13
Analysis of Variance

13.1 General description

In previous chapters, methods to compare population means were presented, based on independent random variables. In many cases, the object of the analysis is to compare a given factor or parameter for different populations, to see if they differ. If three population were compared for a given factor, three independent, random samples would be selected from the three populations, and an *inference* would be obtained. Most likely the sample means would differ, but this does not necessarily imply a difference among the population means for the three groups. What is the statistical method to evaluate if the differences in sample means are large enough to imply the corresponding population means are different ? The method is called **Analysis of Variance**, shortened into **Anova**.

As described, sometime the differences among *samples* means are such that they are not sufficient to *infer* reliable information about differences among *populations*. Therefore the **Analysis of the Variance** must be performed.

When the distribution of the numerical value of a variable, for various groups is of interest, the analysis of the variance is performed. It consists in the separation of the total deviance (the numerator of the variance) in deviance among groups and within groups.

The main idea is the decomposition of the total sum of squares (the numerator of the variance) into a sum of squares within the groups (intra-group) and a sum of squares among groups (inter-groups).

ANOVA estimates three sample variances: a total variance based on all the observation deviations from the grand mean, an error variance based on all the observation deviations from their appropriate treatment means, and a treatment variance. The treatment variance is based on the deviations of treatment means from the grand mean, the result being multiplied by the number of observations in each treatment to account for the difference between the variance of observations and the variance of means.

The fundamental technique is a partitioning of the total sum of squares **TSS** into components related to the effects used in the model. An example is presented. There are three groups of four plants each, who have been applied three treatments to reduce the incidence of a fungal disease, (treatA, treatB and treatC). There are s samples collected from n different populations, nornally distributed and independent, we want to verify that

$$H_0 = \mu_1 = \mu_2 = \mu_3 = ... = \mu_n \tag{13.1}$$

We assume that the variances are equal:

$$\sigma_1 = \sigma_2 = \sigma_3 = ... = \sigma \tag{13.2}$$

The null hypothesis H_0 is tested, therefore test that all the population are equal. There are s samples of numerosity n_k, of the k_{th} sample.
Now the deviation within the sample (**SSW**) must be calculated:

$$\mathbf{SSW} = \sum_{k=1}^{s} \sum_{i=1}^{n_k} (x_{ki} - \bar{x}_k)^2 \tag{13.3}$$

This indicates the elements x_{ki} of the $k-th$ sample, minus the mean of the $k-th$ sample, squared. This was I calculate the sum of square difference for each sample, and then I sum from $k = 1tos$, of all the residuals for each sample.

Then the sum of squares between samples (SSB):

$$\mathbf{SSB} = \sum_{k=1}^{s} (\bar{x}_k - \bar{x})^2 \times n_k \tag{13.4}$$

Given by the difference between the mean of the $k-th$ sample minus the grand mean $\bar{x}$, multiplied by the number of samples.

13.2 Example

The table below shows the reduction of fungal incidence as number of damages per leaf, for each individual for the three treatments. The total value is $n = 12$ and the sample number is $s = 3$:

```
> treatA <- c(2,6,9,4)
> treatB <-c(11,6,14,13)
> treatC<-c(21,1,8,13)
> TREATMENT<-data.frame(treatA,treatB,treatC)
> str(TREATMENT)
'data.frame': 4 obs. of  3 variables:
 $ treatA: num  2 6 9 4
 $ treatB: num  11 6 14 13
 $ treatC: num  21 1 8 13
```

To compute this variable the means are first computed:

```
> mean(TREATMENT$treatA)
[1] 5.25
> mean(TREATMENT$treatB)
[1] 11
> mean(TREATMENT$treatC)
[1] 10.75
```

To calculate the SSW we must take the squared differences:

- treatA $(2-5.25)^2$, $(6-5.25)^2$, $(9-5.25)^2$, $(4-5.25)^2$
- treatB $(11-11)^2$, $(6-11)^2$, $(14-11)^2$,$(13-11)^2$
- treatC $(21-10.75)^2$, $(1-10.75)^2$, $(8-10.75)^2$, $(13-10.75)^2$

These are computed and summed for each population

```
> SSW_treatA=(2-5.25)^2 + (6-5.25)^2 + (9-5.25)^2+ (4-5.25)^2
> SSW_treatA
[1] 26.75

> SSW_treatB=(11-11)^2+ (6-11)^2+ (14-11)^2+ (13-11)^2
> SSW_treatB
[1] 38

> SSW_treatC=(21-10.75)^2+ (1-10.75)^2+ (8-10.75)^2+(13-10.75)^2
> SSW_treatC
[1] 212.75
```

and for all the populations, with a total sum 224.

```
> SSW_treatA+SSW_treatB+SSW_treatC
[1] 277.5
```

