

# **Introduction to Statistics with GraphPad Prism**

## Licence

This manual is © 2008-19, Anne Segonds-Pichon.

This manual is distributed under the creative commons Attribution-Non-Commercial-Share Alike 2.0 licence. This means that you are free:

- to copy, distribute, display, and perform the work
- to make derivative works

Under the following conditions:

- Attribution. You must give the original author credit.
- Non-Commercial. You may not use this work for commercial purposes.
- Share Alike. If you alter, transform, or build upon this work, you may distribute the resulting work only under a licence identical to this one.

Please note that:

- For any reuse or distribution, you must make clear to others the licence terms of this work.
- Any of these conditions can be waived if you get permission from the copyright holder.
- Nothing in this license impairs or restricts the author's moral rights.

Full details of this licence can be found at

<http://creativecommons.org/licenses/by-nc-sa/2.0/uk/legalcode>

## Contents

<b>Introduction to Statistics with GraphPad Prism .....</b>	<b>1</b>
<b>Introduction .....</b>	<b>4</b>
<b>Chapter 1: Power analysis .....</b>	<b>5</b>
<b>Chapter 2: Basic structure of a GraphPad Prism project .....</b>	<b>9</b>
<b>.....</b>	<b>10</b>
<b>Chapter 3: Qualitative data .....</b>	<b>11</b>
<i>Example .....</i>	<i>11</i>
<i>Power Analysis with qualitative data .....</i>	<i>11</i>
<i>A bit of theory: the <math>\chi^2</math> test.....</i>	<i>14</i>
<i>A bit of theory: the null hypothesis and the error types.....</i>	<i>18</i>
<b>Chapter 4: Quantitative data.....</b>	<b>20</b>
4-1 Descriptive stats .....	20
<i>The Mean .....</i>	<i>20</i>
<i>The Median: .....</i>	<i>20</i>
<i>The Variance.....</i>	<i>21</i>
<i>The Standard Deviation (SD) .....</i>	<i>21</i>
<i>Standard Deviation vs. Standard Error .....</i>	<i>22</i>
<i>Confidence interval.....</i>	<i>23</i>
4-2 Assumptions of parametric data .....	24
<i>How can we check that our data are parametric/normal?.....</i>	<i>25</i>
<i>Example .....</i>	<i>26</i>
<i>Power analysis with a t-test.....</i>	<i>26</i>
4-3 The t-test.....	31
<i>Independent t-test .....</i>	<i>32</i>
<i>Paired t-test.....</i>	<i>34</i>
<i>Example .....</i>	<i>34</i>
4-4 Comparison of more than 2 means: Analysis of variance .....	37
<i>A bit of theory .....</i>	<i>37</i>
<i>Example .....</i>	<i>39</i>
<i>Power analysis with an ANOVA .....</i>	<i>39</i>
4-5 Correlation .....	43
<i>Example .....</i>	<i>43</i>
<i>A bit of theory: Correlation coefficient .....</i>	<i>43</i>
<i>Power analysis with correlation.....</i>	<i>46</i>
4-6 Curve fitting: Dose-response .....	47
<i>Example .....</i>	<i>49</i>

## 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 a bit friendlier than with SPSS for instance. SPSS does not hold your hand all the way through the analysis, whereas GraphPad does. On the down side, GraphPad is not as powerful as SPSS - as in we cannot do as many as different analyses with GraphPad as we can with SPSS. If we need to run say a 3-way ANOVA for example, then we would need to use SPSS.

Both GraphPad and SPSS 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!

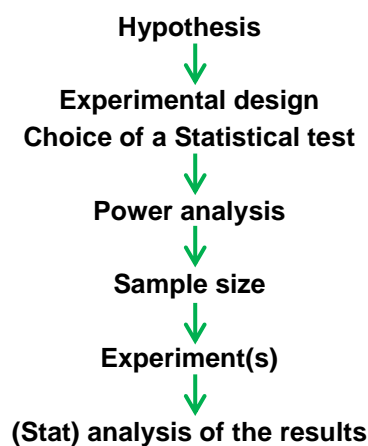
*"To consult a statistician after an experiment is finished is often merely to ask him to conduct a post-mortem examination. He can perhaps say what the experiment died of."* R.A.Fisher, 1938.

## Chapter 1: Power analysis

First, the definition of power: 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. 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**' 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.

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. From a more practical point of view, when we write a grant, we need to justify our sample size which we do through a power analysis. Finally, it is all about the ethics of research really, which is encapsulated in the Home office's **3 Rs: 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' (<https://www.nc3rs.org.uk/>).

When should we run our 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:

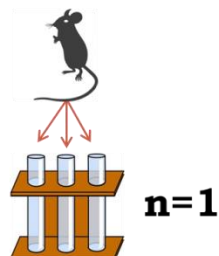


The power analysis depends on the relationship between 6 variables: the **effect size** of biological interest, the **standard deviation**, the **significance level**, the desired **power**, the **sample size** and the **alternative hypothesis**. The significance level is about the p-value, it is generally agreed to be 5% and we will come back to it later. The desired power, as mentioned earlier is usually 80%. The alternative hypothesis is about choosing between one and 2-sided tests, it is a technical thing and we will come back to that later as well. So we are left with the 3 variables on which we have pretty much no control or about which we cannot decide arbitrarily: the effect size, the sample size and the standard deviation. To help us understand what they are and how much they are connected, here's an example.

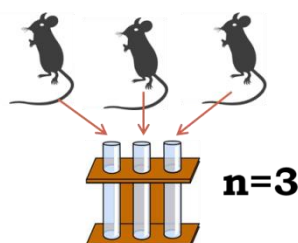
Let's make it simple: say we are studying a gene which is expressed in the brain and we are using a mouse model. Our hypothesis is that knocking out that gene will affect the mouse's behaviour. The next step is to design our experiment. We are going to create a KO mouse in which gene A is inoperative and we are going to compare WT and KO mice's behaviour through a set of tasks. Let's say that the output of one of these tasks is a quantitative variable, the time taken by the mouse to achieve one task, for example. Now we need to translate the hypothesis for this particular task into a statistical question.

The hypothesis is that knocking out gene A will affect KO mice behaviour which can be quantified by a change in the time taken to achieve the task. Statistically we need to know: what type of data we are going to collect (time), which test we are going to use to compare the 2 genotypes and how many mice we will need. When thinking about sample size, it is very important to consider the difference between **technical** and **biological** replicates. Technical replicates involve taking several samples from one tube and analysing it across multiple conditions. Biological replicates are different samples measured across multiple conditions.

### Technical



### Biological



Now remember that a power analysis is about the relationship between 6 variables, 3 of which are directly linked to the experiment: the sample size, the effect size and the standard deviation.

The diagram below explains about these 3 variables.

## Time to think about statistical power

You send your child into the basement to find a tool.

He comes back and says "it isn't there".

What do you conclude? Is the tool there or not? There is no way to be sure.

"If the tool really is in the basement, what are the chances that your child would have found it?"

"If there is a difference between WT and KO, what are the chances that your experiment will pick it up ( $p < 0.05$ )?"

It depends on:

**In the house**

How **long** did he spend looking?



How **big** is the tool?



or



How **messy** is the basement?



or

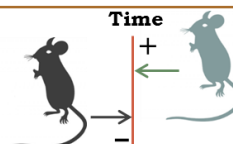


**In the lab**

**The sample size**



**The effect size**



**The Standard Deviation (SD)**



First, the **sample size**, the name itself is self-explanatory. The aim of a power analysis is usually to find the appropriate sample size as in the one which will allow us to detect a specified effect. This effect, also called

**effect size** of biological interest, can only be determined scientifically, not statistically. It is either a difference that would be meaningful biologically, like an increase/decrease of 10% or 20% for a particular variable, or what we expect to get based on preliminary data. The larger the effect size, the smaller the experiment will need to be to detect it.

The **Standard Deviation** (SD) is basically the noise in our data, the variability we observe between our values. This we get ideally from a pilot study or from previous experiments or even the literature.

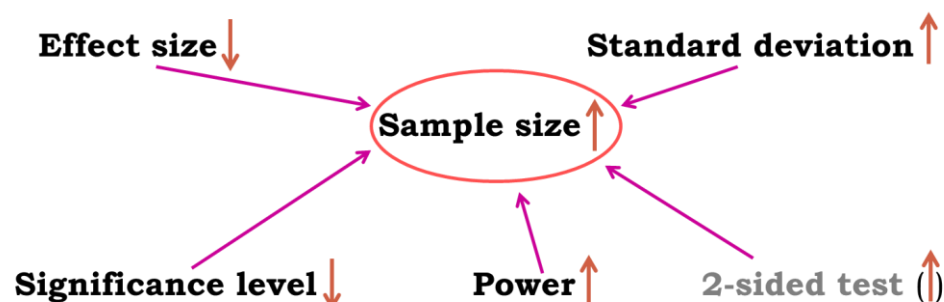
Now, going back to the effect size, there are actually 2 different ones: the absolute one which is basically the difference between the mean value of, say, our control group and the one treatment group, and the relative one, also referred to as Cohen's d. This one is the more useful and more widely used one as it accounts for the variability in our data.

$$\text{Cohen's } d = \frac{\text{Mean 1} - \text{Mean 2}}{\text{Pooled SD}}$$

The **significance level** refers to the famous p-value which, for a test to be 'significant', should be below 0.05 (the 5% threshold). Going back to our experiment about finding out more on gene A, we would define the p-value as the probability that a difference as big as the one observed between the WT and the KO could be found even if the knock-out of gene A does not affect the mouse's behaviour. Basically it means that if we find a significant difference ( $p < 0.05$ ) between our 2 groups of mice, corresponding to the effect size of biological interest, there is less than 5% chance that we would have been able to observe it if the knocking out of gene A did not affect the behaviour of the mouse.

The last variable is the so-called **alternative hypothesis**: one or two-sided test. This refers to the distribution of our variable: are we going to look at one side or at the 2 sides of it. In other words, and again going back to our experiment, do we want to answer the question: does it take longer for KO mice to achieve the task or do we simply want to know if there is a difference at all. Most of the time, in bench science, we go for the second question, even though we might think of one direction more than the other. We don't have enough evidence to 'choose' to look only at one side of the distribution. It is pretty much only in clinical trials that people go for one-sided tests. They have already tested a particular drug, for example, in many systems/species so they have plenty of evidence about the effect of the drug. Finally, it is 2 times easier to reach significance with a one-side test than a 2-side one so a reviewer will always be suspicious if we go for the first one and if they ask for justification, we'd better have one!

The basic idea behind the power analysis is that if we fix any five of the variables, a mathematical relationship can be used to estimate the sixth. So going back one more time to our example, running the power analysis, our question can be: What **sample size** do I need to have an 80% probability (**power**) to detect an average 5 minutes difference (**effect size** and **standard deviation**) at a 5% **significance level** using a **2-sided test**? The variables are all linked and will vary as shown in the following diagram.



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. Mostly, we need to have some idea of the difference we are expecting to see or that would make sense, together with some information on the standard deviation. We will use G\*Power as we go through the statistical tests.

About power analysis, the important message is: after we have designed our **experiment**, run a **power analysis** to estimate the appropriate **sample size** that will allow us to do **good and ethical science**.



## Chapter 2: Basic structure of a GraphPad Prism project

Click on the GraphPad Prism icon and the window below will appear. Before we do anything with GraphPad we need to have in mind the type of graph/analysis we want to do as this will determine the type of table we are going to choose.

