

Springer Series in Statistics

Peter McCullagh

Ten Projects in Applied Statistics



Springer Series in Statistics

Series Editors

Peter Bühlmann, Seminar für Statistik, ETH Zürich, Zürich, Switzerland

Peter Diggle, Dept. Mathematics, University Lancaster, Lancaster, UK

Ursula Gather, Dortmund, Germany

Scott Zeger, Baltimore, MD, USA

Springer Series in Statistics (SSS) is a series of monographs of general interest that discuss statistical theory and applications.

The series editors are currently Peter Bühlmann, Peter Diggle, Ursula Gather, and Scott Zeger. Peter Bickel, Ingram Olkin, and Stephen Fienberg were editors of the series for many years.

Peter McCullagh

Ten Projects in Applied Statistics



Springer

Peter McCullagh
Department of Statistics
University of Chicago
Chicago, IL, USA

ISSN 0172-7397 ISSN 2197-568X (electronic)
Springer Series in Statistics
ISBN 978-3-031-14274-1 ISBN 978-3-031-14275-8 (eBook)
<https://doi.org/10.1007/978-3-031-14275-8>

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Switzerland AG 2022

This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

To Rosa

Preface

Goals

The book begins with ten chapters, each devoted to a detailed discussion of a specific project. Some of these are projects that arose as part of the statistical consulting program at the University of Chicago; others are taken from recent publications in the scientific literature. The discussion specifically covers analyses that might seem superficially plausible, but are in fact misleading.

The areas of application range from medical and biological sciences to animal behavior studies, growth curves, time series, and environmental work. Statistical techniques are kept as simple as is reasonable to do justice to the scientific goals. They range from summary tables and graphs to linear models, generalized linear models, variance-component models, time series, spatial processes, and so on. Recognition of relationships among the observational units and the need to accommodate correlations among responses is a recurring theme.

The second half of the book begins by discussing a range of fundamental considerations that shape my attitude to applied statistics. Once they are pointed out, these matters appear so simple and obvious that it is hard to explain why a detailed discussion should be needed in an advanced-level text. But the simple fact that they are so frequently overlooked shows that this attitude is unhelpful. Most such matters are related to statistical design, the placement of the baseline, the identification of observational units and experimental units, the role of covariates and relationships, initial values, randomization and treatment assignment, and so on. Others are related to the interplay between design and modelling: Is the proposed model compatible with the design? More technical matters related to stochastic processes, including stationarity and isotropy, techniques for constructing spatio-temporal processes, likelihood functions, and so on are covered in later chapters. Parametric inference is important, but it is not a major or primary focus. More important by far is to estimate the right thing, however inefficiently, than to estimate the wrong thing with maximum efficiency.

The book is aimed at professional statisticians and at students who are already familiar with linear models but who wish to gain experience in the application of statistical ideas to scientific research. It is also aimed at scientific researchers in ecology, biology, or medicine who wish to use appropriate statistical methods in their work.

The primary emphasis is on the translation of experimental concepts and scientific ideas into stochastic models, and ultimately into statistical analyses and substantive conclusions. Although numerical computation plays a major role, it is not a driving force.

The aim of the book is not so much to illustrate algorithms or techniques of computation, but to illustrate the role of statistical thinking and stochastic modelling in scientific research. Typically, that cannot be accomplished without a detailed discussion of the scientific objectives, the experimental design, randomization, treatment assignment, baseline factors, response measurement, and so on. Before settling on a standard family of stochastic processes, the statistician must first ask whether the model is adequate as a formulation of the scientific goals. Is it self-consistent? Is it in conflict with randomization? Does it address adequately the sexual asymmetry in *Drosophila* courtship rituals? Is the sampling scheme adequate for the stated objective? A glance at Chaps. 1–5 shows the extent to which the discussion must be tailored to the specific needs of the application, and the unfortunate consequences of adopting an off-the-shelf model in a routine way. As D.R. Cox once remarked at an ISI meeting in 1979, “There are no routine statistical questions—only questionable statistical routines.”

Every analysis and every stochastic model is open to criticism on the grounds that it is not a good match for the application. A perfect match is a rarity, so compromise is needed in every setting, and a balance must be struck in order to proceed. At various points, I have included plausible analyses that are deficient, inappropriate, misleading, or simply incorrect. The hope is that students might learn not only from their own mistakes but from the mistakes of others.

Computation

The R package (R Core Team, 2015) is used throughout the book for all plots and computations. Initial data summaries may use built-in functions such as `apply(...)` or `tapply(...)` for one- and two-way tables of averages, or `plot(...)` for plots of seasonal trends or residuals. Apart from standard functions such as `lm(...)` for fitting linear models, `glm(...)` for generalized linear models, and `fft(...)` for the fast Fourier transformation, two additional packages are used for specialized purposes:

1. `regress(...)` (Clifford and McCullagh, 2006) for fitting linear Gaussian models having non-trivial spatial or temporal correlations;

2. `lmer`(. . .) (Bates et al., 2015) for fitting linear and generalized linear random-effects models.

Either package may be used for fitting the models in Chaps. 1 and 2; `glm()` is adequate for most computations in Chaps. 3 and 7; `regress()` is better suited to the needs of Chaps. 4 and 5; and `lmer()` is better suited for Chap. 9.

Organization

The book is not organized linearly by topic. From a logical standpoint, it might have been intellectually more satisfying to begin at the beginning with Chap. 11 and to illustrate the various statistical design concepts with examples drawn from the literature. That option was considered and quickly abandoned. A deliberate choice has been made to put as much emphasis on the projects as on the statistical methods and to draw upon whatever statistical techniques are required as and when they are required. For that reason, the projects come first. Thus, a reader who is unsure of the distinction between covariates and relationships or between observational and experimental units or the implications of those distinctions may consult the relevant portions of Chap. 11.

Several of the projects are taken from experiments reported in the recent scientific literature. In most cases, the experimental design is fairly easy to understand when the structure of the units is properly laid out. The importance of accommodating correlations arising from temporal or other relationships among the units is a recurring theme. And if that is the only lesson learned, the book will have served a useful purpose.

Acknowledgments

My attitude toward applied statistics has been shaped over many years by discussions with colleagues, particularly David Wallace, Mike Stein, Steve Stigler, Mei Wang, and Colm O’Muircheartaigh. The books on experimental design by Cox (1958), Mead (1988) and Bailey (2008), and the pair of papers Nelder (1965a,b) on randomized field experiments have been particularly influential. Cox and Snell (1981) is a trove of statistical wisdom and unteachable common sense.

For the past 25 years, I have worked closely with my colleague Mei Wang, encouraging graduate students and advising research workers on projects brought to the statistical consulting program at the University of Chicago. Over that period, we have given advice on perhaps 500 projects from a wide range of researchers, from all branches of the Physical and Biological Sciences to the Social Sciences, Public Policy, Law, Humanities, and even the Divinity School. All of the attitudes expressed in these notes are the product of direct experiences with researchers,

plus discussions with students and colleagues. Parts of three consulting projects are included in Chaps. 1, 3 and 10.

Other themes and whole chapters have emerged from an advanced graduate course taught at the University of Chicago, in which students are encouraged to work on a substantial applied project taken from the recent scientific literature. Chapter 5 is based on a course project selected by Wei Kuang in 2020. I have included a detailed analysis of this project because it raises a number of technical issues concerning factorial models that are well understood but seldom adequately emphasized in the applied statistical literature. Chapter 9 is based on course projects by Dongyue Xie, Irina Cristali, Lin Gui, and Y. Wei in 2018–2020. Chapter 16 was motivated by a 2020 masters project by Ben Harris on spatio-temporal variation in summer solar irradiance and its effect on solar power generation in downstate Illinois. All of the attitudes and opinions expressed in these analyses are mine alone.

Over the past few years, several students have tackled various aspects of the Out-of-Africa project. Some have gone beyond the call of duty, including Shane Miao for her Masters project at the University of Oxford in 2016, and Josephine Santoso for her Masters project at the University of Chicago in 2022.

I am indebted to students and colleagues, particularly Heather Battey, Laurie Butler, Emma McCullagh, Mike Stein, Steve Stigler, and Mei Wang, for reading and commenting on an earlier draft of the manuscript.

Chicago, IL, USA

Peter McCullagh

Contents

| | |
|--|-----------|
| 1 Rat Surgery | 1 |
| 1.1 Healing of Surgical Wounds | 1 |
| 1.2 An Elementary Analysis | 3 |
| 1.3 Two Incorrect Analyses | 4 |
| 1.4 Model Formulae | 5 |
| 1.5 A More Appropriate Formal Analysis | 6 |
| 1.6 Further Issues | 7 |
| 1.6.1 Exclusions | 7 |
| 1.6.2 Missing Components | 8 |
| 1.6.3 Back-Transformation | 9 |
| 1.7 Summary of Statistical Concepts | 10 |
| 1.8 Exercises | 10 |
| 2 Chain Saws | 15 |
| 2.1 Efficiency of Chain Saws | 15 |
| 2.2 Covariate and Treatment Factors | 16 |
| 2.3 Goals of Statistical Analysis | 17 |
| 2.4 Formal Models | 19 |
| 2.5 REML and Likelihood Ratios | 20 |
| 2.6 Summary of Conclusions | 21 |
| 2.7 Exercises | 22 |
| 3 Fruit Flies | 25 |
| 3.1 Diet and Mating Preferences | 25 |
| 3.2 Initial Analyses | 26 |
| 3.2.1 Assortative Mating | 26 |
| 3.2.2 Initial Questions and Exercises | 27 |
| 3.3 Refractory Effects | 28 |
| 3.3.1 More Specific Mating Counts | 28 |
| 3.3.2 Follow-Up Analyses | 30 |
| 3.3.3 Lexis Dispersion | 31 |
| 3.3.4 Is Under-Dispersion Possible? | 32 |

| | | |
|----------|--|-----------|
| 3.3.5 | Independence..... | 33 |
| 3.3.6 | Acknowledgement | 36 |
| 3.4 | Technical Points..... | 36 |
| 3.4.1 | Hypergeometric Simulation by Random Matching | 36 |
| 3.4.2 | Pearson's Statistic | 37 |
| 3.5 | Further Drosophila Project | 38 |
| 3.6 | Exercises | 40 |
| 4 | Growth Curves..... | 43 |
| 4.1 | Plant Growth: Data Description | 43 |
| 4.2 | Growth Curve Models | 45 |
| 4.3 | Technical Points..... | 47 |
| 4.3.1 | Non-linear Model with Variance Components | 47 |
| 4.3.2 | Fitted Versus Predicted Values | 48 |
| 4.4 | Modelling Strategies | 51 |
| 4.5 | Miscellaneous R Functions | 52 |
| 4.6 | Exercises | 52 |
| 5 | Louse Evolution..... | 55 |
| 5.1 | Evolution of Lice on Captive Pigeons | 55 |
| 5.1.1 | Background | 55 |
| 5.1.2 | Experimental Design | 56 |
| 5.1.3 | Deconstruction of the Experimental Design..... | 56 |
| 5.2 | Data Analysis | 58 |
| 5.2.1 | Role of Tables and Graphs | 58 |
| 5.2.2 | Trends in Mean Squares | 59 |
| 5.2.3 | Initial Values and Factorial Subspaces | 62 |
| 5.2.4 | A Simple Variance-Components Model | 63 |
| 5.2.5 | Conformity with Randomization | 64 |
| 5.3 | Critique of Published Claims | 66 |
| 5.4 | Further Remarks | 68 |
| 5.4.1 | Role of Louse Sex | 68 |
| 5.4.2 | Persistence of Initial Patterns..... | 69 |
| 5.4.3 | Observational Units | 70 |
| 5.5 | Follow-Up | 71 |
| 5.5.1 | New Design Information | 71 |
| 5.5.2 | Modifications to Analyses | 73 |
| 5.5.3 | Further Remarks | 75 |
| 5.6 | Exercises | 75 |
| 6 | Time Series I | 81 |
| 6.1 | A Meteorological Temperature Series | 81 |
| 6.2 | Seasonal Cycles | 82 |
| 6.2.1 | Means and Variances | 82 |
| 6.2.2 | Skewness and Kurtosis | 84 |

| | | |
|-------|--|-----|
| 6.3 | Annual Statistics | 86 |
| 6.3.1 | Means and Variances | 86 |
| 6.3.2 | Variance of Block Averages | 88 |
| 6.3.3 | Variogram at Short and Long Lags..... | 89 |
| 6.4 | Stochastic Models for the Seasonal Cycle | 91 |
| 6.4.1 | Structure of Observational Units | 91 |
| 6.4.2 | Seasonal Structure | 92 |
| 6.4.3 | Stationary Periodic Processes | 93 |
| 6.5 | Estimation of Secular Trend | 93 |
| 6.5.1 | Gaussian Estimation and Prediction | 93 |
| 6.5.2 | Application to Trend Estimation | 94 |
| 6.5.3 | Matérn Models | 94 |
| 6.5.4 | Statistical Tests and Likelihood Ratios | 95 |
| 6.5.5 | Rough Paths Versus Smooth Paths | 96 |
| 6.5.6 | Smooth Versus Ultra-Smooth Paths..... | 96 |
| 6.6 | Exercises | 97 |
| 7 | Time Series II | 103 |
| 7.1 | Frequency-Domain Analyses..... | 103 |
| 7.1.1 | Fourier Transformation | 103 |
| 7.1.2 | Anova Decomposition by Frequency | 104 |
| 7.2 | Temperature Spectrum | 105 |
| 7.2.1 | Spectral Plots | 105 |
| 7.2.2 | A Parametric Spectral Model..... | 106 |
| 7.3 | Stationary Temporal Processes | 110 |
| 7.3.1 | Stationarity | 110 |
| 7.3.2 | Visualization of Trajectories | 111 |
| 7.3.3 | Whittle Likelihood | 114 |
| 7.4 | Exercises | 115 |
| 8 | Out of Africa | 117 |
| 8.1 | Linguistic Diversity | 117 |
| 8.2 | Phoneme Inventory | 118 |
| 8.3 | Distances | 119 |
| 8.4 | Maps and Scatterplots | 120 |
| 8.5 | Point Estimates and Confidence Regions | 123 |
| 8.5.1 | Simple Version | 123 |
| 8.5.2 | Accommodating Correlations | 125 |
| 8.6 | Matters for Further Consideration | 127 |
| 8.6.1 | Phoneme Inventory as Response | 127 |
| 8.6.2 | Vowels, Consonants and Tones..... | 128 |
| 8.6.3 | Granularity | 128 |
| 8.7 | Follow-Up Project | 129 |
| 8.7.1 | Extended Data Frame | 129 |
| 8.7.2 | An Elementary Misconception | 130 |
| 8.8 | Exercises | 131 |

| | | |
|-----------|--------------------------------|-----|
| 9 | Environmental Projects | 133 |
| 9.1 | Effects of Atmospheric Warming | 133 |
| 9.1.1 | The Experiment | 133 |
| 9.1.2 | The Data | 134 |
| 9.1.3 | Exercises | 135 |
| 9.2 | The Plight of the Bumblebee | 136 |
| 9.2.1 | Introduction | 136 |
| 9.2.2 | Risk of Infection | 136 |
| 9.2.3 | Mixed Models | 139 |
| 9.2.4 | Exchangeability | 140 |
| 9.2.5 | Role of GLMs and GLMMs | 141 |
| 9.3 | Two Further Projects | 142 |
| 9.4 | Exercises | 143 |
| 10 | Fulmar Fitness | 145 |
| 10.1 | The Eynhallow Colony | 145 |
| 10.1.1 | Background | 145 |
| 10.1.2 | The Eynhallow Breeding Record | 146 |
| 10.1.3 | The Breeding Sequence | 147 |
| 10.1.4 | Averages for Cohorts | 148 |
| 10.1.5 | Averages for Disjoint Subsets | 150 |
| 10.1.6 | Resolution of a Paradox | 151 |
| 10.2 | Formal Models | 152 |
| 10.2.1 | A Linear Gaussian Model | 152 |
| 10.2.2 | Prediction | 154 |
| 10.2.3 | Model Adequacy | 154 |
| 10.3 | Mark-Recapture Designs | 156 |
| 10.4 | Further References | 157 |
| 10.5 | Exercises | 157 |
| 11 | Basic Concepts | 159 |
| 11.1 | Stochastic Processes | 159 |
| 11.1.1 | Process | 159 |
| 11.1.2 | Probability | 160 |
| 11.1.3 | Self-consistency | 161 |
| 11.1.4 | Statistical Model | 162 |
| 11.2 | Samples | 163 |
| 11.2.1 | Baseline | 163 |
| 11.2.2 | Observational Unit | 164 |
| 11.2.3 | Population | 164 |
| 11.2.4 | Biological Populations | 165 |
| 11.2.5 | Samples and Sub-samples | 166 |
| 11.2.6 | Illustrations | 167 |

| | | |
|-------------|---------------------------------------|-----|
| 11.3 | Variables | 168 |
| 11.3.1 | Ordinary Variables | 168 |
| 11.3.2 | Relationship | 172 |
| 11.3.3 | External Variable | 173 |
| 11.4 | Comparative Studies | 175 |
| 11.4.1 | Randomization | 175 |
| 11.4.2 | Experimental Unit | 176 |
| 11.4.3 | Covariate and Treatment Effects | 177 |
| 11.4.4 | Additivity | 178 |
| 11.4.5 | Design | 178 |
| 11.4.6 | Replication | 179 |
| 11.4.7 | Independence | 179 |
| 11.4.8 | Interference | 179 |
| 11.4.9 | State Space | 180 |
| 11.4.10 | State-Space Evolution | 181 |
| 11.4.11 | Longitudinal Study | 182 |
| 11.4.12 | Cemetery State | 182 |
| 11.5 | Non-comparative Studies | 183 |
| 11.5.1 | Examples | 183 |
| 11.5.2 | Stratified Population | 183 |
| 11.5.3 | Heterogeneity | 183 |
| 11.5.4 | Random Sample | 184 |
| 11.5.5 | Stratified Random Sample | 184 |
| 11.5.6 | Accessibility | 184 |
| 11.5.7 | Population Averages | 185 |
| 11.5.8 | Target of Estimation I | 185 |
| 11.5.9 | Inverse Probability Weighting | 185 |
| 11.5.10 | Target of Estimation II | 186 |
| 11.6 | Interpretations of Variability | 188 |
| 11.6.1 | A Tale of Two Variances | 188 |
| 11.6.2 | Which Variance Is Appropriate? | 191 |
| 11.7 | Exercises | 192 |
| 12 | Principles | 197 |
| 12.1 | Sampling Consistency | 197 |
| 12.2 | Adequacy for the Application | 200 |
| 12.3 | Likelihood Principle | 201 |
| 12.4 | Attitudes | 204 |
| 12.5 | Exercises | 206 |
| 13 | Initial Values | 211 |
| 13.1 | Randomization Protocols | 211 |
| 13.2 | Four Gaussian Models | 212 |
| 13.2.1 | Distribution and Likelihood | 215 |
| 13.2.2 | Numerical Comparison of Estimates | 216 |

| | | |
|-----------|---|------------|
| 13.2.3 | Initial Values Versus Covariates | 217 |
| 13.2.4 | Initial Values in an Observational Study | 218 |
| 13.3 | Exercises | 219 |
| 14 | Probability Distributions | 223 |
| 14.1 | Exchangeable Processes | 223 |
| 14.1.1 | Unconditional Exchangeability | 223 |
| 14.1.2 | Regression Processes | 224 |
| 14.1.3 | Block Exchangeability | 224 |
| 14.1.4 | Stationarity | 225 |
| 14.1.5 | Exchangeability | 225 |
| 14.1.6 | Axiomatic Point | 226 |
| 14.1.7 | Block Randomization | 226 |
| 14.2 | Families with Independent Components | 227 |
| 14.2.1 | Parametric Models | 227 |
| 14.2.2 | IID Model I | 227 |
| 14.2.3 | IID Model II | 228 |
| 14.3 | Non-i.d. Models | 229 |
| 14.3.1 | Classification Factor | 229 |
| 14.3.2 | Treatment | 231 |
| 14.3.3 | Classification Factor Plus Treatment | 232 |
| 14.3.4 | Quantitative Covariate Plus Treatment | 233 |
| 14.3.5 | Random Coefficient Models | 234 |
| 14.4 | Examples of Treatment Effects | 236 |
| 14.4.1 | Simple Gaussian Model Without Interaction | 236 |
| 14.4.2 | Additive Interaction | 237 |
| 14.4.3 | Survival Models | 237 |
| 14.5 | Incomplete Processes | 240 |
| 14.5.1 | Gosset Process | 240 |
| 14.5.2 | Factual and Counterfactual Processes | 242 |
| 14.5.3 | Limitations of Incomplete Processes | 244 |
| 14.6 | Exercises | 246 |
| 15 | Gaussian Distributions | 251 |
| 15.1 | Real Gaussian Distribution | 251 |
| 15.1.1 | Density and Moments | 251 |
| 15.1.2 | Gaussian Distribution on \mathbb{R}^n | 252 |
| 15.2 | Complex Gaussian Distribution | 254 |
| 15.2.1 | One-Dimensional Distribution | 254 |
| 15.2.2 | Gaussian Distribution on \mathbb{C}^n | 254 |
| 15.2.3 | Moments | 255 |
| 15.3 | Gaussian Hilbert Space | 256 |
| 15.3.1 | Euclidean Structure | 256 |
| 15.3.2 | Cautionary Remarks | 257 |
| 15.3.3 | Projections | 258 |
| 15.3.4 | Dual Space of Linear Combinations | 261 |