The SSB is then computed by computing the grand mean $\bar{x}$:

```
> sum(treatA,treatB,treatC)/12
[1] 9

> mean(c(treatA,treatB,treatC))
[1] 9
```

The mean of all the population is 9. Now subtracting the value from each sample mean:

- $((5.25-9)^2)\times 4$, $((11-9)^2)\times 4$, $((10.75-9)^2)\times 4 = 84.5$

computed in R as follows:

```
> SSB=((5.25-9)^2)* 4 + ((11-9)^2)* 4 +((10.75-9)^2)* 4
> SSB
[1] 84.5
```

The sum of the squared difference is then multiplied by the total number of elements: Therefore the within sample variability (**SSW**) is 224 and the between sample variability (**SSB**) is 84.5

The null hypothesis is now tested with the ANOVA. The **SSB** describes the systematic difference among populations, therefore the factor for which the samples were subdivided. The **SSW** explains how much of the reduction of fungal incidence is due to accidental factors, therefore factors that we did not considered in our analysis.

If we divide the **SSB** by the factor σ^2

$$\mathbf{SSB} = \frac{\sum_{k=1}^{s}(\bar{x}_k - \bar{x})^2 \times n_k}{\sigma^2} = \chi^2(g = s-1) \tag{13.5}$$

a χ^2 is obtained with $(g = s - 1)$ degree of freedom.

If we divide the **SSW** by the factor σ^2

$$\mathbf{SSW} = \frac{\sum_{k=1}^{s} \sum_{i=1}^{n_k} = (x_{ki} - \bar{x}_k)^2}{\sigma^2} = \chi^2(g = n - s) \tag{13.6}$$

a χ^2 is obtained with $(g = s - 1)$ degree of freedom. When these two variables are divided:

$$\mathbf{F} = \frac{SSB/(s-1)}{SSW/(n-s)} \sim F(s-1, n-s) \tag{13.7}$$

We obtain the ratio between two independent χ^2 variables, related to the respective degree of freedoms. The two variances cancel out, therefore we are left with the equation above. This variable is called the Fisher variable.

If the null hypothesis is reject, it mean that the means of the populations are different. In this case the **SSB** should be singificantly larger than the **SSW**, the numerator should be significanly larger than the denominator and the **F** statistic should be larger than 1, and therefore it will fall into the *rejection range*.

$$\mathbf{F} = \frac{SSB/(s-1)}{SSW/(n-s)} = \frac{84.5/(3-1)}{277.5/(12-3)} \tag{13.8}$$

The numerical value of the **F** factor is

```
> (84.5/(3-1))/(277/(12-3))
[1] 1.37
```

Assuming an α value of 0.95, therefore a confidence of 005. Now we can use the Fisher table with 2 and 9 degrees of freedom.
On the F-table I look for 2 degrees of freedom for the numerator and 9 for the denominator. The value is 4.26. Since the F value is smaller than 4.26 the null hypothesis cannot be rejected, therefore there are no significant differences.

$$\mathbf{F} = 1.37 < F_{(2,9,0.05)} = 4.26 \longrightarrow H_0 \longrightarrow \textit{not rejected} \tag{13.9}$$

Therefore with a 5% significant difference, the three treatments to reduce the incidence of a fungal disease *did not display any statistical difference.*

In general if the **F** value is too high, the null hypothesis is not realistic. If **F** is high it means there are significant differences, while the *p-value* express the probability that the null-hypothesis is true. Therefore with a small number the probability is small and it can be rejected.

There are numerous F Table formats based on α values of 0.10, 0.05, 0.025, 0.01, etc. Listed below is a partial F Table for α = 0.05.

ν_2 \ ν_1	1	2	3	4	5	6	7	8	9	10
1	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	241.9
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98

F Critical Values (α=0.05)

Note: The above critical values for F may be used for a one-tailed test (α=0.05, 95% confidence) or a two-tailed test (α/2=0.05, 90% confidence).

13.3 One way ANOVA

13.4 Test for normality

13.5 Example: Soil properties and Land Cover

The example presented here deals with soil physical and chemical data presented im Chapter 5.

