Stat 210
Applied Statistics and Data Analysis
Analysis of Variance
and Experimental Design

Joaquín Ortega
KAUST

Fall, 2020

Contents

1 Analysis of Variance

1.1 Introduction

In many experiments, we want to explore the effect of several **levels** of one or more categorical variables, known as **factors**. A factor may be the type of fertilizer for a crop, a particular treatment for a patient suffering from a given disease, the gender or race of a subject, the type of surface over which a race is run, or the diet of an animal in an experiment. Factors are frequently called **treatments**. The method we will use for the analysis of the effect that these factors have is known as Analysis of Variance or Anova for short. Anova was initially proposed by Sir Ronald Fisher nearly a century ago in the context of experimental analysis.

1.2 Analysis of Variance

The analysis of variance (Anova) is particularly important in statistically designed experiments, where, depending on the characteristics of the process being tested, a scheme is adopted, which determines how the experiment must be carried out and how the results should be analyzed.

The main idea of Anova is to compare several means, which are related to the levels of the different factors. This will allow us to disentangle the effects of several factors being varied together in a single experiment. However, to do this, we need to analyze the variances associated with the different means.

1.2.1 One-way Analysis of Variance

Let us start with a simple example to introduce the main ideas. This example is elaborated from *The R Book, 2nd. Edition*, M.J. Crawley, Wiley 2013.

Suppose we have only one factor and two levels, L_1 and L_2, and we have made five measurements with each level:

$$y_{11}, \dots, y_{15}, \quad y_{21}, \dots, y_{25}.$$

In general, if we have n_1 measurements for the first level and n_2 for the second, we would have

$$y_{11}, \dots, y_{1n_1}, \quad y_{21}, \dots, y_{2n_2}.$$

1.2.2 Notation

An index replaced by a $\bullet$ indicates that we sum the values corresponding to that index, leaving all other indices fixed:

$$y_{i\bullet} = \sum_{j=1}^{n_i} y_{ij}, \qquad y_{\bullet j} = \sum_{i=1}^{2} y_{ij}, \qquad y_{\bullet\bullet} = \sum_{i=1}^{2}\sum_{j=1}^{n_i} y_{ij}.$$

A bar over the variable indicates that we divide by the number of terms that are being added:

$$\bar{y}_{i\bullet} = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}, \qquad \bar{y}_{\bullet j} = \frac{1}{2}\sum_{i=1}^{2} y_{ij}, \qquad \bar{y}_{\bullet\bullet} = \frac{1}{n}\sum_{i=1}^{2}\sum_{j=1}^{n_i} y_{ij},$$

where $n = n_1 + n_2$.

This notation can be used with any number of levels.

Figure 1 shows the ten values obtained in the experiment. The horizontal line corresponds to the overall mean $\bar{y}_{\bullet\bullet}$ for the `Response`. The vertical segments are the differences between the observed values and the average, $y_{ij} - \bar{y}_{\bullet\bullet}$, $i = 1, 2, j = 1, \dots, 5$.

The **total sum of squares** SST is defined as

$$SST = \sum_{i=1}^{2}\sum_{j=1}^{5}(y_{ij} - \bar{y}_{\bullet\bullet})^2.$$

If we color the points in the graph according to the level of the factor, we get

We see that the blue points, which correspond to L_1, tend to be lower than the red points. What we would like to know is whether this is evidence of a real difference between the effect of the two levels or is simply due to chance. Anova will help us decide which case it is.

Next, let us consider the data for each level and fit a separate mean, $\bar{y}_{i\bullet}$ for the values corresponding to $L_i, i = 1, 2$. The graph is

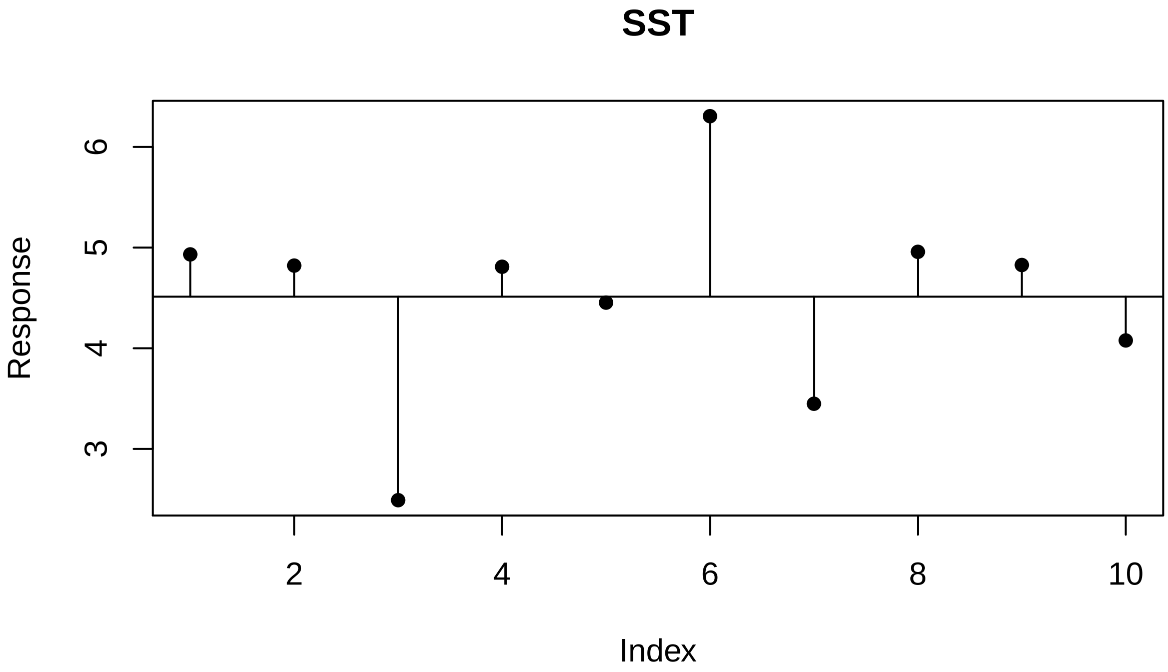


Figure 1: Total sum of squares.

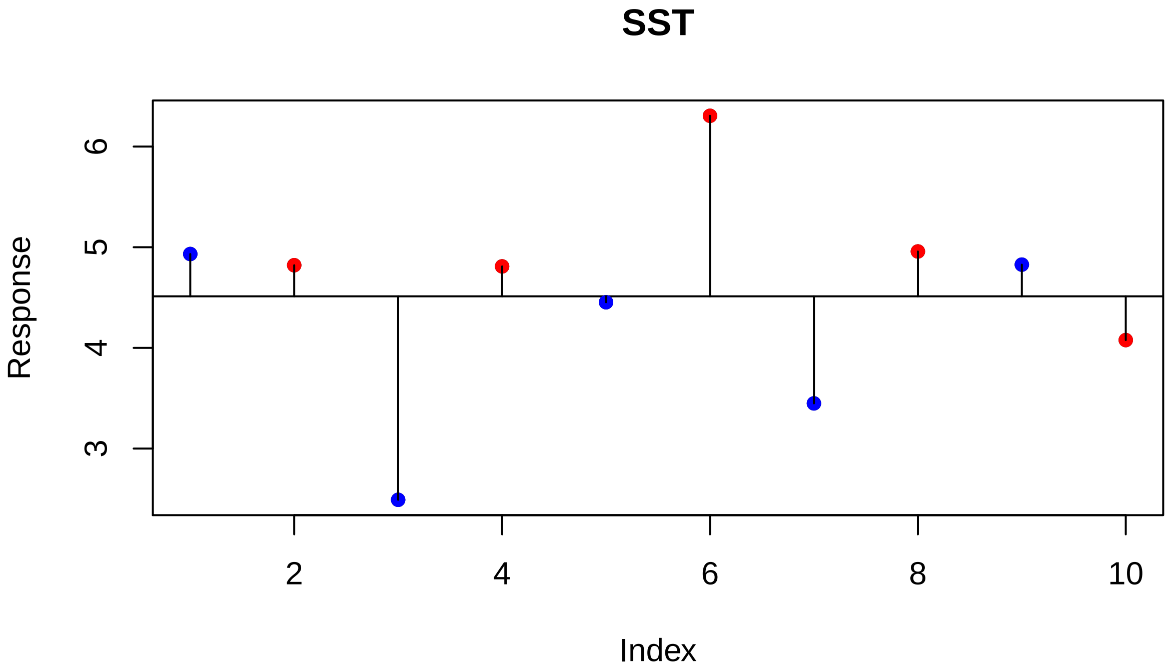


Figure 2: Total sum of squares.

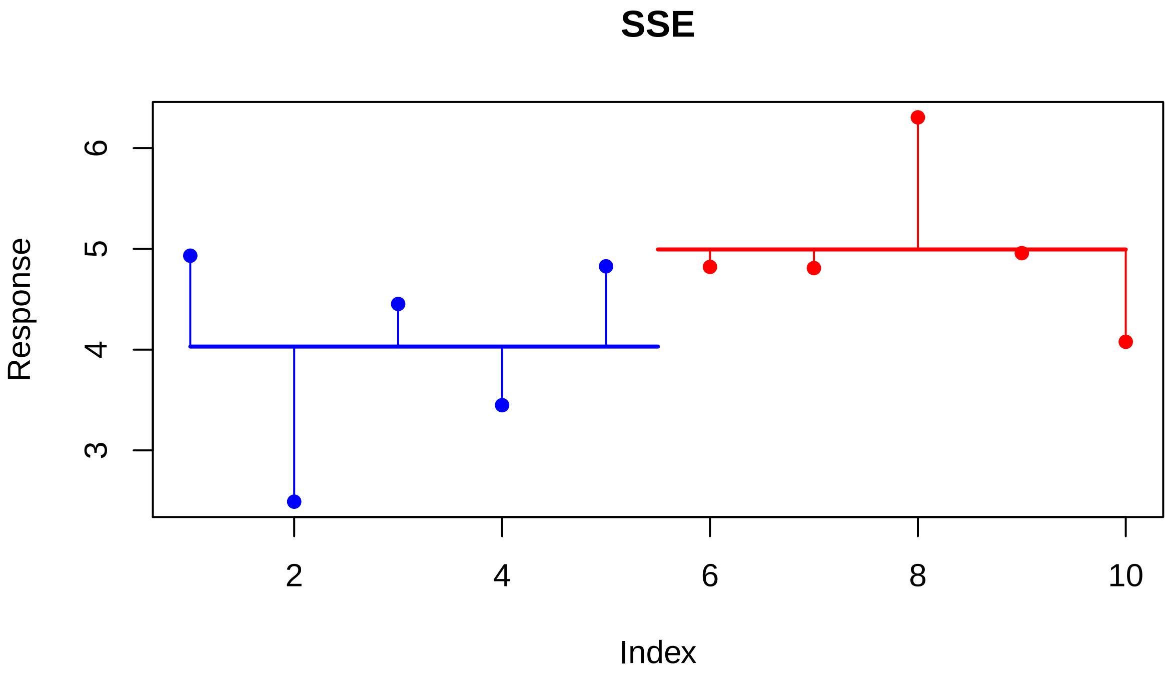


Figure 3: Error sum of squares.

