How to Reduce the Number of Animals Used in Research by Improving Experimental Design and Statistics

Michael FW Festing

C/o Understanding Animal Research
25 Shaftsbury Avenue, London, UK

anzccart
HUMANE SCIENCE

Fact Sheet

ANZCCART
C/o The University of Adelaide
South Australia 5005
Tel: +61883037585 Fax +61883037587
Email: ANZCCART@adelaide.edu.au

Abstract

Animal experiments are expensive and time consuming and raise a number of ethical issues. It is therefore important to ensure that when animals have to be used, that use is both scientifically and ethically justifiable and that the experiments are well designed so that they give the correct answers using the minimum number of animals and other resources. There is a perception that this may not always the case.

There are three major types of experiments. "Pilot studies", involving small numbers of animals are used to test the logistics of a proposed larger study and gain some preliminary information. "Exploratory studies" are used to generate data which might be used to formulate hypotheses which may then need to be tested in "Confirmatory experiments", discussed in this paper, in which groups of animals receive different treatments with the aim of testing some relatively simple hypothesis.

There are five requirements for a well designed confirmatory experiment. First, there must be an absence of bias. There should be no systematic difference between the different treatment groups apart from the administered treatment. This is achieved by correctly identifying the experimental unit, which is the smallest division of the experimental material that can receive a different treatment. Often this would be a cage of animals if the treatment is to be given in the water or diet, but it may be an individual animal if any two animals can receive a different treatment. The experimental units then need to be assigned to the treatment groups at random and as far as possible, all subsequent manipulations and measurements need to be done in random order and blind with respect to the treatment group.

Second, an experiment needs to have high power, so that there is a good chance that it is able to detect a biologically or clinically important effect. Power is obtained by controlling inter-individual variation ("noise") and ensuring that the subjects are capable of responding, to give a good "signal" so that there is a high signal/noise ratio. The group size then needs to be determined using either a power analysis or the resource equation method.

Third, the range of applicability of the results should be explored where possible. Does the treatment affect both sexes equally; are there strains differences in response or does the response depend on the presence of a drug or other factor? This is achieved by using factorial experimental designs in which two or more factors are varied simultaneously. This can usually be done without increasing the total number of animals. Fourth and fifth, experiments should be simple, to minimise the chance of making a mistake, and they should be amenable to a statistical analysis. Finally, once the data is available it should be screened for mistakes, summarised and subjected to appropriate statistical analysis.

Introduction

It is important to reduce the number of animals used in biomedical experiments. This is important for both ethical reasons and because research using animals is expensive and time consuming. One way of doing this is to improve the design and statistical analysis of such experiments[1]. Statisticians have noted that experiments are often poorly designed with inadequate or incorrect statistical analysis. In 1988 Mead[2] claimed that "...the standard of design of experimental investigations is poor and the basic principles of design are widely ignored...". This is sometimes still true in both clinical research and in the design and analysis of animal experiments[3,4]. Good experimental design can save time because the experiments do not need to be repeated and they can save money because experiments which give the wrong results leading, for example, to the failure of a new drug, will result in a waste of money. They can also save animals by ensuring that the experiments are neither unnecessarily large so that animals are wasted or too small so that the experiment fails to detect clinically important effects. Finally, if a scientist gambles his or her scientific reputation on the results of a badly designed experiment, they may lose that scientific reputation.

The aim of this paper is to review the principles of experimental design for scientists using laboratory animals, but with only brief details of how these designs can be statistically analysed.

Types of experiment

There are at least three major types of experiment:

- Pilot studies are usually small scale experiments that aim to test the logistics of carrying out a larger experiment and to generate preliminary data used to plan those experiments. While the results of such studies are not normally published, the work helps to ensure that when a larger study is done, all staff are adequately skilled and all the apparatus, reagents, etc are available. They may indicate whether the proposed dose levels are too high, such as when all the animals die unexpectedly, or too low if an expected response is not observed. This type of experiment should be encouraged as it can save a lot of time, money and animals.
- Exploratory experiments are often carried out in new situations in order to generate data from which scientific hypotheses may be formulated. Sometimes they are the sort of experiment that either "works" or "doesn't work". Often with this kind of study, many outcomes are measured. The statistical analysis of data from this type of experiment may be problematical because the investigator chooses which means to compare after summarising the results and because if many tests are done, some of them will be "statistically significant" just by chance. Thus, there is no assurance that an exploratory experiment will give the correct results. It is never-the-less, useful for exploring biological responses.
- Confirmatory experiments are used to test a formal, and preferably quite simple, hypothesis which is specified before starting the experiment. In most cases there will be a number of treatment groups and the aim will be to determine whether the treatment affects the mean, median or some other parameter of interest. In this case it is essential that the experiments give the correct result. It is this type of experiment which is discussed in more detail here.