As described above, when starting a statistical analysis, the command `summary`, when applied to a `data.frame` object, allows for obtaining a first description of the variables studied. If it is a (partial) one factor distribution of the absolute frequencies is reported, while, if the variable is numerical, they are indicate basic statistics such as mean, median, quartiles and extremes.

```
##   PROF COOR_X COOR_Y ALT VEG
## C1 : 1  Min. :635763 Min. :4894668 Min. :1553 CON :6
## C10 : 1 1st Qu.:636386 1st Qu.:4895264 1st Qu.:1680 FAG :8
## C11 : 1 Median :636851 Median :4895858 Median :1726 MIRT :6
## C12 : 1 Mean :636754 Mean :4895906 Mean :1771 PRATO:8
## C13 : 1 3rd Qu.:637192 3rd Qu.:4896572 3rd Qu.:1862
## C14 : 1 Max. :637590 Max. :4896925 Max. :2134
## (Other):22
## LIT Ca_stock C_stock_30 C_stock_0
## CEV:23 Min. : 2.000 Min. : 69.0 Min. : 5.00
## MOD: 5 1st Qu.: 2.850 1st Qu.:131.2 1st Qu.: 37.50
## Median : 3.700 Median :165.5 Median : 56.00
## Mean : 5.625 Mean :167.5 Mean : 57.71
## 3rd Qu.: 6.700 3rd Qu.:189.2 3rd Qu.: 80.25
```

```
## Max. :18.800 Max. :327.0 Max. :116.00
##
```

The variables can then be considered individually, also in this case distinguishing between the variable type. If the variable is categorical, the frequency distribution can be obtained with the table command and, dividing by the number of observations, the distribution of relative frequencies is also obtained. The concepts of frequency distribution and their representation were presented in Chapter 9, when descriptive statistics was introduced. With the command `table` it is possible to obtain the relative frequencies for the categorical variable.

```
table(Profiles$VEG)/nrow(Profiles)
```

```
BEECH       CON     GRASS       MYR
0.2857143 0.2142857 0.2857143 0.2142857 The command table also allows for
```

obtaining bi−variate relationship among variables, by generating contingency tables. A *contingency table* (also known as a *cross tabulation* or *crosstab*) is a type of table in a matrix format that displays the (multivariate) frequency distribution of the variables. By specifying two variables, in this case vegetation (VEG) and lithology (LIT) a table is obtained

```
> veg_lit<-table(Profiles$VEG,Profiles$LIT)
> veg_lit
```

```
        CEV MOD
BEECH   8   0
CON     6   0
GRASS   4   4
MYR     5   1
```

For instance, this table shows that the beech and coniferous trees were found on the CEV lithology but not on the MOD. The relative frequency is then obtained by applying the instruction `prop.table()` to the table:

```
#Relative frequency
prop.table(veg_lit)
```

```
                CEV        MOD
   BEECH 0.28571429 0.00000000
   CON   0.21428571 0.00000000
   GRASS 0.14285714 0.14285714
   MYR   0.17857143 0.03571429
```

The command allows for compute the relative frequency by row

```
> #Relative frequency by row
> prop.table(veg_lit,margin = 1)
```

```
              CEV       MOD
BEECH 1.0000000 0.0000000
CON   1.0000000 0.0000000
GRASS 0.5000000 0.5000000
MYR   0.8333333 0.1666667
```

Again showing a zero frequency for Beech and Con for the MOD substrate and a 50 % distribution for Grass.

The command allows also for compute the relative frequency by column

```
> #Relative frequency by column
> prop.table(veg_lit,margin = 2)
```

```
             CEV       MOD
BEECH 0.3478261 0.0000000
CON   0.2608696 0.0000000
GRASS 0.1739130 0.8000000
MYR   0.2173913 0.2000000
```

Showing that on the CEV substrate the frequency of the Beech tree was the highest.

As described above the descriptive values (point estimators can also be computed) as well as quantiles for the *numerical values*.

```
### Numerical variables
#Mean
mean(Profiles$Ca_stock)
#standard deviation
sd(Profiles$Ca_stock)
#Variance
var(Profiles$Ca_stock)
#Min
min(Profiles$Ca_stock)
#Max
max(Profiles$Ca_stock)
#Quantiles
quantile(Profiles$Ca_stock,probs=c(0.25,0.5,0.75))

# More variables simultaneously: use of function apply
apply(X = Profiles[,7:9],MARGIN = 2,FUN = mean)
apply(X = Profiles[,7:9],MARGIN = 2,FUN = sd)
```

