Babraham
Bioinformatics

Introduction to Statistics with R

Version 2019-03

Licence

Table of Contents

Introduction

R is a popular language and environment that allows powerful and fast manipulation of data, offering many statistical and graphical options.
Graphical representation of data is pivotal when one wants to present scientific results, in particular in publications. R allows you to build top quality graphs (much better than Excel for example).

In this manual, however, we are going to focus on the statistical possibilities of R. Whatever package you use, you need some basic statistical knowledge if only to design your experiments correctly, so there is no way out of it!

And don’t forget: you use stats to present your data in a comprehensible way and to make your point; this is just a tool, so don’t hate it, use it!

“To consult a statistician after an experiment is finished is often merely to ask him to conduct a post-mortem examination. He can perhaps say what the experiment died of." R. A. Fisher, 1938.

Chapter 1: sample size estimation

It's practically impossible to collect data on an entire population of interest. Instead we examine data from a *random sample* to provide support for or against our hypothesis. Now the question is: how many samples/participants/data points should we collect?

Power analysis allows us to determine the sample sizes needed to detect statistical effects with high probability.
Experimenters often guard against false positives with statistical significance tests. After an experiment has been run, we are concerned about falsely concluding that there is an effect when there really isn't. Power analysis asks the opposite question: supposing there truly is a treatment effect and we were to run our experiment a huge number of times, how often will we get a statistically significant result? Answering this question requires informed guesswork. We'll have to supply guesses as to how big our treatment effect can reasonably be for it to be biologically/clinically relevant/meaningful.

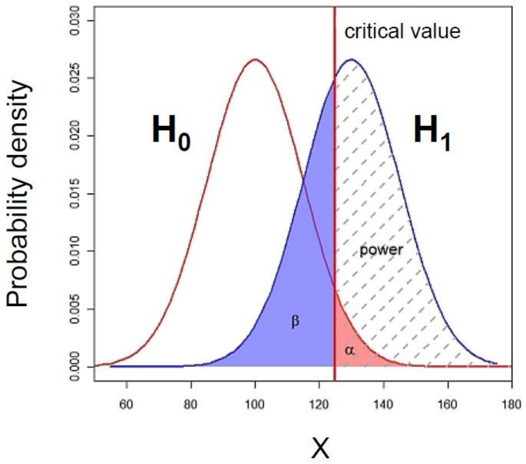
What is Power?

First, the definition of power: probability that a statistical test will reject a false null hypothesis (H_0) when the alternative hypothesis (H_1) is true. We can also say: it is the probability of detecting a specified effect at a specified significance level. Now 'specified effect' refers to the effect size which can be the result of an experimental manipulation or the strength of a relationship between 2 variables. And this effect size is 'specified' because prior to the power analysis we should have an idea of the size of the effect we expect to see. The 'probability of detecting' bit refers to the ability of a test to detect an effect of a specified size. The recommended power is 0.8 which means we have an 80% chance of detecting an effect if one genuinely exists.

Power is defined in the context of hypothesis testing. A hypothesis (statistical) test tells us the probability of our result (or a more extreme result) occurring, if the null hypothesis is true. If the probability is lower than a pre-specified value (alpha, usually 0.05), it is rejected.
The null hypothesis (H_0) corresponds to the absence of effect and the aim of a statistical test is to reject or not H_0. A test or a difference are said to be "significant" if the probability of type I error is: $\alpha =< 0.05$ (max $\alpha=1$). It means that the level of uncertainty of a test usually accepted is 5%.

Type I error is the incorrect rejection of a true null hypothesis (false positive). Basically, it is the probability of thinking we have found something when it is not really there.

Type II on the other hand, is the failure to reject a false null hypothesis (false negative), so saying there is nothing going on whereas actually there is. There is a direct relation between Type II error and power, as Power = $1 – \beta$ where $\beta=0.20$ usually hence power = 0.8 (probability of drawing a correct conclusion of an effect). We will go back to it in more detail later.
Below is a graphical representation of what we have covered so far. H_1 is the alternative hypothesis and the critical value is the value of the difference beyond which that difference is considered significant.



Statistical decision	True state of H_0	
	H_0 True (no effect)	H_0 False (effect)
Reject H_0	Type I error (False Positive) **α**	Correct (True Positive)
Do not reject H_0	Correct (True Negative)	Type II error (False Negative) **β**

The ability to reject the null hypothesis depends upon alpha but also the sample size: a larger sample size leads to more accurate parameter estimates, which leads to a greater ability to find what we were looking for. The harder we look, the more likely we are to find it. It also depends on the effect size**:** the size of the effect in the population: the bigger it is, the easier it will be to find.

What is Effect Size?

Power analysis allows us to make sure that we have looked hard enough to find something interesting. The size of the thing we are looking for is the effect size. Several methods exist for deciding what effect size we would be interested in. Different statistical tests have different effect sizes developed for them, however the general principle is the same. The first step is to make sure to have preliminary knowledge of the effect we are after. And there are different ways to go about it.

Effect size determined by substantive knowledge

One way is to identify an effect size that is meaningful i.e. biologically relevant. The estimation of such an effect is often based on substantive knowledge. Here is a classic example: It is hypothesised that 40 year old men who drink more than three cups of coffee per day will score more highly on the Cornell Medical Index (CMI: a self-report screening instrument used to obtain a large amount of relevant medical and psychiatric information) than same-aged men who do not drink coffee. The CMI ranges from 0 to 195, and previous research has shown that scores on the CMI increase by about 3.5 points for every decade of life. Therefore, if drinking coffee caused a similar increase in CMI, it would warrant concern, and so an effect size can be calculated based on that assumption.

Effect size determined from previous research

Another approach is to base the estimation of an interesting effect size on previous research, see what effect sizes other researchers studying similar fields have found. Once identified, it can be used to estimate the sample size.

Effect size determined by conventions

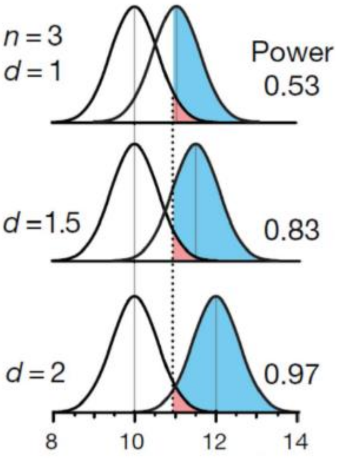
Yet another approach is to use conventions. Cohen (author of several books and articles on power analysis) has defined small, medium and large effect sizes for many types of test. These form useful conventions, and can guide you, if you know approximately how strong the effect is likely to be.

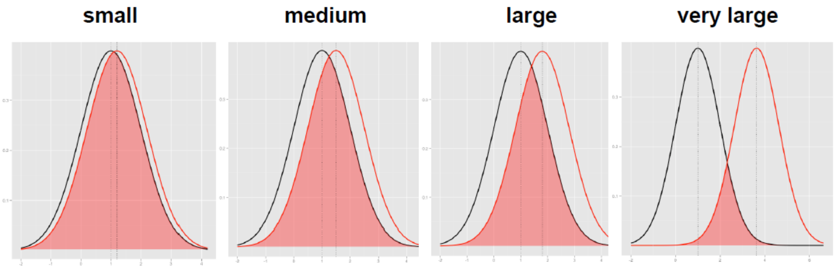
Table 1: Thresholds/Convention for interpreting effect size

Test	**Relevant effect size**		**Effect Size Threshold**	
		Small	Medium	Large
t-test for means	d	0.2	0.5	0.8
F-test for ANOVA	f	0.1	0.25	0.4
t-test for correlation	r	0.1	0.3	0.5
Chi-square	w	0.1	0.3	0.5
2 proportions	h	0.2	0.5	0.8

Note: The rationale for these benchmarks can be found in Cohen (1988), Rosenthal (1996) later added the classification of very large.

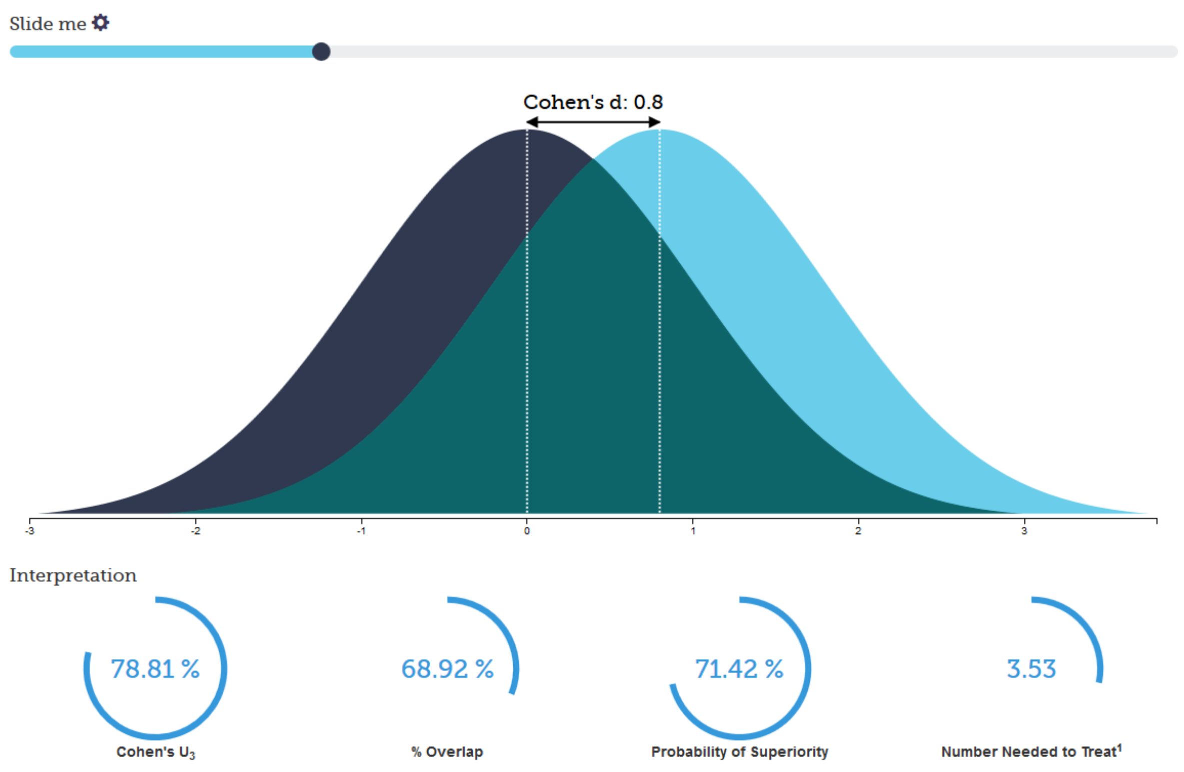
The graphs below give a visual representation of the effect sizes.





Krzywinski and Altman 2013 (Nature Methods)

Below is a link to a sliding tool providing a visual approach to Cohen’s effect size: http://rpsychologist.com/d3/cohend/



The point is sample size is always determined to detect some hypothetical difference. It takes huge samples to detect tiny differences but tiny samples to detect huge differences, so you have to specify the size of the effect you are trying to detect.

So how is that effect size calculated anyway?

Let's start with an easy example. If we think about comparing 2 means, the effect size, called Cohen's *d*, is just the standardised difference between 2 groups:

$$\text{Effect Size} = \frac{\text{[Mean of experimental group] - [Mean of control group]}}{\text{Standard Deviation}}$$

The standard deviation is a measure of the spread of a set of values. Here it refers to the standard deviation of the population from which the different treatment groups were taken. In practice, however, this is almost never known, so it must be estimated either from the standard deviation of the control group, or from a 'pooled' value from both groups.

McGraw and Wong (1992) have suggested a 'Common Language Effect Size' (CLES) statistic, which they argue is readily understood by non-statisticians (shown in column 5 of Table 2). This is the probability that a score sampled at random from one distribution will be greater than a score sampled from another. They give the example of the heights of young adult males and females, which differ by an effect size of about 2, and translate this difference to a CLES of 0.92. In other words 'in 92 out of 100 blind dates among young adults, the male will be taller than the female'.

Table2: Interpretation of Effect Size (Robert Coe, 2002)

Effect Size	Percentage of control group below average person in experimental group	Rank of person in a control group of 25 equivalent to the average person in experimental group	Probability that you could guess which group a person was in from knowledge of their 'score'.	Probability that person from experimental group will be higher than person from control, if both chosen at random (=CLES)
0.0	50%	13th	0.50	0.50
0.2	58%	11th	0.54	0.56
0.5	69%	8th	0.60	0.64
0.8	79%	6th	0.66	0.71
1.2	88%	3rd	0.73	0.80
1.4	92%	2nd	0.76	0.84
2.0	98%	1st	0.84	0.92

Doing power analysis

The main output of a power analysis is the estimation of a sufficient sample size. This is of pivotal importance of course. If our sample is too big, it is a waste of resources; if it is too small, we may miss the effect (p>0.05) which would also mean a waste of resources. On a more practical point of view, when we write a grant, we need to justify our sample size which we can do through a power analysis. Finally, it is all about the ethics of research really which is encapsulated in the UK Home office’s 3 R: Replacement, Refinement and Reduction. The latter in particular relates directly to power calculation as it refers to ‘methods which minimise animal use and enable researchers to obtain comparable levels of information from fewer animals’ (NC3Rs website).

When should we run a power analysis? It depends on what we expect from it: the most common output being the sample size, we should run it before doing the actual experiment (*a priori* analysis). The correct sequence from hypothesis to results should be:

Hypothesis

↓

Experimental design
Choice of a Statistical test

↓

Power analysis

↓

Sample size

↓

Experiment(s)

↓

(Stat) analysis of the results

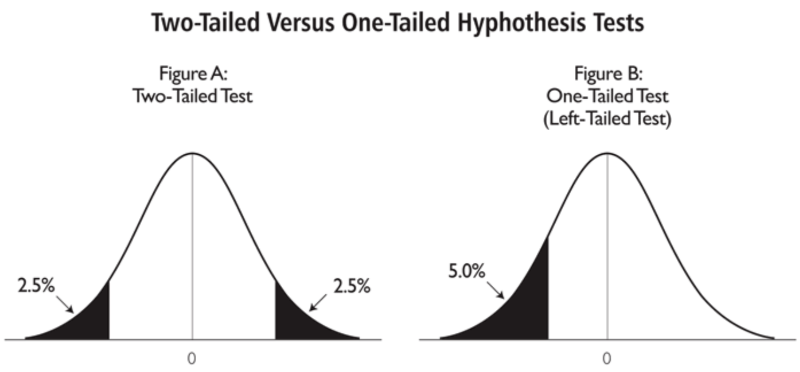
Practically, the power analysis depends on the relationship between 6 variables: the significance level, the desired power, the difference of biological interest, the standard deviation (together they make up for the effect size), the alternative hypothesis and the sample size. The significance level is about the p-value (α =< 0.05), the desired power, as mentioned earlier is usually 80% and we already discussed effect size.

Now the alternative hypothesis is about choosing between one and 2-sided tests (= one and 2-tailed tests). This is both a theoretical and a practical issue and it is worth spending a bit of time reflecting on it as it can help understanding this whole idea of power.

We saw before that the bigger the effect size, the bigger the power as in the bigger the probability of picking up a difference.

Going back to one-tailed vs. 2-tailed tests, often there are two alternatives to H_0, and two ways the data could be different from what we expect given H_0, *but we are only interested in one of them*. This will influence the way we calculate *p*. For example, imagine a test finding out about the length of eels. We have 2 groups of eels and for one group, say Group 1, we know the mean and standard deviation, for eels length. We can then ask two different questions. First question: 'What is the probability of eels in Group 2 having a different length to the ones in Group 1?' This is called a **two-tailed** test, as we'd calculate *p* by looking at the area under both '**tails'** of the normal curve (See graph below).

And second question: 'What is the probability of eels in Group 2 being longer than eels in Group 1?' This is a **one-tailed** test, as we'd calculate *p* by looking at the area under only one end of the normal curve. The one-tailed *p* is just one half of the two-tailed *p*-value. In order to use a one-tailed test *we must be only interested in one of two possible cases, and be able specify which in advance.*

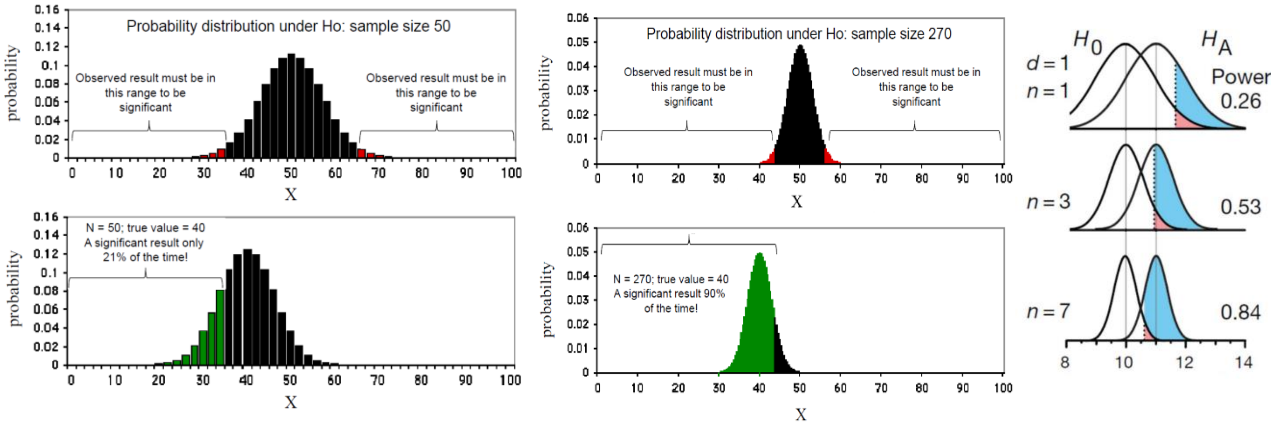


If you can reasonably **predict** the direction of an effect, based on a scientific hypothesis, a 1-tailed test is more powerful than a 2-tailed test. However, it is not always rigidly applied so be cautious when 1-tailed tests are reported, especially when accompanied by marginally-significant results! And reviewers are usually very suspicious about them.

So far we have discussed 5 out of the 6 variables involved in power analysis: the effect size (difference of biological interest + the standard deviation), the significance level, the desired power and the alternative hypothesis. We are left with the variable which we are actually after when we run a power analysis: the sample size.

To start with, we saw that the sample size is related to power but how does it work? It is best explained graphically.

The graph below on the left shows what happens with a sample of n=50, the one of the right what happens with a bigger sample (n=270). The standard deviation of the sampling distribution (= SEM so standard error of the mean) decreases as N increases. This has the effect of reducing the overlap between the H_0 and H_1 distributions. Since in reality it is difficult to reduce the variability inherent in data, or the contrast between means, the most effective way of improving power is to increase the sample size.

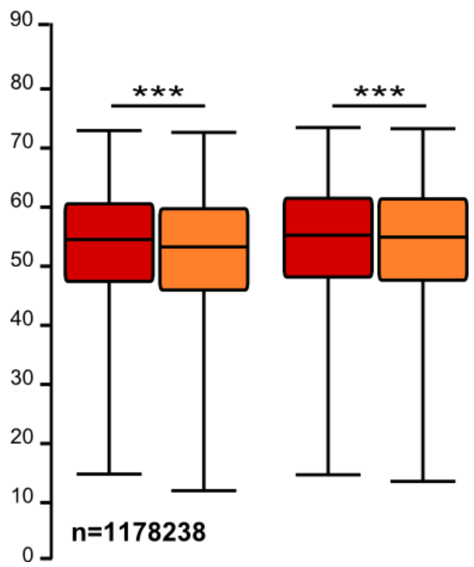


Krzywinski and Altman 2013 (Nature Methods)

So the bigger the sample, the bigger the power and the higher the probability to detect the effect size we are after.

The problem with overpower

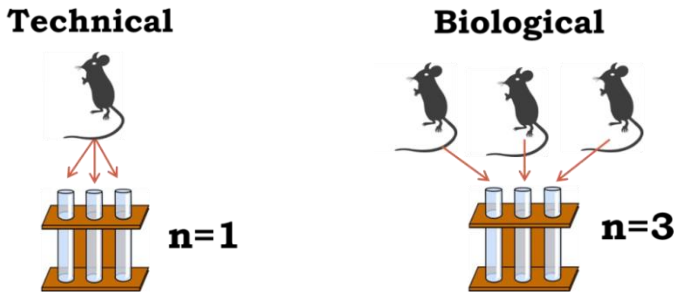
As we saw, power and effect size are linked so that the bigger the power the smaller the effect size that can be detected, as in associated with a significant p-value. The problem is that there is such a thing as overpower. Studies or experiments which produce thousand or hundreds of thousands of data, when statistically analysed will pretty much always generate very low p-values even when the effect size is minuscule. There is nothing wrong with the stats, what matters here is the interpretation of the results.



When the sample size is able to detect differences much finer than the expected effect size, a difference that is correctly statistically distinct is not practically meaningful (and from the perspective of the "end-user" this is effectively a "false positive" even if it's not a statistical one). Beyond the ethical issues associated with overpower, it all comes back to the importance of having in mind a meaningful effect size before running the experiments.

Sample size (n): biological vs. technical replicates (=repeats)

When thinking about sample size, it is very important to consider the difference between technical and biological replicates. For example, technical replicates involve taking several samples from one tube and analysing it across multiple conditions. Biological replicates are different samples measured across multiple conditions. When the experimental unit is an animal, it is pretty easy to make the distinction between the 2 types of replicates.



To run proper statistical tests so that we can make proper inference from sample to general population, we need biological samples. Staying with mice, if we randomly select one white and one grey mouse and measure their weights, we will not be able to draw any conclusions about whether grey mice are, say, heavier in general. This is because we only have two biological samples.

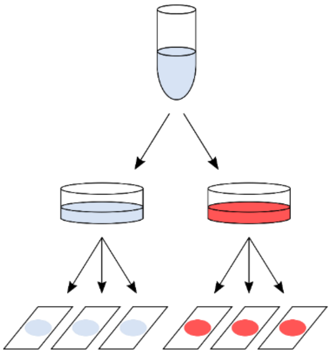
If we repeat the measurements, let’s say we weigh each mouse five times then we will have ten different measurements. But this cannot be used to prove that grey mice are heavier than white mice in general, we still have only looked at one white and one grey mouse. Using the terminology above, the five measurements of each mouse are technical replicates.
What we need to do is to select five different white mice and five different grey mice. Then we would have more than two biological samples and be able to say if there is a statistical difference between white and grey mice in general.

So the concept of biological replicates is quite easy to understand when dealing with animals. But what is "n" in cell culture experiments?

(The examples below are extracts from *Statistics for Experimental Biologists*)
One of the difficulties in analyzing cell culture experiments is determining what the experimental unit is, or what counts as a replicate, or "n". This is easy when cells are derived from different individuals, for example if a blood sample is taken from 20 individuals, and ten serve as a control group while the other ten are the treated group. It is clear that each person is a biological replicate and the blood samples are independent of each other, so the sample size is 20. However, when cell lines are used, there isn't any biological replication, only technical replication, and it is important to have this replication at the right level in order to have valid inferences. The examples below will mainly discuss the use of cell lines. In the figures, the tubes represent a vial of frozen cells, the dishes could be separate flasks, separate culture dishes, or different wells in a plate, and represent cells in culture and the point at which the treatment is applied. The flat rectangular objects could represent glass slides, microarrays, lanes in a gel, or wells in a plate, etc. and are the point at which something gets measured. The control groups are grey and the treated groups are red.

Design 1: As bad as it can get

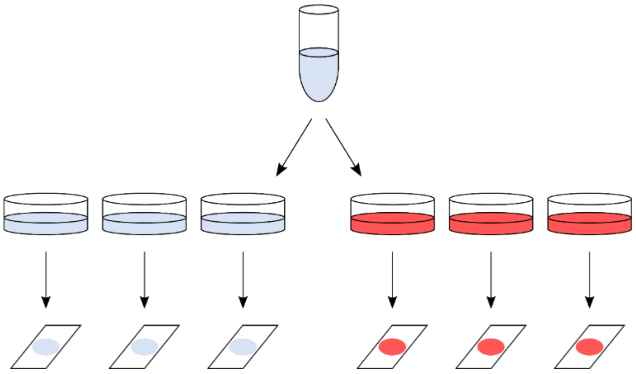
In this experiment a single vial is thawed, cells are divided into two culture dishes and the treatment (red) is randomly applied to one of the two dishes. The cells are allowed to grow for a period of time, and then three samples are pipetted from each dish onto glass slides, and the number of cells are counted (yes there are better ways to count cells, the main point is that from each glass slide we get just one value, in this case the total number of cells). So after the quantification, there are six values--the number of cells on the three control and three treated slides. So what is the sample size--there was one vial, two culture dishes, and six glass slides?



The answer, which will surprise some people, is one, and most certainly not six. The reason for this has to do with the lack of independence between the three glass slides within each condition. A nonlaboratory example will clarify why. Suppose I want to know if people gain weight over the Christmas holidays, so I find one volunteer and measure their weight three times on the morning of Dec 20th (within a few minutes of each other). Then, on the morning of Jan 3rd I measure this same person's weight three times. So I have six data points in total, and I can calculate means, SEMs, 95%CIs, and can even do a t-test. But with these six values, can I address the research question? No, because the research question was **do people** gain weight over the holidays, but I have observations on only one person, and taking more and more observations on this single person will not enable me to make better estimates of weight changes in people. The key point is that the variability from slide-to-slide within a condition is only **pipetting error** (just like measuring someone's weight three times within a few minutes of each other), and therefore those values do not constitute a sample size of three in each condition.

Design 2: Marginally better, but still not good enough

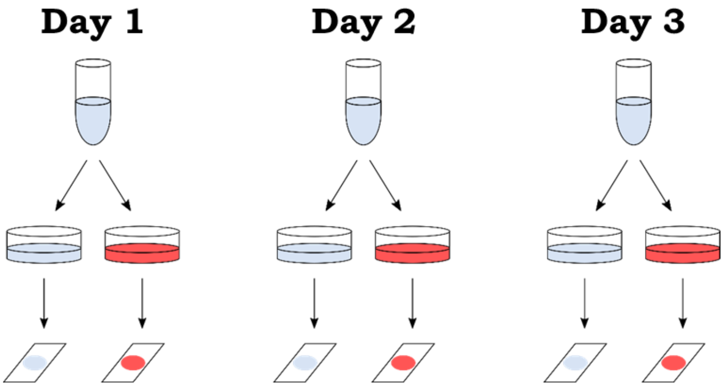
In this modified experiment, the vial of cells is divided into six separate culture dishes, and then cells from each culture dish are pipetted onto a single glass slide. Similar to the previous experiment, there are six values after quantifying the number of cells on each slide. So now is the sample size six?



Unfortunately not, because even though the cells were grown in separate dishes, they are not really independent because they were all processed on the same day, they were all sitting in the same medium, they were all kept in the same incubator at the same time, etc. Cells in two culture dishes from the same stock and processed identically do not become fully independent just because a bit of plastic has been placed between them. However, one might expect some more variability within the groups compared to the first design because the samples were split higher up in the hierarchy, but this is not enough to ensure the validity of the statistical test. To keep with the weight gain analogy, you can think of this as measuring a person's weight in the morning, afternoon, and evening on the same day, rather than taking measurements a few minutes apart. The three measurements are likely to be a bit more variable, but still highly correlated.

Design 3: Often, as good as it can get

In this design, a vial of cells is thawed, divided in two culture dishes, and then eventually one sample from each dish is pipetted onto a glass slide. The main (and key) difference is that the whole procedure is repeated three separate times. Here, they are listed as Day 1, 2, and 3, but they need not be consecutive days and could be weeks or even months apart. This is where independence gets introduced, even though the same starting material is used (i.e. same cell line), the whole procedure is done at one time, and then repeated at another time, and then a third time. There are still six numbers that we get out of the experiment, but the variability now includes the variability of doing the experiment more than once. Note that this is still technical variability, but it is done at the highest level in the hierarchy, and the results of one day are (mostly) independent of the results of another day. And what is the sample size now?



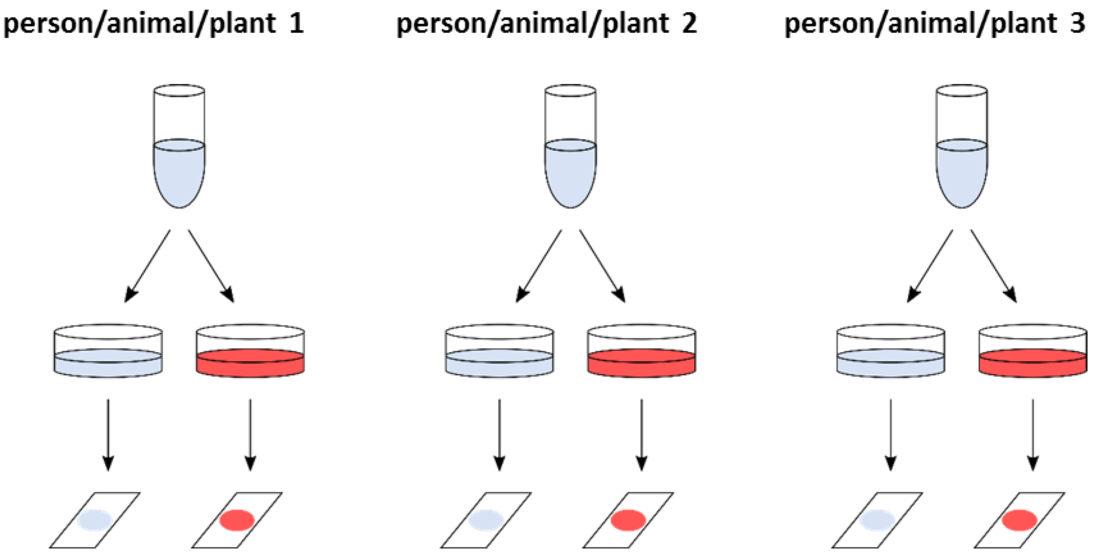
The "independent" aspect of the experiment are the days, and so n = 3. Note, that the two glass slides from the same day can (and should) be treated as paired observations, and so it is the difference between treated and control within each day that is of interest (a paired-samples t-test could be used). An important technical point is that these three replications should be made as independent as possible. This means that it is better to complete the first experiment before starting the second. For example, if the cells will be grown in culture for a week, it is better to do everything over three weeks rather than starting the first experiment on a Monday, the next on Tuesday, and the third on Wednesday. If the three experiments are mostly done in parallel, they will not be as independent as when done back-to-back. Ideally, different media should be made up for each experiment, but this is where reality often places constraints on what is statistically optimal.

Continuing with the weight-gain example, this design is similar to measuring a person's weight before and after the holidays over three consecutive years. This is still not ideal for answering the research question (which was determining whether ***people*** gain weight over the holidays), but if we have only one volunteer at our disposal then this is the best we can do. But now at least we can see whether the phenomenon is reproducible over multiple years, which will give us a bit more confidence that the phenomenon is real. We still don't know about other people, and the best we could do was repeated experiments on this one person.

Design 4: The ideal design

Like many ideals, the ideal experiment is often impossible to attain. With cell lines, there are no biological replicates, and so Design 3 is the best that can be done. The ideal design would have biological replicates (i.e. cells from multiple people or animals), and in this case the experiment need only be done once. I hope it is now clear (and after reading the two references) why Design 1 and Design 2 do not provide any reason to believe that the results will be reproducible. Some people may object that it is a weak analogy, and say that they are only interested in whether compound X increases phosphorylation of protein Y, and are not interested in other proteins, other compounds, other cell lines, etc., and so Design 1 or 2 are sufficient. Unfortunately, this is not the case and it has to do with lack of independence, which is a fundamental assumption of the statistical analysis (see Lazic, 2010 and references therein). But even if you don't appreciate the statistical arguments, this analogy might help:

if you claim to be a superstar archer and hit the bullseye to prove it, this is certainly evidence that you have some skill, but let's see if you can do it three times in a row.

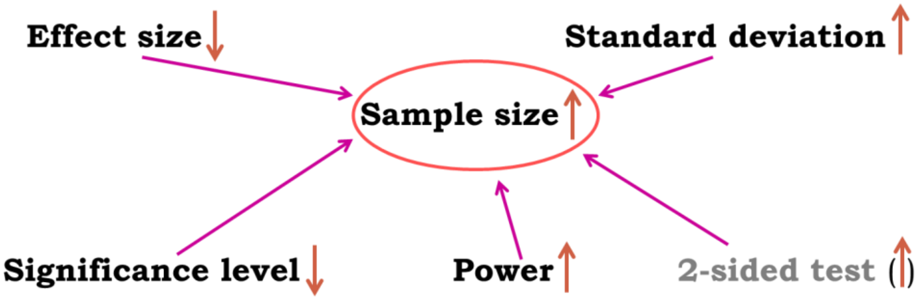


Replication at multiple levels

The analysis of such cell culture experiments in many published studies is inappropriate, even if there were replicate experiments. You will probably have noticed the hierarchical nature of the data: the experiment can be conducted on multiple days, there can be replications of cell cultures within days, there can be replications of more than one glass slide per culture dish, and often multiple measurements within each glass slide can be taken (in this example the total number of cells was measured, but the soma size of 20 randomly selected cells on each glass slide could have been measured, which would give many more data points). This hierarchical structure needs to be respected during the analysis, either by using a hierarchical model (also known as a mixed-effects or multi-level model) or by averaging the lower level values (see Lazic, 2010). Note that it is NOT appropriate to simply enter all of the numbers into a statistics program and run a simple t-test or ANOVA. It is really important to remember that you should never mix biological and technical replicates.

Two more things to note. First, it is possible to have replication at multiple levels, in the previous examples replication was only introduced at one level at a time to illustrate the concepts. However, it is often of interest to know at which level most of the variation comes from, as this will aid in designing future experiments. Cost considerations are also important, if samples are difficult to obtain (e.g. rare clinical samples) then technical replication can give more precise estimates for those precious few samples. However, if the samples are easy to get and/or inexpensive, and you want to do a microarray study (substituting expensive arrays for the glass slides in the previous examples), then there is little point in having technical replicates and it is better to increase the number of biological replicates. Second, if you want to increase the power of the analysis, you need to replicate the "days", not the number of culture dishes within days, or the number of glass slides within a culture dish, or the number of cells on a slide. Alternatively, if biological replicates are available, increasing these will increase power, but not more technical replicates.

Going back to the basic idea behind the power analysis that if you fix any five of the variables, a mathematical relationship can be used to estimate the sixth. The variables are all linked and will vary as shown in the following diagram.



Now here is the good news, there are packages that can do the power analysis for us … providing of course we have some prior knowledge of the key parameters.

We are going to go through 2 examples of power calculations:

- Comparing 2 proportions
- Comparing 2 means

Examples of power calculation

We have previously mainly mentioned quantitative variables but it is also possible to think about power in the context of qualitative variable. All statistical tests, regardless of the type of outcome variables they are dealing with, are associated with a measure of power. Statistics are about confidence in the inferential potential of the results of an experiment so when comparing 2 proportions the question becomes: What makes me believe that 35% is different from 50%? The answer is: a sample big enough, and the 'big enough' is estimated by a power analysis. What makes me believe that 35% is statistically different from 45%? The answer is: a bigger sample!

Comparing 2 proportions

(Data from: http://www.sciencealert.com/scientists-are-painting-eyes-on-cows-butts-to-stop-lions-getting-shot)

Scientists have come up with a solution that will reduce the number of lions being shot by farmers in Africa - painting eyes on the butts of cows. It sounds a little crazy, but early trials suggest that lions are less likely to attack livestock when they think they're being watched - and fewer livestock attacks could help farmers and lions co-exist more peacefully.

Pilot study over 6 weeks: 3 out of 39 unpainted cows were killed by lions, none of the 23 painted cows from the same herd were killed.

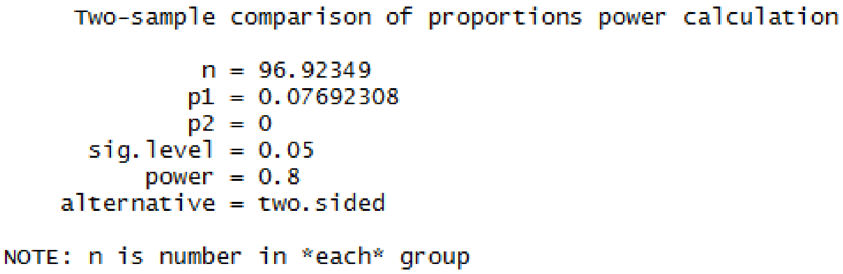
- Do you think the observed effect is meaningful to the extent that such a 'treatment' should be applied? Consider ethics, economics, conservation …
- Run a power calculation to find out how many cows should be included in the study.

We will be using the function below:

```
power.prop.test(n=NULL, p1=NULL, p2=NULL, sig.level=NULL, power=NULL,
alternative=c("two.sided", "one.sided")
```

Exactly one of the parameters n, p1, p2, power and sig.level must be passed as NULL, and that parameter is determined from the others. "two-sided" is the default.

```
power.prop.test(n=NULL, p1=3/39, p2=0, sig.level=0.05, power=0.8,
alternative="two.sided")
```



```
     Two-sample comparison of proportions power calculation 

              n = 96.92349
             p1 = 0.07692308
             p2 = 0
      sig.level = 0.05
          power = 0.8
    alternative = two.sided

NOTE: n is number in *each* group
```

To be able to pick such a difference between the 2 groups, as in to reach significance, we will need about 97 cows in each group. In other words: if we want to be at least 80% confident to spot a treatment effect, if indeed there is one, we will need about 200 cows altogether.

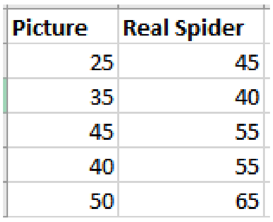
Always remember, power calculations are guessing exercises, the sample sizes found are never absolute. Our data might show an effect a bit bigger ☺ or smaller ☹ than expected. By doing a power calculation, providing the effect size we are after is meaningful, we want to know if we can afford to run the experiment. Afford in all possible ways: money, time, space … and ethically. In our case here, we wanted to know how many-ish animals were needed:90-ish happens to be OK but if it had been 900-ish, maybe the cost-benefit of the experiment would not have been worth it.

One last thing: be careful with small samples sizes. If our power calculation tells you that we need n=3 or 4, try to add one or 2 experimental units if you possible. With n=3, we cannot afford any mistake, so if something goes wrong with one of our animals for instance, we will end up with n=2 and be in trouble statwise.

Comparing 2 means

(Data from *'Discovering Stats with SPSS'* by Andy Field)

Pilot study: 10 arachnophobes were asked to perform 2 tasks:
Task 1: Group1 (n=5): to play with a big hairy tarantula spider with big fangs and an evil look in its eight eyes.
Task 2: Group 2 (n=5): to look only at pictures of the same hairy tarantula.

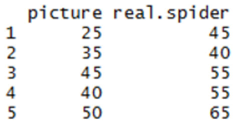


Picture	Real Spider
25	45
35	40
45	55
40	55
50	65

Anxiety scores were measured for each group (0 to 100).

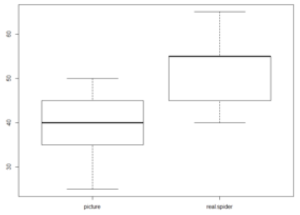
- Use R to calculate the values for a power calculation
  - Enter the data in R
  - Hint: you will need data.frame() and apply()
- Run a power calculation to find out how many subjects should be included in the study.

```
## enter data
picture <- c(25,35,45,40,50)
real.spider <- c(45,40,55,55,65)

spider.data<-data.frame(picture=picture, real.spider=real.spider )
head(spider.data)
```

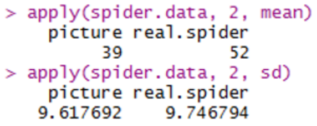


```
  picture real.spider
1      25          45
2      35          40
3      45          55
4      40          55
5      50          65
```

```
## quick look
boxplot(spider.data)
```

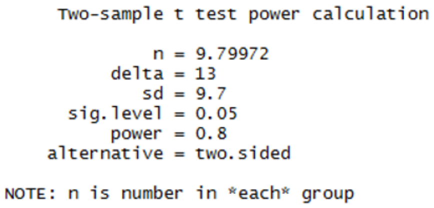


```
## extract data for power
apply(spider.data, 2, mean)
apply(spider.data, 2, sd)
```



```
> apply(spider.data, 2, mean)
    picture real.spider 
         39          52 
> apply(spider.data, 2, sd)
    picture real.spider 
   9.617692    9.746794 
```

```
## power
power.t.test(n = NULL, delta = 52-39, sd = 9.7, sig.level = 0.05,
power = 0.8, type = "two.sample",
             alternative = "two.sided")
```



```
     Two-sample t test power calculation 

              n = 9.79972
          delta = 13
             sd = 9.7
      sig.level = 0.05
          power = 0.8
    alternative = two.sided

NOTE: n is number in *each* group
```

Unequal sample sizes

So far we have only considered balanced design as in groups of equal sizes. However, more often than not, scientists have to deal with unequal sample sizes for a wide variety of reasons. The problem is that there is not a simple trade-off as in if one needs 2 groups of 30 for a particular comparison, going for 20 and 40 will be associated with decreased power.

The best approach is to run the power calculation based on a balanced design and then apply a correction. The tricky bit is that we need to have an idea of the unbalance and express it as a ratio (k) of the 2 sample sizes.

The formula to correct for unbalanced design is then quite simple.

With k, the ratio of the samples sizes in the 2 groups after adjustment (=n1/n2)

$$N = \frac{2n(1+k)^2}{4k}$$

$$n_1 = \frac{N}{(1+k)}$$

$$n_2 = \frac{kN}{(1+k)}$$

Power calculation for non-parametric tests

Nonparametric tests are used when we are not willing to assume that our data come from a Gaussian distribution. Commonly used nonparametric tests are based on ranking values from low to high, and then looking at the distribution of sum-of-ranks between groups.

Now if we want to run a proper power calculation for non-parametric tests, we need to specify which kind of distribution we are dealing with. This would imply more advanced approach to the data and it is not the purpose of this manual.

But if we don't know the shape of the underlying distribution, we cannot do proper sample size calculation. So we have a problem here.

Fortunately, there is a way to have a rough idea of the sample size needed. First of all, non-parametric tests are usually said to be less powerful than their parametric counterparts. It is not always true and depending on the nature of the distribution, the non-parametric tests might actually require fewer subjects. And when they need more, they never require more than 15% additional subjects providing these 2 assumptions are true: we are looking at reasonably high numbers of subjects (say at least n=30) and the distribution is not too unusual.

So the rule of thumb is: if we plan to use a nonparametric test, we compute the sample size required for a parametric test and add 15%.

Chapter 2: Some key concepts

A bit of theory: the null hypothesis and the error types.

The null hypothesis (H_0) corresponds to the absence of effect (e.g.: the animals rewarded by food are as likely to line dance as the ones rewarded by affection) and the aim of a statistical test is to accept or to reject H_0. As mentioned earlier, traditionally, a test or a difference are said to be “significant” if the probability of type **I** error is: $\alpha =< 0.05$ (max α=1). It means that the level of uncertainty of a test usually accepted is 5%. It also means that there is a probability of 5% that we may be wrong when we say that our 2 means are different, for instance, or we can say that when we see an effect we want to be at least 95% confident that something is significantly happening.

Statistical decision	True state of H_0	
	H_0 True (no effect)	H_0 False (effect)
Reject H_0	Type I error (False Positive)	Correct (True Positive)
Do not reject H_0	Correct (True Negative)	Type II error (False Negative)

Tip: if our p-value is between 5% and 10% (0.05 and 0.10), I would not reject it too fast. It is often worth putting this result into perspective and ask ourselves a few questions like:

- what the literature says about what am I looking at?
- what if I had a bigger sample?
- have I run other tests on similar data and were they significant or not?

The interpretation of a border line result can be difficult so it is important to look at the whole picture.

The specificity and the sensitivity of a test are closely related to Type I and Type II errors.

Specificity = Number of True Negatives / (Number of False Positives + Number of True Negatives). A test with a high specificity has a low type I error rate.

Sensitivity = Number of True Positives / (Number of False Negatives + Number of True Positives). A test with a high sensitivity has a low type II error rate.

