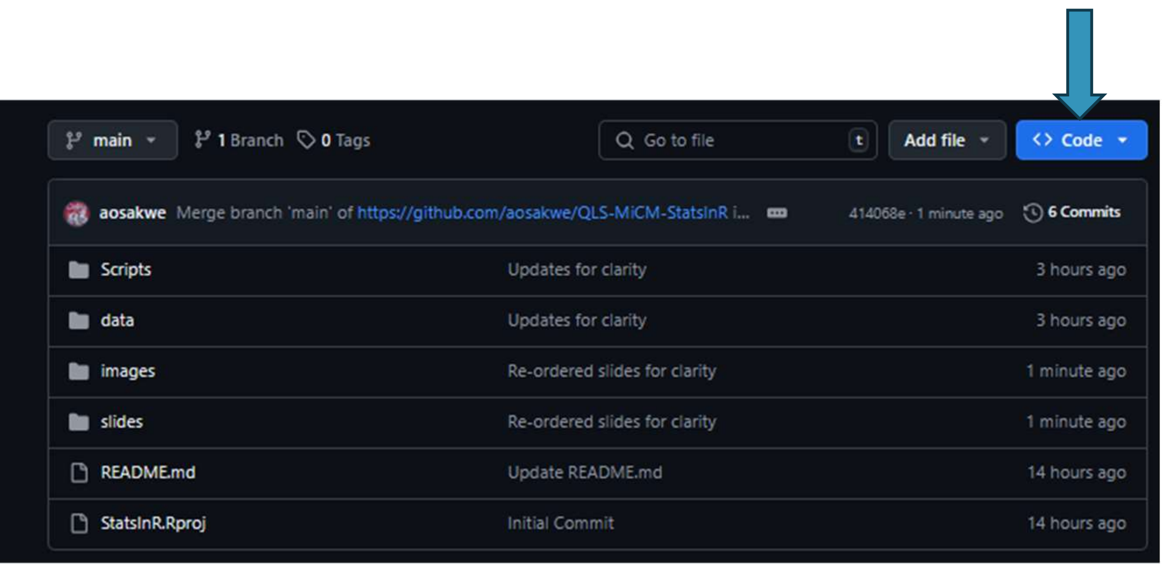
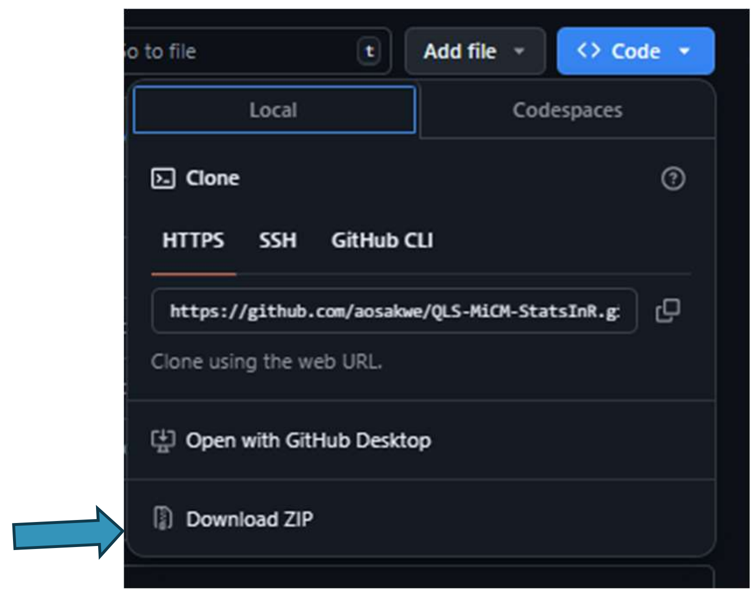
Workshop materials

https://github.com/aosakwe/QLS-MiCM-StatsInR

use ‘git clone’ command OR



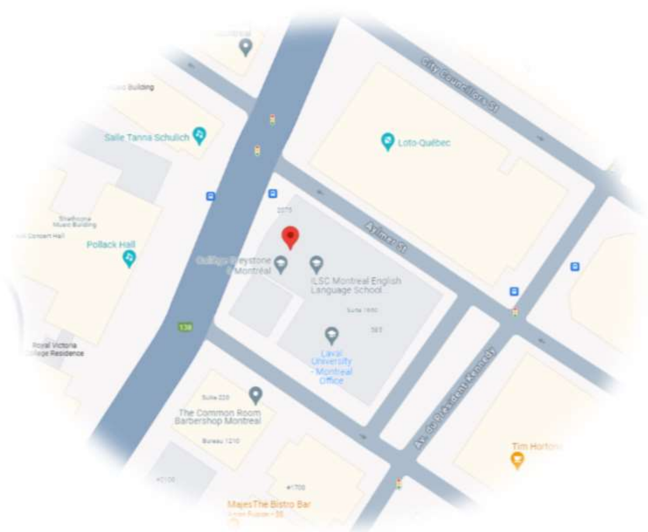


Statistical Analysis in R

Workshop Lead: Adrien Osakwe
Facilitator: Bangli Cao
February 20, 2025

McGill initiative in Computational Medicine

QLS-MiCM mission statement: deliver quality workshops designed to help biomedical researchers develop the skills they need to succeed.