Histograms can then be plotted to obtain a first visual assessment of the variables distribution. In some cases it is useful to log transform the data, if the distribution is particularly skewed. We must always remember that the fundamental assumption for application of ANOVA is a normal distribution and homogeneity of the variance (homoscedasticity).

```
#Histograms
```

```
hist(Profiles$Ca_stock,main = "Istogramma Ca stock", xlab = "Ca stock")

#Log transform
hist(log(Profiles$Ca_stock),
main = "Istogramma log Ca stock", xlab = "log Ca stock")

#create new log transformed variable
Profiles$log_Ca_stock<-log(Profiles$Ca_stock)
```

To elucidate the behavior of a numeric variable with respect to sub−populations identified by a particular factor (in this case for instance the vegetation cover), the ANOVA is applied as detailed above. Before applying ANOVA, a first exploratory phase is the evaluation of mean and standard deviation in the different subgroups. The function `aggregate` allows for computing different function (FUN), such as mean and standard deviation, to multiple variables aggregated for a specific factor. The `by` argument requires the specification of the factors that identify the groups within the list structure:

```
aggregate(x = Profiles[,c(4,7:9)],by = list(Profiles$VEG),FUN = mean)
```

```
  Group.1      ALT Ca_stock C_stock_30 C_stock_O
1   BEECH 1661.625 6.975000   147.1250  57.87500
2     CON 1649.833 7.200000   210.5000  41.50000
3   GRASS 1915.500 4.412500   137.5000  47.50000
4     MYR 1847.000 3.866667   191.6667  87.33333
```

and the same operations is performed for the standard deviation:

```
aggregate(x = Profiles[,c(4,7:9)],by = list(Profiles$VEG),FUN = sd)
```

```
  Group.1       ALT Ca_stock C_stock_30 C_stock_O
1   BEECH  66.48724 5.330974   50.62590  23.05545
2     CON  42.61181 6.017641   58.51068  31.00806
3   GRASS 161.52576 3.587055   46.38965  21.62010
4     MYR  69.08256 1.782882   75.70381  26.95676
```

Evaluation of these properties can also be performed by plotting boxplots:

```
boxplot(Profiles$log_Ca_stock~Profiles$VEG)
```

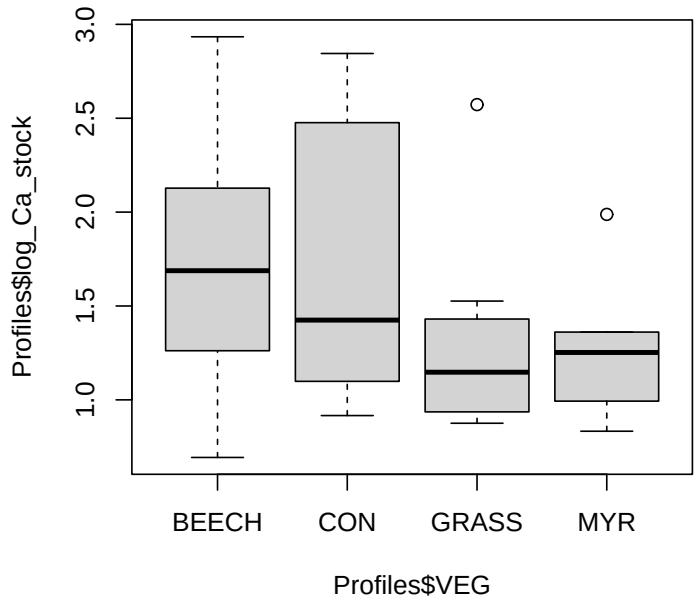

Fig. 13.1 Boxplot for Calcium stock as function of vegetation coverage

By using the tilde symbol, the instructions specifies that the variable *log_Ca_stock* is plotted as function of the VEG variable. The box plot depicts distributions which are not symmetric except for the Beech tree. The GRASS and the MYR present two outliers but smaller variations. These results are clearly visible in the data where the CON displayed a higher standard deviation.

Now an ANOVA is performed. As described above the two fundamental conditions for applying an ANOVA are **normality** of the distribution and **homoskedasticity**.

```
#One way ANOVA on Ca_stock

m_anova1<- aov(formula = Ca_stock ~ VEG,data = Profiles)
#results
summary(m_anova1)
```

```
             Df Sum Sq Mean Sq F value Pr(>F)
VEG           3   59.8   19.93   0.984  0.417
Residuals    24  486.0   20.25
```

The command `summary` print the results of the analysis with the degree if freedom (Df), the Sum of squared residuals, the mean of squared residuals, the F and the p-values. It is called the ANOVA table, in which a greater p-value of the confidence threshold set (usually = 0.05) indicates that there are no statistically significant differences between the different groups.

