ANOVA

2022-03-15

Hypothesis Testing

- hypothesis testing is formally examine two opposing conjectures (hypotheses), H0 and H1
- these two hypotheses are mutually exclusive and exhaustive so that one is true to the exclusion of the other
- We accumulate evidence - collect and analyze sample data - for the purpose of determining if theres enough evidence to reject the null if favour of the alternative hypothesis

so we can test for the difference in means of two samples

so we can test for the difference in means of two samples

- t.test
- wilcox.test
- permutation testing

we measure a quantitative trait (effective warfarin dose) in a group of individuals and also genotype a SNP (AA, Aa, aa) in our favourite gene. We then divide these individuals into the three genotype categories to test whether the mean value differs among genotypes.

what do we do if we have more than two groups?

what the problem with doing lots of pairwise t-tests?

what do we do if we have more than two groups?

what the problem with doing lots of pairwise t-tests?

what is a Type I error?

what do we do if we have more than two groups?

what the problem with doing lots of pairwise t-tests?

what is a Type I error?

what is the probability of making a Type I error if I do pairwise t-tests for three groups?

ANOVA

what is the effect of one or more categorical variable on a quantitative (continuous) outcome variable

Qualitative variables are referred to as **factors** - effective warfarin dose, BMI, height, size

Characteristics that differentiates factors are referred to as **levels** (i.e., genotypes, conditions, treatment)

ANOVA

H_0: $\mu_1 = \mu_2 = \mu_3 \ldots = \mu_k$ H_1: Means are not all equal.

where k = the number of independent comparison groups.

one-way anova

each factor is linked to only level - its specified only one way

ANOVA

Assumptions

- Independence of observations
- the distributions of the residuals are normal
- the variance of data in groups should be the same.

ANOVA

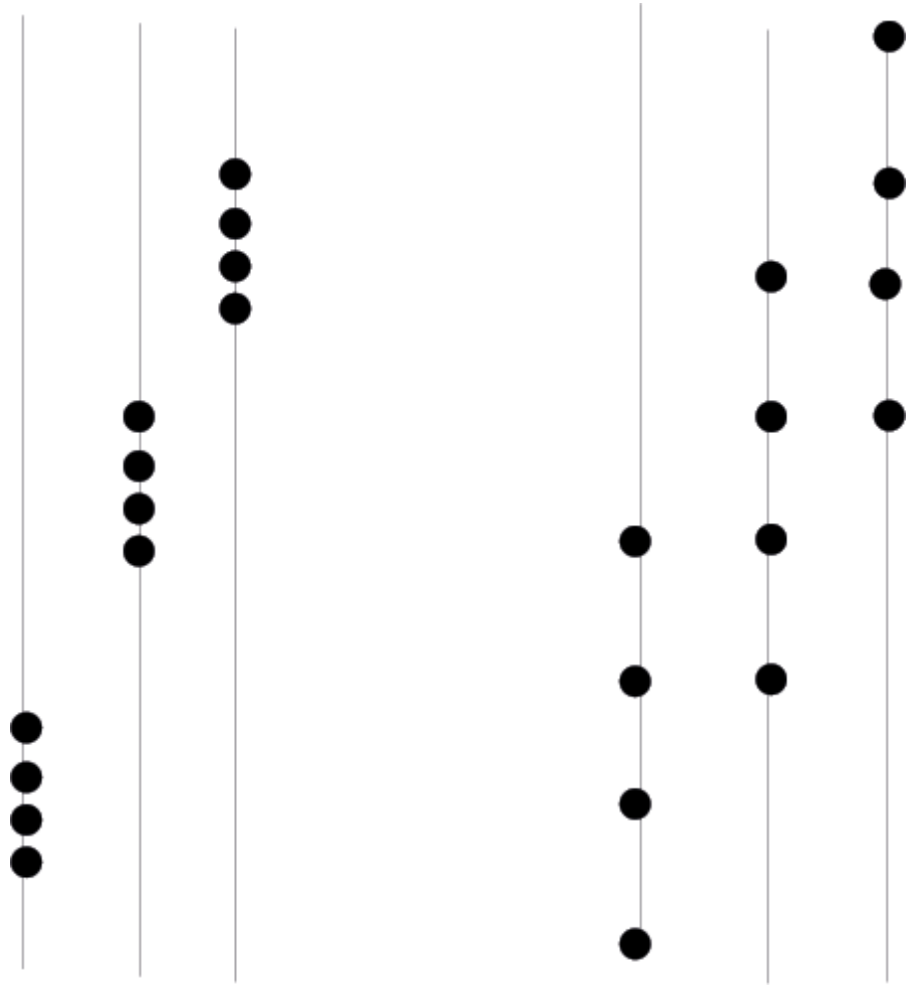
partition total variation of the data as coming from two distinct sources

