Statistical Methods in Biology: Design and Analysis of Experiments and Regression

Book · May 2014

DOI: 10.1201/b17336

4 authors, including:

Sue Welham
Rothamsted Research
79 PUBLICATIONS **2,906** CITATIONS

SEE PROFILE

Salvador A. Gezan
Trigen Improvement
178 PUBLICATIONS **1,283** CITATIONS

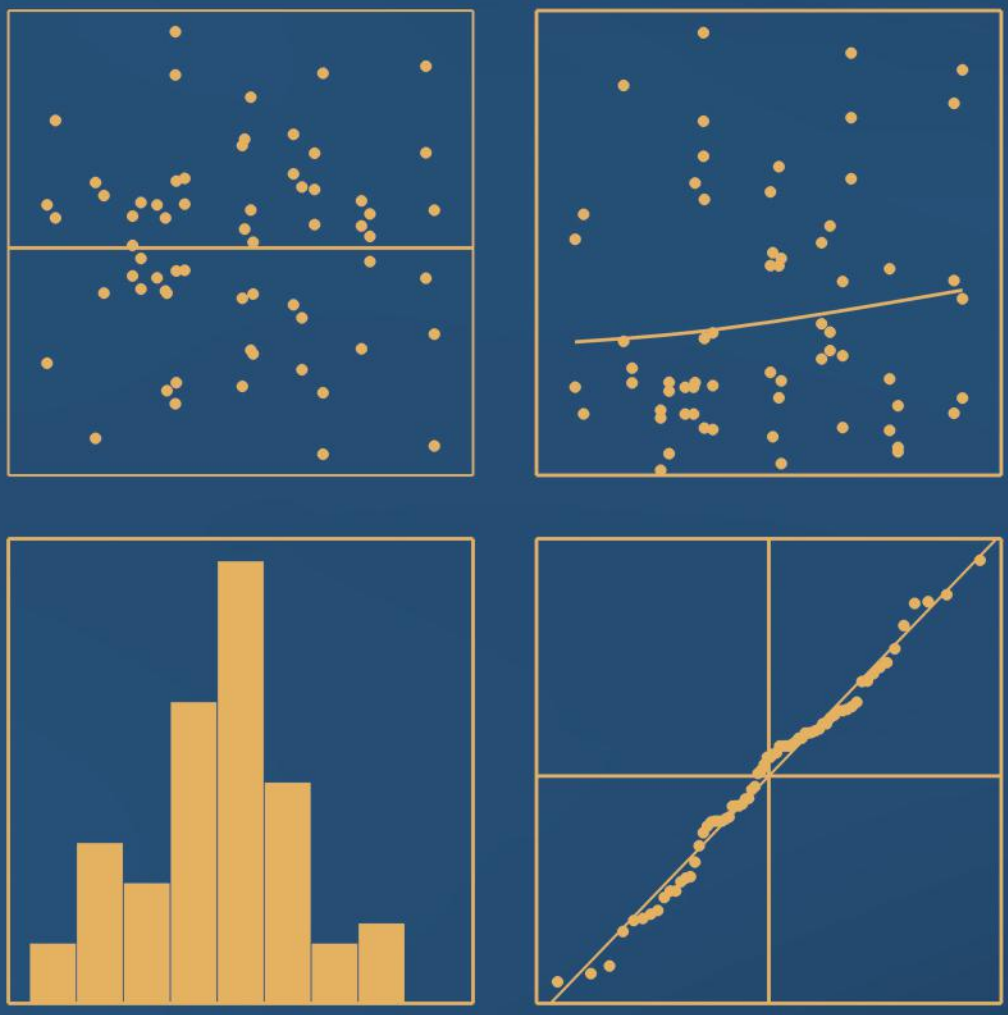
SEE PROFILE

Andrew Mead
Rothamsted Research
163 PUBLICATIONS **5,491** CITATIONS

SEE PROFILE

STATISTICAL METHODS IN BIOLOGY

Design and Analysis of Experiments and Regression



S. J. Welham, S. A. Gezan, S. J. Clark and A. Mead

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

STATISTICAL METHODS IN BIOLOGY

Design and Analysis of Experiments and Regression

STATISTICAL METHODS IN BIOLOGY

Design and Analysis of Experiments and Regression

S. J. Welham

Rothamsted Research, Harpenden, UK

S. A. Gezan

University of Florida, USA
(formerly Rothamsted Research, Harpenden, UK)

S. J. Clark

Rothamsted Research, Harpenden, UK

A. Mead

Rothamsted Research, Harpenden, UK
(formerly Horticulture Research International, Wellesbourne, UK & University of Warwick, UK)

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business

A CHAPMAN & HALL BOOK

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Version Date: 20140703

International Standard Book Number-13: 978-1-4398-9805-5 (eBook - PDF)

To my parents and Simon, with love and thanks

SJW

To Diablita, Psycha and Luna for all their love and unconditional support

SAG

For Mum, Dad and Tony, with love. For Joe, Mike, Moira and Sue, with thanks

SJC

To Sara, Tom and my parents, with love and thanks for your continuing support

AM

Contents

Preface

This book provides an introductory, practical and illustrative guide to the design of experiments and data analysis in the biological and agricultural plant sciences. It is aimed both at research scientists and at students (from final year undergraduate level through taught masters to PhD students) who either need to design their own experiments and perform their own analyses or can consult with a professional applied statistician and want to have a clear understanding of the methods that they are using. The material is based on courses developed at two British research institutes (Rothamsted Research and Horticulture Research International [HRI – then Warwick HRI, and now the School of Life Sciences, University of Warwick]) to train research scientists and post-graduate students in these key areas of statistics. Our overall approach is intended to be practical and intuitive rather than overly theoretical, with mathematical formulae presented only to formalize the methods where appropriate and necessary. Our intention is to present statistical ideas in the context of the biological and agricultural sciences to which they are being applied, drawing on relevant examples from our own experiences as consultant applied statisticians at research institutes, to encourage best practice in design and data analysis.