| | | |
|--------|---|-----|
| 15.4 | Statistical Interpretations | 262 |
| 15.4.1 | Canonical Norm | 262 |
| 15.4.2 | Independence | 262 |
| 15.4.3 | Prediction and Conditional Expectation | 264 |
| 15.4.4 | Eddington's Formula | 267 |
| 15.4.5 | Linear Regression | 269 |
| 15.4.6 | Linear Regression and Prediction | 270 |
| 15.5 | Additivity | 272 |
| 15.5.1 | 1DOFNA Algorithm | 272 |
| 15.5.2 | 1DOFNA Theory | 273 |
| 15.5.3 | Scope and Rationale | 274 |
| 15.6 | Exercises | 275 |
| 16 | Space-Time Processes | 279 |
| 16.1 | Gaussian Processes | 279 |
| 16.2 | Stationarity and Isotropy | 281 |
| 16.2.1 | Definitions | 281 |
| 16.2.2 | Stationarity on Increments | 282 |
| 16.2.3 | Stationary Process on $\mathbb{Z} \text{ (mod } k)$ | 283 |
| 16.3 | Stationary Gaussian Time Series | 284 |
| 16.3.1 | Spectral Representation | 284 |
| 16.3.2 | Matérn Class | 285 |
| 16.4 | Stationary Spatial Process | 286 |
| 16.4.1 | Spectral Decomposition | 286 |
| 16.4.2 | Matérn Spatial Class | 288 |
| 16.4.3 | Illustration by Simulation | 291 |
| 16.5 | Covariance Products | 296 |
| 16.5.1 | Hadamard Product | 296 |
| 16.5.2 | Separable Products and Tensor Products | 297 |
| 16.6 | Real Spatio-Temporal Process | 298 |
| 16.6.1 | Covariance Products | 298 |
| 16.6.2 | Examples of Covariance Products | 300 |
| 16.6.3 | Travelling Wave | 303 |
| 16.6.4 | Perturbation Theory | 305 |
| 16.7 | Hydrodynamic Processes | 306 |
| 16.7.1 | Frame of Reference | 306 |
| 16.7.2 | Rotation and Group Action | 307 |
| 16.7.3 | Action on Matrices | 309 |
| 16.7.4 | Borrowed Products | 309 |
| 16.7.5 | Hydrodynamic Symmetry | 310 |
| 16.8 | Summer Cloud Cover in Illinois | 311 |
| 16.9 | More on Gaussian Processes | 314 |
| 16.9.1 | White Noise | 314 |
| 16.9.2 | Limit Processes | 315 |
| 16.10 | Exercises | 320 |

| | |
|--|-----|
| 17 Likelihood | 327 |
| 17.1 Introduction | 327 |
| 17.1.1 Non-Bayesian Model | 327 |
| 17.1.2 Bayesian Resolution | 328 |
| 17.2 Likelihood Function | 329 |
| 17.2.1 Definition | 329 |
| 17.2.2 Bartlett Identities | 330 |
| 17.2.3 Implications for Estimation | 332 |
| 17.2.4 Likelihood-Ratio Statistic I | 333 |
| 17.2.5 Profile Likelihood | 334 |
| 17.2.6 Two Worked Examples | 335 |
| 17.3 Generalized Linear Models | 337 |
| 17.4 Variance-Components Models | 339 |
| 17.5 Mixture Models | 339 |
| 17.5.1 Two-Component Mixtures | 339 |
| 17.5.2 Likelihood-Ratio Statistic | 341 |
| 17.5.3 Sparse Signal Detection | 342 |
| 17.6 Inferential Compromises | 343 |
| 17.7 Exercises | 345 |
| 18 Residual Likelihood | 349 |
| 18.1 Background | 349 |
| 18.2 Simple Linear Regression | 351 |
| 18.3 The REML Likelihood | 351 |
| 18.3.1 Projections | 351 |
| 18.3.2 Determinants | 352 |
| 18.3.3 Marginal Likelihood with Arbitrary Kernel | 352 |
| 18.3.4 Likelihood Ratios | 353 |
| 18.4 Computation | 354 |
| 18.4.1 Software Options | 354 |
| 18.4.2 Likelihood-Ratios | 355 |
| 18.4.3 Testing for Interaction | 356 |
| 18.4.4 Singular Models | 358 |
| 18.5 Exercises | 358 |
| 19 Response Transformation..... | 363 |
| 19.1 Likelihood for Gaussian Models | 363 |
| 19.2 Box-Cox Transformation | 364 |
| 19.2.1 Power Transformation | 364 |
| 19.2.2 Re-scaled Power Transformation | 365 |
| 19.2.3 Worked Example | 366 |
| 19.2.4 Transformation and Residual Likelihood | 368 |
| 19.3 Quantile-Matching Transformation | 370 |
| 19.4 Exercises | 371 |

| | |
|---|-----|
| Contents | xix |
| 20 Presentations and Reports | 375 |
| 20.1 Coaching Tips I | 375 |
| 20.2 Coaching Tips II | 380 |
| 20.3 Exercises | 383 |
| 21 Q & A | 385 |
| 21.1 Scientific Investigations | 385 |
| 21.1.1 Observational Unit | 385 |
| 21.1.2 Clinical Trials | 387 |
| 21.1.3 Agricultural Field Trials | 391 |
| 21.1.4 Covariates | 393 |
| 21.1.5 Matched Design | 395 |
| 21.1.6 The Effect of Treatment | 397 |
| References | 401 |
| Index | 407 |

Chapter 1

Rat Surgery



1.1 Healing of Surgical Wounds

The data shown in Table 1.1 were obtained in an experiment by Dr. George Huang of the Department of Surgery at the University of Chicago, the purpose of which was to investigate the effect of hyperbaric O₂ treatment on the healing of surgical wounds in diabetic rats. (Diabetics, both human and animal, tend to have more complications following surgery than non-diabetics, and these rats made the ultimate murine sacrifice by serving as the surgical model for diabetic effects in humans.) Thirty rats were first given a drug that has the effect of destroying the pancreas, with the goal of making the rats diabetic. All the rats underwent surgery, during which an incision was made along the entire length of the back. This was immediately sewn up with surgical staples, and the rats were returned to their cages.

The treatment group of fifteen rats was subjected to hyperbaric O₂ treatment, i.e., a 100% oxygen environment at two atmospheres of pressure, for 90 minutes per day following surgery. The control group also received a similar treatment for 90 minutes daily, but at standard oxygen concentration and normal atmospheric pressure. Six rats had glucose levels that were deemed too low to be considered diabetic, and were excluded from the analysis. (You may assume initially that these exclusions are unrelated to the O₂ treatment.) After a 24 day recuperation period, the 24 rats still participating in the experiment were sacrificed, i.e., killed. Strips of skin were taken from five sites labelled A–E on each rat, each site crossing the surgical scar in a right angle. The strips were put on a tensiometer, stretched to the breaking point, and the energy required to break the specimen was recorded. Unfortunately some prepared specimens were deemed sub-par for procedural reasons unconnected with skin strength, and in such cases no observation could be made: the unmeasured

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-14275-8_1.

Table 1.1 Strength of skin specimens 24 days after surgery

| Rat | Site on the back: shoulder to tail | | | | | Mean |
|-----------------|------------------------------------|--------|--------|--------|--------|--------|
| | A | B | C | D | E | |
| 1 [†] | 3.8300 | 7.3788 | 44.353 | 19.555 | – | 18.779 |
| 2 | 27.861 | 29.974 | 15.470 | 23.455 | – | 24.190 |
| 3 | 56.996 | 60.960 | 20.306 | – | 28.123 | 41.596 |
| 4 | – | 38.043 | 68.080 | 42.425 | 30.335 | 44.721 |
| 5 | 16.276 | – | 59.033 | 73.891 | – | 49.733 |
| 6 | 38.267 | 33.702 | 35.558 | 44.598 | 32.678 | 39.961 |
| 7 | 9.0384 | 11.259 | 27.121 | 31.984 | – | 19.851 |
| 8 | 16.728 | 27.590 | 13.238 | 12.139 | 6.3865 | 15.216 |
| 9 | 11.866 | 27.983 | 26.226 | 15.594 | 19.225 | 20.179 |
| 10 | 23.352 | 34.790 | 27.556 | 35.883 | 22.848 | 28.888 |
| 11 | 16.444 | 31.928 | 21.495 | 15.590 | 7.0750 | 18.506 |
| 12 | 23.342 | 46.313 | 33.810 | 15.686 | – | 29.788 |
| 13 | 15.267 | 14.452 | 10.635 | 22.156 | 6.8062 | 13.863 |
| 14 | 21.732 | 20.746 | 12.293 | 17.295 | 10.301 | 16.473 |
| 15 [†] | 82.508 | 13.645 | 49.187 | – | 53.432 | 49.693 |
| 16 | – | 45.919 | 63.090 | 68.137 | 36.500 | 53.412 |
| 17 | 80.147 | 29.943 | 71.928 | – | 46.609 | 57.157 |
| 18 | 31.938 | – | 36.211 | 49.815 | 44.468 | 40.608 |
| 19 | 15.453 | 31.384 | 27.127 | 27.961 | 9.9035 | 22.366 |
| 20 | 21.183 | 27.429 | 20.058 | – | – | 22.890 |
| 21 | 20.445 | 12.532 | 15.661 | 28.694 | – | 19.333 |
| 22 | 16.928 | 59.579 | 29.407 | 18.626 | 8.8352 | 26.675 |
| 23 | 35.631 | 21.613 | 23.155 | 42.379 | 16.203 | 27.796 |
| 24 | 20.523 | 24.621 | 16.292 | – | 18.680 | 20.029 |
| Mean | 27.534 | 29.627 | 31.970 | 31.888 | 23.436 | 29.126 |

[†]Rats 1–14 are hyperbaric O₂-treated; 15–24 are the controls
(Data published with permission from Dr. George Huang.)

specimens are indicated by – in the table. Rats 1–14 received the hyperbaric treatment; rats 15–24 were the controls.

Handling by humans is known to be stressful for rats, and stress is associated with poor health and shorter lifetimes. The experiment was designed to ensure that treated and control rats were handled in a similar manner throughout the experiment, so that any systematic differences between the groups could confidently be attributed to treatment rather than to differences in the way that the rats were handled. Hence, the control rats were inserted daily into the hyperbaric chamber so that they might experience the same stress levels as the treated rats.

The main objective is to determine whether or not hyperbaric O₂ treatment has an effect on the healing of surgical wounds, and if so, whether the effect depends on the position of the wound on the back. It was anticipated that the increased oxygen effect would be beneficial, or at least not detrimental, for healing, and that the site

effects would be small or negligible. Confidence intervals or posterior distributions for the size of any such effects are a part of the answer.

1.2 An Elementary Analysis

It is unclear whether in fact the treatment was assigned to the rats by a scheme that could be described by a statistician as objective randomization. But we can be assured that no effort was made to select rats differentially for the two groups so, after tut-tutting, it is reasonable to proceed as if the assignment were randomized. In principle, this is a completely randomized design with no blocking. Each rat-site pair is one observational unit, and each rat is one experimental unit, so each experimental unit consists of five observational units. The distinction between observational units and experimental units is crucial in the design, in the analysis and in the interpretation of results: see Sect. 11.4.2.

If there were no missing values, the analysis would be relatively straightforward, so we first illustrate the reasoning behind the simpler analysis. First, the observations are strictly positive strength measurements having a moderately large dynamic range from 3.8 to 82.5, so a log transformation is more or less automatic. Normality of residuals is important, but it is not nearly so important as additivity assumptions that are made in a typical linear model—in this case additivity of rat effects, site effects and treatment effects. Although it is easy to check for symmetry in the distribution of residuals or in marginal histograms, it is arguably misleading to point to skewness or non-normality as the principal reason for transformation.

To each experimental unit there corresponds an average response, giving 14 values for O₂-treated rats, and ten values for control rats. In the absence of missing components, the site effects contribute equally to rat averages, so the averages are not contaminated by additive differences that may be present among sites. The sample means for control and treated rats are 3.372 and 3.113 on the log scale, the sample variances are 0.174 and 0.200 respectively, and the pooled variance is

$$\frac{9 \times 0.174 + 13 \times 0.200}{22} = 0.189$$

on 22 degrees of freedom. This analysis, which is based on the rat averages, leads to an estimated treatment effect

$$\text{average for treated rats} - \text{average for control rats} = -0.259$$

with standard error $\sqrt{0.189(1/10 + 1/14)} = 0.180$. The estimated effect is only 1.4 standard deviations from the null value of zero, and the deviation is in the direction not anticipated. The conclusion from this analysis is that there is no evidence of a treatment effect—positive or negative.

This elementary arithmetic analysis is standard for a design that is complete with no missing observational units. It is open to criticism in this setting because the

comparison may be unfair if site effects are appreciable and the pattern of missing treated units is substantially different from the pattern for controls. For example, site D is missing for 40% of the control rats, but only for 7% of treated rats. If the response at site D were appreciably higher than that at other sites, the pattern of missing units would create a bias in the treatment comparison. However, the site averages on the log scale

$$3.093, 3.263, 3.317, 3.326, 2.928,$$

show little indication of appreciable differences or a strong trend, so the preceding analysis appears reasonably sound. Nonetheless, it is only natural to ask for a more definitive analysis taking into account the possibility of additive site effects.

1.3 Two Incorrect Analyses

One way to adjust for site effects is to fit a simple linear Gaussian model in which site and treatment effects are additive on the log scale:

$$E(Y_{is}) = \beta_0 + \beta_{t(i)} + \beta_s; \quad \text{var}(Y_{is}) = \sigma^2, \quad (1.1)$$

where s is the site, and $t(i)$ is the treatment indicator for rat i . The least-squares treatment-effect estimate is -0.298 with standard error 0.119 , which is computed from the residual sum of squares of 34.30 on 98 degrees of freedom. According to this analysis, the treatment estimate is 2.5 standard errors away from its null value, a magnitude that is sufficient to make a case for publication in certain scientific journals, even if its direction is opposite to that anticipated.

Although the error in this analysis may seem obvious, the glib partial description in (1.1) is extremely common in the scientific literature. Very often, the model is stated additively in the form

$$Y_{is} = \beta_0 + \beta_{t(i)} + \beta_s + \varepsilon_{is}.$$

Implicitly or explicitly, the errors ε_{is} are assumed to be independent Gaussian with constant variance. Failure to account for correlations between different observations on the same experimental unit has little effect on the point estimate of the treatment effect, but it has a more substantial effect on the variance estimate.

It is good to bear in mind that there cannot be more degrees of freedom for the estimation of treatment contrasts than there are experimental units in the design. This design has 24 experimental units split into two subsets, so there cannot be more than 22 degrees of freedom for the estimation of inter-unit experimental variability. Thus, failure to mention covariances in the linear model specification, and the claim of 98 degrees of freedom are two red-flag indicators of gross statistical transgressions.

One way to adjust for rat-to-rat variability is to include an additive rat effect:

$$E(Y_{is}) = \beta_0 + \beta_{t(i)} + \beta_s + \gamma_i; \quad \text{var}(Y_{is}) = \sigma^2.$$

For example, this model can be fitted in R using the commands

```
fit <- lm(log(y)~site+treat+rat); anova(fit)
```

The ANOVA function reports a very substantial F -ratio of 10.45 for treatment, with a p -value of 0.2%. However, the treatment effect estimate is only -0.600 ± 0.33 , and the p -value for the hypothesis of no effect is a more modest 7%. We will not attempt here to explain this apparent contradiction because the displayed code points to a serious lack of understanding of linear algebra, geometry, orthogonal projections, and their connection with statistical models. Neither part of the code or the computation is appropriate, and the fitted model is not suited for its intended purpose.

1.4 Model Formulae

In discussions concerning least-squares coefficients and statistical model formulae, it is good to remember that each term in a linear-model formula is first and foremost a vector subspace of \mathbb{R}^n , where n is the number or set of observational units. In this context, the binary operator $+$ denotes the span of subspaces, not a vector sum.

Each subspace associated with a factor has a natural basis consisting of one indicator vector for each factor level. However, certain statistical questions are concerned with the subspace, in which case statistical conclusions are, or should be, unaffected by the choice of basis: see Exercise 1.8. For example, `site`, `treat` and `rat` are subspaces of dimensions 5, 2 and 24 respectively, which is the number of levels of the factor that occur in the design. Every factor subspace includes the one-dimensional subspace **1** of constant vectors, which is the span of the sum of the indicator vectors. In most situations, the intersection of a pair of factor subspaces such as `site` and `rat`, or `site` and `treat`, is precisely this one-dimensional subspace. However, the fact that treatment is assigned to rats means that `treat` is a subspace of `rat`, so `treat+rat = rat`. Treatment effects are said to be confounded with rat effects.

Confounding may be complete or partial. Numerical linear-algebra algorithms detect this confounding, and they resolve it by picking the most convenient subset of the basis vectors on offer. This subset is invariably rather arbitrary, which explains part of the problem in the paragraph at the end of the preceding section. As a result, the numbers reported there are statistically uninteresting, and they are potentially misleading if the algebraic issues are not fully understood.

Each subspace associated with a *block factor* such as `rat` also has a natural indicator basis. Since each rat is one experimental unit, it is implicit that the associated effects are not entirely arbitrary, but are judged *a priori* to be statistically

exchangeable. In a sense, randomization guarantees exchangeability of effects. The effects referred to in this setting are the coefficients of the basis vectors, and specifically the coefficients of the indicator basis for the experimental units, so the indicator basis for the subspace `rat` is not on an equal statistical footing with any other basis. Thus `regress(y~rat)` means that the coefficients are arbitrary real numbers to be estimated, so the basis for the subspace is immaterial. By contrast, `regress(y~1, ~rat)` means that the coefficients of the indicator basis are exchangeable Gaussian random variables, so the covariance includes `rat` as a block-factor in matrix form.

The exchangeability argument does not apply with equal force to a classification factor such as `site`.

For the most part, these remarks are unaffected by replication or by the pattern of missing components in the design.

1.5 A More Appropriate Formal Analysis

The default Gaussian model for the log-transformed measurements Y_{is} incorporates site and treatment effects additively as follows:

$$E(Y_{is}) = \beta_s + \beta_{t(i)}; \quad \text{cov}(Y_{is}, Y_{jt}) = \sigma_0^2 \delta_{ij} \delta_{st} + \sigma_1^2 \delta_{ij}, \quad (1.2)$$

where δ_{ij} is the Kronecker symbol for equality of subscripts. Computationally speaking, `treat` and `site` are two classification factors, which determine subspaces of dimensions two and five respectively, whereas `rat` is encoded as a block factor or symmetric indicator matrix with (is, jt) -component $rat(is, jt) = \delta_{ij}$, and coefficient σ_1^2 . The overall covariance matrix in (1.2) is invariant with respect to permutation of rats and permutation of sites, but, unlike (1.1), it is not invariant with respect to arbitrary permutation of observational units.

Expression (1.2) is equivalent to the distributional statement $Y \sim N_n(\mu, \Sigma)$ in which $\mu = E(Y)$ belongs to the subspace `site+treat`, and $\Sigma = \text{cov}(Y)$ belongs to the convex cone spanned by the identity and `rat` as a block factor. The equivalent expression in terms of additive effects and random variables is

$$Y_{is} = \beta_s + \beta_{t(i)} + (\sigma_1 \epsilon_i + \sigma_0 \epsilon_{is})$$

in which all effects contributing only to variances and covariances are shown in parentheses. Independence of components is not to be taken for granted, so it is necessary to state explicitly that the rat effects $\epsilon_1, \dots, \epsilon_{24}$ are independent and identically distributed with zero mean, and are independent of the 120 standard Gaussian residual effects ϵ_{is} , which are also mutually independent.

All told, there are five site parameters, one treatment parameter, and two variance components whose estimates are $\hat{\sigma}_0^2 = 0.211$ and $\hat{\sigma}_1^2 = 0.148$. Observations on distinct rats are independent, but the covariance between observations at different

sites on the same rat is σ_1^2 , and the correlation is $\sigma_1^2/(\sigma_0^2 + \sigma_1^2)$, which is estimated as 0.41.

In standard software, the site and treatment parameters are estimated by weighted least squares using the inverse of the fitted covariance matrix as weights. The treatment effect estimate, which is automatically adjusted for additive site effects, is -0.294 with standard error 0.184. If the design were complete with no missing values, the null distribution of the ratio would be t_{22} , so the observed effect corresponds to a two-sided p -value of about 12%. The likelihood-ratio statistic of 2.51 on one degree of freedom gives an essentially identical conclusion. To be clear, this is the version recommended in Sects. 18.3–18.4: see Exercise 1.14. The less recommended version using ordinary maximum likelihood is typically somewhat larger—2.62 in this instance.

The code shown in Exercise 1.4 reports fitted site effects

$$0.000, 0.158, 0.181, 0.260, -0.271$$

in head-to-tail order with the anterior site (nearest to the head) as reference. The standard errors of pairwise site contrasts are in the range 0.14–0.15, so it appears that skin from the caudal site is appreciably weaker than that from other sites. The REML log likelihood ratio statistic (Chap. 18) for testing equality of site effects is 13.92, which is beyond the 99th percentile of the limiting null distribution, which is χ_4^2 . Although they appear to be non-zero, the site effects are not sufficiently large to change appreciably the conclusions about treatment reached by the more elementary analysis based on rat averages.

It is mathematically and biologically possible that the treatment could have an effect on either the mean or on the variance or on both, but the standard default formulation assumes that treatment affects only the mean of the distribution. Equality of the two variances is an entirely reasonable assumption in practice, but it is also an assumption that can easily be checked by including two different variance components, one for treated rats and one for the controls. If the variance for treated rats were appreciably different from the variance for controls, it would not be possible to encode the treatment effect in a single number. For these data, there is absolutely no evidence of an effect of treatment on variances: see Exercise 1.14.

