Babraham Bioinformatics

Introduction to Sample Size Estimation

Version 2018.02

Licence

Table of Contents

Introduction

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

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

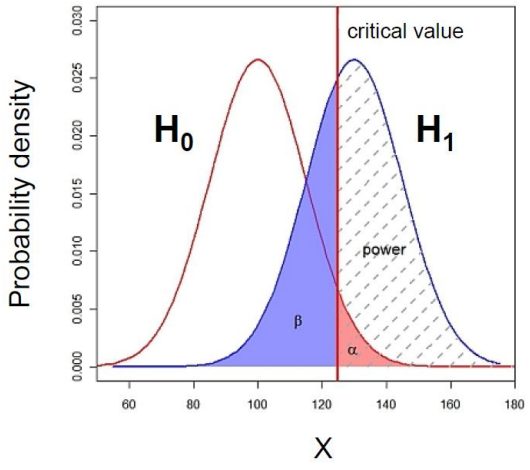
What is Power?

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

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

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

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



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

The ability to reject the null hypothesis depends upon alpha but also the sample size: a larger sample size leads to more accurate parameter estimates, which leads to a greater ability to find what we were looking for. The

harder we look, the more likely we are to find it. It also depends on the effect size**:** the size of the effect in the population: the bigger it is, the easier it will be to find.

What is Effect Size?

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

Effect size determined by substantive knowledge

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

Effect size determined from previous research

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

Effect size determined by conventions

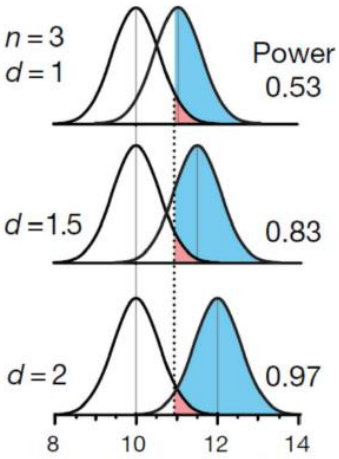
Yet another approach is to use conventions. Cohen (author of several books and articles on power analysis) has defined small, medium and large effect sizes for many types of test. These form useful conventions, and can guide you, if you know approximately how strong the effect is likely to be.

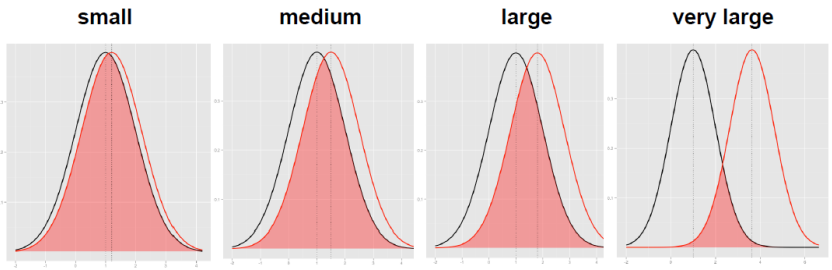
Table 1: Thresholds/Convention for interpreting effect size

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

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

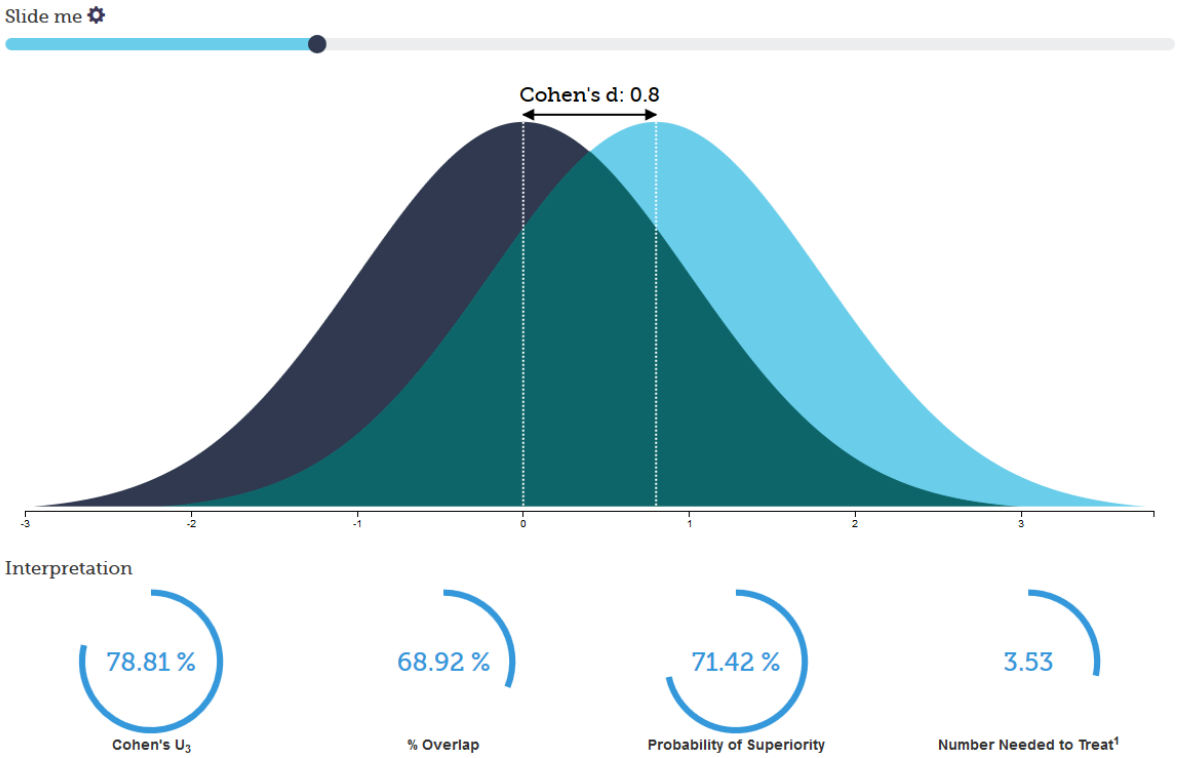
The graphs below give a visual representation of the effect sizes.





Krzywinski and Altman 2013 (Nature Methods)

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



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

So how is that effect size calculated anyway?

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

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

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

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

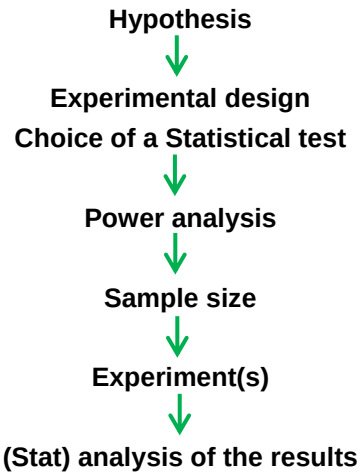
Table2: Interpretation of Effect Size (Robert Coe, 2002)

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

Doing power analysis

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

When should we run a power analysis? It depends of 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:

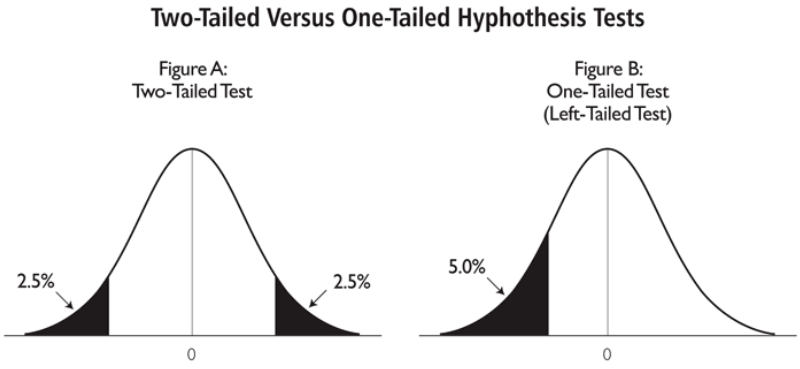


Practically, the power analysis depends on the relationship between 6 variables: the significance level, the desired power, the difference of biological interest, the standard deviation (together they make up for the effect size), the alternative hypothesis and the sample size. The significance level is about the p-value ($\alpha =< 0.05$), the desired power, as mentioned earlier is usually 80% and we already discussed effect size.
Now the alternative hypothesis is about choosing between one and 2-sided tests (= one and 2-tailed tests). This is both a theoretical and a practical issue and it is worth spending a bit of time reflecting on it as it can help understanding this all idea of power.

