Lab week 5

ANOVA in R

Consider the following example that we have already discussed in class.

Example. 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).

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

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

The numerical details of the ANOVA method are included in the lecture notes from week 4. We are now interested in performing the analysis of variance in R. 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.

$> \ y_1 \ <- \ c(23, 23, 20, 21)$

$> \ y_2 \ <- \ c(27, 29, 25, 23, 24)$

$> \ y_3 \ <- \ c(24, 26, 24)$

Next we combine them into one long vector, $\mathbf{y}$:

$> \ y \ <- \ c(y_1, y_2, y_3)$

or we can create the long vector directly as:

$> \ y \ <- \ c(23, 23, 20, 21, 27, 29, 25, 23, 24, 24, 26, 24)$

We also need to tell R which group these observations belong to. We create a second long vector, called **group**, identifying group membership:

$> \ group \ <- \ c(1, 1, 1, 1, 2, 2, 2, 2, 2, 3, 3, 3)$

The value 1 indicates membership to group 1, 2 indicates membership to group 2 and 3 indicates membership to group 3. 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 factor 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:

$> \ group \ <- \ factor(group)$

Place the two long vectors together in a unifying dataframe called **flordata**

$> \ flordata = data.frame(y, group)$

such that each row will have a value with the observed fluorescence signal and its corresponding group level.

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

$> \ model \ = \ aov(y \sim group, flordata)$

$> \ summary(model)$

where the parameters of the *aov* function that need specified are

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

The output from R confirms the results obtained in the class.

Analysis of Variance Table

Response: y

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
group	2	34.3007	17.1500	4.7322	0.03941	⋆
Residuals	9	32.617	3.6241			

Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

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

- degrees of freedom = number of groups-1 = k-1 = 3-1=2
- Sum of Squares Between Groups = SSB = 34.30007
- Mean Square Between Groups = MSB = $\frac{SSB}{df} = \frac{34.0007}{2} = 17.15$

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

- degrees of freedom =$\sum_{j=1}^{k} n_j - 1$ = (4-1)+(5-1)+(3-1)=9
- Sum of Squares Within Groups = Sum of Squared Errors = SSE = 32.617
- Mean Square Within Groups = MSE =$\frac{SSE}{df} = \frac{32.617}{9} = 3.6241$

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. To test wether it is significantly greater, a **one-sided F-test** is used. Using the mean squares in the $MeanSq$ column of this table, we do a variance ratio test to obtain the test statistic:

- F value = $\frac{MSB}{MSE} = \frac{17.15}{3.6241} = 4.7322$

The critical value $F_{2,9;0.05}$= 4.256495 is read from the F distribution with 2 and 9 degrees of freedom respectively and for significance level α=0.05. The test statistics can be read in R using the **qf**

function:

> qf(0.05, 2,9, lower.tail = FALSE)

or

> qf(0.95, 2,9, lower.tail = TRUE)

The test statistic 4.7322 is greater than the critical value of 4.256495, hence we reject the null hypothesis.

We can draw the same conclusion from comparing the p-value and the significance level $\alpha = 0.05$. The associated p-value for this test statistic is 0.03941 which is statistically significant if we compare it to the significance level $\alpha = 0.05$. The p-value associated with the 4.7322 test statistic can be obtained using the **pf** function.

> pf(4.7322,2,9,lower.tail=FALSE)

By either approach we reject the null hypothesis that the means are equal across different treatments. The mean of at least one group is different from the other means.