Introduction to Statistics and Machine Learning

Oliver M. Crook

DAMTP, Department of Biochemistry, MRC Biostatistics Unit
University of Cambridge

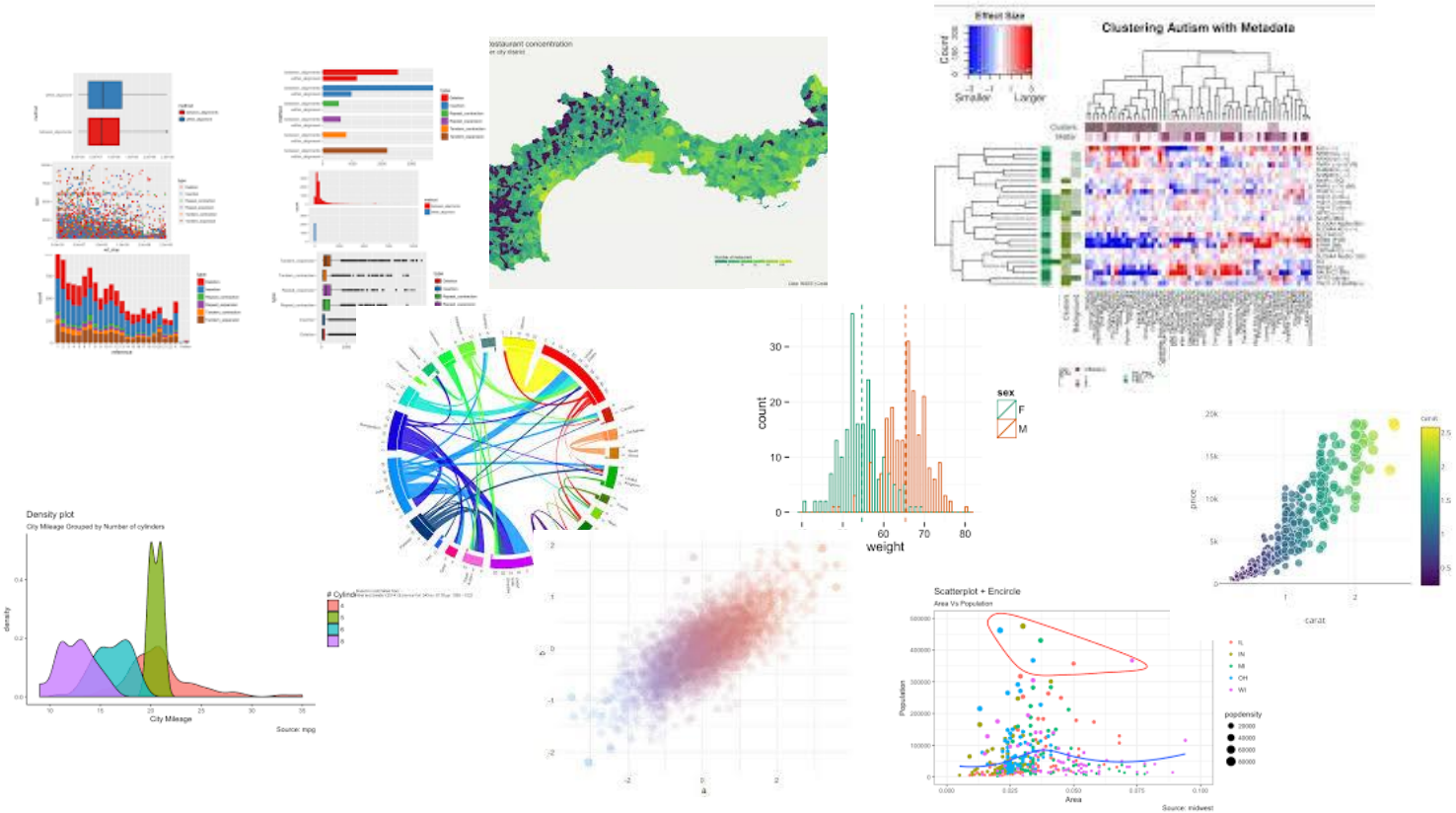
23 July 2019

Summary

What is statistics/machine learning ?

- Learn from data
- Draw inference from data to make scientific conclusions
- Test scientific hypothesis quantitatively
- Make sound predictions
- Visualising raw data, visualise processed data and visualise conclusions

Visualisation, visualisation, visualisation



What does a hypothesis test aim to do ?

Hypothesis Testing

- We want to draw conclusions or make decisions based on data.
- Screening of potentially millions of possible avenues to follow up
- Example genome-wide association studies test millions of variants
- Infeasible to looks at every variant functionally
- Only follow up variants that statistics "recommends"

Example : Is my coin biased ?

I think my coin is probably fair.

I *expect* roughly a 1:1 ratio of heads to tails or 50 heads for every 50 tails

Hypothesis: My coin is fair.
Question: Is my coin fair?

Observe (my experiment):

HHHTHHTHHTTH….

Observe (my data):

Heads	Tails
59	41

Maybe my coin is biased? But how do I know?

Question: How likely is it that this observation is a fluke?
Question: What is the probability of this event happening, given that my coin was fair?

First observations

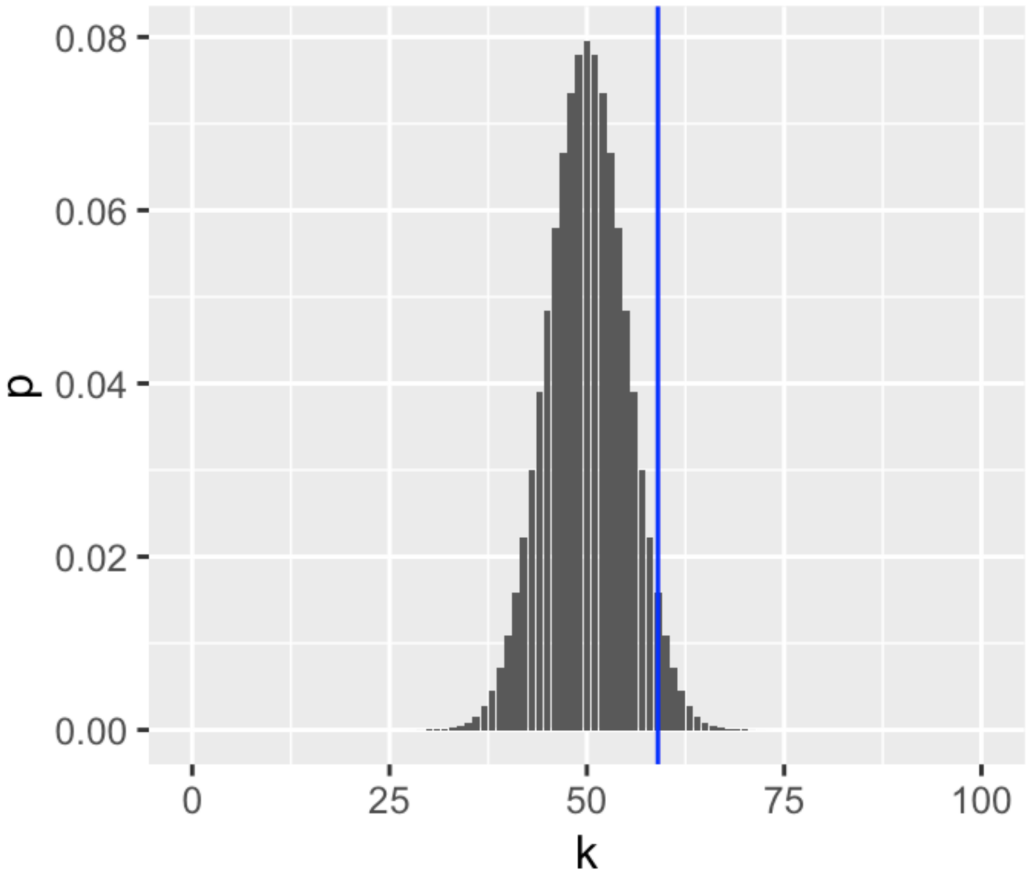
Observations

- Statistics started before we did any experiments
- We made some statements before any experimentation
- We formed a hypothesis (which we call the null hypothesis)
- We designed an experiment and observed data
- We asked a clear statistical question

...yes, but is my coin biased ?