A bit of theory: Statistical inference

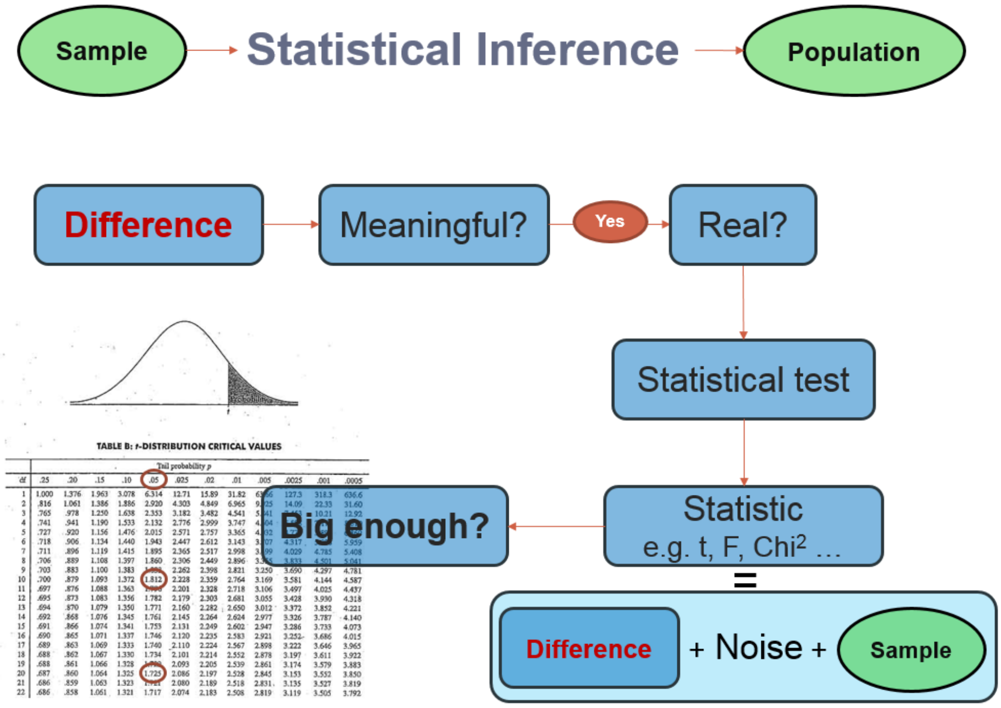
This is such an obvious concept that people tend to forget about it. The whole point of looking at a sample of data and analysing it is because we assume that that sample is a fair representation of the population it is coming from. As such, the findings from the sample can be inferred to that population, they can be generalised.

With that in mind, if we observe a difference between 2 groups in our sample, we get excited because we think that what we are observing can be what is happening in the general population, as in 'for real'.

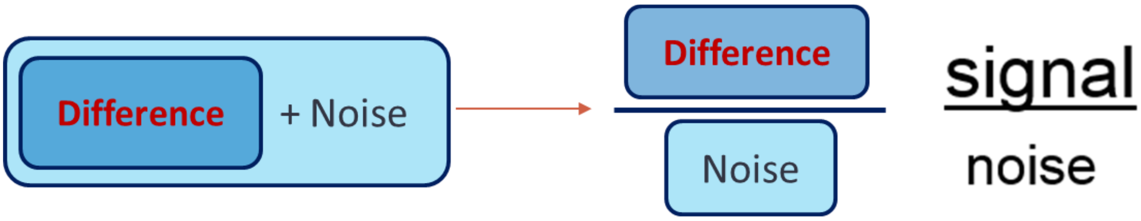
Now when we observe a difference, we get excited if that difference is meaningful or, rather, we should. We should only get excited by a difference which is biologically relevant in the context of our study, and not by any difference.

So, let's say that the difference is meaningful, the next question is: is it real? And for that we need to apply a statistical test, which will allow us to quantify the confidence we have in our difference. All statistical tests produce a statistic (e.g. t-value, F …) and statistics are all about the difference observed but also about the variability of the data (the noise) and the sample size. We need all three to know how confident we are, to be able to infer from our sample to the population.

Then the final question is: is the statistic big enough? Because it will almost never be 0, there will always be a difference, but when does this difference start to be real, meaningful, significant? Statistical tests allow us to draw a line, the critical value, beyond which the result is significant, the difference is real.



The signal-to-noise ratio



Statistics are all about understanding and controlling variation. In pretty much all quantitative tests, the statistic is a variation on the theme of the so-called signal-to-noise ratio, in effect the difference over the variability. We want this ratio to be as big as possible because if the noise is low then the signal is detectable but if the noise (i.e. inter-individual variation) is large then the same signal will not be detected. So in a statistical test, the signal-to-noise ratio determines the significance.

Chapter 3: Descriptive statistics

When it comes to quantitative data, more tests are available but assumptions must be met before applying them. In fact, there are 2 types of stats tests: parametric and non-parametric ones. Parametric tests have 4 assumptions that must be met for the tests to be accurate. Non-parametric tests are based on ranks and they make few or no assumptions about populations parameters like normality (e.g. Mann-Whitney test).

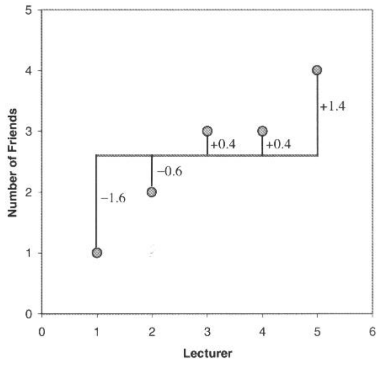
3-1 A bit of theory: descriptive stats

The median: The median is the value exactly in the middle of an ordered set of numbers.

Example 1: 18 27 34 52 54 59 61 68 78 82 85 87 91 93 100, Median = 68
Example 2: 18 27 27 34 52 52 59 61 68 68 85 85 85 90, Median = 60

The mean (or average) μ = average of all values in a column

It can be considered as a model because it summaries the data.
- Example: number of friends of each member of a group of 5 lecturers: 1, 2, 3, 3 and 4
Mean: (1+2+3+3+4)/5 = 2.6 friends per lecturer: clearly a hypothetical value!
Now if the values were: 1, 1, 1, 1 and 9 the mean would also be 2.6 but clearly it would not give an accurate picture of the data. So, how can we know that it is an accurate model? We look at the difference between the real data and our model. To do so, we calculate the difference between the real data and the model created and we make the sum so that we get the total error (or sum of differences).



$\sum(x_i - \mu) = (-1.6) + (-0.6) + (0.4) + (0.4) + (1.4) = 0$ And we get no errors !

Of course: positive and negative differences cancel each other out. So to avoid the problem of the direction of the error, we can square the differences and instead of sum of errors, we get the Sum of Squared errors (SS).
- In our example: SS = $(-1.6)^2 + (-0.6)^2 + (0.4)^2 + (0.4)^2 + (1.4)^2 = 5.20$

The variance

This SS gives a good measure of the accuracy of the model but it is dependent upon the amount of data: the more data, the higher the SS. The solution is to divide the SS by the number of observations (N). As we are interested in measuring the error in the sample to estimate the one in the population, we divide the SS by N-1 instead of N and we get the *variance* (S^2) = SS/N-1

- In our example: Variance (S^2) = 5.20 / 4 = 1.3

Why N-1 instead N?

If we take a sample of 4 scores in a population they are free to vary but if we use this sample to calculate the variance, we have to use the mean of the sample as an estimate of the mean of the population. To do that we have to hold one parameter constant.

- Example: mean of the sample is 10

We assume that the mean of the population from which the sample has been collected is also 10. If we want to calculate the variance, we must keep this value constant which means that the 4 scores cannot vary freely:

> - If the values are 9, 8, 11 and 12 (mean = 10) and if we change 3 of these values to 7, 15 and 8 then the final value must be 10 to keep the mean constant.

If we hold 1 parameter constant, we have to use N-1 instead of N. It is the idea behind the *degree of freedom*: one less than the sample size.

The Standard Deviation (SD)

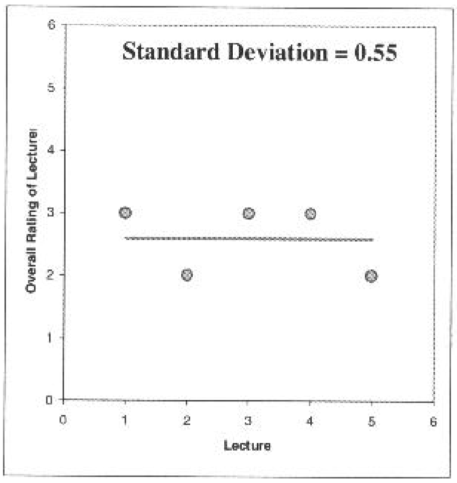
The problem with the variance is that it is measured in squared units which is not very nice to manipulate. So for more convenience, the square root of the variance is taken to obtain a measure in the same unit as the original measure: the *standard deviation.*

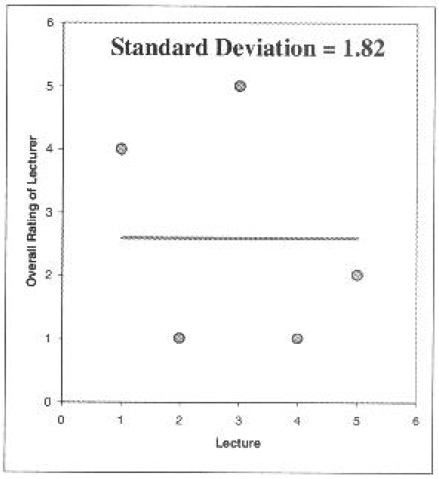
- S.D. = $\sqrt{(SS/N-1)}$ = $\sqrt{(S^2)}$, in our example: S.D. = $\sqrt{(1.3)}$ = 1.14

So you would present your mean as follows: μ = 2.6 +/- 1.14 friends.

The standard deviation is a measure of how well the mean represents the data or how much our data are scattered around the mean.

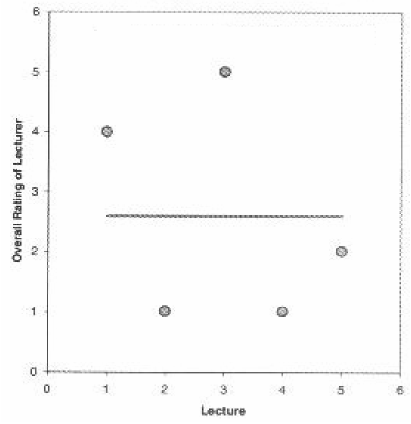
- small S.D.: data close to the mean: mean is a good fit of the data (graph on the left)
- large S.D.: data distant from the mean: mean is not an accurate representation (graph on the right)

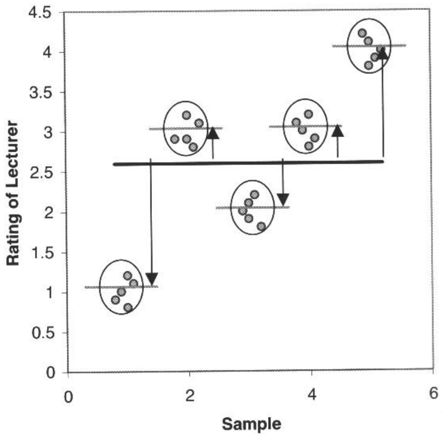




Standard Deviation vs. Standard Error

Many scientists are confused about the difference between the standard deviation (S.D.) and the *standard error of the mean* (S.E.M. = S.D. / √N).
- The S.D. (graph on the left) quantifies the scatter of the data and increasing the size of the sample does not decrease the scatter (above a certain threshold).
- The S.E.M. (graph on the right) quantifies how accurately we know the true population mean, it's a measure of how much we expect sample means to vary. So the S.E.M. gets smaller as our samples get larger: the mean of a large sample is likely to be closer to the true mean than is the mean of a small sample.





A big S.E.M. means that there is a lot of variability between the means of different samples and that our sample might not be representative of the population.
A small S.E.M. means that most samples means are similar to the population mean and so our sample is likely to be an accurate representation of the population.

Which one to choose?

- If the scatter is caused by biological variability, it is important to show the variation. So it is more appropriate to report the S.D. rather than the S.E.M. Even better, we can show in a graph all data points, or perhaps report the largest and smallest value.
- If we are using an in vitro system with theoretically very little biological variability, the scatter can only result from experimental imprecision (no biological meaning). It is more sensible then to report the S.E.M. since the S.D. is less useful here. The S.E.M. gives the readers a sense of how well we have determined the mean.

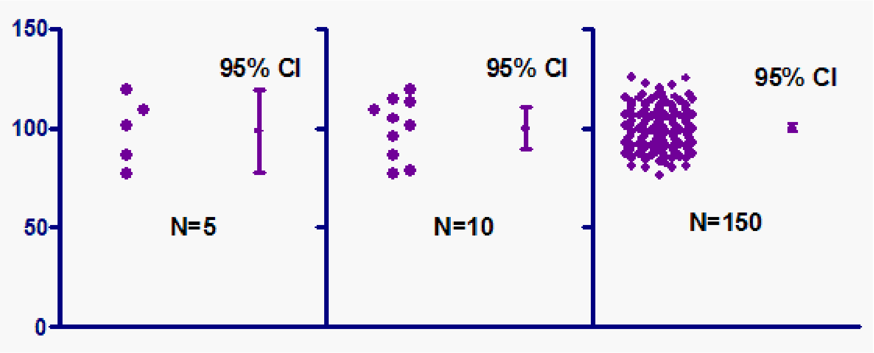
Choosing between SD and SEM also depends on what we want to show. If we just want to present our data on a descriptive purpose then we go for the SD. If we want the reader to be able to infer an idea of significance then you should go for the SEM or the Confidence Interval (see below). We will go into a bit more detail later.

Confidence interval

The confidence interval quantifies the uncertainty in measurement. The mean we calculate from our sample of data points depends on which values we happened to sample. Therefore, the mean we calculate is unlikely to equal the true population mean. The size of the likely discrepancy depends on the variability of the values and the sample size. If we combine those together, we can calculate a 95%

confidence interval (95% CI), which is a range of values. If the population is normal (or nearly so), there is a 95% chance that the confidence interval contains the true population mean.
95% of observations in a normal distribution lie within +/- 1.96*SE

One other way to look at error bars:

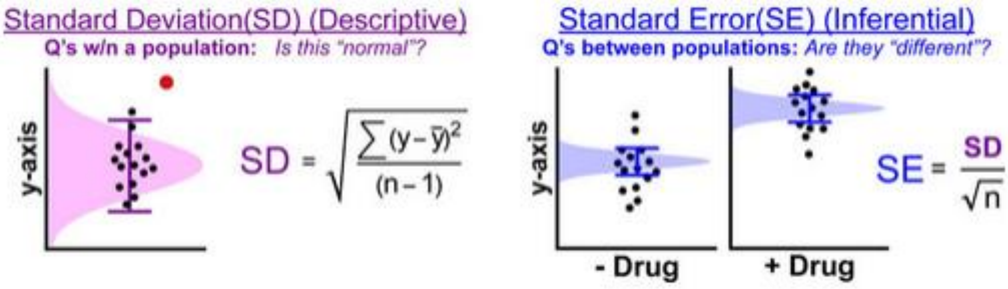


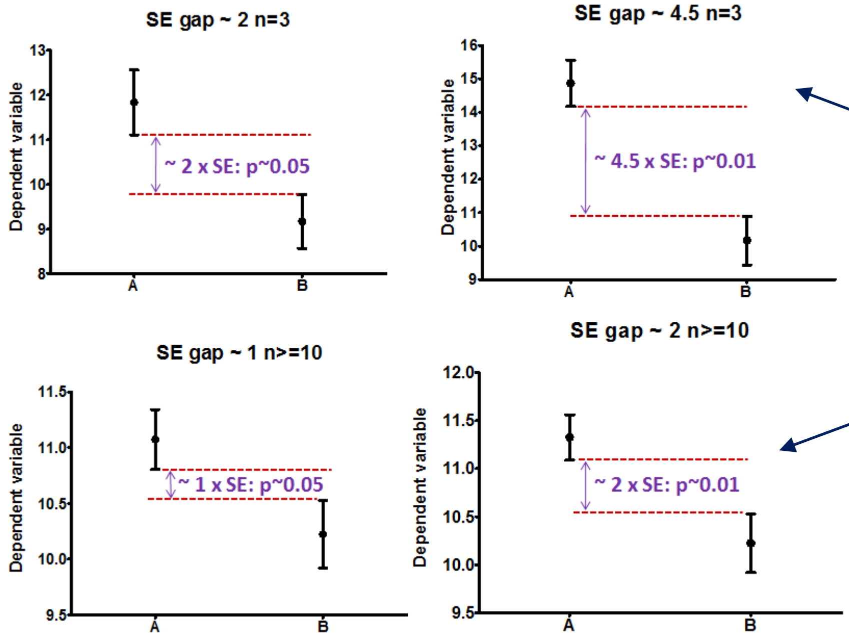
Error bars	Type	Description
Standard deviation (SD)	Descriptive	Typical or average difference between the data points and their mean.
Standard error (SEM)	Inferential	A measure of how variable the mean will be, if you repeat the whole study many times.
Confidence interval (CI), usually 95% CI	Inferential	A range of values you can be 95% confident contains the true mean.

From Geoff Cumming *et al*.

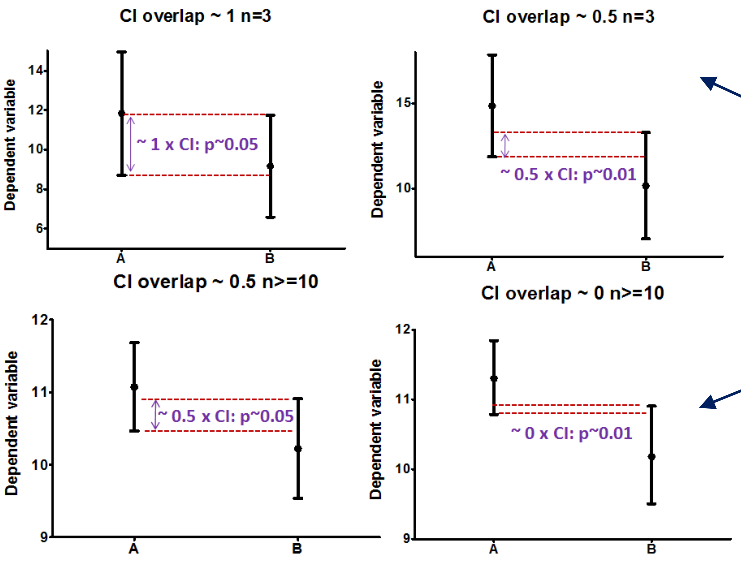
If we want to compare experimental results, it could be more appropriate to show inferential error bars such as SE or CI rather than SD. If we want to describe our sample, for instance its normality, then the SD would be the one to choose.

However, if n is very small (for example n=3), rather than showing error bars and statistics, it is better to simply plot the individual data points.


Standard Deviation(SD) (Descriptive)
Q's w/n a population: Is this "normal"?
y-axis
SD = $\sqrt{\frac{\sum(y-\bar{y})^2}{(n-1)}}$
Standard Error(SE) (Inferential)
Q's between populations: Are they "different"?
y-axis
- Drug
+ Drug
SE = $\frac{SD}{\sqrt{n}}$



We can estimate statistical significance using the overlap rule for SE bars.

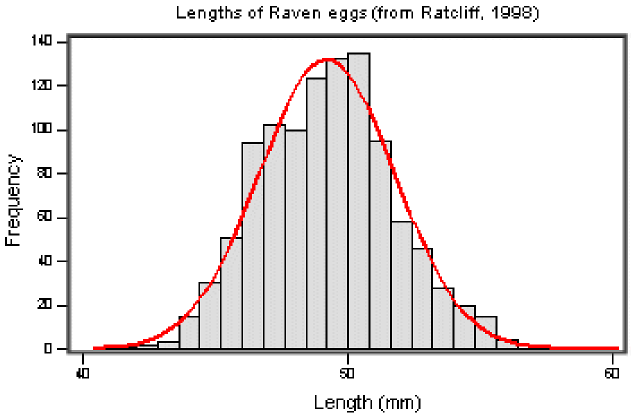


In the same way, you can estimate statistical significance using the overlap rule for 95% CI bars.

3-2 A bit of theory: Assumptions of parametric data

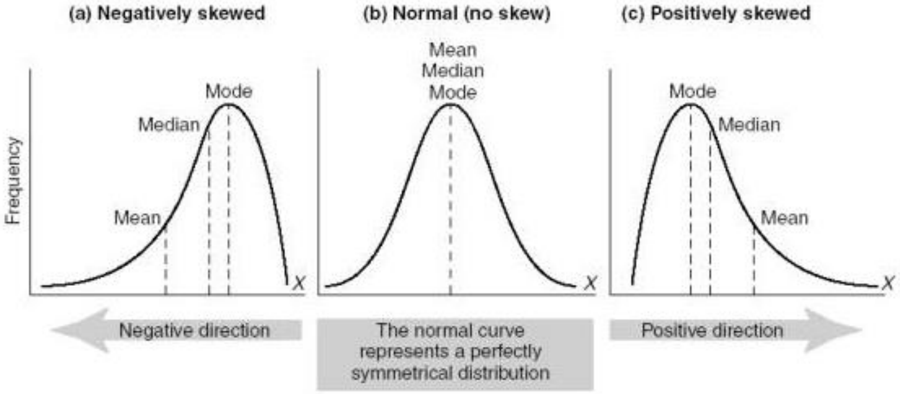
When we are dealing with quantitative data, the first thing we should look at is how they are distributed, what they look like. The distribution of our data will tell us if there is something wrong in the way we collected them or enter them and it will also tell we what kind of test we can apply to make them say something.
T-test, analysis of variance and correlation tests belong to the family of parametric tests and to be able to use them our data must comply with 4 assumptions.

1) The data have to be normally distributed (normal shape, bell shape, Gaussian shape).
Example of normally distributed data:



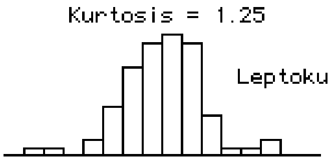
There are 2 main types of departure from normality:

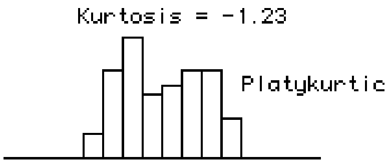
- Skewness: lack of symmetry of a distribution

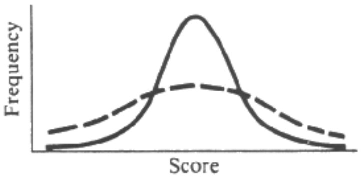


- Kurtosis: measure of the degree of 'peakedness' in the distribution

The two distributions below have the same variance approximately the same skew, but differ markedly in kurtosis.







(e) Platykurtic and leptokurtic

2) Homogeneity in variance: The variance should not change systematically throughout the data.

3) Interval data: The distance between points of the scale should be equal at all parts along the scale

4) Independence: Data from different subjects are independent so that values corresponding to one subject do not influence the values corresponding to another subject. Basically, it means it means one measure per subject. There are specific designs for repeated measures experiments.

Chapter 4: Comparing 2 means

How can we check that our data are parametric/normal?

Let's try it through an example.

Example (File: `coyote.csv`) csv as in 'comma-separated values'

We want to know if male coyotes are bigger than female coyotes. Of course, before doing anything else, we design our experiment and we are told that to compare 2 samples we need to apply a t-test (we will explain this test later). So basically we are going to catch coyotes and hopefully we will manage to catch males and females. Now, the tricky bit here is how many coyotes do we need?

4-1 Power analysis with a t-test

Let's say, we don't have data from a pilot study but we have found some information in the literature. In a study run in similar conditions as in the one we intend to run, male coyotes (n=20) were found to measure on average: 92cm+/- 7cm (SD). We expect a 5% difference between genders with a similar variability in the female sample.

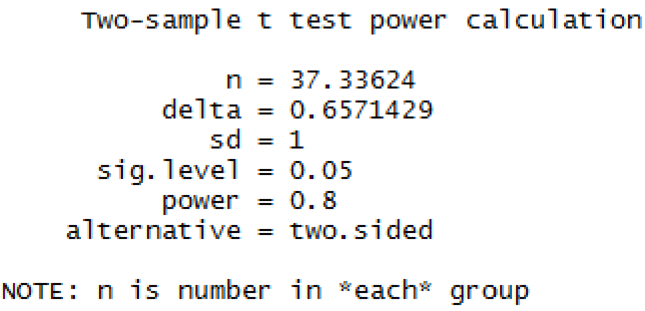
The function in this case will be:

```
power.t.test(n = NULL, delta = NULL, sd = 1, sig.level = NULL, power = NULL,
type = c("two.sample", "one.sample", "paired"),
alternative = c("two.sided", "one.sided"))

mean1<-92
mean2<-87.4 (5% less than 92cm)

delta<-92-87.4
sd<-7
```

So now:

```
power.t.test(delta=92-87.4, sd = 7, sig.level = 0.05, power = 0.8)
```

The default 'Type' is "two.sample" so no need to specify it.

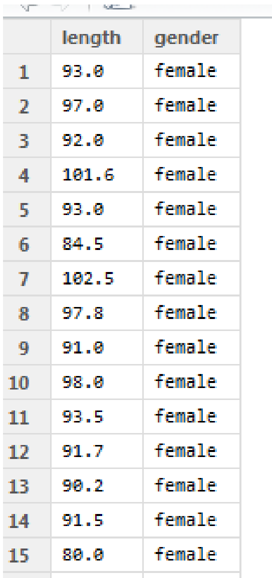


```
     Two-sample t test power calculation

              n = 37.33624
          delta = 0.6571429
             sd = 1
      sig.level = 0.05
          power = 0.8
    alternative = two.sided

NOTE: n is number in *each* group
```

We need a sample size of n=76 (2*38).

Once the data are collected, we check for normality.

Though normality tests are good, the best way to get a really good idea of what is going on is to plot our data. When it comes to normality, there are 3 ways to plot our data: the box plot, the scatter plot and the histogram. We are going to do them all with R.



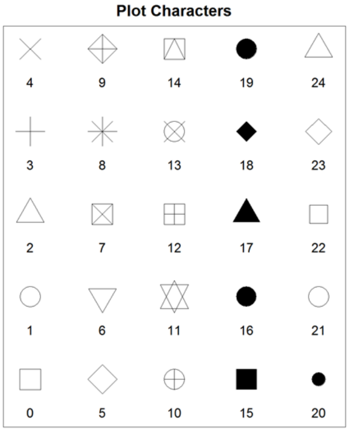
	length	gender
1	93.0	female
2	97.0	female
3	92.0	female
4	101.6	female
5	93.0	female
6	84.5	female
7	102.5	female
8	97.8	female
9	91.0	female
10	98.0	female
11	93.5	female
12	91.7	female
13	90.2	female
14	91.5	female
15	80.0	female

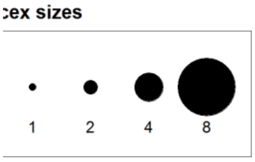
To get the data in R we go:

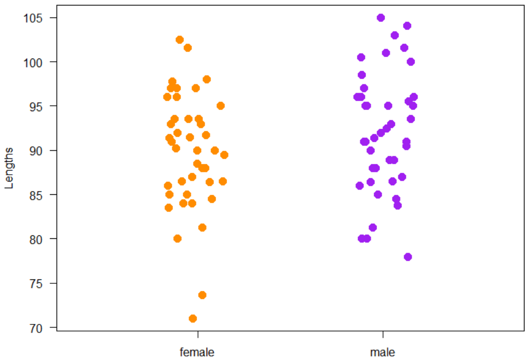
```
coyote<-read.csv("coyote.csv",header=TRUE)
View(coyote) or head(coyote)
```

Now let's start with the stripchart.

```
stripchart(coyote$length~coyote$gender,vertical=TRUE,
method="jitter", las=1, ylab="Lengths",pch=16,
col=c("darkorange","purple"), cex=1.5, at=c(1.2,1.8))
```

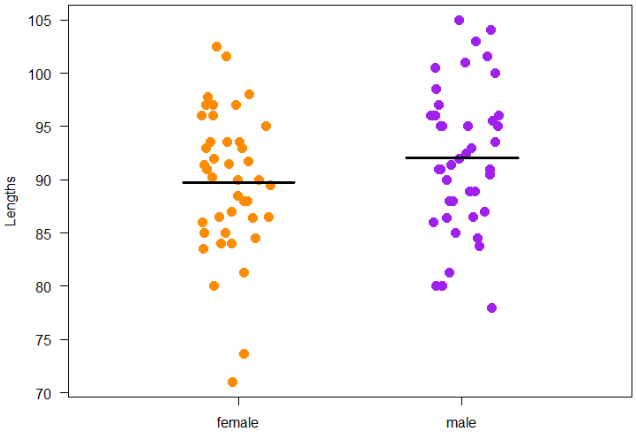






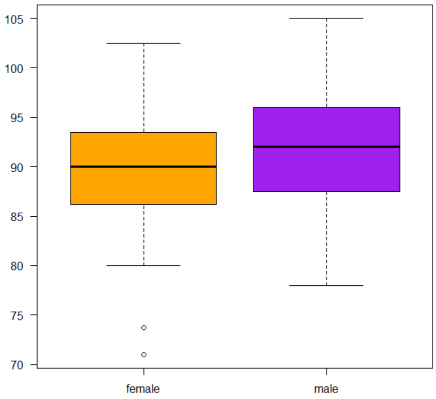
Now you may want to improve this graph by adding the groups'means.

```
length.means <- tapply(coyote$length,coyote$gender,mean)
loc.strip<- c(1.2,1.8)
segments(loc.strip-0.15,length.means, loc.strip+0.15, length.means,
col="black", lwd=3)
```



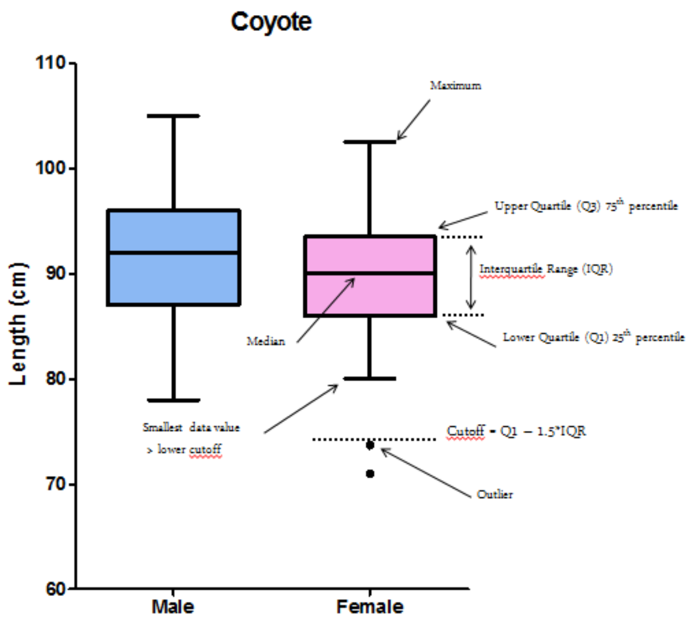
One other way to explore the data is the boxplot.

```
boxplot(coyote$length~coyote$gender, col=c("orange","purple"))
```



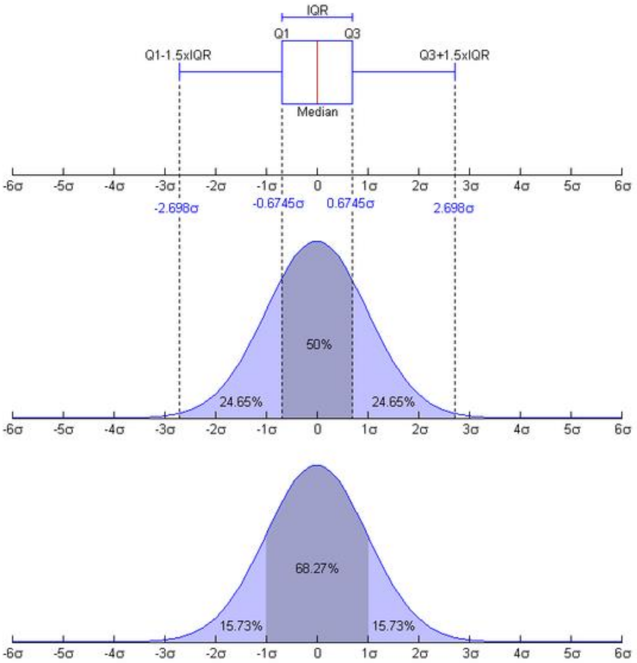
The anatomy of a boxplot is explained in the graph below. It is very important that you know how a box plot is built. It is rather simple and it will allow us to get a pretty good idea about the distribution of your data at a glance.

If the distribution is normal-ish then the box plot should be symmetrical-ish and if both (like in our case) are of the same size-ish, then we can be confident that the variances are about the same.

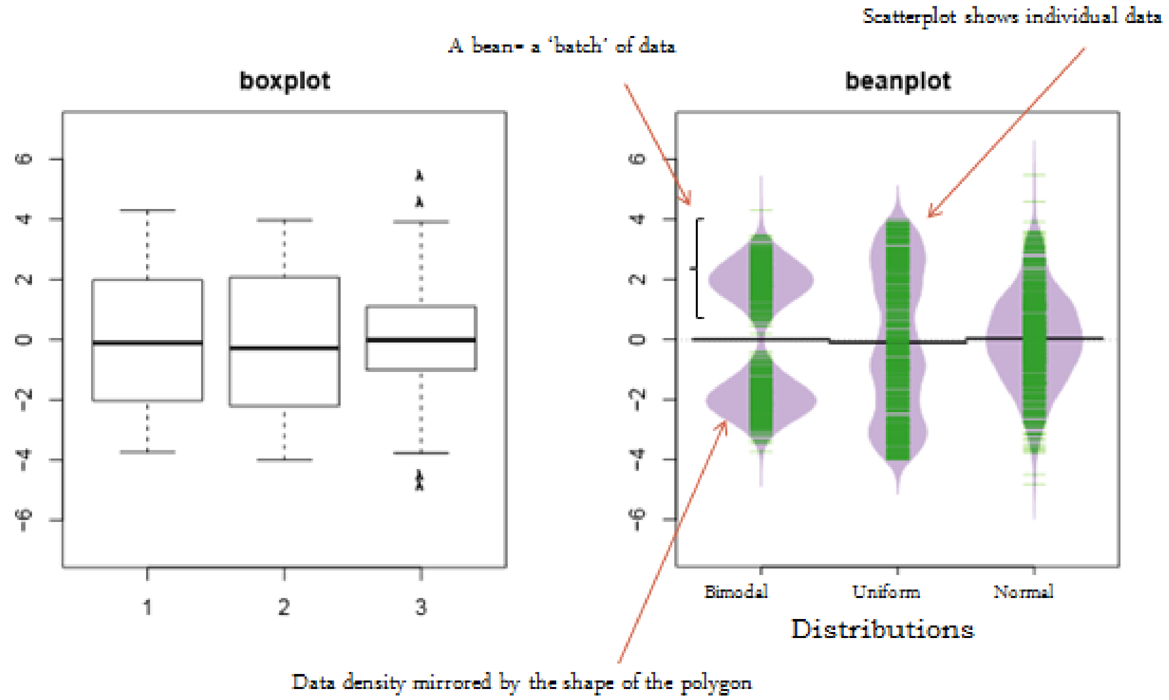


Regarding the outliers, there is really no right or wrong attitude. If there is a technical issue or an experimental problem, you should remove it, of course, but if there is nothing obvious it is up to you. I would always recommend keeping outliers if we can; we can run the analysis with and without it for instance and see what effect it has on the p-value. If the outcome is still consistent with our hypothesis, then we should keep it. If not, then it is between you and your conscience!

Below we can see the relationship between box plot and histogram.

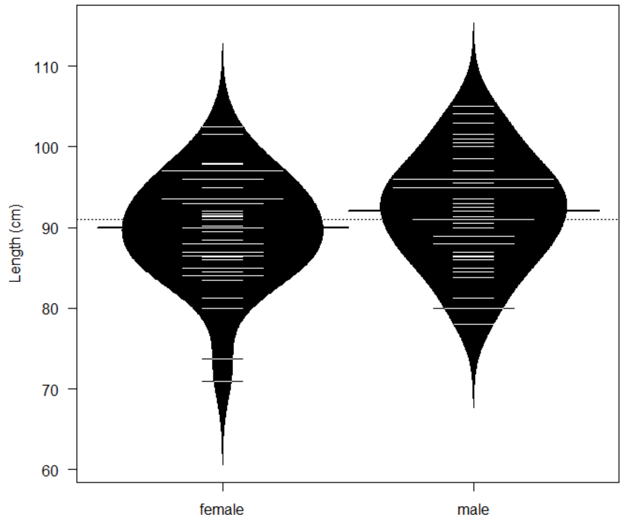


Beanplots can be more informative than boxplot in terms of ‘hidden’ distribution especially with big datasets but they do not identify outliers.



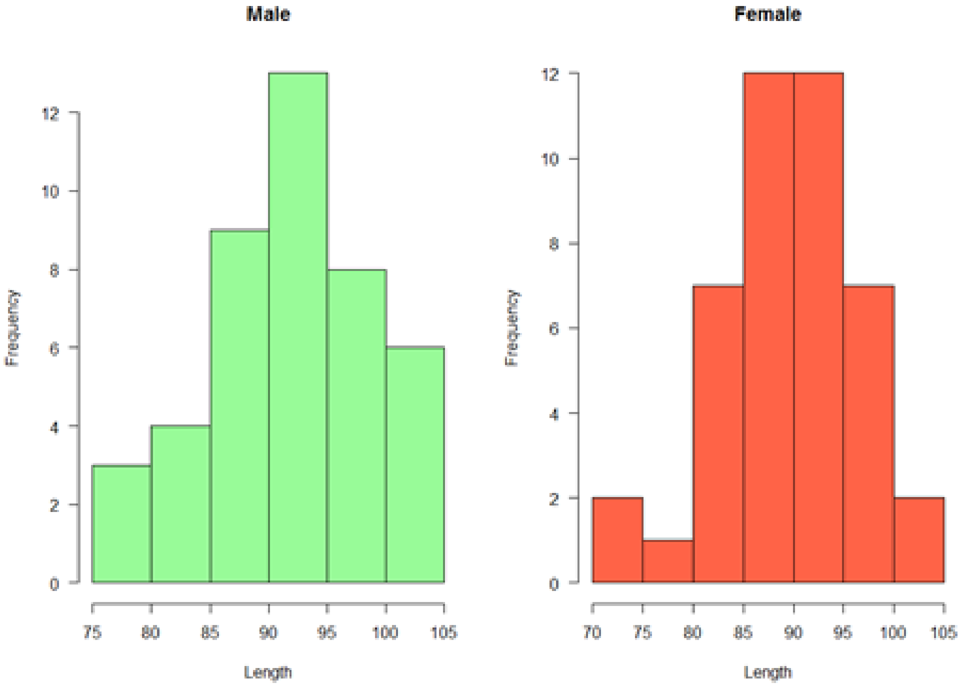
Beanplots look quite cool!

```
beanplot(coyote$length~coyote$gender, las=1,
ylab="Length (cm)")  ## beanplot package ##
```



Finally, the histograms.

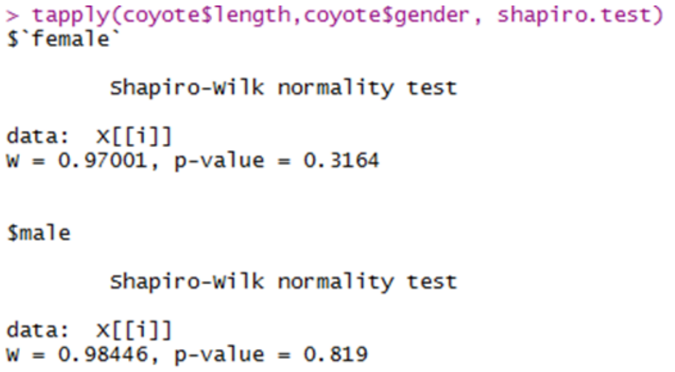
```
par(mfrow=c(1,2))
hist(coyote[coyote$gender=="male",]$length,main="Male",
xlab="Length",col="lightgreen")
hist(coyote[coyote$gender=="female",]$length, main="Female",
xlab="Length",col="tomato1")
```



So from the graphs we have plotted, we can tell that the first and second assumptions are likely to be met: the data seem normal enough (symmetry of the graphs) and the variability seems comparable between groups (spread of the data). My preference goes to the box plot as it tells you in one go anything you need to know: where you are with the first 2 assumptions and it shows you outliers.

Still we may come across cases where it is not that obvious so can ask R to test for the normality (Shapiro-Wilk test or D'agostino and Pearson tests) and homogeneity of variance (Levene test). Here we are going to use 2 new functions: `shapiro()` which gives access to a normality test (Shapiro-Wilk test) and `tapply()` which allows to do it for males and females separately in one go.

```
tapply(coyote$length,coyote$gender, shapiro.test)
```



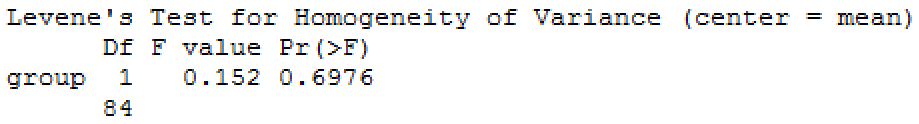
```
> tapply(coyote$length,coyote$gender, shapiro.test)
$`female`

        Shapiro-Wilk normality test

data:  X[[i]]
W = 0.97001, p-value = 0.3164


$male

        Shapiro-Wilk normality test

data:  X[[i]]
W = 0.98446, p-value = 0.819
```

There is no significant departure from normality for females (p=0.316) or males (p=0.819).

That was the first assumption; now we can check the second assumption (homogeneity of variances) using the Levene test.

Second assumption:

```
leveneTest(coyote$length, coyote$gender,centre=mean) ## car package ##
```



```
Levene's Test for Homogeneity of Variance (center = mean)
      Df F value Pr(>F)
group  1   0.152 0.6976
      84
```

So, good again but not surprising news: the variances of the 2 genders do not differ significantly (p=0.698).

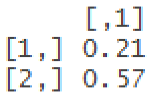
Don't be too quick to switch to nonparametric tests. While they do not assume Gaussian distributions, these tests do assume that the shape of the data distribution is the same in each group. So if your groups have very different standard deviations and so are not appropriate for a parametric test, they should not be analysed with its non-parametric equivalent either. However, parametric tests like ANOVA and t-tests are rather robust, especially when the samples are not too small so you can get away with small departure from normality and small differences in variances. Often the best approach is to transform the data and logarithms or reciprocals does the trick, restoring equal variance.

Finally, we may want to represent the data as a classical bar chart. To achieve that, we can type the lines below.

```
bar.length<-barplot(length.means,
      col=c("darkslategray1","darkseagreen1"),
      ylim=c(50,100),
      beside=TRUE,
      xlim=c(0,1),
      width=0.3,
      ylab="Mean length",las=1,
      xpd=FALSE,
      las=1)

## plotrix package ##
length.se<-tapply(coyote$length,coyote$gender,std.error)
```

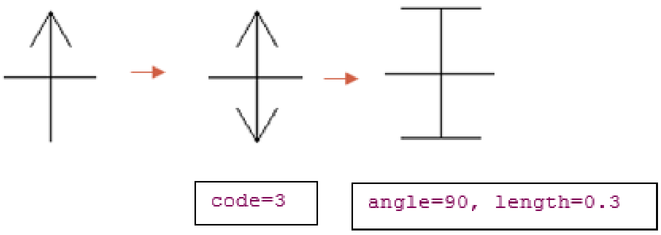
Now to plot the error bars, we are going to use `arrow(`. We need to specify the coordinates (x,y). `barplot ()` returns the values of the centre of the bars.

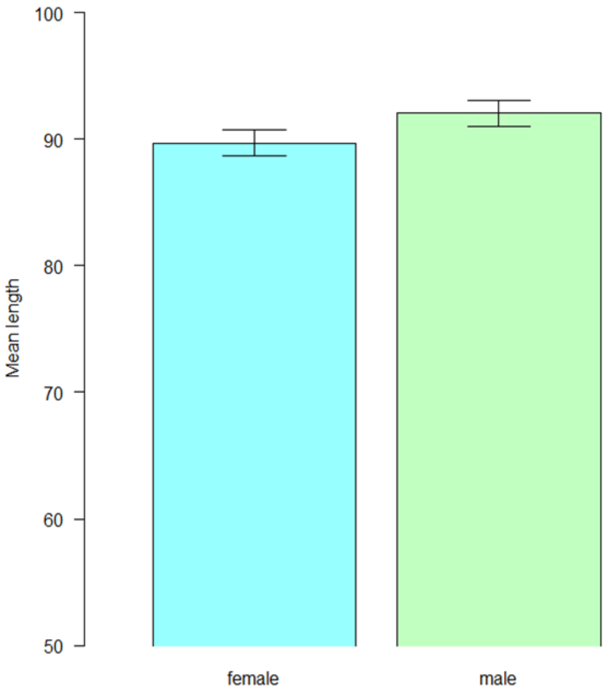
```
bar.length
```



```
     [,1]
[1,] 0.21
[2,] 0.57
```

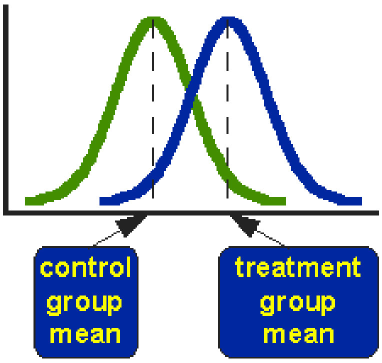
```
arrows(x0=bar.length,
      y0=length.means-length.se,
      x1=bar.length,
      y1=length.means+length.se,
      length=0.3,
      angle=90,
      code=3)
```





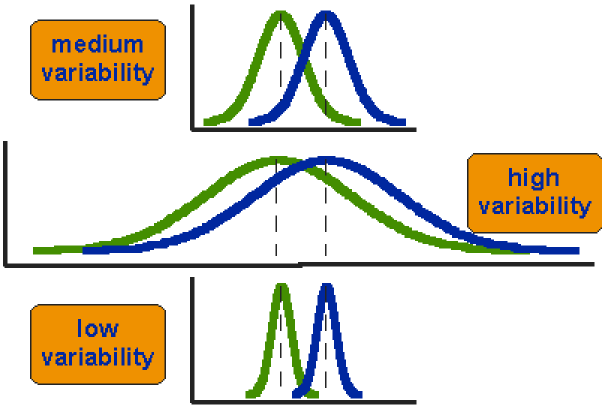
A bit of theory: the t-test

The t-test assesses whether the means of two groups are *statistically* different from each other. This analysis is appropriate whenever you want to compare the means of two groups.



