Instantie voor Dierenwelzijn Utrecht

Reviewed by department of statistics
Info: statistics@Lumc.nl
Leids Universitair Medisch Centrum

Guide to experiment design and statistics

Nelleke Verhave
Paul Westers
January 2017

Table of Contents

I. Introduction

Statistics are an important element for guaranteeing the quality of any experiment using animals. The discipline provides instruments for determining the minimum number of animals you will need to demonstrate a given expected effect with enough strength of evidence (power analysis), as well as instruments for analysing the measurements (descriptive and inferential statistics), so that sound conclusions can be drawn.

This usually makes statistics an essential element of responsible research that takes the 3 Rs into consideration. Although statistics cannot answer the question of whether a study is useful and/or ethically responsible, with the research question in mind, statistics can indeed contribute to an optimal research design.

It is always a good idea to consult a statistician when setting up new research involving laboratory animals. Remember that a statistician him- or herself is usually not an expert in that field of research. The communication between the two of you will be an interaction between the knowledge of the researcher and that of the statistician, so it's crucial that you have good information exchange. The final results will be in balance: a well-run, sound experiment and an optimal design that has already taken the final analyses into account, in which the research question and the welfare of the animals will always be the main elements.

To streamline this interaction, this guide will briefly explain basic statistical terms. There are a few references for more information at the end.

II. Key points to consider

To properly set up and conduct an experiment using animals, and to analyse and present the results, there are various points to take into consideration, many of which are connected:

1. What are the research questions? Make a distinction between primary and secondary research questions.
2. What is the design of the experiment? Think about what you are going to measure, how you will measure it and what animals you need. Also, will be it a paired or unpaired design, how will you organise the randomisation and blinding, how will you analyse the data later, what practical elements are required, etc.
3. How will the data be analysed? In other words, what is the statistical analysis plan? Think of what results you will want to describe, test or model; what hypotheses you intend to test; what statistical analyses you intend to use, whether you intend to do one or two-tailed tests, etc.
4. What is the minimum number of animals you will need? Determine your sample size using the design and the statistical analysis plan, and think about whether it may result in any practical problems. If necessary, modify the design, statistical analysis plan or the power analysis.
5. To what extent can you take the 3 R's into consideration?
6. Are you prepared for unanticipated situations? What will you do with extreme measurements? What will you do if animals become ill or die? What will you do if you do not satisfy the requirements for the statistical analyses? Etc.
7. Will you be following the ARRIVE or GSP guidelines? If so, keep in mind that every aspect of the experiment, from design to writing the article, must meet the guidelines.
8. How will the data ultimately be displayed in a database? Every animal is given a unique code. This code must appear in the database. Sensitive information about an animal must be in a separate file that is only accessible to a very limited group of the people involved. All data is

stored in the database, even data not being used in the statistical analysis. The structure of the data file will be determined by the statistical analysis. The standard is 1 line per animal, with the information about it in the columns. This means not only the information about the individual animal (breed, weight, age etc.) as well as the treatment, time of measurement, the measurements themselves, etc. Keep a code book for the database and keep it up to date, so that everyone can know exactly how data should be input.

9. What results do you intend to present in the article, and how? It can be very frustrating if after all you cannot present particular results that are interesting or important, because you weren't thinking about them when planning the experiment or did not measure for them.
10. Do you have the statistical skills to be able to understand and defend the analyses? Familiarise yourself with the basic principles and concepts of the statistical analyses you will be using. This includes learning how to use the statistical software.

Tip: keep a logbook, in which you not only can find but justify all the choices and agreements you have made. Don't hesitate to stop by and ask the animal welfare officer and/or statistician for advice, even if it's only to check something.

III. Designing animal experiments

Research question

The foundation of any experiment or series of experiments using animals is one or more research questions. The research questions are formulated on the basis of existing or new theories, or as a continuation of previous research. Usually there are 1 or 2 main or 'primary' research questions, and in addition, less important but still interesting research questions ('secondary' research questions). Properly formulating the research questions is essential for the success of an experiment. The research question must neither be too vague, nor too many detailed. It is better to have a series of simple research questions than one question with a lot of complicated ones.

The primary research question often contains the essence of the research question, and will involve the most important measurement. The secondary research questions will be related to the other measurements, or are refinements of the primary research question. The primary research question is the basis for the power analysis.

