Babraham Bioinformatics

Introduction to Statistics with GraphPad Prism

Version 2019-11

Licence

Table of Contents

Introduction

GraphPad Prism is a straightforward package with a user-friendly environment. There is a lot of easy-to-access documentation and the tutorials are very good.

Graphical representation of data is pivotal when we want to present scientific results, in particular for publications. GraphPad allows us to build top quality graphs, much better than Excel for example and in a much more intuitive way.

In this manual, however, we are going to focus on the statistical menu of GraphPad. The data analysis approach is much friendlier than with R for instance. R does not hold your hand all the way through the analysis, whereas GraphPad does. On the down side, GraphPad is not as powerful as R - as in we cannot do as many as different analyses with GraphPad as we can with R. If we are interested in linear modelling for example, we would need to use R.

Both GraphPad and R work quite differently. Despite this, whichever program we choose we need some basic statistical knowledge if only to design our experiments correctly, so there is no way out of it!

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

“*Forget about getting definitive results from a single experiment; instead embrace variation, accept uncertainty, and learn what you can*." Andrew Gelman, 2018.

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.

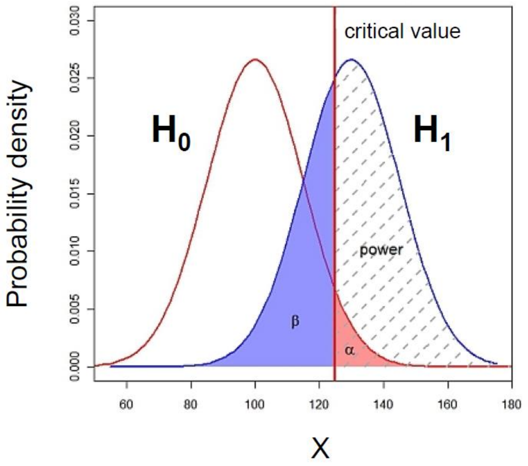
What is Power?

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 α=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 – β where β=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

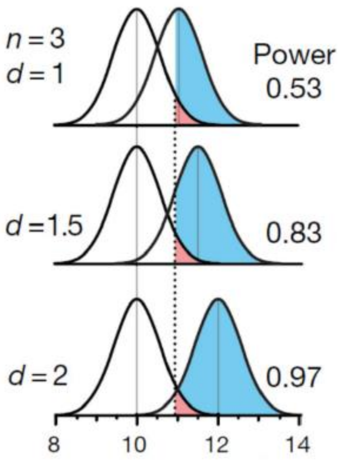
Yet another approach is to use conventions. Cohen (author of several books and articles on power analysis) has defined small, medium and large effect sizes for many types of test. These form useful conventions, and can guide you if you know approximately how strong the effect is likely to be.

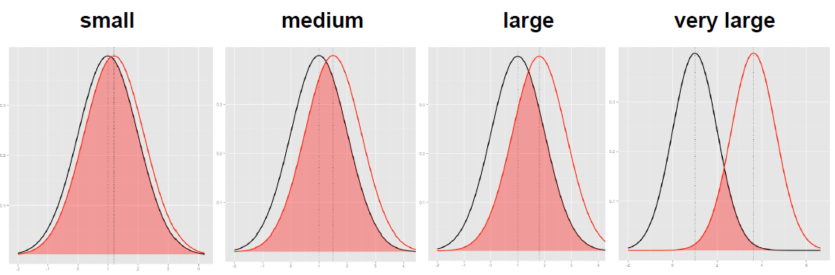
Table 1: Thresholds/Convention for interpreting effect size

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.

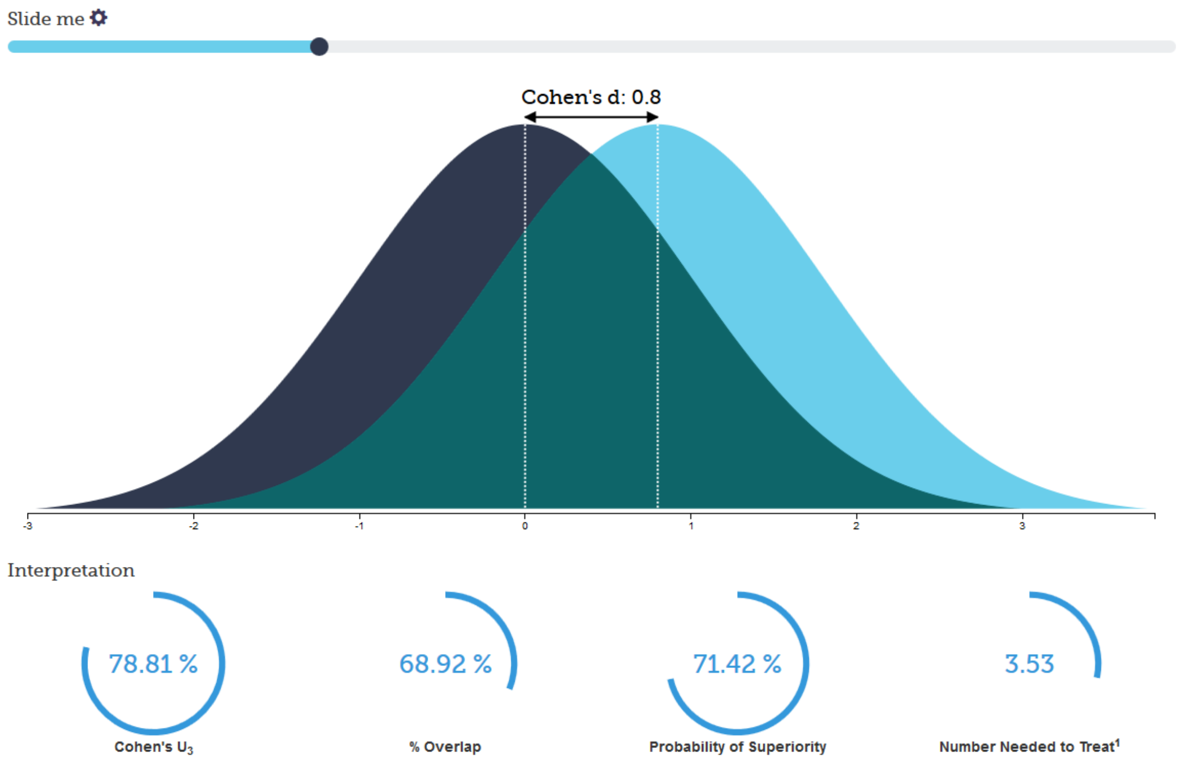
The graphs below give a visual representation of the effect sizes.





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'.

Table 2: 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)

↓

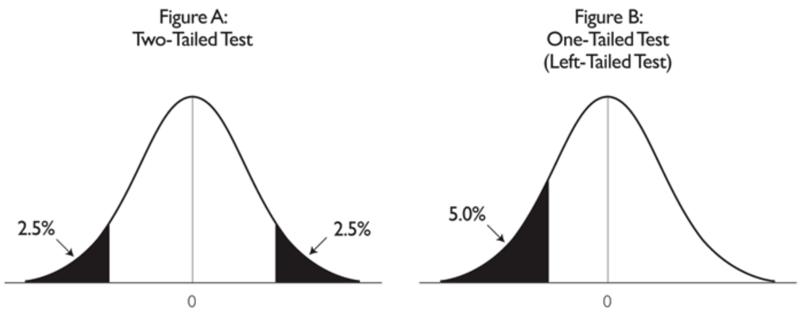
(Stat) analysis of the results

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.*

Two-Tailed Versus One-Tailed Hyphothesis Tests

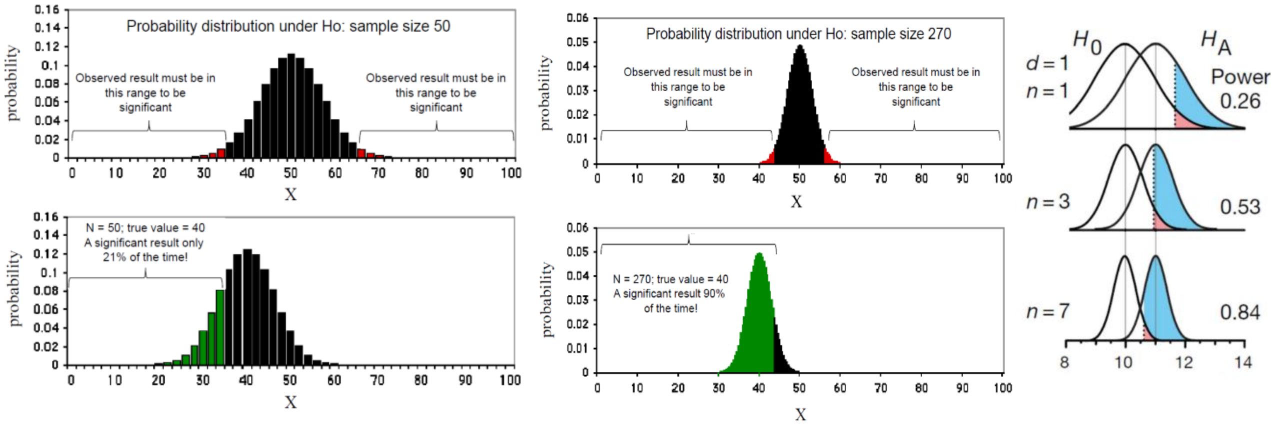


If you can reasonably **predict** the direction of an effect, based on a scientific hypothesis, a 1-tailed test is more powerful than a 2-tailed test. However, it is not always rigidly applied so be cautious when 1-tailed tests are reported, especially when accompanied by marginally-significant results!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

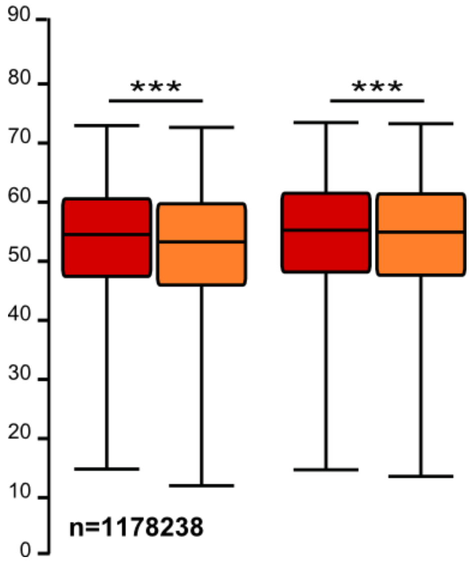


Krzywinski and Altman 2013 (Nature Methods)

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

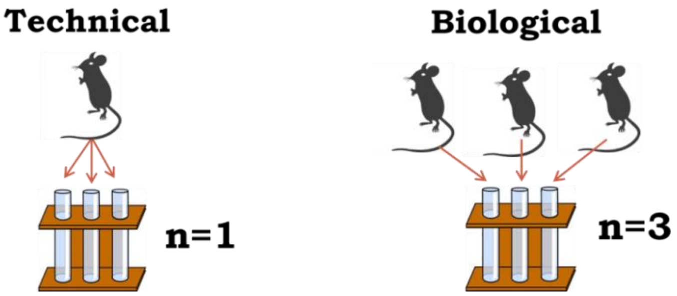
As we saw, power and effect size are linked so that the bigger the power the smaller the effect size that can be detected, as in associated with a significant p-value. The problem is that there is such a thing as overpower. Studies or experiments which produce thousand or hundreds of thousands of data, 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.

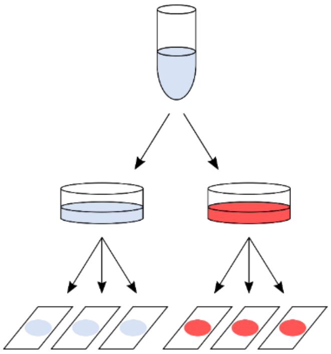
If we repeat the measurements, let’s say we weigh each mouse five times then we will have ten different measurements. But this cannot be used to prove that grey mice are heavier than white mice in general, we still have only looked at one white and one grey mouse. Using the terminology above, the five 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

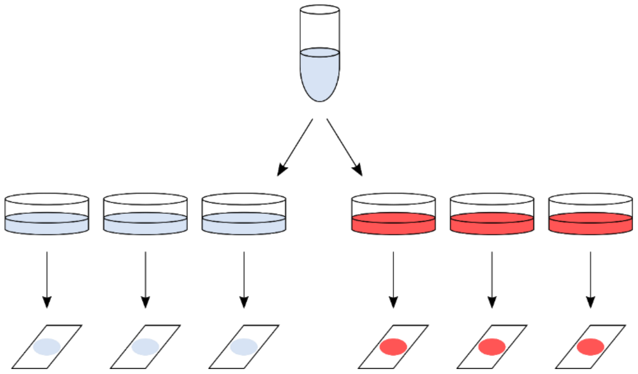
In this experiment a single vial is thawed, cells are divided into two culture dishes and the treatment (red) is randomly applied to one of the two dishes. The cells are allowed to grow for a period of time, and then three samples are pipetted from each dish onto glass slides, and the number of cells 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

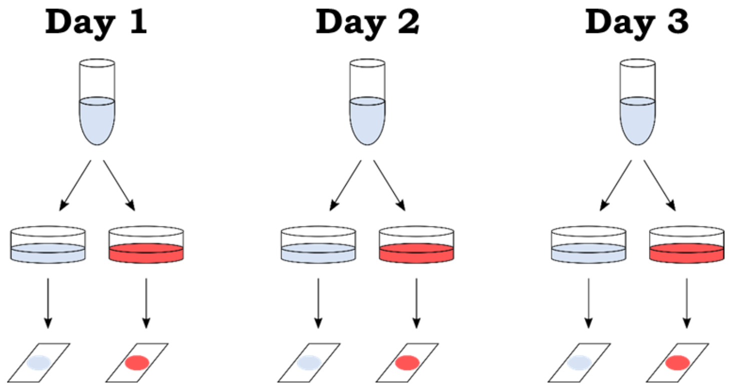
In this modified experiment, the vial of cells is divided into six separate culture dishes, and then cells from each culture dish are pipetted onto a single glass slide. Similar to the previous experiment, there are six values after quantifying the number of cells on each slide. So now is the sample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



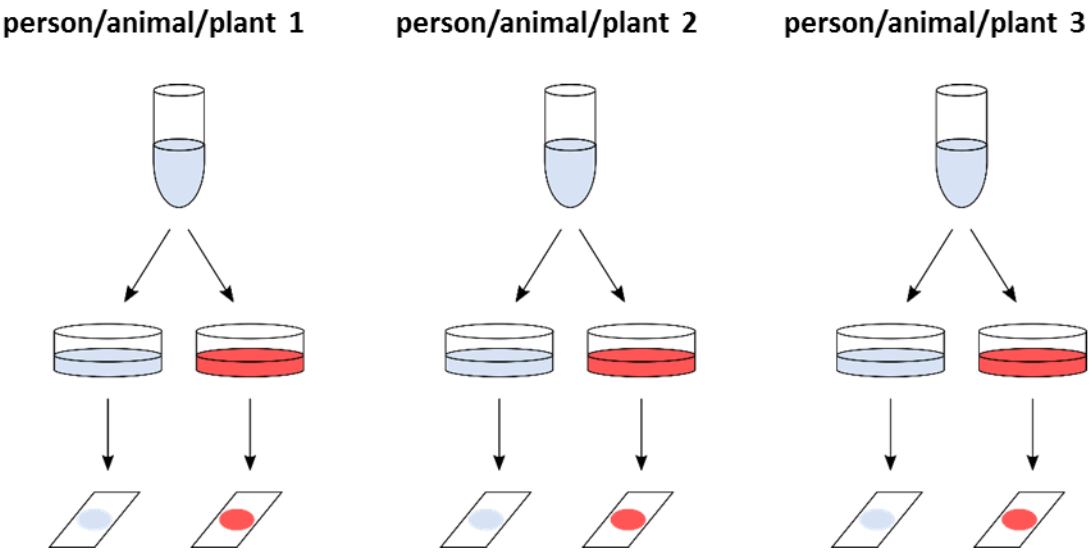
The "independent" aspect of the experiment are the days, and so n = 3. Note, that the two glass slides from the same day can (and should) be treated as paired observations, and so it is the difference between treated and control within each day that is of interest (a paired-samples t-test could be used). 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.

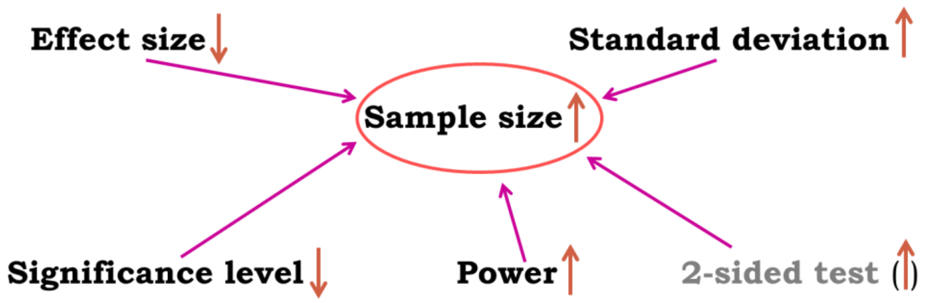


Replication at multiple levels

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!

There are many free resources available to help with sample size estimation. We are going to use G*Power which is a reasonably friendly package.

Comparing 2 proportions

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

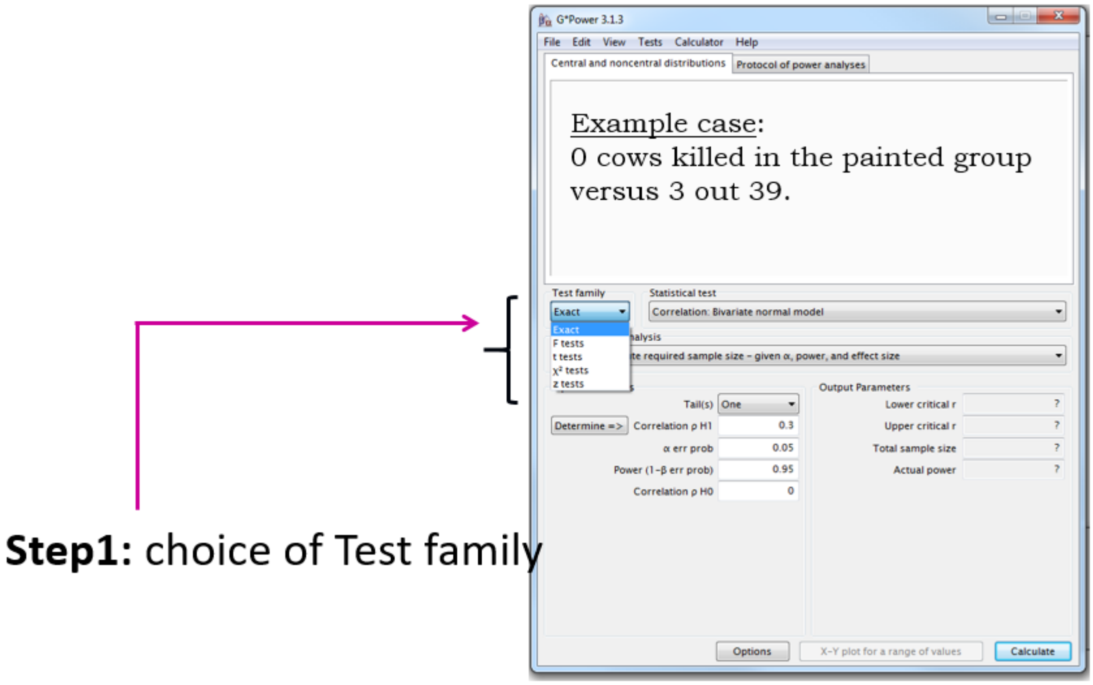
Scientists have come up with a solution that will reduce the number of lions being shot by farmers in Africa - painting eyes on the butts of cows. It sounds a little crazy, but early trials suggest that lions are less likely to attack livestock when they think they're being watched - and fewer livestock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

Using G*Power, we need to follow 4 steps.

First step: we need to choose a Test family. We are going to use a Fisher's exact test on the cow data so we pick the 'Exact' family. As part of the experimental design, we have chosen a statistical test, which means that when the time comes to run a power calculation, we know which test we are going to use.


G*Power 3.1.3
File Edit View Tests Calculator Help
Central and noncentral distributions
Protocol of power analyses
Exact
F tests
t tests
χ² tests
z tests
Determine =>
Options
X-Y plot for a range of values
Calculate

Example case:
0 cows killed in the painted group versus 3 out 39.

Test family: Exact

Statistical test: Correlation: Bivariate normal model

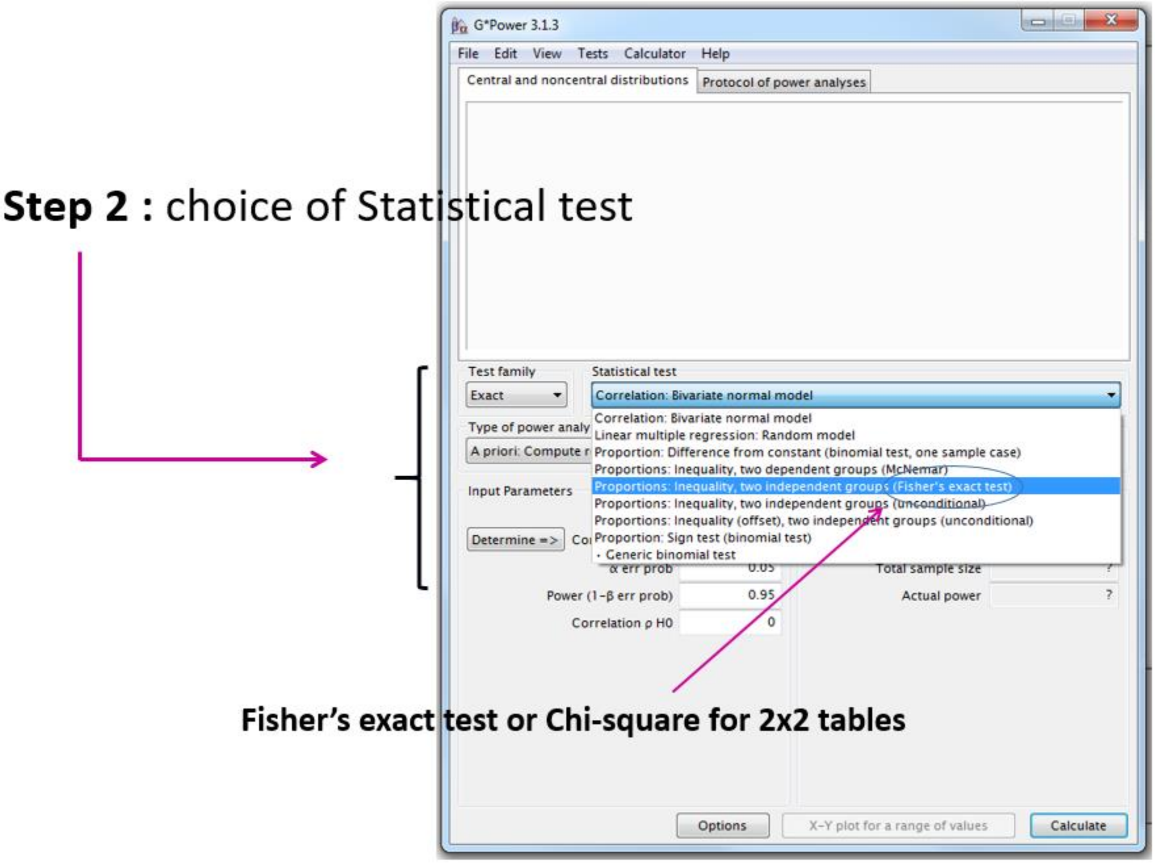
...te required sample size – given α, power, and effect size

Input Parameters: Tail(s) One; Correlation ρ H1 0.3; α err prob 0.05; Power (1-β err prob) 0.95; Correlation ρ H0 0

Output Parameters: Lower critical r ?; Upper critical r ?; Total sample size ?; Actual power ?

Step1: choice of Test family

Second step: we choose the actual test. Again, that choice should have been made at the experimental design stage.


G*Power 3.1.3
File Edit View Tests Calculator Help
Central and noncentral distributions
Protocol of power analyses
Determine =>
Options
X-Y plot for a range of values
Calculate

Step 2 : choice of Statistical test

Test family: Exact

Statistical test: Correlation: Bivariate normal model

- Correlation: Bivariate normal model
- Linear multiple regression: Random model
- Proportion: Difference from constant (binomial test, one sample case)
- Proportions: Inequality, two dependent groups (McNemar)
- Proportions: Inequality, two independent groups (Fisher's exact test)
- Proportions: Inequality, two independent groups (unconditional)
- Proportions: Inequality (offset), two independent groups (unconditional)
- Proportion: Sign test (binomial test)
- Generic binomial test

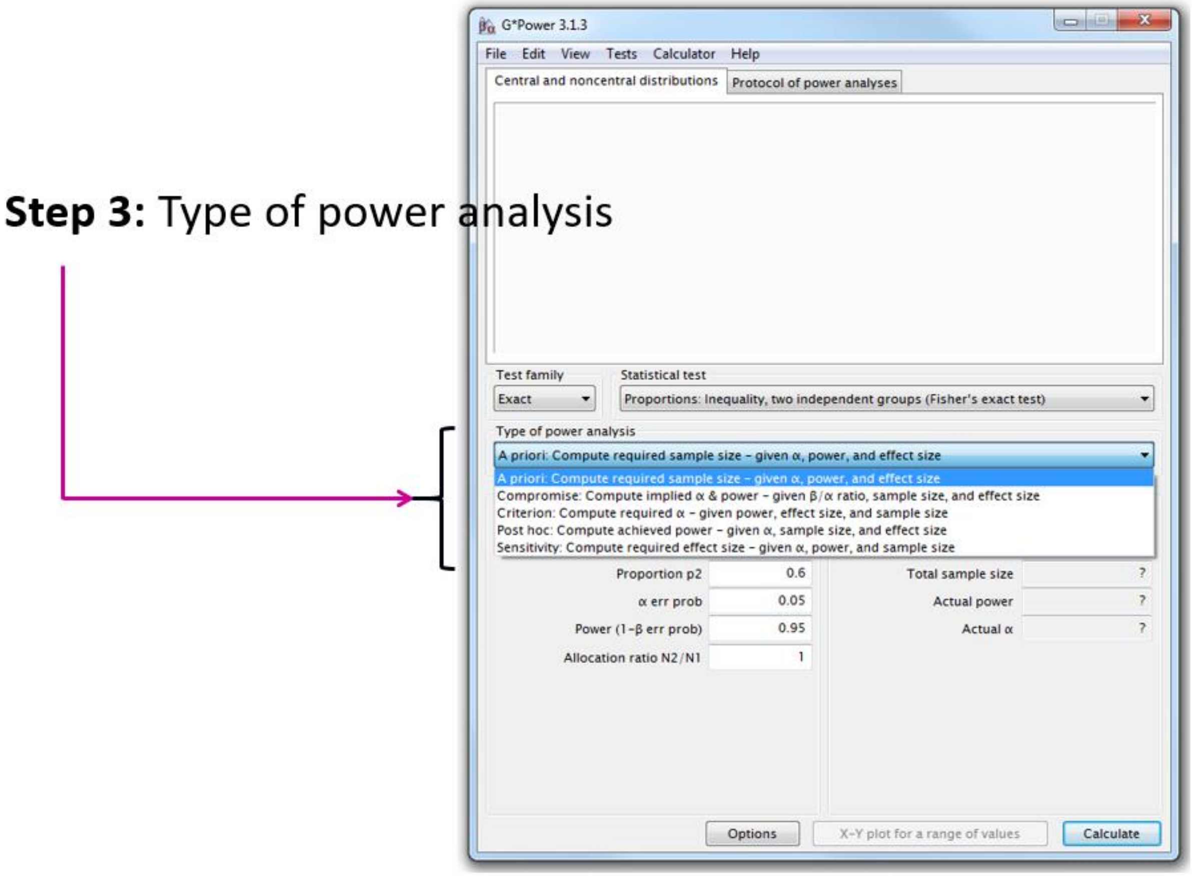
Type of power analy... A priori: Compute r...

Input Parameters: α err prob 0.05; Power (1-β err prob) 0.95; Correlation ρ H0 0

Total sample size ?; Actual power ?

Fisher's exact test or Chi-square for 2x2 tables

Then, step 3, we choose the type of power analysis we want to run. Pretty much always, we will go for the default choice: A priori: as we will want to compute the required sample size, given α, power and effect size.



Step 3: Type of power analysis

G*Power 3.1.3
File Edit View Tests Calculator Help
Central and noncentral distributions
Protocol of power analyses

Test family: Exact

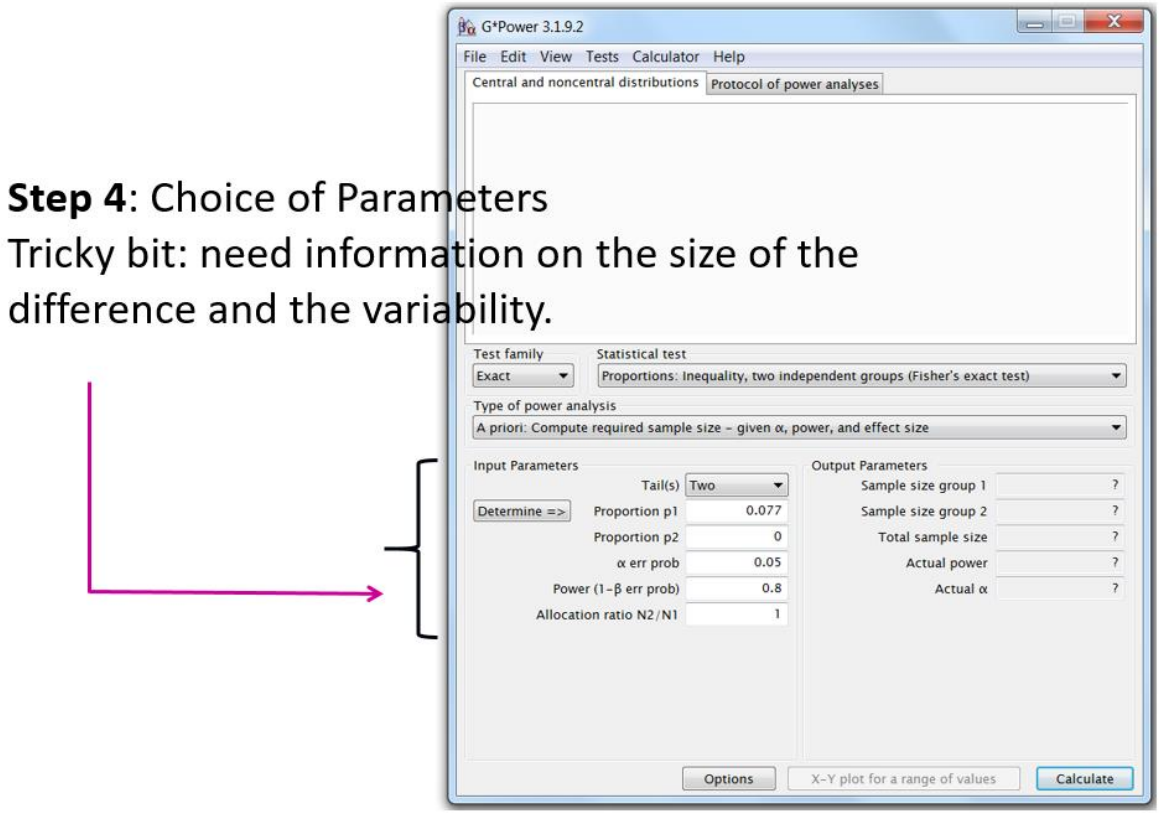
Statistical test: Proportions: Inequality, two independent groups (Fisher's exact test)

Type of power analysis: A priori: Compute required sample size – given α, power, and effect size

- A priori: Compute required sample size – given α, power, and effect size
- Compromise: Compute implied α & power – given β/α ratio, sample size, and effect size
- Criterion: Compute required α – given power, effect size, and sample size
- Post hoc: Compute achieved power – given α, sample size, and effect size
- Sensitivity: Compute required effect size – given α, power, and sample size

Proportion p2	0.6	Total sample size	?
α err prob	0.05	Actual power	?
Power (1-β err prob)	0.95	Actual α	?
Allocation ratio N2/N1	1		

Finally, the last bit is the more difficult one. We need to have/find reliable information about the data we are going to analyse. Luckily, in our case, we have data from a preliminary study.



Step 4: Choice of Parameters
Tricky bit: need information on the size of the difference and the variability.

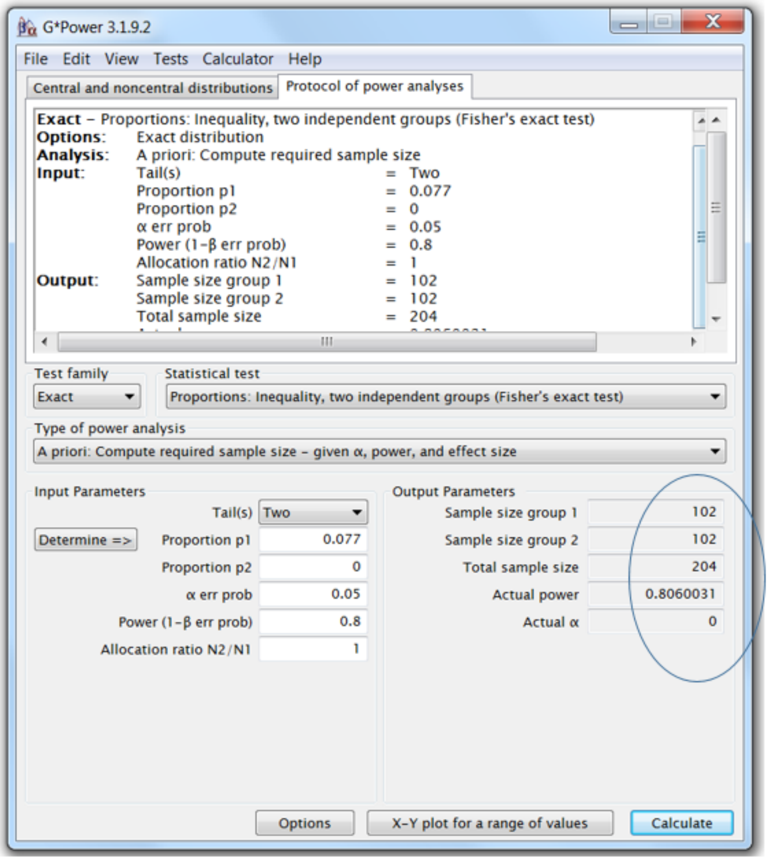
G*Power 3.1.9.2
File Edit View Tests Calculator Help
Central and noncentral distributions
Protocol of power analyses
Options
X-Y plot for a range of values
Calculate
Determine =>

Test family: Exact
Statistical test: Proportions: Inequality, two independent groups (Fisher's exact test)
Type of power analysis: A priori: Compute required sample size – given α, power, and effect size

Input Parameters
- Tail(s): Two
- Proportion p1: 0.077
- Proportion p2: 0
- α err prob: 0.05
- Power (1-β err prob): 0.8
- Allocation ratio N2/N1: 1

Output Parameters
- Sample size group 1: ?
- Sample size group 2: ?
- Total sample size: ?
- Actual power: ?
- Actual α: ?

We hit ‘Calculate’ and:



G*Power 3.1.9.2
File Edit View Tests Calculator Help
Central and noncentral distributions
Protocol of power analyses
Options
X-Y plot for a range of values
Calculate
Determine =>

```
Exact – Proportions: Inequality, two independent groups (Fisher's exact test)
Options:    Exact distribution
Analysis:   A priori: Compute required sample size
Input:      Tail(s)                       =  Two
            Proportion p1                 =  0.077
            Proportion p2                 =  0
            α err prob                    =  0.05
            Power (1-β err prob)          =  0.8
            Allocation ratio N2/N1        =  1
Output:     Sample size group 1           =  102
            Sample size group 2           =  102
            Total sample size             =  204
```

Test family: Exact
Statistical test: Proportions: Inequality, two independent groups (Fisher's exact test)
Type of power analysis: A priori: Compute required sample size – given α, power, and effect size

Input Parameters
- Tail(s): Two
- Proportion p1: 0.077
- Proportion p2: 0
- α err prob: 0.05
- Power (1-β err prob): 0.8
- Allocation ratio N2/N1: 1

Output Parameters
- Sample size group 1: 102
- Sample size group 2: 102
- Total sample size: 204
- Actual power: 0.8060031
- Actual α: 0

To be able to pick such a difference between the 2 groups, as in to reach significance, we will need about 102 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 a bit more than 200 cows altogether.

