1

Introduction

Many researchers, either directly or indirectly, rely on statistical ideas when carrying out animal experiments. While some statistical tools are well known and are applied routinely, other tools are less well understood and so are less well used. The overall aim of this book is to discuss statistical methodologies that can be applied throughout the many stages of the experimental process. Researchers should be able to carry out most of the techniques described, although the advice of a professional statistician is advisable for some of the more advanced topics. Making use of these techniques will ensure that experiments are conducted in a logical and efficient way, which should result in reliable and reproducible decisions.

The particular types of study addressed in this book, as the title suggests, are studies involving animals. We attempt to cover all of the statistical tools that the animal researcher should use to run successful studies. Of course many of the problems faced by the animal researcher are common to other disciplines, and hence the ideas contained within this book can be applied to other areas. It should be noted that certain topics described in the text have been simplified to allow non-statisticians to apply the ideas without professional statistical support. Such pragmatic descriptions, while simplifying the technical details, are not universal and will not be applicable in all scientific disciplines.

There has been much interest in the use of statistics in animal research, in particular in the application of the 3Rs, replacement, reduction and refinement, as described by Russell and Burch (1959):

Every time any particle of statistical method is properly used, fewer animals are employed than would otherwise have been necessary.

Many authors have since highlighted how important the use of good experimental design is when conducting animal experiments; see Festing (1994, 2003a, 2003b) and the references contained within. Some of the more practical, as well as statistical, aspects of experimental design and statistics when applying the 3Rs are described in the book by Festing *et al.* (2002). There have also been surveys into the use of statistics in refereed journals; see McCance (1995) and more recently Kilkenny *et al.* (2009). The latter draws attention to some of the mistakes that can be made by researchers when designing and analysing animal experiments. The reliability of the reporting of animal experiments has been considered in, for example, Macleod *et al.* (2009) and Rooke *et al.* (2011). These articles highlight that papers describing experiments that do not employ suitable randomisation techniques and/or blinding may contain biased results.

The main goal of this text is to demonstrate how statistics can aid the reduction and refinement of animal studies. The efficient use of statistics, both in terms of complex experimental design and powerful statistical analysis, can reduce the number of animals required. Statistics can also help the researcher understand the processes that underpin the animal model and help identify factors that are influencing the experimental results. Such an understanding will inevitably lead to a refinement in the experimental process and a reduction in the total number of animals used.

Statistics, as a discipline, provides researchers with tools to help them arrive at valid conclusions. However, statistics, along with the application of some common sense, can also increase the understanding of the animal model through the application of graphical and mathematical techniques. For example, graphical tools play an important role in helping the researcher understand the effect of the features of the experimental design and also uncover overall patterns present in the data. The application of a formal statistical test, without first investigating the data graphically, can lead to the researcher drawing incorrect conclusions from the data. Consider the following real-life case study, which used graphical, as well as statistical, tools. If a conventional statistical analysis had been carried out, without first investigating all of the information gathered within the experiment, then the conclusions would have been misleading.

Example 1.1: Reducing blood cholesterol levels in mice

A scientist wanted to test the hypothesis that a novel compound had a beneficial effect on reducing high-density lipoprotein (HDL) cholesterol levels in a transgenic C57Bl/6J strain of mice. A blood sample was taken pre-treatment and the baseline cholesterol level for each animal measured. The mice were then randomised to either the drug treatment group or the control group and dosed with either the drug treatment or vehicle twice daily for two weeks. At the end of this period, a terminal blood sample was taken and the HDL cholesterol level measured.

As the scientist wanted to make use of the baseline information in the statistical analysis, it was decided that the percentage change from baseline would be a suitable response to investigate. This would, the scientist hoped, effectively remove the animal-to-animal differences by normalising to the baseline level. While there was evidence of a decrease in HDL cholesterol level in the group of animals administered the drug treatment (a 20% decrease from baseline in the drug treatment group compared to a 10% decrease in the control group) this was not deemed statistically significant using an unpaired t-test ($p = 0.191$). A means with standard errors of the mean (SEMs) plot of the data (see Section 5.3.5) is presented in Figure 1.1.