The sum of the squares of the differences between the observed values and the corresponding treatment mean is known as the **error sum of squares**

$$SSE = \sum_{j=1}^{n_1} (y_{1j} - \bar{y}_{1\bullet})^2 + \sum_{j=1}^{n_2} (y_{2j} - \bar{y}_{2\bullet})^2.$$

SSE is the sum of the squares with respect to the red and blue lines in figure 2.

If the treatment we are considering has no effect on the outcome, we would expect the red and blue lines to be equal and also to coincide with the overall mean represented in the first figure by a horizontal black line. This is the same as saying that $SST = SSE$. If the means for the two treatment groups were the same, we would get a graph like the following figure

On the other hand, if the treatments affect on the response, we would expect the red and blue lines to be different and SSE to be less than SST. In fact, SSE could be zero if all the values were equal to their treatment means, as in figure 5 (r). We see that the difference in the sums of squares SST and SSE is linked to the differences in the treatment means.

To see that $SSE \le SST$ always, start with the definition of SST:

$$\begin{aligned} SST &= \sum_{i=1}^{2} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{\bullet\bullet})^2 \\ &= \sum_{j=1}^{n_1} (y_{1j} - \bar{y}_{\bullet\bullet})^2 + \sum_{j=1}^{n_2} (y_{2j} - \bar{y}_{\bullet\bullet})^2 \\ &= \sum_{j=1}^{n_1} (y_{1j} - \bar{y}_{1\bullet} + \bar{y}_{1\bullet} - \bar{y}_{\bullet\bullet})^2 + \sum_{j=1}^{n_2} (y_{2j} - \bar{y}_{2\bullet} + \bar{y}_{2\bullet} - \bar{y}_{\bullet\bullet})^2 \qquad (1) \end{aligned}$$

Let us look at the first sum; the second one is similar.

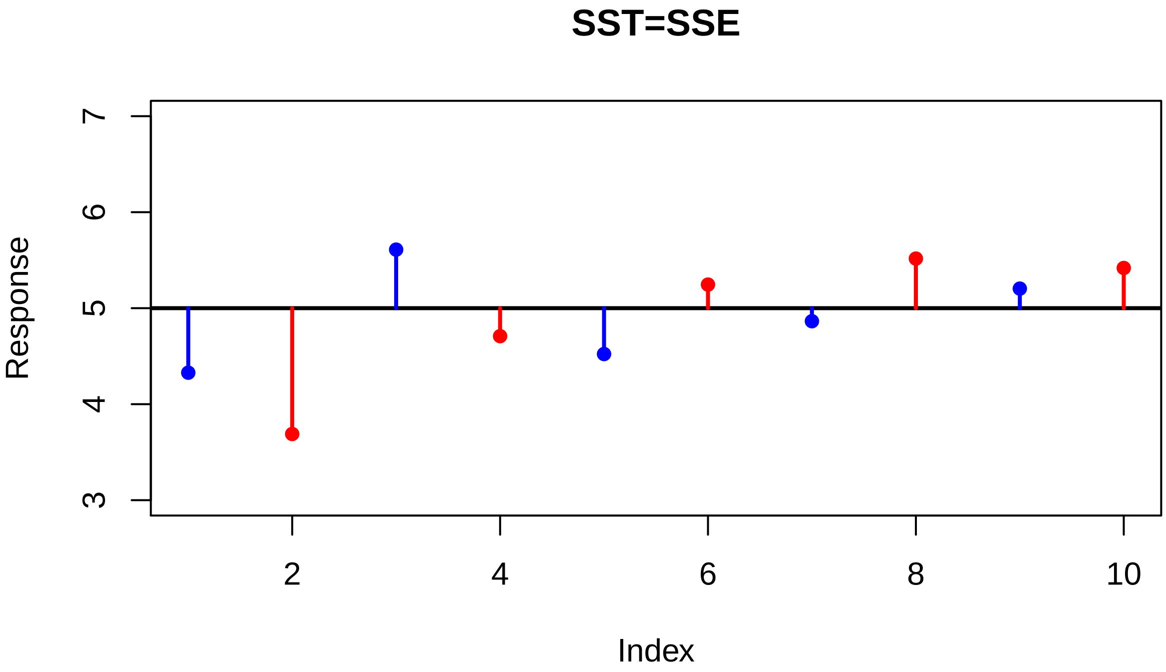


Figure 4: Total and error sum of squares.

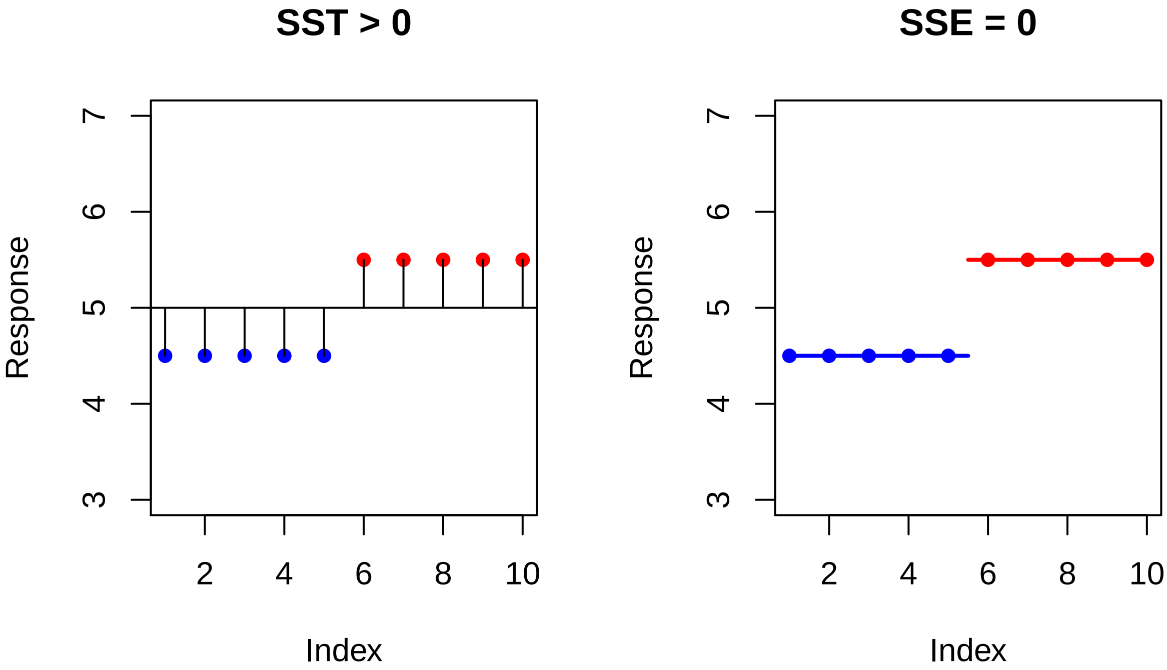


Figure 5: Total and error sum of squares.

$$\begin{aligned}\sum_{j=1}^{n_1}(y_{1j}-\bar{y}_{1\bullet}+\bar{y}_{1\bullet}-\bar{y}_{\bullet\bullet})^2 = \sum_{j=1}^{n_1}(y_{1j}-\bar{y}_{1\bullet})^2 \\ +\sum_{j=1}^{n_1}(\bar{y}_{1\bullet}-\bar{y}_{\bullet\bullet})^2 \\ +\sum_{j=1}^{n_1}2(y_{1j}-\bar{y}_{1\bullet})(\bar{y}_{1\bullet}-\bar{y}_{\bullet\bullet})\end{aligned}$$

The terms in the second sum above do not depend on the index of summation j, and the sum is equal to

$$n_1(\bar{y}_{1\bullet}-\bar{y}_{\bullet\bullet})^2.$$

For the third term in the sum we have

$$\sum_{j=1}^{n_1}2(y_{1j}-\bar{y}_{1\bullet})(\bar{y}_{1\bullet}-\bar{y}_{\bullet\bullet}) = 2(\bar{y}_{1\bullet}-\bar{y}_{\bullet\bullet})\Big(\sum_{j=1}^{n_1}y_{1j}-n_1\bar{y}_{1\bullet}\Big) = 0$$

because $\sum_{1j}^{n_1} y_i = n_1\bar{y}_{1\bullet}$. A similar argument is true for the second sum in (1) and we get

$$\begin{aligned}SST &= \sum_{j=1}^{n_1}(y_{1j}-\bar{y}_{1\bullet})^2+n_1(\bar{y}_{1\bullet}-\bar{y}_{\bullet\bullet})^2 \\ &\quad +\sum_{j=1}^{n_2}(y_{2j}-\bar{y}_{2\bullet})^2+n_2(\bar{y}_{2\bullet}-\bar{y}_{\bullet\bullet})^2 \\ &= \sum_{j=1}^{n_1}(y_{1j}-\bar{y}_{1\bullet})^2+\sum_{j=1}^{n_2}(y_{2j}-\bar{y}_{2\bullet})^2+\sum_{i=1}^{2}n_i(\bar{y}_{i\bullet}-\bar{y}_{\bullet\bullet})^2 \\ &= SSE+\sum_{i=1}^{2}n_i(\bar{y}_{i\bullet}-\bar{y}_{\bullet\bullet})^2. \qquad (2)\end{aligned}$$

Since the second sum in (2) is positive, we see that $SSE \leq SST$.

The difference between SST and SSE is known as the **treatment sum of squares** and will be denoted by SSA.

$$SSA = \sum_{i=1}^{2}n_i(\bar{y}_{i\bullet}-\bar{y}_{\bullet\bullet})^2.$$

When differences in mean between treatments are significant, SSA will be big with respect to SSE. If, on the contrary, the means are similar, then SSA will be small with respect to SSE. Figure 4 presents extreme cases of these situations. On the left, the treatment means are equal and $SST = SSE$, so $SSA = 0$, and the treatment has no effect on the response. On the right, all the variation in the response is explained by the difference in the treatment means, and $SSE = 0$, so $SST = SSA$.

Observe that SSE represents the variability *within* each group or level of the treatment factor, while SSA represents the variability *between* different groups or factor levels. This equation can be seen as a decomposition of the observed variability in the sample into the variability within each factor level and the difference between the factors.

As we will see later, to estimate variances, we divide sums of squares by the corresponding degrees of freedom (d.f.). In our example, we have two levels for the treatment, and we lose one degree of freedom estimating the overall mean, so there are $2 - 1 = 1$ degrees of freedom for the treatments. In general, if we had k treatment levels, we would have $k - 1$ degrees of freedom for treatments. If each factor level were replicated r times, then there would be $r - 1$ degrees of freedom for each level, since we lose one for each treatment mean. Considering that there are k levels, there are $k(r - 1)$ d.f. for error in the whole experiment. Finally, the total number of data points in the experiment is $n = rk$, and we lose one for the overall mean, so there are in total $rk - 1$ d.f. As a verification, it is always useful to check that the degrees of freedom for the components add up to the correct total:

$$k - 1 + k(r - 1) = k - 1 + kr - k = kr - 1.$$

The following expressions give the empirical variance for treatment, MSA, and for the errors MSE

$$MSE = \frac{SSE}{k(r-1)}; \qquad MSA = \frac{SSA}{k-1}.$$

The usual way to sum up these results is through an Analysis of Variance (Anova) table.