1.6 Further Issues

1.6.1 Exclusions

Six rats that were deemed non-diabetic on the basis of post-baseline glucose measurements were excluded from the main analysis. For the main goal of this study, this exclusion was judged to be scientifically reasonable on the basis of an argument that implies that the probability of exclusion is unrelated to treatment.

However, the excluded rats consisted of five controls and only one treated rat. How extreme is that allocation relative to expectation? Does it suggest that treated rats are less likely to be excluded than the controls? If so, treatment may have an effect of an entirely different nature.

Given that six rats were excluded, the number of excluded controls is a random variable whose distribution is central hypergeometric

$$\binom{6}{y} \binom{24}{15-y} / \binom{30}{15};$$

the numerical values are 0.8, 7.6, 24.1, 34.9, 24.1, 7.6, 0.8 in percentages for $y = 0, \dots, 6$. The probability of an allocation at least as extreme as that observed is 8.4% in each tail. Exclusions and other departures from protocol must always be described and included as a part of the discussion. The imbalance in this study is greater than we might have wished for, but it is not sufficiently extreme to imply a systematic bias.

1.6.2 Missing Components

What are the reasons for certain components to be recorded as missing? In a consulting setting, it is important to resolve this question at the outset because the answer could render the preceding analysis entirely inappropriate. We consider here four possibilities—not equally plausible in a laboratory setting.

1. The skin specimens were prepared by a lab assistant with the target dimension 1.0×3.5 cm running perpendicular to the scar. Many of the initial patches had ragged or crooked edges. After trimming, a few of the prepared specimens were below 0.9 cm in width; others had experienced a surgical nick in the preparation. Sub-par specimens were excluded from the analysis.
2. Rats vary in size. On some of the smaller rats, it was not possible to obtain sufficient material to prepare five specimens of the required dimension.
3. Some specimens slipped out of the clamps without breaking; others broke at a point near the clamps without breaking at the scar. In neither case was it possible to obtain a satisfactory measurement of scar strength.
4. Some specimens did not break when the tensiometer reached maximum force, so no value was recorded.

Provided that standard laboratory protocols were followed, the first explanation is relatively benign. Similar remarks apply also to the second. In the latter case, however, it would have been better to record rat size or weight pre-surgery and post-surgery. Pre-surgery weight can be used as a covariate; post-surgery weight is not a covariate because it is measured post-baseline. Furthermore, weight loss or gain may be affected by treatment, and may be correlated with healing. If both weights were recorded, the conclusions about skin strength might not be greatly affected, but a

strong correlation with weight or weight loss might lead to a different understanding of the biology.

The fourth explanation implies that each component reported as missing is associated with a large number, at least $Y > 82.5$. Given the pattern of missing components observed in Table 1.1, this explanation is implausible. Site C with the highest mean value has zero missing components; site E with the lowest mean has the highest fraction. If the description were partly true, perhaps in the reverse direction, the implications for analysis would be very substantial.

The third explanation is unlikely but not entirely implausible. In principle, it would have been better to record the observation as a pair (y_u, s_u) in censored-data format. In other words, y_u is the value at which the breakage occurred for specimen u , $s_u = 1$ if the breakage occurred at the scar, and $s_u = 0$ otherwise. For example, the value $(20.3, 0)$ implies a skin-strength measurement strictly greater than 20.3, which is informative. If censoring is admitted, the state space must be extended to the Cartesian product $\mathbb{R} \times \{0, 1\}$ rather than \mathbb{R} or $\mathbb{R} \cup \{-\}$: see Sects. 11.1 and 11.4.9.

1.6.3 Back-Transformation

The analysis for this experiment used additive effects on the log scale. How should the conclusions be reported? The great majority of physical measurements are made on a scale that is strictly positive. In such cases, it is always better to report effects as multiplicative factors or as multiplicative percentages rather than additive increments. Examples 2, 4 and 5 are typical.

The moment generating function of a Gaussian variable $Z \sim N(\mu, \sigma^2)$

$$M(t) = E(e^{tZ}) = e^{\mu t + \sigma^2 t^2 / 2}$$

implies that the log-Gaussian variable $Y = \exp(Z)$ has median e^μ and moments

$$\begin{aligned} E(Y) &= M(1) = e^{\mu + \sigma^2 / 2}; \\ \text{var}(Y) &= M(2) - M^2(1) = M^2(1)(e^{\sigma^2} - 1); \\ \text{cv}^2(Y) &= \text{var}(Y)/E(Y)^2 = e^{\sigma^2} - 1, \end{aligned}$$

where $\text{cv}(Y) = \text{sd}(Y)/E(Y)$ is the coefficient of variation of a positive variable. If σ is small, $\text{cv}^2(Y) \simeq \text{var}(\log Y)$.

If moments \bar{Z}, s^2 are estimated on the log-transformed scale, the first moment on the original Y -scale is not $\exp(\bar{Z})$, but $\exp(\bar{Z} + s^2/2)$ with multiplicative variance correction. In the two-sample case with means \bar{Z}_0, \bar{Z}_1 and pooled variance s^2 , the fitted means on the original scale are $\exp(\bar{Z}_0 + s^2/2)$ and $\exp(\bar{Z}_1 + s^2/2)$, so the fitted mean ratio is $\exp(\bar{Z}_1 - \bar{Z}_0)$, a simple exponential factor with no variance

correction. Every effect in a comparative study is a contrast for which no variance correction is needed.

If the value $y = 0$ has zero density and does not occur in the sample, we can transform to the log scale, and make direct use of linear Gaussian models. Examples 2, 4 and 5 are typical. Additive effects on the log scale are transformed to multiplicative effects on the original scale. If small values cause problems for transformation, a gamma model with constant coefficient of variation may be preferred.

1.7 Summary of Statistical Concepts

The analysis of this experiment illustrates a number of important concepts both in the design of the experiment and in the analysis of the data. The design aspects, which are discussed in greater detail in Chap. 11, include the following:

- Protocol for randomization and rat handling;
- Baseline values (`site` as covariate and `rat` as relationship);
- Immediate post-baseline values (treatment);
- Subsequent outcomes (exclusion, response, censoring);
- Observational unit versus experimental unit;
- Treatment factor versus classification factor versus block factor;
- Implications of the protocol for formal and informal analyses.

Apart from specifics of computer code, concepts illustrated in the analysis include the following:

- Role of exchangeability in model formulation;
- Response transformation;
- Correlation and variance components;
- Model formulae; each term as a vector subspace and `+` as span;
- Model formulae for block factors: Exercise 1.4;
- Parameter estimation: least squares and maximum likelihood;
- Standard errors for linear contrasts;
- Likelihood analysis and likelihood-ratio statistics: Exercise 1.12.

1.8 Exercises

1.1 To compute the control and treatment averages, the R command

```
tapply(log(y), treat, mean)
```

returns the pair of averages 3.360, 3.085, which is not the pair reported in the text. The alternative log-scale computation

```
tapply(tapply(log(y), rat, mean), trt, mean)
tapply(tapply(log(y), rat, mean), trt, var)
```

returns the numbers 3.372, 3.113 for means, and 0.174, 0.200 for variances. Under what circumstances do the two mean calculations return the same pair of averages? Explain the difference between the factors `trt` and `treat`.

1.2 In the balanced case with no missing cells, the standard analysis first reduces the data to 24 rat averages $\bar{Y}_{i\cdot}$, the treatment and control averages \bar{Y}_T , \bar{Y}_C , and the overall average $\bar{Y}_{..}$. The sum of squares for treatment effects is

$$SS_T = 70\bar{Y}_T^2 + 50\bar{Y}_C^2 - 120\bar{Y}_{..}^2 = (\bar{Y}_T - \bar{Y}_C)^2 \frac{5 \times 10 \times 14}{24}.$$

The total sum of squares for rats splits into two orthogonal parts

$$5 \sum_i (\bar{Y}_{i\cdot} - \bar{Y}_{..})^2 = SS_T + SS_R,$$

which are independent on one and 22 degrees of freedom respectively. If treatment effects are null, the mean squares $SS_T / 1$ and $SS_R / 22$ have the same expected value, and the mean-square ratio

$$F = \frac{SS_T / 1}{SS_R / 22}$$

is distributed as $F_{1,22}$. Simulate a complete design with additive effects, and check that the two terms shown above agree with parts of the decomposition reported by `anova(lm(y~site+treat+rat))`.

1.3 The F -ratio reported by `anova`(...) for treatment effects is not the ratio shown above. At least one is misleading for this design. Which one? Explain your reasoning.

1.4 For the model (1.2), verify that the R commands

```
reg_fit <- regress(log(y)~site+treat, ~rat)
lme_fit <- lmer(log(y)~site+treat+(1|rat))
```

return the same parameter estimates and the same standard errors in a slightly different format.

1.5 The sub-model with zero rat variance can be fitted by omitting the relevant term from the model formula in either syntax. The conventional log likelihood ratio statistic is twice the increase in log likelihood for the larger model relative to the sub-model. Check that these functions report different numbers for the log likelihood, but they return the same log likelihood ratio. Report the value. (Recall that the log

likelihood is defined up to an arbitrary additive constant, which may depend on the response or the design matrix.)

1.6 REML, or residual maximum likelihood, is the standard method for the estimation of variance components: see Chap. 18 for details. Both `regress()` and `lmer()` allow other options, but both use REML as the default. However, `lmer()` constrains the coefficients to be positive, whereas `regress()` allows negative coefficients unless otherwise requested. If $\hat{\sigma}_1^2 > 0$, both functions should report the same values for all coefficients; otherwise if the unconstrained maximum occurs at a negative value, differences are to be expected, both in the fitted variance components and in the regression coefficients.

For regular problems in which the null model is not a boundary subset, the null distribution of the conventional log likelihood ratio statistic is distributed asymptotically as χ_1^2 . Assuming that the unconstrained version is regular with fitted coefficients approximately unbiased, what is the asymptotic distribution of the log likelihood ratio statistic for the constrained problem? Using this null distribution, report the tail *p*-value for the hypothesis of zero rat variance.

1.7 In the balanced case with no missing cells, show that the REML likelihood-ratio statistic for treatment effects is

$$\text{LLR} = (n - 1) \log(1 + F/(n - 2)),$$

where $n = 24$ is the number of rats, and F is the treatment-to-rat mean-square ratio shown in Exercise 1.2. Compute the F -value and the associated tail probability for $\text{LLR} = 2.51$ and $\text{LLR} = 3.86$. Comment briefly on the relevance of this calculation for the calibration of likelihood-ratio statistics in the present setting.

1.8 A quantitative factor x with four equally-spaced levels 0, 1, 2, 3 may be coded using either the indicator basis e_0, e_1, e_2, e_3 (such that $e_r(i) = I(x_i = r)$) or the polynomial basis x^0, x^1, x^2, x^3 (with $x^0 = 1, x^1 = x$). Show that, if every level occurs with equal frequency in the design, the polynomials $1, z = 2x - 3, (z^2 - 5)/4, (5z^3 - 41z)/12$ are orthogonal with respect to the standard inner product in \mathbb{R}^n . Show that the components of the 4×4 transformation matrix that expresses this polynomial basis in terms of the indicator basis are all integers.

1.9 Use polynomials up to degree four to re-parameterize the site effects, and repeat the fitting procedure for (1.2) using the unnormalized orthogonal polynomial basis. Check that the treatment effect estimate and its standard error are unaffected by site re-parameterization. What effect does the change of basis have on the log likelihood?

1.10 How can we be assured that the log transformation is really needed or substantially beneficial? (Chap. 19).

1.11 Extend the model (1.2) so that it contains one variance component for treated rats and another for untreated rats. Show your code for fitting the extended model, report the two fitted variance components, and the REML likelihood ratio statistic for comparison with the simpler model.

1.12 Parameter estimates reported in Sect. 1.5 were computed using the code in Exercise 1.4. Following recommendations in Sect. 18.5, the likelihood ratio statistic for treatment effects was computed using the code

```
K <- model.matrix(~site)
fit0 <- regress(log(y)~site, ~rat)
fit1 <- regress(log(y)~treat+site, ~rat, kernel=K)
llr <- 2*(fit1$llik - fit0$llik)
```

Modify this code to obtain the likelihood-ratio statistic for site effects.

1.13 It is mathematically possible that treatment could have a positive effect at some sites and a negative effect at other sites, so that the average over sites is negligible. Investigate this possibility by computing the appropriate likelihood ratio statistic.

1.14 It is possible that treatment could have an effect on variances in addition to its effect on the mean. Investigate this possibility by replacing the identity matrix with two diagonal matrices D_0 and D_1 such that $D_0 + D_1 = I_n$, and using some version of the code

```
fit0 <- regress(log(y)~treat+site, ~rat)
fit1 <- regress(log(y)~treat+site, ~rat+D1)
fit2 <- regress(log(y)~treat+site, ~rat+D0)
```

Report the two estimated variances. Use the log likelihood ratio statistic in support of your conclusion.

1.15 Let Y be an $n \times m$ array of real-valued random variables with zero mean and covariance matrix

$$\text{cov}(Y_{ir}, Y_{js}) = \sigma_0^2 \delta_{ij} \delta_{rs} + \sigma_1^2 \delta_{ij} + \sigma_2^2 \delta_{rs} + \sigma_3^2$$

for some non-negative coefficients $\sigma_0^2, \dots, \sigma_3^2$. Show that the covariance matrix is invariant with respect to the product group consisting of $n!$ permutations applied to rows and $m!$ permutations applied to columns. In other words, show that the $n \times m$ matrix whose (i, s) -component is $Y_{\sigma(i), \tau(s)}$, has the same covariance matrix as Y .

1.16 For an $n \times m$ array of real numbers, show that the four quadratic forms, $mn\bar{Y}_{..}^2$,

$$\text{Row SS : } m \sum_i (\bar{Y}_{i.} - \bar{Y}_{..})^2;$$

$$\text{Col SS : } n \sum_r (\bar{Y}_{.r} - \bar{Y}_{..})^2;$$

$$\text{Resid SS : } \sum_{ir} (Y_{ir} - \bar{Y}_{i.} - \bar{Y}_{.r} + \bar{Y}_{..})^2,$$

are invariant with respect to row and column permutations. Here, $\bar{Y}_{i.}$ is the i th row total, and $\bar{Y}_{i.}$ is the row average.

1.17 Each of these quadratic forms is non-negative definite. In each case, the expected value is a non-negative linear combination of the four variance components, in which the coefficient of σ_0^2 is the rank of the quadratic form. Find the expected value of each quadratic form as a linear combination of the four variance components.

1.18 Let $\mathbf{1}_n$ be the vector in \mathbb{R}^n whose components are all one. Show that $J_n = \mathbf{1}_n \mathbf{1}'_n / n$ is a projection matrix, i.e., that $J_n^2 = J_n$, and that it has rank one: $\text{tr}(J_n) = 1$. Show also that $I_n - J_n$ is the complementary projection of rank $n - 1$.

Each of the quadratic forms in Exercise 1.15 can be expressed in the form $Y' M_r Y$, where each M_r is a projection matrix of order $mn \times mn$. Show that each matrix is a Kronecker product

$$J_n \otimes J_m; \quad (I_n - J_n) \otimes J_m; \quad J_n \otimes (I_m - J_m); \quad \text{and} \quad (I_n - J_n) \otimes (I_m - J_m).$$

Find the rank of each matrix.

1.19 The set of linear functionals $\mathbb{R}^{nm} \rightarrow \mathbb{R}$ is called the dual vector space; it has dimension mn . Show that the column and row totals $Y \mapsto Y_{.r}$ and $Y \mapsto Y_{i.}$ are linear functionals, and that they are linearly independent. Show that the subspace spanned by $\{Y_{.1}, \dots, Y_{.m}\}$ is closed with respect to row and column permutations. What is its dimension? Show that the subspace spanned by $\bar{Y}_{..}$, and the subspace spanned by $\{\bar{Y}_{.1} - \bar{Y}_{..}, \dots, \bar{Y}_{.m} - \bar{Y}_{..}\}$ are both closed with respect to row and column permutations. What are their dimensions?

1.20 The space of quadratic forms in the $n \times m$ array Y is a vector space of dimension $mn(mn - 1)/2$. Exhibit a basis. Show that the four quadratic forms in Exercise 1.3 are invariant with respect to row and column permutations. Deduce that every invariant quadratic form is a linear combination of these four.

Chapter 2

Chain Saws



2.1 Efficiency of Chain Saws

This example, taken from Bliss (1970, p. 440–441), is a description of an experiment by Zehnder et al. (1951) which was designed to compare the performance of different brands of chain saw. The design is elegant and carefully controlled, but it is moderately complicated in structure, and it repays careful study.

The woodcutting efficiencies of three brands of saw were compared in a fractional factorial design using six cutting teams, three species of softwood (spruce, pine and larch) both with bark and without bark. The response variable is the time in minutes taken to complete a designated cutting task. The fractional factorial is embedded in a 6×6 Latin square whose columns correspond to six teams of workmen covering the range from experienced woodcutters to seasonal labourers. The Roman letters correspond to six distinct saws, where A,D are duplicates of brand 1, B,E are duplicates of brand 2, and C,F are duplicates of brand 3. Table 2.1 shows the design and the response in standard readable format. Before computation, the values must be rearranged in computer-digestible spreadsheet format.

To clarify matters for subsequent discussion, the design consists of 12 spruce logs, 12 pine logs, and 12 larch logs. All logs are presumed to be approximately equal in length and diameter, so the task demands a fixed number of cuts. Six spruce logs, six pine logs, and six larch logs were selected uniformly at random for debarking. Each row of the table is one species/bark combination. The assignment of teams to logs is done uniformly at random subject to the condition that each team is required to cut one log of each species/bark combination. The assignment of saws to logs is done uniformly at random subject to the Latin-square condition that each

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-14275-8_2.

Table 2.1 Time in minutes taken by six teams to complete a woodcutting task using one of six available saws A–F. From Bliss (1970) with one correction in row 1, col 6

| Species | Bark | Team | | | | | |
|---------|------|--------|--------|--------|--------|-------|-------|
| | | I | II | III | IV | V | VI |
| Spruce | No | 6.4 F | 10.9 E | 9.8 D | 7.5 B | 4.6 A | 4.9 C |
| Pine | No | 6.8 B | 6.2 C | 7.9 E | 6.0 A | 4.0 D | 4.2 F |
| Larch | No | 12.7 E | 13.4 A | 12.5 B | 7.3 C | 6.1 F | 7.4 D |
| Spruce | Yes | 8.8 C | 10.2 D | 12.5 A | 8.6 F | 6.1 E | 6.0 B |
| Pine | Yes | 7.4 D | 10.0 B | 8.3 F | 6.4 E | 4.3 C | 5.6 A |
| Larch | Yes | 13.1 A | 12.0 F | 12.0 C | 11.3 D | 6.1 B | 9.7 E |

team gets to use each saw exactly once, and each saw is used exactly once for each species/bark combination.

2.2 Covariate and Treatment Factors

Each observational unit is an ordered pair consisting of one log and one saw, so the units available at the outset may be arranged in a 36×6 array of (log, saw)-pairs. However, each measurement is destructive of the log, so it is necessary to choose a subset or subsample of 36 observational units, one from each row. The Latin square design also calls for six units from each column or saw.

Despite the restrictions, the observational units are log-saw pairs arranged in a 36×6 array. Recall that a covariate is an intrinsic property of the observational units, as opposed to a treatment which is, or may be, assigned to the units. By definition, each marginal component *log* and *saw* is a covariate. In addition, *brand* is a covariate or classification factor, which is a property of the saws, and *species* is also a covariate, which is a property of the logs; the design consists of two saws of each brand, and twelve logs of each species.

By contrast, *team* is a treatment that is assigned by the investigator to each of the 36 selected units only. In the description as given, *debarking* is also a treatment that is assigned to the logs. However, if the logs were initially segregated by bark status, it could plausibly be argued that no random assignment has occurred, in which case *bark* is a covariate or classification factor.

Regardless of its status, *bark* is a Boolean function $[36] \rightarrow \{0, 1\}$ on logs such that six logs of each species are debarked, and six are left intact. There are 924^3 functions of this type. If *bark* is a treatment factor assigned by randomization, it is a random variable selected according to some specified distribution from the indicated set of 924^3 functions. In most instances, the randomization distribution is uniform on functions having the desired balance.

Algebraically, *team* is a function from the 36 selected units into the set of teams; statistically, it is a random function chosen uniformly from a subset of such

functions satisfying certain Latin-square constraints. Restricted randomization can be rather complicated, so this aspect is omitted from discussion here. For further details, see Chap. 9 of Bailey (2008).

Each of the recorded factors

$$\log, \quad \text{species}, \quad \text{bark}, \quad \text{saw}, \quad \text{brand}, \quad \text{team},$$

is a function on the sample units. The number of levels is 36, 3, 2, 6, 3, 6 respectively. In general, the way in which a covariate such as *species*, *saw* or *brand* is accommodated in a statistical analysis or formal model is not the same as the way in which a treatment is accommodated. Consider a particular unit u , a (log, saw) pair, which happens to be of the type (larch, brand 3). The model associates with u a probability distribution for the response-treatment pair; the treatment components are determined by randomization and are not independent. Hence, the model associates with u a conditional distribution $P_u(\cdot \mid T)$, one distribution on \mathbb{R} for every treatment level that has positive probability of being assigned to u . It does not associate with u a [non-trivial] probability distribution for each species or for each brand because the species and the brand are both properties of u that are recorded at baseline.

2.3 Goals of Statistical Analysis

The chief purpose of the study is to compare the relative efficiencies of the three brands of saw, i.e., to compare one brand with another. There are 12 observations for each brand, and the sample averages are 8.78, 8.55 and 7.42 in minutes, or 2.10, 2.11 and 1.95 in log minutes, so brand 3 appears to be the most efficient. The main statistical challenge is to come up with a reasonable assessment of the standard error for brand contrasts. Is it better to do the analysis additively on the time scale or on the logarithmic scale? If we do the analysis on the log scale, how do we report effects on the time scale? Regardless of which scale is used, how do we calculate a standard error for brand effects?

