Design and Statistical Methods in Studies Using Animal Models of Development

Michael F. W. Festing

Abstract

Experiments involving neonates should follow the same basic principles as most other experiments. They should be unbiased, be powerful, have a good range of applicability, not be excessively complex, and be statistically analyzable to show the range of uncertainty in the conclusions. However, investigation of growth and development in neonatal multiparous animals poses special problems associated with the choice of "experimental unit" and differences between litters: the "litter effect." Two main types of experiments are described, with recommendations regarding their design and statistical analysis: First, the "between litter design" is used when females or whole litters are assigned to a treatment group. In this case the litter, rather than the individuals within a litter, is the experimental unit and should be the unit for the statistical analysis. Measurements made on individual neonatal animals need to be combined within each litter. Counting each neonate as a separate observation may lead to incorrect conclusions. The number of observations for each outcome ("n") is based on the number of treated females or whole litters. Where litter sizes vary, it may be necessary to use a weighted statistical analysis because means based on more observations are more reliable than those based on a few observations. Second, the more powerful "within-litter design" is used when neonates can be individually assigned to treatment groups so that individuals within a litter can have different treatments. In this case, the individual neonate is the experimental unit, and "n" is based on the number of individual pups, not on the number of whole litters. However, variation in litter size means that it may be difficult to perform balanced experiments with equal numbers of animals in each treatment group within each litter. This increases the complexity of the statistical analysis. A numerical example using a general linear model analysis of variance is provided in the Appendix. The use of isogenic strains should be considered in neonatal research. These strains are like immortal clones of genetically identical individuals (i.e., they are uniform, stable, and repeatable), and their use should result in more powerful experiments. Inbred females mated to males of a different inbred strain will produce F1 hybrid offspring that will be uniform, vigorous, and genetically identical. Different strains may develop at different rates and respond differently to experimental treatments.

Key Words: experimental design; experimental unit; litter effect; neonates; statistical analysis

Introduction

The principles of experimental design are universal. They apply equally, for example, to experiments in the life sciences involving humans, animals, plants, and cell cultures. However, in some areas of research, the experimental subjects may have characteristics that necessitate special attention if the experiments are to be designed well and analyzed correctly. Experiments involving neonates of multiparous species are just such a special case. Investigators must identify the correct "experimental unit" (EU[1]) and take "litter effect" into account for the experiments to afford correct results. These critical aspects of experimental design are discussed below.

Determining a Suitable Research Strategy

There are several different types of investigation, which include but are not limited to the following: observational studies, pilot studies, exploratory experiments, confirmatory studies, and experiments that seek parameter estimates. The first example, **observational studies**, do not involve the imposition of an experimental treatment. The comparison of animals of two different genotypes is an observational study even though it may have the appearance of being an experiment. Because it is not possible to assign a genotype to an individual at random, it is the investigator's responsibility to ensure that the animals are, to the extent possible, identical in all other respects apart from their genotype. However, the statistical methods used for observational and experimental studies are essentially the same.

Pilot studies are usually small investigations, sometimes involving only a single animal, with the aim of testing the logistics of a proposed study, and sometimes of gaining

Michael F.W. Festing, Ph.D., has retired from the MRC Toxicology Unit, University of Leicester, UK. Dr. Festing continues to lecture, publish, and consult on statistics and genetics.

[1] Abbreviations used in this article: ANOVA, analysis of variance; 95% CI, 95% confidence interval; EU, experimental unit.

preliminary data to be used in the design of a more definitive experiment. For example, a pilot study could be used to assess whether dose levels are appropriate, and to gain information on likely responses and variability.

Exploratory experiments look at the pattern of response to some treatment but are not based on a formal, testable hypothesis. Often many outcomes (characters) are measured, requiring multiple statistical tests. Even though one may use a correction of the *p* values (e.g., Bonferroni's method of dividing the chosen critical value [usually 0.05] by the number of statistical tests) (Roberts and Russo 1999), exploratory experiments tend to generate more questions than they provide answers. They are usually used to generate hypotheses to be tested in a **confirmatory study,** where the aim is to test some formal, prestated, preferably quite simple hypothesis. Experiments may also be done to estimate parameters such as dose-response curves, means, and proportions.

Choosing a Model

There is surprisingly little discussion of the concept of "models" in biomedical research despite their extensive use (Festing 2004). According to the American philosopher Marx Wartofsky, "Theories, hypotheses, models and analogies I take all to be species of a genus, and my thesis is best stated directly by characterizing this genus, as representation (although "imaging" or "mirroring" will do quite as well)" (Wartofsky 1979). He goes on to say, "There is an additional trivial truth, which may strike some people as shocking: anything can be a model of anything else! This is to say no more than that between any two things in the universe there is some property they both share…."

Although the preceding statements are of little help in deciding whether or not a particular animal or in vitro system is a good model of humans, it does at least clarify the fact that models do not have to resemble the thing being modeled in every respect. Indeed, in some cases it is essential for the model to be *different* from the thing being modeled. Rodents are used widely as models of humans because they are small and economical. The availability of isogenic strains is also an advantage because they make it possible to do efficient experiments using fewer animals and scientific resources. The critical factor is whether the model is like humans for the specific system being modeled, such as the growth and differentiation of some organ or biochemical characteristic, or the response of neonates to xenobiotics.