- Variation within groups
- Variation between groups

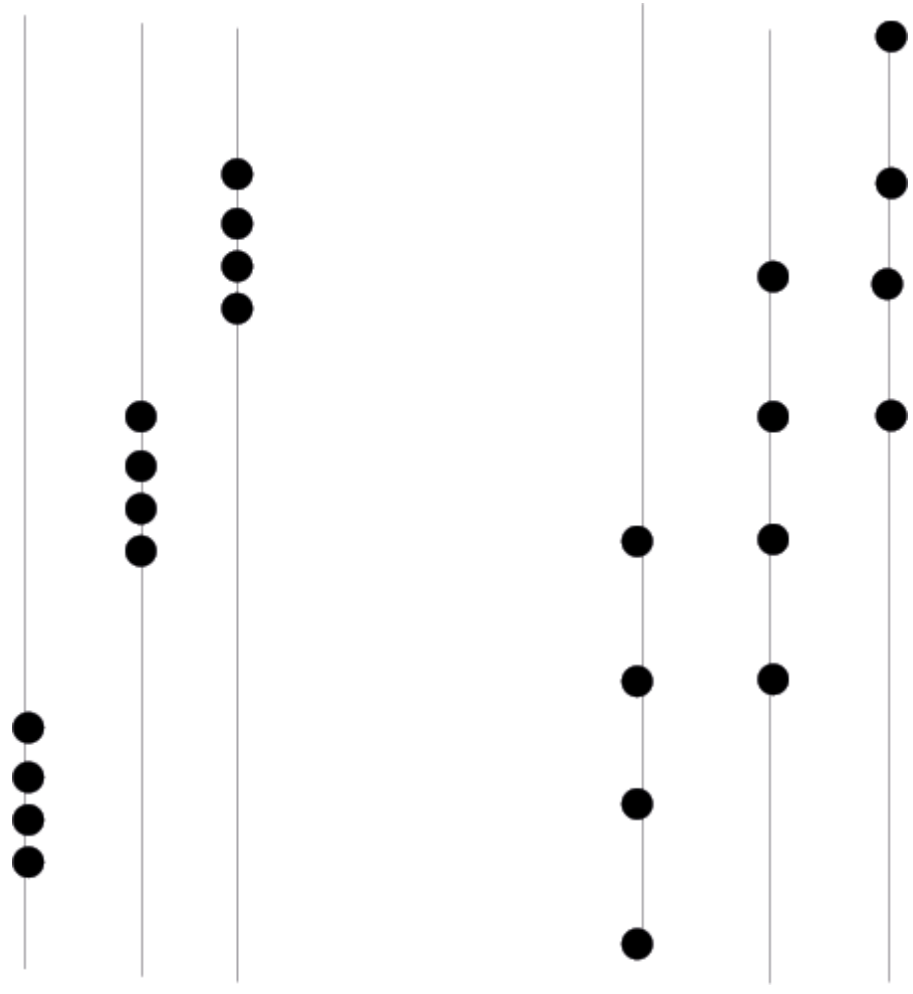
If variation among sample means is large relative to variation within samples, then there is evidence against H0 : $\mu_1 = \mu_2 = \cdots = \mu_k$.

