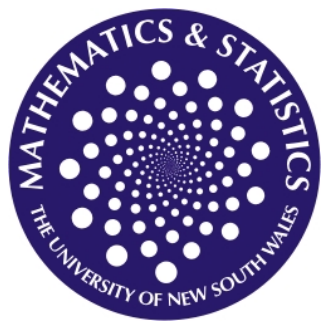
Statistics

MATH2089



Semester 1, 2018 – Lecture 12

This lecture

11. Analysis of Variance (ANOVA)

Additional reading: Sections 9.1, 9.2 and 9.4 in the textbook

11. ANOVA

Introduction

- In Chapter 10, we introduced testing procedures for comparing the means of two different populations, having observed two random samples drawn from those populations (two-sample z- and t-tests)
- However, in applications, it is common that we want to detect a difference in a set of **more than two populations**
- Imagine the following context: four groups of students were subjected to different teaching techniques and tested at the end of a specified period of time. Do the data shown in the table below present sufficient evidence to indicate a difference in mean achievement for the four teaching techniques?

Tech. 1	Tech. 2	Tech. 3	Tech. 4
65	75	59	94
87	69	78	89
73	83	67	80
79	81	62	88
81	72	83	
69	79	76	
	90		

Introduction: randomisation

- To answer this question, we should first note that the method of division of the students into 4 groups is of vital importance
- For instance, some basic visual inspection of the data suggests that the members of group 4 scored higher than those in the other groups. Can we conclude from this that teaching technique 4 is superior? Perhaps, students in group 4 are just better learners

$\rightarrow$ it is essential that we divide the students into 4 groups in such a way to make it very unlikely that one of the group is inherently superior to others (regardless of the teaching technique it will be subjected to)

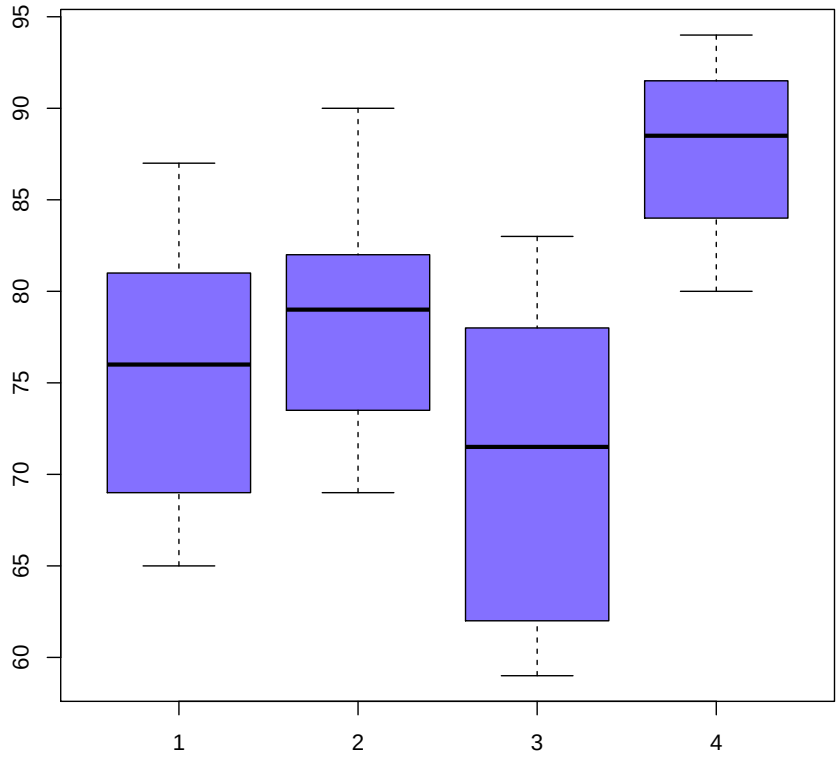
$\rightarrow$ the only reliable method for doing this is to divide the students **in a completely random fashion**, to balance out the effect of any nuisance variable that may influence the variable of interest

- This kind of consideration is part of a very important area of statistical modelling called **experimental design**, which is not addressed in this course (Chapter 10 in the textbook). In this course, we will always assume that the division of the individuals into the groups was indeed done "at random"

Introduction

Now, numerical summaries and a graphical display of the data are always useful:

	Tech. 1	Tech. 2	Tech. 3	Tech. 4
	65	75	59	94
	87	69	78	89
	73	83	67	80
	79	81	62	88
	81	72	83	
	69	79	76	
		90		
$\bar{x}$	75.67	78.43	70.83	87.75
s	8.17	7.11	9.58	5.80

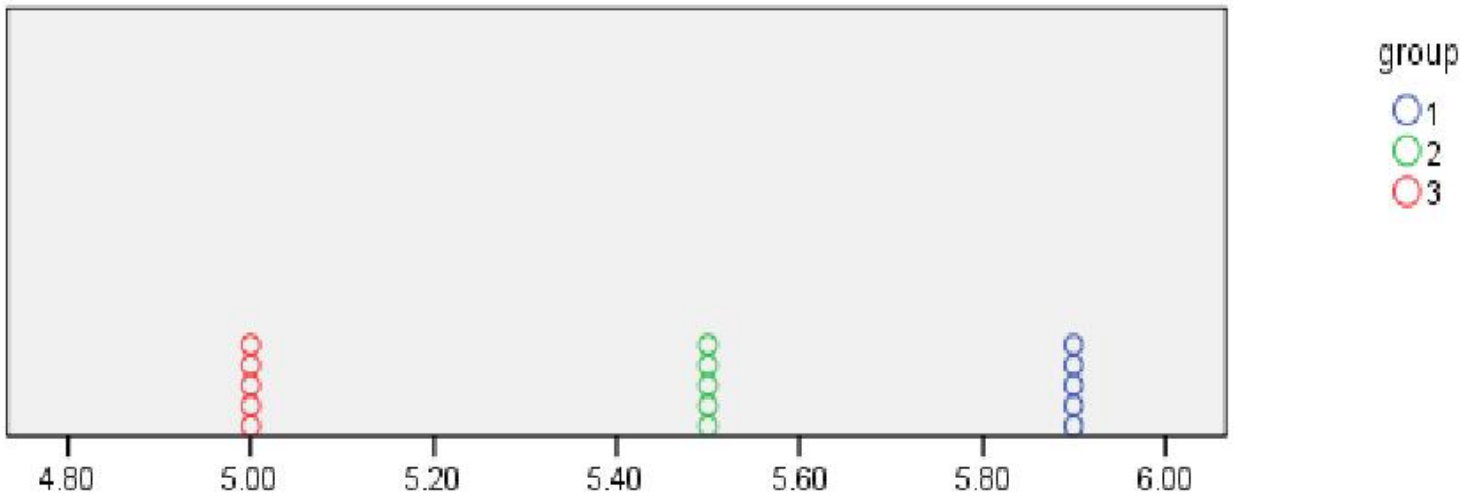


$\rightarrow$ the boxplots show the variability of the observations within a group and the variability between the groups

$\rightarrow$ **comparing the between-group with the within-group variability** is the key in detecting any significant difference between the groups

