Data Wrangling and Data Analysis
Missing Data and Imputation

Daniel L. Oberski

Department of Methodology & Statistics

Utrecht University

Utrecht University

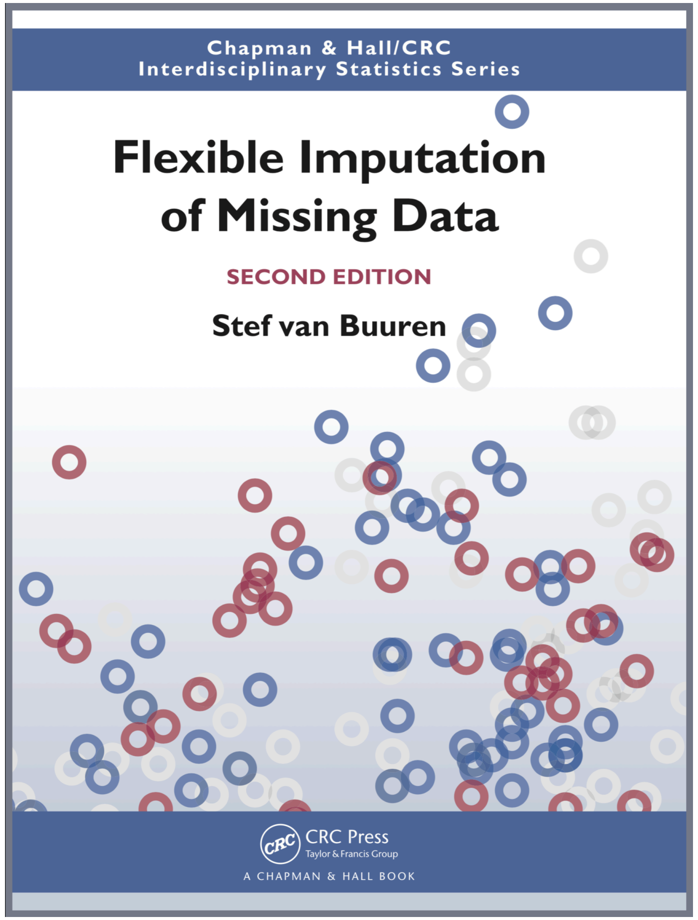
This week

- What is missing data
- Sources of missingness
- Missing data mechanisms
- Missingness patterns

Reading materials for this week

“Flexible Imputation of Missing Data”
https://stefvanbuuren.name/fimd

- Chapter 1
- Optional:
    - Ch 3 (practical, recommended)
    - Ch 4 (practical, more technical)
    - Ch 2 and 6 (more theoretical)



Some of this week’s materials are adapted from Gerko Vink & Stef van Buuren’s courses on multiple imputation

Assignments this week

- Monday: Exercise on missing data in python/R, understanding and visualising missingness.
- Tuesday: Correcting for missingness in databases in R
- Wednesday: Multiple choice test on missingness mechanisms and solutions
- Thursday: either (a) resit for the test, or (b) assignment on multiple imputation in R

