Principles: The need for better experimental design

Michael F.W. Festing

c/o FRAME (Fund for the Replacement of Animals in Medical Experiments), Russell and Burch House, 96–98 North Sherwood Street, Nottingham NG1 4EE, UK

Many experiments could be improved with better experimental design and statistical analysis. Badly designed experiments can lead to incorrect conclusions and wasted time and scientific resources. Such experiments are unethical if they involve animals or humans. Good experimental design requires clearly defined objectives and control of the major sources of variation. In this article, a small mouse experiment involving the response of a liver enzyme to the administration of an antioxidant is used to illustrate some important design concepts such as the control and partitioning of sources of variation using factorial and randomized block designs and the estimation of appropriate sample sizes. Scientists clearly need better training in experimental design with better access to consultant statisticians for more complex situations.

Many scientists ignore the basic principles of experimental design, analyse the resulting data badly, and in some cases reach the wrong conclusions. When such experiments involve animals or humans they are unethical and waste both money and scientific resources [1–5]. For example, a meta-analysis (statistical summary) of 44 animal experiments involving fluid resuscitation found that: (1) none of the experiments had sufficient power to detect reliably a halving in the risk of death, a clinically relevant outcome; (2) only two experiments explained how the animals had been allocated to the treatment groups; (3) there was substantial scope for bias; and (4) the results were highly heterogeneous because of the method of inducing the bleeding and thus the odds ratios were impossible to interpret [6]. The authors who carried out the meta-analysis questioned whether these experiments had any relevance to human medicine [6]. Although such meta-analyses can often be used to summarise the weight of evidence, it would be better to do fewer well-designed and ethically justified experiments instead of many small experiments of often doubtful scientific validity.

The findings of this meta-analysis are not surprising. Although there appears to be no published information on future research workers being trained in statistics, a poll of 59 PhD students at two UK universities showed that a third of the students had undergone no formal statistical training at the undergraduate level and two-thirds of the students had studied a course of <20 hours. In many countries there is often no formal teaching of statistical methods at the PhD level, and therefore future life science research workers are being denied the intellectual tools that they need to operate effectively. Even in the USA where course work is usual in PhD training, badly designed experiments are not uncommon.

Over the past three years FRAME (the Fund for the Replacement of Animals in Medical Experiments) through its 'Reduction' Committee has been raising people's awareness of this problem, and now the UK Medical Research Council, working through the Centre for Best Practice in Animal Research, is convening a working party to identify what can be done to improve the situation.

Common errors in experimental design

Experiments often have the potential for bias because subjects are not allocated at random and/or the treated and control groups are kept separately, for example, on different shelves in an animal room. Measurements taken from the treatment groups are sometimes performed at different times or even by different people from those of the control group. Some experiments even seem to be done in an *ad hoc* manner, with additional treatment groups being added during the course of the experiment. After subjects have been randomized to the treatment groups they and the treatments that they receive should, as far as possible, be coded so that the investigator does not know the group to which they belong. All subsequent manipulations and measurements should be done in random order.

Experiments often lack power: that is, they are unable to detect a clinically or biologically important response. This might be due to small sample sizes but is exacerbated by measurement error, the use of heterogeneous material and the use of statistical methods that lack power. The repeated use of the t test or Mann-Whitney tests, two statistical tests widely used by scientists, in experiments with several treatment groups is a prime example. One recent published study involved killing four treated rats and four control rats at 11 time periods post-treatment, with the aim of detecting the time-course of differences in three liver enzymes between the two groups. However, the experiment was analysed as 11 separate experiments using 33 t tests. Unfortunately, each of these tests was based on such small numbers of animals that they would be unlikely to detect any treatment differences, and among 33 tests some tests would be expected to show 'significance'

Corresponding author: Michael F.W. Festing (Michaelfesting@AOL.com).

http://tips.trends.com 0165-6147/03/$ - see front matter doi:10.1016/S0165-6147(03)00159-7

simply by chance. Because this was clearly a single experiment it should have been analysed using an analysis of variance (ANOVA) for each enzyme to determine the effect of the treatment, time and the interaction between treatment and time. This would show whether there were treatment differences only at certain time points. Each ANOVA would be based on 88 observations and according to the 'resource equation' method of determining sample size (see below) the experiment could then have been done using about half the number of rats.