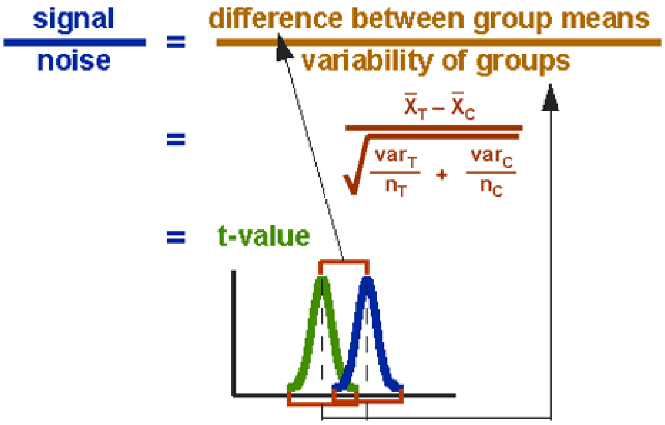
The figure above shows the distributions for the treated (blue) and control (green) groups in a study. Actually, the figure shows the idealized distribution. The figure indicates where the control and treatment group means are located. The question the t-test addresses is whether the means are statistically different.

What does it mean to say that the averages for two groups are statistically different? Consider the three situations shown in the figure below. The first thing to notice is that the difference between the means is the same in all three. But, we should also notice that the three situations don't look the same -- they tell very different stories. The top example shows a case with moderate variability of scores within each group. The second situation shows the high variability case. The third shows the case with low variability. Clearly, we would conclude that the two groups appear most different or distinct in the bottom or low-variability case. Why? Because there is relatively little overlap between the two bell-shaped curves. In the high variability case, the group difference appears least striking because the two bell-shaped distributions overlap so much.



This leads us to a very important conclusion: when we are looking at the differences between scores for two groups, we have to judge the difference between their means relative to the spread or variability of their scores. The t-test does just that.

The formula for the t-test is a ratio. The top part of the ratio is just the difference between the two means or averages. The bottom part is a measure of the variability or dispersion of the scores. We can see below the formula for the t-test and how the numerator and denominator are related to the distributions.


signal
noise
=
difference between group means
variability of groups
=
$\dfrac{\bar{X}_T - \bar{X}_C}{\sqrt{\dfrac{var_T}{n_T} + \dfrac{var_C}{n_C}}}$
=
t-value

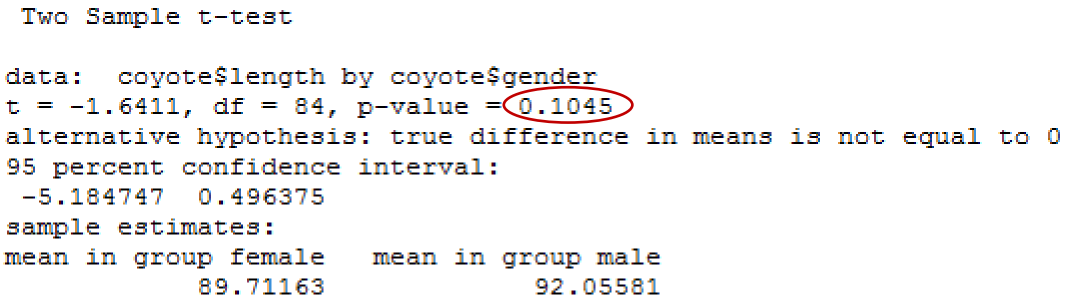
The t-value will be positive if the first mean is larger than the second and negative if it is smaller.

There are 2 types of t-test: Independent and Paired. The choice between the 2 is very intuitive. If we measure a variable in 2 **different populations**, we choose the independent t-test as the 2 populations are independent from each other. If we measure a variable 2 times in the **same population**, we go for the paired t-test.

So, say we want to compare the weights of 2 breeds of sheep. To do so, we take a sample of each breed (the 2 samples have to be comparable) and we weigh each animal. We then run an Independent-samples t-test on our data to find out if the difference is significant.
We may also want to test the effect of a diet on the level of a particular molecule in sheep's blood: to do so we choose one sample of sheep and we take a blood sample at day 1 and another one say at day 30. This time we would apply a Paired-Samples t-test as we are interested in each individual difference between day 1 and day 30.
Now, we want to compare the body length between males and females in coyotes so we are going to go for an Independent-test.

Independent t-test

```
t.test(coyote$length~coyote$gender, var.equal=T)
```



```
  Two Sample t-test

data:  coyote$length by coyote$gender
t = -1.6411, df = 84, p-value = 0.1045
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -5.184747  0.496375
sample estimates:
mean in group female   mean in group male 
            89.71163             92.05581 
```

So males coyotes are bigger than females but not significantly so (p=0.1045).

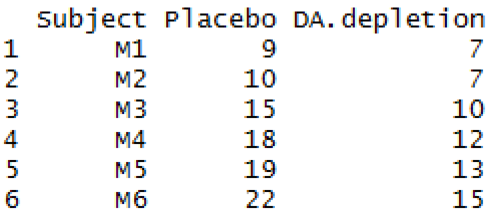
You might have noticed that despite having collected the 'recommended' sample size, we did not reach significance. This is because the difference observed in the collected sample is smaller than expected. If we now consider the data as the one of a pilot study and run the power analysis again, using the actual 2 means and we would need a sample 3 times bigger to reach a power of 80%. But of course, one should wonder whether a 2.3 cm difference (<3%) is biologically meaningful.

Paired t-test

Now let's try a Paired t-test. As we mentioned before, the idea behind the paired t-test is to look at a difference between 2 paired individuals or 2 measures for a same individual. For the test to be significant, the difference must be different from 0.
A researcher studying the effects of dopamine (DA) depletion on working memory in rhesus monkeys, tested working memory performance in 15 monkeys after administration of a saline (placebo) injection and again after injecting a dopamine-depleting agent.

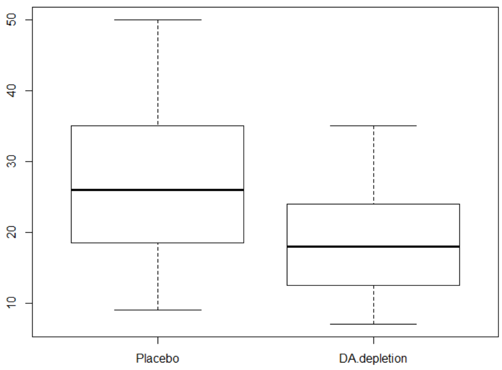
Example (File: `working.memory.csv`)

```
working.memory<-read.csv("working.memory.csv", header=T)
head(working.memory)
```

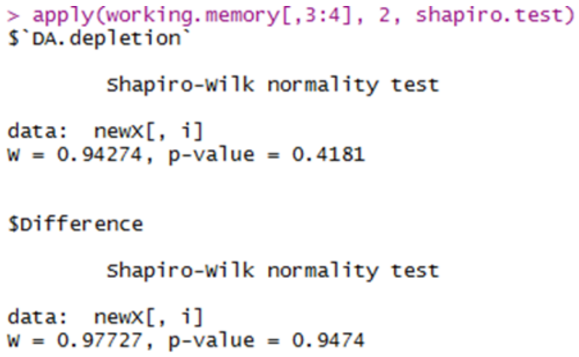


```
  Subject Placebo DA.depletion
1      M1       9            7
2      M2      10            7
3      M3      15           10
4      M4      18           12
5      M5      19           13
6      M6      22           15
```

```
boxplot(working.memory[2:3])
```



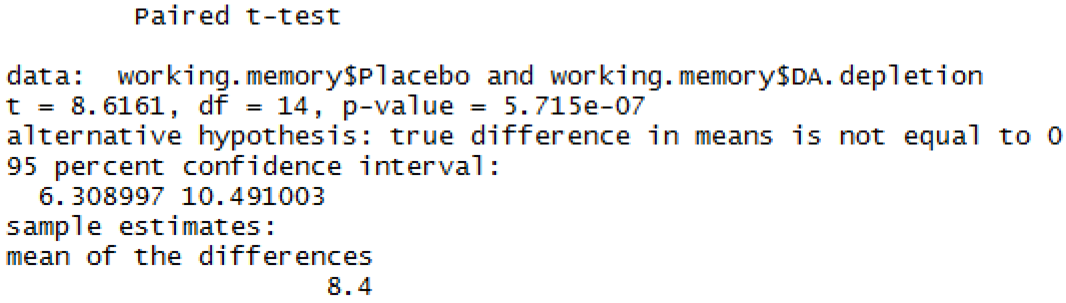
From the graph above, we can conclude that if placebo scores are higher than DA depletion ones, the difference does not seem very big. Before running the paired t-test to get a p-value we are going to check that the assumptions for parametric stats are met. The graphs seem to indicate that there is no significant departure from normality and this is confirmed by the Shapiro-Wilk test.



```
> apply(working.memory[,3:4], 2, shapiro.test)
$`DA.depletion`

        Shapiro-Wilk normality test

data:  newX[, i]
W = 0.94274, p-value = 0.4181


$Difference

        Shapiro-Wilk normality test

data:  newX[, i]
W = 0.97727, p-value = 0.9474
```

Let's run the paired t-test:

```
t.test(working.memory$Placebo, working.memory$DA.depletion, paired=T)
```



```
        Paired t-test

data:  working.memory$Placebo and working.memory$DA.depletion
t = 8.6161, df = 14, p-value = 5.715e-07
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  6.308997 10.491003
sample estimates:
mean of the differences
                    8.4
```

You will have noticed that we did not run a test for the equality of variances in the paired t-test; this is because it is actually looking at only one sample: the difference between the 2 conditions.

The paired t-test turns out to be highly significant (see Table above). So, how come the graph does not look more exciting?
The problem is that we don't really want to compare the mean scores for both groups, we want to look at the difference pair-wise, in other words we want to know if, on average, a given monkey who has received the drug performs less than when he was given only the drug. So we are interested in the mean difference between the 2 treatments.

We need to plot the difference. To do so, we need to create a new column containing the differences.

```
working.memory$Difference<- working.memory$Placebo-
working.memory$DA.depletion
head(working.memory)
```

We can check the distribution of the differences.

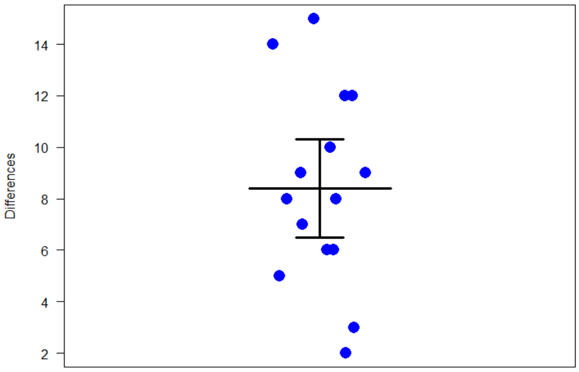
```
stat.desc(working.memory$Difference, basic=F, desc=F, norm=T)
```

And then plot the differences:

```
stripchart(working.memory$difference,
       vertical=TRUE,
       method="jitter",
       las=1,
       ylab="Differences",
       pch=16,
       col="blue",
       cex=2)

diff.mean <- mean(working.memory$difference)
centre<-1
segments(centre-0.15,diff.mean, centre+0.15, diff.mean, col="black", lwd=3)
```

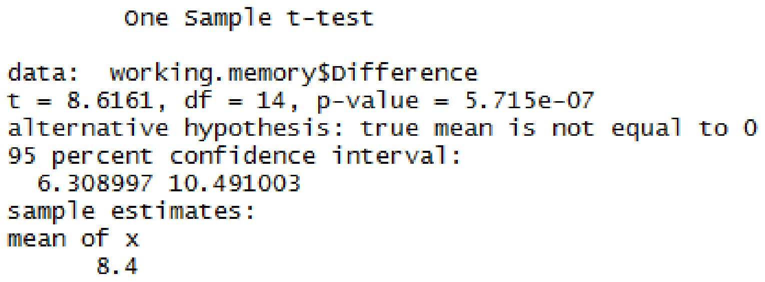
This time we can add error boars, like confidence interval to the graph. This is very informative as if 0 is well out the confidence interval limits (like here) it means the difference will be significant.

```
diff.se <- std.error(working.memory$difference) ## plotrix package ##
lower<-diff.mean-1.96*diff.se
upper<-diff.mean+1.96*diff.se

arrows(x0=loc.strip,
       y0=lower,
       x1=centre,
       y1=upper,
       length=0.3,
       code=3,
       angle=90,
       lwd=3)
```



We can test that the difference is significantly different from 0 using a one-sample t-test with 0 as a hypothetical value. The result will be exactly the same as the paired t-test as it is in effect looking at the same thing.

```
t.test(working.memory$difference ,mu=0)
```



```
        One Sample t-test

data:  working.memory$Difference
t = 8.6161, df = 14, p-value = 5.715e-07
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
  6.308997 10.491003
sample estimates:
mean of x 
      8.4 
```

Chapter 5: Comparing more than 2 means

5-1 Comparison of more than 2 means: One-way Analysis of variance

A bit of theory

When we want to compare more than 2 means (e.g. more than 2 groups), we cannot run several t-test because it increases the **familywise error rate** which is the error rate across tests conducted on the same experimental data.

To understand the following, it helps to remember one of the basic rules ("law") of probability: the Multiplicative Rule: The probability of the joint occurrence of 2 or more independent events is the product of the individual probabilities.

$$\boxed{P(A,B) = P(A) \times P(B)}$$

For example:

$$P(2\ \text{Heads}) = P(\text{head}) \times P(\text{head}) = 0.5 \times 0.5 = 0.25$$

Now, let's take an example: say we want to compare 3 groups (1, 2 and 3) and we carry out 3 t-tests (groups 1-2, 1-3 and 2-3), each with an arbitrary 5% level of significance, the probability of not making the type I error is 95% (= 1 - 0.05). The 3 tests being independent, we can multiply the probabilities (multiplicative rule), so the overall probability of no type I errors is: 0.95 * 0.95 * 0.95 = 0.857. Which means that the probability of making at least one type I error (to say that there is a difference whereas there is not) is 1 - 0.857 = 0.143 or 14.3%. So the probability has increased from 5% to 14.3%. If we compare 5 groups instead of 3, the family wise error rate is 40% (= 1 - $(0.95)^n$)

To overcome the problem of multiple comparisons, we need to run an **Analysis of variance (ANOVA)** followed by **post-hoc tests**. Actually, there are many different ways to correct for multiple comparisons and different statisticians have designed corrections addressing different issues (e.g.: unbalanced design, heterogeneity of variance, liberal vs conservative). However, they all have **one thing in common**: the more tests, the higher the familywise error rate: the more stringent the correction.
Tukey, Bonferroni, Sidak and others went for the FamilyWise Error Rate (FWER) mentioned above while others like Benjamini-Hochberg chose the False Discovery Rate (FDR) approach.
In the former, as already mentioned, the stringency of the correction will be a direct function of the number of comparisons (α_{adjust} = 0.05/n comparisons). The problem with this approach is that it is quickly very conservative, leading to a loss of power (lots of false negative). With only 10 comparisons, the threshold for significance is down to 0.005 (0.05/10), so when running pairwise comparisons across 20.000 genes, the correction becomes over conservative.
One way to address this issue is to use the FDR approach which controls the expected proportion of "discoveries" (significant tests) that are false (false positive). This allows for a less stringent control of Type I Error than FWER procedures which control the probability of at least one Type I Error. It results in more power but at the cost of increased numbers of Type I Errors.
The difference between FWER and FDR is that, with the former, a p-value of 0.05 implies that 5% of all tests will result in false positives whereas a FDR adjusted p-value (or **q-value**) of 0.05 implies that 5% of significant tests will result in false positives.

The ANOVA is an extension of the 2 groups comparison of a t-test but with a slightly different logic. If we want to compare 5 means, for example, we can compare each mean with another, which gives you 10 possible 2-group comparisons, which is quite complicated! So, the logic of the t-test cannot be directly transferred to the analysis of variance. Instead the ANOVA compares variances: if the variance amongst the 5 means is greater than the random error variance (due to individual variability for instance), then the means must be more spread out than we would have explained by chance.

The statistic for ANOVA is the F ratio:

$$F = \frac{\text{variance among sample means}}{\text{variance within samples (=random. Individual variability)}}$$

also:

$$F = \frac{\text{variation explained by the model (systematic)}}{\text{variation explained by unsystematic factors}}$$

If the variance amongst sample means is greater than the error variance, then F>1. In an ANOVA, we test whether F is significantly higher than 1 or not.
Imagine we have a dataset of 78 data points, we make the hypothesis that these points in fact belong to 5 different groups (this is our hypothetical model). So we arrange the data into 5 groups and we run an ANOVA.

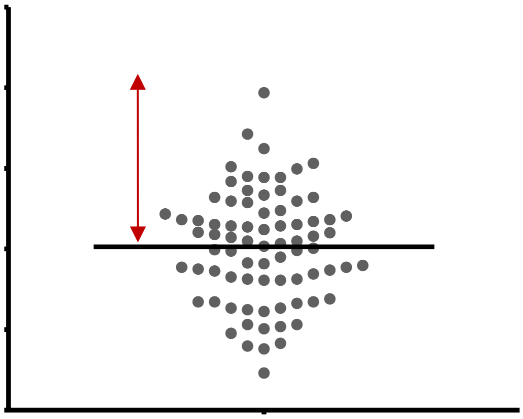
Below, is a typical example of analyse of variance table

Source of variation	**Sum of Squares**	**df**	**Mean Square**	**F**	**p-value**
Between Groups	2.665	4	0.6663	8.423	<0.0001
Within Groups	5.775	73	0.0791		
Total	8.44	77			

Let’s go through the figures in the table. First the bottom row of the table:

Total sum of squares = $\sum(x_i - \text{Grand mean})^2$

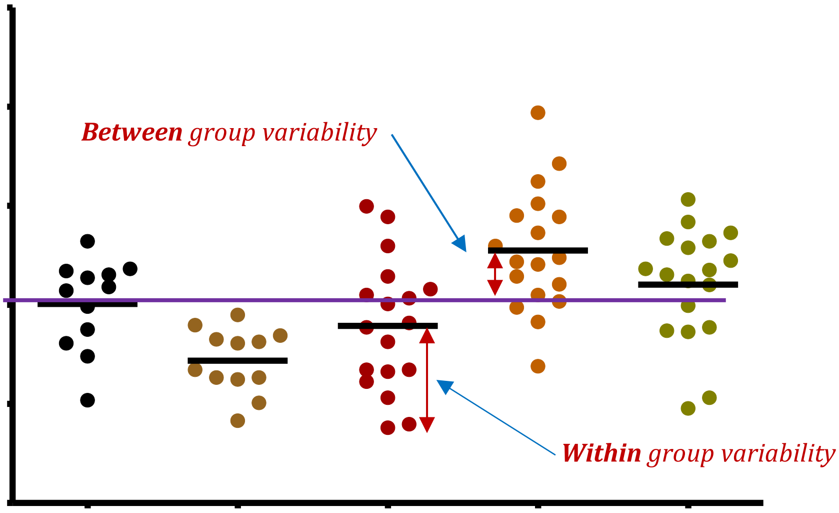
In our case, Total SS = 8.44. If we were to plot your data to represent the total SS, we would produce the graph below. So the total SS is the squared sum of all the differences between each data point and the grand mean. This is a quantification of the overall variability in our data. The next step is to partition this variability: how much variability between groups (explained by the model) and how much variability within groups (random/individual/remaining variability)?



According to our hypothesis our data can be split into 5 groups because, for instance, the data come from 5 cell types, like in the graph below.

So we work out the mean for each cell type and we work out the squared differences between each of the means and the grand mean ($\sum n_i$ (Mean$_i$ - Grand mean)^2). In our example (second row of the table): Between groups SS = 2.665 and, since we have 5 groups, there are 5 – 1 = 4 df, the mean SS = 2.665/4 = 0.6663.

If you remember the formula of the variance (= SS / N-1, with df=N-1), you can see that this value quantifies the variability between the groups' means: it is the between-groups variance.



There is one row left in the table, the within-groups variability. It is the variability within each of the five groups, so it corresponds to the difference between each data point and its respective group mean: Within groups sum of squares = $\sum (x_i - \text{Mean}_i)^2$ which in our case is equal to 5.775.

This value can also be obtained by doing 8.44-2.665 = 5.775, which is logical since it is the amount of variability left from the total variability after the variability explained by the model has been removed.

In our example the 5 groups sizes are 12, 12, 17, 17 and 17 so df = 5 x (n – 1) = 73

So the mean variability within groups: SS = 5.775/73 = 0.0791. This quantifies the remaining variability, the one not explained by the model, the individual variability between each value and the mean of the group to which it belongs according to the hypothesis. From this value can be obtained what is often referred to as the Pooled SD (=SQRT(MS(Residual or Within Group)). When obtained in a pilot study, this value is used in the power analysis.

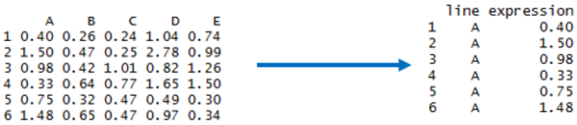
At this point, we can see that the amount of variability explained by our model (0.6663) is far higher than the remaining one (0.0791).
We can work out the F-ratio: F = 0.6663 / 0.0791 = 8.423

The level of significance of the test is calculated by taking into account the F ratio and the number of df (degree of freedom) for the numerator and the denominator. In our example, p<0.0001, so the test is highly significant and we are more than 99% confident when we say that there is a difference between the groups' means. This is an overall difference and even if we have an indication from the graph, we cannot tell which mean is significantly different from which.
This is because the ANOVA is an "omnibus" test: it tells us that there is (or not) an overall difference between our means but not exactly which means are significantly different from which. This is why we apply post-hoc tests. Post hoc tests could be compared to t-tests but with a more stringent approach, a lower significance threshold to correct for familywise error rate. We will go through post-hoc tests more in detail later.

Example (File: protein.expression.csv)

Let's do it! We want to find out if there is a significant difference in terms of protein expression between 5 cell types.

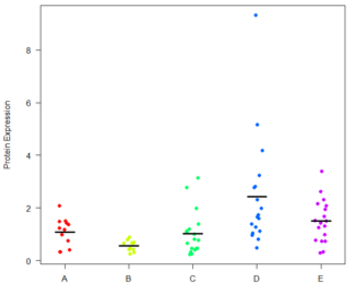
```
protein<-read.csv("protein.expression.csv",header=T)
protein.stack<-melt(protein) ## reshape2 package ##
colnames(protein.stack)<-c("line","expression")
protein.stack.clean <- na.omit(protein.stack
head(protein.stack.clean)
```

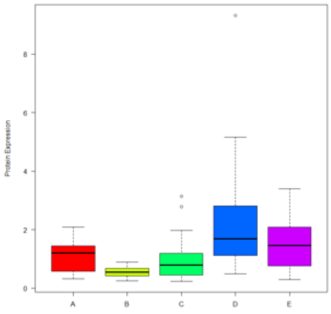


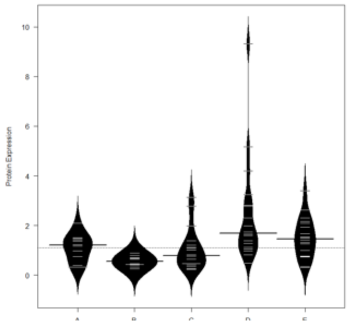
```
     A    B    C    D    E
1 0.40 0.26 0.24 1.04 0.74
2 1.50 0.47 0.25 2.78 0.99
3 0.98 0.42 1.01 0.82 1.26
4 0.33 0.64 0.77 1.65 1.50
5 0.75 0.32 0.47 0.49 0.30
6 1.48 0.65 0.47 0.97 0.34
```

```
  line expression
1    A       0.40
2    A       1.50
3    A       0.98
4    A       0.33
5    A       0.75
6    A       1.48
```

```
stripchart(protein.stack.clean$expression~protein.stack.clean$line,vertical=TRUE,
method="jitter", las=1, ylab="Protein Expression",pch=16,col=1:5)
expression.means<-tapply(protein.stack.clean$expression,protein.stack.clean$line,mean)
segments(1:5-0.15,expression.means, 1:5+0.15, expression.means, col="black", lwd=3)

boxplot(protein.stack.clean$expression~protein.stack.clean$line,col=rainbow(5),
ylab="Protein Expression",las=1)

beanplot(protein.stack.clean$expression~protein.stack.clean$line, log="",ylab="Protein Expression",las=1)
## beanplot package ##
```

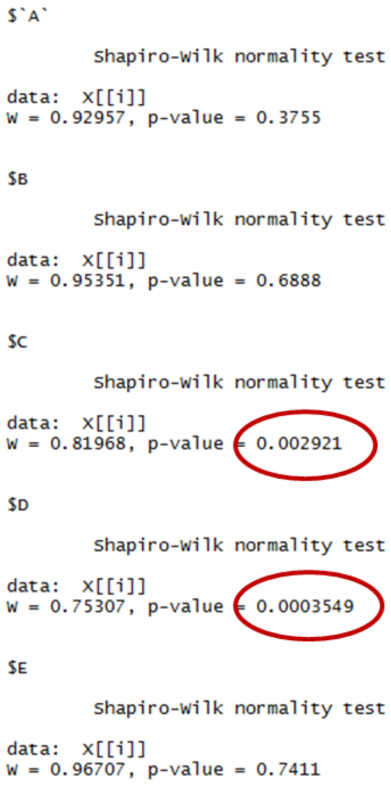






First we need to see whether the data meet the assumptions for a parametric approach. Well it does not look good: 2 out of 5 groups (C and D) show a significant departure from normality. As for the homogeneity of variance, even before testing it, a look at the scatter plots and box plots (see Graphs above) tells us that there is no way the second assumption is met. The data from groups C and D are quite skewed and a look at the raw data shows more than a 10-fold jump between values of the same group (e.g. in group A, value line 4 is 0.17 and value line 10 is 2.09).

```
tapply(protein.stack.clean$expression,protein.stack.clean$line, shapiro.test)
```



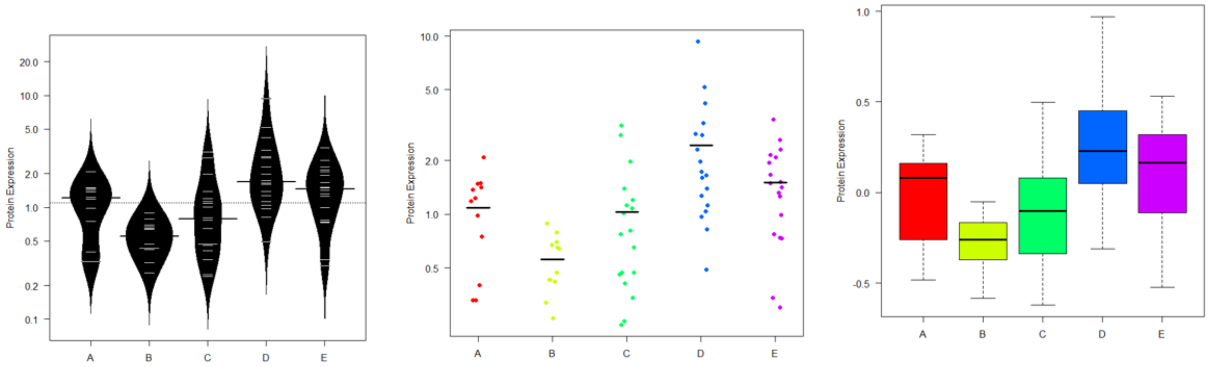
```
$`A`

        Shapiro-Wilk normality test

data:  X[[i]]
W = 0.92957, p-value = 0.3755


$B

        Shapiro-Wilk normality test

data:  X[[i]]
W = 0.95351, p-value = 0.6888


$C

        Shapiro-Wilk normality test

data:  X[[i]]
W = 0.81968, p-value = 0.002921


$D

        Shapiro-Wilk normality test

data:  X[[i]]
W = 0.75307, p-value = 0.0003549


$E

        Shapiro-Wilk normality test

data:  X[[i]]
W = 0.96707, p-value = 0.7411
```

A good idea would be to log-transform the data so that the spread is more balanced and to check again on the assumptions. The variability seems to be scale related: the higher the mean, the bigger the variability. This is a typical case for log-transformation.

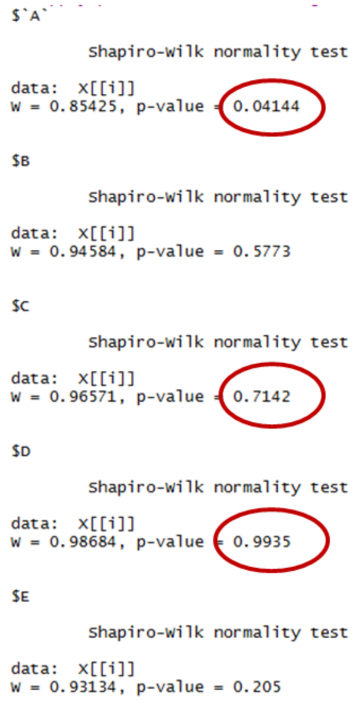
```
protein.stack.clean$log10.expression<-log10(protein.stack.clean$expression)
```

Speaking of log-transformation, `beanplot` has a built-in procedure to automatically determine whether a log transformation of the response axis is appropriate or not, to get rid of it, we need: `log=""`. In our case, since we are thinking log we might as well let the function choose.

```
beanplot(protein.stack.clean$expression~protein.stack.clean$line, ylab="Protein Expression", las=1)

stripchart(protein.stack.clean$expression~protein.stack.clean$line,vertical=TRUE,
method="jitter", las=1, ylab="Protein Expression",pch=16,col=rainbow(5),log="y")

expression.means<-tapply(protein.stack.clean$expression,protein.stack.clean$line,mean)
segments(1:5-0.15,expression.means, 1:5+0.15, expression.means, col="black", lwd=3)

boxplot(protein.stack.clean$log10.expression~protein.stack.clean$line,col=rainbow(5),
ylab="Protein Expression",las=1)
```



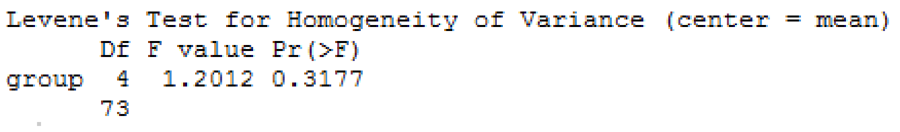
```
tapply(protein.stack.clean$log10.expression,protein.stack.clean$line,shapiro.test)
```



```
$`A`

        Shapiro-Wilk normality test

data:  x[[i]]
W = 0.85425, p-value = 0.04144

$B

        Shapiro-Wilk normality test

data:  x[[i]]
W = 0.94584, p-value = 0.5773

$C

        Shapiro-Wilk normality test

data:  x[[i]]
W = 0.96571, p-value = 0.7142

$D

        Shapiro-Wilk normality test

data:  x[[i]]
W = 0.98684, p-value = 0.9935

$E

        Shapiro-Wilk normality test

data:  x[[i]]
W = 0.93134, p-value = 0.205
```

One last thing before we run the ANOVA: we need to check for the second assumption: the homogeneity of variance. To do so, we do what we did before running the t-test: we run a Levene test:

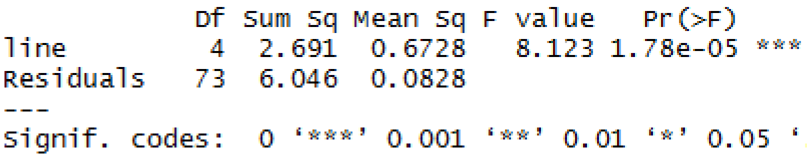
```
leveneTest(protein.stack.clean$log10.expression,protein.stack.clean$line)
## car package ##
```



```
Levene's Test for Homogeneity of Variance (center = mean)
      Df F value Pr(>F)
group  4  1.2012 0.3177
      73
```

Now that we have sorted out the data, we can run the ANOVA: to do so, you go:

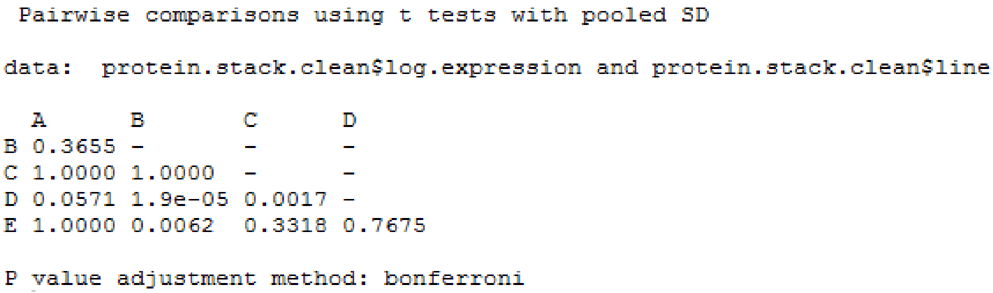
```
anova.log.protein<-aov(log10.expression~line,data=protein.stack.clean)
summary(anova.log.protein)
```



```
             Df Sum Sq Mean Sq F value   Pr(>F)    
line          4  2.691  0.6728   8.123 1.78e-05 ***
Residuals    73  6.046  0.0828                     
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The overall p-value is significant (p=1.78e-05) so the next thing to do is to choose a post-hoc test. There are 2 widely used: the Bonferroni test which is quite conservative so we should only choose it when we are comparing no more than 5 groups and the Tukey which is more liberal. First let’s try the Bonferroni test. It is built into R:

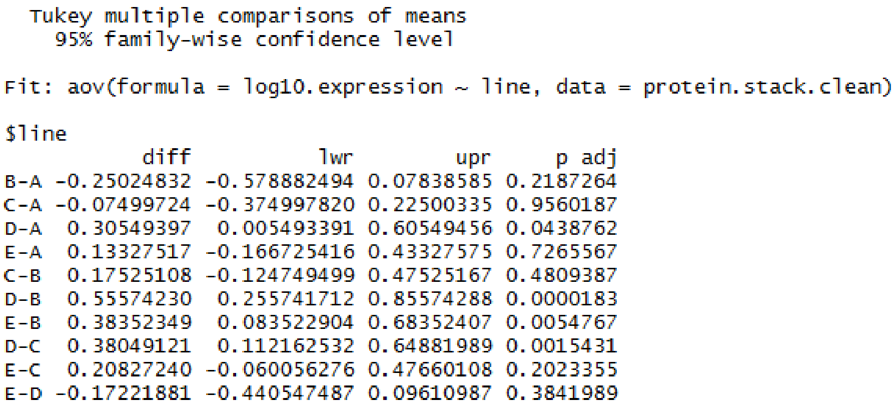
```
pairwise.t.test(protein.stack.clean$log10.expression,
protein.stack.clean$line, p.adj = "bonf")
```



```
        Pairwise comparisons using t tests with pooled SD 

data:  protein.stack.clean$log.expression and protein.stack.clean$line 

  A      B      C      D     
B 0.3655 -      -      -     
C 1.0000 1.0000 -      -     
D 0.0571 1.9e-05 0.0017 -     
E 1.0000 0.0062 0.3318 0.7675

P value adjustment method: bonferroni
```

Then Tukey:

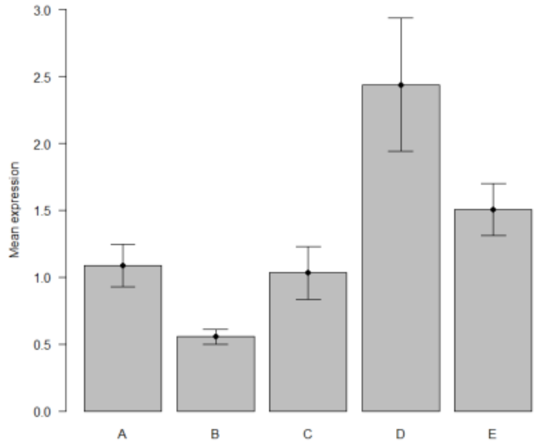
```
TukeyHSD(anova.log.protein,"line")
```



```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = log10.expression ~ line, data = protein.stack.clean)

$line
           diff          lwr        upr     p adj
B-A -0.25024832 -0.578882494 0.07838585 0.2187264
C-A -0.07499724 -0.374997820 0.22500335 0.9560187
D-A  0.30549397  0.005493391 0.60549456 0.0438762
E-A  0.13327517 -0.166725416 0.43327575 0.7265567
C-B  0.17525108 -0.124749499 0.47525167 0.4809387
D-B  0.55574230  0.255741712 0.85574288 0.0000183
E-B  0.38352349  0.083522904 0.68352407 0.0054767
D-C  0.38049121  0.112162532 0.64881989 0.0015431
E-C  0.20827240 -0.060056276 0.47660108 0.2023355
E-D -0.17221881 -0.440547487 0.09610987 0.3841989
```

Again, from the table above we can find out which pairwise comparison reaches significance and which does not.

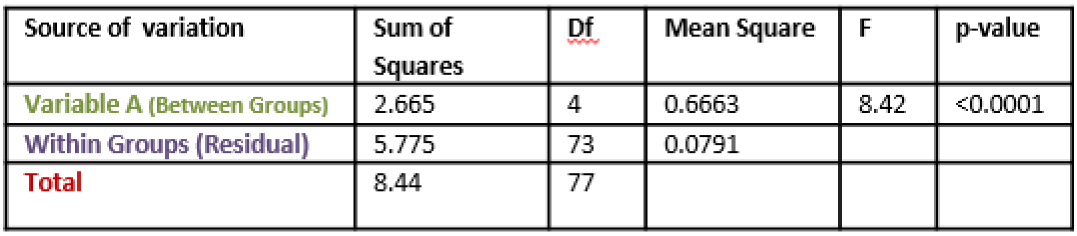
```
bar.expression<-barplot(expression.means, beside=TRUE, ylab="Mean
expression",ylim=c(0,3), las=1)
expression.se <- tapply(protein.stack.clean$expression,
protein.stack.clean$line,std.error)