https://www.menti.com/n1nbvdj3jv

Why a week on missing data

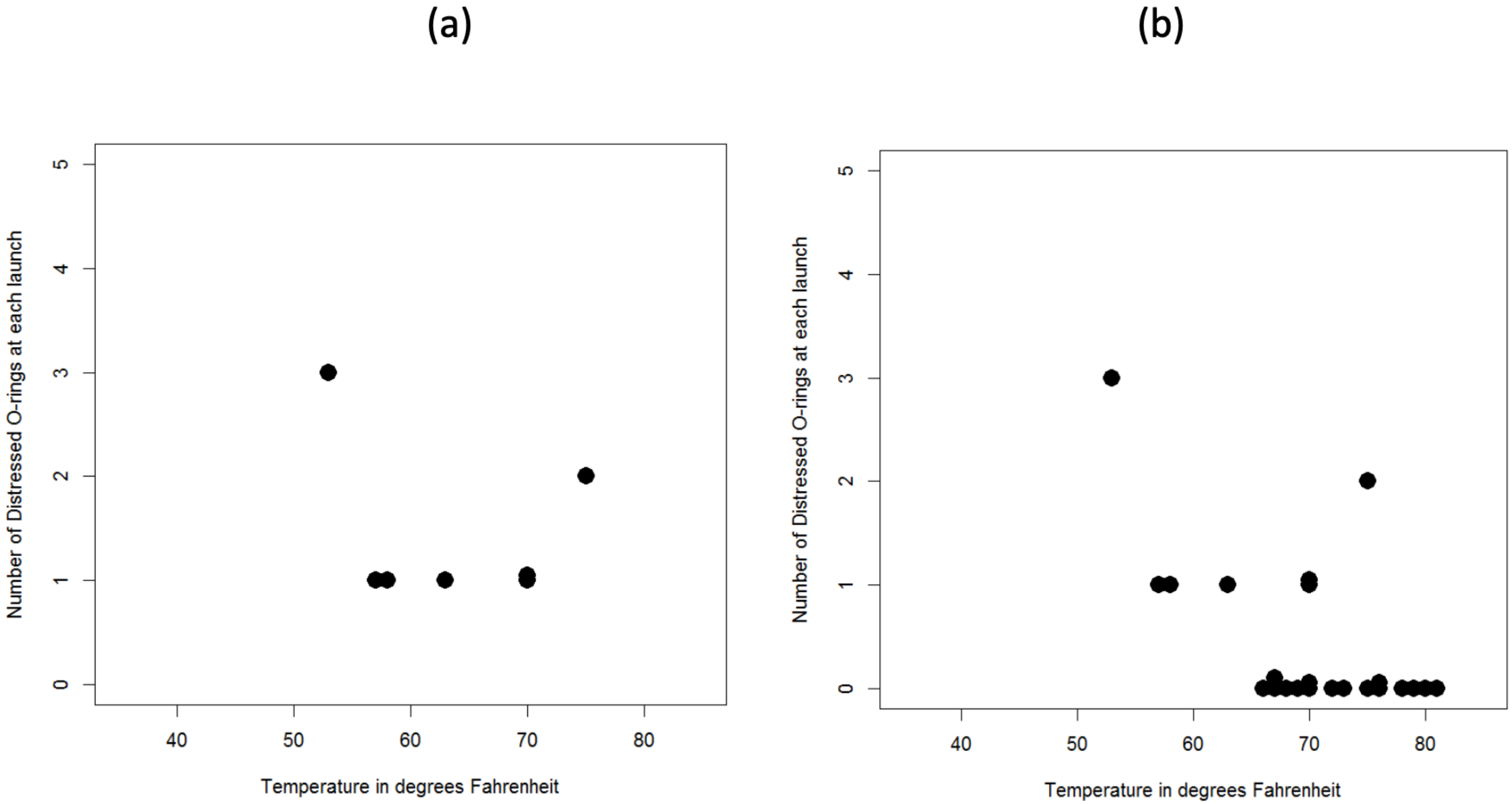
- Missing data are everywhere
- Ad-hoc fixes do not (always) work
- A data scientist needs to understand when which methods do and do not work
- Goal of the week: get comfortable with solving missing data problems

Why is missing data important?

- “Obviously, the best way to treat missing data is not to have them.” (Orchard and Woodbury 1972)
- “Sooner or later (usually sooner), anyone who does statistical analysis runs into problems with missing data” (Allison, 2002)
- Missing data problems are the heart of data analysis

Example: Challenger (1986)

Figure 1.1 (a) Data examined in the pre-launch teleconference; **(b)** Complete data.



Why is missing data important?

1. **Just annoying**
   - Most procedures don't deal with missings by default
2. **Less information than planned**:
   - Uncertainty of estimates (e.g. "standard errors", "power", "C.I", etc.)
   - Accuracy of predictive models
3. **Systematic biases**:
   - Estimates of interest wrong on average
   - Prediction error seems better than it will be in reality

How to think about missing values

- Missing values are those values that are not observed
- Values do exist in theory, but we are unable to see them
- One possible reason is non-response

