Laboratory Animals Handbook No. 14

2nd Edition

The Design of Animal Experiments

Reducing the use of animals in research through better experimental design

(Revised and Updated Edition)

Michael F.W. Festing
Philip Overend
Mario Cortina Borja
Manuel Berdoy

laboratory animals limited

SAGE was founded in 1965 by Sara Miller McCune to support the dissemination of usable knowledge by publishing innovative and high-quality research and teaching content. Today, we publish over 900 journals, including those of more than 400 learned societies, more than 800 new books per year, and a growing range of library products including archives, data, case studies, reports, and video. SAGE remains majority-owned by our founder, and after Sara's lifetime will become owned by a charitable trust that secures our continued independence.

Los Angeles | London | New Delhi | Singapore | Washington DC | Melbourne

Laboratory Animals Handbook No. 14

2nd Edition

The Design of Animal Experiments

Reducing the use of animals in research through better experimental design

(Revised and Updated Edition)

Michael F.W. Festing
Philip Overend
Mario Cortina Borja
Manuel Berdoy

SAGE

Los Angeles | London | New Delhi
Singapore | Washington DC | Melbourne

Los Angeles | London | New Delhi
Singapore | Washington DC | Melbourne

SAGE Publications Ltd
1 Oliver's Yard
55 City Road
London EC1Y 1SP

SAGE Publications Inc.
2455 Teller Road
Thousand Oaks, California 91320

SAGE Publications India Pvt Ltd
B 1/I 1 Mohan Cooperative Industrial Area
Mathura Road
New Delhi 110 044

SAGE Publications Asia-Pacific Pte Ltd
3 Church Street
#10-04 Samsung Hub
Singapore 049483

Typeset by: C&M Digitals (P) Ltd, Chennai, India
Printed in Great Britain by Henry Ling Limited at
the Dorset Press, Dorchester DT1 1HD

Second edition first published 2016

First edition published 2002 (reprinted 2005, 2005, 2006, 2007, 2008, 2010 (twice), 2011, 2013 (four times), 2014 (three times), 2015

British Library Cataloguing in Publication data

A catalogue record for this book is available from
the British Library

ISBN 978-1-4739-7463-0 (pbk)

Contents

Preface to the second edition

The first paper demonstrating the scope for improving the design of animal experiments was published more than 20 years ago and the first edition of this book was published largely in response to that finding. Animal welfare organisations such as the Fund for the Replacement of Animals in Medical Experiments (FRAME) and the Universities Federation for Animal Welfare (UFAW) as well as the UK Home Office have recognised that poor design and inadequate statistical analysis are both an animal welfare and a scientific issue. The funding organisations, which also would have been in a position to demand progress, have been slow to realise the need for improvement. They have largely relied on peer review of submitted papers. But this only works if the referees have adequate training, which clearly has not always been the case. Indeed, the publication of novel meta-analysis and systematic reviews of animal experiments in the early 21st century has highlighted many additional defects in animal research. These included inefficient experimental design, incorrect statistical analysis and poor reporting of important information on the experiment, sometimes all at once.

Science depends on the assumption that experiments give results which can be repeated. The discovery in recent years that a large proportion of papers using laboratory animals have produced false-positive results has therefore been a shock. At the time of writing, it has been estimated that the cost of these failures amounts to about US$28 billion per year in the USA alone. This is clearly ethically and economically unacceptable. The funding bodies, as well as newer organisations (e.g. NC3Rs), have however now taken note and it seems likely that there will be much more emphasis in the next few years on ensuring that all scientists using laboratory animals are adequately trained in experimental design and statistics.

In writing this edition, we have kept an eye on, and indeed have sometimes participated in, the changing landscape in education and training worldwide, and in Europe specifically. We have expanded some aspects to cover important concepts and learning outcomes in more detail and reorganised the structure of the book with an additional six short sections. We have also done all the statistical analyses presented in the book using R-Commander (Rcmdr), a free statistical menu-driven front end to the 'R' statistical programming language.

In short, the evidence suggests that there is still a long way to go but that there is hope on the horizon. We hope that this 2nd edition of the book will continue to contribute to this journey.

Michael F. W. Festing, Philip Overend,
Mario Cortina Borja and Manuel Berdoy
March 2016

Preface to the first edition

It is universally accepted that persons who aspire to use animals for experimental purposes should receive proper training. At the heart of this training is the application of Russell and Burch's 3Rs, i.e. Replacement, Reduction and Refinement (Russell & Burch, 1959). Of these, it is probable that least emphasis is placed on, and least success has been obtained in, reducing the number of animals that are currently used in animal experiments.

A major factor in bringing about a reduction in animal use is the correct application of experimental design and statistical analysis, both of which are, arguably, poorly taught or understood by undergraduates and postgraduates in the biomedical sciences. This deficiency is a major concern to a number of bodies, not least to the UK Home Office, which mandates courses (Modules 1–5) for the training of persons in preparation for applying for a licence to use animals; to the Institute of Biology (IOB) which is one of the two accrediting bodies for these courses; and to the Fund for the Replacement of Animals in Medical Experiments (FRAME) which has set up a 'Reduction Committee' to recommend a resolution to the problem. Therefore, in order to redress this imbalance, the IOB decided to commission a book to provide, in a clear, concise and simple way, an understanding of the benefits to be derived by investigators in the proper design of their studies. It is expected that, in many cases, this will lead to a reduction in the number of animals needed, but in all cases it would lead to the optimum use of animals that would provide valid results.

It is not intended that this book should emulate the many textbooks that are available which give a detailed description of experimental design and statistical analysis. Rather, its purpose is to act as a teaching aid to impart an understanding of what types of experimental design and statistics should be considered when developing an experimental study protocol. References to textbooks and statistical packages are then given to enable these to be carried out. Undoubtedly, the best course of action when contemplating designing a study is to seek out and employ a statistician knowledgeable in the field of interest. He or she will want to know many of the details discussed in this book. It is hoped that this book will be of interest in all fields of scientific research and that it will achieve its aim of helping to improve studies such that animal use is reduced, whilst still ensuring that the maximum benefits are derived from the study.

Bryan Waynforth
Chairman, Editorial Group

1

Introduction and basic principles

Progress in medical and biological research is heavily dependent on the use of laboratory animals, although replacements for the use of animals are continually being sought. For ethical, legal and indeed economic reasons animal experiments should be designed to use the minimum number of animals needed to achieve the objectives of the study. It is a fundamental principle in science that experiments should give results which are reproducible; but scientific progress can, and is, impeded by two sorts of failure: false-negative and false-positive results. In the context of experimental design and statistics, false-negative results arise from experiments where inter-individual variation is poorly controlled or the number of subjects is too few to be able to detect an effect which would be of clinical or scientific importance. This could lead to the loss of a useful treatment. Conversely, false-positive effects occur when a response is claimed but it is actually the result of a bias or other error in the conduct of the experiment and not due to the treatment. This can be extremely damaging as scientists attempt to repeat the work, or take the results at their face value, and continue with further futile research. It has been estimated that the cost of non-reproducible studies in preclinical research is a staggering US$28 billion per annum in the USA alone (Freedman, Cockburn & Simcoe, 2015).

Purpose of this book

This book aims to cover the broad principles of experimental design for scientists using laboratory animals, in a non-mathematical manner. ‘Experimental design’ is considered here in a wide context, including the choice of animals and the control of biological variation, topics not normally covered in a statistics textbook. For those who already have some knowledge of statistics and experimental design it should help them to design better, more efficient, experiments which can reduce the number of animals needed without reducing scientific output. For others, it should provide background information, which will help them to consult statistical textbooks and professional statisticians/biometricians more effectively.

By striking this balance we also hope that the book can be helpful as a core text in the formal training of researchers under the new education and training landscape

Table 1.1 Learning Outcomes for EU Modules 10 and 11.

Learning outcomes for EU Module 10 and part (ii) of Module 11 from the working document (European Commission, 2014) on the development of a common education and training framework to fulfil the requirements under the Directive 2010/63/EU on the protection of animals used for scientific purposes.

The main sections of the book addressing the learning outcomes are listed within square brackets.

Module 10: Design of procedures and projects (level 1)

This module is a pre-requisite for people who will be designing projects (Function B) but it is also beneficial for scientists who have some involvement in designing the procedures that they carry out (Function A). The module comprises information about experimental design concepts, possible causes and elimination of bias, statistical analysis and information about where expertise can be found to assist with the procedure, design, planning and the interpretation of results.

Learning outcomes: trainees should be able to:

10.1. Describe the concepts of fidelity and discrimination (e.g. as discussed by Russell and Burch and others) [see Chapter 1].
10.2. Explain the concept of variability, its causes and methods of reducing it (uses and limitations of isogenic strains, outbred stocks, genetically modified strains, sourcing, stress and the value of habituation, clinical or subclinical infections, and basic biology) [see Chapter 2].
10.3. Describe possible causes of bias and ways of alleviating it (e.g. formal randomisation, blind trials and possible actions when randomisation and blinding are not possible) [see Chapter 3].
10.4. Identify the experimental unit and recognise issues of non-independence (pseudo-replication) [see Chapters 1, 3, 4 and 12].
10.5. Describe the variables affecting significance, including the meaning of statistical power and *P*-values [see Chapter 11].
10.6. Identify formal ways of determining sample size (power analysis or the resource equation method) [see Chapter 11].
10.7. List the different types of formal experimental designs (e.g. completely randomised, randomised block, repeated measures [within subject], Latin square and factorial experimental designs) [see Chapters 4–8].
10.8. Explain how to access expert help in the design of an experiment and the interpretation of experimental results [this is a local issue and beyond the remit of this book, although see Chapter 12].

Module 11: Design of procedures and projects (level 2)

[Function Specific for Function B] (ii) Good scientific practice

11.3. Describe the principles of a good scientific strategy that are necessary to achieve robust results, including the need for definition of clear and

unambiguous hypotheses, good experimental design, experimental measures and analysis of results. Provide examples of the consequences of failing to implement sound scientific strategy [see Chapters 1 and 12].

11.4. Demonstrate an understanding of the need to take expert advice and use appropriate statistical methods, recognise causes of biological variability, and ensure consistency between experiments [this is the general goal of the book as a whole].

11.5. Discuss the importance of being able to justify on both scientific and ethical grounds, the decision to use live animals, including the choice of models, their origins, estimated numbers and life stages. Describe the scientific, ethical and welfare factors influencing the choice of an appropriate animal or non-animal model [see Chapter 2].

11.6. Describe situations when pilot experiments may be necessary [see Chapters 1, 2, 3, 11 and 12].

11.7. Explain the need to be up to date with developments in laboratory animal science and technology so as to ensure good science and animal welfare [see Chapter 2].

11.8. Explain the importance of rigorous scientific technique and the requirements of assured quality standards such as good laboratory practice (GLP) [see Chapter 3].

11.9. Explain the importance of dissemination of the study results irrespective of the outcome and describe the key issues to be reported when using live animals in research, e.g. ARRIVE guidelines [see Chapters 12 and 13].

worldwide, and in Europe specifically. For example, the book targets, and more than covers, the specific learning outcomes which form part of the training on the design of procedures and projects to fulfil the requirements under Article 23 of Directive 2010/63/EU on the protection of animals used for scientific purposes (European Commission, 2014). The list of learning outcomes and a description of how they map to the contents of this book, are listed in Table 1.1. Relevant material for the book can also be found on the book's website (https://uk.sagepub.com/en-gb/eur/Design_Animal_Experiments_Handbook)

Experimental design is an interdisciplinary subject involving both biology and statistics. All too often scientists using animals in research have too little training in statistics, or any training that they are given comes too early in their career so that they have forgotten most of it by the time they need it. Unfortunately courses in statistics often emphasise methods of statistical analysis of data, rather than the design of the experiments needed to collect the data. Statisticians often have a mathematical background and may not understand the biology of laboratory animals, making communication between the two disciplines difficult. All too often the result is that the scientists who have tried to consult a statistician find it unrewarding so they revert to repeating the type of experiments taught to them by their graduate supervisor, or which they see in the literature. They are often concerned that if they submit papers with unusual experimental designs (say with fewer than six rats per treatment group) these will be rejected by the referees, who themselves often have

little training in statistics. Thus, a vicious circle develops which prevents progress and leads to a waste of animals, money, time and effort. We hope this book will help scientists avoid this.

The need to improve the design and statistical analysis of animal experiments

The principles of good experimental design are relatively simple, but failure to adhere to them is common, and can have serious consequences. Early surveys of published papers showed that many animal experiments were poorly designed and inadequately analysed (Festing, 1994a), and with little use of randomised block and factorial designs (Festing, 1992), even though these provide the most powerful and economical ways of designing experiments. Too many experiments give false-positive results which cannot be reproduced by subsequent experiments. They may also produce many false-negative results, but attempts to repeat these experiments are rare.

The results of a survey of 271 animal experiments performed between January 1999 and March 2005 are shown in Table 1.2 (Kilkenny et al., 2009). The survey was restricted to original research papers involving mice, rats or non-human primates reported from academic institutions in the UK or the USA. There is no comparable information on non-academic institutions.

The survey found that 13% did not correctly identify the *experimental unit*. This is the subject of the experiment. By definition any two experimental units must be capable of receiving different treatments. Often the experimental unit is a single animal; but if there are two animals in a cage and the treatment is given in the diet or water then the animals cannot receive different treatments. In this latter case the cage is the experimental unit and the statistical analysis should be based on

Table 1.2 Some results of a survey of a random sample of 271 published papers involving laboratory animals.

Of the papers studied:

- 87% did not report random allocation of subjects to treatments
- 86% did not report 'blinding' where it seemed to be appropriate
- 100% failed to justify the sample sizes used
- 5% did not clearly state the purpose of the study
- 6% did not indicate how many separate experiments were done
- 13% did not correctly identify the experimental unit
- 26% failed to state the sex of the animals
- 24% reported neither age nor weight of animals
- 4% did not mention the number of animals used
- 35% reported numbers used but these differed in the materials and methods and the results sections

See Kilkenny et al. (2009).

the mean of whatever is measured in the two animals. If it is mistakenly assumed that the animal is the experimental unit, then a false-positive result may occur due to 'pseudo-replication', or overestimation of '*n*', the number of subjects in the experiment. An experiment in which all the controls were put into one cage and the treated animals into another cage would be invalid because there would be an *n* of only two. A similar situation would arise if controls were placed in, say, cages 1–10 and treated animals in cages 11–20. In such a case, cages 1 and 2 could not receive different treatments, so they would not be the experimental units. The experiment would again have an *n* of two. The problem with keeping each treatment in a numbered group is that the group would probably be housed, treated and measured in numerical order and there may be environmental and time effects that would change over time and space. This could lead to bias and false-positive results. Treatments need to be assigned to experimental units at random in such a way as to ensure that they are intermingled throughout an experiment. Suitable methods are described in Chapter 3.

The survey found that 87% of papers failed to report randomisation of treatments to the experimental units (the subjects of the experiments being often, but not always, animals), and 86% failed to report 'blinding' of the investigator when measuring the outcomes. For reasons given above, randomisation and blinding are techniques which are fundamental to good experimental design. Failure to use them can lead to bias and false-positive results.

None of the studies justified the sample sizes which were used. Use of too few subjects risks a false-negative outcome, but if too many animals are used there is a waste of animals and other scientific resources. Failure to report relevant variables such as the age, weight and sex of the animals or husbandry details makes it difficult for others to repeat the work. An attempt to reproduce the results of 53 landmark papers in cancer research was successful in only six (11%) cases. The authors (Begley & Ellis, 2012) concluded that 'even knowing the limitations of preclinical research, this was a shocking result'.

Systematic reviews and meta-analysis have been used for many years in clinical trials, but only in the last 10–15 years have the techniques been used for preclinical animal studies. Their use in animal research is starting to reveal many defects in published papers. An early example was a meta-analysis of 44 papers studying fluid resuscitation in animals following removal of a substantial fraction of their blood (Perel et al., 2007). Only two of these papers described how the animals had been allocated (i.e. with no mention of randomisation), and none had sufficient power to reliably detect a halving of the risk of death, a response which would probably be of clinical importance. There was substantial scope for bias and there was heterogeneity of the results due to the method of bleeding. The authors queried whether these animal experiments made any contribution to human medicine.

Another systematic review of six interventions where the outcome was well established in humans was used to investigate whether the animal studies predicted the human outcome (Perel et al., 2007). There was agreement in only three cases. Some lack of agreement was probably due to publication bias as it is often difficult to get papers with negative results accepted for publication. In other cases Perel et al. (2007)

claimed that the papers were of such poor quality that it was unclear whether the lack of agreement with the human results was because the model was inappropriate or because of the poor quality of the research.

Poor design, rather than unsuitable animal models, has been shown to be the expensive culprit in some drug screening studies: for example, the standard model for screening drugs for the treatment of amyotrophic lateral sclerosis (ALS) or motor neuron disease is a strain of genetically modified (GM) mice carrying 23 copies of the human $SOD1^{G93A}$ gene. More than 50 papers have described therapeutic agents which extend lifespans in these mice, but only one (riluzole) has any effect in humans. However, a detailed review of this model (Scott et al., 2008) showed that there are a number of confounding factors such as gender, copy number, litter and censoring which need to be taken into account when using the model. The experimental protocol was redesigned using a power analysis to determine a suitable sample size (the 'power' of an experiment is the probability of detecting an effect larger than a predetermined size, see Chapter 11). All of the 50 drugs, which had previously been shown to have an effect in mice, together with another 20 drugs, were rescreened. This process took five years and used 18,000 mice. The investigators had expected to be able to reproduce the results for the 50 drugs which had previously been tested and had found positive effects, but found that none of them prolonged lifespans in these mice. They concluded that 'the majority of published effects are most likely measurements of noise in the distribution of survival means as opposed to actual drug effects'. In short, they represented false-positives caused by poor experimental design.

The origins of the randomised controlled experiment

The basic principles of modern experimental design, and the statistical tools needed to analyse the resulting data, are not new. They were largely developed by R. A. Fisher and colleagues at the Rothamsted agricultural experimental station in the 1920s. The aim was to study crop husbandry in order to increase crop yield. The methods have since been adapted for use in virtually all scientific disciplines but some of the terms continue to reflect their agricultural origins (e.g. treatment, block and split plot).

Basic principles

If two plots in a field could be found which were in every way identical, then a fertilizer could be applied to one of the plots and the other could be kept as a control. Any differences in yield could then be attributed to the effect of the fertiliser. However, no two plots are ever identical. But if enough control and fertilised plots were to be used then the averages of the two groups would give a good indication of any differences, depending on the uniformity of the field and the number of plots (*sample size*) in each group. *Treatments* need to be applied to the plots in such a way as to avoid any bias as a result of more fertile plots being assigned to one group. As it is difficult to forecast the fertility of individual plots, the treatments were applied to the plots *at random*.

The statistical analysis of the results

The results of such an experiment needed to be statistically analysed to take account of the *intrinsic variation* among plots in order to try to ensure that the observed differences (the *effect size*) were not just due to chance. If there were only two groups, then the data could be analysed using a '*t*-test' developed by W.S. Gosset, a statistician using the pseudonym 'Student' (hence 'Student's *t*-test') who worked for the Guinness brewery in Dublin. This test results in a '*t*-value' which can be interpreted as a signal-to-noise ratio which is converted to a '*P*-value', namely the probability that a difference as great as, or greater than, that which is observed could have arisen simply by chance, rather than as a result of the treatment.

As early as the 1920s Fisher recognised that people need to make decisions based on the outcome of an experiment and suggested that if the *P*-value in a comparison of group means is less than 1/20 (0.05) it is probably safe to assume that the differences are real or *statistically significant*. He went on to develop a generalisation of the *t*-test, the analysis of variance (ANOVA), which could be used with any number of treatment groups, rather than just two. It gives an overall estimate of the *P*-value (see Chapter 4).

Factorial experiments

If a new variety of, say, wheat was developed, it might be important to know how it responded to a fertiliser (treatment). So instead of having, for example, 12 plots of the old variety (control) and 12 plots of the new variety of wheat, a better experiment would have four groups: six of the plots of each variety (2 groups) would have fertiliser and six of both varieties (2 more groups) would have no fertiliser. In this way the two factors, variety of wheat and the effect of the fertiliser, could be tested in a single experiment without increasing the total number of plots. This design would have the added advantage of showing:

1. The effect of the fertiliser, averaged across both varieties.
2. The difference between the varieties.
3. Whether the two varieties averaged across fertilisers would respond equally to the fertiliser.

This *factorial design* allows the testing of two or more factors without having to increase the overall sample size (still using 12 plots per variety). These factorial experiments, which are of great value in animal research, are discussed in detail in Chapter 6. In many cases both males and females can be included in an experiment without increasing the total number of animals which are used.

Randomised block designs

Sometimes the experimental material is quite heterogeneous. This reduces the ability of the experiment to detect any effect. In a randomised block design the material is split up into a number of small groups or 'blocks' which are matched in some way.

Typically each block consists of a single plot for each treatment. Accurate treatment comparisons can then be made separately for each block and these can be averaged across all the blocks. These randomised block designs are usually more powerful (i.e. better able to detect an effect) than the completely randomised experiments because the partitioning of variation into blocks allows us to bring out the effect of the treatment that we are interested in. Blocks can even be placed in separate fields in order to test the generality of the response in different environments. Most agricultural experiments use randomised block designs. Because of these valuable properties, a strong case can be made for their more widespread use in experiments involving laboratory animals.

Comparison of clinical trials with laboratory animal experiments

Humans are very variable and even small effects are usually of clinical importance, so clinical trials need to be large (100s to 1000s of subjects), particularly if there is a binary (cured/not cured) rather than a measurement outcome.

Participants are gathered over a period of time and are started on the trial as they are recruited, so the trial has a staggered start. Allocation bias is a potential problem so participants are assigned to treatments strictly at random with the evaluation of results being double-blinded (whereby neither the patient nor the investigator knows the treatment group to which a patient belongs until the end of the experiment). Relatively simple, completely randomised designs are usually used with sample size being determined by a power analysis (see Chapter 11).

By contrast, samples of laboratory animals can be extremely uniform. As the statistical power of an experiment is largely determined by the variation among the experimental subjects (or 'units'), even small experiments (usually of fewer than 50 subjects) with many treatment groups can be powerful. There are many factors (such as gender, age and additional treatments) which may influence the outcome, so factorial designs, say involving both sexes, can be used very effectively. With laboratory animals the main sources of variability are the physical environment, time (due to biological rhythms), and human factors associated with the interactions between the animals and humans when the outcomes are measured. Randomised block designs (see Chapter 7) can be used to control such variation and thereby increase power. An internal indication of repeatability is given if blocks are set up over a period of time such as days or weeks. Such designs are also much less likely to produce biased results due to incorrect or unlucky randomisation.

There is also scope for the experimental unit to be something other than an animal. It might be a dish of cells from an animal, an animal used for a period of time in a crossover design, or a cage of animals in which all the animals in the cage receive the same treatment. As a result, animal experiments can combine small numbers of subjects with complex experimental designs. In clinical trials a power analysis is usually used to determine sample size. This method can also be used

in relatively simple animal experiments, but it is not so useful for complex experiments. A different method, the 'resource equation' may sometimes be more useful (see Chapter 11).

Two additional types of experiment

Pilot experiments

The aim of a pilot experiment is to test the logistics of, and gain preliminary information for, a proposed future experiment. Is the experiment feasible? What are the bottlenecks in following the protocol? Is further training needed? Are the proposed dose levels (or equivalent) appropriate? What are the major sources of variability?

These are usually small experiments of no more than 5–20 experimental units, although there is no formal way of determining the appropriate sample size for this type of experiment. They are particularly important when an investigator is starting a new project involving techniques and protocols which have not previously been used by him/her. Failure to use pilot experiments can lead to a waste of resources and/or biased results due to unforeseen complications in an experiment. The results of pilot studies should not normally be published until they have been confirmed by further experimentation.

Exploratory experiments

Sometimes an investigator does an experiment 'just to see what happens', without having any formal pre-specified hypothesis. Often this will involve measuring many outcomes. If these are then subjected to a statistical analysis some may be found to be statistically significant. If a hypothesis is formulated to account for any observed differences then, by definition, this post hoc hypothesis will fit the data, leading to a potential false-positive. So any hypothesis derived from an exploratory experiment *must* be tested in a confirmatory experiment before claiming that the effects are real.

Legal requirements and the 3Rs

In the European Union (EU) all vertebrates and cephalopods are protected animals under Directive 2010/63/EU. In the UK this is implemented through an amendment to the 1986 Animals (Scientific Procedures) Act. Similar local legislation has been implemented in each EU member state. A scientist planning to use animals must have the necessary qualifications and training to allow them to carry out such work. The legislation relies heavily on the 3Rs (Replacement, Refinement and Reduction), introduced in 1959 by Russell and Burch in their book, *The principles of humane experimental technique* (Russell and Burch, 1959).

In the USA research workers can use the *Guidelines for the care and use of laboratory animals* as guidance on legal requirements. This is published by the Institute of Laboratory Animal Research (ILAR) (Committee for the Update of the Guide for the Care and Use of Laboratory Animals, 2011).

Replacement

The EU Directive requires that scientists should first consider whether the objectives of their experiment could be achieved using non-animal alternative methods (Replacement). Although this legal requirement may not exist in other countries, it is obviously ethically and economically important to consider alternatives. Could the scientific knowledge be obtained from research done in part or altogether using tissue culture or even mathematical modelling? If so, then those methods should be chosen in preference to the use of vertebrates. In most cases they will also be cheaper. In practice these non-animal methods are often complementary to any animal studies, and form part of an overall strategy. Considerable progress has been made in replacing the use of animals by *in vitro* chemical or immunological assays for biologicals such as insulin. Strenuous efforts are also being made to develop replacement alternatives in toxicity testing. In many cases chemicals and potential drugs which are likely to be excessively toxic to humans can be identified using these methods, so they can be rejected without the need for animal testing.

Refinement

If it is impossible to use a replacement alternative, the next step is to consider 'Refinement'. The aim is to minimise pain, suffering, distress or lasting harm to each animal. The animals need to be cared for by trained staff with ready access to a veterinary surgeon. They should be free of clinical and subclinical diseases. Animals should be handled regularly and sympathetically so that they do not feel fear when entered into an experiment. When used in an experiment every care should be taken to minimise pain and suffering. Where substances are administered to animals, procedures which cause the least possible pain should be used (Morton, 2000). Surgery should be performed with appropriate anaesthesia and analgesia, with good post-operative care. Some experiments are expected to result in substantial pain or the death of some of the animals. In such cases humane endpoints should be used (Stokes, 2000) (see also other papers in the same issue of the ILAR Journal). Animals which develop tumours should be painlessly killed before the tumours become too large. Often it is possible to predict with some degree of confidence that an animal is going to die, from its appearance and behaviour. Such animals should be painlessly killed rather than being left to suffer and die in pain. Finally, when the experiment is over the animals should be killed using an appropriate painless method such as one of the humane methods recommended by the UK Home Office (https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/229022/0193.pdf), or the ILAR (https://grants.nih.gov/grants/olaw/Guide-for-the-Care-and-Use-of-Laboratory-Animals.pdf).

Moreover, good welfare increases uniformity among experimental subjects leading to the need for smaller sample sizes. Hence good animal welfare (see Chapter 2) is essential for good science.

Reduction

This is the main topic of this book. It is concerned with minimising the number of animals used in each experiment, consistent with achieving the desired scientific objectives. It involves having a clear understanding of the objectives of the study, good control of variation, efficient experimental design, a justified sample size, and extraction of all the useful information by appropriate statistical analysis. The results then need careful interpretation.

Common errors in the design and analysis of animal experiments

The surveys mentioned above on the statistical quality of published papers, and the experience of the authors, point to some common errors.

Experiments done on an ad hoc basis

Experiments should be pre-planned, but sometimes additional groups or animals are added during the course of the experiment. Consequently, observed treatment differences may result from time factors such as circadian rhythms or failures in matching groups of subjects. It is not uncommon to see figures with footnotes such as 'n = 3–6' without further explanation being given. Well-designed and correctly randomised experiments normally have equal group sizes with an explanation of how the sample sizes were estimated. Any deviations from this should be explained.

Bias due to faulty randomisation

Bias arises if there are systematic differences among the treated groups which are not due to the effects of the treatments. It can arise as a result of faulty randomisation. In clinical trials the patients will differ in the severity of their symptoms, and randomisation is used to ensure that treatment groups are approximately equal in the levels of severity. If healthy normal animals are used, they will usually be so similar as to be indistinguishable. The purpose of randomisation in this case is to ensure that the environment in the animal house during the course of the experiment and during the measurement of the outcomes is randomised, so that environmental and time variations do not significantly affect groups differently. If surgically or drug-prepared animals are used then randomisation will also ensure that groups are balanced in terms of severity. Experimental subjects receiving different treatments should be intermingled. Once the treatments have been given the subjects should only be identified by a number so that when outcomes are measured the investigators do not know which treatment each individual received.

Even correct randomisation of small experiments may lead to bias if a completely randomised design is used. Animals in different treatment groups may by chance not be equally distributed among the animal house environments. A randomised

block design is usually both more reliable and more powerful than a completely randomised design in situations where the experimental subjects are relatively uniform but the environment is heterogeneous.

Over-statement of *n* due to pseudo-replication

Incorrect identification of the experimental unit can lead to pseudo-replication in which there is an overestimate of the sample size denoted by *n*, leading to differences which are more significant than is fully justified. For example, if there are two animals per cage and the treatment is given in the diet or water, then the two animals cannot receive different treatments; and therefore the cage of two animals, not the individual animal, is the experimental unit. Pseudo-replication can be seen in an extreme form in some *in vitro* studies. Suppose, for instance, there are two dishes of cells, one of which has received a drug treatment while the other was kept as a control. If the diameter of 20 cells is measured in each dish then *n* is 2, not 20. This is because the experimental unit is the dish, not the individual cells, as the cells cannot have received different treatments.

Experiments which are too small

Small experiments or ones where variation is poorly controlled may be unable to detect biologically important results, leading to a false-negative conclusion. However, such results are rarely followed up, so most false-negative results tend to remain undiscovered.

False-positive results due to the choice of significance level

A 5% significance level implies that on average 1/20 tests will show a false-positive significant difference between groups as a result of sampling variation.

The effects of confounding variables leading to false-positive or false-negative results

Inter-individual variation is usually well controlled in laboratory animals and, as a result, time and space effects are more visible. Physiological and behavioural results will be affected by circadian and possibly other rhythms, so measurements made in the mornings and afternoons may differ. Barometric pressure can affect the activity of animals (Sprott, 1967), and this will vary over various time periods. Different locations in the animal house can have different temperature, humidity and light levels. Even the gender and characteristics of the investigators and animal house staff can affect the animals. Investigators doing surgical procedures will themselves vary over a period of time as they become more skilled or more tired. If these factors are not taken into account they can lead to false-positive results due to bias or to false-negative results caused by increased heterogeneity among the experimental units. Many of these sources of variability can be controlled by using randomised block designs (see Chapter 7).

Interactions involving factors such as gender, strain, environment and infection

Different strains and genders of mice, rats and other species can differ in their response to an experimental treatment. If an experiment is done using a sensitive strain or gender, then the results may not be repeatable in animals of less sensitive strains or genders. One of the dangers of using a single genetically undefined outbred stock is that different samples of animals may differ in their responses because they are genetically different. Subclinical infection can also alter the responses of laboratory animals, so results obtained with infected animals may be non-repeatable in healthy ones, or vice versa. Changes in the microbiome may also affect the results. Some protection from this type of failure can be obtained by the use of factorial experimental designs using both genders and/or more than one strain. Pathogenic infections should be eliminated from the animal house.

Lack of repeatability due to the use of the wrong animal

Outbred stocks like Sprague–Dawley rats and ICR mice are genetically undefined. It is not possible to answer the simple question 'what is a Sprague–Dawley rat?' The problem is that these animals vary over both long and short periods of time, so it may be difficult to repeat an experiment because although the animals have the same name or designation they are genetically different. Genetically defined, stable, and identifiable inbred and F1 hybrid strains of mice and rats have been widely used by geneticists and immunologists for several decades, resulting in many important discoveries, some of which have been recognised by the award of more than 22 Nobel prizes (Festing and Fisher, 2000). It is irrational to reject the use of inbred strains when their value in research has been so clearly demonstrated. These strains are discussed in Chapter 2.

Errors in the statistical analysis and interpretation of experimental results

There is enormous scope for making errors in the statistical analysis and interpretation of data (Festing, 1992, 1994b; Ioannidis et al., 2014; Kilkenny et al., 2009; McCance, 1995). Mistakes include:

1. No statistical evaluation of the experimental results. This can lead to false-positive results when differences appear to be large but are in fact not statistically significant. This is common when percentages based on small numbers are being compared.
2. Choice of the wrong statistical method such as trying to analyse an experiment with more than two means using Student's *t*-test instead of the ANOVA.
3. Severe non-normality of the residuals and/or heterogeneity of variances not taken into account when using a parametric test such as the ANOVA.
4. Elimination of outliers without explanation. They may not be wrong just because they are outliers.

5. Choice of non-parametric methods when parametric methods are available. This can lead to false-negative results because non-parametric tests often lack power.
6. Assuming a causal relationship between two variables when using correlation.
7. Misinterpretation of a non-significant *P*-value >0.05 as evidence that there is no effect. Lack of statistical significance may just be a consequence of the sample size being too small to detect an effect.
8. Sometimes a statistical analysis is simply not appropriate. For example, in one study the liver weight of rats was compared with untreated controls, using a *t*-test, immediately after half the liver had been removed. It is trivial to test the hypothesis that half a liver weighs less than a whole liver.

2

Choice of animals and their husbandry

Introduction

Investigators should use high quality, healthy animals which are appropriate models for the question investigated. They should be free of clinical and subclinical infections, fed a nutritionally adequate diet and housed in an enriched environment. Social animals such as mice and rats should be group-housed because single housing is stressful (this is a legal requirement under the EU Directive 210/63/EU). Ideally, mice and rats, should be 'genetically defined' (i.e. isogenic, mutant or genetically modified) rather than be undefined, because samples of genetically undefined 'outbred' mice and rats may be genetically different even if they have the same name, making it more difficult to repeat an experiment. It makes no ethical, scientific or economic sense to use poor quality animals or those housed in an inadequate environment if that reduces the scientific quality of the work.