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

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

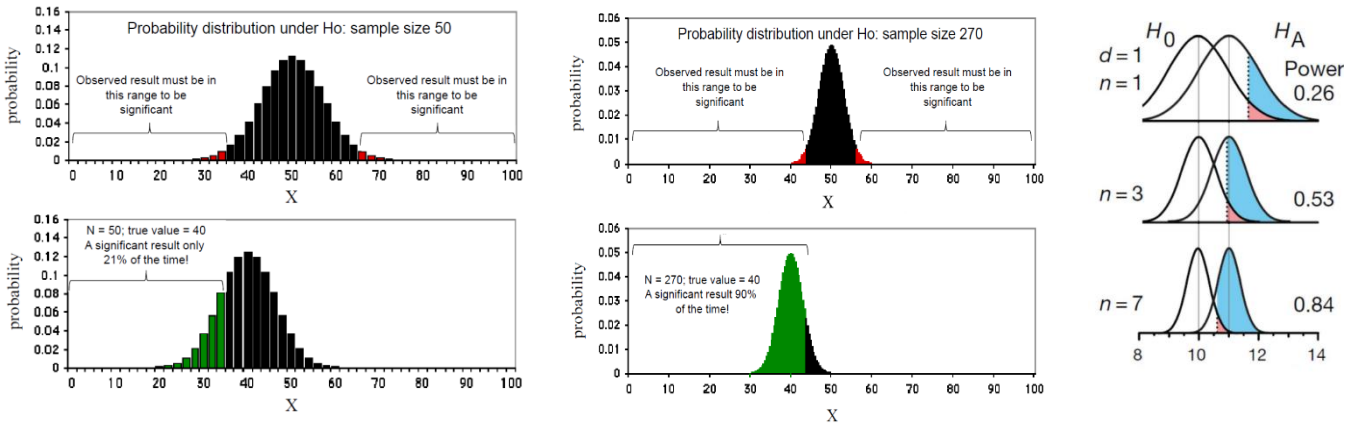
And second question: 'What is the probability of eels in Group 2 being longer than eels in Group 1?' This is a **one-tailed** test, as we'd calculate *p* by looking at the area under only one end of the normal curve. The one-tailed *p* is just one half of the two-tailed *p*-value. In order to use a one-tailed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



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

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

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

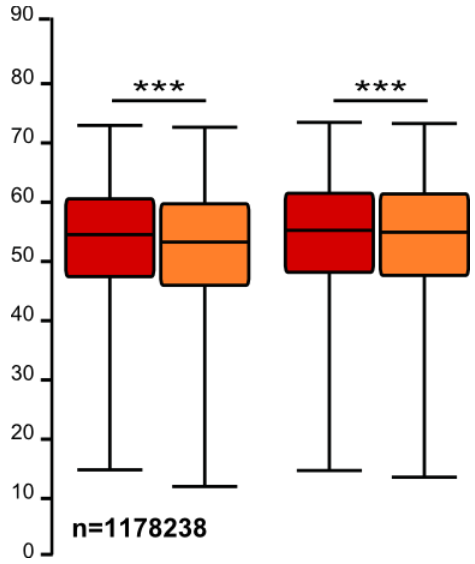


Krzywinski and Altman 2013 (Nature Methods)

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

The problem with overpower

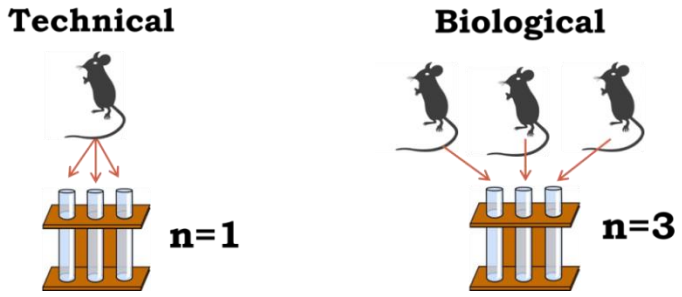
As we saw, power and effect size are linked so that the bigger the power the smaller the effect size that can be detected, as in associated with a significant p-value. The problem is that there is such a thing as overpower. Studies or experiments which produce thousand or hundreds of thousands of data,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



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

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

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



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

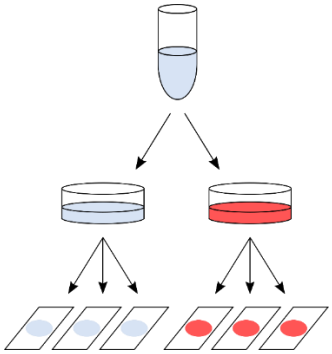
If we repeat the measurements, let's say we weigh each mouse five times then we will have ten different measurements. But this cannot be used to prove that grey mice are heavier than white mice in general, we still have only looked at one white and one grey mouse. Using the terminology above, the five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

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

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

Design 1: As bad as it can get

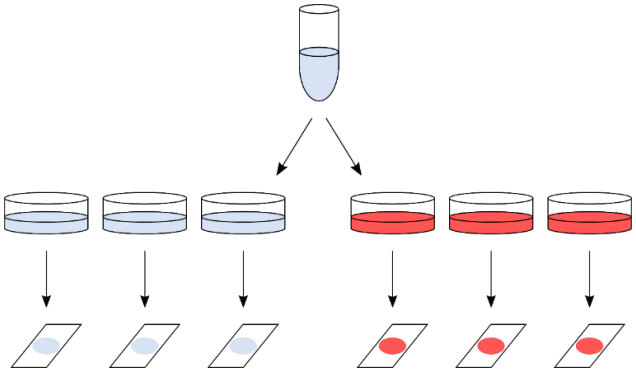
In this experiment a single vial is thawed, cells are divided into two culture dishes and the treatment (red) is randomly applied to one of the two dishes. The cells are allowed to grow for a period of time, and then three samples are pipetted from each dish onto glass slides, and the number of cells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



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

Design 2: Marginally better, but still not good enough

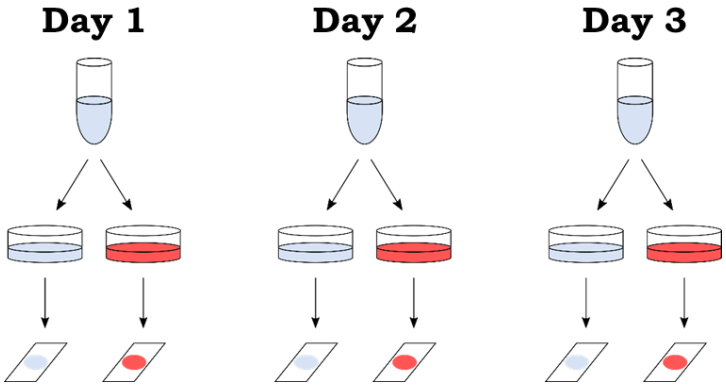
In this modified experiment, the vial of cells is divided into six separate culture dishes, and then cells from each culture dish are pipetted onto a single glass slide. Similar to the previous experiment, there are six values after quantifying the number of cells on each slide. So now is the sample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