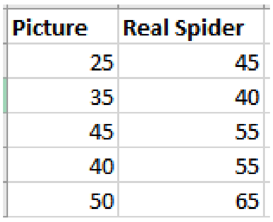
Always remember, power calculations are guessing exercises, the sample sizes found are never absolute. Our data might show an effect a bit bigger ☺ or smaller ☹ than expected. By doing a power calculation, providing the effect size we are after is meaningful, we want to know if we can afford to run the 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: 100-ish happens to be OK but if it had been 1000-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 possible. With n=3, we cannot afford any mistakes, so if something goes wrong with one of our animals for instance, we will end up with n=2 and be in trouble statswise.

Comparing 2 means

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

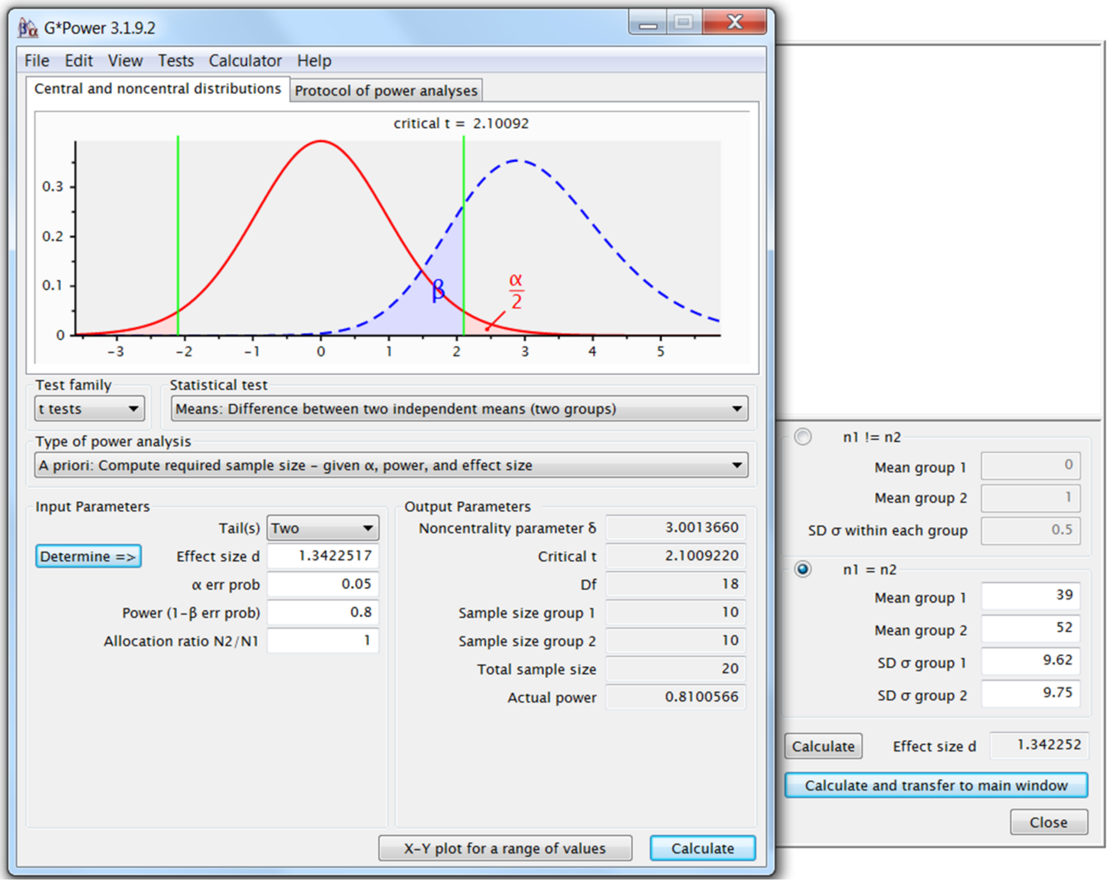
Pilot study: 10 arachnophobes were asked to perform 2 tasks:
<u>Task 1</u>: Group1 (n=5): to play with a big hairy tarantula spider with big fangs and an evil look in its eight eyes.
<u>Task 2</u>: Group 2 (n=5): to look only at pictures of the same hairy tarantula.



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

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

- Use the data to calculate the values for a power calculation.
- Run a power calculation to find out how many subjects should be included in the study.


G*Power 3.1.9.2
File Edit View Tests Calculator Help
Central and noncentral distributions
Protocol of power analyses
critical t = 2.10092
0.3
0.2
0.1
0
-3
-2
-1
0
1
2
3
4
5
β
α/2
X-Y plot for a range of values
Calculate
Determine =>
Calculate and transfer to main window
Close

Test family: t tests

Statistical test: Means: Difference between two independent means (two groups)

Type of power analysis: A priori: Compute required sample size - given α, power, and effect size

Input Parameters

Tail(s)	Two
Effect size d	1.3422517
α err prob	0.05
Power (1-β err prob)	0.8
Allocation ratio N2/N1	1

Output Parameters

Noncentrality parameter δ	3.0013660
Critical t	2.1009220
Df	18
Sample size group 1	10
Sample size group 2	10
Total sample size	20
Actual power	0.8100566

○ n1 != n2

Mean group 1	0
Mean group 2	1
SD σ within each group	0.5

● n1 = n2

Mean group 1	39
Mean group 2	52
SD σ group 1	9.62
SD σ group 2	9.75

Calculate Effect size d 1.342252

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 a 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: α =< 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.

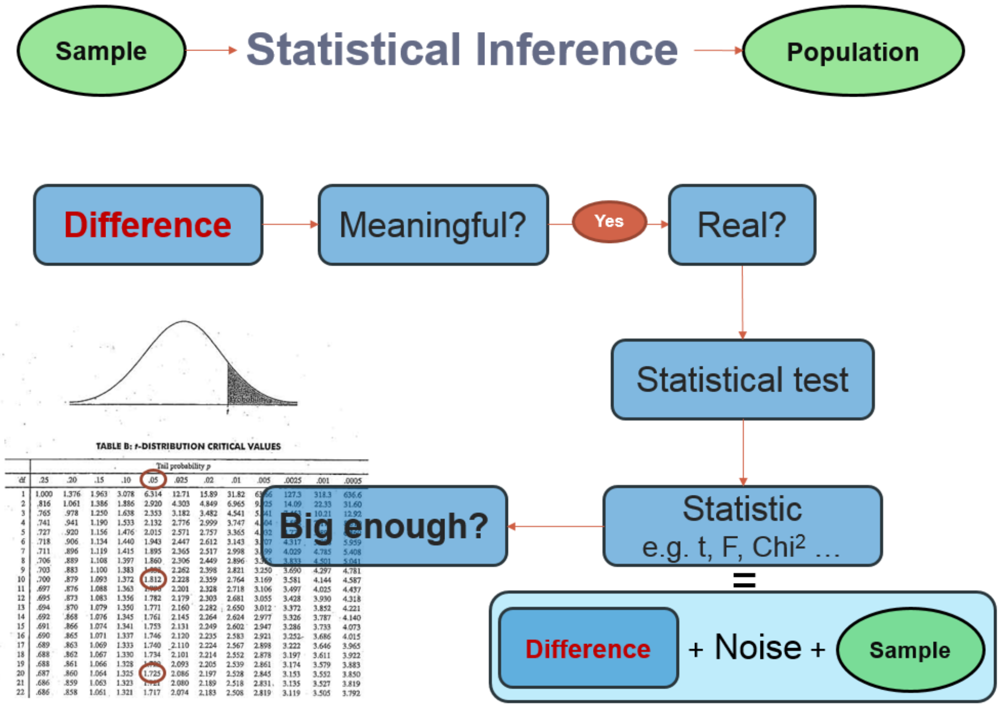
A bit of theory: Statistical inference

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.

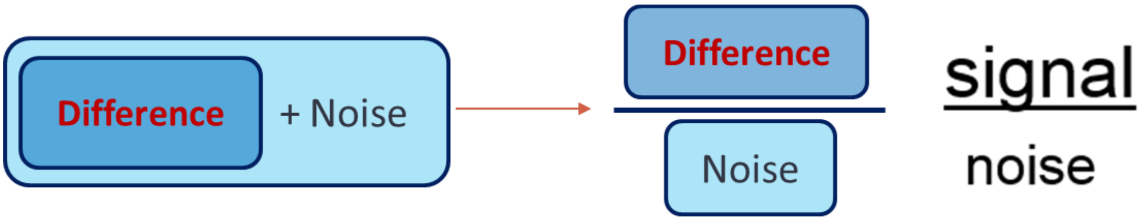
With that in mind, if we observe a difference between 2 groups in our sample, we get excited because we think that what we are observing can be what is happening in the general population, as in ‘for real’. Now when we observe a difference, we get excited if that difference is meaningful or, rather, 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, a lot of tests are available but assumptions must be met before applying them. In fact, there are 2 types of statistical 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 population parameters like normality (e.g. Mann-Whitney test).

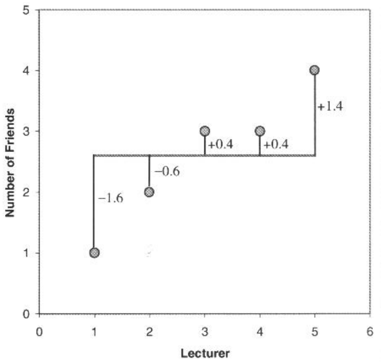
3-1 A bit of theory: descriptive stats

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 summarises 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)

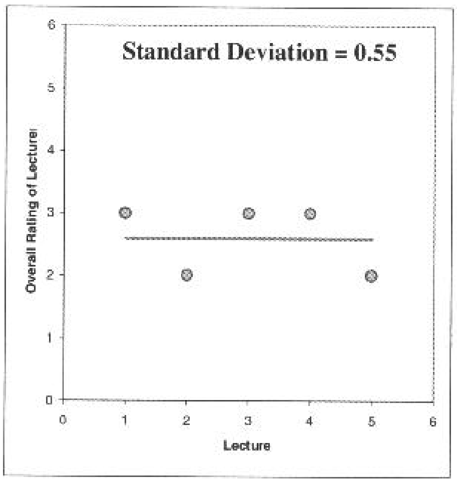
The problem with the variance is that it is measured in squared units which is not very nice to manipulate. So for more convenience, the square root of the variance is taken to obtain a measure in the same unit as the original measure: the *standard deviation.*

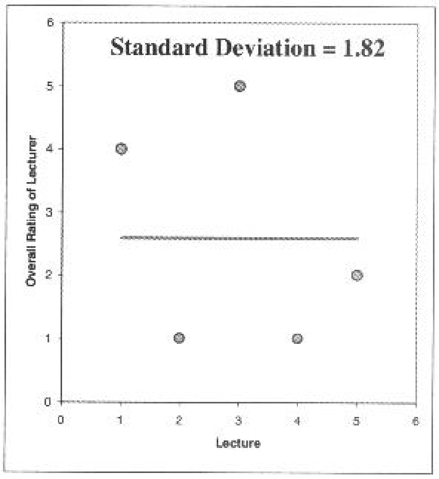
- S.D. = √(SS/N-1) = √(S^2), in our example: S.D. = √(1.3) = 1.14

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.

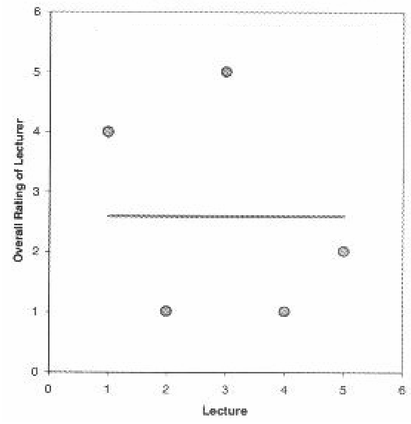
- small S.D.: data close to the mean: mean is a good fit of the data (graph on the left)
- large S.D.: data distant from the mean: mean is not an accurate representation (graph on the right)

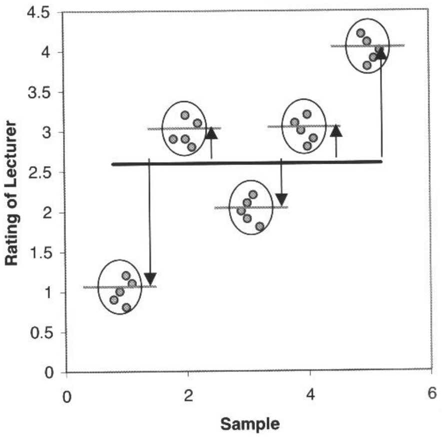




Standard Deviation vs. Standard Error

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?

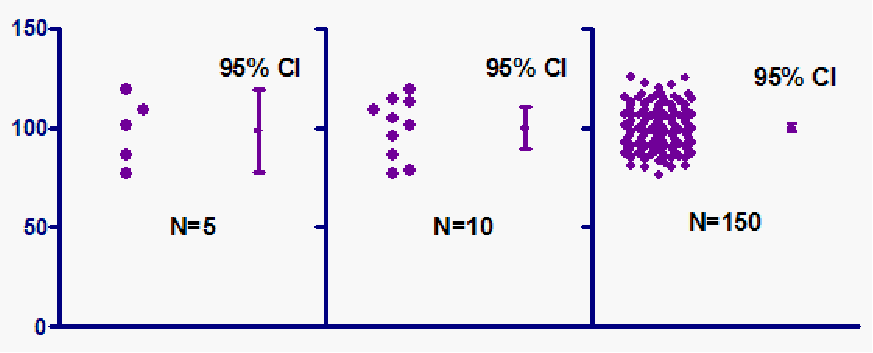
- If the scatter is caused by biological variability, it is important to show the variation. So it is more appropriate to report the S.D. rather than the S.E.M. Even better, we can show in a graph all data points, or perhaps report the largest and smallest value.
- If we are using an in vitro system 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:

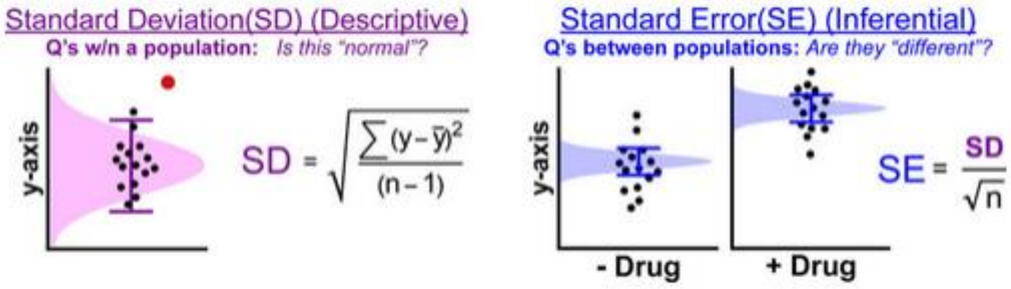


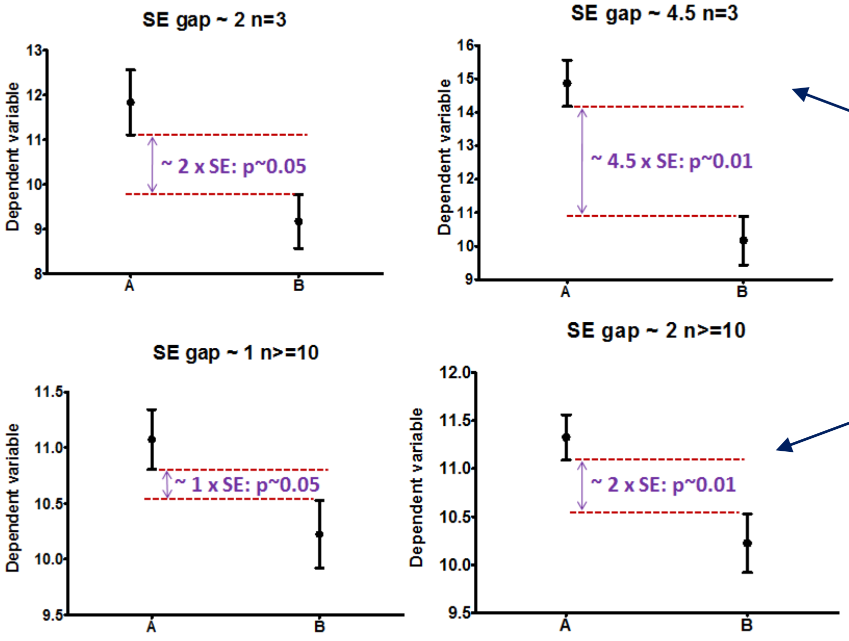
Error bars	Type	Description
Standard deviation (SD)	Descriptive	Typical or average difference between the data points and their mean.
Standard error (SEM)	Inferential	A measure of how variable the mean will be, if you repeat the whole study many times.
Confidence interval (CI), usually 95% CI	Inferential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

From Geoff Cumming *et al*. 2007

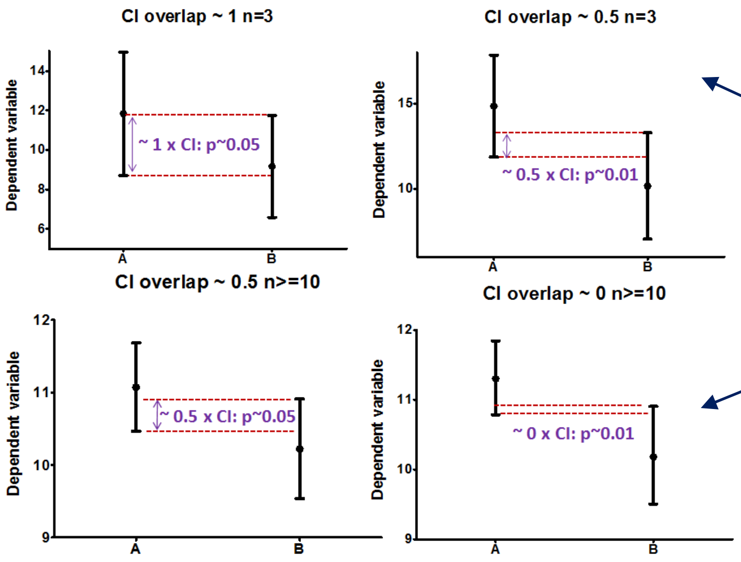
If we want to compare experimental results, it could be more appropriate to show inferential error bars such as SE or CI rather than SD. If we want to describe our sample, for instance its normality, then the SD would be the one to choose.

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.

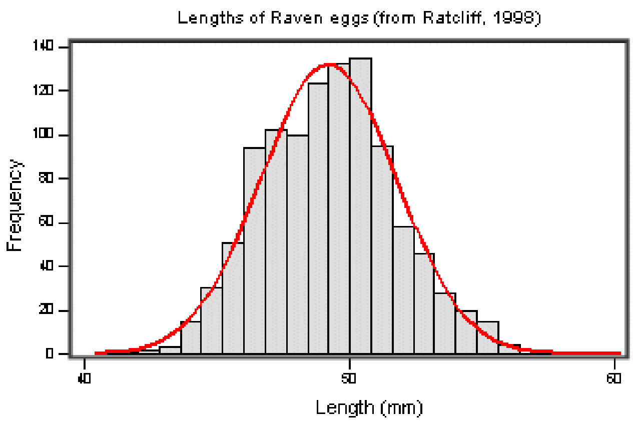


In the same way, you can estimate statistical significance using the overlap rule for 95% CI bars.

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

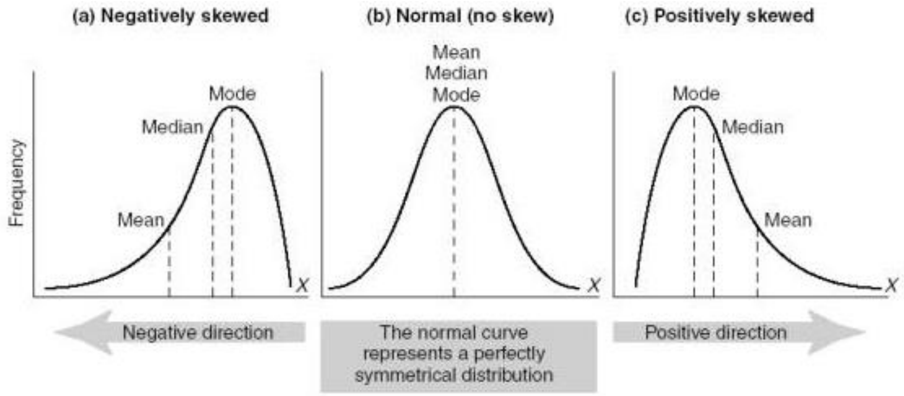
When we are dealing with quantitative data, the first thing we should look at is how they are distributed, what they look like. The distribution of our data will tell us if there is something wrong in the way we collected them or enter them and it will also tell us what kind of test we can apply to make 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:



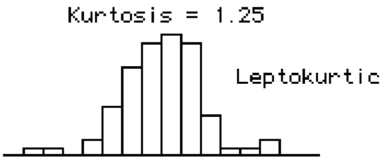
There are 2 main types of departure from normality:

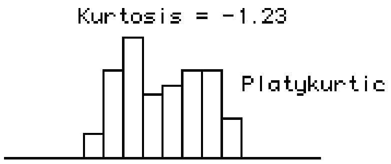
- Skewness: lack of symmetry of a distribution

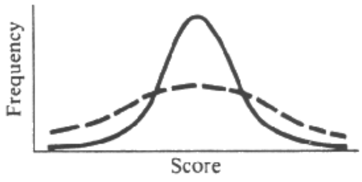


- Kurtosis: measure of the degree of 'peakedness' in the distribution

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 one measure per subject. There are specific designs for repeated measures experiments.

Chapter 4: Comparing 2 groups

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

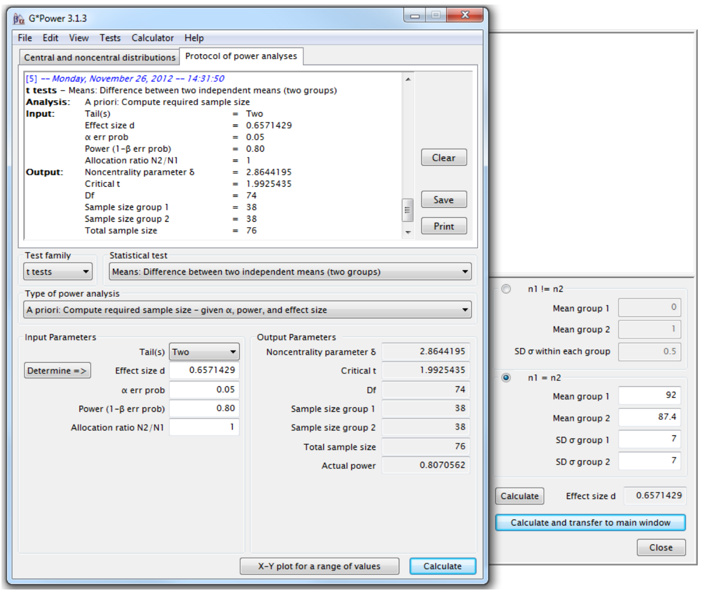
Let's try it through an example.

Example (File: `coyotes.xlsx`) 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 that 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 the one we intend to run, male coyotes were found to measure: 92cm+/- 7cm (SD). We expect a 5% difference between genders.



G*Power 3.1.3
File Edit View Tests Calculator Help
Central and noncentral distributions
Protocol of power analyses
Clear
Save
Print

```
[5] -- Monday, November 26, 2012 -- 14:31:50
t tests – Means: Difference between two independent means (two groups)
Analysis:  A priori: Compute required sample size
Input:     Tail(s)                          =  Two
           Effect size d                    =  0.6571429
           α err prob                       =  0.05
           Power (1-β err prob)             =  0.80
           Allocation ratio N2/N1           =  1
Output:    Noncentrality parameter δ        =  2.8644195
           Critical t                       =  1.9925435
           Df                               =  74
           Sample size group 1              =  38
           Sample size group 2              =  38
           Total sample size                =  76
```

Test family: t tests

Statistical test: Means: Difference between two independent means (two groups)

Type of power analysis: A priori: Compute required sample size – given α, power, and effect size

Input Parameters

Tail(s)	Two
Effect size d	0.6571429
α err prob	0.05
Power (1-β err prob)	0.80
Allocation ratio N2/N1	1

Output Parameters

Noncentrality parameter δ	2.8644195
Critical t	1.9925435
Df	74
Sample size group 1	38
Sample size group 2	38
Total sample size	76
Actual power	0.8070562

○ n1 != n2

Mean group 1	0
Mean group 2	1
SD σ within each group	0.5

● n1 = n2

Mean group 1	92
Mean group 2	87.4
SD σ group 1	7
SD σ group 2	7

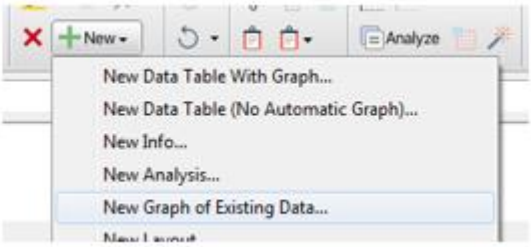
Effect size d: 0.6571429

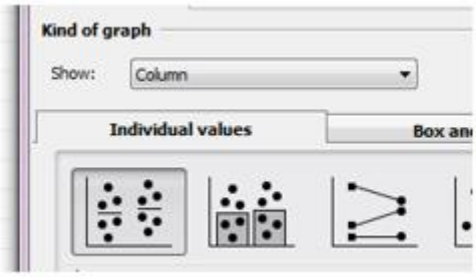
Determine =>
X-Y plot for a range of values
Calculate
Calculate
Calculate and transfer to main window
Close

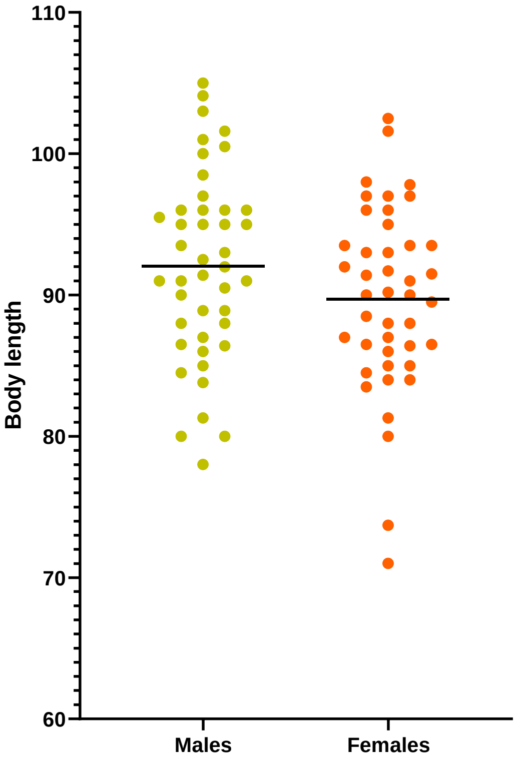
We need a sample size of n=76 (2*38).

4-2 Data exploration

Once the data are collected, we need to check that our data meet the assumptions for parametric tests. 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 GraphPad Prism:


New
Analyze
New Data Table With Graph...
New Data Table (No Automatic Graph)...
New Info...
New Analysis...
New Graph of Existing Data...
New Layout

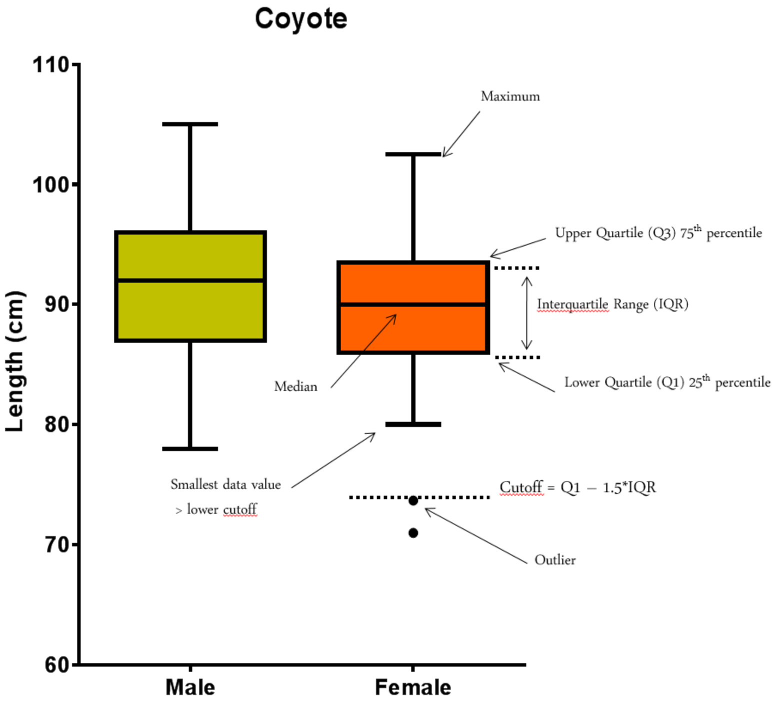




This graphical representation is very informative. It tells us about the difference between the 2 genders, of course, but much more than that: it also tells us about the sample size and the behaviour of the data. The latter is important as we need to know if our data meet the assumptions for the t-test. For that, we are looking for normality and homogeneity of variance; so on the graph, we want symmetry and balance in terms of variability between the 2 genders. Which is pretty much what we see. So far, so good.

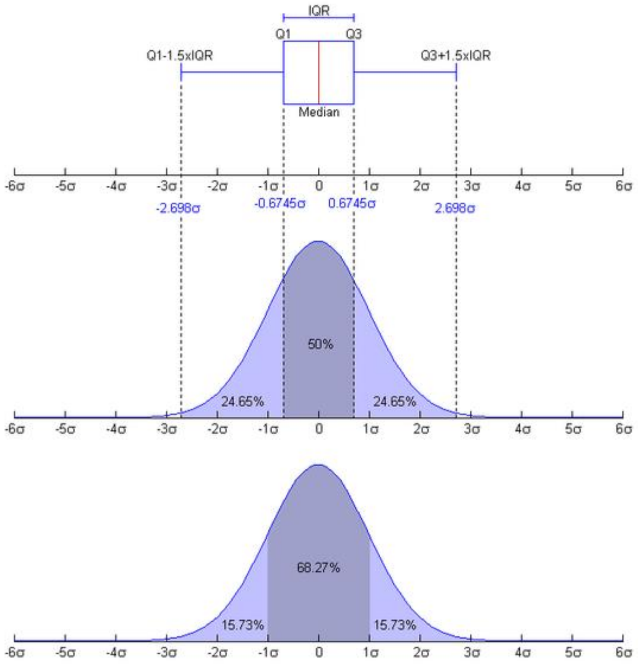
However, we can notice that 2 dots are bit out of range: 2 females seem smaller than their peers. Now the question is: are they smaller, as in just the smallest of the group, or smaller as in outliers. To find out, we need to plot the data in another way: the boxplot.

To draw a box plot we choose it from the gallery of graphs in **Column** and we choose **Tukey** for Whiskers. Tukey was the guy who invented the box plot and this particular representation allows us to identify outliers (which we will talk about later).



It is very important that we 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 our data at a glance. Below we can see the relationship between box plot and histogram. If our distribution is normal-*ish* then the box plot should be symmetrical-*ish*.

Regarding the outliers, there is no really right or wrong attitude. If there is a technical issue or an experimental problem, we should remove it, of course, but if there is nothing obvious, it is up to us. 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 us and our conscience!



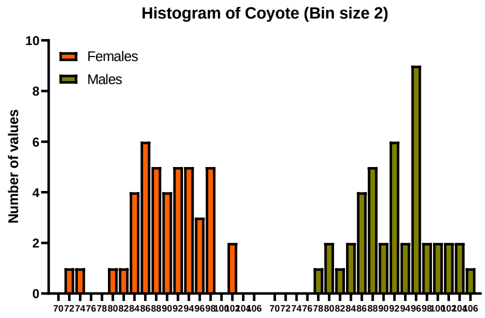
One other way to explore data is with a histogram. I think it works best with big samples, but it is still useful here so let's do it.

To draw such a graph with GraphPad, we first need to calculate the frequency distribution. To do so, we go**: =Analyze>Column Analyses>Frequency** distribution.

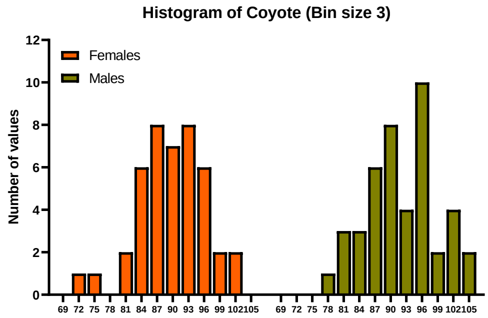
GraphPad will automatically draw a histogram from the frequency. The slightly delicate thing here is to determine the size of the bin: too small, the distribution may look anything but normal, too big, we will not see a thing. The best way is to try 2 or 3 bin size and see how it goes.

Something else to be careful about: by default, GraphPad will plot the counts (in **Tabulate> Number of Data Points**). It is OK when we plot just one group or one data set but when we want to plot several (or just 2 like here) and the groups are not of the same size then we should plot percentages (in **Tabulate> Relative frequencies as percent**) if we want to be able to compare them graphically.

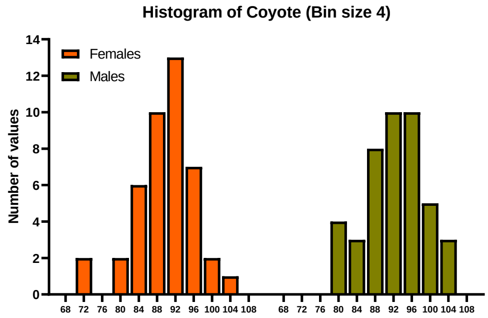
As we can see, depending of the choice of the bin size, the histograms look quite different so again, they work better with a bigger sample size than we have here.



Histogram of Coyote (Bin size 2)

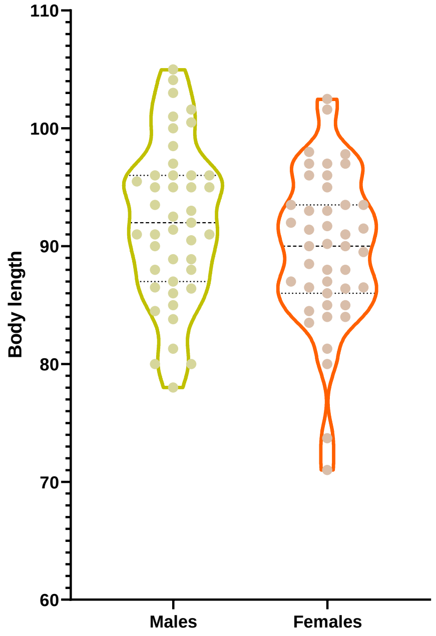


Histogram of Coyote (Bin size 3)



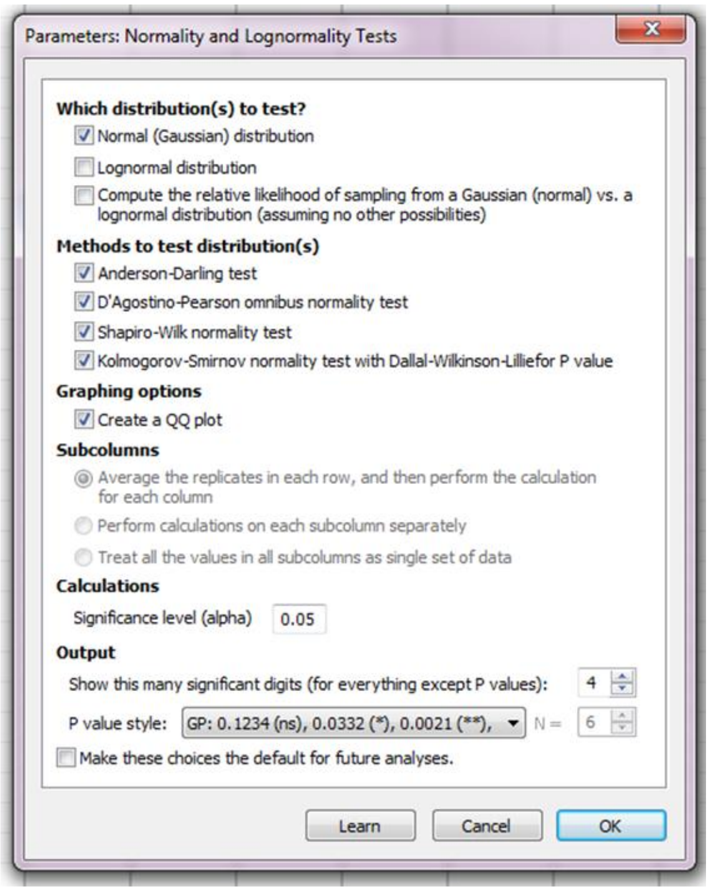
Histogram of Coyote (Bin size 4)

Finally, a totally cool way to explore and present data actually, is the violinplot.