If variation among sample means is small relative to variation within samples, then the data is consistent with H0 : $\mu_1 = \mu_2 = \cdots = \mu_k$.

which of these do you think is more likely that you would accept H1?

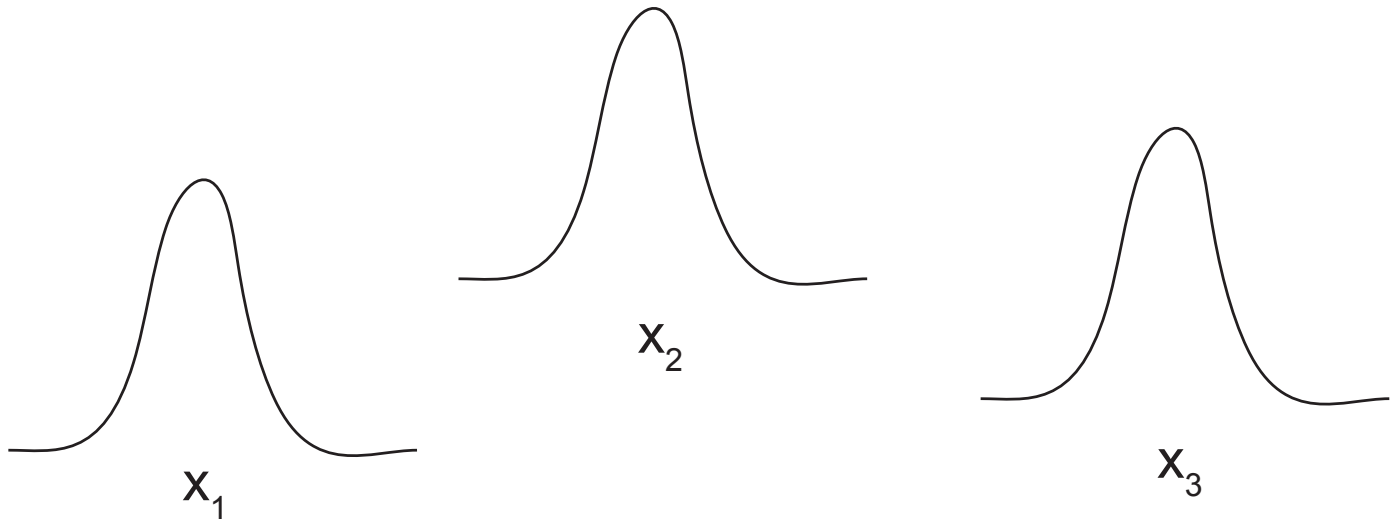


which of these do you think is more likely that you would accept H1?