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

Design 3: Often, as good as it can get

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



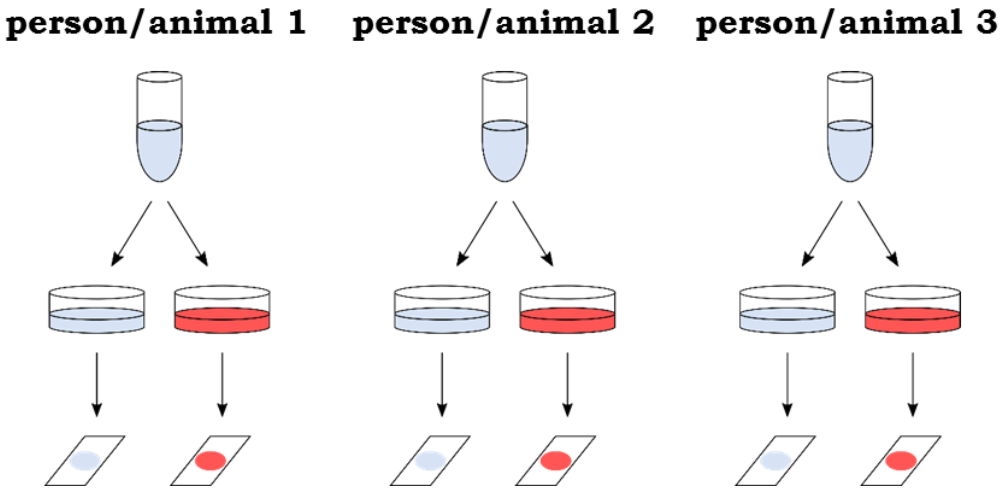
The "independent" aspect of the experiment are the days, and so n = 3. Note, that the two glass slides from the same day can (and should) be treated as paired observations, and so it is the difference between treated and control within each day that is of interest (a paired-samples t-test could be used).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

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

Design 4: The ideal design

Like many ideals, the ideal experiment is often impossible to attain. With cell lines, there are no biological replicates, and so Design 3 is the best that can be done. The ideal design would have biological replicates (i.e. cells from multiple people or animals), and in this case the experiment need only be done once. I hope it is now clear (and after reading the two references) why Design 1 and Design 2 do not provide any reason to believe that the results will be reproducible. Some people may object that it is a weak analogy, and say that they are only interested in whether compound X increases phosphorylation of protein Y, and are not interested in other proteins, other compounds, other cell lines, etc., and so Design 1 or 2 are sufficient. Unfortunately, this is not

the case and it has to due with lack of independence, which is a fundamental assumption of the statistical analysis (see Lazic, 2010 and references therein). But even if you don't appreciate the statistical arguments, this analogy might help: if you claim to be a superstar archer and hit the bullseye to prove it, this is certainly evidence that you have some skill, but let's see if you can do it three times in a row.

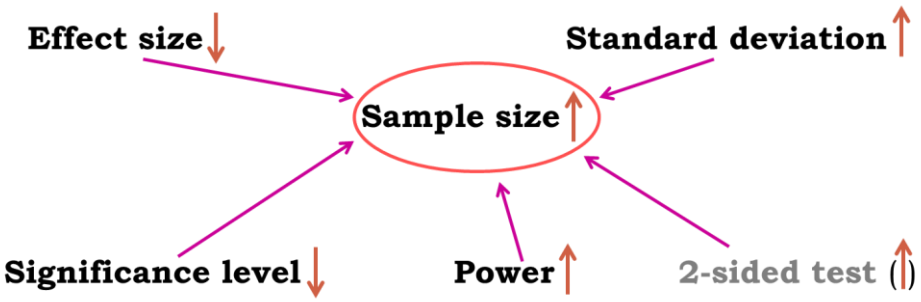


Replication at multiple levels

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

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

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



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

Packages for power calculation

We are going to use G*Power for our power analyses but they are many others, including online resources. For keen R users, the same examples are shown at the end of the manual with R script (See Appendix). A word of warning: though it may change in the future, at the moment, it is pretty cumbersome to do some of the power calculations with R.

G*Power is free to download and to get more information, go to:
http://gpower.hhu.de/

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

- Comparing 2 proportions
- Comparing 2 means
- Comparing more than 2 means
- Correlation

Examples of power calculation

Comparing 2 proportions

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

Research example:

A scientist is looking at a new treatment to reduce the development of tumours in mice. In the control group, 40% of the mice develop tumours and he aims to detect a reduction to 10% at 80% power with 5% significance.

First, we need to know which statistical test we will be applying. In our case, it should be a Fisher's exact test.

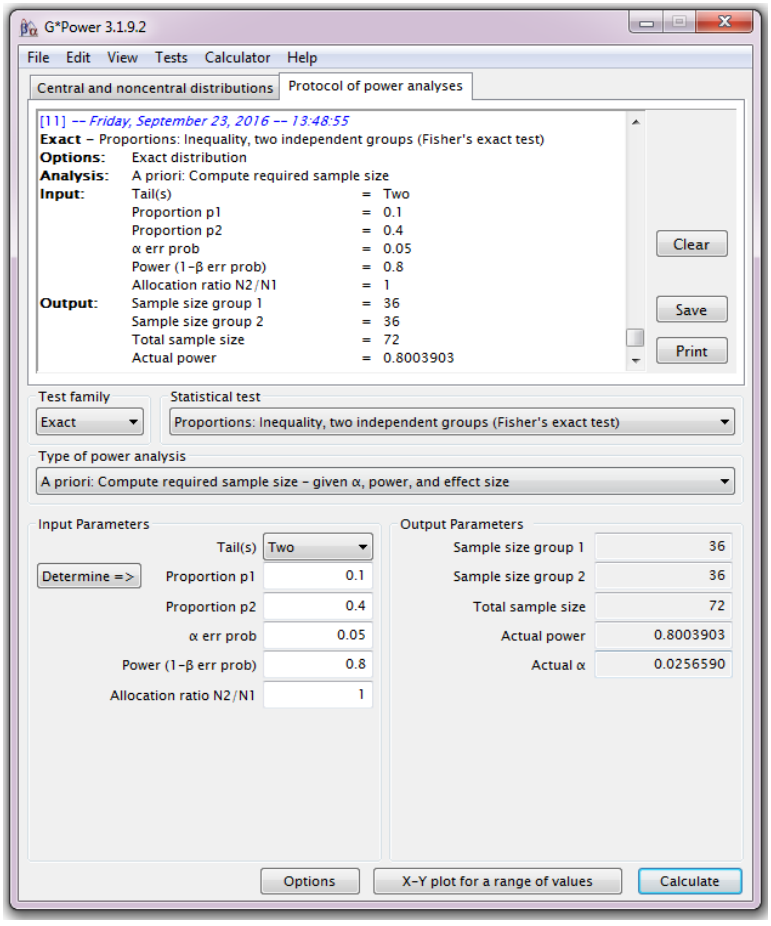
Using G*Power (see below), we follow a 4 steps approach.

-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 group.

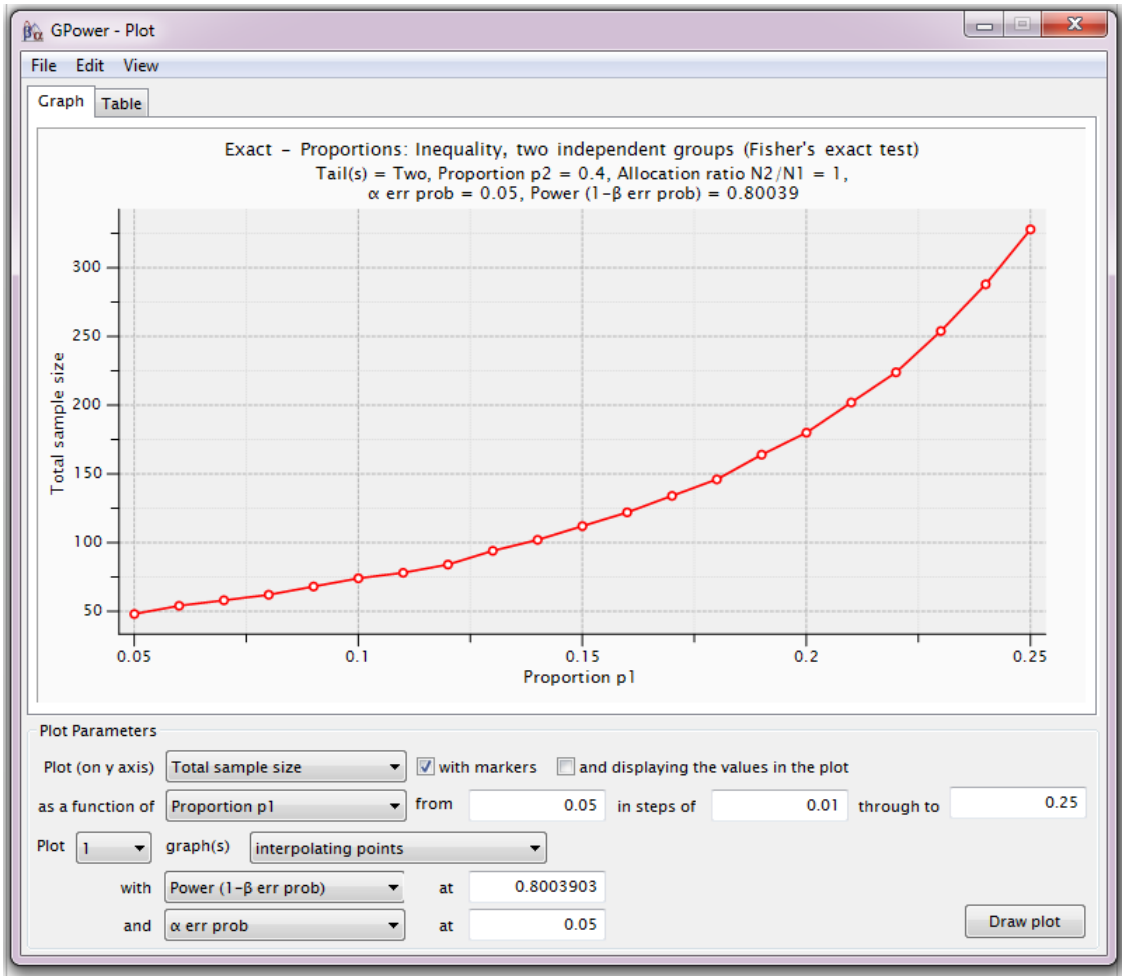
-Step 3: the Type of Power Analysis: we know your significant threshold (α=0.05), the power we are aiming for (80%), we have the results from the pilot study so we can calculate the effect size: we go for an 'A Priori' analysis.