In addition to brand effects, we can also investigate the effect of de-barking. The sample averages with and without bark are 8.80 and 7.70 minutes, or 2.13 and 1.97 on the log scale, so de-barking appears to reduce the cutting time by about 12–15%. How do we compute a standard error? Is the reduction approximately the same for each saw brand?

In addition to brand and de-barking effects, we can also investigate differences between the three species. There are 12 observations for each species, and the average cutting times for spruce, pine and larch are 8.03, 6.42 and 10.30 minutes respectively, or 2.04, 1.82 and 2.29 on the log scale. Larch, one of the few deciduous conifers, is evidently substantially harder or tougher than the other two. Regardless of whether we use the log scale or the time scale for averages, how do we calculate an honest standard error for species contrasts? Do we compute the standard error for

each species contrast in the same way that we compute the standard error for brand contrasts?

Detailed answers to all of these questions are given in subsequent sections. At this stage, we provide brief answers to some of the questions without offering a detailed rationale.

First, the response is a time in minutes as measured by a stopwatch; ordinarily, the appropriate scale for analysis of temporal measurements by linear methods is the log scale. For some, this is obvious and needs no support; others may demand a formal justification (Sect. 19.2). Research workers from an engineering background are accustomed to using logs to the base 10 without comment, but natural logs are used throughout these notes. On the log scale, the effects of bark, species, brand and team are additive, which implies that they are multiplicative on the time scale. An analysis on the log scale does not imply that the conclusions must be reported on the same scale. Thus, in reporting point estimates of effects, we say that de-barking reduces the cutting time by 12–15%, not that de-barking reduces the cutting time by 1.1 minutes. For the particular task in the experiment, both statements are equally true, but, as an isolated statement, one is more sensible than the other. A more careful statement might emphasize that the de-barking reduction applies to the mean of the distribution. Likewise for species contrasts and brand contrasts.

Second, the estimated spruce versus larch contrast is a difference of average cutting times for two disjoint subsets of 12 observations each, the variance is $\sigma^2(1/12 + 1/12)$, and the standard error has the form

$$s\sqrt{1/12 + 1/12}$$

for some suitable estimator s^2 of σ^2 . The estimated brand 3 versus brand 1 contrast is also a difference of averages of two disjoint subsets of 12 observations each, but the variance formula is entirely different and the estimated standard error is about 30% larger than that of the spruce/larch contrast. Why so? The reasons for this difference are subtle, but they are also fundamental and easily overlooked.

The difference is a consequence of the experimental design as described in the third paragraph of this section, rather than a consequence of any parametric or nonparametric model. The crux of the matter is that 36 logs are used in the design, but only six saws. It is one thing to make a statement about the relative efficiencies of two specific saws, C versus A; it is different matter to make a statement about the relative efficiencies of two brands, brand 3 versus brand 1. For a statement of the latter type, or a statement about spruce versus larch, the observed specimens must be typical for the brand or species. But the design includes 12 specimens of each species, and only two specimens of each brand.

The use of the two-sample variance formula $\sigma^2(1/12 + 1/12)$ for the spruce versus larch contrast does not imply that the set of cutting times for spruce and the set of cutting times for larch are independent. They are not independent, and they are not assumed to be so, even conditionally on the design. Nonetheless, the two-sample formula makes good use of orthogonality, additivity and balance associated with the

Latin square, so the analogous formula would not necessarily be appropriate in a less carefully designed experiment.

2.4 Formal Models

Apart from the indicator for distinct logs, which is in 1–1 correspondence with sample units, the factors available pre-baseline in this design are as follows:

$$\text{species, bark, saw.id, brand, team}$$

In addition, *row* in the Latin square is equivalent of *species:bark*, *col* is equivalent to *team*, and *letter* is equivalent to *saw.id*. As always, each term that occurs in a linear model signifies a vector subspace of \mathbb{R}^n , and the additive operator denotes the vector span, not the vector sum. Thus, the statement that *row* is equivalent to *species:bark* is intended as a statement about vector subspaces, not a statement about indicator vectors or basis vectors. In particular, *species:bark* is a space of dimension six containing the additive subspace *species+team* of dimension four.

It is instructive to examine the output from fitting the linear Gaussian model for $\log(\text{time})$, in which the mean response lies in the subspace

$$E(Y) \in \mathcal{X} = \text{species+bark+brand+team}.$$

By default, the variances are constant and the covariances are zero. By assumption, two observations on the same saw are independent, and they are identically distributed if the two logs are of the same species and bark status. This is not the standard Latin-square model because it does not contain either the full row factor or the full letter factor.

The preceding model is contrasted with one in which *saw.id* occurs as a block factor in the variance

$$E(Y) \in \mathcal{X}; \quad \text{cov}(Y_i, Y_j) = \sigma_0^2 \delta_{i,j} + \sigma_1^2 \delta_{s(i),s(j)},$$

where $s(i)$ denotes the saw. Once again, duplicates of the same brand have the same one-dimensional marginal distribution, all observations have the same variance $\sigma_0^2 + \sigma_1^2$, observations on different saws are independent, but observations on the same saw are positively correlated.

The least-squares estimates for both models can be computed from the code

```
fit0 <- regress(log(time) ~ species+bark+brand+team)
fit1 <- regress(log(time) ~ species+bark+brand+team, ~saw_id)
```

Ordinarily, the regression parameter estimates for these two models should be similar but not identical. Because of the balanced design, they are identical, but the standard errors are different, some a little smaller, others appreciably larger. Despite

the fact that the mean square for brand replicates is not significantly larger than the mean square for residuals, the argument for a zero between-replicate variance cannot be regarded as compelling. Accordingly, the second version is preferred. On the other hand, additivity for species and bark effects is plausible on the log scale. Both models assume additivity for species and bark effects, which can be tested in the usual way.

If team effects were not a primary focus, they could reasonably be regarded as independent and identically distributed, in which case, the fitted model is obtained by using `team` as a block factor rather than a treatment factor

```
regress(log(time)~species+bark+brand, ~saw_id+team)
```

Because of orthogonality, the fitted values and standard errors for species and bark contrasts are exactly the same, whether team effects are fixed constants contributing to the mean or independent and identically distributed random variables contributing to the covariances.

2.5 REML and Likelihood Ratios

All of the models described above assume that the effects of species and debarking are additive on the log scale. How do we compute a likelihood ratio statistic for testing additivity in a situation where the model contains more than one variance component? For various reasons, this is a technically complicated question and there is at least one technically incorrect answer. But there is one answer that is both mathematically natural and technically correct, which is the one given by Welham and Thompson (1997): see Chap. 18 for a detailed analysis. The answer that is recommended in the `lmer()` literature, which is to abandon REML and use ordinary maximum likelihood, may be technically defensible, but it is not the most natural for this setting.

The Welham-Thompson likelihood-ratio statistic on two degrees of freedom for testing the null hypothesis of additivity can be computed as follows:

```
K <- model.matrix(~species+bark+team+brand)
fit0 <- regress(log(time)~species+bark+team+brand, ~saw_id, kernel=K)
fit1 <- regress(log(time)~species*bark+team+brand, ~saw_id, kernel=K)
2*(fit1$llik - fit0$llik)      # 1.591
```

The kernel is a subspace of the observation space, which determines the likelihood criterion that is used for estimation purposes. For a valid likelihood-ratio statistic, it is essential that the kernel subspaces be the same for both fits. The kernel shown above is the REML default for the first fit, but it is not the default for the second. If we choose to follow the advice in the `lmer()` literature, we must adjust the argument to `kernel=0` in both `regress()` expressions, giving a likelihood-ratio statistic of 2.17 in place of 1.59. Although the difference is numerically not negligible, the asymptotic null distribution is χ^2_2 , for which the 95th percentile is 6.0, so neither statistic indicates a departure from additivity.

If `team` is removed from the mean model but included as a block factor in the variance, the two likelihood-ratio statistics are 1.59 and 1.81 respectively. In that case, the kernel subspace for the Welham-Thompson statistic is `species+bark+brand`, which is the mean-value subspace under the null model.

2.6 Summary of Conclusions

The principal effects of interest are those related to species and saw brands. Relative to spruce as the reference level, the estimated additive effects on the log scale are

$$\text{spruce: } 0.000; \quad \text{pine: } -0.213; \quad \text{larch: } 0.256.$$

Since $\exp(-0.213) = 0.808$ and $\exp(0.256) = 1.292$, cutting times for pine logs are about 20% less than spruce logs, and cutting time for larch logs exceed those for spruce by approximately 29%. These differences are large enough to be of practical or commercial interest. Qualitatively speaking, they are in accord with on-the-job impressions of anyone who has ever wielded a chain saw on softwood. Standard errors for pairwise contrasts are 0.038, so the observed differences are roughly 5.5 and 6.6 standard deviations respectively.

The estimated bark effect is 0.152 with standard error 0.031. Thus, the effect of bark is to increase average cutting times by an estimated 16%. Or, to put it the other way round, the effect of bark removal is to decrease average cutting times by 14% regardless of species. In both of these comparisons, the standard error is based on the residual mean square on a respectable 25 degrees of freedom.

Relative to brand I as the reference, the the estimated brand effects, or mean differences, are

$$\text{Brand I: } 0.000; \quad \text{Brand II: } 0.012; \quad \text{Brand III: } -0.148.$$

In percentage terms, these amount to 100%, 101% and 86% respectively. The estimated standard errors for pairwise contrasts are 0.05, or about five percentage points. But this figure is based on the between-saw mean square on only three degrees of freedom. In essence, each duplicate pair of saws furnishes one degree of freedom for between-saw contrasts. The Welham-Thompson log likelihood-ratio statistic for brand effects comes in at 8.29 on two degrees of freedom, which is near the 98.4 percentile of the nominal χ^2_2 distribution. It appears that Brand III gives faster cuts than the other two, but the paucity of degrees of freedom for saw replicates makes this comparison less clear-cut than it might otherwise be.

2.7 Exercises

2.1 Suppose that intact logs are numbered 1:36, and that `species` is the species factor. Write code in R that picks uniformly at random a subset of six logs of each species for debarking, and stores the information as a Boolean treatment factor. Explain where the number $924 = 3 \times 4 \times 7 \times 11$ comes from.

2.2 The R commands

```
anova(lm(log(y) ~ row+team+saw_id));
anova(lm(log(y) ~ species*bark+team+brand+saw_id))
```

are designed to decompose the total sum of squares additively into components associated with certain subspaces, which are mutually orthogonal for this design. Explain how to compute the row sum of squares on five degrees of freedom directly from the six row averages

1.943, 1.737, 2.243, 2.129, 1.910, 2.341.

Arrange these six numbers in a 3×2 table, and explain the computation of the sums of squares for `species`, `bark`, and `species:bark` from this table of numbers.

2.3 Use the averages for the six saws A–F

2.122, 2.060, 1.975, 2.070, 2.156, 1.920

to compute the `brand` sum of squares on two degrees of freedom, the `saw` replicate sum of squares on three degrees of freedom, and the F -ratio (ratio of mean squares). Why is this two-part decomposition structurally different from the three-part decomposition in the preceding exercise?

2.4 Use the method described by Welham and Thompson (1997) to compute the REML likelihood-ratio statistic for comparing the two linear models

$$\mathcal{X}_0 = \text{species} + \text{bark} + \text{team}, \quad \mathcal{X}_1 = \text{species} + \text{bark} + \text{team} + \text{brand}$$

in the setting where `saw.id` occurs as a variance component. You may use R code as follows:

```
fit0 <- regress(log(y) ~ species+bark+team, ~saw_id)
K <- model.matrix(~species+bark+team)
fit1 <- regress(log(y) ~ species+bark+team+brand, ~saw_id, kernel=K)
c(fit1$llik, fit0$llik, fit1$llik-fit0$llik)
```

2.5 In the simple linear model setting with $\mu \in \mathcal{X}$ and $\Sigma \propto I_n$, show that the maximum value of the log likelihood is $\text{const} - n \log \|QY\|$, where $Q = I - P$ is the orthogonal projection with kernel \mathcal{X} , and the constant is independent of \mathcal{X} .

2.6 In the simple linear model setting, the F -ratio for testing the hypothesis $\mu \in \mathcal{X}_0$ versus $\mu \in \mathcal{X}_1$ is the ratio of mean squares

$$F = \frac{\|\mathcal{Q}_0 Y\|^2 - \|\mathcal{Q}_1 Y\|^2}{\|\mathcal{Q}_1 Y\|^2} \frac{n - p_1}{p_1 - p_0},$$

where $\mathcal{X}_0 \subset \mathcal{X}_1$, and $p_r = \dim(\mathcal{X}_r)$. Using the expression in the preceding exercise, show that the log likelihood ratio statistic is a monotone increasing function of F :

$$2\Lambda = m \log \left(1 + \frac{(p_1 - p_0)F}{n - p_1} \right).$$

where $m = \dim(\mathbb{R}^n / \mathcal{X}_0) = n - p_0$ for the Welham-Thompson statistic, and $m = n$ for the ordinary likelihood-ratio statistic.

2.7 Check that the F -ratio for brand differences is in approximate agreement with the Welham-Thompson REML statistic computed in Exercise 2.4. Explain why you need $m = 6 - 1$ rather than $m = 36 - 9$ in this comparison.

2.8 Express the random-effects models from the previous section in `lmer()` syntax, and check that the parameter estimates agree with `regress()` output.

2.9 For $y \in \mathbb{R}^n$, a decomposition of the total sum of squares

$$\|y\|^2 = y'y = y'A_1y + \cdots + y'A_ky$$

is called orthogonal if each A_r is an orthogonal projection $A_r = A_r^2 = A_r'$. This implies $\sum A_r = I_n$ and $A_r A_s = 0$ for $r \neq s$. Let $\mathbf{1}_n$ be the vector in \mathbb{R}^n with unit components, and let $J_n = \mathbf{1}_n \mathbf{1}'_n/n$. Show that J_n is a projection matrix of rank one, and that $I_n - J_n$ is a projection of rank $n - 1$.

2.10 For a $m \times n$ array, i.e., for $y \in \mathbb{R}^{mn}$, show that the Kronecker-product matrices

$$J_m \otimes J_n, \quad J_m \otimes (I_n - J_n), \quad (I_m - J_m) \otimes J_n, \quad (I_m - J_m) \otimes (I_n - J_n)$$

are complementary and mutually orthogonal projections $\mathbb{R}^{mn} \rightarrow \mathbb{R}^{mn}$ with ranks one, $m - 1$, $n - 1$, and $(m - 1)(n - 1)$ respectively. Show also that this decomposition is invariant with respect to row and column permutation.

2.11 For a Latin-square design of order m , show that the last term in the preceding decomposition can be split into two parts associated with letters. Show also that the five-part decomposition is invariant with respect to permutation of rows, columns and letters.

Chapter 3

Fruit Flies



3.1 Diet and Mating Preferences

This project concerns the experimental design and the data analysis in the paper titled *Commensal bacteria play a role in mating preference of Drosophila melanogaster*, published in 2010 by Sharon et al.. The experimental design and the initial goals are straightforward in principle: do female flies have a preference for male flies that have been reared on the same diet rather than genetically indistinguishable flies that have been fed a different diet? If so, what is the cause? Some of the finer experimental details are crucial for model formulation, analysis and interpretation, but are easy to miss in a superficial reading. Partial information on the design and analysis is given below, so you are encouraged to read the paper for yourself for additional background.

Two breeding populations of genetically identical fruit flies were raised separately for roughly forty generations on one of two diets, here denoted by C (corn-molasses-yeast) and S (starch). At certain stages, flies destined for experimentation (test matings) were removed from the breeding populations and raised for one intermediate generation on the standard C diet before testing was done. Thus the testing for generation six was done on the virgin offspring, so generation six is really 6+1: see Fig. 1 in Sharon et al. (2010). Mating tests were done on selected generations from two to 37. The mating counts in Table 3.1 are implicit in the authors' Fig. 2. It is not given explicitly in the published paper or in the supplementary online materials, but was provided by the authors on request. It contains five columns of data, generation number, followed by the mating counts for the four types, CxC, CxS, SxC, SxS. Here SxC denotes matings of male flies whose parents were raised on diet S with females whose parents were raised on

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-14275-8_3.

Table 3.1 *Drosophila* mating counts by generation

| Generation | Type of cross | | | |
|------------|---------------|-----|-----|-----|
| | CxC | CxS | SxC | SxS |
| 2 | 12 | 8 | 9 | 16 |
| 6 | 10 | 5 | 9 | 10 |
| 7 | 17 | 9 | 9 | 15 |
| 9 | 8 | 7 | 7 | 9 |
| 10 | 18 | 13 | 5 | 12 |
| 11 | 12 | 5 | 7 | 14 |
| 13 | 14 | 9 | 8 | 12 |
| 15 | 18 | 9 | 7 | 15 |
| 16 | 14 | 5 | 5 | 10 |
| 17 | 31 | 22 | 12 | 27 |
| 20 | 23 | 13 | 10 | 20 |
| 21 | 13 | 7 | 5 | 14 |
| 26 | 30 | 19 | 12 | 21 |
| 31 | 9 | 7 | 3 | 10 |
| 37 | 20 | 14 | 11 | 17 |
| 111 | 18 | 11 | 7 | 16 |
| 112 | 16 | 11 | 8 | 15 |
| 113 | 22 | 13 | 8 | 13 |

diet C. Matings of types CxC and SxS are called homogamic; the other types are heterogamic. The experimental set-up for mating tests consisted of a number of mating wells, from 20 to 70, with four flies in each well, one male and one female of each dietary type. Over a one-hour period, each mating was noted, and the totals for each type were recorded. The number of mating wells was not reported for the first three generations, but the values reported for subsequent generations were 24, 39, 20, 24, 36, 23, 70, 46, 24, 45, 23, 48, 48, 48, 48. The last three rows of data are taken from a parallel experiment run under a similar protocol, so the mating probabilities are expected to be similar, but the generation numbers should be disregarded.

3.2 Initial Analyses

3.2.1 Assortative Mating

The main summary of the experimental data is given in the authors' Fig. 2, which is a barplot of the estimated sexual isolation index (SII) for each of 15 generations. It is similar in style to Fig. 3.1, which also includes the additional three generations. The sexual isolation index is defined as the difference $p_{\text{hom}} - p_{\text{het}} = 2p_{\text{hom}} - 1$ between the probability that a mating is homogamic and the probability that it is heterogamic. The reported values are the empirical relative frequencies observed in each generation. Random mating, or absence of assortative mating, implies

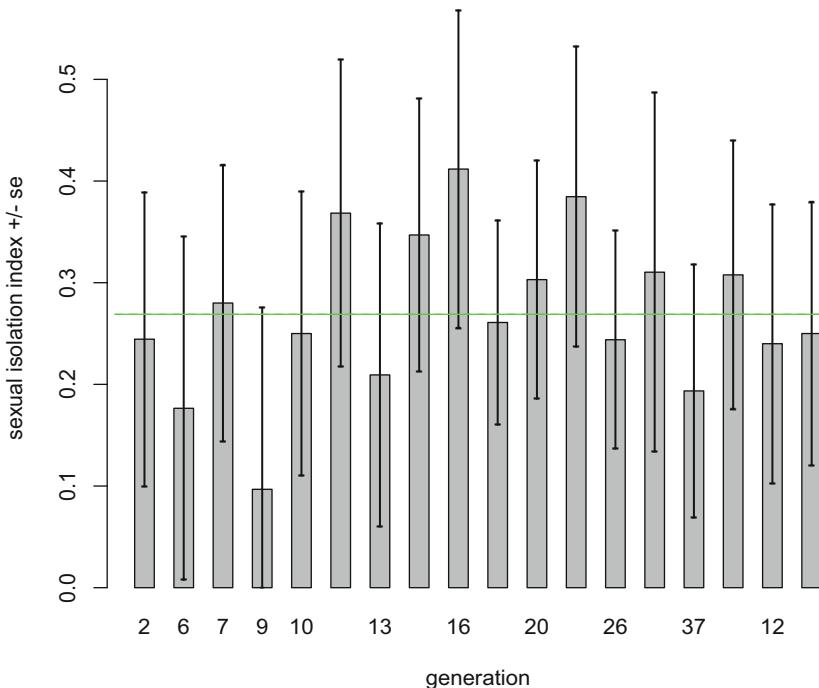


Fig. 3.1 Sexual isolation index plotted against generation number

$p_{\text{hom}} = 1/2$ or $\text{SII} = 0$. Under the assumption of binomial sampling, the estimate \hat{p}_{hom} has variance $p_{\text{hom}}(1 - p_{\text{hom}})/n$, so the observed isolation index has variance $4p_{\text{hom}}p_{\text{het}}/n$, or $(1 - \text{SII}^2)/n$ which reduces to $1/n$ in the absence of assortative mating.

The height of each bar in Fig. 3.1 is the empirical sexual isolation index for that generation, and the whisker length is one binomial standard error, i.e., $\pm\sqrt{(1 - \text{SII}^2)/n}$. The horizontal line at $\text{SII} = 0.27$ is the overall average estimated from the pooled data. Superficially, at least, all of these calculations seem quite standard statistically. However, there are both statistical and non-statistical reasons to have a closer look at the design and the analysis.

3.2.2 Initial Questions and Exercises

1. What are commensal bacteria?
2. A trained observer can distinguish male from female fruit flies. But flies raised on diet C are genetically and morphologically indistinguishable from flies raised on diet S. How did the authors determine the diet type of a mating pair?

