Chris Hallsworth
Statistics Advisory Service Coordinator

c.a.hallsworth@bath.ac.uk
http://www.bath.ac.uk/study/mash/sas/

UNIVERSITY OF BATH

Objectives

Increase familiarity with statistical concepts

- Statistical significance - when are two things different?
- Analysis of variance - protypical statistical analysis
- Diagnostics - how we critique an analysis

Practise reading statistical graphics

- Histograms
- QQ plots
- Residual plots
- Mosaic plots

Introduction to Statistical Concepts

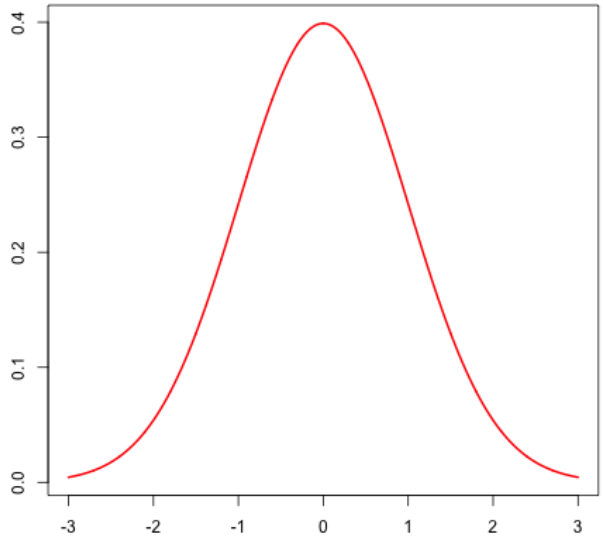
Statistics is all about variability

- Systematic variation due to processes of interest
- Substructure, known or cryptic
- Sampling variation
- Measurement error
- Mistakes

Apportion observed variability to possible sources, building a model that leads to better understanding of the underlying processes.

The normal (or Gaussian) distribution

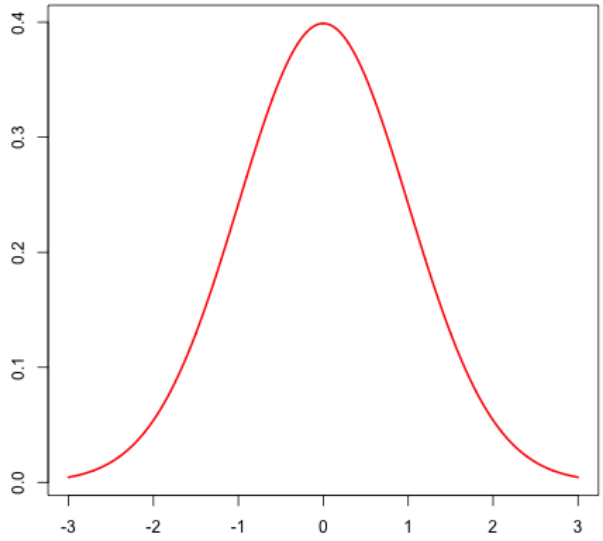
- The normal distribution is a good model for variables that arise as the sum of many small, independent effects
  - biological variables
  - measurement error
  - "noise".
- If we remeasure a normal variable in new units, we still get a normal variable
  - invariant under change of scale and origin
  - if X is normal, so is $Y = aX + b$.
- Characterised by its expectation (location) and standard deviation (spread).



Properties of the normal distribution

- A normal variable with mean μ and standard deviation σ has probability density

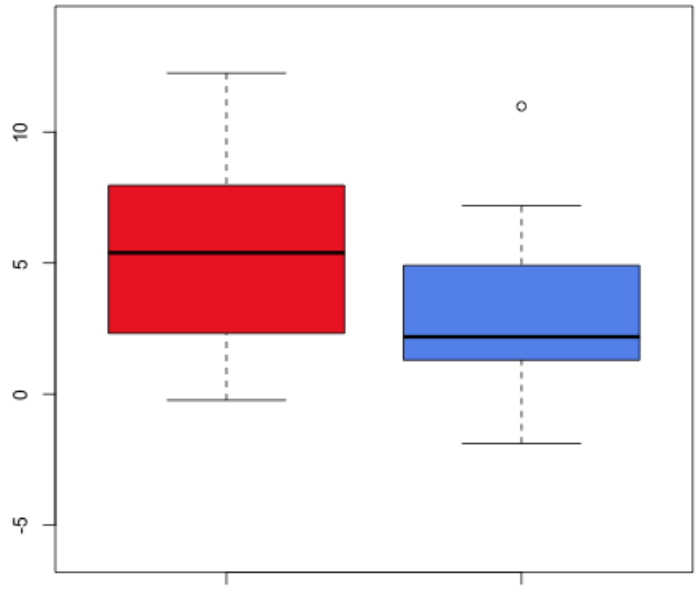
$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$



- The distribution is symmetrical about the mean, which is also the mode.
- The density function has points of inflection at $\mu \pm \sigma$.
- $\approx 95\%$ of the probability lies in the interval $\mu \pm 2\sigma$.

A typical statistical problem: comparing means

We have data on the concentration of a marker in the blood of individuals in two independent samples of size 20.

