**Assignment No. 12.2**

This assignment also considers the same data set as Assignment 12.1 which is related to the scores obtained under various methods regarding the localisation site of a protein.

The following questions need to be answered as part of the assignment

1. a. What are the assumptions of ANOVA, test it out? (Using the data set above)
2. b. Why ANOVA test? Is there any other way to answer the above question?

As per standard Statistics Text Books, the following assumptions underlie the ANOVA

1. The dependent variable(target variable/response variable), which is a numerical type continuous variable, is expected to follow normal distribution, across all the various categories(groups) of a population or across different samples/populations. This also means that the residuals from the ANOVA test should also follow normal distribution

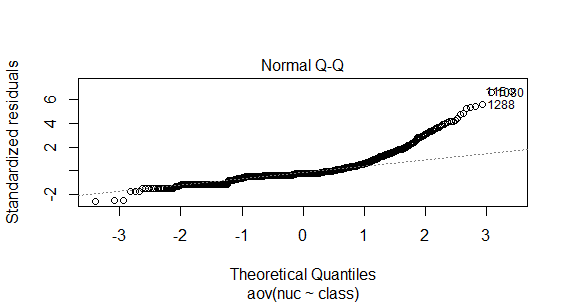
2. The variance of the dependent variable should be same (homoscedastic) across all the various categories(groups) of a population or across different samples/populations. This also means that the residuals from the ANOVA test should also follow homoscedasticity

3. The observations of the dependent variable (target variable/response variable) should independent in its distribution or occurrence. Meaning its observed/recorded value should not be dependent on any of its previous or subsequent values or other events which lead to that value.

We continue the same example in Assignment 12.1 where the ANOVA test is conduct for discriminant scores in the variable “NUC” by “Class” variable. Using the residuals of this ANOVA test, we test for the assumptions of the ANOVA test.

TESTING FOR NORMALITY OF DEPENDENT VARIABLE/RESIDUALS

> plot(result,2)



The above QQ plot shows a clear departure of normality of the dependent variable and the residuals. Hence the data on discriminant scores is not normally distributed across classes

> residuals <- residuals(object = result)

> shapiro.test(x = residuals)

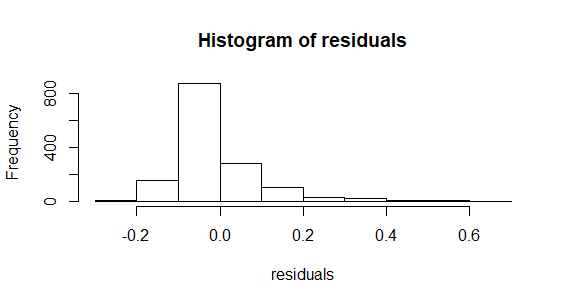
Shapiro-Wilk normality test

data: residuals

W = 0.7959, p-value < 2.2e-16

The Shapiro Wilk test also conforms that dependent variable or the residuals DO NOT FOLLOW NORMAL DISTRIBUTION.

> hist(residuals)



The Histogram of residuals also shows that the dependent variable or the residuals DO NOT FOLLOW A NORMAL DISTRIBUTION.

Given that the normality of dependent variables assumption is violated, we can use the Kruskal Wallis Test on the data to verify consistency of the discriminant scores across classes.

> summary(classnuc$nuc)

Min. 1st Qu. Median Mean 3rd Qu. Max.

0.0000 0.2200 0.2200 0.2762 0.3000 1.0000

> hist(classnuc$nuc)

> kruskal.test(nuc~class,data=classnuc)

Error in kruskal.test.default(c(0.22, 0.22, 0.22, 0.22, 0.22, 0.22, 0.22, :

all group levels must be finite

This error is corrected by observing the class of the “class” variable in the classnuc dataframe. It is found that it belongs to the class character. But the KW test requires that the group variable should have finite values. Hence we convert the character into factor and re-run the test. And it worked.

