14

ANOVA

QUESTION 1.

You have been asked by your local sports club to compare three brands of cricket balls in relation to the distance they travel. The balls are chosen at random and are 'bowled' by a machine used during practice sessions. The distance each ball travels is recorded and shown in the table below.

Brand A	Brand B	Brand C
246	243	265
231	246	260
236	243	265
217	235	253
246	235	291

These data can be analysed using single factor ANOVA.
Construct the null and alternative hypotheses:

H_0: $\mu_A = \mu_B = \mu_C$.
H_1: At least 2 of the means are different.

State three assumptions you can make:

Assumption 1: Samples are randomly selected.
Assumption 2: Probability distributions are normally distributed.
Assumption 3: Population variances are equal.

Complete the following:

The variance for brand A = 145.7
The variance for brand B = 25.8
The variance for brand C = 207.2

The sum of squares between groups = 2870.933
The sum of squares within groups = 1514.8
The total sum of squares = 4385.733

The between groups degrees of freedom, $df_b = 2$
The within groups degrees of freedom, $df_w = 12$
The total degrees of freedom, $df = 14$

The mean sum of squares for the model, $MS_M = \frac{SS_M}{df_M} = 1435$

The mean sum of squares for the residuals, $MS_R = \_\_\_\_\_ = 126$

The value for $F_{test} = \frac{MS_M}{MS_R} = 11.37$

The value for $F_{stat} = 3.89$

The evidence suggests that the null hypothesis should be rejected since $p > 0.002$
The interpretation of this is that at least one of the means is different from the others.

QUESTION 2.

The headteacher of your old school has heard that you are an expert statistician. She has asked you to help her provide some evidence that ethnicity has no impact on the mathematics attainment of year 7 pupils. You decide to look at four ethnic groups (labelled A,B, C and D in the table below) and three classes with each class having its own teacher. The table below shows the data you collected when the children were given a maths test.

	GP A	GP B	GP C	GP D
Teacher 1	4.5	6.4	7.2	6.7
Teacher 2	8.8	7.8	9.6	7.0
Teacher 3	5.9	6.8	5.7	5.2

Your task is find out if there is a significant difference due to the teacher and find out if the ethnicity of the child is significant.

Construct the two null and two alternative hypotheses:

Rows:

$H_0^{row}: \mu_1 = \mu_1 = \mu_2 = \mu_3$

$H_1^{row}: \mu_1 \neq \mu_2 \neq \mu_3$

Columns:

$H_0^{col}: \mu_1 = \mu_1 = \mu_2 = \mu_3 = \mu_4$

$H_1^{col}: \mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4$

The row means = 6.2, 8.3, 5.9
The column mean = 6.4, 7.0, 7.5, 6.3
The grand mean = 6.8

The variation of the row means from the grand mean $v_r = 13.68$
The variation of the column means from the grand mean $v_c = 2.82$
The total variation $v = 23.08$
The random variation $v_e = 6.58$

The row $df = 2$
The column $df = 3$

Complete the following:

For the row means at the 0.05 level of significance $F_{stat} = 5.14$
The calculated F value = 6.24

Since 6.24 > 5.14, the null hypothesis can be rejected.
The conclusion is at the 0.05 significance level there is an significant difference in the test results due to the teacher.

The calculated F value = 0.86
Since 0.86 > 1, the null hypothesis can be accepted.
The conclusion is at the 0.05 significance level there is an insignificant difference in the test results due to ethnicity.

QUESTION 3.

The figure below shows the output from Excel when a two-factor experiment without replacement was carried out.

