Experimental Design and Data Analysis for Biologists

Applying statistical concepts to biological scenarios, this established textbook continues to be the go-to tool for advanced undergraduates and postgraduates studying biostatistics or experimental design in biology-related areas. Chapters cover linear models, common regression and ANOVA methods, mixed effects models, model selection, and multivariate methods used by biologists, requiring only introductory statistics and basic mathematics. Demystifying statistical concepts with clear, jargon-free explanations, this new edition takes a holistic approach to help students understand the relationship between statistics and experimental design. Each chapter contains further-reading recommendations and worked examples from today's biological literature. All examples reflect modern settings, methodology, and equipment, representing a wide range of biological research areas. These are supported by hands-on online resources including real-world datasets, full *R* code to help repeat analyses for all worked examples, and additional review questions and exercises for each chapter.

Gerry Quinn is an honorary professor in the School of Life and Environmental Sciences at Deakin University, having served as Chair in Marine Biology and Head of Warrnambool Campus during his academic career. He has extensive experience in teaching biostatistics at Deakin University and the University of Gothenburg.

Michael J. Keough is an ecologist, environmental scientist, and honorary professor in the School of Biosciences at University of Melbourne. He has taught classes in ecology, experimental design, and environmental science for many years at the University of Melbourne and in the United States.

"The new edition of this 'go-to' text, has been revised to take a more holistic, informative, and up-to-date approach to the subject. This remarkably accessible, practical and – at times – entertaining guide to how we can best translate our questions and ideas into informative experiments and analyses is highly recommended for everyone wanting to investigate, visualise, and analyse biological phenomena."

Professor Jonathon Havenhand, *University of Gothenburg*

"The new book is an excellent resource for researchers, analysts and teachers. The text clearly outlines the important concepts of the tests and models. The examples are all based on published data which makes it easy to source the full manuscript, datasets and replicate the analyses, which is incredibly important and assists with interpretation."

Dr Victoria Goodall, *VLG Statistical Services*

"At last, a book for undergraduates and graduates that distills the complexities of biological data analysis into an easy-to-understand generalized linear model approach – with examples in R! The authors challenge the reader to think critically about data by providing important details on design, summary statistics, power analysis/effect size, model fit, and data visualization. This book will be used in my classes for years to come."

Professor Greg Moyer, *Mansfield University*

"I have been using *Experimental Design and Data Analysis* for teaching and research for 20 years and have been hoping for a second edition for 10. It was worth the wait. The new edition shares many attributes with the original. It is quite easy to read, the examples are varied and interesting, there are ample and revealing box examples and there remains an attitude of Statistics as a tool that will be used by many. There are a number of changes that reflect these attributes – two key ones are the emphasis on Generalized Models as a framework for a broad array of approaches and the conversion of examples to R, with code and additional material available online."

Professor Peter Raimondi, *University of California Santa Cruz*

"I have taught in and coordinated a third-year design/statistics paper for zoology and ecology students for 10 years – an enjoyable and sometimes challenging task. The first edition of Quinn and Keough has been immensely helpful for me in teaching this course. The second edition has been updated and expanded considerably. I'm sure it will continue to inspire my future teaching."

Professor Christoph Matthaei, *University of Otago*

"I was excited to see Quinn and Keough have updated their classic guide to experimental design and data analysis. I read the earlier edition of this book as a graduate student, and the advice it provides on experimental design is the foundation of my own studies, as well as my approach to training graduate students. . . This book is foundational reading for aspiring scientists. Not only does it teach you how to analyse your data, it also provides invaluable advice on how to communicate analyses and write up scientific studies. The book's advice will help give early career scientists the confidence they need to write-up and publish their first studies."

Professor Chris Brown, *University of Tasmania*

Experimental Design and Data Analysis for Biologists

Second Edition

Gerry P. Quinn
Deakin University

Michael J. Keough
University of Melbourne

CAMBRIDGE UNIVERSITY PRESS

Shaftesbury Road, Cambridge CB2 8EA, United Kingdom

One Liberty Plaza, 20th Floor, New York, NY 10006, USA

477 Williamstown Road, Port Melbourne, VIC 3207, Australia

314–321, 3rd Floor, Plot 3, Splendor Forum, Jasola District Centre, New Delhi – 110025, India

103 Penang Road, #05-06/07, Visioncrest Commercial, Singapore 238467

Cambridge University Press is part of Cambridge University Press & Assessment, a department of the University of Cambridge.

We share the University's mission to contribute to society through the pursuit of education, learning and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/highereducation/ISBN/9781107036710

DOI: 10.1017/9781139568173

First published 2002
Second edition 2024

Printed in the United Kingdom by TJ Books Limited, Padstow, Cornwall

A catalogue record for this publication is available from the British Library.

A Cataloging-in-Publication data record for this book is available from the Library of Congress

ISBN 978-1-107-03671-0 Hardback
ISBN 978-1-107-68767-7 Paperback

Additional resources for this publication at www.cambridge.org/quinn-keough2.