The first two chapters of this book provide introductory, review and background material. In Chapter 1, we introduce types of data and statistical models, together with an overview of the basic statistical concepts and the terminology used throughout. The training courses on which this book is based are intended to follow preliminary courses that introduce the basic ideas of summary statistics, simple statistical distributions (Normal, Poisson, Binomial), confidence intervals, and simple statistical tests (including the t-test and F-test). Whilst a brief review of such material is covered in Chapter 2, the reader will need to be comfortable with these ideas to reap the greatest benefit from reading the rest of the book. Some readers may feel that their knowledge of basic statistics is sufficiently comprehensive that they can skip this review chapter. However, we recommend you browse through it to familiarize yourself with the statistical terminology that we use.

The main body of the book follows. Chapters 3 to 11 introduce statistical approaches to the design of experiments and the analysis of data from such designed experiments. We start from basic design principles, introduce some simple designs, and then extend to more complex ones including factorial treatment structures, treatment contrasts and blocking structures. We describe the use of analysis of variance (ANOVA) to summarize the data, including the use of the multi-stratum ANOVA to account for the physical structure of the experimental material or blocking imposed by the experimenter, introduce simple diagnostic methods, and discuss potential transformations of the response. We explain the analysis of standard designs, including the randomized complete block, Latin square, split-plot and balanced incomplete block designs in some detail. We also explore the issues of sample size estimation and the power of a design. Finally, we look at the analysis of unbalanced or non-orthogonal designs. Chapters 12 to 18 first introduce the idea of simple linear regression to relate a response variable to a single explanatory variable, and then consider extensions and modifications of this approach to cope with more complex data sets and relationships. These include multiple linear regression, simple linear regression with groups, linear mixed models and models for curved relationships. We also extend related themes from the earlier chapters, including diagnostic methods specific to regression. We emphasize throughout that the same type of models and principles are used for

both designed experiments and regression modelling. We complete the main body of the book with a discussion of generalized linear models, which are appropriate for certain types of non-Normal data.

We conclude with a guide to practical design and data analysis (Chapter 19), which focuses on the selection of the most appropriate design or analysis approach for individual scientific problems and on the interpretation and presentation of the results of the analysis.

Most chapters include exercises which we hope will help to consolidate the ideas introduced in the chapter. In running the training courses from which this book has been developed, we often find that it is only when students perform the analyses themselves that they fully appreciate the statistical concepts and, most importantly, understand how to interpret the results of the analyses. We therefore encourage you to work through at least some of the exercises for each chapter before moving to the next one. There are fewer exercises in the earlier chapters and the required analyses build in complexity, so we expect you to apply knowledge gained throughout the book when doing exercises from the later chapters. All of the data sets and solutions to selected exercises are available online. Some of the solutions include further discussion of the relevant statistical issues.

We have set up a website to accompany this book (www.stats4biol.info) where we show how to do the analyses described in the book using GenStat®, R and SAS®, three commonly used statistical packages. Whilst users familiar with any of these packages might not refer to this material, others are encouraged to review it and work through the examples and exercises for at least one of the packages. Any errors found after publication will also be recorded on this website.

By the time you reach the end of the book (and online material) we intend that you will have gained

- A clear appreciation of the importance of a statistical approach to the design of your experiments,
- A sound understanding of the statistical methods used to analyse data obtained from designed experiments and of the regression approaches used to construct simple models to describe the observed response as a function of explanatory variables,
- Sufficient knowledge of how to use one or more statistical packages to analyse data using the approaches that we describe, and most importantly,
- An appreciation of how to interpret the results of these statistical analyses in the context of the biological or agricultural science within which you are working.

By doing so, you will be better able both to interact with a consultant statistician, should you have access to one, and to identify suitable statistical approaches to add value to your scientific research.

This book relies heavily on the use of real data sets and material from the original courses and we are hence indebted to many people for their input. Particular thanks go to Stephen Powers and Rodger White (Rothamsted Research) and John Fenlon, Gail Kingswell and Julie Jones (HRI) for their contributions to the original courses; also to Alan Todd (Rothamsted Research) for providing many valuable suggestions for suitable data sets. The majority of real data sets used arose from projects (including PhDs) at Rothamsted Research, many in collaboration with other institutes and funded from many sources; we thank Rothamsted Research for giving us general permission to use these data. We also thank, in alphabetical order, R. Alarcon-Reverte, S. Amoah, J. Baverstock, P. Brookes,

J. Chapman, R. Curtis, I. Denholm, N. Evans, A. Ferguson, S. Foster, M. Glendining, K. Hammond-Kosack, R. Harrington, Y. Huang, R. Hull, J. Jenkyn, H.-C. Jing, A.E. Johnston, A. Karp, J. Logan, J. Lucas, P. Lutman, A. Macdonald, S. McGrath, T. Miller, S. Moss, J. Pell, R. Plumb, P. Poulton, A. Salisbury, T. Scott, I. Shield, C. Shortall, L. Smart, M. Torrance, P. Wells, M. Wilkinson and E. Wright, for specific permission to use data from their own projects or from those undertaken within their group or department at Rothamsted. Rothamsted Research receives grant-aided support from the Biotechnology and Biological Sciences Research Council of the United Kingdom. We thank various colleagues, past and present, at Horticulture Research International, Warwick HRI and the School of Life Sciences, University of Warwick, for permission to use data from their research projects, particularly Rosemary Collier and John Clarkson. We thank M. Heard (Centre for Ecology and Hydrology), A. Ortega Z. (Universidad Austral de Chile) and R. Webster for permission to use data. Examples and exercises marked '*' use simulated data inspired by experiments carried out at Rothamsted Research or HRI. The small remainder of original examples and exercises (also marked '*') were invented by the authors but are typical of the type of experiments we are regularly asked to design and the data we analyse as part of our consultancy work. In the few cases where we have not been able to find examples from our own work we have drawn on data from published sources. We would like to thank Simon Harding for technical help in setting up a repository for our work and our website and Richard Webster, Alice Milne, Nick Galwey, James Bell and Kathy Ruggeiro and an anonymous referee for reading draft chapters and providing many helpful comments and suggestions.

