Guidelines for the Design and Statistical Analysis of Experiments Using Laboratory Animals

Michael F. W. Festing and Douglas G. Altman

Abstract

For ethical and economic reasons, it is important to design animal experiments well, to analyze the data correctly, and to use the minimum number of animals necessary to achieve the scientific objectives—but not so few as to miss biologically important effects or require unnecessary repetition of experiments. Investigators are urged to consult a statistician at the design stage and are reminded that no experiment should ever be started without a clear idea of how the resulting data are to be analyzed. These guidelines are provided to help biomedical research workers perform their experiments efficiently and analyze their results so that they can extract all useful information from the resulting data. Among the topics discussed are the varying purposes of experiments (e.g., exploratory vs. confirmatory); the experimental unit; the necessity of recording full experimental details (e.g., species, sex, age, microbiological status, strain and source of animals, and husbandry conditions); assigning experimental units to treatments using randomization; other aspects of the experiment (e.g., timing of measurements); using formal experimental designs (e.g., completely randomized and randomized block); estimating the size of the experiment using power and sample size calculations; screening raw data for obvious errors; using the *t*-test or analysis of variance for parametric analysis; and effective design of graphical data.

Key Words: animal experiments; experimental design; statistics; variation

Introduction

Experiments using laboratory animals should be well designed, efficiently executed, correctly analyzed, clearly presented, and correctly interpreted if they are to be ethically acceptable. Unfortunately, surveys of published papers reveal that many fall short of this ideal, and in some cases, the conclusions are not even supported by the data (Festing 1994; Festing and Lovell 1995, 1996; McCance 1995). This situation is unethical and results in a waste of scientific resources. In contrast, high-quality methods will help to ensure that the results are scientifically reliable and will not mislead other researchers.

The aim of these guidelines is to help investigators who use animals ensure that their research is performed efficiently and humanely, with the minimum number of animals to achieve the scientific objectives of the study. Some knowledge of statistics is assumed because most scientists will have had some training in this discipline. However, scientists using animals should always have access to a statistician who can help with unfamiliar or advanced methods.

These guidelines and suggestions for further reading are based partly on previously published guidelines for contributors to medical journals (Altman et al. 2000) and for in vitro experiments (Festing 2001). Although a useful set of guidelines for "appropriate statistical practice" in toxicology experiments has previously been published (Muller et al., 1984), with a more extensive set of suggestions for the design and analysis of carcinogenicity studies (Fairweather et al. 1998), general guidelines aimed specifically at experiments using laboratory animals in both academic and applied research do not appear to have been published recently. However, a recent book covers in more detail much of the ground discussed here (Festing et al. 2002).

Although responsibility for the quality of research rests clearly with those who perform it, we believe journal editors should ensure adequate peer review by individuals knowledgeable in experimental design and statistics. They should also ensure that there is a sufficiently full description of animals, experimental designs, and statistical methods used and should encourage the discussion of published papers through letters to the editor and, when possible, by suggesting that authors publish their raw data electronically (Altman 2002).

Ethical Considerations

The use of animals in scientific experiments likely to cause pain, distress, or lasting harm generates important ethical issues. Animals should be used only if the scientific objectives are valid, there is no other alternative, and the cost to the animals is not excessive. "Validity" in this case implies that the experiment has a high probability of meeting the stated objectives, and these objectives have a reasonable

Michael F. W. Festing, M.Sc., Ph.D., D.Sc., CStat., BIBiol., is a Senior Research Scientist at the MRC Toxicology Unit, University of Leicester, UK. Douglas G. Altman, Ph.D., D.Sc., is Director of the Cancer Research UK/NHS Centre for Statistics in Medicine, Institute of Health Sciences, Headington, Oxford, UK.

chance of contributing to human or animal welfare, possibly in the long term.

The following "3Rs" of Russell and Burch (1959) provide a framework for considering the humane use of animals:

- Animals should be *replaced* by less sentient alternatives such as invertebrates or in vitro methods whenever possible.
- Experimental protocols should be *refined* to minimize any adverse effects for each individual animal. For example, appropriate anesthesia and analgesia should be used for any surgical intervention. Death is not an acceptable endpoint if it is preceded by some hours of acute distress, and humane endpoints should be used whenever possible (Stokes 2000). Staff should be well trained, and housing should be of a high standard with appropriate environmental enrichment. Animals should be protected from pathogens.
- The number of animals should be *reduced* to the minimum consistent with achieving the scientific objectives of the study, recognizing that important biological effects may be missed if too few animals are used. Some thought also should be given to the required precision of any outcomes to be measured. For example, chemicals are classified into a number of groups on the basis of their acute toxicity in animals. It may not be necessary to obtain a highly precise estimate of the median lethal dose (LD_{50} value) to classify them. A number of sequential experimental designs that use fewer animals have been developed for this purpose (Lipnick et al. 1995; Rispin et al. 2002; Schlede et al. 1992). Ethical review panels should also insist that any scientist who does not have a good background in experimental design and statistics should consult a statistician.

General Principles

All research should be described in such a way that it could be repeated elsewhere. Authors should clearly state the following:

- The objectives of the research and/or the hypotheses to be tested;
- The reason for choosing their particular animal model;
- The species, strain, source, and type of animal used;
- The details of each separate experiment being reported, including the study design and the number of animals used; and
- The statistical methods used for analysis.

Experiments and Surveys

An **experiment** is a procedure for collecting scientific data on the response to an intervention in a systematic way to maximize the chance of answering a question correctly (confirmatory research) or to provide material for the generation of new hypotheses (exploratory research). It involves some treatment or other manipulation that is under the control of the experimenter, and the aim is to discover whether the treatment is *causing* a response in the experimental subjects and/or to quantify such response. A **survey**, in contrast, is an observational study used to find *associations* between variables that the scientist cannot usually control. Any association may or may not be due to a causal relation. These guidelines are concerned only with experiments.

Experiments should be planned before they are started, and this planning should include the statistical methods used to assess the results. Sometimes a single experiment is replicated in different laboratories or at different times. However, if this replication is planned in advance and the data are analyzed accordingly, it still represents a single experiment.

Confirmatory and Exploratory Experiments

Confirmatory research normally involves formal testing of one or more prespecified hypotheses. By contrast, exploratory research normally involves looking for patterns in the data with less emphasis on formal testing of hypotheses. Commonly, exploratory experiments involve many characters. For example, many microarray experiments in which up or down regulation of many thousands of genes is assayed in each animal could be classified as exploratory experiments because the main purpose is usually to look for patterns of response rather than to test some prespecified hypotheses. There is frequently some overlap between these two types of experiment. For example, an experiment may be set up to test whether a compound produces a specific effect on the body weight of rats—a confirmatory study. However, data may also be collected on hematology and clinical biochemistry, and exploratory investigations using these data may suggest additional hypotheses to be tested in future confirmatory experiments.

Investigations Involving Several Experiments

Scientific articles often report the results of several independent experiments. When two or more experiments are presented, they should be clearly distinguished and each should be described fully. It is helpful to readers to number the experiments.

Animals as Models of Humans or Other Species

Laboratory animals are nearly always used as models or surrogates of humans or other species. A model is a representation of the thing being modeled (the target). It must have certain characteristics that resemble the target, but it can be very different in other ways, some of which are of little importance whereas others may be of great practical importance. For example, the rabbit was used for many years as a model of diabetic humans for assaying the potency of insulin preparations because it was well established that insulin reduces blood glucose levels in rabbits as well as in humans. The fact that rabbits differ from humans in many thousands of ways was irrelevant for this particular application. This was a well-validated model, but it has now been replaced with chemical methods.