Avoiding bias

The experimental unit

A well designed experiment will be unbiased, with no systematic differences between the treatment groups that could be mistaken for the effect of the treatment. This can best be achieved by correct choice of the experimental unit or subject, by blinding where possible and by randomisation of the experimental subjects to the different treatment groups.

The "experimental unit" is defined as "the smallest division of the experimental material that allows any two experimental units to receive different treatments"[5]. It is the unit of randomisation and the unit of statistical analysis. For example, in an investigation of whether or not the breeding performance of mice is affected by having rats housed in the same animal room[6], the animal room is the experimental unit because for such a question it is indivisible. This investigation was actually done as a retrospective survey using breeding records from animal rooms with or without rats, rather than as an experiment. However, if the investigations were to be done as an experiment, an efficient design might be for the experimental unit to be an animal room for a period of time, such as four months. Several rooms would be maintained with or without rats for this period of time, with the breeding performance being recorded and with the same mouse strains being maintained throughout. Such a design is variously called a "within-subject", a "repeated measures", a "crossover" or a "randomised block" design. It is confusing that different disciplines have used different names for essentially the same design. In the rest of this manuscript such a design will be called a "randomised block" design, as this was the first designation of such a design. It is quite common to house several animals in a cage with the treatment being given in the diet or water. In this case the experimental unit is the cage. So if there are four cages each with two animals, "N" (the number of experimental subjects) is four and not eight. However, if the animals can be individually housed or individually treated and there is no good biological reason why animals treated differently can be housed in the same cage, then the individual animal is the experimental unit. In some cases it is possible to do within-cage or within-animal experiments, with each cage or animal receiving different treatments sequentially, in random order. If several animals are used, then this is again a randomised block design. A more confusing situation is a teratology experiment where a pregnant female is treated with the test compound (or a placebo) and the effect on the pups is measured. In this case it is the mother, not the pups, which are the experimental units since two pups in the same litter cannot receive different treatments.

Failure to identify correctly the experimental unit is not uncommon and it can completely invalidate an experiment. For example, if one cage held all the control animals and another all the animals treated with some test substance and a difference was found between the means of the two cages for some character of interest, this might be due to the effect of the treatment, or it might be because the animals in one cage had been fighting, or had a sub-clinical infection not found in the other cage. In such cases any treatment effect is "confounded" or inextricably mixed with an environmental effect.

Randomisation

Once the experimental units have been identified, they need to be randomly assigned to the treatment groups in order to reduce the possibility that one group will have a different environment or be treated differently from another group. Of course, at this stage the sample size should have been determined. This is discussed below. Suppose, for example, an experiment was to be set up to compare three treatments with four animals per treatment, with individual animals being the experimental units. Four ones, four twos and four threes could be written on cards or put in a spreadsheet. They should then be shuffled or randomised using the spreadsheet's random number generator so that animal one receives one of the treatments, at random and so on for all the other animals. As far as possible all further manipulations will be in a random order if the animals are treated in numerical sequence. As far as possible the animals should then only be identified by their number rather than their treatment group. Of course this is not possible if the treatment is being given in the diet or water, but in such cases it is still usually possible for the measurements to be done blind with respect to the treatment.

Caging of the animals

The way in which the animals are caged needs some thought. Single housing is one possibility but this may be stressful for mice and rats. They could be housed with a companion animal which is not part of the experiment, or they could be housed at random in cages with several animals per cage, assuming the treatment is given by injection and there is no scientific objection to housing mice or rats with different treatments in the same box. The disadvantage of housing at random is that if there are cage effects, such as might arise if the animals start fighting, then this will increase the variation ("noise") and reduce the power of the experiment (power is the ability to detect a treatment effect). The animals could be housed with exactly one animal of each treatment group in each box, using a randomised block design. This would be statistically efficient. However, if there is a good biological reason why animals treated differently must be kept separate, for example if they will secrete metabolites into the cage, and the animals are not to be housed singly, then it may be necessary to house, say, two animals of the same treatment group per cage. But in this case the cage becomes the experimental unit because the two animals in the cage cannot have different treatments. On no account should all the animals of each treatment group be housed together as they will no longer be independent experimental units.