Principles of Experimental Design

The basic principles of experimental design were formulated many years ago (Fisher 1960), and they remain unchanged. To understand the ensuing brief discussion of these principles, however, it is first necessary to understand the two special characteristics of neonates that strongly influence the design and statistical analysis of experiments involving them.

"Experimental Unit"

Experiments normally involve a number of subjects, or EUs, in each treatment group to afford information about interunit variation and a comparison with the variation between treatment groups. Each EU must be capable of being assigned to a different treatment group, and the data recorded on the individual EUs are subjected to the statistical analysis.

The EU in animal research is commonly the individual animal. However, in research involving neonates, if the pregnant female or the whole litter is subjected to an experimental treatment, the female or the whole litter, not the individual neonate, is the EU, because individual pups within a litter do not receive different treatments (although see below). It is incorrect to use the data from individual pups because the number of independent observations ("n") would be too large and the results would be incorrect, potentially leading to false-positive results (Raubertas et al. 1999; Zorrilla 1997). Values from individual neonates may be taken into account, for example, by averaging them. Such averaging could improve the precision of the litter mean, although they do not contribute as individual EUs (Haseman and Hogan 1975).

Because litters vary in size, if all the neonates are measured in each litter, the averages will vary in precision according to the number of pups per litter. It may be advantageous to use a weighted statistical analysis when evaluating the results. Pups from large litters may also be smaller and less developed than those from smaller litters, so if size (e.g., crown-rump length) is an important outcome, it may be important to correct for this difference in the statistical analysis. Where the outcome is a binary variable such as "normal/abnormal," a full statistical analysis may require advanced statistical methods (Hunt and Bowman 2004; Yamamoto and Yanagimoto 1994).

If individual pups within a litter are subjected to different treatments either postnatally or as a result of surgical or other intervention on the pregnant female, then "n" will be based on the number of individual pups in a treatment group, and the individual pup is the EU. It is possible to have an experiment that is a mixture of a between-litter and a within-litter design. For example, if pregnant females receive one of two or more treatments (e.g., a drug treatment or a vehicle control), and then after birth the neonates within each litter receive additional individual treatments (e.g., some but not all receive a vitamin supplement), then for the drug treatment the pregnant female is the EU, while for the vitamin supplement the neonate is the EU. This design, known as a "split-plot" experimental design (Cox 1958), is often useful although the statistical analysis probably requires professional advice.

"Litter Effect"

In most cases, individual neonates within a litter are more similar than individuals from different litters; in other words, litters differ in a wide range of characteristics. If genetically heterogeneous animals are being used, then individuals within a litter will be full sibs and genetically more similar than unrelated animals. Both pre- and postnatally, animals also tend to have a similar environment. For example, animals from a large litter may be relatively small and immature. There may even be inaccuracies in recording time of birth so that some litters appear to be older than they really are.

It is important to consider "litter effect" when designing an experiment that involves neonates as the EUs. Suppose, for example, that the experiment involves treating some of the neonates with a hormone, while others receive a placebo. Operationally it would be most convenient to treat whole litters, because then pups would not need to be individually identified before weaning. However, in such a case, the litter rather than the individual pups will be the EU. Each litter will be an "n" of one rather than the number of pups in the litter. In contrast, if pups within a litter can be individually identified and assigned to the treatments, then the pups will be the EUs, and each pup will be an "n" of one. However, in this case although the pups within a litter will tend to be quite similar (e.g., in a character such as weight), there may be large differences between pups having the same treatment but in different litters. It will be necessary to remove these differences between litters in the analysis because otherwise, the power of the experiment to detect treatment effects will be severely reduced.

An additional complication is that litters vary in size, so it may be difficult to obtain a balanced design with equal numbers of animals on each treatment within every litter. As a result, it may even be difficult to calculate treatment means. A numerical example of the analysis of a within-litter experiment illustrating some of these problems is given in the Appendix.

Some litter effects due to the common environment of litter mates may gradually disappear once the animals are weaned and are no longer dependent on milk supply. However, litter effects due to the genetic similarity of full sibs will remain for the life of the animals, assuming studies are performed using genetically heterogeneous animals such as Sprague-Dawley rats or any breed of rabbits.

Cross-fostering soon after birth may reduce but will not entirely eliminate litter effects. For example, cross-fostering did not eliminate a litter effect associated with susceptibility to dental caries (Peeling and Looker 1987), a highly inherited character, in outbred Sprague-Dawley rats, or an effect on growth rate (Raubertas et al. 1999). Standardization of postnatal litter size is a common practice and is likely to reduce, but not eliminate, between-litter variability associated with maternal effects such as limitations in milk yield. One commercial company pooled all 2-day-old Sprague-Dawley pups and made up single sex litters of 12 young. Most female pups were discarded at this age because demand was almost entirely for males. Females left without a litter were returned to the breeding colony where they soon became pregnant again without any apparent problems (Lane-Peter et al. 1968). Such a procedure will reduce but not eliminate litter effects because females will still differ in milk yield. It may increase the variability within a litter because individuals will no longer be full siblings, and the procedure is likely to be practical only in breeding colonies where large numbers of females litter at the same time. Nevertheless, it may be worth investigating for neonatal research because it would be very convenient for all litters to have the same number of pups.

Requirements for a Well-designed Experiment

The principles of good experimental design have been known for many years (Cox 1958). These principles are described very briefly as follows.