Design

The term design means the way in which the experiment is set up. The design is primarily determined by the primary research question, but takes the secondary research questions into account. The research question is further operationalised in the design. It is thus essential to formulate the research questions clearly and unambiguously.

After you have thought thoroughly over the research questions, it is now important to think about the setup of the animal experiments. What treatments (experimental conditions) will you be comparing? Will you be using 2 or more independent groups (a parallel or unpaired design) or dependent groups (a paired or matched design)? What type(s) of data will be measured and how will you measure it/them? What is possible in practice? How many animals will you need, and what requirements must they satisfy? Will the animals survive the experiment, and if so, will they be suitable for use in other experiments?

It's important to think about the statistical analysis when designing the experiment. A minor change to the design can improve the statistical analysis. On the other hand, a statistical analysis also sets requirements for the design.

Paired or unpaired

In an unpaired design the animals are linked at random to an experimental condition (treatment). In a paired design, animals who are often from the same litter are distributed at random over the treatments, or animals are measured at various points over time (repeated measurements). The simplest form of the latter is measuring before and after a procedure. One specific form of paired data is when animals are matched to each other on the basis of characteristics. Paired or matched animals are also called a pair.

The advantage of paired groups is that the statistical analysis can correct for the differences between the animals (biological variability), and thus you will need fewer animals. However, it must be practically feasible, and of course it must be ethically responsible.

Control groups

To determine if an intervention(s), usually a particular treatment(s), have (has) an effect, the results are compared with a reference treatment, usually a control group or a control measurement (for the intervention). It is essential that the reference or control group is similar in composition and treatment with the experimental group(s). This is why randomisation and blinding are so important.

Blinding and Randomisation

In blinding, you ensure that information that refers to the designated treatments remains confidential for all interested parties in the experiment, who otherwise could be influenced consciously or unconsciously by this information.

Randomisation ensures that animals are assigned completely randomly to a treatment. One commonly used controlled method of randomisation is a random assignment. Every animal has the same probability of being assigned to a treatment. In a paired design, you randomise within each pair, so that all treatments are represented in one pair. A randomisation plan is drawn up beforehand, so that it is clear during the experiment what treatment any new animals will be assigned. This prevents certain animals being assigned to a particular treatment. The best way is if the animals are not assigned to a particular condition by the researcher him- or herself, but 'blindly', by an in dependent person. An obvious choice is the person who takes care of the animals. It is necessary to check afterwards to test whether the randomisation plan was followed in making this assignment.

Blinding and randomisation are essential because they prevent any influences that may, consciously or unconsciously, distort your experiment.

Bias

With bias, your results are distorted, for example, because a structurally too high or too low value is measured. This mainly affects subjective parameters, but must certainly not be ignored in supposedly objective measurements. It is thus a good idea to check all the measuring instruments for possible bias beforehand. Trainings and clear agreements about the subjective parameters with those who will be making the measurements are definitely a part of the process.

IV. Statistical Analysis plan

When you are setting up experiments with animals, you must think about how you will ultimately analyse your data. This is written up in the statistical analysis plan. Not only do the research question and the design determine the statistical analysis, but the statistical analysis also sets requirements for the design. Each statistical analysis gives estimates of the effects of a statistical test as well as the results.

Outcome variable

The outcome variable is the variable that reflects the outcome of a measurement or observation. They are used to describe and analyse the results of the experiment. A distinction can be made between primary and secondary outcome variables.

CONTINUOUS VARIABLE

If the measurement or observation is a real number (e.g. length, weight, blood pressure), then the outcome variable is said to be continuous. Most basic statistical analysis (e.g. ANOVA, t-tests, correlation/regression) assumes that a continuous variable has a normal distribution.

DISCRETE VARIABLE

If the data consist of categorical values, these are also known as discrete variables. A discrete variable can be binary (pregnant or not pregnant, dead or living), nominal (blood type, fur colour) or ordinal (hairless, little coat, moderate coat, healthy coat, thick coat). In contrast to a nominal variable, the possible outcomes of an ordinal variable have a logical order.

Counts are a special kind of an outcome variable. In principle, they are not continuous but if the range of the possible counts is large, then they are considered continuous. Whether a count is considered a discrete or/ continuous variable is arbitrary and depends on the experiment.