Choice of species, animals as models and the 'high fidelity fallacy'

The choice of species will often be dictated by the availability and relevance of a model for the condition of interest. A 'model' is a representation of the target of interest, such as a human. The model has to be *like* the target in some ways, but it is always *unlike* the target in other ways. For example, a map is a model of a country or a city; it is like the target only in the relative positions of topological features, but is unlike the target in virtually every other respect, and can only be used for the specific purpose for which it was designed.

A mouse might be a model of a human with respect to its response to a toxic compound, but it is unlike a human in that it is small. This is an advantage because mice are cheap and relatively easy to maintain. Sometimes a model is more useful if it differs from the target in a specific way. For example, inbred strains are widely used in research because, unlike humans, many genetically identical animals can be produced. These are, in a sense, like immortal clones of a single individual. Usually

they are less variable than outbred stocks so sample sizes can be decreased; or if not, the resulting experiment will be more powerful (i.e. there will be a greater chance that a small treatment effect will be detected).

According to Russell and Burch (1959), there are two dimensions that need to be considered when choosing a research model. The 'fidelity' of the model is the extent to which the model resembles the target in every respect. So a high fidelity model of humans might be a non-human primate and a low fidelity model could be a dish of cultured cells. The second dimension is the ability of the model to discriminate between interventions. It may be better to use a low fidelity model if it has good ability to discriminate between interventions of interest. For example when Russell and Burch published their book, human pregnancy was tested using a mouse or rabbit uterus assay or a frog ovulation test. As these are live animals, this was a relatively high fidelity model of the reproductive state of a woman. However, a home pregnancy testing kit has replaced the use of live animals. It is a low fidelity model (very unlike a human), but has excellent ability to detect pregnancy. Russell and Burch showed that it is a fallacy to claim that high fidelity models should always be preferred. Detailed information on the characteristics of individual species and more general information on all aspects of laboratory animal science is given in *The UFAW Handbook on The Care and Management of Laboratory Animals* (Hubrecht and Kirkwood, 2010), *The Handbook of Laboratory Animal Science* (Hau and Shapiro, 2011; Hau and Van Hoosier, 2002) and the *COST Manual of Laboratory Animal Care and Use* (Howard et al., 2011).

Freedom from disease

Before about 1950 all species of laboratory animals carried a wide range of viral, bacterial and metazoan pathogens. These caused clinical or subclinical disease which disrupted research by decreasing lifespan and increasing variability, so that more animals were needed, and in some cases by interacting with the experimental treatments to give spurious results. Thus, the average lifespan of a laboratory rat in the 1930s was only about 12 months compared with 2–3 years today. A major cause was infection with *Mycoplasma pulmonis* and other microorganisms which caused chronic respiratory disease. At one stage, it was thought that vitamin A deficiency caused lung damage because rats which were vitamin A deficient had more serious lung lesions than non-deficient rats. However, it was subsequently found that the physiological stress of vitamin A deficiency was increasing the lung lesions normally associated with chronic respiratory disease (Lindsey et al.,1971).

So-called 'conventional' animals with clinical or subclinical disease are usually more variable than healthy 'specific pathogen free (SPF)' ones, so more are needed in an experiment to achieve a given level of statistical precision. In one study (Gartner, 1990), the standard deviation of the kidney weight of conventional rats suffering from chronic respiratory disease was 43.3 (arbitrary) units, whereas it was only 18.6 units in healthy SPF rats. Using a power analysis (see Chapter 11) with the assumptions that (1) a 10 unit difference between the treated and control groups

would be the minimum difference likely to be of scientific interest; and (2) an 80% chance of detecting it would be acceptable, an experiment using 55 healthy rats per group or 295 unhealthy rats per group would be needed. In addition, there would be no assurance that any observed response would be qualitatively the same as in healthy SPF animals.

Techniques to develop SPF animals were developed in the late 1950s as a means of eliminating pathogenic microorganisms. Fetuses in utero are usually microbiologically sterile. These were removed from their mother just prior to parturition, under sterile conditions, and hand-reared in an isolator using sterile milk and diet. Although these germ-free or 'gnotobiotic' animals survived and bred, they were abnormal in many respects, and were difficult and expensive to maintain. Further research showed that animals infected with a suitable cocktail of non-pathogenic gut bacteria resembled normal conventional animals which were free of disease. Colonies of these animals were later established in 'barrier' animal houses in which all supplies of diet, bedding and equipment were sterilised to prevent reinfection with pathogens.

Although these SPF animals sometimes pick up unwanted microorganisms, often from the staff, they are essentially like conventional animals which are free of all important pathogenic microorganisms.

SPF mice, rats, rabbits, and some other species are now universally available from both commercial breeders and academic institutions in developed countries. These should have been regularly screened and found to be free of certain defined viruses, bacteria and parasites, details of which will normally be provided by the breeder on request. However, not all animal colonies are free from infection, and great care should be taken that animals being imported to an animal house should be quarantined under veterinary supervision even if the parent colony is nominally free of pathogens.

The genetic definition of mice and rats

There is a bewildering range of 'genetic' types of these two species. They fall into three main types: (1) outbred stocks, (2) inbred (isogenic) strains, and (3) genetically modified and mutant strains.

Outbred stocks

Outbred stocks are breeding colonies of genetically heterogeneous animals, usually maintained by some form of random or rotational mating, often avoiding the mating of closely related individuals (the term 'stock' is used for outbred colonies while the term 'strain' is used for inbred ones). These are still widely used in biomedical research. According to one estimate between January 2002 and July 2007 about 33% of all mouse studies and 85% of all rat studies used such stocks (Chia et al., 2005) Many outbred mice and rats are also used in the pharmaceutical industry in the early stages of drug development. Such work is rarely suitable for publication. However, it is not clear whether scientists using these animals fully understand the implications of using these so-called 'genetically undefined' animals in their research.

Nomenclature of outbred stocks

Rules for the nomenclature of outbred stocks can be found at http://www.informatics.jax.org/mgihome/nomen/strains.shtml (accessed 19 January 2016). Stocks are identified by a designation such as 'CD-1' or 'ICR'. The stock name should be preceded by a laboratory code followed by a colon. For example, Tac:ICR is a mouse stock designated 'ICR' (Institute of Cancer Research) now maintained by Taconic Farms. Crl:CD(SD) is an outbred stock of rats designated 'CD' and maintained by Charles River Inc (Crl). It is of Sprague–Dawley origin. Outbred stocks are genetically undefined, i.e. there are no genetic markers which can be used to identify a stock. There is no assurance that stocks with the same name but from a different supplier will be the same or that stocks with different names are genetically distinct. It is not even possible to distinguish genetically between Wistar and Sprague–Dawley, the two most widely used stocks of rats. Rats from a colony designated 'Sprague–Dawley' will probably differ from one designated 'Wistar', but rats from another colony of Sprague–Dawley will probably also differ genetically from the first one.

Origin of mouse and rat outbred stocks

A list of the origin of 38 named mouse stocks was compiled in 2005 (Chia et al., 2005). Many of these are maintained by commercial breeders, and are used for research which does not require any specific characteristics. However, a few of them have special characteristics, sometimes as a result of selective breeding. For example, the ABH and ABL mouse stocks are the result of selective breeding for immune response to sheep red blood cells. The LS and SS mouse stocks were selectively bred for long and short sleep time under alcohol anaesthetic and the SENCAR mouse stock was developed for sensitivity to skin carcinogens. In some cases these selected outbred stocks were subsequently inbred in order to fix their characteristics. There are also a number of genetically heterogeneous mouse stocks such as HSCC, HSCDHG, HSIBG and HSNPT which were developed from crosses between inbred strains. These are used for genetic investigations such as genome wide association (GWA) studies. There are no recent comprehensive lists of rat outbred stocks, but some information is available from commercial breeders.

Genetics of outbred mice

Little was known about the genetics of outbred stocks of laboratory mice and rats until DNA-based genetic markers became available. In 1974, before DNA markers were available, a study of commercially available outbred mice using highly inherited phenotypic markers (bone shape) found many anomalies in such stocks. Stocks with the same name were sometimes phenotypically quite different and stocks with different names were sometimes similar to each other (Festing, 1974). Genetic drift was also found in a mouse stock, which was probably due to genetic contamination from the foster mothers when a colony was rederived after becoming infected (Papaioannou and Festing, 1980).

There have now been two studies investigating the suitability of commercially bred outbred stocks for GWA studies. These associate phenotypes of biomedical interest with gene loci segregating within the stock. They provide a powerful research tool for finding genes of biomedical interest. The ideal population for GWA studies would have a high level of genetic variation, low levels of linkage disequilibrium, few rare alleles, and no substructure within the colony.

In the first study CD-1 mice maintained by a single company at three locations (Aldinger et al., 2009) were found to show patterns of linkage disequilibrium and heterogeneity similar to wild-caught mice. The three populations were genetically distinct, somewhat similar to related human populations, and the differences were consistent with founder effects. Highly significant phenotypic differences in conditioned freezing to an audio tone were found among the three populations. However, it was unclear whether these phenotypic differences were a result of genetic differences or because the mice came from different environments. The ancestry of laboratory mice is known to include a number of subspecies. By including three inbred wild-derived strains representing the three subspecies it could be shown that the genome of CD-1 had 75% *M. m. domesticus*, 19% *M. m. musculus* and 6% *M. castaneus* ancestry.

A second investigation involving 66 outbred mouse colonies (Yalcin et al., 2010) found that genetic variation among all the colonies was surprisingly high, being about 10 times higher than that found in humans, with 45% of the genetic variation being due to differences between colonies. Possibly this is because laboratory mice are derived from three subspecies as noted above.

Over 95% of genetic sequences found in the outbred stocks were also present in inbred mouse strains. However, 4/66 colonies were almost entirely inbred, and a further five were somewhat inbred. These would not be suitable for GWA studies. In most cases gene flow appeared to have occurred between colonies (i.e. genetic contamination). This was not surprising as in the past breeders sometimes crossed their outbred stocks in order to increase breeding performance. Moreover, these animals were mostly maintained by commercial breeders who sometimes trademarked their stock designation so that new colonies derived from one of these colonies had to be renamed. Six of the stocks were resampled at least one year later, and while five of them were unchanged, one of them (HsdOla:MF1) had changed substantially between 2003 and 2007, with heterozygosity declining from 0.30 to 0.05. This change was due to rederivation to eliminate infectious microorganisms.

In the absence of genetic contamination or genetic bottlenecks, the genetic architecture of colonies will remain stable because the colonies are usually maintained in such large numbers.

Genetics of outbred rats

Although there have been many GWA studies in rats, they have been focused on the study of a particular phenotype such as susceptibility to stroke, diabetes or drug abuse. In some cases these have used commercial stocks of Sprague–Dawley rats,

but no comparative studies of Sprague–Dawley versus Wistar rats or stocks of rats from different breeders have been performed which would throw light on the genetics of commercial rat stocks.

Genetically defined strains

These include inbred strains, F1 hybrids, congenic strains, and monozygous twins.

Inbred strains

Inbred strains of mice, rats and a few other species are produced by 20 or more generations of brother-by-sister mating, with all individuals being derived from a single breeding pair in the 20th or a subsequent generation. They are, in a sense, like immortal clones of genetically identical individuals because the same genotype is transmitted to each generation.

Inbred strains tend to be phenotypically uniform (when compared with outbred stocks). They stay genetically constant for many generations, and most have been genotyped at many loci, with the full DNA sequence being known for an increasing number of strains. There is substantial background information on the origin, history, genotype and phenotypic characteristics of each strain. Genetic quality control is relatively easy as, unlike outbred stocks, each animal can be genetically authenticated from a small sample of DNA.

Over 400 separate inbred strains of mice and 200 strains of rats are available throughout the world, and many are commercially available. They represent the nearest thing to a pure analytical grade reagent that is possible with animals, and several Nobel prizes have been awarded for work which depended on their use (Festing and Fisher, 2000). More Nobel prizes have been awarded since then.

About 80% of research involving isogenic strains is performed using the 10 most popular strains, with BALB/c, C3H, C57BL/6, CBA and DBA/2 being among the most widely used mouse strains and F344, LEW, SHR, WKY and BN being among the most widely used rat strains.

Each strain has its own unique characteristics, and strain differences can be found for almost any characteristic which has been studied. Some strains have a high incidence of cancer, others of heart disease, others have neither. Some are active, others are passive. Some like and others detest alcohol. Some learn well and others not so well in a particular learning task. Naturally, care has to be taken when choosing a strain which is appropriate to the particular research project. It would not be sensible to use the AKR mouse strain in a long-term carcinogen screening study because most of these mice would have died of leukaemia before they were one year old.

For general research where there is no requirement to use animals of a specific strain, the best strategy would be to use one or more of the most popular strains such as BALB/c or C57BL/6 mice or F344 or LEW rats. Where a series of experiments is planned it might be useful to do a small pilot study involving several strains to find one which responds most appropriately. It would be unwise to do a long series of experiments using a single strain without occasionally using a different strain to

make sure that the observed results are not unique to the chosen strain, although if the strain is unique in its response, this may itself be of interest.

Lists of inbred strains of mice and rats, with some of their phenotypic characteristics, are available on the Web (www.informatics.jax.org), and the genealogies of inbred mouse strains was updated in 2000 (Beck et al., 2000).

F1 hybrids

F1 hybrids, the first generation cross between two inbred strains, are also isogenic, and have the advantage that they tend to be more robust due to hybrid vigour. They can be used instead of, or in association with, inbred strains. However, unlike inbred strains they will not breed true as they are heterozygous at all the loci at which the parental strains differ, leading to genetic segregation in the F2 and later generations.

The correct genetic nomenclature should always be used so that the work can be repeated in other laboratories. The formal nomenclature rules are available at www.informatics.jax.org. Briefly, strains are known by a code which consists of a few upper-case letters and sometimes a number with a laboratory code. The rules are relatively well observed with mice but not so well with rats. Rat strains should be designated by a code such as F344, BN, LEW, *not* by a name, as this does not conform with the nomenclature, and can cause confusion.

Congenic, consomic and recombinant inbred strains

These are specialised types of isogenic strains developed largely for genetic studies. A pair of congenic strains is produced by backcrossing a defined genetic locus (the differential locus) from a donor strain to an inbred strain (the inbred partner). After about 12 generations of backcrossing the congenic strain will be very similar genetically to its inbred partner, but will differ at the differential locus. Any differences between the congenic strain and its inbred partner can (with some reservations) therefore be attributed to the effects of the differential locus.

Congenic strains have been widely used in transplantation immunology (Snell and Stimpfling, 1966), and many laboratories studying quantitative trait loci are now using these methods to isolate and study quantitative trail loci (QTLs) which control many disorders such as cancer, hypertension and diabetes. Genetic markers can be used to speed up the backcrossing program (Markel, 1997).

Sets of consomic strains differ from an inbred partner strain for a whole chromosome, which is derived from a donor strain. They are used in the genetic dissection of characters controlled by QTLs (Nadeau et al., 2000).

Several sets of recombinant inbred strains have also been developed from a cross between two standard inbred strains followed by 20 or more generations of brother-by-sister mating to produce a large number of new recombinant strains. These can be used to map genetic loci where the phenotypes differ between the two parental strains (Taylor, 1996). Further details of these and other genetic types are given by Silver, 1995.

The choice of inbred strains or outbred stocks in controlled experiments

When designing an experiment it is important to choose the right animals for the proposed study. Genetic variation is a major component of the variation in laboratory animals. GWA studies, for example, need genetically segregating outbred stocks with high levels of heterozygosity, low levels of linkage disequilibrium, and high expression of the phenotypes which are to be studied.

By contrast, randomised controlled experiments need uniform experimental subjects otherwise they will lack statistical power or larger sample sizes will be required. Inbred strains are generally more uniform than outbred stocks, so experiments using them will usually be more powerful and more likely to identify a treatment effect although strains may differ in their responses. A further advantage is that inbred strains are genetically defined so it is possible, from a sample of DNA, to test the animals to see if they are of the correct strain. Moreover, there is much more information about the phenotypic characteristics of inbred strains than of outbred stocks so it is possible to make a better informed choice of which strains to use.

The importance of phenotypic uniformity is illustrated in Table 2.1. This shows sleeping time under hexobarbital anaesthetic in five inbred strains and two outbred stocks of mice (Jay, 1955).

For instance if an experiment were to be set up to determine whether a drug influenced sleeping time in mice, there would be a drug-treated and a vehicle-treated group. Using a power analysis (see Chapter 11) with a 5% significance level, a two-sided test and a 90% power, an experiment designed to detect a 4 min change in sleeping time (mean SD of the 5 inbred strains = 3.2 mins.) between a treated and a control group would require 15 inbred mice or 240 outbred mice (mean SD=13.5) per group using the mean of the standard deviations of each group. Admittedly, this is an extreme example, but it illustrates the over-riding importance of controlling inter-individual variability, some of which is genetically determined.

Table 2.1 Sleeping time (mins) under hexobarbital anaesthetic in inbred and outbred mice.

Type and strain	*n*	Time	SD	Needed
Inbred				
A/N	25	48	4	
BALB/c	63	41	2	
C57BL/HeN	29	33	3	
C3HB/He	30	22	3	
SWR/HeN	38	18	4	
Mean inbred	37	32.4	3.2	15
Outbred				
CFW	47	48	12	
Swiss	47	43	15	
Mean outbred	47	45.5	13.5	240

See text for explanation (Jay, 1955). SD: standard deviation.

Environment and diet

Bedding, diet and physical environment can influence the outcome of an experiment. Reputable suppliers of bedding now generate their own sawdust from chosen timber known not to have been treated with insecticides. Softwood sawdust contains substances which may induce drug metabolising enzymes. These can alter the response of mice to toxic agents. Thus, in one study it halved the sleeping time of mice under barbiturate anaesthetic compared with mice maintained on hardwood bedding (Vesell, 1968), and in another it completely altered the response to a toxic chemical (Malkinson, 1979).

The frequency with which animal cages are cleaned may influence their responses. Mice and rats are uncomfortable if their cage is changed too frequently, but ammonia levels can build up to levels which might predispose the animals to respiratory disease if the cages are not cleaned frequently enough (Hoglund and Renstrom, 2001).

Diet can have an important influence on the characteristics of animals, particularly those on long-term studies. Modern rodent diets should be nutritionally complete and free of contaminants. Diets formulated to maintain pregnancy and lactation may be too nutritionally dense for the long-term maintenance of non-breeding rodents. One result is that these animals may become obese and develop a high incidence of diabetes, cancer and circulatory disorders, with shortened lifespans. Unfortunately, it has proved to be difficult to formulate diets which prevent this obesity. Dietary restriction may be a way of prolonging the life of animals on long-term studies (Masoro, 1993).

The physical environment can also influence the animals. It is a requirement under Directive 63/2010/EU that social animals such as mice and rats are group-housed unless a strong scientific case can be made for single housing.

Caging density may also influence the incidence and type of disease in mice (Les, 1972). On the other hand, male mice housed in groups may start to fight. An experimental protocol should never involve regrouping adult male mice as they are almost certain to fight.

Environmental enrichment is now widely practised. It has been shown to lead to the generation of more hippocampal neurons in mice (Kempermann et al., 1997).

Standardisation, reproducibility and external validity

Environmental enrichment in cages has been found generally not to have an effect on individual variability in mouse behaviour (Wolfer et al., 2004). On the other hand group-housed mice have been found to be more variable than singly-housed ones with regard to some traits (Prendergast et al., 2014). In this case sample sizes need to be increased. But the relationship between environmental enrichment, variability and reproducibility represents an important consideration that researchers should be aware of.

The concept of reproducibility is crucial to research. But in the context of animal studies so is external validity, i.e. the applicability of a result to other species (usually humans) living in inevitably different conditions. This is after all the bedrock of biomedical sciences.

Some authors have argued that efforts to standardise conditions, and therefore reproducibility, by rigorously homogenising environmental conditions within an experiment or a laboratory may, paradoxically, militate against reproducibility. For example, Fisher (1960) notes that 'the exact standardisation of experimental conditions, which is often thoughtlessly advocated as a panacea, always carries with it the real disadvantage that a highly standardised experiment supplies direct information only in respect of the narrow range of conditions achieved by standardisation. Standardisation, therefore, weakens rather than strengthens our ground for inferring a like result, when, as is invariably the case in practice, these conditions are somewhat varied'.

However, this does not mean that conditions should be randomly varied. Cox (1958) explains that 'we should, in designing the experiment, artificially vary conditions if we can do so without inflating the error'. This is done by using factorial and randomised block designs which sample different factors and environments without inflating the error. 'Standardisation fallacy' has recently been discussed in detail in relation to experiments involving laboratory animals (Wurbel, 2000). The effect of variation, and means of controlling it, is the topic of the next chapter.

3

Understanding and controlling variation

Introduction

Experiments can only be designed efficiently and economically if the research worker has a good understanding of biological variability, its causes, and ways in which experiments can be arranged in order to quantify and control it. There are two types of variability that need to be taken into account. Random variability (or random effects) due, for example, to uncontrolled inter-individual differences, and 'fixed effects' such as the sex, strain, age, diet and bedding which can be controlled to a large extent by the investigator. The experimental treatment is also a fixed effect. The ways in which these are handled are discussed in this chapter.

Statistical testing

The majority of statistical tests compare the size of the effect (the biological 'signal') relative to the amount of variability in the data (the 'noise'). Figure 3.1 illustrates what might be concluded about the same biological effect, under two different scenarios. In the first of these, the background variability is relatively low, and in the second, the variability is relatively high. An analogy would be a lecturer trying to speak with a loud radio in the same room.

The lecturer has a number of options:

1. Speak louder (increase the signal).
2. Turn the radio down (decrease the noise).
3. Find a different way to get the message across (e.g. seek an alternative method for delivering the signal such as a silent visual presentation).

In a statistical test, the ratio of signal-to-noise determines the significance. Hence if the variation or noise is large in an experiment, the biological effect, the signal may be hidden by it. There are usually many sources of noise in biological data. Imagine the same lecturer in a room with many radios. The task of turning the radios down

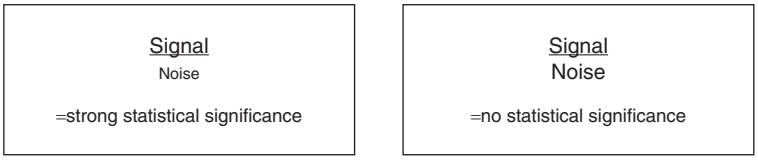


Figure 3.1 If the noise is low then the signal is detectable, but if the noise (i.e. individual variation) is high then the signal will be undetected. The signal may be the difference between two means and the noise could be the pooled standard deviation.

should now involve a first step of identifying the loudest radio(s). Only then can the noise be reduced. The experimenter should not automatically assume that the animal is the main source of noise. Measurement error of various sorts may well be important, and this can often be controlled by better experimental technique, by increasing the number of determinations, or by using blocking (see Chapter 7) or covariance (see Chapter 8) to eliminate otherwise uncontrollable variation. All laboratory determinations and techniques should be done to the highest possible 'GLP' (Good Laboratory Practice) standards.

Sources and types of variation: fixed and random effects

What causes the variability? In a training course, a group of scientists were asked to write down possible sources of variability in their experiments, and these were grouped together into the categories shown in Table 3.1. These variables do not all, necessarily, cause experimental noise. They fall into two main classes designated 'fixed effects' and 'random effects'.

Fixed effects

These variables may affect the outcome of an experiment, but they are largely under the control of the experimenter, and are of particular importance when considering the design of the experiment and in the interpretation of the results. The imposed treatment is the most obvious fixed effect, but the species, strain, sex, age or weight range of the animals, type of caging, type of bedding material and many of the other factors listed in Table 3.1 can be specified by the researcher. Often, it is a matter of scientific judgement whether the conclusions from an experiment using, say, male rats can be generalised to female rats or whether it makes any difference whether the rats were eight or 10 weeks old at the beginning of the experiment. But if it is expected that the response may be different in the two sexes, or age groups, then there are two possible courses of action. First, the scientist may be content to state that the results are only applicable to the sex or age of animals used in the experiment. Second, a factorial experiment (see Chapter 6) involving both sexes or ages

Table 3.1 Some of the variables which may influence the outcome of an experiment.

Environment	Temperature, humidity, season, barometric pressure, lunar cycle, noise, air movement, light level and cycle, smells, room characteristics, cage size and design, bedding material, nest box design, nest materials, number and gender of animals, water quality, diet availability and composition, handling, cleaning
Animals	Species, sex, strain, genotype, health status/microflora and fauna, origin, age, body weight, litter size, oestrus in females, aggression, biological rhythms
People	Gender of technicians and investigators, use of cosmetics, care with handling, investigator personality
Experiment	Type/quality of surgery, route of administration, dose levels, sampling of tissues/organs, time of day, test materials, shelf-life of solutions, calibration of instruments, measurement errors

as well as the treatments of interest could be used to explore whether or not this is the case. Usually this can be done without increasing the total number of animals (or other experimental units).

Some fixed effects such as gender and genotype are classifications. These cannot be randomised in the normal way. For example, an animal can be assigned treatment A or B, but it cannot be assigned male or female. If the experiment involves classification variables, such as genotype, it is up to the investigator to ensure that subjects are comparable in other ways apart from the classification. So, the males and females should be the same age and from the same source, although they are unlikely to be of the same body weight. Weight may be considered to be one part of the gender differences.

All experiments involve the choice of a wide range of fixed effects. So the investigator must decide which effects are likely to be of trivial importance, which may influence the interpretation of the results, and those which might need to be explored using factorial experimental designs.

Random effects

These are the variables that usually contribute noise or unwanted variability among the experimental units. Heterogeneity in body weight within the specified weight range, genotype within an outbred stock (although the experimenter can choose whether to use Wistar or Long–Evans rats, so the stock is usually a fixed effect), accidents of development, social hierarchy and within-group aggression, subclinical variation in pathogen burden, poorly mixed diet, inaccurately administered dose levels, contamination of blood or urine samples, inaccurate measurements or measurements made near the limits of detectability all contribute to variability and noise.

In an experiment, random effects are nearly always to be controlled or avoided, but this is done in a different way from fixed effects.

1. The first step is to try to obtain experimental material with a low intrinsic noise level. Isogenic animals (see Chapter 2), free of pathogens and of a narrow age/weight range, well acclimatised and housed in compatible groups in a good animal house is often the starting point.
2. Variability that occurs over time due to biological rhythms or the shelf-life of reagents can often be partially controlled by using a blocked design or covariance analysis (see Chapters 7 and 8), as can some of the variability due to position of the cages in the animal house.
3. Measurement errors can often be substantially reduced using multiple sample determinations. There may be a whole hierarchy of possibilities. In isolating enzymes or mRNA from a liver, several samples of liver could be taken and several determinations could be done on each. These would be averaged, although it is possible to do a 'components of variance' analysis to estimate the amount of variation associated with each level of sampling (see Chapter 8). Exactly how many liver samples and determinations per sample are chosen will depend on variation among samples and determinations.
4. Finally, if large levels of variability persist, and cannot be identified and controlled, then sample sizes can be increased to provide an experiment with sufficient statistical power to identify the smallest effect likely to be of clinical, biological or scientific interest (see Chapter 11).

The distinction between fixed and random effects is not always clear-cut. Body weight is a fixed effect if a weight range (e.g. 100–120 g) is specified, but within that range it will be a random effect which may cause noise. Likewise, if unsexed animals were used, then sex would be a random effect, but more usually the sex of the animals will be controlled, making it a fixed effect. The important point is whether the variable contributes uncontrolled noise, or whether it can be specified and controlled and/or studied as a factor in a factorial experimental design.

Although treatment is usually considered to be a fixed effect, in some cases an experiment is used to determine the magnitude of a random effect. For example, the effect of variation in commercial mouse diets on body weight of mice maintained on the diets could be determined by taking a random sample of diets from all available diets, and feeding these to mice under specified conditions. The results would normally be expressed in terms of the proportion of variation in mouse weight which could be attributed to variation among the diets. A brief example of this type of analysis estimating variation within and among cages and the associated statistical analysis is given in Chapter 8.

Identification of the important sources of variability

If the variation associated with what are assumed to be the most important random and fixed effects can be quantified, then a rational approach can be used to control

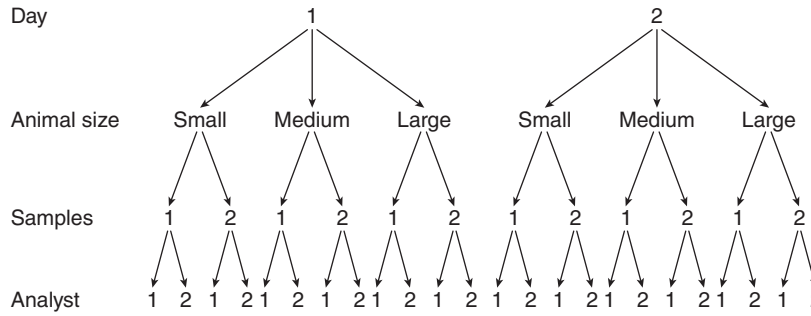


Figure 3.2 The structure of a 'nested' experiment designed to quantify the amount of variation between days, sizes of animals, samples and analysts.

or minimise its effect. For example, a randomised block design may be used if there are large fluctuations in mean results over a period of time or if the experiment has to be split into two rooms or even two shelves which may have different environments. Pilot experiments can be useful if several similar experiments are contemplated. For example, the experiment shown in Figure 3.2 was designed to explore the importance of day-to-day variation, animal size, the number of samples taken, and the analyst who analyses the samples. Note that animal size is a fixed effect, but the other variables are random effects if it is assumed that the two analysts are chosen from a pool of possible analysts.The resulting data could be analysed to show the relative importance of each of these variables. Thus, if there were large differences associated with animal size, then it would imply that body weight would need to be rigidly controlled and possibly included as a factor in a factorial experimental design. The latter would be appropriate if there were suspicions that the response might be different in large and small animals. If the day-to-day variation was relatively large, it could be controlled using blocking (see Chapter 7). If there were larger differences between samples, then triplicate or quadruplicate samples might be appropriate.

Examples

Atherosclerosis

In a study to investigate the effects of drug treatments in a rabbit model of atherosclerosis, there were three drug treatments and four animals on each drug. At autopsy the aorta was removed and five lateral sections were cut. The lesion area in each section was measured using image analysis software. An hierarchical analysis of variance (discussed in Chapter 8) was used to quantify the variability between sections, animals and drug treatments, and it was found that 47% of the variation was due to the variation between sections, 35% due to the effect of the drug (a fixed effect), and 18% due to the inter-animal variation. Thus, the best way of improving the power or precision of such an experiment would probably be to take more sections, rather than using more animals, as this may also be the cheapest alternative both financially and in terms of animal use.

Stroke

In a pilot experiment to study the effect of a new drug on a stroke model caused by occlusion of the median carotid artery in rats, a factorial design was used to explore the effect of the three occlusion periods (30 min, 45 min and 60 min) and two drugs, using five rats per group. Note that the occlusion period and drug treatment are fixed effects, hence the use of a factorial design (see Chapter 6). Most responses were highly variable. However, it was found that the variability was minimised with the 60 min occlusion period. With the 30 min period some rats did not develop stroke lesions, hence they were not responding to the drugs. Thus, rather than increasing the number of animals, it would be better to use the longer occlusion period in order to design more powerful experiments in the future. Factorial designs are often used in this way to determine an optimum set of conditions for obtaining the maximum yield of a product.

Randomisation and blinding

> Randomisation is one of the essential features of most experiments. The investigator who declines to randomise is digging a hole for himself, and he cannot expect the statistician to provide the ladder that will help him out. (Finney, 1978)

The aim of experimental design is, as far as possible, to remove all sources of variation among the experimental units both at the start and during the course of the experiment, apart from the explicit treatment or intervention. Although many differences can be controlled, some variation will always remain.

However similar they may be, no two animals and the environments in which they are maintained are identical. Even if they were identical it would be impossible for the same operator to administer the treatment simultaneously and measure the dependent variable at the same time, thus giving rise to differences in operator and time. Randomisation is essential to ensure that these remaining and inescapable differences are spread among all treatment groups with equal probability, thereby providing a reliable estimate of experimental variation or error and minimising any potential bias.

Note that it is not just a case of randomising the animals to the treatments. It is imperative that the randomisation applies to all aspects of the experiment such as the positions of the cages in the animal rooms and the order in which determinations of the outcome are made. It follows from this that a ‘control’ group, where one of the treatments is considered as a control, cannot be separated from the experiment. It should be studied at the same time as the other treatments. ‘Historical controls’, for example, are unlikely to be valid except in the special case of when many identical tests or ‘screens’ are done over a period of time in the same laboratory, so that it is possible to estimate both the mean and the variation between samples.

Basic randomisation procedures

The best way of ensuring that the randomisation is done correctly is to number the cages (or other experimental units) and then assign the treatments to them at random

using one of the methods described below. All subsequent procedures such as the positioning of the cages in the animal house and the measuring of the outcome should then be done by cage number. This will ensure that there are no systematic differences between treatment groups, for instance in the location in the animal house or in the order in which the observations are made.

Physical randomisation

Physical randomisation is extremely easy. The experimental units (e.g. animals) are numbered 1–n, where n is the total number of experimental units and the numbers are also written on slips of paper which are folded and put in a receptacle. This is thoroughly shaken, and assuming a group size of four, four slips of paper are withdrawn and the numbers drawn are assigned to treatment A, the next four withdrawn to treatment B, and so on.

Using EXCEL and other computer software

Any good statistical package will have procedures to assign units and treatments randomly to each other. However a spreadsheet provides a good way of randomising the experiment.

A completely randomised design (see Chapter 5)

Suppose the experiment is to have two treatments A and B with six animals per treatment in a completely randomised design. The animal numbers 1 to 12 could be put in the first column. Six As and six Bs are put in column two and a random number is put

Table 3.2 Randomisation of a completely randomised design using EXCEL. (1)

ID	Treatment	Rand.No	Treatment randomised	Rand No. sorted
1	A	0.573	B	0.010
2	A	0.440	B	0.016
3	A	0.096	A	0.096
4	A	0.140	B	0.114
5	A	0.368	A	0.140
6	A	0.806	B	0.196
7	B	0.010	B	0.222
8	B	0.016	A	0.368
9	B	0.222	A	0.440
10	B	0.996	A	0.573
11	B	0.196	A	0.806
12	B	0.114	B	0.996

(1) Assuming treatments A and B are to be assigned at random to animals 1–12 using EXCEL. Columns 2 and 3 should be marked and sorted on column 3 to give the random order shown in column 4. Random numbers only shown to three decimal places.

in column 3 using the cell formula"=RAND()' (case independent). These random numbers could be copied and pasted back as values (although this is not strictly necessary). Then columns two and three are sorted on column three. This will put column two into a random order. The spreadsheet before and after the sorting is shown in Table 3.2.