Location: 550 Sherbrooke Street, Montreal, Quebec

Scan the QR code to sign up for our **mailing list**

Contact: workshop-micm@mcgill.ca

Workshop Series

Workshop	Date	Location	Registration
How to think in Code	Jan. 28 1PM-3PM	EDUC 133	Closed
Intro to Git & GitHub	Jan. 30 1PM-5PM	EDUC 133	Closed
Intro to Unix	Feb. 6 1PM-5PM	EDUC 133	Closed
Intro to Python (Part 1)	Feb. 11 1PM-5PM	EDUC 133	Closed
Intro to R (Part 1)	Feb. 13 1PM-5PM	EDUC 133	Closed
Exploring MATLAB	Feb. 18 1PM-5PM	EDUC 133	Closed
Statistics in R (Part 2)	Feb. 20 1PM-5PM	EDUC 133	Open
Data Processing in Python	Feb. 25 1PM-5PM	EDUC 133	Open
Intro to Machine Learning	Mar. 13 1PM-5PM	EDUC 133	TBA
Intro to R (Part 1)	TBA	EDUC 133	TBA
Intro to Python (Part 1)	TBA	EDUC 133	TBA

https://www.mcgill.ca/micm/training/workshops-series

Outline

1. Data Wrangling
2. Linear Regression Models
3. Logistic Regression Models
4. Statistical Testing
5. Study Design

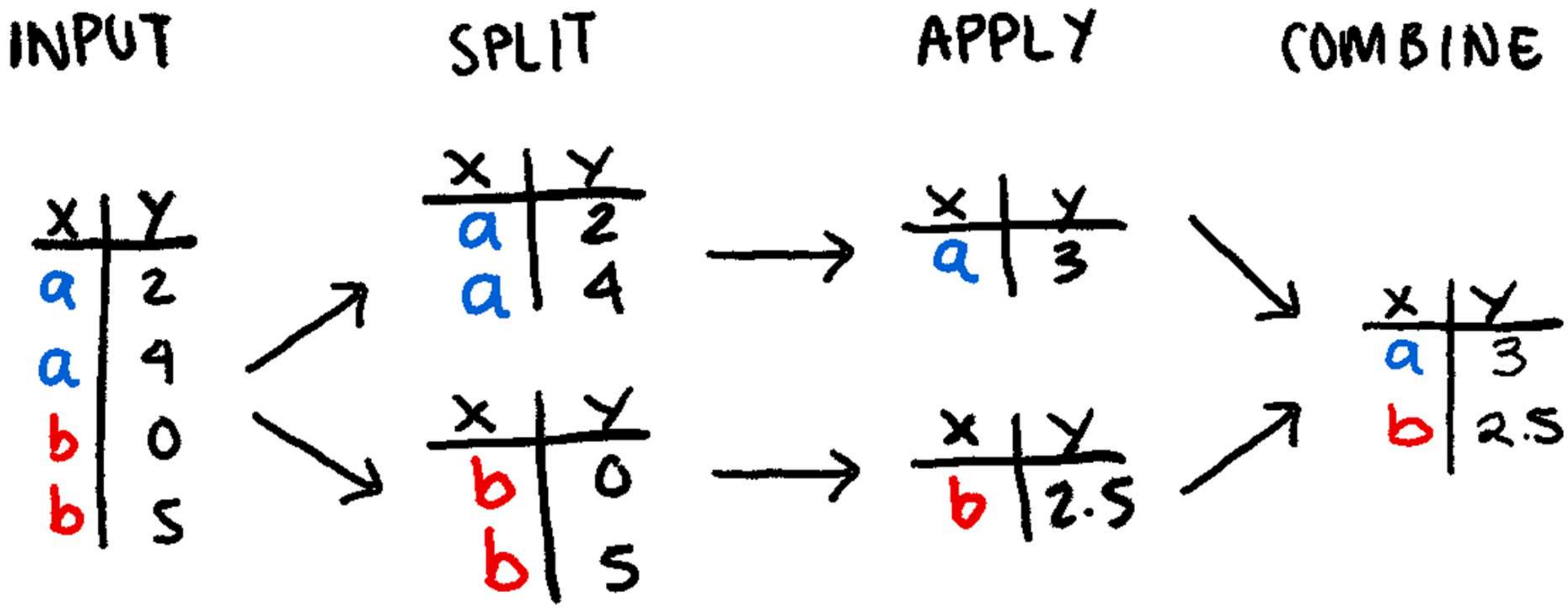
Acknowledgements
Gerardo Martinez - McGill
Alex Diaz-Papkovich - Brown
Larisa Morales Soto - HMS
Lisa Sullivan BUSPH

Data Wrangling

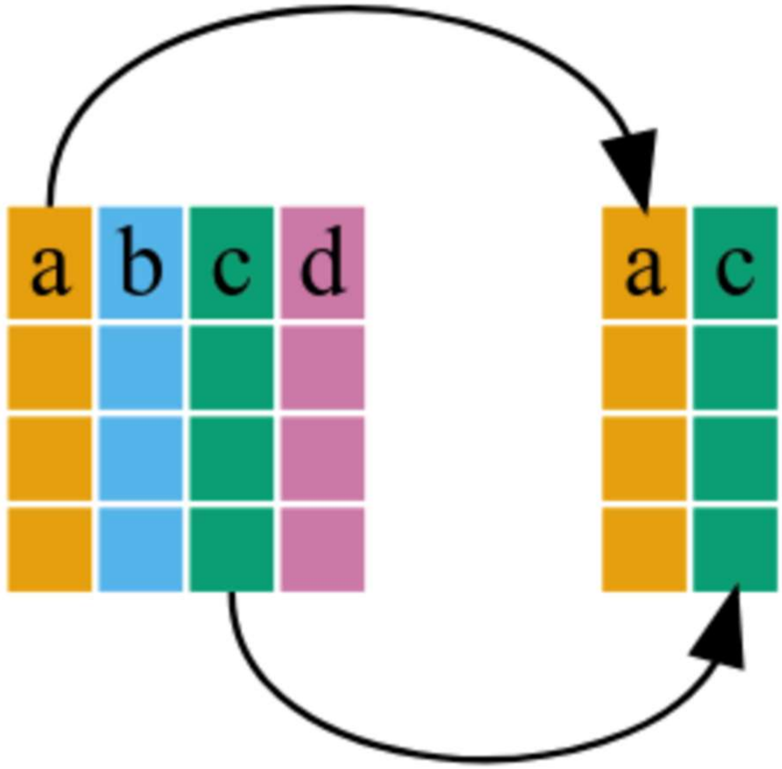
Learning objectives:

- Become familiar with the dplyr syntax
- Create pipes with the operator *%>%*
- Perform operations on data frames using dplyr and tidyr functions
- Examples for how to deal with missing data

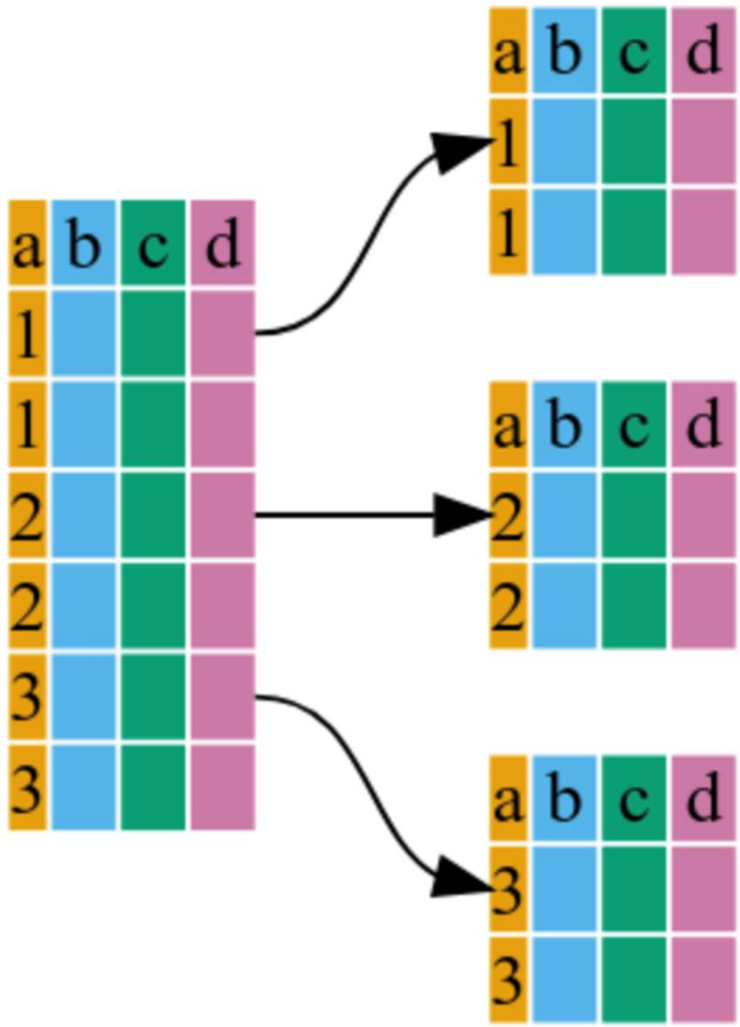
Split-Apply-Combine problem



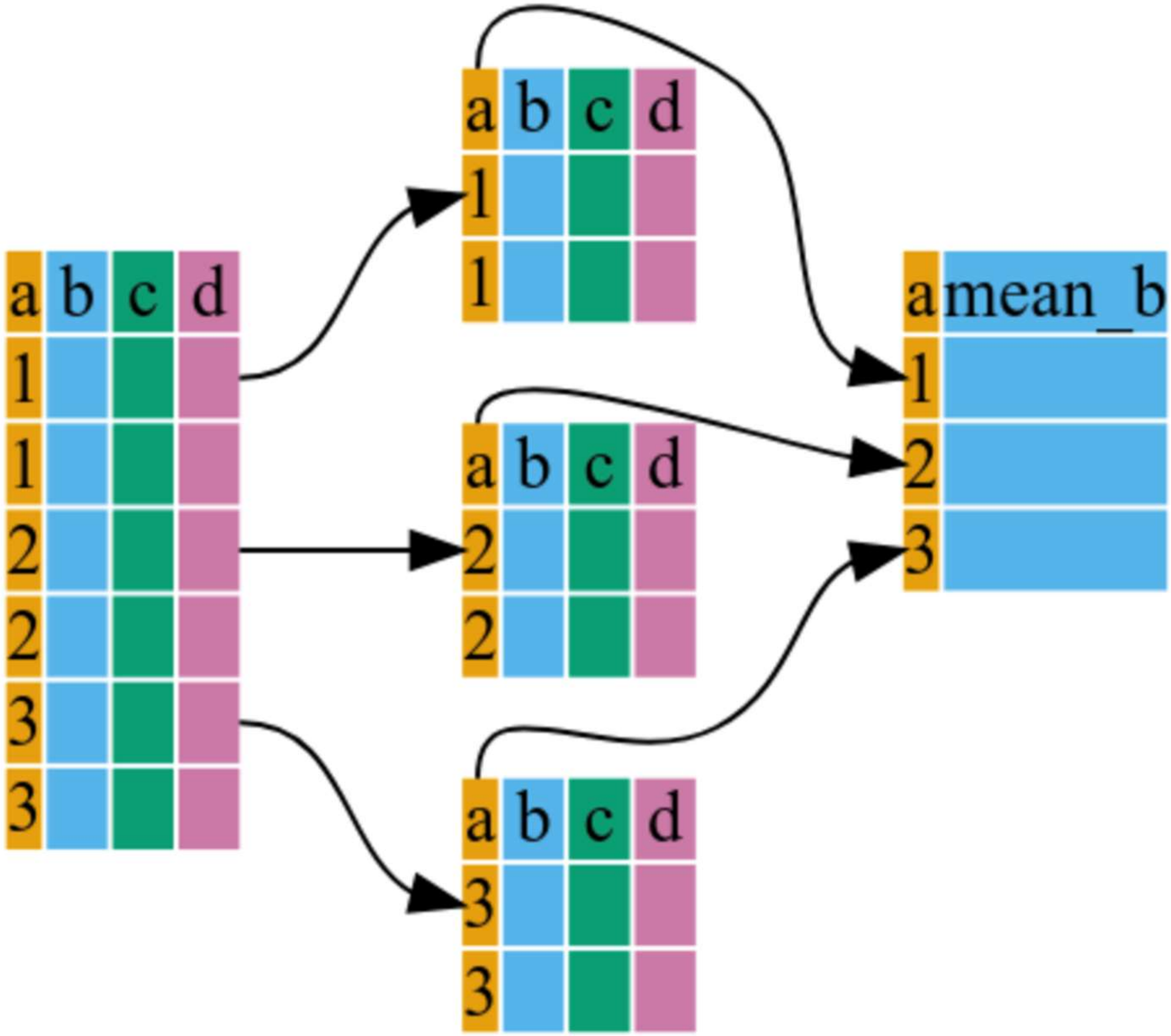
Select



Group by



Summarize



Managing Missing Data

- Most of the time, samples having missing entries
  - Missed medical appointment
  - Error with measuring instruments etc.
  - Represented as ***NA*** in R

Simple Solution

- Drop samples with missing observations

More rigorous solutions

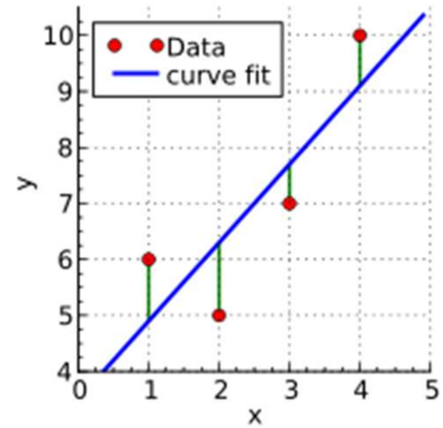
- Fill in with average/median of dataset or condition
- Train a model to 'impute' missing entries based on other features

Linear Regression

Learning objectives:

- Understand the fundamental regression algorithms
- What is a coefficient/effect size

The Linear Model



Credit: Wikipedia

- The fundamental model for statistics & machine learning
- Follows **a Normal distribution**

Imagine a dataset with a dependent variable ***y*** and a set of descriptive features ***x***

We want to learn what features in *x* are good predictors of *y* (what is their **effect size/coefficient** β)

$$\boldsymbol{y} = \boldsymbol{\beta x} + \boldsymbol{\varepsilon} \qquad \varepsilon \sim Normal(0, \sigma^2)$$

$$y \sim Normal(\beta x, \sigma^2)$$

Model Assumptions