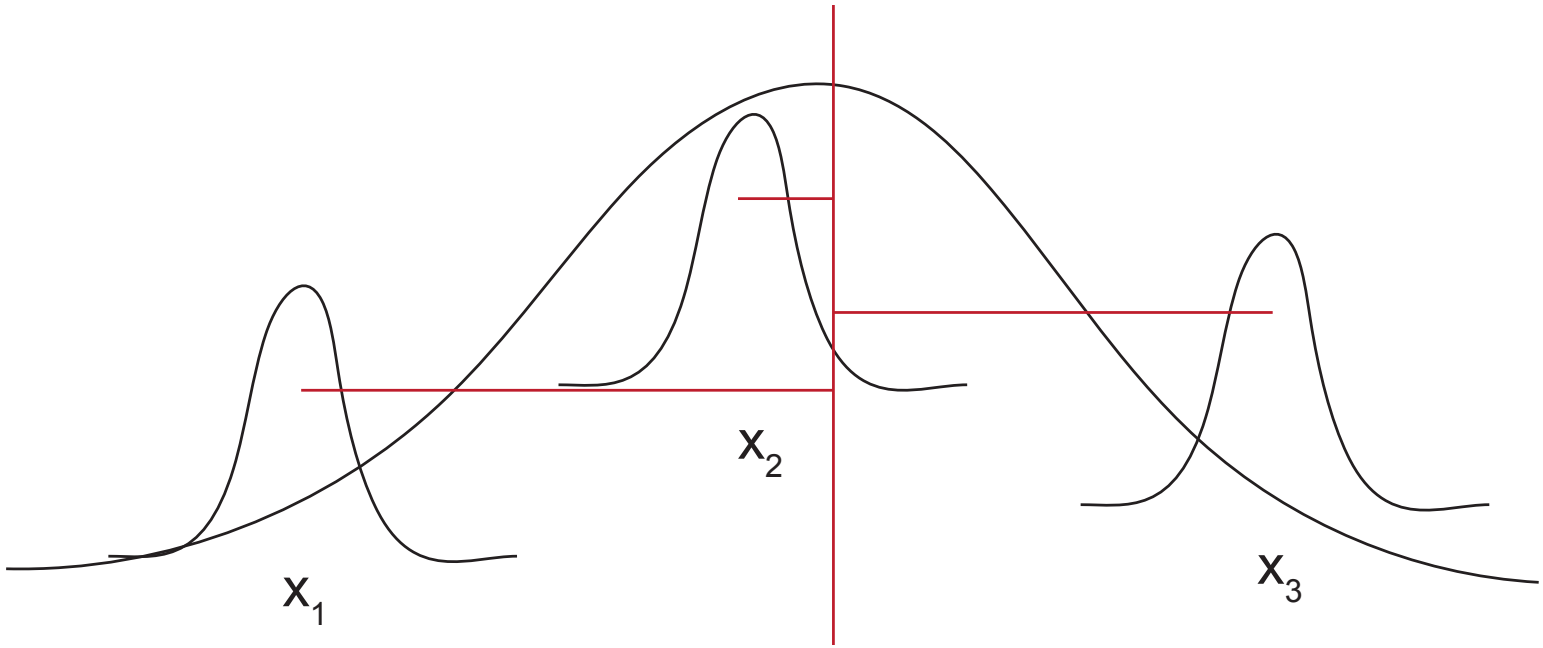
if the variance **between** groups exceeds what is expected in terms of the variance **within** groups, we will reject H0

ANalysis Of VAriance



each sample has mean - do all three come from the same distribution?

ANalysis Of VAriance



whats the distance from the means of each sample to the overall mean?

between-group variability

variance is average squared deviation from the mean

$$\sigma^2 = \frac{\sum(x - \mu)^2}{n - 1} \tag{1}$$

between-group sum of squares

$$\sum(x - \bar{\bar{\mu}})^2 * n_{\text{group}} \tag{2}$$

the between-group sum of squares are weighted by the sample size per group

between-group variability

let play with some data .... what is the between-group sum of squares?

within-group variability

width or spread within groups

sum of squares for error / within-group sum of squares

$$\sum_{group} \sum (x - \bar{\mu}_{group})^2 \tag{3}$$

within-group variability

let play with some data .... what is the within-group sum of squares?