Contents

Preface *page* xiii
List of Acronyms xviii

1 Introduction 1
1.1 Almost Every Biological Theory or Hypothesis Is a Model of How Nature Works 1
1.2 We Use Data to Separate Wrong Models from (Possibly) Correct Ones (or Bad Models from Less Bad Ones) 2
1.2.1 What Kind of Data Do We Need? 2
1.2.2 The Signal and the Noise 2
1.2.3 Random and Representative 2
1.2.4 How Do We Decide If a Model Fits Well? What's the "Best" Model? 3
1.2.5 What Do I Do Next? 3
1.3 There Are Many Ways to Reach Wrong Conclusions! 3
1.4 There Are Also Many Ways to be Right 4
1.5 Our Philosophy in This Book 4
1.5.1 Think Clearly 4
1.5.2 Think in Advance 4
1.5.3 Think before You Analyze 5
1.6 How This Book Is Structured 5
1.7 A Bit of Housekeeping 6

2 Things to Know before Proceeding 8
2.1 Samples, Populations, and Statistical Inference 8
2.2 Probability 9
2.3 Probability Distributions 9
2.3.1 Distributions for Variables 10
2.3.2 Distributions for Statistics 11
2.4 Frequentist ("Classical") Estimation 11
2.4.1 Methods for Estimation 12
2.4.2 Simple Parameters and Statistics 13
2.4.3 Sampling Distribution of the Mean 15
2.4.4 Standard Error of the Sample Mean 15
2.4.5 Confidence Intervals for Population Mean 16
2.4.6 Standard Errors and Confidence Intervals for Other Statistics 16
2.4.7 Resampling Methods for Frequentist Estimation 17
2.5 Hypothesis Testing 18
2.5.1 Frequentist Statistical Hypothesis Testing 18
2.5.2 Decision Errors 20
2.5.3 One- and Two-Tailed Tests 21
2.5.4 Multiple Hypothesis Testing 21

2.5.5 Testing Hypotheses about Means and Variances for Two Populations 23
2.5.6 Parametric Tests and Their Assumptions 23
2.6 Comments on Frequentist Inference 24
2.7 Bayesian Statistical Inference 25
2.7.1 Prior Knowledge and Probability 27
2.7.2 Likelihood Function 27
2.7.3 Posterior Probability 27
2.7.4 Model Comparison and Bayes Factors 28
2.7.5 Final Comments 28

3 Sampling and Experimental Design 30
3.1 Sampling Design 30
3.1.1 Probability Sampling 30
3.1.2 Sample Size for Random Sampling 32
3.1.3 Nonprobability Sampling 33
3.2 Experimental Design 33
3.2.1 Replication of Experimental Units 34
3.2.2 Controls 36
3.2.3 Randomization 36
3.2.4 Independence 38
3.2.5 Reducing Unexplained Variance 38
3.2.6 Limitations of Manipulative Experiments 38
3.3 Sample Size for Detecting Differences: Power Analysis 39
*3.3.1 Using Power to Plan Experiments (*a priori *Power Analysis)* 39
3.3.2 Post Hoc *Power Calculation* 42
3.3.3 Effect Size 42
3.3.4 Using Power Analyses 43
3.4 Key Points 44

4 Introduction to Linear Models 45
4.1 What Is a Linear Model? 45
4.2 Components of Linear Models 46
4.2.1 Types of Response Variables 46
4.2.2 Types of Predictor Variables 48
4.3 Assembling our Linear Model 48
4.4 Estimation for Linear Models 49
4.4.1 Ordinary Least Squares 51
4.4.2 Maximum Likelihood 53
4.4.3 Robust Estimation Methods for Linear Models 54
4.5 How Well Does a Model Fit? 54
4.5.1 OLS Measures of Fit: ANOVA 54
4.5.2 ML Measures of Fit: Log-Likelihood and Deviance 55
4.5.3 Information Criteria 55
4.6 Assumptions for Linear Model Inference 55
4.6.1 Assumptions for OLS 55
4.6.2 Assumptions for ML 57
4.6.3 Model Diagnostics 57
4.7 Types of Linear Models 57
4.7.1 General Linear Models 57
4.7.2 Generalized Linear Models 59

4.7.3 Linear Mixed Models (General and Generalized) 60
4.8 Key Points 61

5 Exploratory Data Analysis 62
5.1 Basic Graphical Tools 62
5.1.1 Some Common Basic Graphs 62
5.1.2 Smoothing 66
5.1.3 Residual Plots 68
5.2 Outliers 69
5.3 Am I Fitting the Right Model? 70
5.3.1 The Underlying Probability Distribution 70
5.3.2 Homogeneity of Variances 70
5.3.3 Is My Linear Model "Linear"? 71
5.4 Is It Normal to Transform Data? 71
5.4.1 Transformations and Distributional Assumptions 72
5.4.2 Transformations and Linearity 72
5.4.3 Transformations and Additivity 73
5.4.4 Do We Really Need a Data Transformation? 73
5.5 Standardizations 73
5.6 Missing Data 73
5.6.1 Missing Data Mechanisms 74
5.6.2 Detecting Missing Data 74
5.6.3 Methods for Missing Data 74
5.7 Key Points 75

6 Simple Linear Models with One Predictor 76
6.1 Linear Model for a Single Continuous Predictor: Linear Regression 76
6.1.1 Linear Model for a Continuous Predictor (Linear Regression Model) 76
6.1.2 Model Parameters 81
6.1.3 Inference for Parameters 83
6.1.4 Inference for Predicted Values 83
6.1.5 Model Comparison and the Analysis of Variance 83
6.1.6 Regression Through the Origin 85
6.1.7 Regression with X Random 85
6.2 Linear Model for a Single Categorical Predictor (Factor) 88
6.2.1 Experimental vs. Observational Studies 88
6.2.2 Linear Model for a Categorical Predictor 88
6.2.3 Model Parameters 90
6.2.4 Inference for Parameters 93
6.2.5 Model Comparison and the Analysis of Variance 95
6.2.6 Unequal Sample Sizes (Unbalanced Designs) 96
6.2.7 Specific Comparisons of Group Means 97
6.3 Predictor Effects 101
6.3.1 Continuous Predictor (Regression) Models 101
6.3.2 Categorical Predictor Models 101
6.4 Assumptions 103
6.4.1 Zero Conditional Mean 103
6.4.2 Independence 103
6.4.3 Variance Homogeneity 103
6.4.4 Normality 103

