STAT5003 – Week 6

Multiple Testing

What Is Real?

Deciding What's Real

How do you know if a significant association you find is real and not just random chance?

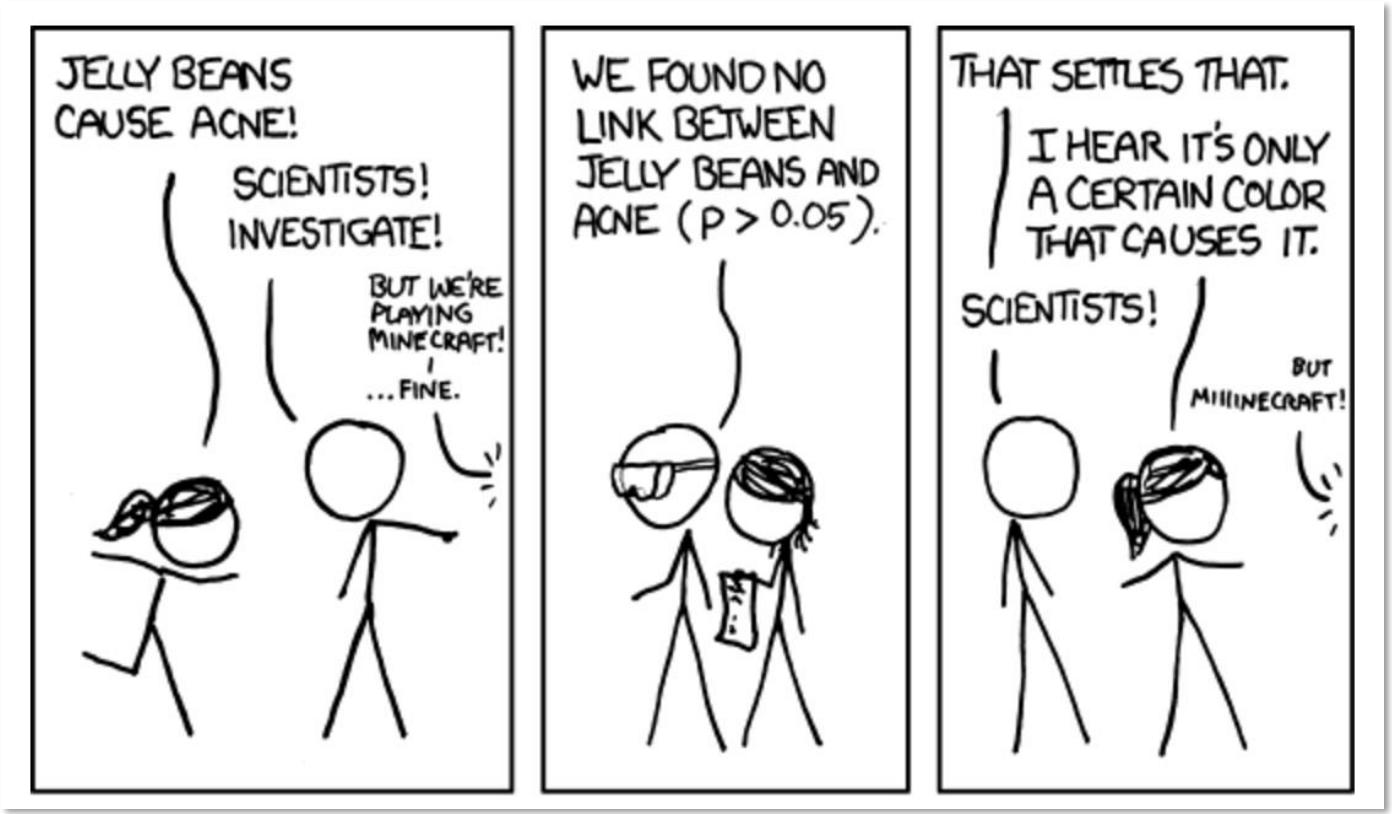
1. Because someone else published it
2. Because *p*-value is less than 0.05
3. Because of all the tests I ran, that one had the lowest *p*-value
4. Because it makes biological sense

The Reality of the Situation

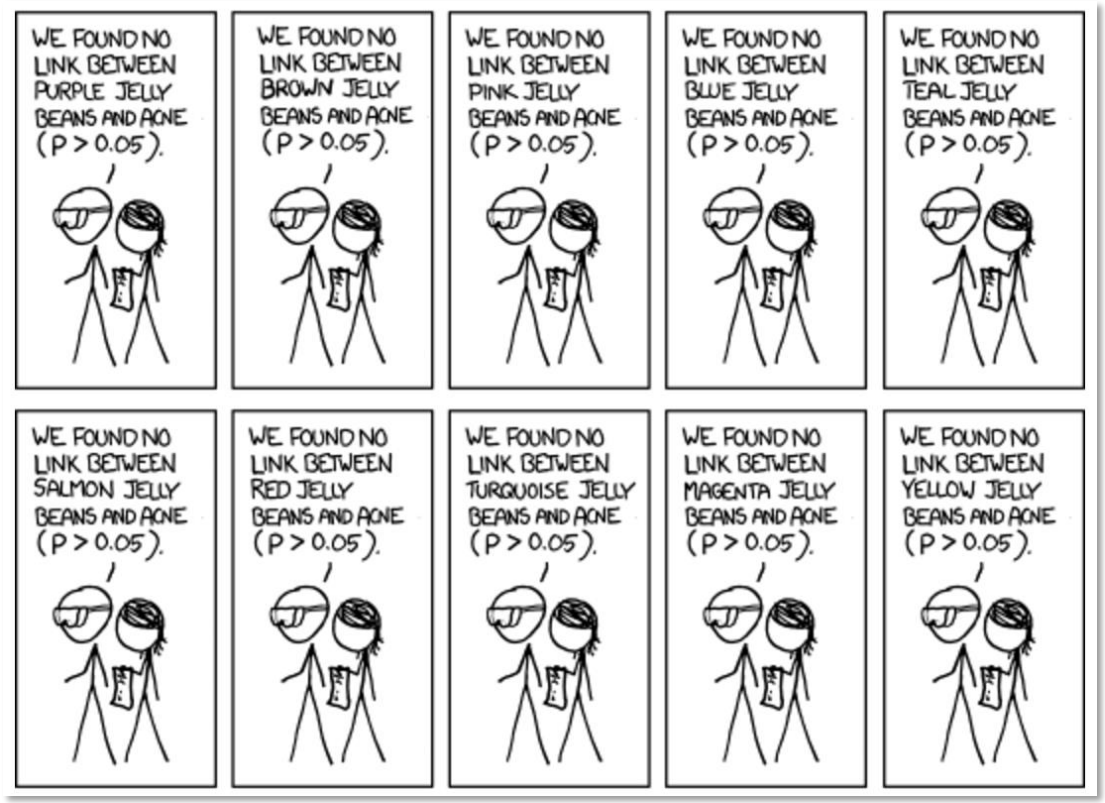
- We never really know what is a *real* association
- A small *p*-value provides some evidence against the null bit it could still be a false positive
- Type 1 error ($\alpha = 0.05$)
- For every model we evaluate at $\alpha = 0.05$, we accept that there is a 5% chance that we reject the null hypothesis when the null hypothesis is actually true