3. Wing vibration appears to be an important part of the *Drosophila* courtship ritual. What are the implications for experimental design? Are there ways in which the design could be improved in this respect?
4. The Bernoulli model and the binomial distribution are used at various points in the authors' analyses, either explicitly or implicitly. In the context of this experiment, what assumptions are required to justify the Bernoulli model?
5. Since the sexual isolation index starts off at zero for generation zero, one might expect the value to increase slowly over generations. Is there any evidence for this, or can it reasonably be taken as constant after 1–2 generations? Construct a suitable test statistic, explain how it addresses the question, and indicate what conclusion is warranted.
6. Are the data consistent with the assumption of independence of mating events in successive generations? Compute a relevant statistic and explain what it tells you.
7. All analyses in the paper are essentially unaffected by switching sexes, which is mathematically fair and even-handed. However, after mating, a female fruit fly is no longer receptive to courtship. By contrast, a male fly may mate a second time if a receptive female is available. Discuss briefly how sexual asymmetry might affect the design or the analysis. You may assume that each well contains four flies, each courtship/mating event lasts up to 10 minutes, the observation period lasts about 40–60 minutes, whereas the female refractory period lasts about 24 hours.
8. Pearson's statistic for testing homogeneity of relative frequencies for each generation has an approximate chi-squared distribution on 3×17 degrees of freedom under standard assumptions. How accurate is this null distributional approximation when applied to Table 3.1. Compute the permutation distribution by simulation, using random matching to keep the row and column totals fixed, and compare the histogram of simulated values with the χ^2_{51} distribution.

3.3 Refractory Effects

3.3.1 More Specific Mating Counts

Table 3.2, which was subsequently provided by the authors, contains a more detailed description of the mating events in each generation. For each mating well, either zero, one or two matings may occur during the observation period. In single-mating wells, the mating is one of four types cc, cs, sc or ss, of which two are homogamic and two heterogamic; in double-mating wells, the female refractory period constrains the set of mating combinations to four, cc.ss, cs.sc, cc.cs and sc.ss, of which one is double homogamic, one is double heterogamic, and two are mixed. The order in which the matings occur is not considered here. The combination cc.cs implies that the c-male mated with both females; the s-male may or may not

Table 3.2 Number of wells having matings of each type

| Gen | Single matings | | | | Double matings | | | | Null | Total wells |
|-----|----------------|----|----|----|----------------|-------|-------|-------|------|-------------|
| | cc | cs | sc | ss | cc.ss | cs.sc | cc.cs | sc.ss | | |
| 2 | 1 | 1 | 0 | 1 | 11 | 6 | 0 | 3 | 0 | 23 |
| 6 | 1 | 0 | 2 | 1 | 7 | 5 | 0 | 2 | 0 | 18 |
| 7 | 2 | 1 | 0 | 3 | 9 | 4 | 4 | 5 | 4 | 32 |
| 9 | 1 | 2 | 3 | 3 | 5 | 3 | 2 | 1 | 4 | 24 |
| 10 | 3 | 3 | 0 | 2 | 10 | 5 | 5 | 0 | 9 | 37 |
| 11 | 2 | 1 | 2 | 1 | 8 | 1 | 2 | 4 | 2 | 23 |
| 13 | 1 | 0 | 0 | 0 | 11 | 7 | 2 | 1 | 2 | 24 |
| 15 | 2 | 1 | 2 | 2 | 11 | 3 | 5 | 2 | 8 | 36 |
| 16 | 2 | 0 | 2 | 1 | 9 | 3 | 2 | 0 | 3 | 22 |
| 17 | 0 | 8 | 3 | 7 | 18 | 6 | 9 | 1 | 17 | 69 |
| 20 | 8 | 4 | 4 | 4 | 8 | 3 | 5 | 0 | 9 | 45 |
| 21 | 1 | 2 | 0 | 2 | 11 | 4 | 1 | 1 | 2 | 24 |
| 26 | 2 | 1 | 2 | 1 | 17 | 7 | 11 | 3 | 3 | 47 |
| 31 | 3 | 2 | 0 | 6 | 4 | 3 | 2 | 0 | 3 | 23 |
| 37 | 5 | 1 | 3 | 7 | 9 | 7 | 6 | 1 | 9 | 48 |
| 111 | 2 | 2 | 1 | 1 | 12 | 4 | 4 | 2 | 14 | 42 |
| 112 | 5 | 2 | 2 | 7 | 7 | 5 | 4 | 1 | 15 | 48 |
| 113 | 9 | 3 | 1 | 3 | 10 | 7 | 3 | 0 | 10 | 46 |

have courted, but did not mate with either female. The other combinations do not occur because of the refractory constraint: a female that has already mated does not mate a second time within about 24 hours. Each courtship ritual and mating takes approximately 10–12 minutes, so the observation period of 40–60 minutes is sufficient for one male to copulate with both females if they are receptive.

It is important to observe the fundamental difference between the two versions of the *Drosophila* data. The objects that are counted in Table 3.1 are matings, which are of four types; the objects that are counted in Table 3.2 are wells of various types, one type for each column. In the first case, each observational unit is a mating, and the response is the mating type; in the second case, each observational unit is a well, and the response is one of nine types.

From a statistical standpoint, it is natural to regard the activity in one well as a multinomial event with nine activity classes that are both disjoint and also exhaustive in the biological sense if not in the mathematical sense. It is also natural to regard flies as exchangeable modulo their sex and diet type, so that events in distinct wells may be taken as independent with identical distributions for all wells in the same generation. Those assumptions justify the reduction of the data to the counts in Table 3.2 as the sufficient statistic. Provided that the activity in one well is independent of that in other wells, each row is an independent multinomial random variable.

Biologically speaking, the multinomial parameters need not be constant from one generation to the next. Apart from the possibility of a monotone increasing sexual isolation index, there are more mundane reasons for distributional heterogeneity that may be related to experimental procedure. One possibility is that the inclination to mate may depend on temperature and other environmental factors that vary with the season, and hence the rates are not constant from one generation to the next. Another very real possibility is that the period set aside for observation is not quite constant from generation to generation, in which case the fraction of null-mating wells is expected to be greater for shorter observation periods. Likewise, for purely bio-mechanical reasons, the fraction of double-mating wells is likely to be low for shorter observation periods.

In principle, each column in Table 3.1 is derivable as a specific linear combination of the columns in Table 3.2. Each linear combination has three unit coefficients and six zeros. For example, the CxS column is the sum of columns cs, cs.sc and cc.cs, while the SxC column is the sum of sc, cs.sc and sc.ss. Both combinations include cs.sc. This linear projection structure implies that the counts in one row of Table 3.1 are correlated in a non-multinomial way, which invalidates the distributional assumptions on which the paper is based. In practice, there are a few numerical discrepancies between the two tables, which is not uncommon in laboratory work. Unless otherwise specified, all subsequent analyses in this chapter use the version in Table 3.2.

3.3.2 Follow-Up Analyses

Given that the main focus is on the excess of homogamic over heterogamic matings, how should we analyze the new version of the data for evidence bearing on the issue of commensally-related assortative mating? Assuming for the moment that there is sufficient homogeneity across generations, it is natural first to examine the aggregate counts or column totals, which are as follows:

| cc | cs | sc | ss | cc.ss | cs.sc | cc.cs | sc.ss | null |
|----|----|----|----|-------|-------|-------|-------|------|
| 50 | 34 | 27 | 52 | 177 | 83 | 67 | 27 | 114 |

Among all events in 163 single-mating wells, 102 are homogamic and 61 heterogamic, so the homogamic sample fraction is 0.626 and the standard error is 0.038. If mating events occurred non-preferentially according to the Bernoulli-1/2 model, we should expect about 81.5 ± 6.4 homogamic and the same number of heterogamic matings, so the observed value is a little more than 3.2 standard deviations away from the non-preferential null. Equivalently, the SII index is 0.251 with standard error $1/\sqrt{163}$ computed under the Bernoulli-1/2 model, and the ratio is $0.251/\sqrt{163} = 3.2$. Three or more standard deviations is usually regarded as moderately strong evidence against the null, so even if we restrict attention to single-mating wells, the evidence for assortative mating is clearly established.

In the double-mating wells, 448 matings out of 708 are homogamic, so the homogamic fraction is 0.633. To obtain a standard error for the sample fraction, the four totals (Y_1, Y_2, Y_3, Y_4) are regarded as multinomial with index $Y = 354$, parameter vector π , and covariance matrix $(\text{diag}(\pi) - \pi\pi')Y$. The number of homogamic matings is the linear combination $2Y_1 + Y_3 + Y_4$, the number of heterogamic matings is $2Y_2 + Y_3 + Y_4$, and the total number of matings is $2Y = 708$. The variance of the linear combination is a quadratic form in the multinomial covariances, whose estimate is 235.04, so the standard error of the homogamic fraction in the sample is $\sqrt{235.04}/708 = 0.022$. The observed value is six standard errors away from the null, so once again the evidence strongly supports assortative mating.

For a slightly cleaner version of the preceding argument, the difference between the number of homogamic and heterogamic matings is $2Y_1 - 2Y_2$, which does not involve the mixed-well counts Y_3 or Y_4 . Arguably, the mixed-event wells are uninformative for testing. The null hypothesis of no assortative mating implies that Y_1 has the same distribution as Y_2 , so it is possible in this setting to construct an exact binomial test by conditioning on the total $Y_1 + Y_2$. However, the *estimate* of the isolation index is not independent of the mixed double-mating well counts.

The probability estimates obtained from single- and double-mating wells, 0.626 ± 0.038 and 0.633 ± 0.022 are in unusually good agreement with one another. The standard error of the difference is the square root of $0.038^2 + 0.022^2$, which is 0.044, whereas the observed difference is only 0.007. A similar analysis on the SII scale gives an equivalent answer. If necessary, the estimates may be pooled or combined in the standard manner with weights inversely proportional to variances.

3.3.3 Lexis Dispersion

The Lexis dispersion statistic, which is the ratio of Pearson's chi-squared statistic to its degrees of freedom, is a natural gauge of variation in a contingency table in which the reference value of unity is the expected value under homogeneous multinomial sampling. For the four single-mating columns in Table 3.2 the value is 48.4/51, and for the four double-mating columns the value is 50.95/51. As we had hoped, both are satisfactorily close to unity, so there is no evidence of inter-generational inhomogeneity in mating behaviour for either the single-mating wells or the double-mating wells. Inter-generational homogeneity of *Drosophila* behaviour is reassuring.

For the 18×2 matrix whose columns are the tallies for single- and double-mating wells in each generation, the Lexis dispersion statistic is $48.2/17 = 2.83$. We conclude that there is substantial heterogeneity in the fraction of single- versus double-mating wells in successive generations. This type of inter-generational inhomogeneity is not a major concern. It does not invalidate the analyses proposed in the preceding section or those in subsequent sections. As mentioned earlier, it

could easily be attributed to environmental variation or to incidental variation in experimental counting procedure.

3.3.4 Is Under-Dispersion Possible?

The dispersion index for Table 3.1 is $19.14/51 = 0.37$, which shows clearly that the counts in that table are substantially under-dispersed. Over-dispersion is common in experimental and observational work, while under-dispersion is rare, so statisticians are naturally on the lookout for phenomena that give rise to under-dispersion. The main explanation for under-dispersion in this instance appears to lie in the experimental design with four flies per mating well and its interaction with the female refractory effect.

This section offers an analysis of whether the under-dispersion that is observed in Table 3.1 should be expected on the basis of its derivation from Table 3.2. The analysis is done under the following ‘multinomial assumption’, which seems mathematically natural for this setting.

1. Given the vector m_1 of single-mating well counts in each generation, the 18×4 table T_1 consisting of the first four columns of Table 3.2 has independent multinomial rows, and the probability vector π_1 is constant across generations.
2. Given the vector m_2 of double-mating well counts in each generation, the 18×4 table T_2 consisting of the columns 5–8 of Table 3.2 has independent multinomial rows, and the probability vector π_2 is constant across generations.
3. The tables T_1 and T_2 are conditionally independent given m_1, m_2 .

Although homogeneity across generations is an important component, we refer to these collectively as ‘the multinomial assumption’.

Let L be the matrix that converts double-mating well counts into mating counts of four types:

$$L = \begin{matrix} & \text{cc} & \text{cs} & \text{sc} & \text{ss} \\ \text{cc.ss} & 1 & 0 & 0 & 1 \\ \text{cs.sc} & 0 & 1 & 1 & 0 \\ \text{cc.cs} & 1 & 1 & 0 & 0 \\ \text{sc.ss} & 0 & 0 & 1 & 1 \end{matrix}$$

so that $T = T_1 + T_2 L$ counts total matings of each type in each generation. Discrepancies between Table 3.1 and T have already been noted, but are not the focus of this analysis. By assumption, the rows of this table are independent. The expected mating count for generation i is the linear combination $m_{1,i}\pi_1 + m_{2,i}L'\pi_2$ of multinomial vectors. However, even if $2\pi_1 = L'\pi_2$, the distribution of T is not multinomial, so Pearson’s statistic does not have its standard χ^2 -reference distribution. The question to be addressed is whether it is possible under the

multinomial assumption for the 18×4 table T to be under-dispersed relative to the multinomial, and to what extent.

The question can be addressed in a variety of ways, either analytically or by simulation. For a partial analytical solution, the mean vector and covariance matrix of the i th row of T are

$$\begin{aligned}\mu_i &= E(T_i) = m_{1,i}\pi_1 + m_{2,i}L'\pi_2 = (m_{1,i} + 2m_{2,i})\pi, \\ \Sigma_i &= \text{cov}(T_i) = m_{1,i}V(\pi_1) + m_{2,i}L'V(\pi_2)L,\end{aligned}$$

where $V(\pi) = \text{diag}(\pi) - \pi\pi'$ is the 4×4 multinomial covariance matrix. Pearson's statistic is the quadratic form

$$X^2 = \sum_{i,j} \frac{(Y_{ij} - \hat{\mu}_{ij})^2}{\hat{\mu}_{ij}}$$

whose expected value is approximately

$$\sum_{i=1}^{18} \text{tr}(\hat{\Sigma}_i \text{diag}(\hat{\mu}_i^{-1})).$$

Using the natural moment estimates for the vectors π_1, π_2 and π , the estimated mean of X^2 is 39.39. For a more accurate approximation, we should multiply by 17/18 to account for parameter estimation. This gives $E(X^2) \simeq 37.2$, so the appropriate null reference level for the Lexis dispersion index is $39.39 \times 17/(18 \times 51) = 0.73$. The conclusion from this analysis is that under-dispersion is not only possible but also expected in this particular situation.

A more accurate estimate of the null distribution of Pearson's statistic can be obtained by simulation along the lines of Sect. 3.4. First generate two conditionally independent hypergeometric tables having the same marginal totals as T_1 and T_2 , combine them into a single table as in $T_1 + T_2L$, and then compute the Pearson statistic. The conclusion from 5000 simulations is that the null mean is 37.2 and the variance is approximately 76.8. Relative to this distribution, the observed value $X^2(T) = 24.1$ falls just below the 5% point; the observed value for Table 3.1 is 19.1, which is below the 1% point. The conclusion is that under-dispersion is expected, though not quite to the extent observed in T or in Table 3.1.

3.3.5 Independence

One fundamental assumption in all of the foregoing analyses is that activities occurring in distinct wells must be independent. Ordinarily, the assumption of independence seems so obvious experimentally that it cannot be called into ques-

tion. After all, distinct wells contain distinct flies whose activities cannot possibly be coordinated. But science rightly demands that assumptions be checked where possible, and the design of this experiment with the data in Table 3.2 provide a rare opportunity to check the independence assumption—at least in part.

Flies in single-mating wells are necessarily distinct from flies in double-mating wells, so in the absence of inter-well communication, we must expect all activity in single-mating wells to be independent of all activity in double-mating wells. This is part of the third component of the multinomial assumption in the preceding section. In particular, we must expect the homogamic fraction in single-mating wells to be statistically independent of the homogamic fraction in double-mating wells. Any failure of independence in this form must have far-reaching consequences for *Drosophila* experimentation. To paraphrase Lord Denning's notorious judgement from 1980, ...*the possibility of coordinated mating activities in distinct wells is such an appalling vista that every sensible Drosophila experimentalist would say 'It cannot be right...'*.

Each generation furnishes a pair of homogamic fractions, one for single-mating wells and one for double-mating wells. Figure 3.2 is a scatterplot of the 18 pairs, one pair for each generation. Contrary to expectation, it shows not only that homogamic fractions in the same generation are correlated, but also that the correlation is negative ($r = -0.64$). The null distribution of sample correlations is symmetric about zero with a standard deviation of about 0.23, so the observed correlation is far removed from the bulk of the null distribution. Hypergeometric simulation by random matching points to a left tail p -value of approximately one in 850, which is equivalent to three standard deviations from expectation on the standard normal scale.

To the extent that the entire edifice of experimental work on *Drosophila* rests on the assumption of independent behaviours for unrelated observational units, lack of independence is an extraordinary claim, and negative dependence even more so. But, as Lord Denning rightly points out, extraordinary claims require extraordinarily strong evidence. By prevailing standards in the literature on animal behaviour, a three- σ deviation is regarded as good evidence supporting a non-zero effect, so the null hypothesis would be firmly rejected. In this instance, however, a three- σ deviation may be very surprising or strongly suggestive, but it cannot suffice to overturn a fundamental tenet or a long-assumed law of behaviour without independent confirmation from other laboratories.

The correlations indicated by Fig. 3.2 are synchronous in time; there is no suggestion that the homogamic fractions in one generation are correlated with homogamic fractions in previous or subsequent generations. Inter-well communication is one potential explanation for synchronous correlations. *Drosophila* courtship rituals are not silent, so sound leakage may be possible. Pheromonal leakage may be more likely. However, in order to achieve the observed *negative* correlation, the communication must be anti-symmetric or conspiratorial, so it is unlikely that pheromonal or sound leakage alone could suffice. On balance, therefore, it seems safe to rule out inter-well communication as a likely explanation. However, it

is difficult to come up with a viable physical or biological explanation, and no alternative has yet been suggested.

The phenomenon analyzed in Sect. 3.3.4 accounts for an under-dispersion factor of 0.73. Negative correlation for pairs of wells has an additional and potentially greater effect on the variance of linear combinations ($1 - r = 0.36$), and this appears to be the main reason for the extreme under-dispersion seen in Table 3.1. For example, the marginal table of homogamic versus heterogamic counts derived from Table 3.1 has a dispersion factor of $4.29/17 = 0.25$, which matches nicely with the product $0.73(1 - r) = 0.26$.

The sample correlation reported above is weighted harmonically with weights w_t satisfying $w_t^{-1} = m_{1,t}^{-1} + m_{2,t}^{-1}$ for generation t , and the points in Fig. 3.2 are enlarged in areal proportion. Either $m_{1,t} = 0$ or $m_{2,t} = 0$ implies $w_t = 0$, which is the main reason for choosing harmonic weights. Some weighting is needed to accommodate the different sample sizes, and this harmonic weighting may not be optimal, but the correlation value is not especially sensitive to the choice of weights.

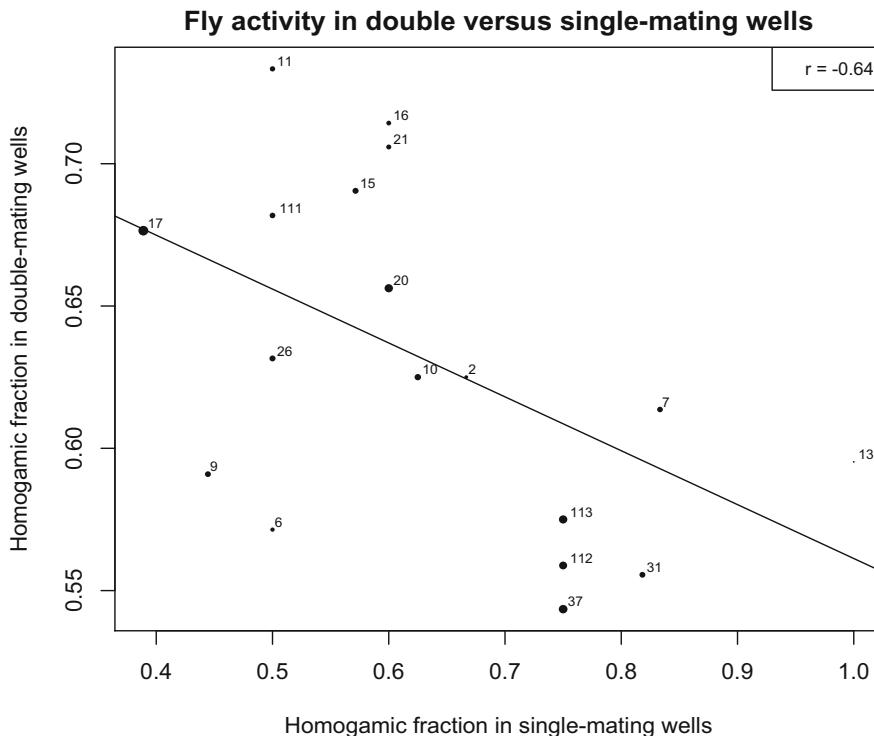


Fig. 3.2 Scatterplot of homogamic mating fractions for 18 generations

3.3.6 Acknowledgement

Table 3.2 and much of the analysis in this section is based on an unpublished report by Daniel Yekutieli, which was provided by the author.