arrows(x0= bar.expression, y0=expression.means-expression.se, x1=
bar.expression, y1=expression.means+expression.se, length=0.2,
angle=90,code=3)
```



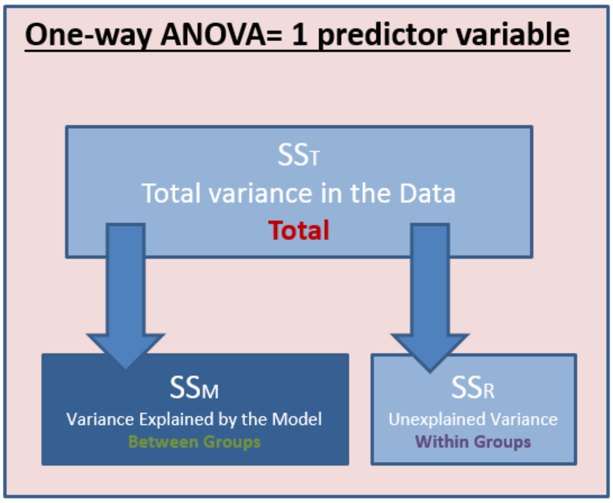
5-2 Two-way Analysis of Variance (File: goggles.dat)

So far, in the context of the t-test and the one-way ANOVA, we have considered the effect of a single independent variable or predictor on some outcome. Like gender on body length for coyotes or cell lines on protein expression. Sometimes, we may want to study the effect of more than one predictor on a given outcome. In that case, we will want to use a multiple factor analysis, sometimes also referred to as factorial ANOVA. In this section, we will see how to deal with 2 factors or 2 predictors and how to do a two-way ANOVA.

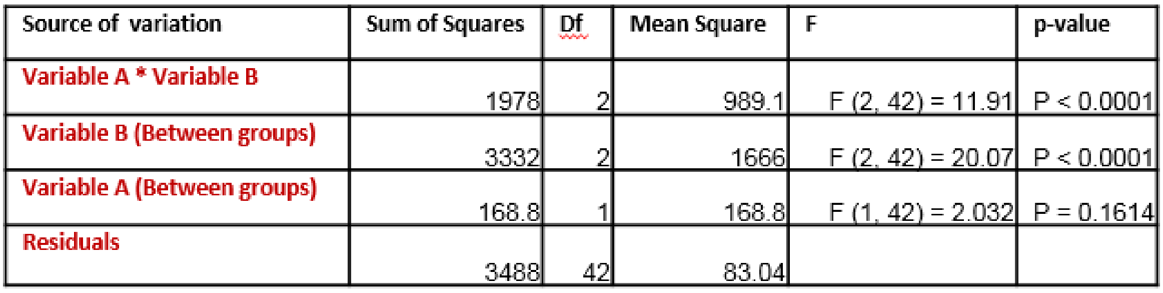
We saw in the previous chapter that running an ANOVA is all about partitioning the variance as seen below.



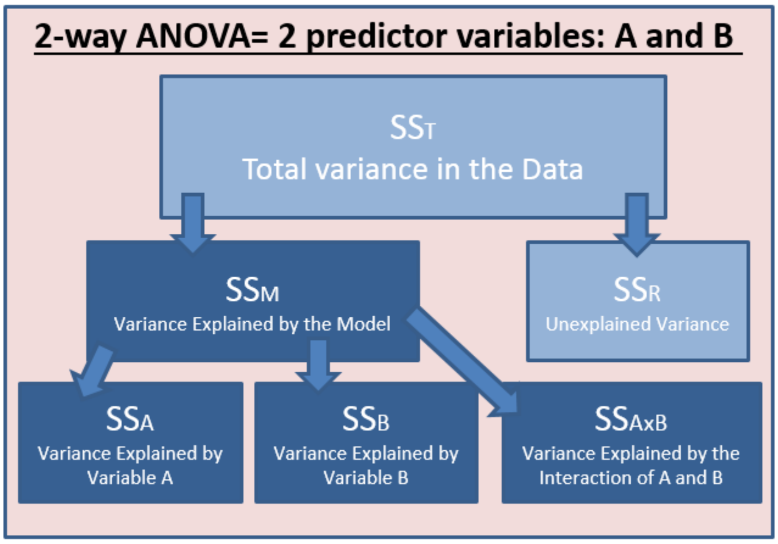
Source of variation	Sum of Squares	Df	Mean Square	F	p-value
Variable A (Between Groups)	2.665	4	0.6663	8.42	<0.0001
Within Groups (Residual)	5.775	73	0.0791		
Total	8.44	77			


One-way ANOVA= 1 predictor variable
SS_T
Total variance in the Data
Total
SS_M
Variance Explained by the Model
Between Groups
SS_R
Unexplained Variance
Within Groups

The logic is pretty much the same for a 2-way ANOVA as seen below.



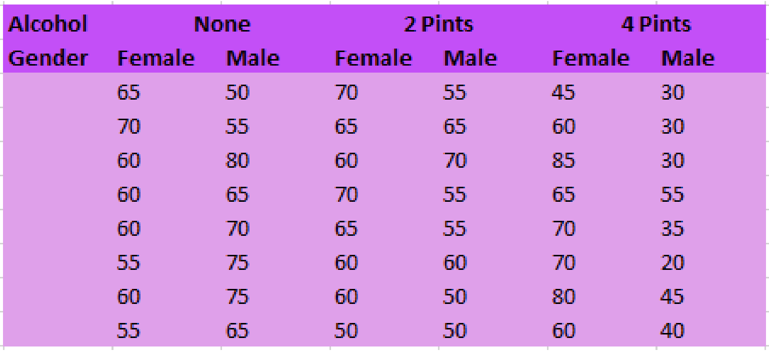
Source of variation	Sum of Squares	Df	Mean Square	F	p-value
Variable A * Variable B	1978	2	989.1	F (2, 42) = 11.91	P < 0.0001
Variable B (Between groups)	3332	2	1666	F (2, 42) = 20.07	P < 0.0001
Variable A (Between groups)	168.8	1	168.8	F (1, 42) = 2.032	P = 0.1614
Residuals	3488	42	83.04		


2-way ANOVA= 2 predictor variables: A and B
SS_T
Total variance in the Data
SS_M
Variance Explained by the Model
SS_R
Unexplained Variance
SS_A
Variance Explained by Variable A
SS_B
Variance Explained by Variable B
SS_{AxB}
Variance Explained by the Interaction of A and B

However, there is an extra layer of complexity with the interaction. Let's see how it works through an example.

Example (File: `goggles.csv`)

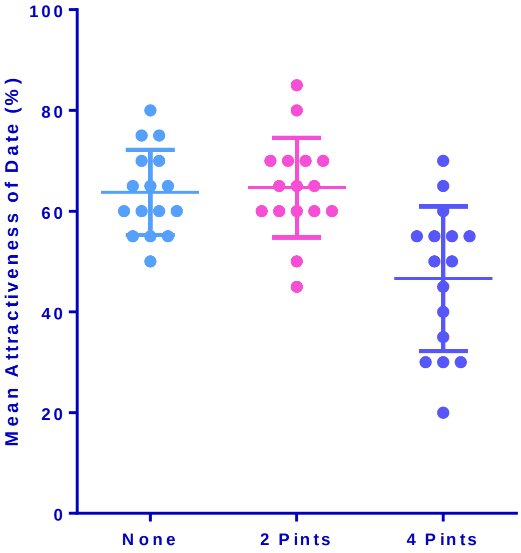
In the UK, there is something known as the beer-goggle effect: it is about subjective perception of physical attractiveness and how it become less accurate after alcohol is consumed. An anthropologist wanted to study the effects of alcohol, so in fact the beer-goggle effect, on mate selection at night-clubs. She was also interested in whether this effect was different for men and women. So she picked 48 students and ran an experiment with results presented below. The scores are the levels of attractiveness (out of 100) from a pool of independent judges of the person that the participant was chatting up at the end of the evening.



Alcohol	None		2 Pints		4 Pints	
Gender	Female	Male	Female	Male	Female	Male
	65	50	70	55	45	30
	70	55	65	65	60	30
	60	80	60	70	85	30
	60	65	70	55	65	55
	60	70	65	55	70	35
	55	75	60	60	70	20
	60	75	60	50	80	45
	55	65	50	50	60	40

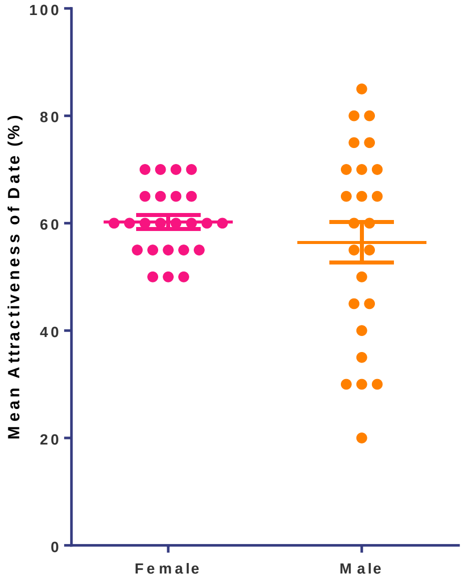
As always, the first thing to do is to explore the data. To understand the concept of interaction, the best way is to follow the logic of the 2-way ANOVAs: to look at the individual effects of the factors and then the interaction between the 2.

Main effect of alcohol:



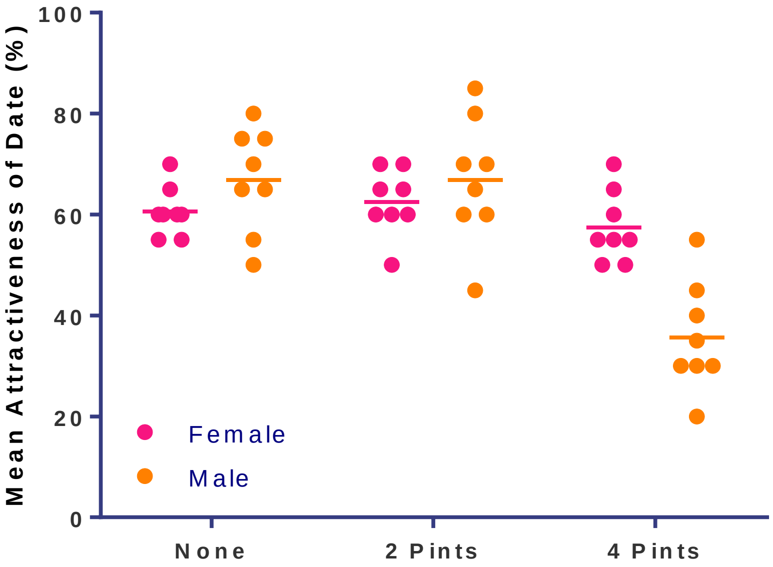
One can see that there is not much happening between None and 2 Pints, but after 4 Pints the level of attractiveness drops quite a bit. We can also notice that there does not seem to be anything too worrying about the data, in terms of distribution or homogeneity of variance.

Main effect of gender:



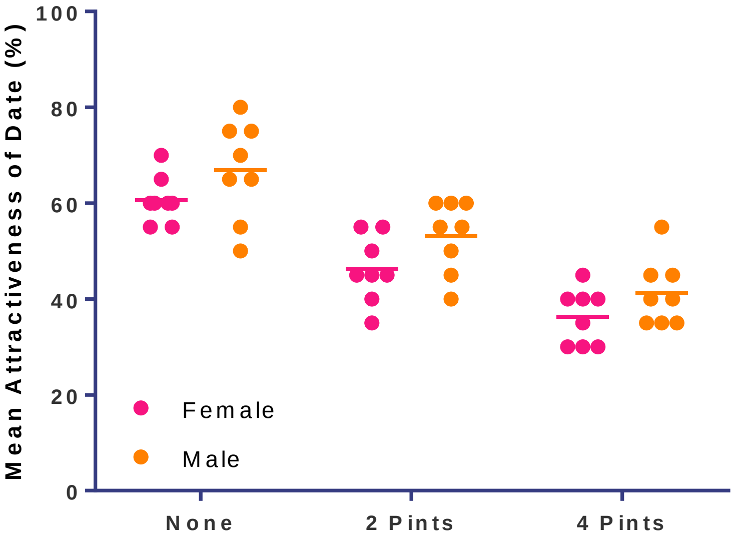
The gender effect does not appear very strong with only a slight decrease for the males. Variability appears higher in males than females.

Now the interaction:



The interaction is about the effect of one factor on the other. Or, put differently: is the effect of one factor the same on all levels of the other? And here the answer is no: males are slightly higher than females for the first 2 levels of alcohol but the gender effect is very different in the highest level of alcohol. This is what an interaction is about and looking at the graph, it is likely that this interaction will be significant. The concept of interaction becomes more intuitive when we try to formulate the answer to our original questions: is there an effect of alcohol consumption on the perception of physical attractiveness? The answer is: yes, but the effect is not the same for boys and girls. It is the 'yes but' that is about the interaction. Similarly, we could say: yes there is gender effect on the perception of physical attractiveness but the effect varies with the level of alcohol consumed.

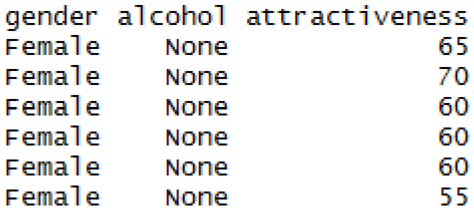
Below is a graph, on fake data, where there would not be an interaction between the 2 factors.



Now try to answer the question about alcohol, looking at the graph above: the answer is just ‘yes’, there is no ‘but’. For both genders, the attractiveness is affected by the consumption of alcohol, <u>in a similar way</u>. And to the question about gender, the answer is ‘yes’, too: levels of attractiveness are higher for males than females, <u>regardless the level of alcohol</u>. There is no interaction here and both factors are independent from one another.

Now let’s try to quantify these effects.

```
goggles<-read.csv("goggles.csv",header=T)
head(goggles)
```

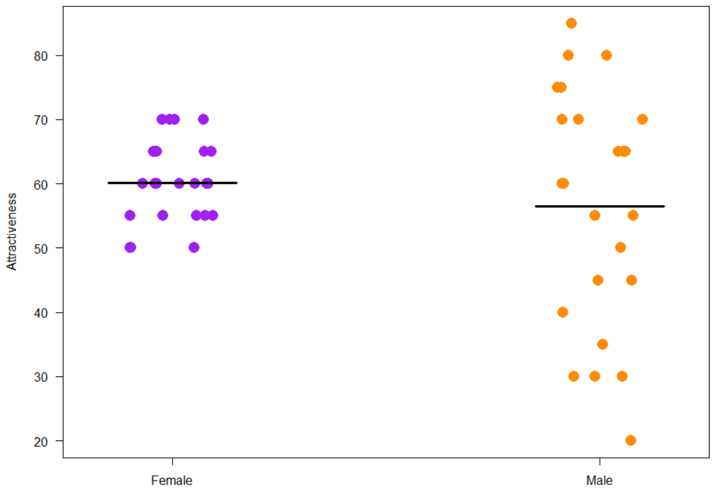


```
gender alcohol attractiveness
Female    None             65
Female    None             70
Female    None             60
Female    None             60
Female    None             60
Female    None             55
```

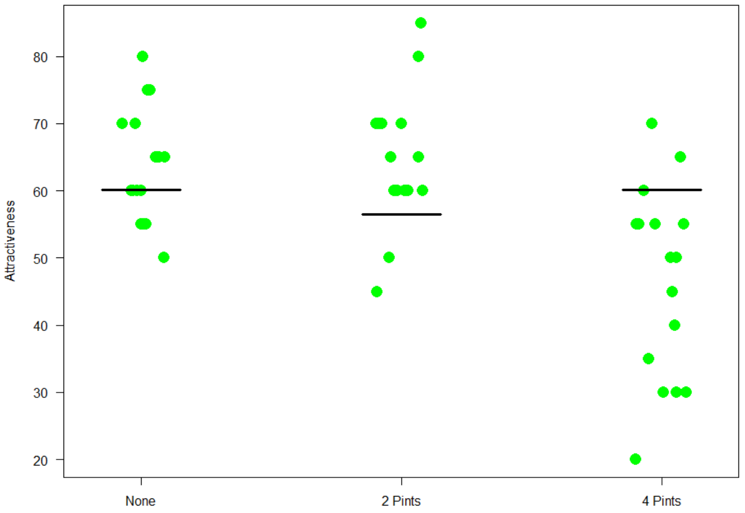
We can now draw the graphs, using the same lines as for the stripchart in the One-way ANOVA section.

```
stripchart(goggles$attractiveness~goggles$gender,
           vertical=TRUE,
           method="jitter",
           las=1,
           ylab="Attractiveness",
           col=c("purple","dark orange"),
           cex=2,
           pch=16)
```

```
attractiveness.gender.means<-
tapply(goggles$attractiveness,goggles$gender,mean)
loc.strip<-1:2
segments(loc.strip-0.15,attractiveness.gender.means,loc.strip+0.15,
attractiveness.gender.means, col="black", lwd=3)
```



```
stripchart(goggles$attractiveness~goggles$alcohol,
           vertical=TRUE,
           method="jitter",
           las=1,
           ylab="Attractiveness",
           col="green",
           cex=2,
           pch=16)

attractiveness.gender.means<-
tapply(goggles$attractiveness,goggles$gender,mean)
loc.strip<-1:3
segments(loc.strip-0.15,attractiveness.gender.means,loc.strip+0.15,
attractiveness.gender.means, col="black", lwd=3)
```



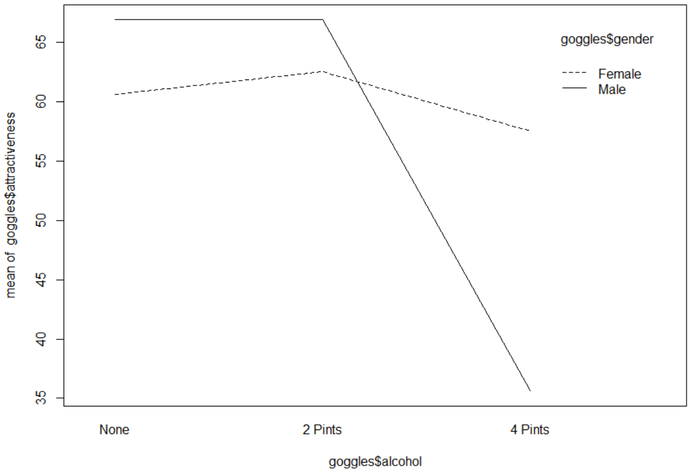
Now, let's plot both factors together. There are 2 ways to go about it. A quick one, which consists of focusing on the potential interaction and not really the individual values, which we have already explored actually. There is a cool function in R `interaction.plot()` which makes it easy for us to have a first look at the possibility of interaction between factors.

Next, we need first to order the levels of the factor alcohol in a sensible fashion.

```
goggles$alcohol <- factor(goggles$alcohol, levels = c("None", "2 Pints", "4
Pints"))
```

Next we want to see what is going on:

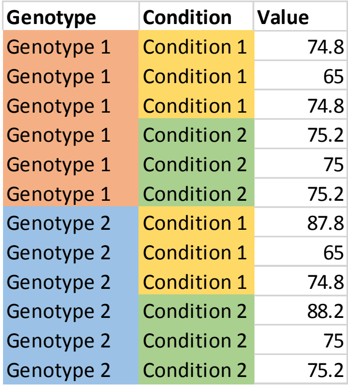
```
interaction.plot(goggles$alcohol,goggles$gender,goggles$attractiveness)
```



It does not give us the prettiest graph but it is very informative. We can see that the difference between genders changes in direction and amplitude across alcohol levels. Now, we have to be a bit careful about the interpretation of this graph as there are no error bars so no information about the behaviour of the data. It is very useful, however, to get an idea of the pattern observed.

Let's take a step back to understand how this interaction business works. Using a fake dataset we are going to go through the different possibilities when it comes to interaction.

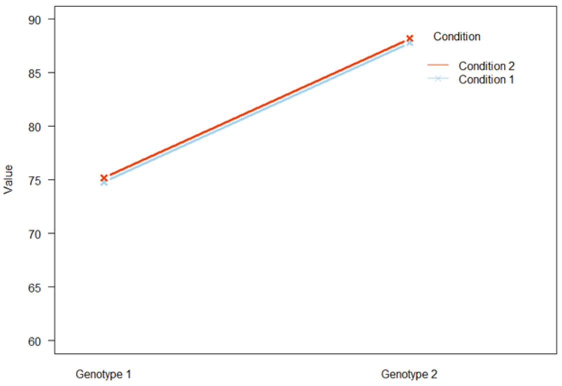
In our fake dataset, we have 2 factors: Genotype (2 levels) and Condition (2 levels).



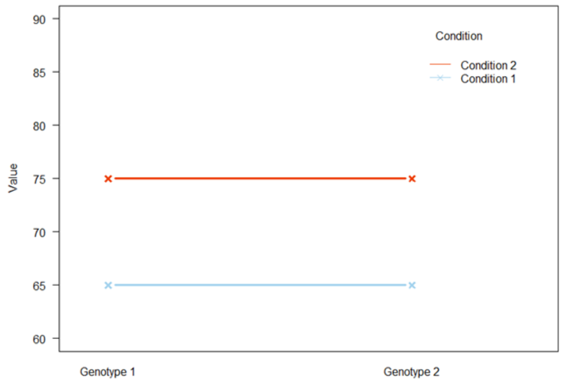
Genotype	Condition	Value
Genotype 1	Condition 1	74.8
Genotype 1	Condition 1	65
Genotype 1	Condition 1	74.8
Genotype 1	Condition 2	75.2
Genotype 1	Condition 2	75
Genotype 1	Condition 2	75.2
Genotype 2	Condition 1	87.8
Genotype 2	Condition 1	65
Genotype 2	Condition 1	74.8
Genotype 2	Condition 2	88.2
Genotype 2	Condition 2	75
Genotype 2	Condition 2	75.2

We are going to use that fake dataset to explore different possible scenarios when it comes to the relationship between 2 factors. The first possible scenario is single effect: either, in our case, Genotype or Condition effect. It would look like below.

Single Effect



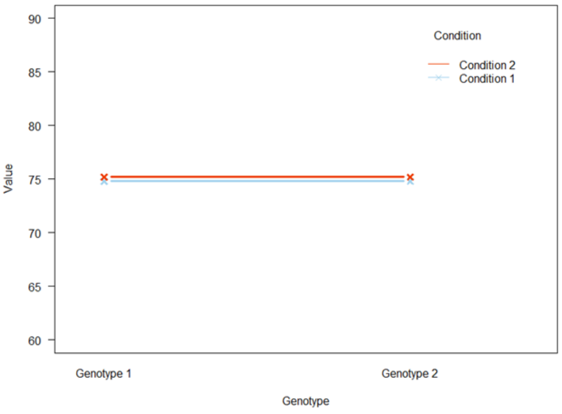
Genotype Effect



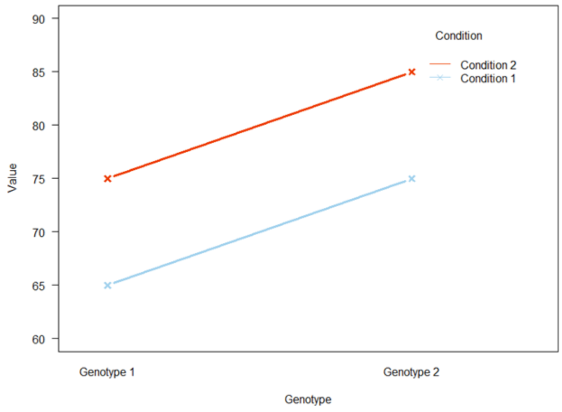
Condition Effect

Then there is the possibility that there is zero effect or both factors have an effect on the outcome variable.

Zero or Both Effect



Zero Effect

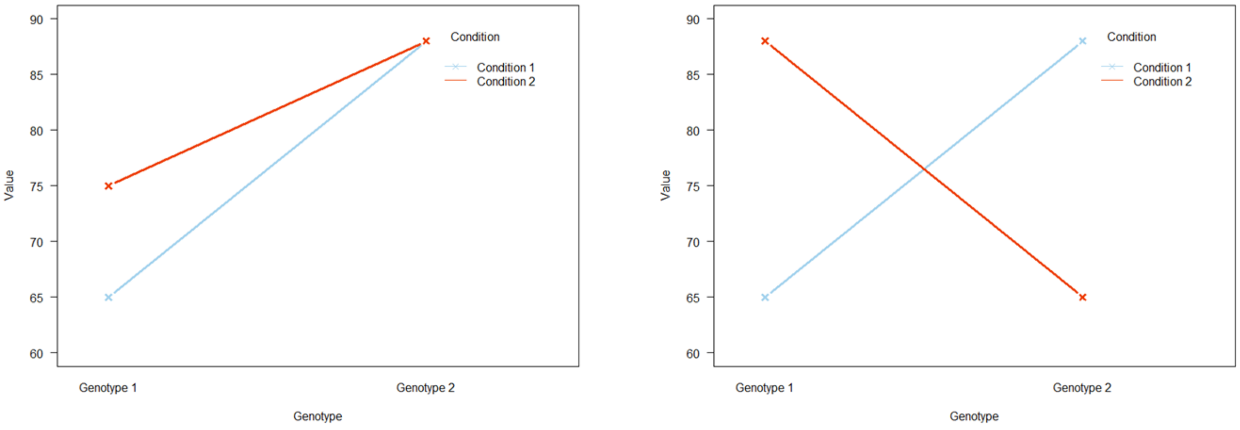


Both Effect

Now, if we look at the 4 graphs above we can answer the question is there an effect of Genotype? For the Genotype effect graph, we can answers ‘yes’ and it is enough as the Genotype effect is the same regardless of Condition. Same thing for the Condition effect graph. Is there an effect of Condition on value, we can just say yes, as again the effect is the same within each Genotype. The same logic applies for Zero and Both effects. In the Both effect graph, even though both factors have an effect on Value, these effects are independent from one another. We can still answer ‘yes’ to both initial question as the Condition effect is the same in both gender and vice-versa.

When there is an interaction, however, the patterns are quite different.

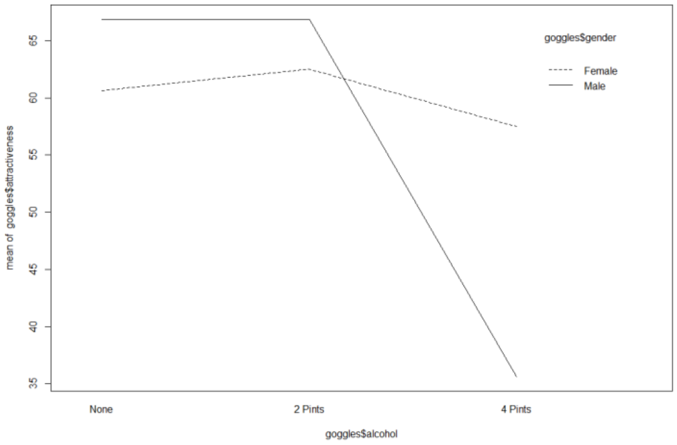
Interaction



On the left, there is an effect of Condition for the first Genotype but not the second one, and there is an overall Genotype effect. On the right, there is a Condition effect but the direction is inversed from Genotype to the other. And there is not Genotype effect. Now if we try to answer the question ‘is there a Condition effect?’ like before, we can no longer answer it by a simple yes or no. On the left, we would have to say yes BUT it depends on the Genotype. And same on the right. This BUT is pretty much the marker for the presence of an interaction (not necessarily significant though).

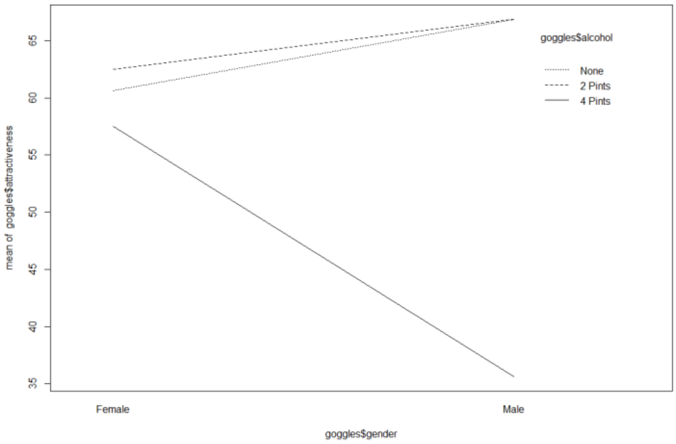
One last thing, there is a kind of symmetry between the factors of a 2-way ANOVA: depending of the scientific question we can choose to say there is an effect of A but this effect varies between levels of B and vice versa.

```
interaction.plot(goggles$alcohol,goggles$gender,goggles$attractiveness)
```



So going back to our goggles data: we can say here that there is an effect of gender but that effect varies depending on alcohol intake.

```
interaction.plot( goggles$gender, goggles$alcohol,goggles$attractiveness)
```

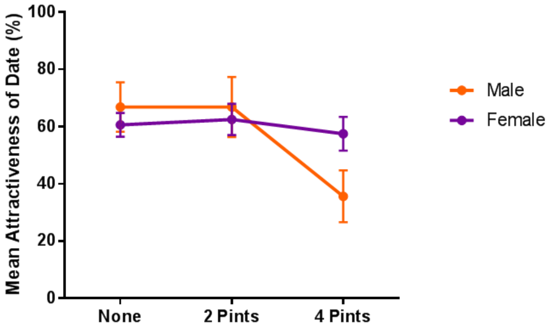


If we consider the data from the other factor’s perspective, we can also say that there is an effect of alcohol but it depends on the gender.

Now because the presence or absence of an interaction will affect our interpretation of the data, it also affects the interpretation of the p-values. Below are the versions of the outcome of our analysis. The first one is the real one, where there is an interaction. The second one is fake and presented for the sake of understanding: how the output would look if there was no interaction.

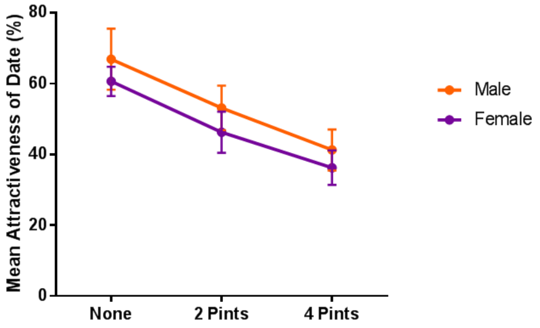
With significant interaction

ANOVA table	SS	DF	MS	F (DFn, DFd)	P value
Interaction	**1978**	**2**	**989.1**	**F (2, 42) = 11.91**	**< 0.0001**
Alcohol Consumption	3332	2	1666	F (2, 42) = 20.07	< 0.0001
Gender	168.8	1	168.8	F (1, 42) = 2.032	0.1614
Residual	3488	42	83.04		



Without significant interaction (fake data)

ANOVA table	SS	DF	MS	F (DFn, DFd)	P value
Interaction	7.292	2	3.646	F (2, 42) = 0.06872	0.9337
Alcohol Consumption	**5026**	**2**	**2513**	**F (2, 42) = 47.37**	**< 0.0001**
Gender	**438.0**	**1**	**438.0**	**F (1, 42) = 8.257**	**0.0063**
Residual	2228	42	53.05		

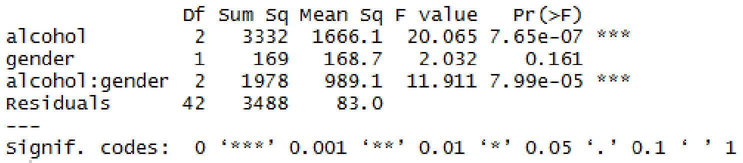


In the case of the real data, we have a significant interaction. It means, we just interpret and report that interaction and we do not report the single effects. The reason is the same as when we are answering the question about the effect of one particular factor: if there is an interaction we cannot simply say, yes, there is a significant effect of alcohol, we also have to mention the fact that this effect is affected

by gender (the BUT thing). Hence, we cannot look at the p-value of Alcohol Consumption as it is meaningless without the gender context (and vice-versa).
On the other hand, if there is no significant interaction, we can just interpret the 2 single effect as we would with a one-way ANOVA. So for instance, we can simply say: there is a significant effect of Gender on Attractiveness because that effect is the same regardless of the alcohol level.

Let's run the actual analysis. To quantify the effects, we are going to use the `aov()` function we used in the previous section.

```
anova.goggles<-aov(attractiveness~alcohol+gender+alcohol*gender,data=goggles)
summary(anova.goggles)
```

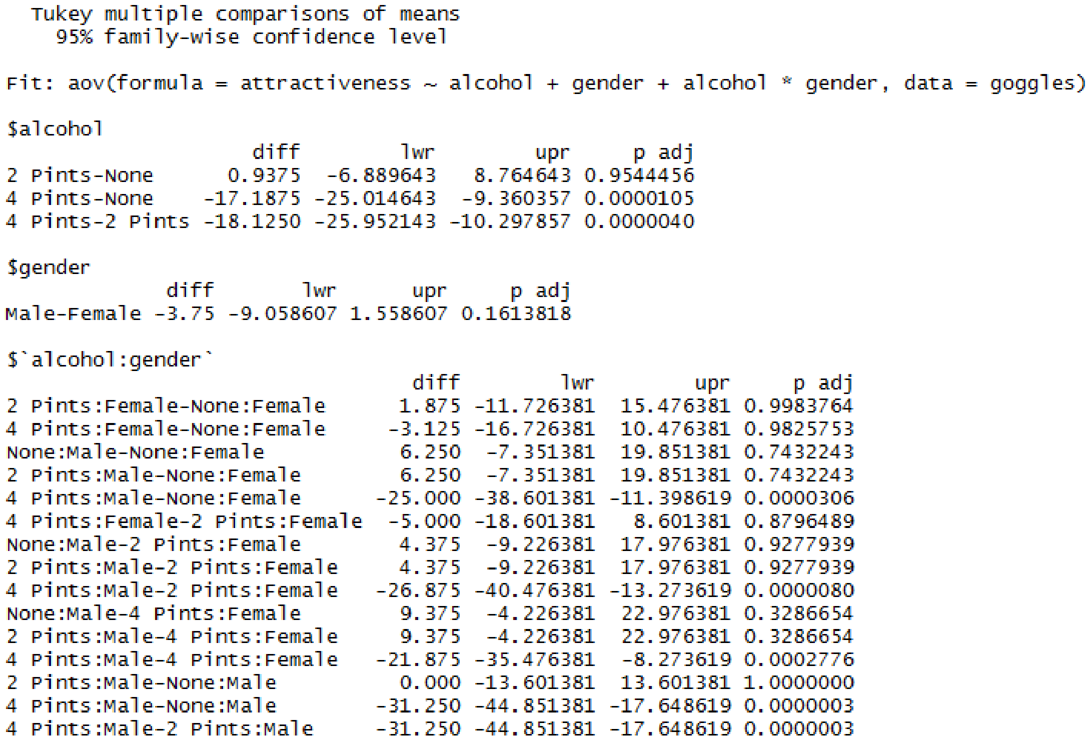


```
               Df Sum Sq Mean Sq F value   Pr(>F)    
alcohol         2   3332  1666.1  20.065 7.65e-07 ***
gender          1    169   168.7   2.032    0.161    
alcohol:gender  2   1978   989.1  11.911 7.99e-05 ***
Residuals      42   3488    83.0                     
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

So, there is a significant interaction between the 2 factors which is consistent with what we observed. Now, in terms of interpretation, when an interaction between 2 factors is significant we don't look at the main effect. It is for the same reason as before: we cannot interpret the effect of one factor without mentioning the other, so it is useless to look at the main effect: all the interpretation is about the interaction.

Now, we may want to quantify the gender effect for each level of alcohol. The most elegant way would be to build contrasts, as in specify the comparisons we want to make, run them and then apply a correction for multiple comparisons. It works very well but it is a bit cumbersome and time consuming. A quicker way is to use Tukey post-hoc tests as we did for the one-way ANOVA. To do that, we write:

```
TukeyHSD(anova.goggles)
```



```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = attractiveness ~ alcohol + gender + alcohol * gender, data = goggles)

$alcohol
                   diff        lwr        upr     p adj
2 Pints-None     0.9375  -6.889643   8.764643 0.9544456
4 Pints-None   -17.1875 -25.014643  -9.360357 0.0000105
4 Pints-2 Pints -18.1250 -25.952143 -10.297857 0.0000040

$gender
             diff       lwr      upr     p adj
Male-Female -3.75 -9.058607 1.558607 0.1613818

$`alcohol:gender`
                                diff        lwr        upr     p adj
2 Pints:Female-None:Female     1.875 -11.726381  15.476381 0.9983764
4 Pints:Female-None:Female    -3.125 -16.726381  10.476381 0.9825753
None:Male-None:Female          6.250  -7.351381  19.851381 0.7432243
2 Pints:Male-None:Female       6.250  -7.351381  19.851381 0.7432243
4 Pints:Male-None:Female     -25.000 -38.601381 -11.398619 0.0000306
4 Pints:Female-2 Pints:Female -5.000 -18.601381   8.601381 0.8796489
None:Male-2 Pints:Female       4.375  -9.226381  17.976381 0.9277939
2 Pints:Male-2 Pints:Female    4.375  -9.226381  17.976381 0.9277939
4 Pints:Male-2 Pints:Female  -26.875 -40.476381 -13.273619 0.0000080
None:Male-4 Pints:Female       9.375  -4.226381  22.976381 0.3286654
2 Pints:Male-4 Pints:Female    9.375  -4.226381  22.976381 0.3286654
4 Pints:Male-4 Pints:Female  -21.875 -35.476381  -8.273619 0.0002776
2 Pints:Male-None:Male         0.000 -13.601381  13.601381 1.0000000
4 Pints:Male-None:Male       -31.250 -44.851381 -17.648619 0.0000003
4 Pints:Male-2 Pints:Male    -31.250 -44.851381 -17.648619 0.0000003
```

It is not very elegant because it compares everything to everything so some comparisons might be irrelevant. Also, by making more comparisons than we should we are overcorrecting, so it is a bit of a conservative approach. On the other hand, it gives us a quick answer to our question.

So, the function looks at the pairwise comparisons for the main effects and for the interaction. Remember: only look at the main effects if the interaction is not significant.

Anyway, we can conclude that there is a significant effect of alcohol consumption on the way the attractiveness of a date is perceived but it varies significantly between genders (p=7.99e-05). With 2 pints or less, boys seem to be very slightly more picky about their date than girls (but not significantly so) but with 4 pints the difference is reversed and significant (p=0.0003).

Chapter 6: Correlation

(File: `Exam Anxiety.dat`)

If we want to find out about the relationship between 2 continuous variables, we can run a correlation.

A bit of theory: Correlation coefficient

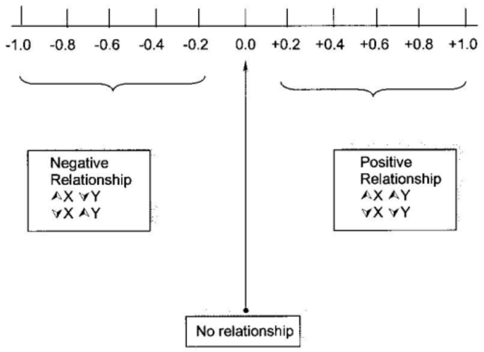
A correlation is a measure of a linear relationship (can be expressed as straight-line graphs) between variables. The simplest way to find out whether 2 variables are associated is to look at whether they co-vary. To do so, we combine the variance of one variable with the variance of the other.

$$\text{cov}(X, Y) = \sum_{i=1}^{N} \frac{(x_i - \bar{x})(y_i - \bar{y})}{N}.$$

A positive covariance indicates that as one variable deviates from the mean, the other one deviates in the same direction, in other words if one variable goes up the other one goes up as well.
The problem with the covariance is that its value depends upon the scale of measurement used, so we would not be able to compare covariance between datasets unless both data are measures in the same units. To standardise the covariance, it is divided by the SD of the 2 variables. It gives us the most widely-used correlation coefficient: the Pearson product-moment correlation coefficient “r”.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Of course, you don’t need to remember that formula but it is important that you understand what the correlation coefficient does: it measures the magnitude and the direction of the relationship between two variables. It is designed to range in value between 0.0 and 1.0.



The 2 variables do not have to be measured in the same units but they have to be proportional (meaning linearly related). Apart from r, there is another important coefficient: the coefficient of determination r^2: it gives the proportion of variance in Y that can be explained by X, in percentage.
Finally, the assumptions for correlation (regression in general) are pretty much the ones we have seen before:

Linearity: The relationship between X and the mean of Y is linear.
Homoscedasticity: The variance of residual is the same for any value of X.
Independence: Observations are independent of each other.

Normality: For any fixed value of X, Y is normally distributed.
When running a regression in general and a correlation in particular, we need to check for problematic points. They can be:

Outliers: an outlier is defined as an observation that has a large residual. In other words, the observed value for the point is very different from that predicted by the regression model.

Leverage points: A leverage point is defined as an observation that has a value of x that is far away from the mean of x.

Influential observations: An influential observation is defined as an observation that changes the slope of the line. Thus, influential points have a large influence on the fit of the model. One method to find influential points is to compare the fit of the model with and without each observation.

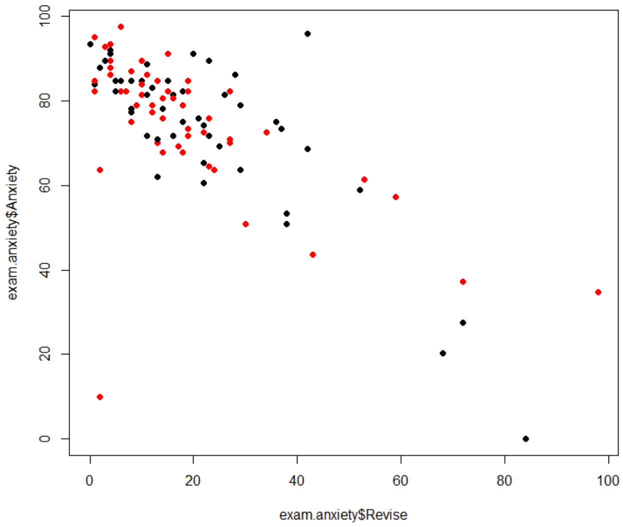
The bottom line is that first we look at the outliers, once we have identified them, we check the influence statistics and if one or more are 'out of line' we can then safely remove the value.

Example (File: `Exam Anxiety.dat`)

```
exam.anxiety<-read.table("Exam Anxiety.dat", sep="\t",header=T)
```

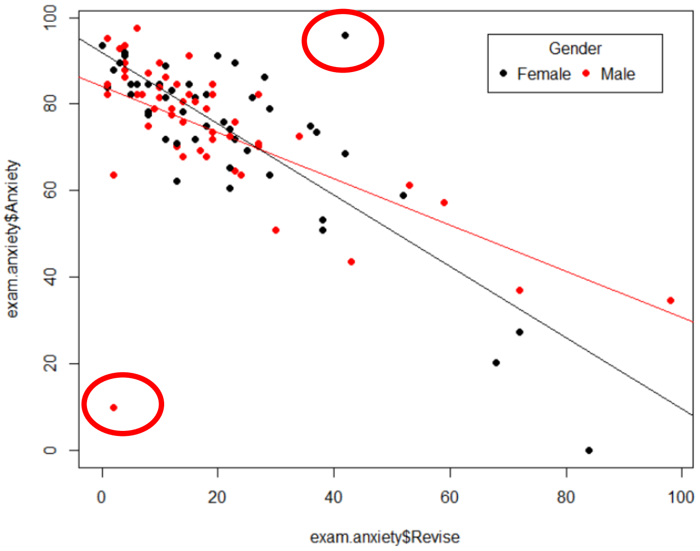
The first thing we are going to do is to plot the data. We will start with revising time vs. anxiety levels.

```
plot(exam.anxiety$Anxiety~exam.anxiety$Revise,col=exam.anxiety$Gender,
pch=16)
legend("topright", title="Gender",inset=.05, c("Female","Male"),
horiz=TRUE, pch=16,col=1:2)
```



By looking at the graph, one can think that something is happening here. To get a better idea we can add lines-of-best fit but to do that we first need to fit the model as the lines-of-best fit's coefficients are one of the outputs of the regression:

```
fit.male<-
lm(Anxiety~Revise,data=exam.anxiety[exam.anxiety$Gender=="Male",])
fit.female<-
lm(Anxiety~Revise,data=exam.anxiety[exam.anxiety$Gender=="Female",])

abline((fit.male), col="red")
abline((fit.female), col="black")
```

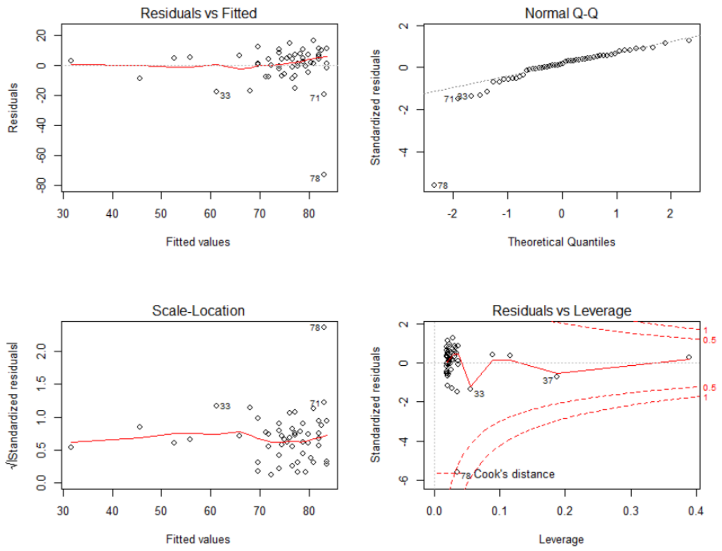


Now, we want to quantify the strength of the relationship between our 2 variables of interest but first we need to check on the data.

Outliers and influential cases

We might have noticed that one point, possibly 2, are really far from the others. So let’s check out our data and keep an eye on our misbehaving cases and in particular the boy (point Code 78) who spent 2 hours revising, did not feel stressed about it (Anxiety score: 10) and managed a 100% mark in his exam. Yeah right …

```
par(mfrow=c(2,2)
plot(fit.male)
```



The first plot depicts residuals versus fitted values. Residuals are measured as follows:
residual = observed y – model-predicted y.

So the further the observed y from the one predicted by the model, the poorer the prediction. The plot of residuals versus predicted values is useful for checking the assumption of linearity and homoscedasticity. If the model does not meet the linear model assumption, we would expect to see residuals that are very large (big positive value or big negative value). To assess the assumption of linearity we want to ensure that the residuals are not too far away from 0. To assess if the homoscedasticity assumption is met we look to make sure that there is no pattern in the residuals and that they are equally spread around the y = 0 line. On our case, R identifies 3 points with high residuals, one of which has a really high one: point 78.

The second plot (QQ-plot) evaluates the normality assumption. It compares the residuals to "ideal" normal observations. We want our observations to lie well along the 45-degree line in the QQ-plot, which is the case here, except for point 78.

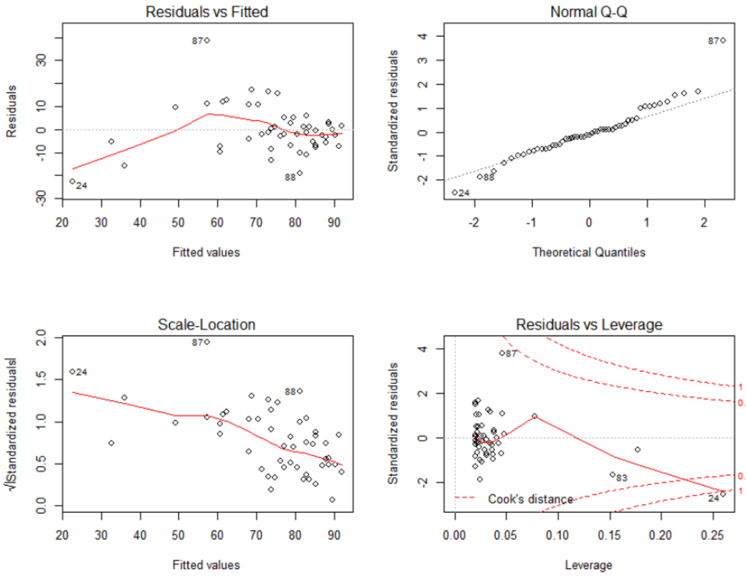
The third plot is a scale-location plot (square rooted standardised residual vs. predicted value). This is useful for checking the assumption of homoscedasticity. In this particular plot we are checking to see if there is a pattern in the residuals. In our case, things look OK. Point 78 is, however, away from the others.

Finally, the fourth plot is of **"**Cook's distance**",** which is a measure of the influence of each observation on the regression coefficients. The Cook's distance statistic is a measure, for each observation in turn, of the extent of change in model estimates when that particular observation is omitted. Any observation for which the Cook's distance is close to 1 or more (above 0.5), or that is substantially larger than other Cook's distances (highly influential data points), requires investigation. Once more, in our case, point 78 is of concern.

Outliers may or may not be influential points**.** As stated before, influential outliers are of the greatest concern. They should never be disregarded. Careful scrutiny of the original data may reveal an error in data entry that can be corrected. They can be excluded from the final fitted model but they must be noted in the final report or paper.

On our case, one point stands out in all 4 graphs: point 78 so we will look at the correlation with and without this value.

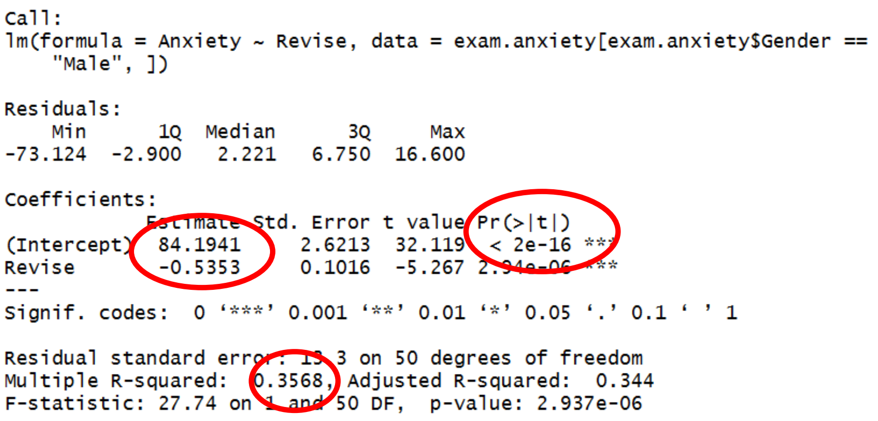
```
plot(fit.female)
```



For the girls, point 87 stand out, though not as strikingly as point 78 for the boys. And it is below the threshold to be identified as an influential case (plot 4). We will, however, keep an eye on it.