So, we have been exploring our data quite thoroughly with scatterplots, boxplots and histograms. We are quite confident that they meet the first and the second assumptions for parametric tests but there will be occasions where data will be a bit more on the dodgy side and thus where it will be more difficult to conclude. It is possible to run tests to quantify whether or not data are departing significantly from the assumptions. Now these tests do not, and should never, replace a proper graphical exploration of the data but they can be useful when said exploration is a bit ambiguous.

First, normality, to test for it, we go: **=Analyze>Column Analyses>Column statistics.**



Parameters: Normality and Lognormality Tests

Which distribution(s) to test?

☑ Normal (Gaussian) distribution

☐ Lognormal distribution

☐ Compute the relative likelihood of sampling from a Gaussian (normal) vs. a lognormal distribution (assuming no other possibilities)

Methods to test distribution(s)

☑ Anderson-Darling test

☑ D'Agostino-Pearson omnibus normality test

☑ Shapiro-Wilk normality test

☑ Kolmogorov-Smirnov normality test with Dallal-Wilkinson-Lilliefor P value

Graphing options

☑ Create a QQ plot

Subcolumns

● Average the replicates in each row, and then perform the calculation for each column

○ Perform calculations on each subcolumn separately

○ Treat all the values in all subcolumns as single set of data

Calculations

Significance level (alpha) 0.05

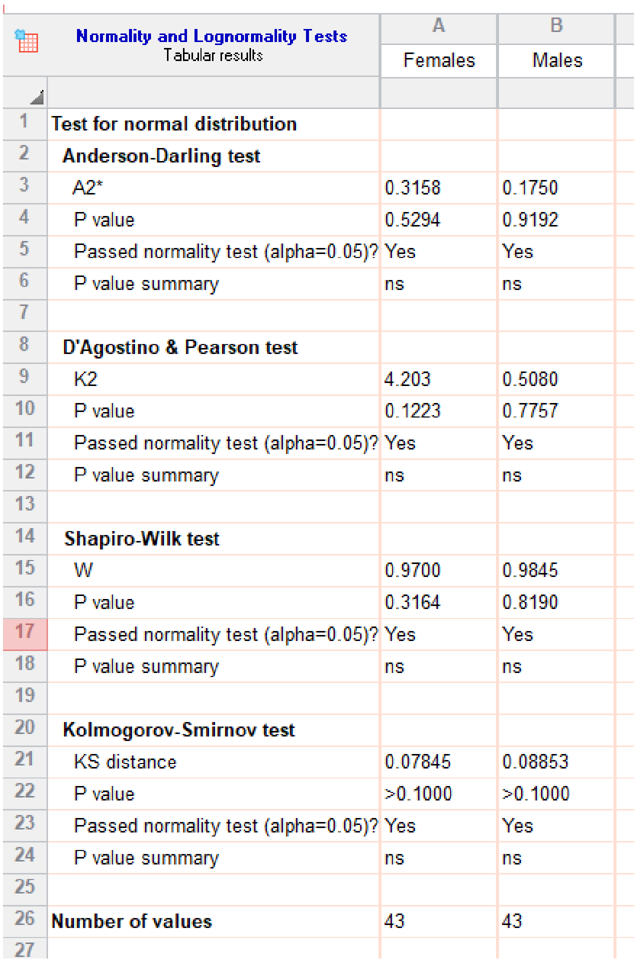
Output

Show this many significant digits (for everything except P values): 4

P value style: GP: 0.1234 (ns), 0.0332 (*), 0.0021 (**), ▾ N = 6

☐ Make these choices the default for future analyses.

GraphPad Prism offers 4 different tests for normality: Anderson-Darling, D'Agostino-Pearson, Kolmogorov-Smirnov and Shapiro-Wilk (the first 2 require n>7). They will produce different p-values but should reach the same conclusion. If we had to choose, I would probably go for D'Agostino-Pearson as it is quite a friendly one. As GraphPad puts it: ‘It first computes the skewness and kurtosis to quantify how far from Gaussian the distribution is in terms of asymmetry and shape. It then calculates how far each of these values differs from the value expected with a Gaussian distribution, and computes a single p-value from the sum of these discrepancies.’ The Kolmogorov-Smirnov test is not recommended, and the Shapiro-Wilk test is only accurate when no two values have the same value.



	Normality and Lognormality Tests Tabular results	A Females	B Males
1	**Test for normal distribution**		
2	**Anderson-Darling test**		
3	A2*	0.3158	0.1750
4	P value	0.5294	0.9192
5	Passed normality test (alpha=0.05)?	Yes	Yes
6	P value summary	ns	ns
7			
8	**D'Agostino & Pearson test**		
9	K2	4.203	0.5080
10	P value	0.1223	0.7757
11	Passed normality test (alpha=0.05)?	Yes	Yes
12	P value summary	ns	ns
13			
14	**Shapiro-Wilk test**		
15	W	0.9700	0.9845
16	P value	0.3164	0.8190
17	Passed normality test (alpha=0.05)?	Yes	Yes
18	P value summary	ns	ns
19			
20	**Kolmogorov-Smirnov test**		
21	KS distance	0.07845	0.08853
22	P value	>0.1000	>0.1000
23	Passed normality test (alpha=0.05)?	Yes	Yes
24	P value summary	ns	ns
25			
26	**Number of values**	43	43
27			

So, no significant departure from normality here which should come as no surprise after our data exploration.

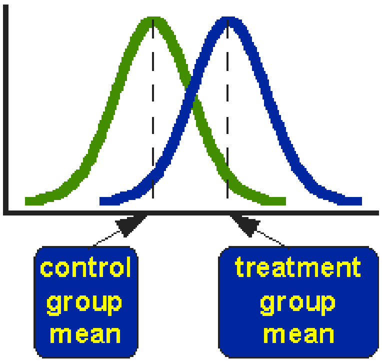
Now if our data had failed the tests, we should not be too quick to switch to nonparametric tests. While they do not assume Gaussian distributions, these tests do assume that the shape of the data distribution is the same in each group. So if your groups have very different standard deviations and so 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.

As for the second assumption, it is tested by default. When we ask for a t-test, GraphPad will calculate an F test to tell us if variances were different or not.

4-3 Student's t-test

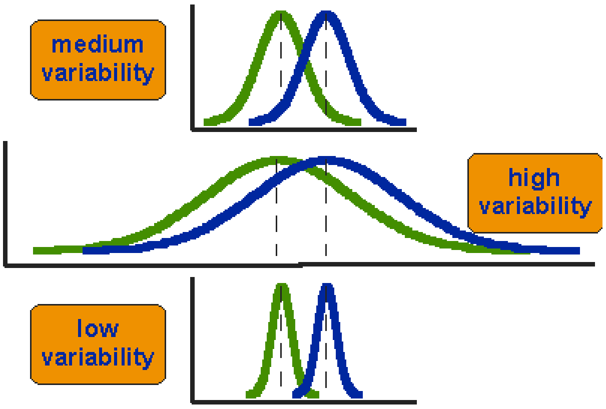
A bit of theory

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.



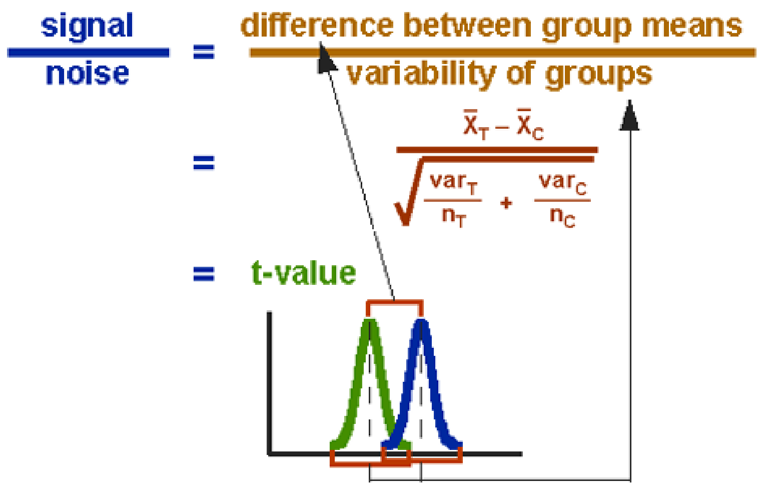
The figure above shows the distributions for the treated (blue) and control (green) groups in a study. Actually, the figure shows the idealised distribution. The figure indicates where the control and treatment group means are located. The question the t-test addresses is whether the means are statistically 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
=
$\bar{X}_T - \bar{X}_C$
$\sqrt{\frac{var_T}{n_T} + \frac{var_C}{n_C}}$
=
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.

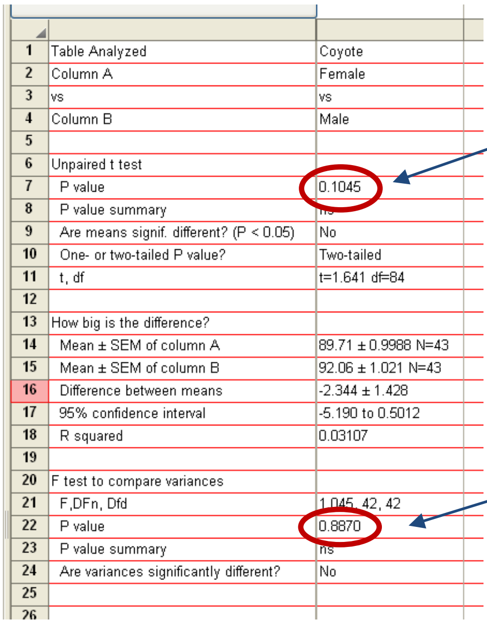
Now, we want to compare the body length between males and females in coyotes so we are going to go for an independent-test.

Independent t-test

Let's go back to our coyotes.

We go **=Analysis>Column analyses> t-tests**.

The default setting here is good as we want to run an unpaired t-test.



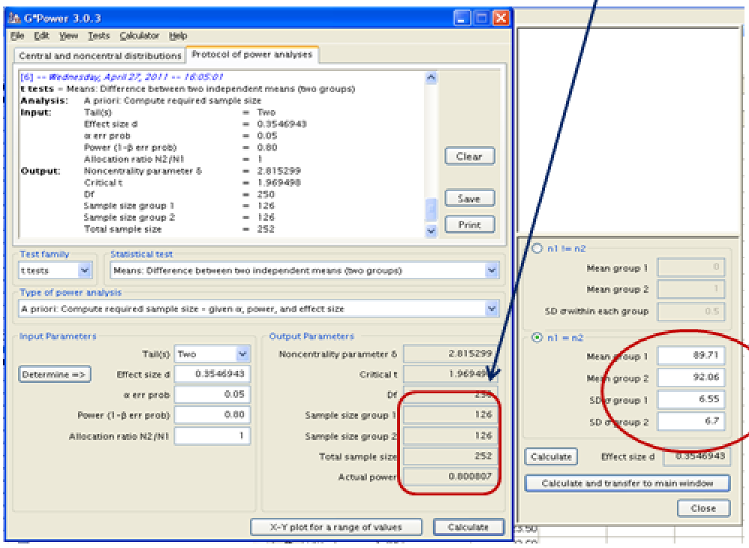
1	Table Analyzed	Coyote
2	Column A	Female
3	vs	vs
4	Column B	Male
5		
6	Unpaired t test	
7	P value	0.1045
8	P value summary	ns
9	Are means signif. different? (P < 0.05)	No
10	One- or two-tailed P value?	Two-tailed
11	t, df	t=1.641 df=84
12		
13	How big is the difference?	
14	Mean ± SEM of column A	89.71 ± 0.9988 N=43
15	Mean ± SEM of column B	92.06 ± 1.021 N=43
16	Difference between means	-2.344 ± 1.428
17	95% confidence interval	-5.190 to 0.5012
18	R squared	0.03107
19		
20	F test to compare variances	
21	F,DFn, Dfd	1.045, 42, 42
22	P value	0.8870
23	P value summary	ns
24	Are variances significantly different?	No
25		
26		

Though the males are bigger than the females, the difference between the 2 genders does not reach significance (p=0.1045).

The variances of the 2 groups are not significantly different (p=0.8870) so the second assumption for parametric test is met.

So, 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 a pilot study and run the power analysis again, we would need a sample 3 times bigger to reach a power of 80%. Now is the time to wonder whether a 2.3cm (<3%) is biologically relevant.

You would need a sample <u>3 times bigger</u>
to reach the accepted power of 80%.



G*Power 3.0.3
File Edit View Tests Calculator Help
Central and noncentral distributions
Protocol of power analyses
Clear
Save
Print
Determine =>
X-Y plot for a range of values
Calculate
Calculate
Calculate and transfer to main window
Close

```
[6] -- Wednesday, April 27, 2011 -- 16:05:01
t tests - Means: Difference between two independent means (two groups)
Analysis:   A priori: Compute required sample size
Input:      Tail(s)                          =  Two
            Effect size d                    =  0.3546943
            α err prob                       =  0.05
            Power (1-β err prob)             =  0.80
            Allocation ratio N2/N1           =  1
Output:     Noncentrality parameter δ        =  2.815299
            Critical t                       =  1.969498
            Df                               =  250
            Sample size group 1              =  126
            Sample size group 2              =  126
            Total sample size                =  252
```

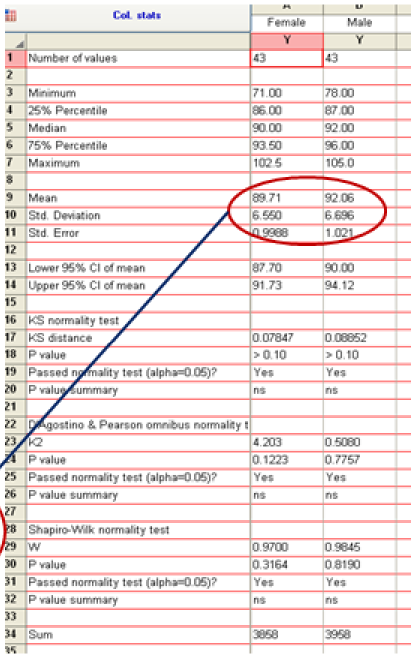
Test family: t tests
Statistical test: Means: Difference between two independent means (two groups)
Type of power analysis: A priori: Compute required sample size - given α, power, and effect size

Input Parameters
Tail(s): Two
Effect size d: 0.3546943
α err prob: 0.05
Power (1-β err prob): 0.80
Allocation ratio N2/N1: 1

Output Parameters
Noncentrality parameter δ: 2.815299
Critical t: 1.969498
Df: 250
Sample size group 1: 126
Sample size group 2: 126
Total sample size: 252
Actual power: 0.800807

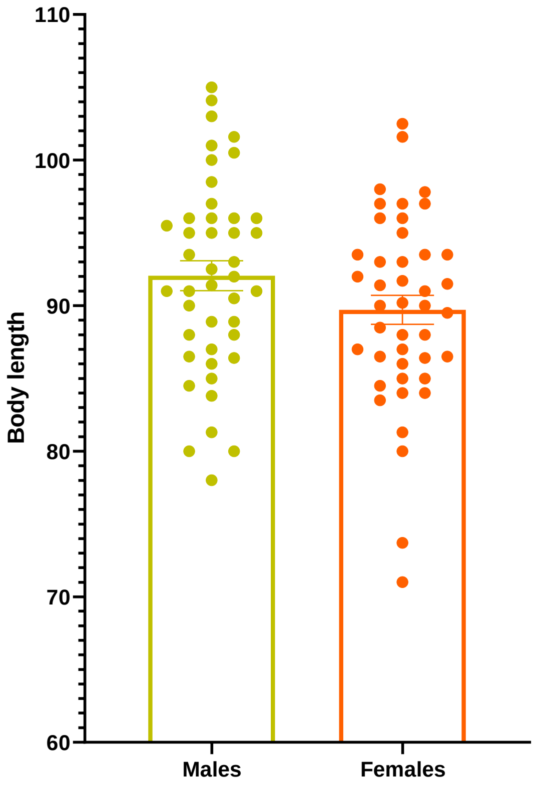
○ n1 != n2
Mean group 1: 0
Mean group 2: 1
SD σ within each group: 0.5

● n1 = n2
Mean group 1: 89.71
Mean group 2: 92.06
SD σ group 1: 6.55
SD σ group 2: 6.7
Effect size d: 0.3546943



	Col. stats	A Female Y	B Male Y
1	Number of values	43	43
2			
3	Minimum	71.00	78.00
4	25% Percentile	86.00	87.00
5	Median	90.00	92.00
6	75% Percentile	93.50	96.00
7	Maximum	102.5	105.0
8			
9	Mean	89.71	92.06
10	Std. Deviation	6.550	6.696
11	Std. Error	0.9988	1.021
12			
13	Lower 95% CI of mean	87.70	90.00
14	Upper 95% CI of mean	91.73	94.12
15			
16	KS normality test		
17	KS distance	0.07847	0.08852
18	P value	> 0.10	> 0.10
19	Passed normality test (alpha=0.05)?	Yes	Yes
20	P value summary	ns	ns
21			
22	D'Agostino & Pearson omnibus normality t		
23	K2	4.203	0.5080
24	P value	0.1223	0.7757
25	Passed normality test (alpha=0.05)?	Yes	Yes
26	P value summary	ns	ns
27			
28	Shapiro-Wilk normality test		
29	W	0.9700	0.9845
30	P value	0.3164	0.8190
31	Passed normality test (alpha=0.05)?	Yes	Yes
32	P value summary	ns	ns
33			
34	Sum	3858	3958
35			

Finally, a way to combine nicely the visual strength of a bar chart to the information given by a scatterplot is the graph below.

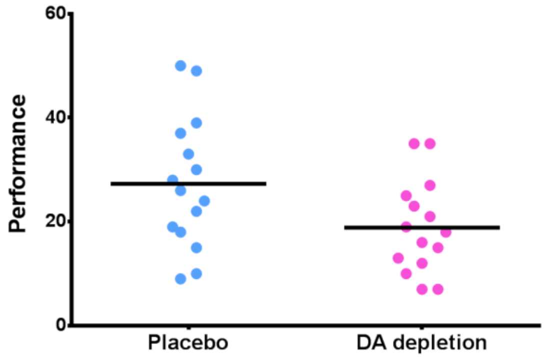


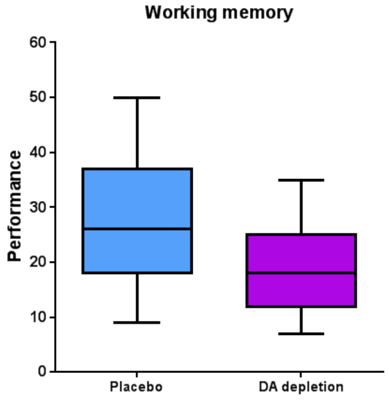
Paired t-test

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.

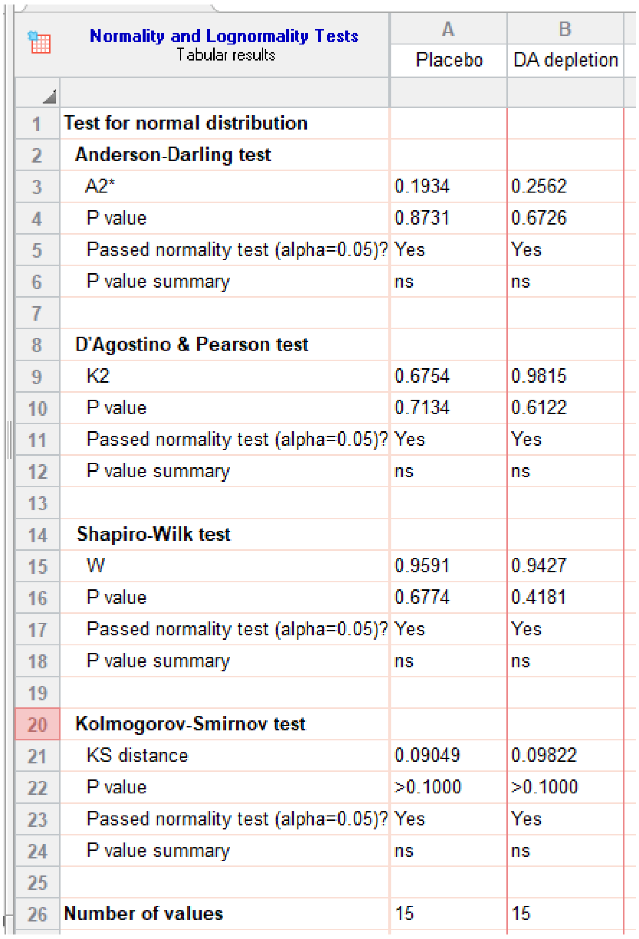
A researcher studying the effects of dopamine (DA) depletion on working memory in rhesus monkeys, tested working memory performance in 15 monkeys after administration of a saline (placebo) injection and again after injecting a dopamine-depleting agent.

Example (File: `working memory.xlsx`)





From the graph above, we observe that performance is lower with DA depletion but the difference is not 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 box plots above seem to indicate that there is no significant departure from normality and this is confirmed by the D'Agostino & Pearson test.



	Normality and Lognormality Tests Tabular results	A Placebo	B DA depletion
1	**Test for normal distribution**		
2	**Anderson-Darling test**		
3	A2*	0.1934	0.2562
4	P value	0.8731	0.6726
5	Passed normality test (alpha=0.05)?	Yes	Yes
6	P value summary	ns	ns
7			
8	**D'Agostino & Pearson test**		
9	K2	0.6754	0.9815
10	P value	0.7134	0.6122
11	Passed normality test (alpha=0.05)?	Yes	Yes
12	P value summary	ns	ns
13			
14	**Shapiro-Wilk test**		
15	W	0.9591	0.9427
16	P value	0.6774	0.4181
17	Passed normality test (alpha=0.05)?	Yes	Yes
18	P value summary	ns	ns
19			
20	**Kolmogorov-Smirnov test**		
21	KS distance	0.09049	0.09822
22	P value	>0.1000	>0.1000
23	Passed normality test (alpha=0.05)?	Yes	Yes
24	P value summary	ns	ns
25			
26	**Number of values**	15	15

Normality ☑

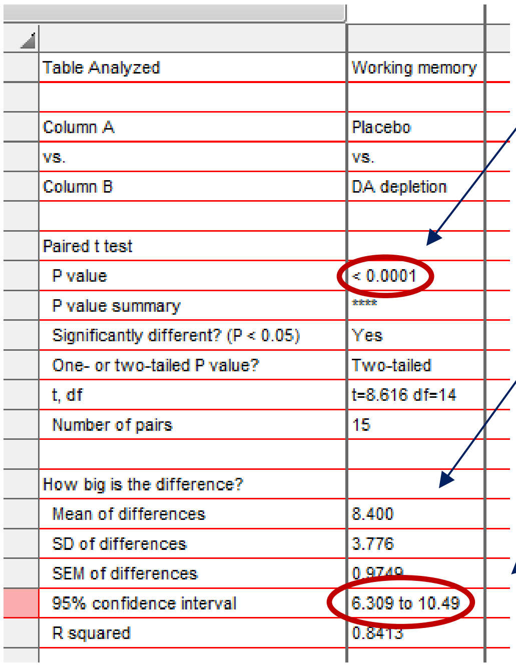


Table Analyzed	Working memory
Column A	Placebo
vs.	vs.
Column B	DA depletion
Paired t test	
P value	< 0.0001
P value summary	****
Significantly different? (P < 0.05)	Yes
One- or two-tailed P value?	Two-tailed
t, df	t=8.616 df=14
Number of pairs	15
How big is the difference?	
Mean of differences	8.400
SD of differences	3.776
SEM of differences	0.9749
95% confidence interval	6.309 to 10.49
R squared	0.8413

There is a significant difference between the 2 groups (p<0.0001).

On average, monkeys lose over 8 points in working memory performance after the injection of the dopamine-depletion agent.

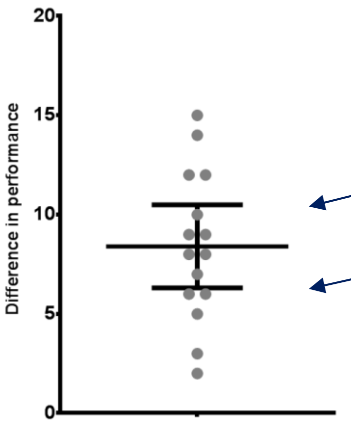
The confidence interval does not include 0 hence the significance.

The paired t-test turns out to be highly significant (see Table above). So, how come the graph and the test tell us different things?
The problem is that we don't really want to compare the mean performance of the monkeys in the 2 groups, we want to look at the difference pair-wise. In other words we want to know if, on average, a given monkey is doing better or worse after having received the dopamine-depleting agent. So, we are interested in the mean difference.

Unfortunately, one of the down sides of GraphPad is that we cannot manipulate the data much. For instance, there is no equivalent of Excel's Function with which one can apply formulae to join several values. In our case, we want to calculate and plot the differences between the 2 conditions. But we can do a few things. Such as work out the difference between values in adjacent columns. To do so we go:

Analyze>Transform>Remove baseline and column math.

The graph representing the difference is displayed below and one can see that the confidence interval does not include 0, meaning that the difference is likely to be significantly different from 0, which we already know by the paired t-test.

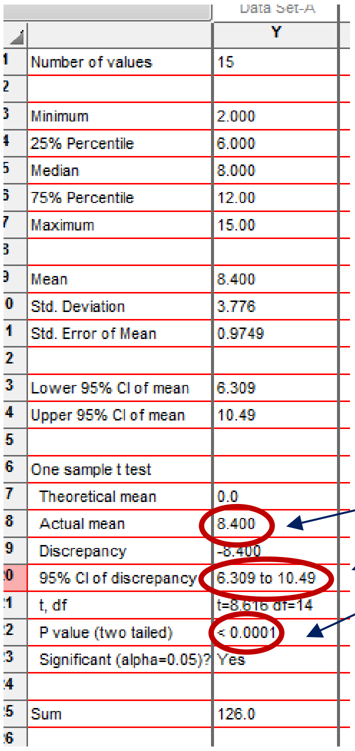


10.49

6.309

Confidence Interval

Now try to run a One Sample t-test which we will find under **Column Analysis > Column Statistics**.



		Data Set-A Y
1	Number of values	15
2		
3	Minimum	2.000
4	25% Percentile	6.000
5	Median	8.000
6	75% Percentile	12.00
7	Maximum	15.00
8		
9	Mean	8.400
10	Std. Deviation	3.776
11	Std. Error of Mean	0.9749
12		
13	Lower 95% CI of mean	6.309
14	Upper 95% CI of mean	10.49
15		
16	One sample t test	
17	Theoretical mean	0.0
18	Actual mean	8.400
19	Discrepancy	-8.400
20	95% CI of discrepancy	6.309 to 10.49
21	t, df	t=8.616 df=14
22	P value (two tailed)	< 0.0001
23	Significant (alpha=0.05)?	Yes
24		
25	Sum	126.0
26		

Same values as for the paired t-test.

We will have noticed that GraphPad does 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 groups of rhesus monkeys.

4-4 Non-parametric data

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.

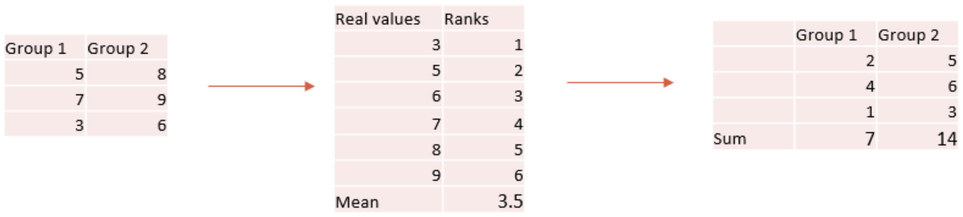
Non-parametric tests are not strictly speaking assumption-free. For instance, they assume a continuous distribution, though scores are OK, but they are far less restrictive than their parametric counterparts. Most of them are based on the principle of ranking: the smallest value has the lowest rank and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

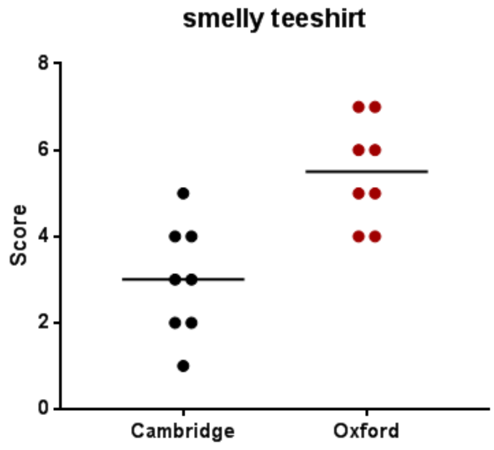
	Group 1	Group 2
	2	5
	4	6
	1	3
Sum	7	14

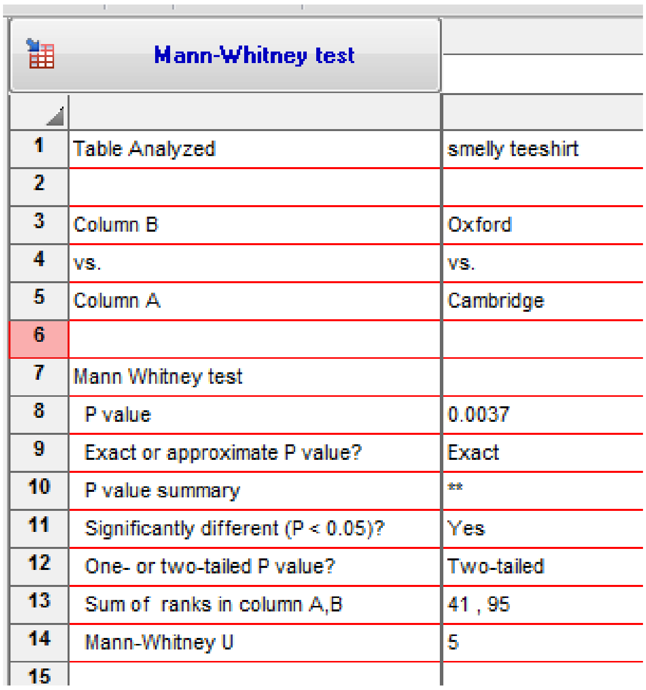
- Statistic of the Mann-Whitney test: **W (U)**
  - W = sum of ranks – mean rank: W_1=3.5 and W_2=10.5
  - Smallest of the 2 Ws: W_1 + sample size → **p-value**

Let’s try an example.

Example (File: `smelly teeshirt.xlsx`)

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.





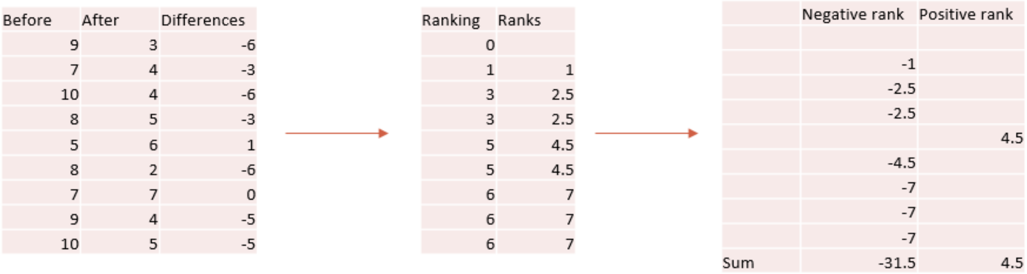
	Mann-Whitney test	
1	Table Analyzed	smelly teeshirt
2		
3	Column B	Oxford
4	vs.	vs.
5	Column A	Cambridge
6		
7	Mann Whitney test	
8	P value	0.0037
9	Exact or approximate P value?	Exact
10	P value summary	**
11	Significantly different (P < 0.05)?	Yes
12	One- or two-tailed P value?	Two-tailed
13	Sum of ranks in column A,B	41 , 95
14	Mann-Whitney U	5
15		

So, not surprisingly (haha!), 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

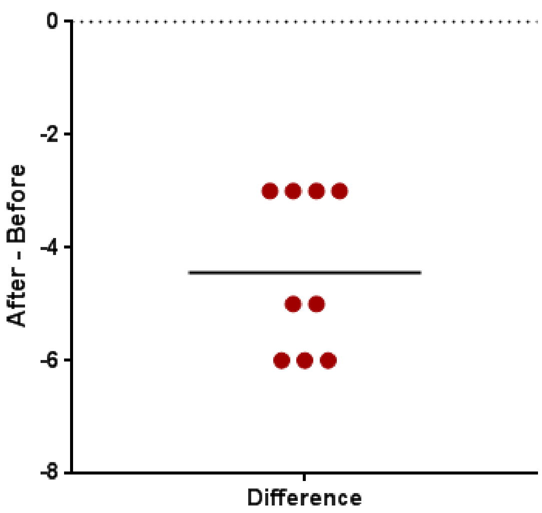
Ranking	Ranks
0	
1	1
3	2.5
3	2.5
5	4.5
5	4.5
6	7
6	7
6	7

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

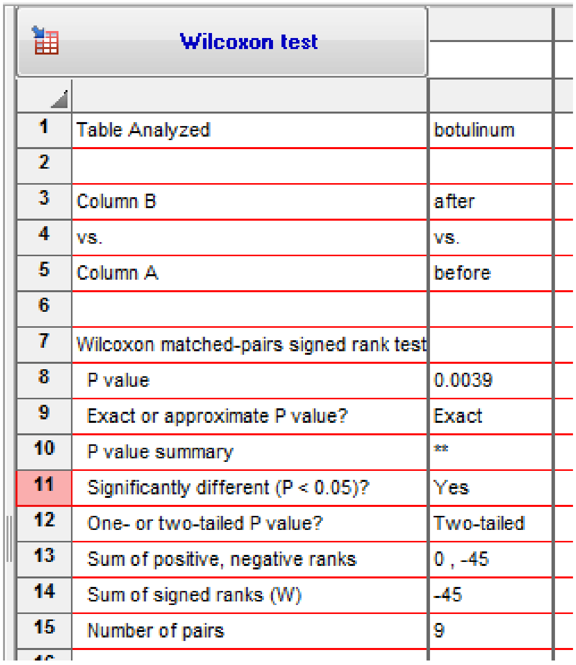
- Statistic of the Wilcoxon's signed-rank test: **T (W)**
  - Here: Wilcoxon's T = 4.5 (smallest of the 2 (absolute value))
  - Sample size N = 9 (we ignore the 0 difference): T + N → **p-value**

Example (File: `botulinum.xlsx`)

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 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 are presented in the file `botulinum. xlsx`.



```
  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
```



1	Table Analyzed	botulinum
2		
3	Column B	after
4	vs.	vs.
5	Column A	before
6		
7	Wilcoxon matched-pairs signed rank test	
8	P value	0.0039
9	Exact or approximate P value?	Exact
10	P value summary	**
11	Significantly different (P < 0.05)?	Yes
12	One- or two-tailed P value?	Two-tailed
13	Sum of positive, negative ranks	0 , -45
14	Sum of signed ranks (W)	-45
15	Number of pairs	9

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.