1. Outcome is **continuous** and a **linear combination of predictors**
2. Outcome is such that $y_i \sim Normal(\beta x_i, \sigma^2)$
3. Predictors must not be perfectly correlated (linear combination)
4. For every observation *i* the error is:
   1. **Normally distributed**
   2. **Mean zero**
   3. **Homoskedastic** (same variance as other observations)
   4. **Independent** (not correlated)

Model Fitting

Probabilistic Approach

- **Maximum Likelihood Estimation** $\underset{\beta}{\operatorname{argmax}}\ log(Normal(\beta x, \sigma^2))$

Machine Learning

- **Mean Squared Error**

$$y = \beta x + \epsilon$$

$$\hat{y} = \hat{\beta} x$$

$$MSE = \frac{\sum_{i=1}^{N} (\widehat{y_i} - y_i)^2}{N}$$

Extensions

Interaction effects

- Outcomes depend on interactions of (e.g.) age*sex

Different Types of Data

- **Multivariate regression (multiple, correlated outcomes)**
- **Logistic regression (Binary outcomes like healthy vs. disease)**
- Multinomial/Poisson regression (Count data)
- ARIMA (Time series—data correlated over time)

Logistic Regression

Learning objectives:

- What is a logistic function
- How LR extends the linear model
- Cross-Entropy

Working with categories

Linear Regression could be used to predict categories (technically) BUT it has many flaws

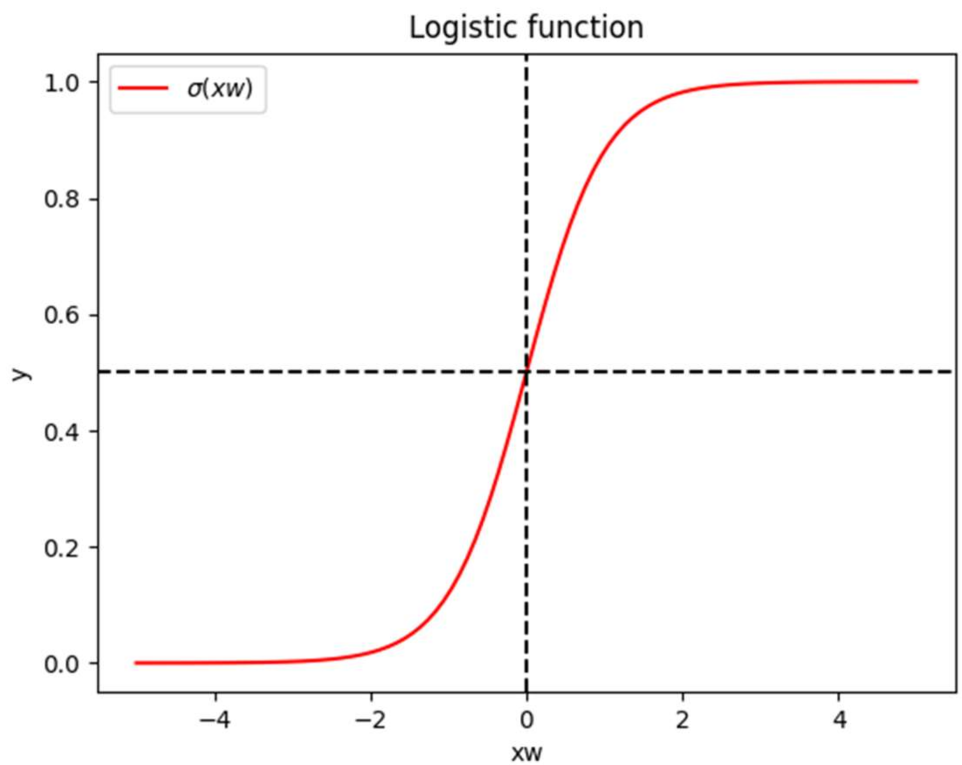
- Formulation isn't bounded (0,1)
- Impractical for predicting multiple classes
- Mean-Squared Error is not an optimal objective function