Another common error is failure to identify correctly the 'experimental units': that is, those entities that can be assigned at random to a treatment. These experimental units could be cages of animals, individual animals, petri dishes, wells in a multi-well plate or even an organ preparation for a period of time if different treatments A, B and C can be applied sequentially in random order. Any two experimental units must be capable of being assigned to different treatments. Each experimental unit provides the metric used in the statistical analysis. For example, if a drug is administered in the diet or water to a cage of mice, with another cage of mice being untreated controls, it is the cage of mice, not the individual mice in the cage, that is the experimental unit. A statistical analysis using individual mouse data will show whether the groups differ with respect to the character of interest, but there is no assurance that this is due to the effect of the treatment. It could be because mice in one cage are fighting.

Designing better experiments

Designing experiments requires clear objectives, careful planning and should ensure that comparisons between groups are unbiased [7]. Each experiment should be large enough to have sufficient power to detect clinically or scientifically important results but should not be so large that they waste scientific resources. The repeatability of the experiment under different conditions needs to be considered, and the experiment should be simple, so as to avoid mistakes. An objective measure of the reliability of the results is also required, often in the form of standard deviations when describing estimates of the population variation, or standard errors or confidence intervals when describing the reliability of estimates of population parameters such as means.

The control of variation

An understanding of types of variation and how they are handled is of crucial importance. Variation, as a result of the species, sex, strain, age, bedding and diet of experimental animals, or the cell type, culture medium and culture conditions for *in vitro* studies, can be controlled directly by the scientist. These sources of variation, known as 'fixed effects', are either set at one level or deliberately varied as part of the design. If, for example, the sex of the animal is considered to be a potentially important factor that controls the response to a drug, the experiment could be split using half males and half females. If the sexes respond similarly any comparison of the effects of the drug will still involve the same number of animals, averaged across both sexes. If the sexes respond differently, this provides important new information, but the treatment means for each sex are still estimated reasonably well because the variation is assessed from the whole experiment. Such 'factorial' experiments are a powerful way of gaining extra information at little or no extra cost.

By contrast, 'random effect' variation consists of inter-individual variation, non-systematic measurement error, and variation associated with time and location. This type of variation can be minimized by careful choice of experimental material and the design of the experiment. Randomized block designs in which the experiment is split over two or more time periods can be used to break up an experiment into smaller, more homogeneous parts. Other things being equal, the more uniform the experimental units, the more powerful the experiment or the smaller the sample size that will be needed to achieve a given level of statistical power. When using rats or mice it is essential to use high-quality, disease-free animals of as uniform a weight and age as possible, although if the animals do vary in weight or age this can sometimes be accommodated using a randomized block design.

The use of isogenic (inbred or F1 hybrid) strains or outbred stocks when using rodents is discussed elsewhere [8,9] but the main disadvantage of outbred stocks is that they are usually more variable than inbred strains, which means that experiments using outbred stocks require larger sample sizes. The use of small numbers of animals of several inbred strains, rather than using the same number of outbred animals, has the advantage of giving an indication of genetic variation in response. Box 1 shows an experiment used to study the effects of an antioxidant on the activity of a liver enzyme in mice. By using healthy isogenic mice and a 2×4 factorial randomized block experimental design it was also possible to gain an idea of any genetic variation in response, even though the whole experiment involved only 16 animals.

Use of additional information

Experiments are commonly set up to test one or a few hypotheses but they often produce large volumes of data. There is a danger of attempting to use such data to answer additional questions that were not considered at the time the experiment was planned. The P values obtained in this way will be unreliable. However, such data can be used in several ways and, in particular, can be used to generate new hypotheses that can be tested in subsequent experiments [10]. In cases where several parameters are being measured it might also be helpful to use so-called 'multivariate methods', which can combine the parameters and reduce the total number of statistical tests that would otherwise be required [11].

Sample size

Scientists sometimes express concern at the apparently small sample sizes often found with factorial designs and wonder whether the results will be accepted by a good journal. In the example shown in Box 1, there were only two mice of each strain in each treatment group. However, the main comparison of butylated hydroxyanisole (BHA)-treated versus control animals involved eight animals in each group, albeit of four different genotypes. Had the experiment been performed with eight outbred animals

Box 1. An example of a well-designed small experiment

Table I shows the results of a small 2 (treatments) × 4 (strains) factorial randomized block experiment aimed at determining whether the antioxidant butylated hydroxyanisole (BHA) induces liver ethoxyresorufin O-deethylase (EROD) enzyme activity in mice, and whether there are important genetic differences in response. This experiment was one of a series of experiments on the effect of antioxidants on carcinogen-induced lung tumours. The individual mouse was the experimental unit. Females were used with the assumption that large sex differences in response are unlikely. In reporting such an experiment, the source of the animals, diet, caging, bedding and other factors should be specified so that the experiment can be repeated by other investigators (not shown here). Liver EROD activity was measured using standard protocols. The experiment also involved taking tissues from organs other than the liver and assessing the activity of several enzymes, and it was not possible to handle more than eight mice at a time. Accordingly, one mouse of each strain was assigned at random to each treatment group using a randomized block design, with blocks being separated by approximately two months.