3.4 Technical Points

3.4.1 Hypergeometric Simulation by Random Matching

A two-way contingency table is a rectangular array Y whose components Y_{ij} are non-negative integers. Usually, Y_{ij} is the number of observational units for which one attribute or factor is equal to i and a second attribute is equal to j . Thus, the table is indexed by attribute values. Let $m_i = Y_{i\cdot}$ be the i th row total, and let $s_j = Y_{\cdot j}$ be the j th column total so that $m_{\cdot} = s_{\cdot} = Y_{\cdot\cdot}$ is the overall total. A random table is said to have the hypergeometric distribution if the joint distribution is

$$\text{pr}(Y = y) = \frac{\prod_i y_{i\cdot}! \prod_j y_{\cdot j}!}{y_{\cdot\cdot}! \prod_{ij} y_{ij}!}.$$

The row and column totals are fixed positive integers, so the probability mass function is inversely proportional to $\prod_{ij} y_{ij}!$ on the space of non-negative integer-valued arrays having the given row and column totals.

If Y has the hypergeometric distribution, so also does the transposed array. If Y is a random matrix whose rows Y_i are independent multinomial vectors, $Y_i \sim M(m_i, \pi)$, which are homogeneous in the sense that they have the same multinomial probability vector, then the conditional distribution given the column totals is hypergeometric.

One way to simulate a hypergeometric random table having given row and column totals is by random matching of the components of two n -component vectors. Suppose that `row` has n components of which m_r are equal to r , and `col` has n components of which s_j are equal to j , with $\sum m_r = \sum s_j = n$. Random matching permutes the components of `row` uniformly at random, does the same independently for `col`, and then tabulates or counts the ordered pairs (r, j) thus generated. Distributionally speaking, it is necessary only to permute one of the vectors as follows:

```
RHG <- function(rowsum, colsum) {
  # rowsum and colsum are integer vectors having the same sum
  row <- rep(1:length(rowsum), rowsum)
  col <- rep(1:length(colsum), colsum)[order(runif(sum(colsum)))]
  table(row, col)
}
```

To simulate the null distribution of Pearson's statistic or any other statistic such as the deviance, we compute the statistic for each table thus generated, and report the histogram. The analysis near the end of Sect. 3.3.4 calls for two independent hypergeometric tables T_1, T_2 , followed by Pearson's statistic computed on the linear combination $T_1 + T_2L$. The analysis in Sect. 3.3.5 also calls for the same pair of independent hypergeometric tables followed by a symmetric correlation statistic $R(\cdot, \cdot)$ computed as a function of the pair (T_1, T_2L) .

3.4.2 Pearson's Statistic

Pearson's statistic is a quadratic form in residuals, $X^2 = (Y - \mu)' \Sigma^{-1} (Y - \mu)$, which is a scalar measure of variability in the response relative to a given reference distribution whose mean vector and covariance matrix are μ and Σ . In most cases, the mean vector is estimated from the data, and Σ is a function of μ .

For counted data in the form of a contingency table, the reference distribution is usually Poisson or binomial with independent components, or multinomial with independent rows. In all of these cases, the algebraic form is the same,

$$X^2 = \sum_i \frac{(Y_i - \hat{\mu}_i)^2}{\hat{\mu}_i},$$

where $\hat{\mu}_i$ is the fitted mean value. In the binomial case, the sum extends over both response classes—failure and success—so that the net contribution from a binomial pair $(Y_{i,0}, Y_{i,1}) \sim B(m_i, \pi_i)$ for which $\hat{\mu}_{i,0} = m_i \hat{\pi}_i$ and $\hat{\mu}_{i,1} = m_i(1 - \hat{\pi}_i)$ is

$$\frac{(Y_{i,0} - \hat{\mu}_{i,0})^2}{\hat{\mu}_{i,0}} + \frac{(Y_{i,1} - \hat{\mu}_{i,1})^2}{\hat{\mu}_{i,1}} = \frac{(Y_{i,0} - m_i \hat{\pi}_i)^2}{m_i \hat{\pi}_i (1 - \hat{\pi}_i)}.$$

The Poisson form of Pearson's statistic differs from the binomial form only in the variance function, $\Sigma = \text{diag}(\mu)$ for the Poisson covariance; and $\Sigma = \text{diag}(m_i \pi(1 - \pi))$ for the binomial. But the Poisson form covers both provided that we sum over both successes and failures.

The sampling distribution of the statistic depends on the distribution of Y and on the degrees of freedom used up in the estimation of μ . Exact moments are available in a few special cases, all of them null in a suitable sense. For a single multinomial $Y \sim M(m, \pi)$ with k classes and given probability vector, we have

$$E(X^2) = k - 1,$$

$$\text{var}(X^2) = 2(k - 1) \frac{m - 1}{m} + \frac{1}{m} \sum \pi_j^{-1} - k^2/m.$$

The third cumulant is given in McCullagh and Nelder (1989, p. 169). The asymptotic distribution for large m is χ_{k-1}^2 .

For an $r \times c$ contingency table that is distributed according to the hypergeometric distribution with strictly positive row and column totals, the Haldane-Dawson formulae give the exact mean and variance. The mean does not depend on the row or column totals, but only on the overall total:

$$E(X^2) = (r-1)(c-1)m/(m-1).$$

The variance of X^2 depends on the sum of reciprocals of the row and column totals: see McCullagh and Nelder (1989, p. 244).

Despite warnings given freely by over-cautious computer software, the nominal $\chi_{(r-1)(c-1)}^2$ approximation is quite accurate even for a large sparse table such as the 12×12 birth-death table where the average cell count is only 2.4. Even for Bortkewitsch's horsekick fatality data for 14 Prussian army corps over 20 years (Andrews and Herzberg, 1985, 17–18), where the mean is only 0.7 fatalities per corps per year, the χ_{247}^2 approximation is reasonably good in the upper tail. The left tail is not so good. In that instance, the Haldane-Dawson values for the mean and variance are 248.3 and 419.8, so the variance-to-mean ratio is only 1.69 as opposed to 2.0 for the χ^2 approximation. Exercise 3.7 shows that the moment-matching approximation $0.85\chi_{294}^2$ is quite accurate in both tails.

Pearson's statistic has a role to play in the analysis of counted data, mainly as a metric for relative dispersion. Over the past 70 years, various authors have pointed out its inferential limitations, and have sought to modify and strengthen it in various ways (Yates, 1948; Cochran, 1954; and Armitage, 1955). Its deficiencies for significance testing are entirely unrelated to the adequacy of any distributional approximation. The discussion in this section focuses on its use as a dispersion index; it is not intended as an endorsement of its widespread use in applications as a test for independence or lack of association.

3.5 Further Drosophila Project

In studies of animal behaviour, the so-called *Coolidge effect* refers to a preference by males for novel females over familiar females. The phenomenon was studied in rats by Fowler and Whalen (1961) and by Wilson et al. (1963). In the latter paper, the coinage is attributed on page 641 to an unpublished paper by R.E. Whalen. Since Beach and Whalen had previously collaborated on related studies, and no explanation was offered, this coinage has all of the hallmarks of an in-joke.

The origin of the term as an ‘elaborate hoax’ was explained subsequently by Beach in a letter to Kimble. The following extract is taken from the undated letter on page 342–3 of Kimble et al. (1980, fifth edition).

The neologism refer[s] to an old joke about Calvin Coolidge when he was President. You will remember that he was an exceedingly laconic individual and was nicknamed Silent Cal... The President and Mrs. Coolidge were being shown [separately] around an experimental government farm. When [Mrs. Coolidge] came to the chicken yard she noticed that a rooster was mating very frequently. She asked the attendant how often that happened and was told, 'Dozens of times each day.' Mrs. Coolidge said, 'Tell that to the President when he comes by.' Upon being told, the President asked, 'Same hen every time?' The reply was, 'Oh, no, Mr. President, a different hen every time.' President: 'Tell that to Mrs. Coolidge.'

The Coolidge phenomenon is of great interest to research workers in evolutionary genetics and animal behaviour. From time to time, new findings are reviewed glowingly in the public press, often out of prurience and with no reference or yardstick to assess the merits of the scientific contribution.

The paper *Sex-specific responses to sexual familiarity, and the role of olfaction in Drosophila* by Tan, Løvlie, Greenway, Goodwin, Pizzari and Wigby, which was published in *Proceedings of the Royal Society, Series B* (2013), addresses a number of facets of the Coolidge effect in fruit flies. The entire focus is on the courtship behaviour of males, and specifically whether males preferentially court novel females over familiar females. Thus, a directly familiar female is the previous mate, and a phenotypically familiar female is a sister of the previous mate. According to the abstract

... we show that male and female *Drosophila melanogaster* respond to the direct and phenotypic sexual familiarity of potential mates in fundamentally different ways. We exposed a single focal male or female to two potential partners. In the first experiment, one potential partner was novel (not previously encountered) and one was directly familiar (their previous mate); in the second experiment, one potential partner was novel (unrelated, and from a different environment from the previous mate) and one was phenotypically familiar (from the same family and rearing environment as the previous mate). We found that males preferentially courted novel females over directly or phenotypically familiar females. By contrast, females displayed a weak preference for directly and phenotypically familiar males over novel males.

As it turns out, the statistical analysis in the original paper is seriously deficient in a number of ways. In a 2014 correction note, the authors remark

... the statistical models we used for analysing male courtship behaviour did not take into account temporal correlations in courtship events within males. Consequently, the variance in courtship events was higher than predicted by the model, and the excess dispersion could potentially result in errors in conclusions. This highlights the general potential for high-frequency sampling of behaviours to give rise to high temporal correlations of event counts within a dataset, and the importance of correcting dispersion factors when analysing this type of data.

In other words, the courtship activity for one male was recorded on multiple occasions over a short period, and the sequence of records was analyzed as if the activities on successive occasions were independent events measured on unrelated flies. To say that high-frequency sampling has the 'potential' to give rise to high temporal correlations is a gross understatement.

Despite the authors' remarks about the potential for and the effect of temporal correlations, it is worth remarking that the revised analysis accounts for over-dispersion, but it makes no attempt to account for serial correlation. If the quoted remark suggests that an excess dispersion factor is adequate to accommodate serial correlation, it is certainly misleading. Indeed, the serial order of events was not reported, so the data provided make it impossible to accommodate such effects in anything like a principled way. A simple over-dispersion factor is better than none, but it does not adequately address statistical issues arising from high-frequency sampling of behaviours.

There is nothing intrinsically wrong with high-frequency sampling provided that the statistical analysis accommodates the inevitable serial correlation in a satisfactory way. If the activity of the focal male were recorded at 24 frames per second, we may mark each frame in which courtship is directed at the novel female by the label 'N' and those in which it is directed at the familiar female by the label 'F'. While it may be reasonable to treat a single marked frame as a Bernoulli random variable, it is obviously unreasonable to treat the *sequence* of frames as a Bernoulli sequence with independent components. For the same reason, it is unreasonable to treat the number of 'N' frames as a Poisson or binomial variable. This statement may be obvious at a sampling rate of 24 frames per second, but it applies equally at a sampling rate of one per minute or one per hour. Doubling the frame rate doubles the computational burden, but has a negligible effect on information pertaining to sexual preferences.

One possibility for analysis is to reduce the frame sequence to the fraction of time spent in each activity, and to regard these temporal fractions as a compositional response in the sense of Aitchison (1986).

The data for three of these experiments are available in the files

```
eyedat <- read.table("CoolEyeColorArchive.dat", header=TRUE)
paintdat <- read.table("CoolPaintArchive.dat", header=TRUE)
decapdat <- read.table("PhenoMaleDecapArchive.dat", header=TRUE)
```

Additional information is available in the file Coolidge.R. Other data files are available online.

3.6 Exercises

3.1 Use the normal approximation to the binomial to compute the probability that the horizontal line in Fig. 3.1 intersects all 18 whiskers at ± 1 standard deviations. Devise a better approximation by simulation that takes account of the fact that the SII index has been computed from the same data.

3.2 Is the total number of matings in Table 3.1 related to the number of mating wells? Is the pattern of variation different for the experiments reported in the last three rows? Explain how you address such questions.

3.3 For the experiment giving rise to the data in Table 3.2, an algebraically natural assumption is that the allowable double matings occur as a Poisson process at a rate proportional to the product of the single-mating rates. It is also natural—physically if not mathematically—to allow separate factors for single and double wells, and a reduced rate for wells in which one male does double duty. Formulate this statement as a Poisson log-linear model or four-class multinomial model, and check whether the data are in compliance with the product assumption. For this exercise, the multinomial assumption in section 3.3.4 may be used. (The computation for this question may involve the entire table, but parameter estimates and other conclusions must be a function of the column totals only. Why so?)

3.4 Hypergeometric simulation in Sect. 3.3.5 implies a symmetric null distribution with standard deviation 0.23 for the weighted sample correlation of homogamic pairs. One suggested alternative to random matching is to generate the null distribution by randomly permuting the vector $(\pi_{2,i}, m_{2,i})$ of double-mating homogamic fractions, keeping the sample-size attached to each fraction. Check that random permutation of generations also gives a symmetric null distribution with standard deviation at least 10% larger than the hypergeometric null. Which of these null distributions is the relevant one to use as a reference in this setting? Explain your reasoning.

3.5 The file `...birth-death.R` contains the data compiled by Phillips and Feldman (1973) on the month of birth and the month of death of 348 ‘famous Americans’. Investigate whether the month of death is or is not independent of the month of birth. The data are given as a 12×12 table of event counts. (This is not a generic contingency table because the row labels and the column labels are not only the same, but also cyclically ordered. Both aspects of the structure are relevant to the question posed, and both should be exploited in your analysis.)

3.6 The advice sometimes given for the validity of the χ^2 approximation to the null distribution of Pearson’s statistic is that the minimum expected value should exceed a suitable threshold, usually in the range 3–5. However, the mean count for the birth-death table is 2.42, so the expected count in every cell falls below the threshold. Compute the null distribution of Pearson’s statistic by hypergeometric simulation. Plot the density histogram of simulated values, and superimpose on it the χ^2_{121} density function. (This is intended as a computational exercise only. It is *not* a suggestion for data analysis aimed to address the question posed by Phillips and Feldman.)

3.7 Check the calculations reported in the penultimate paragraph of Sect. 3.4.2 for Bortkewitsch’s horsekick data. Compute the row and column totals, and simulate the null distribution of X^2 by random matching. Superimpose the χ^2_{247} density on a histogram of the simulated values. Find two positive numbers a, b such that the first two moments of $a\chi_b^2$ coincide with the Haldane-Dawson moments. Superimpose this scaled chi-squared density on your histogram. (The intent of this exercise is

solely to provide insight into distributional approximation. It should *not* be read as an endorsement for data analysis.)

3.8 What was the matter that Lord Denning refused to accept in his 1980 appeals-court judgement when he referred so melodramatically to the ‘appalling vista that every sensible person would reject’? Why was this phenomenon so abhorrent to him?

3.9 Explain where the factor $1 - r$ comes from in the penultimate paragraph of Sect. 3.3.5.

Chapter 4

Growth Curves



4.1 Plant Growth: Data Description

The file `PlantGrowth.dat` contains the heights in mm. of 70 *Arabidopsis* plants measured every four days from day 29 to day 69 following the planting of seeds. The ultimate heights range from 19 mm to 40 mm, and most heights are recorded to the nearest millimetre.

The cumulative number of plants brearded was zero up to and including day 29, eight by day 33, 40 by day 37, and all 70 by day 41. Thus, sprouted plants were first recorded on day 33, and all plants had appeared by day 41. Or, to put it more accurately perhaps, no additional plants emerged after day 41. By day 65 or earlier, the growth was complete; for each plant, the height recorded on day 69 was the same as the height on day 65.

Plant age is most naturally measured from its first emergence at brearding rather than the date on which seed was planted. In this experiment, all seeds were planted on the same date, but the date of brearding varies from plant to plant. The brearding date is deemed to be the last date on which the height was recorded as zero rather than the first date on which the height was positive. In other words, eight plants were deemed to be born on day 29, 32 on day 33, and so on.

The typical growth trajectory for *Arabidopsis* begins at zero on day 0, reaching a plateau whose height varies from plant to plant. Regardless of the ultimate height, the semi-maximum height is attained in about 13 days, which is fairly constant from plant to plant. By inspection of the graphs in Fig. 4.1, it appears that the standard *Arabidopsis* growth trajectory is roughly $h(t) \propto t^2 / (\tau^2 + t^2)$. This is sometimes referred to as ‘inverse quadratic’ because the inverse height is a linear function of

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-14275-8_4.

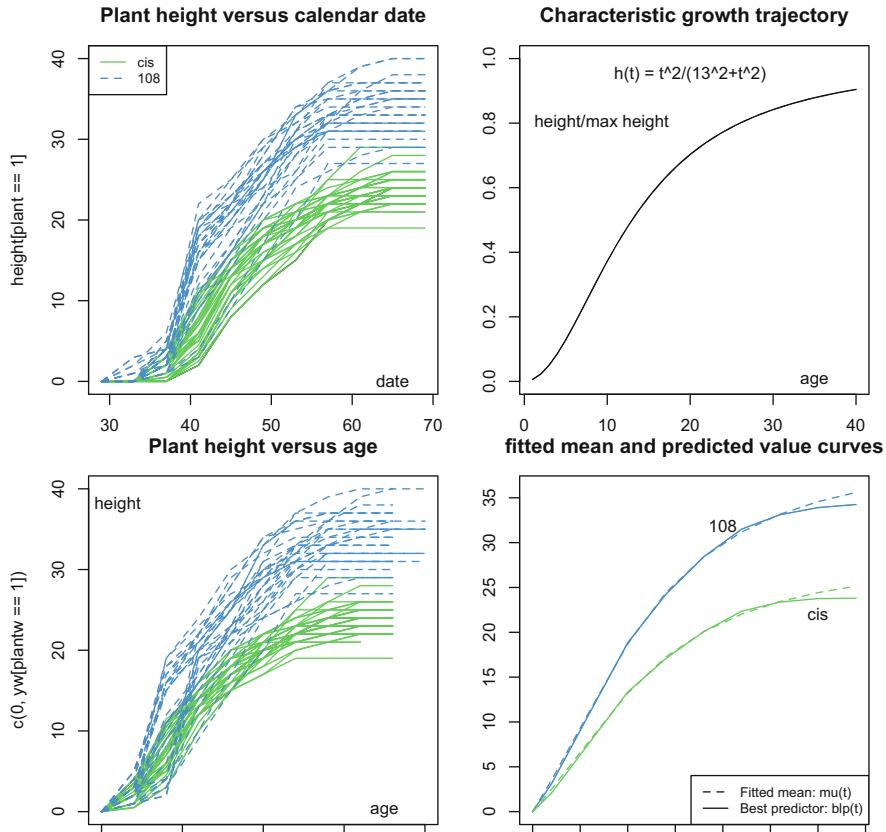


Fig. 4.1 Heights in mm of 70 *Arabidopsis* plants of two strains, plotted against calendar time in panel 1, and against age in panel 3 (lower left). Lower right panel shows the fitted mean functions (dashed lines) together with the best linear predictor (solid lines) of plant height for each strain

inverse squared time, $1/h(t) = \beta_0 + \tau^2/t^2$. The parameterization is such that $h(\tau) = \frac{1}{2}h(\infty)$, so τ is the semi-max age, which is approximately 13 days.

The growth trajectories are plotted against calendar time in the top panel of Fig. 4.1, and against plant age in the lower panel. The graphs give the impression that the number of plants is no more than 30, but there are in fact 69 distinct growth trajectories for 70 plants. The illusion is caused in part by heights being rounded to the nearest millimetre so that, at any given time, there are usually fewer than 20 distinct heights.

Two strains of plant are included in the study, the first 40 called ‘cis’ and the remaining 30 labelled ‘108’. One goal of this project is to compare the two strains, i.e., to characterize the difference between typical trajectories for the two strains, and to assess the significance of the observed differences. The time series plot for all plants in Fig. 4.1 reveals that both types have similar growth trajectories, but that

the ultimate height of the ‘108’ strain is about 40% greater than the ‘cis’ strain. The age-specific ratio of sample mean heights ‘108’/‘cis’ for plants aged 4–32 days is

| Age in days | 4 | 8 | 12 | 16 | 20 | 24 | 28 | 32 |
|---------------|------|------|------|------|------|------|------|-------|
| 108/cis ratio | 1.06 | 1.39 | 1.37 | 1.36 | 1.42 | 1.44 | 1.43 | 1.42. |

The fact that these ratios are remarkably constant from day 8 onwards suggests that a simple multiplicative factor suffices for strain effects.

4.2 Growth Curve Models

The growth curve for plant i is modelled as a random function $\eta_i(t)$ whose value at age zero is, in principle at least, exactly zero, and whose temporal trajectory is continuous. In the analyses that follow, $s(i)$ is the strain of plant i , the mean trajectory is $\beta_{s(i)}h(t)$ with $h(0) = 0$ and $h(\infty) = 1$, so that the plateau levels β_0 , β_1 , or β_{cis} , β_{108} , depend on the strain, and the ratio of means is constant over time. The observation process $Y(t) = \eta(t) + \epsilon(t)$ is also a function of time, but it is not continuous in t ; the additive measurement-error $\epsilon(\cdot)$ is assumed to have mean zero with constant variance σ_0^2 , and to be independent for all times $t > 0$ and all plants.

Brownian motion (BM) starting from zero at time $t = 0$ is a continuous random function with covariance function $\text{cov}(B(t), B(t')) = \min(t, t')$ for $t, t' \geq 0$. We are thus led initially to consider the additive Gaussian model with moments

$$\begin{aligned} E(Y_i(t)) &= \beta_{s(i)}h(t), \\ \text{cov}(Y_i(t), Y_j(t')) &= \sigma_0^2\delta_{ij}\delta_{t,t'} + \sigma_1^2K(t, t') + \sigma_2^2\delta_{ij}K(t, t') \end{aligned} \quad (4.1)$$

where $K(t, t') = \min(t, t')$ for the Brownian-motion contributions.