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:

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

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 <u>not making</u> 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.

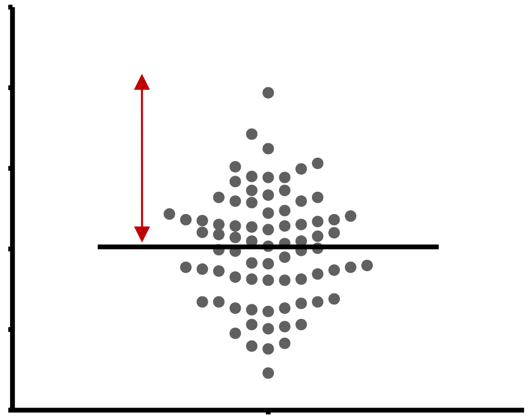
Below, is a typical example of analysis of variance table

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

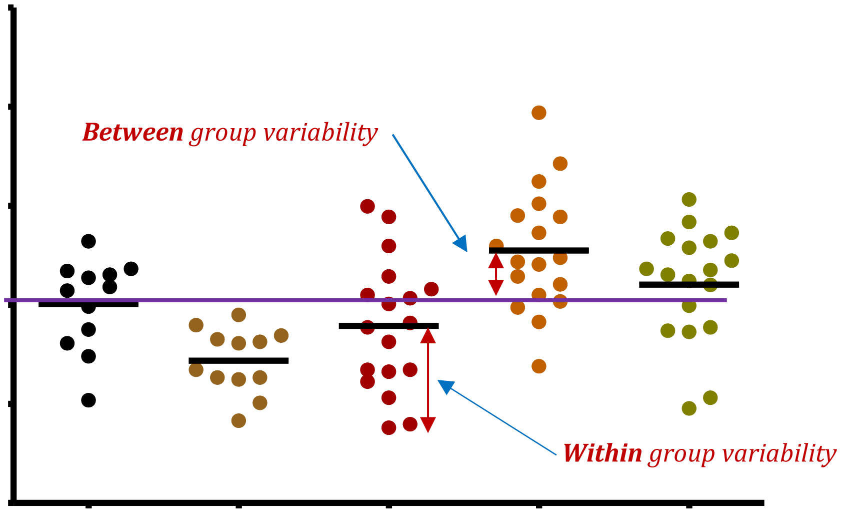
In our case, Total SS = 8.44. If we were to plot your data to represent the total SS, we would produce the graph below. So the total SS is the squared sum of all the differences between each data point and the grand mean. This is a quantification of the overall variability in our data. The next step 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 (∑ 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 = ∑ (x_i - 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

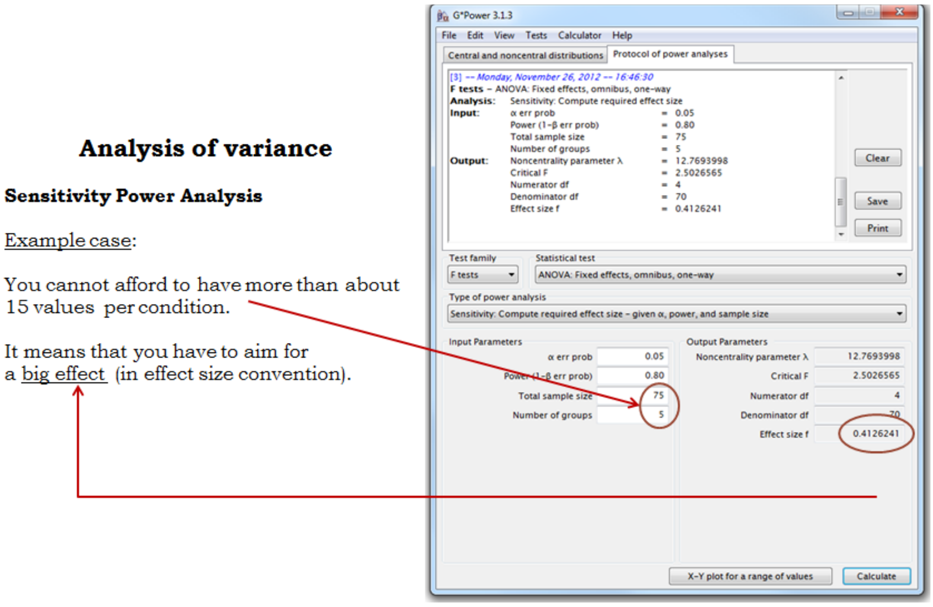
The level of significance of the test is calculated by taking into account the F ratio and the number of df (degree of freedom) for the numerator and the denominator. In our example, p<0.0001, so the test is highly significant and we are more than 99% confident when we say that there is a difference 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 in more detail later.

Example (File: `protein expression.xslx`)

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

As usual, we start by designing our experiment, we decide to go for an analysis of variance and then, we get to the point where we have to decide on the sample size.

Power analysis with an ANOVA



Analysis of variance

Sensitivity Power Analysis

Example case:

You cannot afford to have more than about 15 values per condition.

It means that you have to aim for a big effect (in effect size convention).

G*Power 3.1.3
File
Edit
View
Tests
Calculator
Help
Central and noncentral distributions
Protocol of power analyses
Clear
Save
Print
X-Y plot for a range of values
Calculate

```
[3] -- Monday, November 26, 2012 -- 16:46:30
F tests - ANOVA: Fixed effects, omnibus, one-way
Analysis:  Sensitivity: Compute required effect size
Input:     α err prob                    =  0.05
           Power (1-β err prob)          =  0.80
           Total sample size             =  75
           Number of groups              =  5
Output:    Noncentrality parameter λ     =  12.7693998
           Critical F                    =  2.5026565
           Numerator df                  =  4
           Denominator df                =  70
           Effect size f                 =  0.4126241
```

Test family: F tests

Statistical test: ANOVA: Fixed effects, omnibus, one-way

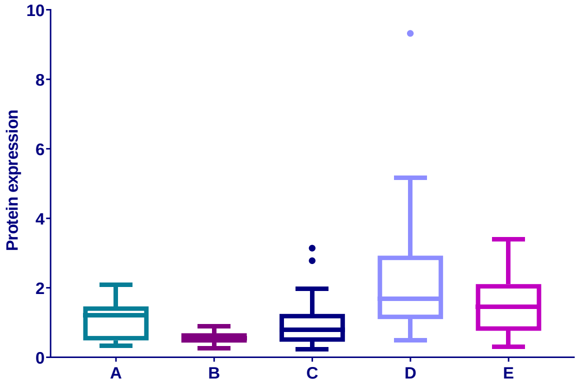
Type of power analysis: Sensitivity: Compute required effect size – given α, power, and sample size

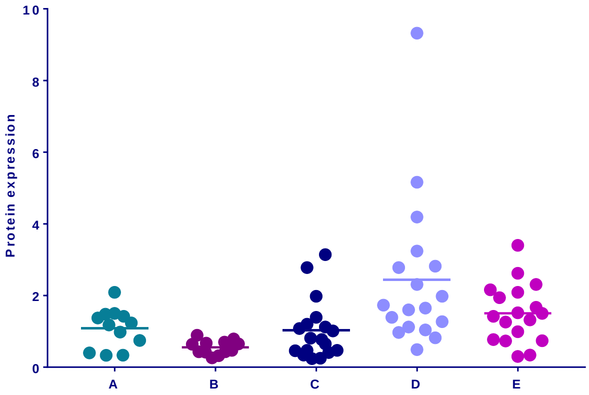
Input Parameters
α err prob: 0.05
Power (1-β err prob): 0.80
Total sample size: 75
Number of groups: 5

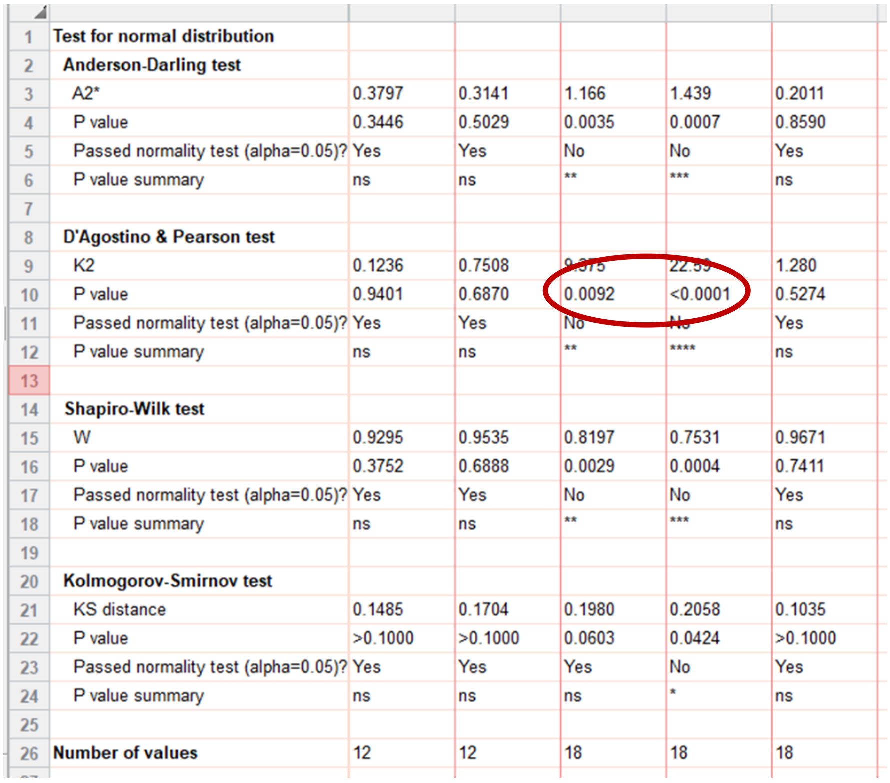
Output Parameters
Noncentrality parameter λ: 12.7693998
Critical F: 2.5026565
Numerator df: 4
Denominator df: 70
Effect size f: 0.4126241

We are quite confident in our hypothesis so we decide to go ahead.

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 (See Table below). As for the homogeneity of variance, even before testing it, a look at the box plots (see Graph) 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).

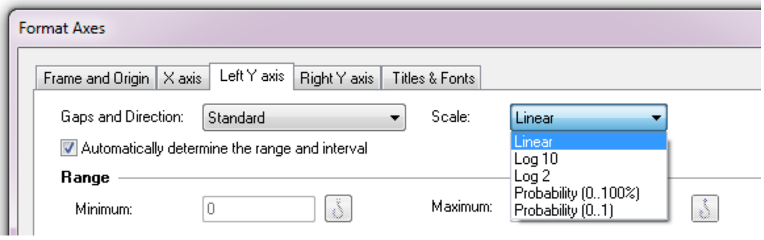






1	Test for normal distribution					
2	Anderson-Darling test					
3	A2*	0.3797	0.3141	1.166	1.439	0.2011
4	P value	0.3446	0.5029	0.0035	0.0007	0.8590
5	Passed normality test (alpha=0.05)?	Yes	Yes	No	No	Yes
6	P value summary	ns	ns	**	***	ns
7						
8	D'Agostino & Pearson test					
9	K2	0.1236	0.7508	9.375	22.59	1.280
10	P value	0.9401	0.6870	0.0092	<0.0001	0.5274
11	Passed normality test (alpha=0.05)?	Yes	Yes	No	No	Yes
12	P value summary	ns	ns	**	****	ns
13						
14	Shapiro-Wilk test					
15	W	0.9295	0.9535	0.8197	0.7531	0.9671
16	P value	0.3752	0.6888	0.0029	0.0004	0.7411
17	Passed normality test (alpha=0.05)?	Yes	Yes	No	No	Yes
18	P value summary	ns	ns	**	***	ns
19						
20	Kolmogorov-Smirnov test					
21	KS distance	0.1485	0.1704	0.1980	0.2058	0.1035
22	P value	>0.1000	>0.1000	0.0603	0.0424	>0.1000
23	Passed normality test (alpha=0.05)?	Yes	Yes	Yes	No	Yes
24	P value summary	ns	ns	ns	*	ns
25						
26	Number of values	12	12	18	18	18

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. Let's see how our data behave on a log-scale. To do that, we simply double-click on the y-axis and change linear for log.



Format Axes

Gaps and Direction: Standard

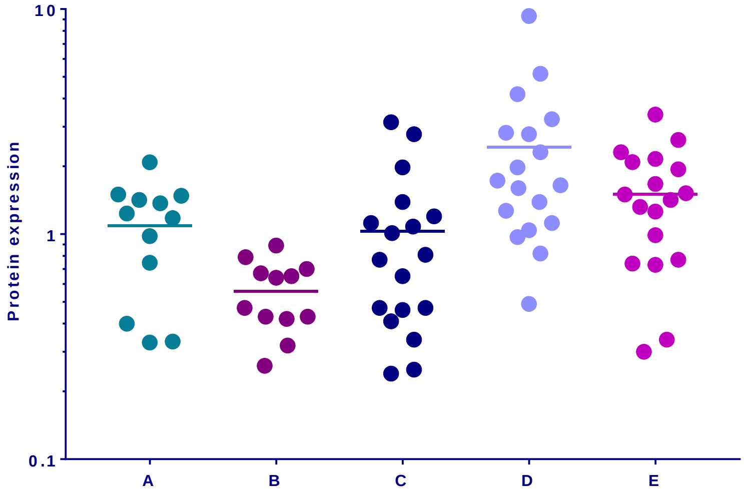
Scale: Linear

☑ Automatically determine the range and interval

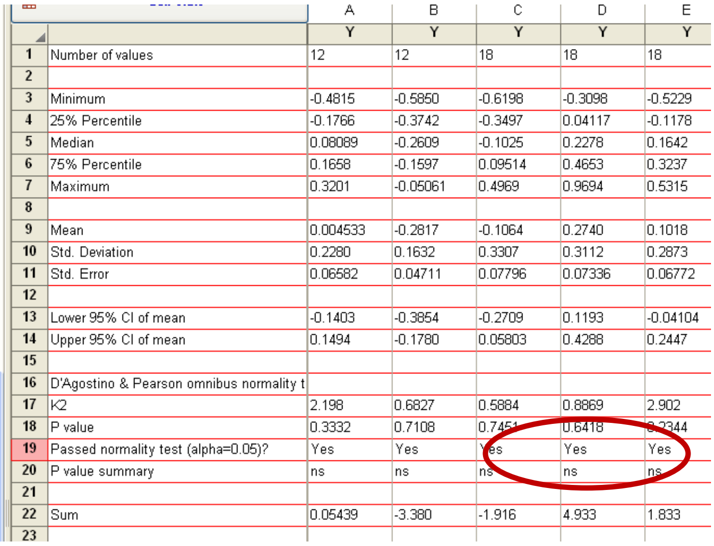
Range

Minimum: 0

Maximum:



It looks much better. Now, the next step is to actually log-transform the data. To do so, we go to = **Analyse> Transform > Transform and we choose Y=Log(Y)**, we can then re-run the analysis.

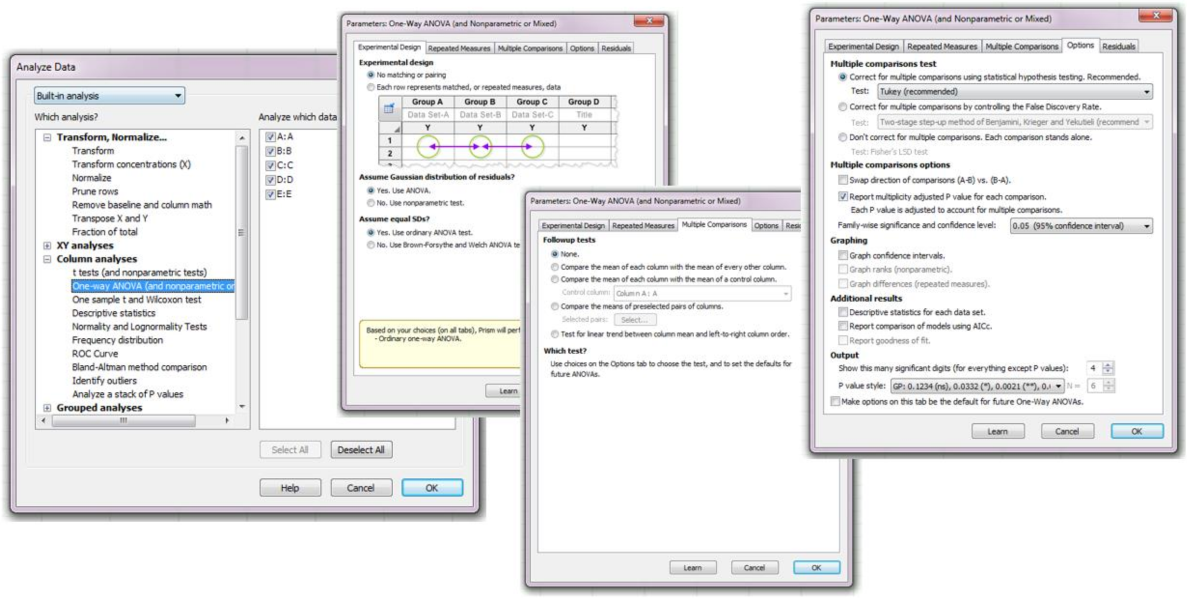


		A	B	C	D	E
		Y	Y	Y	Y	Y
1	Number of values	12	12	18	18	18
2						
3	Minimum	-0.4815	-0.5850	-0.6198	-0.3098	-0.5229
4	25% Percentile	-0.1766	-0.3742	-0.3497	0.04117	-0.1178
5	Median	0.08089	-0.2609	-0.1025	0.2278	0.1642
6	75% Percentile	0.1658	-0.1597	0.09514	0.4653	0.3237
7	Maximum	0.3201	-0.05061	0.4969	0.9694	0.5315
8						
9	Mean	0.004533	-0.2817	-0.1064	0.2740	0.1018
10	Std. Deviation	0.2280	0.1632	0.3307	0.3112	0.2873
11	Std. Error	0.06582	0.04711	0.07796	0.07336	0.06772
12						
13	Lower 95% CI of mean	-0.1403	-0.3854	-0.2709	0.1193	-0.04104
14	Upper 95% CI of mean	0.1494	-0.1780	0.05803	0.4288	0.2447
15						
16	D'Agostino & Pearson omnibus normality t					
17	K2	2.198	0.6827	0.5884	0.8869	2.902
18	P value	0.3332	0.7108	0.7451	0.6418	0.2344
19	Passed normality test (alpha=0.05)?	Yes	Yes	Yes	Yes	Yes
20	P value summary	ns	ns	ns	ns	ns
21						
22	Sum	0.05439	-3.380	-1.916	4.933	1.833
23						

OK, the situation is getting better: the first assumption is met and from what we see when we plot the transformed data (Box-plots and scatter plots below) the homogeneity of variance has improved a great deal.

Now that we have sorted out the data, we can run the ANOVA: to do so we go **=Analyze >One-way ANOVA**.

The next thing we need to do is to choose a post-hoc test. These post-hoc tests should only be used when the ANOVA finds a significant effect. GraphPad is not very powerful when it comes to post-hoc tests as it offers only 3 tests: the Bonferroni and Sidak tests which are quite conservative - so we should only choose them when we are comparing no more than 5 groups - and the Tukey which is more liberal.



Analyze Data: Which analysis? Column analyses – One-way ANOVA (and nonparametric or Mixed). Analyze which data: ☑A:A ☑B:B ☑C:C ☑D:D ☑E:E

Parameters: One-Way ANOVA (and Nonparametric or Mixed) – Experimental Design

Experimental design

● No matching or pairing

○ Each row represents matched, or repeated measures, data

Assume Gaussian distribution of residuals?

● Yes. Use ANOVA.

○ No. Use nonparametric test.

Assume equal SDs?

● Yes. Use ordinary ANOVA test.

○ No. Use Brown-Forsythe and Welch ANOVA te

Based on your choices (on all tabs), Prism will perf
- Ordinary one-way ANOVA.

Parameters: One-Way ANOVA (and Nonparametric or Mixed) – Multiple Comparisons

Followup tests

● None.

○ Compare the mean of each column with the mean of every other column.

○ Compare the mean of each column with the mean of a control column.

Control column: Column A : A

○ Compare the means of preselected pairs of columns.

Selected pairs: Select...

○ Test for linear trend between column mean and left-to-right column order.

Which test?

Use choices on the Options tab to choose the test, and to set the defaults for future ANOVAs.

Parameters: One-Way ANOVA (and Nonparametric or Mixed) – Options

Multiple comparisons test

● Correct for multiple comparisons using statistical hypothesis testing. Recommended.

Test: Tukey (recommended)

○ Correct for multiple comparisons by controlling the False Discovery Rate.

Test: Two-stage step-up method of Benjamini, Krieger and Yekutieli (recommend

○ Don't correct for multiple comparisons. Each comparison stands alone.

Test: Fisher's LSD test

Multiple comparisons options

□ Swap direction of comparisons (A-B) vs. (B-A).

☑ Report multiplicity adjusted P value for each comparison.
Each P value is adjusted to account for multiple comparisons.

Family-wise significance and confidence level: 0.05 (95% confidence interval)

Graphing

□ Graph confidence intervals.

□ Graph ranks (nonparametric).

□ Graph differences (repeated measures).

Additional results

□ Descriptive statistics for each data set.

□ Report comparison of models using AICc.

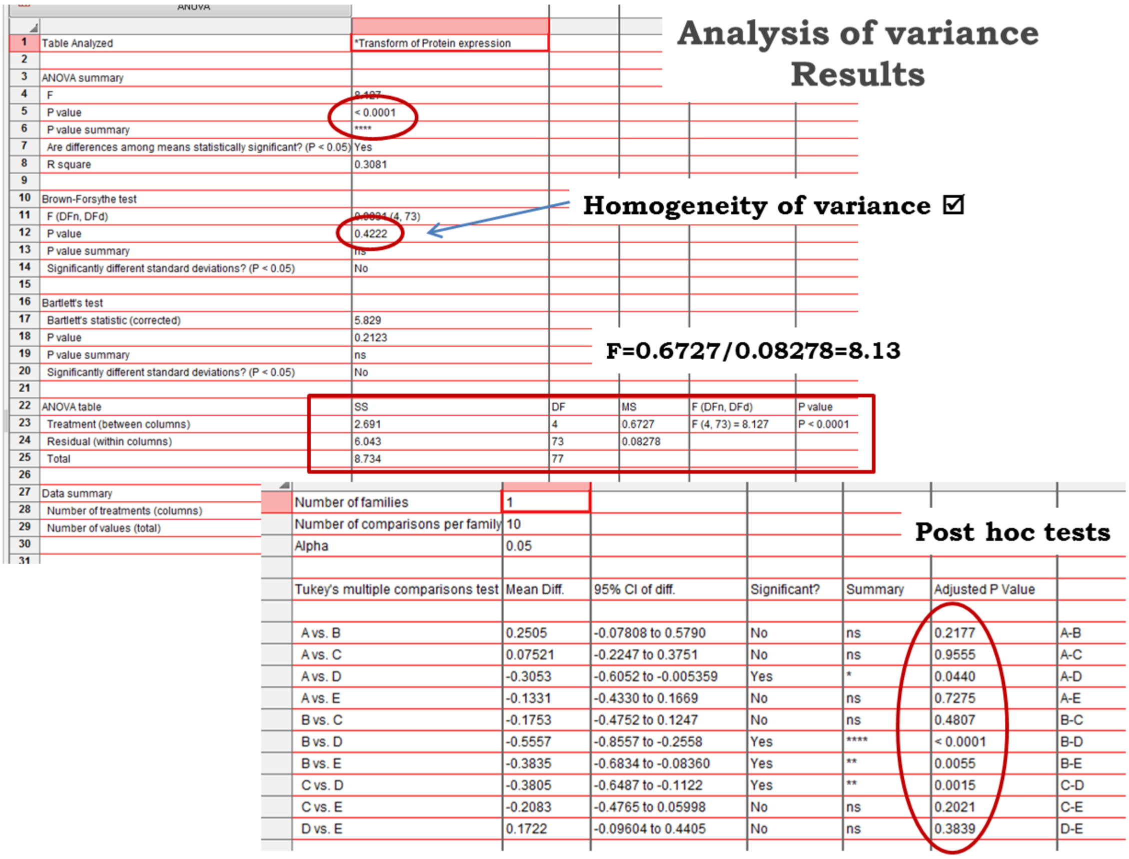
□ Report goodness of fit.

Output

Show this many significant digits (for everything except P values): 4

P value style: GP: 0.1234 (ns), 0.0332 (*), 0.0021 (**), 0.(N = 6

□ Make options on this tab be the default for future One-Way ANOVAs.



Analysis of variance Results

Homogeneity of variance ☑

F=0.6727/0.08278=8.13

Post hoc tests

1	Table Analyzed	*Transform of Protein expression				
2						
3	ANOVA summary					
4	F	8.127				
5	P value	< 0.0001				
6	P value summary	****				
7	Are differences among means statistically significant? (P < 0.05)	Yes				
8	R square	0.3081				
9						
10	Brown-Forsythe test					
11	F (DFn, DFd)	[illegible] (4, 73)				
12	P value	0.4222				
13	P value summary	ns				
14	Significantly different standard deviations? (P < 0.05)	No				
15						
16	Bartlett's test					
17	Bartlett's statistic (corrected)	5.829				
18	P value	0.2123				
19	P value summary	ns				
20	Significantly different standard deviations? (P < 0.05)	No				
21						
22	ANOVA table	SS	DF	MS	F (DFn, DFd)	P value
23	Treatment (between columns)	2.691	4	0.6727	F (4, 73) = 8.127	P < 0.0001
24	Residual (within columns)	6.043	73	0.08278		
25	Total	8.734	77			
26						
27	Data summary					
28	Number of treatments (columns)					
29	Number of values (total)					
30						
31						

Number of families	1					
Number of comparisons per family	10					
Alpha	0.05					
Tukey's multiple comparisons test	Mean Diff.	95% CI of diff.	Significant?	Summary	Adjusted P Value	
A vs. B	0.2505	-0.07808 to 0.5790	No	ns	0.2177	A-B
A vs. C	0.07521	-0.2247 to 0.3751	No	ns	0.9555	A-C
A vs. D	-0.3053	-0.6052 to -0.005359	Yes	*	0.0440	A-D
A vs. E	-0.1331	-0.4330 to 0.1669	No	ns	0.7275	A-E
B vs. C	-0.1753	-0.4752 to 0.1247	No	ns	0.4807	B-C
B vs. D	-0.5557	-0.8557 to -0.2558	Yes	****	< 0.0001	B-D
B vs. E	-0.3835	-0.6834 to -0.08360	Yes	**	0.0055	B-E
C vs. D	-0.3805	-0.6487 to -0.1122	Yes	**	0.0015	C-D
C vs. E	-0.2083	-0.4765 to 0.05998	No	ns	0.2021	C-E
D vs. E	0.1722	-0.09604 to 0.4405	No	ns	0.3839	D-E

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

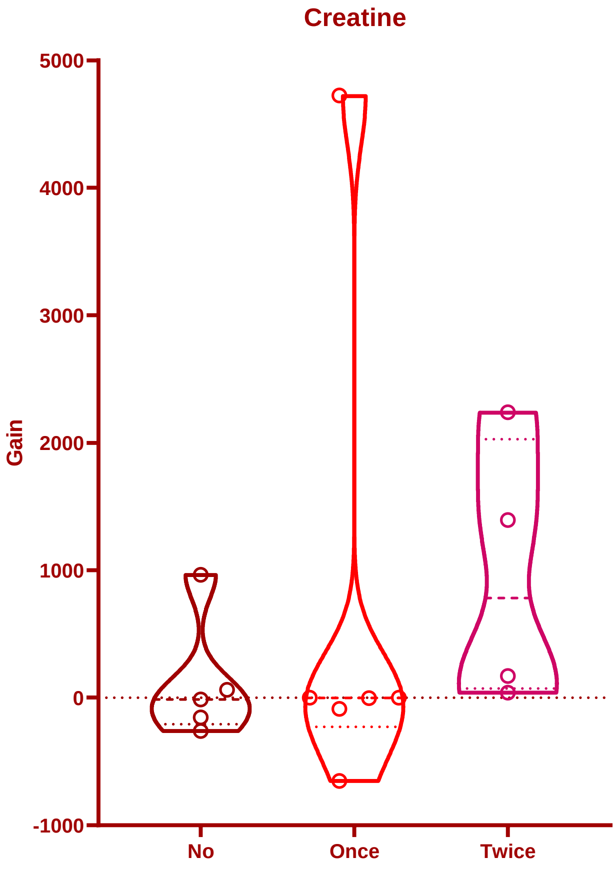
5-2 Non-parametric data: Kruskal-Wallis test

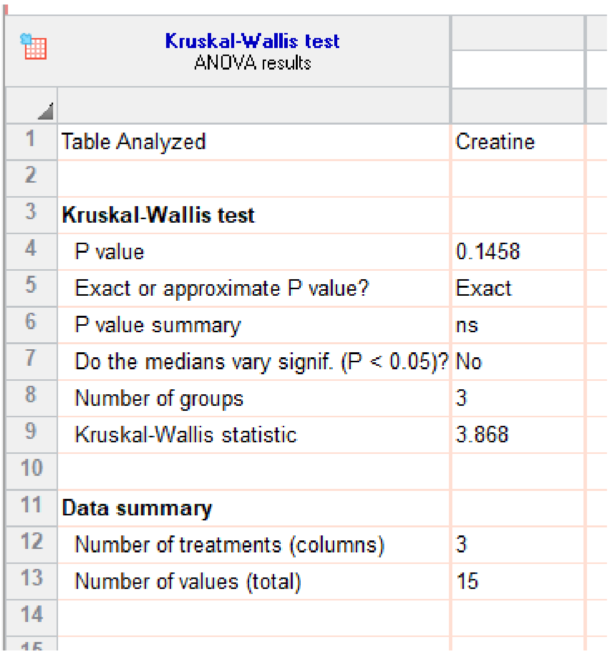
What if the data do not meet the assumptions for ANOVA? We can choose to run the non-parametric equivalent: the Kruskal-Wallis.

Example (File: `creatine.xlsx`)

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?





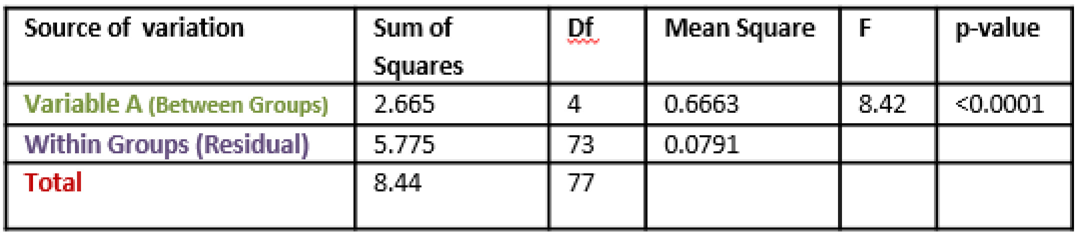
	Kruskal-Wallis test ANOVA results	
1	Table Analyzed	Creatine
2		
3	**Kruskal-Wallis test**	
4	P value	0.1458
5	Exact or approximate P value?	Exact
6	P value summary	ns
7	Do the medians vary signif. (P < 0.05)?	No
8	Number of groups	3
9	Kruskal-Wallis statistic	3.868
10		
11	**Data summary**	
12	Number of treatments (columns)	3
13	Number of values (total)	15
14		

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

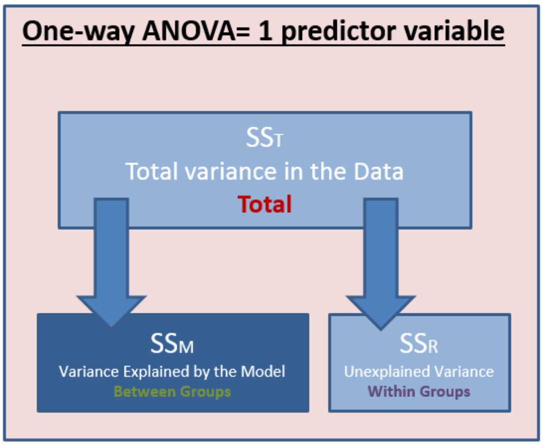
5-3 Two-way Analysis of Variance (File: `goggles.xlsx`)

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.

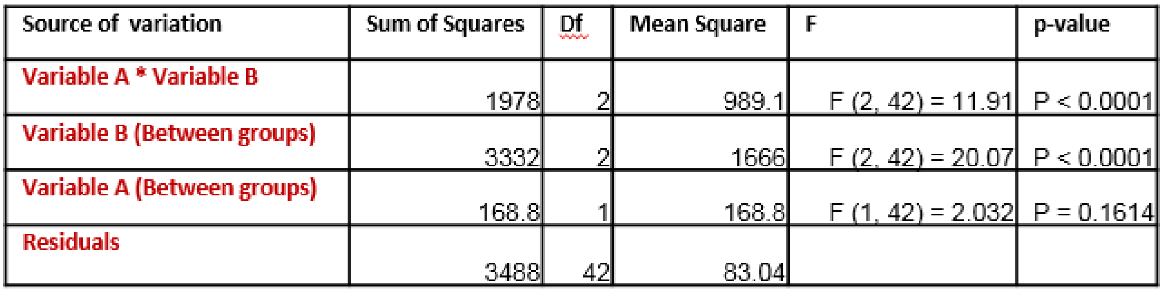
We saw in the previous chapter that running an ANOVA is all about partitioning the variance as seen below.



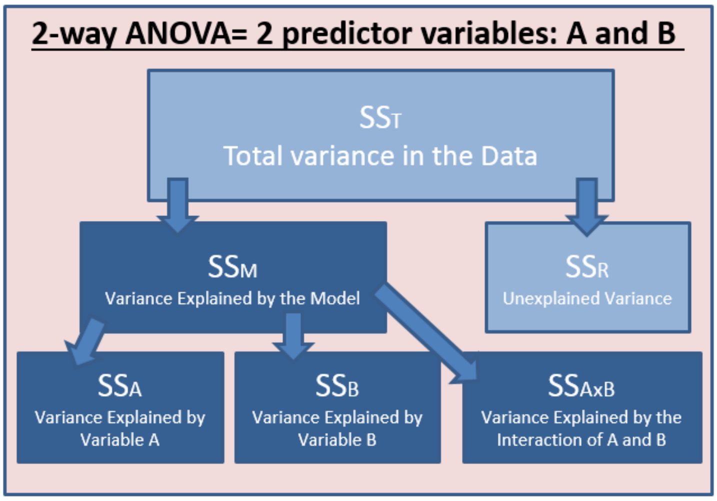
Source of variation	Sum of Squares	Df	Mean Square	F	p-value
Variable A (Between Groups)	2.665	4	0.6663	8.42	<0.0001
Within Groups (Residual)	5.775	73	0.0791		
Total	8.44	77			


One-way ANOVA= 1 predictor variable
SST
Total variance in the Data
Total
SSM
Variance Explained by the Model
Between Groups
SSR
Unexplained Variance
Within Groups

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



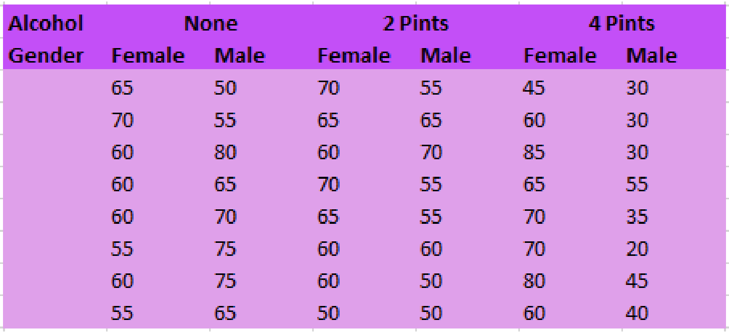
Source of variation	Sum of Squares	Df	Mean Square	F	p-value
Variable A * Variable B	1978	2	989.1	F (2, 42) = 11.91	P < 0.0001
Variable B (Between groups)	3332	2	1666	F (2, 42) = 20.07	P < 0.0001
Variable A (Between groups)	168.8	1	168.8	F (1, 42) = 2.032	P = 0.1614
Residuals	3488	42	83.04		


2-way ANOVA= 2 predictor variables: A and B
SST
Total variance in the Data
SSM
Variance Explained by the Model
SSR
Unexplained Variance
SSA
Variance Explained by Variable A
SSB
Variance Explained by Variable B
SSAxB
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.xlsx`)

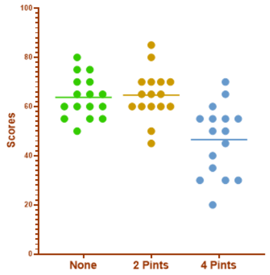
In the UK, there is something known as the beer-goggle effect: it is about subjective perception of physical attractiveness and how it become less accurate after alcohol is consumed. An anthropologist wanted to study the effects of alcohol, so in fact the beer-goggle effect, on mate selection at night-clubs. 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.