Other models may be less well validated; and in some cases it may be difficult, impossible, or impractical to validate a given model. For example, it is widely assumed that many industrial chemicals that are toxic at a given dose in laboratory animals will also be toxic to humans at approximately the same dose after correcting for scale. However, it is usually not possible to test this assumption. Clearly, the validity of an animal model as a predictor of human response depends on how closely the model resembles humans for the specific characters being investigated. Thus, the validity of any model, including mathematical, in vitro, and lower organism models, must be considered on a case-by-case basis.

Need to Control Variation

After choosing a model, the aim of the experiment will be to determine how it responds to the experimental treatment(s). Models should be sensitive to the experimental treatments by responding well, with minimal variation among subjects treated alike. Uncontrolled variation, whether caused by infection, genetics, or environmental or age heterogeneity, reduces the power of an experiment to detect treatment effects.

If mice or rats are being used, the use of isogenic strains should be considered because they are usually more uniform phenotypically than commonly used outbred stocks. Experiments using such animals either should be more powerful and able to detect smaller treatment responses or could use fewer animals. When it is necessary to replicate an experiment across a range of possible susceptibility phenotypes, small numbers of animals of several different inbred strains can be used in a factorial experimental design (see below) without any substantial increase in total numbers (Festing 1995, 1997, 1999). The advantage of this design is that the importance of genetic variation in response can be quantified. Inbred strains have many other useful properties. Because all individuals within a strain are genetically identical (apart possibly from a small number of recent mutations), it is possible to build up a genetic profile of the genes and alleles present in each strain. Such information can be of value in planning and interpreting experiments. Such strains remain genetically constant for many generations, and identification of individual strains is possible using genetic markers. There is a considerable literature on the characteristics of the more common strains, so that strains suitable for each project can be chosen according to their known characteristics (Festing 1997, 1999; <www.informatics.jax.org>).

Animals should be maintained in good environmental conditions because animals under stress are likely to be more variable than those maintained in optimum conditions (Russell and Burch 1959). When a response is found in the animal, its true relevance to humans is still not known. Thus, clinical trials are still needed to discover the effects of any proposed treatment in humans. However, in testing toxic environmental chemicals, it is normally assumed that humans respond in a similar way to animals, although this assumption can rarely be tested. The animals should be adequately described in the materials and methods or other relevant section of the paper or report. The Appendix provides a checklist of the sort of information that might be provided, depending on the individual study.

Experimental Design

The experimental design depends on the objectives of the study. It should be planned in detail, including the development of written protocols and consideration of the statistical methods to be used, before starting work.

In principle, a well-designed experiment avoids bias and is sufficiently powerful to be able to detect effects likely to be of biological importance. It should not be so complicated that mistakes are made in its execution. Virtually all animal experiments should be done using one of the formal designs described briefly below.

Experimental Unit

Each experiment involves a number of experimental units, which can be assigned at random (see below) to a treatment. The experimental unit should also be the unit of statistical analysis. It must be possible, in principle, to assign any two experimental units to different treatments. For this reason, if the treatment is given in the diet and all animals in the same cage therefore have the same diet, the cage of animals (not the individual animals within the cage) is the experimental unit. This situation can cause some problems. In studying the effects of an infection, for example, it may be necessary to house infected animals in one isolator and control animals in another. Strictly, the isolator is then the experimental unit because it was the entity assigned to the treatment and an analysis based on a comparison of individual infected versus noninfected animals would be valid only with the additional assumption (which should be explicitly stated) that animals within a single isolator are no more or no less alike than animals in different isolators. Although individual animals are often the experimental units assigned to the treatments, a crossover experimental design may involve assigning an animal to treatments X, Y, and Z sequentially in random order, in which case the experimental unit is the animal for a period of time. Similarly, if cells from an animal are cultured in a number of dishes that can be assigned to different in vitro treatments, then the dish of cells is the experimental unit.

Split-plot experimental designs have more than one type of experimental unit. For example, cages each containing two mice could be assigned at random to a number of dietary treatments (so the cage is the experimental unit for comparing diets), and the mice within the cage may be given one of two vitamin treatments by injection (so the mice are experimental units for the vitamin effect). In each case, the analysis should reflect the way the randomization was done.

Randomization

Treatments should be assigned so that each experimental unit has a known, often equal, probability of receiving a given treatment. This process, termed randomization, is essential because there are often sources of variation, known or unknown, which could bias the results. Most statistical packages for computers will produce random numbers within a specified range, which can be used in assigning experimental units to treatments. Some textbooks have tables of random numbers designed for this purpose. Alternatively, treatment assignments can be written on pieces of paper and drawn out of a bag or bowl for each experimental unit (e.g., animal or cage). If possible, the randomization method should ensure that there are predefined numbers in each treatment group.

Note that the different treatment groups should be processed identically throughout the whole experiment. For example, measurements should be made at the same times. Furthermore, animals of different treatment groups should not be housed on different shelves or in different rooms because the environments may be different (see Blinding and Block Designs below).

Blinding

To avoid bias, experiments should be performed "blind" with respect to the treatments when possible and particularly when there is any subjective element in assessing the results. After the randomized allocation of animals (or other experimental unit) to the treatments, animals, samples, and treatments should be coded until the data are analyzed. For example, when an ingredient is administered in the diet, the different diets can be coded with numbers and/or colors and the cages can be similarly coded to ensure that the correct diet is given to each cage. Animals can be numbered in random order so that at the postmortem examination there will be no indication of the treatment group. Pathologists who read slides from toxicity experiments are often not blinded with respect to treatment group, which can cause problems in the interpretation of the results (Fairweather et al. 1998).

Pilot Studies

Pilot studies, sometimes involving only a single animal, can be used to test the logistics of a proposed experiment. Slightly larger ones can provide estimates of the means and standard deviations and possibly also some indication of likely response, which can be used in a power analysis to determine sample sizes of future experiments (see below). However, if the pilot experiment is very small, these estimates will be inaccurate.

Formal Experimental Designs

Several formal experimental designs are described in the literature, and most experiments should use one of these designs. The most common are completely randomized, randomized block (see below), and factorial designs; however, Latin square, crossover, repeated measures, split-plot, incomplete block, and sequential designs are also used. These formal designs have been developed to take account of special features and constraints of the experimental material and the nature of the investigation. It is not possible to describe all of the available experimental designs here. They are described in many statistical textbooks.

Investigators are encouraged to name and describe fully the design they used to enable readers to understand exactly what was done. We also recommend including an explanation of a nonstandard design, if used.

Within each type of design there is considerable flexibility in terms of choice of treatments and experimental conditions; however, standardized methods of statistical analysis are usually available. In particular, when experiments produce numerical data, they can often be analyzed using some form of the analysis of variance (ANOVA[1]).

Completely randomized designs, in which animals (or other experimental units) are assigned to treatments at random, are widely used for animal experiments. The main advantages are simplicity and tolerance of unequal numbers in each group, although balanced numbers are less important now that good statistical software is available for analyzing more complex designs with unequal numbers in each group. However, simple randomization cannot take account of heterogeneity of experimental material or variation (e.g., due to biological rhythms or environment), which cannot be controlled over a period of time.

Randomized complete block designs are used to split an experiment into a number of "mini-experiments" to increase

[1]Abbreviations used in this article: ANOVA, analysis of variance; DF, degrees of freedom.

precision and/or take account of some natural structure of the experimental material. With large experiments, it may not be possible to process all of the animals at the same time or house them in the same environment, so it may be better to divide the experiment into smaller blocks that can be handled separately. Typically, a "block" will consist of one or more animals (or other experimental units) that have been assigned at random to each of the different treatment groups. Thus, if there are six different treatments, a block will consist of a multiple of six animals that have been assigned at random to each of the treatments. Blocking thus ensures balance of treatments across the variability represented by the blocks. It may sometimes be desirable to perform within-litter experiments when, for example, comparing transgenic animals with wild-type ones, with each litter being a block. Similarly, when the experimental animals differ excessively in age or weight, it may be best to choose several groups of uniform animals and then assign them to the treatments within the groups. Randomized block designs are often more powerful than completely randomized designs, but their benefits depend on correct analysis, using (usually) a two-way ANOVA without interaction. Note that when there are only two treatments, the block size is two and the resulting data can be analyzed using either a paired *t*-test or the two-way ANOVA noted above, which are equivalent.