As a follow-up the scientist also analysed the terminal HDL cholesterol level. From this analysis it appeared that there was a statistically significant increase in cholesterol level in the drug-treated group compared to the control. A plot of the means with SEMs of the terminal HDL cholesterol level is presented in Figure 1.2.

Based on the results of this experiment, should we conclude the drug increases cholesterol levels? And why did the two analyses give such different conclusions? These questions can be answered

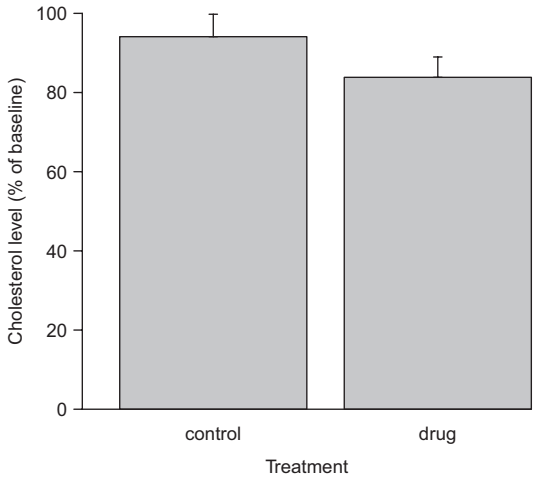

Figure 1.1. Plot of treatment means with standard errors for the percentage of baseline cholesterol response for Example 1.1.

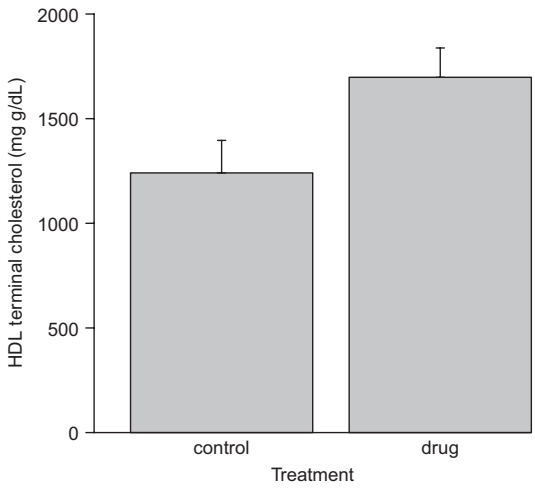

Figure 1.2. Plot of treatment means with standard errors for the terminal HDL cholesterol for Example 1.1.

by a simple scatterplot of the measured HDL cholesterol levels. If we plot terminal vs. baseline HDL cholesterol levels, an underlying problem with the experiment becomes clear. The scatterplot is presented in Figure 1.3.

From Figure 1.3 it can be seen that there are two distinct groupings along the X-axis. The plot reveals that, in terms of the HDL baseline cholesterol level, the animals belong to one of two

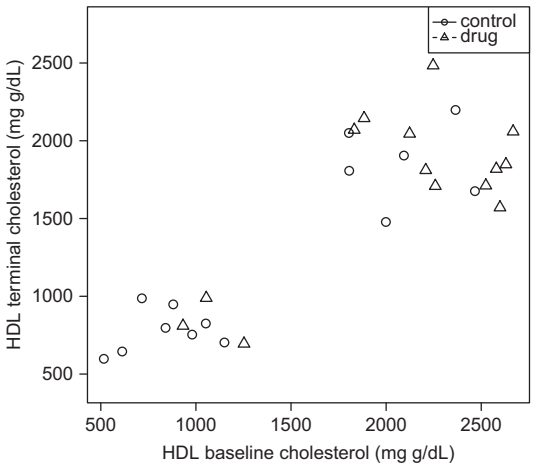

Figure 1.3. Scatterplot of terminal HDL cholesterol vs. baseline HDL cholesterol, categorised by treatment for Example 1.1.