To get the output of the analysis:

```
summary(fit.male)
```

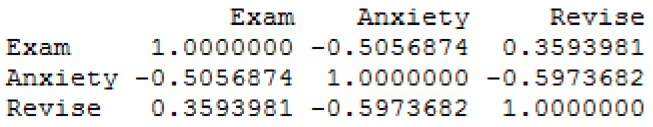


```
Call:
lm(formula = Anxiety ~ Revise, data = exam.anxiety[exam.anxiety$Gender ==
    "Male", ])

Residuals:
    Min      1Q  Median      3Q     Max
-73.124  -2.900   2.221   6.750  16.600

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  84.1941     2.6213  32.119  < 2e-16 ***
Revise       -0.5353     0.1016  -5.267 2.94e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13.3 on 50 degrees of freedom
Multiple R-squared: 0.3568, Adjusted R-squared:  0.344
F-statistic: 27.74 on 1 and 50 DF,  p-value: 2.937e-06
```

From this output we get 4 important pieces of information. First the coefficients of the line of best fit: Intercept: 84.19 and slope: -0.53. So it goes: Anxiety=84.19-0.53*Revise

We can also see that the relationship between the 2 variables is highly significant: p<2e-16. And finally R^2=0.3568: the model explains about 36% of the variability observed in anxiety. We can get the coefficient of correlation by calculating the square root of R^2 or with the line below, if we want to look at the relationships between all variables.

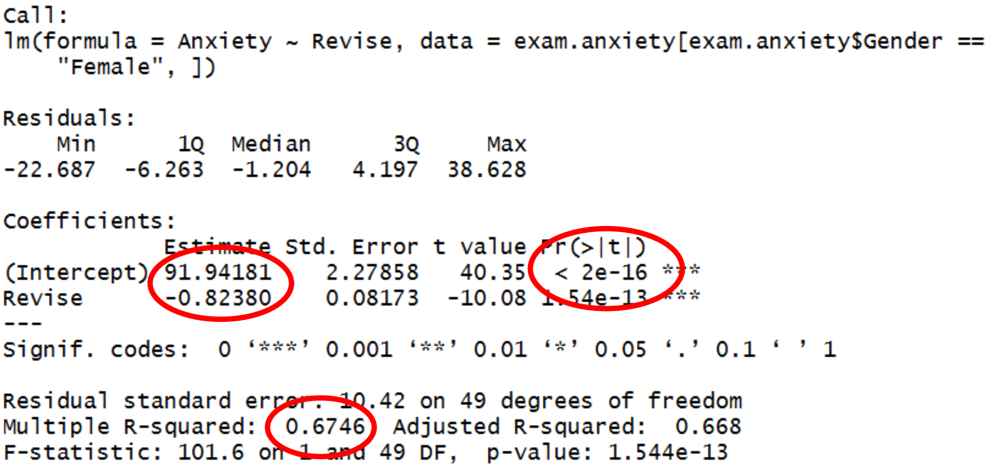
```
cor(exam.anxiety[exam.anxiety$Gender=="Male",c("Exam","Anxiety","Revise")])
```



```
               Exam    Anxiety     Revise
Exam      1.0000000 -0.5056874  0.3593981
Anxiety  -0.5056874  1.0000000 -0.5973682
Revise    0.3593981 -0.5973682  1.0000000
```

For the females:

```
summary(fit.female)
```



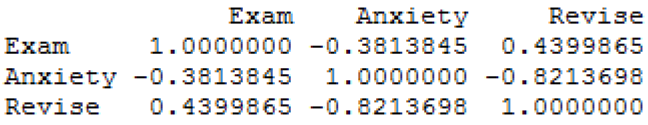
```
Call:
lm(formula = Anxiety ~ Revise, data = exam.anxiety[exam.anxiety$Gender ==
    "Female", ])

Residuals:
    Min      1Q  Median      3Q     Max
-22.687  -6.263  -1.204   4.197  38.628

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 91.94181    2.27858   40.35  < 2e-16 ***
Revise      -0.82380    0.08173  -10.08 1.54e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.42 on 49 degrees of freedom
Multiple R-squared:  0.6746  Adjusted R-squared:  0.668
F-statistic: 101.6 on 1 and 49 DF,  p-value: 1.544e-13
```

We get:

Anxiety=91.94-0.82*Revise with p<2e-16.

So a significant result again, with a higher intercept and a steeper slope as expected. And for the correlations:

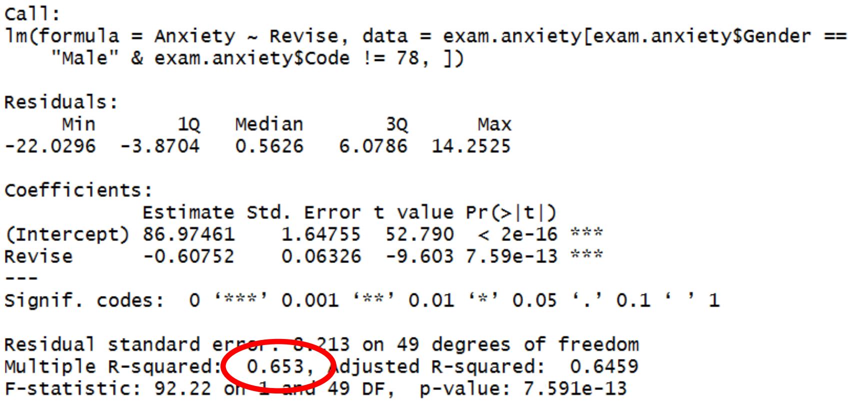
```
cor(exam.anxiety[exam.anxiety$Gender == "Female",
c("Exam","Anxiety","Revise")])
```



```
              Exam    Anxiety     Revise
Exam     1.0000000 -0.3813845  0.4399865
Anxiety -0.3813845  1.0000000 -0.8213698
Revise   0.4399865 -0.8213698  1.0000000
```

Now what happens if we remove point 78 from the males dataset and rerun the analysis?

```
fit.male2<-lm(Anxiety~Revise,
data=exam.anxiety[exam.anxiety$Gender=="Male"&exam.anxiety$Code!=78,])
summary(fit.male2)
```

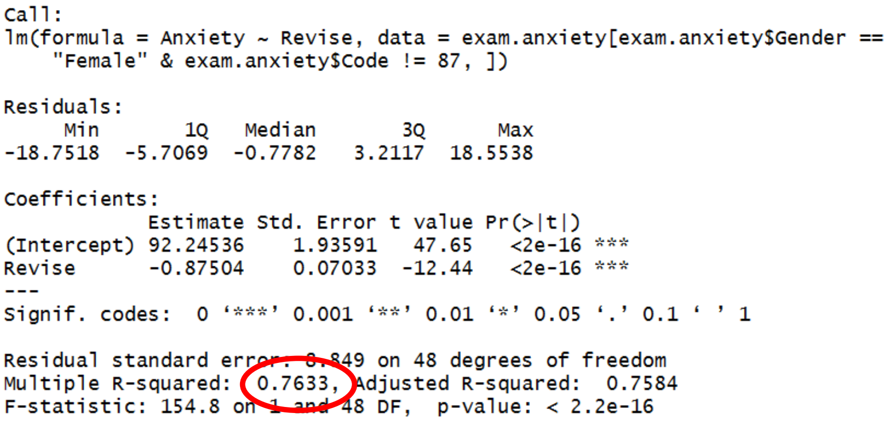


```
Call:
lm(formula = Anxiety ~ Revise, data = exam.anxiety[exam.anxiety$Gender ==
    "Male" & exam.anxiety$Code != 78, ])

Residuals:
     Min       1Q   Median       3Q      Max 
-22.0296  -3.8704   0.5626   6.0786  14.2525 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 86.97461    1.64755  52.790  < 2e-16 ***
Revise      -0.60752    0.06326  -9.603 7.59e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.213 on 49 degrees of freedom
Multiple R-squared:  0.653,	Adjusted R-squared:  0.6459 
F-statistic: 92.22 on 1 and 49 DF,  p-value: 7.591e-13
```

We can notice that, without the influential outlier, the slope is steeper but most importantly R^2 jumps from 36% to 65% so a much better fit.

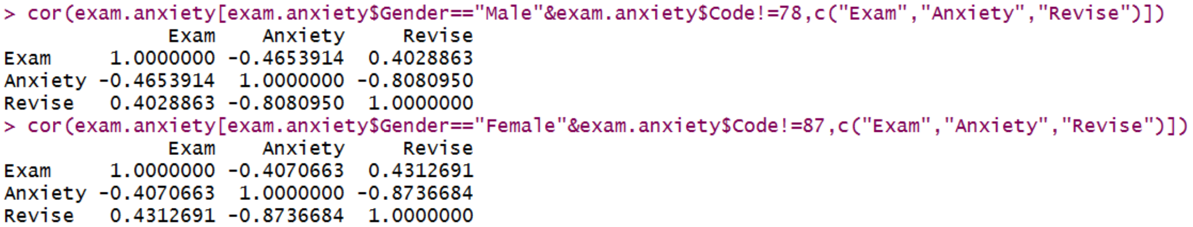
For the females:

```
fit.female2<-lm(Anxiety~Revise,
data=exam.anxiety[exam.anxiety$Gender=="Female"&exam.anxiety$Code!=87,])
summary(fit.female2)
```



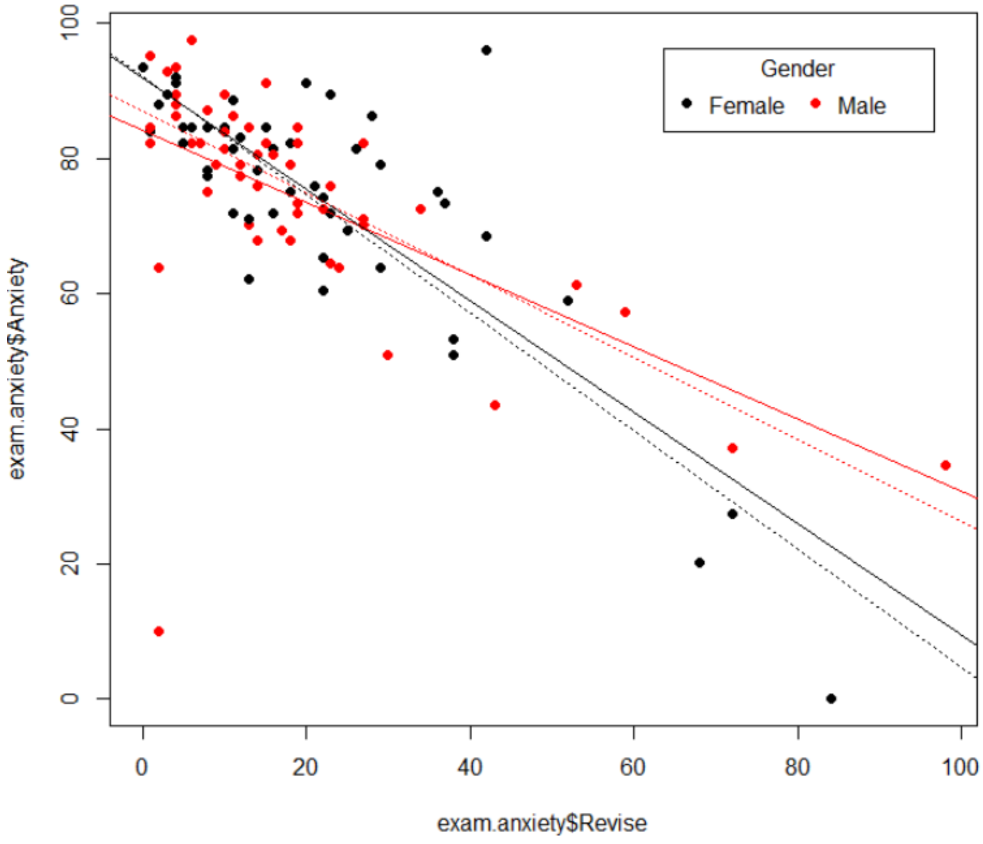
```
Call:
lm(formula = Anxiety ~ Revise, data = exam.anxiety[exam.anxiety$Gender ==
    "Female" & exam.anxiety$Code != 87, ])

Residuals:
     Min       1Q   Median       3Q      Max 
-18.7518  -5.7069  -0.7782   3.2117  18.5538 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 92.24536    1.93591   47.65   <2e-16 ***
Revise      -0.87504    0.07033  -12.44   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.849 on 48 degrees of freedom
Multiple R-squared:  0.7633,	Adjusted R-squared:  0.7584 
F-statistic: 154.8 on 1 and 48 DF,  p-value: < 2.2e-16
```

This model is better than the one with the outlier but the influence of point 87 is not as big. Keeping or removing the value is more debatable.



```
> cor(exam.anxiety[exam.anxiety$Gender=="Male"&exam.anxiety$Code!=78,c("Exam","Anxiety","Revise")])
              Exam    Anxiety     Revise
Exam     1.0000000 -0.4653914  0.4028863
Anxiety -0.4653914  1.0000000 -0.8080950
Revise   0.4028863 -0.8080950  1.0000000
> cor(exam.anxiety[exam.anxiety$Gender=="Female"&exam.anxiety$Code!=87,c("Exam","Anxiety","Revise")])
              Exam    Anxiety     Revise
Exam     1.0000000 -0.4070663  0.4312691
Anxiety -0.4070663  1.0000000 -0.8736684
Revise   0.4312691 -0.8736684  1.0000000
```

```
plot(exam.anxiety$Anxiety~exam.anxiety$Revise,col=exam.anxiety$Gender,pch=1
6)

legend("topright", title="Gender",inset=.05, c("Female","Male"),
horiz=TRUE, pch=16,col=1:2)

abline((fit.male), col="red")
abline((fit.female), col="black")
abline((fit.male2), col="red",lty=3)
abline((fit.female2), col="black",lty=3)
```



Chapter 7: Non-Parametric approach

7-1 Comparing 2 groups

What if the data do not meet the assumptions for parametric tests? Then, we should go for a non-parametric approach. The choice between the two is not always easy, however. If the outcome is a rank or a score, then it is clearly not Gaussian and going for a non-parametric approach is a no-brainer. The difficulty is mostly with small samples, where it is often not easy to determine the distribution. In that case, looking at previous data might help, as what matters is the distribution of the overall population, not the distribution of the sample.

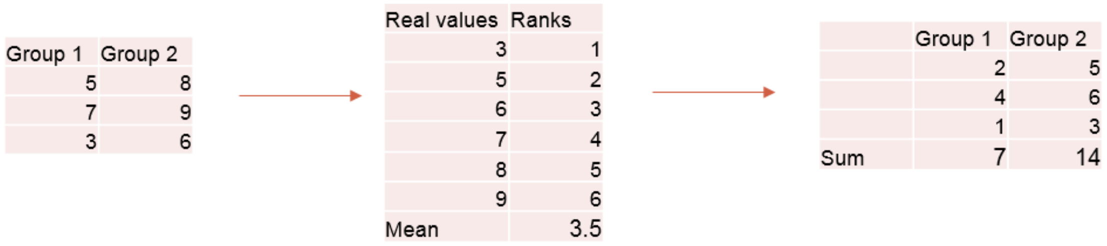
Non-parametric tests are not strictly speaking assumption-free. For instance, they assume a continuous distribution, though scores are OK, but they are far less restrictive than their parametric counterparts. Most of them are based on the principle of ranking: the smallest value has the lowest rank and the highest value the highest rank.

Some may argue that, when in doubt, it is more valid to use nonparametric methods because they are "always valid, but not always efficient," while parametric methods are "always efficient, but not always valid" (Nahm, 2016).

Independent groups: Mann-Whitney (= Wilcoxon's rank-sum) test

The way the Mann-Whitney works is really cool. It groups all the data, regardless of which group it belongs to, and ranks them. The difference, or absence of, between the two groups is based on the sum of the ranks in each group. Hence, if there is a difference, the group with the highest values will have the highest ranks and thus the highest sum of ranks. The Mann-Whitney statistic W is calculated as follows:

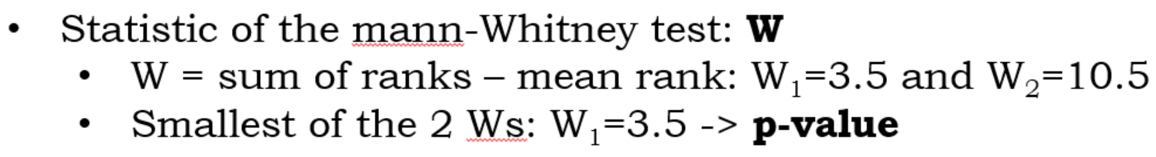
W = sum of ranks (for each group so W_1 and W_2) – mean rank. The smallest of the two Ws is chosen and used to calculate the associated p-value as illustrated below.



Group 1	Group 2
5	8
7	9
3	6

Real values	Ranks
3	1
5	2
6	3
7	4
8	5
9	6
Mean	3.5

	Group 1	Group 2
	2	5
	4	6
	1	3
Sum	7	14



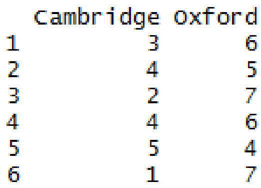
- Statistic of the mann-Whitney test: **W**
  - W = sum of ranks – mean rank: W_1=3.5 and W_2=10.5
  - Smallest of the 2 Ws: W_1=3.5 -> **p-value**

Let's try an example.

Example (File: `smelly teeshirt.csv`)

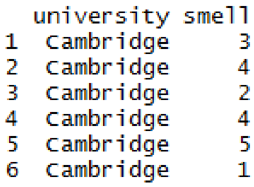
In a study designed to assess whether group body odour is less disgusting when it is associated with an in-group member versus an out-group member, researchers presented two groups of Cambridge University students with one of two smelly, worn t-shirts. One t-shirt bore the logo of Cambridge University and the other bore the logo of Oxford University. The students were asked to rate their disgust on a 7-point ordinal scale. Higher ratings indicate greater levels of disgust. The disgust ratings for each group are presented in the smelly_teeshirt file.

```
smelly.teeshirt<-read.csv("smelly teeshirt.csv",header=T)
head(smelly.teeshirt)
```



```
  Cambridge Oxford
1         3      6
2         4      5
3         2      7
4         4      6
5         5      4
6         1      7
```

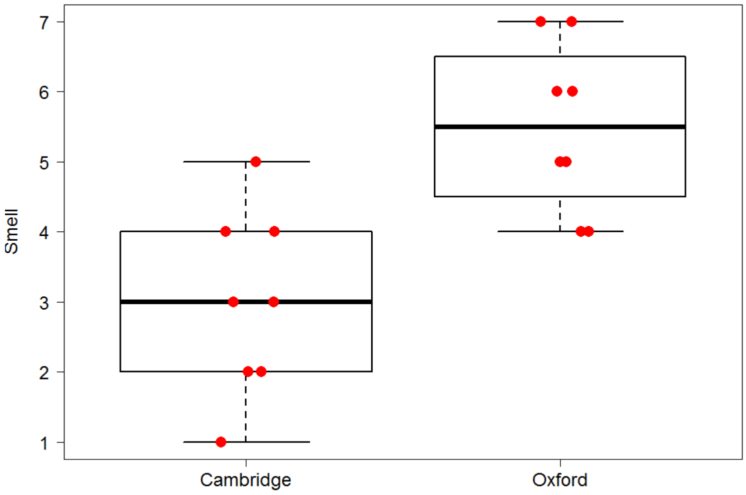
```
smelly.teeshirt.stack <- melt(smelly.teeshirt)
colnames(smelly.teeshirt.stack)<-
c("university","smell")
head(smelly.teeshirt.stack)
```



```
  university smell
1  Cambridge     3
2  Cambridge     4
3  Cambridge     2
4  Cambridge     4
5  Cambridge     5
6  Cambridge     1
```

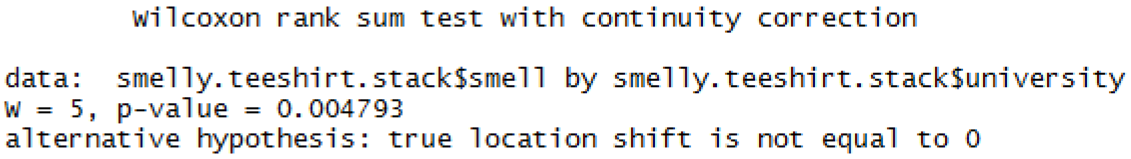
Quick look at the data:

```
boxplot(smelly.teeshirt.stack$smell~smelly.teeshirt.stack$university,
        pch = 20,
        cex=2,
        lwd=2,
        las=1,
        ylab="Smell",
        cex.axis=1.5,
        cex.lab=1.5)

stripchart(smelly.teeshirt.stack$smell~smelly.teeshirt.stack$university,
           vertical=TRUE,
           method="jitter",
           las=1,
           pch=16,
           cex=2,
           col="red",
           add=TRUE)
```



Since we are going for a non-parametric approach, box plot or median are best.
The Mann-Witney test and the Wicoxon's rank one are the same and the function in R is: `wilcox.test()`

```
wilcox.test(smelly.teeshirt.stack$smell~smelly.teeshirt.stack$university)
```



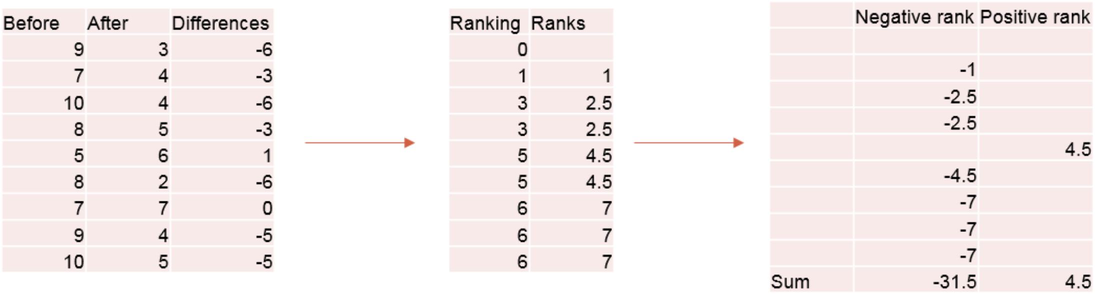
```
        Wilcoxon rank sum test with continuity correction

data:  smelly.teeshirt.stack$smell by smelly.teeshirt.stack$university
W = 5, p-value = 0.004793
alternative hypothesis: true location shift is not equal to 0
```

So, not surprisingly(!), the Cambridge students find the smell of Oxford t-shirts significantly more disgusting than that of Cambridge ones.

Note: depending on which order we enter the data, the function will give either the biggest or the smallest W, which is confusing but does not affect the outcome of the test.

Dependent groups: Wilcoxon's signed-rank test

This test is also based on ranks but this time the differences between the two members of the pair are ranked (see example below). The zero differences are ignored and the sum of the positive and of the negative differences are calculated (T^+ and T^-). Following the same logic as for the Mann-Whitney, the smallest of the two Ts is chosen and becomes the T statistic. This T will allow the calculation of the p-value. Easy.



Before	After	Differences
9	3	-6
7	4	-3
10	4	-6
8	5	-3
5	6	1
8	2	-6
7	7	0
9	4	-5
10	5	-5

Ranking	Ranks
0	
1	1
3	2.5
3	2.5
5	4.5
5	4.5
6	7
6	7
6	7

	Negative rank	Positive rank
	-1	
	-2.5	
	-2.5	
		4.5
	-4.5	
	-7	
	-7	
	-7	
Sum	-31.5	4.5

- Statistic of the Wilcoxon's signed-rank test: **T**
  - Here: Wilcoxon's T = 4.5 (smallest of the 2 (absolute value))
  - N = 9 (we ignore the 0 difference): T + N -> **p-value**

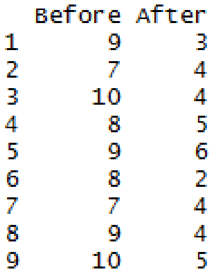
In R, the function is the same as for the independent approach but we have to specify about the matching:

```
wilcox.test(paired = TRUE)
```

Example (File: `botulinum.csv`)

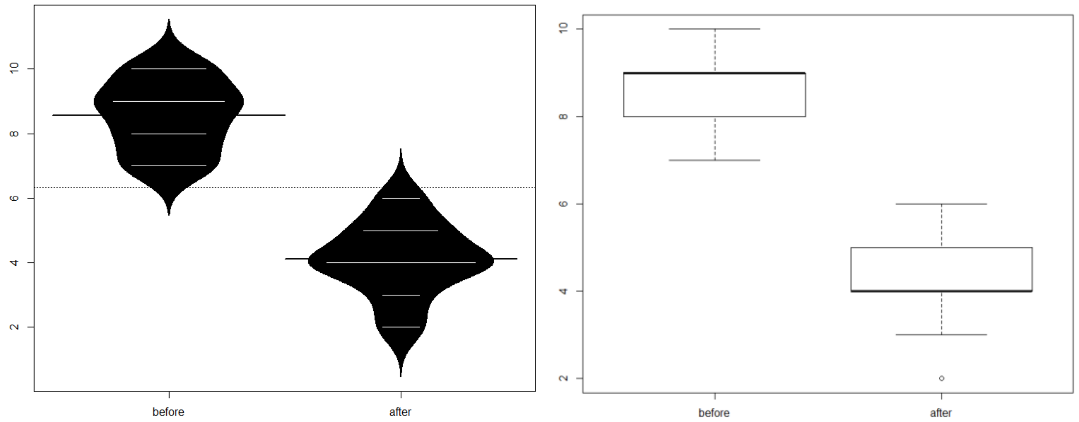
A group of 9 disabled children with muscle spasticity (or extreme muscle tightness limiting movement) in their right upper limb underwent a course of injections with botulinum toxin to reduce spasticity levels. A neurologist (blind to group membership) assessed levels of spasticity pre- and post-treatment for all 9 children using a 10-point ordinal scale. Higher ratings indicated higher levels of spasticity. The ratings for each group are presented in the file `botulinum.csv`.

```
botulinum<-read.csv("botulinum.csv",header=T)
head(botulinum)
```



```
  Before After
1      9     3
2      7     4
3     10     4
4      8     5
5      9     6
6      8     2
7      7     4
8      9     4
9     10     5
```

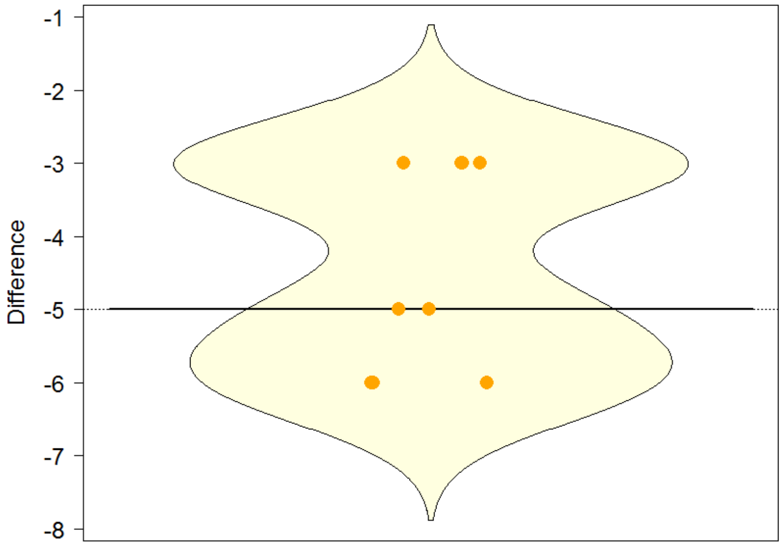
```
beanplot(botulinum$before, botulinum$after, names = c("before", "after"))
boxplot(botulinum$before, botulinum$after, names = c("before", "after"))
```



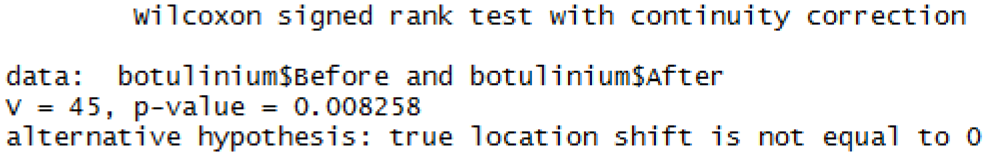
On top of the fact that the dependent variable in both cases is measured using an ordinal scale, the distribution of difference scores in both cases deviates from normality. So, we should definitely go for a non-parametric approach.

Let’s have a look at the difference:

```
botulinum$difference <-botulinum$after - botulinum$before

beanplot(botulinum$difference,
         las=1, overallline = "median",
         ylab = 'Difference',
         cex.lab=1.5,
         col="light yellow",
         what = c(TRUE, TRUE, TRUE, FALSE),
         cex.axis=1.5)



stripchart(botulinum$difference,
           vertical=TRUE,
           method="jitter",
           las=1,
           pch=16,
           cex=2,
           col="orange",
           add=TRUE)
```



```
wilcox.test(botulinum$before, botulinum$after, paired = TRUE)
```



```
        Wilcoxon signed rank test with continuity correction

data:  botulinium$Before and botulinium$After
V = 45, p-value = 0.008258
alternative hypothesis: true location shift is not equal to 0
```

There was a significant difference pre- and post- treatment in ratings of muscle spasticity (p=0.008). Please note that although the test reports T, it calls it V. Go figure.

7-2 Comparing more than 2 groups

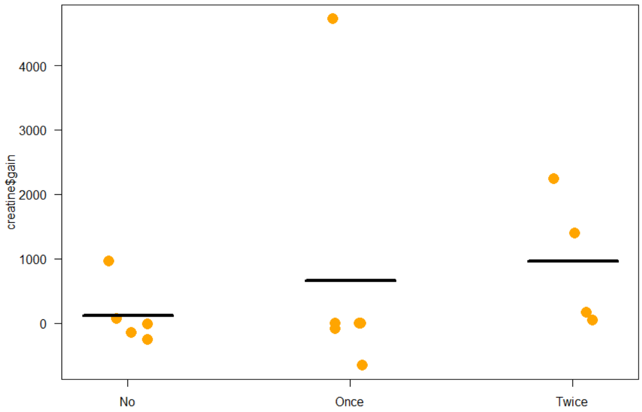
Kruskal-Wallis

What if the data do not meet the assumptions for ANOVA? We can choose to run the non-parametric equivalent: the Kruskal-Wallis.
Now, the `kruskal.wallis()` function only produces the omnibus part of the analysis. However, we will often, but not always, want to look at the pairwise comparisons. The post-hoc test associated with the Kruskal-Wallis is the Dunn test and one of the arguments of the `dunn.test()` function (dunn.test package) is `kw`: the output of the Kruskal-Wallis, so that's the function we will be using.
The statistic associated with the Kruskal-Wallis is H and it follows a Chi^2 distribution hence the statistic reported in the output. The Dunn test works pretty much like the Mann-Whitney one.

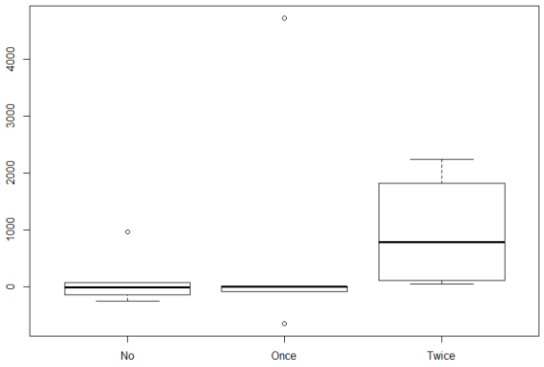
Example (File: creatine.csv)

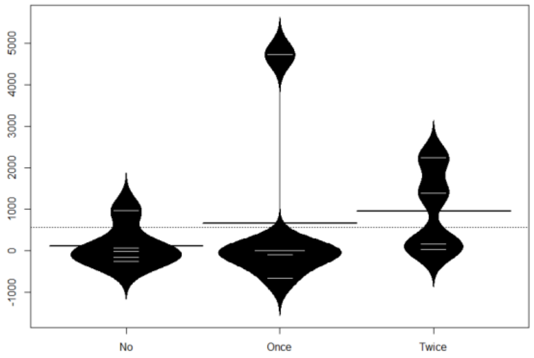
The data contain the result of a small experiment regarding creatine, a supplement that's popular among body builders. These were divided into 3 groups: some didn't take any creatine, others took it in the morning only and still others took it in the morning and evening. After doing so for a month, their weight gains were measured.
The research question is: does the average weight gain depend on the creatine condition to which people were assigned?

```
creatine<-read.csv("creatine.csv",header=T)

stripchart(creatine$gain~creatine$creatine,
           vertical=TRUE,
           method="jitter",
           las=1,
           pch=16,
           cex=2,
           col="orange")

creatine.means<-tapply(creatine$gain,creatine$creatine,mean)
centre<-1:3
segments(centre-0.2,creatine.means,centre+0.2,creatine.means,   col="black",
lwd=4)
```

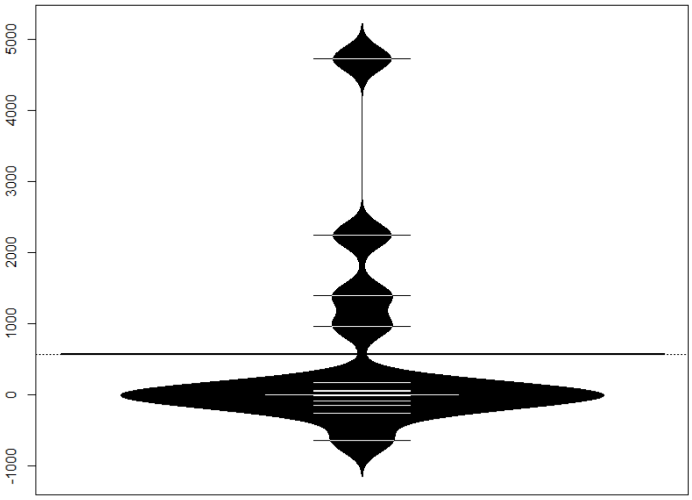


```
boxplot(creatine$gain~creatine$creatine)
beanplot(creatine$gain~creatine$creatine)
```





```
beanplot(creatine$gain)
```



This one looks pretty cool.

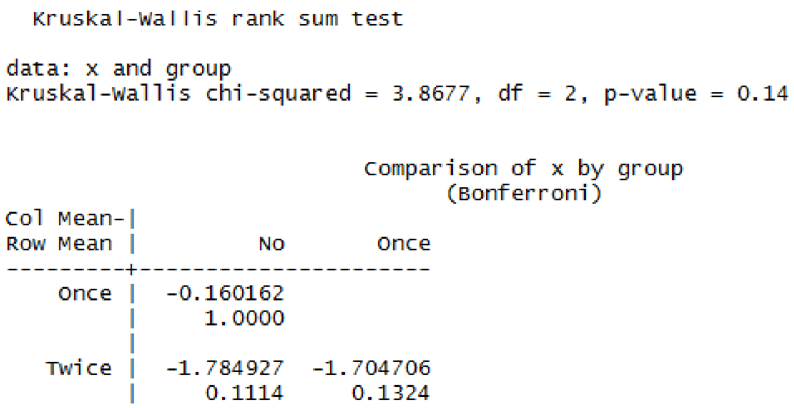
```
tapply(creatine$gain,creatine$creatine,sd)
```

```
       No      Once     Twice 
 488.5317 2005.1585 1047.8519 
```

The graphical exploration of the data, together with a look at the standard deviation tell us that a parametric approach would be unwise.

So, here we go:

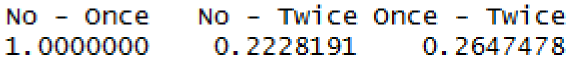
```
dunn.test(creatine$gain,creatine$creatine,kw=TRUE,method="bonferroni",alpha
=0.05)  ## dunn.test package ##
```



```
  Kruskal-Wallis rank sum test

data: x and group
Kruskal-Wallis chi-squared = 3.8677, df = 2, p-value = 0.14


                           Comparison of x by group
                                (Bonferroni)
Col Mean-|
Row Mean |         No       Once
---------+----------------------
    Once |  -0.160162
         |     1.0000
         |
   Twice |  -1.784927  -1.704706
         |     0.1114     0.1324
```

The omnibus part of the analysis tells us that there is no overall difference between the 3 groups (p=0.14). Now because of the absence of significance, we don't need to go any further but there is a little hiccup in the output that needs to be mentioned. The p-values presented in the table are one-sided ones, so it will not do. There are 2 ways to deal with this: the first one, we can just multiply by the p-values and we are good. Second approach, a bit more advanced, we can ask R to do it for us. First, we will create an object, then from that object we will take the P.adjusted and multiply them by 2 (except the first one as it is already 1). Remember, the p-value from a one-tailed test is 2 times lower than that of a one-tailed one. Finally, to make the output clearer, we will assign a name to each comparison.

```
model<-
dunn.test(creatine$gain,creatine$creatine,kw=TRUE,method="bonferroni")
adjust.p.values <- model$P.adjusted*c(1,2,2)
names(adjust.p.values) <- model$comparisons
adjust.p.values
```



```
 No - Once    No - Twice Once - Twice
1.0000000     0.2228191     0.2647478
```

So, this study did not demonstrate any effect from creatine (χ^2 = 3.87, p = 0.14).

Non-parametric two-way Analysis of Variance

What if the data do not meet the assumptions for a 2-way ANOVA? Well, it is a problem as the equivalent test: the Scheirer-Ray-Hare is not well documented nor well regarded, so we will not cover it in this manual.

If we absolutely must, there is a not-so-elegant and a-bit-cumbersome way to deal with such a design: build groups like we did in the 2-way ANOVA above so from a 2 groups by 3 groups design (2-way) we get a 6 groups one (1-way). To this design, we can apply a Kruskal-Wallis approach.

7-3 Non-parametric Correlation: Spearman correlation coefficient

The truth is that Pearson coefficient behaves pretty well even if data are not normal. Spearman is used mostly for ranked data. Spearman's correlation determines the strength and direction of the monotonic relationship between your two variables rather than the strength and direction of the linear relationship between your two variables, which is what Pearson's correlation determines.

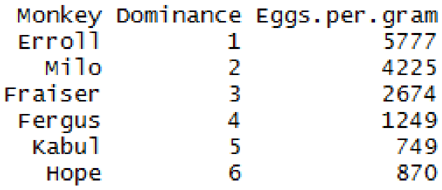
The formula for ρ (rho, the equivalent of r) is the same for Pearson and Spearman, expect that for Spearman the values are ranks.

Example (File: dominance.csv)

After determining the dominance rankings in a group of 6 male colobus monkeys, Melfi and Poyser (2007) counted eggs of *Trichuris* nematodes per gram of monkey faeces, a measurement variable.

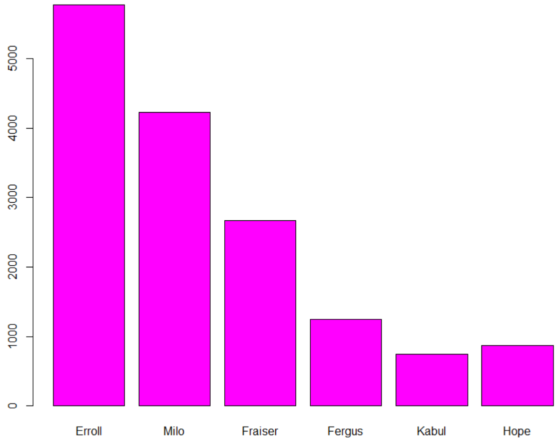
They wanted to know whether social dominance was associated with the number of nematode eggs, so they converted eggs per gram of faeces to ranks and used Spearman rank correlation.

```
dominance <- read.csv("dominance.csv")
head(dominance)
```

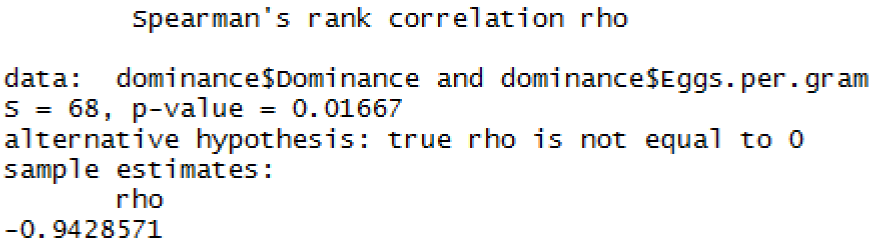


```
  Monkey Dominance Eggs.per.gram
  Erroll         1          5777
    Milo         2          4225
Fraiser          3          2674
  Fergus         4          1249
   Kabul         5           749
    Hope         6           870
```

```
barplot(dominance$eggs.per.gram, col="magenta", names.arg = dominance$monkey)
```



```
cor.test(dominance$dominance,dominance$eggs.per.gram, method = "spearman")
```



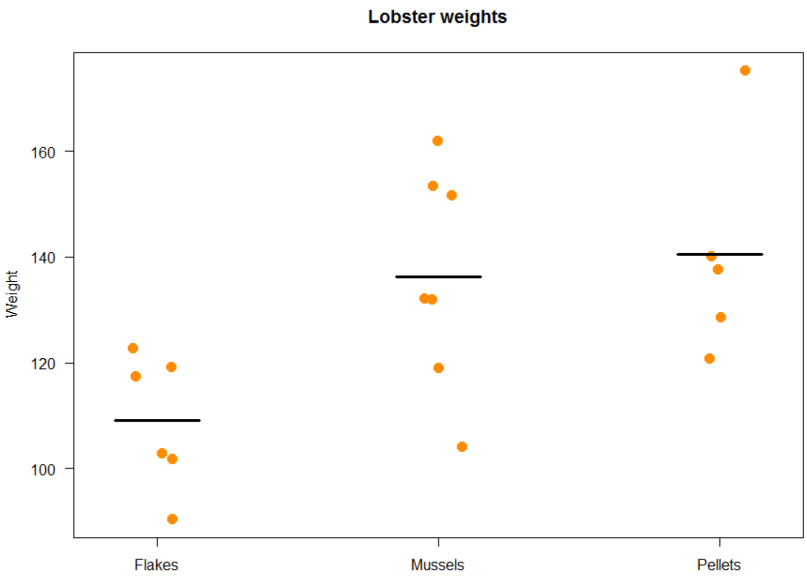
```
        Spearman's rank correlation rho

data:  dominance$Dominance and dominance$Eggs.per.gram
S = 68, p-value = 0.01667
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
-0.9428571
```

We will almost never use a regression line for either description or prediction when you do Spearman rank correlation, so don't calculate the equivalent of a regression line.
As for the relationship between dominance and parasitism, it is significant (p=0.017) with high ranking males harbouring a heavier burden.

Chapter 8: Linear models

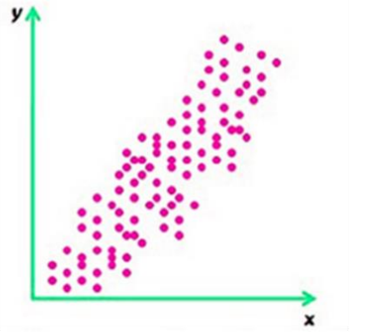
There is another way to look at data, another perspective for the analysis: linear modelling. Now, this approach should not be daunting in any way as linear modelling is really about language.

If we look at the lobster data below, there are 2 ways to formulate the question associated with it. Keeping up with what we have been doing so far, we can ask: Is there a difference between the 3 diets? If we want to take the linear model perspective, we then ask: Can diet predict weight? And this is pretty much all there is to it.

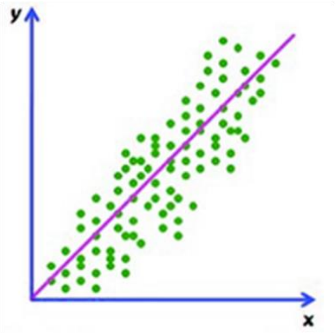


8-1: Simple linear model

One of the simplest versions of linear model is linear regression. We have talked about correlation already, and we know that it is about the association between 2 variables. Regression, on the other hand, not only addresses the question of association, it also asks whether one variable can be used to predict the values of the other.