Failure to randomise and/or blind can lead to seriously biased results. In one study involving 290 papers, those which failed to blind were 3.4 and those which failed to randomise were 3.2 times more likely to report "positive" results. Those which neither randomised nor blinded themselves were 5.3 times more likely to report "positive" results[7]. This is clear evidence that failure to blind and randomise may lead to bias.

Some variables such as the sex or strain of an animal cannot be randomised. These are called "classification variables". If the aim of the experiment is to compare sexes or strains, then the animals should be similar in all other ways, such as age and source.

High power

A well designed experiment needs high statistical power, i.e. it should have a good chance of being able to detect the effect of the treatment. High power is achieved by controlling the variation ("noise"), using sensitive subjects and possibly high dose rates to obtain a strong response (a good "signal") and therefore a high signal/noise ratio. A large sample size can also increase the power of an experiment but an overly large sample size wastes time and money, so needs to be avoided.

Power analysis

The relationship between the signal, noise, sample size, significance level and sidedness of a statistical test is summarised in a so-called "power-analysis". This depends on a mathematical relationship between six variables and is generally used to determine the optimal sample size, particularly in expensive experiments such as clinical trials. The power analysis is only considered here for a two sample experiment (treated and control) with a measurement end-point such as red blood cell counts or organ weights. It can also be used for discrete data, but it becomes more difficult with several groups and if there are several characters then it is usually necessary to decide which is the most important and base the power of the experiment on that one. The variables that need to be considered are as follows:

1. The effect size on the parameter of scientific interest. This is the difference between the means of the treated and control groups which are of clinical or biological significance. A very small difference would be of no interest, but the investigator would certainly want to be able to detect a large response. The effect size is the cut-off between these two extremes. A large response is easy to detect, but a small one is more difficult so needs larger groups.

2. The acceptable chance of obtaining a false positive result. This is controlled by the significance level. For example, a 5% significance level in a statistical test implies that 5% of experiments are expected to produce a false positive result, suggesting that there is a difference between treated and control group means when in fact it is only due to chance. A 5% significance level is usually used.

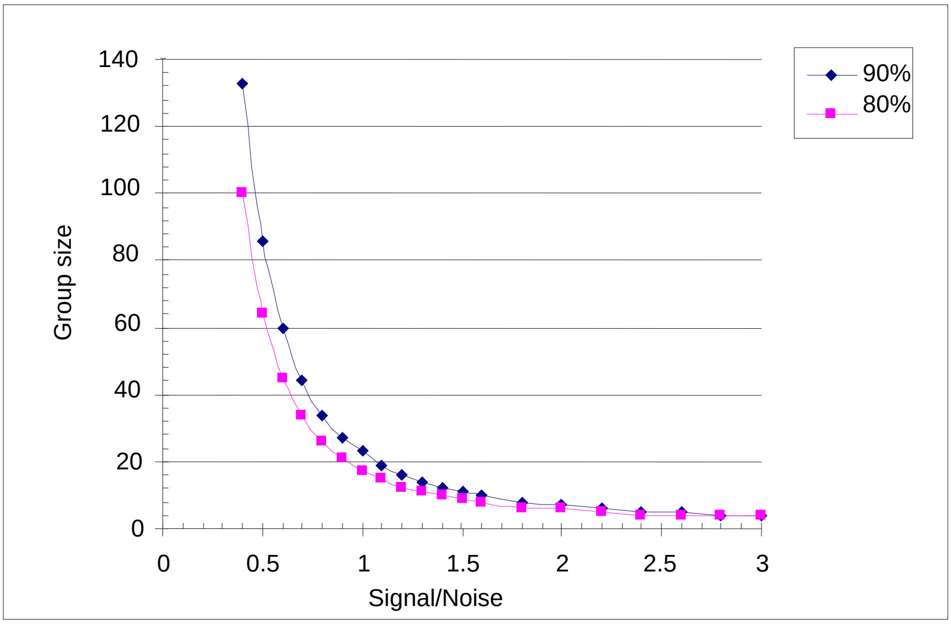
3. The sidedness of the test. A two sided test is used when it is not known in which direction a response will go. A one sided test is used if the response can only go in one direction or is only of interest if it goes in one direction. The default is to use a two sided test.