Table 1: Anova table for example 1.

Source	SS	d.f.	MS	F_{obs}	Critical F
Treatment	SSA	$k-1$	$MSA = \frac{SSA}{k-1}$	$F = \frac{MSA}{MSE}$	qf(1-α, k-1, k(r-1))
Error	SSE	$k(r-1)$	$MSE = \frac{SSE}{k(r-1)}$		
Total	SST	$kr-1$			

The significance of the difference between the means for the different treatment levels is assessed using an F test, which we will describe later on in detail. The null hypothesis is that the means are equal versus the alternative that at least two of them are different.

Another way of thinking about the comparison we made is in terms of the relative amounts of sampling variability between replicates that correspond to the same treatment level and between different levels. If the variation between replicates of a given level is substantial when compared to the variability between treatments, we will probably conclude that the difference between treatments is not significant. On the other hand, if the variability within treatment levels is small compared to the differences between treatments, we will likely reject the null hypothesis of equal treatment means.

1.3 A Worked-Out Example

This example is from M.J. Crawley *The R Book*, J. Wiley 2013.

We have an experiment in which crop yields per unit area were measured from 10 randomly selected fields on each of three soil types. All fields were sown with the same variety of seed and provided with the same fertilizer and pest control inputs. The question is whether soil type significantly affects crop yield, and if so, to what extent.

```r
results <- read.table('yields.txt',header=T)
attach(results)
str(results)
```

```
## 'data.frame':    10 obs. of  3 variables:
##  $ sand: int  6 10 8 6 14 17 9 11 7 11
##  $ clay: int  17 15 3 11 14 12 12 8 10 13
##  $ loam: int  13 16 9 12 15 16 17 13 18 14
```

```
head(results, n=4)
```

```
##   sand clay loam
## 1    6   17   13
## 2   10   15   16
## 3    8    3    9
## 4    6   11   12
```

The function `sapply` is used to calculate the mean yields for the three soils

```
sapply(list(sand,clay,loam),mean)
```

```
## [1]  9.9 11.5 14.3
```

The mean yield was highest on loam (14.3) and lowest on sand (9.9).

It will be useful to have all of the yield data in a single vector called `y`. To create a data frame from a spreadsheet like `results` where the values of the response are in multiple columns, we use the function called `stack` like this:

```
frame <- stack(results)
str(frame)
```

```
## 'data.frame':    30 obs. of  2 variables:
##  $ values: int  6 10 8 6 14 17 9 11 7 11 ...
##  $ ind   : Factor w/ 3 levels "sand","clay",..: 1 1 1 1 1 1 1 1 1 1 ...
```

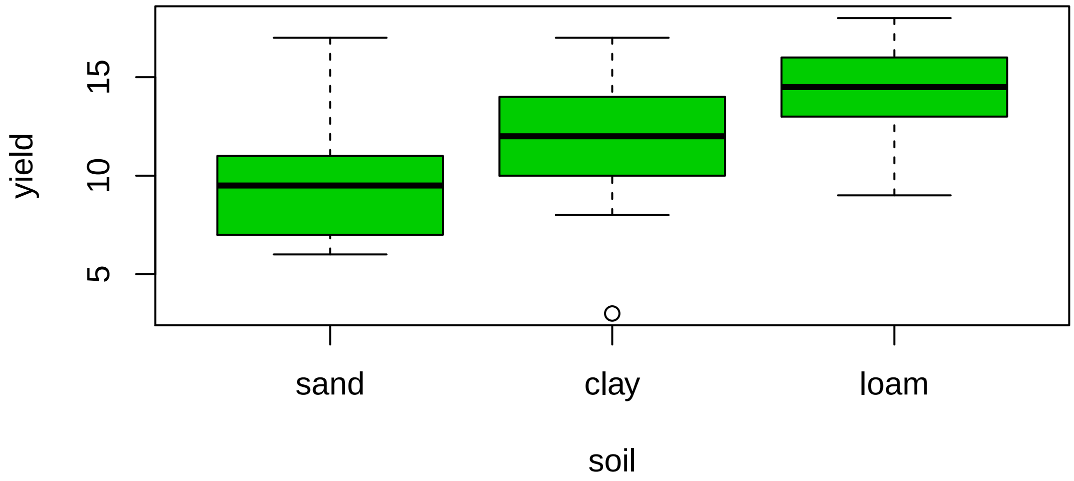
You can see that the stack function has invented names for the response variable (`values`) and the explanatory variable (`ind`). We change these:

```
names(frame) <- c('yield','soil')
attach(frame)
head(frame)
```

```
##   yield soil
## 1     6 sand
## 2    10 sand
## 3     8 sand
## 4     6 sand
## 5    14 sand
## 6    17 sand
```

Because the explanatory variable is categorical (three levels of soil type), initial data inspection involves a box-and-whisker plot of y against soil like this:

```
plot(yield~soil,col='green3')
```



Median yield is lowest on sand and highest on loam, but there is considerable variation from replicate to replicate within each soil type (there is even a low outlier on clay). It looks as if the yield on loam will turn out to be significantly higher than on sand (their boxes do not overlap), but it is not clear whether yield on clay is significantly greater than on sand or significantly lower than on loam. Analysis of variance may be used to answer these questions.

The analysis of variance involves calculating the total variation in the response variable (`yield` in this case) and partitioning it ('analyzing it') into informative components. In the simplest case, we partition the total variation into just two components, explained variation and unexplained variation. Explained variation is called the treatment sum of squares (SSA), and unexplained variation is called the error sum of squares (SSE, also known as the residual sum of squares), as defined earlier.

Let us work through the numbers in `R`. From the formula for SST, we can obtain the total sum of squares by finding the differences between the data and the overall mean:

```
sum((yield-mean(yield))^2)
```

```
## [1] 414.7
```

The unexplained variation, SSE, is calculated from the differences between the yields and the mean yields for that soil type:

```
sand-mean(sand)
```

```
##  [1] -3.9  0.1 -1.9 -3.9  4.1  7.1 -0.9  1.1 -2.9
## [10]  1.1
```

```
clay-mean(clay)
```

```
##  [1]  5.5  3.5 -8.5 -0.5  2.5  0.5  0.5 -3.5 -1.5
## [10]  1.5
```

```
loam-mean(loam)
```

```
##  [1] -1.3  1.7 -5.3 -2.3  0.7  1.7  2.7 -1.3  3.7
## [10] -0.3
```

We need the sums of the squares of these differences:

```
c(sum((sand-mean(sand))^2), sum((clay-mean(clay))^2),
sum((loam-mean(loam))^2))
```

```
## [1] 112.9 138.5  64.1
```

To get the sum of these totals across all soil types, we can use `sapply` like this:

```
sum(sapply(list(sand,clay,loam),
           function (x) sum((x-mean(x))^2) ))
```

```
## [1] 315.5
```

Thus SSE, the unexplained (or residual, or error) sum of squares, is 315.5.

The extent to which SSE is less than SST is a reflection of the magnitude of the differences between the means. The greater the difference between the mean yields on the different soil types, the greater will be the difference between SSE and SST. The treatment sum of squares, SSA, is the amount of the variation in yield that is explained by differences between the treatment means. In our example,

$$SSA = SST - SSE = 414.7 - 315.5 = 99.2$$

Now we can draw up the ANOVA table. There are six columns indicating, from left to right,

- the source of variation,
- the sum of squares attributable to that source,
- the degrees of freedom for that source,
- the variance for that source (traditionally called the mean square rather than the variance),
- the F ratio (testing the null hypothesis that this source of variation is not significantly different from zero) and
- the p-value associated with that F value.

We can fill in the sums of squares just calculated, then think about the degrees of freedom:

Table 2: Anova table for example 2.

Source	Sum of squares	Dof	Mean square	F-ratio	p-value
Soil type	99.2	2	$MSA = 49.6$	4.24	0.025
Error	315.5	27	$MSE = 11.685$		
Total	414.7	29			

There are 30 data points in all, so the total degrees of freedom are $30 - 1 = 29$. We lose 1 d.f. because in calculating SST we had to estimate one parameter from the data in advance, namely the overall mean, $\bar{y}_{\bullet\bullet}$, before we could calculate $SST = \sum(y_{ij} - \bar{y}_{\bullet\bullet})^2$.

Each soil type has $n = 10$ replications, so each soil type has $10 - 1 = 9$ d.f. for error, because we estimated one parameter from the data for each soil type, namely the treatment means $\bar{y}_{i\bullet}$ in calculating SSE. Overall, therefore, the error has $3 \times 9 = 27$ d.f. There were three soil types, so there are $3 - 1 = 2$ d.f. for soil type. The mean squares are obtained simply by dividing each sum of squares by its respective degrees of freedom (in the same row).

The error variance, $\hat{\sigma}^2$, is the residual mean square (the mean square for the unexplained variation); this is sometimes called the 'pooled error variance' because it is calculated across all the treatments. The alternative would be to have three separate variances, one for each treatment:

```
tapply(yield,soil,var)
```

```
##      sand      clay      loam 
## 12.544444 15.388889  7.122222
```

```
mean(tapply(yield,soil,var))
```

```
## [1] 11.68519
```

You will see that the pooled error variance $\hat{\sigma}^2 = MSE = 11.685$ is simply the mean of the three separate variances, because (in this case) there is equal replication in each soil type ($n = 10$). By tradition, we do not calculate the total mean square, so the bottom cell of the fourth column of the ANOVA table is empty.

The F ratio is the treatment variance divided by the error variance, testing the null hypothesis that the treatment means are not significantly different. If we reject this null hypothesis, we accept the alternative hypothesis that at least one of the means is significantly different from the others.