sub-populations. Unless we are careful how these baseline differences are accounted for, we could draw incorrect conclusions from the analysis. Given that there appears to be a correlation between baseline and terminal cholesterol levels, this baseline difference is probably the most important feature of the experiment that influences the conclusion – perhaps more so than the treatment effect itself. The treatment effect observed in the experiment will be influenced by the allocation of the mice (within each sub-population) to the treatment groups. In this case most of the mice that were allocated to the novel drug group were from the sub-population with the high baseline level. So it is not surprising that, when analysing the terminal HDL cholesterol level, it appears that the terminal cholesterol level is higher in the treatment group. Obviously the researcher was unlucky that the randomisation of the mice to the treatment groups produced such an allocation.

The solution is twofold. Firstly, and most importantly, the researcher should try to identify what is causing the baseline differences. We can then account for this effect in the experimental design. However, if we fail to identify what is causing the baseline differences, then the randomisation should be carried out so that the treatment replication is equal in both sub-populations. This will, of course, depend on whether the baseline information is available when allocating the mice to the treatment groups.

If the researcher produces the scatterplot shown in Figure 1.3, then it will become apparent that the treatment effects may be due to baseline cholesterol levels. However, without such a graphical investigation of the data the problem may not have been identified. It is important at this stage of the book to note that a valid statistical investigation of a dataset is more about understanding the information contained within the data. It does not just involve the calculation of p-values. Graphs can be the best and simplest way to achieve this and should always be considered first, ideally by plotting the individual data points.

In reality the treatment allocation observed in this example is quite extreme and perhaps indicates a biased selection process. It should be noted, however, that there is always a chance, however small, that the randomisation will generate a significant treatment effect due to differences at baseline. This will occasionally happen, even when the allocation process is valid. As long as the randomisation is performed correctly, then we should not be too concerned that effects present at the baseline will influence the treatment comparisons.

1.1 Structure of this book

The majority of the remainder of this text is split up into three main chapters. In Chapter 3 we describe families of experimental designs that can be employed when conducting animal research. The chapter consists of a description of each design and practical examples of their use. Also given is an explanation of when and where to apply each design. The section attempts to introduce each experimental design, without overuse of mathematical terminology.

In Chapter 4 some general issues involving randomisation are discussed. We consider why the experimental material should be randomised and describe the influence this has on the statistical analysis. Techniques that can be employed to perform the randomisation are also given.

When conducting a statistical analysis, one of the first steps in the process is to define the statistical model that will be used to explain the observed data. There are several ways to justify the choice of statistical model. Given that the animal researcher has control over the experimental design, it seems sensible to make use of the design when deciding which statistical model to apply. One way of linking the experimental design to the statistical analysis is by considering the randomisation applied to the experimental material. Most analyses, and certainly those considered in this book, make assumptions about the allocation of animals to treatment groups. For a valid statistical analysis of a designed experiment, a suitable randomisation should have been carried out.

Chapter 5 describes the statistical analysis techniques that the researcher should employ when analysing data generated using the designs discussed in Chapter 3. We approach the subject in a practical way, without the use of mathematical formulae. We assume the researcher has access to an advanced statistical package, such as InVivoStat, to compute all numerical results. The tools described in this book are flexible enough to cope with most experimental situations. Any assumptions made during the statistical analysis are also discussed. When these assumptions do not hold alternative approaches are given.

In Chapter 6 we describe how the researcher can perform the analyses discussed in Chapter 5 using the InVivoStat statistical software package. For each InVivoStat module the analysis procedure is described, including input and output options, and a worked example given. Where appropriate, a technical description of the implementation of the analysis methodology is also presented.

This text contains many ideas that the researcher may need to employ during the course of the experimental process. Some of the techniques will be applied frequently during routine work and hence will be of interest to all readers. Other sections describe techniques that are more advanced and would only be used occasionally, for example when setting up a new animal model.