Choice of Dependent Variable(s), Characters, Traits, or Outcomes

Confirmatory experiments normally have one or a few outcomes of interest, also known as dependent variables, which are typically mentioned in the experimental hypotheses. For example, the null hypothesis might be that the experimental treatments do not affect body weight in rats. Ideally there should be very few outcomes of primary interest, but some toxicity experiments involve many dependent variables, any of which may be altered by a toxic chemical. Exploratory experiments often involve many outcomes, such as the thousands of dependent variables in microarray experiments. When there is a choice, quantitative (measurement) data are better than qualitative data (e.g., counts) because the required sample sizes are usually smaller. When there are several correlated outcomes (e.g., organ weights), some type of multivariate statistical analysis may be appropriate.

In some studies, scores such as 0, +, ++, and +++ are used. Such "ordinal" data should normally be analyzed by comparing the number in each category among the different treatment groups, preferably taking the ordering into account. Converting scores to numerical values with means and standard deviations is inappropriate.

Choice of Independent Variables or Treatments

Experiments usually involve the deliberate alteration of some treatment factor such as the dose level of a drug. The treatments may include one or more "controls." Negative controls may be untreated animals or those treated with a placebo without an active ingredient. The latter is normally more appropriate, although it may be desirable to study both the effect of the active agent and the vehicle, in which case both types of control will be needed. Surgical studies may involve sham-operated controls, which are treated in the same way as the tested animals but without the final surgical treatment.

Positive controls are sometimes used to ensure that the experimental protocols were actually capable of detecting an effect. Failure of these controls to respond might imply, for example, that some of the apparatus was not working correctly. Because these animals may suffer adverse effects, and they may not be necessary to the hypothesis being tested, small numbers may be adequate.

Dose levels should not be so high that they cause unnecessary suffering or unwanted loss of animals. When different doses are being compared, three to approximately six dose levels are usually adequate. If a dose-response relation is being investigated, the dose levels (X-variable) should cover a wide range to obtain a good estimate of the response, although the response may not be linear over a wide range. Dose levels are frequently chosen on a log 2 or log 10 scale. If the aim is to test for linearity, then more than two dose levels must be used. If possible, we recommend using dose levels that are equally spaced on some scale, which may facilitate the statistical analysis. More details of choice of dose levels and dilutions in biological assay are given by Finney (1978).

Toxicologists often use fractions (e.g., half to a quarter or less) of the maximum tolerated dose (the largest dose that results in only minimal toxic effects) in long-term studies. The scientific validity of using such high dose levels has been questioned because the response to high levels of a toxic chemical may be qualitatively different from the response to low levels (Fairweather et al. 1998). The possibility of exploring the effects of more than one factor (e.g., treatment, time, sex, or strain) using factorial designs (see below) should be considered.

Uncontrolled (Random) Variables

In addition to the treatment variables, there may be a number of random variables that are uncontrollable yet may need to be taken into account in designing an experiment and analyzing the results. For example, circadian rhythms may cause behavior measured in the morning to be different from that measured in the afternoon. Similarly, the experimental material may have some natural structure (e.g., members of a litter of mice may be more similar than animals of different litters). Measurements made by different people or at different times may be slightly different, and reagents may deteriorate over a period of time. If these effects are likely to be large in relation to the outcomes being investigated, it will be necessary to account for them at the design stage (e.g., using a randomized block, Latin square, or other appropriate design) or at the time of the statistical analysis (e.g., using covariance analysis).

Factorial Experiments

Factorial experiments have more than one type of treatment or independent variable (e.g., a drug treatment and the sex of the animals). The aim could be to learn whether there is a response to a drug and whether it is the same in both sexes (i.e., whether the factors interact with or potentiate each other). These designs are often extremely powerful in that they usually provide more information for a given size of experiment than most single factor designs at the cost of increased complexity in the statistical analysis. They are described in most statistical texts (e.g., Cox 1958; Montgomery 1997).

In some situations, a large number of factors that might influence the results of an experiment can be studied efficiently using more advanced factorial designs. For example, in screening potential drugs, it may be desirable to choose a suitable combination of variables (e.g., presence/absence of the test compound; the sex, strain, age, and diet of the animals; time after treatment; and method of measuring the endpoint). If there were only two levels of each of these variables, then there would be $2^7 = 128$ treatment combinations to be explored. Special methods are available for designing such experiments without having to use excessively large numbers of animals (Cox 1958: Cox and Reid 2000; Montgomery 1997). This type of design can also be used to optimize experiments that are used repeatedly with only minor changes in the treatments, such as in drug development, when many different compounds are tested using the same animal model (Shaw et al. 2002).

Experiment Size

Deciding how large an experiment needs to be is of critical importance because of the ethical implications of using animals in research. An experiment that is too small may miss biologically important effects, whereas an experiment that is too large wastes animals. Scientists are often asked to justify the numbers of animals they propose to use as part of the ethical review process.

Power Analysis

A power analysis is the most common way of determining sample size. The appropriate sample size depends on a mathematical relation between the following (described in more detail below): the (1) effect size of interest, (2) standard deviation (for variables with a quantitative effect), (3) chosen significance level, (4) chosen power, (5) alternative hypothesis, (6) sample size. The investigator generally specifies the first five of these items and these determine the sample size. It is also possible to calculate the power or the effect size if the sample size is fixed (e.g., as a result of restricted resources). The formulae are complex; however, several statistical packages offer power analysis for estimating sample sizes when estimating a single mean or proportion, comparing two means or proportions, or comparing means in an analysis of variance. There are also dedicated packages (e.g., nQuery Advisor [Statistical Solutions, Cork, UK; Elashoff 1997]), which have a much wider range of analyses (Thomas 1997). A number of web sites also provide free power analysis calculations for the simpler situations, and the following sites are currently available: <Http://ebook.stat.ucla.edu/cgi-bin/engine.cgi>; http://www.math.yorku.ca/SCS/Demos/power/;> and <http://hedwig.mgh.harvard.edu/quan_measur/para_quant.html>. Sample size is considered in more detail by Dell and colleagues in this volume (2002), and Cohen (1988) provides extensive tables and helpful discussion of methods.

Effect Size

Briefly, when only two groups are to be compared, the effect size is the difference in means (for a quantitative character) or proportions (for a qualitative, dead/alive character) that the investigator wants the experiment to be able to detect. For example, the investigator could specify the minimum difference in mean body weight between a control group of rats and a treated group that would be of biological importance and that he/she considers the experiment should be able to detect. It is often convenient to express the effect size "D" in units of standard deviations by dividing through by the standard deviation (discussed below). D is a unitless number that can be compared across different experiments and/or with different outcomes. For example, if the standard deviation of litter size in a particular colony of BALB/c strain mice is 0.8 pups (with a mean of ~ 5 pups) and an experiment is to be set up to detect a difference in mean litter size between treated and control groups of, for example, 1.0 pups, then $D = 1.0/0.8 = 1.25$ standard deviation units. If the standard deviation of the total number of pups weaned per cage in a 6-mo breeding cycle is 10 pups (with a mean of ~ 55 pups) and the experiment is set up to detect a difference between a control group and a treated group of 5.0 pups, then $D = 5/10 = 0.5$. This effect size is smaller, so would require a larger experiment than the change in litter size would require. Similarly, if a control group is expected to have, for example, 20% of spontaneous tumors, and the compound is a suspected carcinogen, the increase in the percentage of tumors in the treated group (which would be important to be able to detect) must be specified.