Between-group and within-group variability

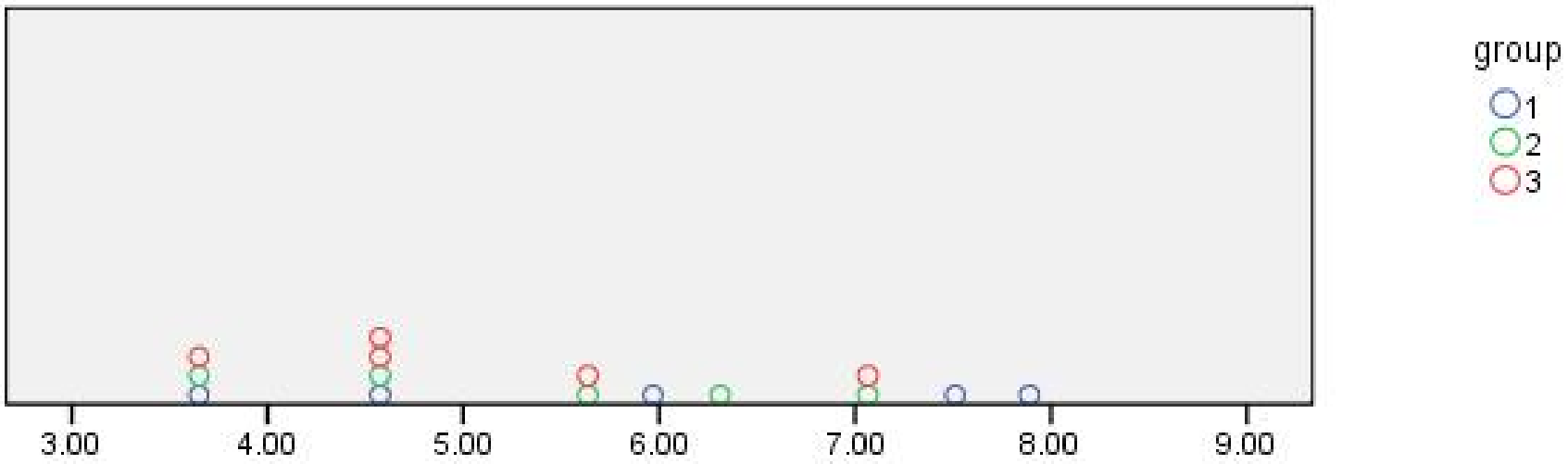
Group 1	Group 2	Group 3
5.90	5.51	5.01
5.92	5.50	5.00
5.91	5.50	4.99
5.89	5.49	4.98
5.88	5.50	5.02



Between-group variance $= 1.017$, within-group variance $= 0.00018$
(ratio $= 5545$)

Between-group and within-group variability

Group 1	Group 2	Group 3
5.90	6.31	4.52
4.42	3.54	6.93
7.51	4.73	4.48
7.89	7.20	5.55
3.78	5.72	3.52



Between-group variance $= 1.017$, within-group variance $= 2.332$
(ratio $= 0.436$)

Analysis of Variance

- Comparing the between-group variability and the within-group variability is the purpose of the **Analysis of Variance**

$\rightarrow$ often shortened to the acronym **ANOVA**

- Suppose that we have k different groups (k populations, or k sub-populations of a population) that we wish to compare. Often, each group is called a treatment or treatment level (general terms that can be traced back to the early applications of this methodology in the agricultural sciences)
- The response for each of the k treatments is the random variable of interest, say X
- Denote X_{ij} the jth observation ($j = 1, \ldots, n_i$) taken under treatment i

$\rightarrow$ we have k independent samples (one sample from each of the treatments)

ANOVA samples

The k random samples are often presented as:

Treatment	1	2	...	k
	X_{11}	X_{21}	...	X_{k1}
	X_{12}	X_{22}		X_{k2}
	$\vdots$	$\vdots$		$\vdots$
	X_{1n_1}	X_{2n_2}	...	X_{kn_k}
Mean	$\bar{X}_1$	$\bar{X}_2$	...	$\bar{X}_k$
St. Dev.	S_1	S_2	...	S_k

where $\bar{X}_i$ and S_i are the sample mean and standard deviation of the ith sample. The total number of observations is

$$n = n_1 + n_2 + \ldots + n_k$$

and the **grand mean** of all the observations, usually denoted $\bar{\bar{X}}$, is

$$\bar{\bar{X}} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} X_{ij} = \frac{n_1\bar{X}_1 + n_2\bar{X}_2 + \ldots + n_k\bar{X}_k}{n}$$

ANOVA model

The **ANOVA model** is the following:

$$\boxed{X_{ij} = \mu_i + \varepsilon_{ij}}$$

where

- μ_i is the mean response for the ith treatment ($i = 1, 2, \ldots, k$)
- ε_{ij} is an individual random error component ($j = 1, 2, \ldots, n_i$)

As usual for errors, we will assume that the random variables ε_{ij} are normally and independently distributed with mean 0 and variance σ^2:

$$\boxed{\varepsilon_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma)} \quad \text{for all } i, j$$

Therefore, each treatment can be thought of as a normal population with mean μ_i and variance σ^2:

$$\boxed{X_{ij} \overset{\text{ind.}}{\sim} \mathcal{N}(\mu_i, \sigma)}$$

Important: the variance σ^2 is common for all treatments

ANOVA hypotheses

We are interested in detecting differences between the different treatment means μ_i, which are population parameters

$\rightarrow$ hypothesis test!

The null hypothesis to be tested is

$$\boxed{H_0 : \mu_1 = \mu_2 = \ldots = \mu_k}$$

versus the general alternative

$$\boxed{H_a : \text{not all the means are equal}}$$

Careful! The alternative hypothesis should be that at least two of the means differ, not that they are all different !

As pointed out previously, the primary tool when testing for equality of the means is based on a comparison of the variances within the groups and between the groups.

Variability decomposition

The ANOVA partitions the total variability in the sample data, described by the **total sum of squares**

$$SS_{\text{Tot}} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \bar{\bar{X}})^2 \qquad (\text{df} = n - 1)$$

into the **treatment sum of squares** (= variability between groups)

$$SS_{\text{Tr}} = \sum_{i=1}^{k} n_i (\bar{X}_i - \bar{\bar{X}})^2 \qquad (\text{df} = k - 1)$$

and the **error sum of squares** (= variability within groups)

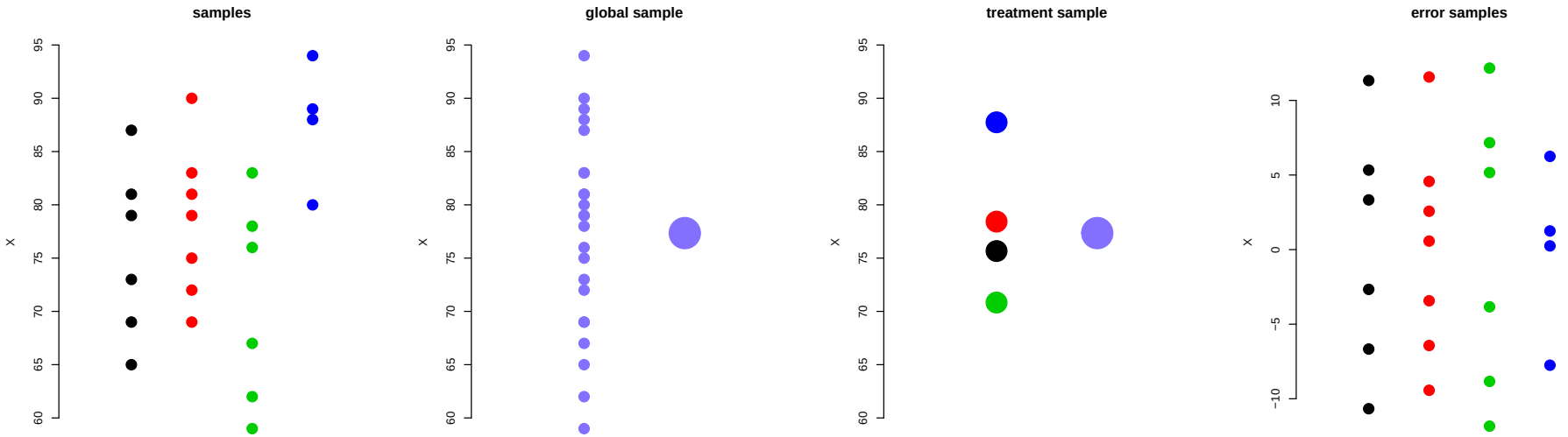
$$SS_{\text{Er}} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2 \qquad (\text{df} = n - k)$$

The sum of squares identity is

$$\boxed{SS_{\text{Tot}} = SS_{\text{Tr}} + SS_{\text{Er}}}$$

Variability decomposition

- The total sum of squares SS_{Tot} quantifies the total amount of variation contained in the global sample
- The Treatment sum of squares SS_{Tr} quantifies the variation 'between the groups', that is the variation between the means of the groups
- The Error sum of squares SS_{Er} quantifies the variation within the groups