6.5 Model Diagnostics 103
6.5.1 Residuals 103
6.5.2 Leverage 104
6.5.3 Influence Measures 105
6.5.4 Diagnostic Plots 105
6.5.5 Transformations 107
6.6 Robust Linear Models 108
6.6.1 Rank-Based ("Nonparametric") Methods 108
6.6.2 Generalized (Weighted) Least Squares 109
6.6.3 Other Robust Methods 109
6.6.4 Resampling and Permutation Methods 110
6.7 Power of Single-Predictor Linear Models 110
6.7.1 Regression Models 111
6.7.2 Categorical Predictor Models 111
6.8 Key Points 112

7 Linear Models for Crossed (Factorial) Designs 115
7.1 Two-Factor Fully Crossed (Factorial) Designs 115
7.1.1 Completely Randomized (Experimental) Designs 115
7.1.2 Observational (Nonexperimental) Designs 115
7.1.3 Designs That Combine Completely Randomized Factors with Nonrandomized (Observational) Factors 118
7.1.4 The Factorial Linear Effects Model 118
7.1.5 Model Parameters 121
7.1.6 Inference for Parameters 121
7.1.7 Model Comparison and Analysis of Variance 123
7.1.8 More on Main Effects and Interactions 126
7.1.9 Interactions and Transformations 127
7.1.10 Specific Comparisons of Marginal Means 127
7.1.11 Interpreting Interactions 128
7.1.12 Predictor Effects 128
7.1.13 Assumptions 129
7.1.14 Robust Factorial ANOVAs 129
7.2 Complex Factorial Designs 129
7.2.1 Missing Cells 130
7.2.2 Fractional Factorial Designs 134
7.3 Power and Sample Size in Factorial Designs 135
7.4 Key Points 136

8 Multiple Regression Models 137
8.1 Linear Model for Multiple Continuous Predictors: Multiple Regression 137
8.1.1 The Multiple Linear Regression Model 137
8.1.2 Model Parameters 139
8.1.3 Inference for Parameters 144
8.1.4 Model Comparison and Analysis of Variance 145
8.1.5 Assumptions of Multiple Linear Regression Models 147
8.1.6 Model Diagnostics 147
8.1.7 Diagnostic Graphics 148
8.1.8 Transformations 148
8.1.9 Collinearity 148

8.1.10 Interactions in Multiple Regression 151
8.1.11 Regression Models with Polynomial Terms 154
8.1.12 Other Issues in Multiple Linear Regression 156
8.1.13 Categorical Predictors in Multiple Regression Models 158
8.2 Analysis of Covariance 158
8.2.1 Linear Models for Simple Analyses of Covariance 160
8.2.2 Model Comparison and the Analysis of (Co)variance 165
8.2.3 Assumptions of ANCOVA Models 166
8.2.4 Homogeneous Within-Group Regression Slopes 167
8.2.5 Robust ANCOVA 170
8.2.6 Unequal Sample Sizes (Unbalanced Designs) 170
8.2.7 Specific Comparisons of Adjusted Means 170
8.2.8 Factorial Designs 170
8.2.9 Designs with Two or More Covariates 172
8.3 Key Points 173

9 Predictor Importance and Model Selection in Multiple Regression Models 174
9.1 Relative Predictor Importance 174
9.1.1 Single Model Methods 174
9.1.2 Multiple Model Methods 181
9.1.3 Recommendations of Relative Importance 182
9.2 Model Selection 182
9.2.1 Model Selection Criteria 183
9.2.2 Traditional Stepwise Selection 184
9.2.3 All Subsets and Information Criteria 184
9.2.4 Model Averaging 185
9.2.5 Model Validation 186
9.3 Regression Trees 187
9.3.1 Standard Regression Trees 187
9.3.2 Bagging and Boosted Regression Trees 189
9.4 Key Points 193

10 Random Factors in Factorial and Nested Designs 194
10.1 Fixed vs. Random Effects and Mixed Models 194
10.1.1 Designs Applicable to Mixed Models 194
10.2 Fitting Linear Models with Fixed and Random Factors 199
10.2.1 Traditional OLS "ANOVA" Models Approach 200
10.2.2 Linear Mixed Effect (or Multilevel) Models Approach 201
10.2.3 Modeling Strategies 204
10.3 Simple Random Factor Designs 205
10.3.1 Traditional OLS Approach 206
10.3.2 Linear Mixed Effects (Multilevel) Models 207
10.4 Multilevel Regressions 207
10.5 Nested (Hierarchical) Designs 209
10.5.1 Two-Level Nested Designs 210
10.5.2 More Complex Nested Designs 214
10.5.3 Sample Size and Nested Designs 215
10.6 Crossed (Factorial) Mixed Designs 215
10.6.1 Types of Factorial Mixed Designs 216
10.6.2 Analysis of Crossed Designs with One Fixed and One Random Factor 217