setting $\alpha = 0.05$ could also mean that we are potentially going to incorrectly reject the null hypothesis 5% of the time

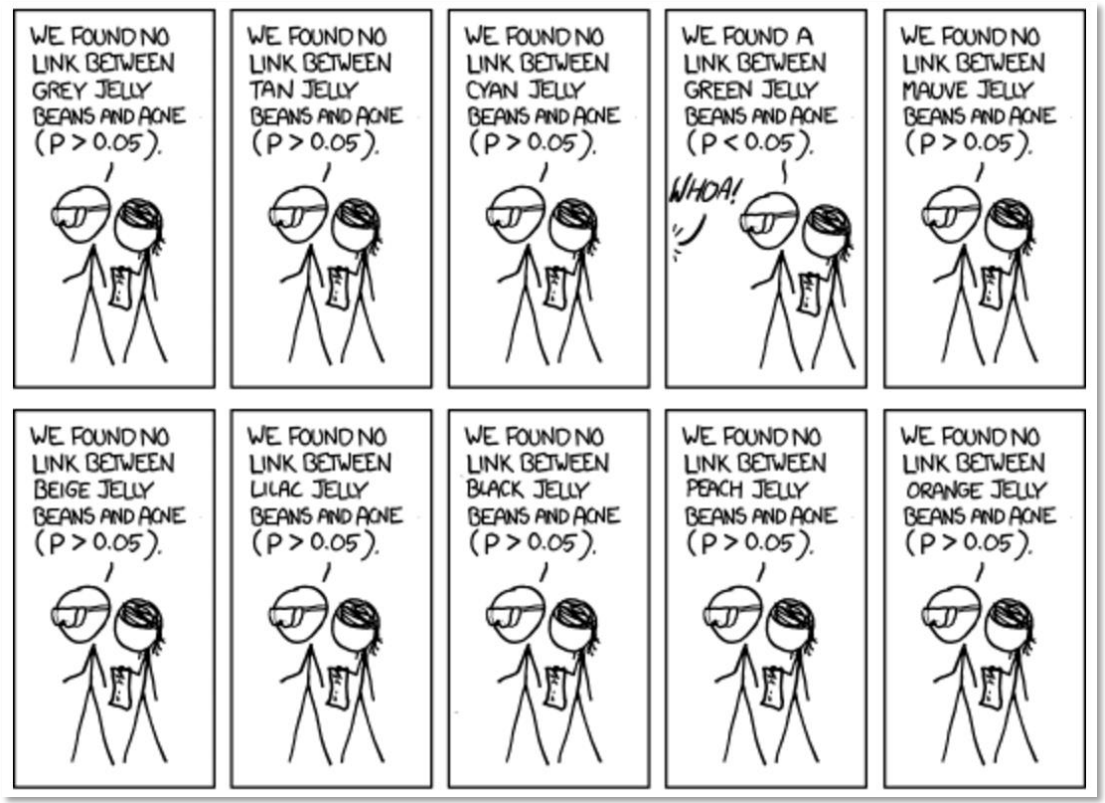
Jelly Beans and Acne



Jelly Beans and Acne (cont.)

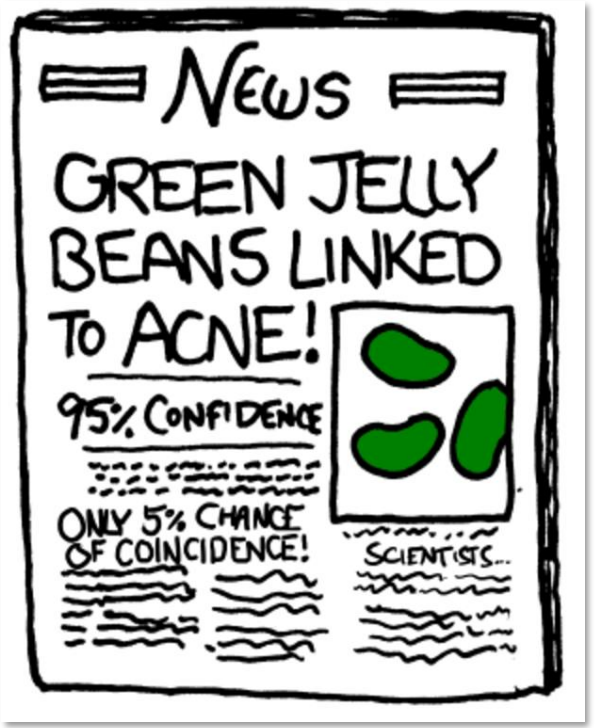


Jelly Beans and Acne (cont.)



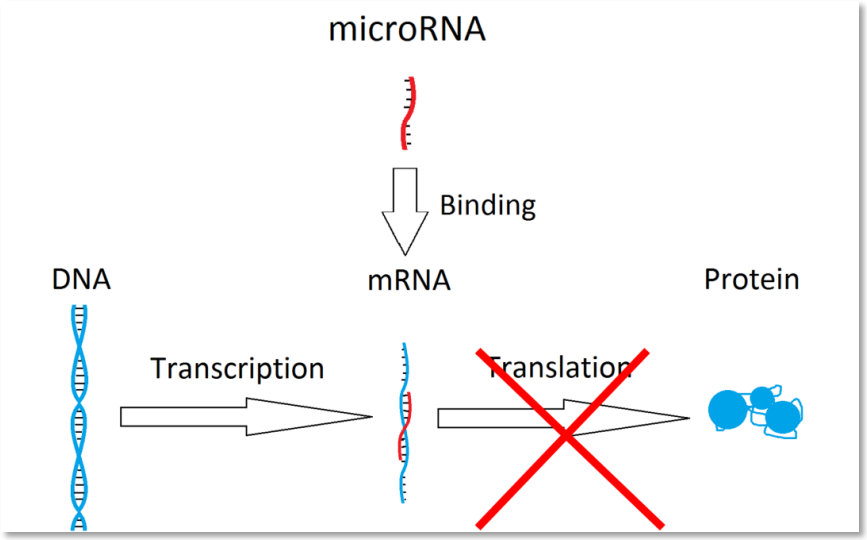
Jelly Beans and Acne (cont.)