Correlation = Association



Regression = Prediction

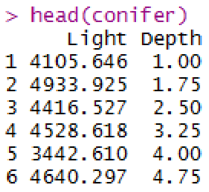
So linear regression models the dependence between 2 variables: a **dependent x** and **an independent y**.

response predictor

$$y = \beta_0 + \beta_1 * x$$

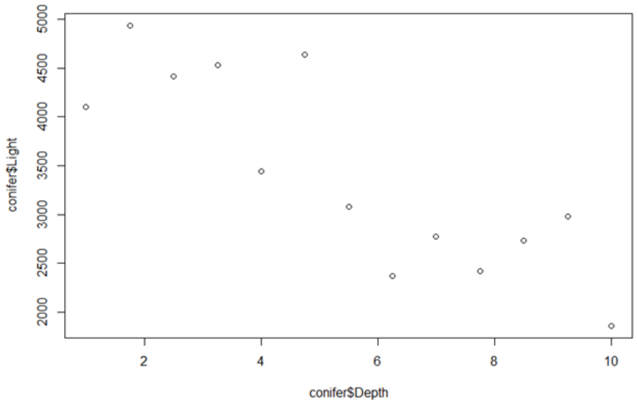
Example: (File: `coniferlight.csv`)

Let's take an example. In an experiment that looks at light intensity in woodland, how is light intensity (cts: lux) affected by the height at which the measurement is taken, recorded as depth measured from the top of the canopy (cts:m).



```
> head(conifer)
     Light Depth
1 4105.646  1.00
2 4933.925  1.75
3 4416.527  2.50
4 4528.618  3.25
5 3442.610  4.00
6 4640.297  4.75
```

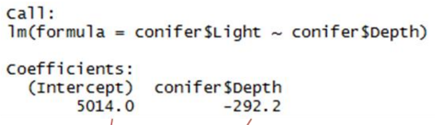
```
plot(conifer$Light~conifer$Depth)
```



In R, we can start with the `lm()` function.

```
lm(conifer$Light~conifer$Depth)
```

If we run this command, we will get the output below:



```
Call:
lm(formula = conifer$Light ~ conifer$Depth)

Coefficients:
  (Intercept)  conifer$Depth
       5014.0         -292.2
```

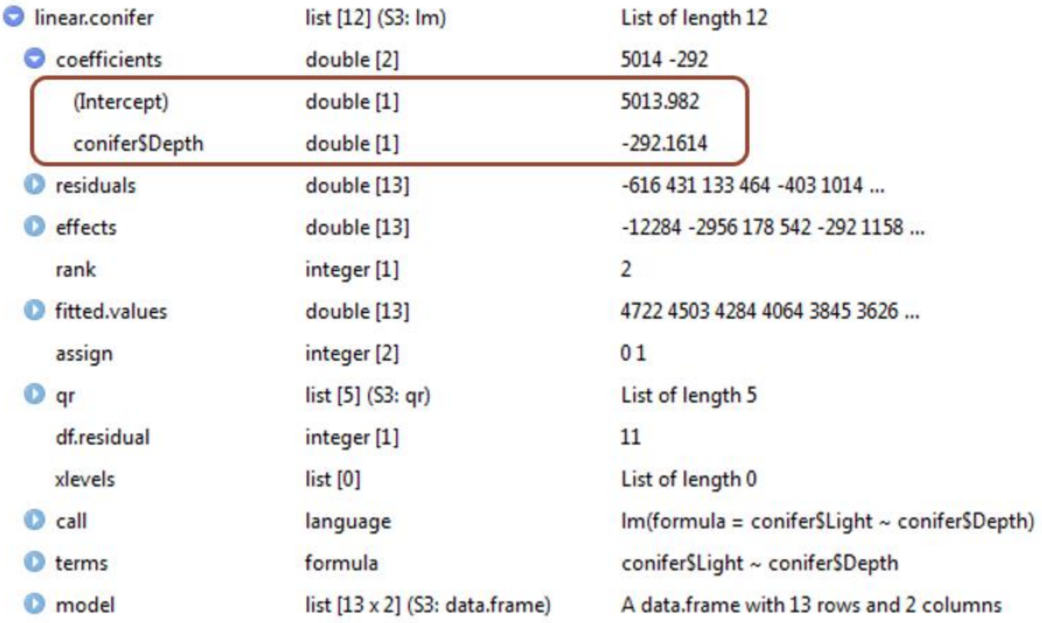
$$light = \beta_0 + \beta_1 * depth$$

So we get: **light = 5014 - 292*depth**

We are going to use linear models to do a little bit of navigation in R. So if we store our analysis in an object, we can extract or use the information in it to do all sorts of cool stuff.

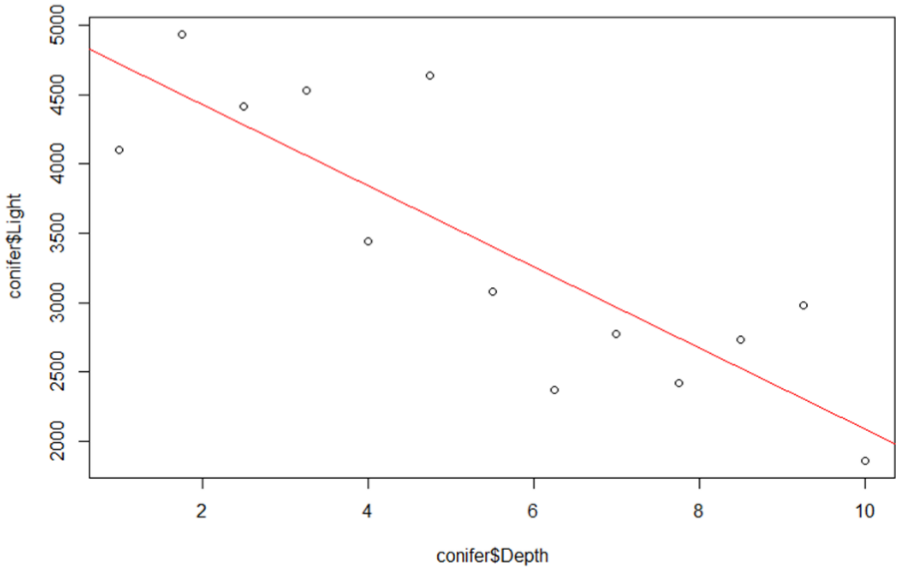
```
linear.conifer<-lm(conifer$Light~conifer$Depth)
```

If we now look into the Global Environment in R Studio (top right panel), we will get something like:



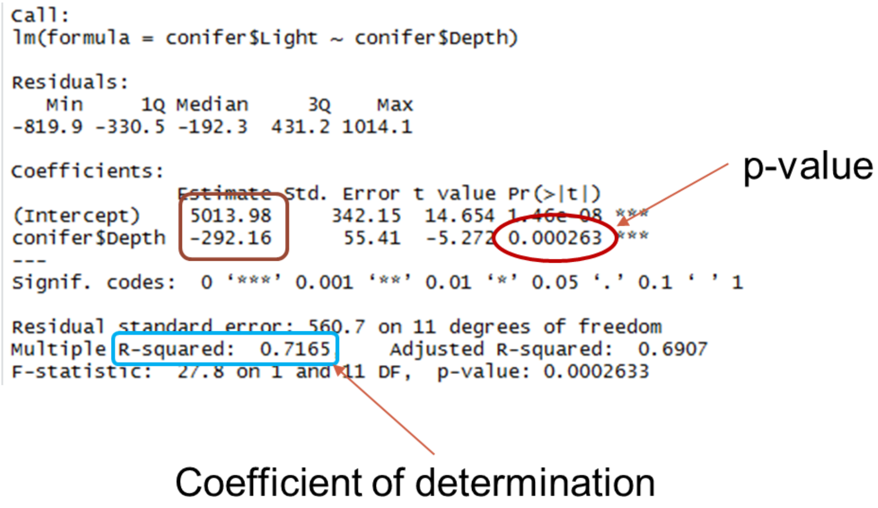
```
linear.conifer      list [12] (S3: lm)              List of length 12
  coefficients      double [2]                      5014 -292
    (Intercept)     double [1]                      5013.982
    conifer$Depth   double [1]                      -292.1614
  residuals         double [13]                     -616 431 133 464 -403 1014 ...
  effects           double [13]                     -12284 -2956 178 542 -292 1158 ...
  rank              integer [1]                     2
  fitted.values     double [13]                     4722 4503 4284 4064 3845 3626 ...
  assign            integer [2]                     0 1
  qr                list [5] (S3: qr)               List of length 5
  df.residual       integer [1]                     11
  xlevels           list [0]                        List of length 0
  call              language                        lm(formula = conifer$Light ~ conifer$Depth)
  terms             formula                         conifer$Light ~ conifer$Depth
  model             list [13 x 2] (S3: data.frame)  A data.frame with 13 rows and 2 columns
```

We recognise the 2 values that we used to write the linear model equation but there is a lot more. First, with the function `abline()`, we can ask R to draw a line of best-fit on top of our data. The function will use the coefficients values from `linear.conifer` to draw the line. How cool is that?



The relationship between light and depth appears negative and quite strong. Next, we may want to put a p-value on the relationship.

```
summary(lm.conifer)
```



```
Call:
lm(formula = conifer$Light ~ conifer$Depth)

Residuals:
   Min     1Q Median     3Q    Max 
-819.9 -330.5 -192.3  431.2 1014.1 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)    5013.98     342.15  14.654 1.46e-08 ***
conifer$Depth  -292.16      55.41  -5.272 0.000263 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 560.7 on 11 degrees of freedom
Multiple R-squared:  0.7165,     Adjusted R-squared:  0.6907 
F-statistic:  27.8 on 1 and 11 DF,  p-value: 0.0002633
```

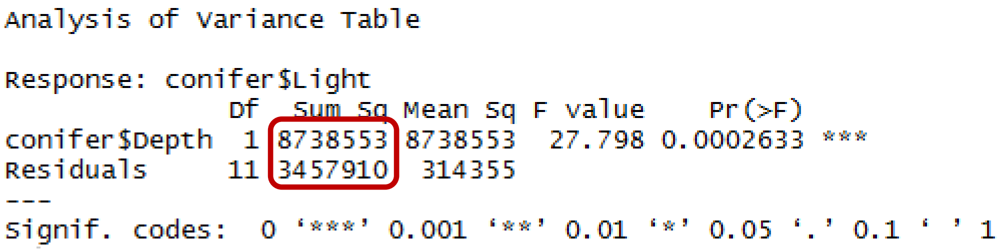
p-value

Coefficient of determination

In this output, we recognise the coefficients and we can also identify the p-value which is significant here. We also have access to the coefficient of determination. We mentioned it before and we know that it quantifies the proportion of variance in y that can be explained by x and it can be expressed as a percentage. So here, **71.65%** of the variability observed in light is explained by the depth at which it is measured in a conifer tree, which is quite a lot.

We can also get the results in an ANOVA-like way with `anova()`.

```
anova(linear.conifer)
```



```
Analysis of Variance Table

Response: conifer$Light
              Df  Sum Sq Mean Sq F value    Pr(>F)    
conifer$Depth  1 8738553 8738553  27.798 0.0002633 ***
Residuals     11 3457910  314355                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

This output allows us to get a better understanding about how R2 is calculated. The total amount of variability is 8738553 + 3457910 = 12196463. So the proportion explained by depth is 8738553/12196463 = **0.716**. Ta-dah!

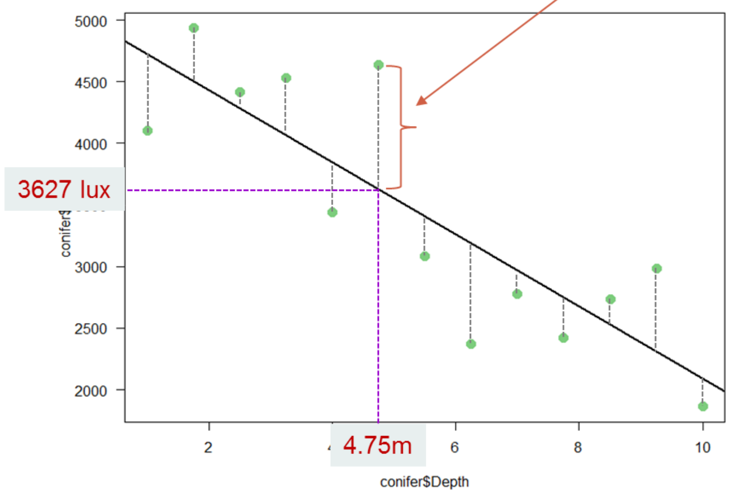
So depth predicts about 71% of the variability of light meaning that 29% is explained by other factors (e.g. individual variability …). For example, the model predicts 3627 lux at a depth of 4.75m in a conifer. There is an error associated with each prediction as the regression equation is actually:

$$y = \beta_0 + \beta_1 * x + \varepsilon$$

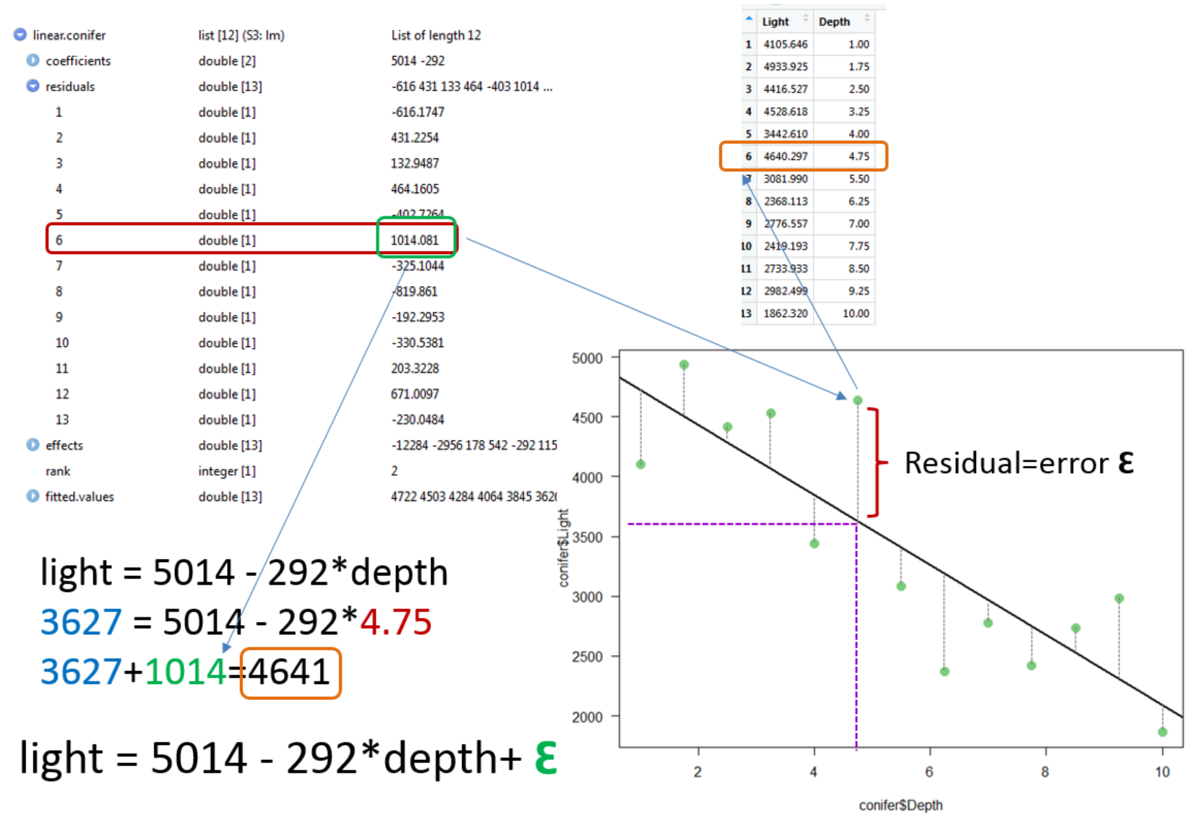
The error is often referred to as the residual, literally what is left after the prediction, what is left unexplained.

3627 = 5014 - 292*4.75

$$y = \beta_0 + \beta_1 * x + \varepsilon$$



If we go back to `linear.conifer`, we have access to these values and we can understand better what is happening.



```
linear.conifer      list [12] (S3: lm)   List of length 12
  coefficients      double [2]           5014 -292
  residuals         double [13]          -616 431 133 464 -403 1014 ...
    1               double [1]           -616.1747
    2               double [1]           431.2254
    3               double [1]           132.9487
    4               double [1]           464.1605
    5               double [1]           -402.7264
    6               double [1]           1014.081
    7               double [1]           -325.1044
    8               double [1]           -819.861
    9               double [1]           -192.2953
    10              double [1]           -330.5381
    11              double [1]           203.3228
    12              double [1]           671.0097
    13              double [1]           -230.0484
  effects           double [13]          -12284 -2956 178 542 -292 115
  rank              integer [1]          2
  fitted.values     double [13]          4722 4503 4284 4064 3845 362(
```

	Light	Depth
1	4105.646	1.00
2	4933.925	1.75
3	4416.527	2.50
4	4528.618	3.25
5	3442.610	4.00
6	4640.297	4.75
7	3081.990	5.50
8	2368.113	6.25
9	2776.557	7.00
10	2419.193	7.75
11	2733.933	8.50
12	2982.499	9.25
13	1862.320	10.00

light = 5014 - 292*depth

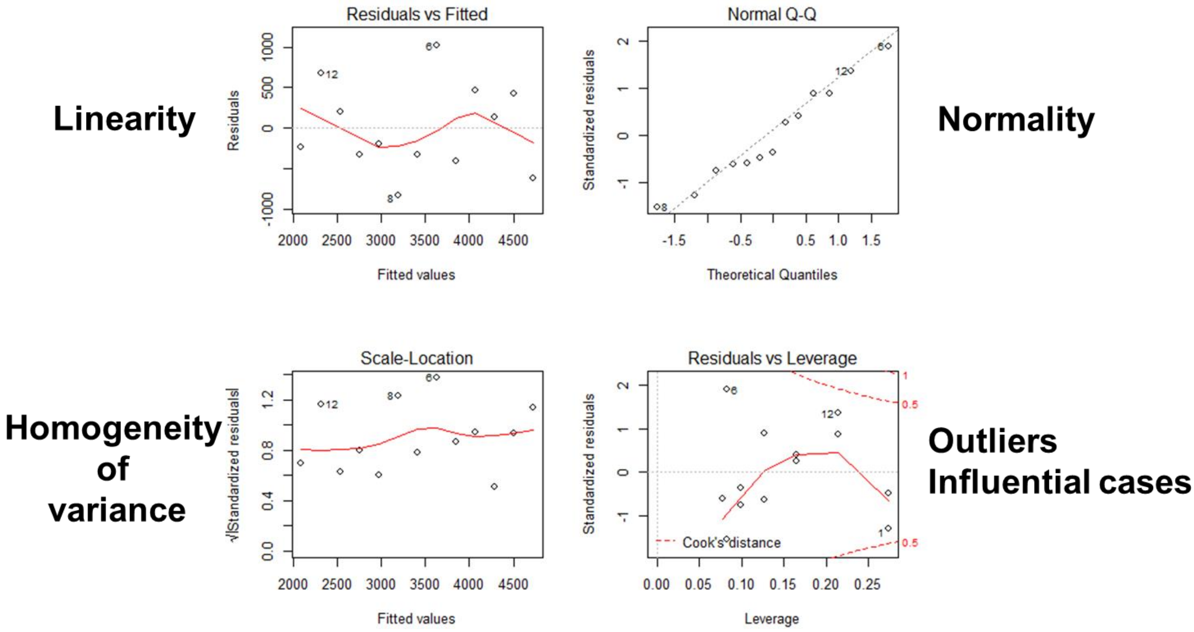
3627 = 5014 - 292*4.75

3627+1014=4641

light = 5014 - 292*depth+ ε

These residuals are also used to help us assess how good or bad our model is. We saw in the correlation chapter that we could check the assumptions and identify influential outliers, this is all thanks to the residuals. Once more we can query `linear.conifer()`.

```
par(mfrow=c(2,2))
plot(linear.conifer)
```

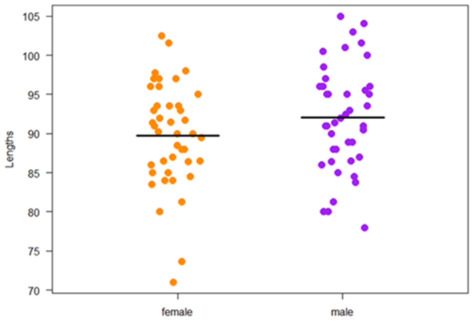


It is not perfect but there is nothing wrong either, we can accept the predictions of our model.

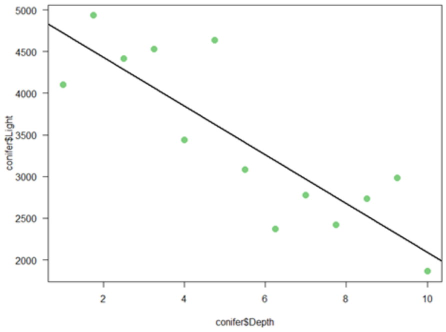
8-2: The linear model perspective

Comparing 2 groups

Now we are going to revisit some of the datasets we used already but with a linear model perspective. The first thing important to understand, is that it does not really matter is the predictor is categorical or continuous, the reasoning is pretty much the same.



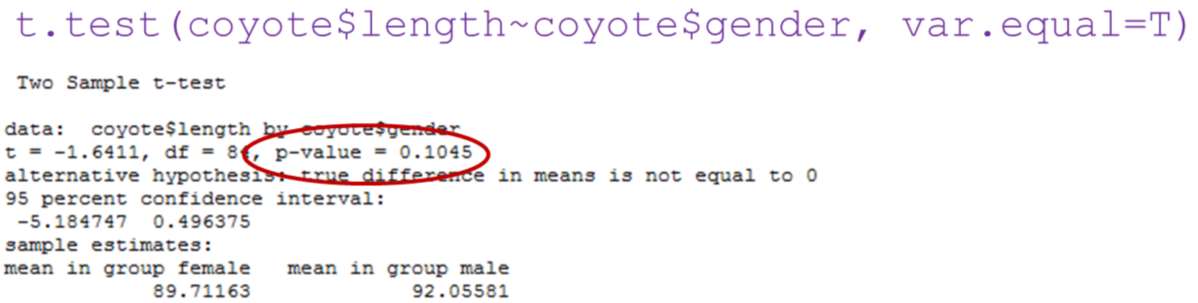
Categorical predictor



Continuous predictor

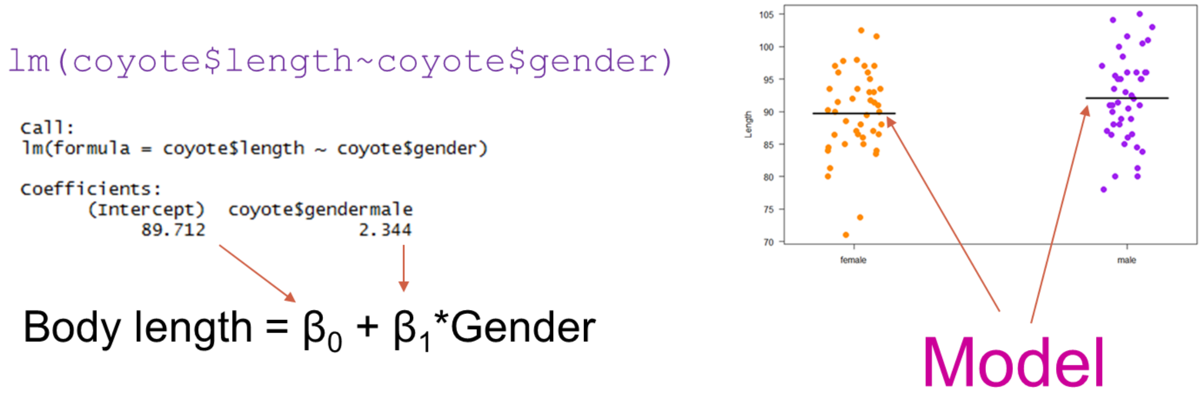
If we start with the coyotes, the question we asked before was: Is there a difference between the 2 genders? In the linear model world, it becomes: Does gender predict body length? But it is really the same thing, isn't it? So again, linear modelling is all about language and vocabulary.

Back in Chapter 3, we did run a t-test to quantify the relationship between gender and body length.



```
t.test(coyote$length~coyote$gender, var.equal=T)

 Two Sample t-test

data:  coyote$length by coyote$gender
t = -1.6411, df = 84, p-value = 0.1045
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -5.184747  0.496375
sample estimates:
mean in group female   mean in group male 
            89.71163             92.05581 
```

Let's try linear model.



```
lm(coyote$length~coyote$gender)

Call:
lm(formula = coyote$length ~ coyote$gender)

Coefficients:
  (Intercept)  coyote$gendermale  
       89.712              2.344  
```

$$\text{Body length} = \beta_0 + \beta_1{*}\text{Gender}$$

Model

So, according to our model, females are expected to measure 89.712 cm and males are expected to be 2.344 cm longer 89.712 + 2.344 = 92.05. The model will use the baseline as the first alphabetical category unless told otherwise .Our means are indeed our model. We can formulate it in different ways:

$$\text{Body Length} = \begin{pmatrix} 89.71 \\ 92.06 \end{pmatrix} \begin{pmatrix} \text{If Female} \\ \text{If Male} \end{pmatrix}$$

Or:

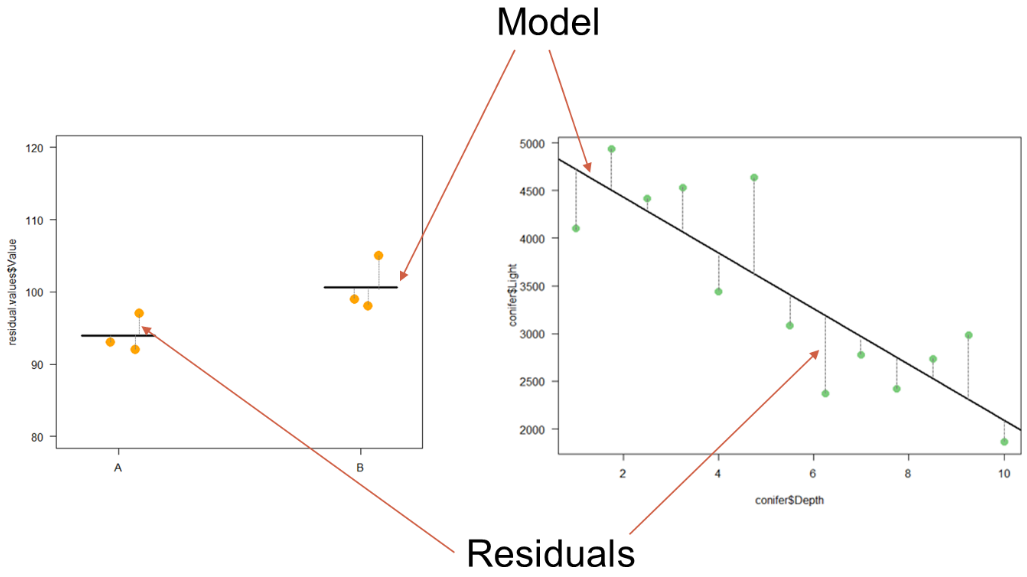
$$\text{Body Length} = 89.71 \begin{pmatrix} 0 \\ 2.344 \end{pmatrix} \begin{pmatrix} \text{If Female} \\ \text{If Male} \end{pmatrix}$$

Finally:

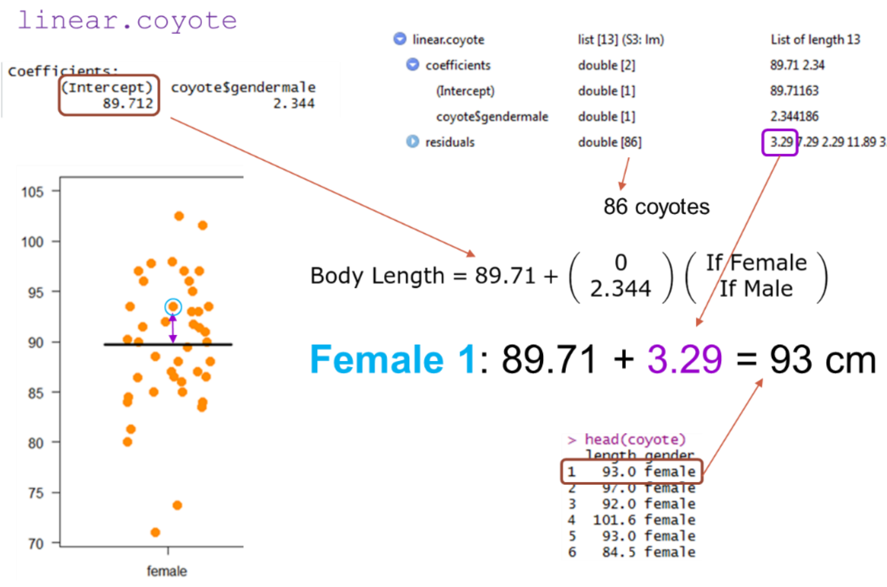
Body length = 89.712 + 2.344*Gender

We can see that the only real difference between a linear model where the predictor is continuous and one where the predictor is categorical is that in the former the coefficient is a number (e.g. the conifer example) whereas in the latter, it is a vector, here: (89.71, 92.06).

In both cases, the residuals are what is left after the prediction.



And these residuals can be used in a similar way in the coyote example as we used them in the conifer one.



```
linear.coyote
```

```
Coefficients:
(Intercept)  coyote$gendermale
     89.712              2.344
```

```
linear.coyote          list [13] (S3: lm)    List of length 13
  coefficients         double [2]            89.71 2.34
    (Intercept)        double [1]            89.71163
    coyote$gendermale  double [1]            2.344186
  residuals            double [86]           3.29 7.29 2.29 11.89 3.29 -5.21 ...
```

86 coyotes

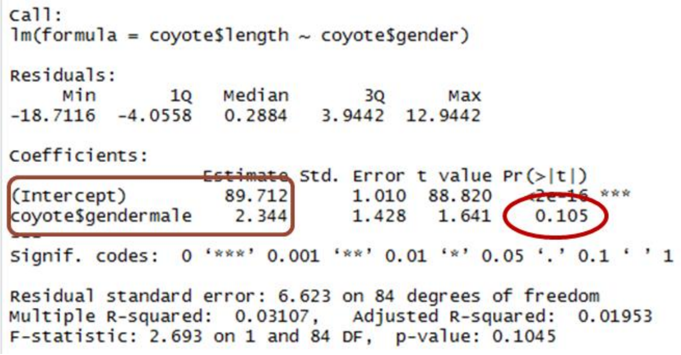
$$\text{Body Length} = 89.71 + \begin{pmatrix} 0 \\ 2.344 \end{pmatrix} \begin{pmatrix} \text{If Female} \\ \text{If Male} \end{pmatrix}$$

Female 1: 89.71 + 3.29 = 93 cm

```
> head(coyote)
  length gender
1   93.0 female
2   97.0 female
3   92.0 female
4  101.6 female
5   93.0 female
6   84.5 female
```

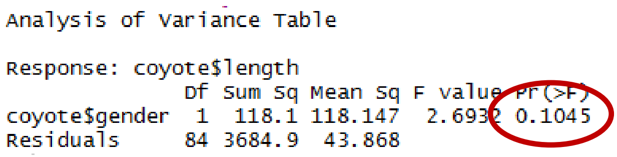
Finally, we can use the 2 functions `summary()` and `anova()` as we did with the conifer data.

```
summary(linear.coyote)
```



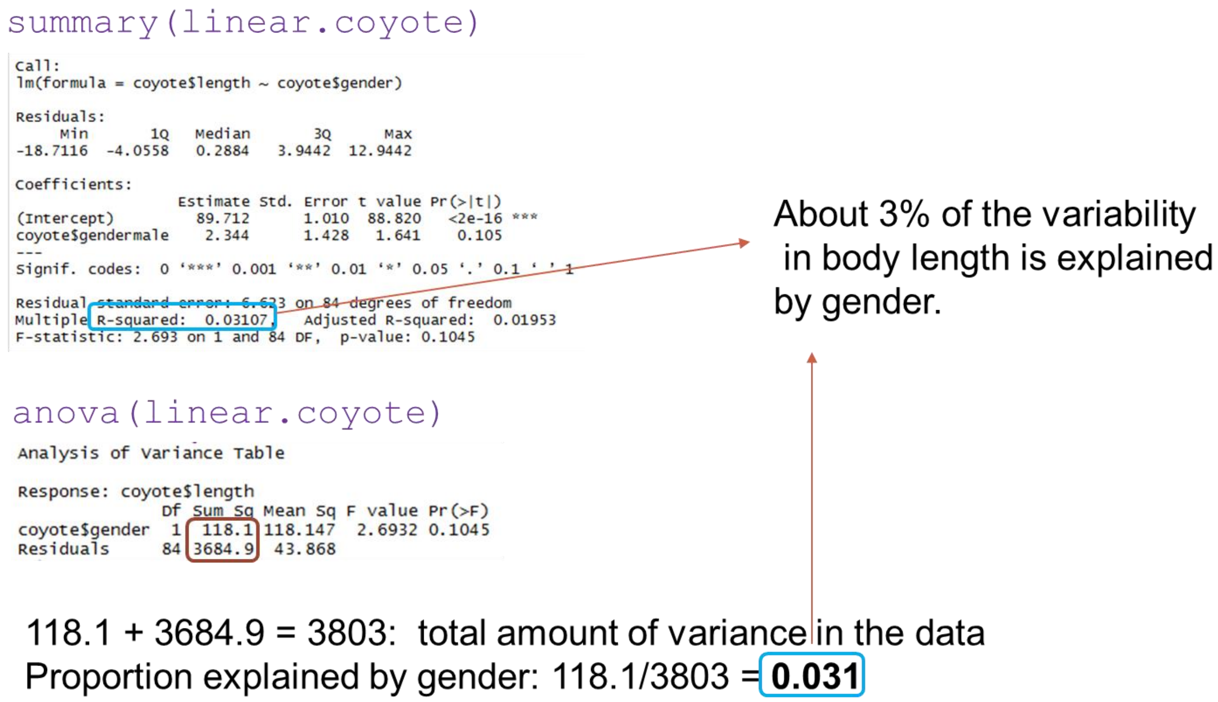
```
Call:
lm(formula = coyote$length ~ coyote$gender)

Residuals:
     Min       1Q   Median       3Q      Max
-18.7116  -4.0558   0.2884   3.9442  12.9442

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)         89.712      1.010  88.820   <2e-16 ***
coyote$gendermale    2.344      1.428   1.641    0.105
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.623 on 84 degrees of freedom
Multiple R-squared:  0.03107,	Adjusted R-squared:  0.01953 
F-statistic: 2.693 on 1 and 84 DF,  p-value: 0.1045
```

```
anova(linear.coyote)
```



```
Analysis of Variance Table

Response: coyote$length
              Df Sum Sq Mean Sq F value Pr(>F)
coyote$gender  1  118.1 118.147  2.6932 0.1045
Residuals     84 3684.9  43.868
```

And like for the conifer data, we can quantify how predictive our model is thanks to R^2.



```
summary(linear.coyote)
```

```
Call:
lm(formula = coyote$length ~ coyote$gender)

Residuals:
     Min       1Q   Median       3Q      Max 
-18.7116  -4.0558   0.2884   3.9442  12.9442 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)         89.712      1.010  88.820   <2e-16 ***
coyote$gendermale    2.344      1.428   1.641    0.105    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.623 on 84 degrees of freedom
Multiple R-squared:  0.03107,	Adjusted R-squared:  0.01953 
F-statistic: 2.693 on 1 and 84 DF,  p-value: 0.1045
```

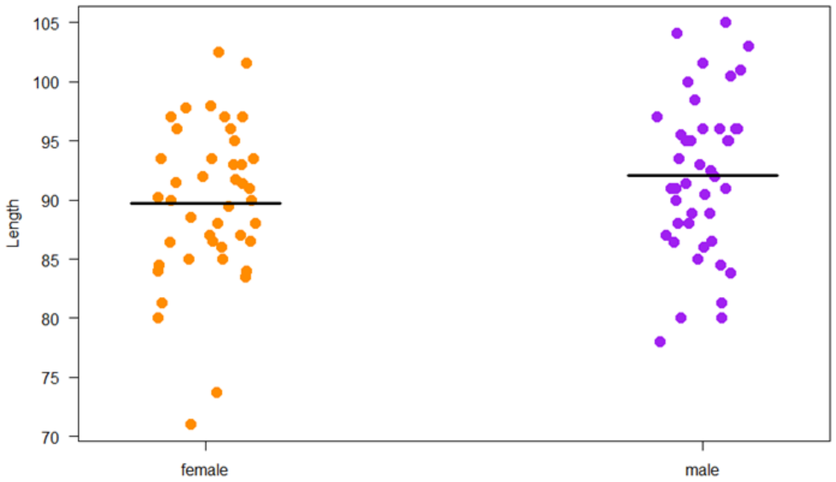
```
anova(linear.coyote)
```

```
Analysis of Variance Table

Response: coyote$length
              Df Sum Sq Mean Sq F value Pr(>F)
coyote$gender  1  118.1 118.147  2.6932 0.1045
Residuals     84 3684.9  43.868
```

About 3% of the variability in body length is explained by gender.

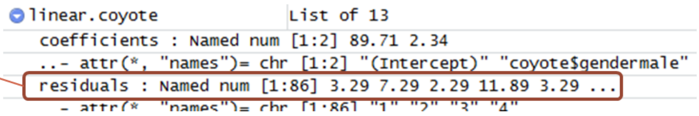
118.1 + 3684.9 = 3803: total amount of variance in the data
Proportion explained by gender: 118.1/3803 = **0.031**

The fact that the model explains only 3% makes sense when we look at the data: even though males are longer than females, there is a lot more about body lengths than gender; a lot as in … 97%.



Finally, rather than using the Shapiro-Wilk and the Levene tests to check for the assumptions, we can use the graphical diagnostics. We can do that because just like for the conifer model, we have the residuals.

linear.coyote



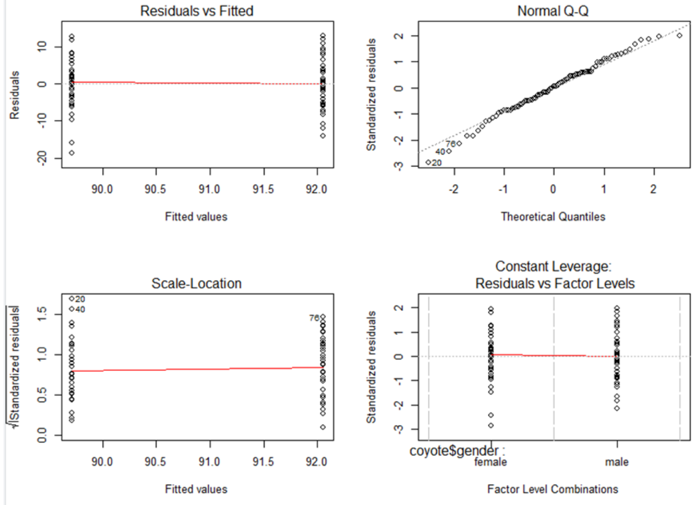
```
linear.coyote                    List of 13
  coefficients : Named num [1:2] 89.71 2.34
  ..- attr(*, "names")= chr [1:2] "(Intercept)" "coyote$gendermale"
  residuals : Named num [1:86] 3.29 7.29 2.29 11.89 3.29 ...
  - attr(*, "names")= chr [1:86] "1" "2" "3" "4"
```

Assumptions

```
par(mfrow=c(2,2))
plot(linear.coyote)
```

Linearity

Normality
~ shapiro.test()



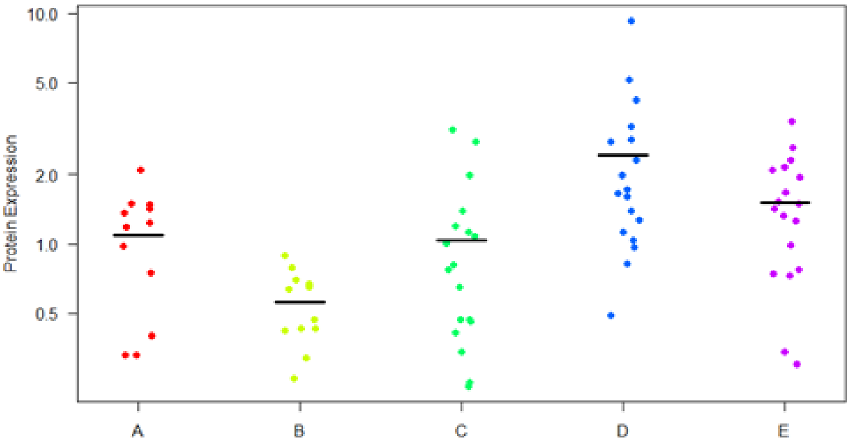
Equality of Variance
leveneTest()

Outliers

So to conclude on this analysis, we can say that gender does not significantly predict body length in coyotes (p=0.105) and it actually only predicts 3% of the variation.

Comparing more than 2 groups: one factor

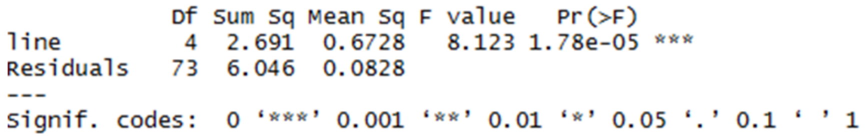
Let’s revisit the protein expression data with the linear model perspective. If we plot the data from `protein.stack.clean.csv`, we get the graph below.



When we first looked at the data we asked: Is there a difference in protein expression between the 5 cell lines? Now we can ask: Do the cell lines predict protein expression and how much of it if they do?

Previously, we got the results below:

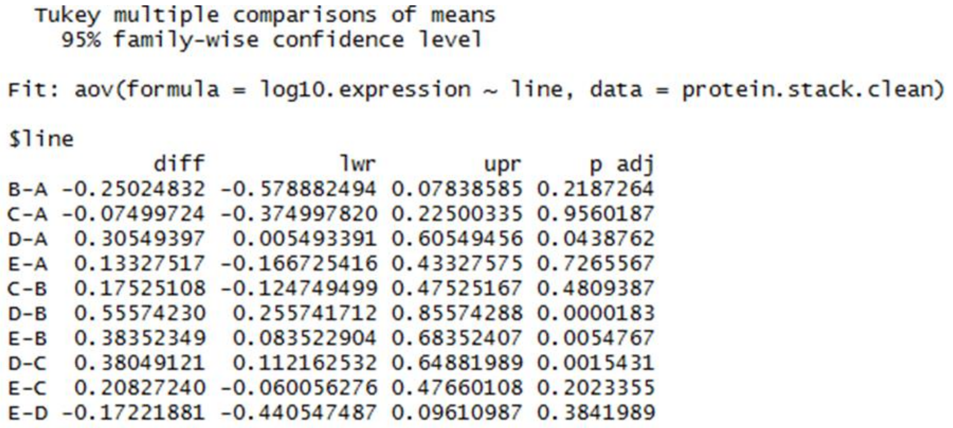
```
anova.log.protein<-aov(log10.expression~line,data=protein.stack.clean)
summary(anova.log.protein)
```



```
            Df Sum Sq Mean Sq F value   Pr(>F)    
line         4  2.691  0.6728   8.123 1.78e-05 ***
Residuals   73  6.046  0.0828                     
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

So there was overall a cell line effect. Next, we looked at the pairwise comparisons.