-Step 4: the tricky one, we need to **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.



If aiming for a decrease from 40% to 10% for tumour development, we will need 2 samples of about **36 mice** to reach significance (p<0.05) with 80% power.
If we look at the top of the window, into 'Protocol of power analyses', we find a summary of the power calculation which we can then export as an .rtf file, very useful for a report!

Now, this sample size is based on a specific effect size. It might be interesting to have a range of sample sizes based on a range of effect sizes.
G*Power produces a graph illustrating the sample size/effect size relationship: to do that we need to fix one of the 2 proportions and 'play' with the other.



So if say, the decrease was smaller, 40% to 20% for example, then, with the same parameters, we would need about 175 mice.

Always remember, power calculations are guessing exercises, the sample sizes found are never absolute. Our data might show an effect a bit bigger ☺ or smaller ☹ than expected. By doing a power calculation, providing the effect size we are after is meaningful, we want to know if we can afford to run the experiment. Afford in all possible ways: money, time, space … and ethically. In our case here, we wanted to know how many-ish mice were needed: 30-ish happens to be OK but if it had been 100 or 200, maybe the cost-benefit of the experiment would not have been worth it.

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

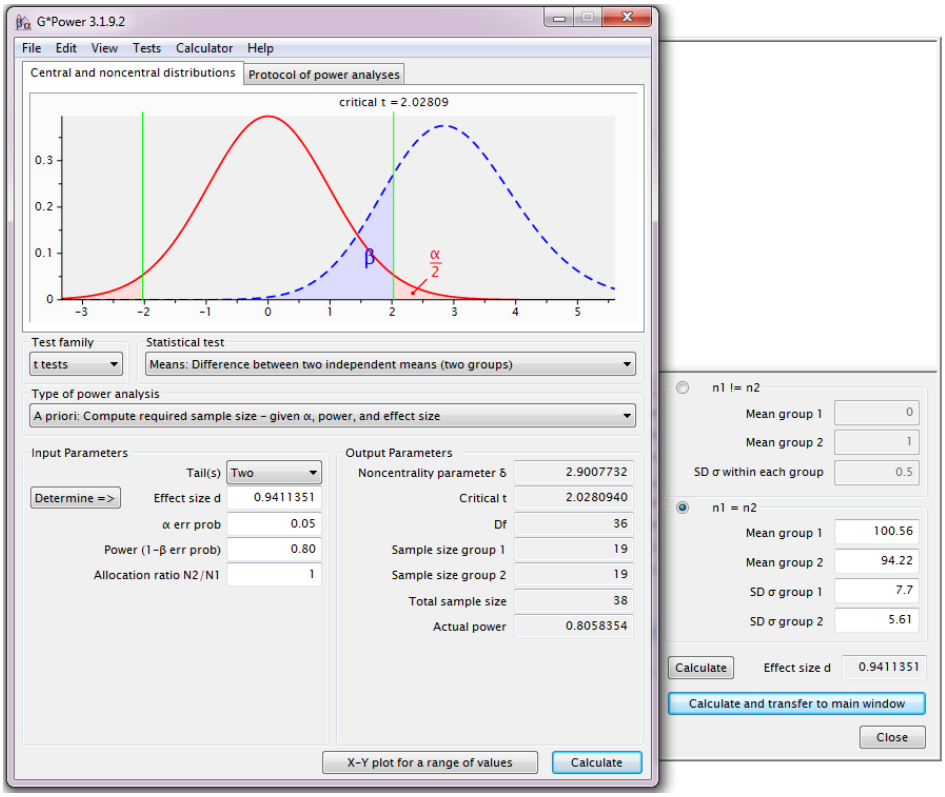
Comparing 2 means

Research example:

A study of the effect of caffeine on muscle metabolism used eighteen male volunteers who each underwent arm exercise tests. Nine of the men were randomly selected to take a capsule containing pure caffeine one hour before the test. The other men received a placebo capsule. During each exercise the subject's Respiratory Exchange Ratio (RER) was measured. RER is the ratio of CO_2 produced to O_2 consumed and is an indicator of whether energy is being obtained from carbohydrates or fats.
The question of interest to the experimenter was whether, on average, caffeine changes RER.
The two populations being compared are “men who have not taken caffeine” and “men who have taken caffeine”.
If caffeine has no effect on RER the two sets of data can be regarded as having come from the same population.

The aim of this module is not to cover statistical tests so we will not go into any details, suffice to say that when using a t-test, we compare 2 means accounting for the variability of each sample.
It makes complete sense, since we want the effect size of interest to about the absolute difference between the means, accounting for the variability in each group. It is exactly what the t-test looks at and the level of significance associated with that observed effect size is a function of the sample size. In other words, how much data do we need to be convinced that the difference we observe is genuine? The smaller the observed effect, the bigger the sample we will need to confident about it.

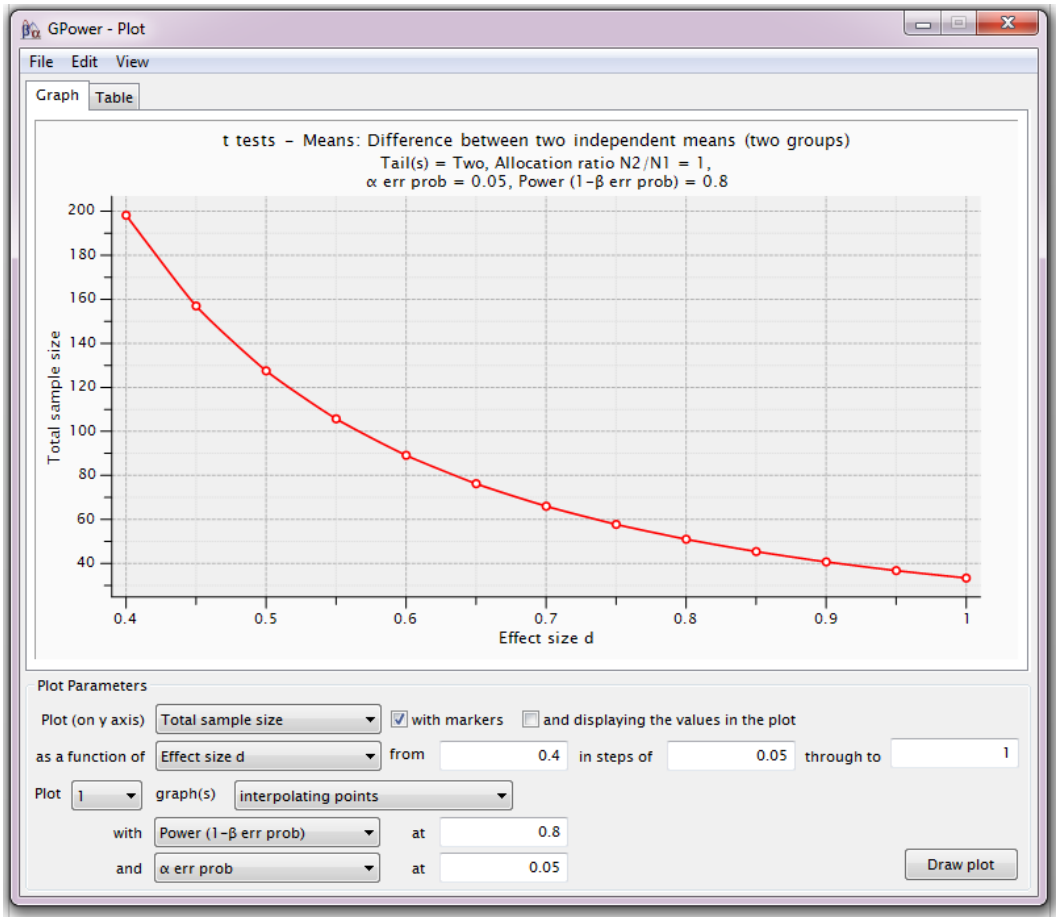
RER(%): Placebo: Mean=100.56, SD=7.70 and Caffeine: Mean=94.22, SD=5.61.