Mean Squared Error

In sample i, the sample variance is given by $S_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2$

which is an unbiased estimator for σ^2: $\mathbb{E}(S_i^2) = \sigma^2$

Since,

$$SS_{\text{Er}} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2 = \sum_{i=1}^{k} (n_i - 1) S_i^2$$

hence

$$\mathbb{E}(SS_{\text{Er}}) = \sum_{i=1}^{k} (n_i - 1) \mathbb{E}(S_i^2) = \sigma^2 \sum_{i=1}^{k} (n_i - 1) = (n - k)\sigma^2$$

$\rightarrow$ another unbiased estimator for σ^2 is the **Mean Squared Error** MS_{Er}

$$\boxed{MS_{\text{Er}} = \frac{SS_{\text{Er}}}{n - k}}$$

(generalisation of the 'pooled' sample variance, Slide 21 Lecture 10)

$\rightarrow$ the number of degrees of freedom for this 'error' estimator of σ^2 is

$$\boxed{n - k}$$

Treatment mean square

Now **if H_0 is true**, that is if $\mu_1 = \mu_2 = \ldots = \mu_k = \mu$, we have $\bar{X}_i \sim \mathcal{N}(\mu, \frac{\sigma}{\sqrt{n_i}})$, that is $\sqrt{n_i}(\bar{X}_i - \mu) \sim \mathcal{N}(0, \sigma)$, for all $i = 1, \ldots, k$

$\rightarrow \sqrt{n_1}(\bar{X}_1 - \mu), \sqrt{n_2}(\bar{X}_2 - \mu), \ldots, \sqrt{n_k}(\bar{X}_k - \mu)$, is a random sample whose sample variance

$$\frac{1}{k-1}\sum_{i=1}^{k} n_i(\bar{X}_i - \bar{\bar{X}})^2 = \frac{SS_{\text{Tr}}}{k-1}$$

is an unbiased estimator for σ^2

$\rightarrow$ the **Treatment Mean Square** MS_{Tr}, defined by

$$\boxed{MS_{\text{Tr}} = \frac{SS_{\text{Tr}}}{k-1}}$$

is also an unbiased estimator for σ^2

$\rightarrow$ the number of degrees of freedom for this 'treatment' estimator of σ^2 is $\boxed{k-1}$

ANOVA test

Thus we have two potential estimators of σ^2:

1. MS_{Er}, which always estimates σ^2
2. MS_{Tr}, which estimates σ^2 only when H_0 is true

Actually, if H_0 is not true, MS_{Tr} tends to exceed σ^2, as we have

$$\mathbb{E}(MS_{\text{Tr}}) = \sigma^2 + \text{ 'true' variance between the groups}$$

$\rightarrow$ the idea of the ANOVA test now takes shape

Suppose we have observed k samples $x_{i1}, x_{i2}, \ldots, x_{in_i}$, for $i = 1, 2, \ldots, k$, from which we can find through calculations the observed values ms_{Tr} and ms_{Er}. Then:

- if $MS_{\text{Tr}} \simeq MS_{\text{Er}}$, then H_0 is probably reasonable
- if $MS_{\text{Tr}} \gg MS_{\text{Er}}$, then H_0 should be rejected

$\rightarrow$ this will thus be a one-sided hypothesis test

We need to determine what "$MS_{\text{Tr}} \gg MS_{\text{Er}}$" means so as to obtain a hypothesis test at given significance level α.

Sampling distribution

It can be shown that, if H_0 is true, the **ratio**

$$F = \frac{MS_{\text{Tr}}}{MS_{\text{Er}}} = \frac{\frac{SS_{\text{Tr}}}{k-1}}{\frac{SS_{\text{Er}}}{n-k}}$$

follows a particular distribution known as the **Fisher's F-distribution** with $k-1$ and $n-k$ degrees of freedom, which is usually denoted by

$$\boxed{F \sim \mathbf{F}_{k-1,n-k}}$$

Note: Ronald A. Fisher (1890-1962) was an English statistician and biologist. Some say that he almost single-handedly created the foundation for modern statistical science. As a biologist, he is also regarded as the greatest biologist since Charles Darwin

The Fisher's F-distribution

A random variable, say X, is said to follow Fisher's F-distribution with d_1 and d_2 degrees of freedom, i.e.

$$\boxed{X \sim \mathbf{F}_{d_1,d_2}}$$

if its probability density function is given by

$$f(x) = \frac{\Gamma((d_1+d_2)/2)(d_1/d_2)^{d_1/2}x^{d_1/2-1}}{\Gamma(d_1/2)\Gamma(d_2/2)((d_1/d_2)x+1)^{(d_1+d_2)/2}} \qquad \text{for } x > 0$$

for some integers d_1 and d_2 $\qquad \rightarrow S_X = [0, +\infty)$

Note: the Gamma function is given by

$$\Gamma(y) = \int_0^{+\infty} x^{y-1}e^{-x}\,dx, \qquad \text{for } y > 0$$

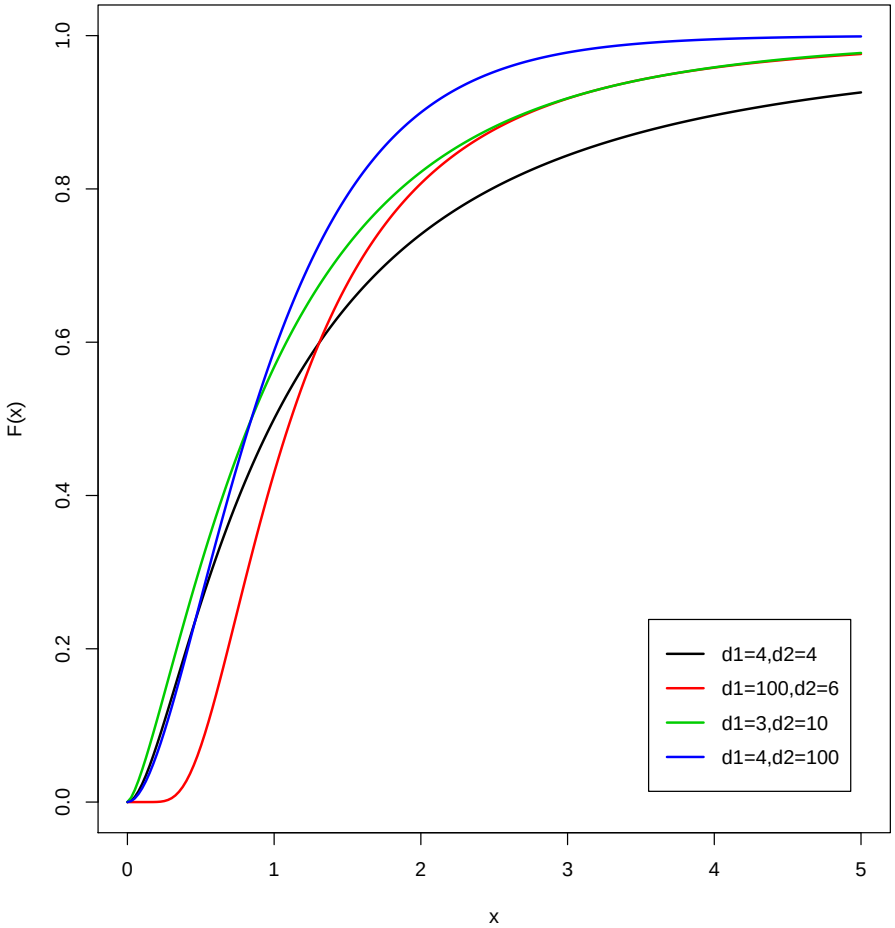
It can be shown that $\Gamma(y) = (y-1) \times \Gamma(y-1)$, so that, if y is a positive integer n,