- **Are jelly beans associated with acne?**
- If we performed 20 tests with $\alpha = 0.05$, how many tests are likely to be significant by chance alone?
- **Does this make you question the conclusions from any of the tests you've performed?**



Microrna and Alzheimer's Disease

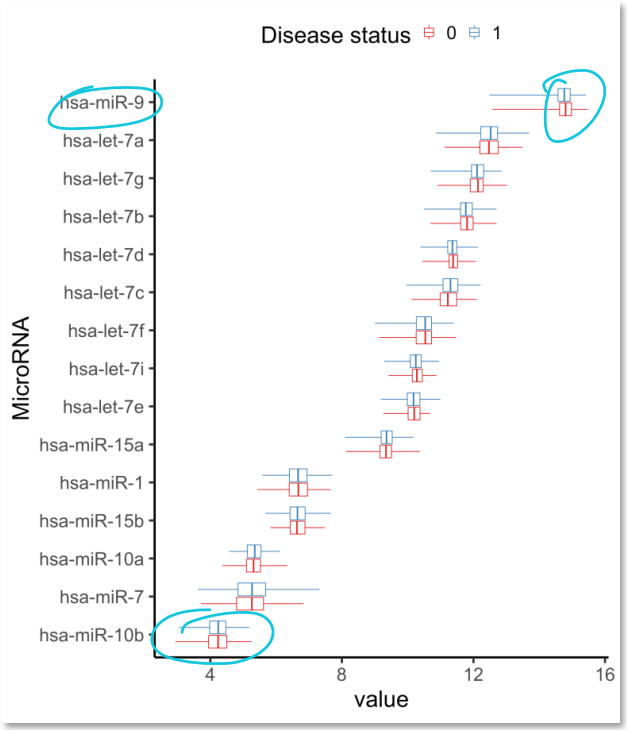
- MicroRNA are small non-coding RNA molecules that regulate gene expression.
  - Experiment by measuring the amount of 309 microRNAs in 701 subjects.
  - Test for significant differences between the means of subjects with and without Alzheimer's disease for each microRNA.
- Is there any evidence that microRNA behaviour in the brain might be associated with Alzheimer's disease (Patrick et al., 2017)?



What Does the Data Look Like?

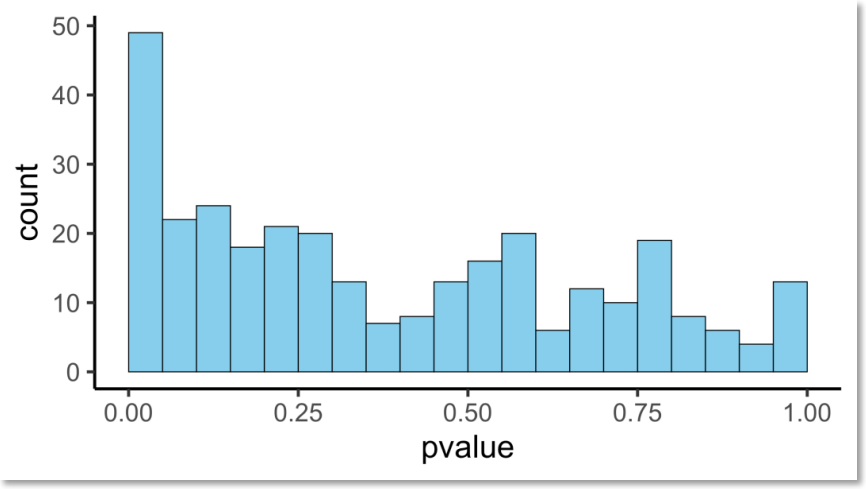
- We have a whole lot of two-sample *t*-tests
- One for each of the 309 MicroRNA

with / without Alzheimer's



Distribution of Observed *p*-Values

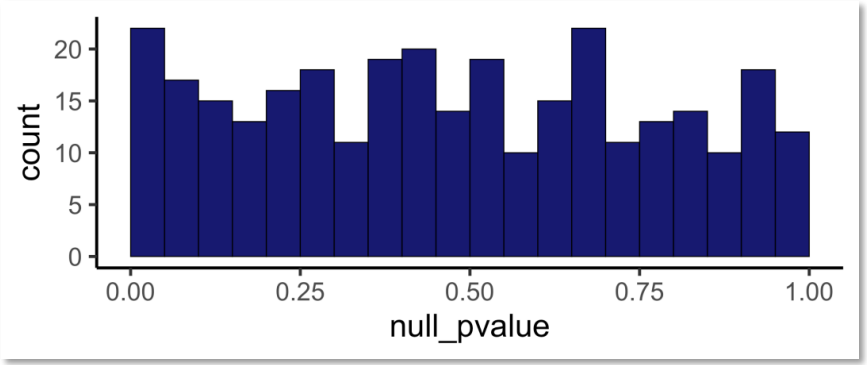
- Let's visualise the distribution of *p*-values for all 309 microRNA
- Of the 309 microRNA tested, 49 with *p*-values less than 0.05
- Are all of these important?



Distribution of Null *p*-Values

- If there was no association between any microRNAs and Alzheimer's disease we would expect our *p*-values to follow a uniform distribution.
- We can generate a set of *p*-values *knowing* that there is no association and visualise this.
- When we **know** that there are no truly important microRNAs, we still see 22 "significant" *p*-values in this simulated example.

Type I error



22 p-values
< 0.50

but in reality no relationship between expression level and Alzheimer's disease.


THE UNIVERSITY OF
SYDNEY

Multiple Testing

p-value: controls FP rate at α.