Absence of bias must be ensured through the use of the use of randomization and blinding. Animals must be selected and assigned to the treatment groups in such a way that there is no systematic difference among groups before starting or during the conduct of the experiment. These factors may be mistaken for the effects of the treatment. This goal is usually achieved by assigning animals (or other experimental subjects) to the treatment groups using a formal randomization system. Subsequent housing and necessary measurements should be in random order. Randomization distributes uncontrolled variation among the groups with equal probability.

The exact method of randomization depends on the design of the experiment. In the most simple "completely randomized" design (i.e., in a between-litter experiment), subjects (e.g., pregnant females) are simply assigned to treatments regardless of their characteristics. Thus, if a teratology experiment involves 20 treated and 20 control pregnant rats, 20 bits of paper could have the letter "C" and 20 the letter "T" written on them. These would be placed in a receptacle and thoroughly shaken. A piece of paper would then be withdrawn, and the first rat would be assigned to the indicated treatment. This process would be repeated with all of the remaining rats. When the neonate is to be the EU in a within-litter experiment, randomization must be done separately within each litter. Again, it is possible to use physical randomization, tables of random numbers, or random numbers generated by a computer.

Ideally, subjects should be identified by codes so that the investigator and other staff members are blind with respect to the treatment groups to the extent possible. Blinding is likely to be particularly important when there is a subjective element to recording observations (e.g., when reading and scoring histological preparations). It would be very unacceptable, for example, to score, measure, or record data from all of the controls first, and subsequently from each treatment group, because standards may change as the scorer becomes more expert. Thus, all manipulations and recording of information should be done either in random order or in such a way as to take account of any time trends with treatment groups equally represented at each time point.

Designing Powerful Experiments: Controlling Variation and Choosing an Appropriate Sample Size

A powerful experiment is one that has a high probability of detecting a difference between treatment groups, assuming that a difference exists. Power depends on the relationship between the variability of the experimental subjects, the size of the treatment effect, and the sample size (discussed in more detail below). Large experiments are likely to be expensive and may exceed the available resources of a facility, so it is worth spending some time and effort to choose uniform experimental material that is sensitive to the effects of the treatment. Thus, if the experimental subjects are adult animals (as in a teratogenesis experiment), they should be closely matched for age, weight, genotype (e.g., by using an isogenic strain where practical), and previous history.

Choosing the Strain or Breed

There are many different strains of mice (www.informatics.jax.org) and rats (www.rgdb.mcw.org) as well as several breeds of rabbits, dogs, and other species. It may be possible to choose one or more strains that are sensitive to the proposed treatments, although for the larger species it is usually necessary to use whatever is available.

Isogenic strains (inbred strains and F1 hybrids between two such strains) of mice and rats are widely available and have many useful properties (Beck et al. 2000; Festing 1999a,b; Festing and Fisher 2000). They resemble immortal clones of genetically identical individuals in some respects. Tissue and organ grafts between individuals of the same isogenic strain are not immunologically rejected and therefore such strains could be of particular value for studies involving such procedures.

Isogenic strains remain genetically constant for many generations and have an international distribution, so that work involving the same strains can be replicated throughout the world. A single individual can be genotyped at loci of interest, which will serve to genotype all animals of that strain. Thus, a genetic profile of the genes present in each strain can be built up by all investigators working on that strain. The genetic authenticity of the animals can be tested using a small sample of DNA. Each strain has a unique set of characteristics, which may make a particular strain valuable for a particular type of study. Some care must be taken in interpreting results if a single inbred strain is used because it represents only a single genotype. However, the interpretation of results is also not easy when using an outbred stock because generally little is known about its genotype.

One disadvantage of inbred strains for neonatal research is that they often have a poor breeding performance, which may limit their use. When the individual neonate is the EU (in a within-litter experiment), it may be worth using inbred mothers mated to a male of a different inbred strain. The pups will then be F1 hybrids, which are vigorous and uniform. Litter size is about 30% larger than when pure isogenic strains are used. When the mother is the EU, it may be worthwhile to use F1 hybrids, which breed exceptionally well as a result of hybrid vigor (Festing 1976). The sire could be either another F1 hybrid of the same strain, in which case the pups will be genetically heterogeneous F2 hybrids, or the females could be backcrossed to one of her parental strains so that the pups would be backcross individuals that, although genetically heterogeneous, are less variable than F2 hybrids.

Outbred stocks such as Sprague-Dawley or Wistar rats and Swiss mice are used widely, but the scientific case for doing to is questionable (Festing 1999b). Animals from different breeders will be genetically different even though they may have the same name. The genotype of any individual will be unknown, the stock is subject to genetic drift over a period of time, the actual degree of genetic heterogeneity is usually unknown, and few methods of genetic quality control are available. It is not even possible to distinguish genetically between Wistar and Sprague-Dawley rats (Festing 1999b). Thus, it is necessary to balance the advantage of better breeding performance against these disadvantages.