The question naturally arises at this point as to whether 4.24 is a big number or not. If it is a big number, then we reject the null hypothesis. If it is not a big number, then we accept the null hypothesis. We calculate the p value associated with our test statistic of 4.24 using the function `pf` for cumulative probabilities of the F distribution like this:

```
1-pf(4.24,2,27)
```

```
## [1] 0.02503987
```

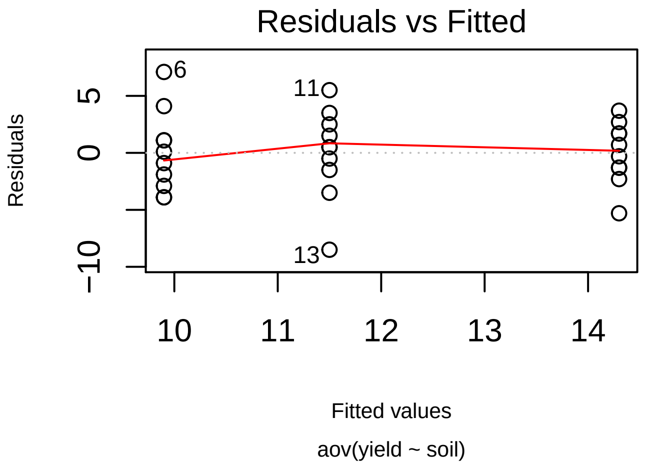
That was a lot of work. `R` can do the whole thing in a single line:

```
summary(aov(yield~soil))
```

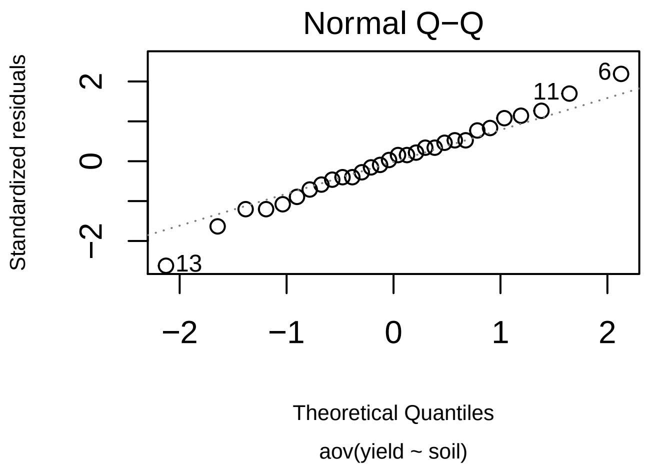
```
##             Df Sum Sq Mean Sq F value Pr(>F)
## soil         2   99.2   49.60   4.245  0.025 *
## Residuals   27  315.5   11.69
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The next thing we would do is to check the assumptions of the `aov` model. This is done using `plot`.

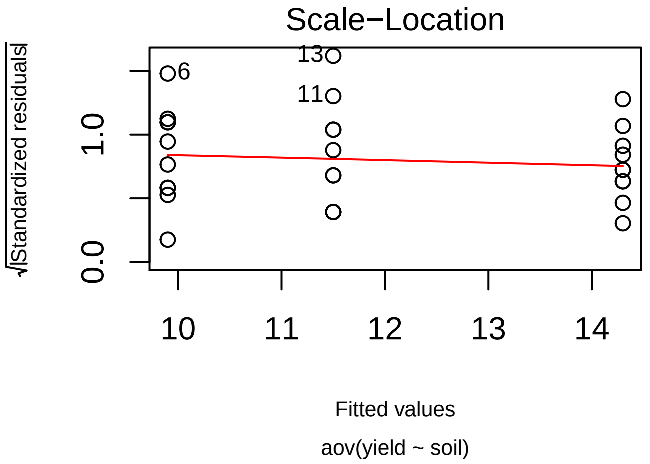
```
plot(aov(yield~soil), which = 1, cex.lab=0.7, cex.sub=0.7)
```



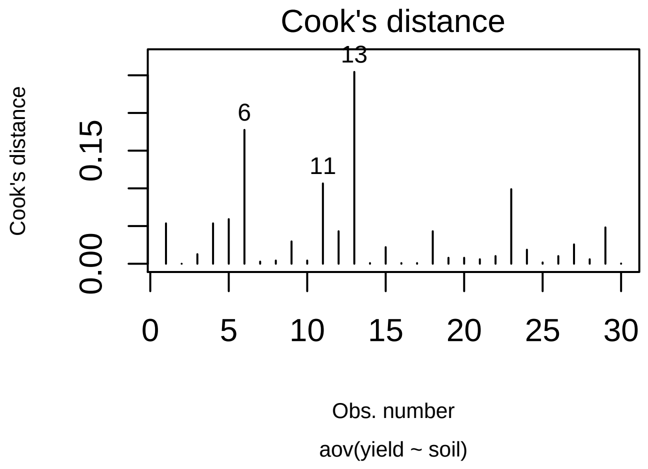
```
plot(aov(yield~soil), which = 2, cex.lab=0.7, cex.sub=0.7)
```



```
plot(aov(yield~soil), which = 3, cex.lab=0.7, cex.sub=0.7)
```



```
plot(aov(yield~soil), which = 4, cex.lab=0.7, cex.sub=0.7)
```



2 Design of Experiments

This section is based partly on Chapter 11, Ugarte, Militino and Arnholt, *Probability and Statistics with R*, Chapman & Hall, 2008. and Chapter 2, J. Lawson, *Design and Analysis of Experiments with R*, Chapman and Hall, 2015.

2.1 Design of Experiments

Experiments are usually carried out to determine the effect of a factor or factors in some response of interest. We will assume that the response is a continuous variable. The factors are independent variables whose levels are set by the experimenter. They may be categorical variables or continuous variables that have been set to a fixed number of discrete levels. These different configurations of the factors are known as treatments, and they are applied to experimental units. The response variable is measured for each treatment, and the objective is to compare the observed responses.

When two or more factors are involved, the experiment is known as a factorial design. For example, a researcher may be interested in the effect of adding a certain amount of carbon on steel hardness, and the experiment is designed so that three different amounts are considered. In this case, the treatments are the three levels of carbon that the experimenter is interested in comparing, carbon is the factor being considered, and strength is the response. In the event a second factor is of interest, such as temperature with two levels, the experiment will consist of $2 \times 3 = 6$ treatment combinations and is known as a factorial design.

Suppose first that the experimenter has one furnace to carry out the experiment. There may be variations in the performance of the furnace at different times. To minimize their effect, the researcher should randomly assign the order in which the experiments are run. By randomizing the assignment of treatments, the possibility of confounding differences due to levels of carbon with differences due to the performance of the furnace is minimized. When the assignment of treatments is done in a completely random way, we speak of a *completely randomized design* (CRD).

Suppose now that the experimenter has two furnaces to run the experiment. Even though the furnaces may have the same characteristics, there may be operational differences that may affect the result of the experiment. In this case, experiments run on the same furnace may be more homogeneous than experiments run on different furnaces. Experiments run on the same furnace are grouped into **blocks**. In this case, treatments are randomly assigned to experimental units within a block, i.e., the order in which experiments are performed is assigned randomly *for each furnace*. Such a design is known as a randomized complete block design (RCBD).

2.1.1 Definitions

- **Treatments** are levels of a factor or combinations of factor levels the experimenter wants to compare.
- **Experimental units** are the objects or persons to which treatments are applied, for example, animals, plots, plants, or people.
- **Responses** are outcomes observed after the application of a treatment to an experimental unit.
- **Experimental error** is random variation present in the experiment, not under the control of the experimenter. Experimental error may be due to many things, including but not limited to: measurement error, different responses from measuring the same quantity in separate trials, and different responses from experimental units given the same treatment.
- **Treatment structure** specifies the set of factors the experimenter has selected to study or compare.
- **Design structure** defines how experimental units are assigned to treatment groups.
- **Randomization** is the use of some well-defined probabilistic mechanism to assign treatments to experimental units. Randomization reduces the possibility of bias and confounding. Randomization should also be used, if possible, with any variable not under the direct control of the experimenter that may influence the measured response.
- **Replication** is the independent assignment of several experimental units to each treatment (factor combination), resulting in independent observations. Replication shows the results are reproducible and allows the experimenter to estimate the experimental error. When the number of experimental

units is the same for all treatments, the design is referred to as a balanced design. Unbalanced designs do not have an equal number of experimental units for all treatments.

When the experimenter fixes all treatment levels, we speak of a **fixed-effects** model or experiment. If treatment levels are random, we talk of **random effects**. Experiments which include both kinds are **mixed-effects** experiments.

The fixed-effects model assumes:

1. The measured responses are independent of one another.
2. The model errors are independent of one another and follow a normal distribution.
3. The variance is homogeneous across treatments.

The experimenter seeks a model that explains how the response varies in terms of the independent variables in the experiment, the predictors. There may be more than one model that adequately describes this relation. In this case, we are guided by the principle of parsimony (Occam's razor), and we should choose the simplest model that reasonably describes the experimental results.

Models are expressed in `R` with the syntax