A factorial design (see Chapter 6)

A 2 × 2 factorial design with, say, two treatments 1 and 2 and both sexes, F and M will have four treatment combinations: F1, F2, M1, M2. If there are to be three experimental units per treatment combination, then column one will have the numbers 1–12, column 2 will have F1, F2, M1, M2, three times and column 3 will have 12 random numbers, columns 2 and 3 should then be sorted on column 3.

A randomised block design (see Chapter 7)

In this design the experiment is split up into several 'mini-experiments', each with one experimental unit per treatment. Each block is randomised separately. So using EXCEL in the above factorial experiment there will be four blocks designated 1, 2, 3, 4 each with the four F1, F2, M1, M2 treatments. Randomisation is then done by sorting firstly by the random number column and then by the block column. The result is shown in Table 3.3. The treatments are now in random order in each block.

Table 3.3 Example showing randomisation of a randomised block design using EXCEL assuming four treatments M1, M2, F1, F2 and four blocks. Treatments are now in random order.

Treatment	Block	Rnum	Animal
M1	1	0.012198	1
M2	1	0.193397	2
F1	1	0.339560	3
F2	1	0.856625	4
F2	2	0.105376	5
M1	2	0.125236	6
F1	2	0.603448	7
M2	2	0.859552	8
M2	3	0.017276	9
F2	3	0.175645	10
M1	3	0.177146	11
F1	3	0.737313	12
F1	4	0.126645	13
F2	4	0.719067	14
M2	4	0.817535	15
M1	4	0.927656	16

Treatments were initially entered as four M1s, four F1s, four M2s, etc. and the block as 1, 2, 3, 4 four times. Treatments were then sorted first by the random number column (Rnum) and second by the Block column. Each block now has one animal of each treatment in random order in each block.

Improving the randomisation

In small experiments randomisation may not always be satisfactory. By chance all of one treatment group may be located in, say, the first half of the experiment. In this case re-randomisation is acceptable (press F9 in EXCEL to update random numbers, and then re-sort the EXCEL sheet). A better alternative would be to use a randomised block design (see Chapter 7).

Some people advocate 'improving' on the random allocation of animals to the treatment groups by moving animals from one group to another so as to have exactly the same mean body weight in each group before starting the experiment. The problem with this approach is that by minimising the mean differences between the groups, the variation within groups is increased. This may result in a reduction in the power of the experiment. If body weight in the available animals is quite variable, then a randomised block design should be considered, with blocking on body weight.

Blinding

When the outcome is being assessed, the person doing the assessment or measurement should not know which treatment was performed on the experimental unit being measured. Failure to blind can lead to one treatment group being favoured either intentionally or unconsciously, giving a biased result. If the above methods of randomisation are used then following the treatment the cages or experimental units should only be identified by a number. This will ensure that the investigator is blinded to the treatment at the time that the outcome is measured.

4

The analysis of variance

Introduction to the analysis of variance (ANOVA)

Despite its name, the ANOVA is a method for comparing means, not variances. According to Fisher, who developed the ANOVA:

> The arithmetical discussion by which the experiment is to be interpreted is known as the analysis of variance. This is a simple arithmetical procedure, by means of which the results may be arranged and presented in a single compact table which shows both the structure of the experiment and the relevant results, in such a way as to facilitate the necessary tests of their significance. (Fisher, 1960)

The ANOVA can be used for simple experiments such as those comparing the means of two groups (in which case it gives identical results to Student's *t*-tests), as well as to more complex designs such as randomised blocks, and factorial designs. It is essential for anyone using experimental animals to have a basic understanding of this powerful technique because it is the only sensible method for analysing the majority of experiments.

One-way ANOVA

An example of a simple one-way ANOVA table is shown in Table 4.1 (using data from the example in Chapter 5) and diagrammatically in Figure 4.1. The column headed 'Source' indicates the source of the variation, which in this case consists of 'Treatment', 'Residuals' and 'Total'. Other or slightly different labels may be used depending on the software package (e.g. 'Residuals' may be labelled 'Error', and R and Rcmdr do not label the first column or show the 'Total' row). In a two-way ANOVA, additional sources of variation such as gender, blocks, rows and columns (say in a Latin square design) and covariates (in the analysis of covariance) may be present.

In the one-way ANOVA table (Table 4.1) DF stands for the 'degrees of freedom' in the experiment. This is defined as 'the sample size, n, minus the number of parameters, p, estimated from the data' Crawley, 2005). The variance of a sample of size n is

the sum of squared deviations of each observation from the mean divided by $n - 1$. In order to estimate this, one parameter (the sample mean) must first be calculated. Hence the DF is $n - 1$. Thus in this experiment the total number of independent observations available to estimate the overall variance is $n - 1 = 26$. As there are three groups the DF associated with the differences among group means is $3 - 1$. Finally, the residual DF is obtained by the difference between the total and the treatment DF: $26 - 2 = 24$.

The column headed 'SS' gives the sum of the squared deviations from the respective means. These deviations quantify the variation associated with each source. It is possible to express the treatment as a percent of the total variation (in this case 52%). These SSs are converted to the mean squares (MSs) by dividing them by the DF.

The 'Residual' or 'Error' mean square is of particular importance because it is an estimate of the pooled within-group variance. So the square root of this is the pooled standard deviation (SD).

The column headed *F*-value shows the 'variance ratio' statistic, being the treatment MS divided by the residual (error) MS. It is designated as 'F' in honour of Fisher. Before computers were available the *F*-values were looked up in tables according to the error and treatment DF to determine the *P*-values. However, this is now done by computers. The *P*-value (0.00014) in the ANOVA table gives the probability that a difference in means as great as or greater than that observed could have arisen as a result of chance sampling variation when there is no true difference between groups (i.e. the null hypothesis of no differences among means is true). By convention, a difference among means with a *P*-value of >0.05 is often considered to be 'not significant', one of $P < 0.05$ is considered 'significant' and one of $P < 0.01$ is 'highly significant'. However, these cut-off points are entirely arbitrary, dating back to the time when exact *P*-values could not be calculated. It is important to quote the actual *P*-value rather than to just state whether it is less than or greater than 0.05, thus allowing readers to make up their own mind on how to interpret the result. Phrases such as 'borderline significance' should be avoided.

A *P*-value of 0.06 does not mean that the treatment has had no effect. It might just be that the experiment was too small or the variation among individuals too great to be able to detect an effect. Absence of evidence is not evidence of absence. When there are more than two treatments the ANOVA only gives an indication of whether overall the means are different. Further calculations are needed to determine which of these means differ. This is often done using post hoc comparisons.

Table 4.1 Example of an analysis of variance table. (1)

Source	DF	SS	MS	*F*-value	*P*-value
Treatment	2	2861	1430.5	13.13	0.00014
Residuals	24	2614	108.9		
Total	26	5475			

DF: degrees of freedom, SS: sum of the squared deviations, MS: mean squares.

(1) The headings in all tabular output from Rcmdr has been edited to conform to that found in most other packages and textbooks.

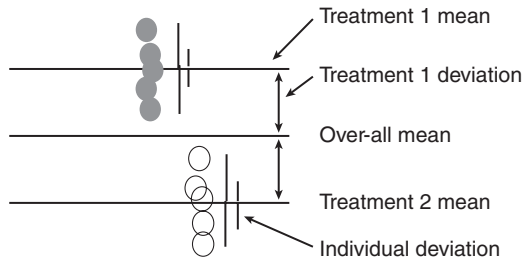


Figure 4.1 Treatment and individual deviations from the overall mean. The treatment sum of squares is the sum of the squared deviations from the overall mean and the error sum of squares is the sum of the squared deviations of each observation from its treatment mean. Jitter has been added so that the points are separated horizontally.

Two-way ANOVA

In an experiment the 'factors' (often called, in a wider context, explanatory variables) can either be fixed or random effects, as discussed in Chapter 2. The one-way ANOVA has a single fixed effect factor, the 'treatment'. By contrast, a randomised block design, in which the experiment is split up into a number of mini-experiments (see Chapter 7), has one or more fixed effects (the treatments) as well as a random effect (the block). So the ANOVA table has an additional row called 'Blocks'. A Latin square design will have two random effect factors often designated as 'Rows' and 'Columns' as well as one or more fixed effect factors. A factorial experiment has two or more fixed effect factors such as 'treatments', 'gender' and/or 'strain'. It will also have some interactions such as 'treatment × gender' which will indicate whether males and females respond in the same way to the treatments. Examples of these types of experiments and the associated ANOVA tables are given in later chapters.

Assumptions about the data when using an ANOVA

The validity of the ANOVA depends on three assumptions, which should normally be examined as part of the statistical analysis, as is done in most of the examples given in later chapters.

1. The observations are statistically independent of one another. This assumption depends on correct identification of the experimental unit as discussed in Chapter 1, and correct randomisation.
2. The residuals (i.e. the deviation of each observation from its group mean, see Figure 4.1) should have a normal distribution (bell-shaped). This is often the case, but with some types of data this assumption may be seriously violated, in which case either the data need to be transformed to another scale (see below) or a non-parametric test (which is based on ranks instead of the observed data) needs to be used. The ANOVA is quite tolerant of small deviations from this assumption.
3. The variances within each group should be approximately equal (homogeneous). Again, a scale transformation may also be appropriate if this assumption is seriously violated.

Assumptions 2 and 3 can be examined using residual diagnostic plots which are introduced in Chapter 5.

Scale transformations

A scale transformation may sometimes be used to make the data satisfy the assumptions for a statistical analysis using the ANOVA. The three most common transformations are:

1. The logarithmic transformation – If the data are skewed with a long tail of high numbers or if the variability increases with larger values of the dependent variable then transforming the data to $\log(x)$ can be used. If there are any zero or negative numbers in the observations, a value should be added that makes them larger than zero because the logarithm is undefined for values ≤ 0.
2. The square root transformation – Counts, with a low mean (resembling a Poisson distribution) will usually have heterogeneous variances. Observations can be transformed to the square root of the counts ($\sqrt{(x)}$) as this exaggerates differences among low values and deflates differences among higher values.
3. With percentage data an arcsine transformation is appropriate for percentage data when many of the observations are <20% or >80%. This is $\sin^{-1} \times \sqrt{(0.01 \times p)}$ where p is the percentage. The calculations can be done in EXCEL.

If none of these transformations work, an alternative is to rank all the data and then perform an ANOVA on these ranks (Montgomery, 1997).

Note that a more general approach would be to use generalised linear models (GLMs) for data that are unlikely to be directly suitable for an ANOVA (Crawley, 2005). However, most data from designed biological experiments can be analysed using an ANOVA.

On the rare occasions when there is no suitable transformation an alternative is to use a non-parametric test. These are not discussed here, but are covered in most statistical texts or specialised monographs. Their main disadvantage is that they may lack power and they are not available for complex designs such as factorial experiments.

Comparisons of treatment group means following the ANOVA

When more than two treatments are being compared, the ANOVA indicates whether there is a significant overall difference among them, but not which ones differ.

There are two main ways of making such comparisons. The first is to plan exactly which comparisons are of interest, and to use an appropriate set of what are called ‘orthogonal comparisons’. This is a fairly flexible method which can be used, for example, to study different comparisons of treatment means. The method can also determine whether there is a linear and/or curved trend in the response to the dose levels of a compound. Unfortunately, the methods are only supported by the more

advanced statistical packages; and although the calculations are not difficult, and can be done by hand, it is not always easy to understand from the textbooks on how they are done. More details are given elsewhere (Altman, 1982; Crawley, 2005; Snedecor and Cochran, 1980).

The more common and easier alternative is to use post hoc ('after the event') comparisons of treatment means. These are available in most statistical packages. They work on the premise that the more statistical tests that are done on the same set of data, the greater the risk that at least one of these tests will be significant purely by chance (a false-positive result). Post hoc tests control this risk. However, they should only be used if the ANOVA indicates that there are statistically significant overall differences among the groups.

Post hoc comparisons

The following tests are commonly used:

1. Dunnett's test is appropriate for comparing several dose groups with a control group (R-Commander (Rcmdr) automatically compares the first group, identified alphabetically by treatment name, with all other groups. So treatments could be named 'A.control', 'B.Apples', 'C.Pears', etc. if this test is to be used).
2. Methods for comparing several treatments include Fisher's least significant difference (LSD), Scheffe's test, Duncan's multiple range test, Neuman–Keuls' test, and Tukey's test. This last test is used by Rcmdr. These differ slightly. It is suggested that investigators should use the ones available on the statistical software which they are using.
3. Bonferroni's method can be used for making a small subset of treatment comparisons with the *P*-value for each treatment comparison being declared significant if it is less than α/k, where α is the chosen significance level (say 0.05) and k is the number of pair-wise treatment comparisons. This method is very conservative and can lead to too many false-negative results if k is large.
4. In situations where there are many comparisons, such as in a microarray analysis, a false discovery rate method is used (Pawitan et al., 2005), but this is not discussed here.

In general, experiments should be designed to test relatively simple hypotheses, and large numbers of post hoc comparisons should be avoided.

Internal and external validity

The concepts of 'internal' and 'external' validity are to do with the generality of the results which are produced. The internal validity of an experiment is concerned with whether the design is appropriate to answer a specific question. By contrast, external validity is concerned with whether the results are likely to be general across a range of other conditions. When designing an experiment it is often sensible to consider trying to increase the external validity of the results as this can often be

achieved at little extra cost by using factorial and randomised block designs which sample different environments and treatments. It could, for example, be useful to know whether the results depend on the gender or strain of the animals which are to be used. This is done by including both genders and/or more than one strain in the experiment using a factorial design.

Choice of an experimental design

The range of available designs

Table 4.2 summarises some of the more common experimental designs used in biology. These include the following six designs.

The completely randomised (CR) single factor design

This is the design most widely used both in clinical trials and in experiments with laboratory animals with a single factor, namely the 'treatment'. There can be any number of treatments such as differing dose levels, genotypes or diets. One of the treatments may be designated as the 'control'. However, the design is not very efficient. It may lack power due to poor control of inter-individual variability; and it may be subject to bias due to faulty or unlucky randomisation, particularly with small experiments. In most cases when using laboratory animals a randomised block (RB) design would be a better choice.

Randomised block and Latin square designs

These can reduce some sources of variation found in CR designs. This is done by choosing subsets of experimental units either matched for some physical characteristics or grouped in time. Typically, these subsets (called blocks) are mini-experiments and have one experimental unit per treatment. Randomisation is done separately within each block. These designs can often take account of some natural structure of the experimental material such as litters, using a within-litter design. Within-animal and crossover designs where the experimental unit is a part of an animal or an animal for a period of time are also blocked designs, with the individual animals being the block.

Factorial 'designs'

Strictly, these are not 'designs'. They are an arrangement of treatments. True designs such as CR and RB designs are mutually exclusive. It is not possible to have an experiment which is both an RB and a CR design. However both CR and RB designs can have a factorial arrangement of treatments (with two or more fixed effect factors).

The US National Institutes of Health (NIH) now requires scientists to use animals (and even cell lines) of both sexes in their experiments. This can usually be done quite simply in either a CR or an RB design by using half males and half females without increasing the total number of animals. So in addition to the factor 'treatment',

there is another factor 'gender'. The result is a 'factorial' experiment with two factors, treatment and gender. This principle of splitting the available material to add extra factors could be extended to include two or more strains, two or more diets, etc. The main limitation is that the experiment should not be too large to handle and it should have sufficient DF in the resulting multi-way ANOVA to obtain a good estimate of error. These designs show the effects of each factor separately, as well as their joint effects (interaction). If males and females differ in their response to the treatments, this would be an 'interaction'. Advanced versions of factorial designs can be used, for example, to find the optimum combination of a large number of factors. For a given input of resources (animals, reagents, time, etc.), factorial designs will normally provide more information than a single factor design, at little or no extra cost. These designs are discussed in Chapter 6.

Repeated measures designs

There is some confusion in the literature on the definition of a 'repeated measures' design. Some statisticians have used this term for what others call a 'crossover' design in which the experimental unit is an animal for a period of time, and the animal receives different treatments over time. Mathematically this is really an RB design with the animal being the block (see Chapter 7). Other investigators use the term for an experiment where the same individual is measured several times without receiving a different treatment. In this case the statistical analysis is controversial. Here it is suggested that the observations are combined into either a mean response, a trend in response, time to peak response, or the area under a curve (see Chapter 8).

Split plot designs

Currently, these designs (see Chapter 8) are rare but could become more common in work with laboratory animals following the requirement by the NIH to use both sexes in an experiment. They could be the natural design in situations where animals with different treatments or conditions (such as genetically modified animals with different modifications) can be kept in the same cage. So the animal is the experimental unit and the cage represents a block of an RB design. However, in order to include both sexes half the cages could be male and half female. So for comparing males and females the cage becomes the experimental unit but for comparing treatments (or genotypes) the animal is the experimental unit. These designs are usually defined as an RB design in which blocks are confounded with a fixed effect factor.

Sequential designs

In most experiments the sample size is determined before starting the experiment. However, if the response is much larger than expected the sample sizes may be unnecessarily large, so subjects will be wasted. Conversely, if the response is lower than expected it may be missed. In the sequential design sample size is determined as the experiment progresses. Consider a simple experiment such as whether surgical treatment A is 'better' in some way than treatment B. The first pair of animals is

Table 4.2 Summary of some of the more common experimental designs.

Type of design	When to use	Advantage	Disadvantage	Notes
The completely randomised (CR) designs	Fixed effects: when uncontrollable sources of variation are unlikely to be important.	Simplest 'designs'. Easy to use and analyse. Less affected by unequal sample size.	No control of additional 'nuisance' variation caused by uncontrolled (i.e. random time and space variation) which might affect results.	Typical designs used when doing simple t-tests.
Randomised block (RB) designs (including Latin squares, within-subject and crossover designs)	Should be the default design. Usually more powerful than a CR design as it helps to control space and time variation. It is under used in the biomedical literature. In crossover designs the experimental unit is an animal for a period of time. Within-subject designs are when several experimental units are provided by a single animal.	Deals with time and space heterogeneity by breaking the experiment into a series of subunits which are combined in the analysis.	Unequal sample sizes present some problems. Can be sensitive to missing values. Becomes increasingly complex when dealing with two (Latin square) or three sources (Graeco–Latin squares) of unwanted variation.	These designs aim to improve the power of the experiment. They can have a factorial arrangement of treatments. The analysis of covariance is another way of increasing power by correcting for initial variation.
Factorial 'designs'	Although called 'designs' they are really an arrangement of treatments and can be used with both CR and RB designs. They can provide extra information at no extra cost by testing *simultaneously* the potential effect of several factors (e.g. treatment, sex, strain) and their interactions on the response (dependent) variable(s).	A much more powerful alternative to doing several smaller experiments for each factor. Allows testing for interactions between these factors.	Three-way or higher interactions are rare and often difficult to interpret. But if they are found it may be due to faulty techniques.	These 'designs' increase the *amount of information* yielded by the experiment. They can be used both with CR and RB designs.

(Continued)

Table 4.2 (Continued)

Type of design	When to use	Advantage	Disadvantage	Notes
Sequential designs	When results with individual animals can be obtained quickly, and where the aim is to minimise animal use.	Results are analysed as the experiment progresses, enabling the experimenter to stop as soon as a pre-defined outcome is obtained.	Requires expert advice. Logistics may be a problem.	Up-and-down method may be replacing the classical LD50 test in the USA.
Repeated measures designs	When measuring the same individual several times following treatment. If the experimental unit is an animal for a period of time and treatments follow sequentially, then this is really an RB design (the animals is the block). Otherwise it is a repeated measures design	It is often necessary to follow an experimental subject over a period of time.	The statistical analysis is controversial. It is probably best to combine the observations into a single response such as shown in Figure 8.3.	The term 'repeated measures' should probably be abandoned as it has been used for two different experimental designs.
The split plot design	An RB design in which a major factor is confounded with the blocks. Can be used to include both sexes in experiments where animals receiving different treatments can be housed in the same cage or if whole blocks use animals of different genders. May occur as a within-subject design using both genders	If animals with different treatments can be housed together it will maximise the use of space. But if both sexes are to be included some cages will need to be only female and others only male.	Housing animals with different treatments in the same cage is not always possible. Has two different experimental units which require separate analyses.	Sometimes investigators use this design without understanding how it is to be analysed.

treated. In the absence of a real difference there will be a 50% chance that A appears to be better than B. With the second pair the chance that A is better than B both times will be $1/2 \times 1/2 = 0.25$. By the time that the sixth pair is processed the probability that A appears better than B every time when in fact there is no difference is $1/64 = 0.016$. It might be decided to stop at this stage and claim that A is significantly better than B. But that is the simplest situation. If treatment A is only a little bit better than B, the outcome may not be so clear cut. A detailed consideration of these designs is beyond the scope of this book, and a useful source of further information is given in http://www.sumo.intec.ugent.be/SED (accessed December 2015).

Experiments versus surveys

Experiments investigate whether an independent variable causes changes in a dependent variable. By contrast a survey gathers information about processes which cannot be controlled by the investigator. In a survey, the variable of interest, such as level of smoking, cannot be deliberately varied, and inferences about its effects on health in a population can only be based on associations or correlations, but not on causation. This can lead to spurious findings such as the well-known example of the strong correlation between the number of storks in German towns and the birth rate: larger towns have more humans, and more chimneys for storks to nest on, but more storks do not cause more babies.

5

The completely randomised single factor design

Introduction

The completely randomised (CR) design is relatively simple and is widely used in clinical trials and laboratory animal research. A strong case can be made for the more widespread use of the randomised block (RB) design in laboratory animal research as it is usually more powerful and less liable to bias through faulty or unlucky randomisation.

In the CR design experimental units are allocated to treatments completely at random. There is a single factor the 'treatment', which can have any number of levels and there can be any number of replicates (sample sizes) per treatment group. Group numbers may be equal or unequal. However, if there are several treatments, each of which will be compared with the control group, then the sample size in the controls might be increased. The design can also have a factorial layout with two or more fixed effect factors, as discussed separately in Chapter 6.

This design is easy to use, it is relatively unaffected by unequal numbers in each treatment group, and it is easy to analyse. However, a CR design does not take account of any identifiable variability among the experimental units (e.g. due to litter effects) and will be inefficient if the experimental material or the environment during the experiment is heterogeneous. It may not be a good design with large experiments because measurements may need to be made over an extended time period, and large numbers of similar experimental units and a large uniform environment may be difficult to find.

A fictitious example

Does jogging improve memory? Suppose that the aim is to test the hypothesis that running has an effect on learning and memory. Contrary to the old belief that the number of neurons can only decrease after birth, it is now known that new neurons

can be added to the adult brain. In some bird species, the part of the brain associated with singing grows during the breeding season, and shrinks after that (Tramontin and Brenwitz, 2000). In mice, exposure to an enriched environment increases the production of new neurons (neurogenesis) in some parts of the brain (Kempermann et al., 1997) and so does general physical activity.

Suppose, therefore, that an experiment is planned to test whether physical activity (e.g. wheel running) enhances learning ability and memory in mice. How should it be designed? This depends on the questions being asked, i.e. what should be measured, and how much additional ('nuisance') variation is likely to be caused by, say, sex, genotype or weight and which may mask the treatment effects. Are there interactions between the independent variables that are used? The first point about design is that the experimenters must be very clear about the hypotheses to be tested, but also be aware of variables which may alter the conclusions.

In this example, three levels of running might be chosen:

1. No running (a non-rotating wheel is placed in the cage, say for 30 min each day).
2. Moderate running (mice are allowed access to a running wheel for 30 min per 24 h).
3. Marathon running (access for 3 h per 24 h).

These treatments are to last for three weeks before the mice are tested for learning ability in a maze. There are many ways that learning ability can be measured in practice. For simplicity, the response is designated 'learning ability' without detailing the specific measure that was used to determine it. Low values represent good learning ability (indicated by a short response time).

Formally, the hypotheses of interest are:

- H_0 (the null hypothesis): all mean learning scores are equal (i.e. there is no effect of running activity on learning ability).
- H_1 (the alternative hypothesis): at least one mean differs from the rest (i.e. running has some effect).

The data comprise nine observations for each running treatment, and are shown in Table 5.1. Note that the treatments should be assigned to the animals at random, using EXCEL as described earlier, and the cages should be distributed in the animal house in order of cage number, not treatment. As the treatments are to last for five days a week for three weeks it might be advisable to use coloured labels to identify which animals have which treatments during the treatment stage. However, once the treatment regime is completed those labels should be removed, leaving only animal numbers during the assessment of the final learning scores so that the investigators doing the scoring are blinded to the treatment.

There are two kinds of variation: between- and within-running regimes. The former refers to the deviation of the treatment mean levels from the grand (overall) mean. This will be large if there is a treatment effect. The deviation of each observation from its corresponding treatment mean ('noise') is assumed to be approximately

Table 5.1 Raw data for the running experiment.

Animal	Group	Score	Animal	Group	Score	Animal	Group	Score
1	B	212	10	C	232	19	A	250
2	A	238	11	C	218	20	C	229
3	A	259	12	C	212	21	B	228
4	B	216	13	A	246	22	A	230
5	A	258	14	B	227	23	C	233
6	C	219	15	C	205	24	B	242
7	B	221	16	A	251	25	B	241
8	C	218	17	C	230	26	B	229
9	B	238	18	A	252	27	A	231

Note that the data are still in a random order. It is split over three columns for presentation.

Group A = None, Group B = Moderate, Group C = Marathon running.

the same in each group. If the null hypothesis is false, the ratio of between to within variances should be large: this ratio is the *F*-value from the analysis of variance (ANOVA) table.

Data analysis

The raw data in randomised order is given in Table 5.1. The statistical analysis is done using R-Commander (Rcmdr) (see Appendix 1) which can be used to analyse all the examples in this book. Most other commercial packages such as MINITAB and SPSS will also do these calculations, although some of the lower-end packages will not perform a three-way ANOVA, which is required for some of the analyses in Chapter 6 and later.

Importing the data into Rcmdr and graphical screening the data for obvious errors

The data are read from EXCEL into Rcmdr via the clipboard (*Data*, *import data*, *from clipboard*). The default name 'Dataset' is given but can be altered if more than one set of data is to be analysed.

It is always good practice to plot the data to show individual observations. Any outliers can be checked for errors. ***Rcmdr*** provides a 'stripchart' for this purpose (*Graphs*, *Stripchart*). Jitter can be added so that there is less chance that the points fall on top of each other. The resulting plot is shown in Figure 5.1. In this case there are no obvious points that need to be checked and the variation seems to be about the same in each group.

Box and whisker plots (*Graphs*, *boxplot*) will also show outliers if they exist and may pick up other abnormalities. The bar across the box is set at the median. The box covers the interquartile range the whiskers show the maximum and minimum

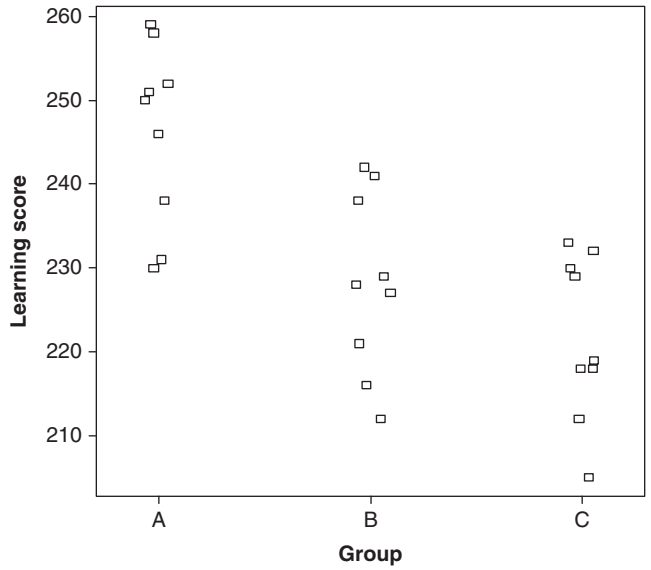


Figure 5.1 A stripchart of learning scores. Jitter has been added so that the points are separated. A plot of individual points, as above, can show any obvious outliers and give a general impression of the distribution of the data. The term 'stripchart' is used by R-Commander for a plot where the X-axis is a discrete factor. This is in contrast to a 'scatterplot' where both X and Y axes are continuous variables.

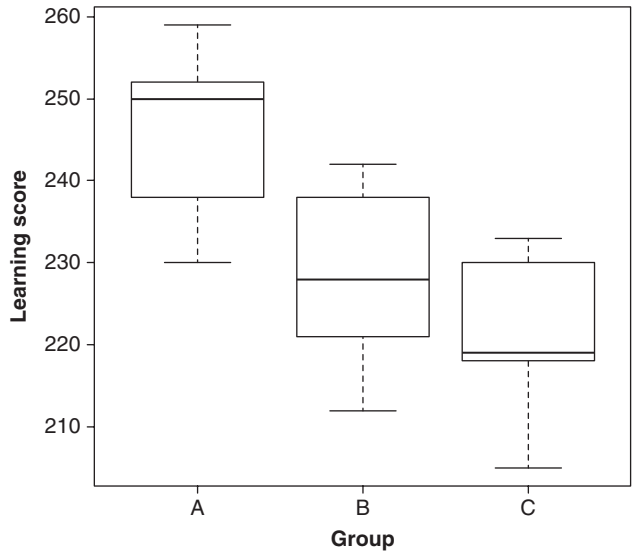


Figure 5.2 Box and whisker plot of the learning data. A = None, B = Moderate, C = Marathon running. Note that the box covers the interquartile range which includes half the numbers. The horizontal line is at the median. The whiskers show the maximum and minimum values, respectively. The results in this case are unusual with the median bar close to the interquartile edge of the boxes in both groups A and C. This implies some bunching at the top and bottom.

values. Any points considered to be outliers are shown as asterisks. In this case (see Figure 5.2) there are no outliers, but in groups A and C the median is quite near the edge of the box. There seems to be some bunching of the results in the 'None' and 'Marathon' treatment groups. This hypothetical example does not say how learning was measured but this sort of result could be observed if many animals either learned or did not learn, with just a few of them half-learning the task. Learning score might be an average of several runs in a maze, for example.

A plot of means (see Figure 5.3) with confidence intervals or other bars is also easily generated (*Graphs*, *plot of means*). Clearly there is good evidence of a treatment effect.

A range of other statistics such as means and standard deviations (SDs) in each group are also available.

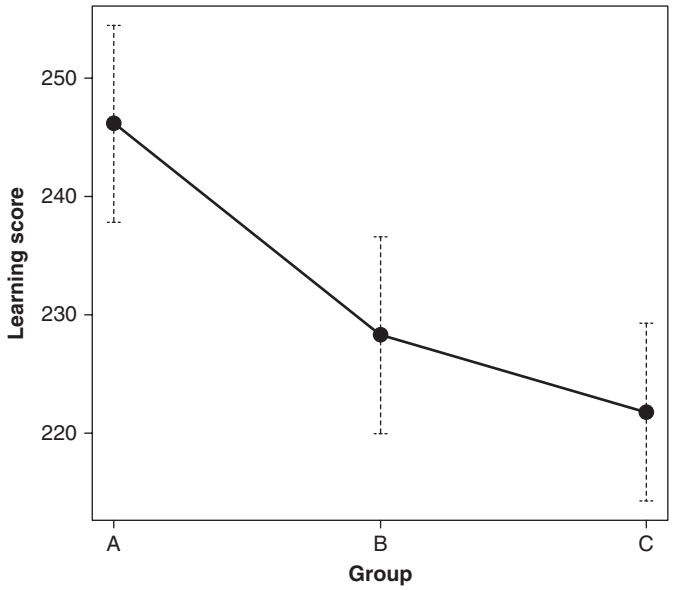


Figure 5.3 Plot of means with 95% confidence intervals. There is clearly a large response.

ANOVA and post hoc comparisons

The data can be statistically analysed in Rcmdr using statistics, means, one-way ANOVA, with the results shown in Table 5.2. There are highly significant differences between group means. If the pairwise comparison box is ticked, then post hoc comparisons will be generated using Tukey's method to account for multiple testing. These can also be shown as a plot (Figure 5.4). This shows group A differs from groups B and C ($P < 0.05$), but these latter two groups do not differ ($P > 0.05$). Group means and SDs are shown in Table 5.3. These analyses again show that the two running groups differ significantly ($P < 0.05$) from the controls, but Moderate does not differ significantly from Marathon. An alternative, more general, way of performing an ANOVA in

Rcmdr is discussed in Chapter 6. This will carry out the equivalent of Dunnett's test, comparing each group with the first group (determined alphabetically).

Table 5.2 ANOVA table for the running data as produced by R-Commander (Rcmdr). Summary (AnovaModel.1).

	DF	SS	MS	*F*-value	*P*-value (>F)
Group	2	2861	1430.5	13.13	0.00014***
Residuals	24	2614	108.9		

Note that Rcmdr does not label the first column 'Source', nor does it include a row of totals as in many other statistical packages. Also the headings (SS, MS etc) have been altered to conform to the more usual output from a statistical package. DF: degrees of freedom, SS: sum of squared deviations, MS: mean squares.

***indicates statistical significance at $p<0.001$

Table 5.3 Group means and standard deviations (SDs).

Treatment	Mean	SD	Significance[1]
A	246	10.8	a
B	228	10.7	b
C	221	9.8	b
Pooled SD		10.4	

These have been expressed to three significant digits. $n = 9$ in each case. [1]Means with the same letters are not significantly different ($P > 0.05$).

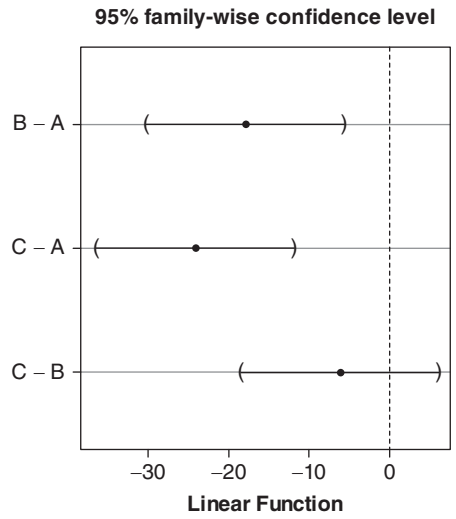


Figure 5.4 Post hoc comparisons using Tukey's method of correcting for multiple testing. Both groups B and C differ from group A, but groups B and C do not differ at the 5% level of probability.