Accounting for Multiple Testing

- If *p*-values are correctly calculated, calling all *p*-values less than α significant will control the false positive rate at level α, on average.
- Suppose that you perform 10,000 tests and the reality is that $\theta = 0$ for all of them.
- Suppose that you call all *p*-values less than 0.05 significant.
- The expected number of false positives is: $10{,}000 \times 0.05 = 500$ false positives.
- **How do we avoid so many false positives?**
- Consider two approaches.
  1. Controlling the **family-wise error rate (FWER)**
  2. Controlling the **false discovery rate (FDR)**

$FP = 10000 \times \alpha$

} never heard of this.

Bonferroni Correction

MATH2831 concept

- The Bonferroni correction is the oldest multiple testing correction
- Given that the number of false positives for m tests is $m\alpha$ then consider defining a new threshold for significance:

$$\alpha^* = \frac{\alpha}{m}$$

  - This is conservative but keeps FWER $< \alpha$
  - For example, for $m = 20$
  - $1 - (1 - \alpha^*)^m = 1 - (1 - 0.05/20)^{20} = 0.0488$

$1 - \left(1 - \frac{\alpha}{m}\right)^m$

$\alpha^* = \frac{\alpha}{m}$

Family Wise Error Rate

Bonferroni Correction (cont.)

- **Basic idea**
  - Suppose you do m tests
  - You want to control FWER at level α so $P(V \geq 1) < \alpha$
  - Calculate *p*-values in the usual way
  - Set $\alpha^* = \alpha/m$ (or alternatively calculate adjusted *p*-values: $p^* = p\text{−value} \times m$)
  - Call all *p*-values less than α^* significant (or all adjusted *p*-values less than α significant)
- **Pros**: easy to calculate, conservative
- **Cons**: may be very conservative

any p-values $< \alpha^*$ is significant

brings threshold down too far

Controlling False Discovery Rate (FDR)

- The Benjamini–Hochberg (BH) procedure is the most popular correction when performing *lots* of tests say in genomics, imaging, astronomy, or other signal-processing disciplines
- **Basic idea**
  - Suppose you do m tests
  - You want to control FDR at level α
  - Calculate *p*-values normally
  - Order the *p*-values from smallest to largest $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$
  - Find $j^* = \max j$ such that $p_{(j)} \leq \frac{j}{m}\alpha$
  - Reject all H_{0i} where $p_{(i)} \leq \frac{j^*}{m}\alpha$
- **Pros**: still pretty easy to calculate, less conservative (maybe much less)
- **Cons**: allows for more false positives, may behave strangely under dependence

each time we adjust the threshold.

need to see example.

$j^* = \max j \,,\ p_{(j)} \leq \frac{j}{m}\alpha$

reject H_{0i} where

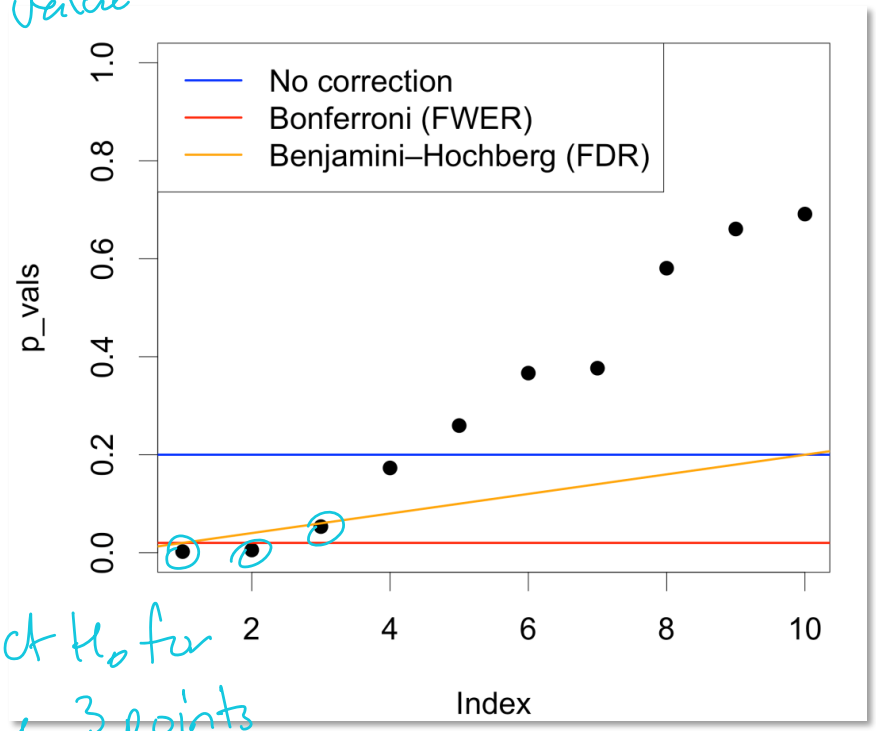
$p_{(i)} \leq \frac{j^*}{m}\alpha$

$p_{(i)} \leq \frac{1}{m}\alpha \leq p_{(2)} \leq \frac{2}{m}\alpha \leq \cdots \leq p_{(m)}\frac{m}{m}\alpha$

$p_{(m)}\alpha$

Comparing Bonferroni and BH

- Controlling all error rates at $\alpha = 0.20$ and using a sample of 10 microRNAs
- The lines are the significance thresholds for the three methods; if a point is below the line, the method would consider it "significant"