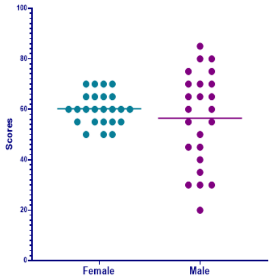
Alcohol	None		2 Pints		4 Pints	
Gender	Female	Male	Female	Male	Female	Male
	65	50	70	55	45	30
	70	55	65	65	60	30
	60	80	60	70	85	30
	60	65	70	55	65	55
	60	70	65	55	70	35
	55	75	60	60	70	20
	60	75	60	50	80	45
	55	65	50	50	60	40

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.

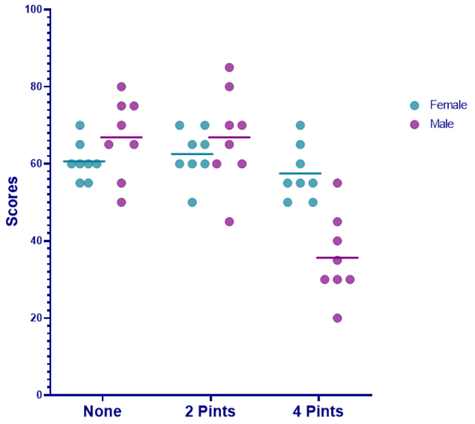
Main effect of Alcohol



Main effect of Gender



Interaction between Alcohol and Gender

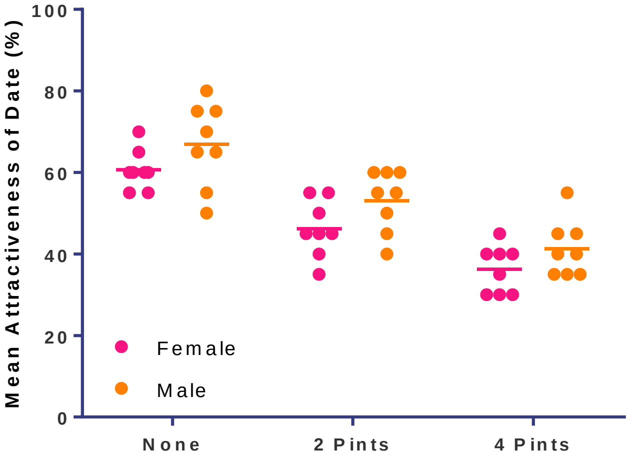


Main effect of alcohol:
One can see that there is not much happening between None and 2 Pints, but after 4 Pints the level of attractiveness drops quite a bit. We can also notice that there does not seem to be anything too worrying about the data, in terms of distribution or homogeneity of variance.

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 though.

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

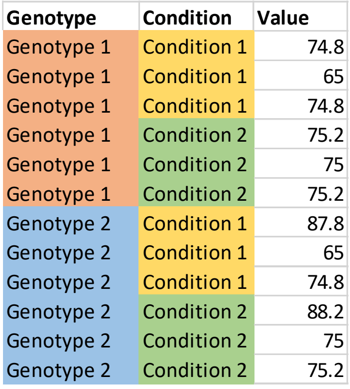
answer is: yes, but the effect is not the same for boys and girls. It is the 'yes but' that is about the interaction. Similarly, we could say: yes there is gender effect on the perception of physical attractiveness but the effect varies with the level of alcohol consumed.
Below is a graph, on fake data, 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, in a similar way. And to the question about gender, the answer is 'yes', too: levels of attractiveness are higher for males than females, regardless the level of alcohol. There is no interaction here and both factors are independent from one another.

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.

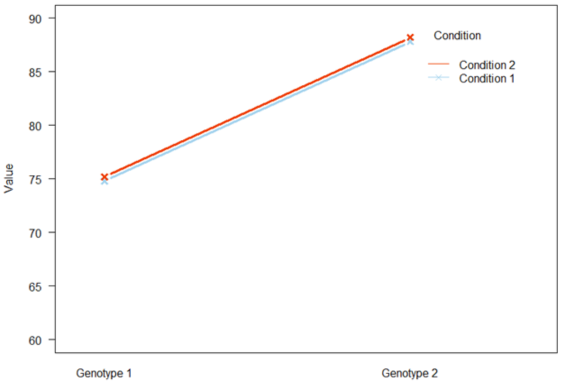
In our fake dataset, we have 2 factors: Genotype (2 levels) and Condition (2 levels).



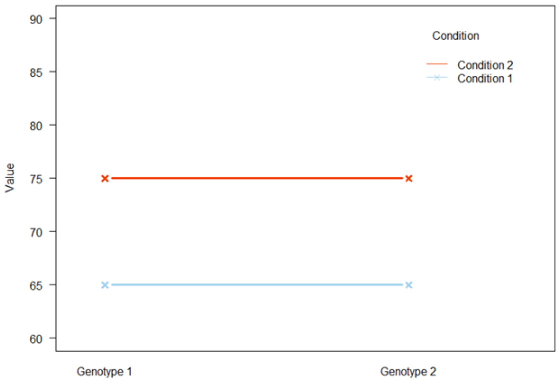
Genotype	Condition	Value
Genotype 1	Condition 1	74.8
Genotype 1	Condition 1	65
Genotype 1	Condition 1	74.8
Genotype 1	Condition 2	75.2
Genotype 1	Condition 2	75
Genotype 1	Condition 2	75.2
Genotype 2	Condition 1	87.8
Genotype 2	Condition 1	65
Genotype 2	Condition 1	74.8
Genotype 2	Condition 2	88.2
Genotype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



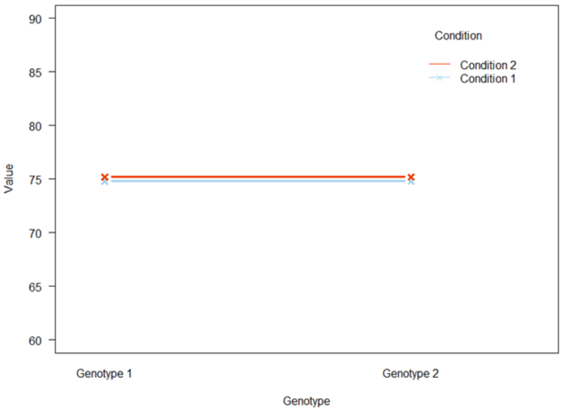
Genotype Effect



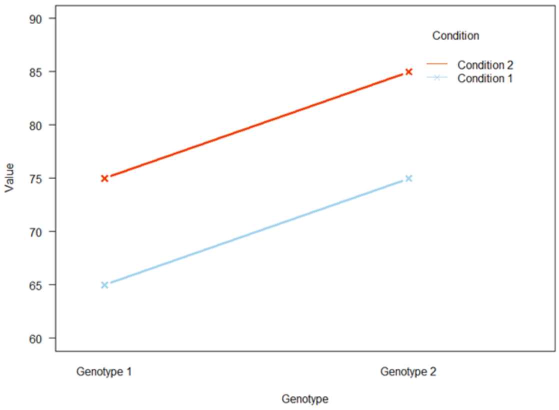
Condition Effect

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

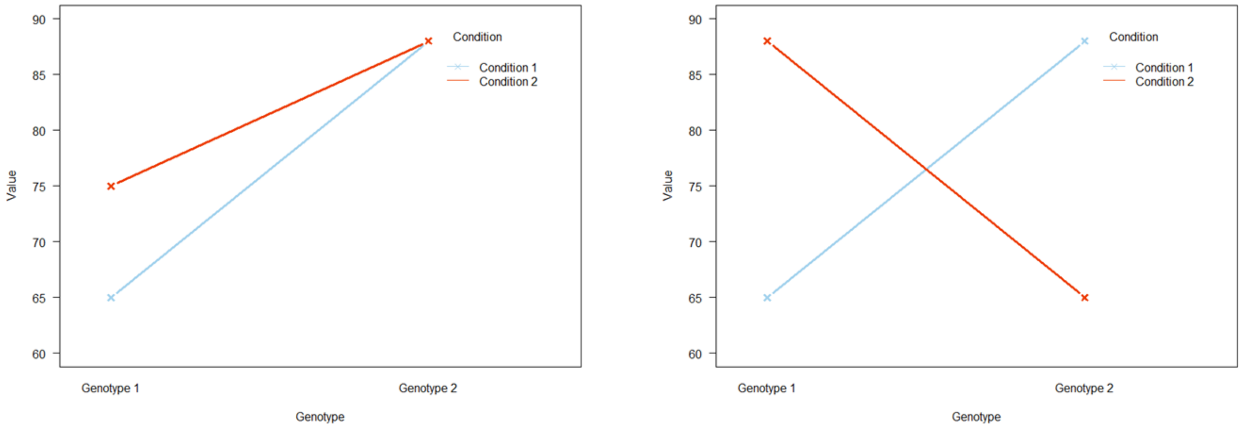


Both Effect

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



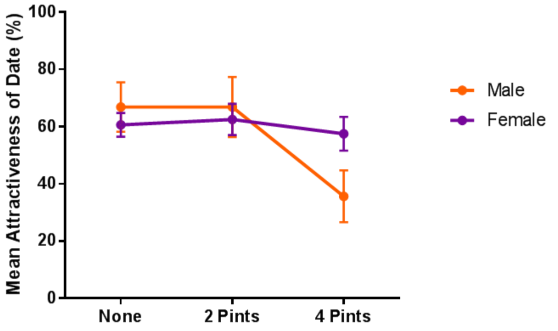
On the left, there is an effect of Condition for the first Genotype but not the second one, and there is an overall Genotype effect. On the right, there is a Condition effect but the direction is inversed from Genotype to the other. And there is not Genotype effect. Now if we try to answer the question ‘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).

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.

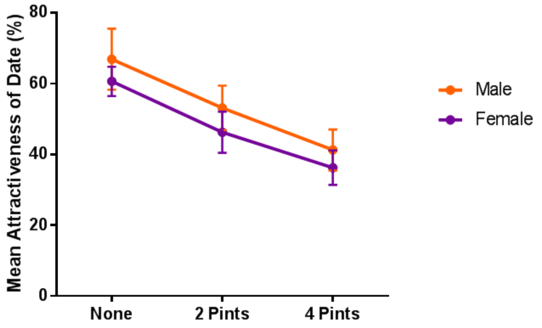
With significant interaction

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)

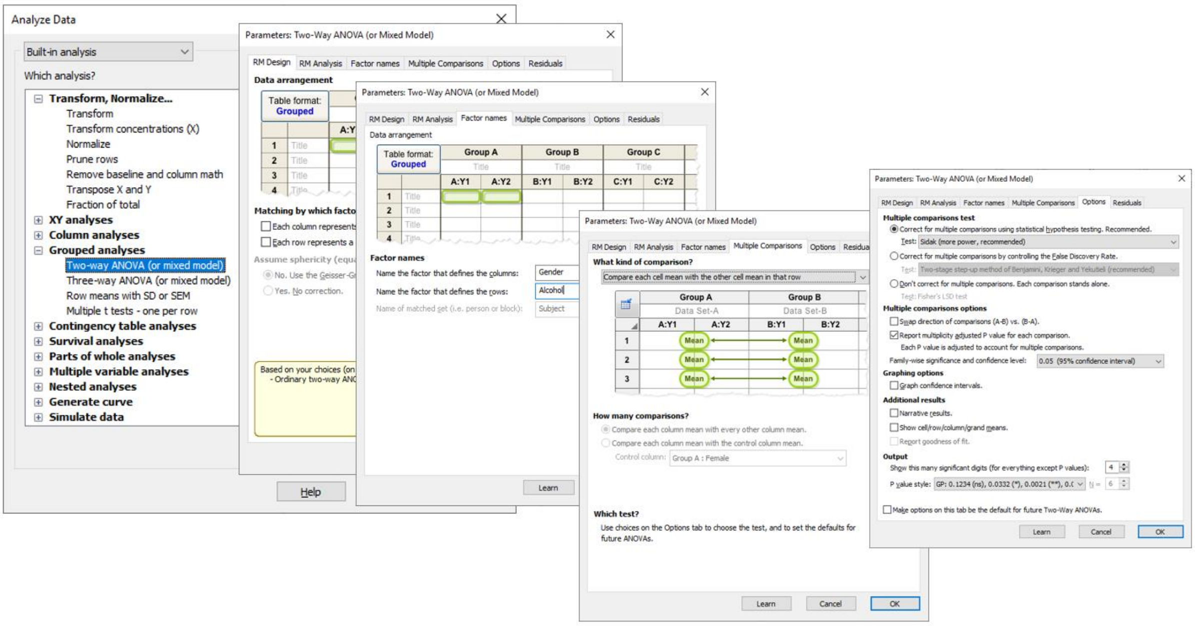
ANOVA table	SS	DF	MS	F (DFn, DFd)	P value
Interaction	7.292	2	3.646	F (2, 42) = 0.06872	0.9337
Alcohol Consumption	**5026**	**2**	**2513**	**F (2, 42) = 47.37**	**< 0.0001**
Gender	**438.0**	**1**	**438.0**	**F (1, 42) = 8.257**	**0.0063**
Residual	2228	42	53.05		



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 effects 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.



Analyze Data

Built-in analysis

Which analysis?

- ⊟ Transform, Normalize...
  - Transform
  - Transform concentrations (X)
  - Normalize
  - Prune rows
  - Remove baseline and column math
  - Transpose X and Y
  - Fraction of total
- ⊞ XY analyses
- ⊞ Column analyses
- ⊟ Grouped analyses
  - Two-way ANOVA (or mixed model)
  - Three-way ANOVA (or mixed model)
  - Row means with SD or SEM
  - Multiple t tests - one per row
- ⊞ Contingency table analyses
- ⊞ Survival analyses
- ⊞ Parts of whole analyses
- ⊞ Multiple variable analyses
- ⊞ Nested analyses
- ⊞ Generate curve
- ⊞ Simulate data

Parameters: Two-Way ANOVA (or Mixed Model)

Data arrangement

Table format: Grouped

Matching by which facto

□ Each column represent

□ Each row represents a

Assume sphericity (equ

● No. Use the Geisser-G

○ Yes. No correction.

Based on your choices (on
- Ordinary two-way ANC

Parameters: Two-Way ANOVA (or Mixed Model)

Factor names

Name the factor that defines the columns: Gender

Name the factor that defines the rows: Alcohol

Name of matched set (i.e. person or block): Subject

Parameters: Two-Way ANOVA (or Mixed Model)

What kind of comparison?

Compare each cell mean with the other cell mean in that row

How many comparisons?

● Compare each column mean with every other column mean.

○ Compare each column mean with the control column mean.

Control column: Group A : Female

Which test?

Use choices on the Options tab to choose the test, and to set the defaults for future ANOVAs.

Parameters: Two-Way ANOVA (or Mixed Model)

Multiple comparisons test

● Correct for multiple comparisons using statistical hypothesis testing. Recommended.

Test: Sidak (more power, recommended)

○ Correct for multiple comparisons by controlling the False Discovery Rate.

Test: Two-stage step-up method of Benjamini, Krieger and Yekutieli (recommended)

○ Don't correct for multiple comparisons. Each comparison stands alone.

Test: Fisher's LSD test

Multiple comparisons options

□ Swap direction of comparisons (A-B) vs. (B-A).

☑ Report multiplicity adjusted P value for each comparison.

Each P value is adjusted to account for multiple comparisons.

Family-wise significance and confidence level: 0.05 (95% confidence interval)

Graphing options

□ Graph confidence intervals.

Additional results

□ Narrative results.

□ Show cell/row/column/grand means.

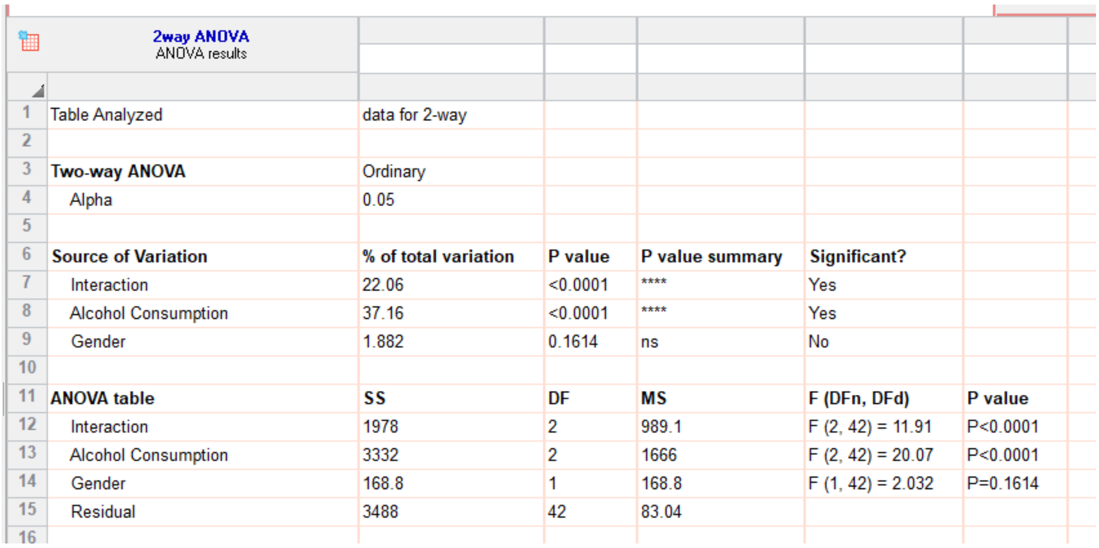
□ Report goodness of fit.

Output

Show this many significant digits (for everything except P values): 4

P value style: GP: 0.1234 (ns), 0.0332 (*), 0.0021 (**), 0.(N = 6

□ Make options on this tab be the default for future Two-Way ANOVAs.

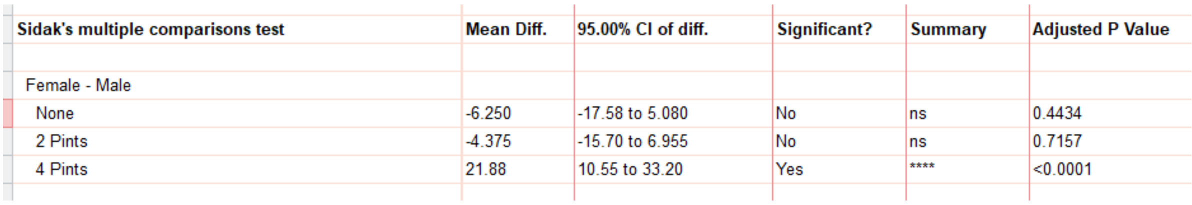


2way ANOVA
ANOVA results

1	Table Analyzed	data for 2-way				
2						
3	**Two-way ANOVA**	Ordinary				
4	Alpha	0.05				
5						
6	**Source of Variation**	**% of total variation**	**P value**	**P value summary**	**Significant?**	
7	Interaction	22.06	<0.0001	****	Yes	
8	Alcohol Consumption	37.16	<0.0001	****	Yes	
9	Gender	1.882	0.1614	ns	No	
10						
11	**ANOVA table**	**SS**	**DF**	**MS**	**F (DFn, DFd)**	**P value**
12	Interaction	1978	2	989.1	F (2, 42) = 11.91	P<0.0001
13	Alcohol Consumption	3332	2	1666	F (2, 42) = 20.07	P<0.0001
14	Gender	168.8	1	168.8	F (1, 42) = 2.032	P=0.1614
15	Residual	3488	42	83.04		
16						

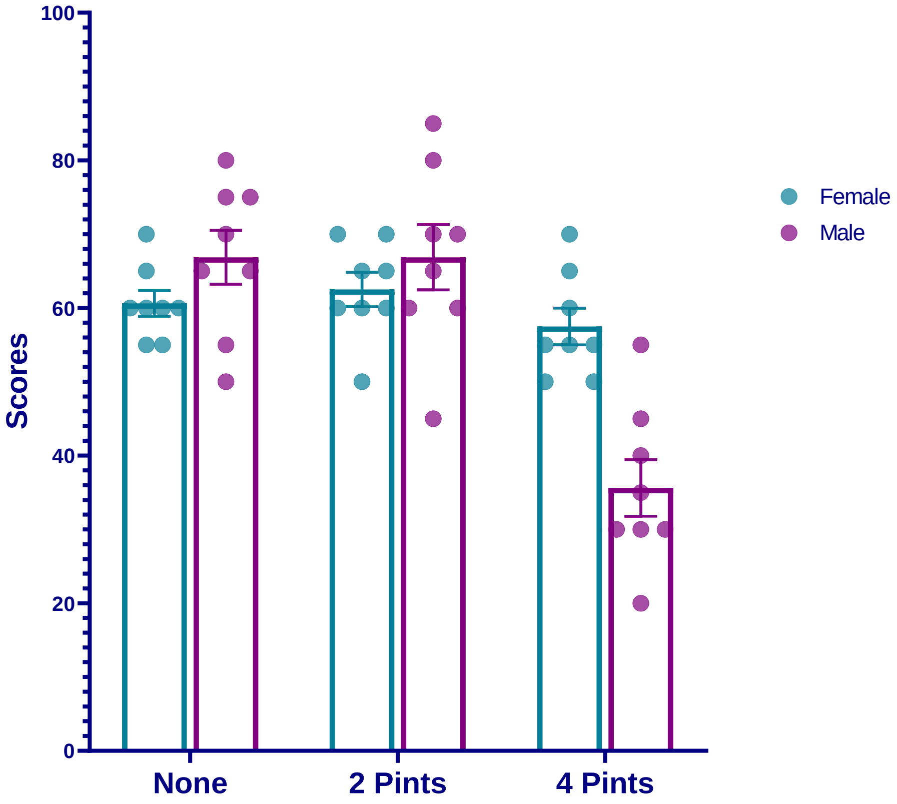
So, there is a significant interaction between the 2 factors which is consistent with what we observed. Now, in terms of interpretation, when an interaction between 2 factors is significant we don't look at the main effect. It is for the same reason as before: we cannot interpret the effect of one factor 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. Let's look at the Tukey post-hoc tests.



Sidak's multiple comparisons test	Mean Diff.	95.00% CI of diff.	Significant?	Summary	Adjusted P Value
Female - Male					
None	-6.250	-17.58 to 5.080	No	ns	0.4434
2 Pints	-4.375	-15.70 to 6.955	No	ns	0.7157
4 Pints	21.88	10.55 to 33.20	Yes	****	<0.0001

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<0.0001).



5-4 Non-parametric data

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.

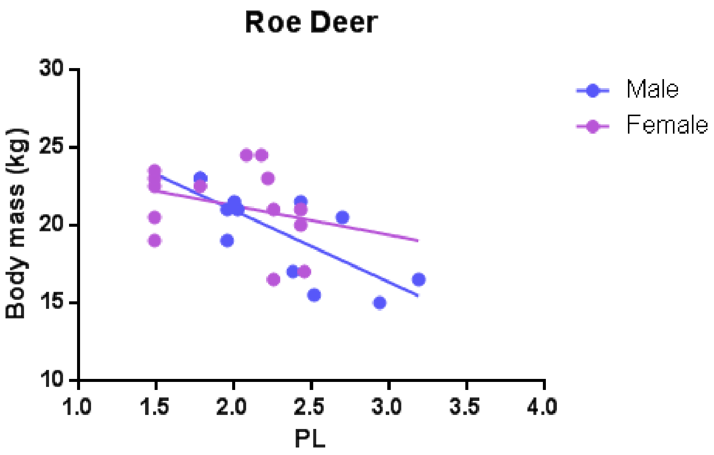
Chapter 6: Correlation

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

Example (File: `roe deer.xlsx`).

When we want to plot data from 2 quantitative variables between which we suspect that there is a relationship, the best choice to have a first look at our data is the scatter plot. So, in GraphPad, we go > choose an XY table.
We have to choose between the x- and the y-axis for our 2 variables. It is usually considered that “x” predicts “y” (y=f(x)) so when looking at the relationship between 2 variables, we must have an idea of which one is likely to predict the other one.

In our particular case, we want to know how an increase in parasite load (PL) affects the body mass (BM) of the host.



By looking at the graph, one can think that something is happening here. Now, if we want to know if the relationship between our 2 variables is significant, we need to run a correlation test.

6-1: Pearson coefficient

A bit of theory

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.

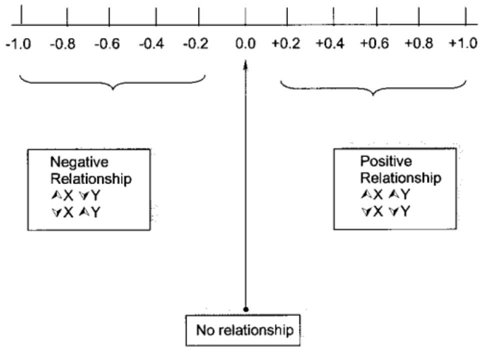
$$\operatorname{cov}(X,Y) = \sum_{i=1}^{N} \frac{(x_i - \bar{x})(y_i - \bar{y})}{N}.$$

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 won’t 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”.

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

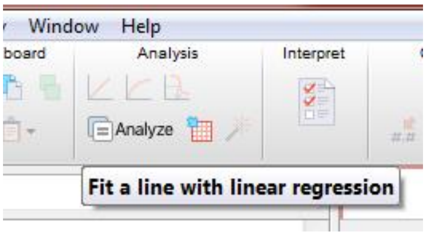
Of course, we don’t need to remember that formula but it is important that we understand what the correlation coefficient does: it measures the magnitude and the direction of the relationship between two variables. It is designed to range in value between 0.0 and 1.0.



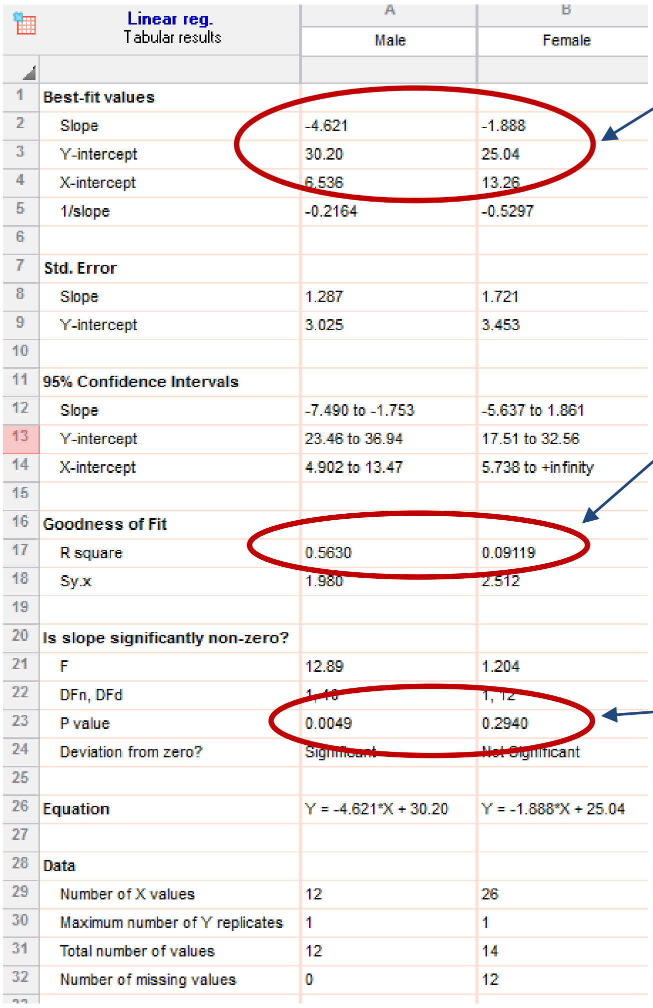
The 2 variables do not have to be measured in the same units but they have to be proportional (meaning linearly related).

One last thing before we go back to our example: the coefficient of determination r^2: it gives us the proportion of variance in Y that can be explained by X, as a percentage.

One way to run a correlation with GraphPad is simply to click on the little icon that represents a regression line in the Analysis window but before that, don’t forget that we need to check the normality of our data. In our case, we are good: D’Agostino and Pearson tests: males: p=0.3083 and females: p=0.5084).



If we look into the results section, we will find that there is a strong negative relationship (for the males) and a weak one (for the females) between the 2 variables, the body mass decreasing when the parasite burden increases (negative slopes).



	Linear reg. Tabular results	A Male	B Female
1	Best-fit values		
2	Slope	-4.621	-1.888
3	Y-intercept	30.20	25.04
4	X-intercept	6.536	13.26
5	1/slope	-0.2164	-0.5297
6			
7	Std. Error		
8	Slope	1.287	1.721
9	Y-intercept	3.025	3.453
10			
11	95% Confidence Intervals		
12	Slope	-7.490 to -1.753	-5.637 to 1.861
13	Y-intercept	23.46 to 36.94	17.51 to 32.56
14	X-intercept	4.902 to 13.47	5.738 to +infinity
15			
16	Goodness of Fit		
17	R square	0.5630	0.09119
18	Sy.x	1.980	2.512
19			
20	Is slope significantly non-zero?		
21	F	12.89	1.204
22	DFn, DFd	1, 10	1, 12
23	P value	0.0049	0.2940
24	Deviation from zero?	Significant	Not Significant
25			
26	Equation	Y = -4.621*X + 30.20	Y = -1.888*X + 25.04
27			
28	Data		
29	Number of X values	12	26
30	Maximum number of Y replicates	1	1
31	Total number of values	12	14
32	Number of missing values	0	12

*For the males the equation would be: Body Mass = 30.2 - 4.621*Parasite Burden. It tells us that each time the parasite burden increases by 1 unit, the body mass decreases by 4.621 units and that the average male roe deer in that sample weights 30.2 kg.*

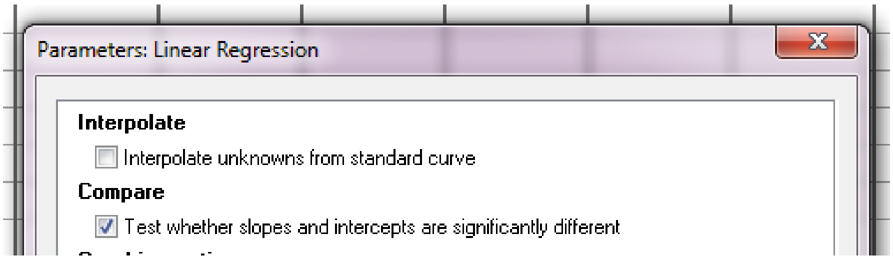
A coefficient of determination $r^2 = 0.56$ means that 56% of the variability observed in the body mass can be explained only by the parasite burden.

The relationship between body mass and parasite burden is significant for males (p=0.0049) but not for females (p=0.2940).

We may want to test whether there is a significant difference in the strength of the correlation between males and females. Some packages, like SPSS, allow us to run an ANCOVA which is a cross between the correlation and the ANOVA. It tests together the difference in body mass between males and females, the strength of the relationship between the body mass and the parasite burden and finally the ‘interaction’ between parasite burden and gender i.e. the difference in the relationship between body mass and parasite burden. We cannot run this analysis with GraphPad Prism.

However, we can test whether the 2 slopes are significantly different.

When we click on the regression line, we can choose to compare the slopes and the intercepts.



Parameters: Linear Regression

Interpolate

☐ Interpolate unknowns from standard curve

Compare

☑ Test whether slopes and intercepts are significantly different

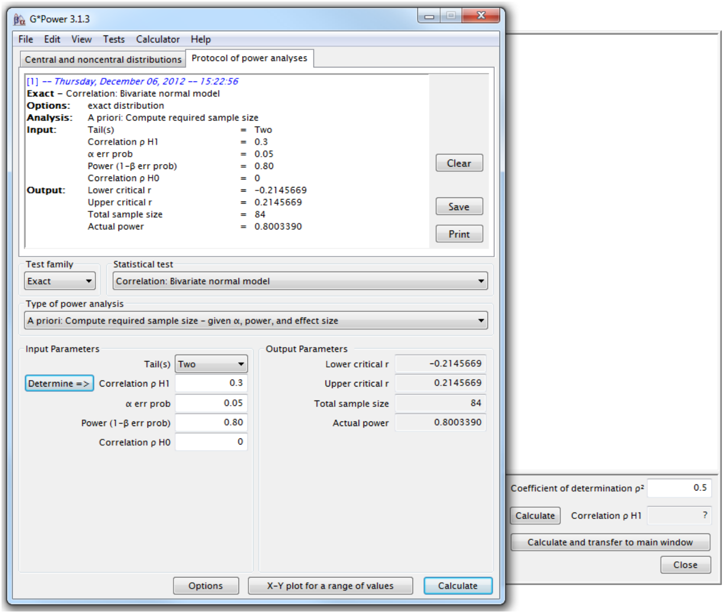
Are the slopes equal?

F = 1.60371. DFn=1 DFd=22
P=0.2186

A key thing to remember when working with correlations is never to assume a correlation means that a change in one variable *causes* a change in another. Sales of personal computers and athletic shoes have both risen strongly in the last several years and there is a high correlation between them, but we cannot assume that buying computers causes people to buy athletic shoes (or vice versa).

Power analysis with correlation

The data we analysed were actually from a pilot study. If we assume that the correlation observed in the female group is a good representation, then to reach significance we would need: n=84.


G*Power 3.1.3
File
Edit
View
Tests
Calculator
Help
Central and noncentral distributions
Protocol of power analyses
Clear
Save
Print
Determine =>
Options
X-Y plot for a range of values
Calculate
Calculate
Calculate and transfer to main window
Close

```
[1] -- Thursday, December 06, 2012 -- 15:22:56
Exact - Correlation: Bivariate normal model
Options:    exact distribution
Analysis:   A priori: Compute required sample size
Input:      Tail(s)                  =  Two
            Correlation ρ H1         =  0.3
            α err prob               =  0.05
            Power (1-β err prob)     =  0.80
            Correlation ρ H0         =  0
Output:     Lower critical r         =  -0.2145669
            Upper critical r         =  0.2145669
            Total sample size        =  84
            Actual power             =  0.8003390
```

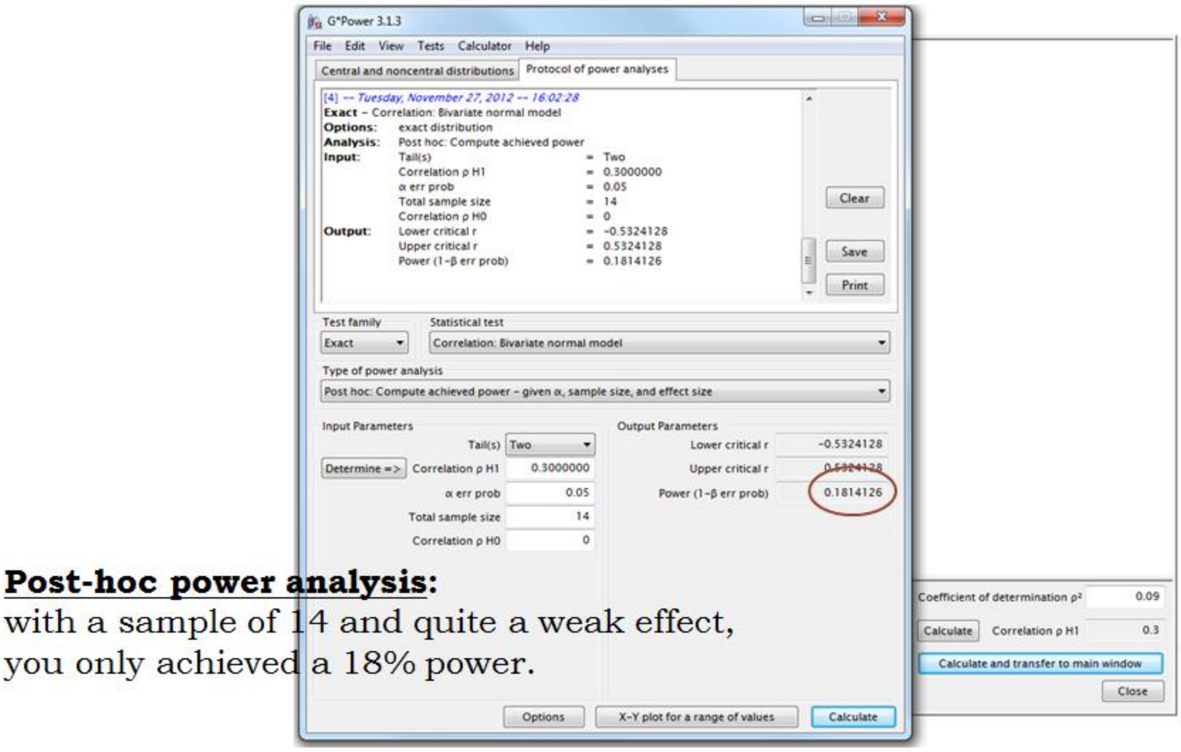
Test family: Exact
Statistical test: Correlation: Bivariate normal model
Type of power analysis: A priori: Compute required sample size - given α, power, and effect size

Input Parameters
Tail(s): Two
Correlation ρ H1: 0.3
α err prob: 0.05
Power (1-β err prob): 0.80
Correlation ρ H0: 0

Output Parameters
Lower critical r: -0.2145669
Upper critical r: 0.2145669
Total sample size: 84
Actual power: 0.8003390

Coefficient of determination ρ²: 0.5
Correlation ρ H1: ?