$$\Gamma(n) = (n-1)!$$

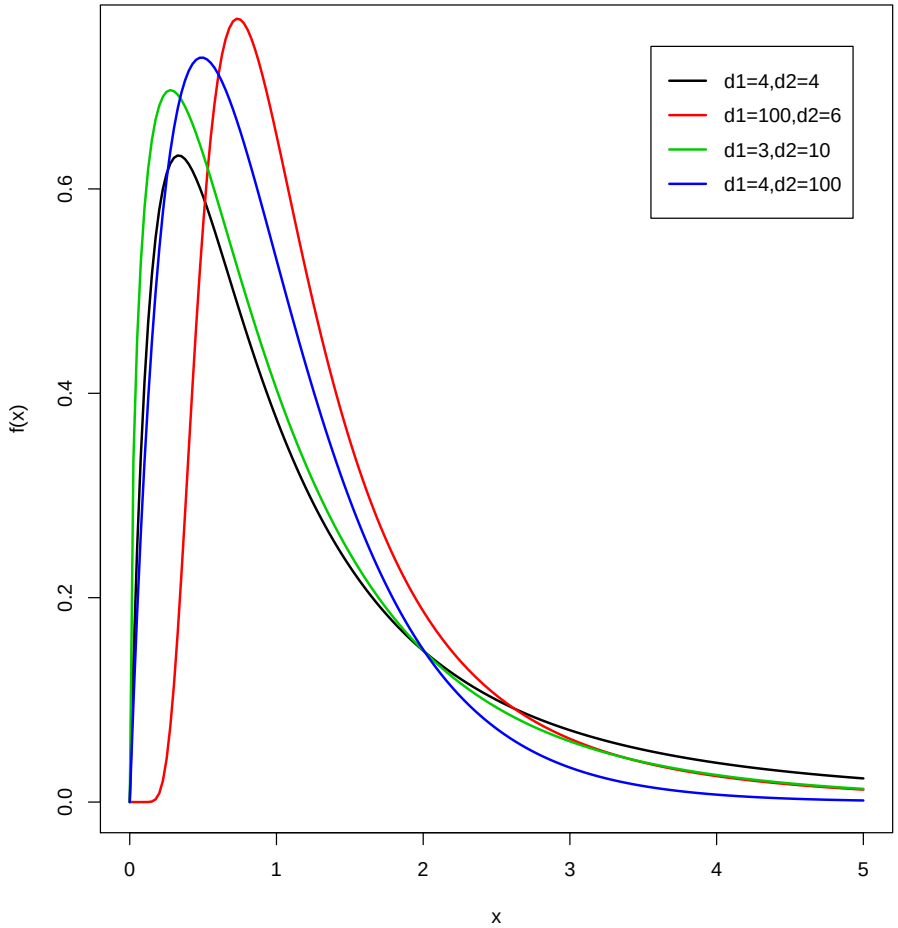
There is usually no simple expression for the F-cdf.

The Fisher's F-distribution

Some F-distributions



cdf $F(x)$



pdf $f(x) = F'(x)$

The Fisher's F-distribution

It can be shown that the mean and the variance of the F-distribution with d_1 and d_2 degrees of freedom are

$$\mathbb{E}(X) = \frac{d_2}{d_2 - 2} \qquad \text{for } d_2 > 2$$

and

$$\mathbb{V}\text{ar}(X) = \frac{2d_2^2(d_1 + d_2 - 2)}{d_1(d_2 - 2)^2(d_2 - 4)} \qquad \text{for } d_2 > 4$$

Note that a F-distributed random variable is **nonnegative**, as expected (ratio of two positive random quantities) and the distribution is **highly skewed to the right**.

The Fisher's F-distribution: quantiles

Similarly to what we did for other distributions, we can define the quantiles of any F-distribution:

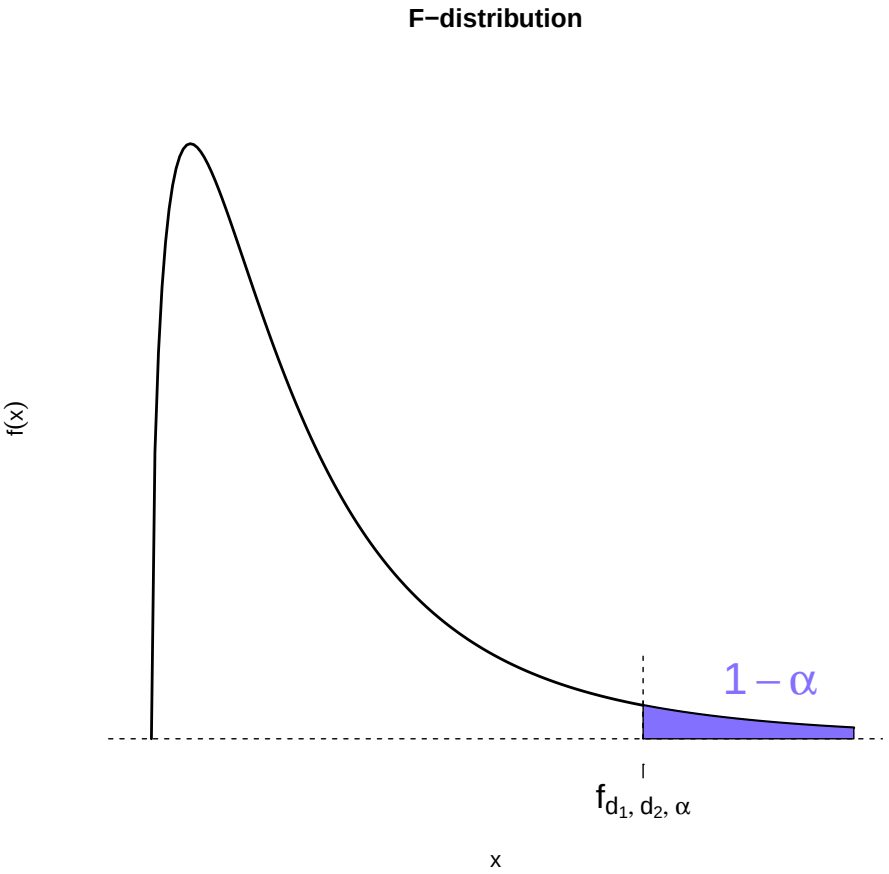
Let $f_{d_1,d_2;\alpha}$ be the value such that

$$\mathbb{P}(X > f_{d_1,d_2;\alpha}) = 1 - \alpha$$

for $X \sim \mathbf{F}_{d_1,d_2}$

The F-distribution is not symmetric, however it can be shown that

$$f_{d_1,d_2;\alpha} = \frac{1}{f_{d_2,d_1;1-\alpha}}$$



$f_{d_1,d_2;\alpha}$ is also referred to as F critical value.

ANOVA test

The null hypothesis to test is H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$

versus the general alternative H_a: not all the means are equal

Evidence against H_0 is shown if $MS_{\text{Tr}} \gg MS_{\text{Er}}$, so we will reject H_0 whenever MS_{Tr} is much larger than MS_{Er}, i.e. $\frac{MS_{\text{Tr}}}{MS_{\text{Er}}}$ much larger than 1

$\rightarrow$ for testing H_0 at significance level α, we need a constant c such that

$$\alpha = \mathbb{P}\left(\frac{MS_{\text{Tr}}}{MS_{\text{Er}}} > c \text{ if } H_0 \text{ is true}\right)$$

We know that, if H_0 is true, $F = \frac{MS_{\text{Tr}}}{MS_{\text{Er}}} \sim \mathbf{F}_{k-1,n-k}$

$\rightarrow$ we have directly that $c = f_{k-1,n-k;1-\alpha}$

From observed values ms_{Tr} and ms_{Er}, the decision rule is:

$$\text{reject } H_0 \text{ if } \quad \frac{ms_{\text{Tr}}}{ms_{\text{Er}}} > f_{k-1,n-k;1-\alpha}$$

ANOVA test: p-value

The observed value of the test statistic is

$$f_0 = \frac{ms_{\text{Tr}}}{ms_{\text{Er}}}$$

and thus the p-value is given by

$$p = \mathbb{P}(X > f_0),$$

where $X \sim \mathbf{F}_{k-1,n-k}$

("the probability that the test statistic will take on a value that is at least as extreme as the observed value when H_0 is true", definition on Slide 21 Lecture 9)

This test is also often called the ***F*-test** or **ANOVA *F*-test**.