Another statistic that may be useful is the observed standardised effect size (SES). This is the response in SD units. In this case for Group B it is (246 – 228)/ √ 108.9 = 1.72, i.e. the difference between the means divided by the pooled SD (the square root of the residual mean square). For group C it is 2.30. Prospective SESs are used in estimating sample sizes, discussed in Chapter 8.

Assumptions for a valid ANOVA

As described in chapter 4, the ANOVA relies on the assumptions of (1) independent observations – this depends on the correct identification of the experimental unit (a mouse) and correct randomisation; (2) normality of the residuals; and (3) homogeneity of the variance. The last two of these can be assessed using 'residual model diagnostic plots'. These are found in Rcmdr under *Models*, *Graphs*, *Basic diagnostic plots*.

The resulting plots are shown in Figure 5.5 (only the top two plots are discussed). The top left-hand plot of the Residuals versus Fitted values is designed to show any heterogeneity of variation among the groups. It should show a random scattering of values around zero, with no obvious systematic patterns across the range of fitted values as is the case here. Thus, there is no evidence of heterogeneity of the variances. Bartlett's or Levene's test (*Statistics*, *variances*) can also be used for the same purpose although they tend to be over sensitive (see below). The most common deviation is when large things vary more than small things. In that case the scattering of points on the left of the plot would be less than on the right (which is not the case here). In extreme cases a transformation of scale may be necessary.

The Q–Q plot (second top) should show a straight line if the residuals have a normal, bell-shaped distribution. In this case there is a slight S-shaped line. Apparently, there is an excess of animals which are poor learners, and an excess of animals which are fast learners, with relatively few animals which are intermediate learners. Thus there are slightly heavy tails to the normal distribution. Although these plots show some deviation from the required assumptions for a valid ANOVA, the deviation is probably not sufficient to invalidate the conclusions. The ANOVA is quite robust against deviations from the assumptions. An example where the assumptions are clearly not met and a scale transformation is needed is discussed in Chapter 6. The interpretation of these diagnostic plots requires some experience.

A criticism of this experiment

The data in this example are fictitious but they serve to show how such an experiment can be summarised (with means, SDs and confidence intervals) and analysed using an ANOVA.

An experiment needs to take account of the biology of the species, the logistics of the way that it is to be conducted, and whether it makes economical use of the available resources.

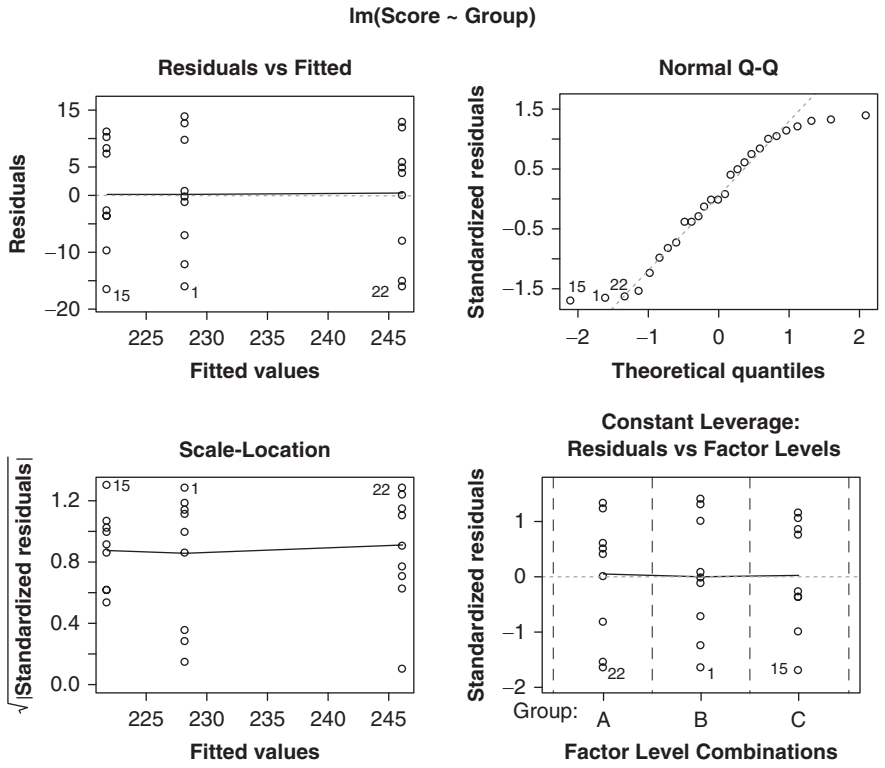


Figure 5.5 Residuals model diagnostic plot of the learning data. Note that in the first plot there is a good scattering of points in all four corners, so there is no evidence of heterogeneity of variance. The slight S-shaped distribution in the second plot indicates some deviation from a normal distribution of the residuals, as is also shown in the box plots.

There are two possible problems with this experiment. First, it may not make good use of the experimental material, and second there could be logistical difficulties in carrying it out. The mice need to be individually housed, which will have welfare consequences. Will the exercise wheels fit in a standard cage? Will the wheels need a counting mechanism to assess how much they are used? There is wide inter-individual and strain variation in the use of exercise wheels (Festing, 1976). Is it going to be expensive to have separate wheels for 27 individual cages? Then there is the logistics of putting the wheels into the cages and taking them out again after 30 min or 3 h. How easy will that be with 27 cages over a period of three weeks? Another problem is that mice are nocturnal. Is it reasonable to put the wheels into the cages during the day in their sleep period? Would it be more sensible to change the lighting so that it comes on at, say, 04:00 h and goes off at 13:00 h. At that point the room could be lit by dim red light so that staff could see where to put in and remove the exercise wheels.

Assessing learning ability in a maze may also introduce logistical problems. How easy is it to assess learning ability in 27 mice? If each mouse took one-quarter of

an hour to assess, that would require nearly 7 h. Would it all be done in one day? If so, will there be biological rhythms that affect mice differently in the morning and afternoon? Would the work need to be staggered in some way? If so, how? Given that mice are nocturnal, will they be tested in the daytime?

All these factors suggest that it might be better to split the experiment up as a Randomised Block design (see Chapter 7) over a period of time. True, the experiment would take more time to complete, but it would provide better control of the variability and would therefore be more powerful, and some idea of repeatability could be obtained from the individual blocks. It would also need fewer cages and exercise wheels, which might save some money.

Moreover, the experiment does not make efficient use of the animals. Methods of determining sample size are discussed in Chapter 11. The 'resource equation' method states that an experiment analysed using an ANOVA should usually have between 10 and 20 degrees of freedom (DF) for the error term. Fewer than 10 DF implies that the experiment will lack power and more than 20 implies either that the experiment may be excessively large or it is asking too few questions. This experiment is quite large with 24 DF for error. There is scope to ask more questions. For example, it could include both males and females at no extra cost using a factorial experimental design, the topic of the next chapter.

In conclusion, careful thought would need to be given to the logistics of this experiment and to whether it would have been possible to obtain more information, by including both sexes or more than one strain of mice.

6

Factorial experiments

Introduction

Factorial experiments provide one of the best ways of implementing 'Reduction' as defined by (Russell and Burch, 1959). According to Fisher (1960):

> [By using a factorial design] … an experimental investigation, at the same time as it is made more comprehensive, may also be made more efficient, if by more efficient we mean that more knowledge and a higher degree of precision are obtainable by the same number of observations.

A 'factor' is a discrete variable used for classification purposes such as gender, treatment, strain, age (young or old), etc. A factorial experiment is one which has at least two fixed effect factors, each of which can have any number of levels. For example gender could be a factor with levels male and female; strain could be a factor with levels Strain A, B, C, etc.; and age could be a factor with levels young and old. Factors can also be numerical with, for example, dose levels 0, 5, 10 mg/kg.

The aims of a factorial experiment are usually:

1. To investigate the effect of each treatment separately on the dependent variable.
2. To assess whether there are interactions between the factors, i.e. whether the magnitude of response to one factor depends on the level of another factor.

Strictly, factorial designs are an arrangement of treatments rather than a 'study design', but they are commonly called 'designs'. True designs are mutually exclusive. It is not possible to have a completely randomised (CR) design which is also a randomised block (RB) design. However, CR designs, RB designs and several other designs can have factorial arrangements of treatments.

The fictitious experiment discussed in Chapter 5 to determine the effect of wheel running on learning ability in mice was technically correct and provided a valid introduction to CR designs, but it was open to criticism on two grounds.

1. It made inefficient use of animals (and other resources) because it only involved a single fixed effect factor, the three levels of wheel running.
2. The logistics of assessing learning ability in 27 mice in a short period (to avoid too much variation due to circadian rhythms) could be a problem, leading to increased inter-individual variation and a reduction in statistical power.

This chapter deals with the first and Chapter 7 addresses the second of these two criticisms. The experiment involved three treatment groups (None, Moderate and Marathon running) with nine mice per group. An experiment of this size could easily incorporate both sexes. For example, instead of using nine animals of the same sex in each group it could have used four males and four females in each group (one less per group in order to have a balanced design). Treatment differences would be estimated from the mean across both sexes, with a total of eight animals per treatment. Gender differences in learning ability would be assessed by averaging across treatments, and any differences in response between the two sexes could also be assessed.

Example 6.1: Statistical analysis of a 2 × 2 factorial design using R-Commander

Table 6.1 shows the red blood cell (RBC) count ($\times 10^{12}$/L) in two mouse strains (BALB/c and C57BL/6) when administered chloramphenicol at two dose levels (vehicle and 2000 μg/kg). This is a 2 (strains) × 2 (dose levels) factorial experiment with four mice in each of the four treatment groups. It is real data extracted from a larger CR factorial experiment (Festing et al., 2001) used here to demonstrate the statistical analysis using R-Commander (Rcmdr) in a situation where there is no interaction. The data are in random order, as if it had been collected in a stand-alone experiment. The objectives of the experiment were:

1. To determine the effect of chloramphenicol at this dose level on RBC counts in mice, averaged across both strains.
2. To determine whether the two strains differ in RBC counts (averaged across chloramphenicol treatments).
3. To determine whether the two strains respond in a similar way to chloramphenicol.

Statistical analysis

To import the data into Rcmdr and prepare it for the analysis, the data (Table 6.1) are copied to EXCEL then to the clipboard and read into Rcmdr (*Data*, *import data*, *from textfile*, *clipboard*, *click clipboard*). The columns show the animal ID numbers, treatment, strain, RBC count and groups. The dose levels are given as 0 and 2000. Rcmdr needs to be told that these are factors, not variables. This is done by using the menu commands *Data*, *Manage variables in the data set*, *Convert numerical variables to factors*. Choose *Dose*, click '*Use numbers*' and *OK*.

Screening the raw data

The raw data are screened visually for obvious errors using a stripchart (Figure 6.1) as explained in Chapter 5. Any obvious outliers should be checked to make sure that they are not errors, but they should not be altered unless there is clear and unambiguous evidence of a mistake.

Table 6.1 Red blood cell count (RBC) in a 2 (treatments) × 2 (strains) factorial experiment.

Animal	Treatment	Strain	RBC	Group
1	B	BALB/c	8.45	C
2	B	C57BL/6	8.82	D
3	A	C57BL/6	9.20	B
4	B	BALB/c	8.89	C
5	A	BALB/c	10.08	A
6	A	C57BL/6	9.60	B
7	B	BALB/c	8.68	C
8	B	C57BL/6	8.24	D
9	A	BALB/c	10.09	A
10	A	BALB/c	10.10	A
11	B	C57BL/6	8.18	D
12	B	BALB/c	8.95	C
13	A	C57BL/6	9.14	B
14	B	C57BL/6	8.10	D
15	A	BALB/c	9.73	A
16	A	C57BL/6	9.56	B

Treatments have been coded A = control, B = chloramphenicol treated.

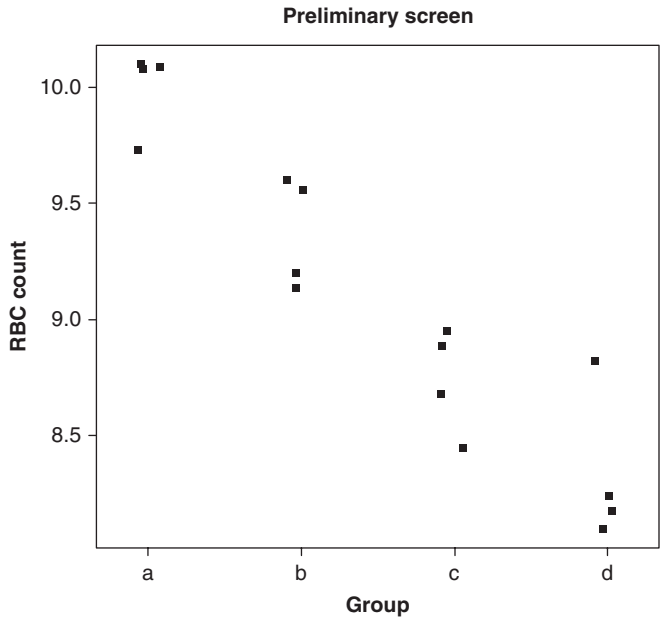


Figure 6.1 Stripchart for preliminary screening by group. The Jitter option separates the points on the X-axis. Any outliers should be checked to make sure that they are correct. Group d has one slight outlier which could be checked to ensure that it is not an error (it is not). RBC: red blood cell.

Carry out an ANOVA with residual plots to check assumptions

The Rcmdr commands *Statistics*, *Fit model*, *Linear model* open up an input box. The RBC count is the dependent variable. This is double-clicked to put it in the first box. The tilde sign ~ means 'depends on', so the Treatment, *, and Strain are clicked. The asterisk implies factor multiplication (a 2 × 2 design). The Return is clicked and Rcmdr gives some results comparing the first group (control, BALB/c) with all the other groups (see Table 6.2). This table is slightly complicated but it is useful for more complex experiments such as this one.

The linear model output is displayed and some details of the residuals are given (Min, IQ, Median, etc., these can be ignored at this stage). The first column of the table shows the mean of the first group (BALB/c, controls) which is 10.000 units. The effect of chloramphenicol is to reduce this by a statistically significant 1.2575 units. Changing the strain to C57BL/6 reduces the RBC count by a statistically significant 0.625 units. The interaction term increases the value by 0.2175 units which is not statistically significant. The residual standard error is 0.2493. This is the same as is obtained by taking the square root of the error mean square in an ANOVA table. The multiple R-squared of 0.8955 confirms that fitting this linear mathematical model accounts for almost 90% of the total variation.

The ANOVA table is obtained from *Models*, *Tests of hypotheses*, *ANOVA table* and click 'Type I'. This provides the ANOVA table (see Table 6.3) in the usual form

Table 6.2 Output by R-Commander (Rcmdr) from fitting a linear model (LM) analysis to a 2 × 2 factorial experiment.

Call:
LM (formula = RBC ~ Treatment * Strain, data = Dataset)

Residuals:

Min	1Q	Median	3Q	Max
–0.29250	–0.19000	0.00875	0.15688	0.48500

Coefficients:

	Estimate	Std error	*t*-value	*P* (>|*t*|)
(Intercept)	10.0000	0.1247	80.221	< 2e-16***
Treatment [T.B]	–1.2575	0.1763	–7.133	1.19e-05***
Strain [T.C57BL/6]	–0.6250	0.1763	–3.545	0.00403**
Treatment [T.B]: Strain [T.C57BL/6]	0.2175	0.2493	0.872	0.40011

Significant codes: ***p<0.001, **p<0.01, *p<0.05

Residual standard error: 0.2493 on 12 degrees of freedom (DF).

Multiple R-squared: 0.8955, adjusted R-squared: 0.8694.

F-statistic: 34.28 on 3 and 12 DF, P-value: 3.643e-06.

1Q and 3Q represent the first and third interquartile range of the numbers. The adjusted R-squared is the proportion of the total variation accounted for by fitting the mathematical model.

Table 6.3 Analysis of variance table. Response is red blood cell (RBC) counts.

Source[1]	DF	SS	MS	*F*-value	*P* (>F)
Strain	1	1.0661	1.0661	17.1512	0.001367**
Treatment	1	5.2785	5.2785	84.9232	8.60e-07***
Strain:Treatment	1	0.0473	0.0473	0.7611	0.400108
Residuals	12	0.7459	0.0622		
Total[1]	*15*				

[1]R-Commander (Rcmdr) does not put in the heading 'Source' or the 'Total' shown here in italics, although they are usual in most other statistical packages. Rcmdr uses the convention that 1, 2 or 3 asterisks are shown to indicate significance levels of $p<0.05$, $p<0.01$ and $p<.001$

The 8.60E-07 means 8.6 with 7 zeros before the 8. DF: degrees of freedom, SS: sum of squared deviations, MS: mean squares.

with heading Source not given in Rcmdr, followed by DF, SS, etc. It provides similar information to Table 6.2, but in a more familiar style. There are highly significant strain and treatment effects but the Strain * Treatment *P*-value is only 0.4. Therefore, although the strains differ in RBC counts and the counts are reduced by the chloramphenicol, there is no evidence that the response to chloramphenicol at this dose depends on the strain of mice.

Calculate the means

The treatment means averaged across strains and the strain means averaged across treatments can be calculated using *Statistics*, *Summaries*, *Numerical summaries* (see results in Table 6.4). When presenting the means it is better to use the pooled standard deviation (SD). This is easily justified because the statistical analysis assumes that the variation is the same in each group, and this is checked as part of the statistical analysis.

Residual diagnostic plots

The residual diagnostic plots discussed in Chapter 5 can be obtained in Rcmdr from *Models*, *Graphs*, *Basic diagnostic plots* (Figure 6.2). The first plot of Residuals versus Fitted should have a scattering of points with no obvious pattern. Animal No. 2

Table 6.4 Treatment and strain means.

Strain	Control	Treated	Strain means
BALB/c	10.00	8.74	9.37
C57BL	9.38	8.34	8.86
Treatment means	9.69	8.54	

Pooled standard deviation (SD) = 0.25. Both strain and treatment means are significantly different ($P < 0.01$) but the interaction is not significant.

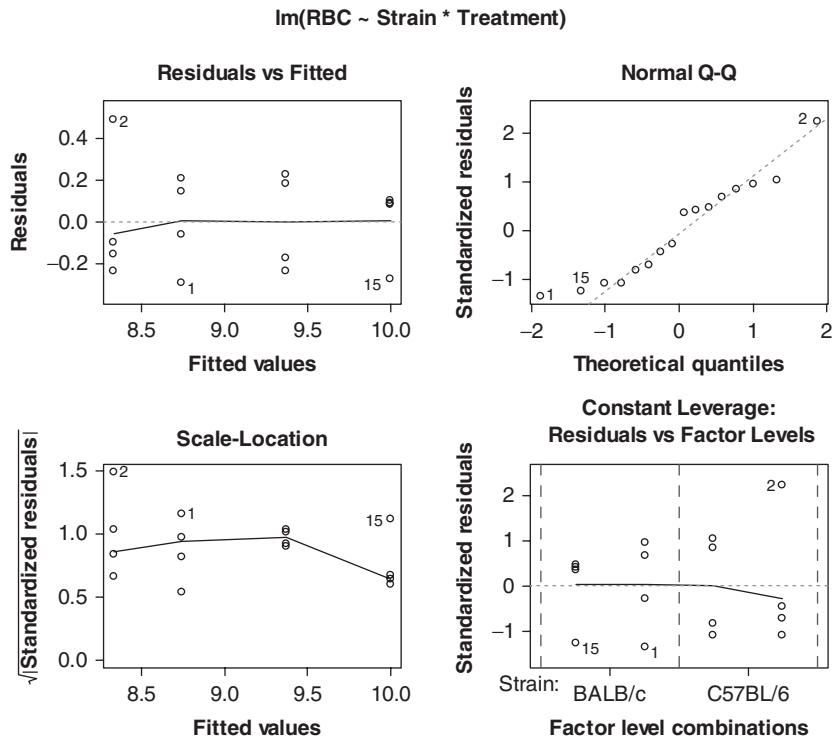


Figure 6.2 Residual diagnostic plots for the first chloramphenicol example. Only the top two plots are discussed. Note the good scattering of points in the first subplot and the reasonable straight line fit in the second plot. Animal 2, which is the slight outlier noted in Figure 6.1 is also identified in these plots. Experience is required to decide on how well the data fit the assumptions of normality of residuals and homogeneity of variance. Here the fit is good.

is identified as a slight outlier but there is no evidence of heterogeneity of variances. The Normal Q–Q plot should show approximately a straight line (as in Figure 6.2) if the residuals have a normal distribution. The remaining two plots offer a more detailed analysis of the residuals, but they are not discussed here. The interpretation of these plots needs some experience as it is a matter of judgement of whether the points are well scattered and the line is reasonably straight. The ANOVA is quite robust against deviations from the assumptions of homogeneity of variance and normality of the residuals.

A plot of means

A plot of the means (Figure 6.3) can be obtained from *Graphs*, *plot of means*. Choose both Treatment and Strain. Options can be used to set the labels of the axes and error bars. Note that the two lines are parallel, which indicates that the two strains are responding in the same way and there is no interaction between strain and treatment.

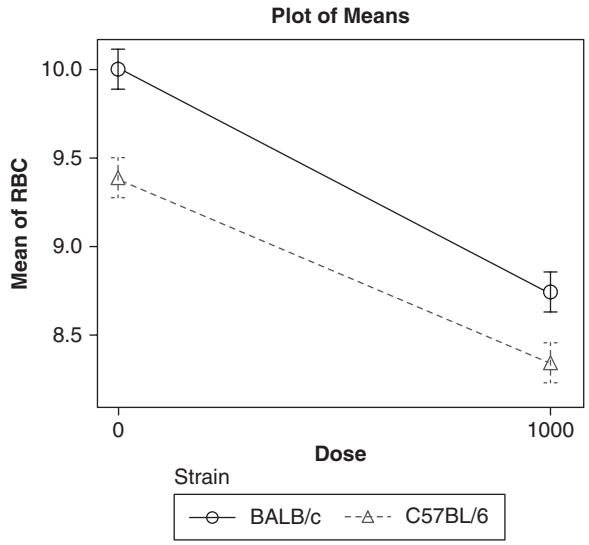


Figure 6.3 Means and standard errors of red blood cell (RBC) counts. Note that chloramphenicol at the 1000 dose level reduces RBC count in both strains equally (the lines are parallel).

Conclusion

This experiment had four treatment groups so it would have been possible to analyse it by a one-way analysis of variance (ANOVA) using post hoc comparisons. However each group would only consist of four experimental units and such an analysis would not provide a test of whether the two strains responded similarly to the effect of the chloramphenicol (the interaction). So it would have been a much less useful way of analysing the data.

The overall conclusion is that chloramphenicol administered at 2000 μg/kg reduced RBC count from 9.69 to 8.40 units averaged across strains (SD = 0.25, $P < 0.001$), and C57BL/6 mice had a lower RBC count (8.86) than BALB/c mice (9.37) (SD = 0.25, $P = 0.001$) averaged across treatments (all units $\times 10^{12}$/L). There is no statistically significant strain × treatment interaction so there is no evidence that the two strains responded differently to chloramphenicol.

Note that by using two instead of a single strain the experiment has supplied extra information at no extra cost because if a single strain had been used, about the same total number of animals would have been needed.

Example 6.2: A 2 (strains) × 3 (doses) factorial experiment

The data in this experiment have been taken from the same large experiment as Example 6.1. It illustrates the statistical analysis of a factorial design where there is significant interaction. The raw data are given in Table 6.5 (note the random order).

Table 6.5 Raw data to illustrate the statistical analysis of a 2 (strains) × 3 (doses) completely randomised (CR) factorial experiment.

Animal	Strain	Dose	RBC	Group	Animal	Strain	Dose	RBC	Group
1	CD-1	1000	8.11	g5	13	CD-1	0	9.01	g4
2	CD-1	1000	9.19	g5	14	BALB/c	0	10.09	g1
3	CD-1	0	8.27	g4	15	BALB/c	2000	8.95	g3
4	BALB/c	0	10.08	g1	16	CD-1	2000	8.31	g6
5	CD-1	2000	9.07	g6	17	BALB/c	2000	8.45	g3
6	CD-1	1000	9.09	g5	18	CD-1	1000	9.40	g5
7	CD-1	2000	9.51	g6	19	CD-1	2000	9.18	g6
8	BALB/c	2000	8.68	g3	20	BALB/c	1000	10.06	g2
9	BALB/c	2000	8.89	g3	21	CD-1	0	9.10	g4
10	BALB/c	1000	9.99	g2	22	BALB/c	1000	9.38	g2
11	CD-1	0	7.76	g4	23	BALB/c	0	10.10	g1
12	BALB/c	0	9.73	g1	24	BALB/c	1000	9.91	g2

For details see text. RBC: red blood cell count.

In this experiment there were two strains, BALB/c and CD-1 and three doses 0, 1000 and 2000 mg/kg, with red blood count (RBC) ($\times 10^{12}$/L) being the dependent variable. The data are read into Rcmdr in the same way as in the previous example. Note that dose is a factor not a variable and Rcmdr needs to be given this information, as is explained in the previous example (*Data*, *Manage variables*…, *Convert*…). A stripchart should be used to give a preliminary screen of the data (use Group as the independent variable). The data are analysed using a two-way ANOVA as previously described and the residual diagnostic plots should be studied to see if the assumptions are reasonably well met (not shown).

Table 6.6 shows the ANOVA for this experiment. Note that there is a highly significant Dose * Strain interaction ($P = 0.003$) and a significant strain difference ($P = 0.001$) but the dose effect is not statistically significant ($P = 0.077$). Figure 6.4 shows clearly what is happening. BALB/c control mice have high RBC counts which are sharply reduced by chloramphenicol whereas almost the opposite is true with CD-1. As a result there is little change in the mean across both strains when the dose is increased. When interactions are present the means of each group should be examined and presented separately. The residual diagnostic plots in this case give no cause for concern (not shown) and with 18 degrees of freedom (DF) for the error term this experiment is about the right size, using the resource equation method of determining sample size, as is explained in Chapter 11.

2^n factorial experiments

Factorial designs with several factors each at two levels provide an economical way of investigating factors which may affect the outcome of an experiment. A 2^3 ($2 \times 2 \times 2$) factorial design will have three factors designated here for convenience A, B and

Table 6.6 Analysis of variance for Example 6.2, a 2 (strains) x 3 (doses) factorial experiment. Response: red blood cell count.

	DF	SS	MS	*F*-value	*P* (>F)
Dose	2	1.1383	0.56913	2.9168	0.079939
Strain	1	2.8773	2.87734	14.7462	0.001199**
Dose:Strain	2	3.1417	1.57084	8.0505	0.003181**
Residuals	18	3.5122	0.19512		

DF: degrees of freedom, SS: sum of squared deviations, MS: mean squares.

Rcmdr uses the convention that 1, 2 or 3 asterisks are shown to indicate significance levels of $p<0.05$, $p<0.01$ and $p<.001$

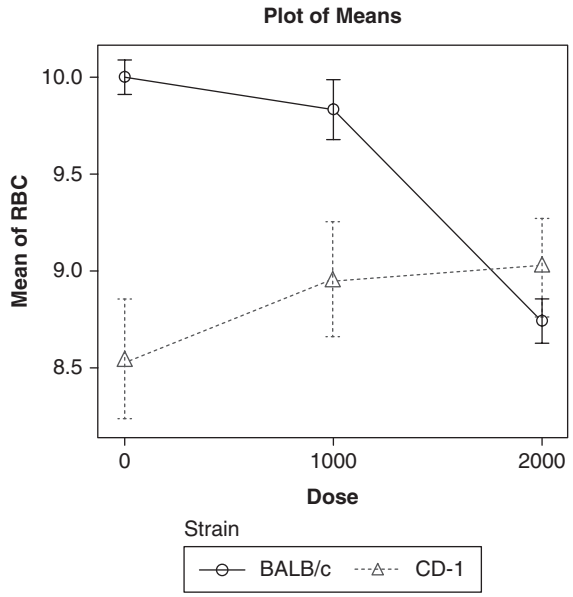


Figure 6.4 Plot of means with standard deviations for a 2 (strains) × 3 (doses) factorial design. Note the strong strain × dose interaction. BALB/c control mice have a high red blood cell (RBC) count which is reduced by chloramphenicol whereas CD-1 control mice have a low count which increases marginally (though probably not significantly) with chloramphenicol.

C, each at two levels giving a total of eight groups. There will be three main effects (A, B, C), three two-way interactions (A*B, A*C, B*C) and one three-way interaction (A*B*C). If it has a three-fold replication it will need $3 \times 8 = 24$ experimental units and each main effect mean will be based on 12 animals per group.

A 2^4 factorial design will have four factors each at two levels. With two-fold replication it will have 16 groups and 32 animals (or other experimental units). Each main effect will be based on a comparison of two groups each of 16 animals. Interactions of 2, 3, and 4 factors can be assessed.

A 2^5 factorial design will have five factors each at two levels. It will have 32 treatment groups, and with two-fold replication per group will require 64 animals. In the absence of interaction each main effect will be based on 32 animals. Interactions of 2, 3, 4 and 5 factors can be assessed.

If even more factors are to be included, the experiment could become excessively large and difficult to handle. In order to reduce the size of the experiment, high order interactions (3–5 way) can be used as the error term because they are rarely statistically significant, in which case the experiment can be done as a single replication, although in this case missing observations may have an exaggerated influence on the results. Even more factors can be included without the experiment becoming excessively large using fractional factorial designs and a carefully chosen 'confounding' (i.e. mixing) of interactions that are unlikely to be statistically significant. More details can be found in major statistical textbooks, e.g. Montgomery, 1997.

Example 6.3: A 2^4 factorial design

The data in Table 6.7 comes from a 2^4 ($2 \times 2 \times 2 \times 2$) factorial experiment. The aim was to explore the effects of several factors on a mouse model for assessing whether antioxidants might protect against cancer. Diallyl sulphide, a chemical found in garlic, or the vehicle was administered by gavage at 0.2 mg/g body weight for three days prior to and three days following treatment with a carcinogen. Carcinogens used were either urethane or 3-methylcholanthrene (3MC) given by intraperitoneal injection. Half of the mice were males and half females, and half of the animals were strain A/J and half were NIH/Ola. The mice were kept for five months, then humanely sacrificed and the number of adenomas on the surface of the lungs was counted as a measure of tumour susceptibility. There were two animals per group, although one animal (an NIH male given 3MC and the vehicle) was missing in group 6, therefore only data on 31 mice are given.

The data are read into Rcmdr as explained previously. These data are counts of tumour numbers rather than measurements. Counts where the mean is low often have Poisson distributions in which the mean and variance are equal, leading to heterogeneous variances. So the first step is to see whether the data need to be transformed to another scale. A trial ANOVA to produce residual diagnostic plots (see Figure 6.5) clearly shows that the basic assumptions of homogeneity of variances and a normal distribution of the residuals are *not* met with this set of data. In the first plot most of the low values in the left of the plot are close to the mid-line whereas the higher values are more scattered. In the Normal Q–Q plot the points do not fall in a straight line. Clearly a scale transformation is needed.

The transformation of scale was discussed in Chapter 4. Where the raw data are not suitable for a statistical analysis using the ANOVA it is often possible to transform the data to a different scale. As the data are counts, and the count is low (compared with, for example, RBC counts which do not need a transformation) it is known that such data often have a Poisson distribution requiring a square root transformation.

Table 6.7 Raw data from a 2^4 factorial experiment studying the effect of strain, sex, carcinogen (Carc) and antioxidant (Antiox) on lung tumours in mice.

Animal	Strain	Sex	Carc	Antiox	Tumours	Group
1	A/J	Male	Urethane	Vehicle	24	1
2	A/J	Male	Urethane	Vehicle	30	1
3	NIH	Male	Urethane	Vehicle	16	2
4	NIH	Male	Urethane	Vehicle	16	2
5	A/J	Male	Urethane	DS	4	3
6	A/J	Male	Urethane	DS	4	3
7	NIH	Male	Urethane	DS	3	4
8	NIH	Male	Urethane	DS	3	4
9	A/J	Male	3MC	Vehicle	51	5
10	A/J	Male	3MC	Vehicle	24	5
11	NIH	Male	3MC	Vehicle	2	6
12	A/J	Male	3MC	DS	37	7
13	A/J	Male	3MC	DS	52	7
14	NIH	Male	3MC	DS	0	8
15	NIH	Male	3MC	DS	4	8
16	A/J	Female	Urethane	Vehicle	19	9
17	A/J	Female	Urethane	Vehicle	23	9
18	NIH	Female	Urethane	Vehicle	4	10
19	NIH	Female	Urethane	Vehicle	13	10
20	A/J	Female	Urethane	DS	7	11
21	A/J	Female	Urethane	DS	4	11
22	NIH	Female	Urethane	DS	3	12
23	NIH	Female	Urethane	DS	1	12
24	A/J	Female	3MC	Vehicle	66	13
25	A/J	Female	3MC	Vehicle	45	13
26	NIH	Female	3MC	Vehicle	2	14
27	NIH	Female	3MC	Vehicle	4	14
28	A/J	Female	3MC	DS	37	15
29	A/J	Female	3MC	DS	59	15
30	NIH	Female	3MC	DS	0	16
31	NIH	Female	3MC	DS	29	16

3MC: 3-methylcholanthrene, DS: diallyl sulphide.

Accordingly, a new variable 'Roottums' was calculated. In Rcmdr this is done using *Data*, *Manage variables in the active data set*, *Compute new variable*. The new variable Roottums and the command to produce it (Sqrt (Tumours)) are then given.