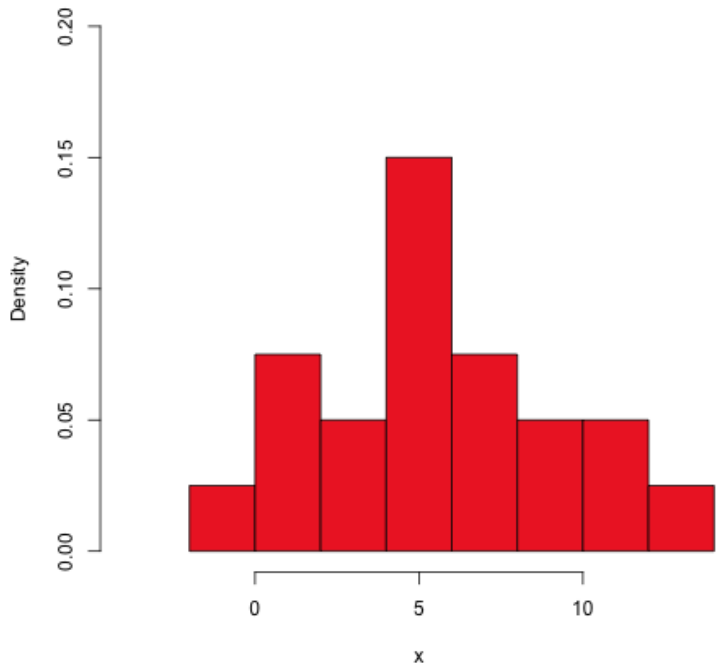


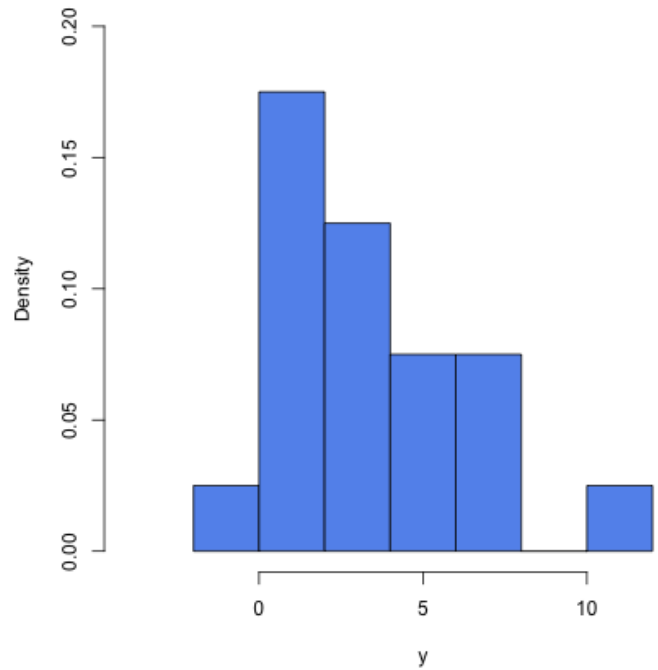
Is there any evidence that the samples come from populations with different means?

Looking at the data

Both seem to follow the normal distribution, roughly.

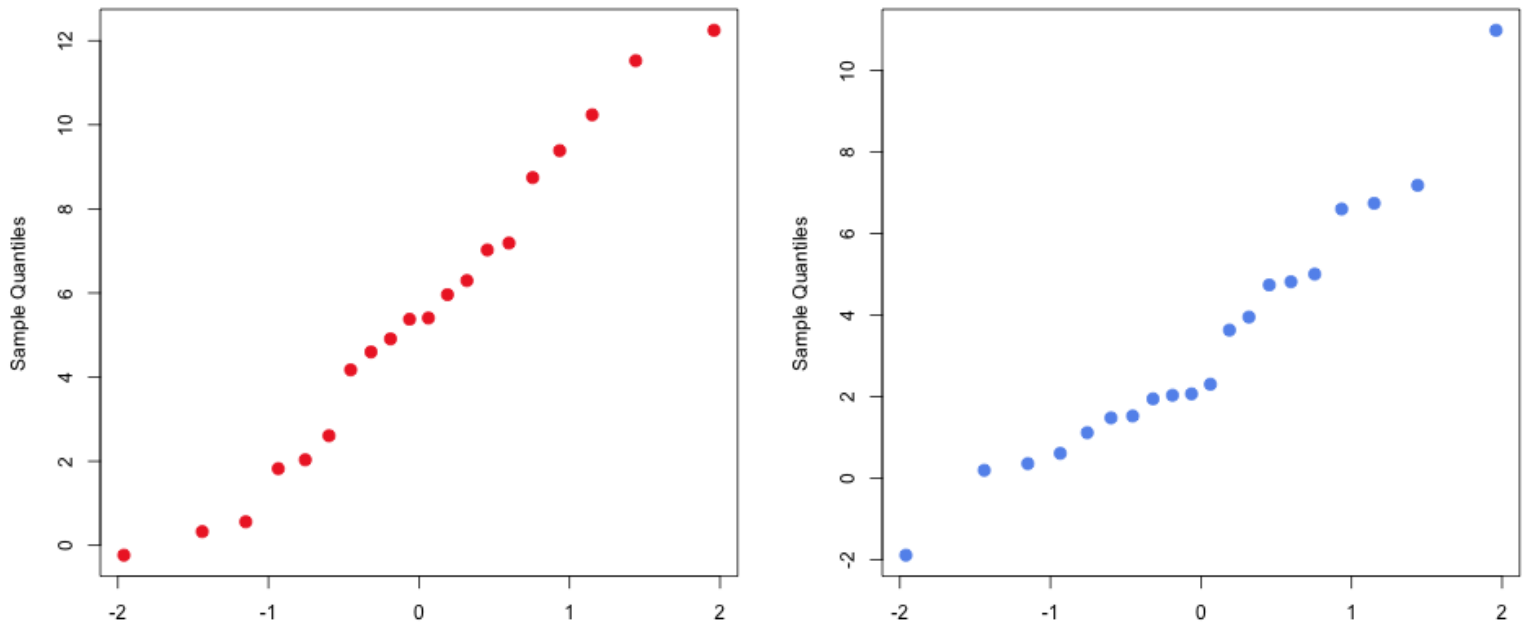
Both samples have roughly the same standard deviation.





QQ plots

A normal QQ (quantile-quantile) plot is better than a histogram for assessing the shape of a sample distribution.



Compares the *quantiles* of a sample distribution to those of a standard normal distribution.

- A straight line suggests that the normal distribution is a good model for the data.

Framework for evaluating the evidence

The null hypothesis

- Specify the simplest conceivable model for the samples.
- General scientific principle: parsimony / Ockham's razor.

In this case:

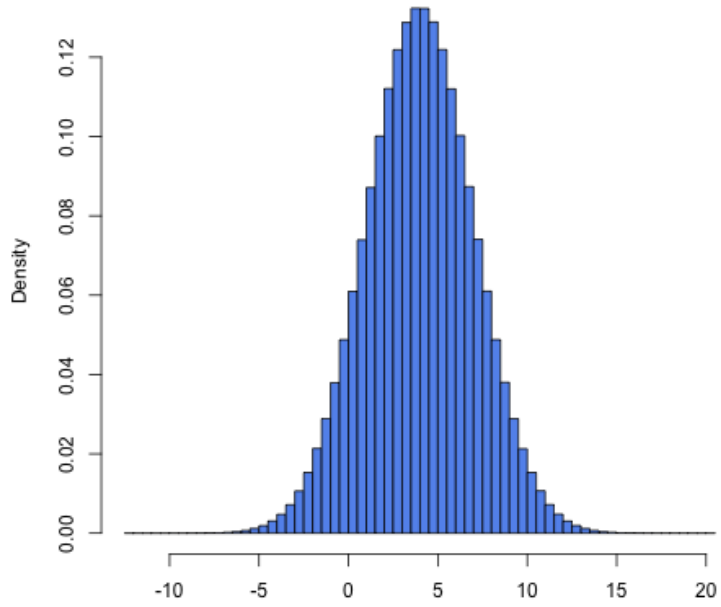
The samples are drawn from normal distributions with the same mean and standard deviation.