10.6.3 Crossed Designs with Two or More Fixed Factors and One Random Factor 223
10.6.4 Design and Analysis Issues with Crossed Mixed Designs and Their Models 223
10.7 Key Points 227

11 Split-Plot (Split-Unit) Designs: Partly Nested Models 228
11.1 Simple Split-Plot Designs 228
11.1.1 Analysis for Simple Split-Plot Designs 230
11.1.2 Assumptions 235
11.1.3 Unbalanced Split-Plot Designs 236
11.1.4 Model Building 236
11.2 More Complex Designs 236
11.2.1 Additional Between-Plots Factors 236
11.2.2 Additional Within-Plots Factors 239
11.2.3 Including Continuous Covariates 239
11.3 Key Points 241

12 Repeated Measures Designs 242
12.1 Simple Repeated Measures Designs 243
12.1.1 Analysis of Simple Repeated Measures Designs 243
12.1.2 Assumptions for Simple Repeated Measures Models 247
12.1.3 Missing Observations 249
12.2 More Complex Repeated Measures Designs 249
12.2.1 One Between-Subjects Factor 249
12.2.2 Two or More Between-Subjects Factors 250
12.2.3 Two or More Within-Subjects Factors 256
12.2.4 Model Selection in Complex Repeated Measures Designs 256
12.3 Key Points 256

13 Generalized Linear Models for Categorical Responses 257
13.1 Logistic Regression 258
13.1.1 Binary Response with a Single Continuous Predictor: Simple Logistic Regression 258
13.1.2 Categorical Predictors in GLMs 258
13.1.3 Binary Response with Multiple Predictors: Multiple Logistic Regression 260
13.1.4 Nominal and Ordinal Multinomial Response Variables 260
13.1.5 Proportion Response 260
13.2 Count Responses: Poisson Regression 260
13.3 Goodness-of-Fit for GLMs 263
13.4 Inference for Parameters in GLMs 265
13.5 Assumptions and Diagnostics for Binomial and Poisson GLMs 265
13.6 Overdispersed Data 266
13.6.1 Identifying Overdispersion 267
13.6.2 Correcting for Overdispersion 267
13.6.3 Too Many Zeroes (Zero-Inflated Data) 268
13.6.4 Zero-Truncated Data 268
13.6.5 Binomial Overdispersion 268
13.7 Contingency Tables 269
13.7.1 Two-Way Tables 269
13.7.2 Three-Way and Higher Tables 274
13.7.3 More Complex Tables 279
13.7.4 Loglinear vs. Logistic Models for Tables 279

13.8 Generalized Linear Mixed Models 279
13.9 Generalized Additive Models 280
13.10 Key Points 283

14 Introduction to Multivariate Analyses 285
14.1 Distributions and Associations 286
14.2 Linear Combinations, Eigenvectors, and Eigenvalues 287
14.2.1 Linear Combinations of Variables 287
14.2.2 Eigenvalues 287
14.2.3 Eigenvectors 287
14.2.4 Derivation of Components 287
14.3 Multivariate Distance and Dissimilarity Measures 291
14.3.1 Dissimilarity Indices for Continuous and Count Variables 292
14.3.2 Dissimilarity Indices for Dichotomous (Binary) Variables 292
14.3.3 General Dissimilarity Indices for Mixed Variables 294
14.3.4 Choosing Dissimilarity Indices 294
14.4 Data Transformation and Standardization 295
14.5 Standardization, Association, and Dissimilarity 296
14.6 Screening Multivariate Datasets 296
14.7 Introduction to Multivariate Analyses 297
14.8 Key Points 298

15 Multivariate Analyses Based on Eigenanalyses 299
15.1 Principal Components Analysis 299
15.1.1 Deriving Components 299
15.1.2 Interpreting the Components 300
15.1.3 How Many Components to Retain? 305
15.1.4 Which Association Matrix to Use? 305
15.1.5 Simplifying Component Structure 306
15.1.6 PCA Assumptions and "Fit" 306
15.1.7 Ordination and Biplots for PCA 307
15.1.8 Principal Components Regression 308
15.1.9 Factor Analysis 308
15.2 Correspondence Analysis 309
15.2.1 Deriving the Axes 309
15.2.2 Ordination and Biplots for CA 311
15.2.3 Reciprocal Averaging 311
15.3 Use of PCA and CA with Ecological Abundance (Count) Data 311
15.4 Constrained (Canonical) Multivariate Analysis 312
15.4.1 Redundancy Analysis 312
15.4.2 Canonical Correspondence Analysis 315
15.5 Linear Discriminant Function Analysis 316
15.5.1 Deriving Discriminant Functions 316
15.5.2 Classification and Prediction 319
15.5.3 Assumptions of Discriminant Function Analysis 320
15.5.4 Multivariate Analysis of Variance 320
15.6 Key Points 320

16 Multivariate Analyses Based on (Dis)similarities or Distances 322
16.1 Multidimensional Scaling or Ordination 322
16.1.1 Metric (Classical) Scaling: Principal Coordinates Analysis 323
16.1.2 Nonmetric (Enhanced) Multidimensional Scaling 323
16.2 Cluster Analysis 329
16.2.1 Agglomerative Hierarchical Clustering 329
16.2.2 Divisive Hierarchical Clustering 330
16.2.3 Nonhierarchical Clustering 330
16.3 Analyses Based on Dissimilarities 331
16.3.1 Contributions of Original Variables to Ordination 331
16.3.2 Relating Dissimilarities to Other Variables 331
16.3.3 Multivariate Linear Models 334
16.4 Key Points 335