Standard Deviation

The standard deviation among experimental units appropriate to the planned experimental design must be specified (for quantitative characters). For a randomized block or crossover design the appropriate estimate will usually be the square root of the error mean square from an analysis of variance conducted on a previous experiment. When no previous study has been done, a pilot study may be used, although the estimate will not be reliable if the pilot study is very small.

Significance Level

The significance level is the chance of obtaining a false-positive result due to sampling error (known as a Type I error). It is usually set at 5%, although lower levels are sometimes specified.

Power

The power of an experiment is the chance that it will detect the specified effect size for the given significance level and standard deviation and be considered statistically significant. Choice of a power level is somewhat arbitrary and usually ranges from 80 to 95%. However, when testing some vaccines for virulence, a power as high as 99% may be specified because of the serious consequences of failure to detect a virulent batch. Note that (1-power) is the chance of a false-negative result, also known as a Type II error.

Alternative Hypothesis

The alternative hypothesis is usually that two means or proportions differ, leading to a two-tailed test; but occasionally, the direction of the difference is specified, leading to a one-tailed test. A slightly larger sample size is required for a two-tailed test.

Sample Size

The sample size is usually what needs to be determined, so all of the other quantities listed above should be specified. However, there are occasions when the sample size is fixed and the aim is to determine the power or effect size, given sample size.

Estimated sample sizes for an experiment involving two groups with measurement data that would be analyzed using a two-sample *t*-test are given in Table 1 as a function of D (see above). For the two examples above, D was 1.25 for the litter size effect, which would require approximately 14 cages in each group; whereas for the total production, example D was 0.5, which would require approximately 86 cages in each group.

When experiments are set up to compare two proportions using a chi-squared test the effect size is the difference in the proportion of "successes" in the two groups and the standard deviation is specified by the two proportions. In Table 2 are shown the estimated number required in each group to compare two proportions for various proportions ranging from 0.2 to 0.8 assuming a power of 90%, a significance level of 5%, and a two-sided test. Note that larger sample sizes are required to detect a given difference between two proportions if they are both high or low (i.e., less than 0.3 or more than 0.7) than if they are near 0.5.

Power analysis can also be used to estimate the required sample sizes for estimating parameters such as a mean, a regression coefficient, survival, or a genetic linkage (recombination proportion) with a specified confidence interval, although dedicated power analysis software may be needed for these more advanced calculations. Note that large numbers of animals are needed to estimate genetic linkage between tightly linked genetic markers if a narrow confidence

Table 1 Sample size as a function of D[a] for a two-sample t-test comparison assuming a significance level of 5%, a power of 90%, and a two-sided test

D	No. per group
0.2	527
0.3	235
0.4	133
0.5	86
0.6	60
0.7	44
0.8	34
0.9	27
1.0	23
1.2	16
1.4	12
1.6	10
1.8	8
2.0	7
2.5	5

[a]D = (difference in means)/(standard deviation)

Table 2 Number required in each group for comparing two proportions (based on a normal approximation of the binomial distribution) with a significance level of 0.05 and a power of 90%

Proportion in each group	0.2	0.3	0.4	0.5	0.6	0.7
0.2	—					
0.3	392	—				
0.4	109	477	—			
0.5	52	124	519	—		
0.6	90	56	130	519	—	
0.7	19[a]	31	56	124	477	—
0.8	13[a]	19[a]	30	52	109	392

[a]Assumptions may lead to some inaccuracy.

interval is wanted. However, several specialized approaches to such studies exist (Silver 1995).

Resource Equation Method for Determining Sample Size

When there is no information about the standard deviation and/or it is difficult to specify an effect size, an alternative method that depends on the law of diminishing returns has been suggested (Mead 1988). This method may also be of value for some exploratory experiments when testing hypotheses is not the main objective.

For quantitative characters that are analyzed using the analysis of variance, it is suggested that the degrees of freedom (DF[1]) for the error term used to test the effect of the variable should be approximately 10 to 20. With less than 10 DF, good returns can be expected from adding more experimental units. However, with more than 20 DF, adding additional units provides little extra information. This rule-of-thumb method seems to work quite well for whole animal experiments, although it tends to assume quite large effect sizes.

Statistical Analysis

The results of most experiments should be assessed by an appropriate statistical analysis even though, in some cases, the results are so clear-cut that it is obvious that any statistical analysis would not alter the interpretation. The analysis should reflect the purpose of the study. Thus, the goal of an exploratory analysis is to identify patterns in the data without much emphasis on hypothesis testing, the goal of a confirmatory experiment is to test one or a few prestated hypotheses, and experiments aimed at estimating a parameter such as a genetic linkage require appropriate estimates and standard errors. The general aim, however, is to extract all of the useful information present in the data in a way that it can be interpreted, taking account of biological variability and measurement error. It is particularly useful in preventing unjustified claims about the effect of a treatment when the results could probably be explained by sampling variation. Note that it is possible for an effect to be statistically significant but of little or no biological importance. The materials and methods section should describe the statistical methods used in analysing the results. The aim should be to "describe statistical methods with enough detail to enable a knowledgeable reader with access to the original data to verify the reported results" (ICMJE 2001—<www.icmje.com>).

Examining the Raw Data

Raw data and data entered into statistical software should be studied for consistency and any obvious transcription errors. Graphical methods, which are available in most statistical packages, are helpful, particularly if individual observations can be seen clearly. "Outliers" should not be discarded unless there is independent evidence that the observation is incorrect, such as a note taken at the time the observation was recorded expressing doubt about its credibility. Exclusion of any observations should be stated explicitly, with reasons. It is sometimes useful to analyze the data with and without the questionable data to learn whether they alter the conclusions. A clear distinction must be made between missing data (caused, for instance, by an animal dying prematurely or being killed due to excessive suffering) and data with a value of zero.

Thought should be given at the design stage to dealing with unexpected deaths, particularly if they are related to the treatments. Details will depend on the nature of the study and the number of animals that die. For example, the death of only one or two animals in a relatively large study may have relatively little effect on the results. In some long-term studies, it may be possible to replace animals that die early, but this replacement is not usually practical when they die late. In other cases, some useful information (e.g., body weights and DNA) can be obtained from the carcasses, provided they are preserved. Treatment-related deaths may bias some of the results. For example, experimental stress could result in some of the smaller animals in one of the treatment groups dying, thereby causing the average weight of the group to appear heavier than it should be. In all cases, the number and treatment groups of any animals that die should be noted in the published paper or report. In principle, data should be kept as "raw" as possible. For example, expressing some numbers as percentages of other numbers should be avoided because it may complicate the statistical analysis and interpretation of the results and/or reduce precision.

Quantitative Data: Parametric and Nonparametric Methods

The method of statistical analysis depends on the purpose of the study, the design of the experiment, and the nature of the resulting data. For example, an analysis involving a test of an hypothesis should not be used if the aim is to estimate the slope of a regression line. Quantitative data are often summarized in terms of the mean, "n" (the number of subjects), and the standard deviation as a measure of variation. The median, n, and the interquartile ranges (i.e., the 25th and 75th centiles) may be preferable for data that are clearly skewed. Nonparametric methods are discussed separately below.