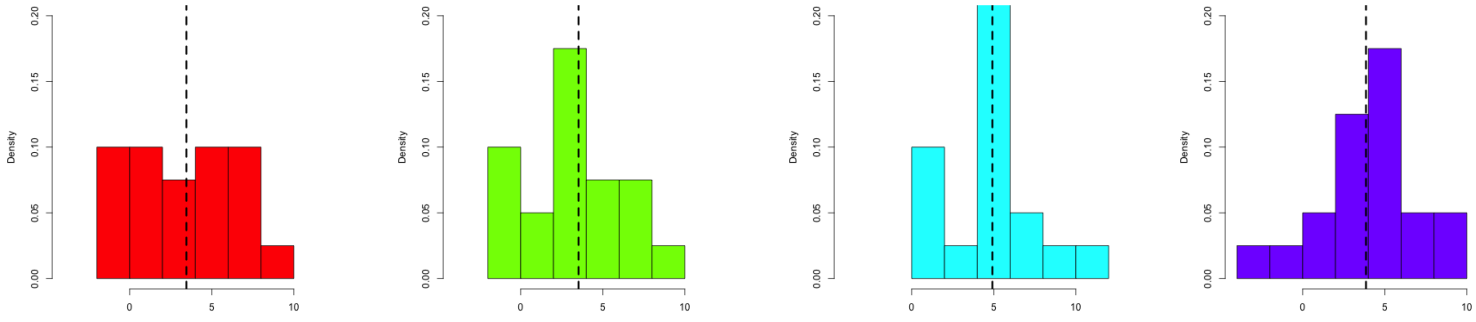
Do the data support the null hypothesis?

- Look for statistical properties of the samples that are inconsistent with this hypothesis.
- Experimental science framework: experiments generally discredit, rather then confirm, hypotheses.
- Only ever reject the null hypothesis in favour of an alternative.

The population

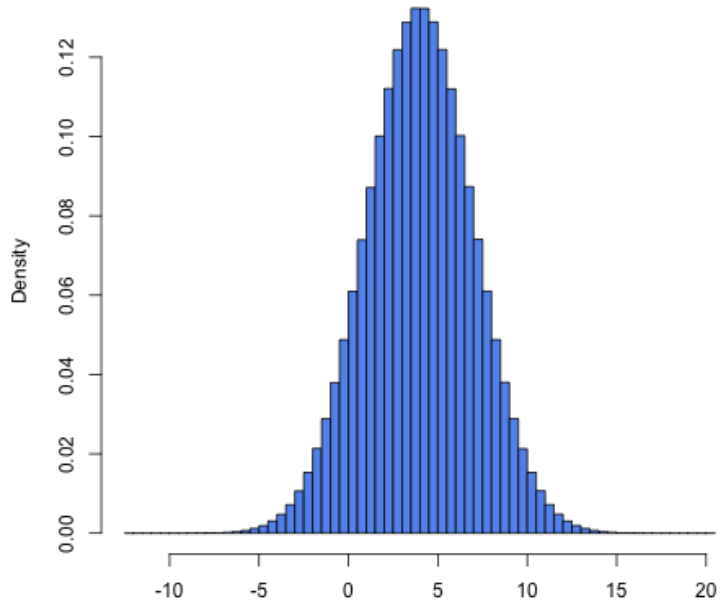
How different would we expect samples from the same distribution to be?

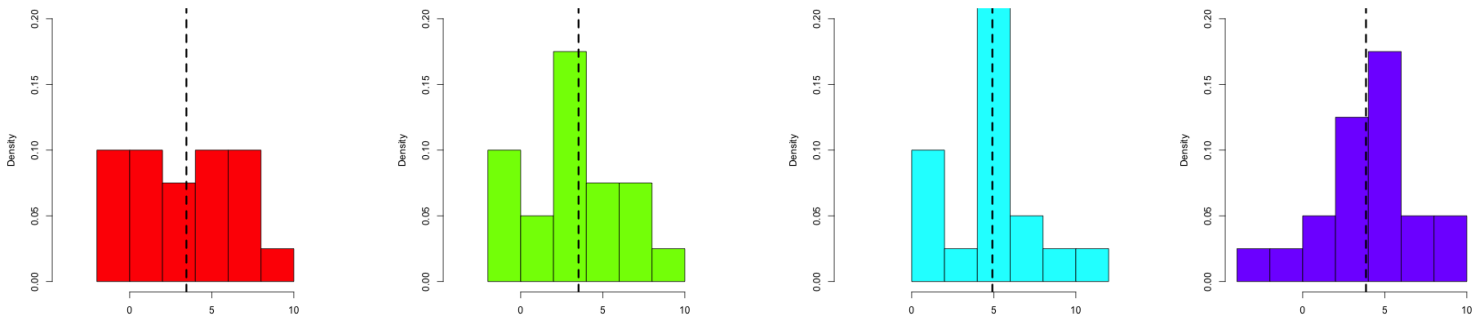




Sampling variation

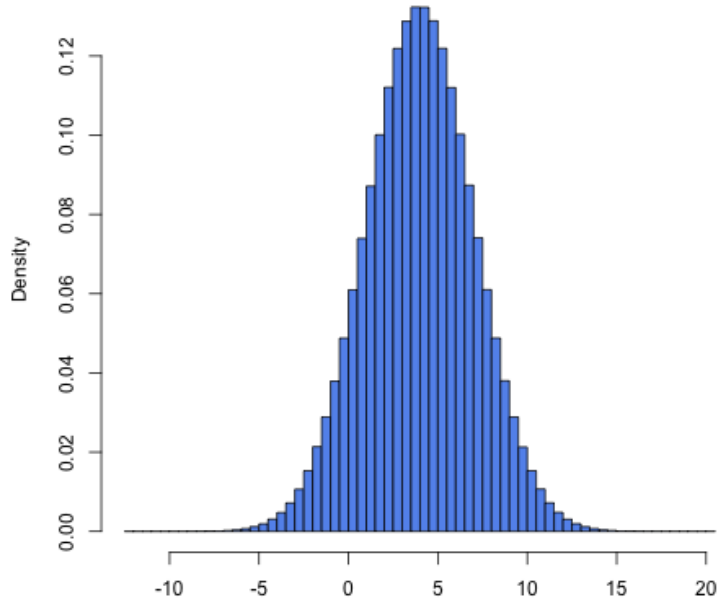
Samples from the same population have different means due to *sampling variation*