17 Telling Stories with Data 337
17.1 Research Doesn't Exist Until You Tell Someone 337
17.1.1 Telling Better Stories: The Importance of Narrative 337
17.2 Summarizing Data Analyses 338
17.2.1 Linear Models 338
17.2.2 Other Analyses 341
17.3 Visualizing Data 341
17.3.1 Just Show Us the Numbers! 343
17.3.2 Tables 343
17.4 Graphical Summaries of the Data 343
17.4.1 Some Basic Principles for Visualizing Data 344
17.4.2 An Appropriate Visual Display 350
17.5 Error Bars: Visualizing Variation and Precision 359
17.5.1 Possible Solutions 360
17.6 Horses for Courses: What You Present Depends on Who's Listening 360
17.6.1 Know Your Audience as Well as Possible 363
17.7 Software and Other Sources 364
17.8 Key Points 364

Glossary 365
References 367
Index 383

Color plates can be found between pages 364 and 365.

Preface

Statistical analysis is at the core of most modern biological research, and many biological hypotheses, even deceptively simple ones, can require complex statistical models.

The landscape can be daunting, with a forest of acronyms, a bewildering array of terms – linear models, generalized linear models (GLMs), generalized linear mixed models (GLMMs), covariates, randomized blocks, mixed models, multilevel analysis, multivariate analysis, and so on, and apparently different statistical tribes. Adding to this complexity is a statistical package (*R*) that has become the standard but often offers several options for the same task and challenging online help. How can a biological researcher, particularly a beginner, make sense of this landscape?

Much of this complexity is unnecessary or illusory. Statistical analyses with different names are sometimes synonyms rather than new techniques to be learned. More importantly, many different analyses are better viewed as linear models built using a common framework rather than distinct tools. Understanding the framework prepares you to deal with a range of unfamiliar situations you may encounter.

As biologists, we try to explain natural phenomena as best we can, given our current knowledge. A good explanation has survived challenges from alternative explanations. We think of these explanations as models – simplified descriptions of nature – and we assess their adequacy by comparing them to data. Ideally, we challenge a particular explanation by finding or creating a novel situation in which our focal model will produce a pattern in the data that differs from the pattern produced by the alternatives.

We use statistical analysis to assess the fit of the data to a particular model. We look for a signal from that model against background noise in the data and estimate that signal's strength. The confrontation between models and data requires us to translate a broad, even qualitative model that is our biological explanation (or hypothesis) into a more precise statistical model. The specific way we compare the model to data depends on our statistical approach, but ultimately, we decide based on that comparison. The decisions may be formal, as in hypothesis tests, or informal but no less important – what do we do next in our research, or what do we advise end-users of the knowledge to do?

We can be led astray in many ways during this process:

- The sampling and experimental design doesn't match the biological question.
- The statistical model doesn't match the biological one.
- The data don't have properties assumed by the statistical model.
- The data aren't sufficient to distinguish signal from noise, between competing models or to allow you to estimate biological effects with confidence.
- You might have unrepresentative data because of poor design or bad luck.

These problems can cause you to expend lots of energy or resources, only to get an unclear or wrong answer to the original biological question. We aim to reduce the likelihood of this happening, and we use three simple principles:

- Think clearly about the biological problem, the different models or explanations in play, and what kind of data we need to distinguish these models unambiguously.
- Think in advance about the statistical model corresponding to each biological model and how the match between model and data will be assessed. Decide how much data you need to make confident decisions.
- Think before you analyze. Make sure that the data you've collected are consistent with assumptions of your statistical model(s); if not, transform the data or respecify the model.

These principles require most of the hard work before a single statistical analysis is attempted. They also require us to be our own harshest critics, to make sure that we have challenged each biological model severely.

We apply this approach to a wide range of statistical models that biologists use, from the simplest to some that are very complex. Most of these models are variants of GLMs. We start with simple models and gradually make

them more complex. We show how different "methods" such as analysis of variance (ANOVA), multiple regression, logistic regression, etc. are closely related.

We need to:

- know the pitfalls and assumptions of particular statistical models;
- be able to identify the model appropriate for the sampling or experimental design and the data we plan to collect;
- be able to interpret the output of analyses using these models; and
- be able to design experiments and sampling programs optimally – that is, with the best possible use of our limited time and resources.

In This Book

Our approach encourages readers to understand the models underlying the most common designs in biology. We assume that our readers have completed some basic training in experimental design and data analysis, and we begin with reminders of some essential basic concepts. We then build the linear models common to so many biological situations. We begin with the structure of a linear model, the variables used in these models, and how we fit models to data with an emphasis on ANOVA. Because model-fitting is crucial, we outline the exploratory methods used to ensure an appropriate model is being used. The next few chapters provide detailed accounts of increasingly complex models. We start with simple models with a single predictor and a single response variable, and outline models for when these variables are continuous or categorical (ANOVA/regression/logistic regression/loglinear models). From there, we consider models with a continuous response variable and multiple predictors, starting with predictors that are only continuous or categorical. These chapters cover familiar approaches of factorial ANOVA and multiple regression and can be expanded to mixtures of continuous and categorical predictors. We then introduce models with fixed and random effects – linear mixed models – and models for correlated data, including nesting, multilevel modeling, and repeated measures. We extend these models to situations with categorical responses in generalized linear and mixed models. Our emphasis is on learning to build and fit linear models for particular situations rather than applying cookbook techniques.