This formulation is incompatible with baseline information available at planting, which consists solely of two covariates, date and strain. Two deviations are noted. First, the observational units at planting are seeds, not growing plants. Second, it is most unlikely that every seed germinates, so the set of germinating plants is a subset of planted seeds—a random subset that is impossible to identify at planting time. This implies that `plant_id` is not a baseline factor nor is the germination date a baseline covariate. Formulation (4.1) with plants as observational units and t representing plant age, is compatible with baseline taken to be the plant-specific date of brearding or visible sprouting. Nonetheless, as the discussion in the preceding section shows, the determination of that date is slightly inaccurate.

One objection sometimes raised to Brownian motion as a model for a growth curve is that it is not sufficiently smooth; with probability one, a Brownian trajectory is everywhere continuous but nowhere differentiable. If a compelling argument could be made that physical growth is a differentiable function, one would have to reconsider the Brownian component, perhaps replacing it with a smoother random

function or a family of random functions having varying degrees of smoothness. But in the absence of a compelling demonstration of smoothness, the lack of differentiability of Brownian trajectories is not a strong argument against its use for growth curves. The Brownian component of the model can be replaced by any continuous random function deemed suitable, such as fractional Brownian motion (FBM), and the data can be permitted to discriminate among these. Despite the perception that physical growth curves must be smooth in time, trajectories rougher than BM are favoured over smooth trajectories: see Exercise 4.3.

Leaving aside the measurement error component, the growth-curve model (4.1) has two additional variance components, one Brownian motion with volatility coefficient σ_1 that is common to all plants regardless of strain, and another with coefficient σ_2 that is plant-specific and independent for each plant. In other words, for $t > 0$ the measured value on plant i is a sum of one non-random term and three independent random processes

$$Y_i(t) = \beta_{s(i)}h(t) + \epsilon_{it} + \sigma_1\eta_0(t) + \sigma_2\eta_i(t), \quad (4.2)$$

where $\eta_0, \eta_1, \dots, \eta_{70}, \dots$ are independent and identically distributed Brownian trajectories starting from zero at time zero. In this model, the variances

$$\begin{aligned}\text{var}(Y_i(t)) &= \sigma_0^2 + (\sigma_1^2 + \sigma_2^2)t \\ \text{var}(Y_i(t) - Y_j(t)) &= 2\sigma_0^2 + 2\sigma_2^2t\end{aligned}$$

are both increasing linear functions of plant age.

In (4.1) or (4.2), the average height at age t of a very large number of plants of strain s is $\beta_s h(t) + \sigma_1\eta_0(t)$. This infinite average is not a deterministic function of t ; it is the value of a Gaussian process with mean $\beta_s h(t)$ and covariance $\sigma_1^2 K(t, t')$. That is to say, $\sigma_1\eta_0(t)$ is the deviation of $\beta_s h(t)$ from the mean trajectory averaged over infinitely many plants of strain s . From the fitted model, the estimated, or predicted, plant height trajectory $E(Y_{i^*}(t) \mid \text{data})$ for a new plant i^* is shown for both strains in the fourth panel of Fig. 4.1. Each fitted trajectory is the sum of the fitted mean $\hat{\beta}_s h(t)$ plus the conditional expected value of $\sigma_1\eta_0(t)$ given the data. The latter term $E(\eta_0(t) \mid \text{data})$ is linear in the data; as a function of t it is a C^∞ -spline with knots at observation times, i.e., continuous at all points and infinitely differentiable at all non-observation times.

Only the 628 response values at strictly positive plant ages are included in the likelihood computations, the heights at $t \leq 0$ being exactly zero by construction. For the mean model, $\hat{\tau} = 12.782$ days is used throughout. The three variance-components estimated by maximum likelihood are

| parameter | estimate | S.E. |
|--------------|----------|-------|
| σ_0^2 | 1.047 | 0.150 |
| σ_1^2 | 0.050 | 0.029 |
| σ_2^2 | 0.428 | 0.052 |

with asymptotic standard errors as indicated. Asymptotic standard errors of variance components are worth reporting, but are often less reliable as indicators of significance than standard errors of regression coefficients (Dickey, 2020). The first coefficient implies that the standard deviation of the measurement error is around 1 mm, which is about right for laboratory measurements of plant height. The small value of σ_1^2 implies that $\beta_{s,h}(t)$ is a close approximation to the mean trajectory averaged over plants, and the relatively large standard error suggests that this term may be unnecessary. Nevertheless, the reduced model with only two variance components is demonstrably inferior: the increase in log likelihood is 12.14, i.e., the likelihood ratio chi-squared statistic is 24.28 on one degree of freedom. In this instance, the comparison of $\hat{\sigma}_1^2$ with its asymptotic standard error gives a misleading impression of the significance of that term.

The regression parameters governing the mean, τ included if necessary, are estimated by weighted least squares. For the 70 plants in this study, the plateau estimates in mm for the two strains are as follows:

| strain | coefficient | S.E. |
|-----------|-------------|------|
| cis | 28.29 | 1.56 |
| 108 | 40.04 | 1.72 |
| 108 – cis | 11.75 | 0.98 |

Although this is a two-sample comparative design, the variance of the 108/cis-contrast estimate is substantially less than the sum of the two variances.

Section 4.3.1 describes the Box-Tidwell method (Box & Tidwell, 1962), which has been used here for the calculation of asymptotic standard errors to make allowance for the estimation of τ . (For comparison, the unadjusted standard errors are 1.43, 1.48 and 0.94 respectively, which are too small.) The parametric bootstrap is a viable alternative for the assessment of standard errors, but is not needed here. This analysis makes it plain that the expected difference between the ultimate heights of the two strains is around 10–14 mm, and is certainly not zero.

4.3 Technical Points

4.3.1 Non-linear Model with Variance Components

The inverse quadratic model for the mean growth curve

$$\mu_{it} = E(Y_{it}) = \beta_{s(i)} t^2 / (\tau^2 + t^2)$$

has three parameters to be estimated, the two asymptote heights β_0 , β_1 and the semi-max temporal parameter τ such that $\mu_\tau = \mu_\infty/2$. Two options for the estimation of parameters are available as follows.

The most straightforward option is to use ordinary maximum likelihood (not REML) for the estimation of all parameters jointly. Since the model for fixed τ is linear in β , this can be done by computing the profile likelihood for a range of τ values, say $10 \leq \tau \leq 15$ in suitably small steps, and using the `kernel=0` option in `regress()` as follows.

```
h <- age^2/(tau^2 + age^2)
fit0 <- regress(y~h:strain-1, ~BM+BMP, kernel=0)
fit0$llik
```

Although all ages used in the computation are strictly positive, the model formula is such that the mean height at age zero is exactly zero. This constraint is enforced by exclusion of the intercept in the model formula `h : strain-1`. We find that the log likelihood is maximized at $\hat{\tau} \simeq 12.782$. A plot of the profile log likelihood values against τ can be used to generate an approximate confidence interval if needed: the 95% limits are approximately (11.7, 14.2) days.

A follow-up step is needed in order for the standard errors of the β -coefficients to be computed correctly from the Fisher information. To compensate for the estimation of τ , the derivative of the mean vector with respect to τ at $\hat{\tau}$ must be included as an additional covariate, as described by Box and Tidwell (1962)

```
deriv <- -2*tau * fit$fitted * h / age^2
fit0a <- regress(y~deriv+h:strain-1, ~BM+BMP,
                  start=fit0$sigma, kernel=0)
```

The `start` option takes the initial variance components from the previous fit.

It is a property of maximum likelihood estimators for exponential-family models that the residual vector $y - \hat{\mu}$ is orthogonal to the tangent space of the mean model (with respect to the natural inner product $\hat{\Sigma}^{-1}$). Consequently, the coefficient of `deriv` is exactly zero by construction, and all other coefficients β , σ^2 are unaffected. The ordinary maximum-likelihood estimates of the variance components are (1.0467, 0.0496, 0.4283), the plateau coefficients are (28.293, 40.042) mm, and the standard error of the difference is 0.975. In this instance, the unadjusted standard error is 0.941, so the effect of the adjustment is not great.

In more complicated settings where partially linear models are employed, the mean-value space for fixed τ is a linear subspace $\mathcal{X}_\tau \subset \mathbb{R}^n$. The intersection of these is another subspace $\mathcal{K} = \cap_\tau \mathcal{X}_\tau$. In the growth-curve example, $\mathcal{K} = \mathbf{0}$ is the zero subspace, but, in general, \mathcal{K} may be non-zero. In that setting, it is better to use \mathcal{K} as the kernel subspace for model comparisons.

4.3.2 Fitted Versus Predicted Values

The mean functions for the two strains are $\beta_0 h(t)$ and $\beta_1 h(t)$, and the fitted curves with β_s replaced by $\hat{\beta}_s$ are shown as dashed lines in the lower right panel of Fig. 4.1.

The fitted mean is not to be confused with the predicted growth curve for an extra-sample plant i^* of strain s , which is deemed to have a response

$$Y_{i^*}(t) = \beta_s h(t) + \sigma_1 \eta_0(t) + \sigma_2 \eta_{i^*}(t) + \epsilon_{i^* t}.$$

In settings such as this, it is good to bear in mind that a stochastic formulation such as (4.1) or (4.2) is not a model for the sample or the data recorded in the experiment; it is a sampling model for the process. As such, it is understood to hold for arbitrary fixed samples, both in-sample units and extra-sample units. Since the extra-sample unit i^* has a growth trajectory that is not independent of the in-sample responses, the covariances

$$\rho_t(i, t') = \text{cov}(Y_{i^*}(t), Y_i(t')) = \sigma_1^2 \text{cov}(\eta_0(t), \eta_0(t')) = \sigma_1^2 K(t, t')$$

are not all zero. The conditional distribution given the data is Gaussian with conditional mean

$$E(Y_{i^*}(t) \mid \text{data}) = \beta_s h(t) + \rho_t' W(y - \mu), \quad (4.3)$$

where μ is the n -component vector of means for all observations on in-sample plants, ρ_t is the n -component vector of covariances, and $W = \Sigma^{-1}$ is the inverse $n \times n$ covariance matrix. The fitted conditional distribution, or the fitted predictive distribution, has a mean $\hat{\beta}_s h(t) + \hat{\rho}_t' \hat{W}(y - \hat{\mu})$, called the best linear predictor (BLUP). This is shown as a pair of solid curves in the lower right panel in Fig. 4.1, one curve for each strain; the fitted means are shown as dashed lines.

For the growth of arabidopsis plants, the inverse linear curve $h(t) = t/(\tau + t)$ is not nearly so effective as the inverse quadratic. The fitted curves are shown as dashed lines, one for each strain, in Fig. 4.2. The fitted semi-max age is an implausible $\hat{\tau} = 50.7$ days. For such a large semi-max age, the fitted curves are nearly linear over the observed range, and the heights for both strains are clearly underestimated in the middle of the age range. To accommodate this deviation, the fitted volatility of the Brownian-motion term $\eta_0(t)$ in (4.2) is $\hat{\sigma}_1 = 0.96$, which is four times as large as the corresponding term in the inverse-quadratic model. Despite the poor fit, the curve of predicted values for the inverse linear model is not appreciably different from the curve of predicted values for the inverse quadratic model. Both are shown as solid lines in Fig. 4.2. The maximum difference between predicted curves is approximately one millimetre (or 3%) at age 40 days.

The average heights at each age are also shown in Fig. 4.2. The fitted inverse-quadratic curve over-estimates the average height at ages 30 days and above, so the term $\eta_0(t)$ is needed, and the predicted curve tracks the averages reasonably well. Across the age range, the inverse quadratic deviations are much smaller than the inverse-linear deviations.

If we had taken the plant age to be one or two days rather than four at the time of the first positive measurement, the comparison would be more favourable for the inverse-linear model. The reduced offset for the temporal origin substantially

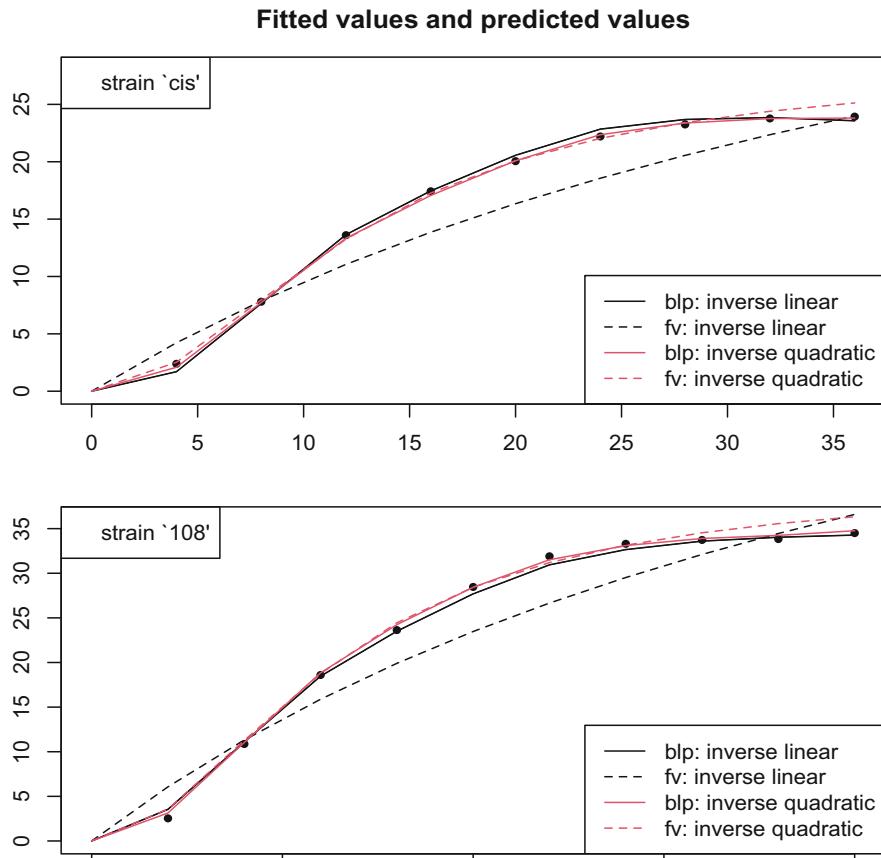


Fig. 4.2 Fitted mean growth curves (dashed lines) and best linear predictors (solid lines) of plant height for two strains, using an inverse linear or inverse quadratic model for the mean and Brownian motions for the deviations. Sample average heights at each age are indicated by dots

improves the inverse-linear fit, but even so, it is less satisfactory than the inverse quadratic.

In certain areas of application such as animal breeding, the chief goal is to make predictions about the meat or milk production of the future progeny of a specific individual bull. This bull is not an extra-sample individual, but one of those experimental animals whose previous progeny have been observed and measured. Such predictions are seldom called for in plant studies, but may be required in animal growth studies. From a probabilistic viewpoint, the procedure for in-sample units is no different. If i^* is one of the sampled plants and t is an arbitrary time point, the covariance of $Y_{i^*}(t)$ and $Y_i(t')$ is

$$\rho_{i^*t} = \sigma_1^2 K(t, t') + \sigma_2^2 \delta_{i,i^*} K(t, t'),$$

which involves two of the three variance components. The conditional expected value (4.3) yields a continuous temporal curve specific to each plant. The observed values for plant i do not lie on this curve, so (4.3) is not an interpolation.

The third variance component does not occur in the preceding calculation because the phrase ‘ t is an arbitrary time point’ is interpreted to mean that plant i^* was not measured at time t . Thus the contribution ϵ_{t^*t} is independent of all observations, and the conditional expectation is

$$E(Y_{i^*}(t) | \text{data}) = E(Y_{i^*}(t)) + E(\sigma_1 \eta_0(t) + \sigma_2 \eta_{i^*}(t) | \text{data}).$$

Section 15.4.3 discusses in more detail the problem of prediction in a linear Gaussian setting, and Sect. 15.4.4 covers Eddington’s formula for prediction in a partially Gaussian setting.

4.4 Modelling Strategies

1. Choice of temporal origin. The distinction between calendar time and plant age is fundamental. The decision to measure plant age relative to the time of brearding is crucial, and has a greater effect on conclusions than any subsequent choice.
2. Selection of a characteristic mean curve. The mean curve must pass through the origin at age zero, so a logistic function $e^t/(1+e^t)$ cannot be used. The graphs in Fig. 4.1 suggest an inverse quadratic curve, which may or may not be appropriate for other plants.
3. Use of a non-stationary covariance model. Plant growth curves are intrinsically non-stationary because they are tied to the origin at age zero. For obvious biological reasons, animal growth curves using weight in place of height are not similarly constrained.
4. Brownian motion. It seems reasonable that every growth curve should be continuous in time. It seems reasonable also to model the response process as a sum of the actual height plus independent measurement error, thereby making a distinction between plant height and the measurements made at a finite set of selected times. The particular choice (BM) is not crucial, and can be improved using fractional Brownian motion. It is also possible to mix these by using FBM for the plant-specific deviation, and BM for the common deviation, or vice-versa.
5. Positivity. Plant heights are necessarily positive—or at least non-negative—at all positive ages, whereas every Gaussian model puts positive probability on negative heights. This is one of those compromises, some major, some minor, that are frequently needed in applied work. Provided that the probability of a negative value is sufficiently small, this compromise is a good bargain: see point 9 below and Sect. 12.2.
6. Response transformation, usually $y \mapsto \log(y)$, is an option that must always be considered: see Chap. 19. The log transformation might be reasonable for

animal growth curves, but it was rejected here because of the role of zero height in determining the age origin.

7. Limiting behaviour. Plants do not grow indefinitely or live for ever, so the capacity of the growth model for prediction is limited to the life span of a typical plant.
8. Other issues. The emphasis on growth curves overlooks the possibility that the two strains may differ in other ways. In fact, the average brearding time for strain ‘108’ is two days less than the time for strain ‘cis’, with a standard deviation of 0.43 days. No single summary tells the whole story.

4.5 Miscellaneous R Functions

The following is a list of various R functions used in the construction of covariance matrices, and in the fitting of variance-components models.

```

BM <- outer(age, age, "pmin")      (BM covariance matrix)
nu <- 0.25; p <- 2*nu;      (ν is the Hurst index for FBM)
FBM <- outer(age^p, age^p, "+") - abs(outer(age, age, "-"))^p
Plant <- outer(plant, plant, "==")   (plant block factor)
BMP <- BM * Plant      (component-wise matrix multiplication)
FBMP <- FBM * Plant     (i.i.d. FBM for each plant)
mht0 <- tapply(height[strain==0], age[strain==0], mean)
mht1 <- tapply(height[strain==1], age[strain==1], mean)
tapply(brearded, strain, mean)
L <- t(chol(FBM))    (Choleski factorization)
fit <- regress(y~h:strain-1, ~BM+FBMP, kernel=0)

```

Computer files

PlantGrowth.dat PlantGrowth.R

4.6 Exercises

4.1 In the inverse quadratic model, the height of plant i at age t is Gaussian with mean $\beta_{s(i)}h(t)$ whose limit as $t \rightarrow \infty$ is $\beta_{s(i)}$. What is the variance of the ultimate height of plant i ?

4.2 For the inverse linear model in which brearding is deemed to have occurred two days prior to the first positive measurement, estimate τ together with the plateau coefficients. Obtain the standard error for the estimated limiting difference of mean heights for the two strains.

4.3 The Brownian motion component of the model can be replaced with fractional Brownian motion with parameter $0 < \nu < 1$, whose covariance function is

$$\text{cov}(Y(s), Y(t)) = s^{2\nu} + t^{2\nu} - |s - t|^{2\nu},$$

where $s, t \geq 0$. The index ν is called the Hurst coefficient, and $\nu = 1/2$ is ordinary Brownian motion. Show that the fit of the plant growth model can be appreciably improved by taking $\nu \simeq 1/4$.

4.4 Bearing in mind that the heights are measured to the nearest millimetre, comment briefly on the magnitude of the estimated variance components for the FBM model.

4.5 In the fractional Brownian model with $\nu < 1/2$, the temporal increments for non-overlapping intervals are negatively correlated. Suggest a plausible mechanism that could lead to negative correlation.

4.6 For 1000 equally spaced t -values in $(0, 10]$ compute the FBM covariance matrix K and its Choleski factorization $K = L'L$. (If $t = 0$ is included, K is rank deficient, and the factorization may fail.) Thence compute $Y = L'Z$, where the components of Z are independent and identically distributed standard Gaussian, and plot the FBM sample path, Y_t against t . Repeat this exercise for various values of ν in $(0, 1)$ and comment on the nature of FBM as a function of the Hurst coefficient.

4.7 Several plants reach their plateau well before the end of the observation period. How is the analysis affected if repeated values are removed from the end of each series?

4.8 Explain the purpose and the implementation of the Box-Tidwell method. Why must the unmodified REML criterion be avoided?

4.9 Investigate the relation between brearding date and ultimate plant height. Is it the case that early-sprouting plants tend to be taller than late-sprouting plants?

Chapter 5

Louse Evolution



5.1 Evolution of Lice on Captive Pigeons

5.1.1 Background

Understanding the mechanisms responsible for the origin of new species is a fundamental topic in evolutionary biology that has been the focus of numerous experiments and much speculation dating back at least to Darwin, who argued that differential natural selection in a range of environments leads to reproductive isolation and thence, eventually, to the formation of new species. Chapter 3 is concerned with speciation induced by differential diets over approximately 40 generations of *Drosophila*. This chapter considers another experiment on the same theme, but with a different system and different environmental pressures.

The paper *Rapid experimental evolution of reproductive isolation from a single natural population* published by Villa et al. (PNAS 2019) is concerned with reproductive isolation developing in response to body-size evolution in isolated lineages of pigeon lice. Each lineage evolved over 60 generations on a different host pigeon. Half of the experimental hosts were captive feral pigeons; the other half were giant runts, a domesticated breed that is roughly three times as large by weight as a feral pigeon.