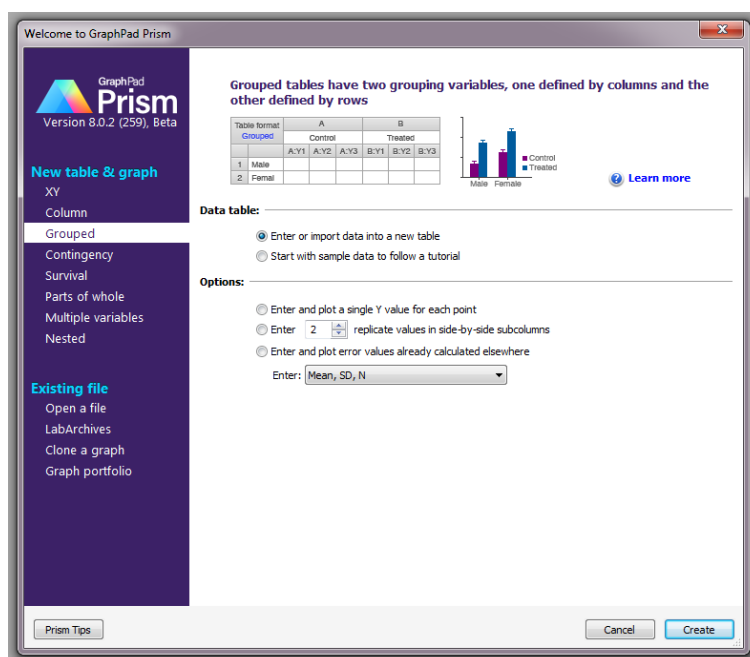
Then there are 2 scenarios:

- Enter our data directly into GraphPad. In which case, depending on the type of table we are choosing, we may need to know exactly how many data points we are going to deal with.
- Our data are already in Excel. In which case, it seems that we cannot import from its latest version. Even with the previous one it is not easy and it does not work for Mac. So whenever possible, as the Prism Help suggests, transfer data from Excel using copy and paste.

As mentioned previously, unlike other software, we need to choose a type of table before doing anything else which will be dependent upon the type of graph/analysis we want to do. Unlike in Excel, for instance, the 'worksheets' don't all have the same structure. We can choose from 5 different types:

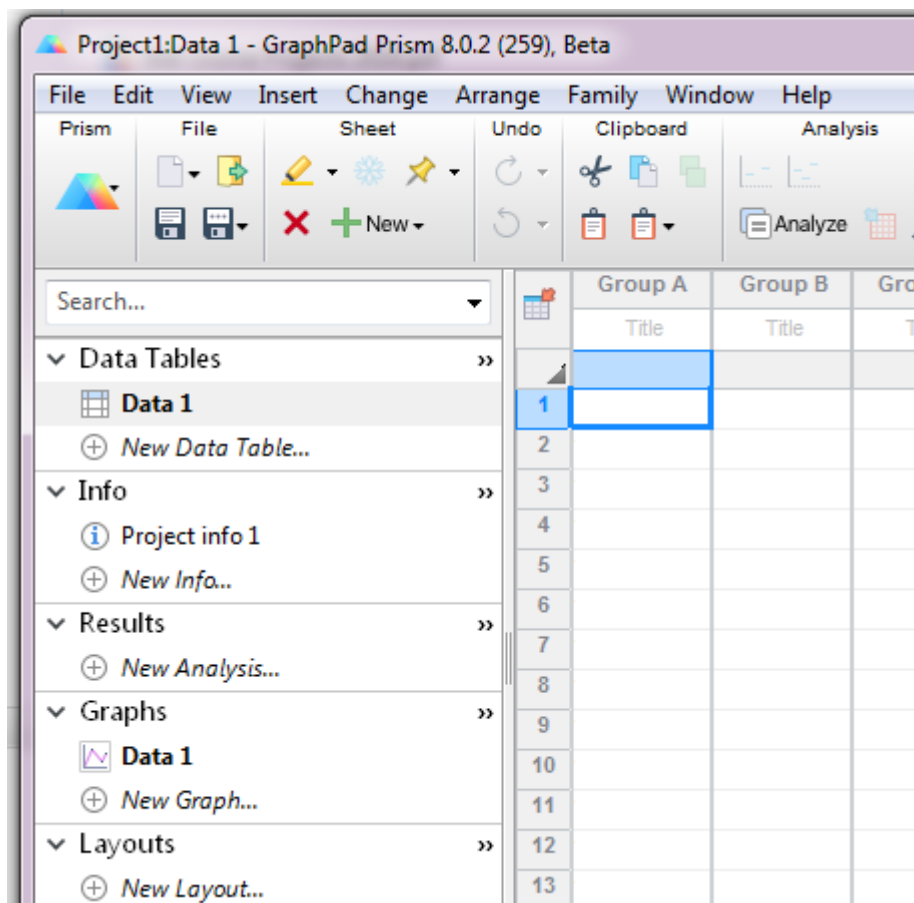
- **XY table** in which each point is defined by both an X and a Y value, though for one X we can have several Y like replicates which will be used to calculate error bars. Replicates are in side-by-side sub columns. This type of table allows us to run linear regression, correlation and to calculate area under the curve.
- **Column table** in which each column defines a treatment group. From this type of table, one can run a t-test and a one-way ANOVA or one of the non-parametric equivalent tests.
- **Grouped table** in which we can have 2 grouping variables, hence running 2-way ANOVAs.
- **Contingency table** in which one can enter categorical data suitable for Fisher's exact test or Chi-square.
- **Survival table** for ... survival analysis!

In this manual we will cover only XY, column and contingency tables.



Whatever type of tables we have chosen, each 'Project' contains the 5 folders:

- **Data Tables** in which are the worksheets containing the data,
- **Info** section in which we can enter information about the technical aspect of the experiment like the protocol or who was the experimenter,
- **Results** in which are the outputs of the statistical analysis
- **Graphs** in which are ... the graphs! They are usually automatically generated from our data but we can make them pretty afterwards. Graphical representation is actually one of the strengths of GraphPad as it is very easy and intuitive to plot data; and we can have fun with colours, as the graphs in the manual will show!
- **Layouts** in which we can present our graphs and analysis.



## Chapter 3: Qualitative data

Let's talk about the important stuff: our data. The first thing we need in order to do good stats is to know our data inside out. They are generally organised into variables, which can be divided into 2 categories: *qualitative* and *quantitative*.

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 variable 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 Test family.** We are going for the Fisher's exact test, we should go for 'Exact'.

**-Step 2: the Statistical Test:** we are looking at proportions and we want to compare 2 independent groups.

**-Step 3: the Type of Power Analysis:** we know our significant threshold ( $\alpha=0.05$ ), 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 **Input Parameters**. 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.

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

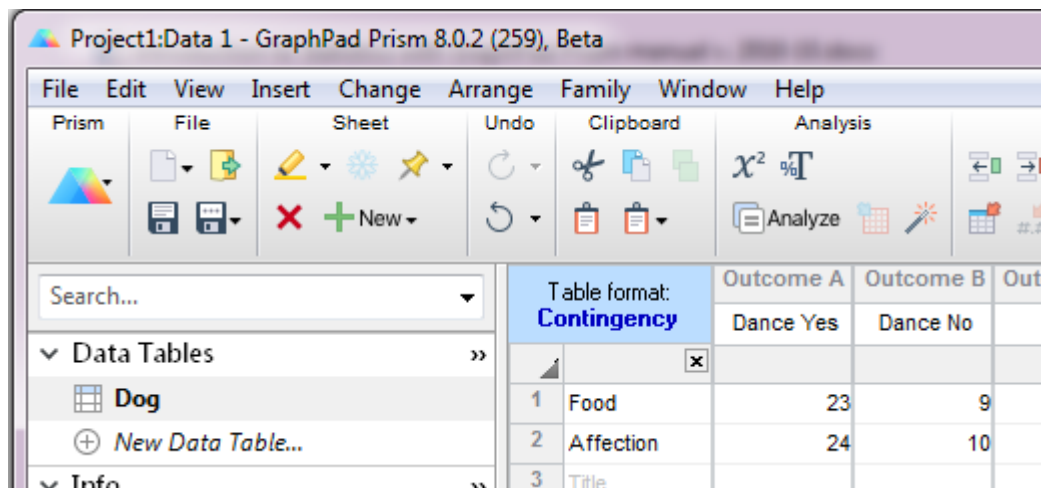
Cat

Count

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

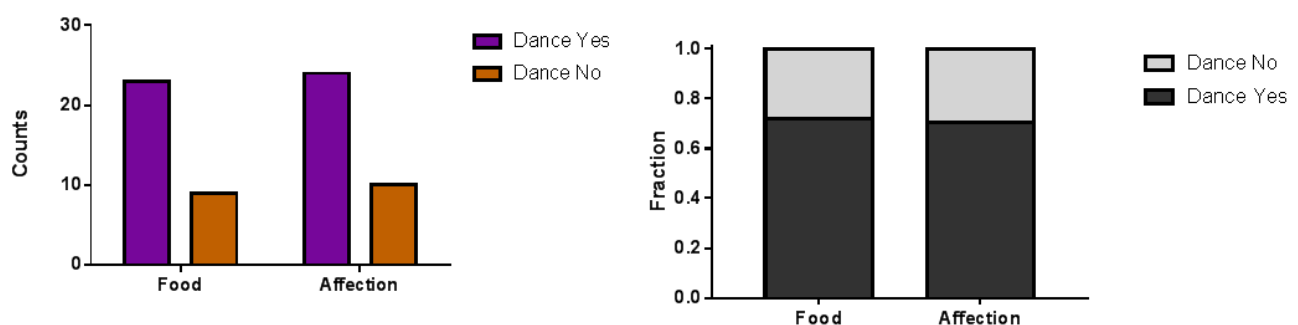
Dog

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.



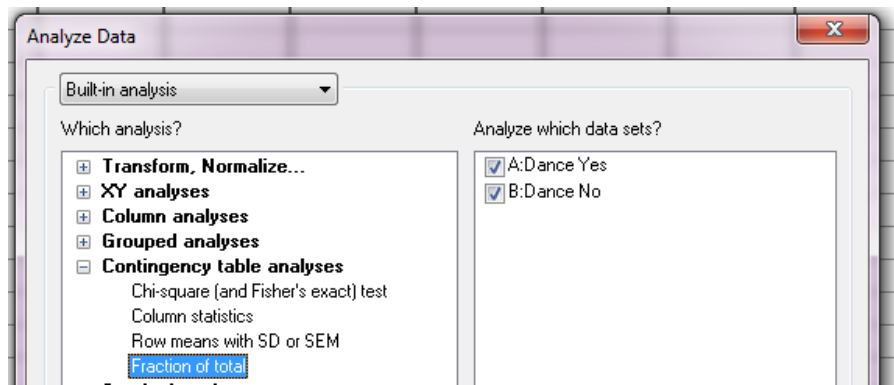
The first thing we want to do is to look at a graphical representation of the data. GraphPad will have prepared it for us and if we go into 'Graphs' we will see the results.

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.



As mentioned before, to analyse such data we need to use a Fisher's exact test but we could also use a  $\chi^2$  test. 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  $\chi^2$  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  $\chi^2$  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 variable's 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  $\chi^2$  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  $\chi^2$  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  $\chi^2$  tells us whether or not there is an association between 2 categorical variables.

An important thing to know about the  $\chi^2$ , 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:

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

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**

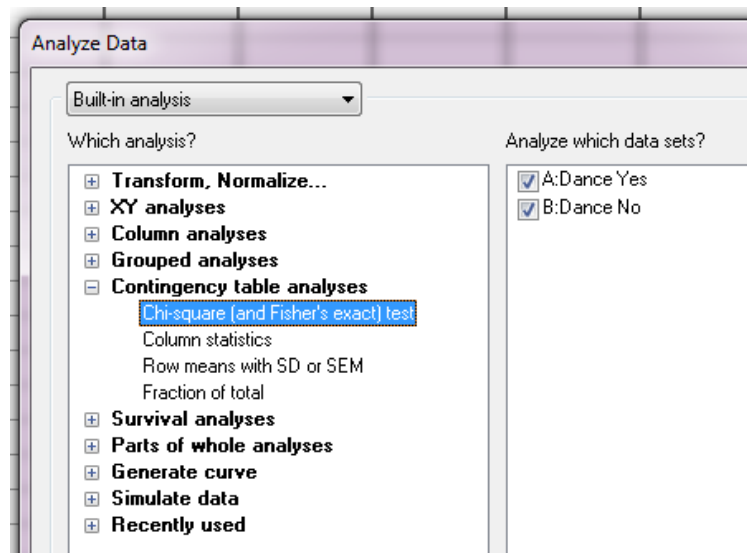
Animal				Type of Training		Total
				Food as Reward	Affection as Reward	
Cat	Did they dance?	Yes	Count	26	6	32
			Expected Count	15.1	16.9	32.0
		No	Count	6	30	36
			Expected Count	16.9	19.1	36.0
	Total		Count	32	36	68
			Expected Count	32.0	36.0	68.0
Dog	Did they dance?	Yes	Count	23	24	47
			Expected 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  $\chi^2$  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.