After considering situations with single response variables, we briefly describe techniques, such as principal components, discriminant function analysis, and multidimensional scaling, used with multiple response variables. We also show how linear models can be applied to multivariate data.

Our emphasis for most of the book is on thinking clearly about collecting, analyzing, and interpreting data. We conclude with some thoughts about communicating this process clearly to our end-users, particularly those unfamiliar with the statistical approaches.

Each chapter includes a further reading section, where we list books or book chapters that provide background to or expand on particular statistical topics. These lists are clearly not exhaustive and simply represent books/authors that we, and more importantly our students, have found helpful. Other relevant books and papers from the primary statistical and biological literature are cited in all chapters.

Learning by Example

One of our strongest beliefs is that most biologists understand statistical principles much better when we see how they are applied to situations in our own discipline. Examples let us link statistical models and formal statistical terms (blocks, plots, etc.) or papers written in other fields and the biological situations we are dealing with. For example, how is our analysis and interpretation of an experiment repeated several times helped by reading literature about blocks of agricultural land? How does literature developed for psychological research help us measure changes in plants' physiological responses?

Throughout this book, we illustrate the statistical techniques with examples from the current biological literature. We describe how to analyze the data. We focus on specifying the appropriate model, assessing assumptions, and interpreting the statistical output. These examples appear as boxes throughout each chapter. We use examples where the raw data are available, mainly through public online repositories (e.g. individual journals, datadryad.org) but sometimes on our website, courtesy of the study's authors. Although we focus on data analysis rather than software, we have implemented all analyses in *R*, with the code available online.

The other value of published examples is that we can see how particular analyses can be described and reported. It is easy to allow the biology to be submerged beneath a mass of statistical output when fitting complex statistical models. We hope that the examples and our thoughts on this subject in the final chapter will help prevent this from happening.

This Book Is a Bridge

We assume that readers have some introductory training, but their research questions will usually require statistical models far beyond that introduction. Even a simple decision such as using each experimental animal more than once triggers an additional complexity in the analysis. Even if we use simple designs, we will be reading papers with more complex analyses. We have tried to cover the most common biological designs we have found students and colleagues using, and to provide readers with the tools to tackle more complex or unusual questions.

Biological data are often messy, and some readers will find that their research questions require more complex models than we describe here. The primary statistical literature provides ways of dealing with messy data or solutions to complex problems. We try to point the way to critical pieces of that statistical literature, providing the reader with the essential tools to deal with that literature or get help. The help might come from formally trained statisticians or more knowledgeable peers, but we need to speak a common language to describe the question and interpret the answer. If help comes from outside our specific discipline, we need to be aware of biological considerations that may cause statistical problems. We can't expect a statistician to know the biological idiosyncrasies of our particular study, but we may get misleading or incorrect advice if they lack that information. This book provides a bridge to these situations.

We hope this book will be used in two ways, as for the first edition. In formal classes, it is the base for a graduate, or perhaps advanced undergraduate subject. It should be preceded by an introductory-level class. There is too much material for a single-semester class, and instructors can choose subsets of material that are right for their students. This has been our approach in teaching. The book also functions as a reference source for individual researchers who need to deal with actual data collection. Researchers can use it for self-directed learning and get practical suggestions for dealing with data. They also have a launching point to delve more deeply into the relevant literature or to start conversations about more complex analyses.

Always remember that for biologists, *statistics is a tool* we use to illuminate and clarify biological problems. We aim to use these tools efficiently without losing sight of the biology that is the motivation for most of us entering this field.

What's Different This Time Around?

In many ways, our first edition was reactive. We were influenced by the experimental and sampling designs we encountered among colleagues and research students. We focused on the common, named analyses – regressions, ANOVA, analyses of covariance, etc. We did focus on mixed models, particularly those involving nesting, because designs for which these statistical models are appropriate were disproportionately common. While acknowledging a range of model-fitting approaches, we used ordinary least squares (OLS) approaches to illustrate the major analyses.

We also presented the hypothesis-testing approaches as the main criterion for deciding whether a hypothesis was supported or not, though we described others. We used this approach because it was (and remains) common and provides an easy link to planning studies with sufficient replication (through power analysis). Our view was that there are several approaches, and all have strengths and shortcomings. Careful statistical and philosophical issues underpin the different approaches, and we presented readers with an introduction to this literature. Our view was and continues to be that we should know these deeper issues. Whether or not we use *P*-values, confidence intervals, likelihoods, etc., we understand precisely what each method does and does not do. We have tried to encompass this diversity of views, using the signal/noise concept – confidence intervals, *P*-values, etc. are tools to avoid getting fooled by noise (Gelman & Loken 2014).

This Edition Differs from Its Predecessor in Several Important Ways

A holistic approach to linear models vs. a box of named tools. In recent years, we've lost count of how often it's been recommended (by advisors, reviewers, etc.) that a researcher use linear mixed models as an alternative to, for example, the planned partly nested ANOVA. After our initial confusion, given that a partly nested ANOVA is almost always a linear mixed model, we realized that three issues were conflated. First, just because two analyses have different names, they may not actually be different (see also partly nested vs. repeated measures designs in Chapters 11 and 12). Second, the linear mixed models were often GLMs, fitted using maximum likelihood (ML) estimation and emphasizing parameter estimation and model comparison. Third, sometimes, "new" approaches did not represent dramatic changes but could be something as simple as changing one predictor variable from categorical to continuous. We moved away from just describing the different named methods. Instead, the analyses done by most biologists share a common core of fitting a statistical model to data. It is usually a linear model. Biologists measure biological responses that may be discrete or continuous, and those

variables may be distributed in many ways. Generally, they follow a distribution from the exponential family (Chapter 4), so we can fit a GLM. The form of the GLM depends on the distribution followed by the response variable and the nature of the predictor variable(s).