Propose to use the linear equation with a **sigmoid function**

Logistic Regression

$$\hat{y} = \sigma(\mathbf{xw}) = \frac{1}{1 + exp(-\mathbf{xw})}$$

- Sigmoid/Logistic function provides us with a bounded output
- Range is also easy to interpret as a probability



Can also use the **softmax** if working with multiple classes

$$\sigma(\mathbf{z})_i = \frac{e^{\beta z_i}}{\sum_{j=1}^{K} e^{\beta z_j}}$$

Cross-Entropy

$$CE(\hat{y}, y) = -ylog(\hat{y}) - (1-y)log(1-\hat{y})$$

Most commonly used loss function for classification

- More intuitive meaning for classification than MSE
- Faster model convergence than MSE
- Is compatible with a **multi-class dataset** (more than two classifications)

Statistical Testing

Learning objectives:

- Understand the formulation an assumption of standard statistical tests
- Understand the interpretation of a p-value
- Considerations when designing an experiment

Hypothesis Tests

Goal

- Determine if two sets of data are **different**
- Can approach this with **test statistics**
  - Is the difference **significant** or not?

Many Types of Tests

- Parametric tests – **make assumptions on the data distribution**
  - Z-score vs. t-test
  - ANOVA
- Non-parametric – **does not make assumptions on data dist.**
  - Permutation test

Hypothesis Formulation

Null Hypothesis

- Effect size is 0
- Difference between conditions is 0

Alternative Hypothesis

- Effect size is NOT 0
- Difference between conditions is NOT 0

We **accept or reject** the null based on the generated **p-value (< 0.05)**

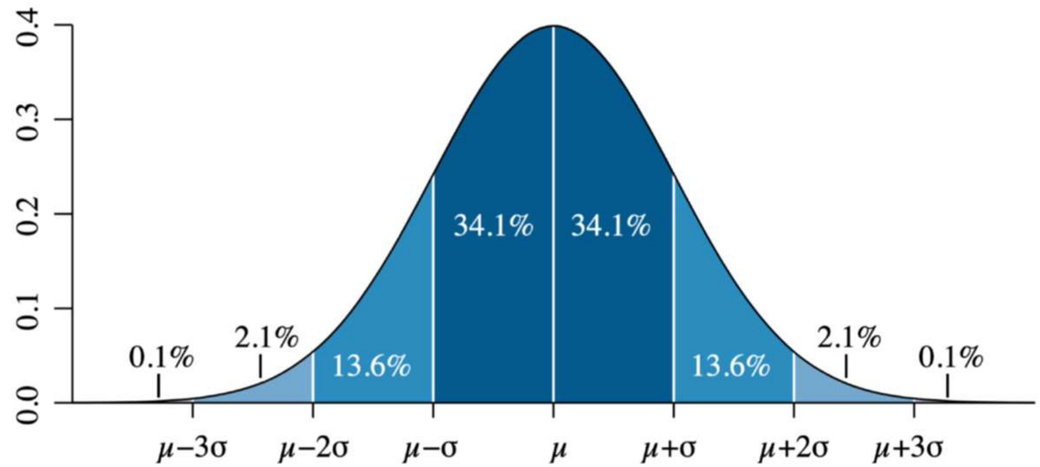
Understanding Significance

Statistical significance **DOES NOT** imply causality

It is a measurement of the likelihood of the observed data happening by chance.

A p-value tells us the probability of seeing the observation **if the null hypothesis is true.**