For various reasons, there is sometimes a desire to transform continuous variables into categorical variables (e.g. ‘age in years’ becomes ‘young, adult or old’). This is not a bad thing in itself, if you have kept this in mind when formulating your research question and designing the experiment. In any case, it will cause information loss and consequently loss of power.

NORMAL DISTRIBUTION

The standard parametric statistical tests (e.g. ANOVA, t-test, correlation/regression) assume a continuous variable with a normal distribution. A normal distribution means that all the measured data are distributed along an axis in a bell-like shape, or ‘bell curve’. Most of the data lie around the average (mean) measurement, and the further away from the mean, the fewer values there are.

There is a misconception that the data must be normally distributed. Actually, in most standard statistical analysis the requirement is that the residual values have a normal distribution. In everyday practice, with the t-test and ANOVA, each group is checked to see if its data are normally distributed. It should be clear that this has less power than seeing whether the overall residual values of all groups together are normally distributed.

Most statistical programmes provide an option with a parametric statistical test to determine whether the residual values are normally distributed. Parametric statistical tests are robust. That means that minor deviations from normal distributions have no effect on the test.

If a continuous variable is not normally distributed, then the measurement may be transformed (e.g. the logarithm or root) or non-parametric statistical techniques may be used (e.g. Mann-Whitney, Kruskal-Wallis, Spearman rank correlation). It is a misunderstanding to suppose that non-parametric test has no other requirements. The Mann-Whitney and Kruskal-Wallis tests also have the condition that the shapes of the distributions of the various groups are the same, and the Wilcoxon signed-rank test requires that the distribution of the differences is symmetrical.

Besides the normal distribution, every statistical test has other requirements. The most important of these are independency of the measurements and homogeneity of the variances.

Statistical analysis

In statistical analysis, there is testing for expected effects, as well as describing possible associations. In the first case, you can think of comparing averages and proportions. If we speak of associations, then we usually are thinking of connections between 2 or more categorical variables (e.g. chi-square test or log-linear model) or between 2 or more continuous variables (e.g. correlation and (possibly multiple) linear regression). Survival statistics and (possibly multiple) logistic regression are examples of a mix of continuous and categorical variables.

Testing is an important element of both testing effects and for modelling associations.

NULL HYPOTHESIS AND ALTERNATIVE HYPOTHESIS IN A STATISTICAL TEST

In terms of statistics, the null hypothesis expresses in an exact and quantitative form, what you do not expect as the outcome of your experiment. Thus, the null hypothesis must be formulated such that it can be rejected or not with the parameters from your experiment. In short, the null hypothesis is that there is no effect. The alternative hypothesis is that there is an effect.

The testing procedure is that, assuming that the null hypothesis is true, you attempt to make it plausible that the null hypothesis cannot be true on the basis of the data. This is comparable to a court case in which the assumption is that the defendant is innocent and must then be proven guilty with evidence.

If the null hypothesis cannot be rejected, you can conclude that the expected effect has not been found, or at least has not been demonstrated. In a court case, this would mean that there has been insufficient evidence to declare the defendant guilty. However, this does not mean that the null hypothesis has been confirmed and that there is no effect: a defendant who has not been found guilty is not necessarily innocent.

If the null hypothesis can be rejected, you can assume that the alternative hypothesis is true ('there is an effect').

Depending on the research question, the effect can involve the difference between 2 or more averages or proportions, but also something like the relationship between 2 or more measured outcome variables. The result of testing the effects is often expressed as: statistically significant (null hypothesis rejected) or not statistically significant (null hypothesis not rejected). In addition to the

statistical significance, there is another aspect that is often ignored or underestimated, namely relevance (including clinical relevance). A found effect can be statistically significant, but not relevant for everyday practice. On the other hand, an effect may not be statistically significant, but it will be clinically relevant given its size.

When publishing the results of the experiments, it is thus essential to, addition to stating the p-value (for the benefit of statistical significance) to state the size of the effect with a confidence interval (for the benefit of clinical or other relevance).

Up to now, we have assumed that the aim of the experiment was: to demonstrate that there is an effect. However, you may instead be aiming to demonstrate that there is no effect. Otherwise, the principle remains the same, except that now the null hypothesis is that there is an effect and the data must furnish the evidence that the effect is negligible. These kinds of designs are also called bio-equivalence studies. As a rule, they require many more (laboratory) animals.