```
response ~ predictors
```

Finding an adequate model is an iterative process that starts by

1. Identifying an appropriate model based on the treatment and design structure of the experiment.
2. Validating the model's assumptions using diagnostic plots.
3. Selecting a different model or transforming the response variable when the model's assumptions are not satisfied until a plausible model is found.

Once a model has been validated, formal inference to test for no treatment effects (equality of treatment means) and estimation of the model's parameters can be undertaken. If the analysis shows that not all treatment means are equal, multiple comparisons may be used to determine which treatment produces the 'best' result.

2.2 A General Model for One-way Anova.

Suppose we have an experiment where one factor or treatment T has k levels and there are r_i replications for configuration $i, i = 1, \ldots, k$. By this, we mean that for each configuration, the experiment is repeated r_i times, and the results obtained are the replications. It does not mean that the experiment is performed once, and r_i measurements are taken.

Let y denote the response variable, then a model for this experiment would be

$$y_{ij} = \mu + \tau_i + \epsilon_{ij}; \qquad j = 1, \ldots, r_i, i = 1, \ldots k, \tag{3}$$

where

- y_{ij} represents the outcome of the j-th replication of level i,
- μ is the overall mean,
- τ_i is the **effect** of treatment i and
- ϵ_{ij} represents the error for the j-th replication of level i.

The errors are assumed to be independent and normally distributed with mean 0 and equal variance σ^2. The total number of data points is $n = kr$.

In this model, μ is the global mean for the experiment, and $\mu + \tau_i$ represents the average response for the i-th group. Equation (3) is usually known as the *effects model.*

An alternative model for this experiment would be

$$y_{ij} = \mu_i + \epsilon_{ij}; \qquad j = 1, \ldots, r_i, i = 1, \ldots k, \tag{4}$$

where μ_i represents the average response for level i of the treatment factor and the same assumptions are made for the errors ϵ_{ij}.

Comparing μ_s with μ_t is equivalent to comparing τ_s with τ_t:

$$\mu_t - \mu_s = (\mu + \tau_t) - (\mu + \tau_s) = \tau_t - \tau_s$$

Model (4) is sometimes known as the *cell means model.* The two models are equivalent. These models can be used in two different scenarios.

When the experimenter specifically chooses the treatments, and there is no desire to extend the results to other treatments, the model is referred to as a **fixed effects model**. When the treatments are selected at random from a larger population of possible treatments and the experimenter would like to extend the conclusions of the experiment of all treatments in the population, the model is called a **random effects model**. We will only consider the fixed effects model.

2.3 Least Squares Estimation

Least square estimators for the one-way Anova model are values for the parameters $\hat{\mu}, \hat{\tau}_1, \ldots, \hat{\tau}_k$ that minimize the error sum of squares

$$\sum_{i=1}^{k} \sum_{j=1}^{r_i} \epsilon_{ij}^2 = \sum_{i=1}^{k} \sum_{j=1}^{r_i} (y_{ij} - \mu - \tau_i)^2. \tag{5}$$

The resulting model $y_{ij} = \hat{\mu} + \hat{\tau}_i$ is the best-fitting model in the sense of minimizing (5).

The procedure for minimizing this expression is the usual. The expression in (5) is differentiated with respect to the parameters $\mu, \tau_1, \ldots, \tau_k$ in turn and each of the resulting expressions is set equal to zero, yielding a set of $k + 1$ equations. These are known as the *normal equations.* It is an exercise in calculus to verify that these equations are

$$y_{\bullet\bullet} - n\hat{\mu} - \sum_{i=1}^{k} r_i \hat{\tau}_i = 0, \tag{6}$$

$$y_{i\bullet} - r_i(\hat{\mu} + \hat{\tau}_i) = 0, \quad i = 1, \ldots, k, \tag{7}$$

where the hat notation indicates that these are the values that minimize (5).

From (7) we get that

$$\hat{\mu} + \hat{\tau}_i = \frac{1}{r_i} y_{i\bullet} = \bar{y}_{i\bullet}$$

for $i = 1, \ldots, k$, so the least squares estimate for the i-th treatment mean is the corresponding sample mean $\bar{y}_{i\bullet}$.

However, there is a problem with the normal equations. If we add up the equations in (7) we get (6). The k equations in (7) are linearly independent, but if we add (6) we get an undetermined system of equations that does not have a unique solution. This means that the $k + 1$ parameters in model (3) are not all estimable.

2.4 Least Squares in Matrix Notation

Consider model (3) with $k = 3$ factor levels and $r = 2$ replicates for each level. We can write the effects model using matrix notation as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{8}$$

where

$$\mathbf{y} = \begin{pmatrix} y_{11} \\ y_{12} \\ y_{21} \\ y_{22} \\ y_{31} \\ y_{32} \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \mu \\ \tau_1 \\ \tau_2 \\ \tau_3 \end{pmatrix}, \quad \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_{11} \\ \epsilon_{12} \\ \epsilon_{21} \\ \epsilon_{22} \\ \epsilon_{31} \\ \epsilon_{32} \end{pmatrix}.$$

The least squares estimators are the solution to the normal equations $\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{y}$. The problem is that the matrix $\mathbf{X}'\mathbf{X}$ is singular and cannot be inverted.

The `R` function `lm` makes the matrix $\mathbf{X}$ full rank by dropping the column that corresponds to the first level of the factor:

$$\mathbf{X} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{pmatrix}.$$

This coding makes the first level of the treatment the standard, and all other levels are compared to it. For example, with $k = 3$ levels the solution to the normal equations is

$$(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \hat{\boldsymbol{\beta}} = \begin{pmatrix} \hat{\mu} + \hat{\tau}_1 \\ \hat{\tau}_2 - \hat{\tau}_1 \\ \hat{\tau}_3 - \hat{\tau}_1 \end{pmatrix}. \tag{9}$$

This is equivalent to adding the equation $\hat{\tau}_1 = 0$ to the equations in (7).

There are other alternatives. For practical purposes, any one of the infinite number of solutions will be satisfactory, since they lead to identical solutions for the estimable parameters. In fact, any extra equation can be added, provided that it is not a linear combination of the equations already present. The trick is to add whichever equation will aid most in solving the entire set of equations.

A common solution is obtained by adding the extra equation $\sum_i r_i\hat{\tau}_i = 0$. In this case the normal equations become

$$\begin{aligned} \sum_i r_i\hat{\tau}_i &= 0 \\ y_{\bullet\bullet} - n\hat{\mu} &= 0 \\ y_{i\bullet} - r_i(\hat{\mu} + \hat{\tau}_i) &= 0, \quad i = 1, \ldots, k \end{aligned}$$

from which we get the least squares solutions

$$\hat{\mu} = \bar{y}_{\bullet\bullet}, \qquad \hat{\tau}_i = \bar{y}_{i\bullet} - y_{\bullet\bullet}.$$

Parameter	Estimator
μ	$\overline{Y}_{\bullet\bullet}$
μ_i	$\overline{Y}_{i\bullet}$
τ_i	$\overline{Y}_{i\bullet} - \overline{Y}_{\bullet\bullet}$
ϵ_{ij}	$Y_{ij} - \overline{Y}_{i\bullet}$
σ^2	$\frac{1}{N-a}\sum_{i=1}^{a}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{i\bullet})^2$

2.5 Example

A tire manufacturer is interested in investigating the braking performance for different types of tread patterns. There are four different tread patterns identified with the letters A, B, C, and D. Six measurements were taken with each one.

Measurements (StopDist) correspond to the braking distance in feet of a medium-sized car from a speed of 60 miles per hour. The same driver and car were used for all the experiments. The order of the treatments was assigned at random. The data can be found in the Tire file in the PASWR package.

```
library(PASWR)
str(Tire)
```

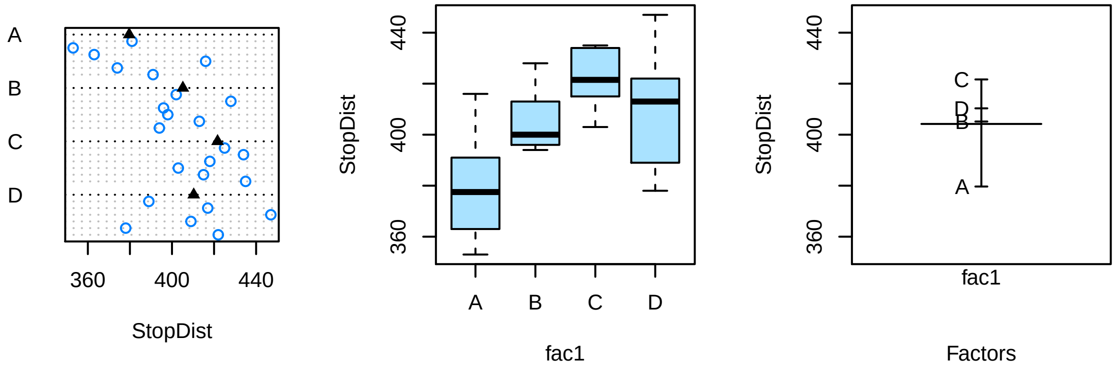
```
## 'data.frame':    24 obs. of  2 variables:
##  $ StopDist: int  391 374 416 363 353 381 394 413 398 396 ...
##  $ tire    : Factor w/ 4 levels "A","B","C","D": 1 1 1 1 1 1 2 2 2 2 ...
```

```
head(Tire, n=4)
```

```
##   StopDist tire
## 1      391    A
## 2      374    A
## 3      416    A
## 4      363    A
```

For an initial graphical exploration of the results we use the function oneway.plots() from the PASWR package.

```
with(Tire, oneway.plots(StopDist, tire))
```



We now use the lm function to fit the linear model. The results are stored in the file mod0.

```
mod0 <- lm(StopDist ~ tire, data = Tire)
summary(mod0)
```

```
##
## Call:
## lm(formula = StopDist ~ tire, data = Tire)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -32.333  -9.667  -2.250  11.417  36.667
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  379.667      7.691  49.363  < 2e-16
## tireB         25.500     10.877   2.344 0.029497
## tireC         42.000     10.877   3.861 0.000973
## tireD         30.667     10.877   2.819 0.010594
##
## (Intercept) ***
## tireB       *
## tireC       ***
## tireD       *
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.84 on 20 degrees of freedom
## Multiple R-squared:  0.4442, Adjusted R-squared:  0.3608
## F-statistic: 5.328 on 3 and 20 DF,  p-value: 0.007316
```

To interpret these results recall from (9) that the default coding for treatment in `R` means that the first value (`Intercept`) corresponds to the average value for the first treatment level, $\hat{\mu}+\hat{\tau}_1$ while `tireB`$=\hat{\tau}_2-\hat{\tau}_1$, `tireC`$=\hat{\tau}_3-\hat{\tau}_1$, and `tireD`$=\hat{\tau}_4-\hat{\tau}_1$. Thus

$$\begin{aligned}
\hat{\mu}+\hat{\tau}_1 &= 379.7;\\
\hat{\mu}+\hat{\tau}_2 &= 379.7+25.5=405.2\\
\hat{\mu}+\hat{\tau}_3 &= 379.7+42.0=421.7\\
\hat{\mu}+\hat{\tau}_4 &= 379.7+30.7=410.4\\
&.
\end{aligned}$$

The **residuals** $\hat{\epsilon}_{ij}, j=1\ldots,r_i, i=1,\ldots,k$ are defined as

$$\hat{\epsilon}_{ij}=y_{ij}-(\hat{\mu}+\hat{\tau}_i)$$

and represent the difference between the j-th replication of the i-th treatment and the estimated treatment mean $\hat{\mu}+\hat{\tau}_i=\bar{y}_{i\bullet}$.

The sum of squares for error or error sum of squares is

$$\begin{aligned}
SSE &= \sum_i\sum_j \hat{\epsilon}_{ij}^2=\sum_i\sum_j(y_{ij}-(\hat{\mu}+\hat{\tau}_i))^2\\
&= \sum_i\sum_j(y_{ij}-\bar{y}_{i\bullet})^2\\
&= \sum_i\sum_j y_{ij}^2-\sum_i r_i\bar{y}_{i\bullet}^2. \qquad (10)
\end{aligned}$$

2.6 Variance Estimation

The sample variance for the i-th treatment is given by

$$\hat{\sigma}_i^2=\frac{1}{r_i-1}\sum_{j=1}^{r_i}(y_{ij}-\bar{y}_{i\bullet})^2 \qquad (11)$$

which is an unbiased estimator for the common variance of the errors: $E(\hat{\sigma}_i^2)=\sigma^2$. From (11) we get that

$$\sum_{j=1}^{r_i}(y_{ij}-\bar{y}_{i\bullet})^2=(r_i-1)\hat{\sigma}_i^2.$$

Using this in (10) and taking expectations we get

$$E(SSE) = \sum_{i=1}^{k}(r_i - 1)E(\hat{\sigma}_i^2) = \sigma^2 \sum_{i=1}^{k}(r_i - 1) = \sigma^2(n-k)$$

and

$$\hat{\sigma}^2 = \frac{SSE}{n-k} = MSE \tag{12}$$

is an unbiased estimator for the variance σ^2.

2.7 Hypothesis Test of No Treatment Effect

In an experiment involving k treatment levels, a hypothesis of interest is whether the treatments are different in terms of their effect on the response variable. Thus, the null hypothesis would be

$$H_0 : \tau_1 = \tau_2 = \cdots = \tau_k \quad vs. \quad H_1 : \text{ at least two of the } \tau_i \text{ differ.}$$

Even though it looks like the null hypothesis involves non-estimable parameters, it can easily be recast in terms of estimable contrasts:

$$H_0 : \tau_2 - \tau_1 = 0, \text{ and } \tau_3 - \tau_1 = 0 \text{ and } \ldots \text{ and } \tau_k - \tau_1 = 0.$$

There are other ways of rewriting H_0 in terms of estimable contrasts, but they will always depend on $k-1$ distinct contrasts, which is the number of degrees of freedom for treatment.

As we saw before, the basic idea of the analysis of variance is that the error sum of squares measures how well the model fits the data. One way of testing whether the treatments have an effect is to compare the sums of squares for the complete model with that obtained with a model that assumes that the null hypothesis is true. This last model is known as the *reduced model* and is

$$y_{ij} = \mu + \epsilon_{ij}$$

with the same hypothesis as before for the noise.

The least-squares estimator for μ in this model is the overall mean:

$$\hat{\mu} = \bar{y}_{\bullet\bullet}$$

and the error sum of squares for the reduced model is

$$\begin{aligned} SST &= \sum_i \sum_j (y_{ij} - \bar{y}_{\bullet\bullet})^2 = \sum_i \sum_j (y_{ij}^2 - 2y_{ij}\bar{y}_{\bullet\bullet} + \bar{y}_{\bullet\bullet}^2) \\ &= \sum_i \sum_j y_{ij}^2 - n\bar{y}_{\bullet\bullet}^2. \end{aligned} \tag{13}$$

If the null hypothesis is false and treatment effects differ, SSE should be smaller than SST.

The test is based on the difference $SST - SSE$, relative to the size of the SSE, i.e., $(SST - SSE)/SSE$ and we reject H_0 if this quantity is large. We used before the notation SSA for $SST - SSE$, and called it the treatment sum of squares. Using (10) and (13), it is given by

$$SSA = \sum_i \sum_j y_{ij}^2 - n\bar{y}_{\bullet\bullet}^2 - \Big(\sum_i \sum_j y_{ij}^2 - \sum_i r_i \bar{y}_{i\bullet}^2\Big) = \sum_i r_i \bar{y}_{i\bullet}^2 - n\bar{y}_{\bullet\bullet}^2.$$

Equivalently,

$$SSA = \sum_i r_i (\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet})^2. \tag{14}$$

As we mentioned before, this sum represents the difference between treatment means and corresponds to the comparison between groups.

It can be shown that

- SSE/σ^2 has a χ^2_{n-k} distribution
- SSA/σ^2 has a χ^2_{k-1} distribution under H_0,
- these variables are independent.

In consequence, under H_0 the quotient

$$\frac{SSA/\sigma^2(k-1)}{SSE/\sigma^2(n-k)} \sim F_{k-1,n-k} \tag{15}$$

and we can use this relation to test H_0.

Recall that we defined $MSE = SSE/(n-k)$, and define $MSA = SSA/(k-1)$, then (15) becomes

$$\frac{MSA}{MSE} \sim F_{k-1,n-k}$$

and if msE and msA represent the observed values of these variables, the decision rule for testing H_0 at level of significance α is

$$\text{reject } H_0 \text{ if } \quad \frac{msA}{msE} > F_{1-\alpha,k-1,n-k} \tag{16}$$

As we have seen, the values for the sums of squares, degrees of freedom, mean squares, and F test are usually written in an Analysis of Variance table, such as Table 3 below.

Table 3: Anova table for the one-way analysis of variance

Source	SS	d.f.	MS	F_{obs}	Critical F
Treatment	SSA	$k-1$	$MSA = \frac{SSA}{k-1}$	$F = \frac{MSA}{MSE}$	`qf(1-`α`, k-1, n-k)`
Error	SSE	$n-k$	$MSE = \frac{SSE}{n-k}$		
Total	SST	$n-1$			
			Computational Formulae		
$SSA = \sum_i r_i \bar{y}_{i\bullet}^2 - n\bar{y}_{\bullet\bullet}^2$				$SSE = \sum_i \sum_j y_{ij}^2 - \sum_i r_i \bar{y}_{i\bullet}^2$	
$SST = \sum_i \sum_j y_{ij}^2 - n\bar{y}_{\bullet\bullet}^2$					

An Anova table can be obtained in `R` using the `aov` function. The inputs for this function are the same as those for the `lm` function we used previously, but the `summary` of an object created with the `aov` function is the Anova table. For the tire tread example, the code is