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  $\chi^2$  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  $\chi^2$ : 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  $\chi^2$ '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  $\chi^2$ , the bigger the difference between observed and expected frequencies and the more likely the difference is to be significant.

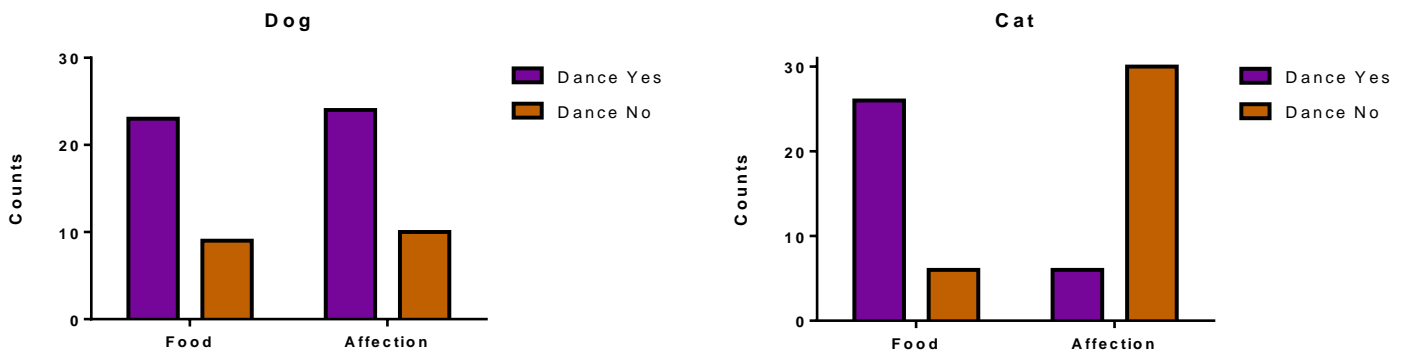
As we can see here the p-values vary slightly between the 2 tests (>0.99 vs.0.9081) though the conclusion remains the same: the type of reward has no effect whatsoever on the ability of dogs to line dance. Though the samples are not very big here, the assumptions for the  $\chi^2$  are met so we can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

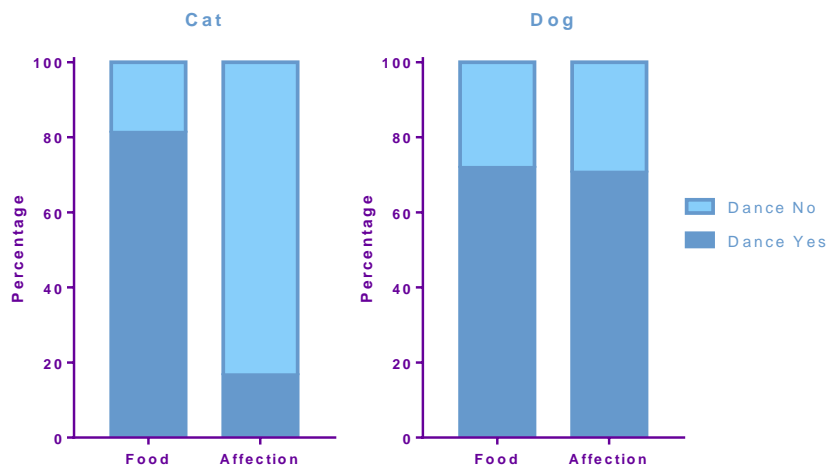
1	Table Analyzed	Cat
2		
3	Fisher's exact test	
4		
5	P value	< 0.0001
6	P value summary	***
7	One- or two-sided	Two-sided
8	Statistically significant? (alpha<0.05)	Yes

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.



### 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 is said to be 'significant' if the probability of type I error is:  $\alpha \leq 0.05$  (max  $\alpha=1$ ). It means that the level of uncertainty of a test of 5% is usually accepted. 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% sure 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 asks a few questions, such as:

- What does the literature say 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.

## Chapter 4: Quantitative data

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

### 4-1 Descriptive stats

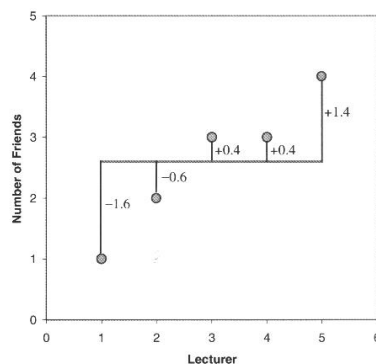
**The Mean** (or average)  $\mu$  = average of all values in a column

It can be considered as a model because it summaries the data.

Example: number of friends of each members 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, 2, 4 and 5 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).



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

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 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 Variance

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

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

Why N-1 instead N?

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

Example: mean of the sample is 10

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

- If the values are 9, 8, 11 and 12 (mean = 10) and if we change 3 of these values to 7, 15 and 8 then the final value must be 10 to keep the mean constant.
- If we hold 1 parameter constant, we have to use N-1 instead of N.
- It is the idea behind the *degree of freedom*: one less than the sample size.

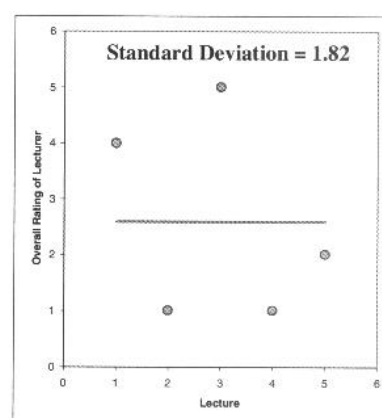
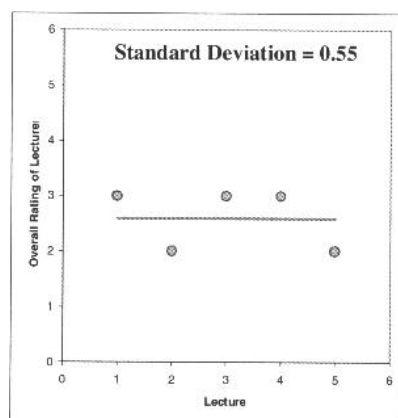
## The Standard Deviation (SD)

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

- S.D. =  $\sqrt{SS/N-1} = \sqrt{S^2}$ , in our example: S.D. =  $\sqrt{1.3} = 1.14$
- So we would present our mean as follows:  $\mu = 2.6 \pm 1.14$  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)

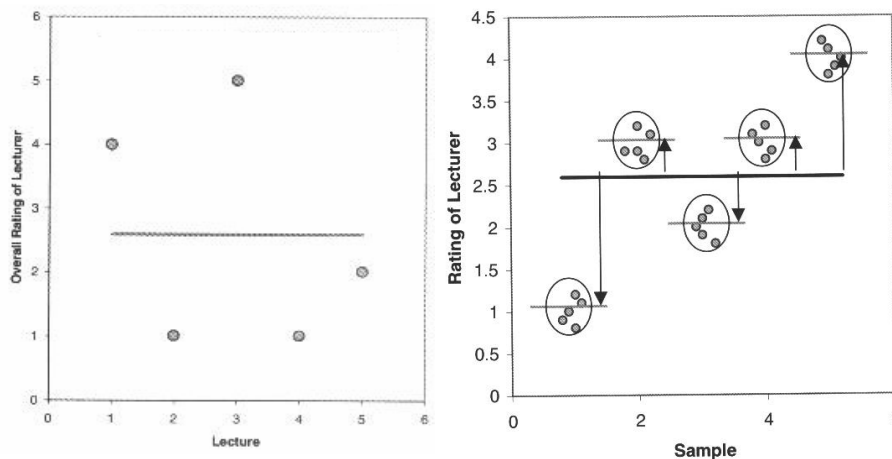


## Standard Deviation vs. Standard Error

Many scientists are confused about the difference between the standard deviation (S.D.) and the **standard error of the mean** (S.E.M. =  $S.D. / \sqrt{N}$ ).

- **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 no 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 our readers a sense of how well we have determined the mean.

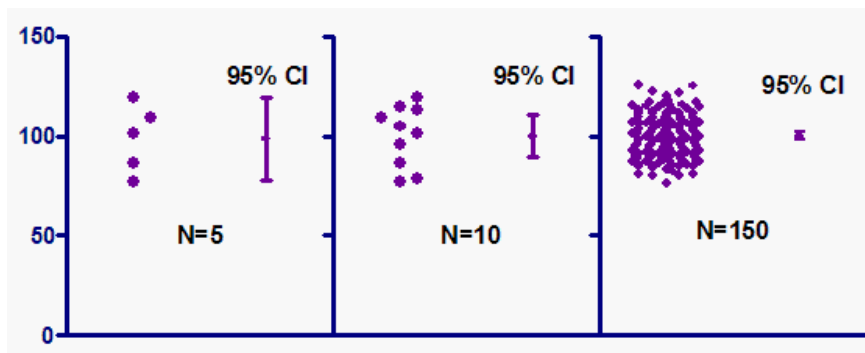
Choosing between SD and SEM also depends on what we want to show. If we just want to present our data on a descriptive purpose, then we go for the SD. If we want the reader to be able to infer an idea of significance, then we should go for the SEM or the Confidence Interval (see below). We will go a bit more in details 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  $\pm 1.96 \times \text{SE}$

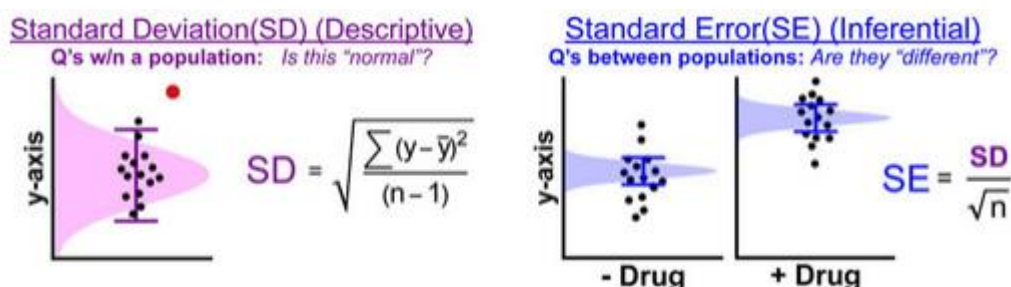
One other way to look at error bars:



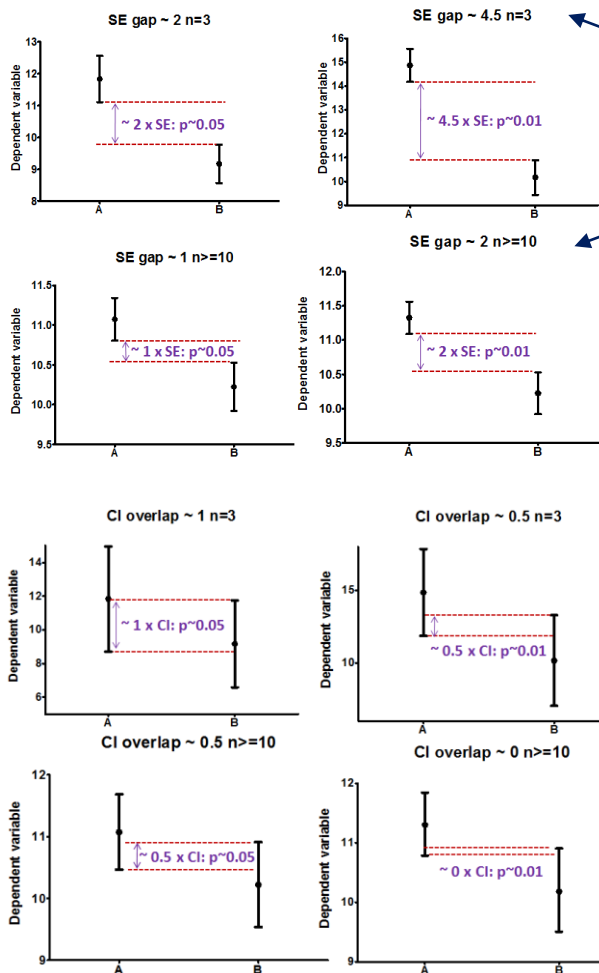
Error bars	Type	Description
Standard deviation (SD)	Descriptive	Typical or average difference between the data points and their mean.
Standard error (SEM)	Inferential	A measure of how variable the mean will be, if we repeat the whole study many times.
Confidence interval (CI), usually 95% CI	Inferential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



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

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

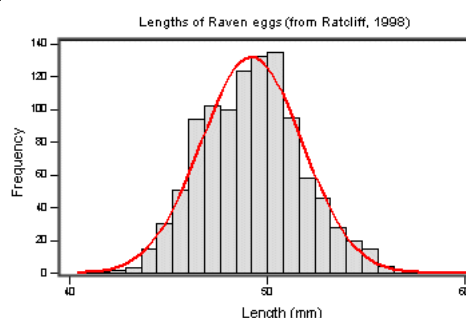
## 4-2 Assumptions of parametric data

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 entered 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).

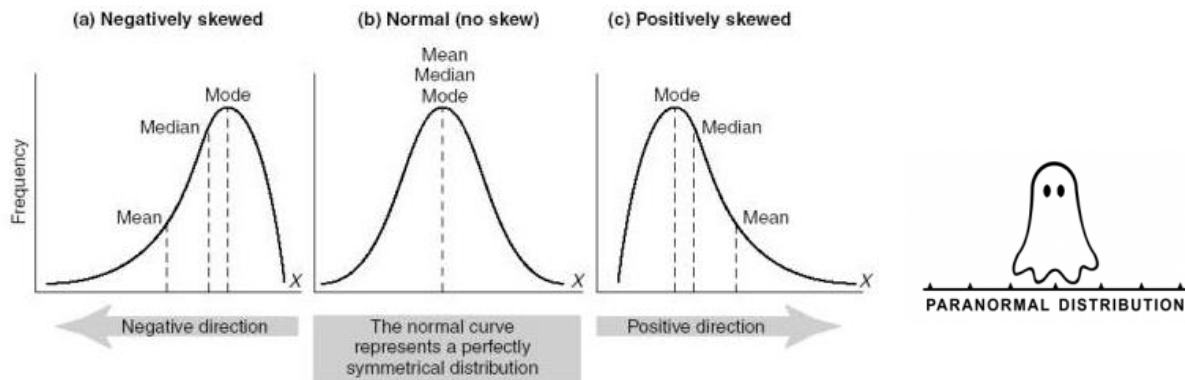
Example of normally distributed data:





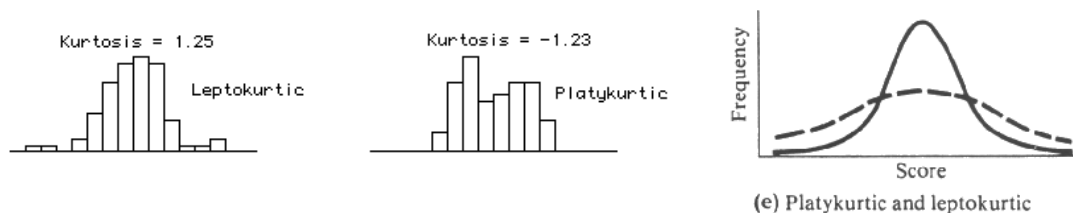
There are 2 main types of departure from normality:

- **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.



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

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

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

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

Let's try it through an example.

## Example (File: coyote.xlsx)

We want to know if there is a difference of biological interest between male and female coyotes. Of course, before doing anything else, we design our experiment, keeping in mind that, to compare 2 samples, we need to apply a t-test (we will explain this test a bit later). So, basically we are going to catch coyotes and hopefully we will manage to catch males and females. When our samples are big enough, we compare them. Now, the tricky bit here is the 'big enough' bit.



## 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 in the one we intend to run, male coyotes were found to measure: 92cm $\pm$ 7cm (SD). We expect a 5% difference between genders with a similar variability in the female sample.

**G\*Power 3.1.3**

File Edit View Tests Calculator Help

Central and noncentral distributions Protocol of power analyses

[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
- $\alpha$  err prob = 0.05
- Power (1- $\beta$  err prob) = 0.80
- Allocation ratio N2/N1 = 1

**Output:**

- Noncentrality parameter  $\delta$  = 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  $\alpha$ , power, and effect size

**Input Parameters**

Tail(s): Two

Determine => Effect size d: 0.6571429

$\alpha$  err prob: 0.05

Power (1- $\beta$  err prob): 0.80

Allocation ratio N2/N1: 1

**Output Parameters**

Noncentrality parameter  $\delta$ : 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

X-Y plot for a range of values Calculate

**Calculation Parameters**

☐ n1 != n2

Mean group 1: 0

Mean group 2: 1

SD  $\sigma$  within each group: 0.5

☒ n1 = n2

Mean group 1: 92

Mean group 2: 87.4

SD  $\sigma$  group 1: 7

SD  $\sigma$  group 2: 7

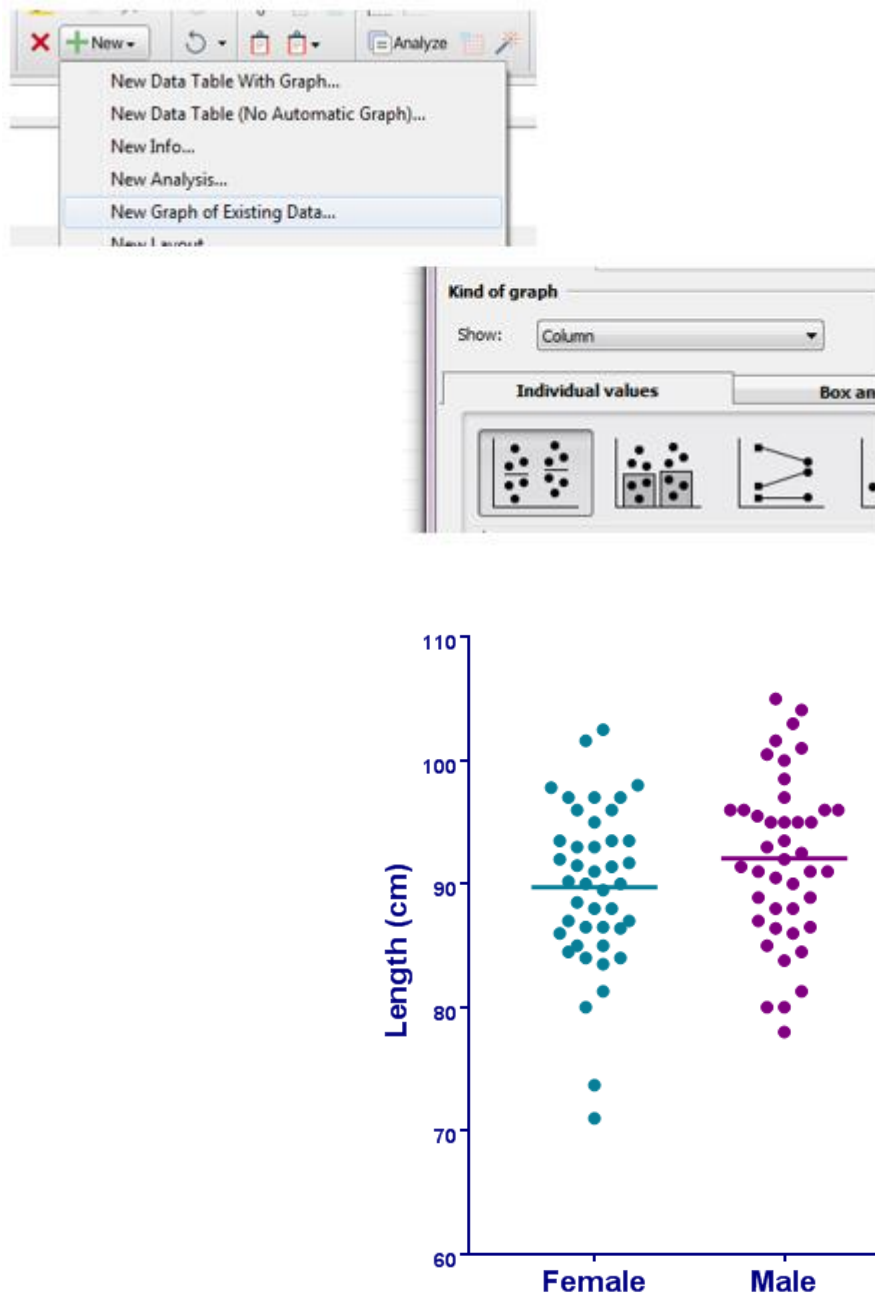
Calculate Effect size d: 0.6571429

Calculate and transfer to main window

Close

We need a sample size of  $n=76$  ( $2 \times 38$ ). Once the data are collected, we need to check that our data meet the assumptions for parametric tests. This should be done first through data exploration and one of the best way to explore data is to plot them as a scatterplot.

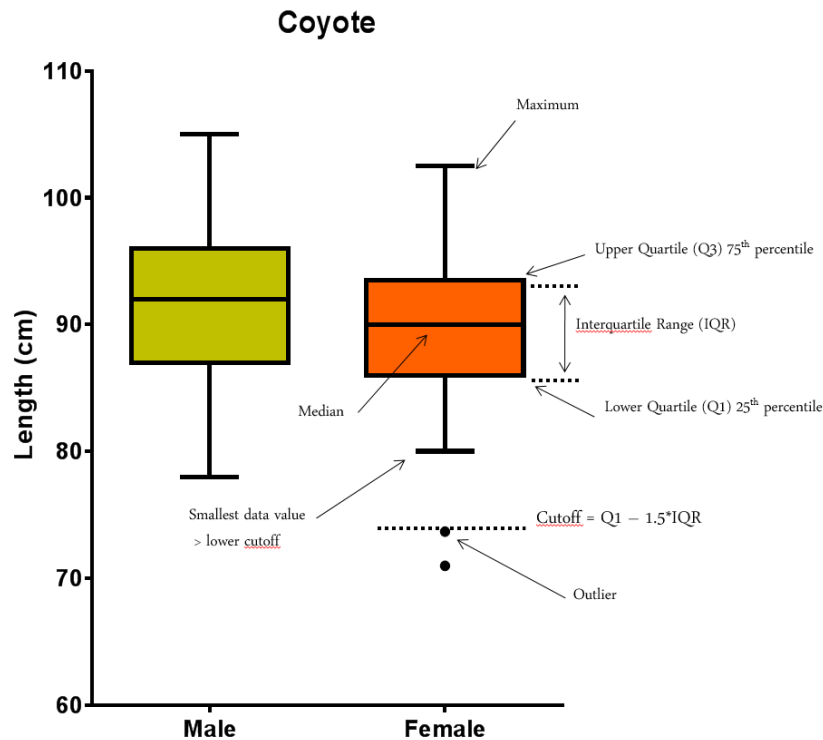
In Graphpad Prism:



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.