Our test statistic

$$P(K = k|n, p) = \binom{n}{k} p^k (1-p)^{n-k} \tag{1}$$

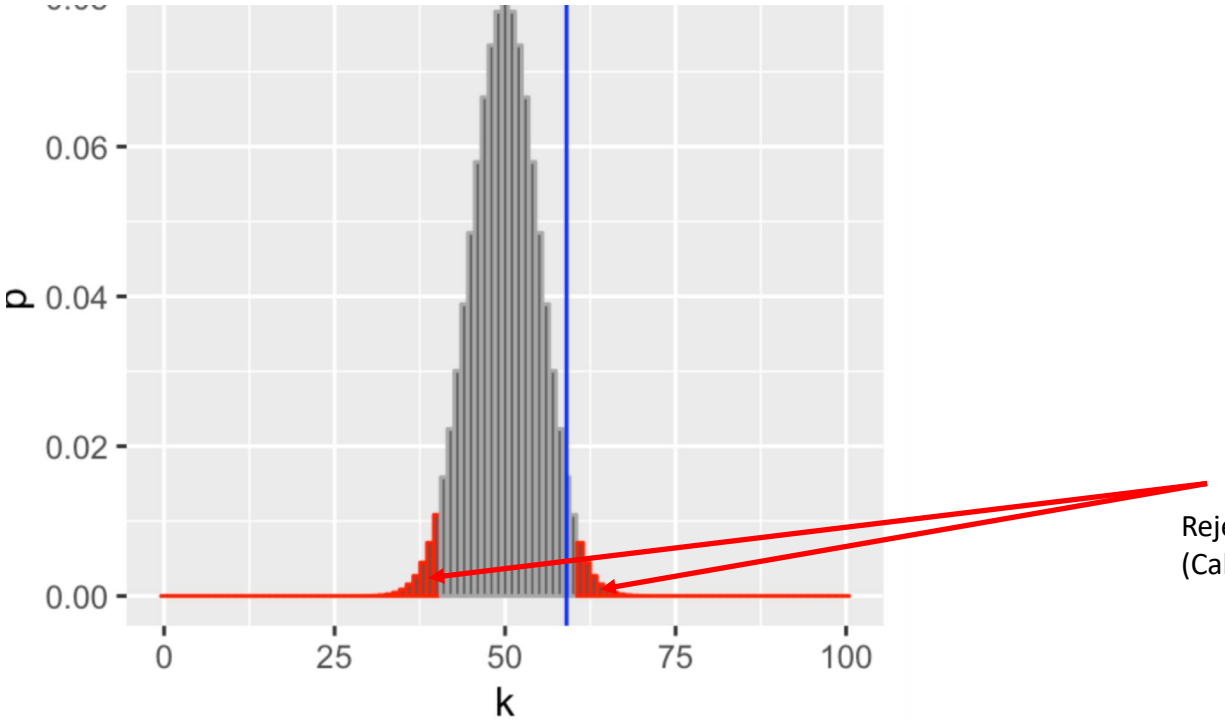


...yes, but is my coin biased ?

- The blue line seems unlikely
- What do I need to do to concluded the coin is biased ?
- Form a region, called the reject region.
- Fill the reject region with as many possible values of k such that total probability is less than 0.05.
- These events are unlikely by definition (they have probability less than 0.05)

the reject region

(Explicit example in practical)



The blue line does not fall in the reject region. We do not reject the hypothesis that the coin is fair. If we saw 75, say, heads we would have rejected the hypothesis that the coin is fair.

The steps to hypothesis test success

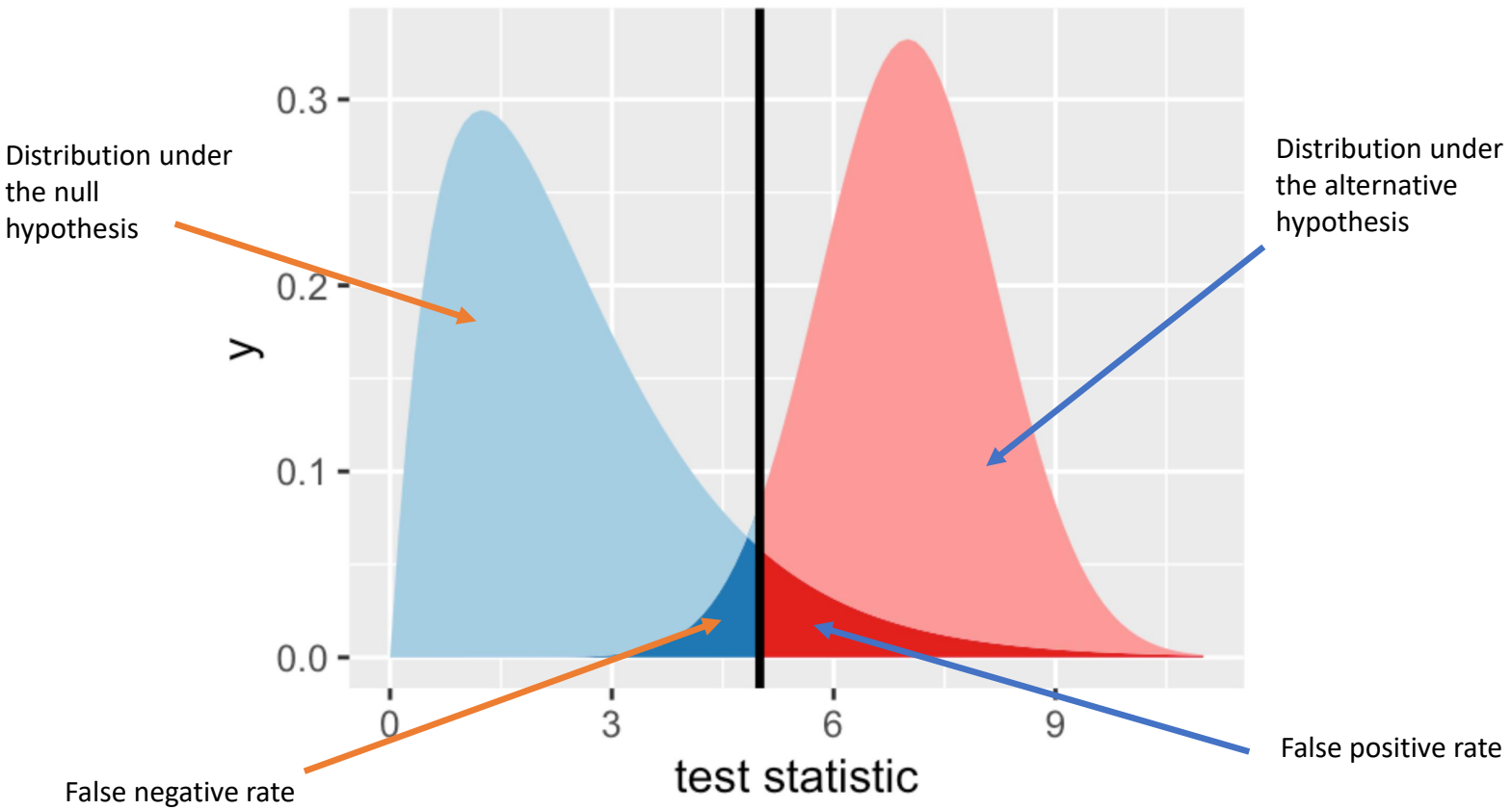
The hypothesis test journey

- Decide on a quantity you are interested in.
- Design a suitable experiment.
- Choose a data summary.
- Decide on a test statistic.
- Set up a simple null hypothesis. (You need to be able to do computations)
- Decide on the rejection region.
- Do the experiment and collect the data.
- Compute the test statistic.
- Make a decision.