However, eve before a numerical assessment of the ANOVA table, it is necessary to generate diagnostic plots to verify the conditions as depicted in Figure 13.6.

```
#diagnostic plots
par(mfrow=c(2,2))
plot(m_anova1)
```

The diagnostic plots are showing problems with both assumptions: there are obvious deviations from normality (the Q-Q plot is not a straight line) and the residues of the different groups appear to have different ranges of variation (heteroskedastic). There are several leverage values.

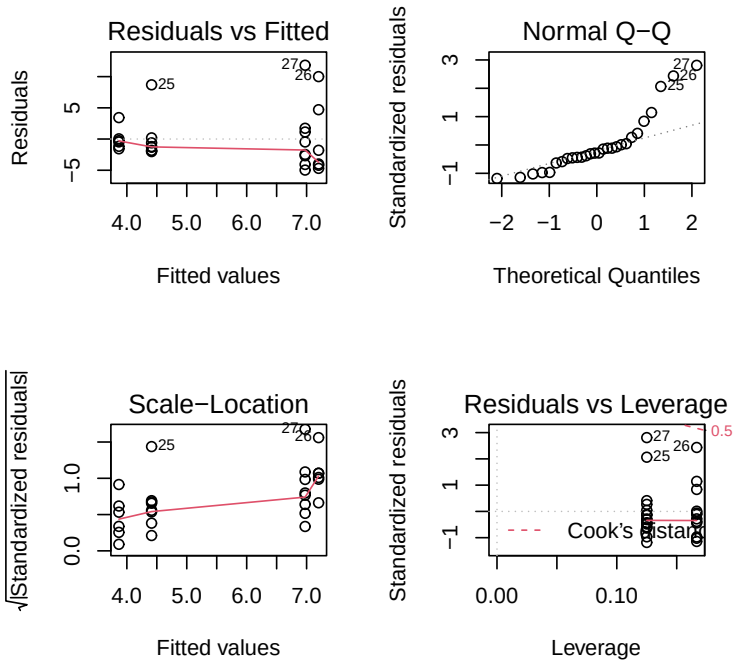

Fig. 13.2 Diagnostic plots for ANOVA

In particular,the Q–Q (quantile-quantile) plot is a probability plot, which is a graphical method for comparing two probability distributions by plotting their quan-

tiles against each other. First, the set of intervals for the quantiles is chosen. A point (x, y) on the plot corresponds to one of the quantiles of the second distribution (y-coordinate) plotted against the same quantile of the first distribution (x-coordinate). Thus the line is a parametric curve with the parameter which is the number of the interval for the quantile. If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x.

In this case, there are two solution to apply:

1. apply of the transformations to the numerical variable (logarithm, square root, reciprocal ,...) aimed at stabilizing it

2. or applying non-parametric methods which do not exploit the assumptions mentioned above

13.6 Transformations

A common transformations, as described in the Chapter on distributions, is a logarithmic transformation. In the example below the ANOVA is performed by using the logarithm of the variable.

```
m_anova2<- aov(formula = log_Ca_stock ~ VEG,data = Profiles)
summary(m_anova2)
plot(m_anova2)
```

After the log transformation, the plots are now showing that conditions are met, with a Q-Q plot showing the points approximately on the line y = x and the residuals having homogenous distribution. However, the ANOVA results still displayed no statistically significant differences with large p-value.