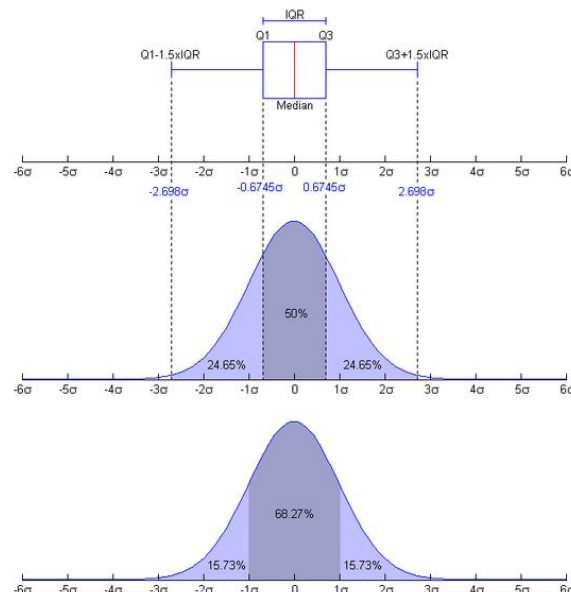
However, we can notice that 2 dots are bit out of range: 2 females seem smaller than their peers. Now the question is: are they smaller, as in just the smallest of the group, or smaller as in outliers. To find out, we need to plot the data in another way: the boxplot.

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!



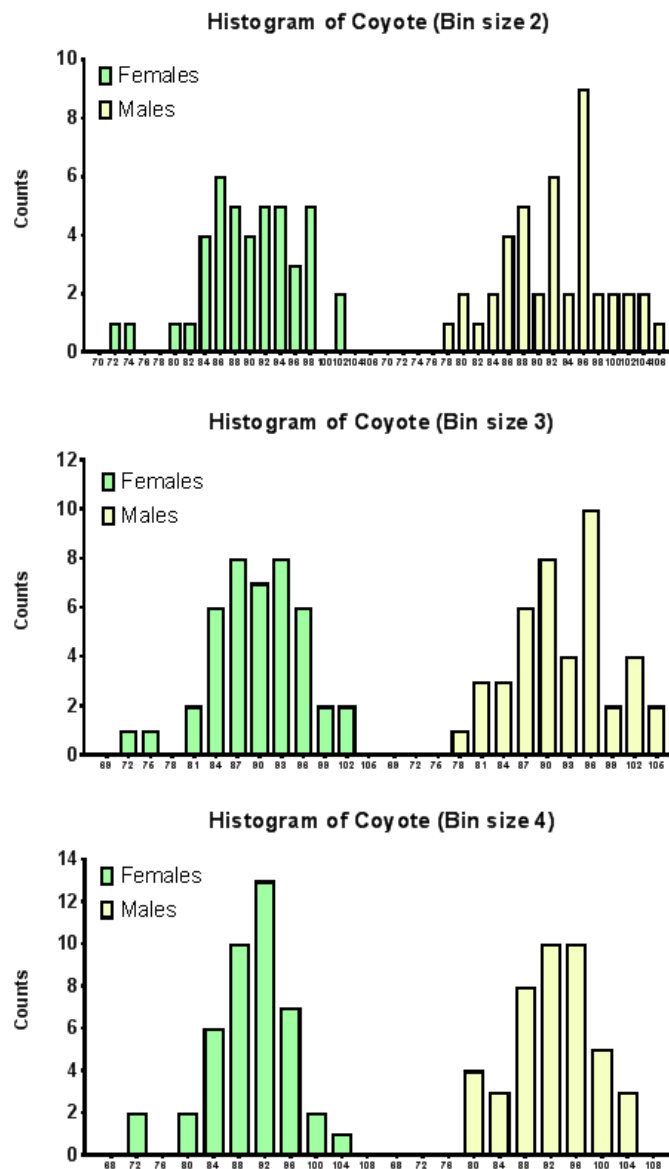
One other way to explore data is with a histogram. I think it works best with big samples, but it is still useful here so let's do it.

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.



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 the said exploration is a bit ambiguous.

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

We are given the choice between 3 tests for normality: D'Agostino and Pearson, Kolmogorov-Smirnov and Shapiro-Wilk. These tests require  $n \geq 7$  and the D'Agostino and Pearson test is the one to go for. 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.

Col. stats		A	B
		Female	Male
		Y	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	K-S normality test		
17	K-S 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 test		
23	K2	1.285	0.3668
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			

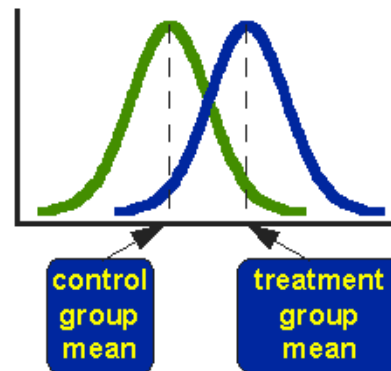
Actually, the test does not tell us that our data are normally distributed, it tells us that they are not significantly different from normality (♀  $p=0.1223$  and ♂  $p=0.7757$ ).

In GraphPad, the second assumption 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. Don't be too quick to switch to nonparametric tests. While they do not assume Gaussian distributions, these tests do assume that the shape of the data distribution is the same in each group. So, if our groups have very different standard deviations and so are not appropriate for a one-way ANOVA for instance, they should not be analysed by a Kruskal-Wallis (non-parametric equivalent of the ANOVA) either. However, ANOVA and t-tests are rather robust, especially when the samples are not too small so we can get away with small departure from normality and small differences in variances.

Often the best approach is to transform the data and transforming to logarithms or reciprocals does the trick, restoring equal variance.

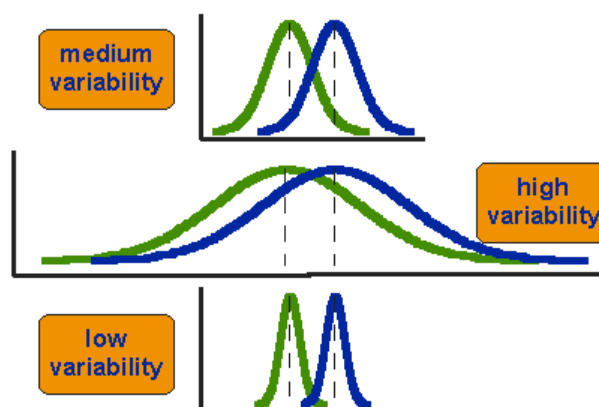
### 4-3 The t-test

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



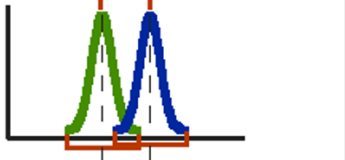
The figure above shows the distributions for the treated (blue) and control (green) groups in a study. Actually, the figure shows the idealised distribution. The figure indicates where the control and treatment group means are located. The question the t-test addresses is whether the means are statistically 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 about the three situations 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. Figure 3 shows the formula for the t-test and how the numerator and denominator are related to the distributions.

$$\begin{aligned}
 \frac{\text{signal}}{\text{noise}} &= \frac{\text{difference between group means}}{\text{variability of groups}} \\
 &= \frac{\bar{X}_T - \bar{X}_C}{\sqrt{\frac{\text{var}_T}{n_T} + \frac{\text{var}_C}{n_C}}} \\
 &= \text{t-value}
 \end{aligned}$$


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

To run a t-test in GraphPad, we go: **=Analysis> Column analyses>t-tests** and then we have to choose between 2 types of t-tests: **Unpaired and Paired t-test**.

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 apply a Paired-Samples t-test as we are interested in each individual difference between day 1 and day 30.

## 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 a Unpaired t-test.



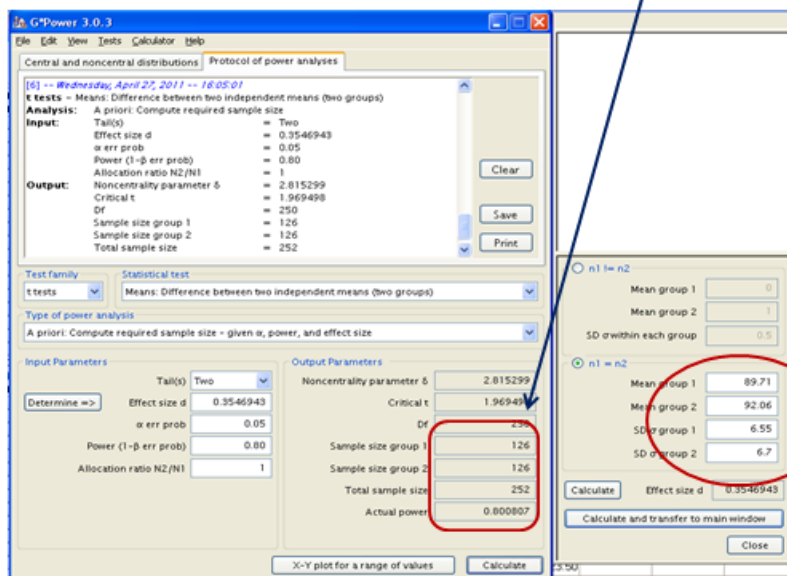
1	Table Analyzed	Coyote
2	Column A	Female
3	vs	vs
4	Column B	Male
5		
6	Unpaired t test	
7	P value	0.1045
8	P value summary	ns
9	Are means signif. different? (P < 0.05)	No
10	One- or two-tailed P value?	Two-tailed
11	t, df	t=1.641 df=84
12		
13	How big is the difference?	
14	Mean $\pm$ SEM of column A	89.71 $\pm$ 0.9988 N=43
15	Mean $\pm$ SEM of column B	92.06 $\pm$ 1.021 N=43
16	Difference between means	-2.344 $\pm$ 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.

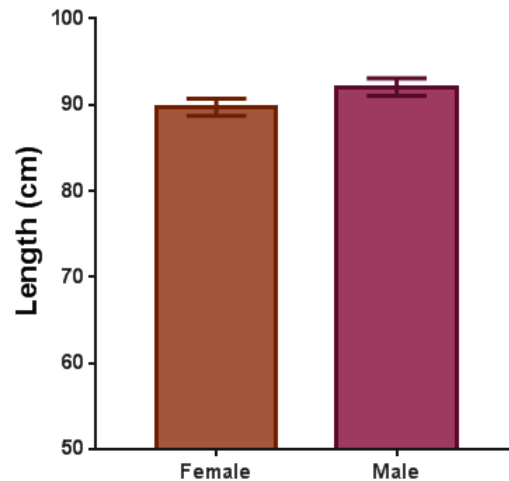
So, despite having collected the 'recommended' sample size, we did not reach significance. This is because the difference observed in the collected sample is smaller than expected. If we now consider the data as a pilot study and run the power analysis again, we would need a sample 3 times bigger to reach a power of 80%. Now is the time to wonder whether a 2.3cm (<3%) is biologically relevant.

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



	Female	Male
1	Y	Y
2		
3		
4		
5		
6		
7		
8		
9		
10		
11		
12		
13		
14		
15		
16		
17		
18		
19		
20		
21		
22		
23		
24		
25		
26		
27		
28		
29		
30		
31		
32		
33		
34		
35		

Finally, even though I much prefer to plot data such as the coyote ones as scatterplots, barcharts are a classic way to present results.

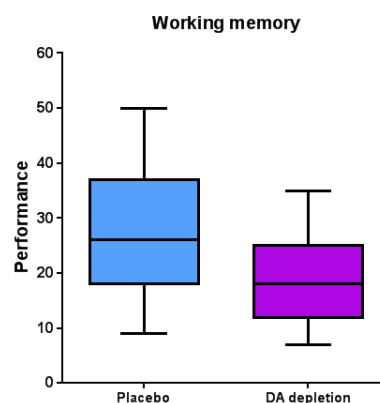
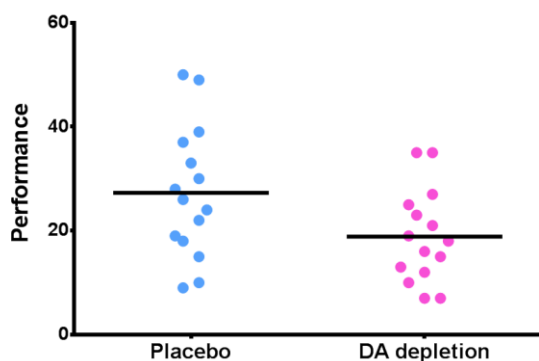


## Paired t-test

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 below seem to indicate that there is no significant departure from normality and this is confirmed by the D'Agostino & Pearson test.