INTRODUCTION TO TESTING THEORY

The principle of testing is to see if there is sufficient evidence that a given supposition ('there is an effect') is demonstrable. The basic assumption (the null hypothesis) is precisely what you hope not to find, i.e. 'there is no effect'. However, there is uncertainty about the real situation, since only a limited number of data are being collected (random sampling). Thus, chance may be causing the difference between the collected data and the null hypothesis. If the probability of it being chance is too small, then chance is seen as improbable, something we call a significant result: the null hypothesis is rejected.

Two kinds of errors can be made in testing:

1. The null hypothesis is erroneously rejected (an innocent person is declared guilty), which is called a type I error. The probability of a type I error is indicated with α. This error is also called the unrealibility of the test.
2. The null hypothesis is erroneously not rejected (a guilty person is declared innocent). This is called a type II error, and the probability of it is indicated with ß.

The probability that the null hypothesis is being rightly rejected (a guilty person is found guilty) is called the power. The power is usually indicated with Π (= 1- ß)

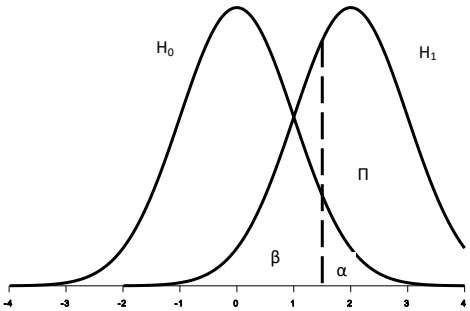


Figure 1: Overview of type I/II errors and power

The following steps are taken when conducting a test:

1. Formulate the null hypothesis and the alternative hypothesis.
2. Choose the desired significance level of the test. Usually α is 5%.

3. A different α can be chosen, e.g. in risky research α = 1% or in exploratory research α = 10%.
4. Determine the test statistic and its distribution under the null hypothesis.
5. Calculate the outcome of the test statistic.
6. Determine its p-value, or the critical value, or the (1-α)*100% confidence interval.
7. Reject the null hypothesis if
   a. the p-value is less than α, or
   b. the test statistic is greater than the critical value, or
   c. the value under the null hypothesis is not in the confidence interval.
8. Formulate your conclusion, always in terms of the context of the research and never with statistical jargon.

Remember that the desired power is not important when testing the null hypothesis.

ONE- OR TWO-TAILED TESTING

If you can expect with great certainty that the effect can only go in one direction (e.g. painkillers will not make the pain worse) you can consider making the test one-tailed. This can mean that you need fewer animals to be able to confirm your alternative hypothesis.

However, if it turns out in practice that the effect is actually going in the other direction, then you have not thoroughly thought through the experiment, and moreover, you have to conclude that, regardless of the size of the effect, the null hypothesis cannot be rejected. Thus, only use a one-tailed test if you are absolutely certain that the effect can only go in one direction. Deciding afterwards to carry out a two-tailed test or, even worse, a one-tailed test in the other direction, is not permitted.