Finally, we would like to make some individual acknowledgements. SJW, SJC and SAG thank Rothamsted Research, and in particular Chris Rawlings, for support and encouragement to pursue this project. AM thanks his colleagues at Horticulture Research International and the University of Warwick, particularly John Fenlon, for support in the development of the original training courses, and hence the development of this project, and his co-authors for the invitation to join this project. SJW thanks Simon Harding for his support, help and long-term forbearance. SAG thanks Emma Weeks for her encouragement, and the other co-authors for their patience and the fruitful discussions we had on this project. SJC thanks Tony Scott for his patience and support, Elisa Allen for her contribution to the presentation of our courses and useful comments on some chapters, and past students for their enthusiasm and constructive feedback which led to improvements in our courses and ultimately this book. AM also thanks his family, Sara and Tom, for their continuing support and understanding.

S J Welham
Welwyn Garden City, UK

S A Gezan
Harpenden, UK and Gainesville, Florida, USA

S J Clark
Harpenden, UK

A Mead
Leamington Spa, UK

Authors

Suzanne Jane Welham obtained an MSc in statistical sciences from University College London in 1987 and worked as an applied statistician at Rothamsted Research from 1987 to 2000, collaborating with scientists and developing statistical software. She pursued a PhD from 2000 to 2003 at the London School of Hygiene and Tropical Medicine and then returned to Rothamsted, during which time she coauthored the in-house statistics courses that motivated the writing of this book. She is a coauthor of about 60 published papers and currently works for VSN International Ltd on the development of statistical software for analysis of linear mixed models and presents training courses on their use in R and GenStat.

Salvador Alejandro Gezan, PhD, is an assistant professor at the School of Forest Resources and Conservation at the University of Florida since 2011. Salvador obtained his bachelor's from the Universidad of Chile in forestry and his PhD from the University of Florida in statistics-genetics. He then worked as an applied statistician at Rothamsted Research, collaborating on the production and development of the in-house courses that formed the basis for this book. Currently, he teaches courses in linear and mixed model effects, quantitative genetics and forest mensuration. He carries out research and consulting in statistical application to biological sciences with emphasis on genetic improvement of plants and animals. Salvador is a long-time user of SAS, which he combines with GenStat, R and MATLAB® as required.

Suzanne Jane Clark has worked at Rothamsted Research as an applied statistician since 1981. She primarily collaborates with ecologists and entomologists at Rothamsted, providing and implementing advice on statistical issues ranging from planning and design of experiments through to data analysis and presentation of results, and has coauthored over 130 scientific papers. Suzanne coauthored and presents several of the in-house statistics courses for scientists and research students, which inspired the writing of this book. An experienced and long-term GenStat user, Suzanne has also written several procedures for the GenStat Procedure Library and uses GenStat daily for the analyses of biological data using a wide range of statistical techniques, including those covered in this book.

Andrew Mead obtained a BSc in statistics at the University of Bath and an MSc in biometry at the University of Reading, where he spent over 16 years working as a consultant and research biometrician at the Institute of Horticultural Research and Horticulture Research International at Wellesbourne, Warwickshire, UK. During this time, he developed and taught a series of statistics training courses for staff and students at the institute, producing some of the material on which this book is based. For 10 years from 2004 he worked as a research biometrician and teaching fellow at the University of Warwick, developing and leading the teaching of statistics for both postgraduate and undergraduate students across a range of life sciences. In 2014 he was appointed as Head of Applied Statistics at Rothamsted Research. Throughout his career he has had a strong association with the International Biometric Society, serving as International President and Vice

President from 2007 to 2010 inclusive, having been the first recipient of the 'Award for Outstanding Contribution to the Development of the International Biometric Society' in 2006, serving as a Regional Secretary of the British and Irish Region from 2000 to 2007 and on the International Council from 2002 to 2010. He is a (co)author of over 80 papers, and coauthor of *Statistical Principles for the Design of Experiments: Applications to Real Experiments* published in 2012.

1

Introduction

This book is about the design of experiments and the analysis of data arising in biological and agricultural sciences, using the statistical techniques of analysis of variance (ANOVA) and regression modelling. These techniques are appropriate for analysis of many (although not all) scientific studies and form an important basic component of the statistician's toolbox. Although we provide some of the mathematical formulae associated with these techniques, we have also tried to interpret the equations in words and to give insight into the underlying principles. We hope that this will make these useful statistical methods more accessible.

This chapter presents an introduction to the different types of data and statistical models that are considered in this book, together with an overview of the basic statistical concepts and terminology which will be used throughout. In particular, we discuss

- Types of scientific study
- Populations and samples
- Mathematical and statistical models used to describe biological processes
- The linear model – which underlies all the models and methods introduced in this book
- Parameter estimation and statistical inference
- ANOVA – the major statistical tool used to evaluate and summarize linear models

At the end of this chapter, we preview the contents of the remaining chapters.

1.1 Different Types of Scientific Study

We shall be concerned with data arising from both experimental and observational studies. Although they have many common features, there are some subtle differences that influence the conclusions that can be drawn from the analyses of data from these two types of study.