Col. stats		Placebo	DA depletion
		Y	Y
1	Number of values	15	15
2			
3	Minimum	9.000	7.000
4	25% Percentile	18.00	12.00
5	Median	26.00	18.00
6	75% Percentile	37.00	25.00
7	Maximum	50.00	35.00
8			
9	Mean	27.27	18.87
10	Std. Deviation	12.65	8.911
11	Std. Error of Mean	3.265	2.301
12			
13	Lower 95% CI of mean	20.26	13.93
14	Upper 95% CI of mean	34.27	23.80
15			
16	D'Agostino & Pearson omnibus normality test		
17	K2	0.6754	0.9815
18	P value	0.7134	0.6122
19	Passed normality test (alpha=0.05)?	Yes	Yes
20	P value summary	ns	ns
21			
22	Sum	409.0	283.0

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	0.8415

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.

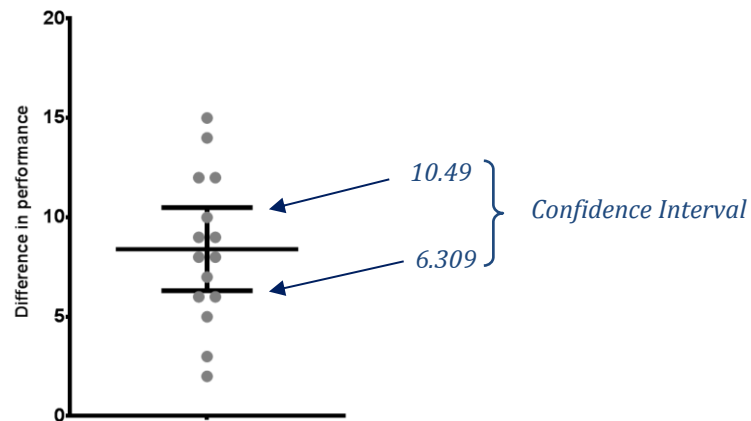
The paired t-test turns out to be highly significant (see Table above). So, how come the graph and the test tell us different things?

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.



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
0	Std. Deviation	3.776
1	Std. Error of Mean	0.9749
2		
3	Lower 95% CI of mean	6.309
4	Upper 95% CI of mean	10.49
5		
6	One sample t test	
7	Theoretical mean	0.0
8	Actual mean	8.400
9	Discrepancy	-8.400
0	95% CI of discrepancy	6.309 to 10.49
1	t, df	t=8.616 df=14
2	P value (two tailed)	< 0.0001
3	Significant (alpha=0.05)?	Yes
4		
5	Sum	126.0
6		

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 Comparison of more than 2 means: 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.

Example: if we want to compare 3 groups (1, 2 and 3) and we carry out 3 t-tests (groups 1-2, 1-3 and 2-3), each with an arbitrary 5% level of significance, the probability of not making the type I error is 95% (= 1 - 0.05). The 3 tests being independent, we can multiply the probabilities, 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)**, which 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 us 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 would be 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 mean 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 our data into 5 groups and we run an ANOVA.

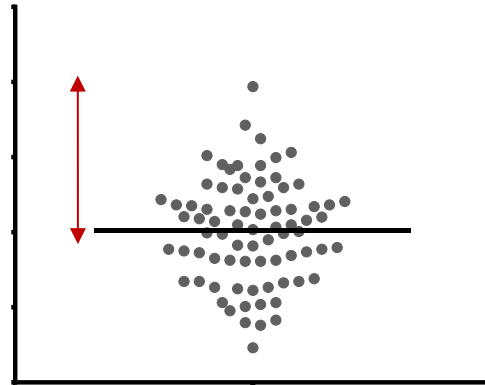
Below is a typical example of an analysis of variance table.

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:

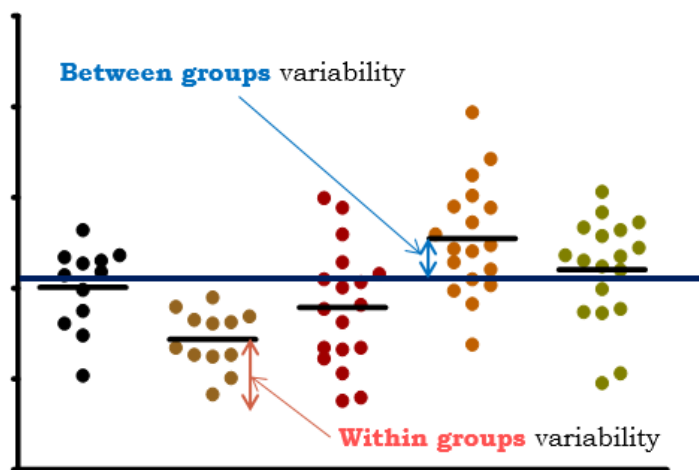
$$\text{Total sum of squares} = \sum (x_i - \text{Grand mean})^2$$

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



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

So, we work out the mean for each cell type and we work out the squared differences between each of the means and the grand mean ( $\sum n_i (\text{Mean}_i - \text{Grand mean})^2$ ). 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 we remember the formula of the variance ( $= \text{SS} / N-1$ , with  $\text{df}=N-1$ ), we can see that this value quantifies the variability between the groups' means: it is the between group 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:

$$\text{Within groups sum of squares} = \sum (x_i - \text{Mean}_i)^2 \text{ which in our case is equal to } 5.775.$$

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 our model has been removed.

In our example, the 5 groups sizes are 12, 12, 17, 17 and 17 so  $df = 5 \times (n - 1) = 73$

So the mean 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 our hypothesis. From this value can be obtained what is often referred to as the Pooled SD ( $=\text{SQRT}(\text{MS}(\text{Residual or Within Group}))$ ). 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).

So, 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 other ones. This is why we apply post-hoc tests. Post hoc tests could be compared to t-tests but with a more stringent approach, a lower significance threshold to correct for familywise error rate. We will go through post-hoc tests more in detail later.

### Example (File: `protein expression.xlsx`)

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

[1] -- Monday, November 26, 2012 -- 16:46:30

F tests -- ANOVA: Fixed effects, omnibus, one-way

Analysis: Sensitivity: Compute required effect size

Input:  $\alpha$  err prob = 0.05  
Power (1- $\beta$  err prob) = 0.80  
Total sample size = 75  
Number of groups = 5

Output: Noncentrality parameter  $\lambda$  = 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  $\alpha$ , power, and sample size

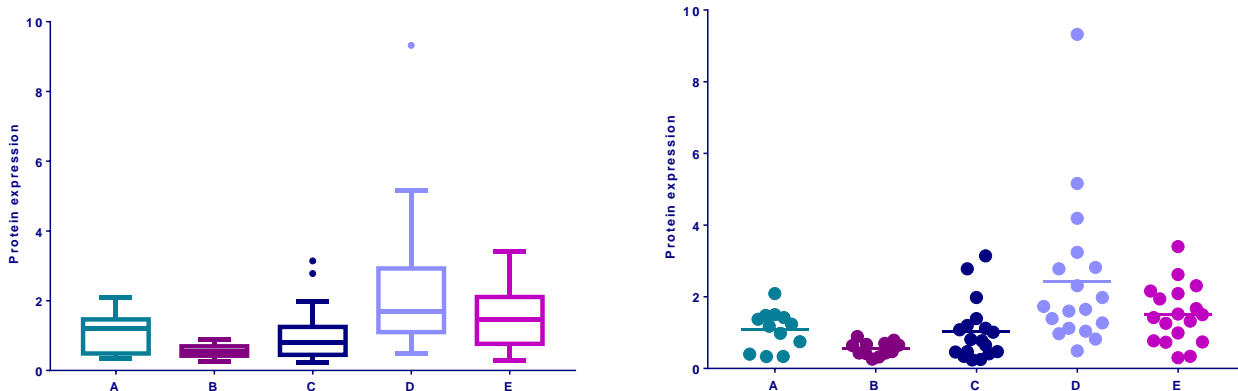
Input Parameters:  $\alpha$  err prob = 0.05  
Power (1- $\beta$  err prob) = 0.80  
Total sample size = 75  
Number of groups = 5

Output Parameters: Noncentrality parameter  $\lambda$  = 12.7693998  
Critical F = 2.5026565  
Numerator df = 4  
Denominator df = 70  
Effect size f = 0.4126241

X-Y plot for a range of values Calculate

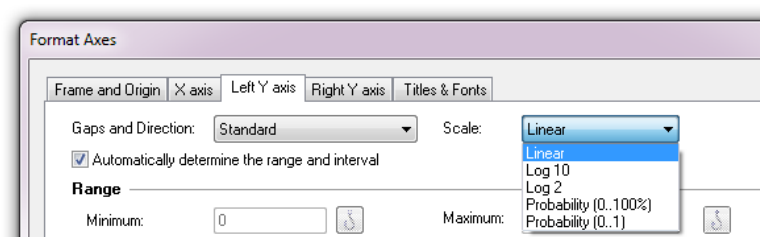
We are quite confident in our hypothesis so we decide to go ahead.

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).

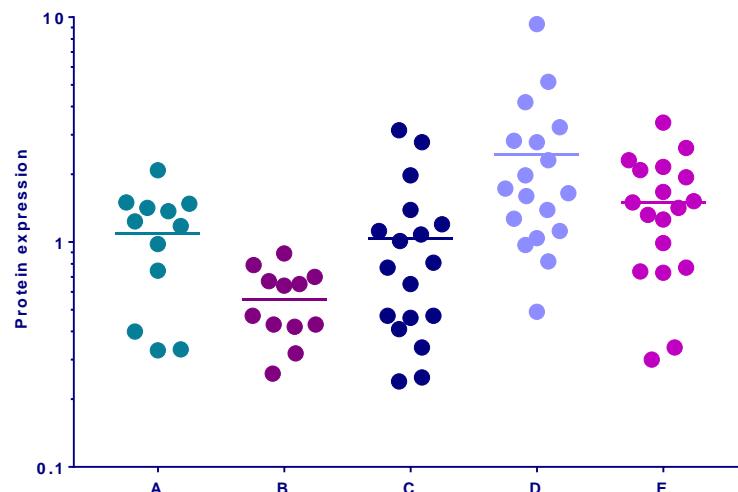


	A	B	C	D	E
	Y	Y	Y	Y	Y
1 Number of values	12	12	18	18	18
2					
3 Minimum	0.3300	0.2600	0.2400	0.4900	0.3000
4 25% Percentile	0.6675	0.4225	0.4475	1.100	0.7625
5 Median	1.205	0.5550	0.7900	1.690	1.460
6 75% Percentile	1.465	0.6925	1.248	2.925	2.108
7 Maximum	2.090	0.8900	3.140	9.320	3.400
8					
9 Mean	1.128	0.5558	1.032	2.438	1.504
10 Std. Deviation	0.4985	0.1947	0.8364	2.108	0.8179
11 Std. Error	0.1439	0.05620	0.1971	0.4968	0.1928
12					
13 Lower 95% CI of mean	0.8116	0.4321	0.6157	1.390	1.098
14 Upper 95% CI of mean	1.445	0.6795	1.448	3.486	1.911
15					
16 D'Agostino & Pearson omnibus normality test					
17 K2	0.05247	0.7508	9.375	22.59	1.280
18 P value	0.9741	0.6870	0.0092	< 0.0001	0.5274
19 Passed normality test (alpha=0.05)?	Yes	Yes	No	No	Yes
20 P value summary	ns	ns	**	***	ns
21					
22 Sum	13.54	6.670	18.57	43.88	27.08
23					

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.