We may also be interested in the level of power we have achieved with our sample.


G*Power 3.1.3
File Edit View Tests Calculator Help
Central and noncentral distributions
Protocol of power analyses
Clear
Save
Print
Determine =>
Options
X-Y plot for a range of values
Calculate
Calculate
Calculate and transfer to main window
Close

```
[4] -- Tuesday, November 27, 2012 -- 16:02:28
Exact - Correlation: Bivariate normal model
Options:    exact distribution
Analysis:   Post hoc: Compute achieved power
Input:      Tail(s)                  = Two
            Correlation ρ H1         = 0.3000000
            α err prob               = 0.05
            Total sample size        = 14
            Correlation ρ H0         = 0
Output:     Lower critical r         = -0.5324128
            Upper critical r         = 0.5324128
            Power (1-β err prob)     = 0.1814126
```

Test family: Exact
Statistical test: Correlation: Bivariate normal model
Type of power analysis: Post hoc: Compute achieved power – given α, sample size, and effect size

Input Parameters
Tail(s): Two
Correlation ρ H1: 0.3000000
α err prob: 0.05
Total sample size: 14
Correlation ρ H0: 0

Output Parameters
Lower critical r: -0.5324128
Upper critical r: 0.5324128
Power (1-β err prob): 0.1814126

Coefficient of determination ρ²: 0.09
Correlation ρ H1: 0.3

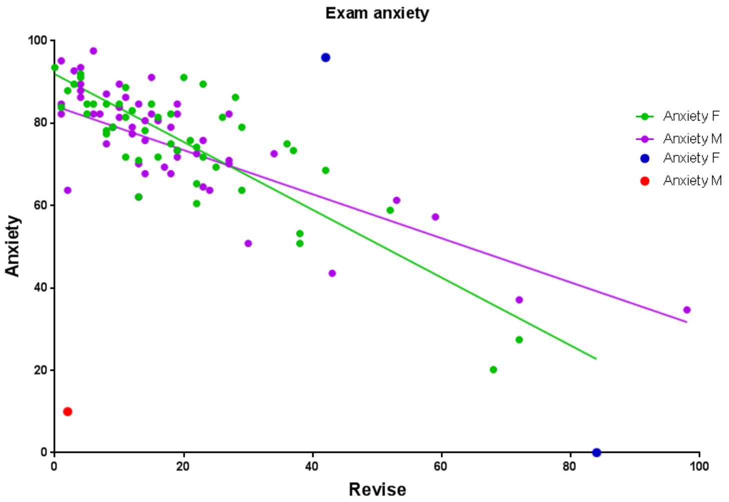
Post-hoc power analysis:
with a sample of 14 and quite a weak effect,
you only achieved a 18% power.

6-2 Linear correlation: goodness-of-fit of the model

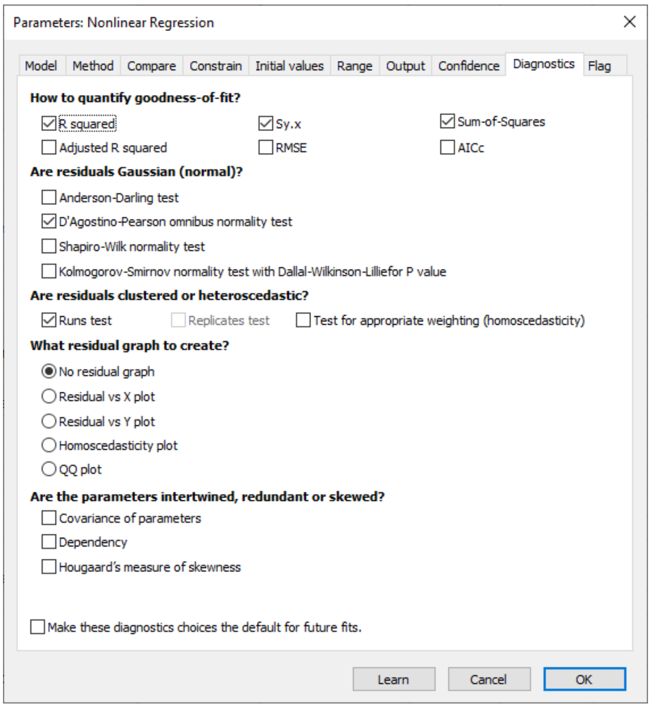
Now, R^2 helps with understanding how good the model is but what happens when there are values which really misbehave.
We can identify them with Prism using **XY analyses>Nonlinear regression**. This allows us to access the diagnostic tools we need, which are not accessible through Correlation or Linear regression. Even though it is the non-linear section, one of the possible models is Line, which is exactly what we want.

Example (File: `exam anxiety.xlsx`).

In this dataset, we try to quantify the realtionship between time spent revising and exam anxiety. We also want to know is there is a difference between boys and girls in that matter.
Visually, we can identify at least 2 values which look very much like outliers.



From the nonlinear section, we go into Diagnostics.



How to quantify goodness-of-fit?

☑ R squared ☑ Sy.x ☑ Sum-of-Squares

□ Adjusted R squared □ RMSE □ AICc

Are residuals Gaussian (normal)?

□ Anderson-Darling test

☑ D'Agostino-Pearson omnibus normality test

□ Shapiro-Wilk normality test

□ Kolmogorov-Smirnov normality test with Dallal-Wilkinson-Lilliefor P value

Are residuals clustered or heteroscedastic?

☑ Runs test □ Replicates test □ Test for appropriate weighting (homoscedasticity)

What residual graph to create?

● No residual graph

○ Residual vs X plot

○ Residual vs Y plot

○ Homoscedasticity plot

○ QQ plot

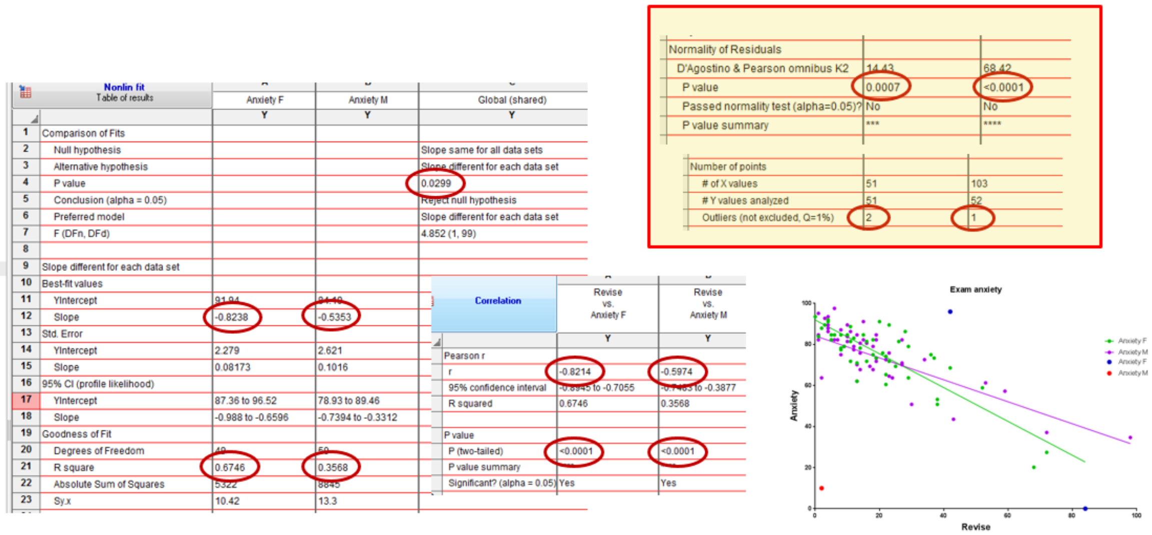
Are the parameters intertwined, redundant or skewed?

□ Covariance of parameters

□ Dependency

□ Hougaard's measure of skewness

□ Make these diagnostics choices the default for future fits.



Nonlin fit Table of results		Anxiety F	Anxiety M	Global (shared)
		Y	Y	Y
1	Comparison of Fits			
2	Null hypothesis			Slope same for all data sets
3	Alternative hypothesis			Slope different for each data set
4	P value			0.0299
5	Conclusion (alpha = 0.05)			Reject null hypothesis
6	Preferred model			Slope different for each data set
7	F (DFn, DFd)			4.852 (1, 99)
8				
9	Slope different for each data set			
10	Best-fit values			
11	YIntercept	91.94	84.19	
12	Slope	-0.8238	-0.5353	
13	Std. Error			
14	YIntercept	2.279	2.621	
15	Slope	0.08173	0.1016	
16	95% CI (profile likelihood)			
17	YIntercept	87.36 to 96.52	78.93 to 89.46	
18	Slope	-0.988 to -0.6596	-0.7394 to -0.3312	
19	Goodness of Fit			
20	Degrees of Freedom	49	50	
21	R square	0.6746	0.3568	
22	Absolute Sum of Squares	5322	8845	
23	Sy.x	10.42	13.3	

Normality of Residuals		
D'Agostino & Pearson omnibus K2	14.43	68.42
P value	0.0007	<0.0001
Passed normality test (alpha=0.05)?	No	No
P value summary	***	****

Number of points		
# of X values	51	103
# Y values analyzed	51	52
Outliers (not excluded, Q=1%)	2	1

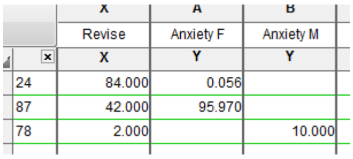
Correlation	Revise vs. Anxiety F	Revise vs. Anxiety M
	Y	Y
Pearson r		
r	-0.8214	-0.5974
95% confidence interval	-0.8945 to -0.7055	-0.7463 to -0.3877
R squared	0.6746	0.3568
P value		
P (two-tailed)	<0.0001	<0.0001
P value summary	****	***
Significant? (alpha = 0.05)	Yes	Yes

Prism tells us several things here:

- there is a negative relationship between the 2 variables, and this relationship is stronger in girls than in boys,
- the R^2 is way smaller for boys than for girls, which seems a bit suspicious looking at the graph,
- the residuals are not normal, which is bad,
- and there are 3 outliers.

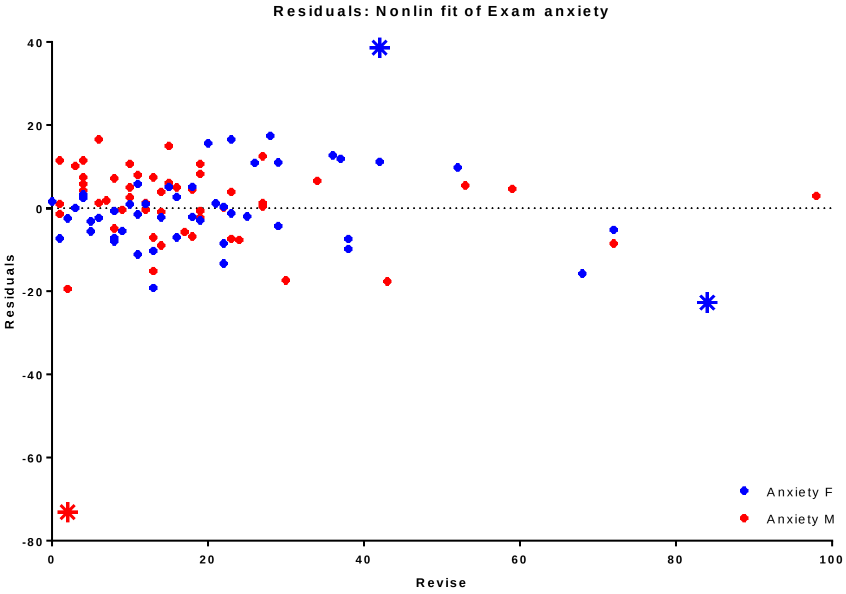
Let's talk about residuals for a minute. That word, residual, literally means what is left. In our case, the lines of best fit are supposed to summarise, as in to model, the relationship between our data. We can also say that it is supposed to predict, for example, the level of anxiety from a given time spent revising. But of course, like in any model, the prediction comes with a measure of error and the individual error associated with a particular prediction is the residual. This is the same error as the one we came across when we went through the Sum of Squared Errors and the Standard Deviation.
So the residual is what is left after the model has been fitted.
Now, the idea is that these residuals should be normally distributed, which will reflect the good fit of the model, with residuals symmetrically distributed on either side of the lines of best-fit.
So these residuals are used in all sorts of ways to check the assumptions.

Prism tells us which the 3 misbehaving students are.

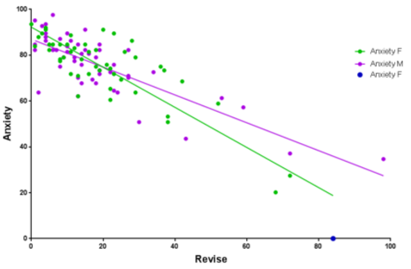


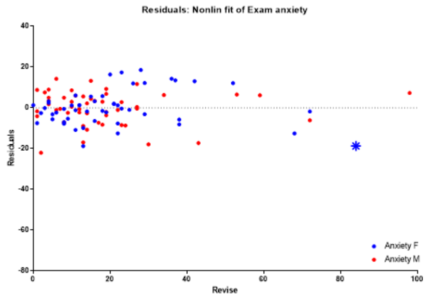
	X	A	B
	Revise	Anxiety F	Anxiety M
☒	X	Y	Y
24	84.000	0.056	
87	42.000	95.970	
78	2.000		10.000

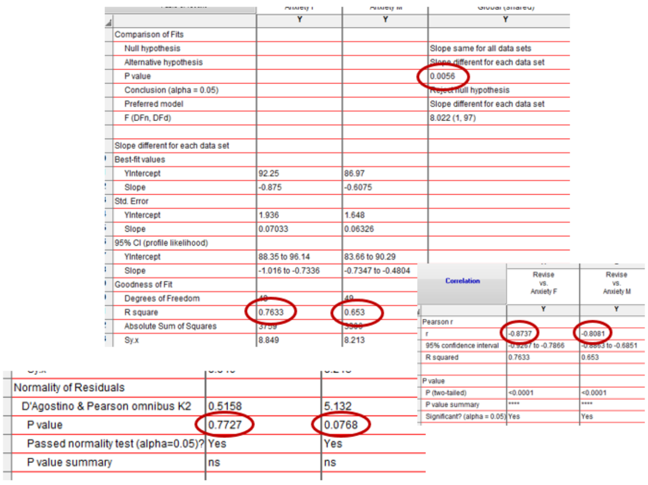
And looking at the residual plot, we can easily spot them. Most of the residuals appear nicely and consistently spread, apart from these 3 values.



The next step is to remove the outliers. We can remove them all or perhaps, like here, keep student 24 as she seems to ‘belong’ more than the others.







	Anxiety F	Anxiety M	Global (shared)
	Y	Y	Y
Comparison of Fits			
Null hypothesis			Slope same for all data sets
Alternative hypothesis			Slope different for each data set
P value			0.0056
Conclusion (alpha = 0.05)			Reject null hypothesis
Preferred model			Slope different for each data set
F (DFn, DFd)			8.022 (1, 97)
Slope different for each data set			
Best-fit values			
YIntercept	92.25	86.97	
Slope	-0.875	-0.6075	
Std. Error			
YIntercept	1.936	1.648	
Slope	0.07033	0.06326	
95% CI (profile likelihood)			
YIntercept	88.35 to 96.14	83.66 to 90.29	
Slope	-1.016 to -0.7336	-0.7347 to -0.4804	
Goodness of Fit			
Degrees of Freedom	48	49	
R square	0.7633	0.653	
Absolute Sum of Squares	3759	[illegible]	
Sy.x	8.849	8.213	

Normality of Residuals		
D'Agostino & Pearson omnibus K2	0.5158	5.132
P value	0.7727	0.0768
Passed normality test (alpha=0.05)?	Yes	Yes
P value summary	ns	ns

Correlation	Revise vs. Anxiety F	Revise vs. Anxiety M
	Y	Y
Pearson r		
r	-0.8737	-0.8081
95% confidence interval	-0.9267 to -0.7866	-0.8863 to -0.6851
R squared	0.7633	0.653
P value		
P (two-tailed)	<0.0001	<0.0001
P value summary	****	****
Significant? (alpha = 0.05)	Yes	Yes

We can see now that the R^2 for the boys is way higher and there is no more departure from normality for the residuals.

There is a strong and significant relationship between revising time and anxiety and that relationship is significantly different between boys and girls.

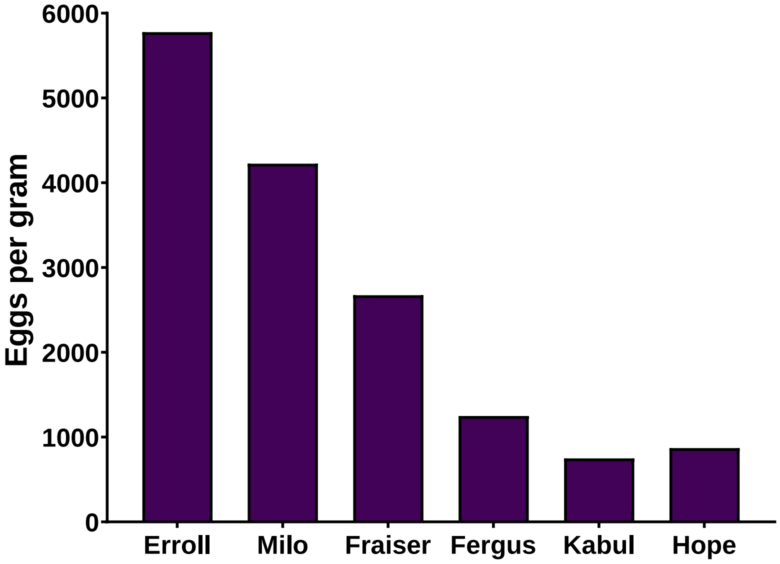
6-2 Non-parametric data: Spearman correlation coefficient

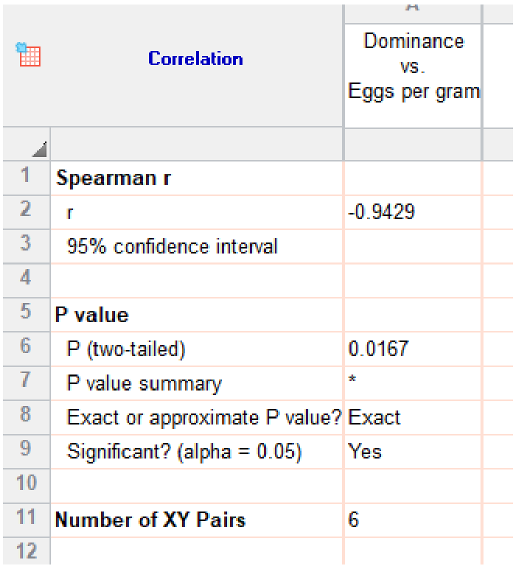
The truth is that Pearson coefficient behaves pretty well even if data are not normal. Spearman is used mostly for ranked data. Spearman's correlation determines the strength and direction of the monotonic relationship between your two variables rather than the strength and direction of the linear relationship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

Example (File: `dominance.xlsx`)

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.





	Correlation	A Dominance vs. Eggs per gram
1	**Spearman r**	
2	r	-0.9429
3	95% confidence interval	
4		
5	**P value**	
6	P (two-tailed)	0.0167
7	P value summary	*
8	Exact or approximate P value?	Exact
9	Significant? (alpha = 0.05)	Yes
10		
11	**Number of XY Pairs**	6
12		

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 7: Curve fitting: Dose-response

Dose-response curves can be used to plot the results of many kinds of experiments. The X axis plots concentration of a drug or hormone. The Y axis plots response, which could be pretty much any measure of biological function.

The term “dose” is often used loosely. In its strictest sense, the term only applies to experiments performed with animals or people, where we administer various doses of drug. We don't know the actual concentration of drug at its site of action—we only know the total dose that was administered.

However, the term “dose-response curve” is also used more loosely to describe *in vitro* experiments where we apply known concentrations of drugs. The term “concentration-response curve” is a more precise label for the results of these types of experiments.

Dose-response experiments typically use around 5-10 doses of agonist, equally spaced on a logarithmic scale. For example, doses might be 1, 3, 10, 30, 100, 300, 1000, 3000, and 10000 nM. When converted to logarithms (and rounded a bit), these values are equally spaced: 0.0, 0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 3.5, and 4.0.

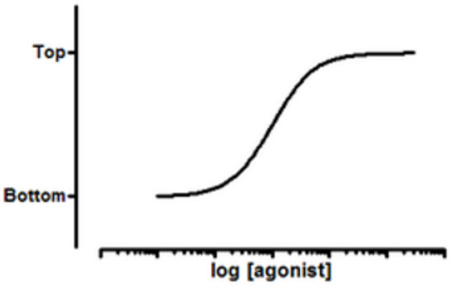
In a dose-response curve, the Y values are responses. For example, the response might be enzyme activity, accumulation of an intracellular second messenger, membrane potential, secretion of a hormone, change in heart rate, or contraction of a muscle.

IC50 or EC50

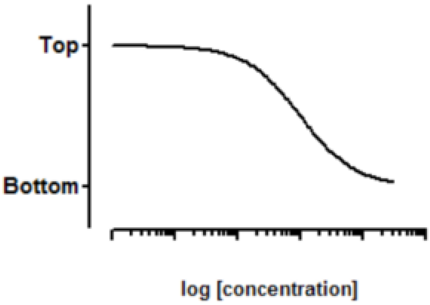
The agonist can be an inhibitor or a stimulator. The higher the concentration of the agonist, the stronger the response.
IC50 (I=Inhibition): concentration of an agonist that provokes a response half way between the maximal (Top) response and the maximally inhibited (Bottom) response.
EC50 (E=Effective): concentration that gives half-maximal response

This is purely a difference in which abbreviation is used, with no fundamental difference.

Many log(inhibitor) vs. response curves follow the familiar symmetrical sigmoidal shape. The goal is to determine the IC50/EC50 of the agonist.



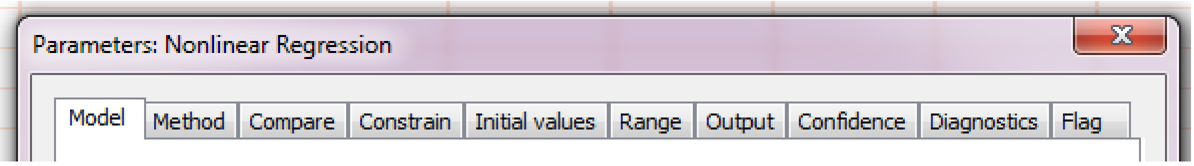
Model: Y=Bottom + (Top-Bottom)/(1+10^((LogEC50-X)*HillSlope))

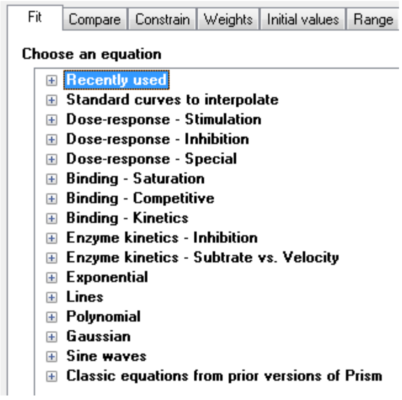


Model: Y=Bottom + (Top-Bottom)/(1+10^((X-LogIC50)))

To run a curve fitting analysis in GraphPad, we first create an XY data table then we enter the logarithm of the concentration of the agonist into X and the response into Y in any convenient units. We enter one data set into column A, and use columns B, C... for different treatments, if needed.
We can enter either biological or technical repeats but we will have to treat them differently during the analysis.

The sigmoid model assumes that the dose response curve has a standard slope, equal to a Hill slope (or slope factor) of -1.0. This is the slope expected when a ligand binds to a receptor following the law of mass action, and is the slope expected of a dose-response curve when the second messenger created by receptor stimulation binds to its receptor by the law of mass action. If we don't have many data points, consider using the standard slope model. If we have lots of data points, pick the variable slope model to determine the Hill slope from the data.



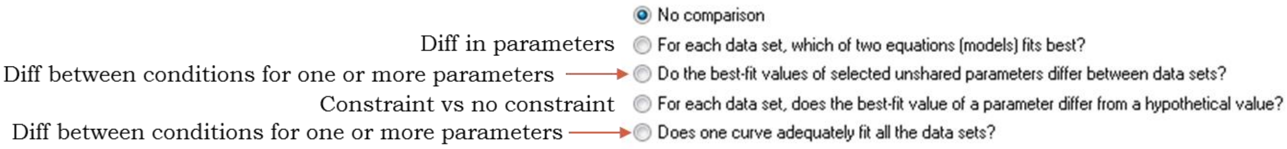


Step by step analysis and considerations:

1- Choose a **Model**. This will come from our knowledge of the experiment we are running. Luckily, GraphPad helps us by listing the possible experiments we might be running. One thing to keep in mind is that it is not necessary to normalise to run a curve-fitting analysis, sometimes it is better to actually show the real data. We should choose it when values for 0 and 100 are precisely defined. Another thing: to go for a variable slope (4 parameters equation) is best when there are plenty of data points.

2- Choose a **Method**: outliers, fitting method, weighting method and replicates.

3- **Compare** different conditions: depending on our questions, we choose between 4 options.



● No comparison
Diff in parameters ○ For each data set, which of two equations (models) fits best?
Diff between conditions for one or more parameters → ○ Do the best-fit values of selected unshared parameters differ between data sets?
Constraint vs no constraint ○ For each data set, does the best-fit value of a parameter differ from a hypothetical value?
Diff between conditions for one or more parameters → ○ Does one curve adequately fit all the data sets?

4- **Constrain**: depends on the experiment, depends if the data define, or not, the top or the bottom of the curve.

5- **Initial values**: defaults usually OK unless the fit looks funny

6- **Range**: defaults usually OK unless we are not interested in the x-variable full range (i.e. time)

7- **Output**: summary table presents same results in a summarised way.

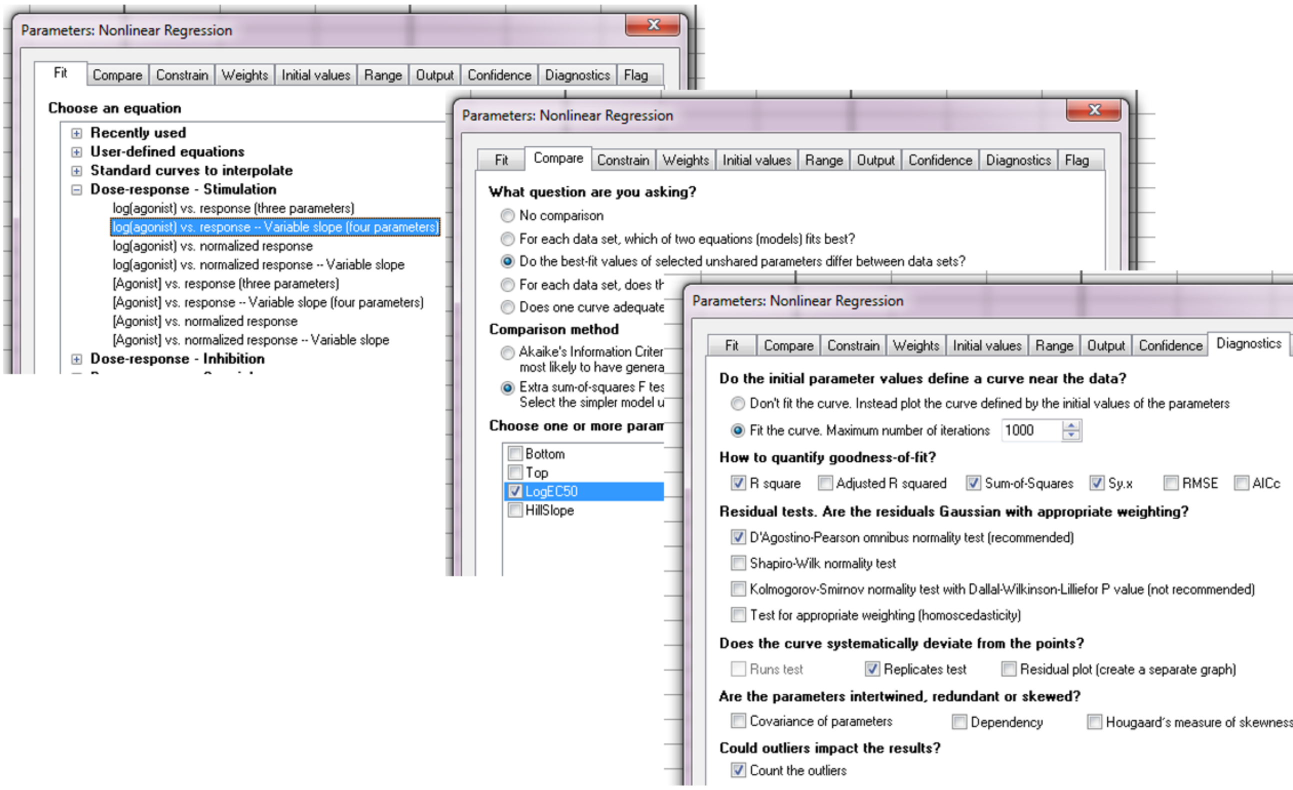
8- **Confidence**: calculate and plot confidence intervals

9- **Diagnostics**: check for normality (weights) and outliers (but keep them in the analysis). Also run the replicates test to see whether the curve gets too far from the points or not. Finally have a look at the residual plots.

Example (File: inhibition data.xlsx).

The Y values are the raw response to the agonist concentrations and the X values are the log of the concentration of the agonist. The replicates are biological replicates.

To run the analysis, we go **>Analysis>Nonlinear regression (curve fit)**. We choose, **log(agonist) vs. response – Variable slope (four parameters)**.



Choose an equation

- ⊞ Recently used
- ⊞ User-defined equations
- ⊞ Standard curves to interpolate
- ⊟ Dose-response - Stimulation
  - log(agonist) vs. response (three parameters)
  - log(agonist) vs. response -- Variable slope (four parameters)
  - log(agonist) vs. normalized response
  - log(agonist) vs. normalized response -- Variable slope
  - [Agonist] vs. response (three parameters)
  - [Agonist] vs. response -- Variable slope (four parameters)
  - [Agonist] vs. normalized response
  - [Agonist] vs. normalized response -- Variable slope
- ⊞ Dose-response - Inhibition

What question are you asking?

○ No comparison
○ For each data set, which of two equations (models) fits best?
● Do the best-fit values of selected unshared parameters differ between data sets?
○ For each data set, does th
○ Does one curve adequate

Comparison method

○ Akaike's Information Criter most likely to have genera
● Extra sum-of-squares F tes Select the simpler model u

Choose one or more param

□ Bottom
□ Top
☑ LogEC50
□ HillSlope

Do the initial parameter values define a curve near the data?

○ Don't fit the curve. Instead plot the curve defined by the initial values of the parameters
● Fit the curve. Maximum number of iterations 1000

How to quantify goodness-of-fit?

☑ R square □ Adjusted R squared ☑ Sum-of-Squares ☑ Sy.x □ RMSE □ AICc

Residual tests. Are the residuals Gaussian with appropriate weighting?

☑ D'Agostino-Pearson omnibus normality test (recommended)
□ Shapiro-Wilk normality test
□ Kolmogorov-Smirnov normality test with Dallal-Wilkinson-Lilliefor P value (not recommended)
□ Test for appropriate weighting (homoscedasticity)

Does the curve systematically deviate from the points?

□ Runs test ☑ Replicates test □ Residual plot (create a separate graph)

Are the parameters intertwined, redundant or skewed?

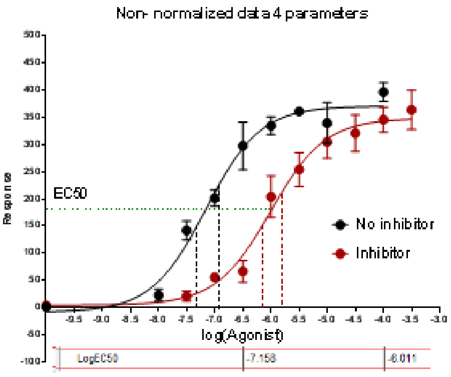
□ Covariance of parameters □ Dependency □ Hougaard's measure of skewness

Could outliers impact the results?

☑ Count the outliers

The results are shown below.

LogEC50 same for all data sets
LogEC50 different for each data set
< 0.0001
Reject null hypothesis
LogEC50 different for each data set
64.86 (1,48)



LogEC50	-7.158	-6.011

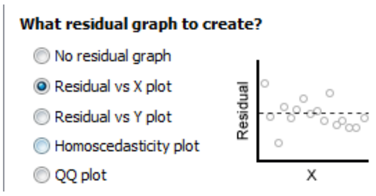
95% Confidence Intervals		
Bottom	-41.39 to 24.94	-22.15 to 31.56
Top	348.3 to 392.6	323.1 to 373.0
LogEC50	-7.324 to -6.991	-6.185 to -5.837
HillSlope	0.6347 to 1.159	0.6095 to 1.186
EC50	4.739e-008 to 1.020e-007	6.538e-007 to 1.455e-006

R square	0.9663	0.9653

The way the graph looks is a good way to check if the model is a good fit. We have also run Replicate tests to check that the curve does not deviate from the data. The p-value is small for the No Inhibitor group indicating that the curve might not describe the data that well.

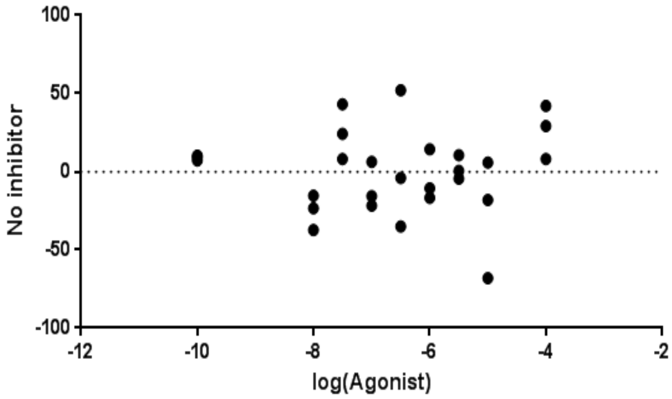
	No inhibitor	Inhibitor
Replicates test for lack of fit		
SD replicates	22.71	25.52
SD lack of fit	41.84	32.38
Discrepancy (F)	3.393	1.610
P value	0.0247	0.1989
Evidence of inadequate model?	Yes	No

One way to check about it is to look at the residuals. In Diagnostics:



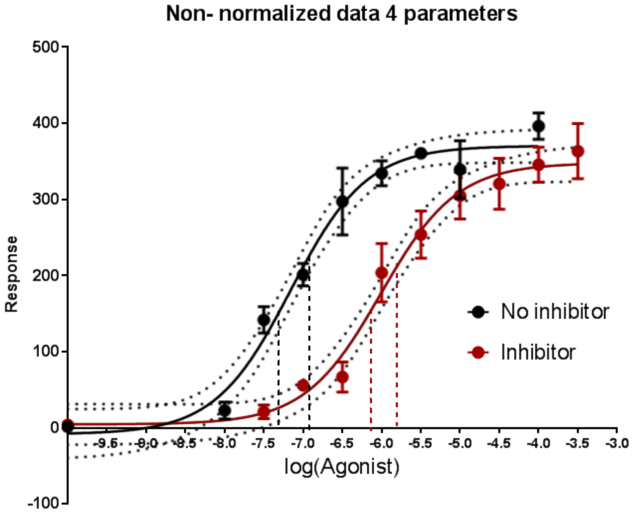
What residual graph to create?