Definition, Definition, definition

- Significance level or false positive rate α : is the total probability of the test statistic falling into this region even if the null hypothesis is true.
- We now its size, but what's it shape ?
- We want the rejection region to be as large possible if the alternative hypothesis is true.
- We want high power (or true positive rate).

The null and the alternative



Types of Errors

Test/Real World	Null hypothesis is true	Null hypothesis if False
Reject null hypothesis	Type I error (false positive)	True positive
Do not reject null hypothesis	True negative	Type II error (false negative)

Difference between groups : The t-test

Imagine the following scenario :

Plant heights

- I grow 100 plants outdoors and they grow to some height
- I wonder if they would grow better in the greenhouse
- I decide to do an experiment
- I take 50 plants and put them in the greenhouse ; 50 remain outside
- I measure their heights after 5 days.
- Did the greenhouse has a positive effect ?

Difference between groups : The t-test

What is my test-statistic ? The t-test is the simplest such statistic

$$t = c\frac{m_1 - m_2}{s} \tag{2}$$

$$c = \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \tag{3}$$

and m_1 is the mean of group 1 and m_2 is the mean group 2 and s is the pooled standard deviation.
To compute a p-value the t-test uses some asymptotic theory.

The t-statistic

This theory states that under the following assumptions

Assumptions of t-test

- assume the null hypothesis of equal means in both groups
- assume that the data are independent and are normally distribution
- assume the standard deviation of both groups are equal

Then the statistic follow a t-distribution with $n_1 + n_2$ degrees of freedom.
(what happens if we break these assumptions ?)

Difference is contingency table : Fisher's exact test

Imagine the following scenario

- My plants have different colour peas : yellow or green
- I have some plants in the greenhouse and some in the garden
- I make the following table

	Outdoors	Greenhouse	total
Yellow	1	9	10
Green	11	3	14
total	12	12	24

Example : Contingency Tables

- Question : Are peas grown in the greenhouse more likely to be yellow ?
- The test statistic in this scenarios is the hyper-geometric distribution.

	Outdoors	Greenhouse	total
Yellow	a	b	a + b
Green	c	d	c + d
total	a + c	b + d	a + b + c + d = n

The statistic

The probability of seeing this table is

$$p = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}} \tag{4}$$

Again, easily to use function in R. The exact refers to the fact that exact p-values are computed (no asymptotic theory).

Bigger tables needs different statistics

In this case, use the chi-squared test.

	Shelf	Bedroom	Outdoors	Greenhouse	Total
Yellow	10	3	4	10	27
Green	7	10	10	4	31
Blue	3	1	0	3	7
Brown	10	12	1	2	25
total	30	26	15	19	

There are many, many, many test

Many tests

- Tukey's test
- Wald test
- Kolmogorov-Smirnov Test
- Test's for normality
- many non-parametric tests

Permutation tests

- A general type of test that build up a sampling distribution by re-sampling from the observed data.
- We do this by shuffling
- Assume we have several greenhouses with lots of plants
- We compute the variance of the heights of plants in each greenhouse σ_1^2, σ_2^2, σ_3^2,...
- Then we permute the plants between the greenhouses and recompute
- Repeat this many times and we get a distribution of variance across each greenhouse
- We can then test whether any of the greenhouse has larger variance than we would expect at random
- Rank the σ_1^2 against the distribution of variances for greenhouse 1.
- The p-value in this scenarios is the rank of σ_1^2 divided by the number samples
- See example in practical

Multiple Testing (a brief tour)

The family wise error rate

The family wise error rate (FWER)

Is the probability that we make one or more false positive errors.

$$P(V > 0) = 1 - P(\text{no rejection of any of } m \text{ null hypothesis}) = 1 - (1 - \alpha)^m \quad (5)$$

This grows to 1 very quickly ! Thus, if we are making a million test ; for example, testing if two people DNA matches, then a false positive is inevitable

Bonferroni Correction

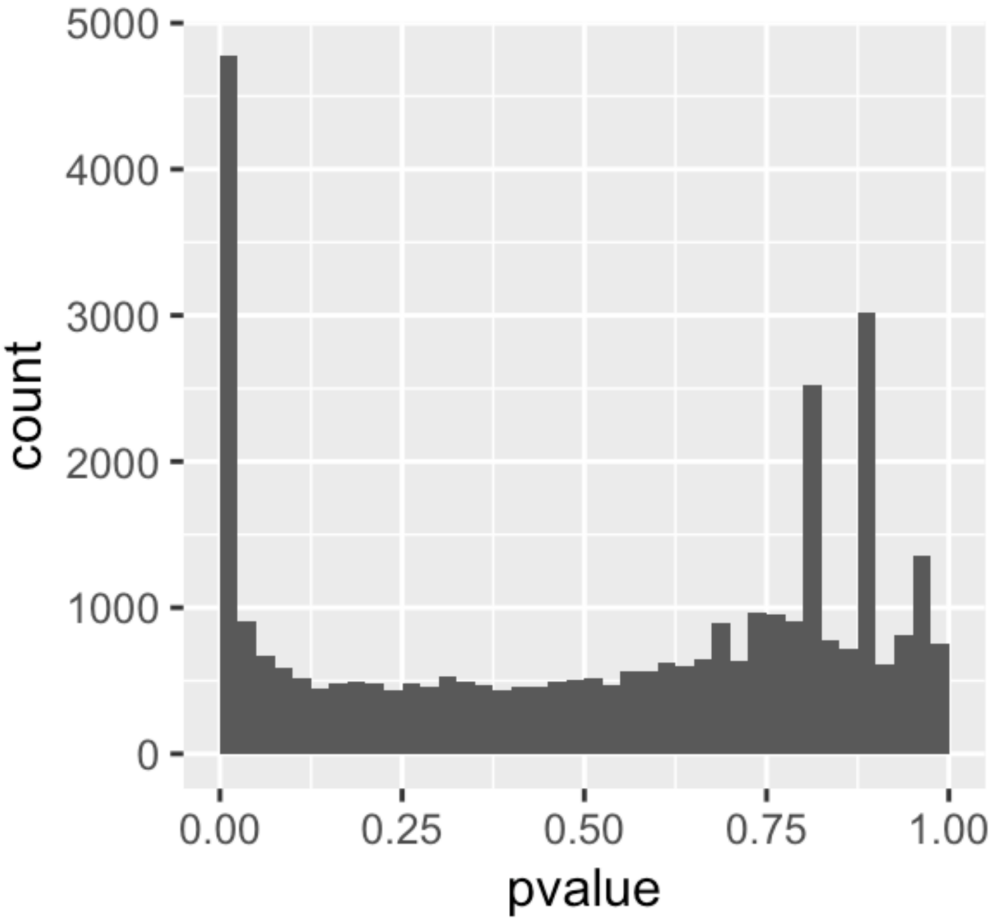
Bonferroni tells us that if we want to control the FWER at level α_{FWER}, then the individual hypothesis threshold is α/m

The false discovery rate

p-value histogram

- plot the p-value histogram
- This is simply a histogram of the computed p-values
- We expect a mixture distribution of two components
- The first component corresponds to the p-values resulting from the tests for which the null hypothesis is true.
- The second to the p-values resulting from the tests for which the null hypothesis is not true