Z-score Test

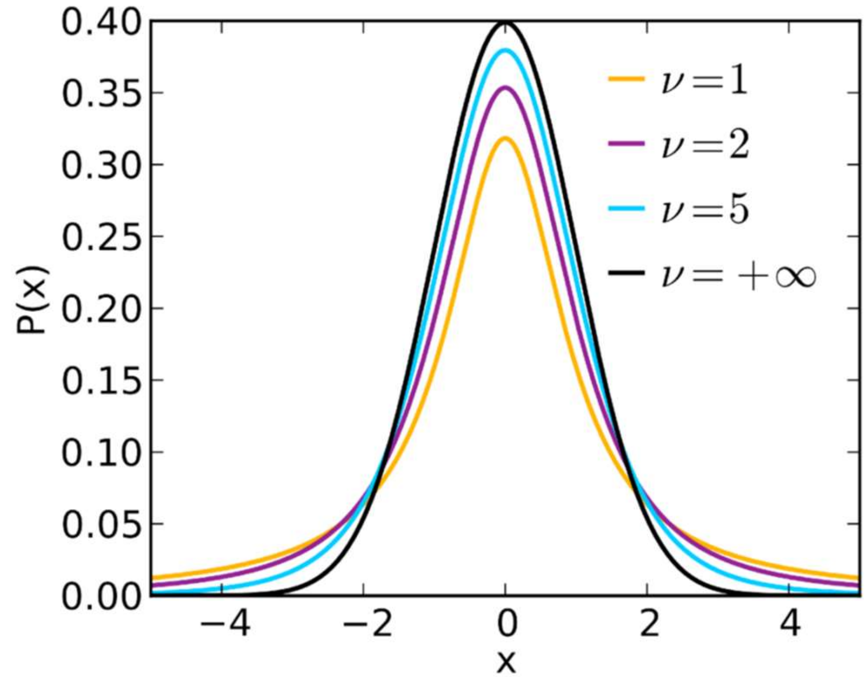


Credit: Wikipedia

- Based on the Normal Distribution
- Generate a Z-score from our effect size estimate
- Compare to Normal(0,1) distribution
  - One-tail or Two-tail
- Used when we have a **large sample size** (Central Limit Theorem)

$$z = \frac{\beta - \mu}{SE(\beta)}$$

t-test



Credit: Wikipedia

Based on the student t distribution

- Defined by degrees of freedom (d.f)
  - n-1 d.f.
- More appropriate for smaller sample sizes (most cases)
- Use our sample mean $\overline{X}$ and sample s.d $\hat{\sigma}$ to calculate test statistic
- Value of μ depends on your null hypothesis
  - One-Sample - 0
  - Two-sample - μ_2

$$t = \frac{\overline{X} - \mu}{s} = \frac{\overline{X} - \mu}{\hat{\sigma}/\sqrt{n}}$$

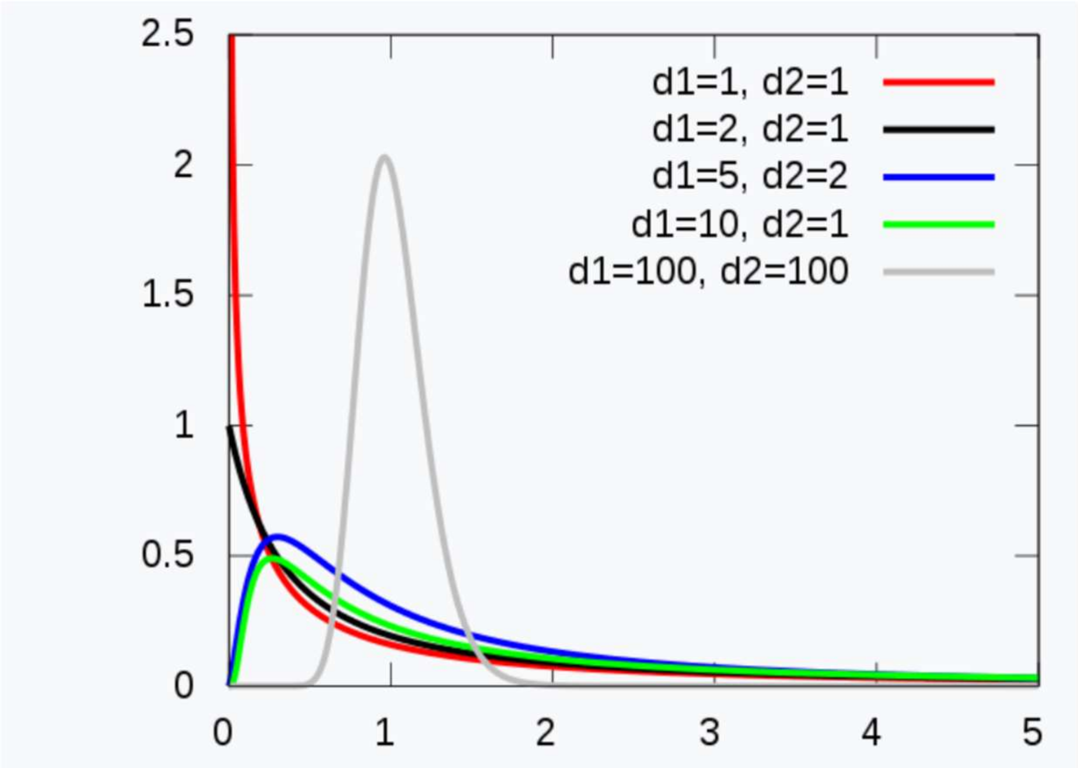
ANOVA

Based on the F-distribution

- Similar to t-statistic
- Statistic is the ratio of two Chi-squareds

Null hypothesis: multiple means are **all** equal

Alternative: At least one group is different



Credit: Wikipedia

ANOVA

How much variation is explained by our treatment condition?

$$SST = \sum_{i=1}^{n} (y_i - \overline{y})^2$$

Total sum of squares

$$SST = SS(Treatments) + SS(Residuals)$$

Take the ratio of the variance (mean sum of squares) between treatments over the variance within treatments