Sampling variation

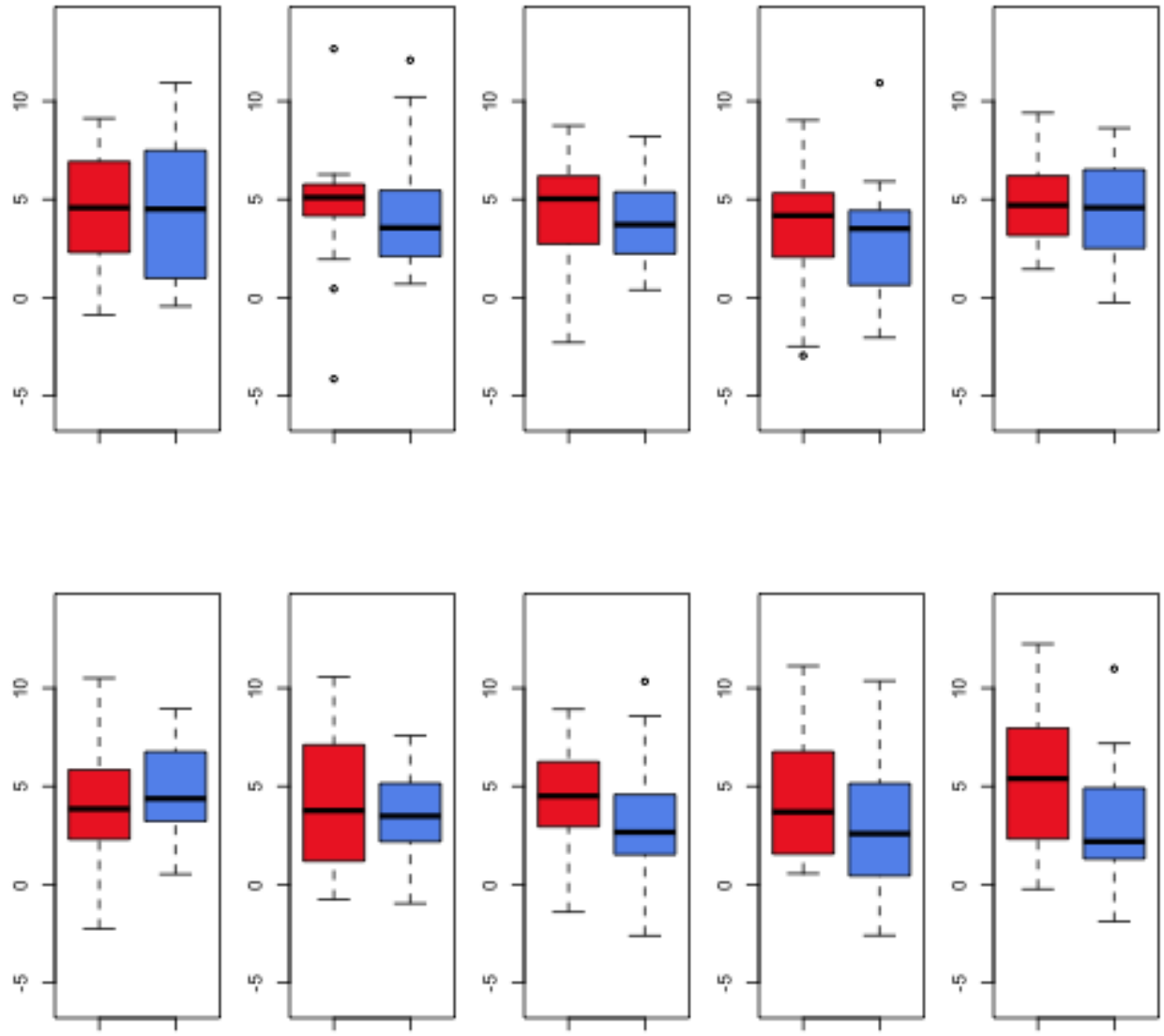
This tells us how to quantify the difference in the means of our two samples.



Take lots of pairs of samples of size 20 from this population and see how often we observe a pair as different from each other as ours.

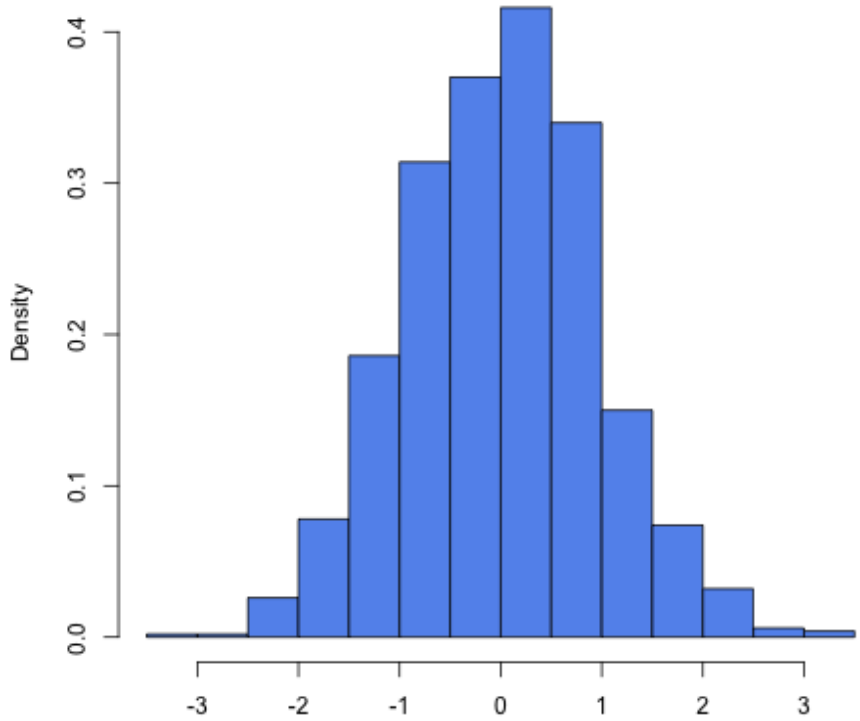
Sampling variation

Ten pairs of samples, each of size 20.



The sampling distribution

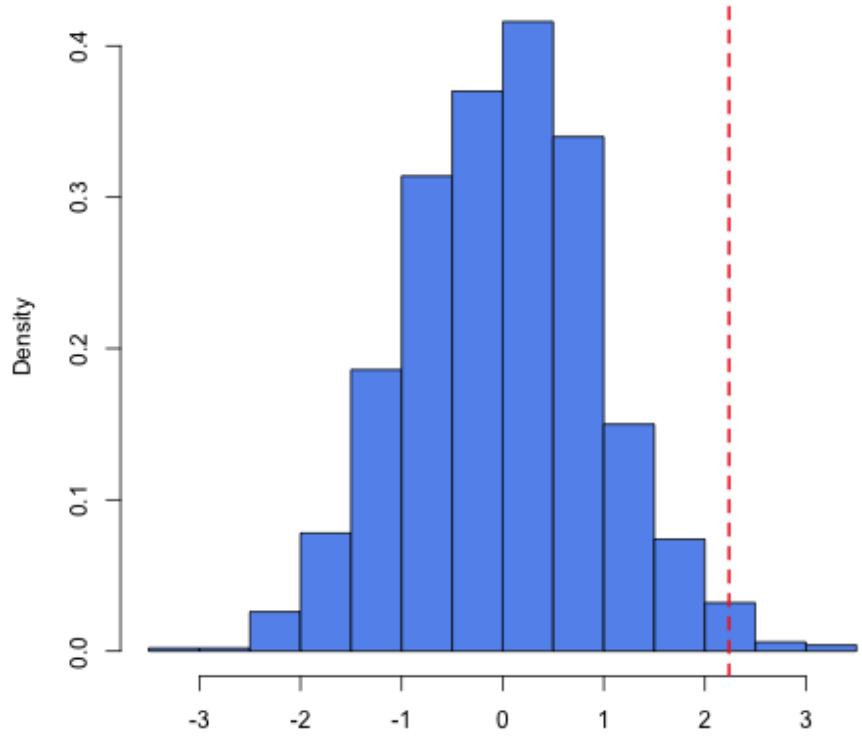
Keep on sampling....