1.1.1 Introductory sections

The following sections should be read by the casual reader who wishes to get a simple overview of the ideas contained within this text. These sections provide a flavour of some of the more advanced sections.

Sections 2.1 and 2.3 – Statistical concepts

Readers should familiarise themselves with these sections as they provide the framework for all following sections on experimental design and statistical analysis.

Section 3.1 – Why design experiments?

This section is an introduction into why we should be designing animal experiments and some information on the benefits that can be gained from an understanding of experimental design.

Section 3.3 – Summary of design types

This section introduces the types of experimental design that are available to the researcher. Information is given on when and why they should be used.

Section 3.7.2 – Sample size and power

This section discusses the factors that influence sample size and gives information on how to calculate suitable sample sizes in animal experiments.

Sections 4.2 and 4.4 – Randomisation

These sections describe why we need to randomise our experimental material and give practical examples of how to carry out the randomisation.

Section 5.1 – Introduction into statistical analysis

This section is a description, including a worked example, of our preferred analysis procedure using the InVivoStat software package.

Section 5.4 – Parametric analysis

This section describes the types of parametric analysis, some of their properties and gives information on when to use them.

1.1.2 Approaches to consider when setting up a new animal model

When setting up a new animal model, or perhaps trying to replicate a model described in the literature, there may be many factors that influence the animals' responses which will need to be quantified. Perhaps the researcher needs to decide which sex to use, how old the animals need to be, how long to dose the animals prior to testing… the list goes on. A common approach taken by researchers is to investigate each of these factors one at a time. However, there are better and more efficient (not to say more informative) ways to conduct these investigations. Use:

- Large factorial designs (Section 3.5.4) to assess factors and factor interactions and to maximise the window of opportunity;
- Nested designs (Section 3.7) to decide on the replication required within the experiment;
- Power analysis to select sample sizes (Section 3.7.2);
- Parametric analysis tools, such as ANOVA (Section 5.4.2), repeated measures analysis (Section 5.4.4) or graphical tools (Section 5.3) to investigate how the factors relate to each other.

1.1.3 Approaches to consider when generating hypotheses

Once the animal model has been set up, the researcher might wish to start generating hypotheses. Consider using:

- Small factorial designs (Section 3.5.3) to assess interactions between factors of interest;
- Block designs to reduce variability and allow the researcher to manage experiments more efficiently (Section 3.4);
- Parametric analysis tools, such as ANOVA (Section 5.4.3), repeated measures analysis (Section 5.4.4) or graphical tools (Section 5.3) to investigate how the factors relate to each other.

1.1.4 Approaches to consider when testing hypotheses

When testing specific hypotheses, the researcher should be stricter (in the analysis) to avoid generating false positive results. Consider using:

- Block designs to reduce variability and manage the experiments more efficiently (Section 3.4);
- Parametric analysis tools, such as ANOVA (Section 5.4.3) or repeated measures analysis (Section 5.4.4);
- Planned comparisons or other suitable multiple comparison procedures to compare individual group means (Section 5.4.8).

1.2 Statistical problems faced by animal researchers

From a statistical point of view the animal researcher faces two major issues. The first problem is that there will usually be substantial animal-to-animal variability. If two animals are given the same treatment regime, then they will respond in subtly different ways. This, combined with the ethical imperative to use as few animals as possible, will cause many problems for the researcher.

The level of animal-to-animal variability varies between different animal models, disease areas, species and even batches of animals. This has, perhaps, been reduced by the advent of inbred isogenic strains, which has meant that the animals themselves are less phenotypically variable (Festing *et al.*, 2002, pp. 17–26). However, it is probably correct to assume that this source of variability will still be large in most animal experiments. The researcher should aim to quantify the size of the animal-to-animal variability but also try to discover any other sources of variability within the experiment that will increase this.