Quantitative data can be analyzed using "parametric" methods, such as the *t*-test for one or two groups or the ANOVA for several groups, or using nonparametric methods such as the Mann-Whitney test. Parametric tests are usually more versatile and powerful and so are preferred; however, they depend on the assumptions that the residuals (i.e., deviation of each observation from its group mean) have a normal distribution, that the variances are approximately the same in each group, and that the observations are independent of each other. The first two of these assumptions should be investigated as part of the analysis by studying the residuals using methods available in most statistical software packages. A normal probability plot of the residuals will show whether the normality assumption is fulfilled (Altman 1991). This type of plot should give a straight line with a normal distribution of residuals. A plot of the fits (estimated group means) versus the residuals will show whether the variation is approximately the same in each group. Both plots also tend to highlight any outliers. Observations must also be independent (i.e., the observations within a treatment group must come from experimental units, which could, in principle, have been assigned to different treatment groups). Thus, if the effect of different diets on mouse body weight are to be compared using several cages with, for example, five mice per cage, the metric to be analyzed should be the mean of all animals in the cage—not the individual mouse weights—because mice within the cage cannot be assigned to different treatment groups, so they are not statistically independent. If the number of mice per cage varies, then this may need to be taken into account in the statistical analysis.

Nesting

Where several observations can be made on an experimental unit (e.g., weights of individual animals within a cage, as above or randomly chosen microscope fields within histological sections from an animal), it may be important to find out whether precision could be increased more effectively by using more experimental units or more observations within each unit. In such situations, the observations are said to be "nested" within the experimental units, and several levels of nesting are possible. A nested ANOVA is usually used with the aim being to estimate the "components of variance" associated with each level of nesting (Dixon and Massey 1983). When this information is combined with the costs of experimental units and observations within a unit, it is possible to estimate the best way to increase precision. In general, extra replication is necessary across the level of nesting with the most variation. Thus, if there are large differences among scores of microscopic fields within an animal, it will usually be better to sample more fields than to use more animals, although this sampling depends on the relative costs of animals and observations.

Nesting may also involve a fixed effect. For example, a number of animals may be assigned at random to some treatment groups and the concentration of a metabolite may then be measured in a number of different organs. A nested statistical analysis can then be used to determine whether there are differences among treatment means, whether there are differences among named organs, and whether there is an organ by treatment interaction. The analysis is somewhat similar to that used for a split-plot design.

Transformations

If the variances are not the same in each group and/or the residuals do not have a normal distribution, a scale transformation may normalize the data. A logarithmic transformation may be appropriate for data such as the concentration of a substance, which is often skewed with a long tail to the right. A logit transformation $\{\log_e(p/(1-p))\}$ where p is the proportion, will often correct percentages or proportions in which there are many observations less than 0.2 or greater than 0.8 (assuming the proportions cannot be < 0 or > 1), and a square root transformation may be used on data with a Poisson distribution involving counts when the mean is less than about five.

Further details are given in most statistics textbooks. If no suitable transformation can be found, a nonparametric test can often be used (see below).

Multiple Comparisons

Student's *t*-test should not be used to compare more than two group means. It lacks power, and multiple testing increases the chance of a false-positive result. When there are two or more groups, and particularly with randomized block or more complex designs, the ANOVA can be used initially to test the overall hypothesis that there are no differences among treatment means. If no significant differences are found, then further comparisons of means should not be done. When the ANOVA results are significant (e.g., at $p < 0.05$) and several groups are being compared, either *post-hoc comparisons* or *orthogonal contrasts* can be used to study differences among individual means.

A range of post-hoc comparison methods are available that differ slightly in their properties. These include Dunnett's test for comparing each mean with the control, Tukey's test, Fisher's protected least-significant difference test, Newman-Keuls test, and several others for comparing all means. Large numbers of post-hoc comparisons should be avoided because some of these tests are "conservative" and fail to detect true differences (Type II errors) whereas others may be too liberal and give false-positive results (Type I errors). It is better to specify those few comparisons of particular interest at the design stage. Authors should state which tests have been used. Note that all of these tests use the pooled within-group standard deviation obtained from the ANOVA. The ANOVA followed by individual *t*-tests to compare means, not using the pooled standard deviation, is not acceptable because each test will lack power due to the low precision of the estimates of individual standard deviations.

Group means can also be compared using so-called "orthogonal contrasts." Depending on the types of treatment, either this method can compare individual means or groups of means or, if the treatments represent dose levels or time and are equally spaced on some scale, these contrasts can be used to test linearity and nonlinearity of response. Unfortunately, the methods are available only in more advanced statistical packages, although the calculations can be done manually. More details are given by Montgomery (1997).

The best estimate of the pooled standard deviation is obtained as the square root of the error mean square in the ANOVA. Indeed, this is the only estimate of the standard deviation that is available for a randomized block design. Thus, when presenting means either in tables or graphically, this estimate of the standard deviation should be used. It will, of course, be the same for each group.

Several Dependent Variables

When there are several dependent variables (characters), each can be analyzed separately. However, if the variables are correlated, the analyses will not be independent of one another. Thus, if sampling variation resulted in a false-positive or false-negative result for one character, the same thing may happen for another character. A multivariate statistical analysis such as principal components analysis could be considered in such cases (Everitt and Dunn 2001).

Serial Measurements

Data on experimental subjects are sometimes collected serially. For example, growth curves, response to pharmaceutical or toxic agents, behavioral measurements, and output from telemetric monitoring may involve repeated measurement on individual animals. Although a repeated measures ANOVA has sometimes been used to analyze such data, this approach is best avoided because the results are difficult to interpret and the assumptions underlying the analysis are rarely met. Appropriate summary measures such as the mean of the observations, the slope of a regression line fitted to each individual, the time to reach a peak or the area under the curve, depending on the type of observed response, offer a better alternative that is easier to interpret (Matthews et al. 1990), although other methods such as a multivariate analysis are also available (Everitt 1995).

Nonparametric Tests

When the assumptions necessary for the *t*-test and the ANOVA of approximately equal variation in each treatment group and approximate normality of the residuals are not valid and no scale transformation is available to correct the heterogeneity of variance and/or non-normality, a nonparametric test can usually be used to compare the equality of population means or medians. For comparing two groups, the Wilcoxon rank sum test and the Mann-Whitney test (which are equivalent) constitute a nonparametric equivalent of the two-sample *t*-test. For comparing several groups, the Kruskal-Wallis is the nonparametric equivalent of the one-way ANOVA. A nonparametric equivalent of a post-hoc comparison can be used, provided the overall test is significant (Sprent 1993). A version of the Wilcoxon test can also be used as the nonparametric version of the paired *t*-test for a randomized block design with two treatment groups.

The Friedman test is the nonparametric equivalent of the randomized block ANOVA for more than two treatment groups. Several other nonparametric tests are appropriate for particular circumstances, and they are described in most statistics textbooks.

Correlation

The most common correlation coefficient is known more formally as the *product-moment correlation,* or Pearson correlation to distinguish it from several other types. It is used for assessing the strength of the *linear* relation between two numerical variables A and B. Both A and B are assumed to be subject to sampling variation. It does not assume that variation in A causes variation in B or vice versa. The correlation can be shown graphically using a scatter plot. Normally a best fitting line should not be shown. The investigator who wishes to fit such a line should remember that the line calculated from the regression of A on B will normally be different from that due to the regression of B on A. The usual hypothesis test is that the correlation is zero; however, in some cases, it may be appropriate to test whether the correlation differs from some other defined value. Note that a change of scale will alter the correlation coefficient and that a nonlinear relation will result in a low correlation even if the two variables are strongly associated. In such circumstances, use of the correlation of ranks may be more appropriate. There are several other forms of correlation coefficient, depending on whether the variables are measurements or ranks or are dichotomous.