ANOVA table

The computations for this test are usually summarised in tabular form

Source	degrees of freedom	sum of squares	mean square	F-statistic
Treatment	$df_{\text{Tr}} = k - 1$	ss_{Tr}	$MS_{\text{Tr}} = \frac{ss_{\text{Tr}}}{k-1}$	$f_0 = \frac{MS_{\text{Tr}}}{MS_{\text{Er}}}$
Error	$df_{\text{Er}} = n - k$	ss_{Er}	$MS_{\text{Er}} = \frac{ss_{\text{Er}}}{n-k}$	
Total	$df_{\text{Tot}} = n - 1$	ss_{Tot}		

Note 1: $df_{\text{Tot}} = df_{\text{Tr}} + df_{\text{Er}}$ and $ss_{\text{Tot}} = ss_{\text{Tr}} + ss_{\text{Er}}$

Note 2: this table is the usual computer output when an ANOVA procedure is run

ANOVA: example

Example

Consider the data shown on Slide 4. Test at significance level $\alpha = 0.05$ the null hypothesis that there is no difference in mean achievement for the four teaching techniques. (**Hint:** You can use the Matlab outputs: $\texttt{finv}(0.95, 3, 19) = 3.1274$, $\texttt{fcdf}(3.77, 3, 19) = 0.9719$)

We have $k = 4$, $n_1 = 6$, $n_2 = 7$, $n_3 = 6$ and $n_4 = 4$, with $\bar{x}_1 = 75.67$, $\bar{x}_2 = 78.43$, $\bar{x}_3 = 70.83$, $\bar{x}_4 = 87.75$ and $s_1 = 8.17$, $s_2 = 7.11$, $s_3 = 9.58$, $s_4 = 5.80$. Besides,

$$n = 6 + 6 + 7 + 4 = 23 \quad \text{and} \quad \bar{\bar{x}} = \frac{1}{n} \sum_{i=1}^{4} n_i \bar{x}_i = 77.35$$

Thus, from the expressions on Slides 15 and 16,

$$ss_{\text{Er}} = 5 \times 8.17^2 + 6 \times 7.11^2 + 5 \times 9.58^2 + 3 \times 5.80^2 = 1196.63$$

$$ss_{\text{Tr}} = 6 \times (75.67 - 77.35)^2 + 7 \times (78.43 - 77.35)^2 + 6 \times (70.83 - 77.35)^2 + 4 \times (87.75 - 77.35)^2 = 712.59$$

ANOVA: example

From there, the ANOVA table can be easily completed:

ANOVA: example

ANOVA: confidence intervals on treatment means

- The ANOVA F-test will tell you whether the means are all equal or not, but nothing more
- When the null hypothesis of equal means is rejected, we will usually want to know **which of the μ_i's are different from one another**
- A first step in that direction is to build confidence intervals for the different means μ_i
- From our assumptions (normal populations, random samples, equal variance σ^2 in each group), we have

$$\bar{X}_i \sim \mathcal{N}\left(\mu_i, \frac{\sigma}{\sqrt{n_i}}\right)$$

- The value of σ^2 is unknown, however we have (numerous!) estimators for it

ANOVA: confidence intervals on treatment means

For instance, the MS_{Er} is an unbiased estimator for σ^2 with $n - k$ degrees of freedom.

This one is based on all the n observations from the global sample

$\rightarrow$ it has smaller variance (i.e. it is more accurate) than any other (like e.g. S_i), and should always be used in the ANOVA framework!

Acting 'as usual', we can conclude that

$$\sqrt{n_i}\,\frac{\bar{X}_i - \mu_i}{\sqrt{MS_{\text{Er}}}} \sim t_{n-k}$$

and directly write a $100 \times (1 - \alpha)\%$ two-sided confidence interval for μ_i, from the observed values $\bar{x}_i$ and MS_{Er}:

$$\left[\bar{x}_i - t_{n-k,1-\alpha/2}\sqrt{\frac{MS_{\text{Er}}}{n_i}}, \bar{x}_i + t_{n-k,1-\alpha/2}\sqrt{\frac{MS_{\text{Er}}}{n_i}}\right]$$

$\rightarrow$ these confidence intervals for each group will tell which values μ_i's are much different from one another and which ones are 'close'

ANOVA: confidence intervals on treatment means

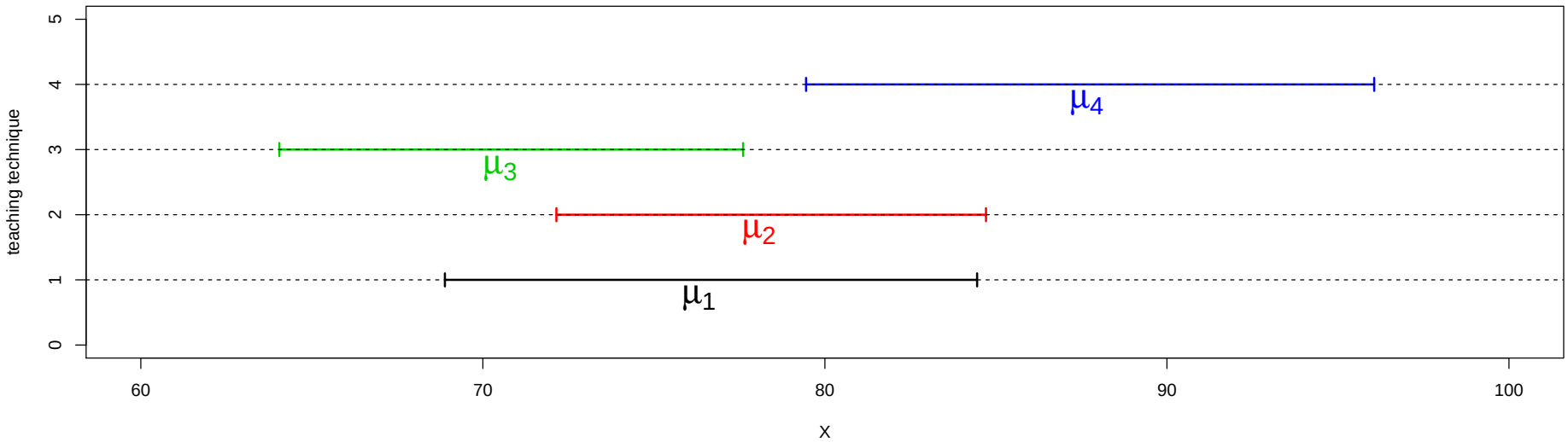
For instance, in the previous example (mean achievement for teaching techniques), we would find, with $t_{19;0.975} = 2.093$ (Matlab) and $ms_{\text{Er}} = 62.98$:

$$95\% \text{ CI for } \mu_1 = [75.67 \pm 2.093 \times \sqrt{\tfrac{62.98}{6}}] = [68.89, 84.45]$$

$$95\% \text{ CI for } \mu_2 = [78.43 \pm 2.093 \times \sqrt{\tfrac{62.98}{7}}] = [72.15, 84.71]$$

$$95\% \text{ CI for } \mu_3 = [70.83 \pm 2.093 \times \sqrt{\tfrac{62.98}{6}}] = [64.05, 77.61]$$

$$95\% \text{ CI for } \mu_4 = [87.75 \pm 2.093 \times \sqrt{\tfrac{62.98}{4}}] = [79.45, 96.06]$$



$\rightarrow$ it seems clear that $\mu_3 \neq \mu_4$ is the main reason for rejecting H_0

ANOVA: pairwise comparisons

It is also possible to build confidence intervals for the differences between two means μ_i and μ_j. From observed values $\bar{x}_i$, $\bar{x}_j$ and MS_{Er}, a $100 \times (1-\alpha)$ % confidence interval for $\mu_i - \mu_j$ is

$$\left[(\bar{x}_i - \bar{x}_j) - t_{n-k;1-\alpha/2} \sqrt{MS_{\text{Er}} \left(\frac{1}{n_i} + \frac{1}{n_j} \right)}, \right.$$
$$\left. (\bar{x}_i - \bar{x}_j) + t_{n-k;1-\alpha/2} \sqrt{MS_{\text{Er}} \left(\frac{1}{n_i} + \frac{1}{n_j} \right)} \right]$$

for any pair of groups (i, j)

Finding the value 0 in such an interval is an indication that μ_i and μ_j are not 'significantly' different. On the other hand, if the interval does not contain 0, that is evidence that $\mu_i \neq \mu_j$.