> classnuc$class <- as.factor(classnuc$class)

> kruskal.test(nuc~class,data=classnuc)

Kruskal-Wallis rank sum test

data: nuc by class

Kruskal-Wallis chi-squared = 198.04, df = 9, p-value < 2.2e-16

The Kruskal-Wallis test shows clearly REJECTS THE NULL HYPOTHESIS that there is no significant difference between the mean discriminant scores of various classes. So we accept the alternate hypothesis that there are significant differences in the mean discriminant scores across the various classes or localization points.

TESTING FOR HOMOGENIETY OF VARIANCE

ANOVA TEST

> result <- aov(nuc~class,data=classnuc);

> summary(result)

Df Sum Sq Mean Sq F value Pr(>F)

class 9 1.993 0.22141 22.01 <2e-16 \*\*\*

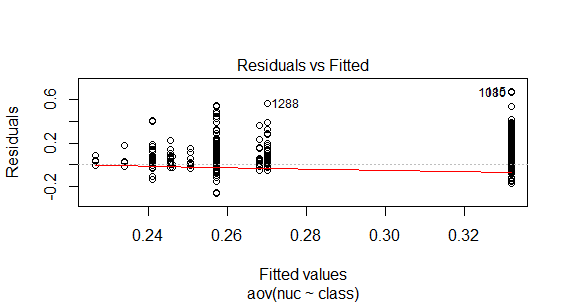
Residuals 1474 14.825 0.01006

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

The ANOVA test REJECTS the NULL hypothesis that there is no significant difference in the mean discriminant scores across the various classes. But we test further to what extent can we base our conclusion on this ANOVA test, since it is based on the above assumptions. In case these assumptions about the data are relaxed then the above test cannot be relied upon.

> plot(result,1) # Plotting the residuals of ANOVA test to see their distribution



The above plot shows that there is a declining red line from left to right and it shows a faint relationship between residuals and the means of each group. So to investigate further we conduct the Levene Test and the results are as follows

> leveneTest(nuc ~ class, data = classnuc)

Levene's Test for Homogeneity of Variance (center = median)

Df F value Pr(>F)

group 9 16.506 < 2.2e-16 \*\*\*

1474

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

The p-value being quite less than 0 clearly suggests that the residuals are not homoscedastic and they are heteroscedastic. It means that the discriminant scores across various classes have shown extreme variation across classes. This is also evident in the box plot itself when we were using descriptives to observe the same.

So another way we can attempt to find out whether there is significant difference in the discriminant scores across classes, relaxing the assumption of unequal variance is to undertake the One-Way Welch Test.

> oneway.test(nuc ~ class, data = classnuc)

One-way analysis of means (not assuming equal variances)

data: nuc and class

F = 23.555, num df = 9.000, denom df = 83.991, p-value < 2.2e-16

The Welch test which does not assume equal variances of the dependent variable or the residuals also produces a p-value less than 0 and hence we can conclude that THERE IS SIGNIFICANT DIFFERENCE IN THE MEAN discriminant scores across the various classes in the data set.

There is a possibility for undertaking pair-wise t-test and one can test across all the various combinations to conclude the specific group which displays the highest or lowest mean, or to identify the group which leads to the distortion in the ANOVA test. It is the difference a particular group/groups which leads to the REJECTION of the ANOVA test, meaning that the differences in the means across the various groups are not same. This also leads to the increase of F test, since the formula of F test goes as Sum of Squares of differences in mean between groups / Sum of Square of differences in mean within the group.

This is how we can accommodate the departures from the assumptions. Since the samples are drawn from various tests and from various centres, their independence can be trusted.

Finally we conclude that ANOVA is not the appropriate test for this data, due to non-normality, and heteroscedasticity. Hence instead we use the Kruskal Wallis Test to accommodate departure from normality, and One-Way Welch test to accommodate departure from homoscedasticity in data and both the tests confirm the classic ANOVA test that there are significant differences in the discriminant scores across the various classes.

The other ways to ANSWER THE QUESTION WITHOUT ANOVA TEST are as follows

1. pair wise t-test
2. One-Way Welch Test
3. Transform the Data into Log values and then perform ANOVA
4. Weighted Least Squares regression and then perform ANOVA on residuals
5. Kruskal-Wallis Test ONLY when the dependent variable is non-continuous
6. Logistic Regression only when the dependent variable is non-continuous