4. Variability of the experimental material. This is measured by the standard deviation. As the experiment has not yet been done the standard deviation must be obtained from a previous study or from a pilot experiment. This is a weakness of the method.

5. The power of the experiment. This is the probability that the specified effect size of scientific interest will be detected, given the other variables. It is usually set at 80-90%. Higher levels should be used if a failure to detect the effect of a treatment could have serious consequences, such as if the experiment were to fail to detect toxicity.

6. Sample size. A power analysis is generally used to determine an appropriate sample size, but in some cases the sample size is fixed, and the aim is to determine power.

The variables can be combined to produce a curve such as that shown in Figure 1. This shows the required sample size as a function of the signal/noise ratio (the effect of size divided by the standard deviation) assuming a 5% significance level and a two sided test. Curves are shown for 80% and 90% power. It shows, for example, that in order to detect a signal/noise ratio of 1.0 (i.e. the response is as large as one standard deviation), a sample size of 17-23 experimental units would be needed. Large sample sizes are needed if the signal/noise ratio is less than 0.5, but small samples are sufficient if it is as large as 1.5.



Assuming 2-sample, 2-sided t-test and 5% significance

Figure 1. Sample size for comparing two means as a function of the signal/noise ratio for 80% and 90% power assuming a 5% significance level and a two-sided test

An example

Suppose the aim of an experiment is to determine whether a new drug causes anaemia. A previous study found that the mean red blood cell count in CD-1 mice was 9.00 ($x10^{12}/l$) with a standard deviation of 0.68. Suppose also that a reduction from 9.0 to 8.5 counts would indicate anaemia, i.e. the signal is 0.5, then the signal/noise ratio would be 0.5/0.68 = 0.73 and the required sample size (from Fig. 1) would be from about 35 to about 43 mice per group for an 80-90% power. However, some data was also available from inbred C57BL/6 mice which had a mean count of 9.60 but a standard deviation of only 0.25 (averaged across four inbred mouse strains). This would give a signal/noise ratio of 0.5/0.25=2.00 and would require a sample size of only about 6-8 mice per group. Thus, by controlling the genetic variation using inbred mice, the sample size can be substantially reduced.

Use of inbred strains

Inbred strains of mice and rats are much like immortal clones of genetically identical individuals, and are of enormous importance in biomedical research. There are several hundred strains of both mice and rats which are available for research and they are also widely used as the genetic background for mutants, transgenes and knockouts. The properties of these strains in comparison with outbred stocks are as follows:

Inbred strains are

- Isogenic:- i.e. all animals of the same strain are genetically identical. Thus, once one animal has been genotyped, the genotype of all other animals of the strain is known.
- Homozygous:- do not carry any recessive genes unless a mutation has occurred very recently. This means that inbred strains breed true and the parents and offspring are genetically identical.
- Phenotypically uniform:- so that sample sizes can be reduced compare with the use of outbred stocks
- Genetically defined:- so that it is possible to obtain the same genotype repeatedly and because it is known what genes each of the common strains carry, it is possible to ensure that the right animals have been used. This is not possible when using outbred stocks because there is no genetic standard for such animals
- Genetically stable:- it is not possible to change the characteristics of an inbred strain by selective breeding. The only way in which it can change is as a result of a new mutation and these are relatively rare. In contrast, the characteristics of outbred stocks can change rapidly as a result of random genetic drift and selection. Moreover, outbred stocks with the same name from different breeders will be genetically different.
- Well characterised:- there is extensive background data on the properties of most inbred strains and this is available on the internet. The complete DNA sequence of several inbred strains of mice has been determined.
- Internationally distributed:- so that investigators in different parts of the world can repeat each other's work with the assurance that they are using the same animals.

In the year 2000 it was noted that there had been at least 16 Nobel prizes for work which probably could not have been done without using inbred strains[8]. Since then a further seven have been awarded, including prizes for the development of embryonic stem cells in mice and for the development of technology for knocking out genes by homologous recombination.