Excel: Why using Microsoft's tool caused Covid-19 results to be lost

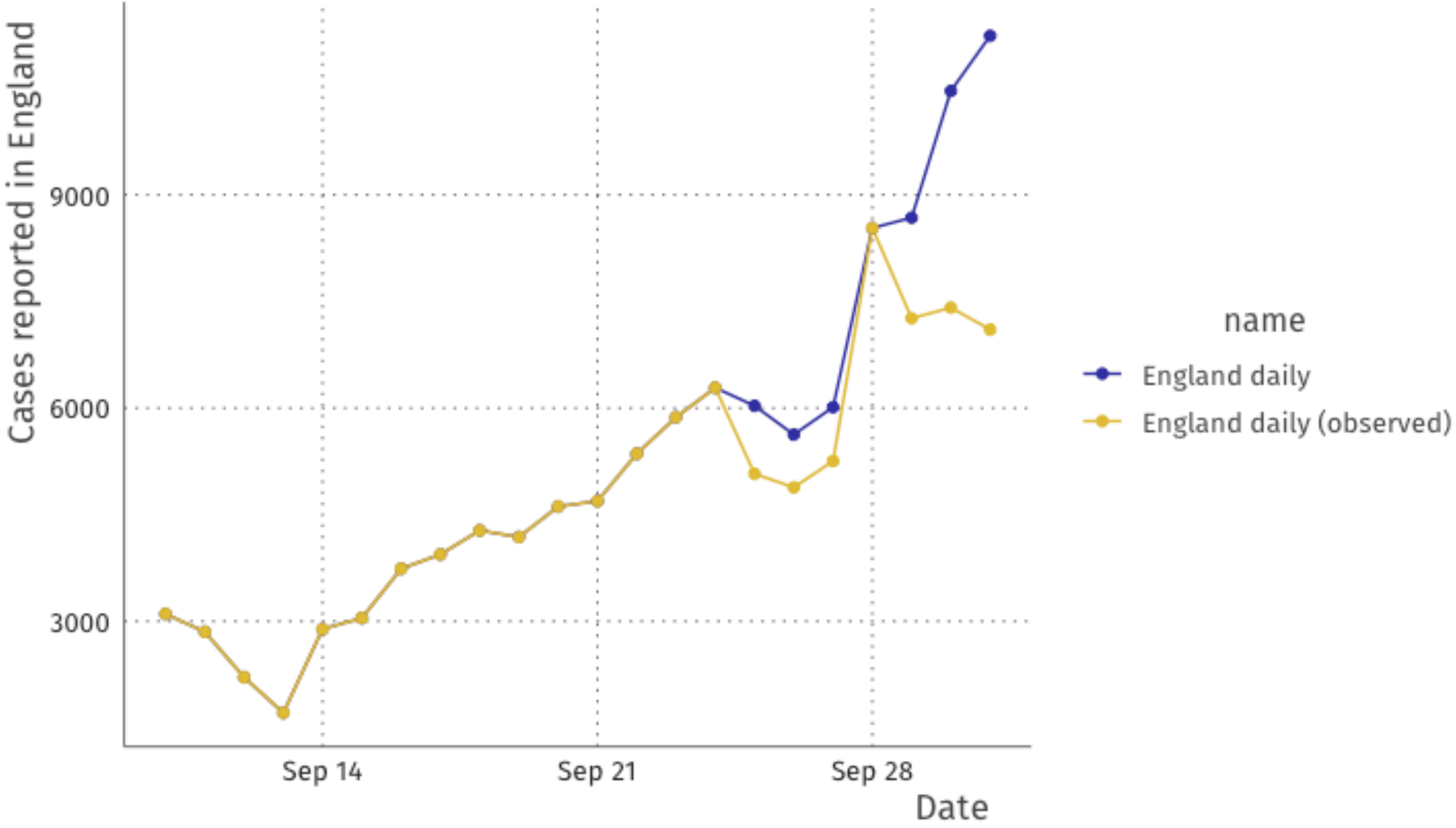
By Leo Kelion
Technology desk editor

5 days ago

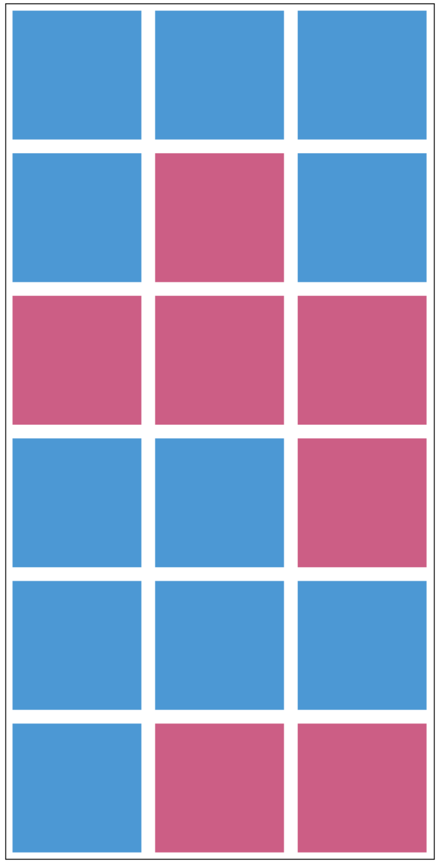
Coronavirus pandemic

Date (recorded – flow though into following day's published numbers)	Expected reported date for GOV.UK	Cases that were not included on the expected data
24/09/2020	25/09/2020	957
25/09/2020	26/09/2020	744
26/09/2020	27/09/2020	757
27/09/2020	28/09/2020	0
28/09/2020	29/09/2020	1415
29/09/2020	30/09/2020	3049
30/09/2020	01/10/2020	4133
01/10/2020	02/10/2020	4786

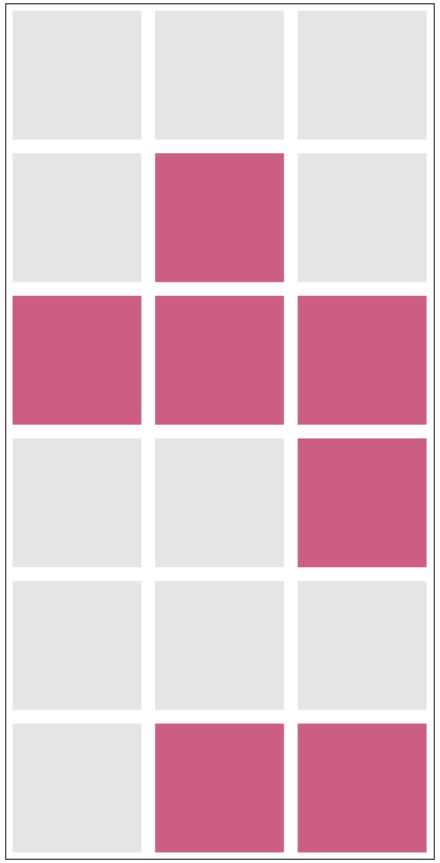
https://www.menti.com/n1nbvdj3jv


Cases reported in England
9000
6000
3000
Sep 14
Sep 21
Sep 28
Date
name
England daily
England daily (observed)