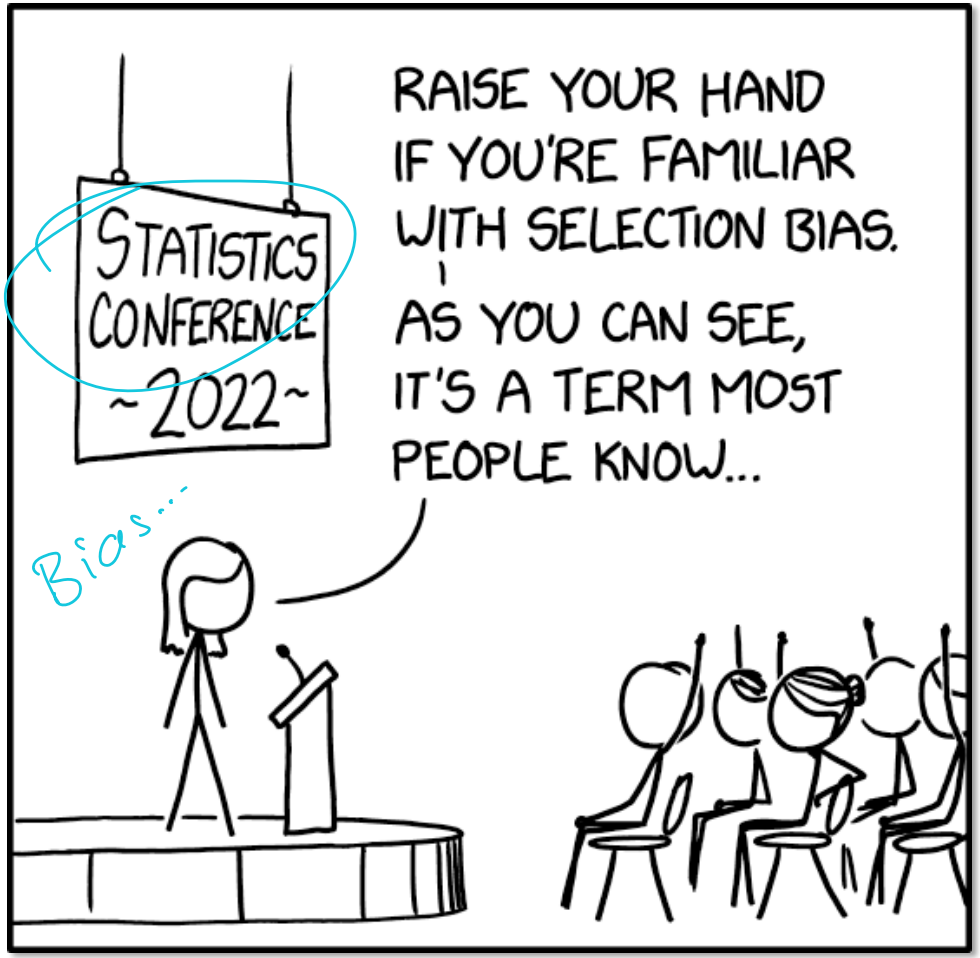
Final Comments

- Multiple testing is an entire subfield of statistics.
- A version of a Bonferroni/BH correction is often sufficient.
- If there is strong dependence between tests, there may be problems.


THE UNIVERSITY OF
SYDNEY

Sampling Issues

Sampling Issues



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% and won 46 of 48 states

Gallup Poll

- George Gallup was setting up his survey organisation.
- He drew 3,000 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?**

Sampling

- 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

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

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)

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

Bias

When looking at data, 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 favourite 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.


THE UNIVERSITY OF
SYDNEY

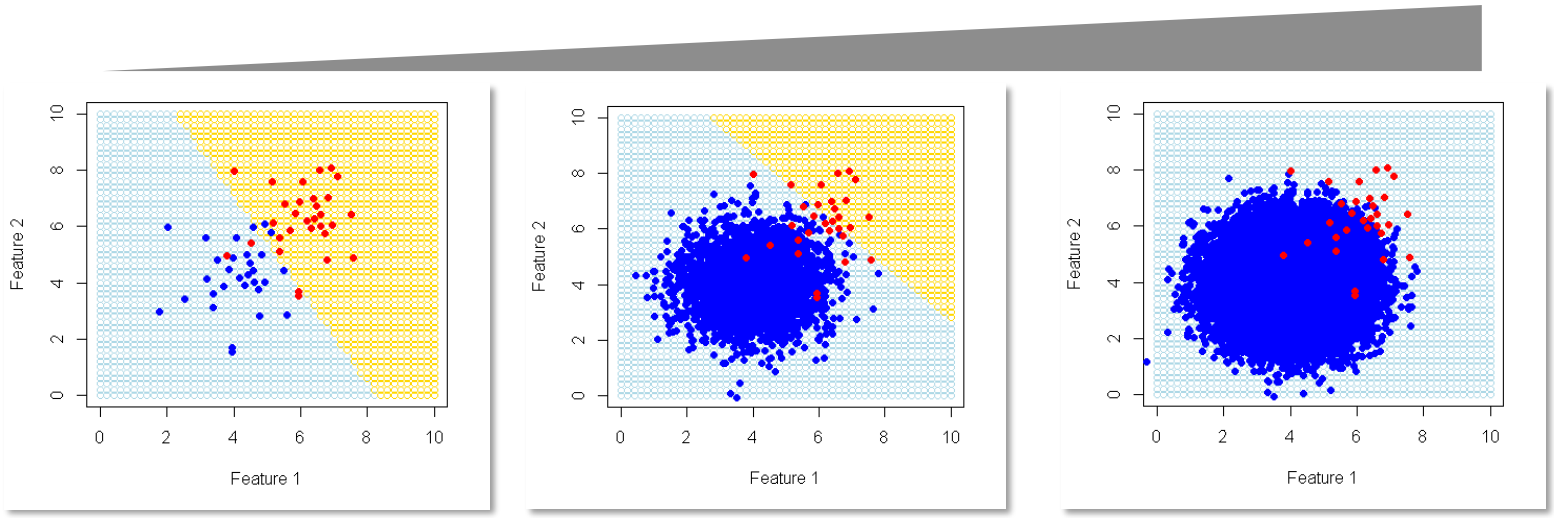
Imbalanced Data

Why Is Class Imbalance a Problem?

- Let's say we have a classification problem to detect credit card fraud, but only 1% of transactions are fraud.
- If you use accuracy as the metric to optimise, then just by classifying every transaction as not-fraud will get you to 99% accuracy!

Class Imbalance

Degree of class imbalance



Decision boundary of a linear SVM

Assume • are positive instances and • are negative instances.