Designing the Experiment

After choosing the EU (the female, and/or litter, or individual neonate), it is necessary to determine the number and types of treatment. It may be useful to perform a small pilot study to define dose levels and clarify logistics. It may be necessary to study male and female neonates separately, in which case a factorial design including both sexes in the one experiment may be appropriate (see below, Increasing the Range of Applicability). Outcomes (characters) to be measured or counted must be decided. Where measurements are possible, they are frequently more precise than a “count” (number of positive/negative), and greater precision requires fewer EUs. Each neonate may provide several numerical observations. For example, one should give thought to methods of analyzing individual growth curves within an overall analysis. A microarray experiment may result in thousands of observations from each individual, so the method of statistical analysis of the resulting data should always be considered at this design stage.

Determining Sample Size

The usual way of estimating sample size is to use a power analysis. The success of using this tool depends on a mathematical relationship between several variables, as shown in Figure 1. However, a serious limitation of this method is that it depends critically on the estimate of the standard deviation. This value is not available because the experiment has not yet been done, so it must be estimated from a previous experiment or from the literature. Unfortunately, because standard deviations can vary substantially between different experiments, the power calculations can provide only an indication of the appropriate size of an experiment. This should be interpreted with common sense and in relation to available facilities.

It is easiest to describe the method for a character where there is a treated and control group with a measurement outcome that can be analyzed using an unpaired t-test, such as a teratology experiment with two treatment groups, treated and control. Six variables are involved. Usually the significance level and sidedness of the test are specified, (often the significance level "α" is set at 0.05 with a two-sided test) and the variability of the material (i.e., standard deviation) is taken from a previous study or the literature. When the neonate is the EU, it is necessary to estimate the standard deviation from the pooled standard deviations within litters and treatment groups. The effect size is the minimum difference in means between the two groups the investigator considers to be of biological or clinical importance. Somewhat arbitrarily, the power (i.e., chance that the study will find a statistically significant effect of the specified size) is usually set somewhere between 80 and 95%. It is then possible to estimate the required sample size.

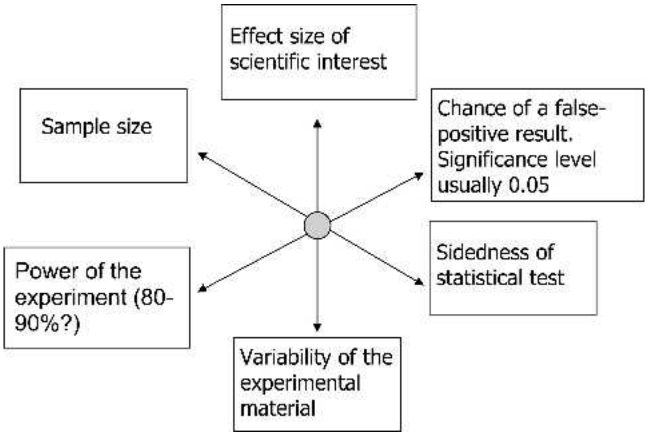
Effect size of scientific interest
Sample size
Chance of a false-positive result. Significance level usually 0.05
Sidedness of statistical test
Power of the experiment (80-90%?)
Variability of the experimental material

Figure 1 The variables involved in a power analysis for a two-sample t-test. Usually the effect size of interest, the significance level, sidedness of the test, variablilty of the material and power are specified, which determines the required sample size. Alternatively, if the sample size is fixed due to resource limitations, the method can be used to assess power or effect size.

For the calculations, a number of dedicated computer programs such as nQuery Advisor (Elashoff 2000) are available. In addition, many statistical packages such as MINITAB have routines for power analysis, and there are a number of free sites on the web (e.g., http://www.biomath.info), where one can enter data to obtain estimates of required sample sizes. In some circumstances, such as when resources are limited, the sample size may be fixed and the power analysis can then be used to estimate the power of the proposed experiment (i.e., the chance that the specified effect is likely to be detected). The calculations are similar for a binary variable (normal/abnormal) with two groups, but the specification becomes more difficult when there are several treatment groups, or when the data are not appropriate for a parametric analysis (Dell et al. 2002).

An alternative method of sample size determination is the so-called "resource equation method," which depends on the law of diminishing returns. This method is useful for small and complex biological experiments that involve several treatment groups for which the results are to be analyzed using the analysis of variance. In such a situation, it is difficult to use a power analysis. The experiment should be of an appropriate size if the error degrees of freedom in an analysis of variance are somewhere between 10 and 20 (Festing et al. 2002; Mead 1988). This case reduces to the very simple equation:

$$X = N - T - B + 1,$$

where N is the total number of observations, T is the number of treatments, B is the number of blocks (litters for a within-litter experiment), and X should be between approximately 10 and 20.

For a within-litter experiment with three treatments, an average litter size of six, and a proposal to use five litters,

$$X = (6 \times 5) - 3 - 5 + 1 = 23.$$

The limits of X being between 10 and 20 can be liberally interpreted, so this proposed experiment would be of an appropriate size, although just beyond the suggested upper limit.

The experiment described in the Appendix has $X = 50$, which is more than twice as large as suggested by this method. A repeated analysis of the data in the Appendix using only the first three litters gives $X = 23$ and a p value for treatments of 0.007 compared with $p = 0.001$ using six litters. Thus, if the experiment had been performed with approximately half the number of animals, the conclusions would have been about the same. Compared with the power analysis, the resource equation method is somewhat crude. Nevertheless, it often seems to work in practice, particularly when relatively large treatment effects are expected.