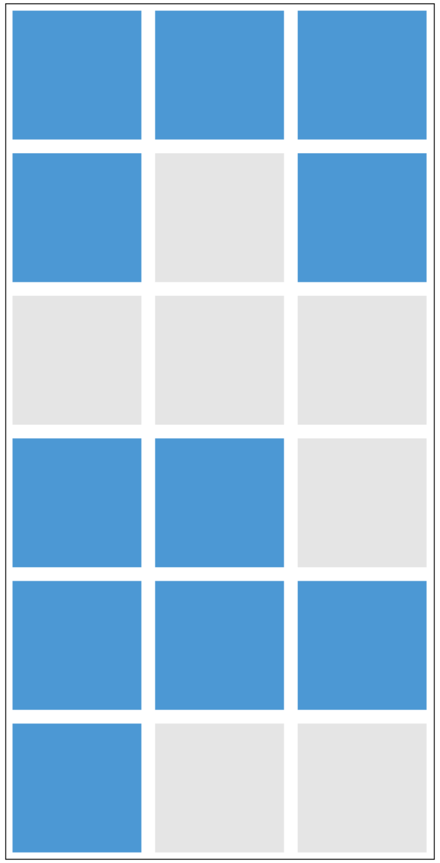
Complete data



Missing data



Observed data



Why values can be missing

- Missingness can occur for a lot of reasons. For example
  - Death
  - Dropout
  - Refusal
  - Programming errors
  - Privacy concerns
  - Too far away (e.g. deep space)
  - Too small to observe (e.g. particles)
  - Bad luck

But more importantly

If not all necessary information is captured, our inference may be wrong. This can be due to errors with respect to

- sampling: does sampling match the research goals?
- coverage: is the target population the same as the targeted population?
- non-contact: unable to reach respondent
- incompetence (interviewer/researcher)
- refusal: respondent does not want to answer

How is missing data coded in databases?