Suppose the response variable follows a normal distribution. Here, we can use least squares methods to fit the model, and it's a general linear model – but that's just a special case of a GLM. We can classify our predictors by whether they, too, are categorical or continuous and whether they are fixed or random (we'll explain this later!). If some of our predictors are fixed, and some are random, we magically have a GLMM. We focused the core of this book (Chapters 4–13) around building GLMs, using our taxonomy of predictors. We start by building simple models with single predictors, extending to more complex models with more complex relationships among the predictors. As we do this, we link the model to the "named" approaches you may find in papers and elsewhere. We've also pointed out how different named methods are related, and sometimes where different approaches have been used, despite the same underlying statistical model. To use a DIY metaphor, it's the change from a shed full of specific mains-powered tools to a core of standard batteries and chargers, matched with task-specific "skins" with the same shared power source.

We have brought the **estimation of effects** very much to the front. They were implicit through much of the first edition and explicit in parts (e.g. power analysis). In this edition, they are front and center – we need tools to identify signals from noise, but the nature of the signal (its strength and form) separates the trivial from the meaningful. We need to know we're not fooled by noise to start with, but we should be telling our audiences what the signal is like. We also need to acknowledge that detecting a signal may tell us little or nothing about its strength.

P-values remain one way of **identifying signals**, but not the only one, and we have tried to illustrate parallel approaches. Published biological papers often show strong effects (which may be why they were published!), and the different approaches often lead us to the same conclusion. While we have views on preferred approaches, it's possible to write a clear justification for several approaches. We encourage you to read thoroughly and critically and decide on your decision-making approach.

We finished the first edition with some thoughts about improving how we talk about our analyses because we felt this topic was underappreciated. We've expanded this section and tried to move it beyond our preliminary thoughts. We're delighted to see how science communication has increased in emphasis, bringing with it a broader diversity of target audiences. There are now many good books on the topic, and many students take a class as part of their training. There is still scope for improvement when reporting analyses, however.

Some Topics Have Been Reduced

We did try to provide an outline of the philosophical underpinnings of biological inference. Since then, there has been lots of published material, renewed attention to poor practices, debates about hypothesis testing, campaigns against using "statistical significance," and so on. Much of this discussion within the biological literature is a little simplistic, and it's the professional domain of others. Rather than presenting our summary, we've reduced this component and encourage you to read clearer accounts of the issues. Statisticians and philosophers write these accounts, but they are intended for practitioners.

We have assumed that you'll already have a basic statistical understanding in venturing into this book. That was the case in the first edition, but despite that, we embarked on at least two chapters of revisionary material. With our expanded coverage of linear models, we've compensated by reducing the recap into a single chapter. In that chapter, we highlight concepts we think you should already know. We'll leave it to you to check that you understand them clearly, and if not, or if by Chapter 4 you feel like you've jumped into the deep end, it may be time to revive your old course notes or do some more reading.

Models for Teaching

This book is a guide for researchers as well as a base for teaching, so it covers far more ground than can be covered in a single-semester subject. There are several options for using this book in class, depending on the course participants, their needs, and their backgrounds. We'll describe a few possibilities, based on our teaching experience.

A First Course for Graduate Students

The biggest challenge in teaching design and analysis to biology students is understanding just what they know as they enter the subject. Their backgrounds are diverse, ranging from recipes without background to math/stats majors. Even when we have an introductory statistics class as a prerequisite, little information may have been retained.

The first step is to assess the background of the students and the complexity of analyses they'll be

expected to do. Our experience is that some of the more complex designs are linked to disciplines in which data analysis is seen as simple and straightforward, so the question is more "what do students need to know to analyze their data appropriately?"

- In an ideal world, with an earlier subject completed, we would assume Chapter 2 and most of 3, and ask students to revise this material before starting. Our aim would then be to use Chapters 4–13 and 17 as the core, selecting relevant sections to keep the content manageable. Split-plot and repeated measures designs are common in biology, so our aim is to cover their analysis at the end. This is challenging for many students, especially those with weaker quantitative backgrounds. Chapters 5 and 17 would not require much time, and Chapter 9 may or may not be required.
- If students are less well prepared, we need to spend more time on the earlier chapters. Early subject time could cover Chapters 2–5 in more depth and build toward Chapter 8. Some parts of Chapter 13 could be included, and Chapters 10–12 are optional. Chapter 17 provides a little light relief at the end.

Do We Teach Multivariate Methods?

In the past, the multivariate introduction in Chapters 14–16 was used if the audience included ecology students, particularly community ecologists. The chapters are a jumping-off point to an extensive literature. In recent years, we have seen multivariate methods embedded in "cookbooks" for genomics, proteomics, etc. We think it's valuable for students using these biological techniques to be aware of how data are being treated.

Ordinary Least Squares or Maximum Likelihood?