Use a Better Performance Metric

- F1 score: harmonic mean of specificity and sensitivity

$$F_1 = \frac{2 * TP}{2 * TP + FP + FN}$$

- Plot the ROC curve and calculate area under curve (AUC)

(new)

- Cohen's Kappa: compares expected to observed accuracy

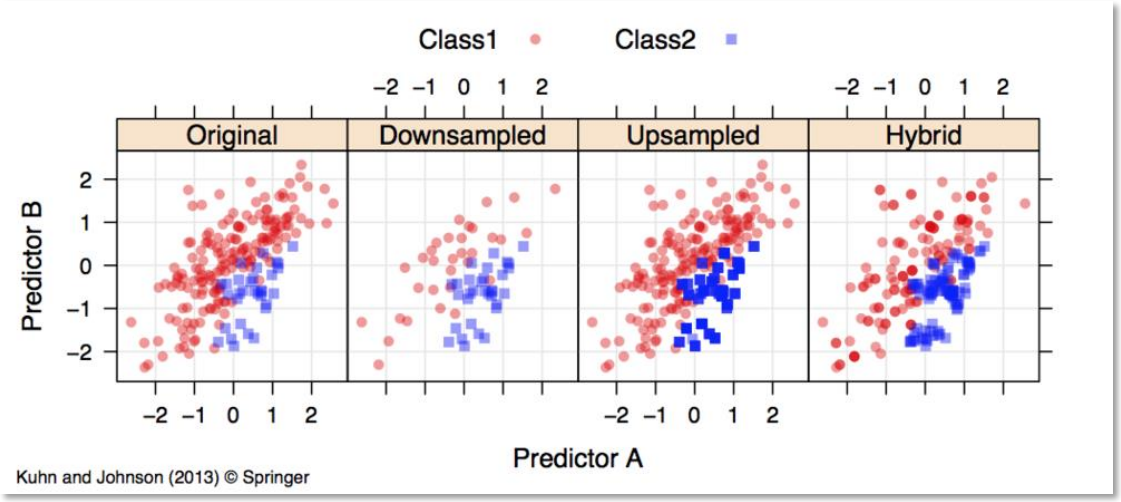
$$\kappa = \frac{p_0 - p_e}{1 - p_e}$$

p_0 : observed accuracy

p_e : expected accuracy

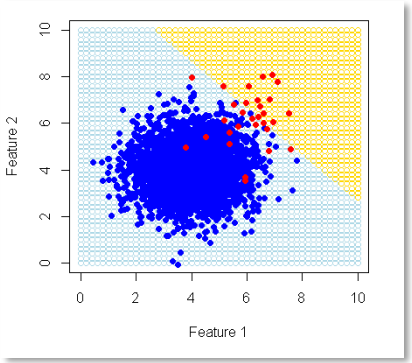
Random Up-Sample to Balance the Data

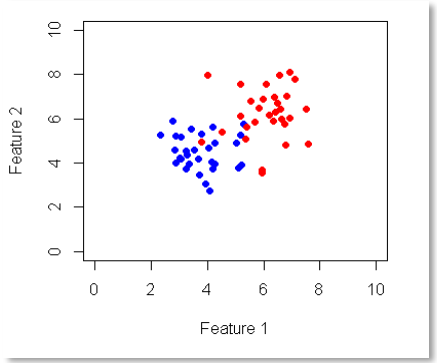
- Random up-sampling
- Advantage: keep and utilise all original data
- Disadvantage: create duplicated and/or artificial instances which may introduce bias and /or noise to the original data

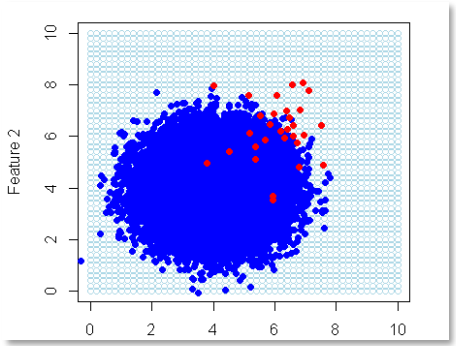


Random Down-Sampling to Balance the Data

- Advantage: do not introduce duplicates and/or artificial instances
- Disadvantage: not all data points are used; potentially removing useful information
- Better choice for data with very high-class imbalance







Create Synthetic Samples of the Minority Class

- Synthetic Minority Over-sampling Technique (SMOTE) is a popular algorithm
- It creates synthetic samples from the minority class by:
  - Finding the *k*-nearest-neighbours for minority class observations
  - Randomly choosing one of the *k*-nearest-neighbours, then using it to create a similar but random new observation
- Be careful you split your data into training/validation **before** doing any oversample/SMOTE; otherwise, you will leak information from training to validation data set
- The R package "*DMwR*" implements SMOTE


THE UNIVERSITY OF
SYDNEY

Missing Data

MCAR, MAR, MNAR

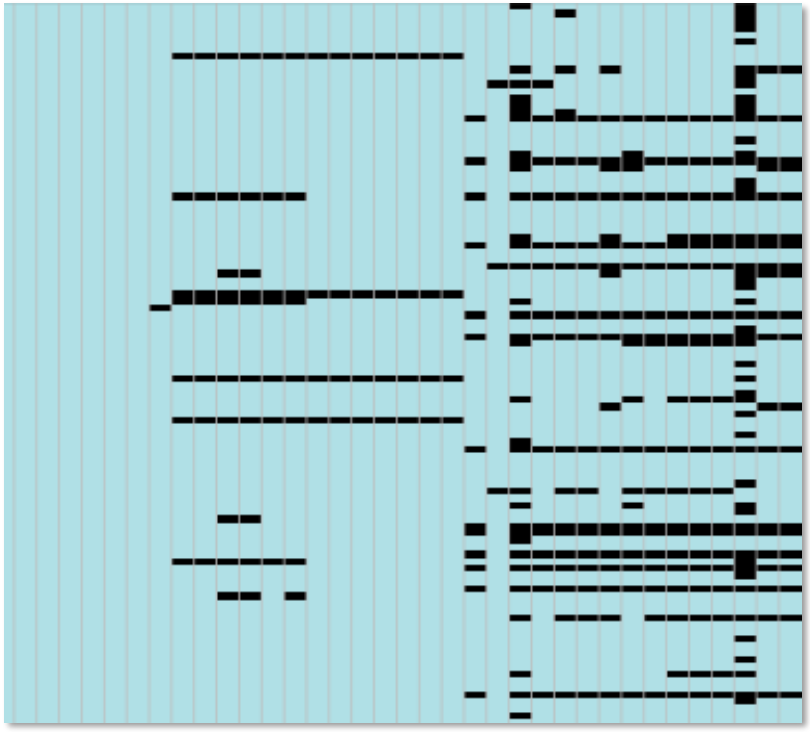
Mechanisms for Missing Data

- **M**issing **C**ompletely **A**t **R**andom **(MCAR)**
  - Pattern of missingness is independent of missing values and the values of any other measured variables.
  - For example, let's say we run a political polling survey and some people don't want to give their age in the questionnaire, but this does not relate to any other variable (including how their party preference).
- **M**issing **A**t **R**andom **(MAR)**
  - Missingness in a variable is not related to the variable but related to some other variables.
  - For example, in a polling survey, if for example women are more likely to disclose how they will vote, missingness could be related to gender but not to party preference.
- **M**issing **N**ot **A**t **R**andom **(MNAR)** — trickiest to deal with
  - Missingness is due to the value of the variable itself even after accounting for other variables.
  - For example, in a polling survey, if liberal voters are less likely to disclose how they intend to vote.