Providing the difference observed in the pilot study is a good estimation of the real effect size, we need a sample size of n=38 (2*19).

Now we saw with the previous example that G*power allowed us to export the power calculation which is quite useful. There is another Tab called' Central and noncentral distribution' which is a graphical representation of the power calculation. It's very useful as it shows a practical example of the concepts we have explored earlier. Whereas the central distribution describes how a test statistic is distributed when the difference tested is null (curve in red), the noncentral distribution describes how t is distributed when the null is false (blue dashed line). This second distribution is centred on the noncentrality parameter **δ** and is calculated with the values we have provided. Take the time to play with the parameters to see how the graphs change.

Then again, we can look at the relationship between sample size and effect size graphically:



Comparing more than 2 means

To compare more than 2 means using a parametric test, you need to run an ANOVA. Like for any other statistical test, it is possible to run a power analyse for it. However determining the sample size for an ANOVA is usually difficult because in theory, one needs to specify all the treatment means. As it is not always possible of course, different approaches have been developed depending on the data available or the package used. We will look at 2 of them here.

But first a quick reminder of what an ANOVA does: it is basically an extension of the t-test as in it compares means accounting for groups variability but because there are more than 2 means, it actually compares the

variance between groups with the one within groups (hence ANalysis Of VAriance). The output of an ANOVA is 2-fold: first, the omnibus part quantifying the overall difference between the groups and second, the pairwise comparisons of interest via post-hoc tests.
Now most of the time, it's the second bit which is really interesting. Remember, the null hypothesis is usually that there is no difference between groups. It quantifies how likely the difference observed was to have occurred by chance. The issue is that if several comparisons are run, there is more chance that a relatively large difference is observed in at least one of them. Hence, with multiples comparisons, the p-values for each pairwise comparison are no longer valid. There are many ways to adjust for multiple comparisons and it is not the purpose of this course.

So back to power. The first approach is to specify the effect size. In the context of the ANOVA, the effect size delta is:

$$\text{delta } (\Delta) = \frac{\text{largest mean - smallest mean}}{\text{sigma } (\sigma)}$$

So one needs to know the largest and the smallest expected mean together with the variability (SD) in each of these group. In fact, this step is very similar in some ways to the comparisons between 2 means: the scientist needs to know at least one mean and SD and among the comparisons he/she wants to make, hypotheses the one which should be the largest and come up with the minimum meaningful value. This is sometimes referred to as the minimum power specification where all means other than the 2 extreme one are equal to the grand mean.

Research example:

We wish to conduct a study in the area of mathematics education involving different teaching methods to improve standardized math scores in local classrooms. The study will include four different teaching methods and use students who are randomly sampled and are then random assigned to the four different teaching methods.

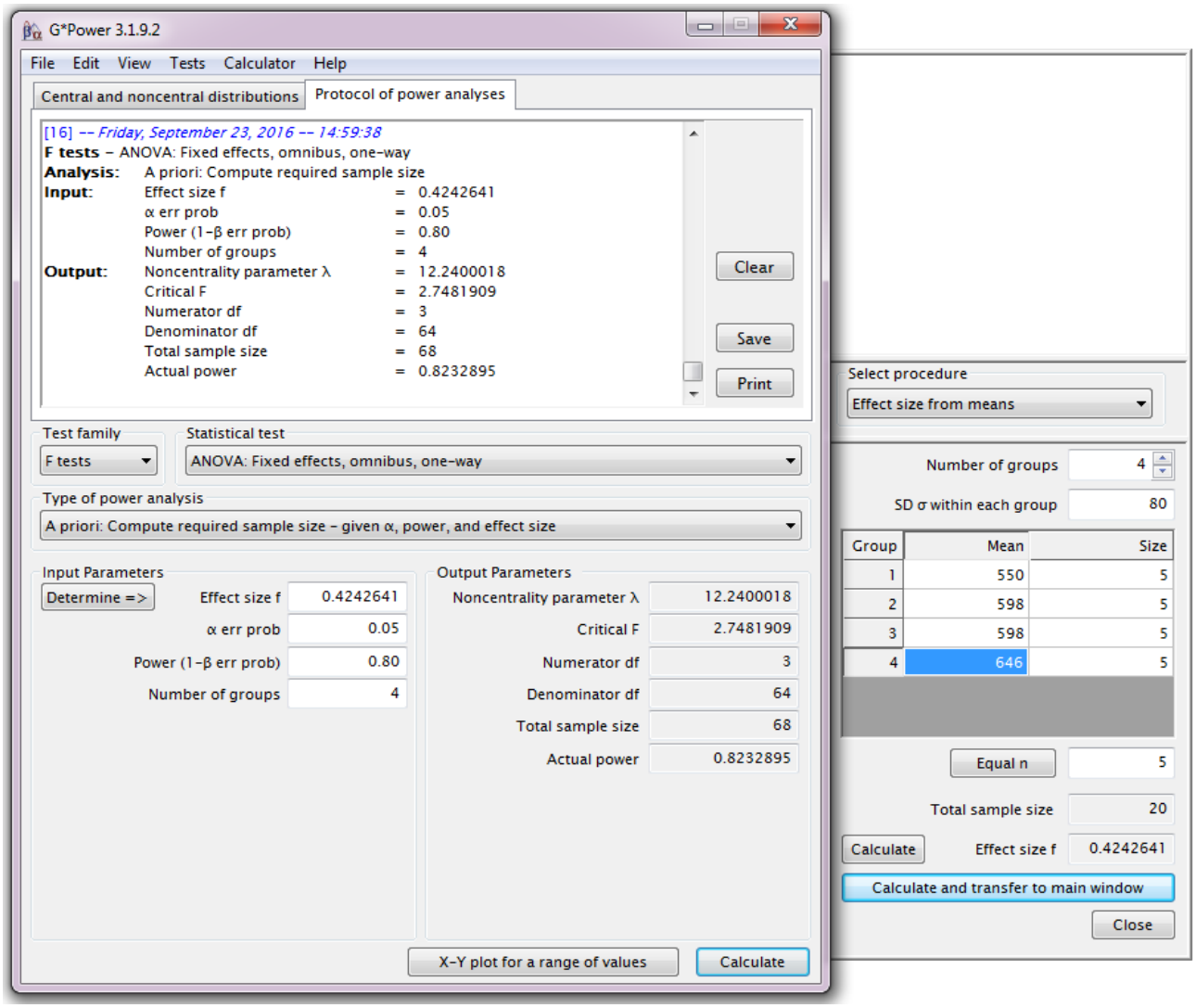
Briefly, here are the four different teaching methods: Group 1: the traditional teaching method, Group 2: the intensive practice method, group 3: the computer assisted method and, Group 4: the peer assistance learning method.
Students will stay in their math learning groups for an entire academic year. At the end of the year, they will take a standardised Maths test. This standardized test has a mean score of 550 with a standard deviation of 80.