Histogram of the differences in means for 1000 pairs of samples of size 20 from the population

The sampling distribution

Keep on sampling....



Our pair of samples differed by about 2.3 units.

Significance

How unusual was our original observation?

Only 1% of pairs of samples of size 20 differ by as much as our pair.

This suggests that sampling variation alone is an implausible explanation for the difference in means we observed.

We *reject* the hypothesis that the two samples come from a distribution with the same mean.

The p-value

What is a p-value?

We say that there is a *statistically significant* difference between the two samples' means.

We quote a p-value or significance level of 1%.

This is the proportion of pairs of samples *from the same distribution* that are as different as the observed pair.

The p-value

What does a p-value measure?

The p-value is a widely misunderstood concept among users of statistics.

Important to note that it is a measure of the strength of *evidence*, not (directly) a measure of the size of the difference.

It is possible to have a lot of evidence for a tiny and uninteresting difference (if there's a large sample size)!

Power

Type 1 error

Incorrectly rejecting the null is called a *Type 1 error*.

If we reject the null hypothesis when $p < 5\%$, this means that we would reject the null hypothesis in 5% of cases in which is is true.

Type 2 error

Failing to reject the null hypothesis when in fact it should have been rejected is a *Type 2 error*.

If the probability of making a type 2 error is β, $1 - \beta$ is the probability of rejecting the null hypothesis when it should be rejected. This is called the *power* of the test.

Factors affecting the power of a test

Sample size

Larger samples lead to more powerful tests.

Effect size

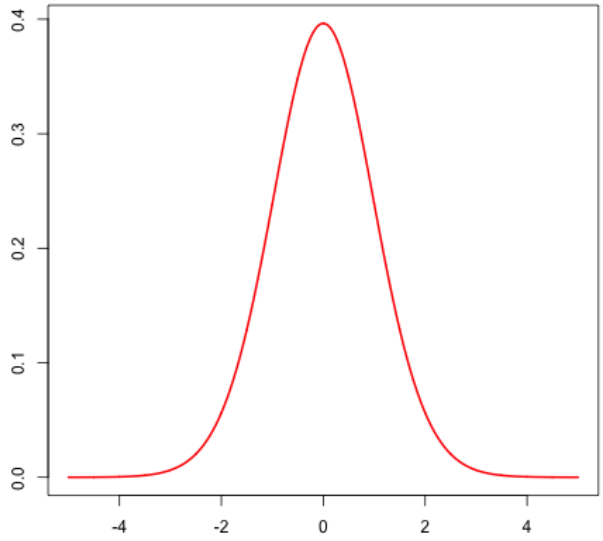
Larger differences between means are easier to detect.

p-value

Decreasing the probability of a type 1 error increases the probability of a type 2 error!

How do we calculate a p-value?

Under the null hypothesis we have $X_1 \dots X_n$ and $Y_1 \dots Y_n \sim N(\mu, \sigma^2)$



It can be shown that the distribution of the standardized difference between the sample means

$$t = \frac{\bar{X} - \bar{Y}}{S}$$

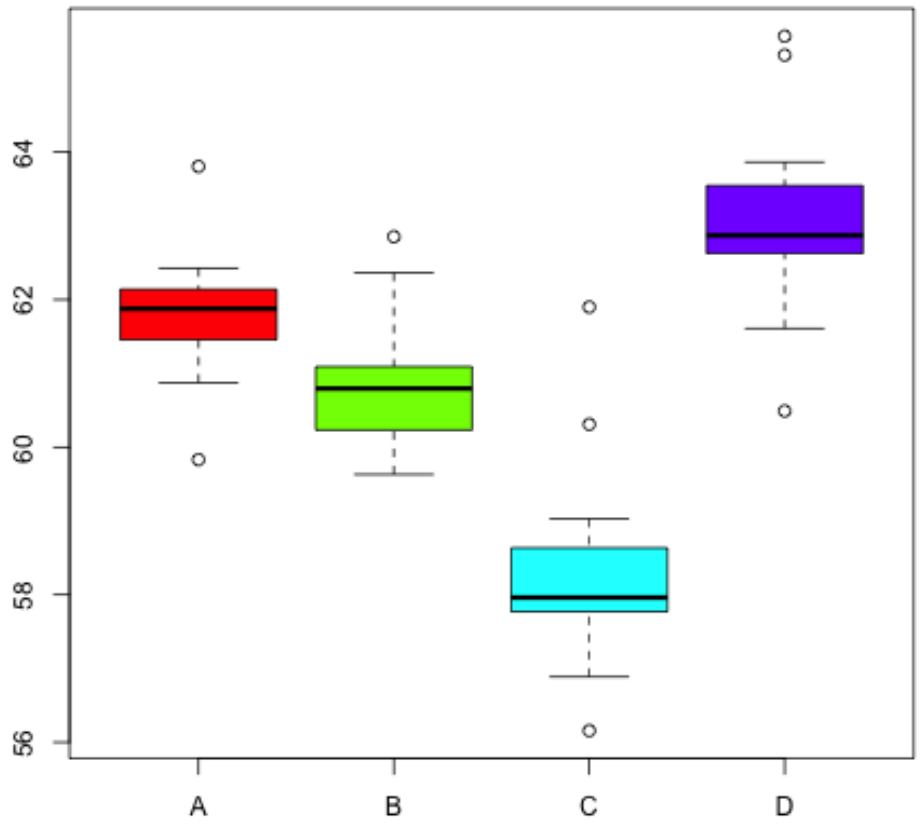
only depends on the sample size n. This is called the t distribution.

S is the standard deviation of the difference in sample means.

Analysis of Variance (ANOVA)

We can ask the same question with more groups - the method of analysis is called ANOVA.

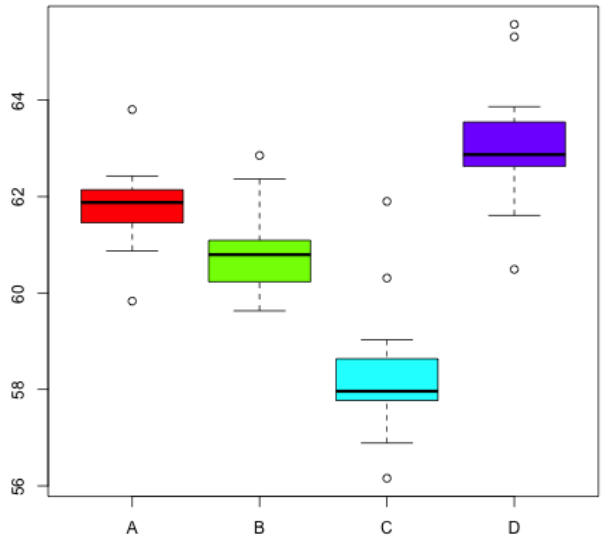
How much of the observed variability is variability *between* groups and how much is just variability *within* groups?