Inter-individual variability

Power analysis methods can be used to explore the importance of controlling inter-individual variability. For example in one study[9] the standard deviation in body weight of mice housed singly was 6.8g, but when housed in pairs it was only 3.9g. If an experiment were to be set up to study the effect of some treatment on mouse body weight with the aim of being able to detect a difference in body weight of 5g, then the signal/noise ratio in the singly house mice would be 0.86 and in the mice housed as pairs it would be 1.28. These translate into sample sizes of about 30 mice per group if the mice are housed singly or 14 mice per group if they are housed in pairs assuming a 90% power, a 5% significance level and a two-sided test.

In another study, 58 groups of rats, each of about 30 individuals were killed and the kidneys were weighed. Some of these rats were infected with *Mycoplasma pulmonis* and the kidney weight had a standard deviation of 43.3 units (the data had been standardised to a mean of 100), while others were free of *Mycoplasma* and had a standard deviation of 18.6 units. Assuming an experiment was to be set up with a compound which it is thought might affect the weight of the kidney and the aim was to be able to detect a 10 unit difference between the means of a treated and control group with a 5% significance level and a power of 80%, then it would require 298 rats per group using the diseased animals but only 55 per group using the disease-free animals. Of course, these are still large sample sizes. Maybe the experiment should be set up to detect a 20 unit rather than a 10 unit change in kidney weight. In this case it is easy to re-calculate the figures and find that using the *Mycoplasma* infected animals would require a sample size of 75 animals per group, but the experiment could be done using only 15 disease-free animals per group.

Randomised block designs

Another method of controlling variability and reducing sample size is to use randomised block experimental designs. These are widely used in agricultural research and should be more widely used in work involving laboratory animals. They are quite common in *in-vitro* work where the investigator repeats the "experiment" (actually a block) on several occasions, but often these workers do not know how to do the correct statistical analysis.

Suppose, for example that the aim was to study the effect of a drug on open field activity or some other quantitative characteristic such as haematology or clinical biochemistry. It may be difficult to obtain sufficient animals in a narrow weight range and there may be a limitation on the handling of the animal in the time available. The solution is to split the experiment up into a number of smaller "mini-experiments" each done with animals chosen to be as similar as possible. Typically, each block of a randomised block design has one animal (or other experimental unit) of each of the treatments. For example, if there is to be one control and two dose levels, then a block will consist of three animals selected to be as similar as possible and these are assigned at random to the three treatments. This is repeated several times, depending on the required sample size. Differences between blocks can then be removed in the statistical analysis using a two-way analysis of variance without interaction.

Clearly, it is possible to reduce sample size, and therefore the number of animals used, by reducing inter-individual variation using inbred, disease-free animals of similar weight and age and randomised block designs where necessary. A high dose can be given to produce a large response, provided it makes biological sense. But this can only be done to a limited extent in toxicity testing as there may be non-linearity in the dose response curve. As will be seen in the next section, it may also be possible to use several strains in the hope of finding one or more which is particularly sensitive, without increasing the total number of animals which are used.

The Resource Equation method of determining sample size

A power analysis is usually the preferred method of determining sample size, particularly for expensive experiments such as clinical trials. However, this can be difficult where more complex experimental designs are employed as can happen in more fundamental research projects. For example, if there are several different treatment groups, it can be difficult to specify the effect size (signal) that would be of scientific interest and if many characters are to be measured, it may be difficult to decide which is the most important. The power analysis also requires a reliable estimate of the standard deviation, so it cannot be used if this is not available. In many disciplines there is a tradition of simply using six or eight animals in each group, but this can be very wasteful if there are more than two groups.

The Resource Equation method provides a much better alternative for experiments with a quantitative outcome (i.e. using measurement data). It depends on the law of diminishing returns. Adding one more experimental unit to a small experiment will give good returns, but as the experiment gets larger the value of adding one additional unit diminishes. The resource equation is:

E = (total number of experimental units)
-(number of treatment groups)

E should normally be between 10 and 20, although it can be greater than 20 if the cost of an experimental unit is low (e.g. if it is a well in a multi-well plate) or in order to ensure a balanced design with equal numbers in each group. As an example, suppose an experiment is to be set up to study the effect of four dose levels of a compound on activity in male and female mice. This is a factorial design (discussed below), and it involves eight groups (4 doses x 2 sexes). How many animals should be used in each group? According to the Resource Equation if there were, say, three mice per group, that would involve a total of 24 mice and with eight groups E=24-8 = 16. So this would be an appropriate number. Of course, these animals should be chosen to be free of disease, of uniform weight and preferably of an inbred strain.