However, these confidence intervals are sometimes misleading and must be carefully analysed, in particular when related to the global null hypothesis $H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$

ANOVA: pairwise comparisons

Suppose that for a pair (i, j), the $100 \times (1 - \alpha)\%$ confidence interval for $\mu_i - \mu_j$ does not contain 0

$\rightarrow$ at significance level $\alpha\%$, you would reject $H_0^{(i,j)} : \mu_i = \mu_j$

If $\mu_i \neq \mu_j$ then automatically $H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$ is contradicted

$\rightarrow$ should you also reject H_0 at significance level $\alpha\%$? **No**

When you reject $H_0^{(i,j)} : \mu_i = \mu_j$ at significance level $\alpha\%$, you essentially keep a $\alpha\%$ chance of being wrong

Successively testing $H_0^{(1,2)}$: $\mu_1 = \mu_2$, and then $H_0^{(1,3)} : \mu_1 = \mu_3$, and then ..., and then finally $H_0^{(k-1,k)} : \mu_{k-1} = \mu_k$, that is

$$K = \binom{k}{2} = \frac{k!}{2!(k-2)!} \quad \text{pairwise comparisons,}$$

greatly increases the chance of making a wrong decision

ANOVA: pairwise comparisons

Suppose that H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$ is true

If the decisions made for each of the K pairwise tests $H_0^{(i,j)} : \mu_i = \mu_j$ were *independent* (which they are not! why?), we would wrongly reject at least one null hypothesis with probability $1 - (1 - \alpha)^K$ (why?)

If the decisions were *perfectly dependent* (which they are not either!), we would wrongly reject at least one null hypothesis with probability α (why?)

$\rightarrow$ if we based our decision about H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$ on the pairwise comparison tests, we would wrongly reject H_0 with a probability strictly between α and $1 - (1 - \alpha)^K$, larger than α !

To fix ideas, suppose $k = 4$ groups, which would give $K = \binom{4}{2} = 6$ pairwise comparisons, and $\alpha = 0.05$

$\rightarrow$ the test based on pairwise comparisons would be of effective significance level between 0.05 and $1 - (1 - 0.05)^6 = 0.265$

Pairwise comparisons: Bonferonni adjustments

It is usually not possible to determine exactly the significance level of such a test: it all depends on the exact level of dependence between the decisions about the different pairwise comparisons.

Several procedures have been proposed to overcome this difficulty, the simplest being the Bonferonni adjustment method.

It is based on the Bonferonni inequality (see Exercise 1 Tut. Week 5):

$$\mathbb{P}(A_1 \cup A_2 \cup \ldots \cup A_K) \leq \mathbb{P}(A_1) + \mathbb{P}(A_2) + \ldots + \mathbb{P}(A_K)$$

Suppose that A_q is the event 'we wrongly reject H_0 for the qth pairwise comparison'. Then, the event $B = (A_1 \cup A_2 \cup \ldots \cup A_K)$ is the event 'we wrongly reject $H_0 : \mu_1 = \ldots = \mu_k$'

$\rightarrow$ if we want $\mathbb{P}(B) \leq \alpha$, it is enough to take $\mathbb{P}(A_q) = \frac{\alpha}{K}$ for all q

Hence, to guarantee an overall significance level of at most α%, the pairwise comparison tests must be carried out at significance level α/K% (instead of α%), where $K = \binom{k}{2}$.

Bonferonni-adjusted t-test

To compare treatment i with treatment j, the null hypothesis is

$$H_0 : \mu_i = \mu_j$$

against the alternative

$$H_a : \mu_i \neq \mu_j$$

The test statistic is

$$t_0 = \frac{\bar{x}_i - \bar{x}_j}{\sqrt{MS_{Er}(\frac{1}{n_i} + \frac{1}{n_j})}}$$

The p-value is computed as

$$p = 2 \times P(T > |t_0|), \quad T \sim t_{n-k}$$

Reject H_0 if p-value is less than α/K, where K is the number of pairwise comparisons.

Pairwise comparisons: example

In our running example, we have $k = 4$ groups, and we can run $K = 6$ pairwise two-sample t-tests. We can find:

- t-test for $H_0 : \mu_1 = \mu_2$ $\rightarrow$ p-value = 0.5388
- t-test for $H_0 : \mu_1 = \mu_3$ $\rightarrow$ p-value = 0.3047
- t-test for $H_0 : \mu_1 = \mu_4$ $\rightarrow$ p-value = 0.0292
- t-test for $H_0 : \mu_2 = \mu_3$ $\rightarrow$ p-value = 0.1017
- t-test for $H_0 : \mu_2 = \mu_4$ $\rightarrow$ p-value = 0.0764
- t-test for $H_0 : \mu_3 = \mu_4$ $\rightarrow$ p-value = 0.0037

At level 5%, we reject $H_0 : \mu_1 = \mu_4$ and $H_0 : \mu_3 = \mu_4$

From this, can we reject $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$ at level 5%? No

$\rightarrow$ we must compare the above p-values to $\alpha/K = 0.05/6 = 0.0083$

The last p-value is smaller than 0.0083 $\rightarrow$ reject $H_0 : \mu_3 = \mu_4$

There is a difference between mean achievement for using teaching techniques 3 and 4 only.

Adequacy of the ANOVA model

The ANOVA model is based on several assumptions that should be carefully checked.

The central assumption here is that the random variables $\varepsilon_{ij} = X_{ij} - \mu_i$, $i = 1, \ldots, k$ and $j = 1, \ldots, n_i$, are (1) independent and (2) normally distributed:

$$\varepsilon_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma),$$

with (3) the same variance in each group.

We do not have access to values for ε_{ij} (μ_i's are unknown!), however we can approximate these values by the observed residuals

$$\hat{e}_{ij} = x_{ij} - \bar{x}_i$$

Note that these residuals are the quantities arising in ss_{Er}.

$\rightarrow$ as for a regression model (see Slides 41-42 Week 11), the adequacy of the ANOVA model is established by examining the residuals $\rightarrow$ **residual analysis**

Residuals analysis

- The **normality** assumption can be checked by constructing a **normal quantile plot** for the residuals
- The assumption of **equal variances** in each group can be checked by plotting the residuals against the treatment level (that is, $\bar{x}_i$)

$\rightarrow$ the spread in the residuals should not depend on any way on $\bar{x}_i$