degrees of freedom

between-group sum of squares

number of conditions - 1

within-group sum of squares

number of observations - number of conditions

F-statistic

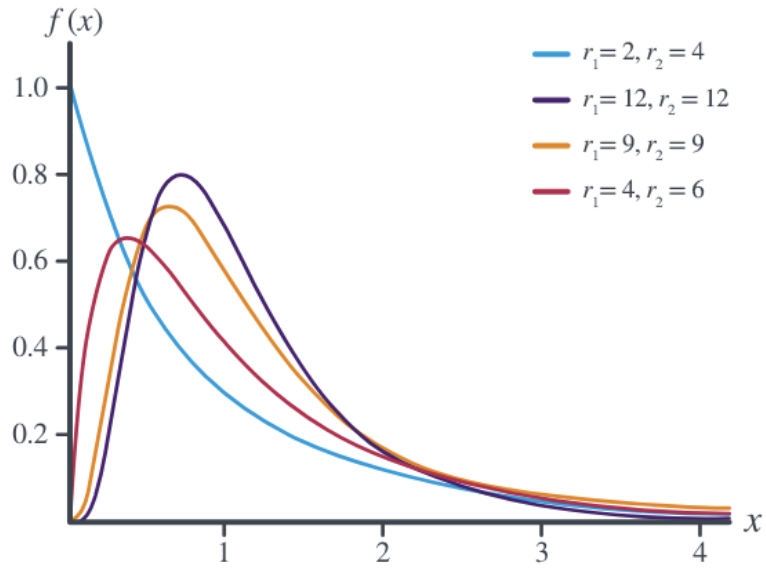
$$F = \frac{\text{between-group variability}}{\text{within-group variability}} \tag{4}$$

$$F = \frac{\frac{U}{r_1}}{\frac{V}{r_2}} \tag{5}$$

F-distribution

when two random samples are taken from two normal distributions with equal variance the ratio of those variances follows an F-distribution

- ?df
- ?qf
- ?pf
- ?rf



F-test

test for equality of variances (?var.test)

so what is the p-value of the differences we observe?

Lets do this in R

```
summary(aov(score ~ year, data=data))
```

```
             Df Sum Sq Mean Sq F value Pr(>F)
genotype      2   88.7   44.33   0.284  0.756
Residuals    18 2812.6  156.25
```

genotype reflects within group, residuals reflects between groups / errors

what happens if we mess with the data?

lets try...

- altering one of the groups to have a larger difference in its mean?
- altering two of the groups to have a large difference in their mean?
- altering the variance of one of the groups?

what happens?

- to the p-values?
- to the sum of squares?
- to the f-statistics?

Partitioning total variation

what happens if you calculate the sum of squares for all data points compared to the global mean?

Partitioning total variation

what happens if you calculate the sum of squares for all data points compared to the global mean?

$$ss_{\text{total}} = ss_{\text{between}} + ss_{\text{within}} \tag{6}$$

Partitioning total variation

what happens if you calculate the sum of squares for all data points compared to the global mean?

$$ss_{\text{total}} = ss_{\text{between}} + ss_{\text{within}} \tag{6}$$

total variance is the variance **explained** by the treatment + the **unexplained** variance

But what is my assumptions aren't met?

Non Parametric Alternative

- Kruskal-Wallis Rank Sum Test - non-parametric version of ANOVA
- ?kruskal.test

Two way ANOVAs

two-way ANOVA - add another factor or group - block - **randomized block design**

subtract out this from the between-groups/error variance

allows us to focus on group differences

$$ss_{\text{total}} = ss_{\text{between}} + ss_{\text{within}} \tag{7}$$

another sum of squares ....

$$ss_{\text{total}} = ss_{\text{between}} + ss_{\text{block}} + ss_{\text{within}} \tag{8}$$

$$F = \frac{\text{between-group variability}}{\text{within-group variability}} \tag{9}$$

calculate the sum of squares for the blocks - each block has mean - how does it compare to the overall mean

two-way ANOVA

so we have treated a tumour with different drugs and measured the levels of a specific protein - but we had to do batch this experiment

how do we take into account the variation caused by the batch?

degrees of freedom

total sum of squares

N - 1

between-group sum of squares

number of conditions - 1

within-group sum of squares

(number of conditions - 1)(number of blocks - 1)

block sum of squares

number of blocks - 1

so we need to calculate the block sum of squares

lets try this out

what happens if we don't control for batch?

ANOVA is a special case of linear regression - its all tied together

Questions?

- Next lecture: linking linear models with t-tests and anovas
- interaction terms