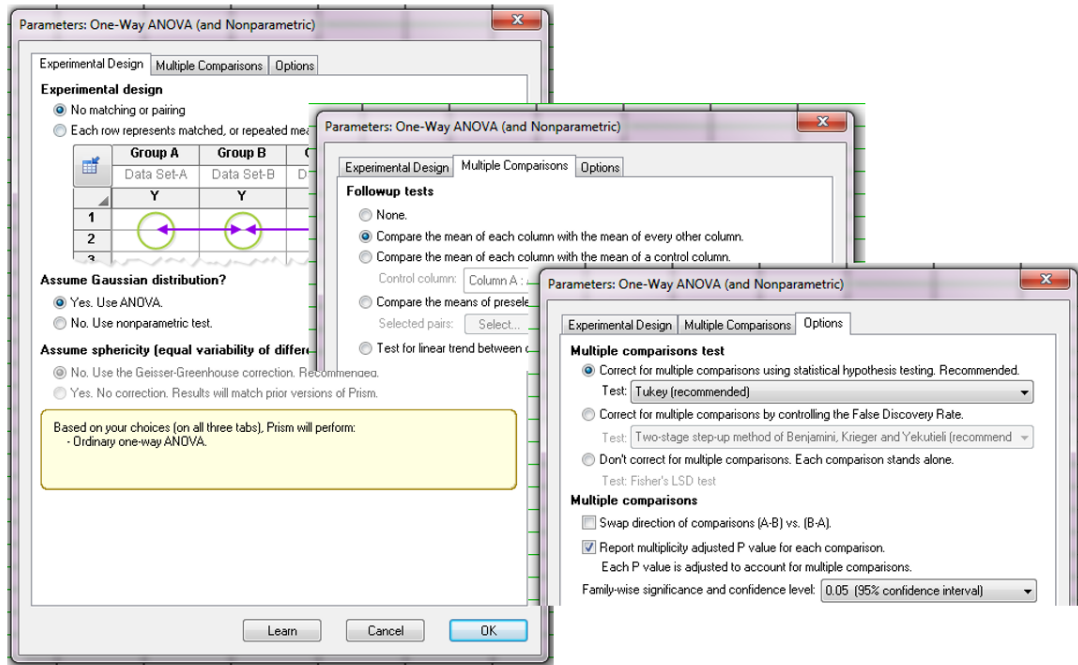


It looks much better. Now, the next step is to actually log-transform the data. To do so, we go to **Analyse > Transform > Transform** and we choose **Y=Log(Y)**, we can then re-run the analysis.

	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	0.4268	0.2447
15					
16 D'Agostino & Pearson omnibus normality test					
17 K2	2.198	0.6827	0.5884	0.8869	2.902
18 P value	0.3332	0.7108	0.7451	0.6416	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 is 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.



ANOVA					
1	Table Analyzed	*Transform of Protein expression			
2					
3	ANOVA summary				
4	F	0.6727			
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)	0.6727 (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	Number of families	1		
28	Number of treatments (columns)	Number of comparisons per family	10		
29	Number of values (total)	Alpha	0.05		
30					
31					

Post hoc tests					
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 vs. C	0.07521	-0.2247 to 0.3751	No	ns	0.9555
A vs. D	-0.3053	-0.6052 to -0.005359	Yes	*	0.0440
A vs. E	-0.1331	-0.4330 to 0.1669	No	ns	0.7275
B vs. C	-0.1753	-0.4752 to 0.1247	No	ns	0.4807
B vs. D	-0.5557	-0.8557 to -0.2558	Yes	****	< 0.0001
B vs. E	-0.3835	-0.6834 to -0.08360	Yes	**	0.0055
C vs. D	-0.3805	-0.6487 to -0.1122	Yes	**	0.0015
C vs. E	-0.2083	-0.4765 to 0.05998	No	ns	0.2021
D vs. E	0.1722	-0.09604 to 0.4405	No	ns	0.3839

## Analysis of variance Results

Homogeneity of variance ☒

$$F=0.6727/0.08278=8.13$$

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

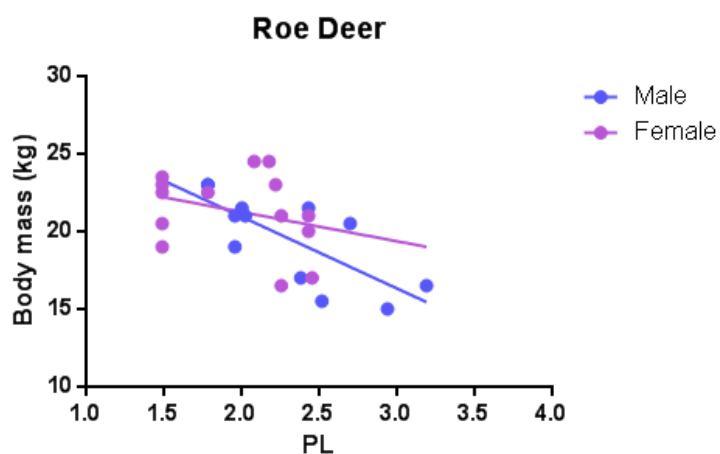
## 4-5 Correlation

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.

### A bit of theory: Correlation coefficient

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

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

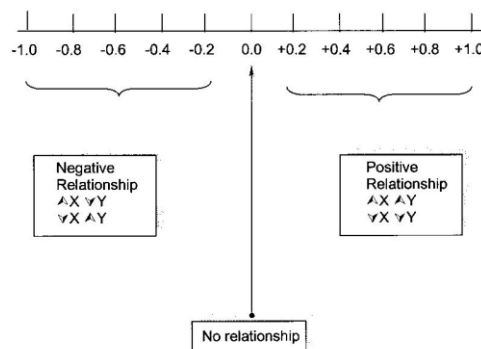
A positive covariance indicates that as one variable deviates from the mean, the other one deviates in the same direction, in other word if one variable goes up the other one goes up as well.

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”.

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

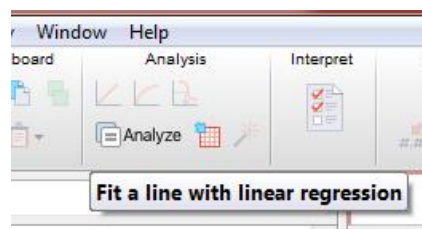
Of course, we don't need to remember that formula but it is important that we understand what the correlation coefficient does: it measures the magnitude and the direction of the relationship between two variables. It is designed to range in value between 0.0 and 1.0.



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	B
		Male	Female
1	<b>Best-fit values</b>		
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	<b>Std. Error</b>		
8	Slope	1.287	1.721
9	Y-intercept	3.025	3.453
10			
11	<b>95% Confidence Intervals</b>		
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	<b>Goodness of Fit</b>		
17	R square	0.5630	0.09119
18	Sy.x	1.980	2.512
19			
20	<b>Is slope significantly non-zero?</b>		
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	<b>Equation</b>	Y = -4.621*X + 30.20	Y = -1.888*X + 25.04
27			
28	<b>Data</b>		
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:  
 $\text{Body Mass} = 30.2 - 4.621 * \text{Parasite Burden}$ .  
 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 weighs 30.2 kg.

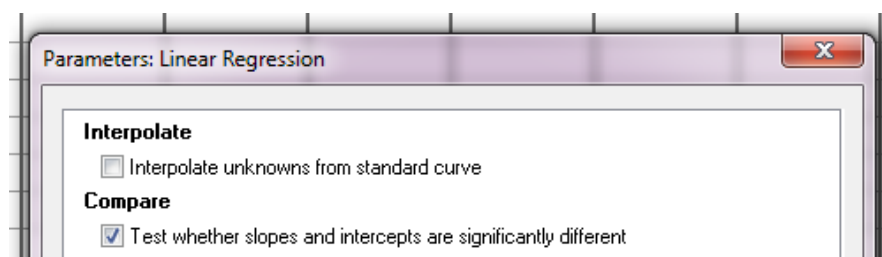
A coefficient of determination  $r^2 = 0.56$  means that 56% of the variability observed in the body mass can be explained only by the parasite burden.

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.



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

Central and noncentral distributions | Protocol of power analyses

[1] -- Thursday, December 06, 2012 -- 15:22:58

Exact -- Correlation: Bivariate normal model

Options: exact distribution

Analysis: A priori: Compute required sample size

Input: Tail(s) = Two  
Correlation p H1 = 0.3  
 $\alpha$  err prob = 0.05  
Power (1- $\beta$  err prob) = 0.80

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  $\alpha$ , power, and effect size

Input Parameters: Tail(s) = Two  
Determine => Correlation p H1 = 0.3  
 $\alpha$  err prob = 0.05  
Power (1- $\beta$  err prob) = 0.80  
Correlation p H0 = 0

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

Options | X-Y plot for a range of values | Calculate

Coefficient of determination  $p^2$  = 0.5  
Calculate | Correlation p H1 = ?  
Calculate and transfer to main window | Close

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

G\*Power 3.1.3

Central and noncentral distributions | Protocol of power analyses

[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 p H1 = 0.3000000  
 $\alpha$  err prob = 0.05  
Total sample size = 14  
Correlation p H0 = 0

Output: Lower critical r = -0.5324128  
Upper critical r = 0.5324128  
Power (1- $\beta$  err prob) = 0.1814126

Test family: Exact  
Statistical test: Correlation: Bivariate normal model

Type of power analysis: Post hoc: Compute achieved power - given  $\alpha$ , sample size, and effect size

Input Parameters: Tail(s) = Two  
Determine => Correlation p H1 = 0.3000000  
 $\alpha$  err prob = 0.05  
Total sample size = 14  
Correlation p H0 = 0

Output Parameters: Lower critical r = -0.5324128  
Upper critical r = 0.5324128  
Power (1- $\beta$  err prob) = 0.1814126

Options | X-Y plot for a range of values | Calculate

Coefficient of determination  $p^2$  = 0.09  
Calculate | Correlation p H1 = 0.3  
Calculate and transfer to main window | Close

### Post-hoc power analysis:

with a sample of 14 and quite a weak effect, you only achieved a 18% power.

## 4-6 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

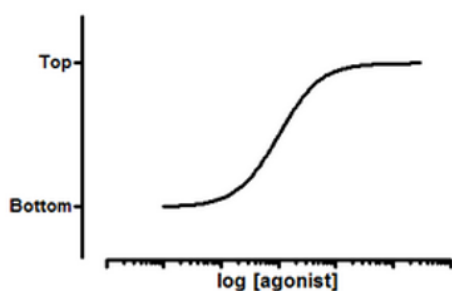
The agonist can be an inhibitor or a stimulator. The higher the concentration of the agonist, the stronger the response.

**IC50 (I=Inhibition):** concentration of an agonist that provokes a response half way between the maximal (Top) response and the maximally inhibited (Bottom) response.