```r
mod1 <- aov(StopDist ~ tire, data = Tire)
summary(mod1)
```

```
##             Df Sum Sq Mean Sq F value  Pr(>F)   
## tire         3   5673  1891.0   5.328 0.00732 **
## Residuals   20   7099   354.9                   
## ---
## Signif. codes:  
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The results for treatment are in the line labeled `tire` while the results corresponding to the errors are in the line marked `Residuals`. Results for *total* can be obtained adding up the corresponding terms in the table. The F value is the ratio MSA/MSE, and the last column labeled `Pr(>F)` is the probability of exceeding the calculated F-value when the null hypothesis is true, i.e., it is the p-value for the F test.

In this example, we would conclude that there are significant differences among the average braking distances for different treads at the 0.05 level and also at the 0.01 level.

The same table is produced using the function `anova()` on the model we obtained with the `lm` function (`mod0`).

```
anova(mod0)
```

```
## Analysis of Variance Table
##
## Response: StopDist
##           Df Sum Sq Mean Sq F value   Pr(>F)
## tire       3 5673.1 1891.04  5.3278 0.007316 **
## Residuals 20 7098.8  354.94
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

To see the complete list of quantities calculated by the `aov` function type `names(mod1)`:

```
names(mod1)
```

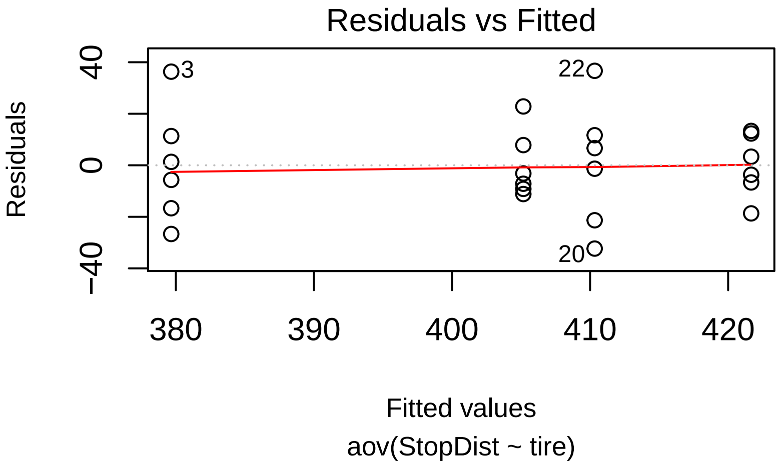
```
##  [1] "coefficients"  "residuals"
##  [3] "effects"       "rank"
##  [5] "fitted.values" "assign"
##  [7] "qr"            "df.residual"
##  [9] "contrasts"     "xlevels"
## [11] "call"          "terms"
## [13] "model"
```

```
names(anova(mod0))
```

```
## [1] "Df"      "Sum Sq"  "Mean Sq" "F value"
## [5] "Pr(>F)"
```

Look at diagnostic plots to check the assumptions of the model.

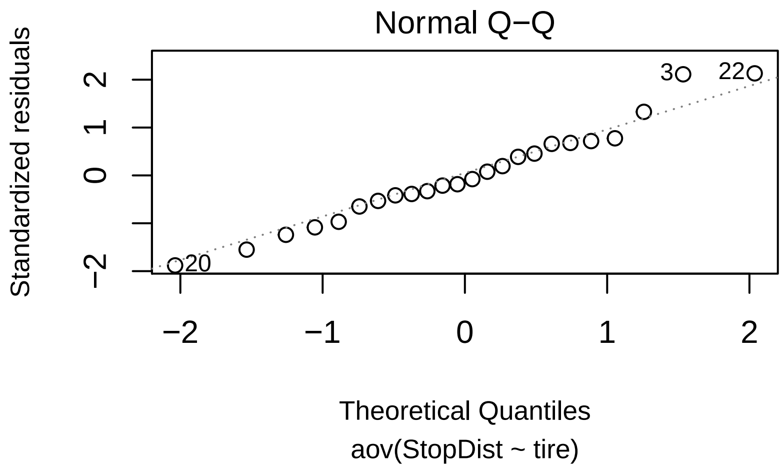
```
plot(mod1, which=1, cex.lab=0.8, cex.sub=0.8)
```



Since we have six replications for each treatment level, and there are only four x-values, the points appear vertically aligned at these values. We look in these graphs for constant variance. We see that, in some cases, values appear to be more spread than in others, and this may be a sign of non-constant variance. However, we only have a few points, and this is difficult to determine.

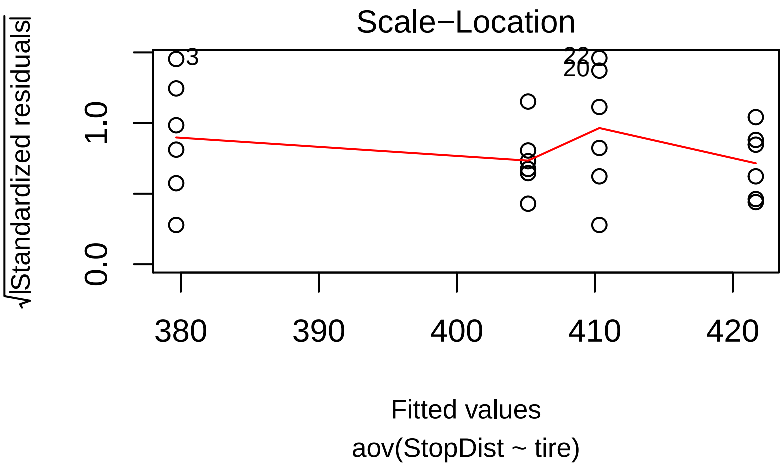
We look for departures from normality in the standardized residuals. Considering the fit we observe, the normality assumption seems justified.

```
plot(mod1, which=2, cex.lab=0.8, cex.sub=0.8)
```

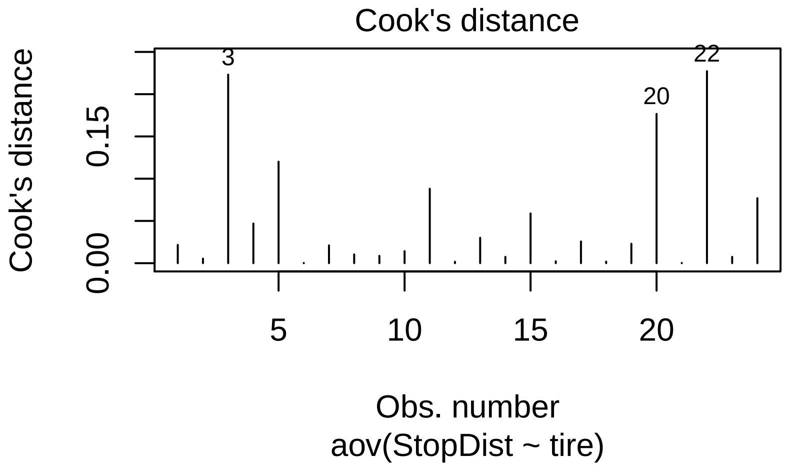


The third graph plots the square root of standardized residuals versus fitted values, and again we look for changes in the variance. As in the first graph, we see that some points seem to have more spread than others, perhaps pointing to heteroscedasticity. However, there are very few points to be certain.

```
plot(mod1, which=3, cex.lab=0.8, cex.sub=0.8)
```

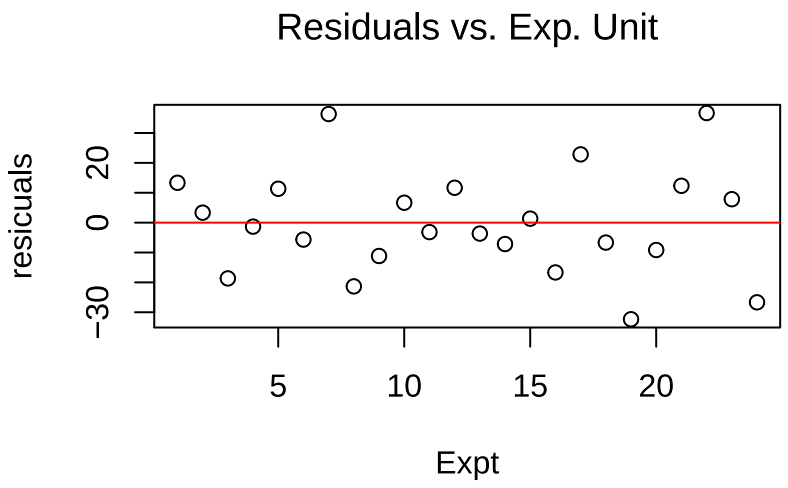


```
plot(mod1, which=4)
```



Another plot that is frequently useful is the plot of residuals versus the order in which the experiments have been performed. This may help to detect if there is a tendency in the results that should be taken into account. In this case, that information is not available in the data set. To illustrate the procedure, we will do the graph using a random ordering of the experiments. To get the residuals for the model use the command `residuals(mode1)`.