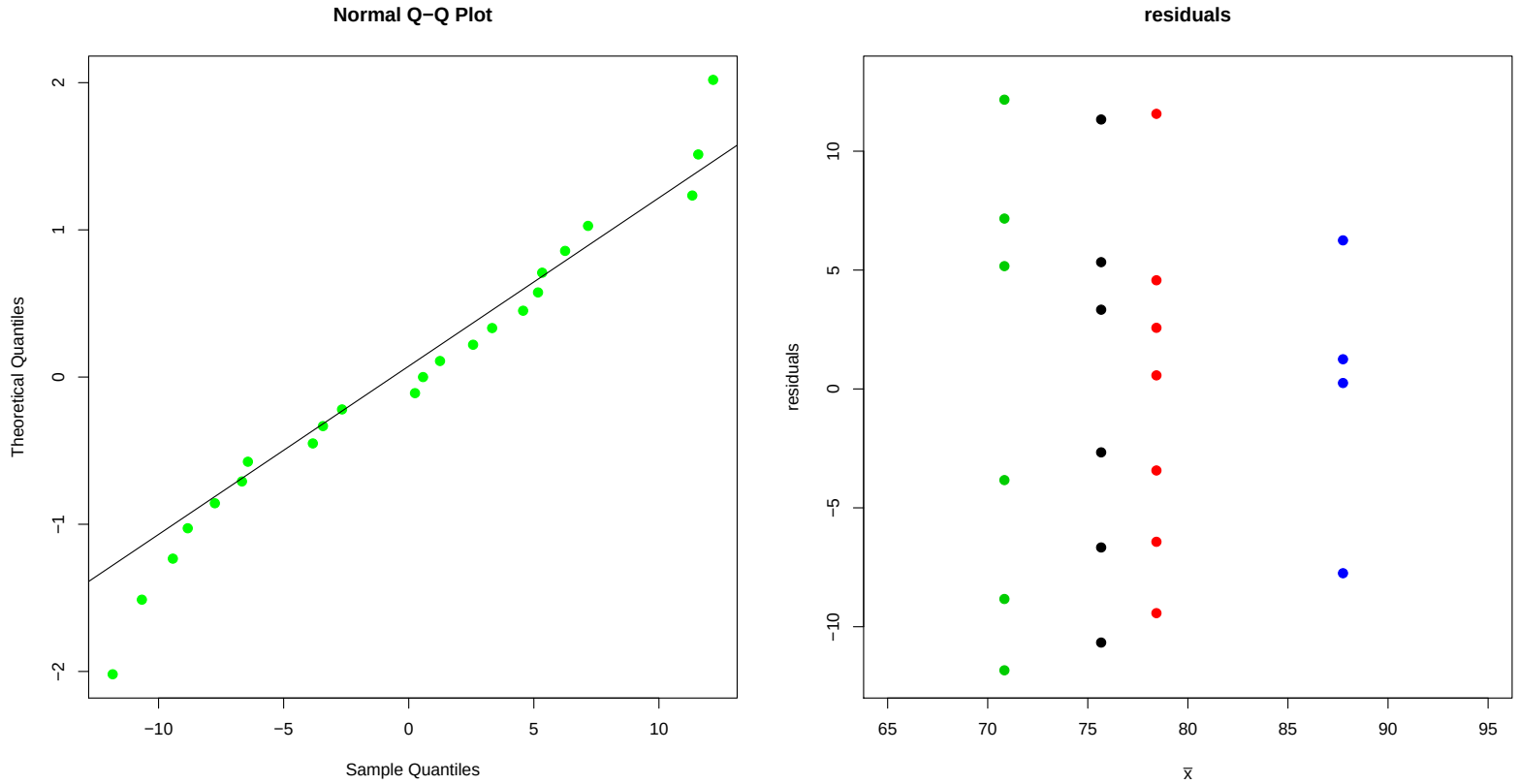
- A rule-of-thumb is that, if the ratio of the largest sample standard deviation to the smallest one is smaller than 2, the assumption of equal population variances is reasonable
- The assumption of **independence** can be checked by plotting the residuals against time, if this information is available

$\rightarrow$ no pattern, such sequences of positive and negative residuals, should be observed

- As for the regression, the residuals are everything the model will not consider $\rightarrow$ no information should be observed in the residuals, they should look like random noise

Residual analysis: example

For our running example, a normal quantile plot and a plot against the fitted values $\bar{x}_i$ for the residuals are shown below:



$\rightarrow$ nothing (obvious) to report

$\rightarrow$ the assumptions we made look valid

Residual analysis: example

Example

To assess the reliability of timber structures, researchers have studied strength factors of structural lumber. Three species of Canadian softwood were analysed for bending strength (Douglas Fir, Hem-Fir and Spruce-Pine-Fir). Wood samples were selected from randomly selected sawmills. The results of the experiment are given below. Is there any significant difference in the mean bending parameters among the three types of wood? (**Hint:** You can use the Matlab outputs: $\texttt{finv}(0.95, 2, 15) = 3.68$, $\texttt{fcdf}(0.33, 2, 15) = 0.274$)

Douglas (1)	Hem (2)	Spruce (3)
370	381	440
150	401	210
372	175	230
145	185	400
374	374	386
365	390	410

$\rightarrow$ an ANOVA was run to test the null hypothesis $H_0 : \mu_1 = \mu_2 = \mu_3$, against the alternative H_a : not all the means are equal

Residual analysis: example

We computed values for the ANOVA table:

Source	degrees of freedom	sum of squares	mean square	F-statistic
Treatment	2	7544	3772	0.33
Error	15	172929	11529	
Total	17	180474		

According to the hint, we know that $f_{2,15;0.95} = 3.68$

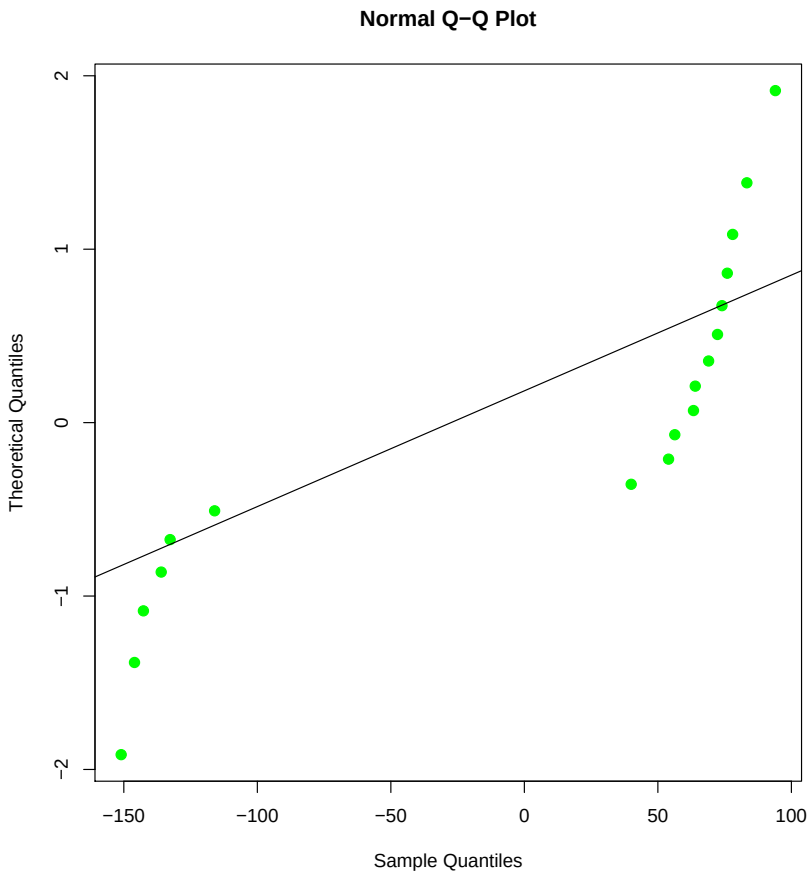
$\rightarrow$ here we have observed $f_0 = 0.33 \rightarrow$ do not reject H_0 !

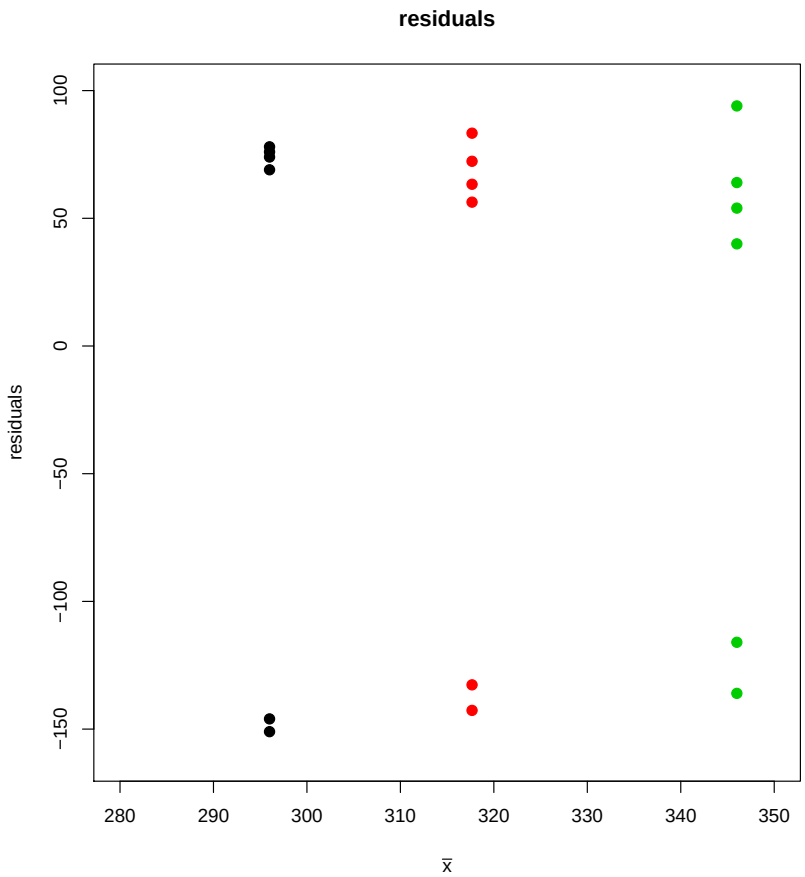
Associated p-value: $p = \mathbb{P}(X > 0.33) = 1 - 0.274 = 0.726$ for $X \sim \mathbf{F}_{2,15}$