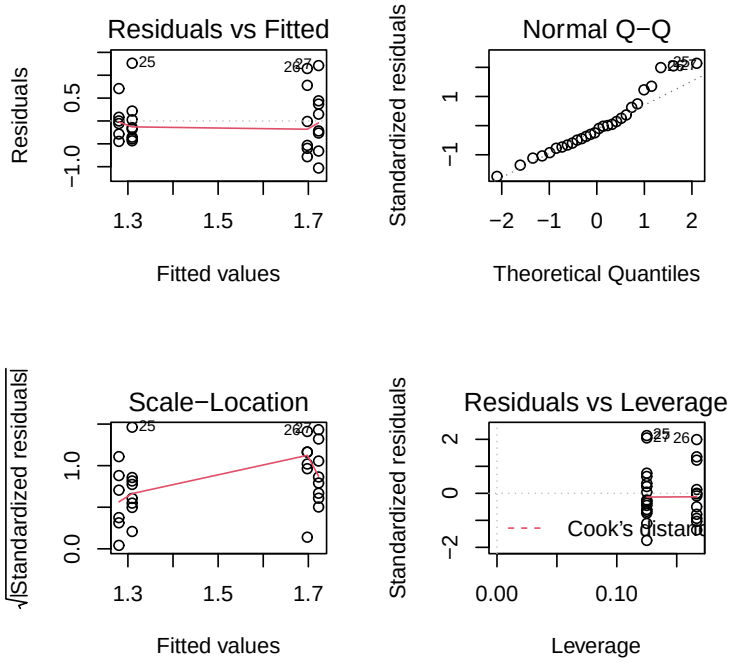

Fig. 13.3 Diagnostic plots for ANOVA for log transformed variable

```
            Df Sum Sq Mean Sq F value Pr(>F)
VEG          3  1.212  0.4039   1.013  0.404
Residuals   24  9.565  0.3986
```

13.7 Kruskal-Wallis

The Kruskal-Wallis test is a non-parametric method for testing whether samples originate from the same distribution. It is used for comparing two or more independent samples of equal or different sample sizes, when the conditions for applying parametric tests, like the ANOVA, are not met.

```
library(agricolae)
kw1<-kruskal(y = Profiles$Ca_stock, trt = Profiles$VEG)
print(kw1)
```

```
$statistics
Chisq Df   p.chisq
2.542648  3 0.4676329

$parameters
test p.ajusted       name.t ntr alpha
Kruskal-Wallis      none Profiles$VEG   4  0.05

$means
Profiles.Ca_stock     rank      std r Min  Max   Q25  Q50    Q75
BEECH          6.975000 17.50000 5.330974 8 2.0 18.8 3.950 5.50  8.250
CON            7.200000 16.16667 6.017641 6 2.5 17.2 3.050 4.30 10.275
GRASS          4.412500 11.87500 3.587055 8 2.4 13.1 2.575 3.15  4.000
MYR            3.866667 12.33333 1.782882 6 2.3  7.3 2.875 3.50  3.825

$comparison
NULL

$groups
Profiles$Ca_stock groups
BEECH          17.50000      a
CON            16.16667      a
MYR            12.33333      a
GRASS          11.87500      a

attr(,"class")
[1] "group"
```

Also in this case the p-value greater than indicates that there is no difference in ranks between the different groups. In the output there are also several statistics calculated in the different ones groups (*means* section), while the *groups* reports the results of multiple comparisons. Such tests, as we shall see, are useful in assessing which groups are showing differences statistically significant in case these are detected.

Now, the same analysis is performed on a different variable, the soil carbon stock.

```
#A different variable: C_Stock_0
m_anova_C<- aov(formula = C_stock_0 ~ VEG,data = Profiles)
summary(m_anova_C)
plot(m_anova_C)
```

A preliminary plot showed that the condition for applying ANOVA were met. The results of the statistical analysis on the Total carbon budget is now significant, showing that the soil carbon content depends on the vegetation coverage.

```
            Df Sum Sq Mean Sq F value Pr(>F)
VEG          3   7676  2558.7   3.979 0.0196 *
Residuals   24  15434   643.1
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

13.8 Tukey test

Another interesting test is the Tukey's HSD (honestly significant difference) test, it is a single-step multiple comparison procedure and statistical test. It can be used to find means that are significantly different from each other. Tukey's test compares the means of every treatment to the means of every other treatment; that is, it applies simultaneously to the set of all pairwise comparisons and identifies any difference between two means that is greater than the expected standard error.

It is performed on the results of the ANOVA to elucidate if there are differences among treatments for the dependent variable (in this case the soil carbon content)

```
tukeyTestVEG <- TukeyHSD(m_anova_C)
plot(tukeyTestVEG)
```

```
$VEG
                         diff        lwr      upr     p adj
CON-BEECH   -16.37500 -54.155094 21.40509 0.6353903
GRASS-BEECH -10.37500 -45.352571 24.60257 0.8452573
MYR-BEECH    29.45833  -8.321761 67.23843 0.1660183
GRASS-CON     6.00000 -31.780094 43.78009 0.9712585
MYR-CON      45.83333   5.444714 86.22195 0.0219996
MYR-GRASS    39.83333   2.053239 77.61343 0.0361426
```

13.9 Two way ANOVA

13.10 Tukey test