The experiment is designed so that each of the four groups will have the same sample size. The important question at this stage of course is: how many students will be needed in each group?

In order to answer it, we will need to make some assumptions and some educated guesses about the data. First, we will assume that the standard deviation for each of the four groups will be equal and will be equal to the national value of 80. Further, because of prior research, we expect that the traditional teaching group (Group 1) will have the lowest mean score and that the peer assistance group (Group 4) will have the highest mean score. In fact, we expect that Group 1 will have a mean of 550 and that Group 4 will have mean that is greater by 1.2 standard deviations, i.e., the mean will equal at least 646. For the sake of simplicity, we will assume that

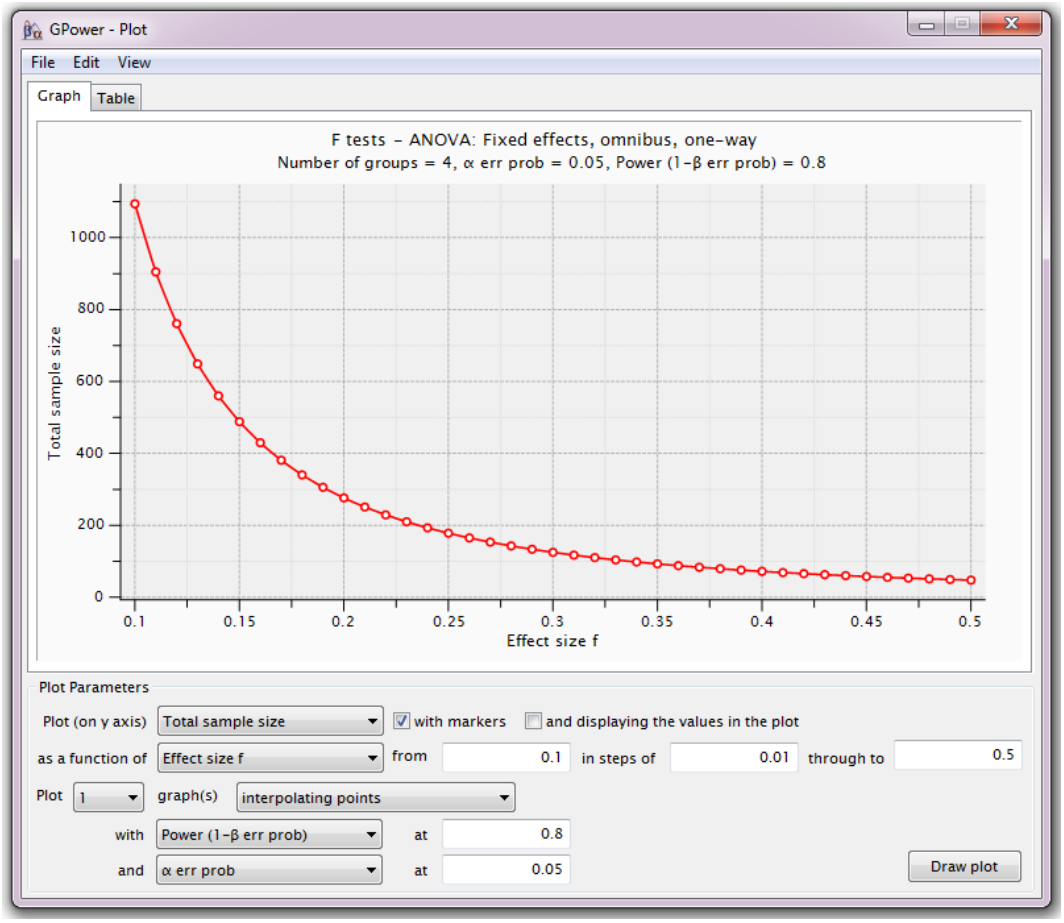
the means of the other two groups will be equal to the grand mean (598=550+646). Of course, if we have a good idea on what these means should be, we should make use of this piece of information in our power analysis. Especially if the other 2 means are more polarized towards the two extreme ends as it will be easier to detect the group effect which will in turns allow for smaller samples being needed.

At this stage, we have done all the hard work, the rest is just calculation and data entry.



A total of 68 students will be required for the test; 17 for each class.

We can also look at the relationship between effect size and sample size:



An effect size of 0.42 is considered big and from the graph above we can see how many students we would need if the effect was medium (n~170) or small (n>1000). In this particular scenario it would be wise to choose sample size closer to a 'medium' effect, to be o the safe side.

The second way to go about power calculation in the context of an ANOVA is more practical and perhaps more useful. The starting point is kind of the opposite of the previous approach: we have to start with the difference we assume will be the smallest (but still relevant). We then consider the number of comparisons of interest we intend to look at and we correct accordingly. This correction can and is usually done manually. It uses the simplest and one of the most widely used correction for multiple comparisons: the Bonferroni one. It is dead easy: basically we divide our significant threshold of choice, usually 5% by the number of comparisons we intend to look at. So if we make 4 comparisons, the significant threshold will be: 0.0125 = 0.05/4. From then, we just follow the same approach as for a t-test power calculation.

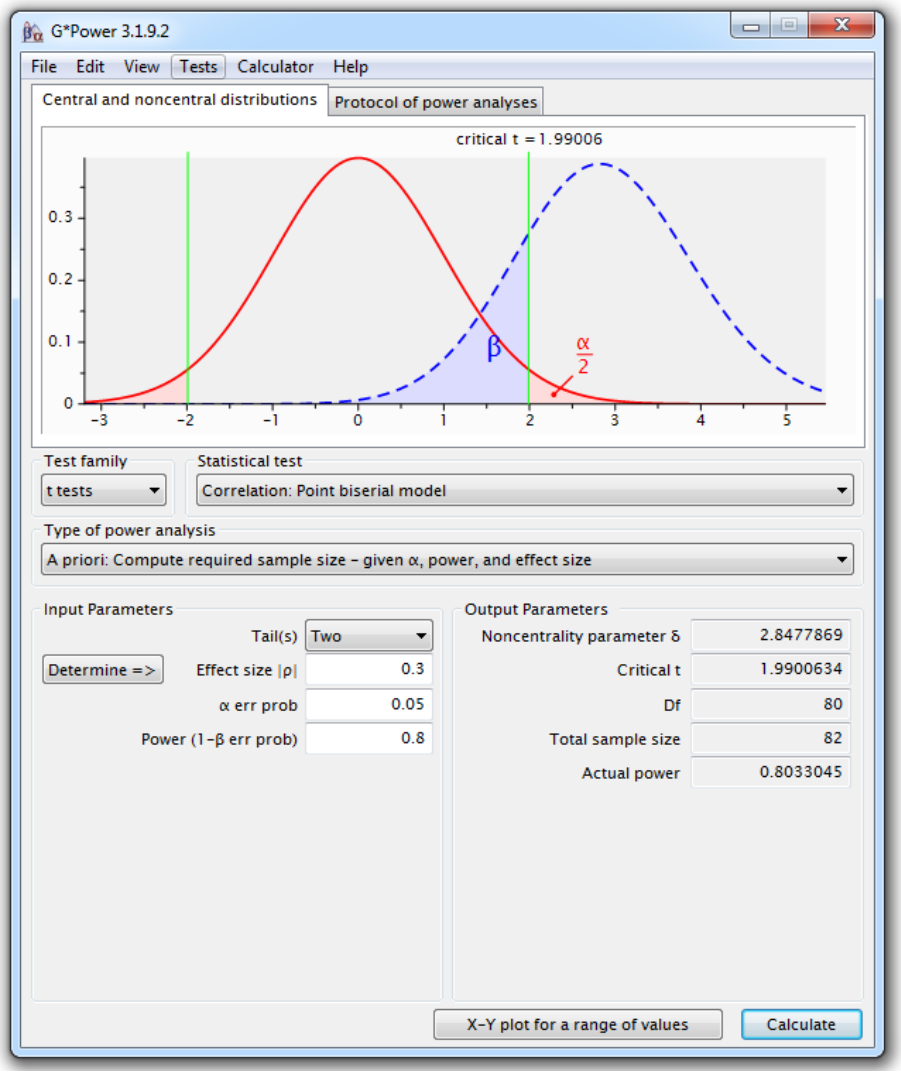
Power calculation for correlation

Doing power calculation for correlation is really easy as the effect size is actually r. So let's do it.

