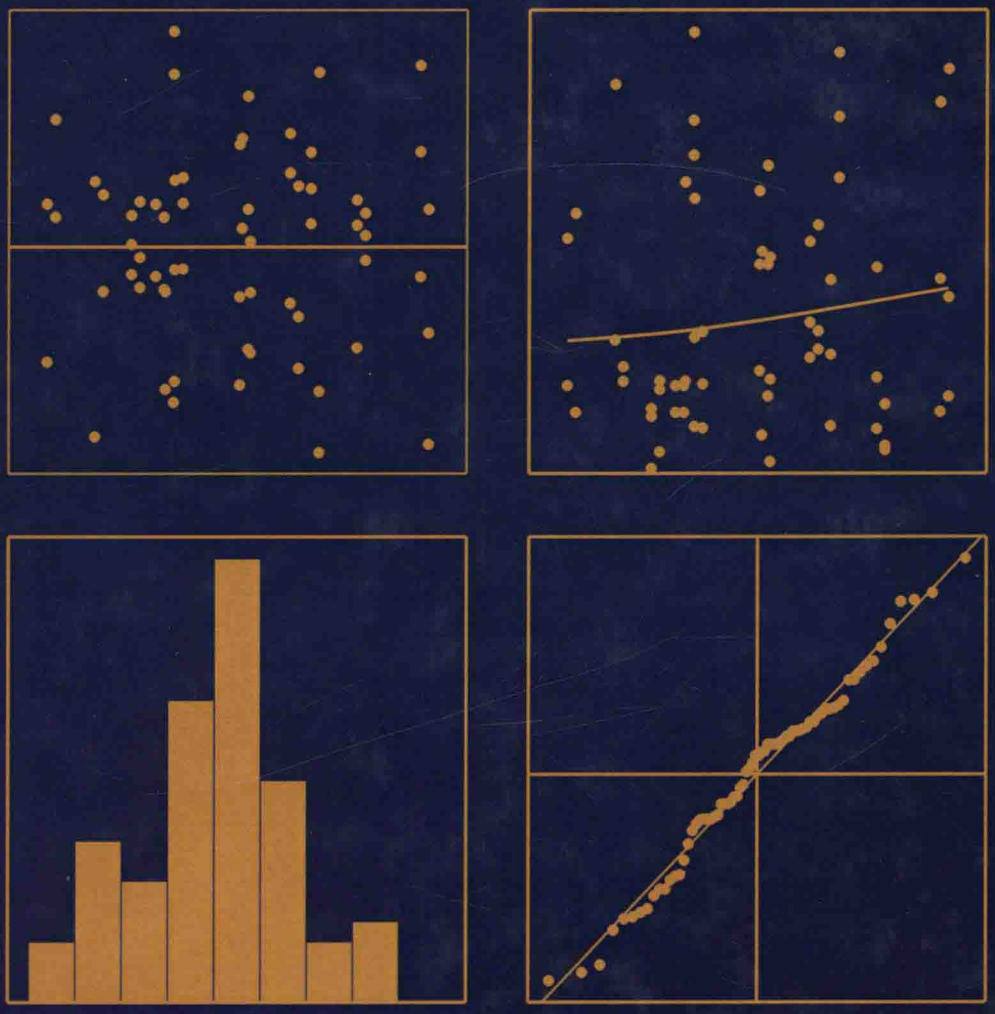
STATISTICAL METHODS IN BIOLOGY

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

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

Printed on acid-free paper
Version Date: 20140703

International Standard Book Number-13: 978-1-4398-0878-8 (Hardback)

Library of Congress Cataloging-in-Publication Data

Welham, S. J. (Suzanne Jane), author.
Statistical methods in biology : design and analysis of experiments and regression / S.J. Welham , S.A. Gezan, S.J. Clark, A. Mead.
pages cm
Includes bibliographical references and index.
ISBN 978-1-4398-0878-8 (hardback : acid-free paper) 1. Biometry. 2. Regression analysis. 3. Experimental design. I. Title.

QH323.5.W45 2014
570.1'5195--dc23 2014016839

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site at
http://www.crcpress.com

STATISTICAL METHODS IN BIOLOGY

Design and Analysis of Experiments and Regression

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

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.

Contents