$\rightarrow$ we confidently claim that there is no significant difference in the mean bending parameters for the different wood types

Residual analysis: example

Residual analysis:



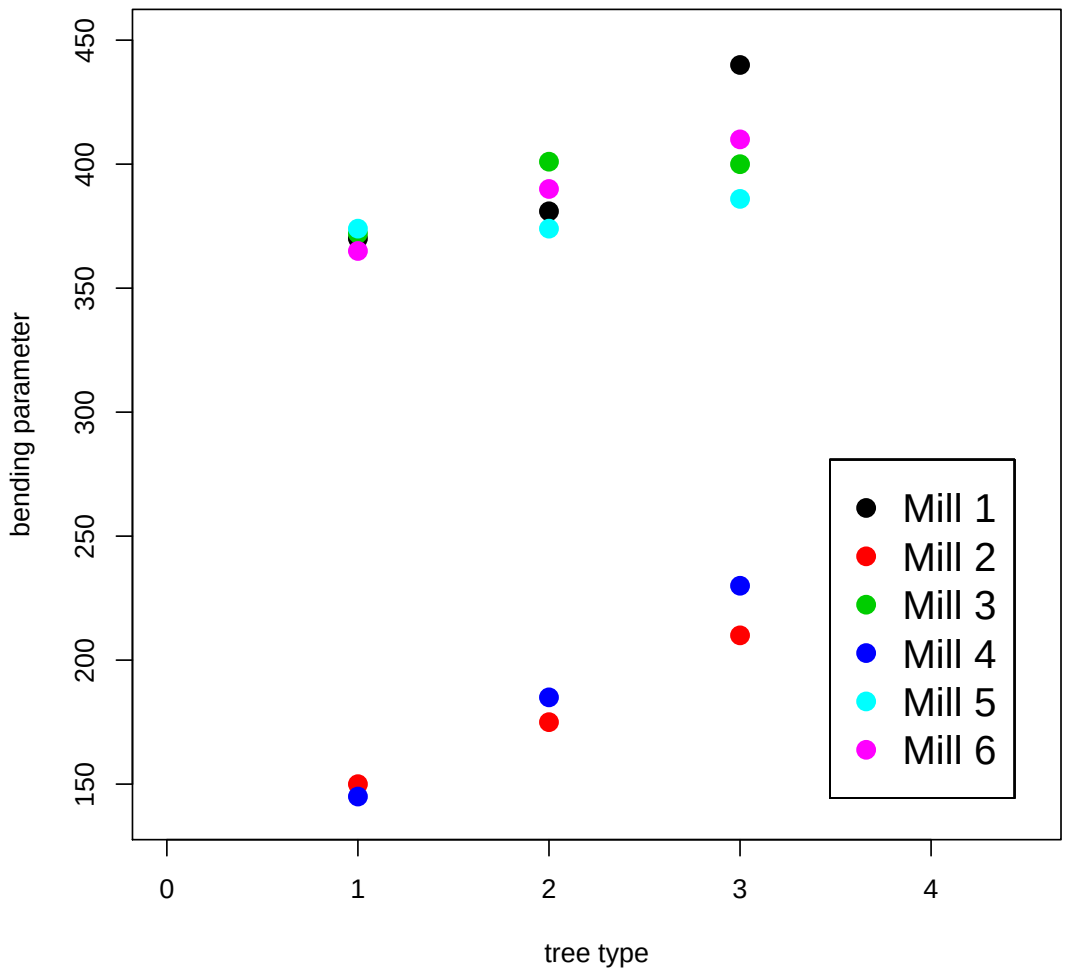


$\rightarrow$ the assumptions are clearly not fulfilled!

$\rightarrow$ the above conclusion is certainly not reliable!

Blocking factor

If we had plotted the data first, we would have seen:



(Bottom line: always plotting the data before analysing them!)

Blocking factor

It is clear that over and above the wood type, the mills where the lumber was selected is **another source of variability**, in this example even more important than the main treatment of interest (wood type).

This kind of extra source of variability is known as a **blocking factor**, as it essentially groups some observations in blocks across the initial groups $\rightarrow$ the samples are not independent! (assumption violation)

$\rightarrow$ a potential blocking factor must be taken into account!

When a blocking factor is present, the 'initial' Error Sum of Squares, say SS^*_{Er}, that is the whole amount of variability not due to the treatment, can in turn be partitioned into:

1. the variability due to the blocking factor, quantified by SS_{Block}
2. the 'true' natural variability in the observations SS_{Er}

We can write that $SS^*_{\text{Er}} = SS_{\text{Block}} + SS_{\text{Er}}$, and thus

$$SS_{\text{Tot}} = SS_{\text{Tr}} + SS_{\text{Block}} + SS_{\text{Er}}$$

Blocking factor

The ANOVA table becomes:

Source	degrees of freedom	sum of squares	mean square	F-statistic
Treatment	$k-1$	ss_{Tr}	$MS_{\text{Tr}} = \frac{ss_{\text{Tr}}}{k-1}$	$f_0 = \frac{MS_{\text{Tr}}}{MS_{\text{Er}}}$
Block	$b-1$	ss_{Block}	$MS_{\text{Block}} = \frac{ss_{\text{Block}}}{b-1}$	
Error	$n-k-b+1$	ss_{Er}	$MS_{\text{Er}} = \frac{ss_{\text{Er}}}{n-k-b+1}$	
Total	$n-1$	ss_{Tot}		

where b is the number of blocks

Note: the test statistic is again the ratio $\frac{MS_{\text{Tr}}}{MS_{\text{Er}}}$ (we have just removed the variability due to the blocking factor first), to be compared with the quantile of the $\mathbf{F}_{k-1,n-k-b+1}$ distribution

Blocking factor

In the previous example, we would have found:

Source	degrees of freedom	sum of squares	mean square	F-statistic
Treatment	2	7544	3772	15.87
Block	5	170552	34110	
Error	10	2378	238	
Total	17	180474		

From MATLAB, we can find that $f_{2,10;0.95} = 4.10$

Here, we have observed $f_0 = 15.87 \rightarrow$ clearly reject H_0!

Associated p-value: $p = \mathbb{P}(X > 15.87) = 0.0008$ for $X \sim \mathbf{F}_{2,10}$

Blocking factor: comments

- The SS_{Er} in the first ANOVA (without block) was 172,929 which contains an amount of variability 170,552 due to mills

$\rightarrow$ about 99% of the initial SS^*_{Er} was due to mill to mill variability, and so was no *natural variability*!

- The second ANOVA (with blocking factor) adjusts for this effect
- The net effect is a substantial reduction in the 'genuine' MS_{Er}, leading to a larger F-statistic (increased from 0.33 to 15.87!)

$\rightarrow$ with very little risk of being wrong ($p \simeq 0$), we can now conclude that there is a significant difference in the mean bending parameters for the three different wood types

- An analysis of the residuals in this second ANOVA would not show anything peculiar $\rightarrow$ valid conclusion

Generally speaking, ignoring a blocking factor leads to a misleading conclusion, and it should always be carefully assessed whether a blocking factor may exist or not (plot the data!)

Objectives

Now you should be able to:

- conduct engineering experiments involving a treatment with a certain number of levels ☐
- understand how the ANOVA is used to analyse the data from these experiments ☐
- assess the ANOVA model adequacy with residual plots ☐
- understand the blocking principle and how it is used to isolate the effect of nuisance factors ☐

Recommended exercises:

→ Q3, Q6 p.406, Q9 p.407, Q10, Q11 p.412, Q13, Q15, Q17 p.413, Q19 p.414, Q22, Q23 p.415, Q35 p.428 (2nd edition)
→ Q3, Q6 p.418, Q9 p.418, Q10, Q11 p.423, Q13, Q15 p.424, Q17, Q19 p.425, Q20 p.426, Q23 p.427, Q35 p.439 (3rd edition)