Regression

Regression analysis can be used to quantify the relation between two continuous variables X and Y, where variation in X is presumed to cause variation in Y. Regression is thus asymmetric with respect to X and Y. The X variable is assumed to be measured without error. Linear regression can be used to fit a straight line of the form $Y = a + bX$, where a and b are constants that are estimated from the data using the least-squares method. In this case, "a" (the intercept) represents the value of Y when X is zero, and "b" is the slope of the regression line. A positive value of b implies that the slope rises from left to right, and a negative value implies that it declines. Confidence intervals can be obtained for the slope and can be fitted around the regression line to give, for example, a 95% confidence interval for the mean value of Y for a given value of X. Prediction intervals can also be fitted to give, for example, a 95% interval for the variation of individual observations of Y for any given value of X. When possible, it is important to quote R^2, which is interpreted as the proportion of the variability in the data explained by regression. This may be low if the X-variable does not have a reasonably large range. The residual (error) variation from the ANOVA table should also be quoted.

If animals are caged in groups and the X variable (e.g., the dose level of a test compound or a dietary ingredient) is administered to whole cages, then the cage becomes the experimental unit and the Y variable will be the mean of all the animals in the cage. If the number of animals per cage varies, then more weight can be given to cages with more animals. A weighted regression analysis, which takes account of possible variation in the precision with which each point is estimated, is available in many statistical packages. Quadratic regression can be used to fit a curve to the data points. Many other types of curve can be fitted, and some have useful biological interpretations.

The effect of several independent X variables can be evaluated simultaneously using multiple regression. Often such an analysis is exploratory, with the aim of identifying which variables are influential. Logistic regression can be used to explore the relation between one or more predictor variables and a binary (e.g., dead/alive) outcome.

Regression and the ANOVA are closely related so that a regression of, for example, response on dose level can sometimes be included as part of the ANOVA using orthogonal contrasts (Altman 1991; Dixon and Massey 1983; Montgomery 1997). The usual statistical test in regression analysis is of the null hypothesis that there is no linear relation between X and Y. Other common tests are of whether two regression lines have the same slope b and/or the same intercept a. A test to determine whether there is a quadratic relation would be a test of whether a quadratic curve gives a significantly better fit than a straight line.

Categorical Data

Categorical data consist of counts of the number of units with given attributes. These attributes can be described as "nominal" when they have no natural order (e.g., the strain or breed of the animals). They are described as "ordinal" when they have a natural order such as low, medium, and high dose levels or scores, which may also be defined numerically. When there are two categories, the data are called binary. Categorical data are often presented in the form of frequency tables and/or as proportions or percentages.

Proportions or percentages should be accompanied by a confidence interval (preferably) or standard error, and n should be clearly indicated. The usual method of comparing two or more proportions is a contingency table chi-squared analysis, which tests the null hypothesis that rows and columns are independent. The method is inaccurate if the numbers in some cells are very low. Fisher's exact test can be used in such cases. Other methods of analysis are available and are described in some texts.

Presentation of the Results

When individual means are quoted, they should be accompanied by some measure of variation. Excess decimal places, often produced by the computer, should be eliminated. It is usually sufficient to quote means to three significant digits (e.g., 11.4 or 0.128). Percentages can often be rounded to the nearest whole number. If the aim is to describe the variation among individuals that contribute to the mean, then the standard deviation should be given. Avoid using the ± sign. When presenting means it is better to use a designation such as "mean 9.6 (SD 2.1)" because it avoids any confusion between standard deviation and standard error. When the aim is to show the precision of the mean, a confidence interval (e.g., 9.6 [95% CI = 7.2-12.0]) should be used (preferably) or a standard error (e.g., 9.6 [SE 1.2]). Actual observed *p* values should be quoted whenever possible, rather than using < or > signs, although if these values are very low, a < sign can be used (e.g., $p < 0.001$). Lack of statistical significance should not be used to claim that an effect does not exist. Nonsignificance may be due to the experiment being too small or the experimental material being too variable.

When two means are compared, the size of the difference between them should be quoted together with a confidence interval. When nonparametric analyses have been done, it is more sensible to quote medians and, for example, 25 and 75% centiles indicating the interquartile range. When proportions or percentages are given, a standard error or confidence interval and n should also be given. When proportions are compared, the confidence interval for the difference (or ratio) should be supplied.

We advise showing tabulated means in columns rather than rows because this arrangement makes it easier to compare values. If the means have been compared using a *t*-test or ANOVA and the standard deviations have been found not to differ materially between groups, use of a pooled standard deviation may be more appropriate than showing the standard deviations separately for each mean. The number of observations should always be indicated.

Graphical Presentation of Data

Graphs are especially valuable to illustrate points that would be difficult to explain in writing or in a table. Presentation of a small number of means can often be done more clearly and using less space using a table than a bar diagram. It is also easier to read numerical values off a table than to read them off a graph.

Graphs showing individual points rather than bar charts or graphs with error bars are strongly encouraged because they provide a much clearer impression of the nature of the data. For example, a dose-response curve with individual points (Figure 1) provides a much clearer impression of individual variation than the example in Figure 2, which tends to give a false impression of uniformity at each dose level.

When means have been compared statistically, it may be better to indicate significant differences on the diagram, rather than adding error bars. When error bars are shown on graphs or bar diagrams, there should be a clear indication of whether these are standard deviations, standard errors, or confidence intervals (preferred), and the number of observations should be clearly indicated in the text or figure caption. With more complex graphs, it may be better not to use error bars but instead to summarize the data in an accompanying table. Regression lines should never be shown without the data points; preferably, they should be shown with a confidence interval and/or prediction interval.

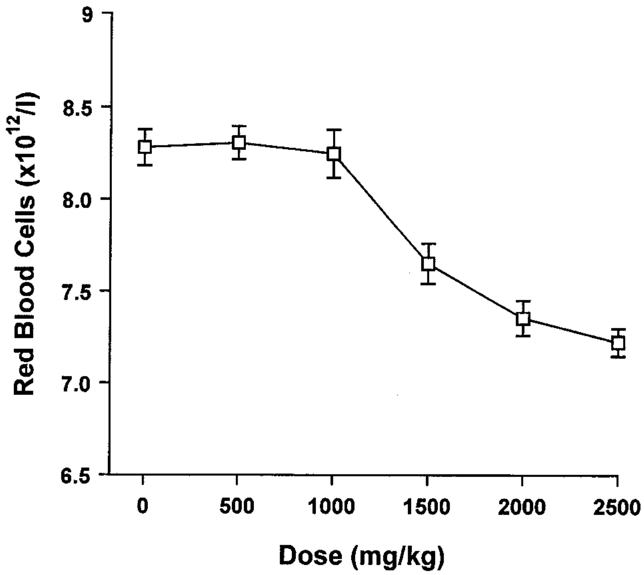

Figure 2 Same data as in Figure 1, but just showing the group means and error bars of one standard error about each mean. This type of presentation is not recommended as it tends to obscure individual variability.

Combining Data from Different Studies: Meta-analysis

Sometimes answers to the same essential questions are sought in several independent experiments or trials from different investigators. Formal methods of "meta-analysis" have been developed that attempt to combine the results of different experiments taking account of sample sizes and apparent quality of the data. Meta-analysis usually forms only part of a systematic review to identify all relevant studies (Egger et al. 2001). There are a number of difficulties in doing such reviews, one of which is publication bias. Many studies are published only if they give positive results because journals are often reluctant to publish studies where differences are not statistically significant. For example, findings that some types of environmental enrichment benefit laboratory mice are more likely to be published than those that find there is no effect. Thus, if only published studies are included in the meta-analysis, the case for environmental enrichment might appear to be overwhelming. Unfortunately, no mechanism exists for finding unpublished data.