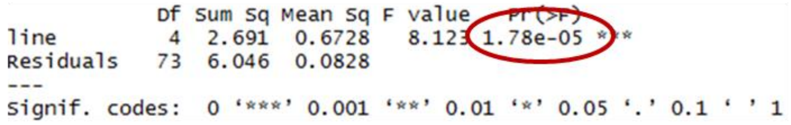
```
    TukeyHSD(anova.log.protein,"line")
```



```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = log10.expression ~ line, data = protein.stack.clean)

$line
           diff          lwr        upr     p adj
B-A -0.25024832 -0.578882494 0.07838585 0.2187264
C-A -0.07499724 -0.374997820 0.22500335 0.9560187
D-A  0.30549397  0.005493391 0.60549456 0.0438762
E-A  0.13327517 -0.166725416 0.43327575 0.7265567
C-B  0.17525108 -0.124749499 0.47525167 0.4809387
D-B  0.55574230  0.255741712 0.85574288 0.0000183
E-B  0.38352349  0.083522904 0.68352407 0.0054767
D-C  0.38049121  0.112162532 0.64881989 0.0015431
E-C  0.20827240 -0.060056276 0.47660108 0.2023355
E-D -0.17221881 -0.440547487 0.09610987 0.3841989
```

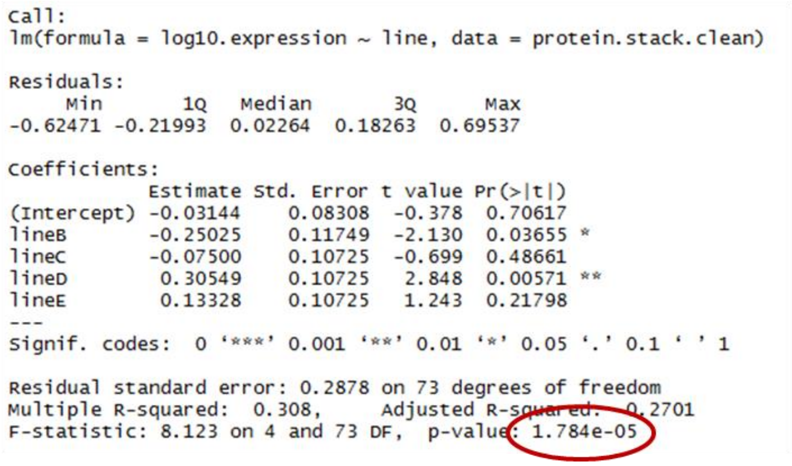
If we take the linear model approach, we get.

```
linear.protein<-lm(log10.expression~line,data=protein.stack.clean)

  anova(linear.protein)
```



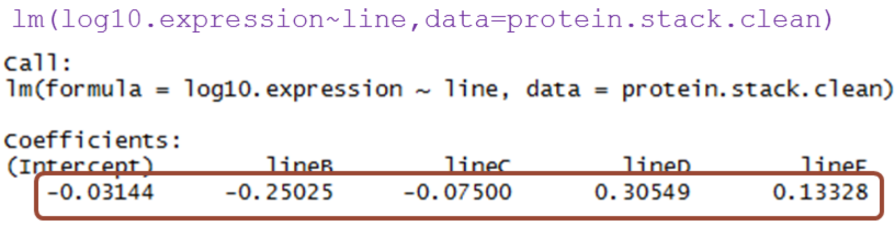
```
            Df Sum Sq Mean Sq F value   Pr(>F)    
line         4  2.691  0.6728   8.123 1.78e-05 ***
Residuals   73  6.046  0.0828                     
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
summary(linear.protein)
```



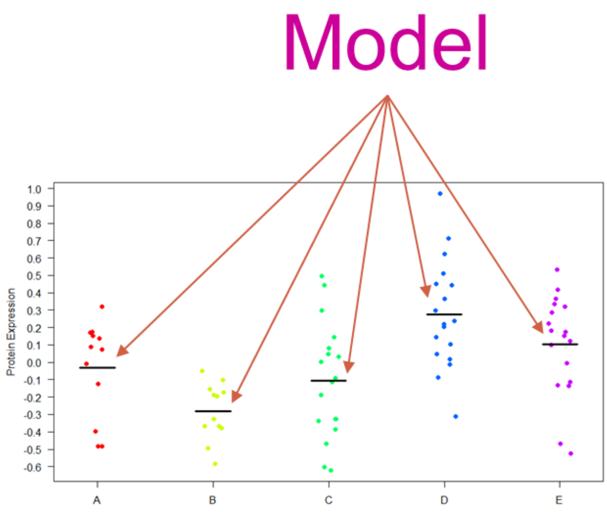
```
Call:
lm(formula = log10.expression ~ line, data = protein.stack.clean)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.62471 -0.21993  0.02264  0.18263  0.69537 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)   
(Intercept) -0.03144    0.08308  -0.378  0.70617   
lineB       -0.25025    0.11749  -2.130  0.03655 * 
lineC       -0.07500    0.10725  -0.699  0.48661   
lineD        0.30549    0.10725   2.848  0.00571 **
lineE        0.13328    0.10725   1.243  0.21798   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2878 on 73 degrees of freedom
Multiple R-squared:  0.308,	Adjusted R-squared:  0.2701 
F-statistic: 8.123 on 4 and 73 DF,  p-value: 1.784e-05
```

Now if we want to identify the coefficients of the model as we did for the coyotes, we can get them from the output above or we can go:

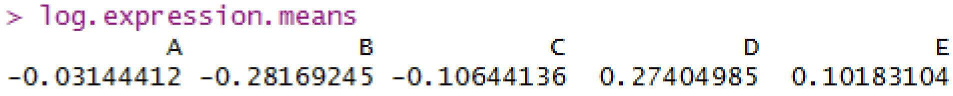


```
lm(log10.expression~line,data=protein.stack.clean)

Call:
lm(formula = log10.expression ~ line, data = protein.stack.clean)

Coefficients:
(Intercept)        lineB        lineC        lineD        lineE
   -0.03144     -0.25025     -0.07500      0.30549      0.13328
```

So far, so good. Now, the means of the 5 cell lines are the model's predictions.



In R, we can get the means:

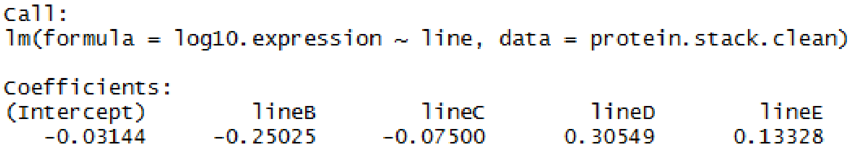


```
> log.expression.means
          A           B           C           D           E
-0.03144412 -0.28169245 -0.10644136  0.27404985  0.10183104
```

We can then formulate these values in a linear model:

$$\text{Expression} = \begin{pmatrix} -0.03144 \\ -0.28169 \\ -0.10644 \\ 0.327405 \\ 0.10183 \end{pmatrix} \begin{pmatrix} \text{Line A} \\ \text{Line B} \\ \text{Line C} \\ \text{Line D} \\ \text{Line E} \end{pmatrix}$$

Now if we look at the coefficients again:



```
Call:
lm(formula = log10.expression ~ line, data = protein.stack.clean)

Coefficients:
(Intercept)        lineB        lineC        lineD        lineE
   -0.03144     -0.25025     -0.07500      0.30549      0.13328
```

So the model is:

$$\text{Expression} = \beta_0 + \beta_1 * \text{Line}$$

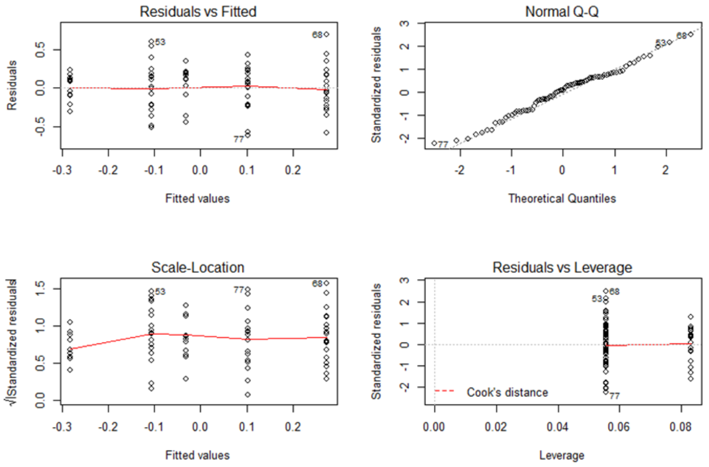
$$\text{Expression} = -0.03144 + \begin{pmatrix} 0 \\ -0.25025 \\ -0.07500 \\ 0.30549 \\ 0.13328 \end{pmatrix} \begin{pmatrix} \text{Line A} \\ \text{Line B} \\ \text{Line C} \\ \text{Line D} \\ \text{Line E} \end{pmatrix}$$

And it works beautifully. For example:

Line B = -0.03-0.25 = -0.28

And we can check graphically, as we did before if the model is to be trusted.

```
par(mfrow=c(2,2))
plot(linear.protein)
```



Linearity

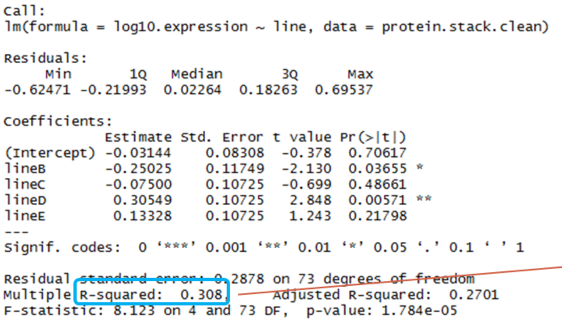
Normality `shapiro.test()`

Equality of Variance `leveneTest()`

Outliers

As for the coefficient of determination, we can get it as we did with the coyote data.

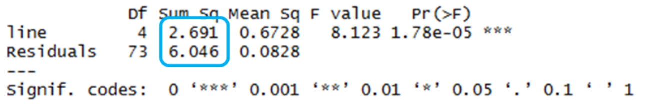
```
linear.protein<-lm(log10.expression~line,data=protein.stack.clean)
summary(linear.protein)
```



```
Call:
lm(formula = log10.expression ~ line, data = protein.stack.clean)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.62471 -0.21993  0.02264  0.18263  0.69537 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)   
(Intercept) -0.03144    0.08308  -0.378  0.70617   
lineB       -0.25025    0.11749  -2.130  0.03655 * 
lineC       -0.07500    0.10725  -0.699  0.48661   
lineD        0.30549    0.10725   2.848  0.00571 **
lineE        0.13328    0.10725   1.243  0.21798   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2878 on 73 degrees of freedom
Multiple R-squared:  0.308,	Adjusted R-squared:  0.2701 
F-statistic: 8.123 on 4 and 73 DF,  p-value: 1.784e-05
```

Proportion of variance explained by cell lines: 31%

```
anova.log.protein<-aov(log10.expression~line,data=protein.stack.clean)
summary(anova.log.protein)
```



```
            Df Sum Sq Mean Sq F value   Pr(>F)    
line         4  2.691  0.6728   8.123 1.78e-05 ***
Residuals   73  6.046  0.0828                     
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

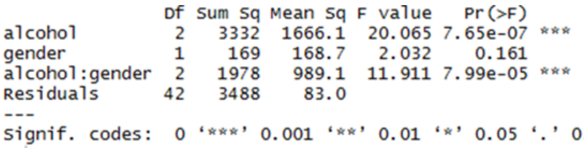
2.691 + 6.046 = 8.737: total amount of variance in the data
Proportion explained by gender: 2.691/8.737 = **0.308**

So to conclude, cell lines do significantly predict protein expression and explain about 31% of its variation. Which is not bad but it also means that almost 70% is explained by some other factors.

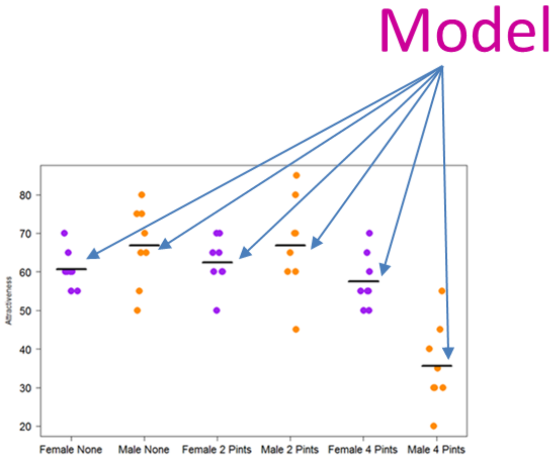
Comparing more than 2 groups: two factors

We can also revisit the goggles data in a similar way. Previously, we got:

```
fit.goggles<-aov(attractiveness~alcohol+gender+alcohol*gender,data=goggles)
summary(fit.goggles)
```

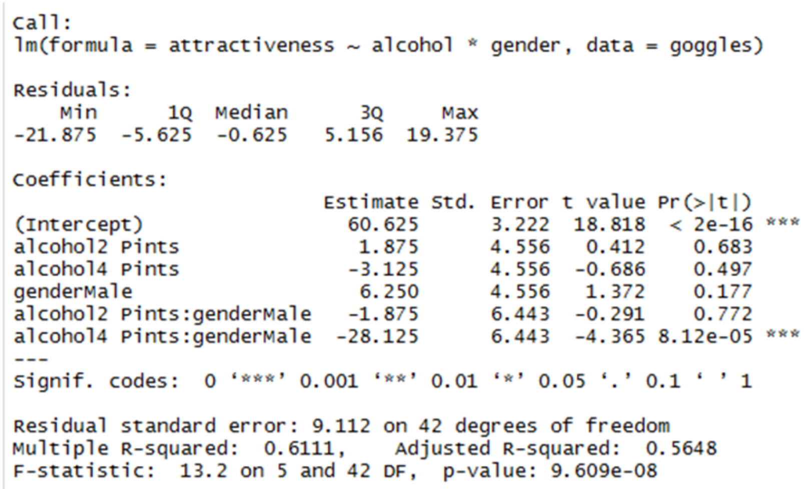


```
               Df Sum Sq Mean Sq F value   Pr(>F)    
alcohol         2   3332  1666.1  20.065 7.65e-07 ***
gender          1    169   168.7   2.032    0.161    
alcohol:gender  2   1978   989.1  11.911 7.99e-05 ***
Residuals      42   3488    83.0                     
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```



Now the way the model is built mirrors the way we write it in the function:

```
linear.goggles<-lm(attractiveness~alcohol+gender+alcohol*gender,data=goggles)
summary(linear.goggles)
```



```
Call:
lm(formula = attractiveness ~ alcohol * gender, data = goggles)

Residuals:
    Min      1Q  Median      3Q     Max 
-21.875  -5.625  -0.625   5.156  19.375 

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)    
(Intercept)                 60.625      3.222  18.818  < 2e-16 ***
alcohol2 Pints               1.875      4.556   0.412    0.683    
alcohol4 Pints              -3.125      4.556  -0.686    0.497    
genderMale                   6.250      4.556   1.372    0.177    
alcohol2 Pints:genderMale   -1.875      6.443  -0.291    0.772    
alcohol4 Pints:genderMale  -28.125      6.443  -4.365 8.12e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.112 on 42 degrees of freedom
Multiple R-squared:  0.6111,	Adjusted R-squared:  0.5648 
F-statistic:  13.2 on 5 and 42 DF,  p-value: 9.609e-08
```

First we deal with the intercept, then each of the single and finally the interaction term.

Attractiveness= β_0 + β_1Alcohol + β_2Gender + β_3Gender*Alcohol

It becomes:

$$\text{Attractiveness} = 60.625 + \begin{pmatrix} 0 \\ 1.875 \\ -3.125 \end{pmatrix} \begin{pmatrix} \text{if None} \\ \text{if 2 Pints} \\ \text{if 4 Pints} \end{pmatrix} + \begin{pmatrix} 0 \\ 6.250 \end{pmatrix} \begin{pmatrix} \text{if Female} \\ \text{if Male} \end{pmatrix} + \begin{pmatrix} 0 \\ -1.875 \\ -28.125 \end{pmatrix} \begin{pmatrix} \text{Otherwise} \\ \text{if Male and 2 Pints} \\ \text{if male and 4 Pints} \end{pmatrix}$$

So a male student who has drunk 4 pints, is predicted to be chatting with a date whose attractiveness score is:

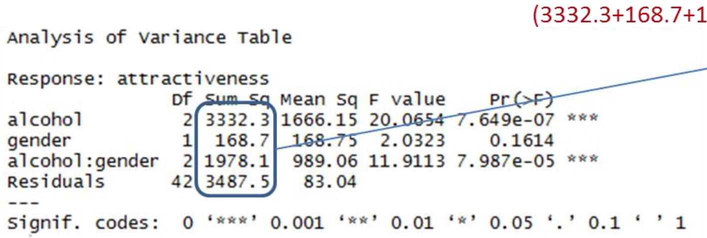
Attractiveness = 60.625 – 3.125 + 6.250 – 28.125 = 35.625

And we can calculate, R^2 as we did before.

```
linear.goggles<-lm(attractiveness~alcohol+gender+alcohol*gender,data=goggles)
anova(linear.goggles)
```

(3332.3+168.7+1978.1)/(3332.3+168.7+1978.1+3487.5) = 0.611

R^2 = 61%



```
Analysis of Variance Table

Response: attractiveness
               Df Sum Sq Mean Sq F value    Pr(>F)    
alcohol         2 3332.3 1666.15 20.0654 7.649e-07 ***
gender          1  168.7  168.75  2.0323    0.1614    
alcohol:gender  2 1978.1  989.06 11.9113 7.987e-05 ***
Residuals      42 3487.5   83.04                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

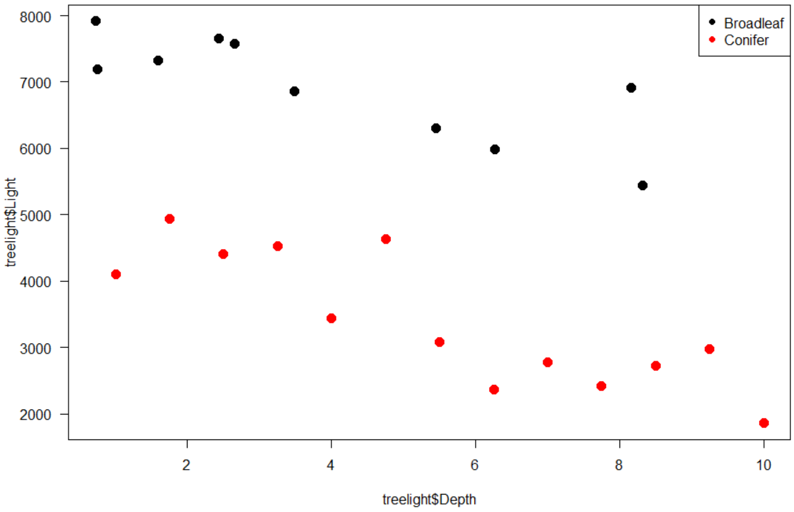
We can then conclude as we did before: there is a significant effect of alcohol consumption on the way the attractiveness of a date is perceived but it varies significantly between genders (p=7.99e-05). With 2 pints or less, boys seem to be very slightly more picky about their date than girls (but not significantly so) but with 4 pints the difference is reversed and significant (p=0.0003).

Comparing more than 2 groups: two factors: one continuous and one categorical

The last thing we are going to cover is the case of models which contain both continuous and categorical predictors. There is nothing particular about it actually, apart from that R does not make it easy for those models to be plotted. So let’s do it together.

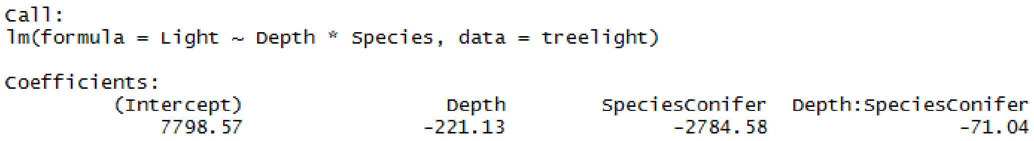
We are going to go back to the treelight data but this time we are going to add another species and thus add a categorical predictor: tree species.

```
treelight<-read.csv("treelight.csv")

plot(treelight$Light~treelight$Depth,col=treelight$Species,pch=16, cex=1.5)
legend("topright", c("Broadleaf", "Conifer"), pch=16, col=1:2)
```



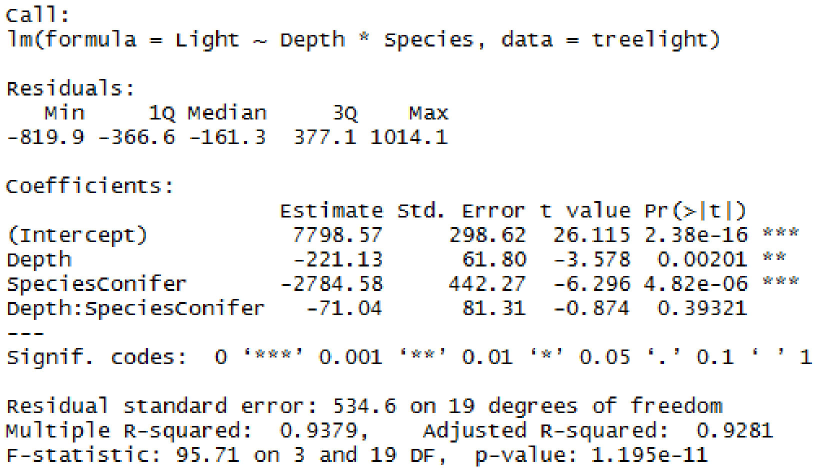
We get the coefficients of the model the usual way:

```
lm(Light~Depth*Species, data=treelight)
```



```
Call:
lm(formula = Light ~ Depth * Species, data = treelight)

Coefficients:
          (Intercept)                  Depth         SpeciesConifer  Depth:SpeciesConifer
              7798.57                -221.13               -2784.58                -71.04
```

```
linear.treelight<-lm(Light~Depth*Species, data=treelight)
summary(linear.treelight)
```

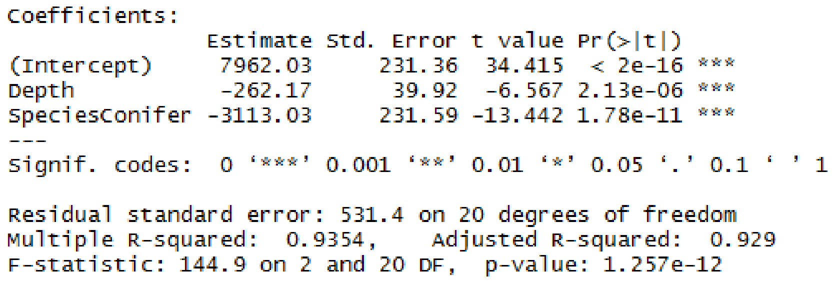


```
Call:
lm(formula = Light ~ Depth * Species, data = treelight)

Residuals:
   Min     1Q Median     3Q    Max 
-819.9 -366.6 -161.3  377.1 1014.1 

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)           7798.57     298.62  26.115 2.38e-16 ***
Depth                 -221.13      61.80  -3.578  0.00201 ** 
SpeciesConifer       -2784.58     442.27  -6.296 4.82e-06 ***
Depth:SpeciesConifer   -71.04      81.31  -0.874  0.39321    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 534.6 on 19 degrees of freedom
Multiple R-squared:  0.9379,	Adjusted R-squared:  0.9281 
F-statistic: 95.71 on 3 and 19 DF,  p-value: 1.195e-11
```

So far, nothing special. We can see which terms are significant and how much of the variability in light is explained by the model: 93.8%, which is definitely a lot.

Now we notice that the interaction is not significant (p=0.39), so it would make sense to remove it from the model since it probably does not improve by much. Then, from the complete model, we go to what is called the additive model, as we just consider the addition of the single effect terms.

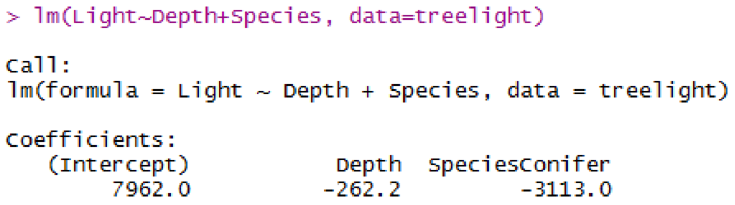
```
linear.treelight.add<-lm(Light~Depth+Species, data=treelight)
summary(linear.treelight.add)
```



```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)     7962.03     231.36  34.415  < 2e-16 ***
Depth           -262.17      39.92  -6.567 2.13e-06 ***
SpeciesConifer -3113.03     231.59 -13.442 1.78e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 531.4 on 20 degrees of freedom
Multiple R-squared:  0.9354,	Adjusted R-squared:  0.929 
F-statistic: 144.9 on 2 and 20 DF,  p-value: 1.257e-12
```

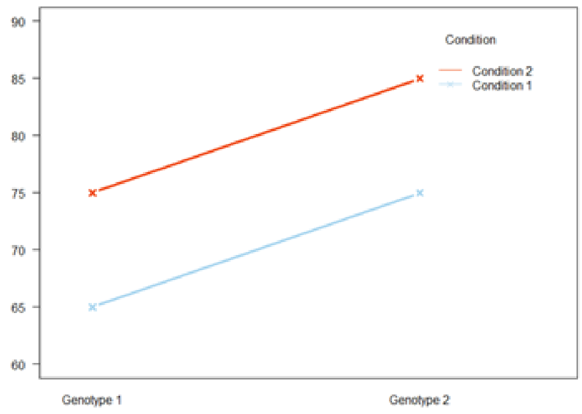
We can see that R-squared is pretty much the same, 93.5%, so the interaction did not make much of a difference. There are other ways to test that and to build models but what we just did works out nicely here.

Now, the next thing that we probably want to do is plot our model and that’s where things get nasty. The function `abline()` we used before wants a vector of 2 values, one for the intercept and one for the slope and the problem here is that we have 3 values. So `abline()` is confused.



```
> lm(Light~Depth+Species, data=treelight)

Call:
lm(formula = Light ~ Depth + Species, data = treelight)

Coefficients:
   (Intercept)           Depth  SpeciesConifer  
        7962.0          -262.2         -3113.0  
```

First, let’s make sure we understand these 3 values. Since we have no interaction, as we saw before, we are expecting graphical representation that should look like something like this:

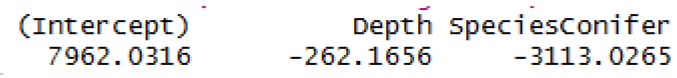


So, the absence of a significant interaction means that the 2 slopes are parallel or rather they are the same. So, going back to the 3 values:

- the first one is the intercept for the broadleaf species (alphabetical order for the levels),
- the second one will be the slope
- and the third, the intercept for the second species (7962.0 – 3113.0)

We need somehow to extract these values in such a way that abline() can use them. First we are going to use the function coefficients() to get the values from the model linear.treelight.add and save them as a vector with 3 values.

```
cf.add<-coefficients(linear.treelight.add)
cf
```

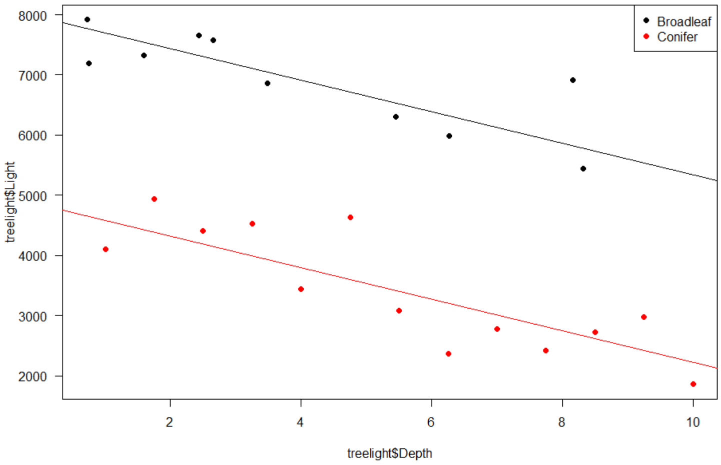


```
 (Intercept)          Depth SpeciesConifer 
   7962.0316      -262.1656     -3113.0265 
```

Then we just go:

```
abline(
       cf.add [1],
       cf.add [2],
       col="black”
)

abline(
       cf.add[1]+cf.add [3],
       cf.add [2]),
       col="red"
)
```

The colours are chosen to match the one on the graph.

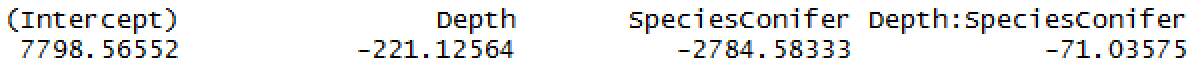


We get:

for broadleaf trees: Light = 7962.03 -262.17*Depth
for conifers: Light = (7962.03-3113.03) -262.17*Depth

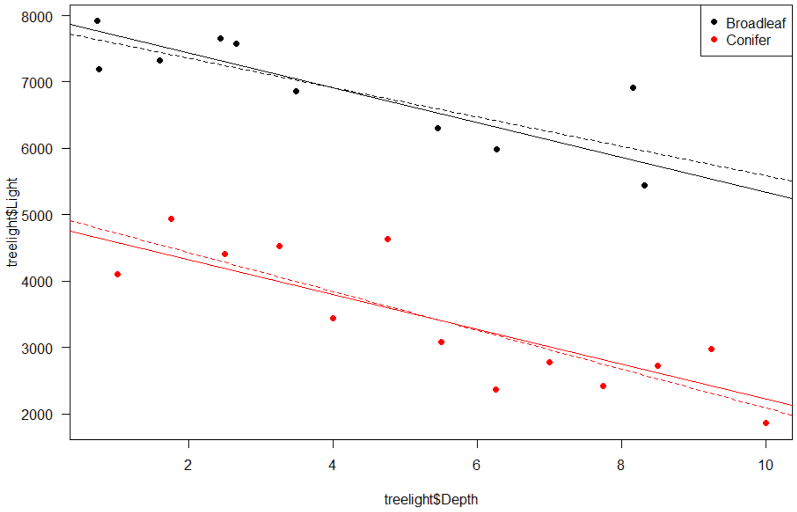
If we want to plot the insignificant difference in slopes, we of course can.

```
 cf<-coefficients(linear.treelight)
```



```
(Intercept)                 Depth        SpeciesConifer Depth:SpeciesConifer 
 7798.56552            -221.12564           -2784.58333            -71.03575 
```

```
abline(
      cf[1]+cf[3],
      cf[2]+cf[4],
      col="red",
      lty="dashed”
)

abline(
      cf[1],
      cf[2],
      col="black",
      lty="dashed"
)

plot(treelight$Light~treelight$Depth, col=treelight$Species, pch=16, las=1)
legend("topright", c("Broadleaf", "Conifer"), pch=16, col=1:2)
```



So to recapitulate, it is very easy to build models by just adding terms (factors) and it is not always necessary/meaningful to consider interactions. The error is always implicit even though we don’t always write it and each x_i is predicted by the model with its own error.

Linear model

Simplest

$$y = \beta_0 + \beta_1{}^*x$$

With 2 factors

$$y = \beta_0 + \beta_1{}^*x_1 + \beta_2{}^*x_2 + \beta_3{}^*x_1x_2$$

With n factors

$$y = \beta_0 + \beta_1{}^*x_1 + \ y = \beta_0 + \beta_1{}^*x\ {}_2 + \ldots + \beta_n{}^*x_n$$

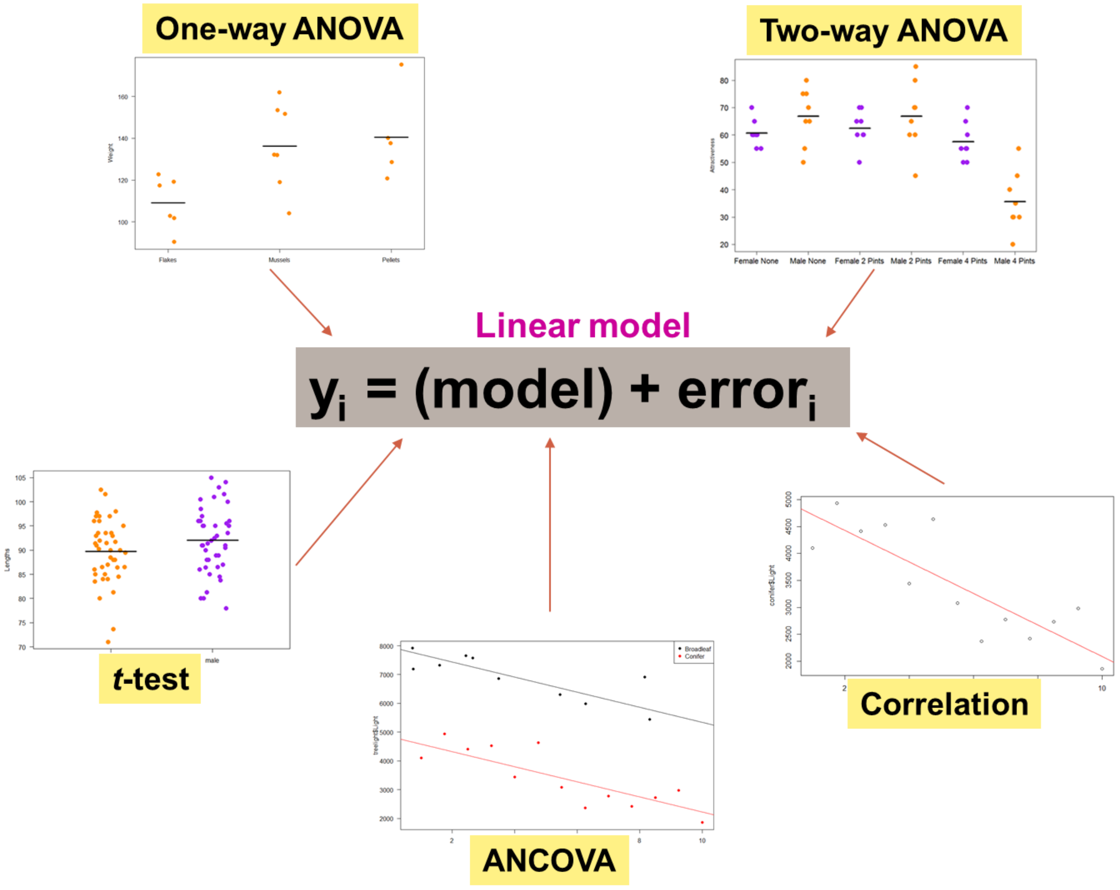
Let's not forget the error

$$y_i = (\beta_0 + \beta_1{}^*x_i) + \varepsilon_i$$

General formula

$$y_i = (\text{model}) + \text{error}_i$$

Finally, let's not forget it is possible to approach data and quantify relationships in different ways.



One-way ANOVA

Two-way ANOVA

Linear model

$$y_i = (\text{model}) + \text{error}_i$$

t-test

ANCOVA

Correlation

Chapter 9: Qualitative data

9-1 Comparison between 2 groups (1 factor)

Example (File: `cats.dat`)

A researcher is interested in whether cats could be trained to line dance. He tries to train them to dance by giving them either food or affection as a reward (**training**) for dance-like behaviour.

At the end of the week a note is made of which animal could line dance and which could not (**dance**). All the variables are dummy variables (categorical).

The pivotal (!) question is: Is there an effect of training on cats' ability to learn to line dance? We have already designed our experiment and chosen our statistical test: it will be a Fisher's exact test or a Chi-square.

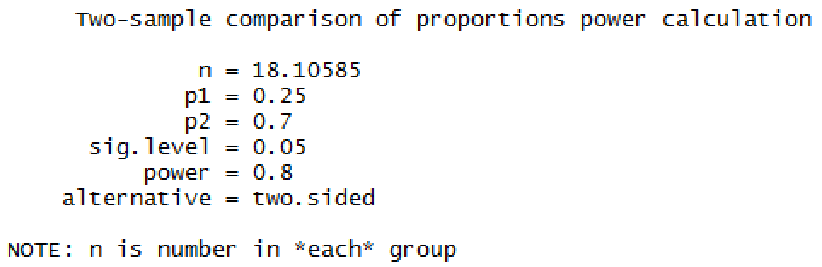
Power Analysis with qualitative data

The next step is to run a power analysis. In an ideal world, you would have run a pilot study to get some idea of the type of effect size you are expecting to see. Let's start with this ideal situation. Let's say, in your pilot study, you found that 25% of the cats did line dance after received affection and 70% did so after they received food.

So we want to compare 2 proportions (0.25 and 0.7), with a power of 80% and a significance threshold of 5%.

So we want to compare 2 proportions (0.25 and 0.7), we will be using the function `power.prop.test` from R:

```
power.prop.test(p1 = .25, p2 = .7, power = .8, sig.level = 0.05)
```



```
     Two-sample comparison of proportions power calculation 

              n = 18.10585
             p1 = 0.25
             p2 = 0.7
      sig.level = 0.05
          power = 0.8
    alternative = two.sided

NOTE: n is number in *each* group
```

R tells us that we need 2 samples of 18 cats to reach a power of 80%. In other words: if we want to be at least 80% confident to spot a reward effect, if indeed there is one, we will need about 36 cats altogether.

Next the experiment is run and the data collected.

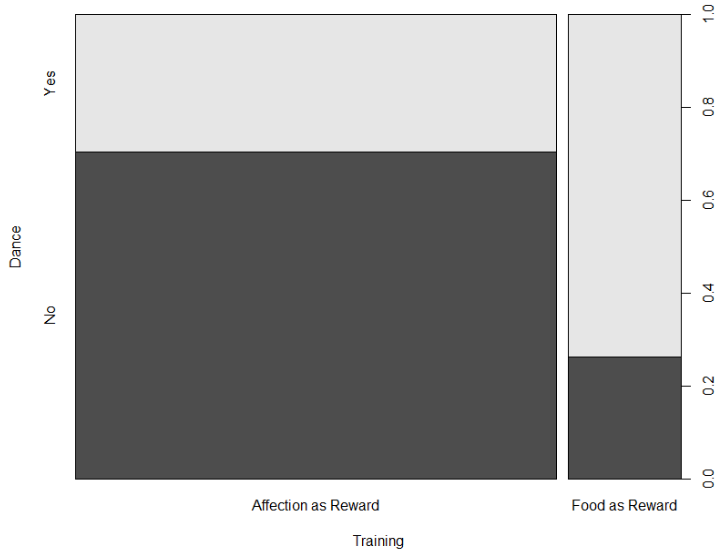
```
cats<-read.table("cats.dat", sep="\t",header=T)
```

Always worth have a quick look:

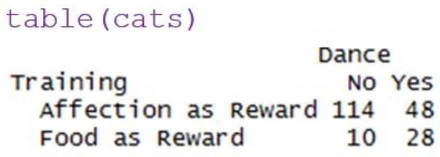
```
head(cats)
View(cats)
```

Next step, plotting the data:

```
plot(cats$Training, cats$Dance, xlab = "Training", ylab = "Dance")
```



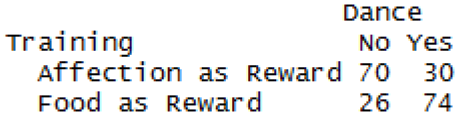
Now this a weird looking graph. It tells us about the different proportions (frequencies) of cats dancing and not dancing in both groups but it has quite an unbalanced look. The reason for that is the unbalanced design of the experiment.



```
table(cats)
                    Dance
Training              No Yes
  Affection as Reward 114  48
  Food as Reward       10  28
```

So that graph above is informative but not very pretty. We can do better than that and what would help is to change the format of the file into a contingency table. For that we can use the table above and change the counts into proportions `prop.table()` and then percentages. They can be more informative visually than raw data. When we are at it, we can round the values `round()` and get a nice neat table.

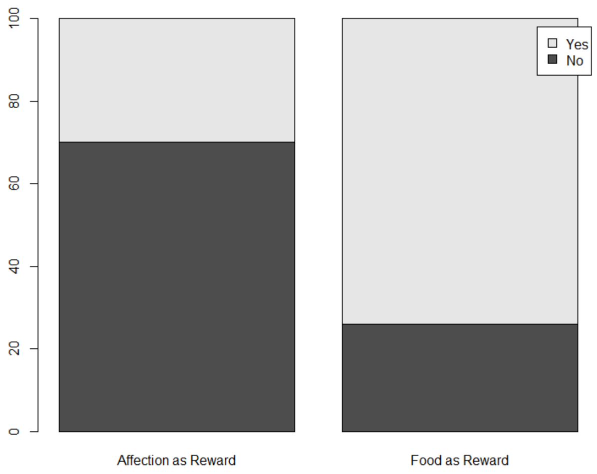
```
contingency.table <- table(cats.data)
contingency.table100<-prop.table(contingency.table,1)
contingency.table100<-round(contingency.table100*100)
contingency.table100
```



```
                     Dance
Training              No Yes
  Affection as Reward 70  30
  Food as Reward      26  74
```

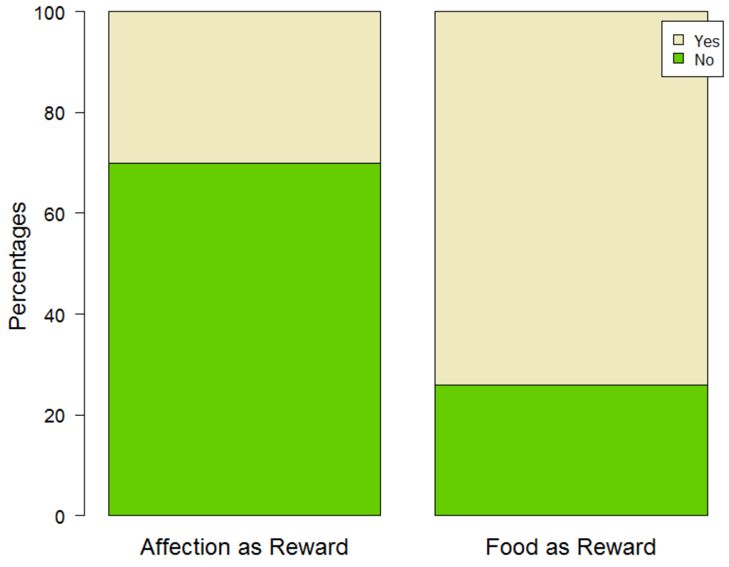
And we can plot the data as percentage/proportion:

```
barplot(t(contingency.table100), legend.text=TRUE)
```



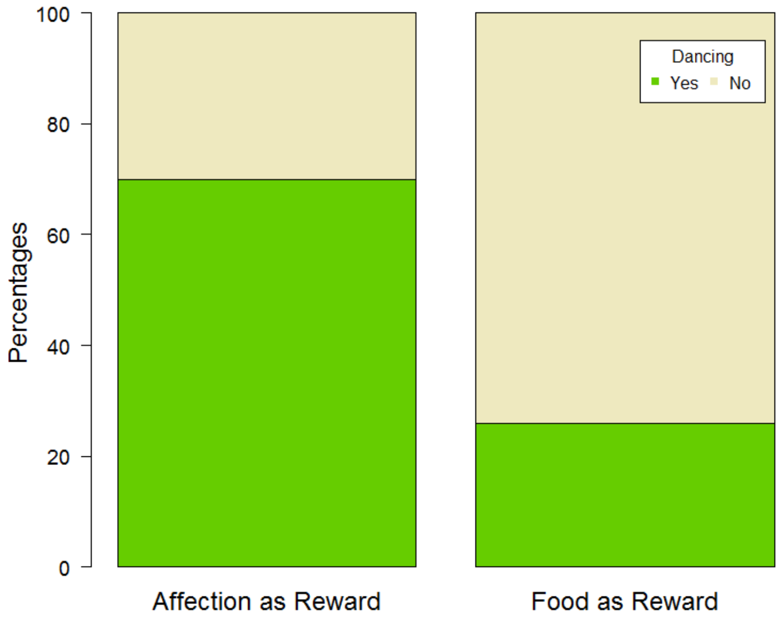
Prettier:

```
barplot(t(contingency.table100),
        col=c("chartreuse3","lemonchiffon2"),
        cex.axis=1.2,
        cex.names=1.5,
        cex.lab=1.5,
        ylab = "Percentages",
        las=1)
```



If we add `legend.text=TRUE`, R will add the legend on the graph but it will be a bit off (see above). If we want to have more control over it:

```
legend("topright",
       title="Dancing",
       inset=.05,
       c("Yes","No"),
       horiz=TRUE,
       pch=15,
       col=c("chartreuse3","lemonchiffon2"))
```



The χ^2 and the Fisher's exact tests

First, there are 2 types of χ^2 test :

- a one-way χ^2 test, which is basically a test that compares the observed frequency of a variable in a single group with what would be the expected by chance (also called goodness-of-fit).

- a two-way χ^2 test, the most widely used, in which the observed frequencies for two or more groups are compared with expected frequencies by chance. In other words, in this case, the χ^2 tells you whether or not there is an association between 2 categorical variables.

An important thing to know about the χ^2, and for the Fisher's exact test for that matter, is that it does not tell us anything about causality; it is simply measuring the strength of the association between 2 variables and it is our knowledge of the biological system we are studying which will help us to interpret the result. Hence, we generally have an idea of which variable is acting on the other.

The χ^2 value is calculated using the formula below:

$$\chi^2 = \Sigma \frac{(\textit{Observed Frequency - Expected Frequency})^2}{\textit{Expected Frequency}}$$

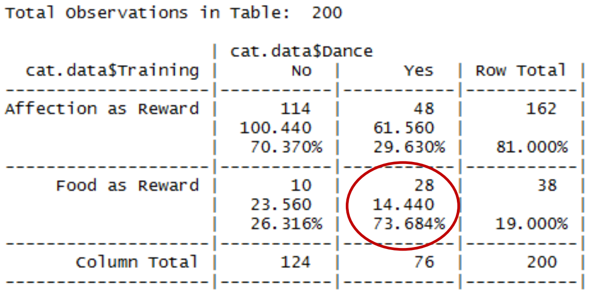
The observed frequencies are the ones we measured, the values that are in our table. Now, the expected ones are calculated this way:
Expected frequency = (row total)*(column total)/grand total

So, for example: the expected frequency of cats that would line dance after having received food as reward is: (76*38)/200=14.44

We can also think in terms of probabilities:

- probability of line dancing: 38/200
- probability of receiving food: 76/200

If the 2 events are independent, the probability of the 2 occurring at the same time (the expected frequency) will be: 38/200*76/200 = 0.072 and 7.2% of 200 is 14.4.



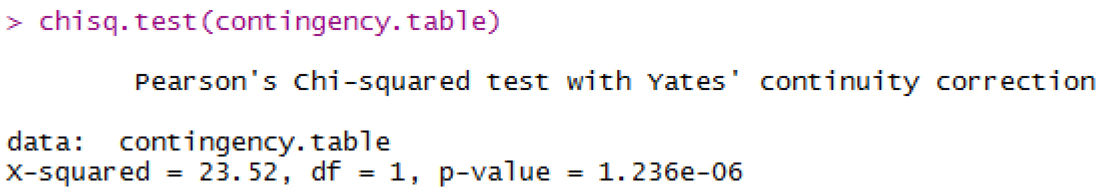
```
Total Observations in Table:  200

                     | cat.data$Dance
    cat.data$Training |         No |        Yes | Row Total |
---------------------|-----------|-----------|-----------|
Affection as Reward |        114 |         48 |       162 |
                     |    100.440 |     61.560 |           |
                     |     70.370% |     29.630% |    81.000% |
---------------------|-----------|-----------|-----------|
      Food as Reward |         10 |         28 |        38 |
                     |     23.560 |     14.440 |           |
                     |     26.316% |     73.684%|    19.000% |
---------------------|-----------|-----------|-----------|
        Column Total |        124 |         76 |       200 |
---------------------|-----------|-----------|-----------|
```

Intuitively, one can see that we are kind of looking for 50-50 (random output) results but accounting for the counts we have. If we work out the values for all the cells, we get the results above.

To run a χ^2 analysis, we are going to use `chisq.test()`.
There is only one assumption that we have to be careful about when we run a χ^2: with 2x2 contingency tables we should not have cells with an expected count below 5 as if it is the case it is likely that the test is not accurate. For larger tables, all expected counts should be greater than 1 and no more than 20% of expected counts should be less than 5.

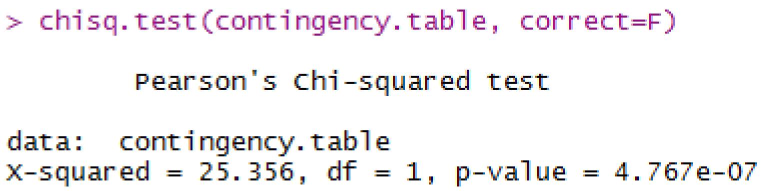
Let's see what we get:

- with the correction