Once all the sources of variability have been identified and quantified, another problem is determining the sample size required for the experiment. Sadly, many assume the advice of a statistician will be to increase sample size, regardless of the practical implications! While there are benefits to be gained from increasing sample size (for purely statistical reasons) there may be other techniques that statistics can offer, other than simply increasing animal numbers, which will improve the reliability and reproducibility of the experimental results.

In many animal experiments there is one statistical 'saving grace' that can be used to reduce the impact of high variability and small sample sizes, and that is the experimental design. Researchers usually have almost complete control over the experimental design used. For example, animals can be ordered so that they arrive from the supplier at set dates, unlike clinical trials where patients need to be enrolled. Researchers can also plan how the study is conducted. If the study is completed over two days, or two pieces of test equipment are used or two surgeons perform the surgery, then these can be taken into account when planning the study. Hypotheses that are to be tested are planned well in advance and so designs can be tailored to suit the questions being answered. Many characteristics of the animals are also recorded before the start of a study for welfare reasons, for example age and body weight. So there

is extra information about the animals themselves that can be used in the analysis of the study data.

In conclusion, if the researcher is to avoid the statistical pitfalls of high variability and small sample sizes, then it can be argued that the use of good experimental design (and the appropriate statistical analysis of the data generated when using such designs) is more crucial in animal research than in many other scientific disciplines.

1.3 Pitfalls encountered when applying statistics in practice

There are many pitfalls that may trap the unwary researcher when carrying out animal research. The following examples are taken from a number of published sources, namely Festing *et al.* (2002, pp. 11–16), McCance (1995), Gaines Das *et al.* (2009) and Kilkenny *et al.* (2009), along with the authors' own experiences.

1.3.1 Pitfalls with experimental design

Using appropriate designs at specific points in the experimental process

Many researchers fail to employ the right design at the right time. This can lead to using more animals than is necessary and can undermine the reliability of the experimental conclusions. For example, when setting up a new animal model, or revising an existing one, there are certain types of design that can be used to investigate the many hypothesised factors that could influence the experimental results. These designs, the so-called large factorial designs, provide a quick, easy and systematic way of developing knowledge of the animal model. If the researcher fails to use these designs, then it may take longer to fully understand the animal model. Factorial designs are discussed later in this book (see Section 3.5).

Failure to account for nuisance effects in the design

Most researchers can probably list nuisance effects that could be accounted for in the study design. For example, animals may be housed in different cages or rooms, be selected from two or more litters or be operated on by one of two surgeons. The list can go on. It is important to check if these nuisance effects have an effect on the measured response. If not then they can be ignored and future designs simplified, otherwise strategies should be developed that take them into account.

If a design is not planned in advance, then comparisons between the treatments may become influenced (or biased) by other unwanted nuisance effects that were not taken into account. It may be the case that a nuisance effect cannot be separated from the treatment effect in the statistical analysis and so the treatment effect cannot be reliably assessed. We say that the two effects are *completely confounded* with each other. In the worst case scenario the observed treatment effect may be wholly due to the nuisance effect.

Consider, for example, an experiment where the control animals were tested on one piece of equipment and the treated animals are tested on a second. Any treatment differences observed could be due to, or influenced by, differences between the two pieces of equipment. Unfortunately if such a design has been used then there is no statistical way of testing for this bias.

As a rule of thumb if the results from an experiment appear unusual, then there may be an underlying nuisance effect that is influencing the results.

Experiments done on an *ad hoc* basis

Some researchers do not take a systematic approach when planning a series of studies. Rather than plan them in advance, studies are carried out in a piecemeal fashion. For example, in a series of drug trials, higher doses of the compound are investigated by conducting extra studies, rather than including all doses in the design at the beginning of the experiment. In this situation, the dose-response relationships are assessed across studies, and hence any study-to-study differences will influence the assessment of the dose-response relationship.