hey idk: impute data

Identifying Different Types of Missingness

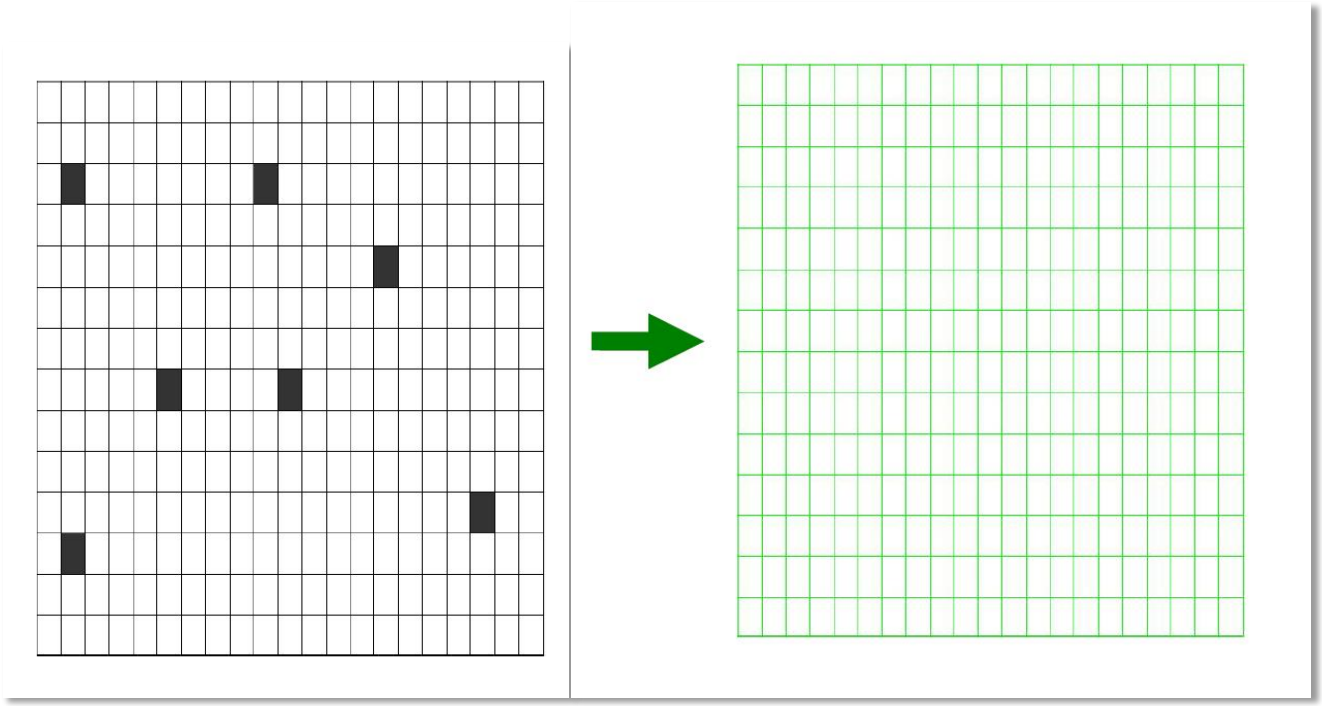
- Unfortunately, there is no statistical method to determine the mechanism of missingness.
- You can guess the mechanism of missingness by knowing something about the data, and something about the data collection method.
- To see if the data is **MAR**, you can try to fit a classification model to predict missingness.



Dealing With Missing Value

- For categorical data, “missing” can be a category
  - For example, in a survey poll, if someone does not want to disclose who they want to vote for, can be in the category “undecided.”
- ***Delete data with missing value;*** two options
  1. Omit the variable with missing data.
  2. Omit the observation with missing data.
  - Drawbacks are that you might be throwing away valuable information, or inadvertently introduce bias into the data.
- ***Impute,*** i.e., fill in the missingness