The p-value histogram



The FDR definition

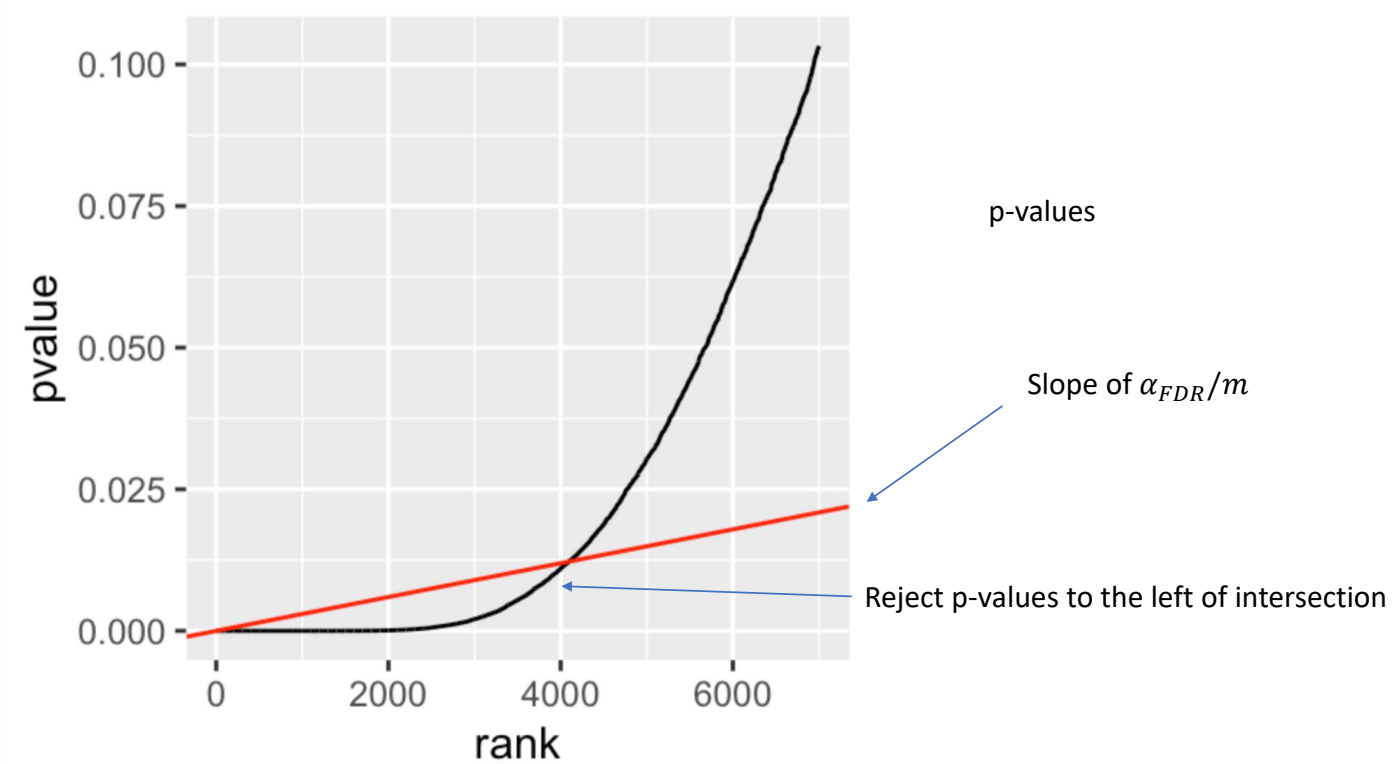
False Discovery rate

$$\text{FDR} = E\left[\frac{\text{number of false positives}}{\max(\text{Number of rejections}, 1)}\right] \tag{6}$$

Benjamini-Hochberg correction

1. First, order the p-value in ascending order $p_1, ..., p_m$
2. For a FDR theshold of level α_{FDR}, find the largest value of k such that $p_k \leq \alpha_{FDR} k/m$
3. Reject all hypothesis $p_1, ..., p_k$

visualisation



Linear models

What is a linear regression model ?
A straightforward approach for predicting a quantitative response Y on the basis of a variable X
We often write

$$Y \approx \beta_0 + \beta_1 X \tag{7}$$

Observations

- This is clearly a linear relationship
- β_0 is referred to as the intercept
- β_1 is the slope
- Often we say that we are regressing Y on X
- β_0, β_1 are coefficients and need to be inferred from the data.

Estimation

notation

- coefficient are unknown need to estimate them
- estimates are given a hat e.g. $\hat{\beta}_0$
- The predicted value of y at some new points x_* is given as

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_* \tag{8}$$

- We have data $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$
- how do we find the parameters ?
- Let $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ be the prediction for Y based on the ith value of X.
- Let $e_i =$ True value $-$ predicted value $= y_i - \hat{y}_i$
- we call e_i the ith residual
- We compute the so-called *residual sum of squares* (RSS) as

$$RSS = e_1^2 + e_2^2 + ...e_n^2 \tag{9}$$

Least squares

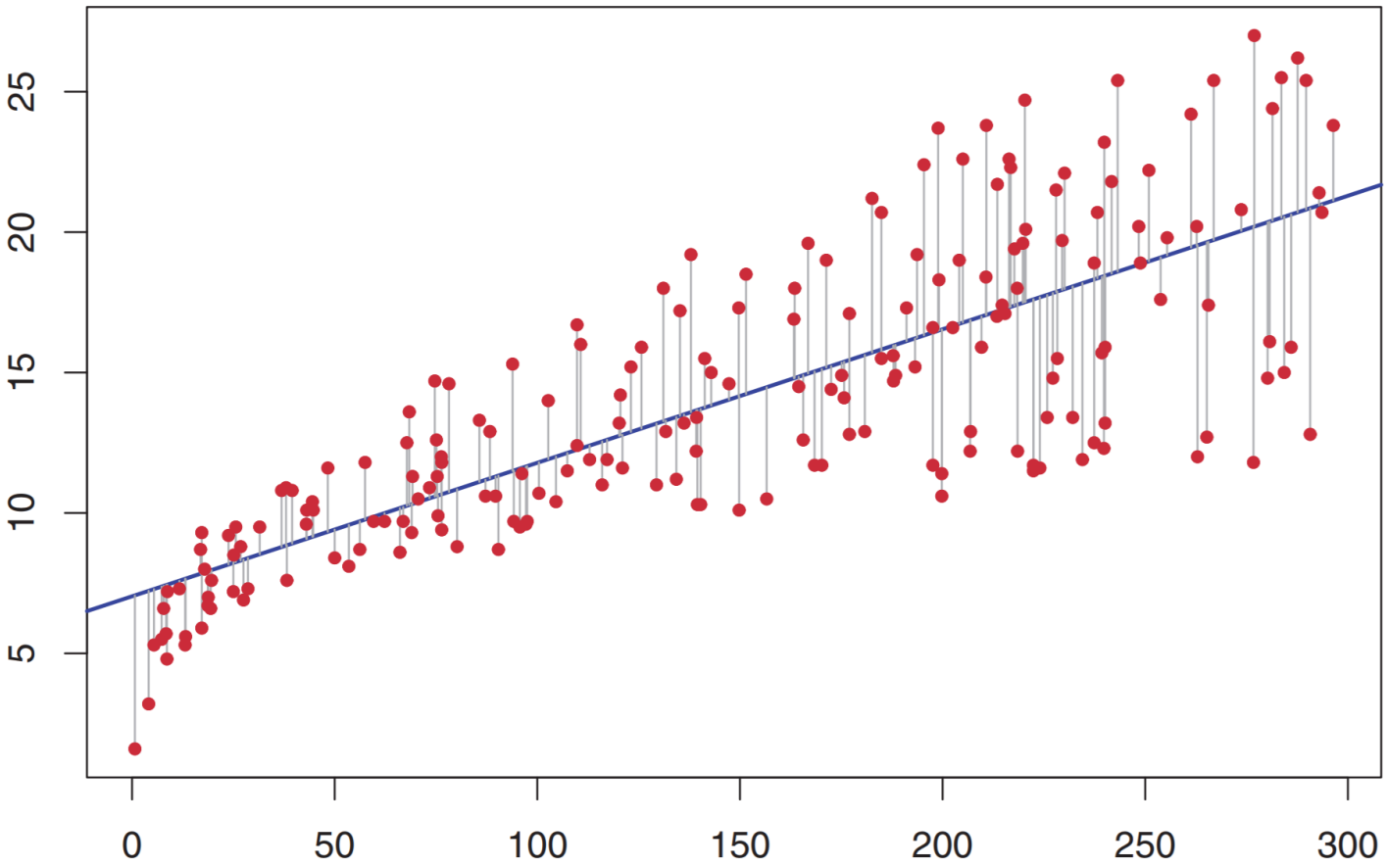
The least squares method choose the coefficient $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimise the RSS. This leads to

$$\hat{\beta}_1 = \frac{\sum_{n=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{10}$$

and

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}, \tag{11}$$

where the bar denotes the sample means.