The analysis of variance (ANOVA) used to analyse the data (Table II) shows the six sources of variation in the data, namely block, strain, treatment, strain x treatment, error and total variation. The block and error variation are due to random effects whereas the other sources of variation are fixed effects. The relative magnitude of each source of variation is shown in the 'sums of squares (SS)' column of Table II. The four *P* values shown in Table II indicate the probability that an effect at least as large as the effect observed could have arisen by chance sampling variation, and an effect is usually judged to be 'significant' if *P* is less than 0.05, although the actual *P* value should be given and the actual magnitude of the response should always be stated. Clearly, there was a massive treatment response ($F_{1,7} = 166$, $P < 0.0005$, where the subscript numbers denote the degrees of freedom in the F-test). The control group had a mean of 7.9 EROD units and the treated group had a mean of 18.1 EROD units, a difference of 10.3 units (95% confidence interval 8.4–12.2 units). There was some evidence of a treatment × strain interaction ($F_{3,7} = 5.19$, $P = 0.03$), suggesting that the strains responded slightly differently. This was mainly due to the slightly greater response in the BALB/c mice but in the context of the massive treatment response the interaction is probably of little biological significance. The block effect was large, showing that even with an identical protocol, measurements taken at different times can differ. Strain and treatment means are shown in Fig. I. Note that such a bar diagram gives a visual impression of the response. Error bars are not necessary because the data have been fully analysed in the ANOVA.

Table I. Effect of BHA on liver EROD activity in four mouse strains[a]

Strain	Treatment[b]	Block 1[c]	Block 2[c]
A/J	Control	7.7	6.4
129/Ola	Control	8.4	6.7
NIH	Control	9.8	8.1
BALB/c	Control	9.7	6.0
A/J	BHA	18.7	16.7
129/Ola	BHA	17.9	14.4
NIH	BHA	19.2	12.0
BALB/c	BHA	26.3	19.8

[a]Abbreviations: BHA, butylated hydroxyanisole; EROD, ethoxyresorufin O-deethylase.
[b]Treated mice were given BHA at 5 g per kg diet for two weeks. Control mice had the same diet without BHA.
[c]Data show enzyme activity (nmol resorufin formed per mg protein per min) in individual mice. Blocks were separated by approximately two months. These data came from Ministry of Agriculture Fisheries and Food Project FS1710 'Mechanisms of modulation of carcinogens by antioxidants: genetic control of the anticarcinogenic response in mice', and the experimental work was done by W.A. Evans (unpublished).

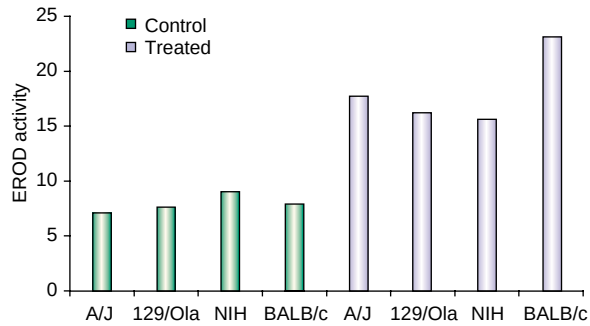

Fig. I. Ethoxyresorufin O-deethylase (EROD) enzyme activity in the livers of different strains of mice is shown. Data from control mice are shown in green, whereas data from mice treated with butylated hydroxyanisole (BHA) are shown in light blue. Each bar represents the mean of two mice. The standard deviation is 1.6 units. There was clearly a large and statistically significant treatment effect, with slight strain differences in response, largely as a result of the slightly greater response in BALB/c mice.