Increasing the Range of Applicability: Factorial Designs

It is often important to know the extent to which a response to a treatment can be generalized. Is the same response found in males and females, or in different strains of animals, or with different diets? Does the presence of some drug or chemical alter response? Factorial experimental designs allow such questions to be examined without requiring any substantial increase in resources. A typical example might be to learn whether alcohol potentiates the effect of a teratogen in rats. If, for example, the basic plan was to have 20 pregnant females as controls and 20 treated with the teratogen, then the effect of alcohol might be studied by administering alcohol to half the rats in each group. There would then be four groups of 10 pregnant females with or without alcohol and with or without the teratogen. It first seems as though the group size has been reduced from two to 10 rats, but in fact the effect of the teratogen is still determined by comparing those receiving the teratogen (20 rats) and those that do not receive it (20 rats). Similarly, the effect of the alcohol is determined by comparing the 20 rats that receive it with the 20 rats that do not receive it. Finally, any potentiating effect of alcohol is determined by seeing whether the difference in fetal weight, number of abnormalities, and other factors between the teratogen-treated and -untreated rats is greater in the group receiving alcohol than in those that do not receive it.

Factorial designs can also be used for within-litter experiments. Pups could be sexed and assigned separately at random to either a control or a treated group. There would then be four groups within each litter: male and female controls and male and female treated. The experiment could then be analyzed (probably using an analysis of variance) to determine whether the pups responded to the treatment, averaging across sexes; whether the measured outcome (e.g., weaning weight) differed between males and females, averaging across treatments; and whether the response to the treatment differed between the two sexes.

Factorial designs provide a way of obtaining more information from the same scientific resources at relatively little extra cost. Any number of factors (e.g., treatments, strain, sex, diet) can be involved, and each can have any number of levels (i.e., there can be any number of dose levels within a factor). The main extra cost is the increase in the complexity of the experiment, which could lead to mistakes, and the increased complexity of the statistical analysis. Splitting groups into a number of subgroups does not lead to any substantial loss of power, provided the experiment is not too small.

Avoiding Excessive Complexity

Complex experiments may lead to mistakes and invalid conclusions. All experiments should be planned ahead, with written protocols and standard operating procedures. It is appropriate to alter experiments while they are in progress only in exceptional cases (e.g., for ethical reasons). Animal care staff should be regarded as integral and valued members of the research team. If mistakes occur, it is vital to acknowledge them, rather than covered them up, so that staff members are not made to fear that they will be in serious trouble if they make a mistake.

Statistical Analysis

No experiment should be started without the investigator having a clear idea of how the results will be analyzed statistically, although it may be necessary to modify the analysis later in the light of actual results. For example, it may be necessary to transform scales and to account for missing observations. However, the statistical analysis is a basic and integral part of the experimental design. Moreover, time (i.e., avoiding delay) is important. Normally, it is important to analyze experiments as soon as they have been completed so that the results can be used in formulating future experiments (e.g., adjusting dose levels or altering the timing of observations in subsequent experiments).

The aim of the statistical analysis is to obtain summarized results that may be easily understood and that clarify the range of uncertainty in the conclusions. Access to a good statistical textbook is highly recommended. A basic assumption is that the EUs are a random sample from a population of such units (real or hypothetical), and the aim is to make inferences about the population from the sample. The accuracy of these inferences will depend mainly on the biological variability of the EUs and the sample size, assuming that the experiment has been designed well to avoid bias. Clearly, if the sample size is very small and/or the variation is large, then only rough estimates of the population characteristics will be available.

It is essential to use a good-quality statistical package. Spread sheets such as EXCEL are adequate for storing and manipulating the raw data, but they should not be used for the main statistical analysis. The output is often not standard, and it fails to provide the range of methods available in a dedicated package. For example, the statistical analysis presented in the Appendix could not be done using EXCEL. Packages such as SPSS, MINITAB, SAS, Statistika, Graphpad, GLIM, Genstat, and BMDP are readily available and have been tested thoroughly for errors. One or more are usually available on most institutional networks.

The first step in the analysis should be to screen the data for errors. Histograms and dotplots showing individual observations (e.g., as in Figure 2 in the Appendix), possibly plotted against dose levels, or plots of two outcomes likely to be correlated will often show whether there are any serious outliers. Any outliers should be individually checked against notebooks or original printouts to ensure that they are not transcription errors, and should be corrected if necessary. Outliers that appear to be valid should not be discarded at this stage. Many outcomes of measurement data, particularly concentrations of a substance, have a log-normal distribution, with most numbers being relatively low but with a few very high. If this is the case, the data can be transformed by taking logarithms or square roots of the raw observations. This step frequently removes outliers and allows parametric statistical methods—usually a t test or an analysis of variance (ANOVA[1])—to be used in the analysis. These parametric methods depend on the assumption that the residuals (deviations of each observation from group means) have a normal distribution and the variation is approximately the same in each group.

One way to deal with one or two persistent outliers is to perform the statistical analysis with and without them. If it makes no difference to the conclusions, then they can be retained. However, if the conclusions depend entirely on one or a few outliers, and these appear to be perfectly valid data points, the results should be treated with caution. Outliers that are more than 3 standard deviations from the mean (assuming an approximately normal distribution) are automatically rejected by some authors; but again, it may be worth seeing what effect the outliers have on the overall conclusions.