A second trial ANOVA of Roottums is used to check whether the transformed data can be safely analysed using an ANOVA. This is shown in Figure 6.6. There

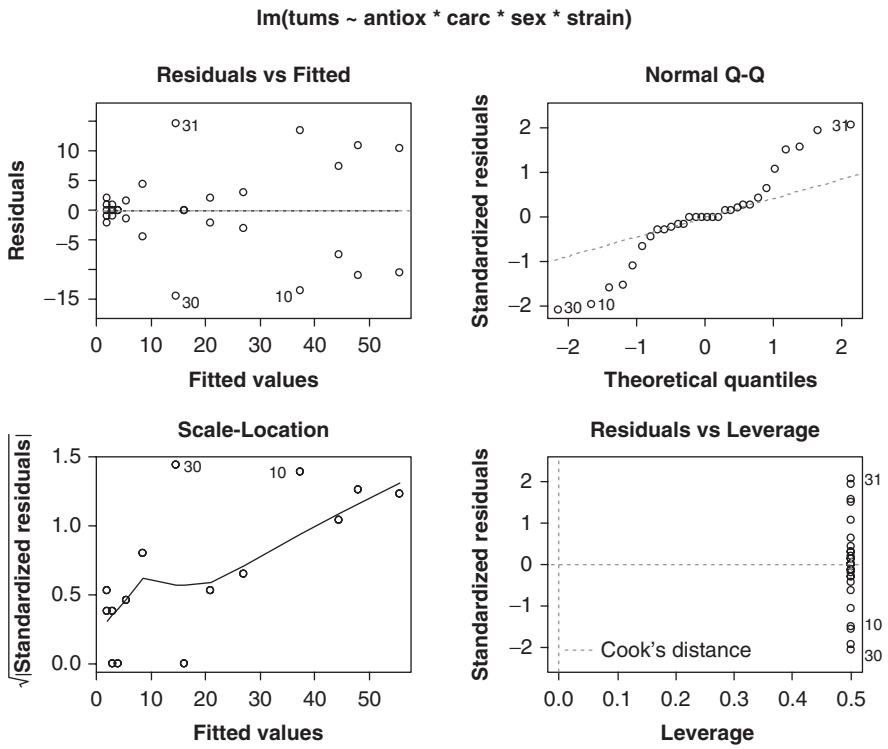


Figure 6.5 Trial residual diagnostic plots. Note the cluster of points around zero in the top left-hand plot with low fitted values. This shows extreme heterogeneity of variance. The second plot also shows that the residuals plot deviates seriously from a straight line. A transformation of scale is clearly needed. The remaining two plots offer a more detailed analysis of the residuals, but they are not discussed here.

is a good scatter of points in the first plot and a good straight line in the second (Normal Q–Q) plot showing normality of the residuals. However, there are two serious outliers labelled in the figure as points 30 and 31. What should be done about these? The answer is that they should be checked to see if there is any evidence of a mistake. In this case the original records were checked and there is no evidence that a mistake has been made. A tumour count of 29 in an NIH mouse is excessively high. If tissue is available it is possible, when using inbred strains, to check whether this really is an NIH rather than an A/J mouse. However, this experiment was done before DNA genetic markers were widely available and no tissue is available. One strategy used in such cases is to perform the statistical analysis with and without the outliers. If it makes no difference to the conclusions, then the outliers can be kept. If the conclusions depend on the outliers, then the experiment probably has to be repeated. In this case the conclusions are not changed when the outliers are deleted, so they have been kept, but they will have inflated the error in the ANOVA.

The full ANOVA is shown in Table 6.8 and the main effect means and the pooled standard deviation are shown in Table 6.9a. There is a statistically significant effect

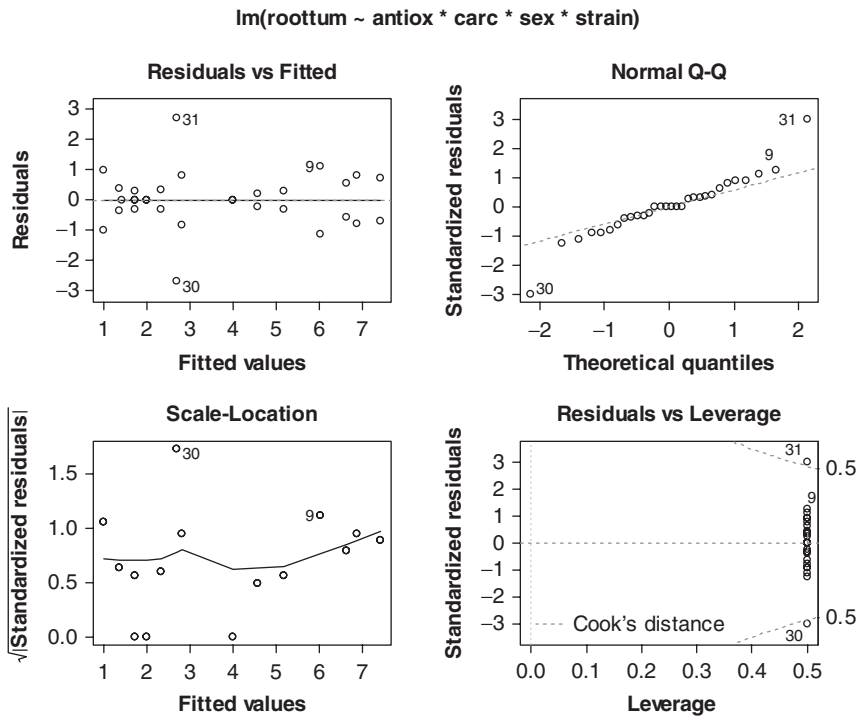


Figure 6.6 Residual diagnostic plots following a square root transformation of the data. Note the good scattering of points in the first plot. However, there are two serious outliers labelled 30 and 31. The second top plot shows a good straight line, implying that the residuals now have a normal distribution, apart from the two outliers. The remaining two plots offer a more detailed analysis of the residuals, but they are not discussed here.

of the antioxidant with a mean Roottums of 3.08 in the antioxidant group versus 4.32 in the saline group. Therefore, the antioxidant appears, on average, to be reducing the number of tumours. The pooled SD from the error mean square is 1.27 on this scale. Means for the other main effects as well as two-way tables can be produced using *Statistics*, *Summaries*, *Tables of statistics*. The carcinogen effect is significant with the 3MC treatment producing 4.41 Roottums compared with 3.00 for the urethane, but this is an effect of the carcinogen dose so is not of direct interest. There was a very large strain effect with A/J getting 5.13 and NIH 2.13 Roottums, but there was no significant difference between the sexes ($P = 0.983$).

There are two significant two-way interactions. First, the two strains differ in their relative sensitivity to the two carcinogens, with 3MC inducing many more tumours in A/J than in NIH mice (Table 6.9b). The strain difference when the mice were treated with urethane was much less. Second, the antioxidant substantially reduced the tumour count when the carcinogen was urethane (from 4.1 to 1.9 on the square root scale), but it failed to significantly reduce the tumour count when the carcinogen was 3MC. This is shown in Figure 6.7. It seems that the benefit of

Table 6.8 Analysis of variance table for the square root of the tumour count. Response: Roottums.

	DF	SS	MS	*F*-value	*P* (>F)
Antioxidant (Antiox)	1	11.946	11.946	7.4262	0.0156502*
Carcinogen (Carc)	1	16.349	16.349	10.1628	0.0061133**
Sex	1	0.001	0.001	0.0005	0.9829563
Strain	1	65.591	65.591	40.7734	1.225e-05***
Antiox:Carc	1	10.051	10.051	6.2483	0.0245224*
Antiox:Sex	1	0.931	0.931	0.5790	0.4585113
Carc:Sex	1	2.861	2.861	1.7784	0.2022397
Antiox:Strain	1	0.347	0.347	0.2159	0.6488533
Carc:Strain	1	31.194	31.194	19.3911	0.0005133***
Sex:Strain	1	0.070	0.070	0.0437	0.8372813
Antiox:Carc:Sex	1	0.370	0.370	0.2302	0.6383281
Antiox:Carc:Strain	1	0.137	0.137	0.0854	0.7741417
Antiox:Sex:Strain	1	0.683	0.683	0.4243	0.5246580
Carc:Sex:Strain	1	0.377	0.377	0.2342	0.6353883
Antiox:Carc:Sex:Strain	1	0.834	0.834	0.5186	0.4825002
Residuals	15	24.130	1.609		

DF: degrees of freedom, SS: sum of squared deviations, MS: mean squares.

Rcmdr uses the convention that 1, 2 or 3 asterisks are shown to indicate significance levels of $p<0.05$, $p<0.01$ and $p<.001$

Table 6.9a Main effect means and *n*.

Treatment	Mean	*n*
DS	3.08[a]	16
Vehicle	4.32	15
3MC	4.41[b]	15
Urethane	3.00	16
Female	3.72	16
Male	3.64	15
A/J	5.13	16
NIH	2.13	15

Pooled standard deviation = 1.27. [a]$P = 0.02$, [b]$P = 0.01$. 3MC: 3-methylcholanthrene, DS: diallyl sulphide.

the antioxidant depended on the nature of the carcinogen. More work would be needed to determine how general a reduction in tumours would be in relation to different carcinogens.

Table 6.9b The strain by carcinogen means.

Treatment	Strain	
	A/J	NIH
3MC	6.74	1.74[a]
Urethane	3.52	2.48

Pooled standard deviation = 1.27. Interaction, $P < 0.001$. [a]$n = 3$ in this group but 4 in the other groups. 3MC: 3-methylcholanthrene.

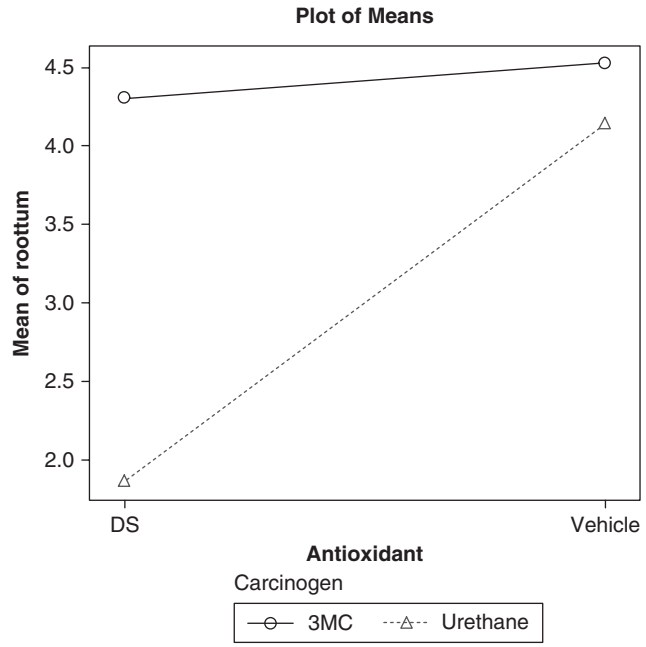


Figure 6.7 A plot of the carcinogen × antioxidant interaction. Note that diallyl sulphide (DS) reduces the root tumour count when urethane is the carcinogen, but hardly does so when 3-methylcholanthrene (3MC) is the carcinogen. This interaction is statistically significant ($P = 0.025$).

Conclusions

This chapter has explained the advantages of using factorial experiments. The experiment to determine the effect of wheel activity on learning in mice, introduced in Chapter 5, used three groups of nine mice in each group. It would probably have been better to use, say, eight mice per group in a factorial design with half being male and half being female. In this way information in the responses of both sexes could have been obtained at no extra cost. This makes the assumption that there is

no difference in variability between the two sexes, and this would be studied by looking for heterogeneity of variance using the residual diagnostic plots. However, a meta-analysis of 293 papers involving both sexes with the stage of the oestrus cycle in the females being uncontrolled gave no evidence for greater variability in females than in males (Prendergast et al., 2014).

Most clinical trials are already logistically quite complex, so factorial designs, which add some complexity, are not widely used. But experiments involving laboratory or farm animals are logistically relatively simple, and offer complete control of the animals, so factorial designs should always be considered. They provide extra information at little or no additional cost. The National Institutes of Health in the USA requires investigators there to use both sexes in their experiments and this can usually be done without increasing the total number of animals if factorial designs are used.

7

Randomised block designs

Introduction

Compared with completely randomised (CR) designs, randomised block (RB) designs are usually more powerful, logistically more convenient, make better use of heterogeneous subjects and environments, and can often take account of the natural structure in experimental subjects. The individual blocks may also provide an internal check on the repeatability of an experiment. These advantages as so compelling that RB designs should replace the CR designs as the default design in experiments involving laboratory animals.

In an RB design the experiment is split up into a number of mini-experiments or 'blocks'. A block usually has a single experimental unit of each treatment. So if there are four treatments each block will consist of four experimental units, each assigned to a different treatment. Each block is separately randomised (see Chapter 3).

Blocks can be separated in time so that the experiment is spread over hours, days or weeks at the convenience of the investigators. They can also be separated in space. So each block can be on a shelf or in a different room, or even in different laboratories. Blocks can often take account of some natural structure in the experimental material. For example, animals from a single litter may be used as a block.

An RB design is sometimes used in *in vitro* studies but the investigators may not realise this, so they may analyse it incorrectly. For example, an investigator may state that 'we repeated the experiment three times' without realising that each replication is a block in an RB design. Then they do not carry out the appropriate statistical analysis (i.e. a two-way analysis of variance without interaction).

The characteristics of RB designs

These designs are distinguished from CR designs in that they have a single random effect factor (the block) in addition to one or more fixed effect factors. For example, 'litter' may be a random effect factor (the block) when performing a within-litter experiment in pre-weaned animals. Several litters would be involved and these could be used at different times. Time alone can also be treated as a factor

when the experiment is split up over a period of hours, days or weeks because there are uncontrollable and/or unobservable rhythms which may influence the experimental units.

There are a number of named designs which, mathematically, are equivalent to RB designs, i.e. they have a random effect factor and one or more fixed effect factors. These include:

1. Crossover designs: In these designs, an animal or other subject is given a set of treatments sequentially in random order. In this case, the experimental unit is an animal for a period of time and the block is the animal.
2. Matched designs: In these designs, *n* experimental units are matched, where *n* is the number of treatments, and then one of these is assigned to each treatment at random. This is repeated *r* times, where *r* is the required sample size.
3. Within-subject designs: In these designs, an animal or other subject can provide several experimental units, such as two eyes or several skin patches, which can receive different treatments. The block is then the animal and the experimental unit is, say, an eye or a patch of skin.
4. Latin square designs: These belong to the same family of designs. They have two random effect factors (often known as rows and columns) as well as one or more fixed effect factors such as 'treatment'.

Advantages of RB designs

These designs are often more convenient than CR designs because they allow the experiment to be split up in time or space, making it easier to handle. Blocking on time is particularly useful for large experiments where it may be difficult to carry out all the procedures and measurement of outcomes within a limited time.

RB designs usually increase the power of an experiment. By matching subjects which receive different treatments they decrease the error variation and increase the power of the experiment (the concept of statistical power is introduced in Chapter 9). Even quite heterogeneous animals (or other experimental units) can be used in RB designs if they can be matched on age, weight, source or other important characteristics.

RB designs can take account of any natural structure of the experimental material. Litters of mice and rats are an obvious example where each litter could be a separate block.

RB designs can increase the repeatability of an experiment. If the blocking factor is time, then they will help to ensure that experiments are repeatable in time. If they give a different result each time, then no statistically significant treatment effects will be detected.

By sampling from a slightly different environment in each block an RB design also slightly increases external validity, i.e. the extent to which the results can be generalised.

Disadvantages of RB designs

The main disadvantages are that a very small experiment with a small or non-existent block effect will lack power, but this is a relatively rare situation. RB designs are also less tolerant of several missing observations than are CR designs.

Statistical analysis of an RB design

In an RB design with a single treatment factor each observation is defined by the block and treatment. The statistical analysis is done using a two-way ANOVA 'without interaction'. Treatments are a fixed effect, i.e. the levels are determined by the investigator but blocks and the block by treatment interaction are random effects. The latter provides an estimate of the error variation. Blocking on fixed effect factors such as gender should be avoided because there may be real gender × treatment interactions. RB designs can also have a factorial set of treatments in which case there will be two or more fixed effect factors plus the single random effect factor.

Examples

Example 1: Removing the effect of rabbit size: a 'matched pairs' design

When comparing two treatments, which may be size-related, pairs of rabbits would be matched for body weight and then one assigned at random to each treatment. Each pair is a block. If there were three treatments, then each block would consist of a trio of rabbits. It is assumed that any response to the treatment will be approximately the same in each block and that differences between blocks are of no interest. Note that blocks can differ by more than one source of variation. For example, a block of large subjects might be treated in the morning and small subjects in the afternoon so the blocks differ in both time and size. Any variation due to size and time of day, which do not need to be separately identified, will be removed in the statistical analysis.

Treatment means are estimated by averaging across all blocks. When there are only two treatments the experiment can be analysed either by a paired *t*-test or by a two-way ANOVA without interactions. These two methods are mathematically identical.

Example 2: Apoptosis in rat thymocytes: blocking in time

The aim of this very small experiment was to check that two drugs designated CPG and STAU cause apoptosis (programmed cell death) in rat thymocytes. The drugs were to be used in further studies and the investigators wished to be certain that they worked. In this respect it could be classified as a pilot study.

Each week, for three weeks they humanely killed a single rat, removed the thymus, and prepared the thymocytes. These were pipetted into three tissue cultures. Solutions of the three treatments (vehicle, CPG and STAU) were added to the dishes. The dishes were incubated for a specified period and the apoptosis was scored. The dishes were the experimental units and the blocking factor was 'week'. The outcome was the apoptosis score.

Statistical analysis using R-Commander (Rcmdr)

The data (Table 7.1) are read into Rcmdr from the clipboard using *Data*, *import data*, checking the clipboard button and naming the data set in the box provided (if wanted). Note that both treatment and blocks are designated alphabetically so that Rcmdr considers them to be factors. If blocks or treatments had been designated 1, 2, 3, then Rcmdr would need to have been told that these are factors, not numerical variables. Note also that Rcmdr sorts and compares groups alphabetically so the name of the control group should start with a letter earlier in the alphabet than the treatment groups (as happens here). Treatment names can be adjusted so that the control group (when present) comes first.

As already noted, the data should be screened for any serious outliers. With two-way (e.g. RB) designs stripcharts are not particularly useful if there are large block differences. Individual observations can be plotted using an index plot but this is perhaps better done using an EXCEL bar chart. Figure 7.1, drawn in EXCEL, shows that the relationships between the three treatments is the same in each block but the scores in block C were all much lower than those in blocks A and B. The reason for this is unknown. The protocols were the same each week and it might be assumed that the experiment is simple enough to ensure that the scores were very similar on each replication. In fact, large block differences are common, suggesting that historical data should be treated with great caution.

Table 7.1 Apoptosis score in rat thymocytes.

Treatment	Block	Score
C	A	365
CPG	A	398
STAU	A	421
C	B	423
CPG	B	432
STAU	B	459
C	C	308
CPG	C	320
STAU	C	329

Treatment (C): control.

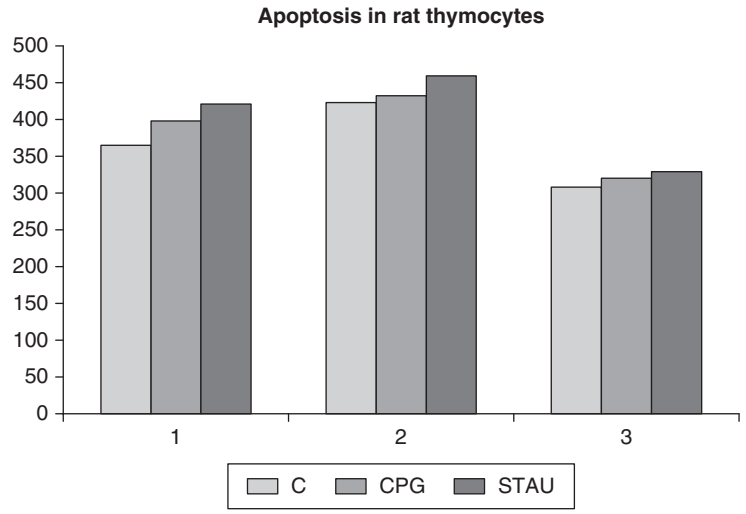


Figure 7.1 Plot of the raw data in the apoptosis experiment (drawn in EXCEL). One block was completed each week for three weeks. Treatments were C, CPG and STAU. Note the good repeatability within each week, but large differences between weeks. These differences are separated out as a block effect in the two-way ANOVA without interaction so do not contribute to the error.

As the effect of the block is removed in the statistical analysis they are of no concern in a properly designed RB experiment.

Another approach to screening the raw data is to do the ANOVA and look at the residual diagnostic plots, which show individual points. These are not shown here as there is no evidence either of heterogeneity of variances or non-normality of the residuals.

The ANOVA is done using *Statistics*, *Fit model*, *Linear model*. The form requires input of the dependent variable ('Score' in this case), the treatment, '+' (instead of the '*' used for a factorial design) and the block.

The resulting linear model analysis is shown in Table 7.2. First is a list of residuals. A serious outlier would show up as a large value. Next is a list of coefficients. These compare the means of the first group (controls, because alphabetically C comes before CPG and STAU) with the other groups. The last column provides a *P*-value for comparison of the controls with the other two groups. Therefore, the *P*-value for CPG is 0.086 and for STAU it is 0.009. Thus only STAU differs statistically from the controls at the 5% level of significance. The effect of each block is also shown, but these are of no direct interest. However, it is worth noting that large block effects imply that there are some uncontrollable factors having an important effect on the measurements. The output also gives the residual standard error of 9.735. This is the square root of the error mean square, and is the standard deviation (SD) that should be quoted with the means. The adjusted R-squared is 0.9688. This is the proportion

Table 7.2 Linear model (LM) analysis of the apoptosis data.

Call:
Lm (formula = Score ~ Treatment + Block, data = Dataset)

Residuals:

1	2	3	4	5	6	7	8	9
–11.111	3.889	7.222	3.556	–5.444	1.889	7.556	1.556	–9.111

Coefficients:

	Estimate	Std. error	*t*-value	*P* (>|*t*|)
(Intercept)	376.111	7.256	51.832	8.29e-07***
Treatment [T.CPG]	18.000	7.949	2.264	0.08625
Treatment [T.STAU]	37.667	7.949	4.739	0.00905**
Block [T.Block]	43.333	7.949	5.451	0.00550**
Block [T.C]	–75.667	7.949	–9.519	0.00068***

Rcmdr uses the convention that 1, 2 or 3 asterisks are shown to indicate significance levels of $p<0.05$, $p<0.01$ and $p<.001$

Residual standard error: 9.735 on 4 degrees of freedom (DF).

Multiple R-squared: 0.9844, adjusted R-squared: 0.9688.

F-statistic: 63.03 on 4 and 4 DF, *P*-value: 0.0007242.

Table 7.3 Analysis of variance of the apoptosis data. Response: score.

	DF	SS	MS	*F*-value	*P*-value
Treatment	2	2129.6	1064.8	11.235	0.0228374*
Block	2	21764.2	10882.1	114.817	0.0002931***
Residuals	4	379.1	94.8		

DF: degrees of freedom, SS: sum of squared deviations, MS: mean squares. *p<0.05, ***p<0.001

Table 7.4 Means and pooled standard deviation for the apoptosis data.

Treatment	Mean
C	365
CPG	383
STAU	403*

Pooled standard deviation = 9.74. *Significantly different from C (control) ($P < 0.05$).

of the total variation accounted for by the fitted model. This is a high value, implying a good fit. There is clearly very little variation in this experiment that is not taken into account by the mathematical model which has been fitted. Therefore, although by any criterion the experiment is excessively small, it can still identify statistically important effects.

Treatment means are obtained from *Statistics*, *Summaries*, *Numerical summaries*. However, the individual SDs should be ignored as they include differences between the blocks. The pooled standard error should always be used when presenting the results of an RB design.

The ANOVA (see Table 7.3) is obtained from *Models*, *Hypothesis tests*, *ANOVA table* and click 'Type I'. The means and the pooled SD are shown in Table 7.4.

Example 3: A crossover experiment

The aim of this experiment was to determine taste preferences in C57BL mice. These data are given in Table 7.5. There were four boxes each containing two C57BL/6 female mice. Each box had two water bottles, one of distilled water and the other containing the test solution. The bottles were re-filled daily and their positions were reversed. Each box had the test solutions for five days with distilled water in both bottles over the weekend. There were five treatments given in random order to each box, including a control group which had both water bottles filled with water, one of which was designated the 'treatment'. The data are the percentages of the total fluid consumed that was the test fluid, by weight. The test solutions were water (the control), saccharine, sodium chloride, sucrose, and ethanol.

Table 7.5 Raw data for the taste preference experiment.

ID	Cage (block)	Treatment[1]	Score
1	B1	B	69.6
2	B1	C	61.9
3	B1	A	54.9
4	B1	E	69.4
5	B1	D	78.3
6	B2	A	48.2
7	B2	D	81.5
8	B2	E	60.9
9	B2	B	61.1
10	B2	C	43.9
11	B3	C	53.4
12	B3	D	74.7
13	B3	A	49.9
14	B3	B	58.2
15	B3	E	68.5
16	B4	E	64.5
17	B4	A	50.4
18	B4	D	73.6
19	B4	B	55.3
20	B4	C	50.7

Note that the experimental unit is a cage of two mice for five days. Each block (cage) has been separately randomised. [1]A = control, B = 0.02% saccharine, C = 0.05 mmol/L sodium chloride, D = 0.04 mmol/L sucrose, E = 10% ethanol.

In this experiment the experimental unit was a cage of two mice for five days, the block was the cage, the treatments were the test solutions and the dependent variable was the weight of the test fluid consumed expressed as a percentage of the total fluid.

The statistical analysis is very similar to Example 2. The data are read into Rcmdr. A stripchart is used to study individual data points (not shown). These may be scattered to some extent by block differences. The statistical analysis involves fitting a linear model as previously described (*Statistics*, *Fit model*, *Linear model*). The residual diagnostic plots should be studied. In this case they give no cause for concern (not shown). The output from Rcmdr is shown in Table 7.6. The control treatment (distilled water) was coded treatment A, and it can be seen that treatments B, D and E all resulted in a statistically significant increase in consumption of the test solutions, whereas there was no preference for treatment C. Strain C57BL/6 is known to like ethanol whereas some other strains avoid it.

Table 7.6 Output from fitting a linear model (LM) to the taste preference data.

Call:
Lm(formula = Score ~ Cage + Treatment, data = Taste)

Residuals:

Min	1Q	Median	3Q	Max
–6.250	–1.951	–0.385	2.165	6.800

Coefficients:

	Estimate	Std error	*t*-value	*P* (>|*t*|)
(Intercept)	56.225	2.482	22.655	3.24e-11***
Cage [T.B2]	–7.700	2.482	–3.103	0.009146**
Cage [T.B3]	–5.880	2.482	–2.369	0.035450*
Cage [T.B4]	–7.920	2.482	–3.191	0.007758**
Treatment [T.B]	10.200	2.775	3.676	0.003172**
Treatment [T.C]	1.625	2.775	0.586	0.568968
Treatment [T.D]	26.175	2.775	9.433	6.70e-07***
Treatment [T.E]	14.975	2.775	5.397	0.000161***

Rcmdr uses the convention that 1, 2 or 3 asterisks are shown to indicate significance levels of p<0.05, p<0.01 and p<.001

Residual standard error: 3.924 on 12 degrees of freedom (DF).

Multiple R-squared: 0.9164, adjusted R-squared: 0.8676.

F-statistic: 18.78 on 7 and 12 DF, *P*-value: 1.418e-05.

Note that treatment means B, D and E differ from the controls ($P < 0.01$), but treatment C does not.

R-squared, the proportion of the variation accounted for by fitting the model was 91.6%, a high value. The ANOVA is shown in Table 7.7 and the treatment means are given in Table 7.8, and are plotted in Figure 7.2. It can be concluded that C57BL/6 mice prefer solutions of saccharine, sucrose and ethanol, but are indifferent to salt at the dose levels used.

Table 7.7 Analysis of variance for the taste preference experiment. Response: score.

	DF	**SS**	**MS**	***F*-value**	***P*-value**
Cage	3	205.14	68.38	4.4407	0.02557*
Treatment	4	1819.17	454.79	29.5350	3.962e-06***
Residuals	12	184.78	15.40		

DF: degrees of freedom, SS: sum of squared deviations, MS: mean squares. *p<0.05, ***p<0.001

Table 7.8 Means for the taste preference experiment.

Treatment[1]	Mean (% drunk from test bottle)
A (water)	50.850
B (saccharine)	61.050*
C (sodium chloride)	52.475
D (sucrose)	77.025*
E (ethanol)	65.825*

[1]$n = 4$, pooled standard deviation = 3.92, *$P < 0.05$.

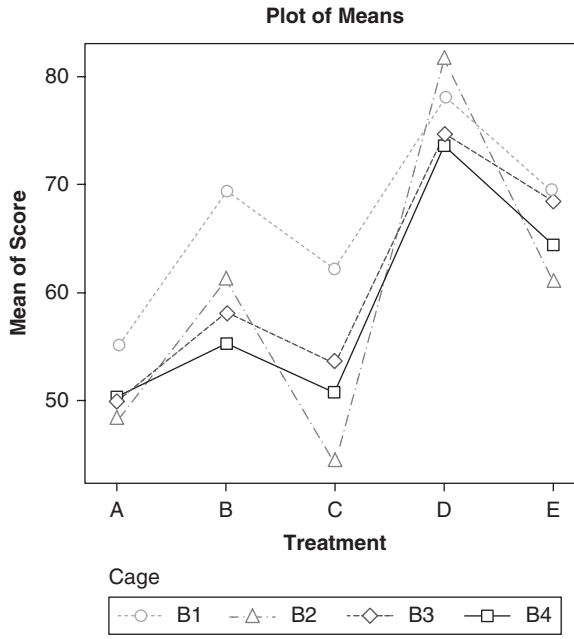


Figure 7.2 Taste experiment. Plot is of individual observations by cage (block) and treatment. Note the good agreement between blocks. Treatments are A = control, B = 0.02% saccharine, C = 0.05 mmol/L sodium chloride, D = 0.04 mmol/L sucrose, E = 10% ethanol.

Example 4: An RB experiment with a factorial treatment structure

As noted in Chapter 6 a factorial treatment structure can be used for a CR design, an RB design as well as other designs.

The aim of this experiment was to discover whether the antioxidant diallyl sulphide, a chemical found in garlic, affects the activity (nmol conjugate formed per minute per mg of protein) of a liver enzyme glutathione-S-transferase (Gst) in four mouse strains. It was of particular interest to see if there were large strain differences

Table 7.9 Raw data for a 2 (treatments) × 4 (strains) experiment in two blocks.

Strain	Treatment	Gst	Block
129/Ola	C	447	A
A/J	C	408	A
129/Ola	T	719	A
BALB/c	C	423	A
BALB/c	T	625	A
NIH	T	614	A
NIH	C	444	A
A/J	T	856	A
A/J	T	1002	B
A/J	C	609	B
BALB/c	T	782	B
NIH	T	831	B
NIH	C	764	B
129/Ola	T	766	B
BALB/c	C	586	B
129/Ola	C	606	B

Note randomisation within each block. Gst: glutathione-S-transferase.

in response. It involved two blocks separated, for logistical reasons, by a period of approximately two months. Each block consisted of eight mice: a treated and a control mouse of each of the four strains. The raw data are given in Table 7.9.

In Rcmdr the linear model box (*Statistics*, *Fit model*, *Linear model*) needs 'Gst ~ Block + Strain * Trt' to be entered into the boxes provided. Note the plus symbol following Block and the multiplication symbol following Strain. This code indicates that Gst levels depend (~) on Block plus the treatment and strain effect and the strain × treatment interaction.

The ANOVA is given in Table 7.10. As is common in RB designs there was a substantial block difference. The reason for this is unknown. The protocols were

Table 7.10 Analysis of variance for glutathione-S-transferase (Gst) activity in a 2 × 4 factorial design in two blocks. Response: Gst.

	DF	SS	MS	*F*-value	*P*-value
Block	1	124256	124256	42.0175	0.0003398***
Strain	3	28613	9538	3.2252	0.0914353
Treatment	1	227529	227529	76.9394	5.041e-05***
Strain:Treatment	3	49590	16530	5.5897	0.0283197*
Residuals	7	20701	2957		

DF: degrees of freedom, SS: sum of squared deviations, MS: mean squares.*$p<0.05$, ***$p<0.001$

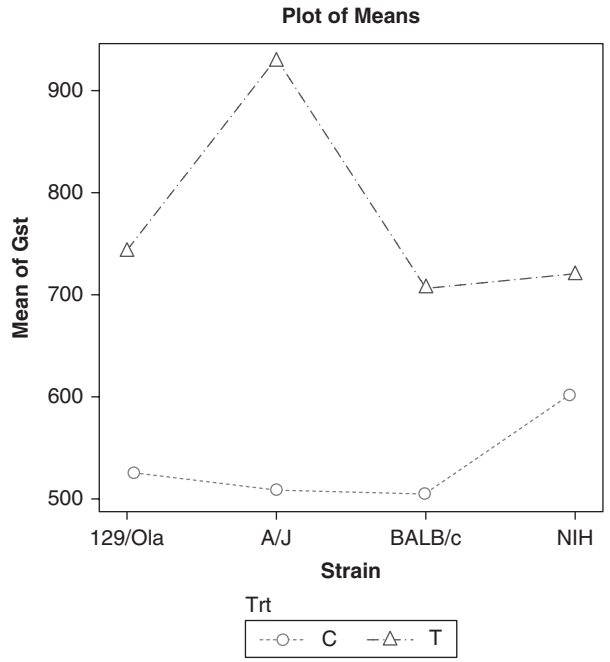


Figure 7.3 Plot of treatment means from a 4 (strains) × 2 (treatments) factorial experiment. Note that there is a significant ($P < 0.05$) treatment effect as well as a strain × treatment interaction ($P = 0.03$) which is due to the larger response in A/J than in the other strains. The results are plotted without error bars in this case because they need to be based on a pooled standard deviation. For publication purposes they should be re-plotted using the pooled standard deviation. T: treatment, C: control.

Table 7.11 Strain and treatment means for the glutathione-S-transferase (Gst) experiment.

Strain	**Mean (n = 4)**
129/Ola	634.5
A/J	718.7
BALB/c	604.0
NIH	663.2

Strain means are not significantly different $P = 0.09$. But there is a significant ($P = 0.03$) strain by treatment interaction. This is due to the increased response in strain A/J.