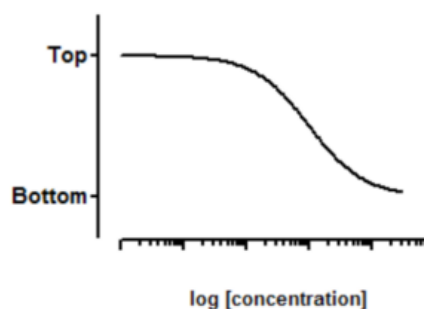
**EC50 (E=Effective):** concentration that gives half-maximal response

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:**  $Y = \text{Bottom} + (\text{Top} - \text{Bottom}) / (1 + 10^{-(\text{LogEC50} - X) \cdot \text{HillSlope}})$



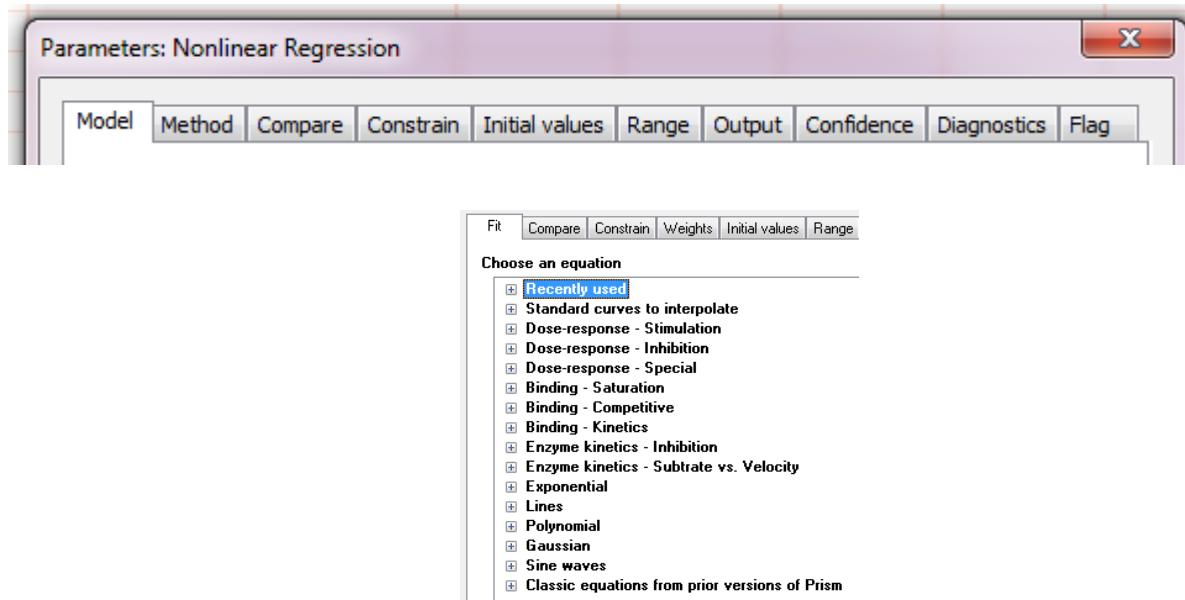
**Model:**  $Y = \text{Bottom} + (\text{Top} - \text{Bottom}) / (1 + 10^{-(X - \text{LogIC50})})$

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.



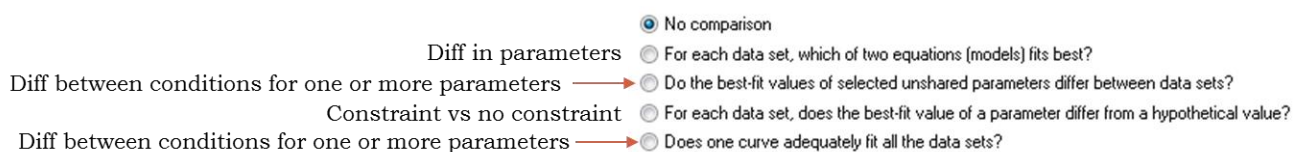
We can enter either biological or technical repeats but we will have to treat them differently during the analysis.

The sigmoid model assumes that the dose response curves 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.



- 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)

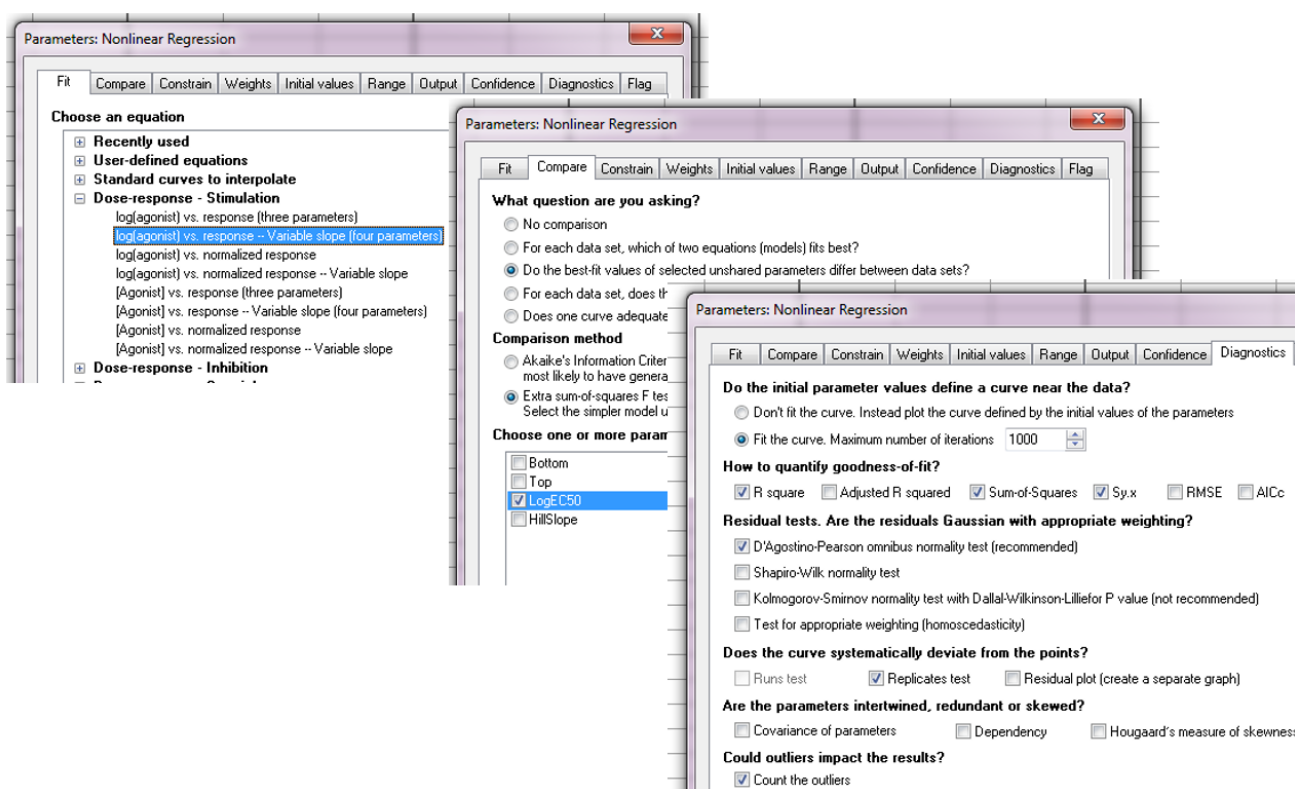


- 7- **Output:** summary table presents same results in a summarised way.
- 8- **Confidence:** calculate and plot confidence intervals
- 9- **Diagnostics:** check for normality (weights) and outliers (but keep them in the analysis). Also run the replicates test to see whether the curve gets too far from the points or not. Finally have a look at the residual plots.

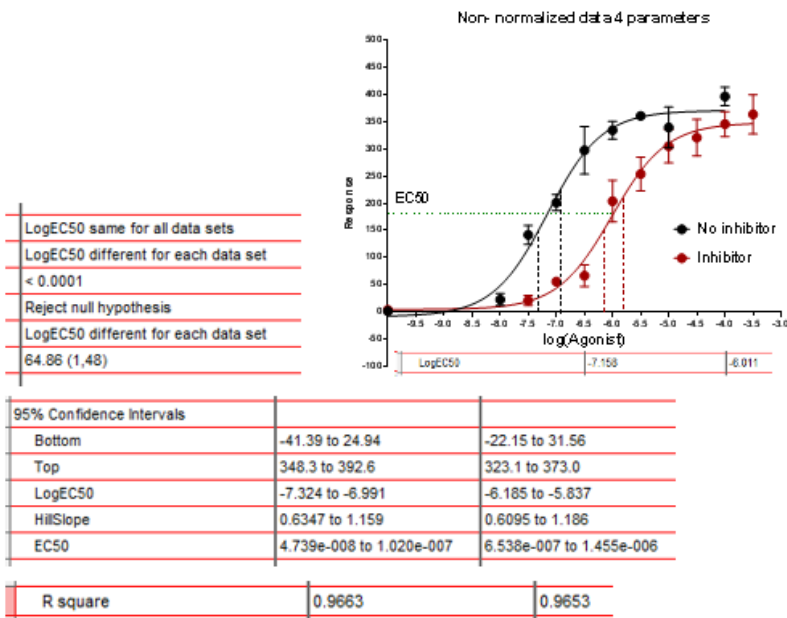
**Example** (File: `Inhibition data.xlsx`).

The Y values are the raw response to the agonist concentrations and the X ones 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)**.



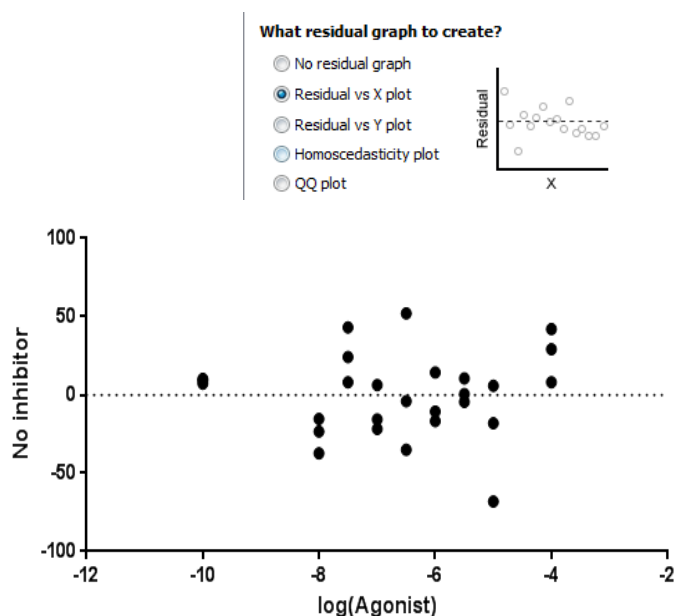
The results are shown below.



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.

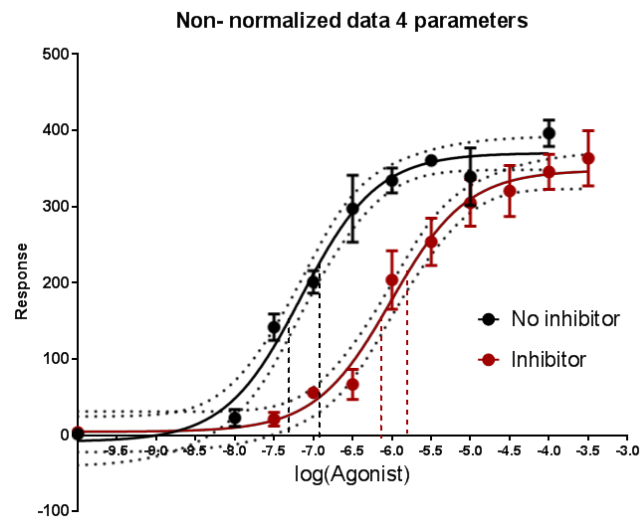
	No inhibitor	Inhibitor
Replicates test for lack of fit		
SD replicates	22.71	25.52
SD lack of fit	41.84	32.38
Discrepancy (F)	3.393	1.610
P value	0.0247	0.1989
Evidence of inadequate model?	Yes	No

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



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

Normality of Residuals		
D'Agostino & Pearson omnibus K2	0.08701	0.08356
P value	0.9574	0.9591
Passed normality test (alpha=0.05)?	Yes	Yes

Number of points		
Analyzed	27	29
Outliers (not excluded, Q=1.0%)	0	0

Goodness of Fit		
Degrees of Freedom	23	25
R square	0.9663	0.9653

So despite a slight departure from the assumptions, the curve looks OK overall and we should trust the results of the comparison that the IC50 are significant different from one another ( $p < 0.0001$ ).

## 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. (2009) *Discovering statistics using SPSS* (3<sup>rd</sup> Edition). London: Sage.

McKillup S. (2005) *Statistics explained*. Cambridge: Cambridge University Press.