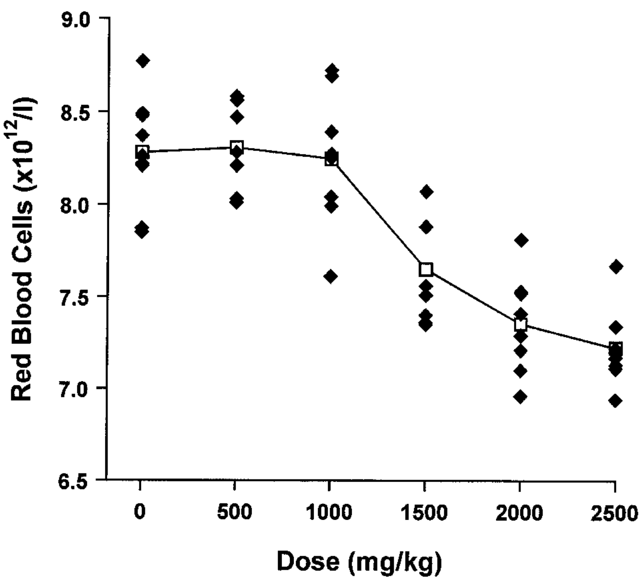

Figure 1 Red blood cell counts in mice as a function of the dose of chloramphenicol showing counts for individual mice with a line connecting the mean count at each dose level. Note that this example provides a better impression of the variability of the data than Figure 2. Raw data from Festing MFW, Diamanti P, Turton JA. 2001. Strain differences in haematological response to chloramphenicol succinate in mice: Implications for toxicological research. Food Chem Toxicol 39:375-383.

Despite this potential difficulty, bringing together all relevant research evidence in a topic should be generally encouraged. A key aspect of such review is to assess the methodological quality of the individual studies. Meta-analysis of the results from several studies may then be done for those studies deemed to be scientifically reliable and addressing the same question. Although various statistical methods are available, meta-analysis may not be straightforward, however, especially for observational studies.

Use of Historical Data

The value of historical data depends on its quality and its reliability. Many factors (e.g., strain, origin, associated mi-

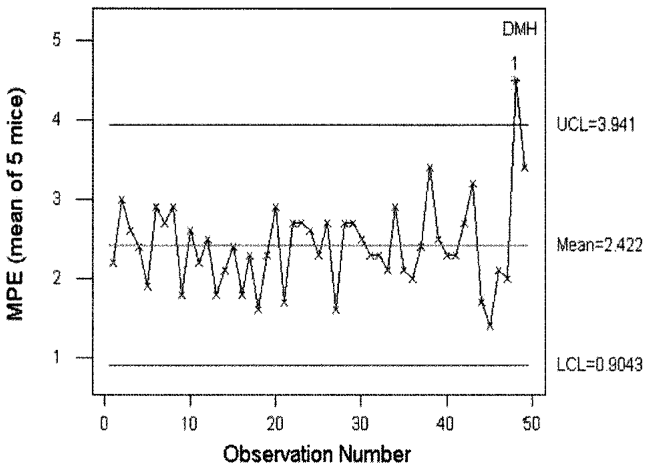

Figure 3 Control chart of micronuclei counts per 1000 polychromatic erythrocytes in 47 batches of five control mice and two batches of mice treated with 1,2-dimethylhydrazine. Such a chart provides one method of making use of relatively homogeneous sets of historical control data collected within a single laboratory over a long period of time. (Data used to illustrate the point, although the real time sequence was not available, raw data was extracted from: Morrison V, Ashby J. 1995. High resolution rodent bone marrow micronucleus assays of 1,2-dimethylhydrazine: Implications of systemic toxicity and individual responders. Mutagenesis 10:129-135.)

croflora, housing, husbandry, and methods of measuring each outcome) can influence individual results so that in nearly all studies, contemporary controls are almost essential, and historical data, particularly from another laboratory should be treated with considerable caution. Methods of meta-analysis may be appropriate in some cases.

However, when similar experiments are performed repeatedly in the same laboratory, there will often be scope for using historical data. For example, chemicals are often routinely tested to determine whether they produce micronuclei in mice when given by injection. Usually a laboratory will standardize on a single strain and sex of mice and use a standard protocol that includes contemporary controls. Quality control charts, often used in industry, provide one method of using such data (Hayashi et al. 1994). In Figure 3 is shown a control chart of the mean number of micronuclei in 47 samples of five control mice collected over a period of several months in one laboratory, with the last two samples of mice treated with 35 mg/kg of 1,2-dimethylhydrazine. The control chart shows the mean number of micronuclei among the control samples with upper and lower control limits. One of the samples of mice treated with the 1,2-dimethylhydrazine clearly exceeds the upper control limit and has been flagged by the computer. Careful use of such techniques, which need further development for use in a biological context, might mean that smaller sample sizes could be used in each study.

Conclusions

The need for improved experimental design and statistical analysis of animal experiments, if they are to be considered ethically acceptable, has already been emphasized. However, a recent example re-emphasizes this. A meta-analysis of 44 animal studies on fluid resuscitation (Roberts et al. 2002) reported that only two of the investigators stated how the animals were allocated to the treatment groups, none of them were sufficiently large to detect a halving in the risk of death reliably, there was considerable scope for bias to enter into the conclusions, and there was substantial heterogeneity in the results due to the method of bleeding. Presumably the latter could have been detected using a factorial design with bleeding method as a design factor. The authors concluded that the odds ratios were impossible to interpret, and they questioned whether these animal data were of any relevance to human health care. If scientists are to have the privilege of being allowed to do painful experiments on animals, they must ensure that their experiments are beyond criticism.

Acknowledgment

Thanks are due to Peter Sasieni, Cancer Research UK, for helpful comments.

References

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

Altman DG. 2002. Poor quality medical research: What can journals do? JAMA (In Press).

Altman DG, Gore SM, Gardner MJ, Pocock SJ. 2000. Statistical guidelines for contributors to medical journals. In: Altman DG, Machin D, Bryant TN, Gardner MJ, eds. Statistics with Confidence. 2nd Ed. London: BMJ Books.

Boisvert DPJ. 1997. Editorial policies and animal welfare. In: van Zutphen LFM, Balls M, eds. Animal Alternatives, Welfare and Ethics. Amsterdam: Elsevier. p 399-404.

Cohen J. 1988. Statistical Power Analysis for the Behavioral Sciences. 2nd ed. Hillsdale: Lawrence Erlbaum Associates.

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

Cox DR, Reid N. 2000. The Theory of the Design of Experiments. Boca Raton: Chapman and Hall/CRC Press.

Dell R, Holleran S, Ramakrishnan R. 2002. Sample size determination. ILAR J 43:207-213. <http://www.national-academies.org/ilar>.

Dixon WJ, Massey FJJ. 1983. Introduction to Statistical Analysis. 4th ed. Auckland: McGraw-Hill International Book Co.

Egger M, Davey Smith G, Altman DG, eds. 2001. Systematic Reviews in Health Care. Meta-analysis in Context. 2nd Ed. London: BMJ Books.

Elashoff JD. 1997. nQuery Advisor Version 2.0 User's Guide. Cork: Statistical Solutions.

Everitt BS. 1995. The analysis of repeated measures: A practical review with examples. Statistician 44:113-135.

Everitt BS, Dunn G. 2001. Applied Multivariate Data Analysis. 2nd Ed. London: Arnold.

Fairweather WR, Bhattacharyya A, Ceuppens PP, Heimann G, Hothorn LA, Kodell RL, Lin KK, Mager H, Middleton BJ, Slob W, Soper KA, Stallard N, Ventre J, Wright J. 1998. Biostatistical methodology in carcinogenicity studies. Drug Infor J 32:401-421.

