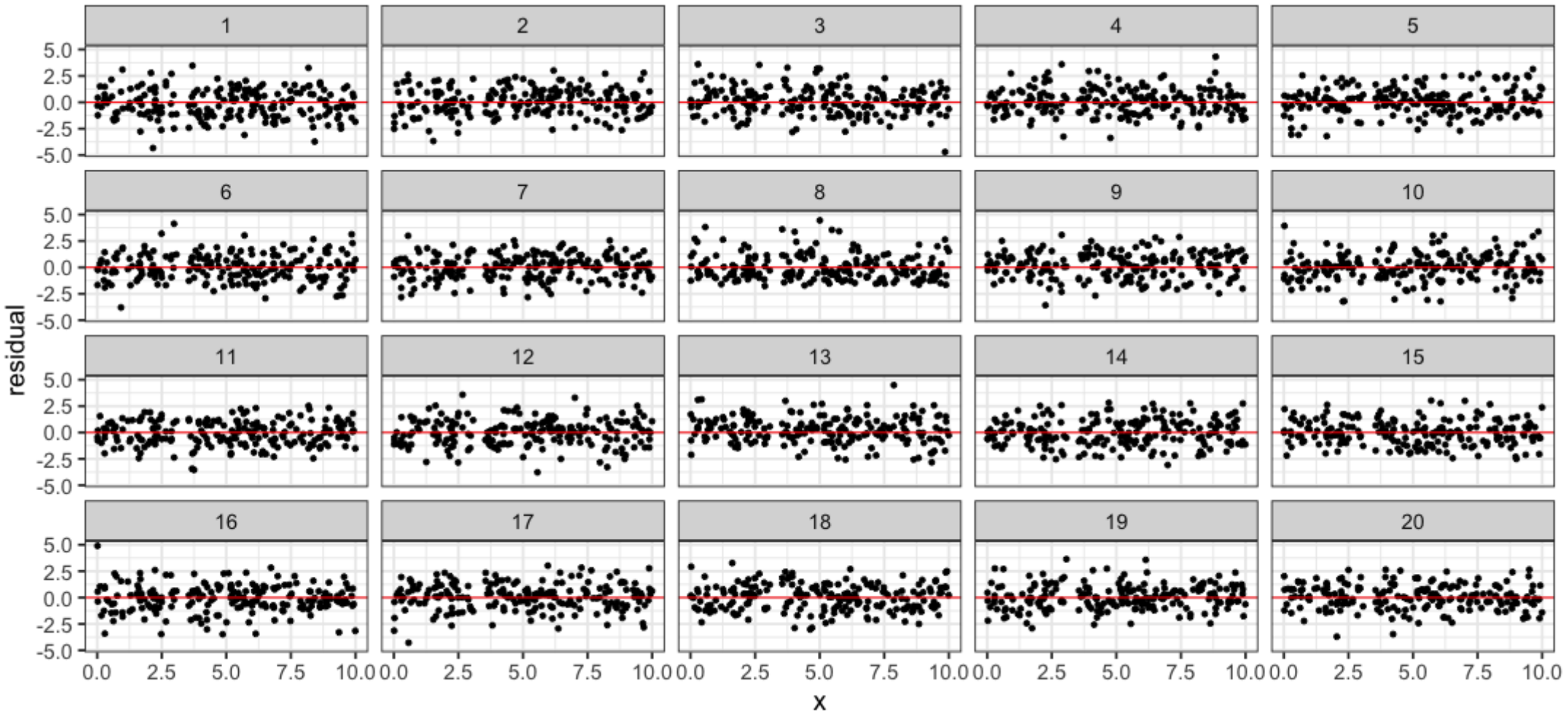
Line up for simulated data 1

```
## decrypt("E0Ui w676 VQ rnqV7VnQ KK")
```