Research example:

An ecologist is looking at the host-parasite relationship in roe deers. Measures of body weight and parasite load will be collected from a group of females: Body weight = f(parasite load).

From a pilot study on a small group, the scientist found: r=0.3. And the other parameters are power=80% and significance level = 5%.



To reach significance, with this level of correlation, the scientist will need n=82 roe deers.

Unequal sample sizes

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

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

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

$$N = \frac{2n(1+k)^2}{4k}$$
$$n_1 = \frac{N}{(1+k)}$$
$$n_2 = \frac{kN}{(1+k)}$$

Research example:

Let's go back to our caffeine example but this time we know that, say, the placebo group will have to be 2 times smaller than the caffeine one, hence k=2. Using the formula above, we get a total:
n=43, placebo=14 and caffeine=29.

Power calculation for non-parametric tests

Nonparametric tests are used when we are not willing to assume that our data come from a Gaussian distribution. Commonly used nonparametric tests are based on ranking values from low to high, and then looking at the distribution of sum-of-ranks between groups.
Now if we want to run a proper power calculation for non-parametric tests, we need to specify which kind of distribution we are dealing with. This would imply more advanced approach to the data and it is not the purpose of this manual.

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

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

Appendix: Power calculations with R:

Comparing 2 proportions:

Cohen's h, is a measure of distance between two proportions or probabilities and is our effect size in this context. Now a transformation needs to be applied to the proportions before calculating the effect size because differences between proportions are not on an equal scale for detectability.
For example, all other things being equal, the power to detect a difference of 0.2 is not the same for: 0.65-0.45 (power = 48%) than the one associated with the 0.25-0.05 (power = 82%).

The transformation needed is:

$$\Phi = 2\arcsin\sqrt{p}$$

From the package **pwr,** we will be using the function `pwr.2p.test`:

```
pwr.2p.test(h = , n = , sig.level =, power = )   ## pwr package ##
with h<-2*asin(sqrt(p1))-2*asin(sqrt(p2))
```

Exactly one of h, n, power and sig. level must be null, so in our case, it will be n. With p1=0.1 and p2=0.4:

```
h<-2*asin(sqrt(0.10))-2*asin(sqrt(0.40))
pwr.2p.test(h, sig.level = 0.05, power = 0.8)
```

```
     Difference of proportion power calculation for binomial distribution (arcsine transformation)

              h = 0.7259373
              n = 29.7878
      sig.level = 0.05
          power = 0.8
    alternative = two.sided

NOTE: same sample sizes
```

R tells us that we need 2 samples of 30 mice to reach a power of 80%. In other words: if we want to be at least 80% confident to spot a treatment effect, if indeed there is one, we will need about 60 mice altogether.

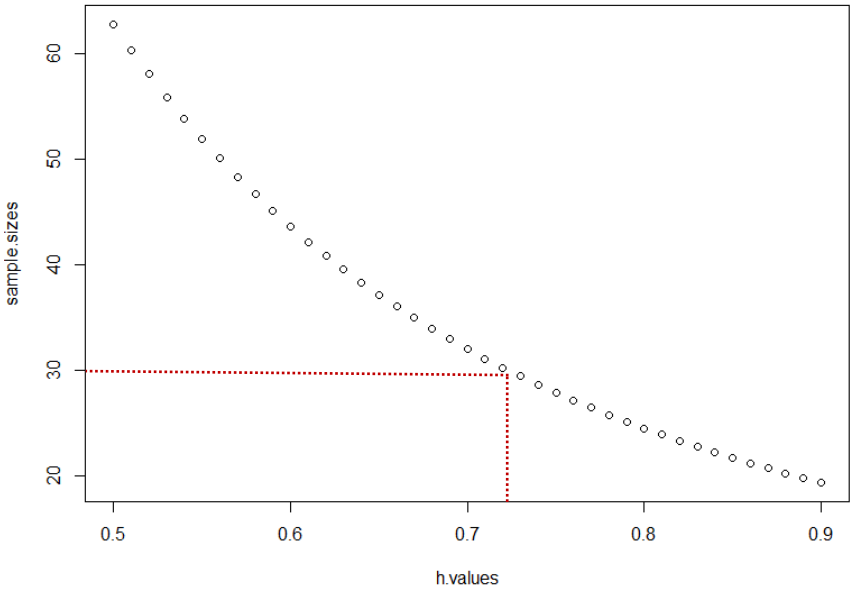
This is based on the specific effect size. It might be interesting to have a range of sample sizes based on a range of effect sizes.

```
h.values <- seq(0.5,0.9,0.01)
sample.sizes <- sapply(h.values, function(x) pwr.2p.test(h=x, power=0.8)$n)
head(sample.sizes)
```

```
[1] 62.79088 60.35264 58.05370 55.88366 53.83306 51.89329
```

Below are the first 6 samples sizes from the vector we have created. We can notice that the sample sizes go down and the effect size goes up.

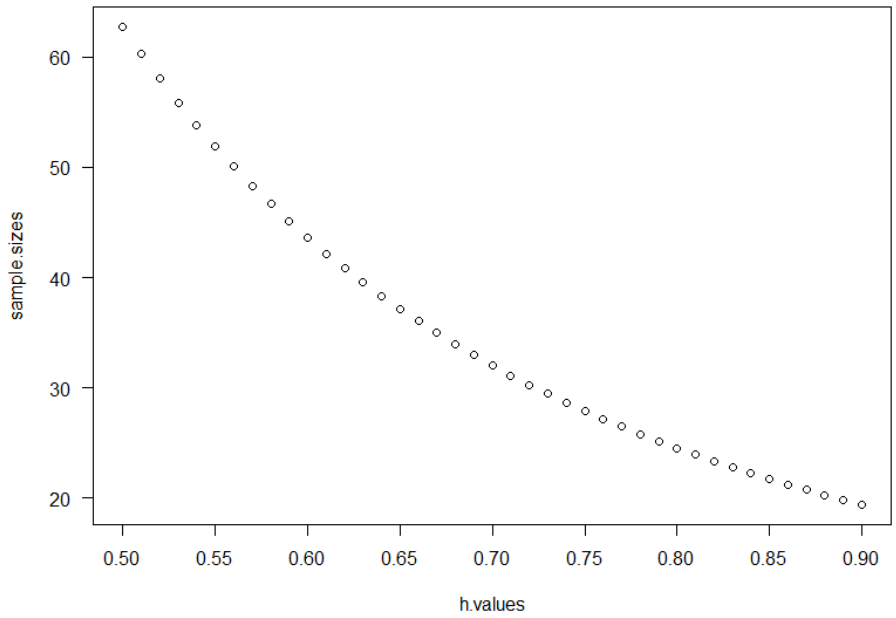
```
plot(h.values, sample.sizes)
```



The relationship between sample size and effect size you can detect is not linear so such a graph allows you to plan your experiment.

If you want to tweak the graph for a better resolution:

```
plot(h.values,sample.sizes,xaxt="n",yaxt= "n")
axis(side=1, at = seq(0.4, 0.9, by = 0.05))
axis(side=2, at = seq(0, 100, by = 10), las=1)
```



This type of graph is very useful if we have a range of effect size in mind. It can be that we are after an optimal effect size of biological interest but an effect 2% or 5% smaller for instance would still be meaningful.

Comparing 2 means:

The function in this case will be:

```
pwr.t.test(n = , d = , sig.level = , power = , type = c("two.sample",
"one.sample", "paired"))
```

So we need to calculate Cohen's d:

$$d = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\frac{s_1^2 + s_2^2}{2}}}$$

*S*1 and *S2* are the standard deviations for the first and the second samples respectively, S1=7.7 and S2=5.61. Together, they are used to calculate the so-called pooled SD (denominator).