ANOVA

The underlying model here is

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$$

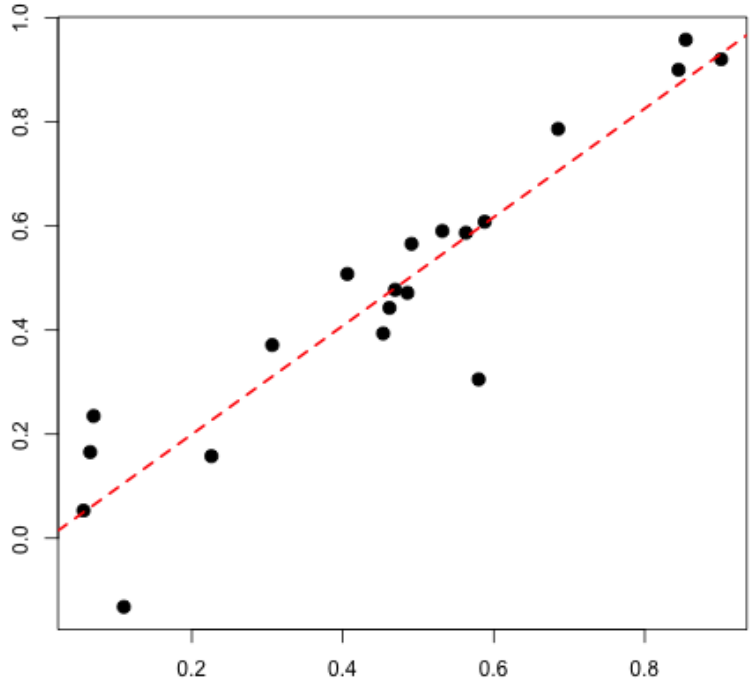


- Y_{ij} measurement of individual j from group i
- μ overall mean
- α_i mean correction for group i
- $\epsilon_{ij} \sim N(0, \sigma^2)$

Regression

Very similar to linear regression

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$



- Y_i response measurement of individual i
- x_i predictor measurement of individual i
- β_0 intercept of regression line
- β_1 gradient of regression line
- $\epsilon_i \sim N(0, \sigma^2)$

The linear model (for the mathematicians!)

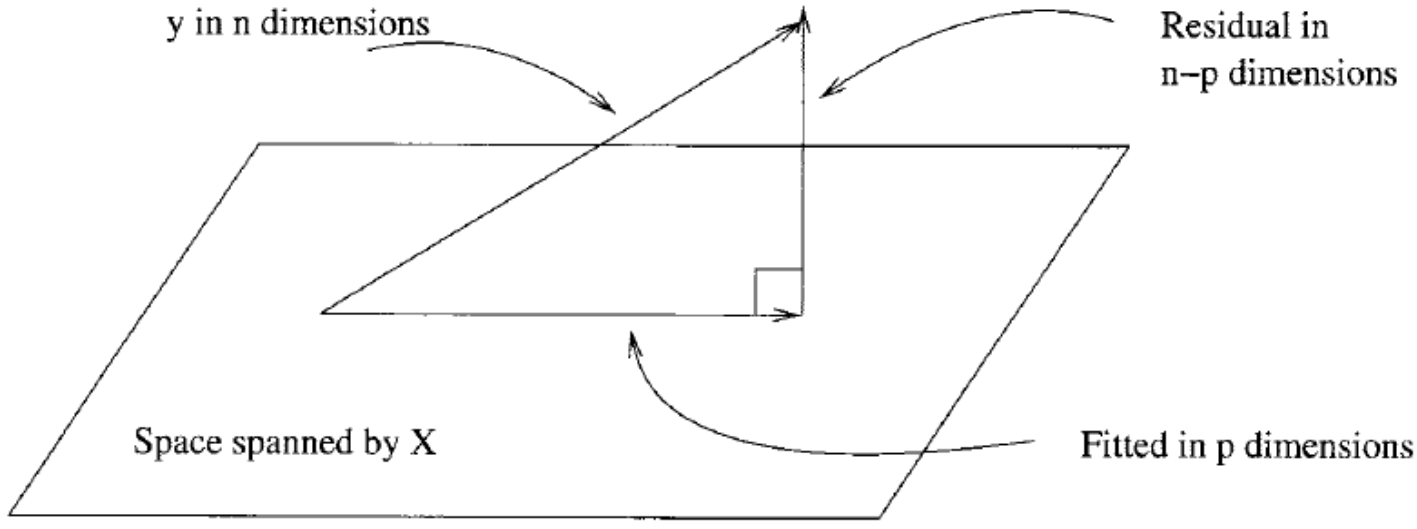
ANOVA and linear regression are both instances of a more general approach to statistics.

In both settings we specify the relationship between a predictor and a response as

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon,$$

where ϵ is a vector of independently distributed normal errors and $\mathbf{X}$ is the *design matrix*.

Find the vector β that minimizes the sum of squares $\epsilon^\top \epsilon$



Assumptions

What assumptions are needed?

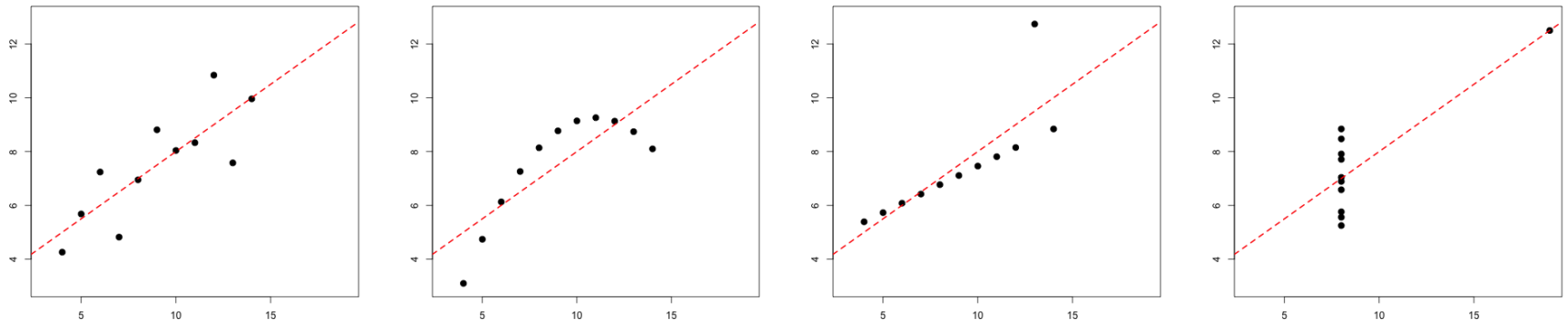
- Continuous data
- Normally distributed
- Homogeneous variance
- Appropriately specified independence structure

What if the assumptions fail to be met?