- R data frames: `NA` `is.na(df)`
- Python/Pandas: `NaN` `df.isna()`
- SQL: `NULL` `WHERE x IS NULL`
- JavaScript/ json: `null` `x === null`
- .csv files: "NA", "", "NULL"
- SPSS: -999, 99999 ☹

- So pay attention during data wrangling!

Illustrating the problem of missing data

- Let’s take this small set of numbers X that represent the body weight of 6 respondents.
    - one respondent has unobserved weight (NA)
    - weight is incomplete

Illustrating the problem of missing data

Because we have missing values, our statistics are not defined. Take for example the mean:

We can not calculate the mean over the cases.

- A missing value is **not** zero (its true counterpart may be, but we do not know that).

We can only calculate the mean over the observed set (minus the missings).

Illustrating the problem of missing data

The unbiased population variance estimator is then also not defined

as is the correlation with any other variable (here Y)

You can see where this leads. Statistics are not defined on incomplete data. Period.

Illustrating the problem of missing data

Because we have a smaller observed set (when compared to the incomplete set), in the analysis we have:

- lower statistical power – it is harder to find a significant difference when such a difference indeed exists.
- larger standard errors and confidence intervals – there is more uncertainty about the statistics of interest

Illustrating the problem of missing data

Remember: parameter estimates are statistics

- Regression parameters
- Spline knots
- Neural network weights
- Support vectors
- Eigenvalues
- Cross-validated hyperparameters

The missing data problem is pervasive

Further example

- Now add another variable weight that correlates with age.

```
##    age weight
## 1 13  42
## 2 40  80
## 3 24  73
## 4 14  NA
## 5 23  70
## 6 18  61
## 7 25  68
```

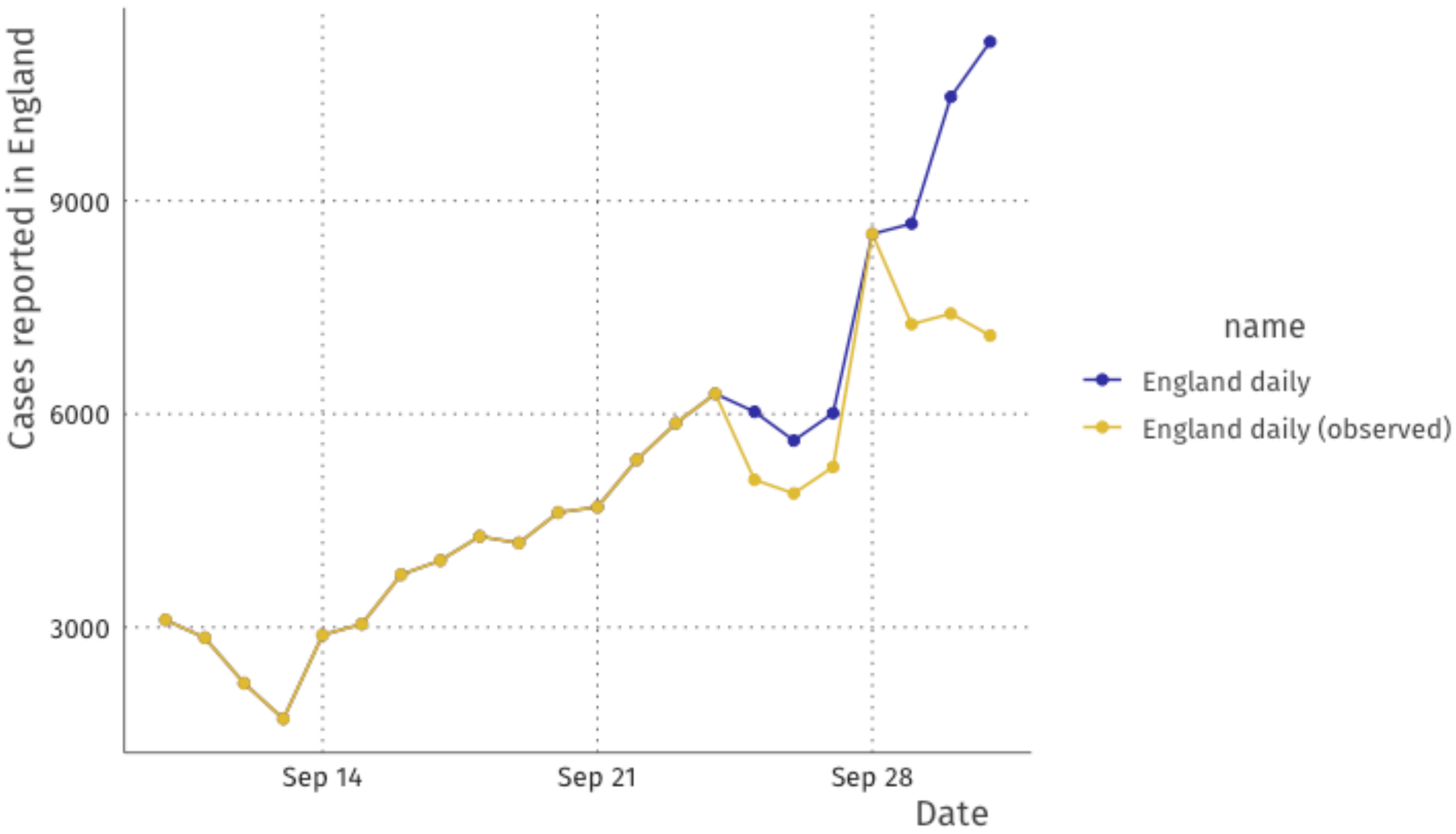
- We now have more information!