Factorial experimental designs

Factorial experimental designs are ones where there are two or more factors (treatments or conditions) being varied at the same time. Hypothetically, a drug treatment may be given at, for example, four dose levels to, for example, three strains of mice. In this case, “drug” is a factor with four “levels”, “strain” is a factor with three “levels” and the experiment is a 4x3 factorial design. Factorial designs can have any number of factors, each at any number of levels. Such a design could theoretically be used to explore interactions between drugs. Drugs A, B and C could each be given at two levels (none and some) in a 2x2x2 factorial design, resulting in eight means to be interpreted. The aim in this case would probably be to study whether the responses in the animals were additive, or whether the drugs potentiated each other. The data would be analysed using a three-way analysis of variance. This would show whether there were significant “main effects”, i.e. whether there was a significant difference between treated and control means for each drug averaged across the other drugs and whether there were interactions between the drugs, i.e. whether the magnitude of the response to each drug depended on whether one or other of the drugs had been administered.

It does not seem to be widely understood that factorial designs can often be used without increasing the total number of animals which are used. For example, instead of doing a simple experiment with a control and a treated group using eight male mice of a particular strain (a single factor design), it would be more informative to use four males and four females in each treatment group in a 2x2 factorial design. This design would show whether there was a response to the treatment averaged across both sexes whether males and females differed averaging across the treatments and whether the response to the treatment was the same in the two sexes.

According to R.A. Fisher[10] (*By using a factorial design*)”.... an experimental investigation, at the same time as it is made more comprehensive, may also be made more efficient if by more efficient we mean that more knowledge and a higher degree of precision are obtainable by the same number of observations."

As an example, the haematological response of mice to chloramphenicol was studied using two parallel experiments[11]. In the first experiment eight outbred CD-1 mice were given the compound at six dose levels and haematological counts were done but for simplicity, only the results for highest dose level are given here. In the second experiment two mice of each of four inbred strains were similarly treated with the compound at the same dose levels. Only the white blood cell counts are presented in Table 1. Originally there were eight inbred mice at each level, but in order to make the two experiments comparable, two inbred mice of each strain were selected at random to represent the 2 (doses) x 4 (strains) factorial design using the same total number of animals as in experiment 1.

Experiment 1 involved 16 CD-1 outbred mice, with eight controls and eight treated with chloramphenicol. The signal (difference between means of treated and control mice) was 0.40 counts and the noise was the pooled within-group standard deviation of 0.86, so the signal/noise ratio was 0.47 and from a t-test the p-value was 0.38 so the difference was not statistically significant at the 5% level.

Experiment 2 also involved eight control and eight treated animals but in this case, each group of eight consisted of two animals of each of the four strains. Looking as the average across strains the signal was 0.73 and the noise estimated as the square root of the error mean square in the two-way analysis of variance used to analyse the experiment was 0.34, giving a signal/noise ratio of 2.15, leading to a p-value of <0.001, i.e. the difference was highly significant. In this experiment the signal was larger than in experiment 1 because the CD-1 mice were relatively resistant compared with the average of the four inbred strains and the noise was lower because the inbred mice were more phenotypically uniform than the outbred mice. However, the response of the inbred mice was not uniform. The BALB/c mice had lower basal counts and did not respond to the chloramphenicol and the strain x dose interaction was highly significant (p<0.001). Thus there was evidence that the inbred strains differed in sensitivity. Strain differences are mostly due to genetic factors.

So in conclusion, the multi-strain study is more powerful because it used inbred strains with a lower noise and on average, a higher signal so was able to detect the effect of chloramphenicol and it was also more informative as it showed that there is genetic variation in response. Had a single inbred strain been chosen, then different conclusions would have been reached, depending on the strain. Had BALB/c been chosen then the effect of chloramphenicol would have been missed.

One point worth noting is that some investigators are doubtful about using what they consider to be group sizes of only two inbred mice. However, each CD-1 mouse is genetically different so if genotype is considered to be a grouping factor, then in the CD-1 mice the group size is only one. This is simply ignored, and the results are average across these genotypes because they are not recognisable.