25
20
15
10
5
0
50
100
150
200
250
300

How good is by approximation ?

A more formal problem set-up is that our observation arise from a linear model but are then corrupted by independent noise

$$Y = \beta_0 + \beta_1 X + \epsilon, \tag{12}$$

where

$$\epsilon \sim \mathcal{N}(0, \sigma^2) \tag{13}$$

(the tilde/squiggle means distributed as)
We can now ask how close are $\hat{\beta}_0$ and $\hat{\beta}_1$ to the "true" β_0 and β_1 ?
We can compute the standard errors (SE) :

$$SE(\hat{\beta}_0)^2 = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right] \tag{14}$$

and

$$SE(\hat{\beta}_1)^2 = \frac{\sigma_1^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{15}$$

Confidence Intervals

For linear regression, we can now compute the 95% confidence interval

$$\hat{\beta}_1 \pm 2SE(\hat{\beta}_1) \tag{16}$$

We say that the the true value of β_1 has a 95% chance of falling into the interval.

$$\left[\hat{\beta}_1 - 2SE(\hat{\beta}_1), \hat{\beta}_1 + 2SE(\hat{\beta}_1)\right] \tag{17}$$

warning

The 2 in front of the standard error means that this is only approximate. We actually need to put the 97.5% quantile of a t-distribution with $n-2$ degrees of freedom (more later)

linear models as a Hypothesis test

We might want to test the following hypothesis

$$H_0 : \text{There is no relationship between } X \text{ and } Y \tag{18}$$

and

$$H_a : \text{There is some relationship between } X \text{ and } Y \tag{19}$$

This is equivalent to

$$H_0 : \beta_1 = 0 \tag{20}$$

and

$$H_1 : \beta_1 \neq 0 \tag{21}$$

As we've seen before we need to see whether our estimate $\hat{\beta}_1$ is sufficiently far from 0. We compute the statistic

$$t = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)} \tag{22}$$

(This is our good old friend the t-statistics).

- If there really is no relationship between X and Y, then the above statistics has a t-distribution with $n - 2$ degrees of freedom.
- You can then calculate the p-value (using pt in R).
- Intuitively we are concerned about whether the magnitude of the coefficients are much bigger than the standard error.

Goodness of fit

How good is our model of the data ?

R^2

- Many ways to measure model fit
- How much variance does our model explain ?
- Want quantity to be independent of the scale of the data
- R^2 measures the proportion of variance explained by a model

$$R^2 = \frac{\text{TSS} - \text{RSS}}{\text{TSS}} \tag{23}$$

where

$$\text{TSS} = \sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{24}$$

- R^2 measure the proportion of variability in Y that can be explained using X

Multiple Regression

We have only worked with one predictor but in reality we have more than 1 predictor. For example, the height of plant is a function of both amount of sun and amount of water (and whether or not nutrients were added etc)

multiple regression

- Suppose we have p possible predictors (variables that may effect our outcome)

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \epsilon \tag{25}$$

- The least squares approach (i.e. minimising the RSS) can be used to estimate parameters

Multiple Regression questions

multiple regression questions

- Is at least one predictor useful for predicting the response ?
- Are a subset of predictors useful ?
- Model fit ?

Answers

- This is part of variable selection and can use criterion such as BIC and AIC
- More sophisticated variable selection methods (time restriction)
- The R^2 can be used to assess model fit as before

Hypothesis for multiple regression

Are of of my predictors useful ?

$$H_0 : \beta_1 = \beta_2 = \beta_3 = ... = \beta_p = 0 \tag{26}$$

against

$$H_a : \text{ at least one of } \beta_j \text{ is non-zero} \tag{27}$$

We need a more complicated test statistics. We compute

$$F = \frac{(\text{TSS - RSS})/p}{RSS/(n-p-1)} \tag{28}$$

large values of F indicate evidence against the null. But what about p-values ? unsurprisingly, the F-statistic follows the F-distribution (pf in R)

warning

Why didn't wee test each variable separately using the t-statistic ?

Linear models with factors

Up until this point we have only considered linear models with quantitative variables. We can also include factors, such as treated or untreated, male or female, Human or mouse etc.

factors

Consider plant height with treatment or no treatment of fungicide
$x_i = 1$ if treated $x_i = 0$ not treated

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \tag{29}$$

β_0 is now interpreted as treated average height of untreated plants
$\beta_0 + \beta_1$ is the average height of treated plants
β_1 is the average difference in heights between the treated and untreated

Question

How do I test if the treatment has an effect on height ?

Linear models with more factors

Suppose there are now two possible treatments.

Multiple factors

Use the 0 or 1 coding for each treatment

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i \tag{30}$$

How do I interpret these coefficients ?
Does a treatment improve height ? Use the F-statistic for

$$H_0 : \beta_1 = \beta_2 = 0 \tag{31}$$

Interaction effects

What if we think that the two treatments together my have an additional *iteraction* effect

Multiple factors and interaction effects

Use the 0 or 1 coding for each treatment, include an additional product term.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \boldsymbol{\beta_3 x_{i1} x_{i2}} + \epsilon_i \tag{32}$$

The coefficient β_3 explains the additional effect of having both treatments over the additive benefit of each individual treatment
We can apply all the machinery we've done before. For example we can test to see whether β_3 is significant and conclude there is an interaction effect.

Warning : The hierarchical principle

If we include an interaction in a model, we also include the main effects, even if the p-values associated with their coefficients are not significant.

Further Models

To cover sufficient breadth, we missed out some important topics.

linear models are linear in the parameters

The following is a linear model, even though there is a non-linear transformation of the predictors

$$y_i = \beta_0 + \beta_1 x_{i1}^2 + \beta_2 \sin(x_{i2}) + \beta_3 x_{i1} \sin(x_{i2}) + \epsilon_i \tag{33}$$