Festing MFW. 1994. Reduction of animal use: Experimental design and quality of experiments. Lab Anim 28:212-221.
Festing MFW. 1995. Use of a multi-strain assay could improve the NTP carcinogenesis bioassay program. Environ Health Perspect 103:44-52.
Festing MFW. 1997. Fat rats and carcinogen screening. Nature 388:321-322.
Festing MFW. 1999. Warning: The use of genetically heterogeneous mice may seriously damage your research. Neurobiol Aging 20:237-244.
Festing MFW. 2001. Guidelines for the design and statistical analysis of experiments in papers submitted to ATLA. ATLA 29:427-446.
Festing MFW, Lovell DP. 1995. The need for statistical analysis of rodent micronucleus test data: Comment on the paper by Ashby and Tinwell. Mutat Res 329:221-224.
Festing MFW, Lovell DP. 1996. Reducing the use of laboratory animals in toxicological research and testing by better experimental-design. J R Stat Soc 58(B-Methodol):127-140.
Festing MFW, van Zutphen LFM. 1997. Guidelines for reviewing manuscripts on studies involving live animals. Synopsis of the workshop. In: van Zutphen LFM, Balls M, eds. Animal Alternatives, Welfare and Ethics. Amsterdam: Elsevier. p 405-410.
Festing MFW, Diamanti P, Turton JA. 2001. Strain differences in haematological response to chloramphenicol succinate in mice: Implications for toxicological research. Food Chem Toxicol 39:375-383.
Festing MFW, Overend P, Gaines Das R, Cortina Borja M, Berdoy M. 2002. The Design of Animal Experiments: Reducing the Use of Animals in Research Through Better Experimental Design. London: Royal Society of Medicine Press Limited.
Finney DJ. 1978. Statistical Method in Biological Assay. 3rd Ed. London: Charles Griffin & Company Ltd.
Hayashi M, Hashimoto S, Sakamoto Y, Hamada C, Sofuni T, Yoshimura I. 1994. Statistical analysis of data in mutagenicity assays: Rodent micronucleus assay. Environ Health Perspect 102(Suppl 1):49-52.
ICMJE [International Committee of Medical Journal Editors]. 2001. Uniform requirements for manuscripts submitted to biomedical journals. <www.jcmje.org>.
Lipnick RL, Cotruvo JA, Hill RN, Bruce RD, Stitzel KA, Walker AP, Chu I, Goddard M, Segal L, Springer JA, Myers RC. 1995. Comparison of the up-and-down, conventional LD_{50}, and fixed-dose acute toxicity procedures. Food Chem Toxicol 33:223-231.
Matthews JNS, Altman DG, Campbell MJ, Royston P. 1990. Analysis of serial measurements in medical research. Br Med J 300:230-235.
Maxwell SE, Delaney HD. 1989. Designing experiments and analyzing data. Belmont CA: Wadsworth Publishing Company.
McCance I. 1995. Assessment of statistical procedures used in papers in the Australian Veterinary Journal. Aust Vet J 72:322-328.
Mead R. 1988. The Design of Experiments. Cambridge: Cambridge University Press.
Montgomery DC. 1997. Design and Analysis of Experiments. 4th Ed. New York: John Wiley & Sons.
Morrison V, Ashby J. 1995. High resolution rodent bone marrow micronucleus assays of 1,2-dimethylhydrazine: Implications of systemic toxicity and individual responders. Mutagenesis 10:129-135.
Muller KE, Barton CN, Benignus VA. 1984. Recommendations for appropriate statistical practice in toxicological experiments. Neurotoxicology 5:113-126.
Obrink KJ, Rehbinder C. 1999. Animal definition: A necessity for the validity of animal experiments? Lab Anim 34:121-130.
Rispin A, Farrar D, Margosches E, Gupta K, Stitzel K, Carr G, Greene M, Meyer W, McCall D. 2002. Alternative methods for the LD_{50} test: The up and down procedure for acute toxicity. ILAR J 43:233-243. <http://www.national-academies.org/ilar>.
Roberts I, Kwan I, Evans P, Haig S. 2002. Does animal experimentation inform human healthcare? Observations from a systematic review of international animal experiments on fluid resuscitation. Br Med J 324: 474-476.
Russell WMS, Burch RL. 1959. The Principles of Humane Experimental Technique. London: Methuen & Co. Ltd. [Reissued: 1992, Universities Federation for Animal Welfare, Herts, England.] <http://altweb.jhsph.edu/publications/humane_exp/het-toc.htm>.
Schlede E, Mischke U, Roll R, Kayser D. 1992. A national validation study of the acute-toxic-class method—An alternative to the LD_{50} test. Arch Toxicol 66:455-470.
Shaw R, Festing MFW, Peers I, Furlong L. 2002. The use of factorial designs to optimize animal experiments and reduce animal use. ILAR J 43:223-232. <http://www.national-academies.org/ilar>.
Silver LM. 1995. Mouse Genetics. New York: Oxford University Press.
Sprent P. 1993. Applied Nonparametric Statistical Methods. 2nd Ed. London: Chapman and Hall.
Stokes WS. 2000. Reducing unrelieved pain and distress in laboratory animals using humane endpoints. ILAR J 41:59-61. <http://www.national-academies.org/ilar>.
Thomas L. 1997. A review of statistical power analysis software. Bull Ecol Soc Am 78:126-139.

Appendix

Specification of the Animals Used in an Experiment

Scientific experiments should be repeatable, so it is important that the animals, their environment, and their associated micro-organisms are described as fully as possible (Obrink and Rehbinder 1999). Often the descriptions of the animals published in scientific papers are totally inadequate (Boisvert 1997). Scientists should also be aware that animals with the same designation from different sources or from one source at different times may be genetically different, and that the microbiological status of animals can influence their response to experimental treatments. The following checklist is based largely on one proposed by Festing and van Zutphen (1997). It should be used to help ensure that all details of the animals relevant to a particular study are fully described.

Specify in the paper as many as possible of the following:

Animals

Source: Species (with Latin name if not a common laboratory species), source, conservation status if wild, age and/or body weight, sex.
Transportation: Length of acclimatization period
Genotype: The breed, strain, or stock name. Inbred strains, mutants, transgenes, and clones should be described using internationally accepted nomenclature when available (see <www.informatics.jax.org> for mouse and rat nomenclature). Any genetic quality assurance verifying the genotype should be mentioned.
Microbiological status: Conventional, specified pathogen-free (SPF), germfree/gnotobiotic. When possible, reference should be made to some agreed-upon standards for microbiological characterization such as the FELASA standards (<www.felasa.org>).

Environment

Housing: Type of housing including whether conventional, barrier, isolator, or individually ventilated cages. Room

temperature (with diurnal variation), humidity, ventilation, light/dark periods, light intensity. Cage type, model, material, type of floor (solid/mesh), type of bedding, frequency of cage cleaning, number of animals per cage, cage enrichments.
Diet: Type, composition, manufacturer, feeding regimen (ad libitum, restricted, pair fed), method of sterilization.
Water: ad libitum, bottles or automatic, quality, sterilization.

Statistical Software

Many good statistical packages are now available, and the choice will often depend on which packages are supported by the particular research organization. Researchers are strongly urged to use one of the dedicated statistical packages, rather than a spreadsheet. Such packages have a wider range of statistical methods, the algorithms have usually been optimized over a period of several years, and the manuals often provide more help with the interpretation of the results than is available with a spreadsheet. In most cases, it is easy to paste material from a spreadsheet into a statistical package, so raw data can be kept in the spreadsheet if preferred.

Suggested Reading

There are numerous textbooks on statistics and experimental design. Most are directed at specific disciplines (e.g., agriculture, psychology, clinical medicine), but the methods are general and applicable to animal experiments. Anyone intending to continue with a research career should invest in a personal copy of a good textbook, which they should be able to consult for many years. A review of available textbooks is beyond the scope of this article, but the following books that are quoted herein (and several others not quoted) may be worth consulting, depending on the exact application: Altman (1991), Cox (1958), Cox and Reid (2000), Dixon and Massey (1983), Everitt and Dunn (2001), Festing et al. (2002), Finney (1978), Maxwell and Delaney (1989), Mead (1988), Montgomery (1997), Sprent (1993). Note that more recent editions of some of these books may be available since publication of this issue of *ILAR Journal*.