Statistical analysis

The statistical analysis of an experiment should be planned at the time that the experiment is designed and no scientist should start an experiment unless he or she knows how the results will be analysed. To do so is asking for trouble. They may find that they do not have the tools or the know-how for the analysis so that it does not get done correctly. They may put off the analysis until they have done several similar experiments, but in this case they will be unable to adjust conditions according to results observed in the earlier experiments.

Table 1. White blood cell counts in control mice and mice treated with chloramphenicol at a dose of 2500 mg/kg in two parallel experiments. Experiment 1 was done using outbred CD-1 mice and experiment 2 was done using two mice of each of four inbred strains at each dose level

Experiment 1. Results of a single factor experiment using eight mice at each dose level

Strain	N	0	2500	Signal (Difference)	Noise (SD)	Signal/noise	p
CD-1	16	2.23	1.83	0.40	0.86	0.47	0.38

Experiment 2. Results of the factorial design using two mice of each of four inbred strains at each dose level.

Strain	N	0	2500	Signal (Difference)	Noise (SD)	Signal/noise	p
CBA	4	2.25	0.30	1.95	0.34	5.73	
C3H	4	2.15	0.40	1.85	0.34	5.44	
BALB/c	4	1.05	1.35	-0.30	0.34	(0.88)	
C57BL	4	2.25	0.95	1.30	0.34	3.82	
Mean	16	1.93	1.20	0.73	0.34	2.15	<0.001
Dose * strain							<0.001

The statistical analysis should be done using a reputable commercial statistical package available to the investigator. Alternatively, investigators may consider learning to use one of the reputable open source (free) software packages available, some of which can be extremely versatile and reliable, but seem to have the serious disadvantage of being more difficult to learn as they are often command driven rather than menu driven and so may not be suitable for a statistical novice, except as part of a formal course on statistics.

The statistical analysis should first involve screening and examining the raw data once it has been put into the computer to make sure that it has no mistakes and to obtain a general feel for it. Graphical methods of presentation are extremely useful at this stage. Next, the data should be summarised with means, medians or counts, depending on the nature of the data. Finally, the statistical analysis should be used to assess the significance of any differences among groups.

In this paper it has mostly been assumed that the data are measurements. In this case there is a good chance that the results can be analysed using an analysis of variance, the most widely used statistical method. All research scientists need a basic understanding of this technique which is described in all good statistical texts if they are to be able to design and analyse their own experiments.

Reference List

1. Russell, W. M. S. and Burch, R. L. (1959). "The principles of humane experimental technique," Special Edition, Universities Federation for Animal Welfare, Potters Bar, England.

2. Mead, R. (1988). "The design of experiments," Cambridge University Press, Cambridge, New York.

3. Altman, D. G. (1982). Statistics in medical journals. *Statistics in Medicine* 1, 59-71.

4. Festing, M. F. W., Overend, P., Gaines Das, R., Cortina Borja, M., and Berdoy, M. (2002). "The Design of Animal Experiments," Laboratory Animals Ltd., London.

5. Cox, D. R. and Reid, N. (2000). "The theory of the design of experiments," Chapman and Hall/CRC Press, Boca Raton, Florida.

6. Pritchett-Corning, K., Chang, F. T., and Festing, M. F. W. (2009). Breeding and Housing Laboratory Rats and Mice in the Same Room Does Not Affect the Growth or Reproduction of Either Species. *Journal of the American Association for Laboratory Animal Science* 48, 1-7.

7. Bebarta, V., Luyten, D., and Heard, K. (2003). Emergency medicine animal research: does use of randomization and blinding affect the results? *Acad. Emerg. Med.* 10, 684-687.

8. Festing, M. F. W. and Fisher, E. M. C. (2000). Mighty mice. *Nature* 404, 815.

9. Chvedoff, M., Clarke, Mr., Faccini, J. M., Irisari, E., and Monro, A. M. (1980). Effects on mice of numbers of animal per cage: an 18-month study. (preliminary results). *Archives of Toxicology, Supplement* 4, 435-438.

10. Fisher, R. A. (1960). "The design of experiments," Hafner Publishing Company, Inc, New York.

11. Festing, M. F. W., Diamanti, P., and Turton, J. A. (2001). Strain differences in haematological response to chloramphenicol succinate in mice: implications for toxicological research. *Food and Chemical Toxicology* 39, 375-383.