```
plot(residuals(mod1) ~ sample(1:24,24), xlab='Expt',
     ylab='resicuals', main='Residuals vs. Exp. Unit',
     font.main=1)
abline(h=0, col='red')
```



2.8 Effect Sizes

To see the effect sizes in tabular form use `model.tables`

```
model.tables(mod1, se=T)
```

```
## Tables of effects
##
##  tire
## tire
##       A       B       C       D
## -24.542   0.958  17.458   6.125
##
## Standard errors of effects
##          tire
##         7.691
## replic.     6
```

Specifying `means` gives

```
model.tables(mod1,'means',  se=T)
```

```
## Tables of means
## Grand mean
##
## 404.2083
##
##  tire
## tire
##     A     B     C     D
## 379.7 405.2 421.7 410.3
##
## Standard errors for differences of means
##          tire
##         10.88
## replic.     6
```

2.9 Pairwise comparisons

If the conclusion of the F test is to reject the null hypothesis of no treatment effects, one is naturally interested in determining where the difference lies. For this, it becomes necessary to compare the individual groups. This can be (partly) done looking at the regression coefficients using `summary` and `lm`.

```
summary(mod0)
```

```
##
## Call:
## lm(formula = StopDist ~ tire, data = Tire)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -32.333  -9.667  -2.250  11.417  36.667
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  379.667      7.691  49.363  < 2e-16
## tireB         25.500     10.877   2.344 0.029497
## tireC         42.000     10.877   3.861 0.000973
## tireD         30.667     10.877   2.819 0.010594
##
## (Intercept) ***
## tireB       *
## tireC       ***
## tireD       *
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.84 on 20 degrees of freedom
## Multiple R-squared:  0.4442, Adjusted R-squared:  0.3608
## F-statistic: 5.328 on 3 and 20 DF,  p-value: 0.007316
```

Recall that default coding for treatment in `R` means that the first value (`Intercept`) corresponds to the average value for the first treatment level $\hat{\mu} + \hat{\tau}_1$ while `tireB` $= \hat{\tau}_2 - \hat{\tau}_1$, `tireC` $= \hat{\tau}_3 - \hat{\tau}_1$, and `tireD` $= \hat{\tau}_4 - \hat{\tau}_1$

The t value in the table for `timeB` gives a t-test for comparing the first two groups (tread A vs. B, p-value = 0.0294), the next row in the table has a t-test for comparing groups 1 and 3 (tread A vs C, p-value=0.0009) and finally `tireD` compares tread A vs D, with p-value 0.011. However, there is not a test for comparing groups 2 and 3, or 2 and 4 in the table.

One could redefine the factor `time` to include some of the missing comparisons, but then some others will disappear. This is not a convenient way to get all the tests, particularly if there are many levels for the treatment.

There exists a function for comparing all groups called `pairwise.t.test`. However, one must correct for multiple testing. The problem is that if we perform many tests, the probability of finding one of them to be significant by chance alone increases.

Consider one hundred statistical tests at the 5% level and assume all null hypotheses are true. We expect to reject 5 of them by chance alone. This is the expected value of Type I errors. If the tests are independent, the probability of rejecting at least one hull hypothesis can be calculated using the binomial distribution:

```r
1 - dbinom(0,100,0.05); 1 - dbinom(0,100,0.01)
```

```
## [1] 0.9940795
```

```
## [1] 0.6339677
```

One simple and frequently used method for correction is based on the Bonferroni inequalities and is known as the Bonferroni correction:

$$P(\cup_1^n B_i) \leq \sum_1^n P(B_i)$$

Thus, by dividing the significance level by the number of tests, we get a test with a true significance level smaller or equal to the nominal significance level. This is equivalent to multiplying the p-values by the number of tests. This test is conservative (tends to produce p values that are larger than they really are). The Bonferroni correction should be applied at the planning stage of the experiment.

```r
with(Tire, pairwise.t.test(StopDist, tire,
                p.adjust.method = 'bonferroni'))
```

```
## 
##  Pairwise comparisons using t tests with pooled SD 
## 
## data:  StopDist and tire 
## 
##   A      B      C     
## B 0.1770 -      -     
## C 0.0058 0.8696 -     
## D 0.0636 1.0000 1.0000
## 
## P value adjustment method: bonferroni
```

We have seen that the Bonferroni inequality can be used in the case of preplanned comparisons. We will consider here briefly another method proposed by Tukey for comparisons of the form $H_0 : \tau_s = \tau_u$ for $s \neq u$ in favor of the alternative $H_1 : \tau_s \neq \tau_u$. The procedure simultaneously considers all pairs of effects and adjusts the critical region by using the studentized range statistic instead of student's t-distribution.

The test is

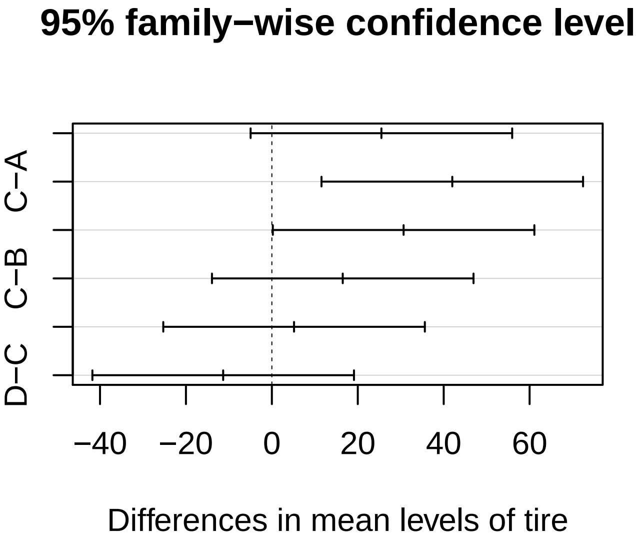
$$\text{reject } H_0 \text{ if } \quad |\hat{\tau}_u - \hat{\tau}_s| > \sqrt{2} q_{I,n-k,1-\alpha/2} \hat{\sigma}(\bar{y}_{s\bullet} - \bar{y}_{u\bullet})$$

where $q_{I,n-k,1-\alpha}$ is the $1-\alpha$ percentile of the studentized range and $\hat{\sigma}(\bar{y}_{s\bullet} - \bar{y}_{u\bullet})$ is the estimated standard error for the difference between the averages. If $X_1, \ldots X_I$ are independent random variables with a $N(\mu, \sigma^2)$ distribution and $R = \max_i X_i - \min_i X_i$ then $R/\hat{\sigma}$ follows the studentized range distribution. This is only approximate when the sample sizes are unequal.

```r
(mod1.tky <-TukeyHSD(mod1))
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
## 
## Fit: aov(formula = StopDist ~ tire, data = Tire)
## 
## $tire
##           diff         lwr      upr     p adj
## B-A  25.500000  -4.9446409 55.94464 0.1213153
## C-A  42.000000  11.5553591 72.44464 0.0049515
## D-A  30.666667   0.2220258 61.11131 0.0479540
## C-B  16.500000 -13.9446409 46.94464 0.4464584
## D-B   5.166667 -25.2779742 35.61131 0.9637307
## D-C -11.333333 -41.7779742 19.11131 0.7273681
```

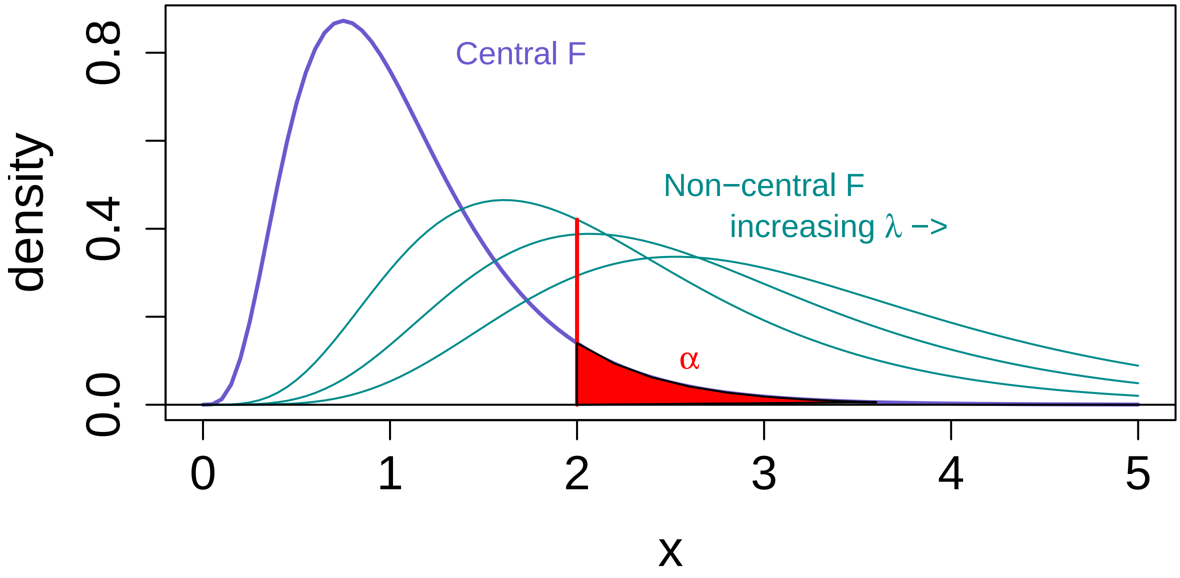
```
plot(mod1.tky)
```



2.10 Power and the number of replicates

The test statistic msT/msE follows the F-distribution when the null hypothesis is true, but when the null hypothesis is false, it follows the noncentral F-distribution. The noncentral F-distribution has a wider spread than the central F-distribution.

The spread in the noncentral F-distribution and the probability of exceeding the critical limit obtained from the central F-distribution is an increasing function of the noncentrality parameter, λ. When the distribution is the noncentral F, the probability of exceeding the critical limit from the central F-distribution is the power of the test.



The power can be computed for any scenario of differing means, if the values of the cell means, the variance of the experimental error, and the number of replicates per factor level are specified. For a constant difference among cell means, represented by $\sum_i(\mu_i - \mu)^2$, the noncentrality parameter and the power increase as the number of replicates increase.

When the differences among cell means are large enough to have practical importance, the experimenter would like to have high power or probability of rejecting the hypothesis of no treatment effects. When the difference among the means has practical importance to the researcher, we call it **practical significance**. Practical significance does not always correspond to statistical significance, as determined by the F-test from the ANOVA.

If there is a difference among the cell means, the power is given by

$$Power(\lambda) = \int_{F_{k-1,k(r-1)\alpha}}^{\infty} f(x, k-1, k(r-1), \lambda)\, dx$$

where

- $F_{k-1,k(r-1)\alpha}$ is the α-th percentile of the central F distribution, with $k-1$ and $k(r-1)$ degrees of freedom
- $f(x, k-1, k(r-1), \lambda)$ is the non-central F density with non-centrality parameter λ and
- $\lambda = \frac{r}{\sigma^2}\sum_{i=1}^{k}(\mu_i - \bar{\mu}_{\bullet})^2$

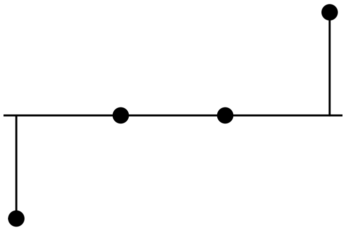
For a fixed value of $\frac{1}{\sigma^2}\sum_{i=1}^{k}(\mu_i - \bar{\mu}_{\bullet})^2$ the power increases with r.

These computations can be carried out with the `Fpower1` function in the `daewr` package.

In the tire example, suppose that the standard tread is `D` and a difference of less than 30 feet in the braking distance is of no interest to the manufacturer, but a difference larger than this value would be of interest. In this case, we regard $\Delta = 30$ as a practical difference in cell means. We need a value, or at least a lower bound for the sum $\sum_{i=1}^{k}(\mu_i - \bar{\mu}_{\bullet})^2$ under the condition that at least one of the values for the stopping distance for treads `A, B` or `C` differ from `D` by at least 30 feet.

The minimum value is attained when the result for one of the cell means for `A, B`, or `C` is lower than the average by $\Delta/2$, the braking distance for `D` is higher than the average by $\Delta/2$, and the other two values are equal to the average.