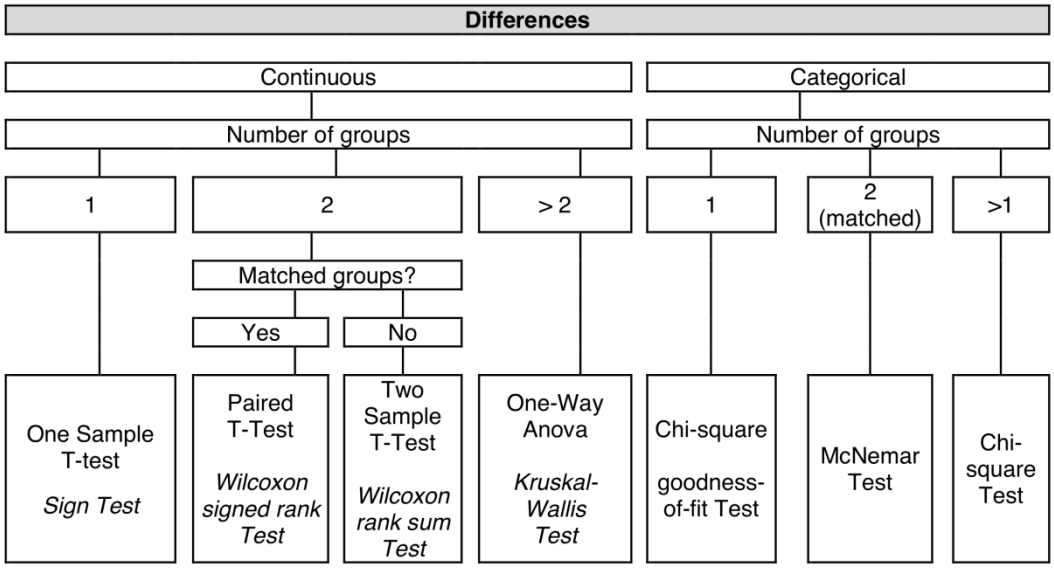
POWER

Power is used to indicate the probability that the null hypothesis is correctly rejected, or to use the example of the court case, that a guilty person is indeed declared guilty. The power is indicated in a formula with: Π=1-β, in which β stands for the probability of a type II error.

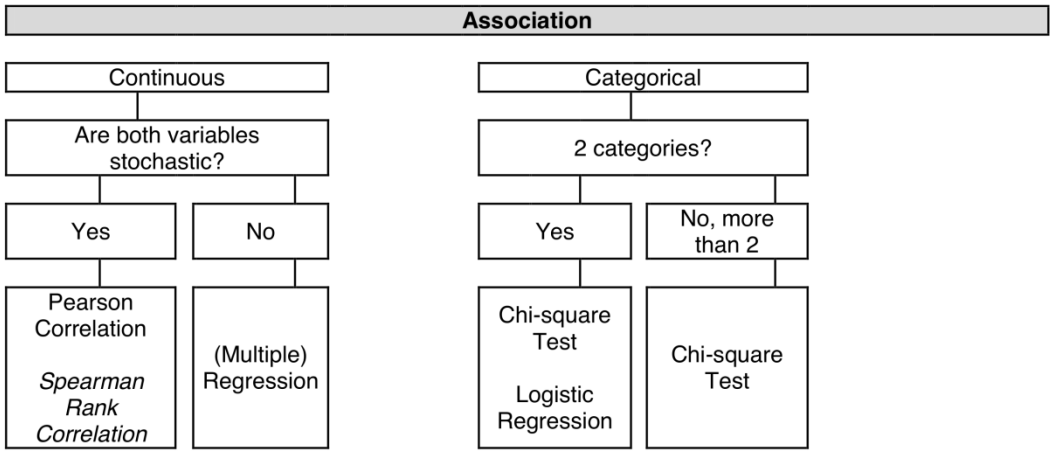
SCHEMATIC DIAGRAM OF BASIC STATISTICAL ANALYSES

Depending on the research question and the design of the experiment, there is a wide range of possible statistical techniques available to you. The standard statistical techniques are shown as a schematic diagram in the figure below. Remember, however, that the range of possible statistical techniques is much greater. There are also survival statistics, repeated measurements, multi-way ANOVA or ANOVA for paired data, techniques for hierarchical structures (Multi level), etc.

Flow chart of the statistical tests[1]



[1] The tests denoted in *italics* are the nonparametric versions of the above mentioned tests



V. Determining sample size

One important element of the design is the number of laboratory animals it requires. The goal is to find a balance between not too many animals (ethically and economically undesirable, legally prevented), and not too few animals (predictive value and power too low). You can determine this optimum number using the power analysis.

Power analysis

There are two points in the research where it can be worthwhile to perform a power analysis:

1. Before starting the study. At this point the goal is to find the optimum between a not too-large and a not too-small sample, as mentioned above.
2. After completing the study. At this point the important things are recognising if the power is too low, and being able to make a distinction between statistical significance and the size of a (possibly clinically relevant) effect or strength of relationship. This is only interesting if a clinically relevant effect has been found, but was not statistically significant.

Factors that determine the power are 1) design of the research; 2) one- or two-tailed test; 3) unreliability of the test (α); 4) effect size (δ); 5) size of the variability or standard deviation (σ) and 6) sample size (n). These factors are shown in the illustration below (note that the standard error is a function of σ and n). In general, it can be said that the power increases as the effect increases, the significance level increases, the variability decreases or the size of the sample increases (see figure 2).