When it is not possible to normalize badly skewed data using a scale transformation, and when the aim is to compare groups, it may be necessary to analyze the data using nonparametric methods such as the Mann-Whitney or Wilcoxon test. Dose response curves are normally estimated using some form of regression analysis. A numerical example illustrating the statistical analysis of a within-litter experiment using the analysis of variance is shown in the Appendix.

Presentation of Results

Scientific papers are often written in such a way as almost to observe exactly what the investigators did. In theory, sufficient information should be given so that others can repeat the studies. Unfortunately, in a surprisingly large proportion of papers, it is difficult or impossible to determine exactly how many animals were used, or how many separate experiments were involved.

Guidelines are available for the design and statistical analysis of experiments using animals (e.g., Festing and Altman 2002), and they include a number of suggestions for presenting results.

- Label and number each experiment;
- State the number of animals used in each experiment, along with the purpose of each experiment;
- Identify the species, breed, and/or strain of animals complying with agreed international nomenclature rules where these are available (e.g., for rats and mice, WWW.informatics.jax.org);
- Provide details of husbandry (e.g., diet and housing) to the extent allowed by the journal editor;
- Describe efforts where possible to minimize pain, distress, or lasting harm to the animals;
- Describe methods of statistical analysis, with references in the case of any unusual methods used;
- Identify the statistical software used;
- Avoid excess decimal places where means, proportions, or differences are presented;
- Include measures of variation (e.g., standard deviations, standard errors, or, preferably, confidence intervals [Altman 1991; Altman et al. 2000]);
- Identify the number of observations for every mean, including those shown graphically. It is not adequate to make statements such as “the number in each group ranged from four to 10.” Where possible, tabulate means in columns for ease of comparison.
- Use graphs to illustrate points that are difficult to show in tables or in the text. Where possible, show individual observations rather than means with error bars because this presentation more clearly indicates the distribution of the observations. If error bars are used, explain clearly whether they are standard deviations, standard error, or confidence intervals.

Again, the main aim in presenting the results should be to state as clearly and succinctly as possible exactly what was done and what results were obtained.

APPENDIX: A NUMERICAL EXAMPLE

Consider the weaning weight of 59 unsexed Sprague-Dawley rats (real data), including one that died as a missing observation (Table 1). When pups were 2 days old, each litter was split, and the pups were assigned at random to a control group, a “low-dose” group, or a “high-dose” group (simulated by subtracting 0.5 g from the low-dose group and 1.0 g from the high-dose group). Within each litter, to the extent possible, the same number of pups were assigned to each treatment, and pups were individually marked for subsequent identification. The sex of the pups was not recorded. The aim of the statistical analysis is to determine whether the treatments altered weaning weight, and if so to what extent. (Note: It should be a reduction of approximately 0.5 g and 1.0 g in the low and high groups, respectively.)

The first step in analyzing such data is to examine it graphically to learn whether there are any obvious outliers and to obtain a visual impression of the situation (see plot in Figure 2). In this case, there are no obvious outliers. However, the litter effect is very obvious and clearly there is considerable variation within each litter. Although there is a tendency for the controls to weigh more than the treated groups (e.g., in litter 6), in litter 2 the lightest pup is a control.

Anyone planning to make a career in animal research is strongly advised to familiarize him- or herself with the analysis of variance as it is the most appropriate statistical

Table 1 Data for the numerical example. The table shows weaning weight (g) of six litters of Sprague-Dawley rats assigned to three treatments: control, low, and high doses. Weights are real data, but treatments are simulated (see text).

	Litter number								
Treatment	1	2	3	4	5	6	Simple mean (1)	Weighted mean (2)	Least squares mean (3)
Control	49.5	45.2	56.8	43.0	42.2	53.8			
	51.5	37.2	59.7	47.7	38.4	47.0			
	48.6	42.9		45.4	43.9	51.6			
		X				50.8			
		45.6							
Mean	49.9	42.7	58.3	45.4	41.5	50.8	48.1	47.4	48.1
Low dose	45.6	40.1	54.1	45.1	40.9	50.0			
	48.6	40.8	57.4	48.7	36.3	47.8			
	47.6	41.7		44.7	40.4	46.9			
		40.5							
		40.7							
Mean	47.3	40.8	55.8	46.2	39.2	48.2	46.2	45.2	46.2
High dose	40.6	43.2	53.8	40.0	42.2	46.3			
	44.1	42.5	55.5	45.6	40.6	44.3			
	45.1	41.4		46.8	39.8	46.8			
		39.3			40.2				
		40.0							
Mean	43.3	41.3	54.7	44.1	40.7	45.8	45.0	43.9	45.2

X, missing observation due to death of animal.

(1) Mean of litter by treatment means. These means are biased (see 4, below).

(2) Mean of all animals in a treatment group, ignoring litter. These means are biased (see 4, below).

(3) Differences between least squares means give the best unbiased estimate of the treatment differences.

(4) Numbers in parenthesis show the size of the treatment effect (control mean-dose mean) estimated from these means. The least squares means give the best unbiased estimate of the size of the treatment effect.

method for dealing with most data arising from formal experiments like this one. A good introduction to the methods is given by Roberts and Russo (1999), and it is also described in detail in most statistical textbooks.