An **experimental study** is a scientific test (or a series of tests) conducted with the objective of studying the relationship between one or more outcome variables and one or more condition variables that are intentionally manipulated to observe how changing these conditions affects the results. The outcome of a study will also depend on the wider environment, and the scientist will endeavour to control other variables that may affect the outcomes, although there is always the possibility that uncontrolled, perhaps unexpected, variables also influence the outcome. Adequate planning is therefore crucial to

experimental success. There are a few key elements that need to be clearly specified and considered for an experimental study (Kuehl, 2000). These are the

- aims of the experiment – usually expressed as questions or hypotheses
- physical structure of the study materials
- subjects or entities to be used
- set of conditions to be investigated
- other variables that might affect the outcome
- outcome variables to be measured
- protocols that define how the measurements are taken
- available resources (e.g. money, time, personnel, equipment, materials)

The aims of an experimental study need to be clearly specified, often in the form of hypotheses to be tested or a set of questions to be answered; this is a vital part of the planning process. The physical structure and subjects to be used should be chosen so that the results of the experiment can be related to a wider context (see Section 1.2). In addition, the set of conditions to be investigated must be chosen to answer directly the scientific questions. Other variables likely to affect the outcome should be identified and evaluated so that they can be controlled, as far as possible, and therefore do not interfere with the measured outcome. If they cannot be controlled then they should be measured. Consideration of the variables to be measured is often overlooked at the planning stage, but is important because it may affect both the statistical analysis and the efficiency of the design. As discussed later (Chapter 18), the analysis required for binary data (e.g. absence or presence of disease) or count data (e.g. numbers of insects or weeds present) may be different from that for a continuous variable (e.g. shoot length). A full definition of measurement protocols is good practice and should reduce differences in procedure between scientists working on the same experiment, and improve repeatability of the results. Finally, the resources available will usually limit the size and scope of the experiment.

Design of experiments is a process that brings together all the elements above to produce an experiment that efficiently answers the questions of interest and aims to obtain the maximum amount of information for the resources available, or to minimize the resources needed to obtain the information desired. The main statistical principles used in constructing a good design are replication, randomization and blocking. These concepts are discussed in detail in Chapter 3.

An **observational study** differs from an experimental study in that the application of conditions that affect the outcome is not directly controlled by the scientist. However, all the elements listed above for experimental studies should still be considered when you plan an observational study, although opportunities for the random allocation of conditions to subjects will be limited and sometimes non-existent. In observational studies, the set of conditions to be investigated is first defined, and then subjects with these characteristics are sought and measurements made. Observational studies are often used in ecology where it is difficult to set up an experiment whilst retaining natural habitats. For example, a study might aim to determine the difference in beetle populations using selected field margins as the subjects under two conditions: with and without hedges. In this context, it is harder than in experimental studies to control other variables that may

affect the outcome. For example, the set of hedges available may be composed of several plant types, which might in turn affect the species and abundance of beetles present. In addition, hedges are already in place, and fields with hedges may differ systematically in other characteristics from fields without hedges – in an extreme case they might be on different farms, with different farming methods used. The scientist should therefore consider that differences between conditions in an observational study might be caused by other unrecorded, or possibly unobserved, variables. In experimental studies, where we have greater control over conditions, this can still be true, but we can use randomization to guard against such unknown differences between subjects. But where there are potential uncontrolled sources of variability, the scientist should be wary of inferring direct causal relationships. Hill (1965) gave criteria that should be satisfied by a causative relationship in the context of epidemiology, and many of these criteria can be applied more widely and may be helpful in deciding whether a causal relationship is plausible for any observational study.

The separation between experimental and observational studies is not complete, as some studies may have both experimental and observational components. However, both types of study incorporate structure, and we should take account of this structure in the planning, design, statistical analysis and interpretation of such studies.

1.2 Relating Sample Results to More General Populations

For most scientific studies there is an implicit assumption that the results obtained can be applied to a population of subjects wider than those included in the study, i.e. that the conclusions will apply more generally (although usually with caveats) to the real world. For example, in a field trial to investigate disease control it will generally not be possible to have very large plots, nor to assess visually every plant in a plot, and so a random sample of plants is selected from each plot. It is assumed that the sampled plants are representative of all the plants in the plot and so the results from the sample are inferred to apply to the whole plot. In turn, we should usually have several plots within the trial with the same treatment applied and hope to infer the results from this sample of plots to the whole field. However, it is well-known that field experiment results vary markedly over years and locations, so the trial would ideally be performed at several locations across several years to provide a representative sample of environments. The combined results from the whole set of trials can then be claimed to apply to the region where they were carried out, rather than to a single field in a single year.

In planning any scientific study, it is therefore important to consider the frame of reference when experimental subjects are selected. The scientist should identify the **population** (wider group of subjects) to which they hope the experimental results will apply. Ideally, the subjects should then consist of a **sample**, or subset, drawn from this population. If the process of selecting a sample, known as **sampling**, is made at random, then it is reasonable to assume that the sample will have similar properties to the whole population, and we can use it to make statistical inferences about the population. Generally, as the number of observations in the sample increases, the inferences made about the population become more secure. If a sample is not taken at random, then this sense of the sample being representative of the population may be lost.

1.3 Constructing Models to Represent Reality

A **model** is an abstract representation of a hypothesized process that underpins a biological or physical phenomenon, that is, a way of describing a real system in words, diagrams, mathematical functions, or as a physical representation. In biology, models usually correspond to a simplification of the real process, as no existing model can represent reality in all details. However, this does not mean that models cannot be useful. A good model summarizes the major factors affecting a process to give a representation that provides the level of detail required for the objective of a particular study.

Mathematical models use mathematical notation and expressions to describe a process. A **statistical model** is a mathematical model that allows for variability in the process that may arise from sampling variation, biological variation between individuals, inaccuracies in measurement or influential variables being omitted (knowingly or not) from the model. Therefore, any statistical model has a measure of uncertainty associated with it.