Table II. ANOVA of the data shown in Table I[a]

Source[b]	DF	SS	MS	F	*P*[c]
ANOVA for liver EROD activity					
Block	1	47.610	47.610	18.37	0.004
Strain (S)	3	32.963	10.988	4.24	0.053
Treatment (T)	1	422.303	422.303	162.96	0.000
S × T	3	40.342	13.447	5.19	0.034
Error	7	18.140	2.591		
Total	15	561.358			
Treatment group	***n***	**Mean EROD**			
Control	8	7.850			
Treated	8	18.125			
Difference	10.28	(95% confidence interval 8.37–12.18)			
Standard deviation	1.61				

[a]Abbreviations: ANOVA, analysis of variance; DF, degrees of freedom; EROD, ethoxyresorufin O-deethylase; F, the variance ratio and a test statistic such as the Student's t; MS, mean square; *n*, the number of animals in each group; *P*, the probability that differences among means of the observed magnitude could have arisen by chance in the absence of a true effect by each specified source of variation; SS, sums of squares.
[b]The source is the factor causing the variation being analysed.
[c]Note that *P* values are only given to three decimal places.

few people would have queried it even though the genetic variation would not then have been quantifiable.

However, experiments do need to have some justification of sample sizes, and so power and sample size calculations are now usually required when designing clinical and animal studies. An estimate of the required sample size depends on: (1) the size of the effect of scientific or clinical interest; (2) the standard deviation; (3) the desired power of the experiment; (4) the chosen significance level; (5) the sidedness of the test [i.e. whether the treated group mean could be either larger or smaller than the control group mean (a two-sided test) or whether there is a biological reason why the difference could only go in one direction (a one-sided test)]; and (6) the type of statistical test to be used (which depends on the design of the experiment). The use of computer software such as nQuery Advisor [Statistical Solutions (http://www.statsol.ie), Cork, Ireland] is almost essential because the formulae are complex. Free software is available on the web for simpler situations (search for 'statistical power'; websites are somewhat ephemeral). Deciding on the effect size and obtaining a reliable estimate of the standard deviation are the main problems.

Suppose the experiment shown in Box 1 is the first in a series of experiments investigating the effects of various compounds on liver ethoxyresorufin O-deethylase (EROD) enzyme activity using similar protocols. All the subsequent experiments should be of an appropriate size. The standard deviation of 1.6 units is obtained from Table I, Box 1 (the square root of the error 'mean square'). The problem is to decide how large an effect (i.e. the difference between treated and control groups) will be of biological interest in future experiments. It is often useful to think in terms of standard deviations. Table 1 shows sample sizes in terms of the number of standard deviations for the stated power and significance levels. For an effect size of one standard deviation (1.6 units) (i.e. increasing the enzyme activity from 7.8 EROD units in the controls to $7.8 + 1.6 = 9.4$ EROD units in the treated group) the experiment should have 23 mice in each group (Table 1). For an increase of three standard deviations (i.e. 4.8 EROD units) only approximately four mice would be required in each group. Keeping the experiment as outlined in Box 1, with eight mice per group, means that the experiment should be able to detect an effect of just under two standard deviations or ~3.2 EROD units. It is up to the investigator to decide whether this seems to be appropriate.

Table 1. Determination of sample size[a]

Difference in means	Sample size
0.2	527
0.4	133
0.6	60
0.8	34
1.0	23
1.5	11
2.0	7
2.5	5
3.0	4

[a]The calculation of the sample size needed per treatment group for comparing two groups using a two-sided unpaired *t* test is shown as a function of the difference (D) in means between the control and treated groups (in standard deviation units), assuming a 5% significance level, a two-sided test and a 90% power. See the main text for further details.

In other animal studies and clinical trials a small fraction of a standard deviation might be worth detecting, so sample sizes are often large. For example, if the experiment were to be set up to detect a difference of 0.6 standard deviations [i.e. the controls have a mean of ~8 EROD units and the treated groups have a mean of $8 + (1.6 \times 0.6) = 8.96$ EROD units], then the experiment would require 60 subjects in each treatment group. By considering the effect size in this way in relation to available facilities it should be possible to design an appropriately sized experiment.

In cases where there is no prior information a pilot study can give a rough estimate of the probable response and standard deviation. Alternatively, the 'resource equation' method, which is applicable to experiments that will be analysed by an ANOVA, can be used [12]. The resource equation method depends on the law of diminishing returns. Enlarging a small experiment should give good returns, whereas enlarging an already large experiment gives little extra information. A useful guide is that E, the error degrees of freedom in an ANOVA, should be between 10 and 20. In the case of a completely randomized design E is the total number of experimental units minus the number of treatments, and in the case of a randomized block design E is the total number of experimental units minus the number of treatments, minus the number of blocks plus one. In Table II, Box 1 E is 7, which suggests that adding a few more animals might give reasonably good returns. However, this method should not be used too rigidly. There is often a good case for going below an E value of 10 if a large response is expected or going above an E value of 10 if the response is likely to be small.