○ No residual graph
● Residual vs X plot
○ Residual vs Y plot
○ Homoscedasticity plot
○ QQ plot



The scatter of the residuals looks random enough so we can accept the model.
We should also look

- at the parameters: are they plausible?
- at the confidence intervals: are they not too wide?



We can also

- Check for outliers and departure from normality (both OK here)
- Finally, we should have a look at the R2 and check that it is as close as possible to 1

Chapter 8: Qualitative data

8-1 Comparing 2 groups

Qualitative data are non-numerical data and the values taken are usually names (also *nominal* data, e.g. variable sex: male or female). The values can be numbers but not numerical (e.g. an experiment number is a numerical label but not a unit of measurement). A qualitative variable with intrinsic order in their categories is *ordinal*. Finally, there is the particular case of qualitative variables with only 2 categories, it is then said to be *binary* or *dichotomous* (e.g. alive/dead or male/female).

We are going to use an example to go through the analysis and the plotting of categorical data.

Example (File: `cats and dogs.xlsx`)

A researcher is interested in whether animals could be trained to line dance. He takes some cats and dogs (**animal**) and 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 dogs' and 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, we would have run a pilot study to get some idea of the type of effect size we are expecting to see. Let's start with this ideal situation and concentrate on the cats. Let's say, in our pilot study, we found that 25% of the cats did line dance after they received affection and 70% did so after they received food.

Using G*Power (see below), we should follow a 4 steps approach:

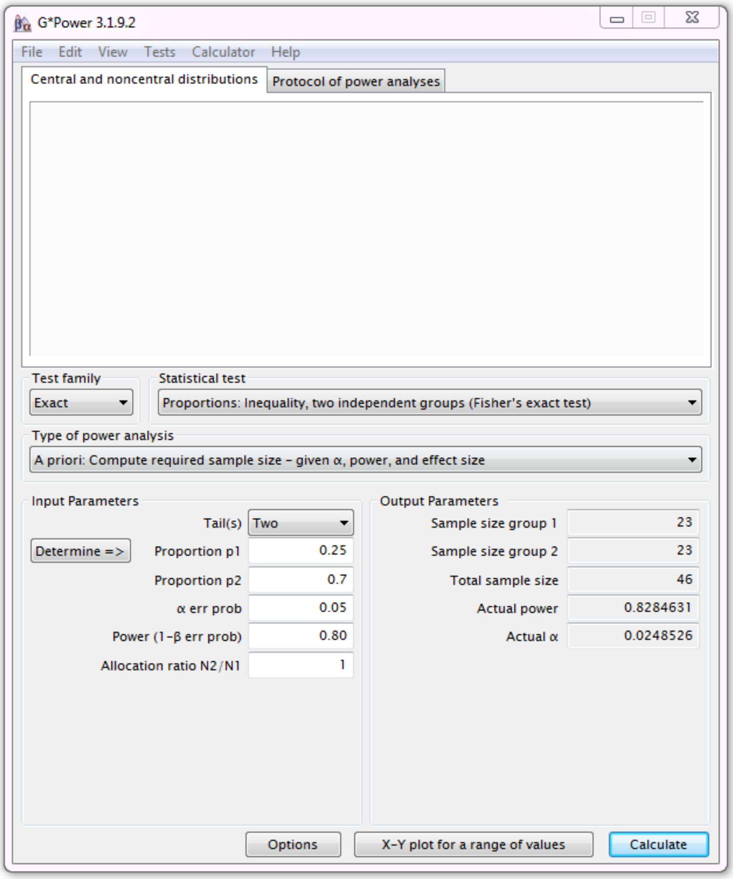
-Step 1: the <u>Test family</u>. We are going for the Fisher's exact test, we should go for 'Exact'.

-Step 2: the <u>Statistical Test</u>: we are looking at proportions and we want to compare 2 independent groups.

-Step 3: the <u>Type of Power Analysis</u>: we know our significant threshold (α=0.05), the power we are aiming for (80%), we have the results from the pilot study so we can calculate the effect size: we go for an '*a priori*' analysis.

-Step 4: the tricky one, we need to **<u>Input Parameters</u>**. Well, it is the tricky one when we have no idea of the effect size but in this case we are OK. Plus if we enter the results for the pilot study, G*Power calculates the effect size for us.

So if we do all that, G*Power will tell us that we need 2 samples of 23 cats to reach a power of 80%. In other words: if we want to be at least 80% sure to spot a treatment effect, if indeed there is one, we will need about 46 cats altogether.


G*Power 3.1.9.2
File Edit View Tests Calculator Help
Central and noncentral distributions
Protocol of power analyses
Determine =>
Options
X-Y plot for a range of values
Calculate

Test family: Exact

Statistical test: Proportions: Inequality, two independent groups (Fisher's exact test)

Type of power analysis: A priori: Compute required sample size – given α, power, and effect size

Input Parameters

Tail(s)	Two
Proportion p1	0.25
Proportion p2	0.7
α err prob	0.05
Power (1-β err prob)	0.80
Allocation ratio N2/N1	1

Output Parameters

Sample size group 1	23
Sample size group 2	23
Total sample size	46
Actual power	0.8284631
Actual α	0.0248526

It is quite intuitive that after having run such an experiment, we are going to end up with a contingency table that is going to show the number of animals who danced or not according to the type of training they received. Those contingency tables are presented below.

Count

		Type of training		
		Food	Affection	Total
Did they dance?	yes	26	6	32
	no	6	30	36
Total		32	36	68

Cat

Count

		Type of training		
		Food	Affection	Total
Did they dance?	Yes	23	24	47
	no	9	10	19
Total		32	34	66

Dog

The first thing to do is enter the data into GraphPad. As mentioned before, while for some software it is OK (or even easier) to prepare our data in Excel and then import them, it is not such a good idea with GraphPad because the structure of the worksheets varies with the type of graph we want to do. So, first, we need to open a New Project which means that we have to choose among the different types of tables mentioned earlier. In our case we want to build a contingency table, so we choose ‘Contingency’ and we click on OK.

The next step is to enter the data after having named the columns and the rows.

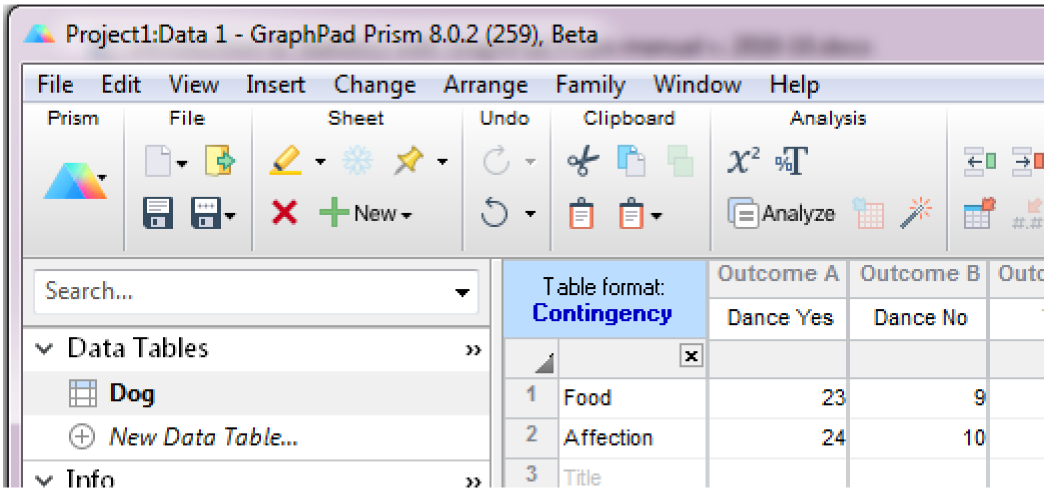
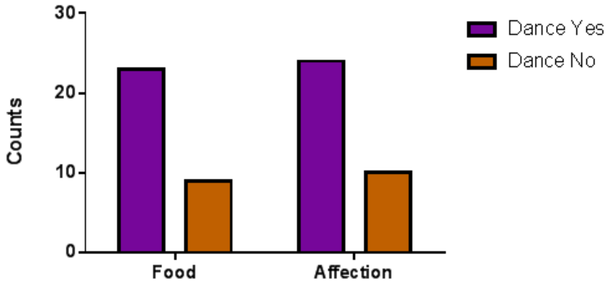
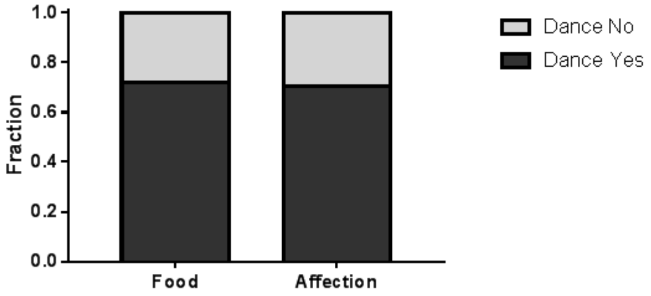

Project1:Data 1 - GraphPad Prism 8.0.2 (259), Beta
File Edit View Insert Change Arrange Family Window Help
Prism
File
Sheet
Undo
Clipboard
Analysis
New
Analyze
Search...
Data Tables
Dog
New Data Table...
Info

	Table format: Contingency	Outcome A Dance Yes	Outcome B Dance No	Outc
1	Food	23	9	
2	Affection	24	10	
3	Title			

The first thing we want to do is to look at a graphical representation of the data. GraphPad will have prepared it for us and if we go into ‘Graphs’ we will see the results.

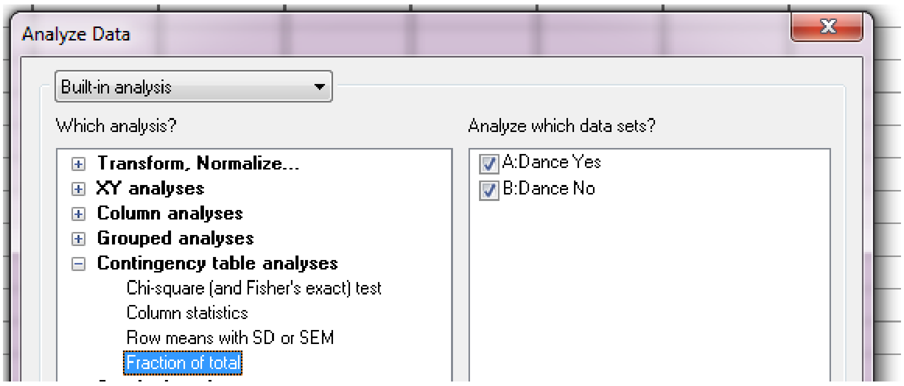
We can change pretty much everything on a graph in GraphPad and it is very easy to make it look like either of the graphs below.





I will not go into much detail in this manual about all the graphical possibilities of GraphPad because it is not its purpose, but it is very intuitive and basically, once we have entered the data in the correct way, we are OK. After that all we have to do is click on the bit we want to change and, usually, a window will pop up.

To get the graph on the right however, we need to add extra step: **Analyze>Contingency table analyses>Fraction of total**. This will produce a data table containing the data as proportions which we can then plot.



Analyze Data

Built-in analysis

Which analysis?

- ⊞ Transform, Normalize…
- ⊞ XY analyses
- ⊞ Column analyses
- ⊞ Grouped analyses
- ⊟ Contingency table analyses
  - Chi-square (and Fisher's exact) test
  - Column statistics
  - Row means with SD or SEM
  - Fraction of total

Analyze which data sets?

☑ A:Dance Yes

☑ B:Dance No

As mentioned before, to analyse such data we need to use a Fisher's exact test but we could also use a χ^2 test (Chi-squared).

Both tests will give us similar p-values for big samples, but for small samples the difference can be a bit more important and the p-value given by Fisher's exact test is more accurate. 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, as in GraphPad for example, 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 can also ask: knowing that 32 of the 68 cats did dance and that 36 of the 68 received affection, what is the probability that those 32 dancers would be so unevenly distributed between the 2 types of reward?

When we want to insert another sheet we have 2 choices. If the second sheet has the same structure and variables' names that the first one, we can right-click on the first sheet name (here 'Dog') and choose 'Duplicate family' and all we have to do is change the values. If the second sheet has different structure, we click on 'New>New data table' in the Sheet Menu.

A bit of theory: the Chi^2 test

It could be either:

- 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.
- 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 us 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 Chi^2 value is calculated using the formula below:

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

The observed frequencies are the one we measured, the values that are in our table. Now, the expected ones are calculated this way:

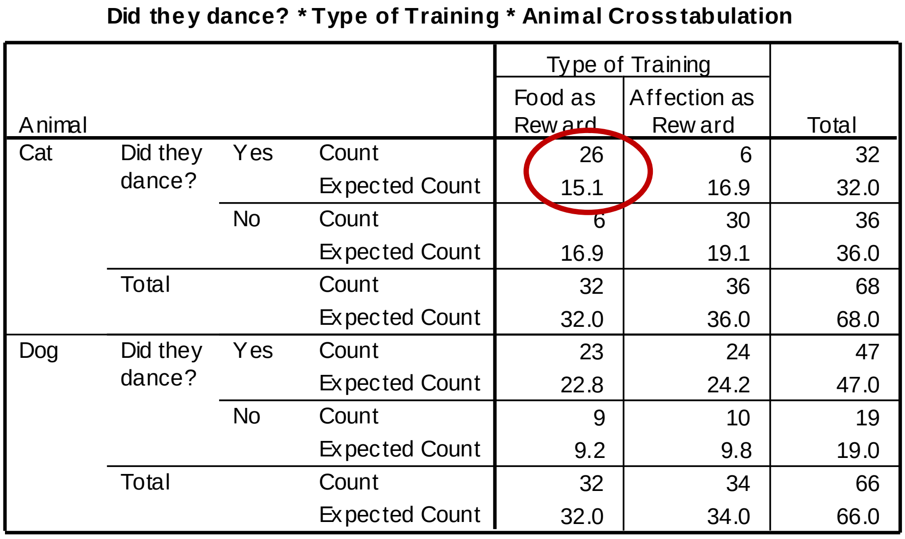
Expected frequency = (row total)*(column total)/grand total

So, for the cat, for example, the expected frequency of cats line dancing after having received food as reward would be : (32*32)/68 = 15.1

Now we can also choose a probability approach:

- probability of line dancing: 32/68
- probability of receiving food: 32/68

If the 2 events are independent, the probability of the 2 occurring at the same time (the expected frequency) will be: (32/68)*(32/68) = 0.22 and 22% of 68 = 15.1



Did they dance? * Type of Training * Animal Crosstabulation

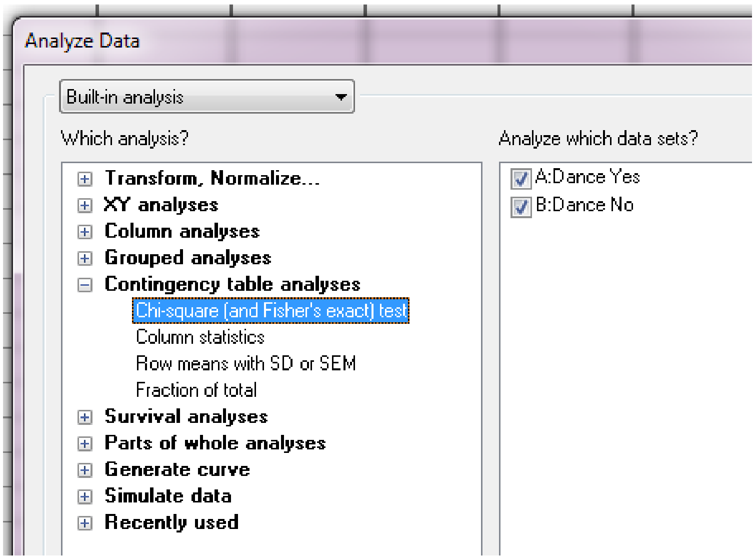
Animal				Type of Training Food as Reward	 Affection as Reward	Total
Cat	Did they dance?	Yes	Count	26	6	32
			Expected Count	15.1	16.9	32.0
		No	Count	6	30	36
			Expected Count	16.9	19.1	36.0
	Total		Count	32	36	68
			Expected Count	32.0	36.0	68.0
Dog	Did they dance?	Yes	Count	23	24	47
			Expected Count	22.8	24.2	47.0
		No	Count	9	10	19
			Expected Count	9.2	9.8	19.0
	Total		Count	32	34	66
			Expected Count	32.0	34.0	66.0

Intuitively, one can see that we are kind of averaging things here, we try to find out the values we should have got by chance. If we work out the values for all the cells, we get:

So for the cat, the χ^2 value is:

$(26-15.1)^2/15.1 + (6-16.9)^2/16.9 + (6-16.9)^2 /16.9 + (30-19.1)^2/19.1 = 28.4$

Let's do it with GraphPad. To calculate either of the tests, we click on = **Analyze** in the tool bar menu, then the window below will appear.



Analyze Data

Built-in analysis

Which analysis?

- ⊞ **Transform, Normalize...**
- ⊞ **XY analyses**
- ⊞ **Column analyses**
- ⊞ **Grouped analyses**
- ⊟ **Contingency table analyses**
  - Chi-square (and Fisher's exact) test
  - Column statistics
  - Row means with SD or SEM
  - Fraction of total
- ⊞ **Survival analyses**
- ⊞ **Parts of whole analyses**
- ⊞ **Generate curve**
- ⊞ **Simulate data**
- ⊞ **Recently used**

Analyze which data sets?

☑ A:Dance Yes
☑ B:Dance No

GraphPad will offer us by default the type of analysis which goes with the type of data we have entered. So, for the question, 'Which analysis?' for Contingency table, the answer is Chi-square and Fisher's exact test.

If we are happy with it, and after having checked that the data sets to be analysed are the ones we want, we can click on OK. The complete analysis will then appear in the Results section.

Below are presented the results for the χ^2 and the Fisher's exact test for the dogs.

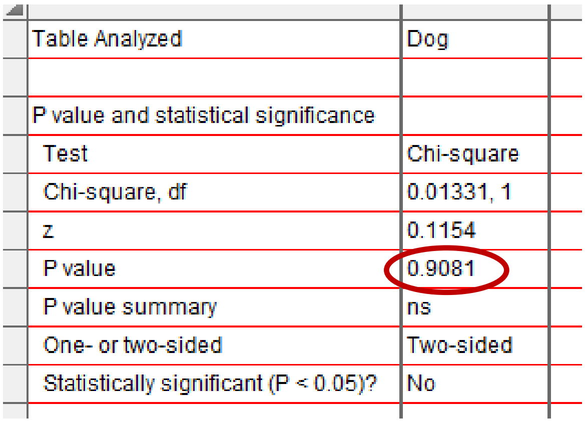


Table Analyzed	Dog
P value and statistical significance	
Test	Chi-square
Chi-square, df	0.01331, 1
z	0.1154
P value	0.9081
P value summary	ns
One- or two-sided	Two-sided
Statistically significant (P < 0.05)?	No

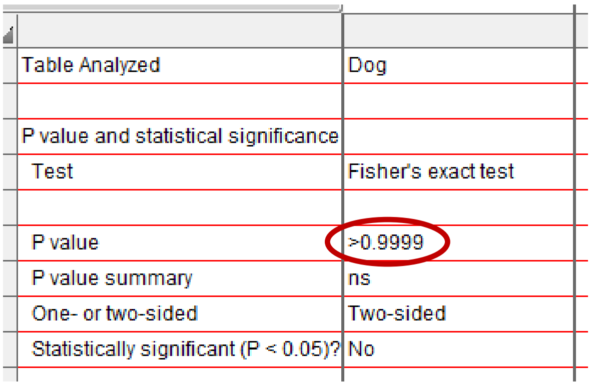


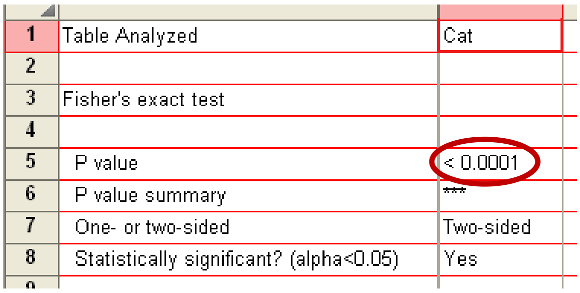
Table Analyzed	Dog
P value and statistical significance	
Test	Fisher's exact test
P value	>0.9999
P value summary	ns
One- or two-sided	Two-sided
Statistically significant (P < 0.05)?	No

Let’s start with the χ^2: there is only one assumption that we have to be careful about when we run it: 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). If we have a high proportion of cells with a small value in it, then we should use a Fisher’s exact test. However, as I said before much software - including GraphPad - only offers the calculation of the Fisher’s exact test for 2x2 tables. So when we have more than 2 categories and a small sample we are in trouble. We have 2 solutions to solve the problem: either we collect more data or we group the categories to boost the proportions.

If you 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.

As we can see here the p-values vary slightly between the 2 tests (>0.99 vs.0.9081) though the conclusion remains the same: the type of reward has no effect whatsoever on the ability of dogs to line dance. Though the samples are not very big here, the assumptions for the χ^2 are met so we can choose either test.

As for the cats, we are more than 99% confident (p< 0.0001) when we say that cats are more likely to line dance when they receive food as a reward than when they receive affection.



1	Table Analyzed	Cat
2		
3	Fisher's exact test	
4		
5	P value	< 0.0001
6	P value summary	***
7	One- or two-sided	Two-sided
8	Statistically significant? (alpha<0.05)	Yes

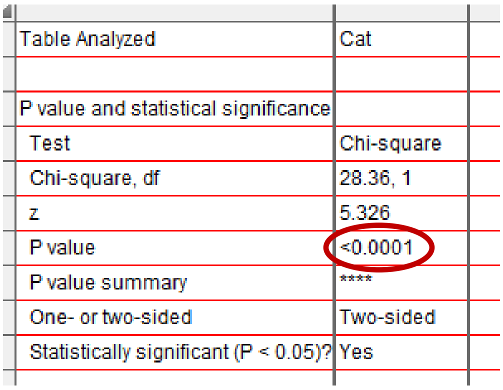
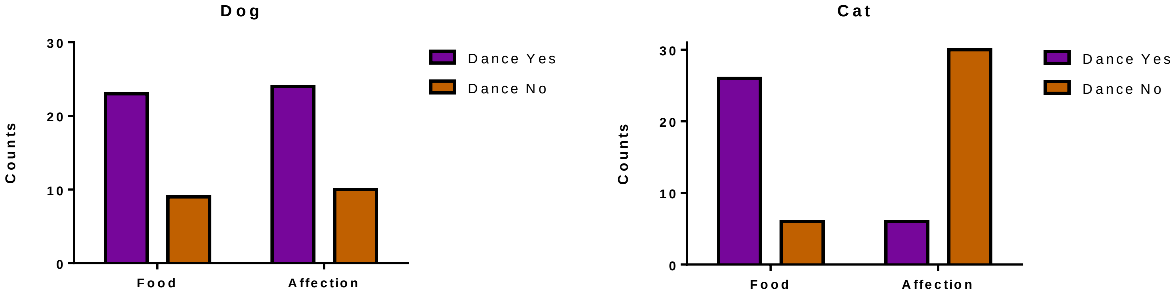
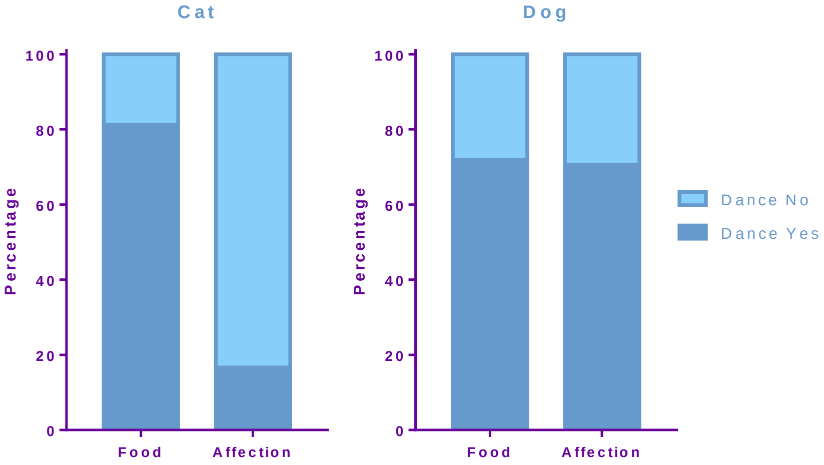


Table Analyzed	Cat
P value and statistical significance	
Test	Chi-square
Chi-square, df	28.36, 1
z	5.326
P value	<0.0001
P value summary	****
One- or two-sided	Two-sided
Statistically significant (P < 0.05)?	Yes

Graphically, we can choose to plot the actual counts:



Or we can plot the percentages or fractions. Though it fails to tell us about the sample size, which is pivotal information for a correct interpretation of the results, it can be more intuitive visually to identify differences.



8-1 Comparing more than 2 groups

Now, if we want to compare more than 2 groups as in more than 2 proportions, it is a bit of an issue as GraphPad Prism does not allow for it. At least not directly not the way we would do it.
The type of analysis we are after here, like a logistic regression for instance, is not supported by Prism, at least not yet.
So let’s go through an example to show how we can overcome that problem.

Example (File: `cane toads`)

In this example, which takes us to Australia, a researcher decided to check the hypothesis that the proportion of cane toads with intestinal parasites was the same in 3 different areas of Queensland.

The question is: Is the proportion of cane toads infected by intestinal parasites the same in 3 different areas of Queensland?

	Infected	Uninfected
Rockhampton	12	8
Bowen	4	16
Mackay	15	5

Easy enough, we enter the data in Prism, and we run a Chi^2. This time, we don't have a choice: Prism does not allows for the Fisher's test with a 3x2 table. We get the result below:

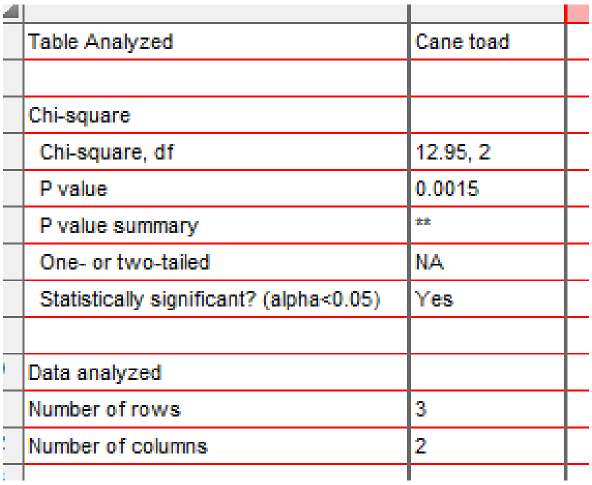
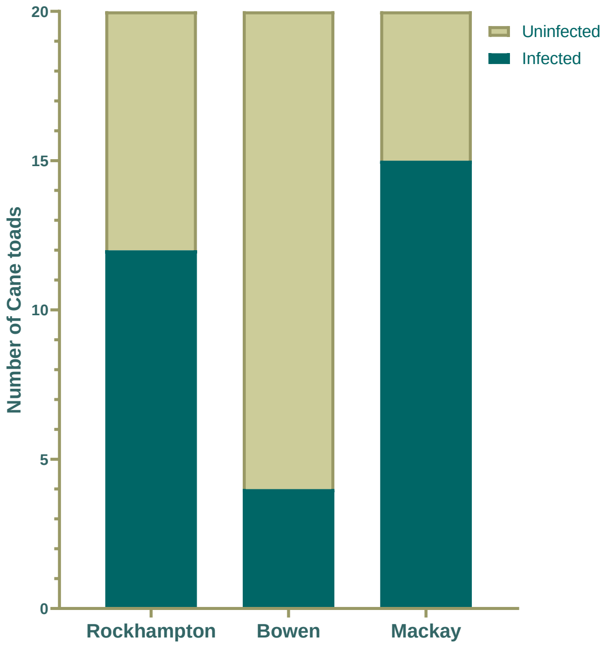
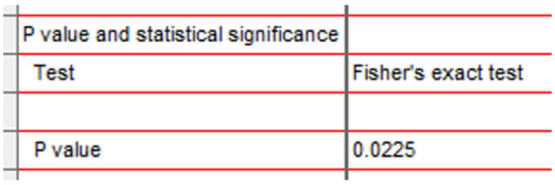


Table Analyzed	Cane toad
Chi-square	
Chi-square, df	12.95, 2
P value	0.0015
P value summary	**
One- or two-tailed	NA
Statistically significant? (alpha<0.05)	Yes
Data analyzed	
Number of rows	3
Number of columns	2

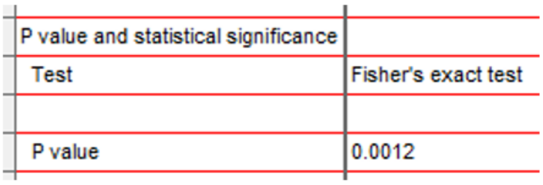
OK, so this p-value tells us that there is a significant difference in infected cane toads in these 3 areas. Which is consistent with what the graph below tells us.



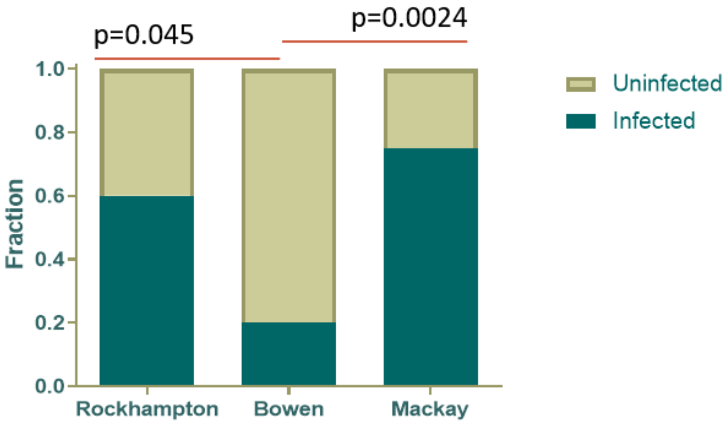
It is basically the equivalent to the omnibus phase of the ANOVA. However Prism does not allow for the post-hoc tests here so we have to do it manually. We have no choice but to run 2 separate analyses to get the pairwise comparisons and then apply a correction for multiple comparisons (here I went for a Bonferroni correction). We can do it by either excluding the values we don't want to include or by creating 2 new tables each containing the data of interest. We could go for the 3 pairwise comparisons but Mackay vs Rockhampton is a pretty done deal.



P value and statistical significance	
Test	Fisher's exact test
P value	0.0225



P value and statistical significance	
Test	Fisher's exact test
P value	0.0012



Chapter 9: Survival analysis

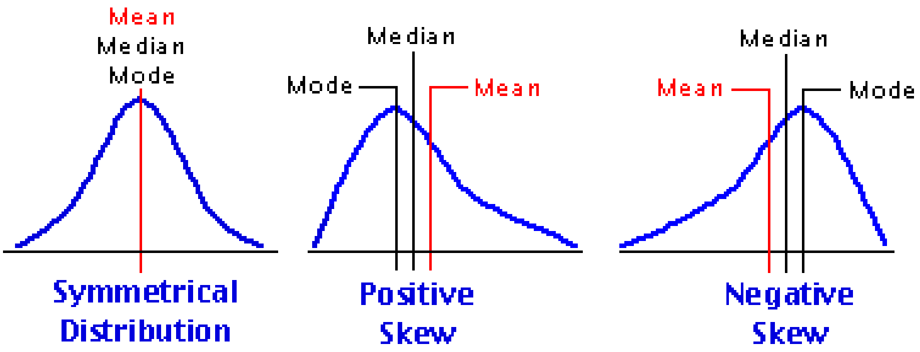
There is a particular type of categorical data: dichotomous outcome over time, namely survival data. Survival analyses are applied to data generated by experiments where the outcome is time until death (or some other one-time event). Most of the time it will be about determining whether a treatment or a condition changes survival.

To run such an analysis, you need to know:
- about time to event data and censoring
- what is a survivor function and a Hazard function and how to plot a Kaplan-Meier estimate
- how to use log-rank tests and simple Cox regression models.

Time to event and censoring

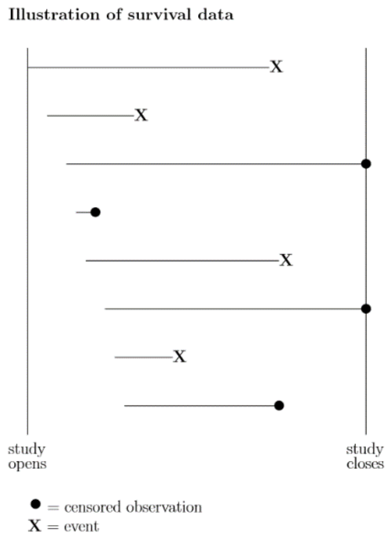
Time to event data can be applied to a wide range of data but it is always about time until a defined event. Examples are: time to death, time to progression of cancer, time to development of diabetes, time to recover from diarrhoea. Time to event data are typically collected in cohort studies (time between study baseline and event of interest) and clinical trials (time between randomisation and event of interest). These data are also referred to as survival data.

Survival data are non-negative values and often not normally distributed (usually positively skewed) which is why the median is much more often used in survival context than the mean.



The event of interest is not usually observed for all individuals during the study. An observation is censored if an individual does not experience it during the study. **Censoring time** is defined as the time from baseline/randomisation until the latest date at which the individual is known to be still alive and event-free. There are several types of censoring:

- Fixed censoring: the event has not occurred when the study has ended or data analysis is performed.
- Loss to follow-up: individual has been lost to follow-up (e.g. he/she no longer wished to take part in the study)
- Competing risks: another event occurs which prevents or changes risk of occurrence of the event of interest.

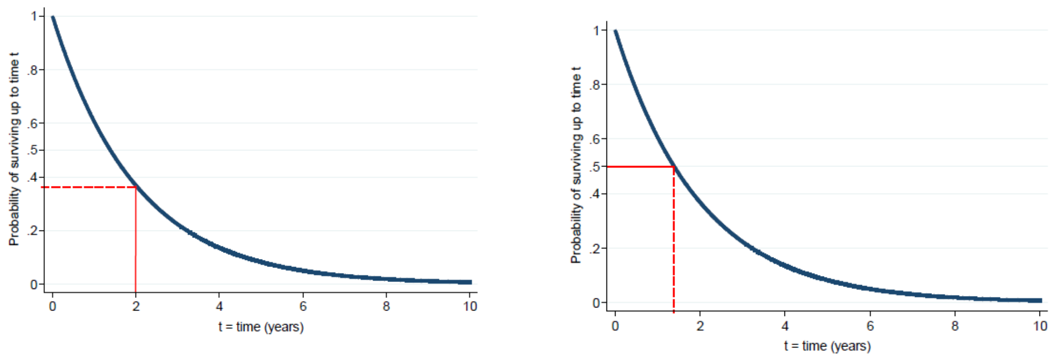


Survival analysis methods make use of information from censoring as in all the subjects of a study (censored or not) are used for probability calculations.
Also survival analyses assume that censoring is non-informative as in if an individual is censored his/hers subsequent risk of the event of interest in unaffected.