Models are additionally often classified as either process (or mechanistic) models or empirical models. A **process model** purports to give a description of the real underlying process. This type of model can be useful in testing our knowledge: if a model can be built to reproduce the behaviour of the system accurately, then our knowledge of the process (theory) is at least consistent with reality. Conversely, and arguably more usefully, failure of a process model may indicate gaps in knowledge that can be pursued by further experimentation. Process models are often complex, with many parameters, but can sometimes be fitted using statistical principles (see e.g. Brown and Rothery, 1993, Chapter 10).

Statistical models usually fall under the category of **empirical models**, which use the principle of correlation to construct a simple model to describe an observed response in terms of one or more explanatory variables. Empirical models use the correlation between the explanatory (input) variable(s) and the measured response (output) variable to build a model without explicit reference to the true underlying process (although knowledge of this process may be used both to select suitable input variables and to identify the appropriate form of the relationship). This can be useful to identify variables that are influential where no detailed knowledge of the process exists, although some care should be taken with interpretation as there may be no direct causative relationship between the input and output variables; instead they may both be driven by some other hidden (latent) or unmeasured variable.

We shall consider statistical models of the general form

$$\text{response} = \text{systematic component} + \text{random component}.$$

This model can exist in abstract form, but we usually relate it to a set of measurements that have been made. The **response**, or response variable, relates to one type of numerical outcome from the study, sometimes also called the **set of observations**. The **systematic component** is a mathematical function of one or more explanatory variables that provide a representation of the experimental conditions. The systematic component describes the relationship between the response and these explanatory variables and hence between the response and the experimental conditions. Where the conditions have a direct numerical evaluation, such as count, weight or height, the explanatory variable is termed **quantitative**. We refer to quantitative variables as **variates**. Where the conditions are classified into groups or categories the explanatory variable is termed **qualitative**. In this case, the explanatory variable indicates the group to which each subject belongs. We shall refer to

qualitative variables as **factors** and identify the distinct groups in the factor as the **factor levels**. For example, sex would be a factor with two levels: male and female. Note that it is sometimes convenient to group a quantitative variable into categories so as to treat it as a qualitative variable, for example, heights can be classified as short, medium or tall. However, this change cannot always be made in reverse; some explanatory variables, such as sex, are inherently qualitative. Similarly, if a scientist had compared three types of fertilizer, or one fertilizer across three different plant varieties, then the levels of the explanatory variable (fertilizer type or plant variety) cannot be translated into meaningful numbers. In the context of experimental studies, the conditions imposed by the experimenter are usually represented as factors and referred to as **treatments**. We also use this term more generally to describe the set of conditions present in observational studies when represented by factors. In some contexts, where it is more natural, we use the alternative term **groups** instead of treatments.

In general, the systematic component of the statistical models that we consider can be partitioned further into explanatory and structural components as

$$\text{systematic component} = \text{explanatory component} + \text{structural component}.$$

The **explanatory component** corresponds to the conditions of interest, or treatments, in the study. The **structural component** is used to account for the structure of the study, such as sub-sampling within an observational study or blocking within a designed experiment. The structural component is not always present: it may be omitted in the (rare) case that the experimental material consists of an unstructured sample. This partition facilitates the accurate specification of the whole model, as it encourages us to consider the two components separately: the explanatory component relates to our hypothesis (or hypotheses) of interest, and the structural component relates to the structure of the experimental material.

The **random component**, also known as error or noise, corresponds to variation in the response that is not explained by the systematic component. This component may have several sources, such as inherent between-subject variability, measurement errors and background variation within the environment of the study. Mathematically, we usually describe the random component in terms of some appropriate probability distribution (see Chapters 2 and 4).

The systematic component is used to predict the response for any set of experimental conditions, and the random component is used to estimate the uncertainty in those predictions. Here are two simple examples of statistical models.

EXAMPLE 1.1: QUALITATIVE EXPLANATORY VARIABLE

Consider an experiment to investigate nutrient feeding strategies for plants grown in pots. A scientist has obtained a new liquid feed and wishes to evaluate its effect on plant growth. The instructions provided by the manufacturer suggest three feeding regimes labelled A, B and C. The scientist decides to grow 12 plants of a single plant variety, each one in a separate pot, and to allocate four plants at random to each of the three suggested regimes. After six weeks, the height of each plant is measured. Here, the response variable is plant height and the only explanatory variable is the feeding regime, which is a qualitative variable with three levels.

We might hypothesize that the plant height for a given feeding regime can be expressed symbolically as

$$\text{height} = \text{overall mean} + \text{effect of feeding regime} + \text{deviation}.$$

This is a simple (empirical) statistical model with height as the response. For an unstructured sample of 12 pots, there is no need for a structural component. So here the systematic part of the model (i.e. overall mean + effect of feeding regime) relates only to the explanatory component, with plant height modelled as an overall mean modified by some specific amount for each feeding regime. The random part (labelled deviation) allows for the deviation of individual observations from the feeding regime value given by the systematic component. Using mathematical notation (see also Section 2.1) we can write this model as

$$y_{jk} = \mu + \tau_j + e_{jk} \ . \tag{1.1}$$

Here, we have identified each plant by labelling it by the treatment applied ($j = 1, 2, 3$ for regimes A, B, C, respectively) and then we number the plants within each treatment group (using $k = 1, 2, 3, 4$). Hence, y_{jk} represents the height of the kth plant with the jth feeding regime. We use μ to represent the population mean height (the 'overall mean'), and τ_j represents the difference in response for the jth feeding regime relative to the overall mean (the 'effect of the feeding regime'). Finally, e_{jk} is the deviation associated with the kth replicate plant under the jth feeding regime.