```
> chisq.test(contingency.table)

        Pearson's Chi-squared test with Yates' continuity correction

data:  contingency.table
X-squared = 23.52, df = 1, p-value = 1.236e-06
```

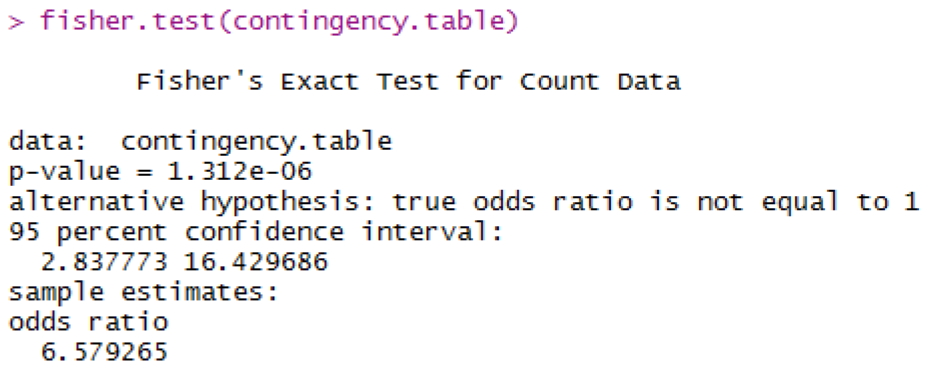
- without the correction:



```
> chisq.test(contingency.table, correct=F)

        Pearson's Chi-squared test

data:  contingency.table
X-squared = 25.356, df = 1, p-value = 4.767e-07
```

If we remember the χ^2's formula, the calculation gives us an estimation of the difference between our data and what we would have obtained if there was no association between our variables. Clearly, the bigger the value of the χ^2, the bigger the difference between observed and expected frequencies and the more likely the difference is to be significant.

Now with a 2x2 table, we can also use a Fisher's exact test: `fisher.test()`



```
> fisher.test(contingency.table)

        Fisher's Exact Test for Count Data

data:  contingency.table
p-value = 1.312e-06
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
  2.837773 16.429686
sample estimates:
odds ratio
  6.579265
```

As we can see here the p-values vary slightly between the 2 tests (p=1.31e-06 vs.p=4.77e-07) though the conclusion remains the same: cats only care about food. Though the samples are not very big here, the assumption for the χ^2 is met so you can choose either test.

So both tests will give us the same-ish p-value for big samples but for small samples the difference can be more important. Fisher's test is more accurate than Chi2 test on small samples but Chi2 test is more accurate than Fisher's test on large samples.

Having said that, the calculation of the Fisher's exact test is quite complex whereas the one for χ^2 is quite easy so only the calculation of the latter is going to be presented here. Also, the Fisher's test is often only available for 2x2 tables, so in a way the χ^2 is more general.
For both tests, the idea is the same: how different are the observed data from what we would have expected to see by chance i.e. if there were no association between the 2 variables. Or, looking at the table we may also ask: knowing that 76 of the 200 cats did dance and that 162 of them received affection, what is the probability that those 76 dancers would be so unevenly distributed between the 2 types of reward?

Now, the last thing we need to discuss is the Yate's correction. If we use the default version of `chisq.test()`, R will give us the p-value with Yates continuity correction.
Misusing a statistical test means that the output (ie the p-value) should not be trusted and therefore it is very likely that there is an increase of the probability of making the type I error. So one solution is to increase the p-value, hence making it more difficult to reach significance thus reducing the probability of making the type I error. So all corrections for statistical tests work the same way: they increase the p-value. There is some danger, however, in using corrections and some people think that it does matter how far their data are from the assumptions, applying a correction will be like sprinkling some magic: it will make the problem go away. But of course, it is not true. So here are 3 scenarios:

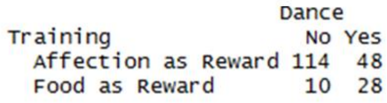
- Data meet the assumptions: all good
- Data almost meet the assumptions: can be still OK but p-values will have to be considered with a pinch of salt
- Data far from meeting the assumptions: no corrections can help and another approach should be considered (non-parametric for instance).

Now the problem is where do we draw the line between almost and far? Where does it stop being OK use a correction? It is a slippery slope and people make all kind of mistakes because of that. So the use of corrections is quite controversial: some people think it should always be used just to be on the safe side, others that it should never be used.

One last thing: we can notice that the output from the tests are quite different. This is because the functions/packages have been written by different people.

The Fisher's test's outcome gives us an extra bit of information: the odds ratio. It is exactly what it says it is: the ratio of the odds: the odds (or probability) of dancing in the food group over the odds of dancing in the affection one. It is basically the effect size and it is dead-easy to calculate. Using the information in the table below:

```
table(cats)
```



```
                       Dance
Training                 No Yes
  Affection as Reward 114  48
  Food as Reward        10  28
```

Odds of dancing: Food group: 28/10
Odds of dancing: Affection group: 48/114

Odds ratio = Food group / Affection group = 6.6.

Cats in the receiving food as a reward are almost 7 times more likely to line dance than those receiving affection (p=1.31e-06).

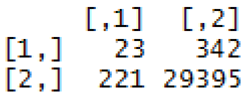
One last thing: the test is always run on the actual counts and not the percentages. If not, the samples are artificially set to 100, which is either overestimating or underestimating the real sample size, which will have an effect on the power of the test and thus on the likelihood of its significance.

Let's have a look at another example.

Example: Gene ontology enrichment

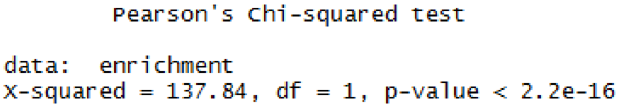
Sometimes the values we are dealing with are just a few numbers, out of which we need to make some sense.
For instance, say that in the entire genome (n=29395), 221 genes are associated with a function (e.g. antigen binding). From an experiment, we identified 342 genes out of which 23 are showing such an association. So here are our numbers: 23, 342, 221 and 29395 and we want to know if the observed enrichment is significant.
For that, we can build a matrix and run a χ^2 on it (or a Fisher's test, let's do both).

```
enrichment <- matrix(c(23, 342, 221, 29395), nrow=2, ncol=2, byrow = TRUE)
```



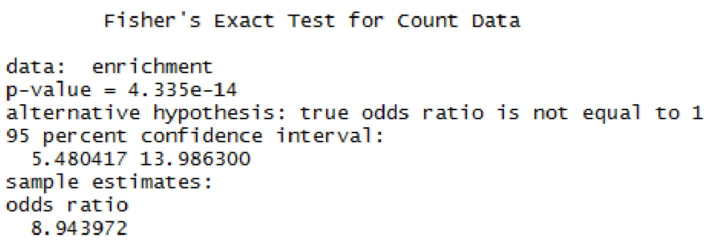
```
     [,1]  [,2]
[1,]   23   342
[2,]  221 29395
```

```
chisq.test(enrichment, correct=F)
```



```
        Pearson's Chi-squared test

data:  enrichment
X-squared = 137.84, df = 1, p-value < 2.2e-16
```

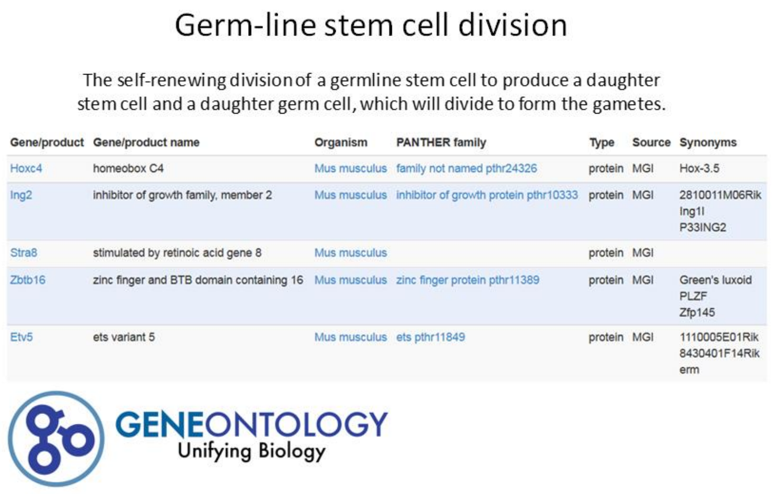
```
fisher.test(enrichment)
```



```
        Fisher's Exact Test for Count Data

data:  enrichment
p-value = 4.335e-14
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
  5.480417 13.986300
sample estimates:
odds ratio
  8.943972
```

There are nine times more genes associated with our function of interest in our selected group than in the general background (p=4.33e-14).

This type of analysis is used in many areas of biology but one common example would be when trying to find functional associations between a list of gene hits from an experiment and functional lists from a database such as gene ontology. In this type of analysis you take a list of genes with a known functional association and compare it to the genes in your experimental hit list. The idea is that if there are more genes from the functional list in your hit list than you'd expect by chance that the associated function might be interesting in your experiment.



Germ-line stem cell division

The self-renewing division of a germline stem cell to produce a daughter stem cell and a daughter germ cell, which will divide to form the gametes.

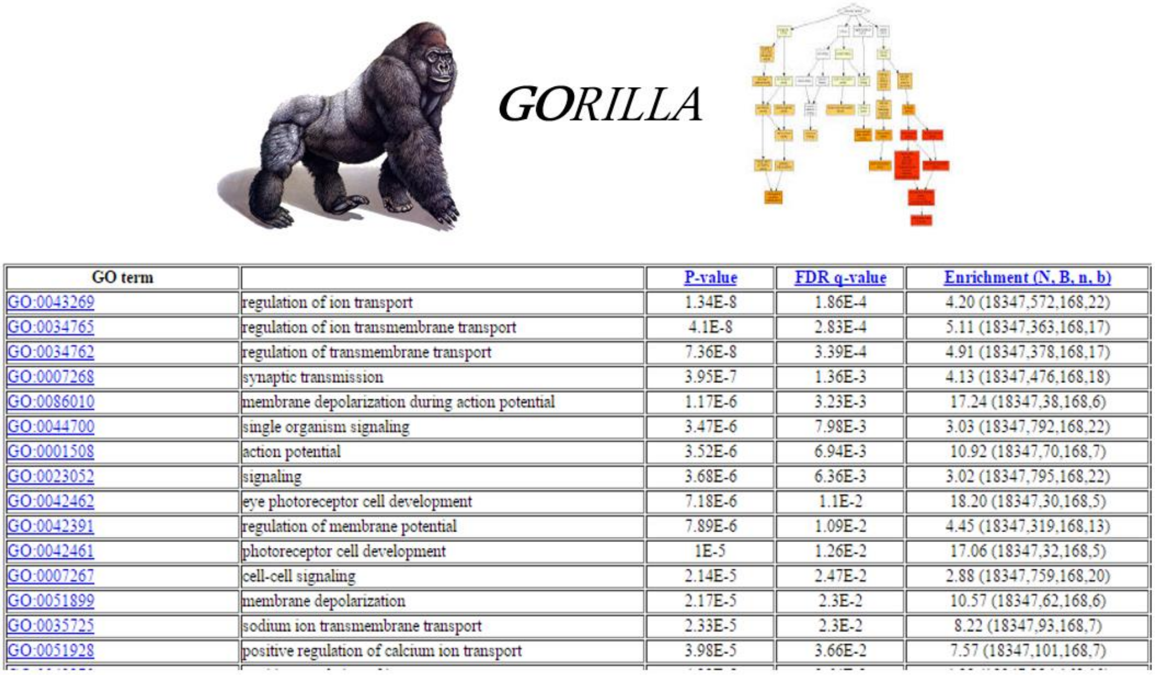
Gene/product	Gene/product name	Organism	PANTHER family	Type	Source	Synonyms
Hoxc4	homeobox C4	Mus musculus	family not named pthr24326	protein	MGI	Hox-3.5
Ing2	inhibitor of growth family, member 2	Mus musculus	inhibitor of growth protein pthr10333	protein	MGI	2810011M06Rik Ing1l P33ING2
Stra8	stimulated by retinoic acid gene 8	Mus musculus		protein	MGI	
Zbtb16	zinc finger and BTB domain containing 16	Mus musculus	zinc finger protein pthr11389	protein	MGI	Green's luxoid PLZF Zfp145
Etv5	ets variant 5	Mus musculus	ets pthr11849	protein	MGI	1110005E01Rik 8430401F14Rik erm

Hits from
Experiment

A4galt
Atl1
Cdk19
Cdon
Cecr2
Etv5
Flywch1
Gnpda2
Hoxc4
Ing2
Iigp1
Map3k9
Mypop
Rnf6
Serinc1
Stra8
Trp73
Zbtb16

The data used in this case is a simple contingency table, the same as we've already seen.

	[genes in genome]	[genes in hit list]
[genes in functional category]	N	B
[genes not in functional category]	n	b



GO term		P-value	FDR q-value	Enrichment (N, B, n, b)
GO:0043269	regulation of ion transport	1.34E-8	1.86E-4	4.20 (18347,572,168,22)
GO:0034765	regulation of ion transmembrane transport	4.1E-8	2.83E-4	5.11 (18347,363,168,17)
GO:0034762	regulation of transmembrane transport	7.36E-8	3.39E-4	4.91 (18347,378,168,17)
GO:0007268	synaptic transmission	3.95E-7	1.36E-3	4.13 (18347,476,168,18)
GO:0086010	membrane depolarization during action potential	1.17E-6	3.23E-3	17.24 (18347,38,168,6)
GO:0044700	single organism signaling	3.47E-6	7.98E-3	3.03 (18347,792,168,22)
GO:0001508	action potential	3.52E-6	6.94E-3	10.92 (18347,70,168,7)
GO:0023052	signaling	3.68E-6	6.36E-3	3.02 (18347,795,168,22)
GO:0042462	eye photoreceptor cell development	7.18E-6	1.1E-2	18.20 (18347,30,168,5)
GO:0042391	regulation of membrane potential	7.89E-6	1.09E-2	4.45 (18347,319,168,13)
GO:0042461	photoreceptor cell development	1E-5	1.26E-2	17.06 (18347,32,168,5)
GO:0007267	cell-cell signaling	2.14E-5	2.47E-2	2.88 (18347,759,168,20)
GO:0051899	membrane depolarization	2.17E-5	2.3E-2	10.57 (18347,62,168,6)
GO:0035725	sodium ion transmembrane transport	2.33E-5	2.3E-2	8.22 (18347,93,168,7)
GO:0051928	positive regulation of calcium ion transport	3.98E-5	3.66E-2	7.57 (18347,101,168,7)

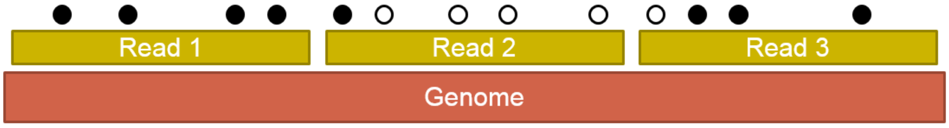
In this sample output from the GOrilla search program (http://cbl-gorilla.cs.technion.ac.il/) you can see the 4 values for the test in brackets in the final column. GOrilla does not use the Fisher's test actually but the hypergeometric one; it is based on the same distribution but works slightly differently. It yields the same p-value though. It also shows the fold enrichment as well as the raw p-value. Because many different tests have been performed (one for each functional category) the output also shows a p-value which has been corrected for multiple testing, which is the value you would actually use in this instance.

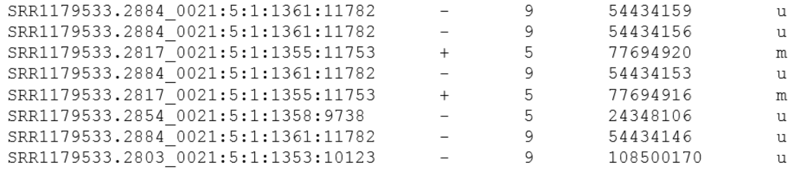
9-2 Comparison with more than 1 factor

A nice example to illustrate this type of analysis is DNA methylation. This is an epigenetic modification which is involved in gene regulation. DNA methylation is measured by a sequencing methodology called bisulphite sequencing. The output of this method is a set of observations of individual cytosines where they can be observed as being either methylated or unmethylated.

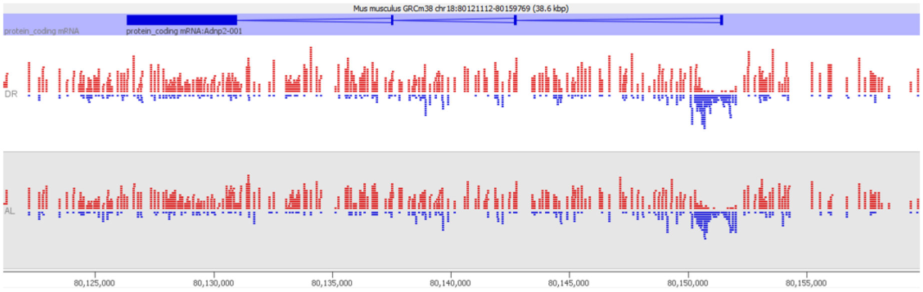
The idea for the example analysis we're going to do is that you would take all of the observations for one gene and count the number of methylated and unmethylated cytosines which fell within the gene body. These counts would then be collected for multiple samples, which could be divided between different experimental conditions.

The question we want to ask would be whether the proportion of methylated reads (and therefore the methylation level) changes significantly between different experimental conditions.





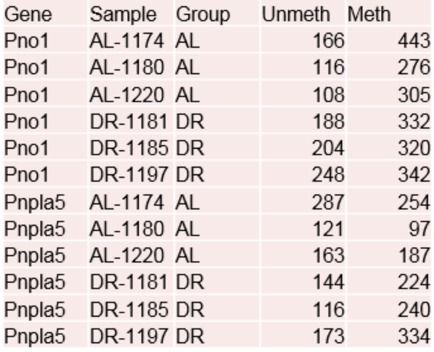
```
SRR1179533.2884_0021:5:1:1361:11782     -      9      54434159       u
SRR1179533.2884_0021:5:1:1361:11782     -      9      54434156       u
SRR1179533.2817_0021:5:1:1355:11753     +      5      77694920       m
SRR1179533.2884_0021:5:1:1361:11782     -      9      54434153       u
SRR1179533.2817_0021:5:1:1355:11753     +      5      77694916       m
SRR1179533.2854_0021:5:1:1358:9738      -      5      24348106       u
SRR1179533.2884_0021:5:1:1361:11782     -      9      54434146       u
SRR1179533.2803_0021:5:1:1353:10123     -      9      108500170      u
```



Example: (File: `dna_methylation_format2.csv)`

We still have a comparison between 2 groups, the comparison of interest shall we say, but for each group/condition, we have 3 biological samples. So here we want to know if something is happening between our groups (factor of interest) but accounting for the variability between samples (random factor, biological/experimental variation).

Values are independent observations of individual DNA bases. A DNA base is either methylated (Meth) or unmethylated (Unmeth).



Gene	Sample	Group	Unmeth	Meth
Pno1	AL-1174	AL	166	443
Pno1	AL-1180	AL	116	276
Pno1	AL-1220	AL	108	305
Pno1	DR-1181	DR	188	332
Pno1	DR-1185	DR	204	320
Pno1	DR-1197	DR	248	342
Pnpla5	AL-1174	AL	287	254
Pnpla5	AL-1180	AL	121	97
Pnpla5	AL-1220	AL	163	187
Pnpla5	DR-1181	DR	144	224
Pnpla5	DR-1185	DR	116	240
Pnpla5	DR-1197	DR	173	334

The question is: to what extent can we predict the behaviour/variation of our outcome with 2 factors (one of interest and one random)? And by the way, the outcome can be continuous or categorical: if it is continuous, we will likely use a 2-way ANOVA and if it is categorical, and in our case binary: a binary logistic regression.

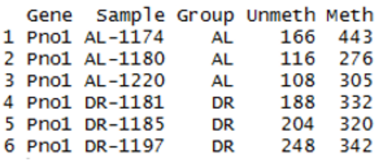
Actually, the question can be summarised as follows:

Outcome ~ factor1 + factor2 + factor1*factor2.

So here the question is: For each gene: Is there a difference in methylation between the 2 conditions (AL and DR) accounting for the variability between samples?

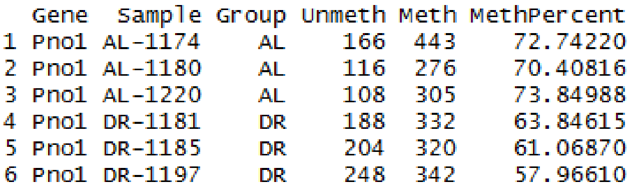
First, let's have a look at the data, for each gene, as percentages of methylation.

```
dna.methylation <- read.csv("dna_methylation_format2.csv")
head(dna.methylation)
```



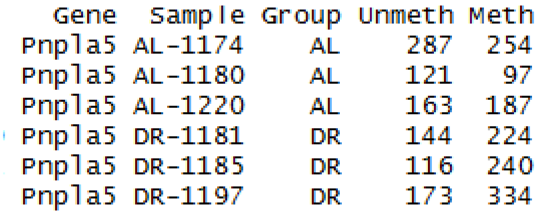
```
  Gene  Sample Group Unmeth Meth
1 Pno1 AL-1174    AL    166  443
2 Pno1 AL-1180    AL    116  276
3 Pno1 AL-1220    AL    108  305
4 Pno1 DR-1181    DR    188  332
5 Pno1 DR-1185    DR    204  320
6 Pno1 DR-1197    DR    248  342
```

```
dna.methylation$MethPercent<-
(dna.methylation$Meth/(dna.methylation$Meth+dna.methylation$Unmeth)*100)
head(dna.methylation)
```



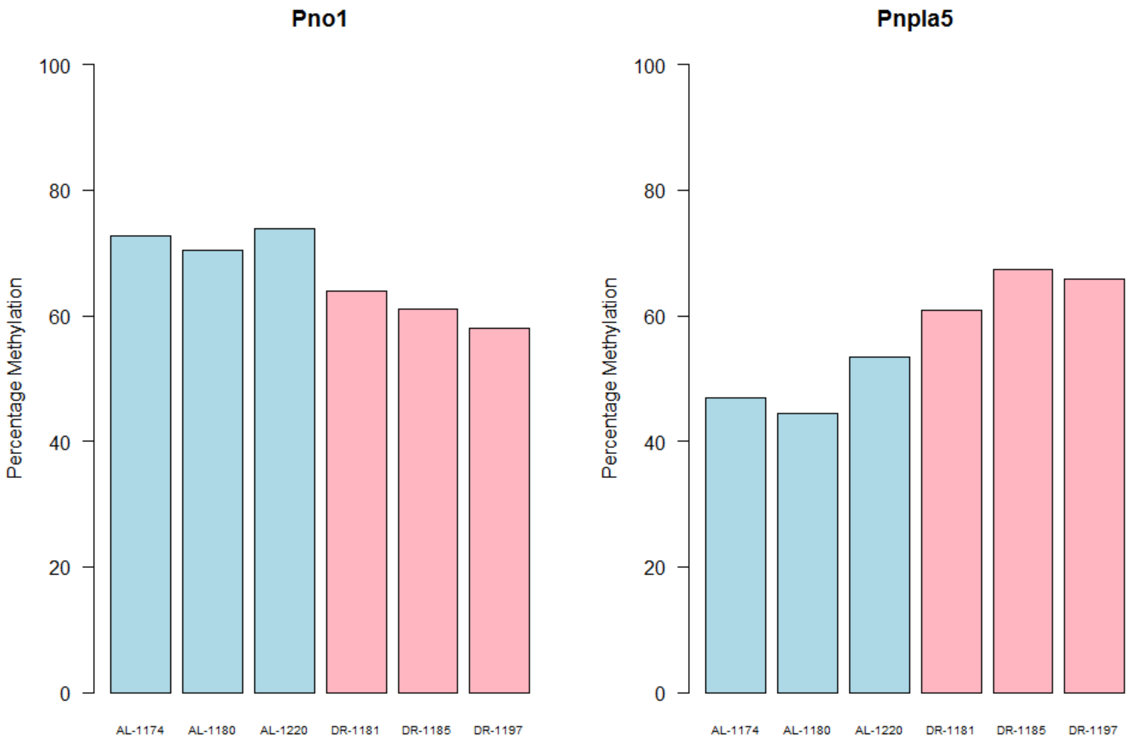
```
  Gene  Sample Group Unmeth Meth MethPercent
1 Pno1 AL-1174    AL    166  443    72.74220
2 Pno1 AL-1180    AL    116  276    70.40816
3 Pno1 AL-1220    AL    108  305    73.84988
4 Pno1 DR-1181    DR    188  332    63.84615
5 Pno1 DR-1185    DR    204  320    61.06870
6 Pno1 DR-1197    DR    248  342    57.96610
```

```
dna.methylation.Pno1<-dna.methylation2[dna.methylation2$Gene=="Pno1",]
dna.methylation.Pnpla5<-dna.methylation2[dna.methylation2$Gene=="Pnpla5",]
head(dna.methylation.Pnpla5)
```



```
   Gene  Sample Group Unmeth Meth
 Pnpla5 AL-1174    AL    287  254
 Pnpla5 AL-1180    AL    121   97
 Pnpla5 AL-1220    AL    163  187
 Pnpla5 DR-1181    DR    144  224
 Pnpla5 DR-1185    DR    116  240
 Pnpla5 DR-1197    DR    173  334
```

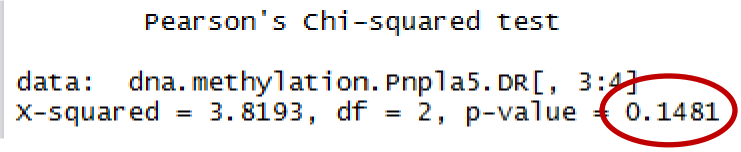
```
par(mfrow=c(1,2))
## dna.methylation.Pno1<-as.data.frame(dna.methylation.Pno1) ##
barplot(dna.methylation.Pno1$MethPercent,
        names.arg = dna.methylation.Pno1$Sample,
        col=rep(c("lightblue","lightpink"), each=3),
        main = "Pno1",
        ylim=c(0,100),
        cex.names=0.6,
        ylab = "Percentage Methylation",
        las=1)
## dna.methylation.Pnpla5<-as.data.frame(dna.methylation.Pnpla5) ##
barplot(dna.methylation.Pnpla5$MethPercent,
        names.arg = dna.methylation.Pnpla5$Sample,
        col=rep(c("lightblue","lightpink"), each=3),
        main = "Pnpla5",
        ylim=c(0,100),
        cex.names=0.6,
        ylab = "Percentage Methylation",
        las=1)
```



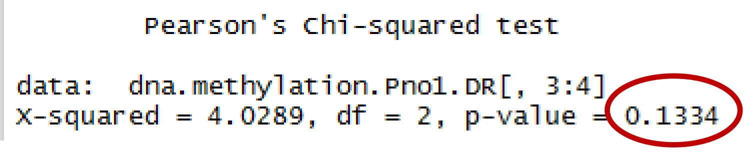
From the graph above, we can see that there is some variability between the samples but also that overall values for the condition AL are higher than the ones in condition DR for Pno1 gene and the reverse for Pnpla5. This is promising: we expect variability between samples within a condition but also consistency between the conditions. This is what we are going to base our confidence on that methylation does differ between our conditions of interest.

There are different ways to go about quantifying the effects we see. We are going to go for a simple and stepwise one. First, let's see if there is a significant difference between the samples within each condition and for each gene. For that, we can use Chi2 test. Each test will tell us if there is an overall difference between the samples. We don't care here about pairwise comparisons, since each sample has no predictive meaning, we just care about consistency: are all the samples telling us the same thing?

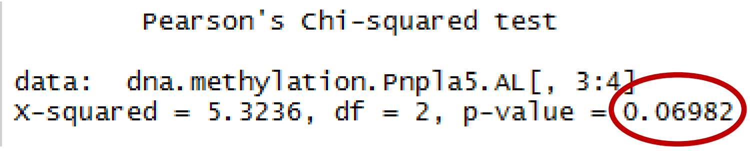
```
## test difference between samples for each gene/condition ###

dna.methylation.Pno1.AL<-
dna.methylation.Pno1[dna.methylation.Pno1$Group=="AL",]
chisq.test(dna.methylation.Pno1.AL[,4:5], correct=F)

dna.methylation.Pno1.DR<-
dna.methylation.Pno1[dna.methylation.Pno1$Group=="DR",]
chisq.test(dna.methylation.Pno1.DR[,4:5], correct=F)

dna.methylation.Pnpla5.AL<-
dna.methylation.Pnpla5[dna.methylation.Pnpla5$Group=="AL",]
chisq.test(dna.methylation.Pnpla5.AL[,4:5], correct=F)

dna.methylation.Pnpla5.DR<-
dna.methylation.Pnpla5[dna.methylation.Pnpla5$Group=="DR",]
chisq.test(dna.methylation.Pno1.DR[,4:5], correct=F)
```



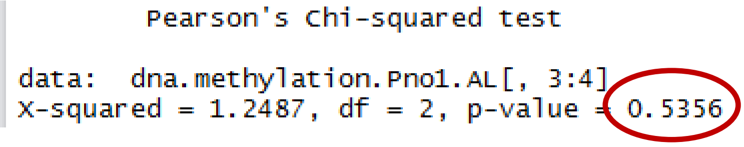
```
        Pearson's Chi-squared test

data:  dna.methylation.Pnpla5.DR[, 3:4]
X-squared = 3.8193, df = 2, p-value = 0.1481
```



```
        Pearson's Chi-squared test

data:  dna.methylation.Pno1.DR[, 3:4]
X-squared = 4.0289, df = 2, p-value = 0.1334
```



```
        Pearson's Chi-squared test

data:  dna.methylation.Pnpla5.AL[, 3:4]
X-squared = 5.3236, df = 2, p-value = 0.06982
```



```
        Pearson's Chi-squared test

data:  dna.methylation.Pno1.AL[, 3:4]
X-squared = 1.2487, df = 2, p-value = 0.5356
```

So this is good news: there is no significant difference between the samples within each condition/each gene.

Now, the important stuff: is there a difference between our conditions of interest, AL and DR, accounting for the variability between the samples? Or: how much can we predict methylation using these 2 conditions?

As mentioned earlier, we are going to use the binary version of a 2-way ANOVA: the binary logistic regression. The 2 approaches belong to the wider family of linear models. Sometimes referred to as general linear models or generalized linear models. The idea is always the same: quantifying the linear relationship between predictors and outcome. The simplest format is:

y = Constant + (coeff * x)

and with 2 factors (like here):

y = Constant + (coeff$_1$ * x$_1$) + (coeff$_2$ * x$_2$)

This last equation can also be:

Outcome ~ factor1 + factor2

And if we want to also look at the interaction, like we did with the quantitative data analysis:

Outcome ~ factor1 + factor2 + factor1*factor2

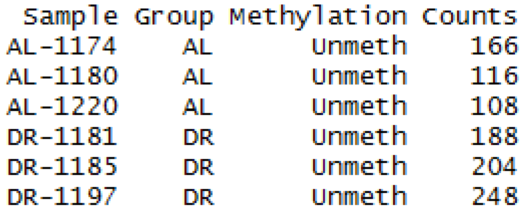
It is not always useful or meaningful to include the interaction term. For instance, in our case, we have already explored the difference between samples and saw that it was no significant. So we are interested in accounting for the sample variability but not how it interacts with our factor of interest.

With R, there are several packages/functions that allow for the analyses of this type of data, `aov()` and `glm()` being 2 of them. The latter works with both quantitative and qualitative data but the first one has the advantage of producing a more simple yet informative output. Ultimately they both do the same thing but since only `glm()` 'caters' for binary data we are going to use it. We just need to specify the family to which our outcome belongs.

```
glm(Outcome ~ factor1 + factor2, family = binomial())
```

Before, running the analysis, we need to restructure the data as `glm()` works on long file format.

```
dna.methylation.Pno1.long<-melt(dna.methylation.Pno1[,2:5],  ID=c("Sample", "Group"))
dna.methylation.Pnpla5.long<-melt(dna.methylation.Pnpla5[,2:5], ID=c("Sample", "Group"))

head(dna.methylation.Pno1.long)

colnames(dna.methylation.Pno1.long)[3:4]<-c("Methylation", "Counts")
colnames(dna.methylation.Pnpla5.long)[3:4]<-c("Methylation", "Counts")

head(dna.methylation.Pno1.long)
```



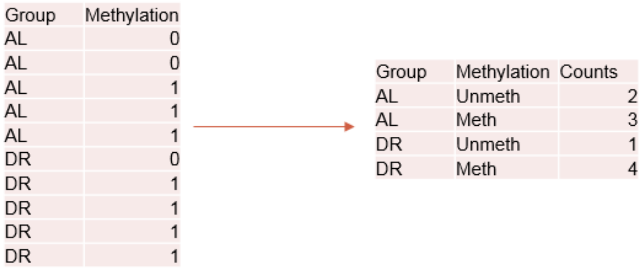
```
  Sample Group Methylation Counts
AL-1174    AL       Unmeth    166
AL-1180    AL       Unmeth    116
AL-1220    AL       Unmeth    108
DR-1181    DR       Unmeth    188
DR-1185    DR       Unmeth    204
DR-1197    DR       Unmeth    248
```

So, with `glm()`, we go:

```
glm(Methylation~Group+Sample,data=dna.methylation.Pno1.long)
```

Now we need to specify the family of the data: `family = binomial()`

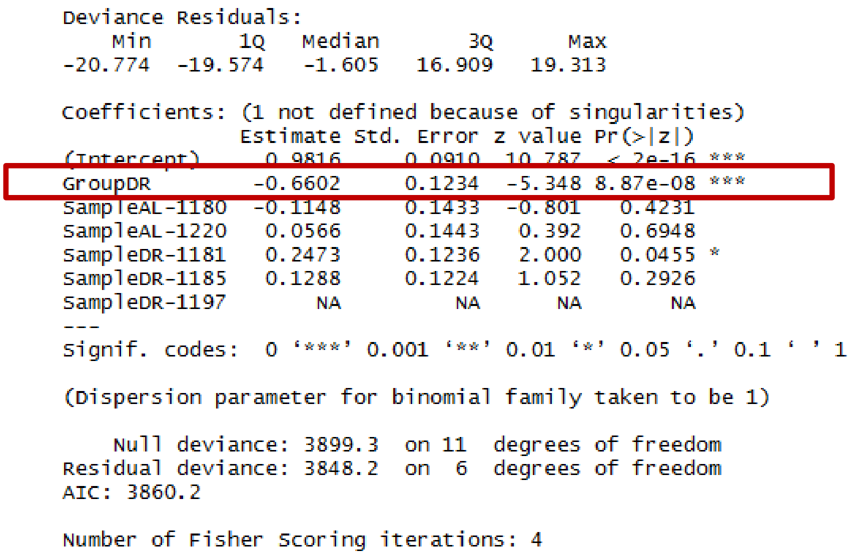
There is one last thing: our outcome is `Methylation` but presented as in our file above, `glm()` 'will not know' that the counts associated with say 'Unmeth' in the first row is 166. If our data was in the format as on the left below, then each row being a single count, the function will handle it by summing the counts. But in our case, this step has been done (like in the file on the right) so we need an argument to specify the associated counts, for instance Counts= 166 for the first row. It is the 'weight' of Unmeth hence `weights=Counts`.



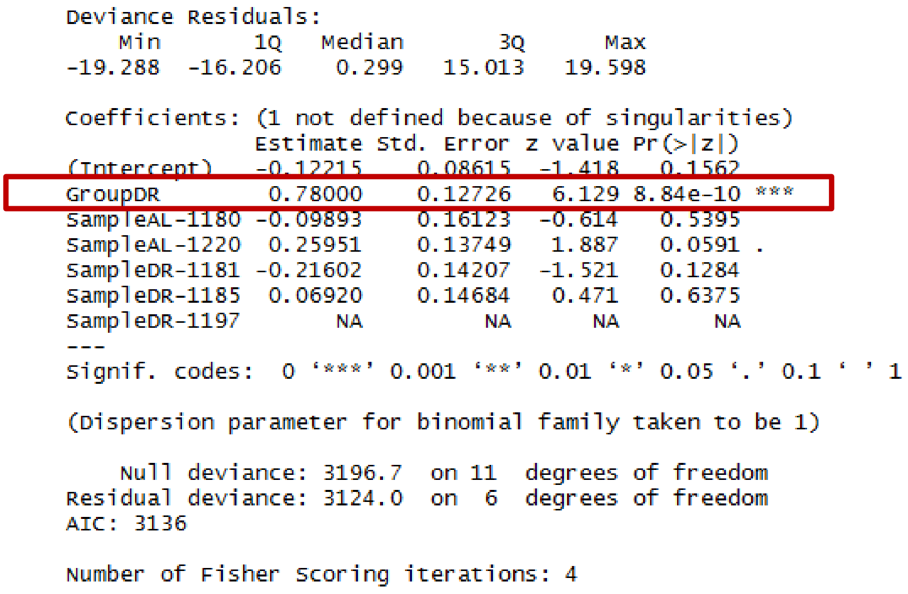
Group	Methylation
AL	0
AL	0
AL	1
AL	1
AL	1
DR	0
DR	1
DR	1
DR	1
DR	1

Group	Methylation	Counts
AL	Unmeth	2
AL	Meth	3
DR	Unmeth	1
DR	Meth	4

```
dna.model.Pno1<glm(Methylation~Group+Sample,data=dna.methylation.Pno1.long,
      family = binomial(),
      weights=Counts)

summary(dna.model.Pno1)
```



```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-20.774  -19.574   -1.605   16.909   19.313

Coefficients: (1 not defined because of singularities)
             Estimate Std. Error z value Pr(>|z|)
(Intercept)    0.9816     0.0910  10.787  < 2e-16 ***
GroupDR       -0.6602     0.1234  -5.348 8.87e-08 ***
SampleAL-1180 -0.1148     0.1433  -0.801   0.4231
SampleAL-1220  0.0566     0.1443   0.392   0.6948
SampleDR-1181  0.2473     0.1236   2.000   0.0455 *
SampleDR-1185  0.1288     0.1224   1.052   0.2926
SampleDR-1197      NA         NA      NA       NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3899.3  on 11  degrees of freedom
Residual deviance: 3848.2  on  6  degrees of freedom
AIC: 3860.2

Number of Fisher Scoring iterations: 4
```

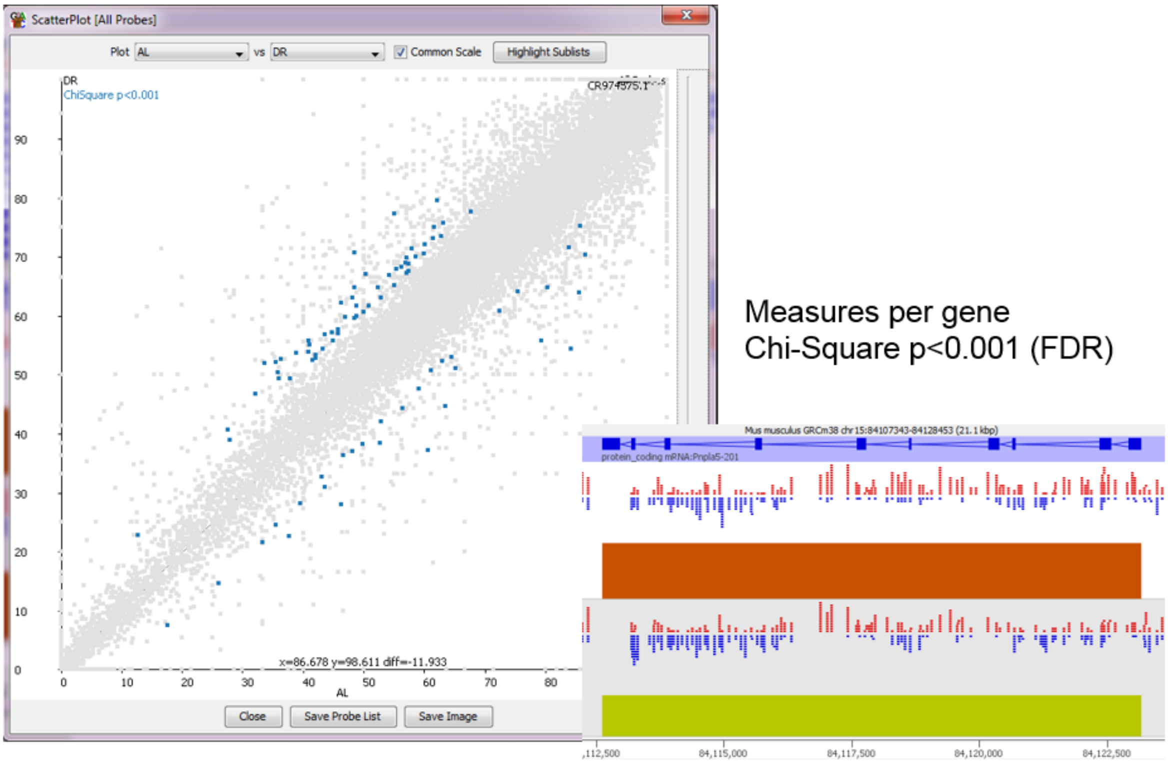


```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-19.288  -16.206    0.299   15.013   19.598

Coefficients: (1 not defined because of singularities)
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   -0.12215    0.08615  -1.418   0.1562
GroupDR        0.78000    0.12726   6.129 8.84e-10 ***
SampleAL-1180 -0.09893    0.16123  -0.614   0.5395
SampleAL-1220  0.25951    0.13749   1.887   0.0591 .
SampleDR-1181 -0.21602    0.14207  -1.521   0.1284
SampleDR-1185  0.06920    0.14684   0.471   0.6375
SampleDR-1197       NA         NA      NA       NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3196.7  on 11  degrees of freedom
Residual deviance: 3124.0  on  6  degrees of freedom
AIC: 3136

Number of Fisher Scoring iterations: 4
```

glm(), rather than presenting the p-value associated with Group factor like aov() does, it gives it as the difference between a given level of a factor vs. a control level. By default the control is alphabetical so here AL. Hence, GroupDR associated with p= 8.84e-10. The rest of the output is not relevant as it shows comparisons between samples regardless of which group they belong to.

So here, we have for both genes a significant difference in methylation levels between the AL and DR groups (Pno1, p=8.87e-08 and Pnpla4, p=8.84e-10).

In this example we analysed only 2 genes, but in a real study we would likely to be analysing every gene in the genome. The basic process remains the same though. Below are the results of doing a genome-wide scan of this same data, showing the observed differences in methylation, and with the blue dots representing those which were identified as statistically significant. You can see that there are many points with apparently large differences which were not identified as significant, probably because of either poor observation (low counts), or possibly because of high variation between samples in the same condition.



Measures per gene
Chi-Square p<0.001 (FDR)

References

Cumming G., Fidler F. and Vaux D.L. Error bars in experimental biology. *The Journal of Cell Biology*, Vol. 177, No.1, 7-11.

Field A. 2012. *Discovering statistics using R* (1st Edition). London: Sage.

McKillup S. 2005. *Statistics explained*. Cambridge: Cambridge University Press.

Cohen J. 1992. A power primer. *Psychological Bulletin*. Jul; 112(1):155-9.