Control groups not used correctly

The purpose of a control group is to allow treatment effects to be assessed in the absence of any other

experimental effects. To achieve this, the control group must be exposed to exactly the same conditions as the treatment groups (to allow the treatment comparisons to be unbiased). For example, in rodent studies it is often the case that the animals are housed in racks of cages. Each rack is allocated to a single treatment to avoid cross-contamination. However, if the racks containing the control animals were placed nearest to the door of the animal room, then these animals may be more disturbed than the treated animals. This could bias the treatment comparisons.

Inefficient choice of treatment groups

If there are two or more factors of interest in a study, then it is recommended that all combinations of the factor levels are included in the design. For example, consider an experiment where there are two factors: Drug (levels: vehicle and compound) and Strain (levels: transgenic and wildtype). It is important that where possible all combinations of the two factors are included in the design, i.e. vehicle + transgenic, vehicle + wildtype, compound + transgenic and compound + wildtype. This is an example of a small full-factorial design, as discussed in Section 3.5.3. If a combination of the factor levels is not included in the final design, then drawing inferences from the analysis can become more difficult. The sensitivity of the statistical analysis to identify significant treatment effects can also be compromised.

Too few animals per group

If the sample size is too small, then the experiment will lack sufficient statistical power to detect a real treatment effect (see Section 3.7.2). Running a study with too few animals is a waste of animals as well as the researcher's time and resources (Button *et al.*, 2013). A power analysis, as described in Sections 3.7.2 and 6.8, should be completed before running a study to confirm the sample size is large enough to achieve meaningful results. There is at least anecdotal evidence that suggests researchers generally underestimate the sample size required when conducting animal experiments. It is preferable to conduct one or two large (and reliable) studies instead of a series of smaller inconclusive ones.

Too many animals per group

It should be remembered that, when running a statistical analysis, it is possible that a biologically irrelevant effect could be declared statistically significant if the sample size is too large. The researcher should begin the planning process by identifying the level of biological relevance. For example, perhaps a drug that causes a 20% change from control is of interest and merits further investigation. If an estimate of variability is available, then an appropriate sample size can be selected so that the statistical analysis should generate a statistically significant result only when a biologically relevant effect has been observed. Failure to take biological relevance into account when designing a study can lead to oversensitive tests. Such tests will declare statistical significance when the biological effect is not large enough to be of practical interest. In practice it is perhaps more likely that the sample size in animal experiments will be too small than too large.

Failure to recognise the true structure of the design

In some experiments complex experimental designs are used and it can be difficult for the researcher to recognise the structure. The replication of the factors in the study may not have been chosen using a suitable technique and hence the statistical analysis may be less powerful than it could otherwise have been. For example, an experiment was planned to assess two types of flooring in guinea pig cages. There were 30 guinea pigs available for inclusion in the study. Animals were group housed and their preference to the floor types assessed individually by measuring the time spent in either half of the cage. By considering the experimental design it was found that the sensitivity of the statistical tests could be improved, without increasing the total numbers of animals used, if the guinea pigs were housed in pairs rather than four per cage, as originally planned.

Trying to do too much with limited resources

Occasionally the researcher will try to achieve too much in a single study. This can cause problems if there are a fixed number of animals available to use. For example, a study was planned to assess the effect of a treatment on plaque deposition in the brains of a strain of transgenic mice. It was hoped that the treated group could be compared to the control group at five distinct time points (2, 3, 4, 6 and 12 months of age). However, only 40 mice from the transgenic strain were available for inclusion in the study. If the ten groups (five time points by two treatments) were included in the study, then there would only be four mice per group per time point. This would not have been enough to detect biologically relevant effects, assuming testing differences between treatment and control was the purpose of the study. Choosing three time points, say 2, 6 and 12 months, would allow either six or seven mice per group and still allow possible differences with age to be detected.

Ignoring the possibility of within-animal testing

In certain situations it is possible to administer more than one treatment to each animal. This can be achieved using a crossover design (i.e. testing a sequence of treatments over time on each animal; see Section 3.4.9) or a dose-escalation design (see Section 3.8.2). With such designs it is important to allow sufficient time gaps between test periods to allow the treatment effects to wash out of the animals' biological systems. In theory each animal should return to approximately its baseline level before receiving the next treatment.