Treatment	**Mean (n = 8)**
Control	535.8
Treated	774.3

Pooled standard deviation = 53.4. The difference is highly significant, $P < 0.01$.

identical and the same person carried out the Gst determinations. The block differences present no problem as they are removed in the statistical analysis.

There were no significant strain differences, but there was a significant effect of the treatment and a significant strain × treatment interaction due to the larger response in strain A/J than the other strains. This is shown in Figure 7.3.

Treatment and strain means are given in Table 7.11.

Reproducibility

One of the advantages of RB designs is that they can provide some internal assurance of the reproducibility of the experiment, particularly if time and/or location are blocking factors, as was the case in the experiment discussed above. If the blocks are not in agreement, then there will be no statistically treatment effects. Figure 7.4 (prepared in EXCEL) shows that there is a good level of agreement between the two blocks for each treatment by strain combination.

Another example shows a lower level of agreement. 'Big Blue' mice are used to assess the possible mutagenicity, and therefore carcinogenicity of a chemical. They are transgenic for a shuttle vector consisting of a stretch of *Escherichia coli* DNA.

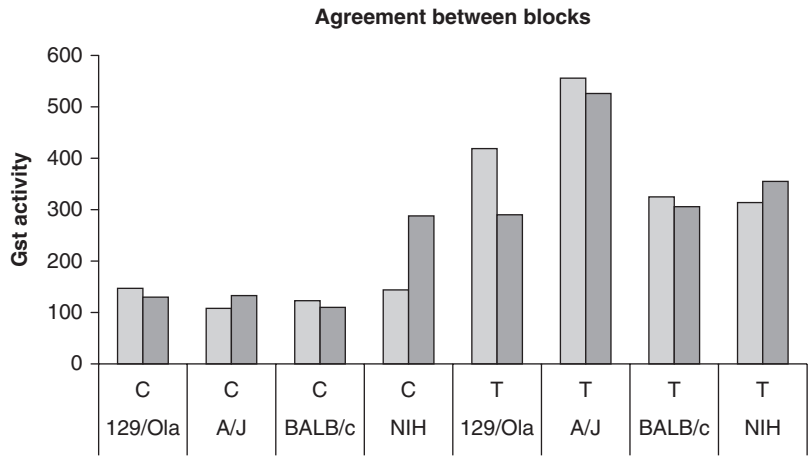


Figure 7.4 Plot showing agreement between blocks A and B. All the observations in block B are substantially higher than those in block A. But in order to show agreement between blocks this plot shows individual observations following subtraction of the mean block difference from block B and subtraction of 300 units. The agreement is good apart from some differences in the NIH controls (C) and 129/Ola treated (T) group. This plot is done using EXCEL. A plot of this type is not normally required. If the agreement is poor, then no statistically significant treatment effects would be observed. Gst: glutathione-S-transferase.

Table 7.12 Raw data from a multi-laboratory study. Data are the number of blue plaques per hundred thousand plaques.

Dose	Laboratory	Plaques
1.Control	A	23.1
1.Control	B	19.8
1.Control	C	10.4
1.Control	D	11.6
1.Control	E	29.4
2.Low	A	39.4
2.Low	B	29.8
2.Low	C	6.3
2.Low	D	23.3
2.Low	E	34.9
3.High	A	39.8
3.High	B	27.8
3.High	C	17.0
3.High	D	24.6
3.High	E	38.3

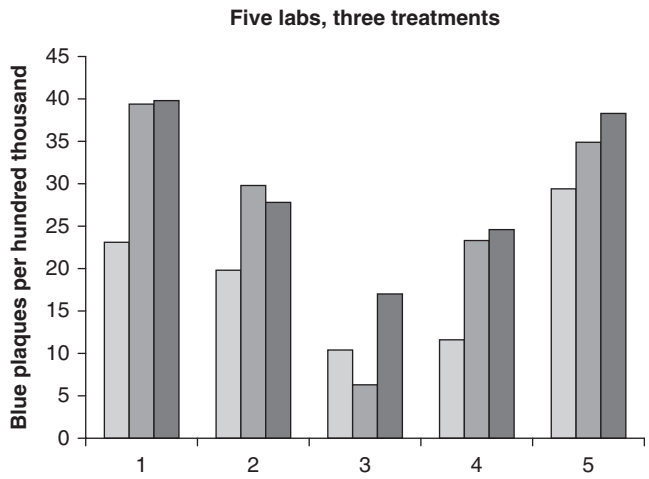


Figure 7.5 Plot of individual observations to indicate agreement.

The mice can be treated with the test chemical and the DNA can be recovered and transfected into *E. coli* and grown on agar plates. Any colony which has a mutation will result in a blue plaque in the lawn of colourless plaques. The data in Table 7.12 are gathered from an experiment in which a new way of counting the plaques was

being investigated. Coded samples of DNA from three mice given none, low, or a high dose of a carcinogen were sent to five laboratories. These all used the same protocol to count the plaques (unfortunately, no further details are available). A statistical analysis shows that the low and high dose counts differ significantly from the control at $P = 0.016$ and $P = 0.003$, respectively.

The reproducibility of these results is shown in Figure 7.5. There are clearly substantial differences between the laboratories/times in the mean response, as seems to be common with RB designs. At least the control is always lower than the high dose and in four out of five cases the control is lower than the low dose. However, the investigators decided that this level of reproducibility was not acceptable and no further development of this method was undertaken inspite of the highly significant differences between the treatments.

8

Split plots, Latin squares, covariance and other techniques

The split plot design

A split plot experiment is officially defined as a randomised block (RB) design with blocks being confounded with a major factor. An alternative definition would be that it is an experiment with two (or more) different levels of experimental units.

This design could arise in situations where animals receiving different treatments can be housed in the same cage, and the investigator wishes to use both males and females. Clearly, the two sexes cannot also be housed in the same cage, so the experiment might consist of several cages of males and the same number of cages of females, with each cage having one animal on each treatment. In this case the cage would be the experimental unit for comparing males and females but the animal would be the experimental unit for comparing the treatments.

The design could also arise when using a crossover or within-subject design in which the experimental unit is an animal for a period of time or, for example, a patch of skin with several patches per animal, and both sexes are used. In this case the experimental unit for comparing treatments would be the animal for a period of time or the skin patch, and the experimental unit for comparing males and females would be the whole animal.

The main problem with this design is that the gender differences (main plot level) will be poorly estimated because the sample size will be low, but the treatment differences and the gender × treatment effects will be well estimated, depending on sample sizes.

Split plot designs are worth knowing about because these situations can sometimes arise without the investigator being aware that they are using a split plot design.

A numerical example of a split plot design

Table 8.1 uses some artificial data to illustrate the analysis. There were six cages each with two animals; one treated and one control. Three cages contained males

Table 8.1 Fictitious data to illustrate the statistical analysis of a split plot experiment.

Cage	Sex	Treatment	BW
A	M	C	34.7
A	M	T	43.6
B	M	C	33.6
B	M	T	52.5
C	M	C	31.4
C	M	T	55.0
D	F	C	24.5
D	F	T	26.1
E	F	C	32.2
E	F	T	27.3
F	F	C	25.8
F	F	T	29.0

BW: body weight (grams).

and three contained females. The two animals within a cage were assigned either to the control or the treatment, given by injection. It is assumed that there could be no cross-contamination between treatments. The measured outcome (dependent variable) was body weight.

The data were read into R-Commander (Rcmdr) and a stripchart showing individual points was used to check the data (not shown).

In R a split plot design can be analysed with a single command but in Rcmdr a two-step statistical analysis is necessary, reflecting the nature of the experiment with two different types of experimental units (animals and cages).

First, the data in Table 8.1 were analysed as a 2 (sexes) × 2 (treatments) factorial design using *Statistics*, *Fit model*, *Linear model*. The 'BodyWt' is the dependent variable and the factors were Treatment * Sex. As usual, diagnostic plots should be used (not shown) to check the assumptions for using a parametric test.

Table 8.2 shows the resulting analysis of variance (ANOVA) (*Models*, *Hypothesis tests*, *ANOVA table*). The effect of Treatment and the Treatment * Sex interaction are each correctly based on 1 and 8 degrees of freedom (DF). However, the statistical significance of the sex difference with one DF is incorrect (and has been struck through). That needs to be tested against the variation between cages using a second one-way ANOVA of cage means (averaged across treatments), reflecting the fact that for comparing the sexes the cage with two animals in it is the experimental unit. This is shown in Table 8.3. A plot of the gender and treatment means is shown in Figure 8.1. As with a factorial design if there is a strong interaction between the main plot and split plot factors, then it is really the interaction which is of most interest.

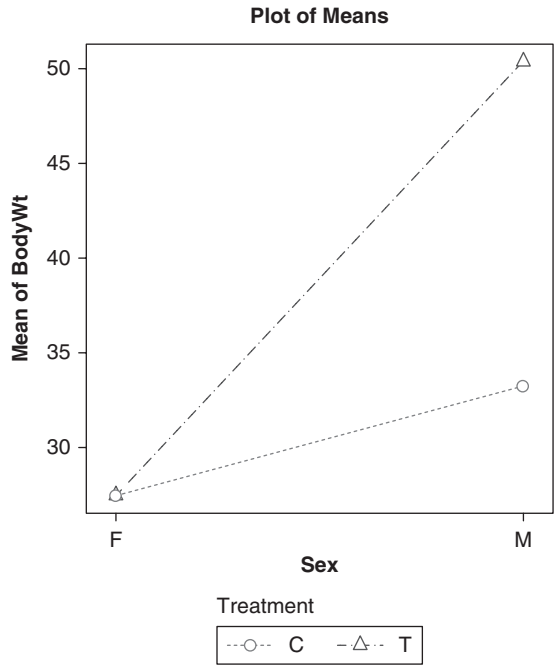


Figure 8.1 Plot of means for the split plot data in Table 8.1 showing the statistically significant interactions between sex and treatment as shown in Table 8.3. T: treatment, C: control.

Table 8.2 Analysis of the data from Table 8.1 as a Sex × Treatment factorial design. Response: body weight.

	DF	SS	MS	*F*-value	*P*-value
~~Sex~~	~~1~~	~~614.9~~	~~614.9~~	~~42.524~~	~~0.000184~~***
Treatment	1	219.31	219.31	15.166	0.004581**
Sex:Treatment	1	221.02	221.02	15.285	0.004484**
Residuals	8	115.68	14.46		

m1 = aov (BodyWt ~ Sex × Treatment + Error (Cage)).

Rcmdr uses the convention that 1, 2 or 3 asterisks are shown to indicate significance levels of $p<0.05$, $p<0.01$ and $p<.001$

Note that the Sex differences (struck through) are not correctly estimated by this analysis. These are obtained from a one-way analysis of variance (ANOVA) of the cage means as given in Table 8.3. DF: degrees of freedom, SS: sum of squared deviations, MS: mean squares.

Conclusion

A split plot could be a sensible design to use in cases where animals receiving different treatments can be housed together. In the special case of a within-animal experiment it is the only possible design if the experiment is to include both sexes. Remember, however, that the differences at the main block level are not estimated with much precision.

Table 8.3 The one-way analysis of variance of cage totals comparing the two sexes.

	DF	SS	MS	*F*-value	*P*-value
Sex	1	307.45	307.45	60.12	0.00149**
Residuals	4	20.46	5.11		

DF: degrees of freedom, SS: sum of squared deviations, MS: mean squares. ** $p<0.01$

Latin square designs

These designs are an extension of the concept of an RB design to an additional dimension. They were developed for agricultural field experiments where a field may vary in fertility in both a North–South and in an East–West direction. The power of the experiment can be increased by taking this into account. The design is rarely used in experiments involving laboratory animals, but there are situations where it may be useful.

In this design the number of experimental units is the square of the number of treatments. For example, if there are four treatments 16 experimental units will be needed and the experiment will be quite small, so may lack power (E, the error DF used in the resource equation method for determining sample size explained in Chapter 11, is only 6 when it should be at least 10). With seven treatments a total of $7 \times 7 = 49$ experimental units will be needed and the experiment will be quite large. Like the completely randomised (CR) and randomised block (RB) designs it can have a factorial arrangement of treatments. For example, a 2×2 factorial experiment has four treatments such as male treated, male control, female treated and female control.

As an example, suppose that there had only been four treatments in the taste experiment described in Chapter 7 (Example 3). It could then have been designed as a 4×4 Latin square. Such a design has exactly one of each treatment in each row and one of each treatment in each column. So in this case the rows would be the cage, each of which would have the four treatments, and the columns would be the weeks in which the treatments would be given.

Randomising a Latin Square design

The treatments in an un-randomised 4×4 Latin square design can be written A, B, C, D in the first row, B, C, D, A in the second row (i.e. the first letter, A in this case, placed at the end) and C, D, A, B in the third row, etc. More letters are needed for larger experiments. The result is that each row and column will have exactly one of each treatment. Randomisation is done by randomising whole rows followed by whole columns. This retains the structure of each row and column having exactly one of each treatment but they are now in a random order. The actual randomisation could be done by using cards with A–C on them which are then shuffled for each row and column.

A numerical example

The same data that were used in Example 3 in Chapter 7 to illustrate a crossover design, but with one treatment omitted, are used here to illustrate the statistical analysis of a 4 × 4 Latin square experiment (although that is not how it was actually designed). These data are shown in Table 8.4. Note that each cage (row) and each week (column) now have one of each treatment. The data are read into Rcmdr in the usual way. First, in Rcmdr the row and column numbers need to be converted to factors (*Data*, *manage variables in the active data set*, *convert numerical variables to factors*). They can be kept as numbers. Then: *Statistics*, *Fit model*, *Linear model*. The percent fluid consumed is the dependent variable. The independent variables are Treatment + Cage + Week. The ANOVA (see Table 8.5) is obtained from *Models*, *hypothesis tests*, *ANOVA table*. In this case there is a highly significant treatment effect, a row effect which approaches significance at $P = 0.07$ and a non-significant column effect (as expected).

Means can be obtained from *Statistics*, *Summaries*, *Numerical summaries*. However, a pooled standard deviation (SD) should be used, the square root of the residual mean square in the ANOVA table, i.e. $\sqrt{(15.4)} = 3.92$, not the individual SDs which contain variation due to the blocking.

Diagnostic plots can be obtained in the usual way. The control treatment (water) was designated as A, so the output from fitting the linear model shows which treatments differ significantly from the control (in this case B at $P = 0.01$ and D at $P < 0.001$). A plot of the means for each week is shown in Figure 8.2.

Table 8.4 Data to illustrate the statistical analysis of a Latin square experiment.

Treatment	Cage	Week	Percent
B	1	1	69.6
C	1	2	61.9
A	1	3	54.9
D	1	4	78.3
A	2	1	48.2
B	2	2	61.1
D	2	3	81.5
C	2	4	43.9
C	3	1	53.4
D	3	2	74.4
B	3	3	58.2
A	3	4	49.9
D	4	1	73.6
A	4	2	50.4
C	4	3	50.7
B	4	4	55.3

A = water, B = saccharine, C = sodium chloride, D = sucrose.

Table 8.5 Analysis of variance table for the Latin square example. Response: percent.

	DF	SS	MS	*F*-value	*P*-value
Treatment	3	1713.26	571.09	37.0359	0.0002885***
Cage	3	186.99	62.33	4.0422	0.0687220
Week	3	65.93	21.98	1.4252	0.3249437
Residuals	6	92.52	15.42		

DF: degrees of freedom, SS: sum of squared deviations, MS: mean squares. ***$p<0.001$

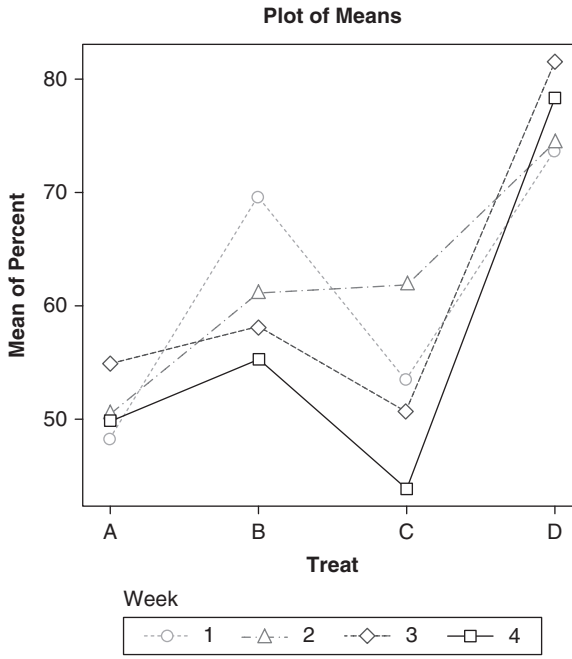


Figure 8.2 Plot of treatment means for each week in the Latin square example. Note some minor changes in the ranking each week, but the results are reasonably consistent.

A Latin square design is most likely to be of benefit if there is a trend in the observations over both rows and columns, which is not the case here.

Repeated measures designs

There is confusion and disagreement in the literature about 'repeated measures' designs. In most textbooks (Montgomery, 1997) this is just another name for a cross-over experiment where a series of treatments is given sequentially to an animal, with some dependent variable being measured after each treatment. The experimental unit is an animal for a period of time.

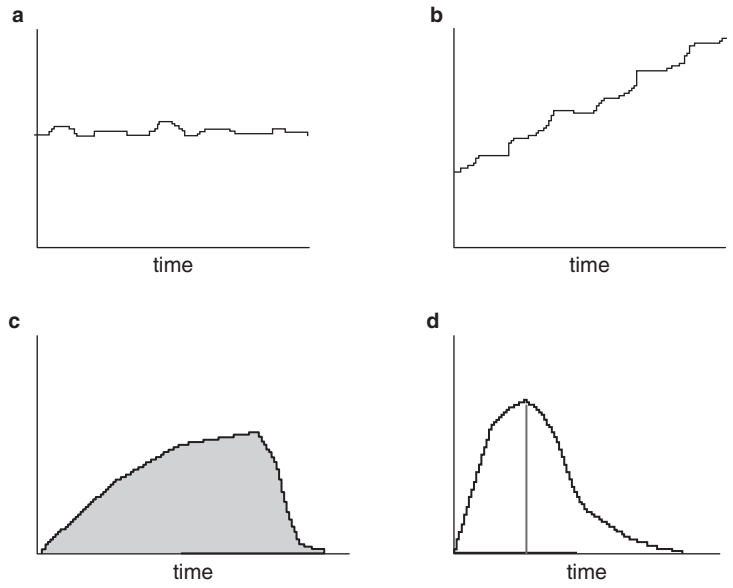


Figure 8.3 In a 'repeated measures' design in which serial measurements are taken from an experimental unit without a change of treatment it may be best to convert the observations on each individual to one of the options shown here: (a) the mean of the set of observations or (b) the slope of the line or (c) the area under the curve or (d) time to reach a peak.

However, in other textbooks the term 'repeated measures' is used in situations where an experimental unit is given a treatment and then several sequential measurements are made without any change of treatment. It is claimed that 'time' is a treatment. But because 'times' cannot be assigned at random, an analysis based on the assumption that they are experimental units is open to criticisms. The repeated measurements are more like outcomes than experimental units (Mead et al., 1993).

Clearly it is sometimes of interest to know what happens in the time following a treatment. But it seems wiser to use an alternative approach to the analysis of such data. The easiest approach is to form a new variable which is a combination of the measurements. Figure 8.3 shows four situations in which the multiple measurements on each individual could be converted to means, the slope of a line, the area under a curve or time to reach a peak. These composite measurements (e.g. the mean across time points) could then be analysed using, for example, an ANOVA.

Another alternative would be to use a 'multivariate' approach such as a principle components analysis, the aim of which is to reduce the dimensionality of the data; or a discriminant function analysis to distinguish between individuals. A discussion of these techniques is beyond the scope of this book.

Covariance analysis

The aim of covariance analysis is to increase the power of an experiment by taking into account a variable which is measured *before* the start of the experiment and

which it is thought might contribute to the outcome of interest. An obvious example in work with laboratory animals would be to use data on initial body weights to correct some outcome which may be related to body weight, such as the weight of various organs. The technique is in some ways similar to blocking. Blocks are discrete factors set up before starting the experiment, but a covariate is a quantitative variable, such as initial body weight. In each case the block or covariate is established or recorded before treatments are given. The specification in Rcmdr is the same as for an RB design except that the covariate is a variable instead of a discrete factor.

A numerical example

Table 8.6 shows data on the effect of two diets (A and B) on the weight of abdominal fat in mice. A *t*-test fails to show any significant difference in mean fat weight between the two diets ($P = 0.07$). Nor do the two groups differ in initial body weight,

Table 8.6 Data to demonstrate the analysis of covariance.

Animal	Initial BW[1]	Fat[2]	Diet
1	23.03	0.433	A
2	26.15	0.728	A
3	27.33	0.875	A
4	24.31	0.846	A
5	24.00	0.451	A
6	26.94	0.800	A
7	21.51	0.366	A
8	26.66	0.744	B
9	23.36	0.691	B
10	21.06	0.593	B
11	24.42	0.913	B
12	24.41	0.922	B
13	25.97	1.143	B
14	23.52	1.007	B

[1]This was recorded before putting the animals on the test diets. [2]Weight of abdominal fat (g). BW: body weight (g).

Table 8.7 Analysis of variance table for the covariance example. Response: fat.

	DF	SS	MS	*F*-value	*P*-value
Diet	1	0.16373	0.163728	7.1656	0.021528*
Initial body weight	1	0.25491	0.254909	11.1562	0.006592**
Residuals	11	0.25134	0.022849		

DF: degrees of freedom, SS: sum of squared deviations, MS: mean squares. *P<0.05, **P<0.01

recorded *before* the experiment was started ($P = 0.6$). However, a covariance analysis adjusts the fat weight for initial body weight (*Statistics*, *Fit model*, *Linear model*, fat ~ diet + Init.body.wt). Table 8.7 shows the resulting ANOVA table. When initial body weight is taken into account there is a statistically significant difference between the diets in the mean fat weight ($P = 0.022$).

Random effects and variance components

Sometimes it would be of interest to quantify different sources of variation. This was discussed briefly in Chapter 3 and Figure 3.2 where variation due to days, animal sizes, samples and analysts was considered. In this simplified example variation due to differences between cages, and variation among animals within cages is considered. This variation will depend to a large extent on the heterogeneity of the environment due to factors such as the level of the cage in a rack and consequent exposure to light and heat, and also to factors such as social interactions within each cage. Knowledge of these sources of variation may influence the design of a future experiment. For example if there is a lot of variation between cages, it is better to increase the size of the experiment by adding more cages. But if most of the variation is due to individual variation within cages, then it might be better to have more animals per cage, although this does not take into account possible interactions between individuals.

A numerical example

The data in Table 8.8 were extracted from a much larger experiment to illustrate this type of analysis. Female rats housed in pairs were maintained on a control diet for three months. The cages were rotated both vertically and horizontally every week in order to average out environmental differences due to their location in the rack. The data were subjected to a one-way ANOVA with the random factor 'cage' being the independent variable and body weight being the dependent variable. As there were eight cages, this resulted in 7 DF for cages and 8 DF for within cages. The ANOVA table with the 'expected mean squares' is shown in Table 8.9. The component of variance associated with cage differences was 149 g and within cages it was 179 g, so the within-cage variability was slightly higher than the between-cage variability.

In the males on the same experiment (not shown) the between-cage variance component was negative. This implies that two rats in a cage differed more in body weight than would have been expected had they been assigned to the cages at random. This is probably due to social interaction between the rats. Similarly, an extensive review of the available literature showed that group-housed mice of both sexes can be more variable than singly-housed mice (Prendergast et al., 2014).

Table 8.8 Numerical example for variance components analysis.

Cage ID	Animal ID	Final weight
1	1	276.27
1	2	249.21
2	3	221.44
2	4	215.17
3	5	218.08
3	6	234.02
4	7	207.03
4	8	235.20
5	9	246.76
5	10	245.45
6	11	215.40
6	12	226.00
7	13	248.05
7	14	223.88
8	15	234.71
8	16	216.00

Female rats aged 12 weeks. Cages were rotated vertically and horizontally in the cage rack each week.

Table 8.9 Variance components model. Response: final weight.

	DF	SS	MS	Expected MS
Cage ID	7	3345.8	477.97	$\sigma^2 + n\sigma^2_B$
Residuals	8	1433.8	179.22	σ^2

In this case n (the number per cage) = 2.

S^2 =179.

S^2_B = (477–179)/2 = 149.

DF: degrees of freedom, SS: sum of squared deviations, MS: mean squares.

9

Counts and proportions

Experiments are most powerful when there is a quantitative outcome such as body weight, the activity of an enzyme or some measure of behavioural activity. However, sometimes the outcome is a binary attribute such as a tumour being present/absent, female rats being pregnant/not-pregnant or alive/dead at the end of the experiment. In such cases the outcome is usually expressed in terms of counts and proportions or percentages and the aim of the statistical analysis is usually to compare the outcome, in treated and control groups. In all other respects the experiment should be conducted as previously described using randomisation and blinding where this is possible.

Sample size can be determined using a power analysis as described in Chapter 11.

Example

In a study of the effects of transportation on pregnancy in F344 strain time-mated rats, 148 rats were kept as controls and 150 rats were transported to another site in a journey of about 24 h (Pritchett et al., 2013). In the control group, 121 (121/148 = 81.8%) of the plugged females produced litters whereas in the transported females only 105 (105/150 = 70%) produced litters. Is this evidence of an effect of transportation on the number that produced litters, or could it just be due to chance?

Using R-Commander (Rcmdr), the data on rats littering and not littering (121/27 in controls and 105/45 in the transported animals) are entered as rows in a 'contingency table' (*Statistics*, *Contingency tables*, *Enter and analyse a two-way table*). The analysis then does a chi-squared test of the null hypothesis that rows and columns vary independently. This produces a chi-squared 23.67 with a *P*-value of well below 0.01, so the null hypothesis should be rejected at the one percent level and we conclude that the transported rats produced fewer litters than the non-transported animals.

A chi-squared test can be used for tables larger than 2 × 2. However, it may be inaccurate if the counts in some of the cells of the input table are less than 5. In such cases it may be possible to group across treatments.

The 'Statistics' tab also offers Fisher's exact test for analysing 2 × 2 tables. This test gives, in this case, a *P*-value of 0.021 and calculates the odds ratio (an estimate of the effect size) with a 95% confidence interval. In this example the odds of being pregnant when not transported are 121/27 = 4.48 and the odds of being pregnant

Table 9.1 Output from R-Commander (Rcmdr) in a comparison of the number of litters in transported and non-transported female rats.

Data: Table
P-value = 0.0213
Alternative hypothesis: true odds ratio is not equal to 1
95% confidence interval: 1.079232, 3.451855
Sample estimates:
Odds ratio: 1.91645

when transported are 105/45 = 2.33. So the odds ratio is 4.48/2.33 = 1.92 with a 95% confidence interval of 1.08 to 3.45. An odds ratio of one would imply no difference between the groups. The output is shown in Table 9.1.

It can be concluded that the chance of not being pregnant is 1.92 times higher in animals which are transported than those which are not transported, with the 95% confidence interval being as quoted above.

10

Regression and correlation

Introduction

Regression and correlation are used to study the relationship between two or more variables. In regression there are one or more independent or 'explanatory' variables, such as dose level, temperature or body weight, often controlled by the investigator; and a dependent variable which is causally associated with it. Correlation, by contrast, studies any association between two variables which may, or may not, be causally related.

Linear regression

The main aim in regression analysis is to quantify any linear relationship between one or more independent X variables and the dependent Y variable. X can be a factor such as a dose level or a measurement variable. When there is just a single X and a single Y variable, the regression analysis will determine the best fitting straight line in the form $Y = a + bX$, using a 'least squares' procedure, where a and b are constants determined from the data in the statistical analysis. In this case a is the value of Y when $X = 0$ and b is the slope of the line when Y is plotted as a function of X. When there are two or more independent X variables, then multiple regression analysis is used (not discussed here).

An example

Some data on red blood cell (RBC) counts in mice (Festing, 2001) treated with chloramphenicol are shown in Table 10.1. The data can be read into R-Commander (Rcmdr) in the usual way and plotted to show the relationship between the dose of chloramphenicol and RBC counts (*Graphs*, *Scatterplot*). The dose is the independent or X variable and RBC is the dependent Y (or response) variable. Under 'Options' all the boxes should be un-ticked except for 'least squares line'. This will result in the best fitting straight line to the data points, as shown in Figure 10.1.

Table 10.1 Red blood cell (RBC) count ($\times 10^{12}$/L) in mice treated with chloramphenicol.

Dose	RBC	Dose	RBC	Dose	RBC
0	9.60	5	9.27	10	9.61
0	9.56	5	9.16	10	9.82
0	9.14	5	9.53	10	9.44
0	9.20	5	9.44	10	9.29
15	9.81	20	8.82	25	7.83
15	9.83	20	8.24	25	8.07
15	9.83	20	8.18	25	8.45
15	9.40	20	8.10	25	7.77

Data are presented over three columns.

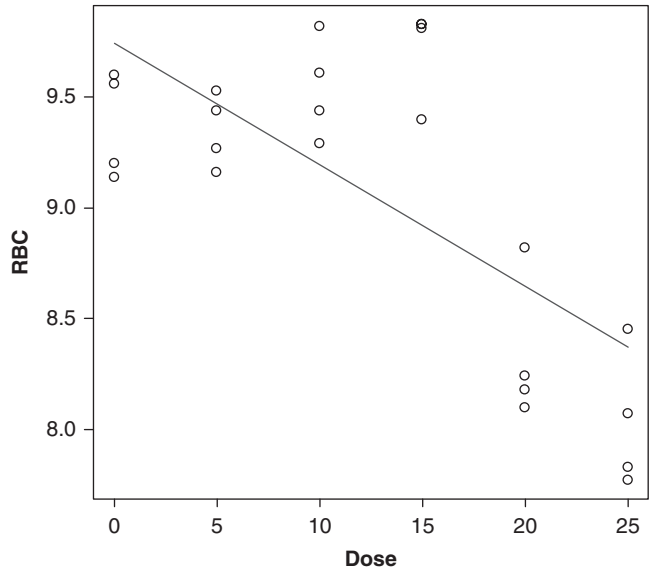


Figure 10.1 Plot of red blood cell (RBC) counts as a function of the dose of chloramphenicol with the best fitting straight line.

The regression analysis to determine the formula for the regression line is done using *Statistics*, *Fit model*, *Linear regression*. The output is shown in Table 10.2.

The section labelled 'Coefficients' shows the intercept (a in the formula above) as 9.74 and the slope b is –0.055. So the resulting formula for the regression line is $Y = 9.74 - 0.055X$. This can be used to estimate the value of Y for any value of X within the range of the X values.

The standard errors, t-values and statistical significances are also given. As might be expected, both of these are significantly different from zero. The multiple

Table 10.2 R-Commander (Rcmdr) linear regression analysis of red blood cell (RBC) counts on chloramphenicol dose.

Call:
Lm (formula = RBC ~ Dose, data = Dataset)

Residuals:

Min	1Q	Median	3Q	Max
–0.60310	–0.42181	–0.08606	0.28755	0.90912

Coefficients:

	Estimate	Std error	*t*-value	*P* (>|*t*|)
(Intercept)	9.74310	0.18299	53.245	<2e-16***
Dose	–0.05481	0.01209	–4.535	0.000163***

Rcmdr uses the convention that 1, 2 or 3 asterisks are shown to indicate significance levels of p<0.05, p<0.01 and p<.001

Residual standard error: 0.5057 on 22 degrees of freedom (DF).

Multiple R-squared: 0.4831, adjusted R-squared: 0.4596.

F-statistic: 20.56 on 1 and 22 DF, *P*-value: 0.0001634.

R-squared, the proportion of the total variation that is accounted for by fitting the straight line, is 0.48. This value is quite low.

Figure 10.1 shows that a straight line may not give an adequate indication of the relationship between dose and response. It appears as though there is no response to chloramphenicol at the lower doses. If at all possible the RBC counts seem to increase. But at the two higher doses there is a sharp decline in RBC counts. A curve might be a better fit.

A second degree polynomial with the formula $Y = a + bX + cX^2$ can be used to fit a suitable curve. In Rcmdr a new variable is formed by *Data*, *Manage variables in the active data set, Compute new variable*. The new variable should be named (e.g. dose.2) and the formula for it is Dose * Dose. Now the regression analysis can be re-done to obtain the best fitting curve. In this case *Statistics*, *Fit model*, *Linear regression*. In the Explanatory variables box both dose and dose.2 should be checked.

The output (see Table 10.3) shows that R-squared is now 0.74 and dose.2 accounts for a significant proportion of the total variation which is explained by the regression analysis. The formula for the line is now $Y = 9.28 + 0.08X - 0.005X^2$, where Y is the RBC count and X is the dose. Fitting the square of the dose alone (dose.2) results in a reasonably straight line, but an R-squared of 0.66 suggests that it is not quite as good as fitting both parameters. R-Commander cannot plot the second degree polynomial curve, but it can produce a scatterplot which shows a fitted line with some indication of variation, as shown in Figure 10.2

The overall conclusions will depend to a large extent on the purpose of the statistical analysis. In some cases it is sufficient to show that there is a statistically

significant relationship between the independent and dependent variables, quantified by R-squared. In other cases it may be useful to be able to predict the value of Y for a given value of X using the formula $Y = a + bX$ with the estimates of a and b. It may also be useful to know whether there is a non-linear response to the dose, as shown in this example.

Table 10.3 Output from a regression analysis fitting a second degree polynomial to the chloramphenicol data.

Call:
Lm (formula = RBC ~ Dose2 + Dose, data = Dataset)

Residuals:

Min	1Q	Median	3Q	Max
–0.63836	–0.17680	–0.05779	0.26280	0.54293

Coefficients:

	Estimate	Std error	*t*-value	*P* (>|*t*|)
(Intercept)	9.285357	0.165747	56.021	< 2e-16***
Dose2	–0.005493	0.001197	–4.588	0.000159***
Dose	0.082507	0.031181	2.646	0.015108*

Rcmdr uses the convention that 1, 2 or 3 asterisks are shown to indicate significance levels of p<0.05, p<0.01 and p<.001

Residual standard error: 0.3658 on 21 degrees of freedom (DF).

Multiple R-squared: 0.7419, adjusted R-squared: 0.7173.