```r
plot(1:4,c(-1,0,0,1), pch=19, axes = FALSE, xlab='',
     ylab='')
abline(h=0); segments(1,0,1,-1); segments(4,0,4,1)
```



This results in

$$\sum_{i=1}^{k}(\mu_i - \bar{\mu}_{\bullet})^2 = \left(\frac{\Delta}{2}\right)^2 + 0 + 0 + \left(\frac{\Delta}{2}\right)^2 = \frac{\Delta^2}{2} = 450.$$

By previous experience, the manufacturer knows that a reasonable estimate for the variance of the braking distance is 225 ft^2. The noncentrality parameter can be calculated as

$$\lambda = \frac{r}{225}450.$$

The power is calculated for $r = 2, \ldots, 10$ using the `Fpower1` function in the `daewr` package.

```r
library(daewr)
rmin <-2 #smallest number of replicates considered
rmax <-10 # largest number of replicates considered
alpha <- rep(0.05, rmax - rmin +1)
sigma <- 15; nlev <- 4; nreps <- rmin:rmax; Delta <- 30
(power <- Fpower1(alpha,nlev,nreps,Delta,sigma))
```

```
##        alpha nlev nreps Delta sigma     power
##  [1,]  0.05    4     2    30    15 0.1698028
##  [2,]  0.05    4     3    30    15 0.3390584
##  [3,]  0.05    4     4    30    15 0.5037050
##  [4,]  0.05    4     5    30    15 0.6442332
##  [5,]  0.05    4     6    30    15 0.7545861
##  [6,]  0.05    4     7    30    15 0.8361289
##  [7,]  0.05    4     8    30    15 0.8935978
##  [8,]  0.05    4     9    30    15 0.9325774
##  [9,]  0.05    4    10    30    15 0.9581855
```

3 Non-Parametric Tests

3.1 Non-Parametric Tests

Non-parametric procedures make no distributional assumptions about the data being analyzed. In this sense, they free us from the possible pitfalls of choosing a wrong distribution. The price to pay is usually having less power associated with these procedures, when the hypotheses upon which the parametric methods are based are true, which seems natural since non-parametric procedures will work for general samples.

3.2 Comparing Two Populations: The Rank Sum Test

We begin by introducing the rank sums test, which was proposed by Wilcoxon in 1945. Later, Mann and Whitney proposed another test, which is equivalent to Wilcoxon's. Therefore, any combination of these three surnames may appear to refer to this test.

In this test, we are interested in differences in means. Assume that we have two samples $x_1, \dots, x_n$ and $y_1, \dots, y_m$ from two continuous symmetric distributions with the same variance but (possibly) different mean. If the distributions are not symmetric, a test of medians is possible. If the distributions are the same, we can consider that the pooled sample $x_1, \dots, x_n, y_1, \dots, y_m$ is a random sample of size $N = n + m$ from the common distribution. Hence, if these observations are ordered according to magnitude, we would expect to see the xs and ys well mixed. If one of these sets tends to cluster together at some extreme of the ordered sample, we would consider this as evidence that the null hypothesis of equal distributions is not true.

Let μ_X be the (population) mean for the X distribution and similarly for μ_Y. Our test is

$$H_0 : \mu_X = \mu_Y \qquad \text{vs.} \qquad H_A : \mu_X \neq \mu_Y.$$

The (Wilcoxon) rank sums test proceeds as follows: The two samples are joined together, giving a sample of size $N = n + m$. This sample is ordered, and the ranks (positions in the ordered sample) for the elements of the x sample are added up. This is the test statistic W. If the null hypothesis of equal distributions is true, all $\binom{N}{n}$ possible assignments of ranks for the x sample are equally likely, each having probability $1/\binom{N}{n}$.

When the samples have equal sizes $n = n_1 = n_2$, the test statistic W takes integer values ranging from $n(n+1)/2$ to $n(2N - n + 1)/2$ when there are no ties in the ranks. The distribution of W is known as Wilcoxon rank-sum distribution, and it can be obtained in `R` with the command `pwilcox`.

3.2.1 Example

Two samples of fish were drawn from different lakes and the fish were weighted. The weights in grams are

```
sampl1 <- c(286, 251, 325, 313, 309, 308)
sampl2 <- c(249, 324, 289, 303, 310, 318)
```

we build a data frame with this information

```
sample <- c(rep(1,6), rep(2,6))
wcx <- data.frame(weight=c(sampl1,sampl2), sample)
str(wcx)
```

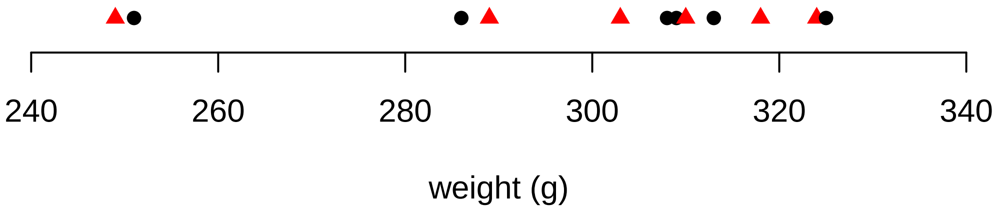
```
## 'data.frame':    12 obs. of  2 variables:
##  $ weight: num  286 251 325 313 309 308 249 324 289 303 ...
##  $ sample: num  1 1 1 1 1 1 2 2 2 2 ...
```

We can order the results by weight using the `order` function. We also add a new column with the rank.

```
ord <- order(wcx[,1])
wcx.ord <- cbind(wcx[ord,],rank=1:12)
head(wcx.ord)
```

```
##    weight sample rank
## 7     249      2    1
## 2     251      1    2
## 1     286      1    3
## 9     289      2    4
## 10    303      2    5
## 6     308      1    6
```

We plot the ordered values



Next, we sum the ranks for the observations in the pooled data:

```
(rank.sum <- c( sum(wcx.ord[,3][wcx.ord[,2]==1]),
                sum(wcx.ord[,3][wcx.ord[,2]==2]) ))
```

```
## [1] 39 39
```

In this example, the rank sums are equal, and this is strong evidence in favor of the null hypothesis of equal means.

The command `wilcox.test()` in `R` performs the rank sums test:

```
wilcox.test(sampl1,sampl2)
```

```
##
##  Wilcoxon rank sum test
##
## data:  sampl1 and sampl2
## W = 18, p-value = 1
## alternative hypothesis: true location shift is not equal to 0
```

Comparing with the t-test, which assumes normality, the rank sums test has less power than the t-test if the distribution is normal, but if departures from normality in the sample are marked, the t test is inferior, and the p-values cannot be trusted.

3.3 The Kruskal-Wallis test

The Kruskal-Wallis test is an extension of the rank-sum test to the case of d multiple samples. The null hypothesis is that all the samples come from the same distribution, while the alternative is that at least two of the samples come from different distributions. The only requisite is that the population distributions have to be continuous.

Because the underlying distributions of the d populations are assumed to be identical under the null hypothesis, this test can be applied to means, medians, or any other quantile. The null and alternative hypotheses are expressed in terms of the means as

$$H_0 : \mu_1 = \cdots = \mu_d \quad \text{vs.} \quad H_A : \mu_i \neq \mu_j \text{ for at least one pair of } i, j.$$

To test the null hypothesis the $n_1, n_2, \ldots, n_d$ observations are pooled together and ordered from 1 to $N = n_1 + \cdots + n_d$ to obtain the ranks. The standardized test statistic used by `R` is

$$H = \frac{12 \sum_{i=1}^{d} n_i(\bar{R}_i - \bar{R})}{N(N+1)},$$

where

- n_i is the number of observations for the i-th treatment,
- $\bar{R}_i$ is the average of the ranks in the i-th treatment and
- $\bar{R}$ is the overall average of the ranks.

When there are ties in the average ranks for the groups, adjustments in the test statistic must be made.

As the size of the smallest group goes to infinity, the test statistic converges in distribution to a χ^2 distribution with $d-1$ degrees of freedom. To use this approximation, it is usually required that the minimum group size be at least five.

In `R`, `kruskal.test()` performs this test with the corresponding corrections when ties are present.

Let us compare the results with this test to those obtained before for the `yields` example. We first reproduce the Anova table and then we do the K-W test:

```r
results <- read.table('yields.txt',header=T)
frame <- stack(results)
names(frame) <- c('yield','soil')
with(frame,summary(aov(yield~soil)))
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## soil         2   99.2   49.60   4.245  0.025 *
## Residuals   27  315.5   11.69
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
with(frame, kruskal.test(yield~soil))
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  yield by soil
## Kruskal-Wallis chi-squared = 7.5813, df = 2,
## p-value = 0.02258
```

We get similar p-values in this case, 0.025 in the Anova table and 0.0226 in the K-W test, so we would reach the same conclusion for the usual values of α.

The other example we considered was the tire experiment:

```
anova(mod0)
```

```
## Analysis of Variance Table
##
## Response: StopDist
##           Df Sum Sq Mean Sq F value   Pr(>F)
## tire       3 5673.1 1891.04  5.3278 0.007316 **
## Residuals 20 7098.8  354.94
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
with(Tire, kruskal.test(StopDist ~ tire))
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  StopDist by tire
## Kruskal-Wallis chi-squared = 9.9133, df = 3,
## p-value = 0.01932
```

In this case, results are not so close. p values are 0.007 and 0.019, so at the 1% level, we would reach different conclusions.