The symbols μ and τ_1, τ_2, τ_3 (usually written as τ_j, $j = 1 \ldots 3$) are unknown population parameters that have to be estimated from the observed sample from the experiment. This model represents the height of a plant under the jth regime using the systematic component $\mu + \tau_j$, so a different value pertains to each regime, as shown in Figure 1.1a.

In Example 1.1, the explanatory variable (feeding regime) is a qualitative variable, or factor, with three levels (A, B and C). Without further information we cannot infer relationships between these factor levels and so we model the response by fitting a separate effect for each level. However, if the different feeding regimes correspond to different application rates for the liquid feed, then the scientist could evaluate the quantities corresponding to each feed rate and turn them into quantitative values (numbers). We can then consider other models for these data as shown in Example 1.2.

EXAMPLE 1.2: QUANTITATIVE EXPLANATORY VARIABLE

Suppose now that the scientist in Example 1.1 has evaluated the volumes (or doses) for feeding regimes A, B and C as 20, 40 and 60 mL per plant, respectively. The explanatory variable now corresponds to a quantitative variable (i.e. dose) with numeric values, and we can reasonably consider the response as a function of this continuous variable, expressed symbolically as

$$\text{height} = \text{f(dose)} + \text{deviation},$$

where f(dose) indicates some mathematical function of dose. Here, we assume the simplest case, namely that the function is a straight line relationship (see Figure 1.1b). We can formally write this simple model as

$$y_{jk} = \alpha + \beta x_j + e_{jk} \ . \tag{1.2}$$

We again label each plant by the treatment applied (here $j = 1, 2, 3$ for doses 20, 40 and 60 mL, respectively) and then number plants within each treatment group (using $k = 1, 2, 3, 4$) so y_{jk} is the height of the kth replicate plant with the jth dose. Now, x_j is the numerical quantity of the jth dose ($x_1 = 20$, $x_2 = 40$, $x_3 = 60$), α is the plant height at zero dose (the intercept of the line in Figure 1.1b with the y-axis at $x = 0$), β is the linear response to increasing the dose by 1 mL (the slope of the line in Figure 1.1b), and e_{jk} is the deviation

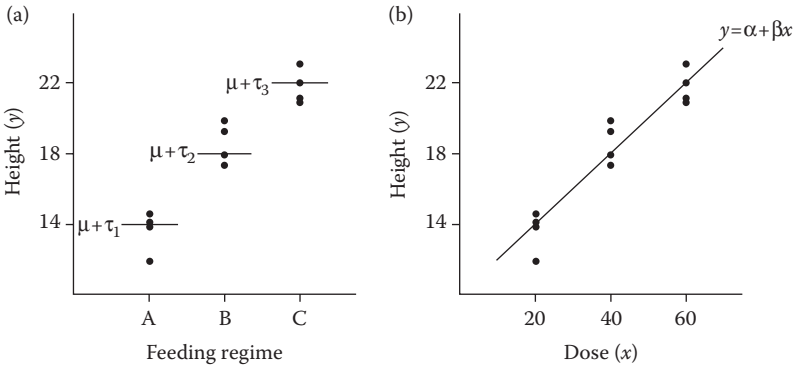


FIGURE 1.1
Two linear models with observed (•) and population (—) responses (heights) for the plant growth experiment for (a) a qualitative explanatory variable representing three feeding regimes (A, B and C, Example 1.1), and (b) a quantitative explanatory variable representing three doses (20, 40 and 60 mL, Example 1.2).

from the linear trend for the kth replicate plant with the jth dose. The symbols α and β are unknown population parameters that have to be estimated from the observed sample.

The model represented by Equation 1.2 differs from the model represented by Equation 1.1 in several important respects, even though it could arise from the same experiment. In Example 1.1, feeding regime was considered to be a qualitative variable (and so here we call Equation 1.1 the qualitative model), and a separate effect was allowed for each level. In Example 1.2, we used additional information, that is the numeric values of dose, to fit height as a linear function of dose (and so here we call Equation 1.2 the quantitative model). The qualitative model might be considered more flexible, as it does not make any assumption about the shape of the relationship. However, the quantitative model has the advantage that it is more **parsimonious**, i.e. that it uses fewer parameters to describe the pattern. It has the further advantage that we can also make predictions at intermediate doses (e.g. 50 mL) using the fitted model (under the assumption that the straight line model is appropriate).

1.4 Using Linear Models

Equations 1.1 and 1.2 are simple examples of **linear models**, an important sub-class of the statistical models introduced in Section 1.3. In this context, the response variable is sometimes called the dependent variable and the explanatory variables are sometimes called independent or predictor variables. The explanatory and structural components of a linear model each consist of a set of terms added together (an **additive structure**) and each term consists of either a single unknown parameter (such as τ_j in Equation 1.1), or an unknown parameter multiplied by a known variable (such as βx_j in Equation 1.2) – this is the **linear structure**. The random component, or deviation, is added to the systematic component to give the full model. In general, linear models might contain terms for several qualitative or quantitative explanatory variables or both. It is important, but slightly confusing, to note that the output

from a complex linear model will generally not be a straight line (e.g. Equation 1.1), although the straight line relationship between a response variable and a single explanatory variable (e.g. Equation 1.2) is the simplest example of a linear model. The class of linear models is a large and flexible one and, although the models themselves are usually approximations, they can adequately represent many real-life situations. The most common uses for linear models are model specification, parameter estimation and prediction.

The main objective in **model specification** is to determine what form of statistical model best describes the relationship between the response and explanatory variable(s). There will often be a biological hypothesis behind a study that suggests a suitable form of model and the explanatory variables that should be included in the model. For example, in Example 1.1 the scientist wanted to investigate whether the different feeding regimes had detectable effects on plant growth. The process of statistical **hypothesis testing** can be used to refine the model by determining whether there is any evidence in the data that the proposed explanatory variables explain patterns in the response. Often several competing models might be compared. If many potential explanatory variables have been measured, **variable screening** may be used to select the variables that best explain the variation in the response. For example, in field studies on insect abundance, many climatic and environmental variables can be measured, and those that are most highly related to insect counts then identified.