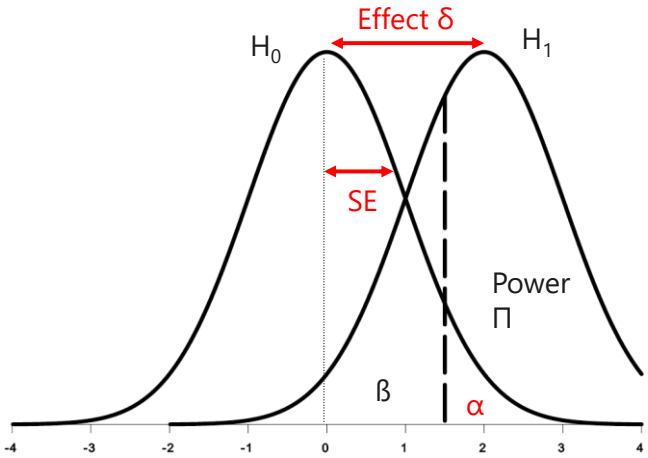


Figure 2: Overview of the factors determining the power

You will need to know the following to determine the minimum number of animals required:

1. what statistical analysis will be conducted (in other words, what is the design);
2. one-or two-tailed testing;
3. probability of a type I error (α);
4. desired power (Π);
5. minimum relevant or expected effect (δ) and
6. expected variability or standard deviation (σ)

Choosing the size of the minimum relevant effect

The biggest problem is often determining the minimum relevant or expected effect and the expected distribution. One aspect that helps to determine the minimum relevant effect is the limit at which an effect is no longer practically interesting or relevant (clinical relevance). To get an indication of the expected distribution you can refer to a pilot study, literature study or general knowledge.

However, to determine the minimum sample size, it is not important to know the effect and the distribution themselves, but their relationship. This relationship is called the effect size (= ES). To calculate a sample size for the unpaired t-test, for example, the ES is defined as ES = δ/σ = (u1-u2)/sd. In some literature (such as the handbook by Van Zutphen) percentages are used to indicate the expected effect and expected variability. This is not important in and of itself, but with absolute values, the effect is much more easily quantifiable in the original scale. An effect of 10% means nothing if it is not known what that 10% is related to.

1. Van Zutphen: *Effect size* (ES) = effect / CV = $\{(u_1-u_0)/u_0\}$ / $\{sd/u_0\}$ with u_0 is mean in control group and u_1 mean in treatment group, sd is the (pooled) standard deviation per group and CV is variance coefficient or the sd expressed as percentage of the mean.
2. Standard literature: Effect size (ES) = difference in means / sd = (u_1-u_0) / sd with u_0 is mean in control group and u_1 mean in treatment group and sd is the (pooled) standard deviation per group.

It may be that you have absolutely no idea what your minimum relevant effect is or what the expected variability can be. In this case, you can resort to the Cohen's effect size measures, although it is recommended that you avoid them if possible.

What if minimum sample size is not practically feasible?

The calculated minimum required number of animals may turn out to be difficult to work with in practice. If it only involves the number of animals, you can try to solve it by modifying the settings of the power analysis or your design. If the problem is not the number of animals, but that not all measurements can be made in one day, it can be done over several days. Remember that this will have consequences for your design, your statistical analysis and your power analysis. We recommend that you consult with a statistician and/or animal welfare officer.

How do you determine the sample size?

Software

There are several programmes that can perform a power analysis and calculate sample size. The best known of them are nQuery, PASS, G*Power and PS. The last 2 can be downloaded free of charge from the internet. There are also numerous sites that can determine the power or sample size for particular statistical techniques. Check first if the calculations from these sites are correct.

The minimum required number of animals can also be determined by the 'pwr' package from R.

Calculating sample size manually

In simple statistical analysis, the minimum number of animals required can also be calculated manually, with met β = 1-power.