- Transform data
- Use non-parametric techniques
- Bootstrap

How things go wrong

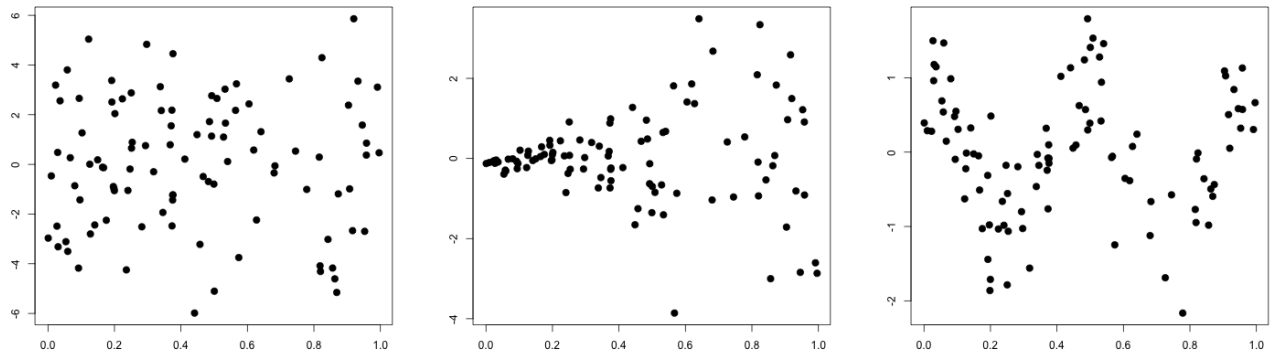
All of the x variables and all of the y variables have the same mean and standard deviation.



What's more, linear regression produces the same line for each pair.

Diagnostic Plots - checking things haven't gone wrong

Plot the residuals $\epsilon_i = y_i - \hat{y}_i$ against x_i. If the assumptions hold, this should be pure noise - so there should be no pattern.



1. Left: no pattern. No reason to suspect any departure from assumptions.
2. Centre: marked increase in variability from left to right. Suggests heterogeneity of variance.
3. Right: strong pattern in x. Suggests a non-linear relationship between x and y.

Are eye colour and hair colour independent?

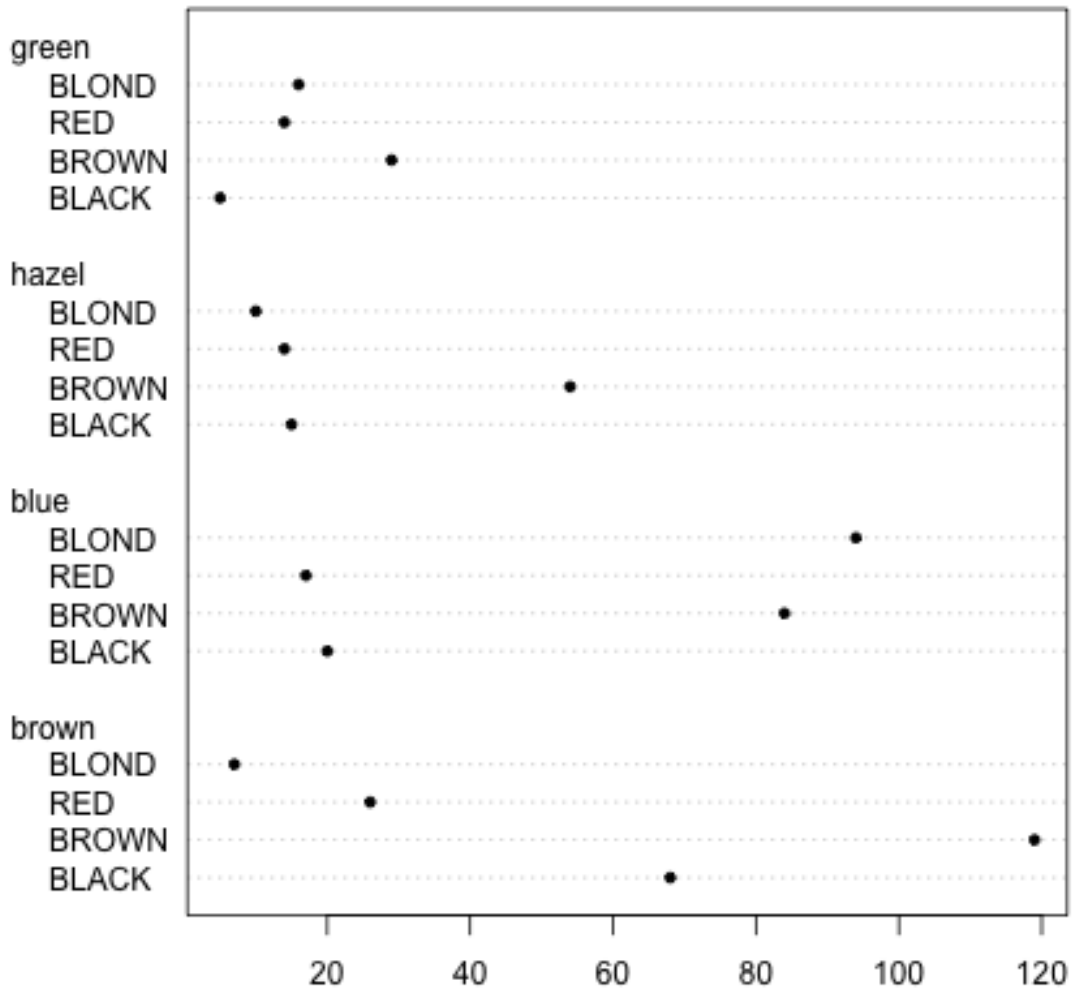
Data taken from Faraway 2006.

	Green	Hazel	Blue	Brown
Black	5	15	20	68
Brown	29	54	84	119
Red	14	14	17	26
Blond	16	10	94	7