Anova: Two-Factor Without Replication						
SUMMARY	*Count*	*Sum*	*Average*	*Variance*		
a	4	24.8	6.2	1.39		
b	4	33.2	8.3	1.29		
c	4	23.6	5.9	0.45		
cr1	3	19.2	6.4	4.81		
cr2	3	21	7	0.52		
cr3	3	22.5	7.5	3.87		
cr4	3	18.9	6.3	0.93		
ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Rows	13.68	2	6.84	6.24	0.03	5.14
Columns	2.82	3	0.94	0.86	0.51	4.76
Error	6.58	6	1.10			
Total	23.08	11				

Looking at the ANOVA table, in the box below give an explanation for each of the sources of variation (think about the grand mean):

Rows Depending on the context and looking at the *F* values and the *p*-value indicates there is a difference and the null hypothesis can be rejected. The null hypothesis being the grand mean is a good model.
Columns Depending on the context and looking at the *F* values and the *p*-value indicates there is a not a difference and the null hypothesis cannot be rejected. The null hypothesis being the grand mean is a good model.
Error Using the randomised block design means the data has been blocked to reduce the variability among the variables to reduce the effects of the random error and hence produce a better interpretation of the data.

In the box below give an explanation of F and F_{crit} and describe their relationship

F This is the computed F statistic from the data.
F_{crit} This is the F value used to compare the computed value to.
F and F_{crit} are used to determine If the computed value is greater than this the null hypothesis can be rejected.

In the box below explain what the *p*-value indicates:

The *p*-value can be used to help make the decision either to reject the null hypothesis or accept it.

QUESTION 4.

The sports club (from question 1) would now like to know which brands of cricket balls are different.

What sort of test could you do to compare the means of the three brands? One-way ANOVA.

What specific test could you use to compare the means? A randomised block design.

The value of the least significant difference (LSD) statistic comparing brand A with brand B is found by:

$$\text{LSD} = t_{\alpha/2}\sqrt{MS_R(\frac{1}{n_i}+\frac{1}{n_j})} = 2.179\sqrt{126\left(\frac{1}{5}+\frac{1}{5}\right)} = 16.34$$

The null hypothesis is rejected if $|\bar{x}_i - \bar{x}_j| > \text{LSD}$.

$|\bar{x}_i - \bar{x}_j| > 16.34$.

$5.2 < \text{LSD}$, therefore the null hypothesis cannot be rejected.

The value of the LSD statistic comparing brand A with brand C is found by:

$$\text{LSD} = t_{\alpha/2}\sqrt{MS_R(\frac{1}{n_i}+\frac{1}{n_j})} = 2.179\sqrt{126\left(\frac{1}{5}+\frac{1}{5}\right)} = 16.34$$

The null hypothesis is rejected if $|\bar{x}_i - \bar{x}_j| > \text{LSD}$.

$31.6 > 16.34$.

The null hypothesis is rejected.

The value of the LSD statistic comparing brand B with brand C is found by:

$$\text{LSD} = t_{\alpha/2}\sqrt{MS_R(\frac{1}{n_i}+\frac{1}{n_j})} = 16.34$$

The null hypothesis is rejected if $|\bar{x}_i - \bar{x}_j| >$ LSD.

$26.4 > 16.34$.

The null hypothesis is rejected.

MINI PROJECT

The university you work for are looking at pre-university preparatory courses. The management are looking at two programmes: one is a 10 day condensed course and the other a 30 day course. They are also considering two delivery options: one is the programme being taught in a classroom, just like any 'traditional' programme and the other delivered on line via the Internet.

They would like you to perform a statistical analysis and advise them which option would potentially improve the courses.

You collect data from 10 randomly selected students studying on each of the programmes.

The table below shows the collected data:

Course delivery	10 day	10 day	30 day	30 day
Traditional	26	18	34	28
Traditional	27	24	24	21
Traditional	25	19	35	35
Traditional	21	20	31	29
Traditional	21	18	28	26
Online	27	21	24	21
Online	29	32	16	19
Online	30	20	22	19
Online	24	28	20	24
Online	30	29	23	25

The management want a report which will tell them if there is a statistically significant interaction between the length of course and the type of course at a 0.05 level of significance.

Your report should describe the statistical analysis technique you have chosen and why you chose to use it. It should also contain sufficient information so the management can decide from four options:

Option 1: Both courses delivered in a 'traditional' classroom.

Option 2: Both courses delivered online.
Option 3: The 10 day course delivered online and the 30 day course delivered in a classroom.
Option 4: The 10 day course delivered in a classroom and the 30 day course delivered online.

And finally…

Interacting with statistics can be ANOVER *F*-test!