Analysis of Variance

The results obtained in an investigation into the stability of a fluorescent reagent stored under three different conditions. The values for the fluorescence signals are:

Group1	Group2	Group3
23	27	24
23	29	26
20	25	24
21	23	
	24	

We tested the null hypothesis that the mean fluorescence signals are the same for the three different conditions(groups).

H_0 : μ1 = μ2 = μ3

H_0: not all the means are equal (at least one mean is different).

We are now interested in perform this hypothesis test by using the ***analysis of variance*** method. There are three groups with n_1=4, n_2=5 and n_3=3 observations per group respectively. We denote group j values by y_j and store them into three vectors in R.

```
y1 = c(23, 23, 20, 21)
y2 = c(27, 29, 25, 23, 24)
y3 = c(24, 26, 24)
y = c(y1, y2, y3)
```

We also specify which group these observations belong to. We create a second long vector, called group, identifying group membership:

```
group = c(1, 1, 1, 1, 2, 2, 2, 2, 2, 3, 3, 3)

#Alternative Approach – Useful for larger data sets

group = c(rep(1,4),rep(2,5),rep(3,3))
```

Categorical Data

The values stored in the vector group are displayed as numeric values ***1, 2*** and ***3***, but in fact they are levels of a categorical variable indicating the group level. We could have used the letters ***a, b*** or ***c*** to define the three groups. In ***R***, we must make the distinction between factors and integers, hence we redefine the group variable as a factor.

```
group = factor(group)
flordata = data.frame(y,group)
```

We can now place the two ***long*** vectors together in a dataframe called ***flordata*** such that each row will have a value with the observed fluorescence signal and its corresponding group level.

Type in the name of the data frame, and compute the dimensions. Constrast this structure to the table presented on the previous page.

The form shown on your screen is called ***long form***, which is easier for a computer to work with, whereas the table on the previous page is known as ***wide form***. Wide form is easier for a human eye to read and interpret.

```
flordata
dim(flordata)
```

To run the analysis of variance we use the command `aov` and then use summary to view the ANOVA output.

```
> model = aov(y ~ group,  flordata)
> summary(model)
```

The parameters of the `aov` function that need specified are

- the response data, ***y***
- the categorical factor ***group*** mentioned after the ~ symbol
- the data frame, ***flordata***, that contains y and group variables

In your submission sheet, write down the values for each of the following terms

The group row in the output contains information about the Variation Between Groups

[a] degrees of freedom = number of groups-1 = k-1
[b] Sum of Squares Between Groups = SSBg
[c] Mean Square Between Groups = MSBg = SSBg/k-1

The Residuals row in the output contains information about the Variation Within Groups

[d] degrees of freedom

$$=\sum_{j=1}^{k} n_j - 1$$

[e] Sum of Squares Within Groups (aka Sum of Squared Errors)
[f] Mean Square Within Groups = MSE =SSE/df

If the null hypothesis is correct, then the two estimates of variance (between and within groups) should not differ significantly. If it is incorrect, the between-groups variance will be greater than the within group variance.

The test statistic is the ratio of the MS values (i.e MSBg/MSW). Also included in the table is the associated p-value. The degrees of freedom for the test statistic are the degrees of freedom already computed.

In your submission sheet write down the test statistic and p-value ([g] and [h] respectively).

Interpreting the p-value is dependent on how stringent we wish to be.

- Using a very stringent threshold of p-value =0.01 we would fail to reject the null.
- Using a less stringent threshold, p-value =0.05, we would reject the null.

Part 2
Five analysts each made 10 determinations of the paracetamol content of the same batch of tables. The results are shown below.

A	84.55	84.61	84.26	84.36	84.66	84.31	84.65	84.41	84.52	84.44
B	84.12	84.04	83.95	84.51	84.08	84.07	84.35	83.99	84.25	84.14
C	84.44	84.48	84.14	84.17	84.31	84.60	84.44	84.24	84.64	84.47
D	84.05	84.14	84.53	84.07	84.45	83.95	84.10	84.29	84.13	83.98
E	84.09	84.53	84.60	84.48	84.42	84.57	84.35	84.30	84.37	84.63

We are interested in finding out whether there is a significant difference between the paracetamol contents obtained by the five analysts. Before testing the null hypothesis that the mean paracetamol contents are the same for the five experimenters:

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$$

$$H_1 : \text{not all the means are equal}$$

```
a = c(84.55, 84.61, 84.26, 84.36, 84.66, 84.31, 84.65, 84.41, 84.52, 84.44)
b = c(84.12, 84.04, 83.95, 84.51, 84.08, 84.07, 84.35, 83.99, 84.25, 84.14)
c = c(84.44, 84.48, 84.14, 84.17, 84.31, 84.60, 84.44, 84.24, 84.64, 84.47)
d = c(84.05, 84.14, 84.53, 84.07, 84.45, 83.95, 84.10, 84.29, 84.13, 83.98)
e = c(84.09, 84.53, 84.60, 84.48, 84.42, 84.57, 84.35, 84.30, 84.37, 84.63)

y = c(a, b, c, d,e)

group = rep(1:5, each = 10)
group
group = factor(group)
```

Before running the analysis of variance, you should graph the means and standard errors of each group of data. Illustrate these results with parallel boxplots (one for each treatment).

```
plot(group,y)
```

The side-by-side boxplots indicate that the data in the five groups might have different means but similar variances.

Place the two long vectors y and group together in a unifying dataframe called ***paracetamol*** and use the analysis of variance function in R to perform the test.

```
paracetamol = data.frame(y, group)
model = aov(y ~ group,paracetamol)
```

Both the commands `summary(model)` or `anova(model)` contain the summary of the ANOVA performed.

Using the output, comment on your conclusion for the null hypothesis.