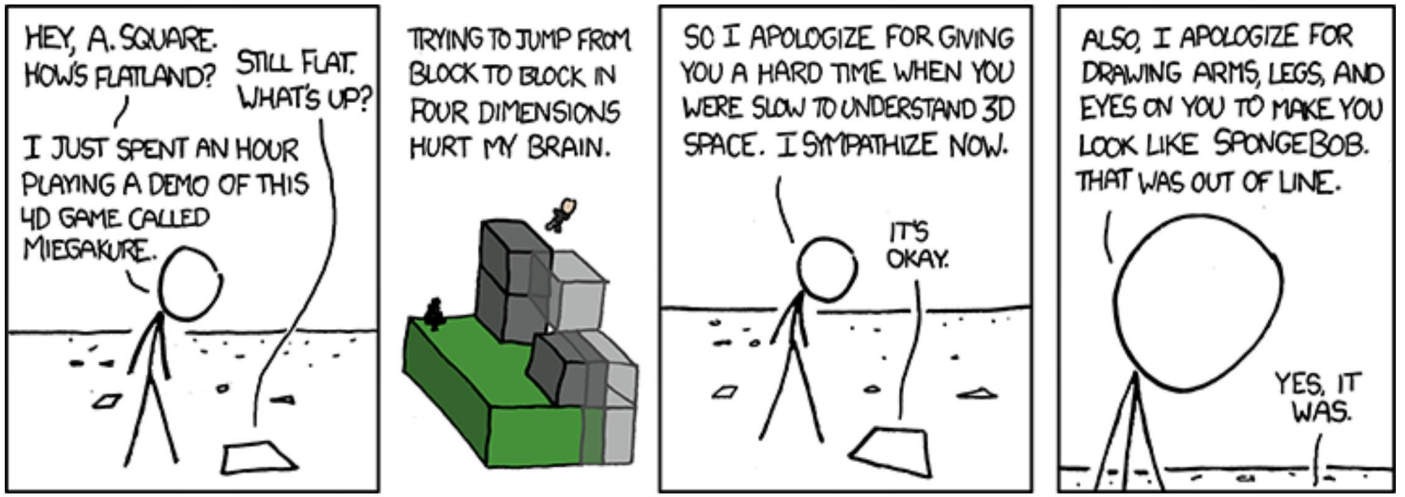
The response may more complex

The following is an example of binomial regression, where the outcome responses are binomially distributed

$$y_i = \text{Binom}\left(n_i, \frac{1}{1 + \exp(-\beta_1 x_{i1} - \beta_2 x_{i2})}\right) \tag{34}$$

Further, complex models can be included in this framework known a *generalised linear models*

Dimensionality Reduction



Dimensionality Reduction

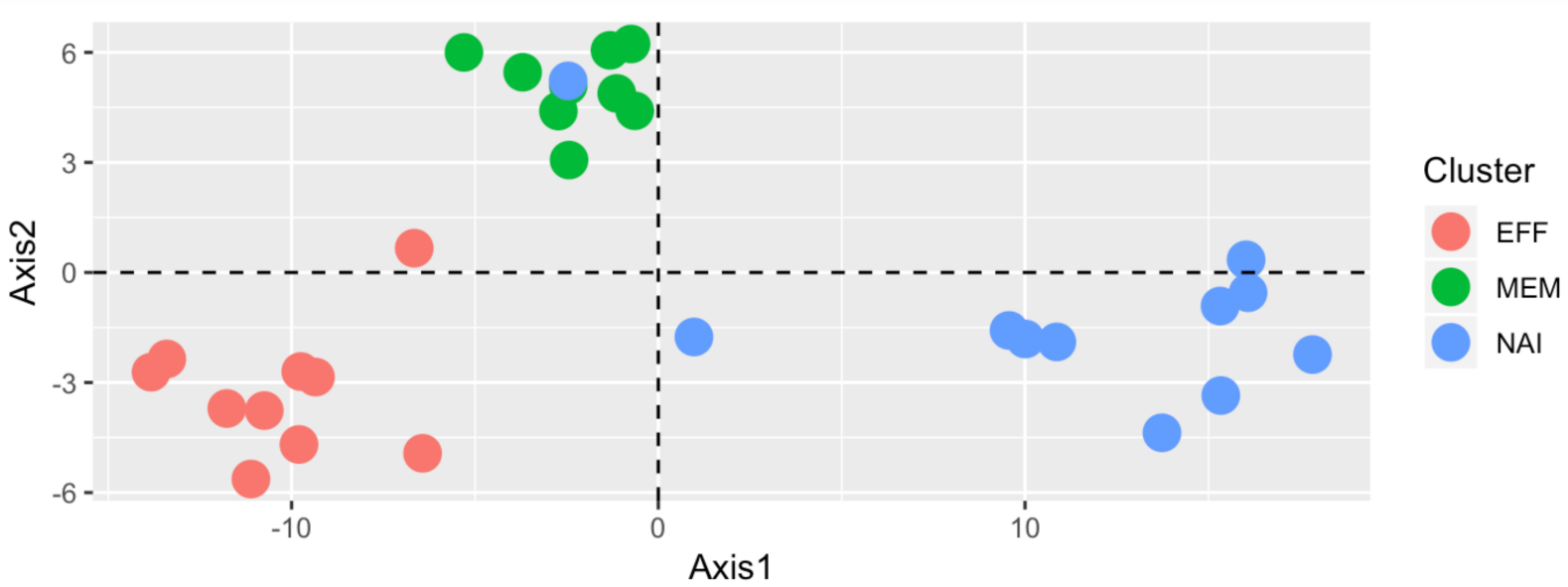
Aims

- Visualise high-dimensional data
- Reduce high-dimensional data into a more manageable size
- Understanding the variation in the data.

Warnings

- Are we applying the correct dimensionality reduction tool ?
- Yes, it looks nice ... but is it useful ?
- Visualisations should tell you something about the data, not what you'd like to data to show you

Principal Component Analysis



Principal Component Analysis

PCA

- Unlike regression PCA only involves features $X_1, X_2, ..., X_p$(and no response)
- PCA can be used to visualise observation or variables.
- Is applicable to continuous data.
- We would like to project the data in direction of greatest variability
- Look at linear combinations, which maximise sample variance

$$z_{i1} = \phi_{11} x_{i1} + \phi_{21} x_{i2} + ... + \phi_{p1} x_{ip} \tag{35}$$

- $z_{11}, ..., z_{n1}$ are known as *scores* (of the first principal component)
- $\phi_1, ..., \phi_p$ are referred to as *loadings*
- Subsequent principal components look for maximal variance but are also orthogonal(uncorrelated) to previous principal components

Principal Component Analysis

Variance and variables

- We can compute how much variance is explained by each principal component
  - Compute the overall variance of the data (for each feature)

$$var(X_1), \ldots, var(X_p) \tag{36}$$

  - Compute the the variance of the scores for each component

$$var(z_1), \ldots, var(z_m) \tag{37}$$

  - The proportion of variance explain by the th jth component is the variance of that component divided by the total variance.

$$VE_j = \frac{var(z_j)}{\sum_{i=1}^{p} var(X_i)} \tag{38}$$

batch Effects

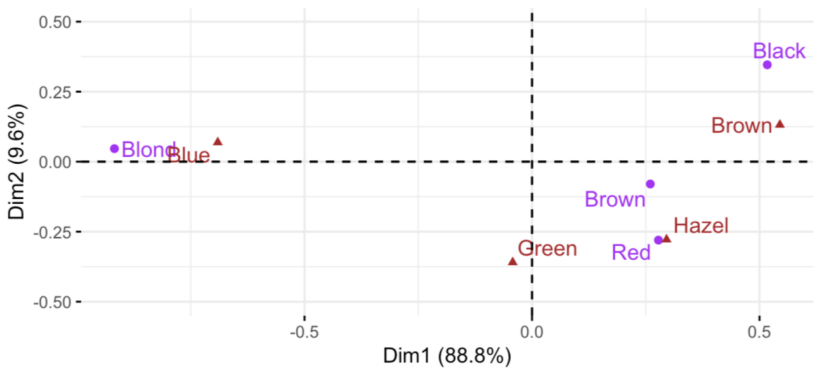
If a large proportion of variance is explained by an experimental day, machine type etc - you have a batch effect. (More later). These can often be found using PCA.