Further example

- The example becomes more apparent when we sort the data on age

```
##   age weight
## 1 13  42
## 4 14  NA
## 6 18  61
## 5 23  70
## 3 24  73
## 7 25  68
## 2 40  80
```

- Guess: NA between 42 and 61, but likely closer to 42 than to 61?



Cases reported in England
9000
6000
3000
Sep 14
Sep 21
Sep 28
Date
name
England daily
England daily (observed)

MCAR, MAR, MNAR

- MCAR: Missing Completely At Random
- MAR: Missing At Random
- MNAR: Missing Not At Random

MCAR, MAR, MNAR

- MCAR: The probability to be missing is constant for all units
- MAR: The probability to be missing depends on observed data
- MNAR: The probability to be missing depends on unobserved data

Missing data mechanism

- The theoretical, true, underlying process by which data goes missing
- Helpful concept: response indicator
  - R=1 if Y is observed
  - R=0 if Y is missing

MCAR, MAR, MNAR

- Formally:
  - MCAR: $P(R, X, U) = P(R)P(X, U)$
  - MAR: $P(R, X, U) = P(R, X)P(U)$
  - MNAR: $P(R, X, U)$ cannot be reduced

Examples of MCAR mechanisms

- Randomly sample people from a population
- Obtaining measurement of a certain process fails randomly

Examples of MAR mechanisms

- Non-response on income, where we do have register data for income from labour and data on wealth

- Measuring has a small probability of failing for high values of

- Branching patterns in questionnaires, e.g., “do you have children?” -> “How old are your children?”

Further example

- Now add another variable weight that correlates with age.

```
##    age weight
## 1 13   42
## 2 40   80
## 3 24   73
## 4 14   NA
## 5 23   70
## 6 18   61
## 7 25   68
```

- We now have more information!