	Input	To calculate sample size (n)
1 mean	sd(σ), effect(δ)	1. $n = \frac{\sigma^2}{\delta^2}\left(z_{1-\alpha/2} + z_{1-\beta}\right)^2$ 2. round n up 3. df = n-1, with df = degrees of freedom 4. $n = \frac{\sigma^2}{\delta^2}\left(t_{1-\alpha/2;df} + t_{1-\beta;df}\right)^2$ 5. repeat steps 2 t/m 4 until n no longer changes
1 probability	P_0 and P_1 (probabilities under the null hypothesis and alternative hypothesis)	$n \geq \frac{\left(Z_\alpha\sqrt{p_0(1-p_0)} + Z_\beta\sqrt{p_1(1-p_1)}\right)^2}{(p_1 - p_0)^2}$
2 means (unpaired)	sd(σ), effect(δ)	1. $n = 2\frac{\sigma^2}{\delta^2}\left(z_{1-\alpha/2} + z_{1-\beta}\right)^2$ 2. round n up 3. df = n-1 4. $n = 2\frac{\sigma^2}{\delta^2}\left(t_{1-\alpha/2;df} + t_{1-\beta;df}\right)^2$ 5. repeat steps 2 t/m 4 until n no longer changes
2 probabilities (unpaired)	p_C = the probability in de control group, p_E = the probability in the treatment group, $\delta_0 = p_E - p_C$	$n \geq \frac{p_C(1-p_C)+p_E(1-p_E)}{\delta_0^2}\left(Z_\alpha + Z_\beta\right)^2$

With z_p the z-value under the standard normal distribution in which Pr(Z<z-value)=p and analogue $t_{p;df}$ the t-value under the t-distribution with degrees of freedom df Pr(T<t-value)=p. For example, if α=5% then $z_{1-.05/2}$ = 1.96.

POWER ANALYSIS IN 1-WAY ANOVA THEORY

The best statistical analysis for comparing the means of 'k' conditions is the 1-way ANOVA. The principle of the 1-way ANOVA is that it tests if there is even a difference at all between the k conditions. The null hypothesis is then 'there is no effect between the k conditions'. If this null hypothesis is rejected, it can be concluded that the means differ from each other for at least 2 conditions. An interesting logical question is then for which 2 this is the case. To answer this question, there are post-hoc tests.

Most post-hoc tests are geared to pair-wise comparisons of conditions, but it is also possible to compare means of subsets (whether or not with weightings) of conditions. In this latter case, however, we speak of contrasts rather than post-hoc tests, but in fact they are special kinds of post-hoc tests. Most post-hoc tests can be roughly described as modified versions of the unpaired t-test.

With k conditions, there are at most k*(k-1)/2 possible pairwise comparisons, each with an unreliability of 5% (α). This means that the probability is greater than 5% that for at least 1 of all these pairwise comparisons, the null hypothesis will be erroneously rejected. If you perform all the possible pairwise comparisons, then this probability is at most 5*k*(k-1)/2. The actual probability partly depends on the dependence between the post-hoc tests and the post-hoc test used. To maintain the overall unrealibility at 5% the significance levels with the post-hoc tests are modified, in other words lowered. However, lowering the probability of a type I error increases the probability of a type II error, and thus the power is reduced (see figure 1).

Over the years, dozens of post-hoc tests have been developed that attempted to keep the overall unrealibility at 5% with as little loss of power as possible. Almost all post-hoc tests, except for the LSD test, are corrected in some way for the number of post-hoc tests. Increasing knowledge has shown however that some of these post-hoc tests are still available in many kinds of statistical software, but it is better not to use them. In everyday practice, the most commonly used are the Tukey (if all comparisons are pairwise), Dunnett (only comparison with a reference group) and Bonferroni (if a selection of pairwise comparisons and/or contrasts is limited). Of course, the choice is also determined by which post-hoc test is commonly used in the research field.

The post-hoc tests in an ANOVA analysis can be approached in several ways:

1. Theoretical: exactly according to the theory. Even the significance level bias is modified completely in accordance with the theory: the modified significance level becomes α/(k*(k-1)7/2). This can result in extremely low modified significance level.
2. Practical: exactly according to the theory, but keeping in mind the fact that the bias must be modified, while simultaneously a maintaining a lower limit for the significance level. For example: regardless of the number of post-hoc tests maintaining a modified significance level of 1%.
3. No correction: without modifying the significance level, because the post-hoc tests are a separate issue.
4. No overall ANOVA: the overall ANOVA is not conducted. You immediately conduct the post-hoc tests, and even use the regular unpaired t-test, with or without correction of the significance level.

The practical method seems the obvious one, especially if you are comparing many conditions. Remember that each approach has its own consequences for the test results and for the power.

POWER ANALYSIS WITH 1-WAY ANOVA IN PRACTICE

Most articles about calculating sample sizes deal with the primary research question 'Is there an effect between the k conditions', in other words the overall ANOVA. But what if the post-hoc tests are the primary research questions?