F-statistic: 30.18 on 2 and 21 DF, *P*-value: 6.673e-07.

$Y = 9.28 + 0.08X - 0.005X^2$

Correlation

Correlation is used to quantify the linear relationship between two variables, without the assumption of any causal relationship between them. It should be treated with some caution as there is a danger of assuming a causal relationship.

The 'product–moment' or Pearson correlation is the one (among several) which is most widely used. The correlation can range from +1, in which there is complete agreement between the two variables, through to 0 in which they are not associated to –1 in which there is complete disagreement with the highest value in one of the variables being the lowest value in the other. Note that Pearson's correlation coefficient assumes that both variables are normally distributed; if this is not the case other, such as Spearman's or Kendall's, coefficients (which are based on ranks and concordant pairs, respectively) should be used.

An example

Table 10.4 shows liver and mean of the left and right kidney weights in 18 genetically heterogeneous mice. The relationship is shown graphically in Figure 10.2 (in *Rcmdr Graphs*, *Scatterplot*). Clearly there is a positive relationship between the two. Such plots can be useful in screening data to identify outliers (one point in this figure is an outlier) when two or more outcomes are measured in an experiment. Any found can subsequently be checked to make sure that they are not mistakes (there was no evidence of a mistake in this case). Note that no regression line has been drawn. There are two possible lines; the regression of liver weight on kidney weight and vice versa. These give different lines. So when presenting a correlation it is best either to give no line or both lines.

Could this correlation be attributed to chance? The result of a test for correlation (*Statistics*, *Summaries*, *Correlation test*) is given in Table 10.5. The correlation is 0.66, with a 95% confidence interval of 0.28 to 0.86 and it is highly significant ($P = 0.002$), so is unlikely to be due to chance sampling variation. Rcmdr gives two alternative correlations: Spearman's and Kendall's correlation coefficients. These are used with skewed data sets.

Table 10.4 Liver and mean kidney weights (g) in genetically heterogeneous female mice.

Animal	Liver	Kidney	Animal	Liver	Kidney	Animal	Liver	Kidney
1	0.99	0.29	7	1.09	0.29	13	0.98	0.24
2	1.19	0.34	8	0.83	0.25	14	0.93	0.24
3	1.28	0.28	9	1.03	0.29	15	0.80	0.26
4	1.19	0.32	10	1.01	0.24	16	0.99	0.29
5	1.06	0.29	11	0.9	0.26	17	1.03	0.26
6	1.04	0.28	12	0.99	0.26	18	0.86	0.24

Table 10.5 Test of the correlation between mean kidney weight and liver weight in female mice.

Pearson's product–moment correlation
Data: kidneys and liver
$t = 3.5191$, DF = 16, P-value = 0.002846
Alternative hypothesis: true correlation is not equal to 0
95% confidence interval: 0.280020, 0.861678
Sample estimates:
Correlation: 0.6605362

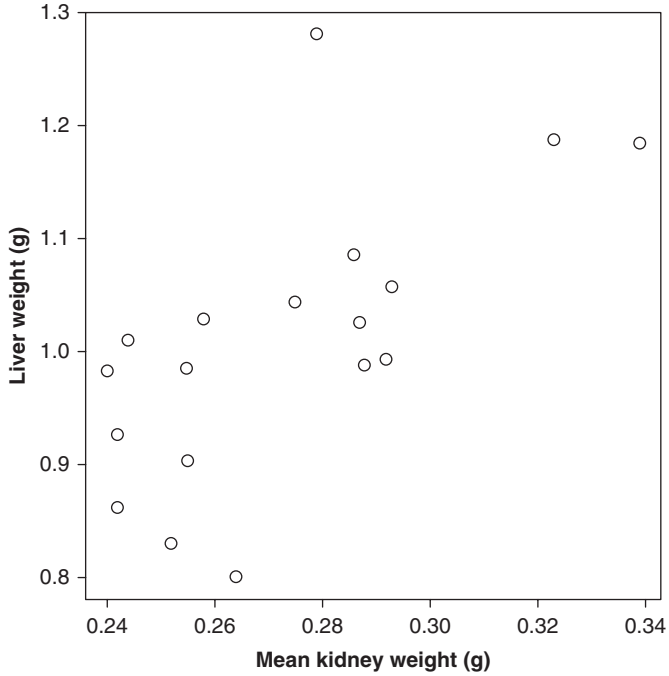


Figure 10.2 Scatter plot showing relationship between mean kidney weight and liver weight in genetically heterogeneous female mice. There is one outlier.

11

The determination of sample size

Introduction

If an experiment is unnecessarily large, then time and scientific resources will be wasted and animals may suffer unnecessary distress. By contrast, if the sample size is too small, effects which are of biological or clinical importance may be missed. Unfortunately, there is no simple way of determining the optimum sample size, which depends on several factors, some of which are unknown at the time the experiment is planned. According to one authority (Cox and Reid, 2000), 'except in rare instances…, a decision on the size of the experiment is bound to be largely a matter of judgement and some of the more formal approaches to determining the size of the experiment have spurious precision'.

Two methods of estimating a suitable sample size are discussed in this chapter: power analysis and the resource equation. Power analysis is most appropriate for relatively simple applied experiments, where a good estimate can be made of the standard deviation (SD) of the characters of interest and there is good information on the magnitude of a response which would to be of clinical or scientific importance if it were to be found.

Power analysis is the method used for clinical experiments where a new treatment may only be worth developing if, it works substantially better than existing treatments. It is difficult to apply with complex experiments and should not be used unless there is a good estimate of the SD of the character of interest and it is realistically possible to estimate an effect size (ES) that is likely to be of scientific importance (see below).

Note that a power analysis is not usually an objective way of determining sample size because it requires an estimate of the ES likely to be of clinical or scientific importance, which usually depends to a large extent on judgement.

By contrast, the resource equation method is more objective and is usually more suitable for fundamental studies with more complex designs where there is little or no baseline information and the main interest is in determining whether there is any treatment response of a magnitude which might justify further investigation.

Power analysis

The aim of power analysis is usually used to estimate the sample size needed to detect the smallest response (difference between groups) which the investigator considers to be of biological or clinical importance. It depends on a mathematical relationship between six variables discussed below. If five of these are specified, the sixth variable can be calculated. This method can be used both for quantitative (measurement) and qualitative data.

Where an experiment involves measuring several different dependent variables such as body weight, haematological parameters and blood pressure, it is necessary to decide which of these characters is of most importance and use this to determine the sample size. The six variables involved are as follows:

1. The effect size of biological or clinical interest

A small difference between the control and the treated group, i.e. a small ES, may not be of much scientific/clinical interest, even if it is statistically significant. However, a large response (i.e. ES) probably would be of interest. The ES used in a power analysis is the cut-off between these two situations. It is the minimum size of the response that is judged to be of clinical or scientific importance. It is *not* an estimate of the actual magnitude of the response.

In some cases it is difficult to estimate an ES. An alternative approach is to specify a sample size that seems reasonable and practical (say 10 animals per group). The formulae can then be used with the other variable being specified to estimate an ES that the experiment is, with a specified probability, capable of detecting. The investigator could then decide whether that ES would be acceptable. If not, another iteration using a different sample size can be carried out.

For categorical variables the ES is the difference likely to be of biological/clinical interest in percent responders between two groups. For example, if 50% of the control group is expected to show some qualitative effect, such as a tumour, sample sizes needed to detect a reduction to, say, 40%, 30% or 20% in the treated group can be estimated for a given level of power. Several iterations of the calculations may be needed to arrive at an answer in which the numbers are reasonably practical and of scientific interest.

2. The standard deviation

The estimated sample size is heavily dependent on the magnitude of the SD. For continuous characters such as body weight or enzyme activity, an estimate of the SD has to come from a previous experiment either carried out by the investigator or from the literature. If this is not available, then a pilot study may be necessary to provide an approximate estimate. It may be worthwhile to do a ‘best case’ and a ‘worst case’ calculation based on the lowest and highest of the available estimates of the SD to determine the effect of this on sample size estimates.

For discrete characters such as dead/alive, the SD is a function of the proportion of animals that are affected, so there is no need to specify it.

3. The significance level

Usually, a significance level of $\alpha = 0.05$ (5%) is used, though other levels such as $\alpha = 0.01$ (1%) could also be used. However, this would mean that the sample size would have to be larger, otherwise the power of the experiment will be reduced.

4. The power of the experiment

The power of the experiment is the probability of being able to detect a specified effect at a specified significance level. The general aim should be for experiments to be powerful, as these will have a high chance of detecting an effect, if it is present. Somewhat arbitrarily, the power is usually set at somewhere between 80% and 90% (0.8 to 0.9). The higher the power, the larger the sample size needed. But power does not increase linearly with an increase in sample size, rather it follows a curve of diminishing returns; so specifying a very high power, such as 99%, may require an unfeasibly large number of animals. This would only be justified if the consequences of failing to detect an effect would be serious.

5. The alternative hypothesis

The usual null hypothesis is that there are no differences among treatment means, with the alternative being that there are differences, but the direction of response is not specified. For example, body weight might be changed in either direction. This leads to a two-sided significance test. However, if a compound is being tested for toxicity, for example, it will be either toxic or non-toxic but it is unlikely to have negative toxicity, so a one-sided test should be used. Of course a toxic substance may increase or decrease body weight, so in that case a two-sided test might be relevant.

6. Sample size

In most cases it is the sample size that is to be determined, with all the other variables being specified. However, as noted above, the sample size could be fixed and the power or the effect size (ES) of the proposed experiment could be estimated. This situation can also arise when only a limited number of animals are available and the aim might be to determine what ES could be detected or what power the experiment would have for specified levels of the other variables.

The standardised effect size (SES) or Cohen's *d*

The effect size(ES) when divided by the standard deviation (SD), i.e. ES/SD, is known as the SES or Cohen's d. Power analysis was developed by Jacob Cohen and SES is explained in his 1969 book and later editions (Cohen et al., 1988).

It is a signal/noise ratio. A common concern of researchers who perform power calculations is that they cannot predict what the SD would be (although also see '2. The standard deviation' above). The advantage of the SES is that it allows researchers to design their experiment to detect a chosen signal/noise ratio (which has incorporated the SD). This is easier to conceptualise. Do we need to detect a large SES or a small one (see Cohen, 1988 for SES conventions in the behavioural sciences)? For clinical studies Cohen suggested that values of 'd' of 0.2, 0.5 and 0.8 would represent a small, medium or large treatment effect. For laboratory

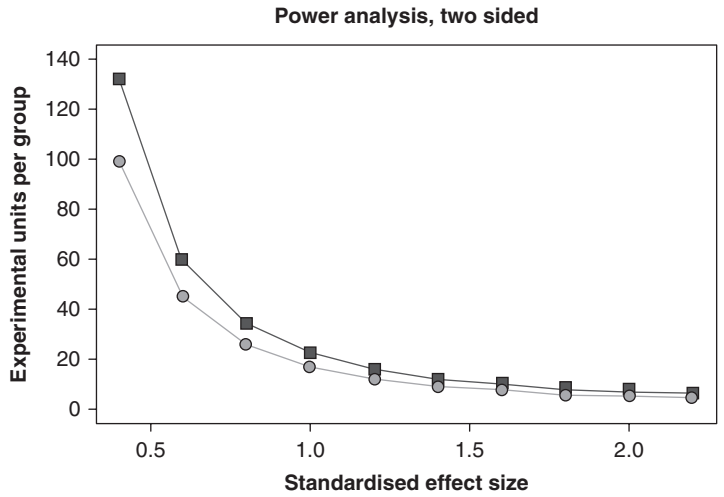


Figure 11.1 Plot of sample size as a function of standardised effect size assuming a two-sample *t*-test, a significance level of 0.05 and a two-sided test. Circles = 80% power, rectangles = 90% power.

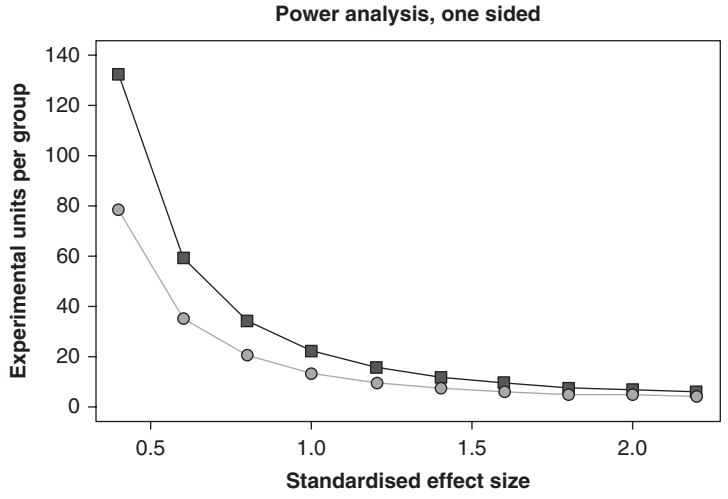


Figure 11.2 Plot of sample size as a function of standardised effect size assuming a two-sample *t*-test, a significance level of 0.05 and a one-sided test. Circles = 80% power, rectangles = 90% power.

animals it is probably more realistic to set these as 0.5, 1.0 and 1.5. The required sample sizes can be seen in Figures 11.1 or 11.2. For example, if an experiment with two groups is planned to be able to detect a medium ES with d = 1.0, then the experiment would require slightly fewer or slightly more than 20 animals in each group, depending on the required power.

Estimating sample size with two treatment groups

Figures 11.1 and 11.2 show sample size as a function of SES for 80% and 90% powers, a 5% significance level, and one- or two-sided tests, respectively. These are probably accurate enough for most purposes given that the SD is being estimated from a previous study. Note that large numbers of animals are needed to detect small effects, with an SES below half an SD in magnitude.

Many software packages are available for power analysis sample size calculations including R and Rcmdr. But in this case researchers need to write a command rather than use a menu. In Rcmdr for a two-sided test with an 80% power and 5% significance level the following command (all with lower case letters) is written in the 'top output box', with the 'hash' marks substituted by appropriate numbers (and power and sig.level altered if desired); 'delta' is the ES and 'sd' is the standard deviation. This is for a two-sided test (the default).

```
power.t.test(delta=##, sd=##, power=0.8, sig.level=0.05)
```

Once the command is written and the hash marks are replaced with the chosen values, the command is marked in the usual way and the 'submit' button is clicked.

If the aim is to estimate the ES for a given value of *n*, with power and sig.level as above, then the command is:

```
power.t.test(n=##, sd=##, power=0.8, sig.level=0.05)
```

If n = 10 and the sd is 0.5 then delta, the ES, is 0.66, as shown below

```
power.t.test(n=10, sd=0.5, power=0.8, sig.level=0.05)

Two-sample t-test power calculation:

     n = 10
     delta = 0.6624728
     sd = 0.5
     sig.level = 0.05
     power = 0.8
     alternative = two.sided

Note: n is the number in *each* group
```

For a one-sided test the command below is used, again with the appropriate numbers substituted, and the power and significance levels altered appropriately.

power.t.test(delta=##, sd=##, power=0.8, sig.level=0.05, alt='one.sided')

Or to estimate the ES for a given n, replace delta=## with n=## and replace the hash marks with appropriate numbers.

An example

Chapter 4 uses a fictitious example of the effect of running in an exercise wheel on learning ability in mice as an example of a completely randomised (CR) design. The mean learning score in the control mice was 246 units and in the Marathon mice it was 221 units, giving a reduction of 25 units. The SD was 10.4 units. The observed SES was therefore 25/10.4 = 2.40 SDs.

Suppose an experiment is now planned to determine the effect of a psychoactive drug on learning ability using the same methods of measuring learning ability and similar mice but with the drug instead of the exercise wheel, what would be an appropriate sample size?

It is assumed here that there will only be two groups. The first problem is to establish the minimum ES likely to be of scientific interest. This has to be determined somewhat subjectively. Based on the previous study, would an ES of 10 units be of scientific/clinical interest? If so, with an SD of 10.4 the SES would be 10/10.4 = 0.96, which could be rounded to 1.0 SD. In this case a one-sided test would be appropriate as it is unclear what negative learning would be. With a one-sided test, according to Figure 9.2 about 14 mice per group would be required to detect a 10 unit response with an 80% power or 20 mice for a 90% power. Using Figures 11.1 and 11.2 it is easy to find the consequences of varying the ES. Greater accuracy is not needed in view of the uncertainty in estimating the SD and ES. But the formula below can be used if this is preferred.

power.t.test(delta=10, sd=10.4, power=0.8, sig.level=0.05, alt='one.sided')

The resulting output is:

```
Two-sample t-test power calculation:
   n = 14.10248
   delta = 10
   sd = 10.4
   sig.level = 0.05
   power = 0.8
   alternative = one.sided
Note: n is the number in *each* group
```

In this case n could justifiably be rounded down to 14 mice per group.

Comparing two or more proportions

A power analysis can be used to determine sample size for comparing two proportions provided there is some prior information on the incidence of the effect in the control group. This is specified as a proportion 'p1'. The ES likely to be of clinical or scientific importance is specified by choosing the percent in the treated group, 'p2'. Suitable sample sizes can then be determined by typing the command below (adjusting p1 and p2) into the 'top output box' in Rcmdr. This should then be marked and submitted. The command, below, shows the sample size when the control group is expected to have an incidence of 15%, and where the aim is to be able to detect an effect of 30% or more, if it is present.

```
power.prop.test(power=0.90,p1=0.15,p2=0.30)
```

The output is:

```
>power.prop.test(power=0.90,p1=0.15,p2=0.30)

Two-sample comparison of proportions power calculation:

   n = 160.7777
   p1 = 0.15
   p2 = 0.3
   sig.level = 0.05
   power = 0.9
   alternative = two.sided

Note: n is the number in *each* group
```

So this would be a large experiment with over 300 animals in total.

More complex situations

With three or more treatment groups it becomes more difficult to specify the ES of clinical or scientific importance. One strategy could be to base the calculations on the control and top dose groups. This would slightly overestimate the required sample size because there would be extra information on inter-individual variation in the middle dose group(s). Remember that great precision is not possible in power calculations because they depend on estimates of the SD and ES which may not be accurate.

It is also possible to do a power analysis for a one-way analysis of variance (ANOVA). There is a command in R (and Rcmdr):

```
power.anova.test(groups = NULL, n = NULL, between.var =
NULL, within.var = NULL, sig.level = 0.05, power = NULL)
```

In this case all except one of the NULL values (usually n = NULL if sample size is to be estimated) will be filled in (exactly as shown), and the value being estimated should be deleted from the command.

As with the two-group case some prior information is needed. For example, if another experiment involving learning ability in mice is proposed, the ANOVA for the learning experiment in Chapter 5 could be used as a template. In such an experiment the 'between.var' is the treatment sum of squares 2861, and the 'within.var' is the residual sum of squares 2614, the number of groups is 3 and the power could be specified as 0.8. The command is therefore:

```
power.anova.test(groups=3,between.var=2862,within.
var=2614,sig.level=0.05, power=0.8)
```

Putting this information into the top box in Rcmdr, marking it and submitting it gives the output.

```
Balanced one-way ANOVA power calculation:

  groups = 3
  n = 5.53158
  between.var = 2862
  within.var = 2614
  sig.level = 0.05
  power = 0.8

Note: n is the number in each group
```

In this case n would be rounded up to 6.

If it were decided to do another learning experiment, but this time with four treatment groups, the command could be tweaked to replace groups = 3 with groups = 4. This would give an estimate of $n = 4.42$. This would need to be rounded up to five mice per group.

If a smaller ES was thought to be of scientific importance, then the 'between.var' could be reduced. If it was reduced to the same as the 'within.var' and with three groups, then the required sample size would be 5.9 mice per group, which would be rounded up to six mice per group.

For even more complex experiments such as randomised block (RB) and factorial designs the resource equation method may be somewhat easier to use.

The resource equation method

This method (Mead, 1988) is useful for experiments which are to be analysed by the ANOVA. It can be used when the power analysis method is either not possible or impractical (e.g. where there is no information on the SD, where the ES cannot realistically be estimated, where the experiment has a complex design, or where there are many outcomes, such as when the outcomes are haematology and biochemistry).

The method depends on the law of diminishing returns. Adding one more experimental unit when the experiment is small can usefully increase power, but adding one more experimental unit to an experiment which is already large may be of little benefit.

The main features of the resource equation method are:

1. It is easy to use even for complex experiments where a power analysis would be difficult.
2. It is only appropriate for experiments producing quantitative data which can be analysed by ANOVA or the *t*-test.
3. It does not require an estimate of the SD or the ES of biological interest.
4. The power, significance level and alternative hypothesis do not need to be specified.
5. It works by ensuring that there is a reasonably good estimate of the pooled SD, against which the differences in means are to be measured.

The resource equation is: $E = N - T - B$, and $10 < E < 20$, where E is the error degrees of freedom, N is the total degrees of freedom (i.e. the total number of experimental units minus one), T is the treatments degrees of freedom (the number of treatment combinations minus one), and B is the blocks degrees of freedom (the number of blocks minus one).

However, a good case can be made for leaving the blocks degrees of freedom out of the equation because although blocking takes degrees of freedom from the error term it nearly always increases the power of the experiment. So in this case:

E = (total number of animals) − (number of treatment groups), and E should be between 10 and 20 (i.e. $10 < E < 20$)

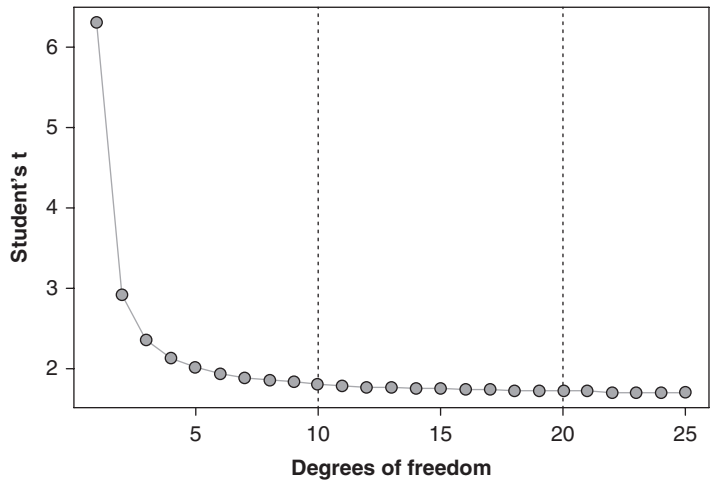


Figure 11.3 The value of Student's *t* for 2–25 degrees of freedom (DF). Note that *t* can be considered to be a measure of uncertainty in an estimate of the standard deviation. This decreases rapidly up to about 10 DF but going beyond 20 DF does not substantially reduce the level of uncertainty. So the resource equation method suggests that experiments should be designed to have 10–20 DF for the error (residual) term.

If E is less than 10, increasing the numbers would lead to worthwhile returns. If it is substantially more than 20, diminishing returns will result in resources being wasted. This can be seen from the shape of the curve of the 5% critical value of the Student's t (see Figure 11.3) which can be regarded as a measure of uncertainty in the estimate of the SD. Increasing the error degrees of freedom (E) from one to 10 leads to a substantial reduction in the level of t which would be declared statistically significant. However, increasing E much beyond 20 hardly alters the critical value of t. Therefore the experiment is approximately optimised if E is between 10 and 20.

This method should be used only as a guide, just as with the power analysis. Some freedom should be allowed to account for special situations. It might sometimes be acceptable for E to be as low as 6 when sufficient numbers of animals are not available, provided it is also realised that the experiment will then lack power. So it may result in a false-negative result. For *in vitro* studies, which are inexpensive and do not involve live animals, E can be substantially higher than 20. When the aim of the experiment is accurate estimation of some parameter such as the magnitude of a treatment effect, other methods of sample size determination may be used. Genetic linkage studies are a good example, where the interest is not so much whether two loci are linked, but how closely they are linked. This may involve large numbers of animals if the linkage estimate is to be accurate.

Example 1

A CR experiment is proposed with five treatment groups and six rats per group so $E = 30 - 5 = 25$. So the group size is slightly too large. With four rats per group $E = 20 - 5 = 15$, which is in the middle of the suggested range.

Example 2

A CR factorial experiment is proposed with three treatments and both sexes with five rats per group. So there are six groups of five rats = 30 rats in total, and six groups so $E = 24$. So the sample size would be marginally too large. Four rats per group would give $E = 24 - 6 = 18$.

Example 3

An experiment is proposed with both sexes and four treatments done as an RB design with three blocks. The number of animals is 2 (sexes) × 4 (treatments) × 3 (blocks) = 24 animals in total. The experiment has 2 × 4 treatments. Here $E = 24 - 8$ which is 16. So this is an acceptable sample size.

12

Seventeen steps in designing a randomised controlled animal experiment

1. Formulate the question to be answered by the experiment

The question must be specific such as 'will this drug alter blood pressure in laboratory rats?', and it must be possible for it to be answered by the experiment.

2. Decide which animals to use (see Chapter 2)

In this case the question specifies that rats are to be used, but gender, strain(s), and age(s) will also need to be specified.

3. Identify the 'experimental unit'

The experimental unit is 'the smallest division of the experimental material such that any two subjects can receive different treatments'. It is the unit of randomisation and of the statistical analysis. If animals within the same cage cannot receive different treatments, then the cage of animals will be the experimental unit, and the statistical analysis should use the mean of the animals within a cage. Treated and control experimental units must be intermingled in the environment.

In a crossover experiment (see Chapter 7), the experimental unit is often the animal or a cage of animals for a period of time. In a split plot experiment (see Chapter 8) there will be two different types of experimental units. An experiment may involve some cages of males and some of females. Within each cage there may be two rats which can receive different treatments. Thus, for comparing the males

and females the cage is the experimental unit and the analysis will be based on cage means, and for comparing the treatments, and any interaction between gender and treatment the individual rat is the experimental unit.

4. Minimise inter-individual variability (see Chapter 3)

Sample size and statistical power depend on the uniformity of the experimental material. Uniformity can be achieved by choosing animals of the same sex, a similar age and/or weight and from the same source. The use of genetically uniform, isogenic (i.e. inbred or F1 hybrid) animals is recommended (see Chapter 2). Specific pathogen free (SPF) rats and mice, which are free from clinical and subclinical diseases should always be used because infection increases variability and may lead to an atypical response. Animals housed in an enriched environment tend to be more uniform than those housed in un-enriched conditions. Social animals should be housed in groups for welfare reasons (see Directive 2010/63/EU), even though this may increase variability. Randomised block designs can increase uniformity by matching animals in each block and eliminating some environmental variability. Each block will also fit into a small space, so minimising environmental variation. The use of RB designs is strongly recommended.

5. Chose independent variables

The independent variables or 'factors' are the variables to be studied. Often one factor will be designated 'treatment' with at least two levels, one of which may be called a 'control'.

A second factor 'gender' could have two levels, 'female' and 'male'. If the experiment has two or more factors (e.g. treatment and gender) it is said to have a factorial design (see Chapter 6). Factorial designs provide extra information, often at little extra cost. Such designs can lead to experiments with many treatment groups, yet they remain relatively easy to analyse and interpret.

6. Choose the dependent variables (characters or outcomes)

These are the 'characters', 'traits' or 'outcomes' which are to be measured/counted. This is usually specified in the question being asked. So in the above question, blood pressure is the outcome of interest. Additional outcomes may also be measured. In some experiments multiple outcomes such as haematology and clinical biochemistry are measured. Gene microarray data may involve thousands of observations per experimental unit. This type of data will require specialised statistical analysis, not covered in this book.

Quantitative (measurement) data are usually more informative than categorical data. However, the main design principles remain the same whatever the nature of the data that are to be collected. Be aware that there may be problems in the separate statistical analysis of many outcome variables due to the number of statistical tests. It may lead to some false-positive results.

7. Decide whether a pilot study is necessary

A pilot study is recommended if the investigator and/or staff are unfamiliar with the species and/or techniques which are to be used.

8. Choose the experimental design

The main options are:

A completely randomised design (see Chapter 5)

With this design the treatments are allocated to the experimental units at random without taking into account the individual characteristics. Animals and their environments should be as uniform as possible. Each cage will need to contain two or more identically treated animals to comply with EU Directive 2012/63/EU; and the cage, not the animal, will be the experimental unit.

If animals receiving different treatments *can* be housed together, then a randomised block design (see below) is suggested.

A randomised block design (recommended, see Chapter 7)

In this design the experiment is split up into a number of mini-experiments or blocks, usually with one experimental unit on each treatment. Each block is repeated several times. It is easier to manage, is usually more powerful, and is less likely to lead to bias than the completely randomised design.

Two alternatives are suggested depending on whether animals receiving different treatments can be housed in the same box.

Animals housed together cannot receive different treatments

In this case animals would need to be housed at least two per cage for welfare and legislative requirements but housing more than two animals together is inefficient in terms of experimental design but may be considered for welfare reasons (see point 12 below). The cage will be the experimental unit. Each block will consist of n cages, where n is the number of treatments. The number of blocks (sample size) needed can

be estimated using the resource equation method, or a power analysis if some previous data are available. If both sexes are to be included, then a $2 \times t$ factorial design will be needed where t is the number of treatments.

Animals housed together can receive different treatments

In this case the experimental unit is the animal and each cage will be a block with n animals of the same sex, where n is the number of treatments. If both genders are to be included then half of the cages should be males and half females. This is a split plot design which requires a slightly more complex statistical analysis, as discussed in Chapter 8.

Other designs (see Chapter 8)

These include Latin square, crossover, and sequential designs. Other designs are important in special circumstances.

9. Determine a suitable sample size (see Chapter 11)

An experiment needs to be large enough to detect any scientifically important treatment response, but not so large that resources and animals may be wasted. Two methods are available.

The power analysis

This is most suitable for large, simple, expensive experiments such as clinical trials or animal studies just prior to clinical trials. It can be used both for measurement and categorical data. There must be a good prior estimate of the standard deviation and it must be possible to make a sensible estimate of the minimum size of response likely to be of interest. It is less useful for fundamental studies where there is no good information on variability and/or it is difficult to determine an appropriate effect size. However, a decision can be made about the minimum standardised effect size (SES) sought. It is also less useful for complex experiments with several treatment groups and a factorial arrangement of treatments because in such circumstances it is difficult to specify an effect size.

The resource equation

This method is suitable for complex smaller experiments with a measurement outcome, particularly where there are many treatment groups, such as in a factorial design.

Assuming that a randomised block design is to be used, Table 12.1 shows the minimum number of blocks for various numbers of treatments in order to have a reasonable number of error degrees of freedom. For example, if there are four treatments,

Table 12.1 Suggested minimum sizes (according to the resource equation) of randomised block designs for 2–10 treatments.

No. of treatments	No. of blocks[1]	Error DF	Total DF	No. of animals[2]	Factorial experiment?[3]	Possible split plot[4]
2	10	10	19	20	–	Yes
3	6	10	17	18	–	Yes
4	5	12	19	20	2 × 2	
5	4	12	19	20	–	
6	3	10	17	18	2 × 3	
7	3	12	20	21	–	
8	3	14	23	24	2 × 4, 2 × 2 × 2	
9	3	16	26	27	3 × 3	
10	2	9	19	20	2 × 5	

[1]A block is either a cage if the animals within it can receive different treatments (so are the experimental units), or a group of cages with one animal per cage and one cage on each treatment.

[2]This assumes that the animal is the unit. If two animals receiving the same treatment must be housed together, to conform with Directive 2010/63/EU, the number of animals needed must be doubled.

[3]Possible factorial layouts. E.g. 2 × 2 could be two treatments and two genders.

[4]At least six blocks are needed if the main plot treatment effect is to be reasonably well estimated.

DF: degrees of freedom.

then five blocks would provide 12 degrees of freedom for the error term. The treatments could be a 2 × 2 factorial with two treatments and both genders for instance. The experiment would need 40 animals if the cage is the experimental unit but only 20 animals if the animal is the experimental unit. In this latter case it would not be possible to include both genders without more blocks.

10. Randomise the treatments to the experimental units (see Chapter 3)

For a completely randomised design the experimental units should be numbered 1 to *n* (the total number of experimental units), then treatments should be assigned to them at random, possibly using EXCEL as explained in Chapter 3. This randomisation is done in the office, not by physical randomisation in the animal house. Note that the treatments should be in a random order with respect to the experimental unit number.

With a randomised block design randomisation is done separately for each block as is also explained in Chapter 3.

11. Plan the statistical analysis (see Chapter 4 and subsequent chapters)

The experiment and the statistical analysis should both be planned at the same time. If necessary, a statistician should be consulted at this planning stage, well before any data have been collected. Each experiment should be analysed before starting the next one so that any new knowledge can be incorporated as the project progresses.

12. Refinement

Consider ways in which pain and discomfort can be minimised. Enriched housing, humane endpoints, anaesthesia and analgesia should be discussed with the animal house veterinarian, the animal house staff and others concerned with animal welfare.

13. Check availability of resources

Ensure that good facilities (such as space, cages and other equipment) and suitably trained staff are available, that the required animals can either be bred in-house or purchased from a commercial supplier (purchased animals may need to be quarantined) and that sufficient funds are available to carry out the project. If special apparatus is required ensure that this will be available.

14. Ensure that all legal aspects are covered

The animal house director, investigators and staff who are directly involved in the project need to have all the necessary legal authority to proceed with the study.

15. Protocols and standard operating procedures (SOPs)

SOPs and protocols should be prepared and discussed with the staff. A diary with notes of any unexpected events should be kept. The reasons for the loss of any experimental units (animals) should be recorded. Decide how the data will be collected and stored prior to the statistical analysis. The ARRIVE guidelines can be used as a checklist before the experiment is started.

If the experimental protocol requires repeated doses then care needs to be taken to avoid mistakes. Possibly two staff should be involved, cross-checking each other. Cage labels could be colour-coded to the treatment. These would be removed prior to measuring the outcome so as to 'blind' the staff to the treatment.

16. Prepare a flow chart showing the main actions required during the course of the experiment

An example for the fictional experiment described in Chapter 5 is given in Figure 12.1. Preparation of such a chart is useful for identifying any factors that have not previously been identified (see also EDA in https://eda.nc3rs.org.uk/). For example, the need for a room with changed lighting times, need for weekend access by the investigator and the potential errors associated with the probable need to measure learning ability over a period of several days. Clearly a randomised block design should have been considered.