Alternatively if treatments can be applied locally, for example when assessing the effect of cream treatments on a skin condition, then more than one cream can be tested at the same time in each animal. In both cases comparisons between treatments can be made within-animal. This removes any animal-to-animal variability from the assessment of the treatment effects and generally provides more sensitive tests. Ignoring the possibility of testing multiple compounds in the same animal could seriously compromise the experimental results and increase overall animal use.

Quality of responses

The type of response measured in the experiment should be considered at the planning stage. As a general rule numerically continuous responses contain the most information, see Section 3.2.1, as they can be observed at many values. The researcher can therefore differentiate subtle effects when using this type of response. A response that is discrete, ordinal or binary (see Section 3.2.1.1) will be measured on a scale that has fewer distinct values. It is therefore more difficult to observe experimentally induced small changes and hence these responses contain less information. Such experiments will require more animals to achieve the same level of statistical sensitivity (Festing *et al.*, 2002). Also the statistical tests available to analyse discrete, ordinal or binary responses can be less powerful than those available for continuous ones (see Sections 5.5.1 and 5.5.2).

For example, consider a study to assess the effect of transporting rats from a supplier to the test establishment on the formation of lesions in the liver. Let us assume the researcher wants to assess the severity of the lesioning. The initial plan was to count the number of animals showing lesions (a yes/no binary response). However, counting the total number of lesions per animal would contain more information (a count response is on a numerical scale). Such responses can be analysed using more powerful statistical analysis techniques, such as ANOVA (see Section 5.4.3), and hence fewer animals would be required. Better still, if an imaging technique such as magnetic resonance imaging (MRI) were used to measure the total lesion volume per animal (a continuous numerical response) then even fewer animals would be required.

Designs chosen through habit

The experimental design being used should always be questioned and may change as new information becomes available or practical techniques are refined. A review should be conducted after the initial study data have been analysed and any nuisance effects (either proven or suspected) should be accounted for

in follow-up studies. A design should not be selected simply because it has been used extensively in the literature.

Unusual designs

It has been suggested that journal referees are unwilling to publish results from unusually designed studies. If the author has included a specific description of the design in the manuscript, and given reasons why it was selected, then we argue that referees should feel more confident in the results obtained. Perhaps in future as researchers become familiar with the benefits of using complex designs, then there will be fewer unusual cases.

Internal validity

Internal validity is defined as the extent to which the design and conduct of the experiment eliminates the possibility of bias (van der Worp *et al.*, 2010). If the experiment is internally valid then any observed treatment effects (when compared to a suitable control) should be purely due to the treatment itself and not other unforeseen effects. To avoid bias, studies should be randomised (to avoid selection bias) and blinded (to avoid performance and detection bias). The latter should ensure that the researcher's beliefs do not, however subtly, influence the outcome of the experiment. There is evidence that failure to blind an experiment correctly can result in an apparent increase in treatment efficacy; see Rooke *et al.* (2011) for example.

External validity

Assuming an experiment has been blinded and the randomisation performed correctly (hence the experiment has good internal validity), then there is still a risk that the external validity of the experiment will be questionable. Van der Worp *et al.* (2010) define external validity as: 'the extent to which the results of an animal experiment provide a correct basis for generalisation to the human condition'. In many areas of animal research there are, perhaps valid, concerns about the reliability of animal models to predict responses in human patients. For example, the use of a model for inducing a disease that is not sufficiently similar to the human disease could result in development of test compounds that work in animals but not humans. While such practical considerations are beyond the scope of this text, a suitable experimental design can help avoid such problems. If both males and females were included in the experimental design and statistical analysis, for example, then this would avoid the problem described in van der Worp *et al.* (2010) where only male or female animals were used in an animal experiment whereas the disease itself occurred in both male and female human patients.