Once an appropriate model has been determined, **parameter estimation** (see Section 1.5) is required to interpret the model and, potentially, the underlying biological process. Associated with each parameter estimate is a measure of uncertainty, known as the **standard error**.

The fitted model can be derived by substitution of estimates in place of the unknown parameter values in the model, and uncertainty in the fitted model is derived from the parameter standard errors. **Prediction** involves the use of the fitted model to estimate functions of the explanatory variable(s) – for example, the prediction of a treatment mean together with some measure of its precision. Again, uncertainty in predictions is derived from uncertainty in the parameter estimates.

1.5 Estimating the Parameters of Linear Models

Any linear model has an associated set of unknown parameters for which we want to obtain estimates. For example, in fitting the models represented by Equations 1.1 and 1.2 to the observed data, our aim is to find the 'best' estimates of the parameters μ, τ_1, τ_2 and τ_3, or α and β, respectively. In Chapters 4 (qualitative model) and 12 (quantitative model) we present detailed descriptions of how to obtain estimates of these parameters; here, we outline the basic principles. Before we consider the estimation process, some basic notation is required. In general, we represent estimated parameter values by placing a 'hat' ($\hat{}$) over the parameter symbol, for example, $\hat{\mu}$ indicates an estimate of μ, the population mean. Then, the fitted value for an observation y_{jk}, denoted $\hat{y}_{jk}$, consists of the systematic component of the model with all parameters replaced by their estimates. So, in the qualitative model represented by Equation 1.1, the fitted value for the kth plant with the jth feeding regime is

$$\hat{y}_{jk} = \hat{\mu} + \hat{\tau}_j \,, \tag{1.3}$$

which is an estimate of the population mean for plants with the jth feeding regime. For the quantitative model in Equation 1.2, the corresponding fitted value is

$$\hat{y}_{jk} = \hat{\alpha} + \hat{\beta} x_j \ . \tag{1.4}$$

For all linear models, parameters are estimated with the **principle of least squares**. This method finds the 'best-fit' model in the sense that it finds estimates for the parameters that minimize the sum, across all observations, of the squares of the differences between the observed data and fitted values. For example, for the qualitative model (Equation 1.1) we minimize

$$\sum_{j=1}^{3} \sum_{k=1}^{4} (y_{jk} - \hat{y}_{jk})^2 \ ,$$

where $\hat{y}_{jk}$ was defined in Equation 1.3. For the quantitative model (Equation 1.2), the quantity minimized has the same generic form, but now Equation 1.4 is used to define the fitted values. The symbol Σ is used to indicate the sum across the specified index (see Section 2.1 for more details). Note that these summations are over all combinations of the three factor levels ($j = 1, 2, 3$) and the four replications ($k = 1, 2, 3, 4$), and hence over the full set of 12 observations. Having found the best-fit model for our observed data, we can calculate fitted values based on the parameter estimates. We can then obtain estimates of the deviations, called **residuals**, from the discrepancy between the observed and fitted values, as

$$\hat{e}_{jk} = y_{jk} - \hat{y}_{jk} \ .$$

If the residuals are relatively small, then our model gives a good description of the data. These residuals can be examined to assess the validity of our model (to diagnose any lack of fit of the model to the data) and the assumptions made in fitting the model to the data (Chapters 4 and 12). One such assumption concerns an underlying probability distribution for the deviations (see Chapter 4), and the estimated variance of this distribution is used to calculate the parameter standard errors. This variance, often called the residual variance, provides a measure of uncertainty which can also be used in hypothesis testing and to form confidence intervals for predictions.

1.6 Summarizing the Importance of Model Terms

The main tool we use for the statistical analysis of any linear model, with either qualitative (factor) or quantitative (variate) explanatory variables, or both, is the **analysis of variance**, usually abbreviated as ANOVA. As the name suggests, the principle behind ANOVA is the separation and comparison of different sources of variation. In its simplest form, ANOVA quantifies variation in the response associated with the systematic component of the model (systematic variation) and compares it with the variation associated with the random component of the model (often called noise or background variation). Informally,

if the ratio of systematic variation to background variation is large then we can conclude that the proposed model accounts for much of the variation in the response, and that the explanatory variables provide a good explanation of the observed response. However, if the ratio of systematic variation to background variation is small, then it does not necessarily indicate that the response is not related to the explanatory variables – it may just be that the background variation is too large to clearly detect any relationship. We can use ANOVA to assess whether the variation associated with different levels, or groups of levels, of a qualitative explanatory variable (factor) is larger than the background variation, which would give evidence that the explanatory variable is associated with substantive changes in the response. Similarly, we can assess whether there is substantive variation in the response associated with some trend in a quantitative explanatory variable (variate). We can often also partition variation associated with different explanatory variables to assess their relative importance, and a well-designed experiment can make this easier. We use ANOVA to summarize model fitting in two related contexts.

We first consider the use of ANOVA in structured scientific studies where we include the experimental conditions as factors, and wish to relate variation in the response to variation in the conditions. For example, consider a traditional field trial to assess the yield response of a set of plant varieties to different levels of fertilizer application. Here, the experimental conditions are combinations of plant variety and fertilizer application, with both considered to be qualitative variables. In a basic analysis, we are interested in identifying whether differences between plant varieties or fertilizer application levels, or particular combinations of these factors, provide an explanation for the observed differences in yield response. Within this context we can then generalize this basic analysis in several different ways: to take account of the physical structure of the experimental units (e.g. to allow for the blocking of experimental units); to take account of any quantitative scale underlying the factor levels (e.g. the nitrogen content of the fertilizer applications); and, in a limited way, to account for other explanatory variables that may have been measured (e.g. perhaps soil pH varies across the field and affects yield). This is the traditional framework for ANOVA and most statistical packages have algorithms tailored to the analysis of data within this framework (e.g. the `ANOVA` command in GenStat, the `aov()` function in R and the `proc glm` procedure in SAS).