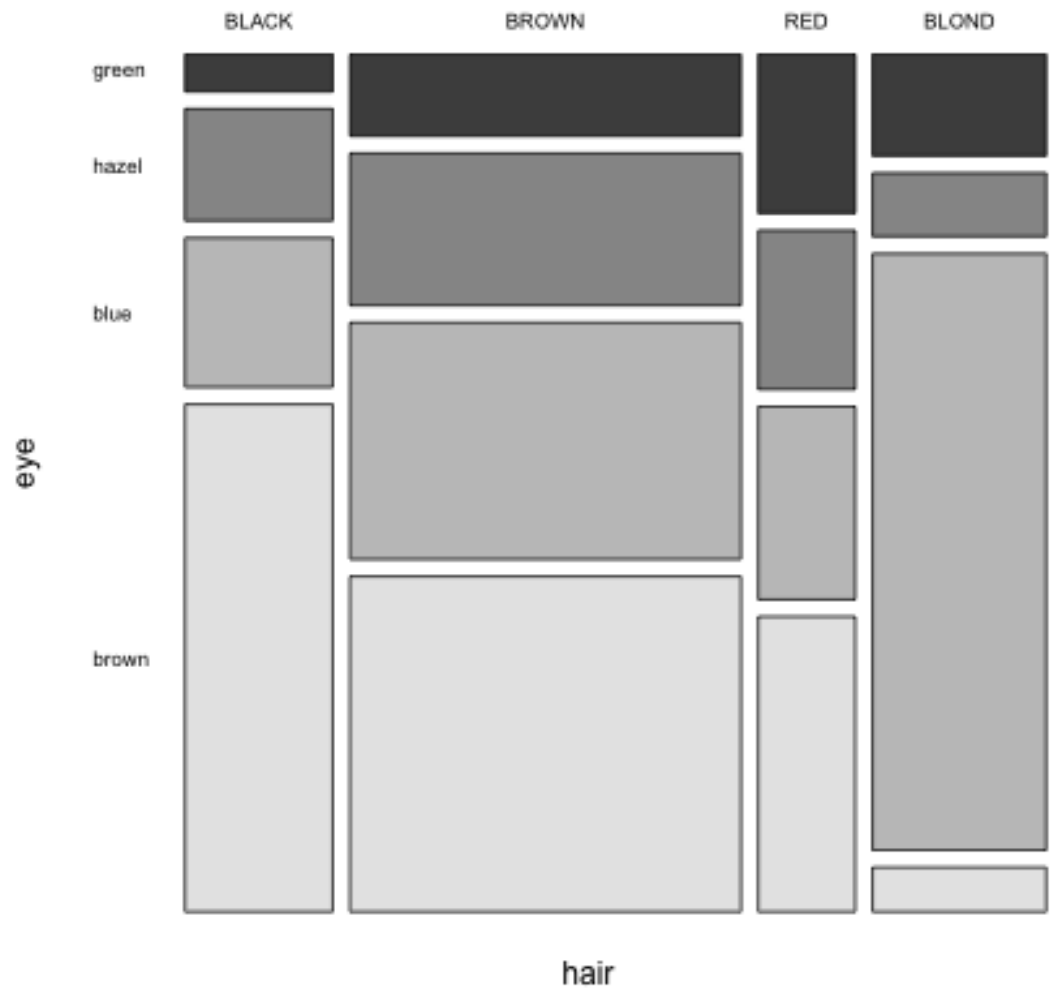
Is there evidence against the hypothesis that the rows and columns of the table are independent?

How best to represent this data graphically?

A Dot plot


green
BLOND
RED
BROWN
BLACK
hazel
BLOND
RED
BROWN
BLACK
blue
BLOND
RED
BROWN
BLACK
brown
BLOND
RED
BROWN
BLACK
20
40
60
80
100
120

A Mosaic plot



The $\chi^2 test$

So long as the cell counts are all reasonably large, the following quantity

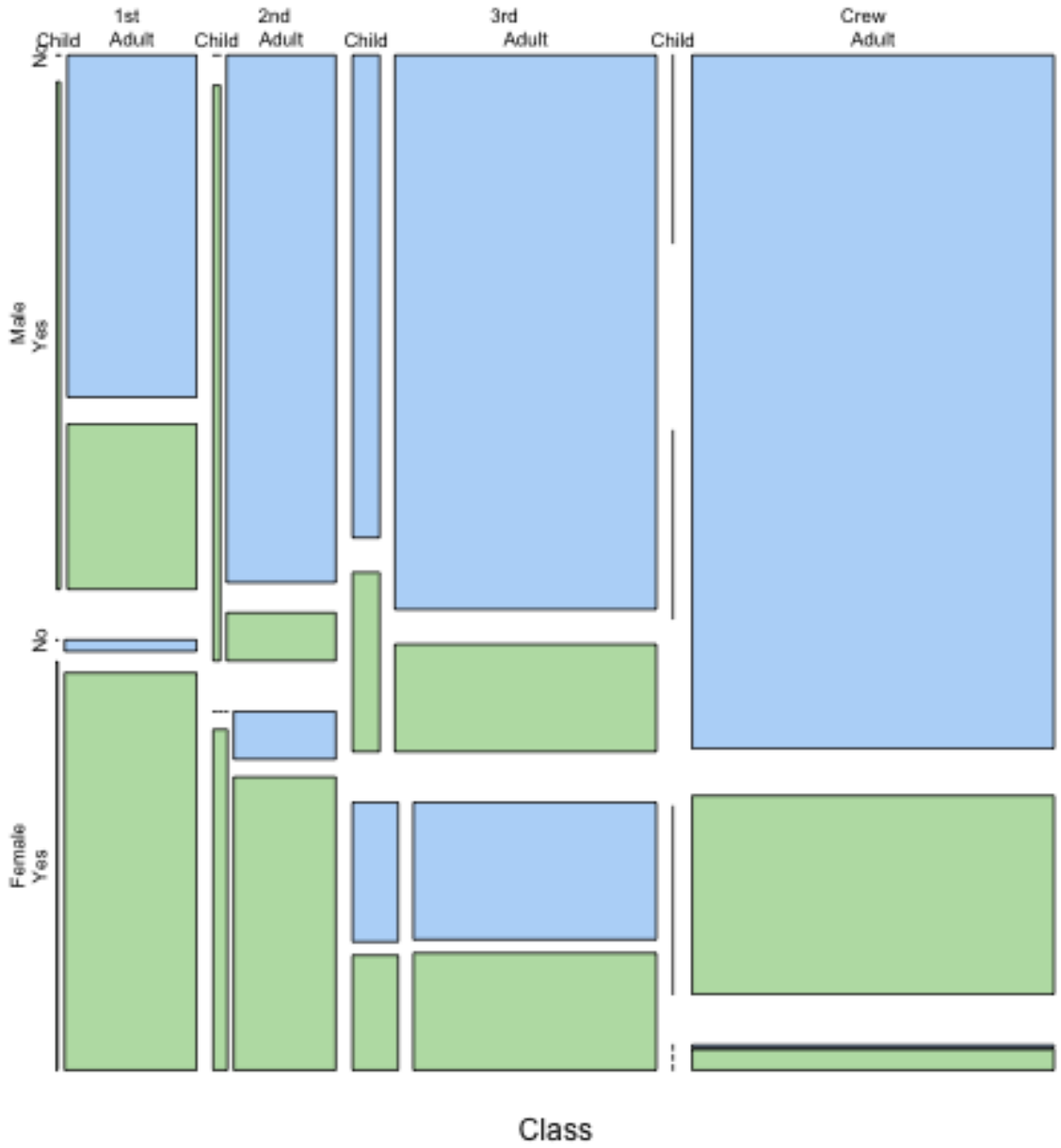
$$\chi^2 = \sum_{i=i}^{r} \sum_{j=1}^{c} \frac{(O_i - E_i)^2}{E_i}$$

has the $\chi^2 distribution$ with $(r-1)(c-1)$ degrees of freedom.

E_i is the expected number of counts under the hypothesis of independence.

For the eye and hair colour dataset, this test gives an extremely small p-value. We reject the hypothesis of independence.

A four way mosaic plot: survival on the Titanic



Multiple linear regression: which factors influence life expectancy in the US states