So in R we go:

```
mean1<-100.56
mean2<-94.22

s1<-7.7
s2<-5.61
```

We are assuming equal sample size.

```
numerator <- abs(mean1-mean2)
denominator<- sqrt(((s1*s1)+(s2*s2))/2)

d<- numerator/ denominator
```

You should get:

```
[1] 0.9411351
```

So now:

```
pwr.t.test(d=d, sig.level = 0.05, power = 0.8)
```

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

```
     Two-sample t test power calculation

              n = 18.73347
              d = 0.9411351
      sig.level = 0.05
          power = 0.8
    alternative = two.sided

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

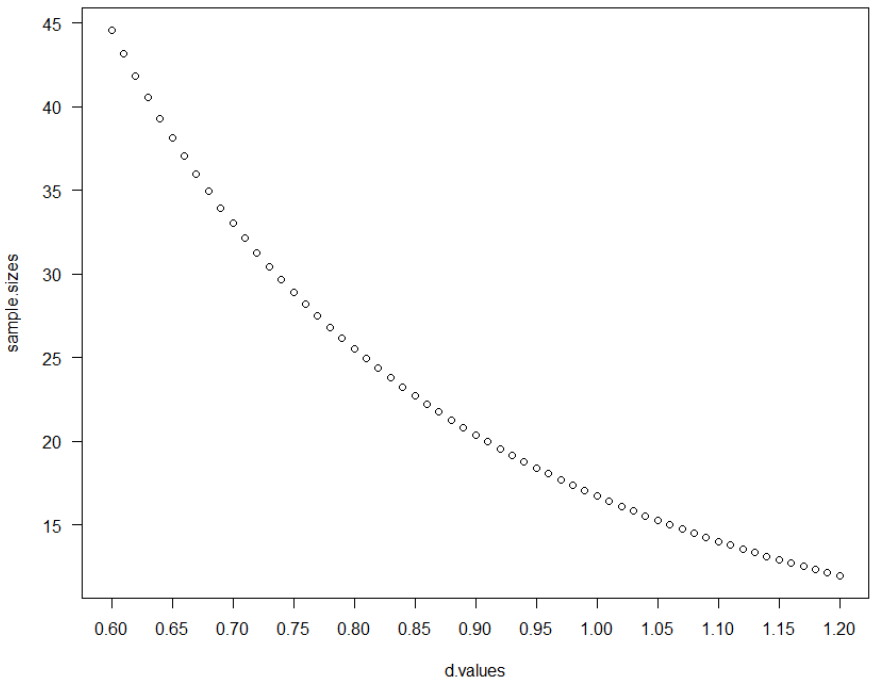
We need a sample size of n=38 (2*19).

For a graphical representation we go:

```
d.values <- seq(0.4, 1, 0.01)
sample.sizes <- sapply(d.values, function(x) pwr.t.test(d=x, power=0.8)$n)
plot(d.values, sample.sizes)
```

Cuter graph:

```
plot(d.values,sample.sizes,xaxt="n",yaxt= "n")
axis(side=1, at = seq(0.6, 1.2, by = 0.01))
axis(side=2, at = seq(5, 45, by = 5), las=1)
```



Comparing more than 2 means:

From core R, the function will be:

```
power.anova.test(groups = , n = ,between.var = , within.var = ,sig.level = 0.05,
power = )
```

Knowing the means, R can calculate the variance between them (between.var) and we know that SD=80, hence within.var=6400 (80*80)

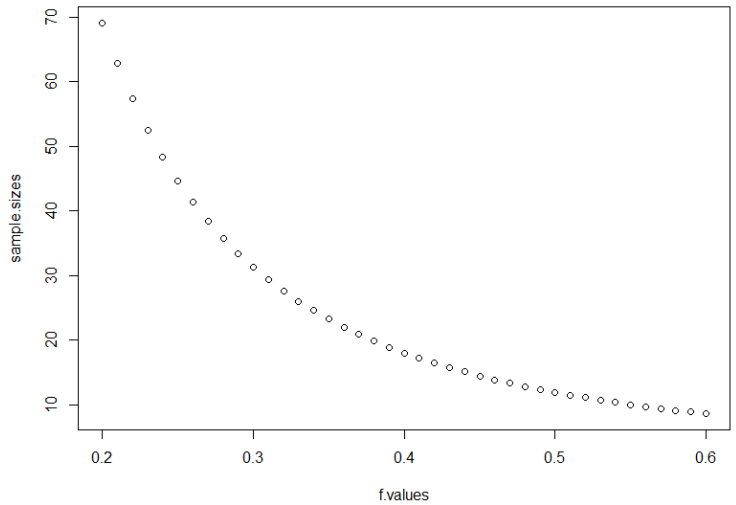
```
groupmeans <- c(550, 598, 598, 646)
power.anova.test(groups = length(groupmeans),
                 between.var = var(groupmeans),
                 within.var = 6400, power = .8)
```

```
     Balanced one-way analysis of variance power calculation

         groups = 4
              n = 16.15347
    between.var = 1536
     within.var = 6400
      sig.level = 0.05
          power = 0.8

NOTE: n is number in each group
```

Now if we want to plot the range of sample sizes as we have done before, it is a bit cumbersome because `power.anova.test()` does not give us the effect size f.

So basically, the easiest way to go about, is to use the sample size returned by the function (n=16) to get the effect size. From then, we simply follow the same approach.

```
pwr.anova.test(k=4, n=16,power=0.8)
f.values <- seq(0.2, 0.6, 0.01)
sample.sizes <- sapply(f.values, function(x) pwr.anova.test(k=4,f=x,
power=0.8)$n)
plot(f.values, sample.sizes)
```



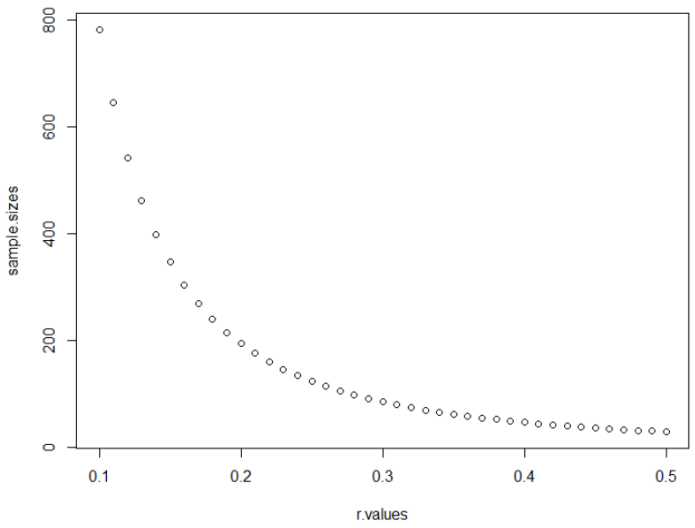
Correlation:

```
pwr.r.test(r =0.3, sig.level = 0.05, power = 0.80)
```

```
     approximate correlation power calculation (arctangh transformation)

              n = 84.07364
              r = 0.3
      sig.level = 0.05
          power = 0.8
    alternative = two.sided
```

So providing the correlation observed in the pilot study was a good reflection of the reality, to reach significance with such value, the scientist will need n=84 animals.
And to get the power for a range of sample sizes:

```
r.values <- seq(0.1, 0.5, 0.01)
sample.sizes <- sapply(r.values, function(x) pwr.r.test(r=x, power=0.8)$n)
plot(r.values, sample.sizes)
```



References

Getting the sample size right: a brief introduction to power analysis by Jeremy Miles.
http://www.jeremymiles.co.uk/misc/power/

Kraemer, H.C. and Theimann, S. (1987). How many subjects? Statistical power analysis in research. Newbury Park, CA: Sage.

http://rpsychologist.com/d3/cohend/

Robert Coe (2002). It's the Effect Size, Stupid. What effect size is and why it is important.
http://www.leeds.ac.uk/educol/documents/00002182.htm

McGraw, K.O. and Wong, S.P. (1992) 'A Common Language Effect Size Statistic'. *Psychological Bulletin*, 111, 361-365.

Martin Krzywinski& Naomi Altman (2013) Points of significance: Power and sample size. Nature methods.

Statistics for Experimental Biologists. What is "n" in cell culture experiments?
http://labstats.net/articles/cell_culture_n.html

Thresholds for interpreting effect sizes. Paul D. Ellis, Hong Kong Polytechnic University