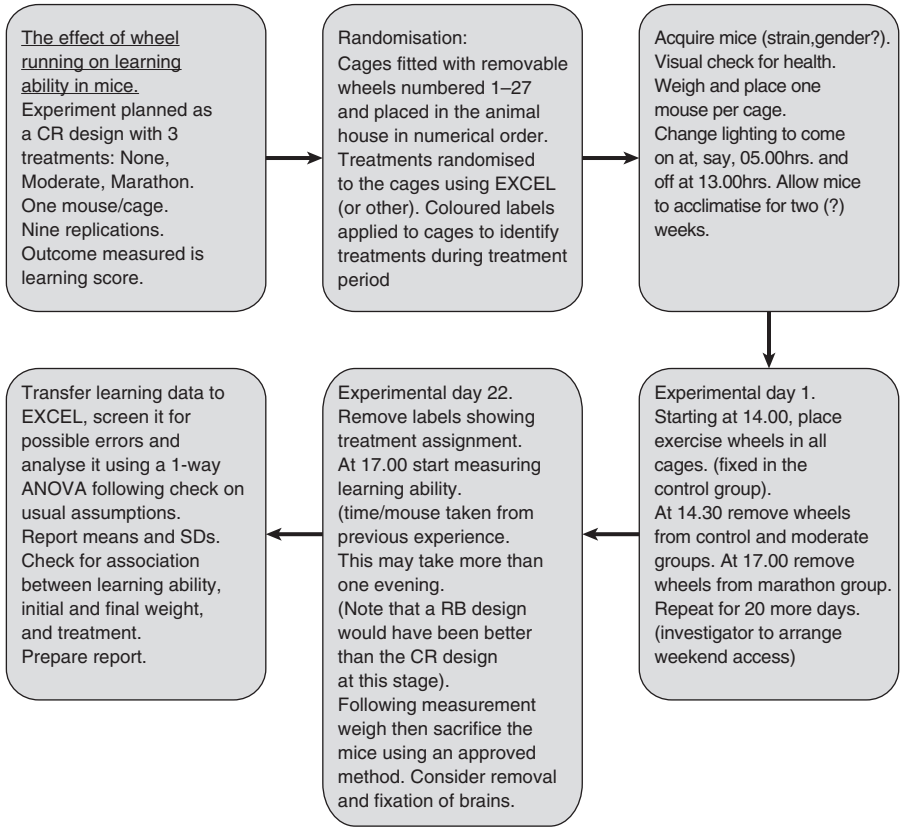


Figure 12.1 Example of a flow chart for the wheel running/learning experiment described in Chapter 5. CR: completely randomised, RB: randomised block, ANOVA: analysis of variance, SDs: standard deviations.

17 Submit the proposals to the local ethics committee (IACUC in the USA)

Leave ample time for the ethical review process to take place and to make any necessary changes to the project suggested by the committee or individuals.

And finally, carry out the experiment, comforted by the thought that you have done everything that you could to ensure that it is well designed, justified, and worthwhile. Some of the aspects that should be taken into consideration when reporting the results of the experiment are addressed in the next, and concluding, chapter.

13

Reporting the results

Any well designed non-trivial experiment should be publishable whether the results are negative or positive. This is partly to avoid unnecessary repetition of negative studies and partly because if the results are later used in a meta-analysis both positive and negative studies are needed. If several studies indicate a positive, but not statistically significant, effect then a meta-analysis may show that the combined effect is likely to be real. It is the scientific quality and validity of the work rather than the nature of the outcome which should be considered.

The ARRIVE guidelines (Kilkenny et al., 2010) provide a checklist to ensure that no important information about the experiment(s) is/are omitted. Table 2 of these guidelines is given in Table 13.1. The following highlights some points worth remembering when publishing.

Presentation of the results

The aim of a manuscript is to show the results as clearly and accurately as possible. If the paper is presenting the results from more than one experiment, then these should be individually numbered so that it is clear which set of results comes from which experiment.

Means and proportions should usually be presented with some measure of variation such as standard deviation (SD), standard error of the mean (SE or SEM), or confidence interval (CI). In order to avoid confusion, it has been suggested that the ± symbol should be abandoned and means should be presented as, for example, 'mean 9.6 (SD 2.1) units'.Where possible plots showing individual data points should be shown in preference or in addition to error bars.

Specifying the variation

The standard deviation

This provides an estimation of the variation among individual experimental units within a group. It does not vary with the numbers in each group. It is used to construct various types of error bars. With reasonable sample sizes the SD can be calculated

Table 13.1 Table 2 of the 'Animal Research: Reporting *In Vivo* experiments: The ARRIVE guidelines'.

	ITEM	RECOMMENDATION
TITLE	1	Provide as accurate and concise a description of the content of the article as possible.
ABSTRACT	2	Provide an accurate summary of the background, research objectives (including details of the species or strain of animal used), key methods, principal findings, and conclusions of the study.
INTRODUCTION Background	3	(a) Include sufficient scientific background (including relevant references to previous work) to understand the motivation and context for the study, and explain the experimental approach and rationale. (b) Explain how and why the animal species and model being used can address the scientific objectives and, where appropriate, the study's relevance to human biology.
Objectives	4	Clearly describe the primary and any secondary objectives of the study, or specific hypotheses being tested.
METHODS Ethical statement	5	Indicate the nature of the ethical review permissions, relevant licences (e.g. Animal [Scientific Procedures] Act 1986), and national or institutional guidelines for the care and use of animals, that cover the research.
Study design	6	For each experiment, give brief details of the study design, including: (a) The number of experimental and control groups. (b) Any steps taken to minimise the effects of subjective bias when allocating animals to treatment (e.g. randomisation procedure) and when assessing results (e.g. if done, describe who was blinded and when). (c) The experimental unit (e.g. a single animal, group, or cage of animals). A time-line diagram or flow chart can be useful to illustrate how complex study designs were carried out.
Experimental procedures	7	For each experiment and each experimental group, including controls, provide precise details of all procedures carried out. For example: (a) How (e.g. drug formulation and dose, site and route of administration, anaesthesia and analgesia used [including monitoring], surgical procedure, method of euthanasia). Provide details of any specialist equipment used, including supplier(s).

(Continued)

Table 13.1 (Continued)

	ITEM	RECOMMENDATION
		(b) When (e.g. time of day). (c) Where (e.g. home cage, laboratory, water maze). (d) Why (e.g. rationale for choice of specific anaesthetic, route of administration, drug dose used).
Experimental animals	8	(a) Provide details of the animals used, including species, strain, sex, developmental stage (e.g. mean or median age plus age range), and weight (e.g. mean or median weight plus weight range). (b) Provide further relevant information such as the source of animals, international strain nomenclature, genetic modification status (e.g. knockout or transgenic), genotype, health/immune status, drug- or test-naıve, previous procedures, etc.
Housing and husbandry	9	Provide details of: (a) Housing (e.g. type of facility, specific pathogen free (SPF), type of cage or housing, bedding material, number of cage companions, tank shape and material, etc. for fish). (b) Husbandry conditions (e.g. breeding programme, light/dark cycle, temperature, quality of water, etc. for fish, type of food, access to food and water, environmental enrichment). (c) Welfare-related assessments and interventions that were carried out before, during, or after the experiment.
Sample size	10	(a) Specify the total number of animals used in each experiment and the number of animals in each experimental group. (b) Explain how the number of animals was decided. Provide details of any sample size calculation used. (c) Indicate the number of independent replications of each experiment, if relevant.
Allocating animals to experimental groups	11	(a) Give full details of how animals were allocated to experimental groups, including randomisation or matching if done. (b) Describe the order in which the animals in the different experimental groups were treated and assessed.

	ITEM	RECOMMENDATION
Experimental outcomes	12	Clearly define the primary and secondary experimental outcomes assessed (e.g. cell death, molecular markers, behavioural changes).
Statistical methods	13	(a) Provide details of the statistical methods used for each analysis. (b) Specify the unit of analysis for each dataset (e.g. single animal, group of animals, single neuron). (c) Describe any methods used to assess whether the data met the assumptions of the statistical approach.
RESULTS Baseline data	14	For each experimental group, report relevant characteristics and health status of animals (e.g. weight, microbiological status, and drug- or test-naive) before treatment or testing (this information can often be tabulated).
Numbers analysed	15	(a) Report the number of animals in each group included in each analysis. Report absolute numbers (e.g. 10/20, not 50%). (b) If any animals or data were not included in the analysis, explain why.
Outcomes and estimation	16	Report the results for each analysis carried out, with a measure of precision (e.g. standard error or confidence interval).
Adverse events	17	(a) Give details of all important adverse events in each experimental group. (b) Describe any modifications to the experimental protocols made to reduce adverse events.
DISCUSSION	18	(a) Interpret the results, taking into account the study objectives and hypotheses, current theory, and other relevant studies in the literature. (b) Comment on the study limitations including any potential sources of bias, any limitations of the animal model, and the imprecision associated with the results. (c) Describe any implications of your experimental methods or findings for the replacement, refinement, or reduction (the 3Rs) of the use of animals in research.
Generalisability/ translation	19	Comment on whether, and how, the findings of this study are likely to translate to other species or systems, including any relevance to human biology.
Funding	20	List all funding sources (including grant number) and the role of the funder(s) in the study.

Source: © Kilkenny et al., 2010

for each group. However, randomised block designs only provide a pooled estimate of the SD obtained from the square root of the error mean square in the analysis of variance (ANOVA) table. With all designs, residual plots (see Chapter 4) should have been used to indicate whether the variation is approximately the same in each group (assuming quantitative data). If it is, then again the best estimate of the SD is the pooled estimate obtained from the error mean square in the ANOVA, and there seems to be little point in quoting a different SD for each mean. However, if the SDs differ among groups, this should be clearly indicated. In such cases the data may need to be transformed to a different scale before statistical analysis.

Standard error of the mean

The SE is an estimation of the variation among means. It is the SD divided by the square root of n, the number in each group. The SEM × 1.96 is known as the margin of error. It is one arm of an approximate 95% confidence interval (CI) for the mean.

Standard error of the difference between two means

This is estimated as the square root of the mean of the sum of the variances of each mean, i.e. $\sqrt{(s_1^2/n_1 + s_2^2/n_2)}$.

95% confidence interval for a mean

A 95% confidence interval (CI) is the interval within which 95% of means would be expected to lie if the experiment were to be repeated many times.

If the CI includes zero, then the mean of a group, or difference between two groups, is not significantly different from zero at the specified level. The margin of error is one arm of the CI. It is approximately 1.96 × SEM.

Least significant difference

This is (Student's t) × (the standard error of the difference between two means), where t has the appropriate number of DF and significance level. When used for error bars, it has the advantage that if the bars overlap then the differences will not be significantly different and if they do not overlap the differences will be statistically significant.

Proportions and percentages

These should also be presented with some estimate of their reliability such as an standard error (SE) or, preferably, a confidence interval (CI), or showing the results of a statistical test to indicate which are significantly different. The variance of a proportion is npq where n is the number of observations and p and q are the number

of positive and negative counts, respectively. For example, in a group of $n = 50$ C57BL/6 mice, if the proportion having chewed whiskers is $p = 0.2$ (20%), then the variance of this is npq or $50 \times 0.2 \times 0.8 = 8$ and the SE is the square root of 8 which is 2.82.

P-values

Exact *P*-values should be quoted in preference to $P < 0.05$ or the use of asterisks unless the *P*-values are very low such as $P < 0.001$. Note that $P = 0.06$ does not mean that there is no difference between groups. It just means that this experiment failed to detect an effect at this level of probability. Had sample size been larger an effect might have been detected at the 5% level.

Many dependent variables (multivariate data)

Some experiments involve more than one dependent variable. For example, an experiment involving measurement of haematology and blood biochemistry may have 20 or more dependent variables such as counts of red blood cells, haematocrit, lymphocytes, etc. Modern studies using gene arrays to assess changes in mRNA may involve several thousand dependent variables measured on each experimental unit. Such studies often require special multivariate statistical analysis such as 'principal components analysis', 'discriminant function analysis', and various clustering methods which take account of any relationships between the variables, and can reduce large quantities of raw data down to a level where the results can be interpreted. These are specialised methods which require statistical advice.

Toxicological data are usually presented as large tables of means and SDs for many separate outcomes (e.g. haematology, organ weights, clinical biochemistry). This is quite difficult to interpret. A novel approach has been suggested, by converting responses to standardised effect sizes. As all characters are then expressed in the same units (SDs), they can then be averaged (Festing, 1974). This approach leads to a number of informative graphical representations of the results as well as a statistical test of the overall response.

Tables and figures

Tables and figures should be designed to convey information as clearly as possible. In most cases three significant digits (i.e. the left-hand digits in a number) will suffice when presenting means and SDs. It is easier to compare a set of means when they are in the same column, rather than in the same row. When means are shown in columns, SDs should be in the next column, not in the same column but on a different row. Bar diagrams should only be used if they illustrate something of particular interest. They often take up a lot of room compared with a table of means and SDs.

Appendix 1

R-Commander: a free menu-driven statistical software package

R-Commander (Rcmdr) is a powerful free menu-driven 'front end' to the 'R' statistical programming software. It has been used in all the examples of this book. However, it is often necessary to prepare the data for entry into Rcmdr, and this can be done using EXCEL.

R is widely used by professional statisticians. It is well established and can be used to produce complex graphics and advanced statistical analyses. It can be downloaded and installed free of charge. However, it is command-driven and its flexibility and power mean that a substantial effort is required to master it. R has many associated packages which can carry out specialised statistical analyses, such as survival analysis or various multivariate analyses.

Rcmdr was designed for teaching statistics (probably to students of statistics!). It will do most of the statistical analyses needed by those using animals for biomedical research. It is free and easy to use, once a few concepts have been understood. It has simple graphics output which are good enough for exploring data and presenting the outcomes, and these can be saved in various formats. Full publication quality graphics can be done using R. Most textbooks on R have a section on graphical methods.

Installing R and Rcmdr

R can be downloaded from http://www.r-project.org/. It can be installed on a wide range of computers and operating systems and can even be installed on, and run from, a memory stick. Full details of how to download and install R are provided on their website.

Note that if Rcmdr is to be used a 'custom' installation of R is necessary. The SDI not the default MDI version of R should be chosen. If the wrong version is installed Rcmdr will not sit permanently on the screen.

Once R is installed with its icon, it should be started. The 'Packages' menu should be clicked and then the 'Install package' item should be clicked. Choose a local mirror site from the list of sites which appears. This will bring up a long (>1000)

list of packages that can be installed. Choose Rcmdr and follow the instructions for installing it. There are many files which will be installed automatically.

Once Rcmdr has been fully downloaded it can be invoked by typing 'library(Rcmdr)' in the R window followed by the return key.

The Rcmdr interface

This has menus across the top (File, Edit, Data, etc. and Help). The Help menu gives extensive information on the use of Rcmdr as well as access to the Rcmdr website. There are now a number of add-on packages for Rcmdr, such as one for survival analysis.

Below the menu there are some buttons showing the name of the active data set, Edit data set, View data set and Models. The editing facilities are minimal. Single numbers can be changed, but any major editing of the data is best done in EXCEL and imported back into Rcmdr. The easiest way of importing data into Rcmdr is via the clipboard.

There are two output windows. The upper one shows the code that is generated by the menu commands. It is possible to write commands for R in this window. These can then be highlighted and executed, using the Submit button. In this book the commands for the power analysis method of determining sample size (see Chapter 11) are run this way.

The lower window shows the numerical output. This can be marked and copied in the usual way, and pasted into Word or EXCEL. A non-proportional font such as Courier New should be used for the data to ensure proper alignment.

Graphical output comes in a separate window. A new graph will overwrite an existing one. The graphs may be copied in various formats.

Using Rcmdr

Rcmdr has limited facilities for manipulating data so it is advisable to do this first, using EXCEL. The data should be in the standard format with columns for ID or 'Animal.number' (note that each column is headed by a single word; if two words are necessary they should be joined with a point '.'), treatment (two or more columns if it is a factorial experiment and/or blocked design), and observation(s). The tables in this book can be used as examples, although some have been split to fit in with the text. The missing observation code is 'NA'.

Output in R and Rcmdr is in alphabetic order, so factors may need to be coded. For example, in the experiment on the effect of wheel running on learning ability in mice the treatments were None, Moderate and Marathon. So output will be in the order Marathon, Moderate and None. If this is not acceptable the factors may be coded A, B and C or A.None, B.Moderate and C.Marathon. If numerical factors are entered as numbers Rcmdr must be told that they are factors not variables (*Data, manage variables in active data set, convert numeric variables to factors*).

Appendix 2
Further reading

Books on laboratory animal science

Hau, J. and Shapiro, S.J. *Handbook of Laboratory Animal Science. Vol 1. Essential Principles and Practices*. Boca Raton FL: CRC Press, 2011
A comprehensive volume covering most topics associated with laboratory animals and their use.

Howard, B. Nevalainen, T. and Perretta, G. *The COST Manual of Laboratory Animal Care and Use*. Boca Raton FL: CRC Press, 2011
This book covers a wide range of topics relevant to the use of animals in research. These range from the design of animal facilities, ethical evaluation of scientific procedures, reduction, animal models, handling, to basic procedures and many more techniques and topics.

Hubrecht, R. and J.Kirkwood. *The UFAW Handbook on the Care and Management of Laboratory and Other Research Animals*, Chichester: Wiley-Blackwell, 2010
A large single volume definitive textbook on all aspects of the care and management of laboratory and some other animals. There are sections on 'Implementing the 3Rs', 'Species kept in laboratories', 'Reptiles' and 'Amphibians and fish'.

van Zutphen, L. F. M.; Baumans, V.; Beynen, A. C., editors. *Principles of Laboratory Animal Science*. Amsterdam, New York, London: Elsevier; 1993
A reference book which covers the principles of laboratory animal science with chapters on legislation, the biology of laboratory animals, standardisation, nutrition, genetics, diseases and microbiology and the design of animal experiments. It does not have chapters on individual species.

Wolfensohn, S. and Lloyd. M. *Handbook of Laboratory Animal Management and Welfare*. 4th Edition. Oxford: Wiley-Blackwell, 2013
This is a general introduction to laboratory animal science often used by animal technicians, but useful to anyone starting work with laboratory animals.

Books on experimental design

Bate, S. T. and Clark, R. *The Design and Statistical Analysis of Animal Experiments.* Cambridge: Cambridge University Press, 2014
A good book covering in a non-mathematical way the design and analysis of *in vivo* experiments. It also describes how to use InvivoStat, a software package, assembled by the authors.

Clarke, G.M. and Kempson, R.E. *Introduction to the Design and Analysis of Experiments.* London: Arnold. 1997
A modern book aimed at undergraduates in statistics and mathematics rather than research scientists.

Cochran, W.G and Cox, G.M. *Experimental Designs*. New York: John Wiley & Sons, 1957
A classical book on experimental design. Though an older edition (before computers were readily available), it still covers the main principles. It can be recommended to anyone using more advanced designs.

Cox, D, R. *Planning Experiments*. New York: John Wiley & Sons, 1958
Although published many years ago, this book is still in print and is not in any way dated. It is a readable book aimed directly at the research worker, with few mathematical formulae. However, it also manages to deal with some advanced concepts and experimental designs. Strongly recommended.

Cox, D.R. and Reid, N. *The Theory of the Design of Experiments*. Boca Raton FL: Chapman and Hall/CRC Press, 2000
An advanced book covering the principles of experimental design, with quite a bit of mathematical notation. However, the text is clear and easy to understand even for the non-mathematician.

The Experimental Design Assistant (EDA) https://eda.nc3rs.org.uk/
A good free online resource from the NC3Rs to support researchers in the planning of animal experiments - Its main thrust is to help build (and critique) a visual representation of the design of experiments but it also contains explanations of good experimental design and statistical principles.

Fisher RA. *The Design of Experiments*. 7th edn. New York: Haffner Publishing, 1960
A classical non-mathematical book which is still capable of providing many insights into the design of experiments and statistical inference.

Mead, R. *The Design of Experiments*. Cambridge: Cambridge University Press, 1988
An advanced textbook on experimental design aimed largely at postgraduate students specialising in statistics.

Mead, R., Gilmour, S.G. and Mead, A. *Statistical Principles for the Design of Experiments*. Cambridge: Cambridge University Press, 2012
A major textbook on experimental design. Only for professional statisticians.

Ruxton, G.D. and Colegrave N. *Experimental Design for the Life Sciences*. 2nd edn. Oxford: Oxford University Press, 2006
A small but excellent non-mathematical text covering many aspects of experimental design in the life sciences.

Books on statistics

Altman, D.G. *Practical Statistics for Medical Research*. London: Chapman and Hall, 1991
A textbook aimed at medical researchers with a bias towards work involving humans, but sufficiently general to cover all the biological sciences, including quite advanced statistical concepts. Readable and not too mathematical.

Crawley, M.J. *Statistics. An Introduction using R*. Chichester: John Wiley & Sons, 2005
Covers statistics from an elementary level right up to the use of advanced techniques using R. Anyone working through this book should end up with a good understanding of the subject and of how to program in R.

Cumming J. *Understanding the New Statistics: Effect Sizes, Confidence Intervals, and Meta-Analysis*. New York: Routledge, 2012
Some scientists claim that too much emphasis has been placed on statistical significance at the expense of estimating effect sizes and confidence intervals. This interesting book argues this case as well as making a case for the more widespread use of meta-analysis.

Dalgaard, P. *Introductory Statistics with R*. New York: Springer-Verlag, 2010
An excellent book for learning statistics using the R statistical programming language.

Howell, D.C. *Fundamental Statistics for the Behavioral Sciences*. Pacific Grove CA: Duxbury Press, 1999
A textbook which recognises that virtually all statistical analyses will be done by computer, so emphasises the importance of understanding the data, choosing appropriate statistical methods, and interpreting the output from statistical packages. However, the book does not specifically cover experimental designs.

Maxwell, S.E. and Delaney, H.D. *Designing Experiments and Analyzing Data*. Belmont CA: Wadsworth Publishing, 1989
A classical reference book/textbook covering both experimental design and statistics, but with a bias towards the behavioural sciences. Although it has some advanced topics, according to the authors '…the necessary background for the book is minimal'.

Mead, R. and Curnow, R.N. *Statistical Methods in Agriculture and Experimental Biology*. London: Chapman and Hall, 1983
The aim of this book '…is to describe and explain those statistical ideas which we believe are an essential part of the intellectual equipment of a scientist working in agriculture or on the experimental side of biology'. There is a strong emphasis on experimental design, quite a few formulae, and lots of worked examples.

References

Aldinger KA, Sokoloff G, Rosenberg DM, Palmer AA and Millen KJ (2009) Genetic variation and population substructure in outbred CD-1 mice: implications for genome-wide association studies. *PLoS One* 4: e4729.

Altman DG (1982) Statistics in medical journals. *Statistics in Medicine* 1: 59–71.

Beck JA, Lloyd S, Hafezparast M, et al. (2000) Genealogies of mouse inbred strains. *Nature Genetics* 24: 23–25.

Begley CG and Ellis LM (2012) Drug development: raise standards for preclinical cancer research. *Nature* 483: 531–533.

Chia R, Achilli F, Festing MF and Fisher EM (2005) The origins and uses of mouse outbred stocks. *Nature Genetics* 37: 1181–1186.

Cohen J (1988) *Statistical Power Analysis for the Behavioral Sciences*. Hillsdale NJ: Lawrence Erlbaum Associates.

Committee for the Update of the Guide for the Care and Use of Laboratory Animals (2011) *Guide for the Care and Use of Laboratory Animals*. 8th ed. Washington DC: The National Academies Press.

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

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

Crawley MJ (2005) *Statistics. An Introduction using R*. 1st ed. Chichester: John Wiley & Sons, Ltd.

European Commission (2014) National competent authorities for the implementation of Directive 2010/63/EU on the protection of animals used for scientific purposes. A working document on the development of a common education and training framework to fulfil the requirements under the Directive.

Festing MF (2014) Extending the statistical analysis and graphical presentation of toxicity test results using standardized effect sizes. *Toxicologic Pathology* 42: 1238–1249.

Festing MFW (1974) The genetic reliability of commercially bred laboratory mice. *Laboratory Animals* 8: 265–270.

Festing MFW (1992) The scope for improving the design of laboratory animal experiments. *Laboratory Animals* 26: 256–267.

Festing MFW (1994a) Are animal experiments well designed? In: *Welfare and science. Proceedings of the 5th Symposium of the Federation of European Laboratory Animal Science Associations* (ed J Bunyon) Brighton, 1993, pp. 32–36. London: Royal Society of Medicine Press.

Festing MFW (1994b) Reduction of animal use: experimental design and quality of experiments. *Laboratory Animals* 28: 212–221.

Festing MFW, Diamanti P and Turton JA (2001) Strain differences in haematological response to chloramphenicol succinate in mice: implications for toxicological research. *Food and Chemical Toxicology* 39: 375–383.

Festing MFW and Fisher EMC (2000) Mighty mice. *Nature* 404: 815.

Festing MFW and Greenwood R (1976) Home-cage wheel activity recording in mice. *Laboratory Animals* 10: 81–85.

Finney DJ (1978) *Statistical Method in Biological Assay*. High Wycombe: Charles Griffin & Company Ltd.

Fisher RA (1960) *The Design of Experiments*. New York: Hafner Publishing.
Freedman LP, Cockburn IM and Simcoe TS (2015) The economics of reproducibility in preclinical research. *PLoS Biology* 13: e1002165.
Gartner K (1990) A third component causing random variability beside environment and genotype. A reason for limited success of a 30 year long effort to standardize laboratory animals. *Laboratory Animals* 24: 71–77.
Hau J and Shapiro SJ (2011) *Handbook of Laboratory Animal Science*. 3rd ed. Vol. 1. Boca Raton FL: CRC Press.
Hau J and Van Hoosier GL Jr (2002) *Handbook of Laboratory Animal Science*. 2nd ed. Boca Raton FL: CRC Press.
Hoglund AU and Renstrom A (2001) Evaluation of individually ventilated cage systems for laboratory rodents: cage environment and animal health aspects. *Laboratory Animals* 35: 51–57.
Howard B, Nevalainen T and Perretta G (2011) *The COST Manual of Laboratory Animal Care and Use*. Boca Raton FL: CRC Press.
Hubrecht R and Kirkwood J (2010) *The UFAW Handbook on the Care and Management of Laboratory and Other Research Animals*. 8th ed. Chichester: Wiley-Blackwell.
Ioannidis JP, Greenland S, Hlatky MA, et al. (2014) Increasing value and reducing waste in research design, conduct, and analysis. *Lancet* 383: 166–175.
Jay GE (1955) Variation in response of various mouse strains to hexobarbitol (Evpal). *Proceedings of the Society of Experimental Biology and Medicine* 90: 378–380.
Kempermann G, Kuhn HG and Gage FH (1997) More hippocampal neurons in adult mice living in an enriched environment. *Nature* 386: 493–495.
Kilkenny C, Browne W, Cuthill IC, Emerson M and Altman DG (2010) Improving bioscience research reporting: the ARRIVE guidelines for reporting animal research. PLoS Biol 8: e1000412.
Kilkenny C, Parsons N, Kadyszewski E, et al. (2009) Survey of the quality of experimental design, statistical analysis and reporting of research using animals. *PLoS One* 4: e7824.
Les EP (1972) A disease related to cage population density: tail lesions of C3H/HeJ mice. *Laboratory Animal Science* 22: 56–60.
Lindsey JR, Baker HJ, Overcash RG, Cassell GH and Hunt CE (1971) Murine chronic respiratory disease. *American Journal of Pathology* 64: 675–716.
McCance I (1995) Assessment of statistical procedures used in papers in the Australian Veterinary Journal. *Australian Veterinary Journal* 72: 322–328.
Malkinson AM (1979) Prevention of butylated hydroxzytoluene-induced lung damage in mice by cedar terpene administration. *Toxicology and Applied Pharmacology* 49: 551–560.
Markel P, Shu P, Ebeling C, et al. (1997) Theoretical and empirical issues for marker-assisted breeding of congenic mouse strains. *Nature Genetics* 17: 280–284.
Masoro EJ (1993) Nutrition, including diet restriction, in mammals. *Aging Clinical and Experimental Research* 5: 269–275.
Mead R (1988) *The Design of Experiments*. Cambridge: Cambridge University Press.
Mead R, Curnow RN and Hasted AM (1993) *Statistical Methods in Agriculture and Experimental Biology*. London: Chapman and Hall.
Montgomery DC (1997) *Design and Analysis of Experiments*. New York: Wiley.
Morton DB (2000) A systematic approach for establishing humane endpoints. *ILAR Journal* 41: 80–86.
Nadeau JH, Singer JB, Martin A and Lander ES (2000) Analysing complex genetic traits with chromosome substitution strains. *Nature Genetics* 24: 221–225.
Papaioannou VE and Festing MFW (1980) Genetic drift in a stock of laboratory mice. *Laboratory Animals* 14: 11–13.
Pawitan Y, Michiels S, Koscielny S, Gusnanto A and Ploner A (2005) False discovery rate, sensitivity and sample size for microarray studies. *Bioinformatics* 21: 3017–3024.
Perel P, Roberts I, Sena E, et al. (2007) Comparison of treatment effects between animal experiments and clinical trials: systematic review. *British Medical Journal* 334: 197.
Prendergast BJ, Onishi KG and Zucker I (2014) Female mice liberated for inclusion in neuroscience and biomedical research. *Neuroscience and Biobehavioral Reviews* 40: 1–5.

Pritchett-Corning KR, Clifford CB and Festing MF (2013) The effects of shipping on early pregnancy in laboratory rats. *Birth Defects Research (Part B) Developmental and Reproductive Toxicology* 98: 200–205.

Russell WMS and Burch RL (1959) *The Principles of Humane Experimental Technique*. Special Edition. Potters Bar: Universities Federation for Animal Welfare.

Scott S, Kranz JE, Cole J, et al. (2008) Design, power, and interpretation of studies in the standard murine model of ALS. *Amyotrophic Lateral Sclerosis* 9: 4–15.

Silver LM (1995) *Mouse Genetics*. Oxford: Oxford University Press.

Snedecor GW and Cochran WG (1980) *Statistical Methods*. Ames IA: Iowa State University Press.

Snell GD and Stimpfling JH (1966) Genetics of tissue transplantation. In: Green EL (ed) *Biology of the Laboratory Mouse*. New York: McGraw-Hill, pp. 457–491.

Sprott RL (1967) Barometric pressure fluctuations: effect on the activity of laboratory mice. *Science* 157: 1206–1207.

Stokes WS (2000) Reducing unrelieved pain and distress in laboratory animals using humane endpoints. *ILAR Journal* 41: 59–61.

Taylor BA (1996) Recombinant inbred strains. In: Lyon MF (ed) *Genetic Variants and Strains of the Laboratory Mouse*. 2nd edn. Oxford: Oxford University Press, pp. 1597–1659.

Tramontin AD and Brenowitz EA (2000) Seasonal plasticity in the adult brain. *Trends in Neuroscience* 23: 251–258.

Vesell ES (1968) Factors altering the responsiveness of mice to hexobarbital. *Pharmacology* 1: 81–97.

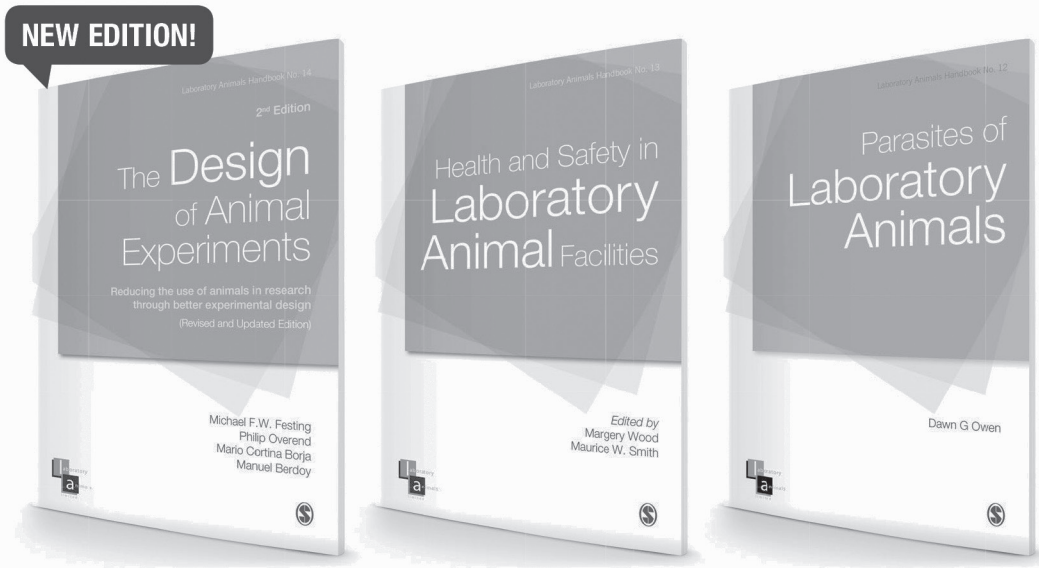
Wurbel H (2000) Behaviour and the standardization fallacy. *Nature Genetics* 26: 263.

Yalcin B, Nicod J, Bhomra A, et al. (2010) Commercially available outbred mice for genome-wide association studies. *PLoS Genetics* 6: e1001085.

Index

[Page numbers in italics refer to tables.]

Explore the Laboratory Animals Handbooks



The Design of Animal Experiments (2nd Edition)

Reducing the Use of Animals in Research through Better Experimental Design

Michael Festing

Health and Safety in Laboratory Animal Facilities

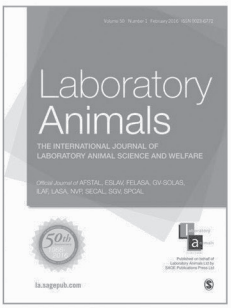
Edited by
Margery Wood and
Maurice W. Smith

Parasites of Laboratory Animals

Dawn G Owen

Find out more and buy online at uk.sagepub.com/lahandbooks

More about Laboratory Animals



The international journal of laboratory animal science and welfare, *Laboratory Animals* publishes peer-reviewed original papers and reviews on all aspects of the use of animals in biomedical research. The journal promotes improvements in the welfare or well-being of the animals used, it particularly focuses on research that reduces the number of animals used or which replaces animal models with in vitro alternatives.

Find out more and sign up for email Table of Contents alerts at la.sagepub.com

SAGE Publishing | laboratory animals limited