We know that most post-hoc tests maintain the overall bias at 5%, but that there is some loss of power. To retain the desired power for the post-hoc tests, the bias (α) must also be modified when determining the minimum sample size, so that the sample size becomes somewhat larger.

> **Example:** Researchers are aiming to demonstrate an effect size of ES = 1.5 (this was the smallest expected *effect size* of the 4 specific post hoc tests) with a power of 90%. For their research, they are only interested in 4 specific post-hoc tests. For their post-hoc test they opt for the Bonferroni method with α=5/4%=1.25%.

If they had not modified the bias when determining the minimum sample size, then n would have been n=11. However, in the analysis this would have resulted in a power of 77% and thus a loss of 13% with the same effect size. If they had also chosen the modified significance level α=1.25% when determining the minimum sample size, then n would have been n=15 with no loss of power.

To sum up, the researcher has three options:

1. If the primary research question is 'is there a difference between the k conditions', then the sample size should be determined based on the overall ANOVA analysis. The post-hoc tests are then only secondary research questions or are considered a nice 'extra' that is interesting for the test but not for the power.
2. However, if your primary research questions are focused on the 'difference between given conditions', then the minimum sample size determination can be based on an unpaired t-test with as distribution the pooled distribution of all conditions and
   a. without a modified significance level
   b. with a modified significance level

Note: this assumes the Bonferroni post-hoc test is used.

If we apply this to the example above, then the minimum sample size in option 1 is n=3, for option 2a n=11 and for option 2b n=15. The choice between 2a and 2b is a weighing of the size of the sample and loss of power. It is a tricky decision, in which loss of power is also determined for the choice of the number of post-hoc tests, as well as the actual effect size.

VI. To conclude

This guide was written for the average researcher with a basic knowledge of statistics. The underlying theories can be found in basic statistics textbooks, Wikipedia or the references below.

VII. More Information:

Websites

You can read a lot about the subjects above on the following website: www.3Rs-reduction.co.uk. The site also allows you to self-test your knowledge.

Literature

Amor, S., Baker, D., (2012) Checklist for reporting and reviewing studies of experimental animal models of multiple sclerosis and related disorders. Mult Scler Relat Disord. 1(3)
An article with a complete list of the information you should provide when you present your results in a scientific journal. Although the focus is op MS, the information is suitable for other fields as well.

Lara-Pezzi, E et al., (2015) Guidelines for Translational Research in Heart Failure
J. of Cardiovasc. Trans. Res. 8(1)
An article with a main focus on the translation of models for heart failure in animals and effective suggestions for designing similar studies. The focus is on heart failure, but the information is suitable for other fields as well.

Steward, O., Balice-Gordon, R. (2014) Rigor or mortis: best practices for preclinical research in neuroscience. Neuron. 84(3)
This article discusses best practices in experimental design and statistics in preclinical studies of neurological and psychiatric disorders. There is also some focus on data management. The focus of the article is on neurological and psychiatric disorders, but the information is suitable for other fields as well.

Festing, M. F. W., Altman, D.G., (2002) Guidelines for the Design and Statistical Analysis of Experiments Using Laboratory Animals. ILAR 43(4)
This article helps you answer your research question with various types of experiments, step by step. It provides ways to prevent errors and to get meaningful data. It is especially oriented to the use of animals in research and emphasises the 3R's and sound statistical analysis.

Aban, I.B., George, B., (2015) Statistical considerations for preclinical studies, Exp. Neurol. 270
This article discusses statistical terms with the goal of improving the quality of animal studies. This article was written especially for people using animals in preclinical studies so that the data are suitable as preparation for the clinical phase of research.

Hirst, J.A., et al. (2014) The Need for Randomization in Animal Trials: An Overview of Systematic Reviews. PLoS ONE 9(6)
This article demonstrates using a meta analysis how important it is to conduct randomised animal studies, with blind assignment of interventions, and conducted blind.

Tweel, I. van der (2006) Sample size determination. Intern Report no. 4
(http://portal.juliuscentrum.nl/Portals/2/Disciplines/Biostatistics/SAMPLE%20SIZE%20DETERMINATION_electronic%20version.pdf)
This report explains the simplest sample size calculation.

Bate, S.T. & R.A. Clark, (2014) The design and statistical analysis of animal experiments, Cambridge University Press
This book discusses many aspects of setting up and analysing animal experiments.