Imputation

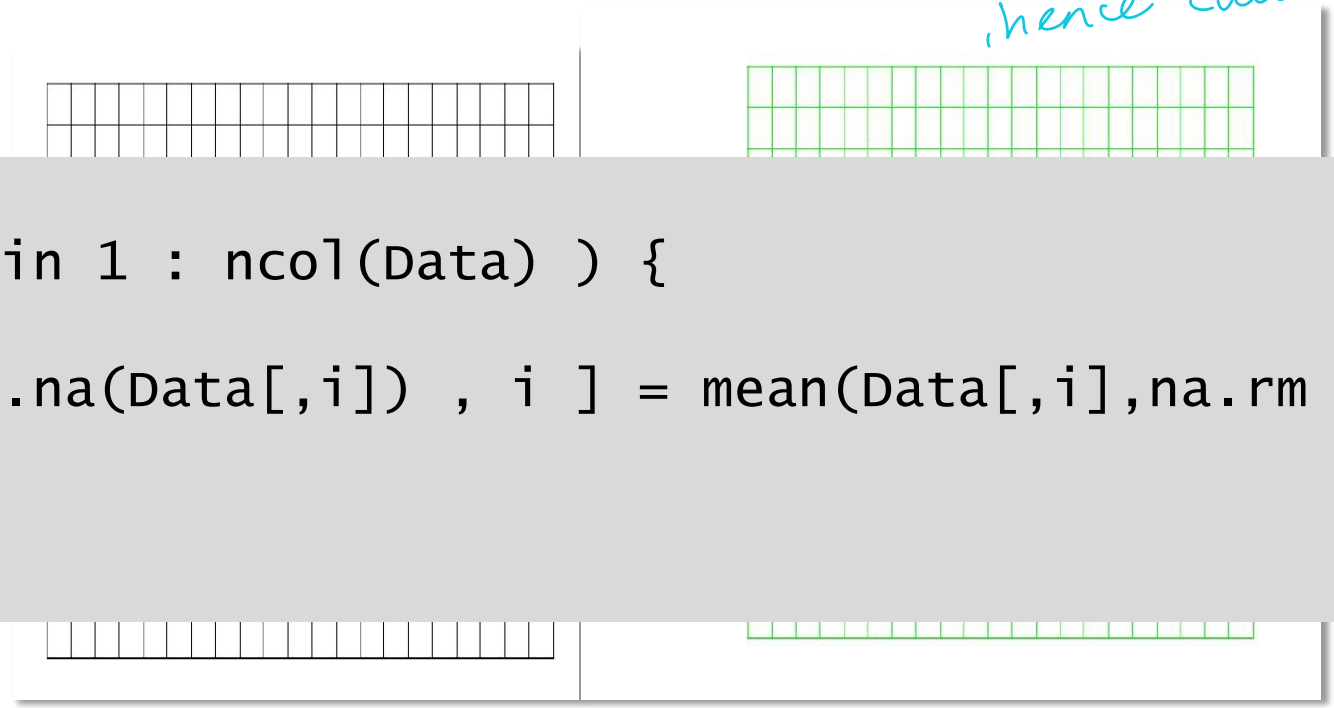


Single Imputation

- Single imputation replaces the missing value with a single value
- Examples
  - Replace the missing values of a feature with the mean/median value of that feature.
  - Use a predictive method for filling in the missing values, e.g., regression trees, *k*NN.
  - Replace the missing value with the last observed value for that feature.
- With single imputation, once the missing data is added back, it is treated as equal to the non-missing data, hence the uncertainty in the missing value data is lost

Mean Imputation

a lot of missingness may cause mean imputation to clump together in the centre, hence causing bias



```
for ( i in 1 : ncol(Data) ) {

Data[ is.na(Data[,i]) , i ] = mean(Data[,i],na.rm = TRUE)

}
```

A Categorical Variable

A	B	NA
180	20	10

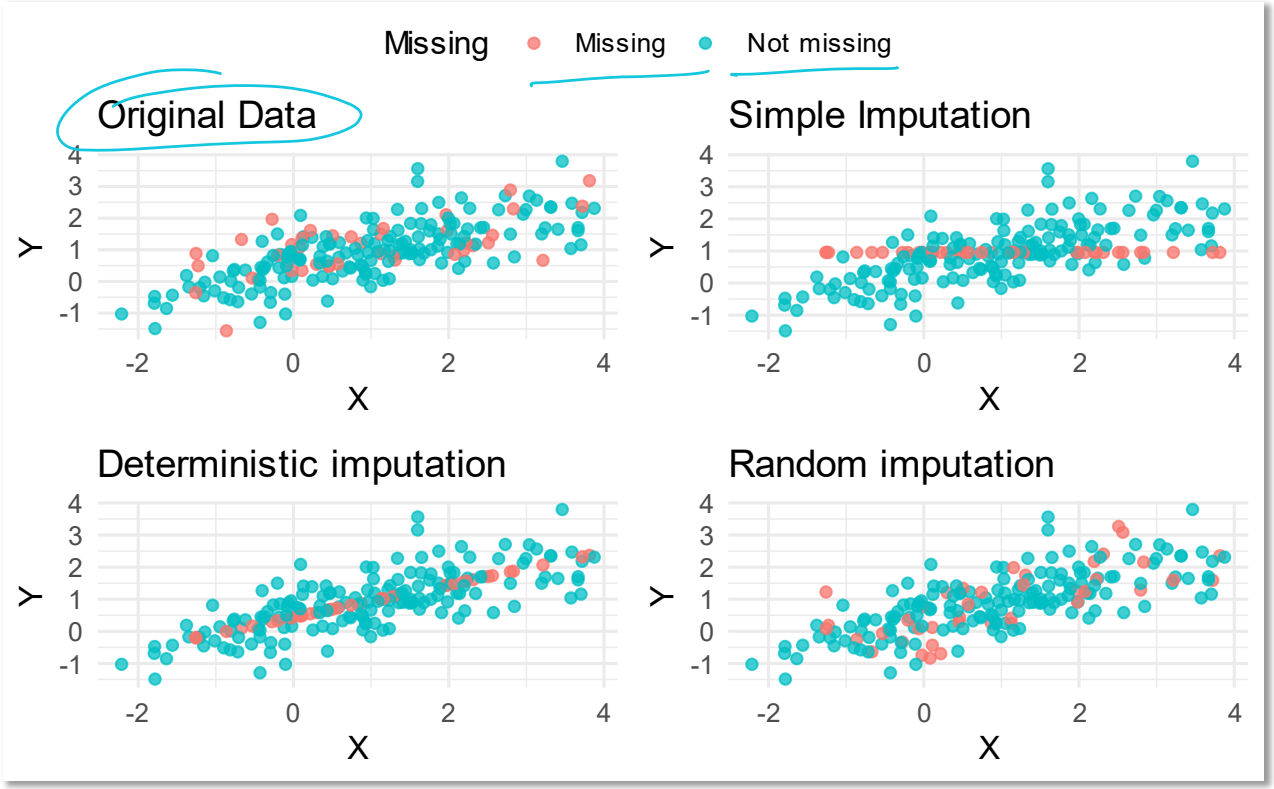
```
sample( c('A','B'), 10, replace=TRUE, prob=c(180,20)/200 )
```

proportions we have observed

Multiple Imputation

- Multiple imputation accounts for uncertainty in the imputation process
- Generally follows three steps:
  1. Impute the data *k* times (this can be done using a single imputation method).
  2. Perform analysis (e.g., regression) on each of the *k* imputed data sets.
  3. Pool the *k* results together.
- Multiple Imputation by Chained Equation (MICE) is a popular method and it is implemented in the R package "MICE"
  - See van Buuren and Groothuis-Oudshoorn (2011).

Imputation Types



← mean imputation.

Practical Suggestions

- It is highly recommend that you visualise your data to look for patterns of missingness.
- Be wary of variables with high proportion of missing data. However, this might not be a problem if imputation is applicable and performs well.
- Some algorithms can cope with missingness (e.g., decision trees) and so you may not need to do imputation.
- If you believe the pattern of missingness is informative, you can include it as a dummy variable.


THE UNIVERSITY OF
SYDNEY