Further example

- The example becomes more apparent when we sort the data on age

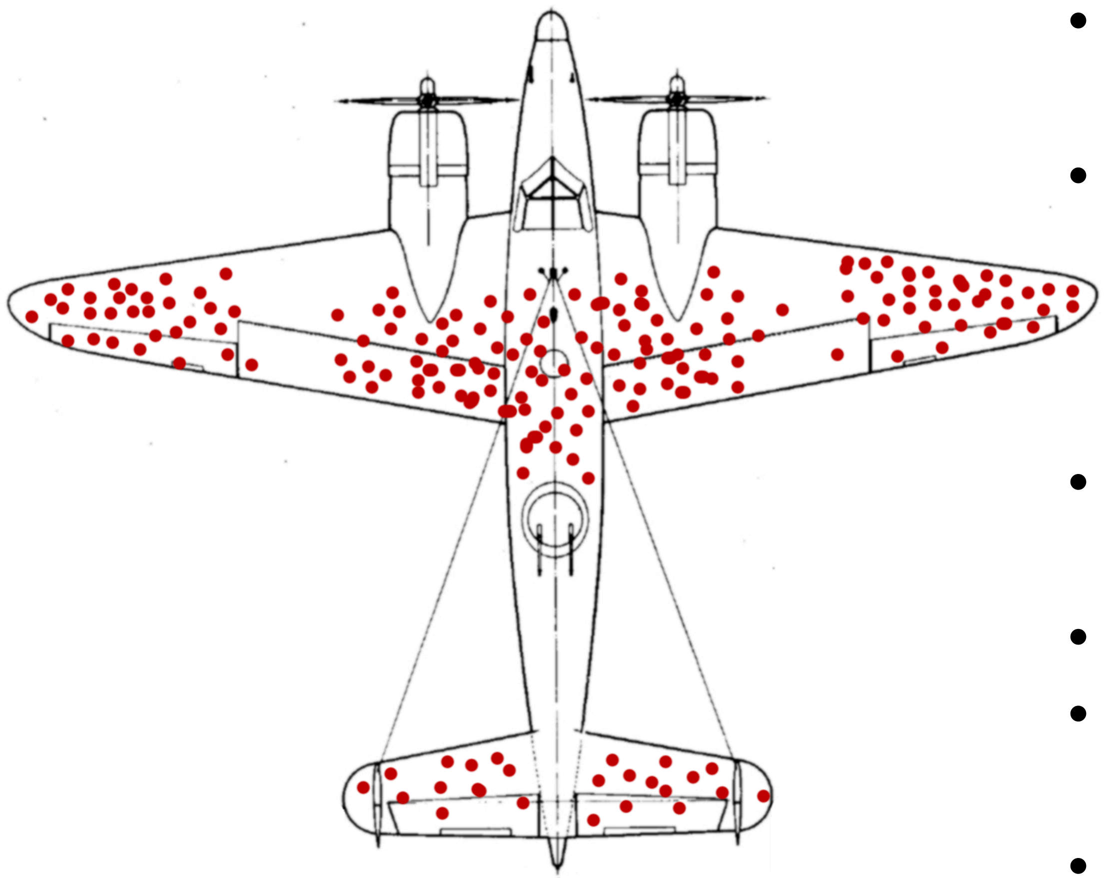
```
##    age weight
## 1 13   42
## 4 14   NA
## 6 18   61
## 5 23   70
## 3 24   73
## 7 25   68
## 2 40   80
```

- Guess: NA between 42 and 61, but likely closer to 42 than to 61?

Examples of MNAR mechanisms

- Non-response for income, but we do not have additional data
- Missingness depends on the missing values themselves!
- Failure probability of measuring depends on a variable which causes

Abraham Wald and the Missing Bullet Holes



- WW2 problem: too many airplanes are being shot down!
- Solution: collect data on where the returning planes are hit → reinforce those locations.

- It did not work. Abraham Wald (a famous statistician) was asked to help.
- Wald's conclusion: survivor bias
- Important info is where the **downed planes** were hit, not the returning planes!
- Missingness related to outcome

https://medium.com/@penguinpress/an-excerpt-from-how-not-to-be-wrong-by-jordan-ellenberg-664e708cfc3d

MCAR, MAR, MNAR

- It is possible to test whether the mechanism is MCAR or MAR, assuming it is either one of those (Little’s MCAR test)
- It is impossible to test whether the mechanism is MNAR or not: this is an assumption!
  - MNAR is a fundamental theoretical issue

Visualizing missing data

- It can be helpful to visualize the missing data
- Summarizes & provides insight into the amount and pattern of missingness
- In the assignment this afternoon you will work in R with missing data and create visualizations

- Visualization can help in understanding the missing data mechanism

Conclusion

- Missing data is not only annoying, but it can also lead to excessive uncertainty and even bias

- When bias occurs, depends on the situation and the thing you're interested in

- It can be challenging to deal with this type of situation.
- Standard advice (e.g. "impute mean") usually does not work!