The aims of a survival analysis are usually to:

- To estimate the probability of not experiencing the event of interest (not dying = “surviving”) over a given period of time (e.g. 5 year survival rate)
- To compare overall survival experience between different groups of individuals (e.g. between groups in a randomised clinical trial).

A survivor function is often presented as in the graphs below. For example, the probability to survive up to 2 years is 0.37. The information people are usually after is the median survival time: it is the time (expressed in days, months, years …) when half the patients are expected to be alive. It means that the chance of surviving beyond that time is 50%. For example, in the diagram below the median survival time = 1.4 years, since the probability of surviving up to 1.4 years is 0.5.

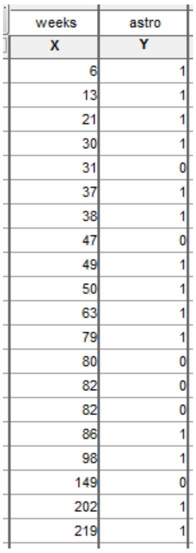


Example of time to event data

Example (File: `'tumours' in Survival data.xlsx`)

Here is a first example of survival data: weeks to deaths or censoring (*) in 20 adults with recurrent astrocytoma (Data reproduced from BMJ 2004; 328:1073).

6	13	21	30	31*	37	38	47*	49	50
63	79	80*	82*	82*	86	98	149	202	219



weeks	astro
X	Y
6	1
13	1
21	1
30	1
31	0
37	1
38	1
47	0
49	1
50	1
63	1
79	1
80	0
82	0
82	0
86	1
98	1
149	0
202	1
219	1

Usually, the first step of a survival analysis is to build a survival curve using the product limit method of Kaplan and Meier.
Here are the steps to build such a curve, often referred to as Kaplan-Meier estimation of survivor function.

<u>First death</u>

There are 20 individuals at t=0 and the first death occurs at t=6 weeks. The probability of dying for each individual is 1/20=0.05. Therefore the probability of surviving beyond t=6 is (1-0.05)=0.95=19/20

6	13	21	30	31*	37	38	47*	49	50
63	79	80*	82*	82*	86	98	149	202	219

Weeks in follow-up (t)	N at risk at time t	N of deaths at time t	Prob. of death at time t	Prob. of no death at time t	Prob. of surviving up to and including time t
0	20	0	0	1	1
6	20	1	0.05	0.95	1 x 0.95 = 0.95

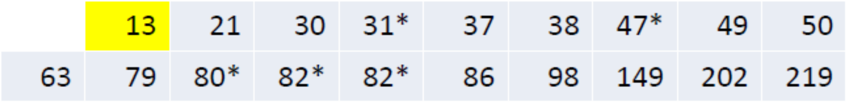
1/20

19/20

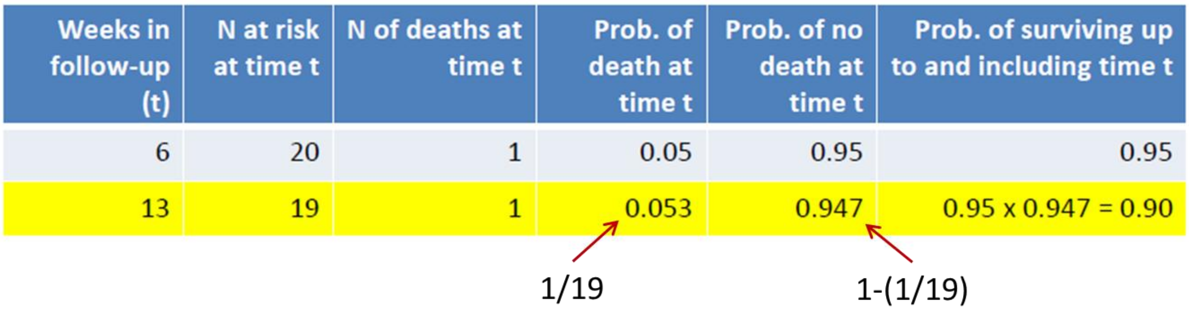
"Risk set" at time t

Second death:

There are 19 individuals in the study between t=6 and t=13 when the second death occurs. No individual is censored during that period so the probability of dying for each individual is 1/19=0.053. Therefore the probability of surviving beyond t=13 is 0.95*0.947=0.9 with 0.95=19/20=(1-(1/20) and 0.947=18/19 =(1-(1/19)).



	13	21	30	31*	37	38	47*	49	50
63	79	80*	82*	82*	86	98	149	202	219



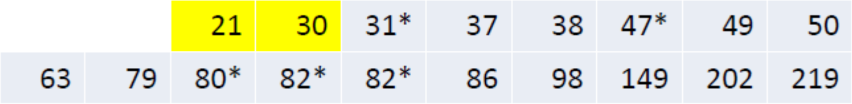
Weeks in follow-up (t)	N at risk at time t	N of deaths at time t	Prob. of death at time t	Prob. of no death at time t	Prob. of surviving up to and including time t
6	20	1	0.05	0.95	0.95
13	19	1	0.053	0.947	0.95 x 0.947 = 0.90

1/19

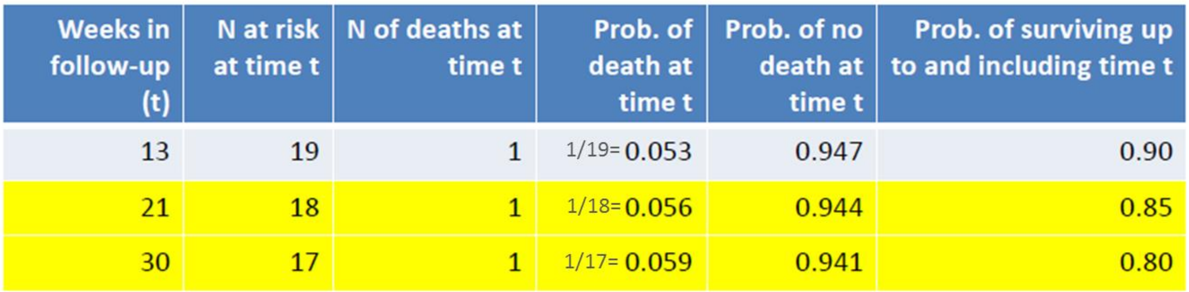
1-(1/19)

Third and fourth death:

Eighteen individuals are in the study between t=13 and t=21 and the probability of dying for each individual is 1/18=0.056. The probability to survive beyond t=21 is 0.9*(1-(1/18))=0.85 with 0.9 coming from t=13. There are 17 individuals in the study between t=21 and t=30 so the probability of death for each individual is 1/17=0.059. Hence the probability of surviving beyond t=30 is 0.85*(1-(1/17))=0.8.



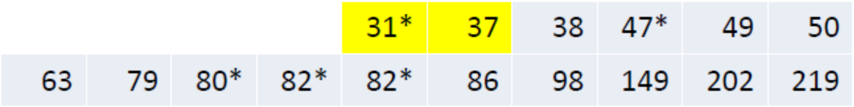
		21	30	31*	37	38	47*	49	50
63	79	80*	82*	82*	86	98	149	202	219



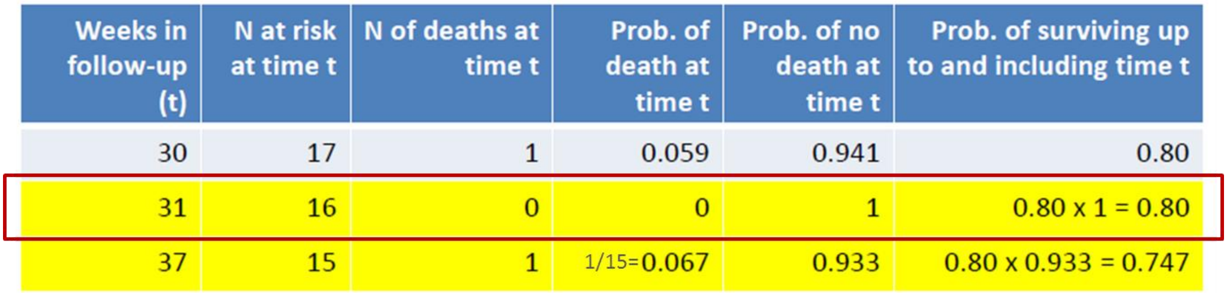
Weeks in follow-up (t)	N at risk at time t	N of deaths at time t	Prob. of death at time t	Prob. of no death at time t	Prob. of surviving up to and including time t
13	19	1	1/19= 0.053	0.947	0.90
21	18	1	1/18= 0.056	0.944	0.85
30	17	1	1/17= 0.059	0.941	0.80

Fifth and sixth death:

Sixteen individuals are in the study between t=30 and t=31 but one individual censored at t=31: the probability of surviving beyond t=31 remains at 0.8. So there are 15 individuals left between t=31 and t=37 and the probability of surviving beyond t=37 is 0.8*(1-(1/15))=0.747.

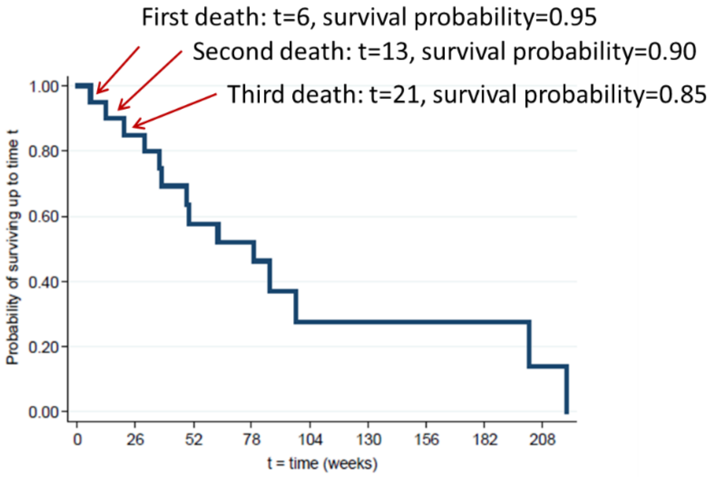


				31*	37	38	47*	49	50
63	79	80*	82*	82*	86	98	149	202	219

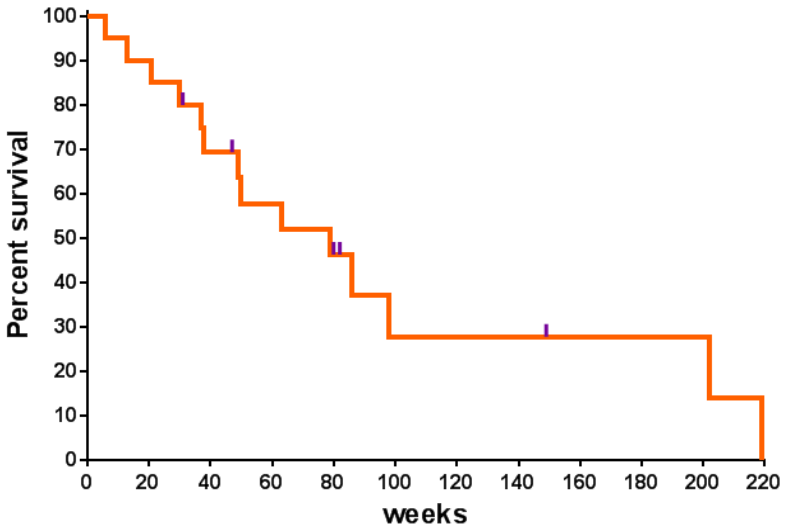


Weeks in follow-up (t)	N at risk at time t	N of deaths at time t	Prob. of death at time t	Prob. of no death at time t	Prob. of surviving up to and including time t
30	17	1	0.059	0.941	0.80
31	16	0	0	1	0.80 x 1 = 0.80
37	15	1	1/15=0.067	0.933	0.80 x 0.933 = 0.747

The calculations continue until reaching the longest event time and K-M curve can be drawn as a step function.



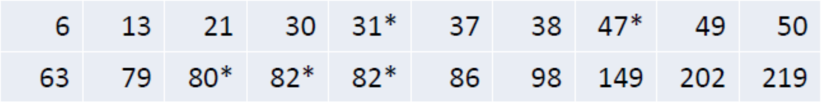
Let’s do it with GraphPad. Create a new table in Survival format and enter the data as in page 18. Then go **Analyze>Survival Curve** (the default is OK). The Data summary tells you the median survival is 79 weeks. And you get a graph similar to the one above.



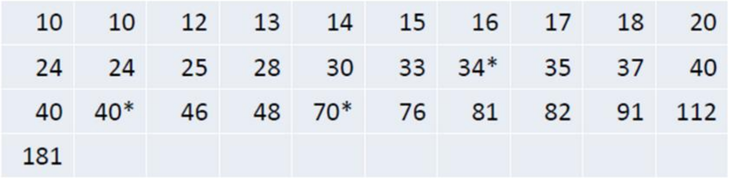
Comparing 2 samples

Example (File: `tumours in Survival data.xlsx`)

We are going to compare survival in adults with recurrent astrocytoma (n=20) to the one of a group of patients with recurrent glioblastoma (n=31).

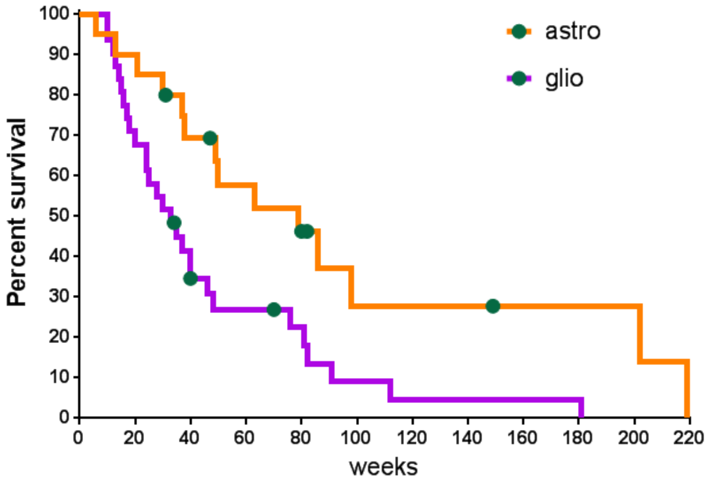


6	13	21	30	31*	37	38	47*	49	50
63	79	80*	82*	82*	86	98	149	202	219

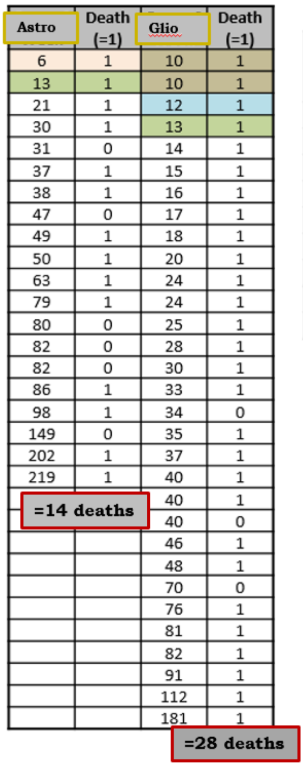


10	10	12	13	14	15	16	17	18	20
24	24	25	28	30	33	34*	35	37	40
40	40*	46	48	70*	76	81	82	91	112
181									

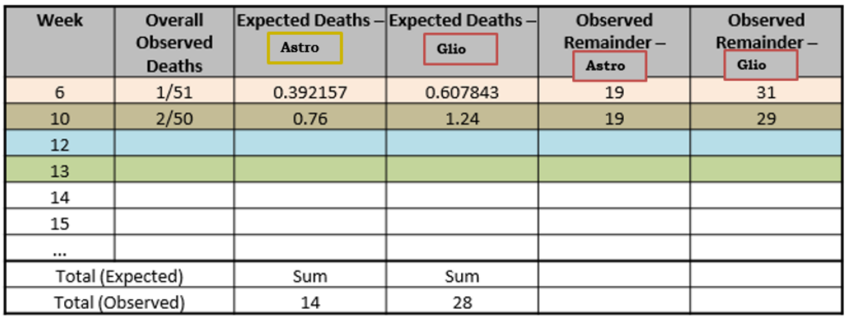
Here is the thing: from the graph below, survival chances appear better in individuals with astrocytoma than with glioblastoma, but is the difference between groups statistically significant?



It is possible to compare the median survival times and the probability of surviving up to any particular time. It is usually better to use a test which compares survivor functions over whole follow-up period. The Log rank test does such a thing: it tests the null hypothesis of no difference between samples in probability of an event (death in this example) at any time point during follow-up. The log rank statistic is basically a χ^2 and is based on calculating expected number of events that would occur under the null hypothesis at each event time and compares it to the observed number of events.



Astro	Death (=1)	Glio	Death (=1)
6	1	10	1
13	1	10	1
21	1	12	1
30	1	13	1
31	0	14	1
37	1	15	1
38	1	16	1
47	0	17	1
49	1	18	1
50	1	20	1
63	1	24	1
79	1	24	1
80	0	25	1
82	0	28	1
82	0	30	1
86	1	33	1
98	1	34	0
149	0	35	1
202	1	37	1
219	1	40	1
=14 deaths		40	1
		40	0
		46	1
		48	1
		70	0
		76	1
		81	1
		82	1
		91	1
		112	1
		181	1
		=28 deaths	

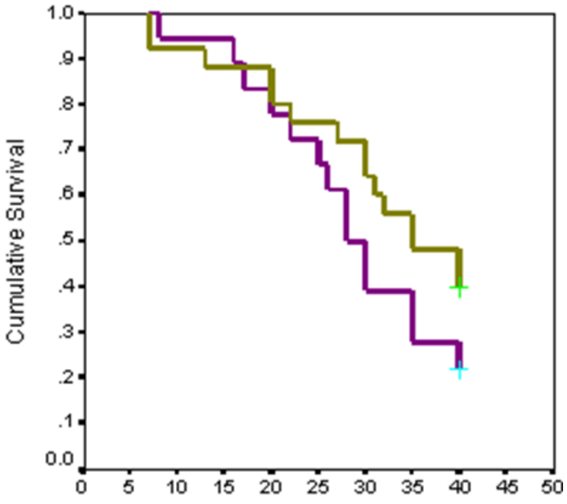


Week	Overall Observed Deaths	Expected Deaths – Astro	Expected Deaths – Glio	Observed Remainder – Astro	Observed Remainder – Glio
6	1/51	0.392157	0.607843	19	31
10	2/50	0.76	1.24	19	29
12					
13					
14					
15					
...					
Total (Expected)		Sum	Sum		
Total (Observed)		14	28		

Log rank test statistic has a Chi² distribution:

$$Z = \frac{\sum_{j=1}^{J}(O_{1j} - E_{1j})}{\sqrt{\sum_{j=1}^{J} V_j}}$$

A log rank-test is unlikely to detect a difference between groups if survivor functions cross over during follow-up (graph below). It assumes non-informative censoring and can be extended to compare more than 2 groups.



But it only provides a p-value, not an estimate of size difference between groups or a confidence interval. To get an estimate of the size of the difference, you need the hazard ratio.

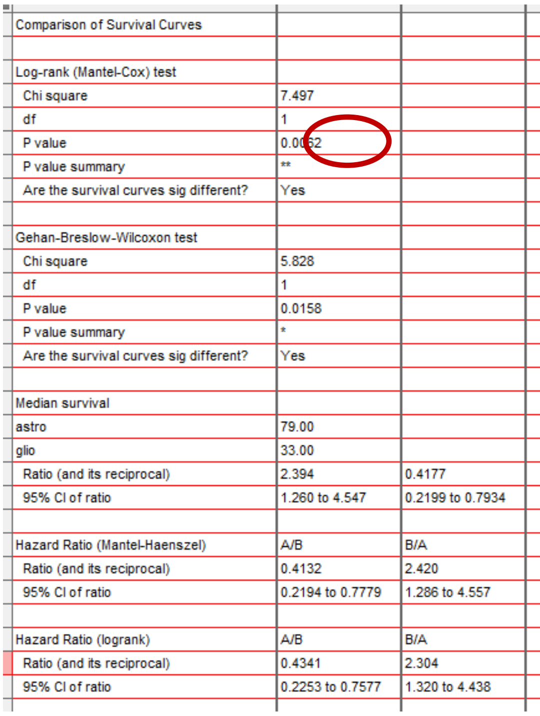
Hazard function

Hazard is defined as the slope of the survival curve: a measure of how rapidly subjects are dying. The hazard function describes how hazard varies over time.

The hazards may vary over time but it is assumed that the hazard ratio (HR) is constant over time. HR is not directly related to the ratio of median survival times as it is calculated in a different way.
In the context of comparing 2 groups (a and b), as part of the Kaplan-Meier calculations, the number of expected events (Ea and Eb) are calculated assuming a null hypothesis of no difference in survival.
With the numbers of observed events (deaths) in each group (Oa and Ob), HR is obtained from Cox regression: HR=(Oa/Ea)/(Ob/Eb)
No assumption is needed about shape of hazard functions or underlying distribution of time to event data.
Let's do it with GraphPad: Analyze>Survival Curve and again the default is OK. We can see that there is a significant difference between the survivals (p=0.0062).
Also:

HR = 2.3 (95% CI [1.32;4.44])
(Hazard$_{\text{Glio}}$ (t) = Hazard$_{\text{Astro}}$(t) x HR)

It means that, at any point in time, hazard (i.e. instantaneous rate) of dying in individuals with recurrent glioblastoma is **2.3 times** higher than in individuals with recurrent astrocytoma.



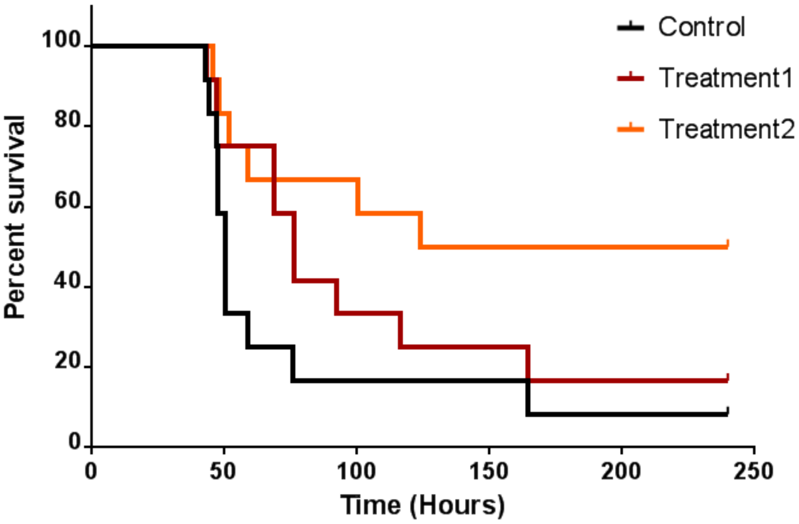
Comparison of Survival Curves		
Log-rank (Mantel-Cox) test		
Chi square	7.497	
df	1	
P value	0.0062	
P value summary	**	
Are the survival curves sig different?	Yes	
Gehan-Breslow-Wilcoxon test		
Chi square	5.828	
df	1	
P value	0.0158	
P value summary	*	
Are the survival curves sig different?	Yes	
Median survival		
astro	79.00	
glio	33.00	
Ratio (and its reciprocal)	2.394	0.4177
95% CI of ratio	1.260 to 4.547	0.2199 to 0.7934
Hazard Ratio (Mantel-Haenszel)	A/B	B/A
Ratio (and its reciprocal)	0.4132	2.420
95% CI of ratio	0.2194 to 0.7779	1.286 to 4.557
Hazard Ratio (logrank)	A/B	B/A
Ratio (and its reciprocal)	0.4341	2.304
95% CI of ratio	0.2253 to 0.7577	1.320 to 4.438

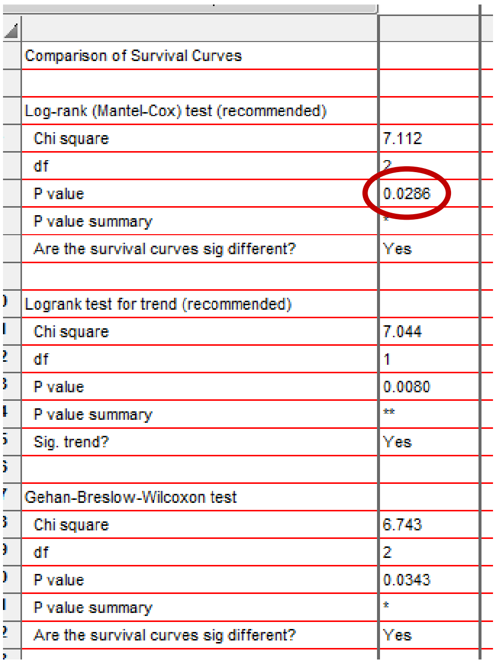
Comparing more than 2 samples

Example (File: `lung infection in Survival data.xlsx`)

Now the issue with GraphPad is that we cannot compare 2 samples directly. It is like when we want to compare more than 2 proportions: the χ^2 test will give us a general p-value (like in the omnibus step for the ANOVA) but will not allow for pairwise comparisons. To get it, we need to be proactive, namely do as many Fisher's tests we need and then apply a correction by hand, usually Bonferroni. This what we have to do with survival analysis as well in GraphPad.

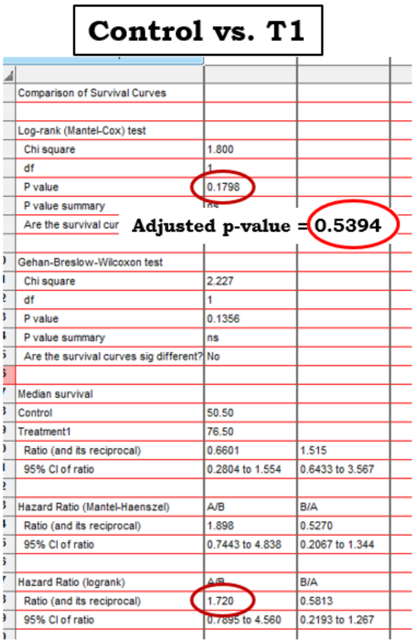
The first step is the same as the one for the 2 samples comparisons: Analyze>Survival Curve





Comparison of Survival Curves	
Log-rank (Mantel-Cox) test (recommended)	
Chi square	7.112
df	2
P value	0.0286
P value summary	*
Are the survival curves sig different?	Yes
Logrank test for trend (recommended)	
Chi square	7.044
df	1
P value	0.0080
P value summary	**
Sig. trend?	Yes
Gehan-Breslow-Wilcoxon test	
Chi square	6.743
df	2
P value	0.0343
P value summary	*
Are the survival curves sig different?	Yes

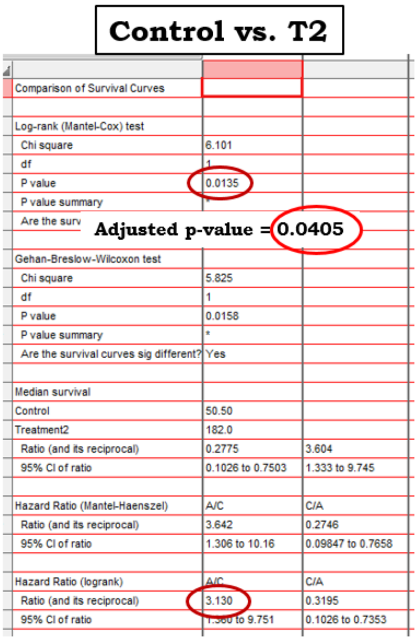
The second is about duplicating the dataset the corresponding number of pairwise comparisons and then run them as in the first step.



Control vs. T1

Comparison of Survival Curves		
Log-rank (Mantel-Cox) test		
Chi square	1.800	
df	1	
P value	0.1798	
P value summary	ns	
Are the survival cur		
Gehan-Breslow-Wilcoxon test		
Chi square	2.227	
df	1	
P value	0.1356	
P value summary	ns	
Are the survival curves sig different?	No	
Median survival		
Control	50.50	
Treatment1	76.50	
Ratio (and its reciprocal)	0.6601	1.515
95% CI of ratio	0.2804 to 1.554	0.6433 to 3.567
Hazard Ratio (Mantel-Haenszel)	A/B	B/A
Ratio (and its reciprocal)	1.898	0.5270
95% CI of ratio	0.7443 to 4.838	0.2067 to 1.344
Hazard Ratio (logrank)	A/B	B/A
Ratio (and its reciprocal)	1.720	0.5813
95% CI of ratio	0.7895 to 4.560	0.2193 to 1.267

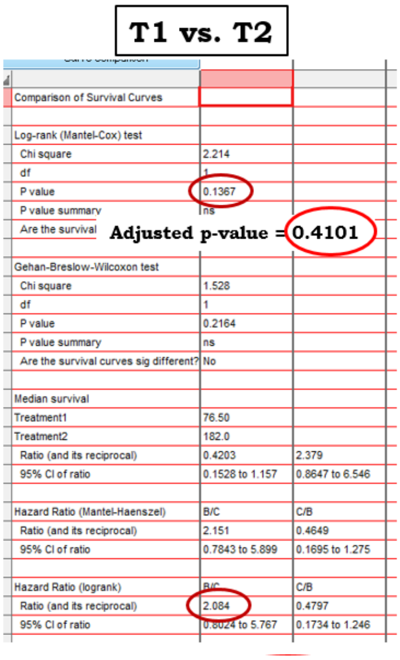
Adjusted p-value = 0.5394



Control vs. T2

Comparison of Survival Curves		
Log-rank (Mantel-Cox) test		
Chi square	6.101	
df	1	
P value	0.0135	
P value summary	*	
Are the surv		
Gehan-Breslow-Wilcoxon test		
Chi square	5.825	
df	1	
P value	0.0158	
P value summary	*	
Are the survival curves sig different?	Yes	
Median survival		
Control	50.50	
Treatment2	182.0	
Ratio (and its reciprocal)	0.2775	3.604
95% CI of ratio	0.1026 to 0.7503	1.333 to 9.745
Hazard Ratio (Mantel-Haenszel)	A/C	C/A
Ratio (and its reciprocal)	3.642	0.2746
95% CI of ratio	1.306 to 10.16	0.09847 to 0.7658
Hazard Ratio (logrank)	A/C	C/A
Ratio (and its reciprocal)	3.130	0.3195
95% CI of ratio	1.360 to 9.751	0.1026 to 0.7353

Adjusted p-value = 0.0405



T1 vs. T2

Comparison of Survival Curves		
Log-rank (Mantel-Cox) test		
Chi square	2.214	
df	1	
P value	0.1367	
P value summary	ns	
Are the survival		
Gehan-Breslow-Wilcoxon test		
Chi square	1.528	
df	1	
P value	0.2164	
P value summary	ns	
Are the survival curves sig different?	No	
Median survival		
Treatment1	76.50	
Treatment2	182.0	
Ratio (and its reciprocal)	0.4203	2.379
95% CI of ratio	0.1528 to 1.157	0.8647 to 6.546
Hazard Ratio (Mantel-Haenszel)	B/C	C/B
Ratio (and its reciprocal)	2.151	0.4649
95% CI of ratio	0.7843 to 5.899	0.1695 to 1.275
Hazard Ratio (logrank)	B/C	C/B
Ratio (and its reciprocal)	2.084	0.4797
95% CI of ratio	0.8024 to 5.767	0.1734 to 1.246

Adjusted p-value = 0.4101

The adjusted p-values are obtained by applying a Bonferroni correction: multiply the p-values by the number of comparisons (e.g. Control vs. T1: p=0.1798 * 3 = p=0.5394).

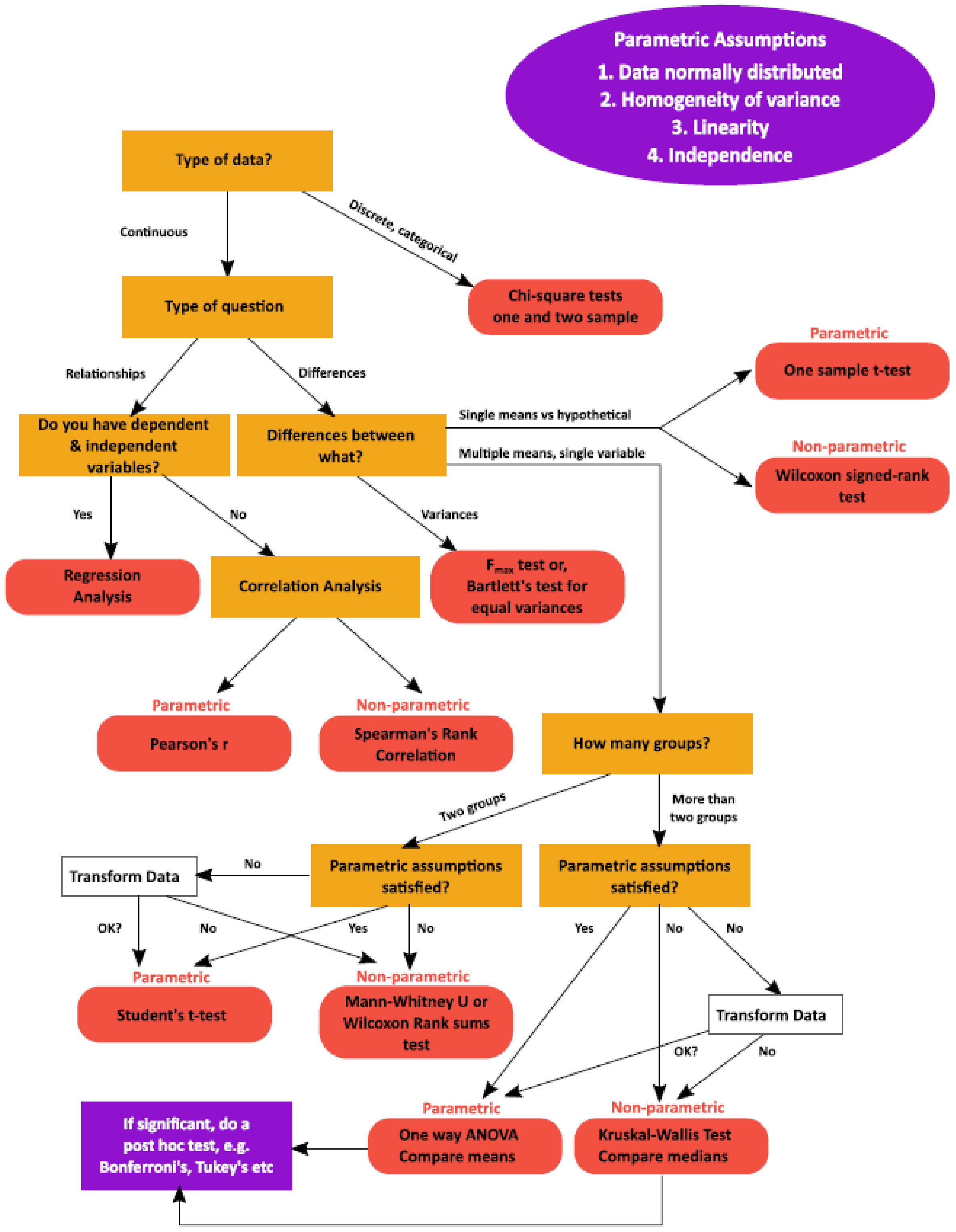
References

Cumming G., Fidler F. and Vaux D.L. 2007. 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.


Parametric Assumptions
1. Data normally distributed
2. Homogeneity of variance
3. Linearity
4. Independence
Type of data?
Continuous
Discrete, categorical
Chi-square tests one and two sample
Type of question
Relationships
Differences
Do you have dependent & independent variables?
Differences between what?
Single means vs hypothetical
Multiple means, single variable
Parametric
One sample t-test
Non-parametric
Wilcoxon signed-rank test
Yes
No
Variances
Regression Analysis
Correlation Analysis
F_{max} test or, Bartlett's test for equal variances
Parametric
Pearson's r
Non-parametric
Spearman's Rank Correlation
How many groups?
Two groups
More than two groups
Transform Data
No
Parametric assumptions satisfied?
Parametric assumptions satisfied?
OK?
No
Yes
No
Yes
No
No
Parametric
Student's t-test
Non-parametric
Mann-Whitney U or Wilcoxon Rank sums test
Transform Data
OK?
No
If significant, do a post hoc test, e.g. Bonferroni's, Tukey's etc
Parametric
One way ANOVA Compare means
Non-parametric
Kruskal-Wallis Test Compare medians