High variance between treatments will give a low p-value

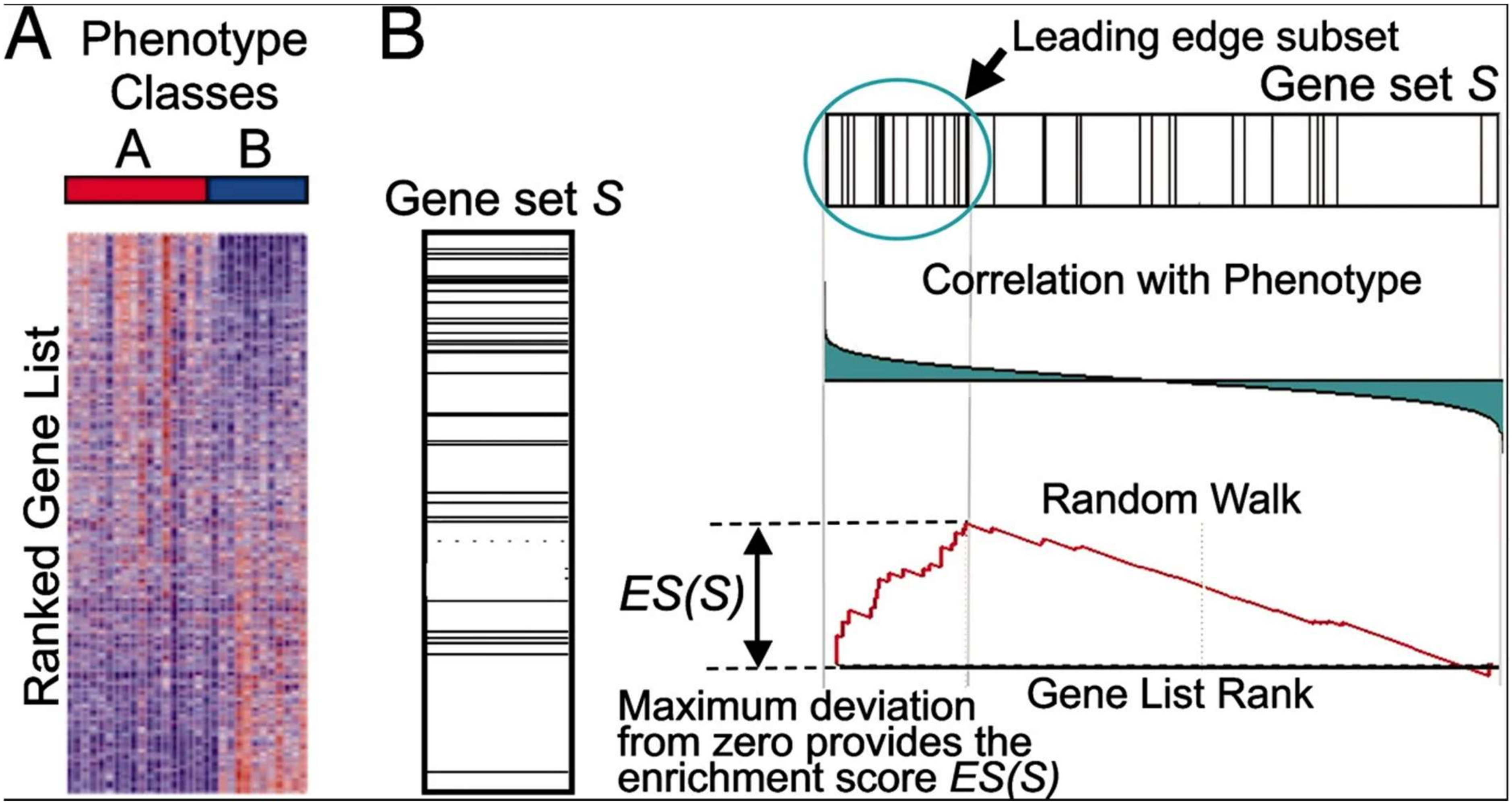
Non-Parametric Test

Similar to what is done for enrichment analyses

- Avoids needing to make assumptions on distributions
- Can be very **slow**

1. Shuffle labels in your data to generate estimates (random baseline)
2. See where your true observation lies on this baseline

Gene Set Enrichment



Credit: GSEA

Multiple Testing

- May test multiple hypothesis (example, effect sizes in linear model)
- If null distribution is true, expect a **uniform distribution** of p-values
    - 5% of tests will give a significant result
- Multiple test correction methods are used to address this
    - Bonferonni Correction
    - **Benjamini & Hochberg Correction**
    - Found in results as **FDR, p-adjusted, q-value**