The data in Table 1 can be analyzed using a two-way (treatment and litter) analysis of variance "without interaction." A t test would be entirely inappropriate because there are more than two groups, and it is necessary to account for

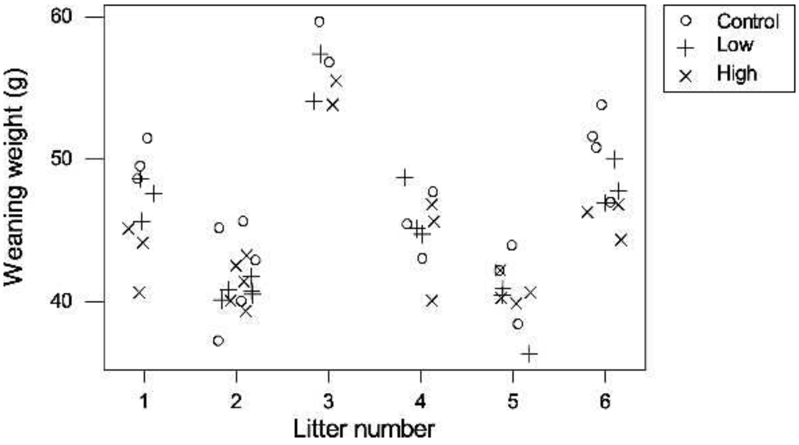

Figure 2 Weaning weight by litter number and treatment for the numerical example. Note that some random variation or "jitter" has been applied on the X-axis to avoid too much overlap between points (see text for details).

the litter effect. The ANOVA quantifies the variation associated with treatments, litters, and the remaining "residual" or "error" variation. It is assumed that the response is the same in each litter apart from sampling variation (hence "without interaction"). However, there is a problem with these data as they stand. The usual two-way ANOVA assumes that there are equal numbers in each treatment group within each litter. In this case, there is one missing observation in litter 2 and two extra animals in litters 5 (high-dose) and 6 (control). The data could be adjusted by discarding at random two animals from groups where there are the extras, and replacing the value for the animal that died by an appropriate value. Missing values can be worked out using formulae available in most of the older textbooks (e.g., Cochran and Cox 1957). In situations where there is more than one animal in a litter by treatment subgroup, as in this case, it would probably be sufficiently accurate (although not strictly correct) to replace the missing value with the mean of the rest of the animals in the group. Having a balanced design used to be almost essential because otherwise the calculations were extremely tedious. However, modern statistical packages now make it possible to do a "general linear model" ANOVA, which is capable of accommodating unequal numbers in each group, so a balanced design is no longer so essential.

A general linear model ANOVA of the data in Table 1 is shown in Table 2. Note that whereas in the normal ANOVA there is a heading labeled "Sums of Squares" (or simply SS), in this case there are two headings "Seq SS" and "Adj SS," with the two being slightly different for the litter effect. The ANOVA shows an F value of 7.99 and a *p* value of 0.001 for the treatment effect (abbreviated Trt). The "least squares means" presented in Table 2 are marginally different from the simple means and weighted means presented in Table 1 (all three types of means are shown in Table 1) inasmuch as they take account of the unequal group sizes.

It is often necessary to use a post hoc comparison to determine which means differ from which. When the aim is to compare the means of the treatment groups with the control, Dunnett's test is appropriate (shown in Table 2). If the aim is to compare each mean with every other mean, it is appropriate to use other available post hoc comparisons (e.g., Tukey's test [Roberts and Russo 1999]). Dunnett's test subtracts the mean of the control group from each of the other groups and then either gives a 95% confidence interval (CI[1]) for the difference, or involves a t test to resolve whether it is different from zero. Both approaches are shown in the case. Note that the differences between the three groups are larger than the simulated treatment effect of

Table 2 General linear model analysis of variance of the data in Table 1

```
Analysis of Variance for Weaning Wt, using Sequential SS[a] for Tests
Source        DF        Seq[a] SS       Adj[a] SS       Seq MS[a]        F[a]         p

Litter         5        1270.1          1232.49         254.18           48.37        0.000
Trt[a]         2          84.02           84.02          42.01            7.99        0.001
Error         50         262.76          262.76           5.26
Total         57        1617.70
Least Squares Means for Weaning Wt
Trt           Mean
1             48.08
2             46.18
3             45.17
Dunnett 95.0% Simultaneous Confidence Intervals
Response Variable Weaning Wt Comparisons with Control Level
Trt = 1 subtracted from:
Trt          Lower           Center          Upper
2            –3.599          –1.903          –0.207
3            –4.585          –2.909          –1.232
Dunnett Simultaneous Tests
Response Variable Weaning Wt. Comparisons with Control
LevelTrt = 1 subtracted from:
Level
Trt           Difference of Means     SE[a] of Difference     T-Value      Adjusted p Value
2             –1.903                  0.7455                  –2.552       0.0258
3             –2.909                  0.7368                  –3.948       0.0005
```

[a] Wt, weight; SS, sums of squares; DF, degrees of freedom; Seq, sequential; Adj, adjusted; F, variance ratio (a test statistic like Student's t); Trt, treatment; SE, standard error; p, probability that a difference as large as or larger than the one observed could have arisen by chance; T-value, Student's t.

-0.5 g and -1.0 g in the low- and high-dose groups, respectively, because the groups already differed by chance.