Correspondence Analysis

Aims

- Dimensionality reduction applied to contingency tables
- Uses the chi-squared distance
- As with PCA can be used to explain variability.
- I won't go into the mathematics here
- Many other dimensionality reduction tools

Correspondence Analysis



	Brown	Blue	Hazel	Green
Black	36	9	5	2
Brown	66	34	29	14
Red	16	7	7	7
Blond	4	64	5	8

Classification

Goals

- Many statistical and machine learning tasks require the prediction of labels from data (and some already labelled data)
- This might be cell type, country of origin, sex, broken or not etc.
- There are MANY methods, logistics regression, K-NN, SVM, Deep learning etc

K-nearest neighbours

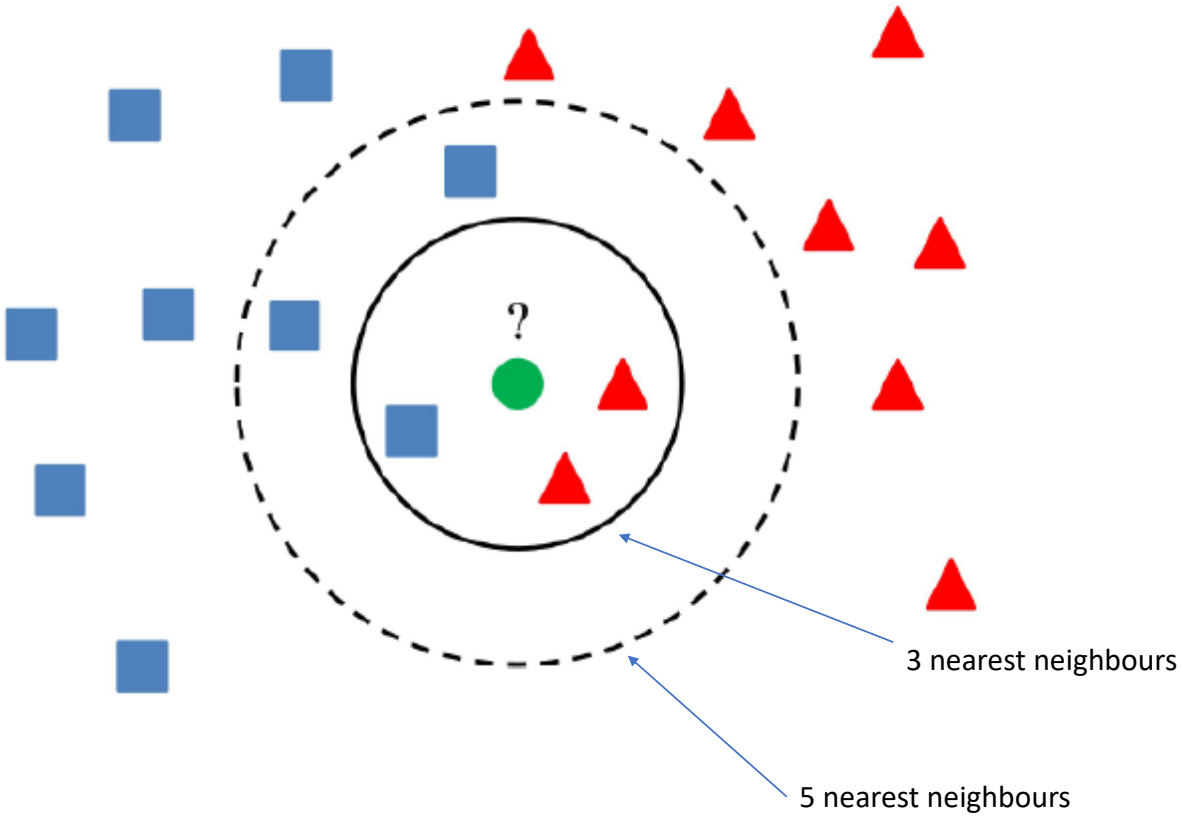
One of the simplest classification algorithms, is the k-nearest neighbours (KNN). It is a non-parametric approach.

KNN

- Suppose our data X can take label $y_i = 1, ..., J$
- Imagine a portion of these data already have labels
- Suppose we have a point x_i without a label
- Assign a label by looking at the K closest points with labels
- Probabilistically we can think of assigning the label with maximal probability

$$Pr(y_i = j | X = x_i) = \frac{1}{K} \sum_{l \in \mathcal{N}_i} I(y_l = j), \tag{39}$$

where $\mathcal{N}_i$ is the set of K closest labelled points to x_i.


?
3 nearest neighbours
5 nearest neighbours

How to choose K

As we've seen different choice of K can lead to different answers.

Cross-validation

- How to do we choose K to minimise the test error
- We call K a parameter
- We can apply a general method called cross-validation
- We split the training data (the data with labels) into random partitions
- Apply K-NN algorithm for different choices of K. E.g. $1, 3, 5, 7, 9, 11$
- Compute the test error of each K
- repeat on different random partitions
- Choose the K that minimises the average test error across random test partitions
- (Generally applicable method)

Cross Validation

Logistic Regression

Logistic Regression

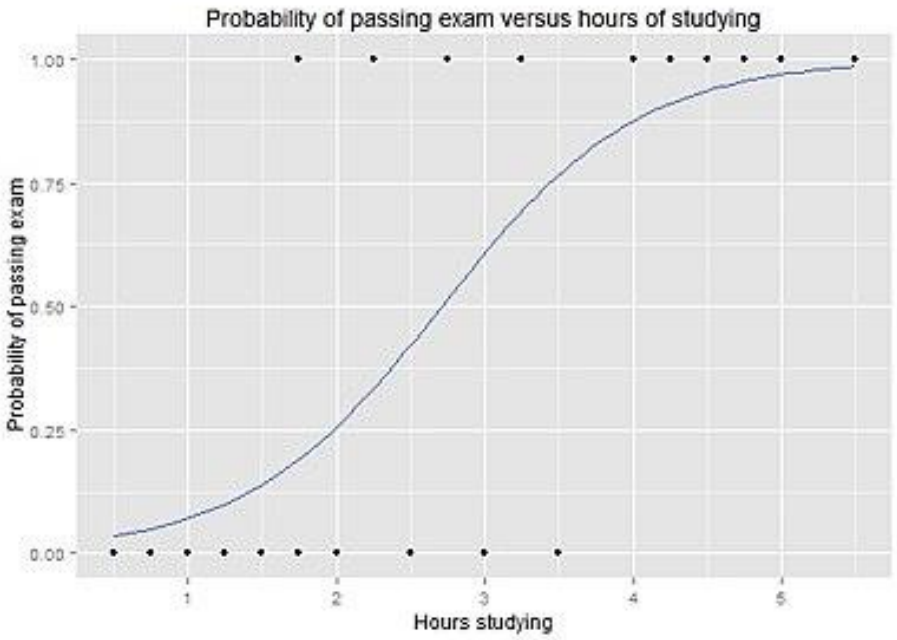
- The class output is related to a linear model
- Recall simple linear regression

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \tag{40}$$

- In logistic regression we want to model $P(Y = 1|X)$ a classification
- The logistic regression model makes the assumption that the log-odds are follow a linear model

$$\log\left(\frac{Pr(Y = 1|X)}{1 - Pr(Y = 1|X)}\right) = \beta_0 + \beta_1 X \tag{41}$$

- Regression coefficients are fitted with a method call maximum likelihood
- This is a general method for fitting model by maximising the probability of the model given the data.


Probability of passing exam versus hours of studying
Probability of passing exam
1.00
0.75
0.50
0.25
0.00
1
2
3
4
5
Hours studying

Is the variable useful for predicting the response ?