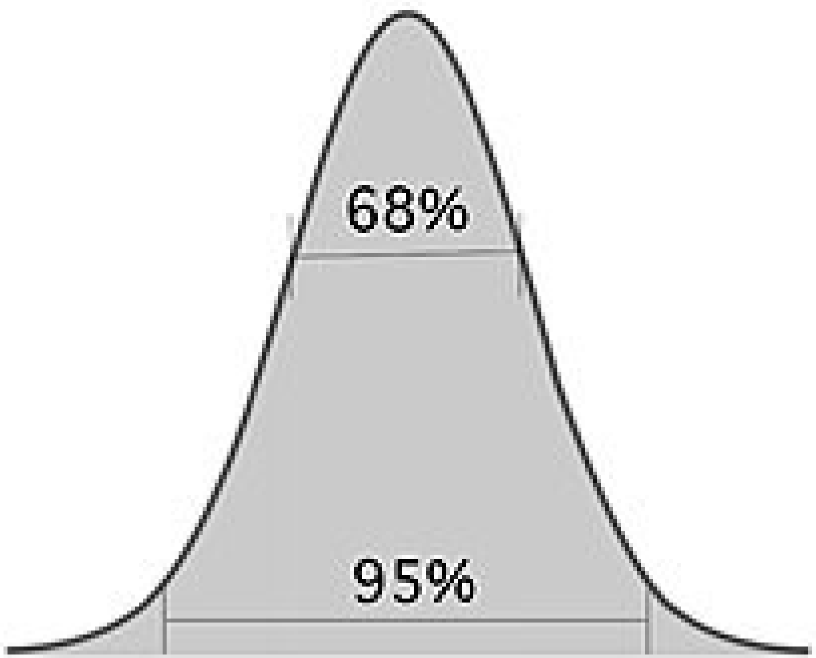
Confidence Intervals

- We may want to attribute a range to our estimate

- Can be calculated with the **margin of error**

$$E = Z\frac{\sigma}{\sqrt{n}}$$

- Where Z is the z-score that defines our confidence level

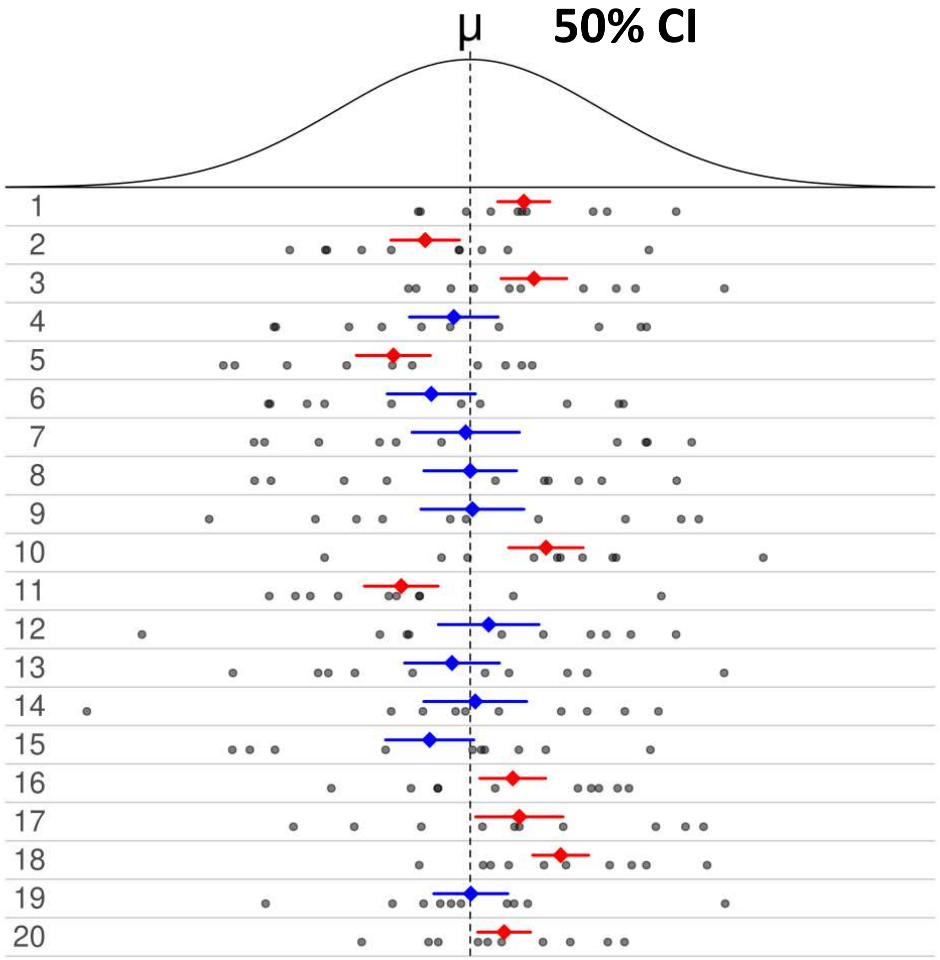


PSA: Confidence Intervals

A 95% CI **DOES NOT mean** there is a 95% chance the true value is in that range

For multiple experiments, 95% of CIs will overlap with the true value

Consider a Bayesian formulation instead (beyond scope)



Study Design

Learning objectives:

- Understand how simulations can be used to help with study design
- How to determine the required sample size

Study Design

Data collection can be expensive and time-consuming

Need to ensure that the collected data has enough **power** to avoid false negatives (type II error)

Can simulate data to help decide on the statistical test used and determine the appropriate **sample size**

Determining sample size

- Statistical test results depend on multiple factors
  - Difference in Samples
  - **Sample Size**
- We want to make sure we have enough samples to avoid False Negatives (type II errors)

- Can leverage the assumptions used for our statistical test to determine sample size

Confidence Intervals

In practice you will want your CI to cover a small range to be informative.

Can achieve this by solving for the **margin of error E**

$$E = Z\frac{\sigma}{\sqrt{n}}$$

$$n = \left(\frac{Z\sigma}{E}\right)^2$$

Note that we may not have a value for σ

- Use estimate from past study
- Make an assumptive guess
- Run a small pilot study

Sample Comparison

$$n_i = 2\left(\frac{Z_{1-\frac{\alpha}{2}} + Z_{1-\beta}}{ES}\right)^2$$

$$\textit{Effect Size (ES)} = \frac{|\mu_1 - \mu_2|}{\sigma}$$

α → Significance Level

Probability of rejecting null when it is true

1-β → Power

Probability of rejecting null when it is false

To summarize

- ✓ Basics for data manipulation & cleaning
- ✓ Explored basic linear modeling for regression & classification
- ✓ Explored statistical testing and intuition behind study design

Now you are ready to:

- Use dplyr to facilitate reproducible data manipulation
- Apply and interpret linear model results
- Interpret confidence intervals and p-values
- Explore more complex models and tests!

Future Statistics Workshops

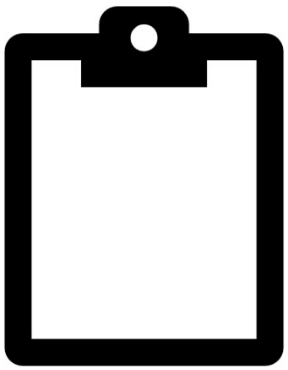
- Dimension Reduction
- **Bayesian Inference**
- Time-series Analysis
- And more…

Thank you for attending!

1

Scan the QR code to confirm you attended today's workshop.

2



Fill out the feedback survey in the next 72h.

3

Get recognition for this workshop on your co-curricular record.