1.3.2 Pitfalls with randomisation

Randomising when designing is actually better

Sometimes it is easier to rely on the randomisation to remove the influence of a nuisance effect, rather than include a factor in the experimental design that will account for it. If a factor is included in the experimental design and subsequent statistical analysis, then the size of the effect can be assessed at the analysis stage and its influence on the experimental results removed.

Consider an experiment where rats are shown a series of visual stimuli over a set period of time, some of which provide a food reward. The stimuli could be shown to the rats in a random order. However, if the order was planned and controlled then the researcher could account for time and learning effects in the statistical analysis.

Failing to randomise

The process of assigning animals to treatment groups should be done at random, preferably using a randomisation technique such as picking balls from a bag. Selecting animals at random from the cage is not truly a random process and could introduce unwanted systematic effects that may influence the outcome of the experiment. For example, consider what happens when animals arrive from the supplier and are assigned to cages. If the inquisitive animals are picked out first,

and these animals are assigned to the control group cages, then you may end up with all the active animals as controls. If one of the responses measured is locomotor activity, then you may already have a group effect present at baseline (caused by the non-random allocation), which will bias any treatment comparisons.

Incorrect randomisation

As we shall see in Chapter 4, the choice of randomisation has implications on the type of analysis that can be performed. For example, the analysis of a full-factorial design (Sections 6.3.3.1 and 6.3.3.2) is different from that of a complete block design (Section 6.3.3.3) even though structurally they may be the same. The analysis of factorial designs includes additional factor interactions in the analysis whereas the analysis of block designs should not include treatment by block interactions. The difference in these two analysis approaches, as we shall see in Section 4.2.2, is based on the different randomisations applied. Failure to employ the correct randomisation may lead to an unreliable statistical analysis.

Blinding studies

One should be careful when randomising studies to ensure that all the scientists involved in the experiment are blinded to the treatment allocation (see Section 4.1.2). If the assessments are qualitative in nature, and the treatment the animal receives is known, then it is difficult for a scientist to remain impartial. There is now evidence that a failure to blind an experiment properly may induce an increased observed treatment effect (Macleod *et al.*, 2009). Observers assessing the treatment effect should be blinded to the treatment allocation, as should those administering the treatments to the animals and anyone performing routine husbandry duties.

1.3.3 Pitfalls with statistical analysis

The *t*-test

The *t*-test is a simple and popular statistical test. This test involves comparing the difference between two treatment means with the variability of the responses from these two groups only. In the authors' experience, animal experiments are usually complicated affairs and hence the *t*-test is rarely the most appropriate test to use; see also Nieuwenhuis *et al.* (2011). The statistical analysis should reflect the experimental design employed and make full use of its properties. This is not to say the conclusions drawn from the results of *t*-tests are incorrect, just that more powerful tests could perhaps have been used, thus allowing sample sizes to be reduced. We contend that journal referees should always question the use of *t*-tests in submitted articles.

Using all information collected

The principal purpose of the statistical analysis of many animal experiments is to test the hypothesis that one group is in some sense different from another. However, there may be more information that can be recovered from the data collected. For example, use of graphical tools to investigate interrelationships between the responses in a study can help the researcher understand more about the underlying processes in the animal model. These insights, gained by an appropriate statistical analysis, may enable the scientist to reduce animal usage in future studies.

Data trawling

Sometimes it is tempting for a researcher to conduct a data-trawling exercise to try to find a statistical result that agrees with a preconceived idea of what the result should be. This strategy can lead to erroneous false positive conclusions. Such approaches are perhaps more likely to occur when the researcher has freedom to choose (and change) the analysis methods, for example in academic research or early drug discovery studies, as opposed to regulatory testing such as safety assessment and toxicology studies, where analysis strategies are predefined in advance in the protocol.

A commonly encountered example of this pitfall occurs when performing multiple comparison procedures. The researcher is confronted (usually by the computer package) with a long list of available tests and little