We then consider the use of ANOVA in scientific studies where the main aim is to model the response as a mathematical function of one or more quantitative explanatory variables. This context is usually called **regression modelling** or **regression analysis**, and we emphasize the particular case of linear regression, where only linear functions of one or more continuous explanatory variables are permitted. For example, suppose a forester wishes to build a model to predict timber volume from easily measured field variables such as tree diameter and height. In a basic analysis, having measured both the field variables and the actual timber volume for a number of trees, we are interested in determining which field variables (or combinations of field variables) explain the observed differences in timber volume. Again, within this context we can generalize the basic analysis to take account of any grouping of observations, such as tree variety or location. Within regression modelling, ANOVA is the main statistical tool used for assessment of the importance of different explanatory variables. Statistical software packages usually contain more general algorithms for regression analyses (e.g. `FIT` in GenStat, the `lm()` function in R and the `proc reg` procedure in SAS).

It should be clear that there is much overlap between these two contexts. For example, both the qualitative model of Example 1.1 and the quantitative model of Example 1.2 could be analysed by either type of algorithm. However, using different algorithms to analyse

the same data set can be confusing, because even when the methods are equivalent, the results may appear to differ if different conventions are used for their presentation. One of the main aims of this book is to explain the rationale behind these different conventions, and so to eliminate this confusion.

1.7 The Scope of This Book

We follow this chapter with a review chapter. Although we minimize the use of mathematical formulae, some are essential, and so we provide a review of mathematical notation in Chapter 2, along with the basic statistical concepts and methods used elsewhere in the book. Many readers will be familiar with these concepts and might treat this chapter as optional.

The early chapters of the book (Chapters 3 to 11) focus on the design of experiments and the analysis of data from designed experiments. In Chapter 3, we concentrate on the essential statistical principles of design: replication, randomization and blocking. We consider the structure of an experiment and describe some common designs. In Chapters 4 to 7, we consider analysis of simple designs. In Chapter 4, we consider in detail the analysis of data from the simplest design – the completely randomized design – to explain the concepts of ANOVA. We explain how the ANOVA table is formed, how it relates to a model for the data and how to interpret it. In Chapter 5, we explore the assumptions underlying the model and analysis and describe the diagnostic tools we can use to check them. We consider how these assumptions might be violated and the possible consequences, and ways to remedy these problems. In Chapter 6, we discuss transformations of the response variable as one remedy for failure to satisfy the model assumptions. In Chapter 7, we extend the analysis to the simplest design that includes blocking, the randomized complete block design, and introduce the concept of strata, or different structural levels, within a design and its analysis. In Chapters 8 to 11, we consider more advanced issues in the analysis of designed experiments. In Chapter 8, we consider how best to extract answers about our experimental hypotheses from our analysis. The advantages of factorial treatment structures, used to test the effects of several treatment factors simultaneously, will be explained. We describe the use of crossed and nested models for factorial structures, and how to make pairwise comparisons of treatments. In Chapter 9, we describe the analysis of some designs with somewhat more complex blocking structures, namely the Latin square, split-plot and balanced incomplete block designs. Then in Chapter 10, we consider how to calculate the replication required to obtain a specified precision for treatment comparisons in simple designs, and we introduce the concept of statistical power. We also discuss the case of equivalence testing, where the interest is in detecting equivalence rather than differences between treatments. Finally, Chapter 11 examines the issues that arise for non-orthogonal designs, where an unambiguous analysis can no longer be obtained.

In the later chapters of the book (Chapters 12 to 18) we turn our attention to regression modelling. In Chapter 12, after a brief general introduction, we concentrate first on simple linear regression, relating the response to a linear function of a single explanatory variate. The diagnostic tools introduced in Chapter 5 can be used for regression modelling, but additional diagnostic tools are available to check the validity of a regression analysis, and these are introduced in Chapter 13. In Chapters 14 and 15, we then extend regression models. In Chapter 14, we introduce multiple linear regression, extending the simple linear

regression model to include several explanatory variates and considering problems of collinearity and variable selection. In Chapter 15, we show how to investigate the best form of a regression model when observations arise from different groups, how to incorporate simple designs into regression models and discuss analysis of covariance. We then move beyond linear regression. In Chapter 16, we introduce linear mixed models for the analysis of unbalanced studies where structure is present. In Chapter 17, we first use functions of explanatory variables to model curved relationships with linear models and then give a brief introduction to non-linear models. This concept is extended in Chapter 18 to the case of the generalized linear model, which can be used to model responses with certain types of non-Normal errors. We introduce two special, but commonly used, cases – the logit model for Binomial (proportion) data, and the log-linear model for Poisson (count) data.

Finally, the concluding chapter (Chapter 19) provides an overview of the full process of design and statistical analysis by way of real examples.

Our website (www.stats4biol.info) provides an overview and basic introduction to three commonly used statistical packages: GenStat, R and SAS. All of the examples are analysed with each of these packages, together with answers to selected exercises. Our personal preference is for the GenStat statistical software, because of its excellent implementation of algorithms for the analysis of designed experiments, and the provision of menus to make analyses easily accessible to all. The R package provides functions for all the standard analysis approaches introduced in this book, and has the benefits and drawbacks associated with being free, open-source software. We include SAS because of its wide user base and general availability. Most results presented in the book can be obtained with any of these packages; we comment where results may differ between packages and output has been obtained from a specific package.