- We want a hypothesis test $H_0 : \beta_1 = 0$
- Compute the z-statistic (which is normally distribution)

$$z = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)} \tag{42}$$

- Then compute p-values from the normal distribution
- Logistic regression can be extended in the same way as we say for linear regression

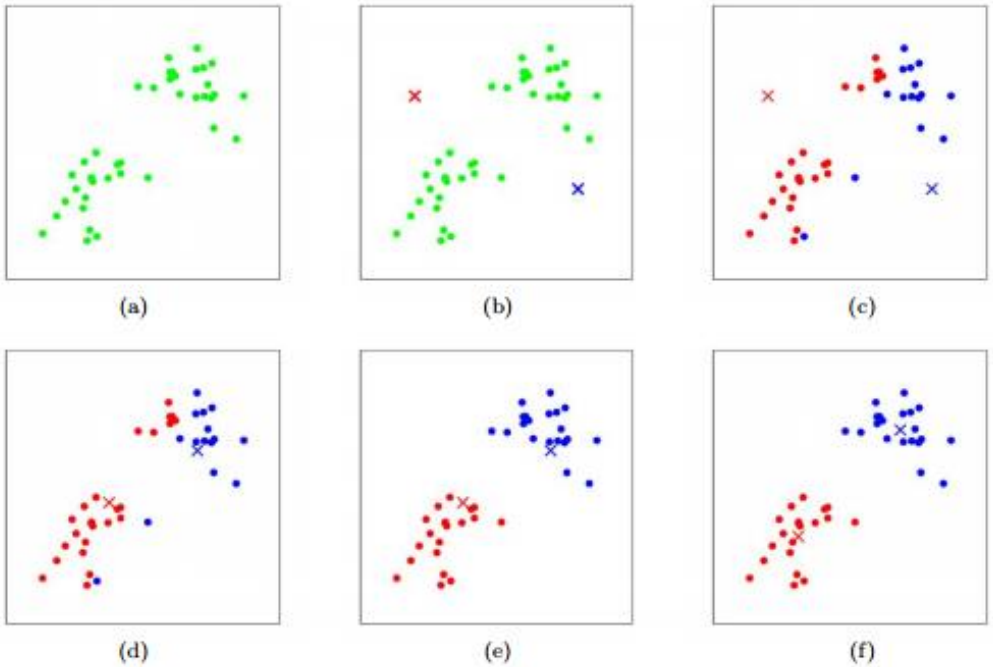
What is clustering

Aims and methods

- We want to group data in a meaningful way
- We want to discover labels when we have no labelled data
- We want to partition data to make it easier to handle
- K-means
- hierarchical clustering
- spectral clustering (not covered)
- mixture models (not covered)

K-means

- K-means is an iterative method
- We start with our data X
- We then select K initial cluster centres
- Assign the points to the cluster with closest centre
- Compute a new centre for each cluster by computing the arithmetic mean of the points in each cluster
- Repeat until stability
- Answer will depend on initialisation so advised to repeat with random starting points


(a)
(b)
(c)
(d)
(e)
(f)

How do we choose K ?

Choosing K

- Prior knowledge about the value of K
- Compute the Within Sum of Squares (WSS) for different values of K
- We can use the Elbow method ; that is, look at where WSS beings to plateau and choose that K
- The Gap statistics (See MSMB)
- Other methods

Hierarchical clustering

Ideas

- A bottom up approach
- Start with by joining close "points" together (into a group)
- As we work up we join groups of points together
- Often visualised as a tree

Distances

- single linkage computes the distance between clusters as the smallest distance between any two points in the two clusters

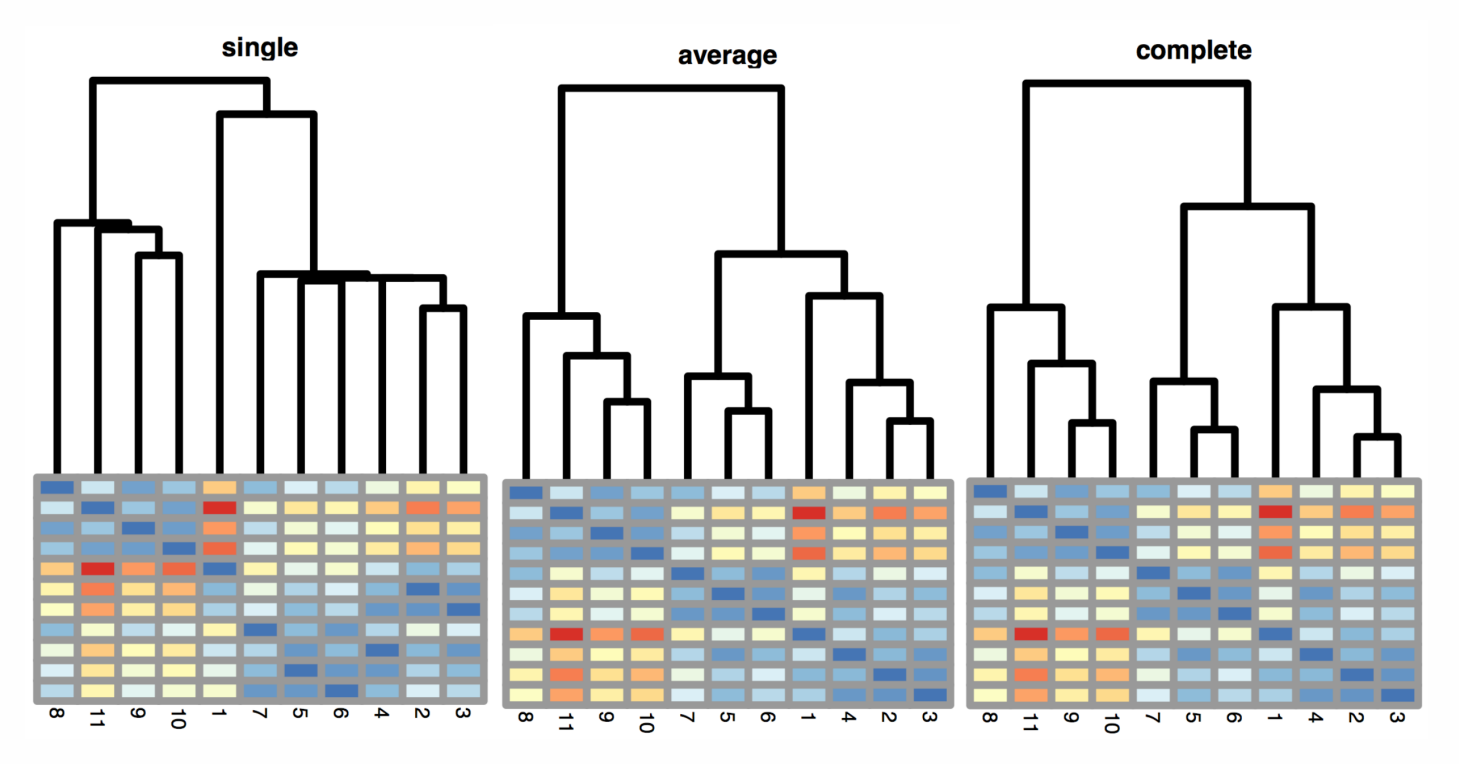
$$d_{12} = \min_{i \in C_1, j \in C_2} d_{ij} \tag{43}$$

- Complete linkage defines the distance between clusters as the largest distance between any two objects in the two clusters

$$d_{12} = \max_{i \in C_1, j \in C_2} d_{ij} \tag{44}$$

- Ward's method minimize the variance within clusters.

hierarchical clustering



Summary

Summary

- We discussed hypothesis tests including some example of tests
- We covered different methods to deal with multiple testing
- We discussed linear models and how to fit them
- We discussed hypothesis tests for linear models and model fit
- We covered multiple regression and regression with factors
- We talked about PCA and CA
- We discussed classification approaches such as KNN and logistic regression
- We discussed clustering methods such as hierarchical clustering and K-means

References

Modern Statistics for Modern Biology ; Susan Holmes and Wolfgang Huber
An Introduction to Statistical Learning ; James, Witten, Hastie, Tibshirnani
The Elements to Statistical Learning ; Friedman, Tibshirani, Hastie