In this case, 95% CIs for the means should be calculated by hand. The error mean square of 5.26 is the pooled within-group variance, so the standard deviation is the square root of this value, or 2.29. Standard errors are calculated by dividing 2.29 by the square root of the number in each mean (19 in the control and low-dose group, 20 in the high-dose group). The 95% CI is estimated from the formulae given below (also shown in most statistical text books):

$$M- SE^*t_{0.05,d.f.} < M < M + SE^*t_{0.05,d.f.,}$$

where *M* is the observed mean, the SE is the standard error of the mean, and $t_{0.05,d.f.}$ is the value of the Student's t for the 0.05 level of significance for the degrees of used in estimating the variance, which is 50 (Table 2). The means can now be presented as follows:

Control mean = 48.1 (95% CI 47.0, 49.1);
Low-dose mean = 46.2 (95% CI 45.1, 47.2);
High-dose mean = 45.2 (95% CI 44.1, 46.2).
These confidence intervals could be used as error bars in a bar diagram.

Finally, if one performed a similar experiment, but treated whole litters rather than doing a within-litter experiment, the EU would be the litter, rather than the individual pup within the litter. To determine how many litters would be needed, assume for simplicity that there would be only a control and a high-dose group. The question can be addressed using a power analysis as described above. The standard deviation of litter means in Table 1 is 5.68 g.

If one decided that a treatment effect (difference between treated and control groups) of 4 g in mean pup weight would be of scientific interest, and the experiment should have a 90% power and a significance level of 0.05, with a two-sided t-test, then using the power calculator in MINITAB, 44 litters in each group would be required to perform this experiment. Thus, the between-litter experiment would involve a total of 88 litters and at an average of 9.7 pups per litter over 850 pups, yet would only be capable of distinguishing an effect of 4.0 g compared with a resolution of 2.9 g in the within-litter experiment involving six litters and only 59 pups. Clearly, between-litter designs should only be used in situations where there is no alternative, such as in teratology experiments.

References

Altman DG. 1991. Practical Statistics for Medical Research. London: Chapman and Hall.

Altman DG, Machin D, Bryant TN, Gardiner MJ. 2000. Statistics with Confidence. London: BMJ Press.

Beck JA, Lloyd S, Hafezparast M, Lennon-Pierce M, Eppig JT, Festing MFW, Fisher EMC. 2000. Genealogies of mouse inbred strains. Nat Genet 24:23-25.

Cochran WG, Cox GM. 1957. Experimental Designs. New York: John Wiley & Sons, Inc.

Cox DR. 1958. Planning Experiments. New York: John Wiley & Sons.

Dell R, Holleran S, Ramakrishnan R. 2002. Sample size determination. ILAR J 43:207-213.

Elashoff JD. 2000. nQuery Advisor Version 4.0 User's Guide. Cork: Statistical Solutions.

Festing MFW. 1976. Effects of marginal malnutrition on the breeding performance of inbred and F1 hybrid mice-a diallel study. In: Antikatzides T, ed. The Laboratory Animal in the Study of Reproduction. Stuttgart: Gustav Fischer. p 99-114.

Festing MFW. 1999a. Introduction to laboratory animal genetics. In: Poole T, ed. The UFAW Handbook on the Care and Use of Laboratory Animals. Harlow: Longman Scientific and Technical. p 61-94.

Festing MFW. 1999b. Warning: The use of genetically heterogeneous mice may seriously damage your research. Neurobiol Aging 20:237-244.

Festing MFW. 2004. Is the use of animals in biomedical research still necessary in 2002? Unfortunately, "yes." Altern Anim Res 32(S1):733-739.

Festing MFW, Altman DG. 2002. Guidelines for the design and statistical analysis of experiments using laboratory animals. ILAR J 43:233-243.

Festing MFW, Fisher EMC. 2000. Mighty mice. Nature 404:815.

Festing MFW, Overend P, Gaines Das R, Cortina Borja M, Berdoy M. 2002. The Design of Animal Experiments. London: Laboratory Animals Ltd.

Fisher RA. 1960. The Design of Experiments. New York: Hafner Publishing Company, Inc.

Haseman JK, Hogan MD. 1975. Selection of the experimental unit in teratology studies. Teratology 12:165-171.

Hunt DL, Bowman D. 2004. A parametric model for detecting hormetic effects in developmental toxicity studies. Risk Anal 24:65-72.

Lane-Peter W, Lane-Petter ME, Boutwell CW. 1968. Intensive breeding of rats. I. Crossfostering. Lab Anim 2:35-39.

Mead R. 1988. The Design of Experiments. Cambridge: Cambridge University Press.

Peeling AN, Looker T. 1987. Problem of standardising growth rates for animals suckled in separate litters. Growth 51:165-169.

Raubertas RF, Davis BA, Bowen WH, Pearson SK, Watson GE. 1999. Litter effects on caries in rats and implications for experimental design. Caries Res 33:164-169.

Roberts MJ, Russo R. 1999. A Student's Guide to the Analysis of Variance. London: Routledge.

Wartofsky MW. 1979. Models: Representation and the Scientific Understanding. Dordrecht: D. Reidel Publishing Company.

Yamamoto E, Yanagimoto T. 1994. Statistical methods for the beta-binomial model in teratology. Environ Health Perspect 102(Suppl 1): 25-31.

Zorrilla EP. 1997. Multiparous species present problems (and possibilities) to developmentalists. Dev Psychobiol 30:141-150.