Line up for simulated data 2

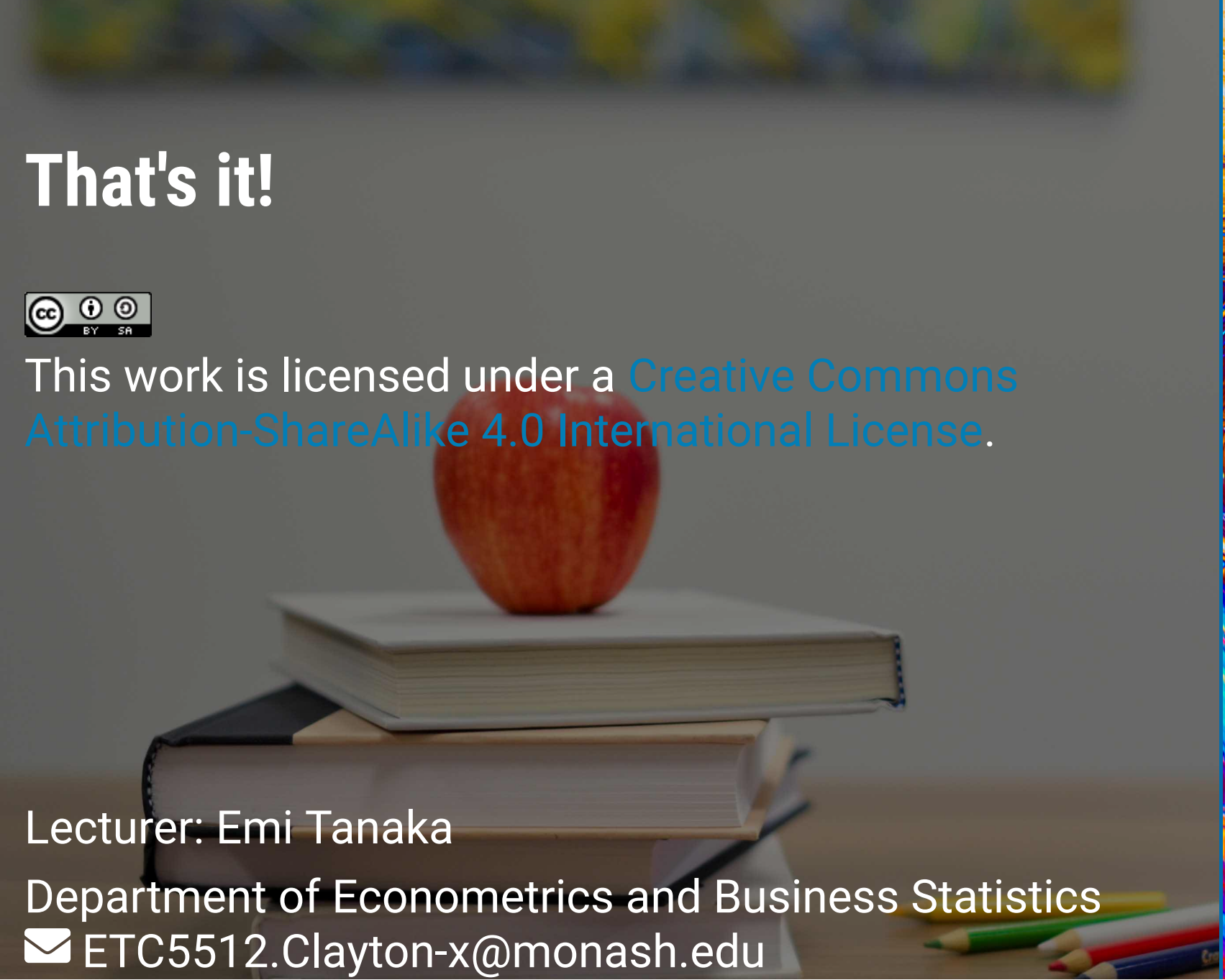
```
## decrypt("E0Ui w676 VQ rnqV7VnQ 28")
```



References

- In practice, you shouldn't show the data plot (also the test statistic in visual inference) before the lineup.
- You can read more about visual inference on the paper by Buja et al. (2009) but also one of your lecturers, Prof Di Cook, is a world leading expert on it! She is also the one of the authors of the paper, as well as, the maintainer of `nullabor` 📦.

That's it!



Lecturer: Emi Tanaka

Department of Econometrics and Business Statistics

ETC5512.Clayton-x@monash.edu