Some of the core chapters are somewhat bulky because we describe OLS and ML approaches to fitting and interpreting specific models. We do this because many traditional approaches taught to biologists are based around normal distributions and OLS, transforming data where necessary. Generalized linear model approaches rely heavily on ML estimation. In our "holistic" approach to linear models, the GLM/ML path is consistent and provides easier entry into GLMMs, additive models (GAMs), etc. However, for now, most students' backgrounds are more likely to be around OLS regression, ANOVA, etc.

We wouldn't teach these approaches in parallel. If students have a good background in OLS approaches, it can be prudent to use OLS, and introduce ML for advanced models where necessary. If we have a blank slate, it's tempting to use ML throughout. More advanced models can be challenging for students, and they can feel overwhelmed when trying to come to grips with ML (and wrestle with *R*) at the same time. We suggest being flexible, depending on how far down the path (e.g. to Chapter 8, 12, or 13) we're trying to get by the end of a course.

Advanced Graduate Students

Students in mid-to-late degree stages may already have completed some training and have experience handling data. They may be ready for more advanced work. In this case, Chapters 10–12, and parts of 13 and 9 could be the basis of an introduction to mixed models, with Chapters 1–8 as assumed knowledge. Other combinations of chapters can provide different emphases.

Acronyms

AIC	Akaike information criterion
AIC_C	small sample adjustment of AIC
ALAN	artificial light at night
ANCOVA	analysis of covariance
ANOSIM	analysis of similarities
ANOVA	analysis of variance
AR	autoregressive
ASE	asymptotic standard error
BENT	bad evidence no test
BF	Bayes factor
BIC	Bayesian information criterion
BLUE	best linear unbiased estimator
BLUP	best linear unbiased predictor
BRT	boosted regression trees
CA	correspondence analysis
CART	classification and regression trees
CB	complete blocks
CCA	canonical correspondence analysis
CI	confidence interval
C-M-H	Cochran–Mantel–Haenszel (test)
CR	completely randomized
CV	coefficient of variation
CWD	coarse woody debris
db-RDA	distance-based redundancy analysis
DC	deciles of risk
DCA	detrended correspondence analysis
df	degrees of freedom
EM	expectation maximization
EMB	expectation maximization with bootstrapping
EMS	expected mean squares
ES	effect size
FA	factor analysis
FCS	fully conditional specification
FDR	false discovery rate
GAM	generalized additive model
GAMM	generalized additive mixed model
GCV	generalized cross-validation
GDM	generalized dissimilarity modeling
GLM	generalized linear model
GLMM	generalized linear mixed model; also GLME
GLS	generalized least squares
HSD	honestly significant difference
i.i.d.	independently and identically distributed
IRLS	iteratively reweighted least squares
LAD	least absolute deviations
LASSO	least absolute shrinkage and selection operator
LDA	linear discriminant analysis
LM	linear model
LME	linear mixed effect model; also LMM
LMG	Lindeman, Merenda, and Gold
LOESS	locally weighted sums of squares; also sometimes LOWESS
LR	likelihood ratio
LSD	least significant difference
MA	major axis
MAD	mean absolute deviation
MANOVA	multivariate analysis of variance
MAR	missing at random
MCAR	missing completely at random
MCMC	Markov chain Monte Carlo
MDA	malondialdehyde
MDES	minimum detectable effect size
MDS	multidimensional scaling
MI	multiple imputation
ML	maximum likelihood
MNAR	missing not at random
MRPP	multi-response permutation procedure
MS	mean squares
NB	negative binomial
NMDS	nonmetric multidimensional scaling
N-P	Neyman–Pearson
NR	Newton–Raphson
OAW	ocean acidification and warming
OLRE	observation-level random effect
OLS	ordinary least squares
OR	odds ratio

PC	principal component
PCA	principal components analysis
PCoA	principal coordinates analysis
PCR	principal components regression
pdf	probability density function
PERMANOVA	permutational analysis of variance
PERMDISP	permutational comparison of dispersion
PEV	proportion of explained variance
PLS	partial least squares
PMVD	proportional marginal variance decomposition
PSG	positively selected gene
PUFAs	polyunsaturated fatty acids
QDA	quadratic discriminant analysis
RA	reciprocal averaging
RCB	randomized complete blocks
RDA	redundancy analysis
REGW	Ryan–Einot–Gabriel–Welsch test
REML	restricted maximum likelihood
RM	repeated measures
RMA	reduced major axis
RMSE	root mean square error
RT	rank transformation
SD	standard deviation
SE	standard error
SIC	Schwarz information criterion
SIMPER	similarity percentages
SMA	standardized (or standard) major axis
SNK	Student–Neuman–Keuls (test)
SOC	soil organic carbon
SPLOM	scatterplot matrix
SS	sum of squares
SSCP	sums of squares and cross products
SVD	singular value decomposition
SW	sum of (Akaike) weights
TWINSPAN	two-way indicator species analysis
UBRE	unbiased risk estimator
UPGMA	unweighted pair-groups method using arithmetic averages
UPGMC	unweighted pair-groups method using centroids
UPMC	unplanned pairwise multiple comparison
VC	variance component
VIF	variance inflation factor
WHC	water-holding capacities
WLS	weighted least squares; see also GLS
WPGMA	weighted pair-groups method using arithmetic averages
ZA	zero-altered; also ZAB (binomial), ZAP (Poisson), and ZANB (negative binomial)
ZI	zero-inflated; also ZIB (binomial), ZIP (Poisson), and ZINB (negative binomial)