Statistical analysis and presentation of results

Because this article is focused on experimental design, the methods of statistical analysis will not be considered in detail. However, the method of statistical analysis should always be decided at the time that the experiment is designed, although some modification might be necessary after the data have been collected. The ready availability of statistical software takes the hard work out of most calculations but choosing methods and interpreting output still needs a good understanding of statistics.

Results also need to be clearly presented. If a study involves several separate experiments each should be described in the materials and methods section of the paper and labeled Experiment 1, Experiment 2, etc. Every experimental subject should be accounted for; otherwise there will be the suspicion that unwelcome results have been suppressed. Actual *P* values should be given and means and proportions need an indication of numbers in addition to a standard deviation, standard error or confidence interval. Statistical significance is not the same as biological importance so the magnitude of any effect should always be given.

Guidelines and textbooks

This article has only touched on two techniques, the use of factorial and randomized block designs, for improving

experimental design. Many other designs and techniques are available. Statistical guidelines often give helpful hints and a gentle introduction to statistical methods for planning, designing, executing, analysing and presenting experimental results. These are available for contributors to medical journals [13], for animal experiments [14] and for *in vitro* experiments [15]. Furthermore, there are several statistical textbooks that emphasize experimental design, ranging from an introductory text for those using animals [12] to more advanced and comprehensive texts [7,16–19].

Concluding remarks

The key to designing good experiments is to have clear objectives and to understand and control the main sources of variation. Fixed effects such as the strain, species and sex of animals are either held at a single level or are varied deliberately using a factorial design. Random effects are controlled by choosing uniform material such as disease-free isogenic strains of laboratory rodents, minimizing measurement error and controlling for time and/or space variation using randomized block designs. All scientists performing clinical or animal research should have a basic understanding of experimental design and enough background in statistical methods to be able to consult a statistician effectively for more advanced studies.

References

1 Altman, D.G. (1982) Statistics in medical journals. *Stat. Med.* 1, 59–71
2 Festing, M.F.W. (1994) Reduction of animal use: experimental design and quality of experiments. *Lab. Anim.* 28, 212–221
3 McCance, I. (1995) Assessment of statistical procedures used in papers in the *Australian Veterinary Journal*. *Aust. Vet. J.* 72, 322–328
4 Festing, M.F.W. and Lovell, D.P. (1996) Reducing the use of laboratory-animals in toxicological research and testing by better experimental-design. *J. Royal Statistical Society Series B-Methodological* 58, 127–140
5 Festing, M.F.W. (1995) Reduction of animal use and experimetnal design. In *World Congress on Alternatives and Animal Use in the Life Science: Education, Research, Testing* (Goldberg, A.M. and van Zutphen, L.F.M., eds) pp. 43–49, Mary Ann Liebert
6 Roberts, I. *et al.* (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
7 Cox, D.R. and Reid, N. (2000) *The Theory of the Design of Experiments*, Chapman and Hall/CRC Press
8 Festing, M.F.W. (1999) Warning: the use of genetically heterogeneous mice may seriously damage your research. *Neurobiol. Aging* 20, 237–244
9 Festing, M.F.W. (1997) Fat rats and carcinogen screening. *Nature* 388, 321–322
10 Gaines Das, R.E. (2002) Role of ancillary variables in the design, analysis and interpretation of animal experiments. *ILAR J.* 43, 214–222
11 Festing, M.F.W. *et al.* (2001) Strain differences in haematological response to chloramphenicol succinate in mice: implications for toxicological research. *Food Chem. Toxicol.* 39, 375–383
12 Festing, M.F.W. *et al.* (2002) *The Design of Animal Experiments*, Laboratory Animals
13 Altman, D.G. *et al.* (1989) Statistical guidelines for contributors to medical journals. In *Statistics with Confidence* (Altman, D.G. *et al.*, eds), pp. 171–190, British Medical Journal
14 Festing, M.F.W. and Altman, D.G. (2002) Guidelines for the design and statistical analysis of experiments using laboratory animals. *ILAR J.* 43, 233–243
15 Festing, M.F.W. (2001) Guidelines for the design and statistical analysis of experiments in papers submitted to ATLA. *Altern. Lab. Anim.* 29, 427–446
16 Mead, R. *et al.* (1993) *Statistical Methods in Agriculture and Experimental Biology*, Chapman and Hall
17 Cox, D.R. (1958) *Planning Experiments*, John Wiley and Sons
18 Montgomery, D.C. (1997) *Design and Analysis of Experiments*, Wiley
19 Altman, D.G. (1991) *Practical Statistics for Medical Research*, Chapman & Hall