To establish their claim, the authors must show evidence of two phenomena: first that louse size evolves rapidly in giant runt hosts relative to that in captive feral hosts, and second that differential louse size induces sexual isolation. The evidence for both of these phenomena is essentially statistical. The mechanism by which size differences lead to reproductive isolation is important from an evolutionary standpoint, but this chapter deals only with size evolution, i.e., whether systematic

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-14275-8_5.

louse-size changes are detectable in a 60-generation span. Our concern is not so much with the evolutionary implications of the authors' findings, but with the experimental design, the data analysis, and the inferences that follow. The goal is solely to examine the data for evidence of systematic body-size changes in response to host size.

5.1.2 Experimental Design

The following synopsis of experimental procedure is taken directly from Villa et al. (2019). Before the start of the experiment, resident lice on all experimental pigeons were eradicated by housing the birds in low-humidity conditions for at least ten weeks. According to the authors, this procedure kills both lice and eggs, while avoiding residues from insecticides. To begin the experiment, 800 lice taken from wild-caught feral pigeons were transferred to 32 lice-free experimental pigeons, 25 lice per host. Pigeons were housed in eight aviaries, each aviary containing four birds of the same breed. Every six months, a random sample of lice from each bird was removed, photographed, and returned to the host. The sex, body length, metathorax width, and head width of each louse was recorded.

One aspect of this design is different from that in Chap. 3. After measurements were made, the lice were returned to their host. This was done in order to minimize the effect of measurement on the host-parasite system. Otherwise, the act of measurement would reduce the resident population, and introduce instability in the lineage, which is not desirable. In the design in Chap. 3, the flies removed for experimental purposes were reared separately for one generation on a standard diet, so it was not possible to return them to the main breeding line. However, the *Drosophila* breeding lines were more easily controlled, so plans could be made in advance to accommodate the numbers needed in any particular generation.

As always in situations of this sort, the phrase 'random sample of lice from each bird' must be treated with caution, particularly with regard to size measurements. Larger lice are more visible than smaller specimens, so it would be naive to expect the random sample to behave like a simple random sample of the resident lice on a given bird. Nonetheless, size-biased sampling need not be a serious concern for this experiment provided that it affects all birds equally.

5.1.3 Deconstruction of the Experimental Design

Since each measurement is made on one louse, it is evident that each observational unit is either one louse or one louse on one occasion, while the response Y_u is a point in the state space, which is $\{M, F\} \times \mathbb{R}^3$ for three size measurements. Since one louse generation is approximately 24–25 days, and measurement occasions are six months apart, we can be sure that no louse was measured on more than one

occasion. While there is no practical distinction between louse and louse-occasion as the observational unit, as a matter of principle the ordered pair is the correct choice.

The lice are arranged in 32 lineages, one lineage to each bird. Thus *lineage* and *pigeon* are equivalent as block factors, and *aviary* is a coarser partition or block factor with eight levels. With respect to birds, *host* or *breed* is a binary classification factor.

The baseline is the time at which de-lousing was complete, and the experiment was ready to commence with new lice lineages on captive birds. The paper mentions randomization only incidentally in the ‘Materials and Methods’ section, and the reference there is a little ambiguous, but two crucial choices appear to have been made at baseline. First, the 800 initial lice were partitioned into 32 lineages with 25 founders for each lineage. Second, each lineage was associated with a particular bird. Regardless of the biological and mechanical constraints in the laboratory, it seems reasonable and mathematically natural to regard each of these steps as the outcome of an independent uniform randomization scheme. Since the objective is to study selective pressure, host size is the principal treatment. If the randomization was done in two steps as indicated, treatment is assigned to lineages in step two, in which case each lineage serves as one experimental unit.

By definition, a covariate is a pre-baseline variable, and it appears that there is only one. Measurement occasion or *time* is a function on the observational units, which is a quantitative factor. However, as indicated in the preceding paragraph, *lineage* could be regarded as a pre-baseline block factor, and it should certainly be used as the experimental unit to assess variability of the treatment effect estimate.

In addition to *time* and *lineage*, pre-baseline vital measurements including louse sex are available on the 800 founder lice. All pre-baseline variables are available for use as covariates as if the values were fixed and non-random, and initial response values are no different in that respect from any other pre-baseline measurements. Randomization ensures that the distribution of initial values is the same for all treatment groups, so the initial response values are uninformative for treatment effects. Generally speaking, when the response is a time series or temporal process, it is more convenient and mathematically more natural to treat initial response measurements as an integral part of the response process. A crucial point is that the probability model for the response at $t = 0$ must be consistent with the randomization: see Sects. 5.2.3, 5.2.5 and 13.2. The joint distribution implies a conditional distribution, which is available if needed for purposes of estimation or prediction.

The paper does not discuss how birds were assigned to aviaries, but it seems reasonable to regard that too as the outcome of a balanced randomization applied to birds, subject to restrictions mentioned earlier. We presume here that birds were quarantined in their aviaries during de-lousing, in which case *aviary* is a pre-baseline block factor. Since all birds in one aviary are of the same breed, a strong argument can be made that *aviary* is the experimental unit, not *lineage* as stated earlier. Both seem to be relevant. Whether they are pre-baseline or immediately post-baseline, *time*, *lineage*, *aviary*, and *treatment* are available for purposes of analysis and model construction.

Apart from the founders, louse sex is a post-baseline variable, and thus one of four components in the response. Genetic theory leads one to expect the sex ratio should remain steady at 50:50 for most species, and post-baseline counts in Table 5.3 confirm this. But the same table also shows that the baseline F:M ratio is 464:336, which is significantly in excess of 50:50.

Each lineage was associated with a particular pigeon at baseline, which means that *lineage* and *pigeon* are equivalent as block factors. A subsequent remark in the paper shows that this statement is not quite correct. When a bird died during the experiment, all lice from the dead bird were transferred to a new parasite-free bird of the same type. Thus, one lineage could span two or more birds. Unfortunately the data file does not indicate when deaths might have occurred, so we have no way to check the effect on lineages of host transfers.

5.2 Data Analysis

5.2.1 *Role of Tables and Graphs*

However it is measured, the response of evolutionary interest is louse size. To keep matters as simple as reasonably possible, we focus on the single response, body length. Since we plan to use additive decompositions, the log transform is more or less automatic, so Y_u is the log body length for louse u . However, the range of variation in all size measurements is only a few percent of the average, so the log transformation is essentially linear and has little effect on conclusions.

The purpose of a table or graph is to advance the narrative thread by drawing attention to the most important patterns or features in the data such as the nature and direction of various effects. It is natural enough to emphasize the effects of scientific interest—but not at the cost of misleading the reader. Every table or graph invites the question ‘What is the point of this table?’ or ‘What feature does this graph illustrate?’. If the answer is not clearly apparent, the narrative is not advanced, and the reader is likely to be confused. Generally speaking, the data analyst examines numerous tables and graphs. Only the most useful of these are retained for presentation.

The first four rows of Table 5.1 show the average log body length of all lice measured on each occasion. Most impressive is the stability of body length for both louse sexes over 60 generations. If anything, there is a slight decrease in length for lice on both hosts, with a slightly greater decrease for captive feral pigeons.

The numbers in Table 5.1 are accurate to three decimal places or four decimal digits, but the first three digits are essentially constant at 7.88 for females and 7.72 for males, so we say that there are only 1–2 significant decimal digits. Usually, one digit is not enough to gauge accurately the statistical variation in the process. However, we have chosen to leave the table in its present form to emphasize how tiny are the size differences between lice on the two hosts.

Table 5.1 Average log body length (in μm) of lice on two pigeon hosts

| Sex | Host | Time in months | | | | | | | | |
|---|-------|----------------|-------|-------|-------|-------|-------|-------|-------|-------|
| | | 0 | 6 | 12 | 18 | 24 | 30 | 36 | 42 | 48 |
| F | Feral | 7.883 | 7.883 | 7.883 | 7.874 | 7.866 | 7.886 | 7.880 | 7.872 | 7.864 |
| F | G.R. | 7.885 | 7.894 | 7.882 | 7.882 | 7.882 | 7.895 | 7.894 | 7.899 | 7.886 |
| M | Feral | 7.720 | 7.716 | 7.705 | 7.700 | 7.702 | 7.712 | 7.709 | 7.713 | 7.700 |
| M | G.R. | 7.720 | 7.718 | 7.717 | 7.716 | 7.713 | 7.723 | 7.726 | 7.731 | 7.720 |
| Differences $\times 100$: Giant runt – Feral | | | | | | | | | | |
| F | G–F | 0.2 | 0.1 | -0.1 | 0.8 | 1.7 | 0.9 | 1.4 | 2.6 | 2.2 |
| M | G–F | 0.0 | 0.2 | 1.2 | 1.7 | 1.1 | 1.1 | 1.7 | 1.8 | 2.0 |

The sex difference $7.88 - 7.72 = 0.16$ on the log scale means that female lice are about 17% longer than males: ($e^{0.16} \simeq 1.17$). The last two rows show that the mean difference for hosts tends to increase over time, reaching around 2% for both sexes after 48 months. It is remarkable that such a small size difference could have a detectable effect on sexual coupling.

The first panel of Fig. 5.1 shows a plot of the same data with sexes combined. Automatic centering and re-scaling of the y -axis has the effect of exaggerating the variation and the magnitude of the divergence between the two groups. In other words, that which is emphasized by the table of averages is eliminated by the plot.

The remaining panels show similar plots for the head width, the metathorax width, and the first principal component, which is a roughly equally-weighted positive linear combination of the three standardized size variables. For all size variables, the temporal trajectory for louse size on giant runts is surprisingly similar to that for feral pigeons, and lice on giant runts are larger on average than those on feral pigeons. Apart from the uniform decrease in all size measurements in the initial and final intervals, no clear temporal trend is visible.

Ideally, it would be good to show error bars for every point. But size measurements for different lice on one pigeon are not independent, so honest error assessment is not straightforward. On balance, it is better to show no error bars than to show the naive default based on independence, which is misleading in this setting: contrast Fig. 5.2 with Table 5.4 in Sect. 5.5.2.

5.2.2 Trends in Mean Squares

Table 5.1 and Fig. 5.1 illustrate temporal trends in average body size. To get a comparable impression of trends in variance, it is helpful to compute mean-squares associated with *louse sex*, *host size*, *aviary*, *lineage* and residuals at each of the nine time points.

The dominant mean square is that for *louse sex* which starts off at 5.25 at baseline, drops to half that value at six months and decreases slowly to 1.64 at

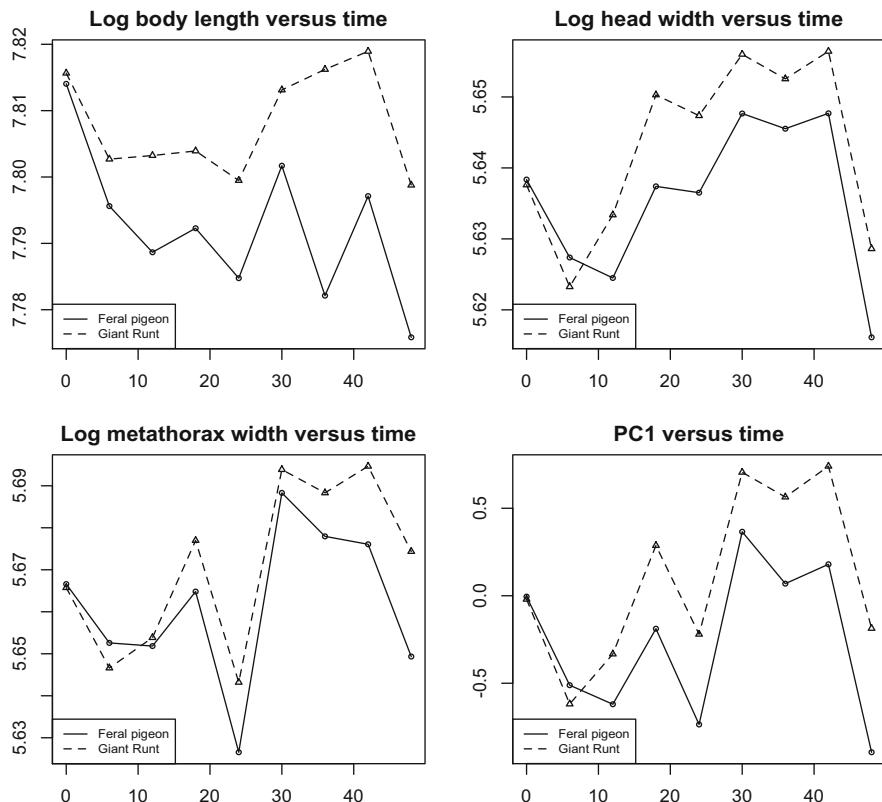


Fig. 5.1 Average body sizes of lice for two hosts over time

48 months. For the other factors, the mean squares are shown in the top half of Table 5.2, together with the REML variance components for *aviary*, *lineage* and residual in the second half. For this fit, *host* and *sex* were eliminated as fixed effects, so the mean-squared residual does not coincide exactly with $\hat{\sigma}_0^2$.

Some of the following points are accommodated in subsequent analyses, but others are merely noted.

1. The residual variability at baseline is twice that on all subsequent occasions. One plausible explanation is that founder lice collected from wild pigeons are more variable in size than those resident on captive pigeons.
2. The lineage mean square is remarkably constant from baseline onwards. Relative to the residual mean square, it is below expectation at baseline, but not significantly so. After baseline, it is uniformly larger than the residual mean square, but not by a large factor.
3. The host mean square at baseline seems artificially low. There is strong evidence in the data, for example in the sex ratios, that the randomization scheme was

Table 5.2 Trends in mean squares and variance components $\times 10^5$

| MS | Time t in months | | | | | | | | |
|--|--------------------|------|------|------|------|------|------|------|------|
| | 0 | 6 | 12 | 18 | 24 | 30 | 36 | 42 | 48 |
| Host | 12 | 340 | 190 | 996 | 1095 | 690 | 1024 | 3448 | 1506 |
| Aviary | 290 | 204 | 408 | 575 | 493 | 333 | 324 | 528 | 350 |
| Lineage | 83 | 85 | 85 | 94 | 87 | 80 | 100 | 79 | 152 |
| Residual | 111 | 52 | 60 | 68 | 56 | 52 | 56 | 57 | 64 |
| Variance-component estimates ($\times 10^5$) | | | | | | | | | |
| $\hat{\sigma}_{\text{aviary}}^2$ | 2.1 | 3.2 | 10.7 | 10.4 | 9.5 | 6.5 | 6.7 | 10.6 | 6.6 |
| $\hat{\sigma}_{\text{lineage}}^2$ | -1.1 | 2.8 | 3.3 | 3.8 | 2.8 | 2.2 | 5.5 | 0.3 | 12.1 |
| $\hat{\sigma}_{\text{resid}}^2$ | 111.3 | 53.1 | 59.4 | 67.3 | 56.9 | 52.5 | 56.1 | 58.4 | 63.3 |

more complicated than that depicted in the preceding section, so this may be a consequence of an effort to balance the randomization.

4. The between-aviary mean square at baseline is a little larger than expected from uniform random assignment: the F -ratio is 2.6, which is at the upper 1.6% point of the reference null distribution.
5. Variance-component estimates on few degrees of freedom, such as those for aviary and lineage, have notoriously high variances.

The main issue to be addressed at this point is the size of the aviary mean square at baseline, and whether the mean square provides sufficient probative evidence to cast doubt on the randomization or to declare it inadequate or biased. The question is not whether the initial lice were labelled 1–800 and lots drawn to determine which lice would be assigned to which birds, but whether the laboratory procedures actually employed are a reasonable facsimile of objective randomization. The only evidence before the court is shown in Table 5.2.

One traditional view is that the aviary mean square is selected for attention as the largest of three or four, so the p -value, or measure of extremity, is closer to 5%. That calculation tells us something, but it does not answer directly the question of interest to the court: ‘Given the data, what is the probability that the allocation to aviaries was biased?’ From another viewpoint in which sparseness prevails at odds level ρ , the odds against aviary bias given the mean-square ratio $F = 2.6$ on 6, 367 degrees of freedom are approximately $\rho \zeta_6(2.6)$, where $\zeta_6(2.6) = 3.81$. This calculation uses a modification for F -ratios of the sparsity argument in McCullagh and Polson (2018). The strength of the evidence is such that the initial presumption of innocence with probability $1/(1 + \rho)$ is changed to $1/(1 + \rho \zeta_6(2.6))$. For $\rho = 0.1$, which is not a strong prior presumption for this setting, the probability of a no aviary bias is changed by the evidence from 0.91 to $1/1.38 = 0.72$. So we take note and proceed with caution, giving the randomization a provisional pass. This point is revisited in Sect. 5.4.2.

5.2.3 Initial Values and Factorial Subspaces

If host size has an effect on louse size, it is an evolutionary development, so the effect is not immediate. Thus, *treatment* and *time* are the principal covariates whose effects are to be studied. In addition, the body length for *C. columbae* male lice is approximately 85% of that for females, so louse sex must also be taken into consideration. The effects of *lineage* and *aviary* are assumed to be additive random variables with independent and identically distributed components for each pigeon and each aviary respectively. Since their effects are additive zero-mean random variables, *lineage* and *aviary* do not contribute to the mean-value subspace.

Setting the two block factors aside temporarily, the factors *treatment* or *host size*, *time* and *sex* are to be taken into account. If we proceed to use factorial models in the naive manner, we may begin with all three main effects and check which interactions are needed. Or we may follow the authors' practice in their Tables S2–S5, which is to report the coefficients in the full three-factor interaction model. Both approaches are technically incorrect. Fitting either of the suggested models is a pointless exercise that serves only to confuse the narrative thread for this experiment.

The problem with the naive application of factorial models to this design lies in the role of time, and the fact that $t = 0$ corresponds to the experimental baseline. If Y_{ut} denotes the log body length of louse u at time t , the additive main-effects model for the conditional mean given sex and host has the form

$$E(Y_{ut} | s, h) = \beta_0 + \beta_1 t + \beta_2 s(u) + \beta_3 h(u),$$

in which $h(u)$ is a code for the host size, and $s(u)$ is the louse sex. At baseline, the additive model implies

$$E(Y_{u0} | s, h) = \beta_0 + \beta_2 s(u) + \beta_3 h(u),$$

with three coefficients to be estimated. The presumption of randomization, which is that lice are assigned to hosts independently of their size, implies $\beta_3 = 0$. Thus, randomization contradicts both the additive model and any other factorial model that contains it as a subspace.

Whether or not randomization was explicitly employed in this experiment, it is reasonable to imagine or suppose that the initial assignment of lice was effectively randomized. Randomization has implications. The use of a model that contradicts those implications is a source for confusion; the use of a model that conforms with randomization is strongly advised.

Only the most cynical reader would seriously consider the possibility that the researchers had deliberately assigned the lice differentially to hosts or to aviaries in an inappropriate manner. However, there might well be sound biological arguments for balancing the design in certain ways or for favouring females in the establishment of lineages. Deviations of this sort are normal practice, but they

must be clearly reported. Nonetheless, unintentional biased assignment can occur, so it is routine in many areas of application to check whether the baseline values are in conformity with randomization. That can be done here. While there is no indication of bias in Fig. 5.1, randomization implies that the mean squares for aviary, lineage and residual have the same expected value at baseline. However, the aviary-to-residual mean-square ratio in Table 5.2 is 2.61, which falls near the upper 98.5 percentile of the null distribution. This is not proof positive of randomization bias, but it is a little troubling and calls for an explanation.

The phenomenon described in this and in the following two sections—of time in relation to treatment and initial values—is fundamental in its own way. Although the implications are both substantive and substantial, the conflict with factorial models is seldom emphasized in methodological work—possibly because the issue seems to arise infrequently. As a result, the matter is not widely appreciated by those most likely to encounter it in scientific research. A simple example is given in Exercise 3.11 of McCullagh and Nelder (1989).

A related issue arises in connection with factorial models for the effect of varying doses of multiple fertilizers or medications. Zero dose is special in that a zero dose of one medication is physically indistinguishable from a zero dose of another. In one instance, the exchangeability of responses is induced by baseline randomization; in the other, it is a consequence of chemistry or biochemistry. All of this may be obvious common sense, but the implications do bear repetition (Bailey, 2008).

5.2.4 A Simple Variance-Components Model

The following linear models address directly the question that is of principal interest to an evolutionary biologist. Without straying from linearity in time, the null and alternative may be formulated as linear subspaces.

$$H_0: E(Y_{ut}) = \beta_0 + \beta_1 t + \beta_2 s(u); \quad (5.1)$$

$$H_A: E(Y_{ut}) = \beta_0 + \beta_h(u)t + \beta_2 s(u). \quad (5.2)$$

The model formulae `time+sex` and `host : time+sex` generate basis vectors for the two subspaces whose dimensions are three and four respectively. The alternative model has two linear trends in time, one for captive feral hosts $h(u) = 0$, and one for giant runts $h(u) = 1$.

For covariances, we start out following the authors' suggestion with three variance components

$$\text{cov}(Y_u, Y_{u'}) = \sigma_0^2 \delta_{u,u'} + \sigma_1^2 \delta_{l,l'} + \sigma_2^2 \delta_{a,a'}, \quad (5.3)$$

where l, l' and a, a' are the lineages and aviaries respectively. This is a linear combination of three identity matrices, one on the lice, one on the lineages or pigeons

with 32 blocks, and one on the aviaries with eight blocks. It is usually justified either by appeal to exchangeability based on recorded similarities of observational units, or, if that argument fails to convince, by appeal to randomization. Although neither argument carries weight in this instance, computation is cheap so we proceed.

For the log body length, the REML variance components in (5.3) paired with (5.2) are

$$\begin{array}{lll} \text{lice} & \hat{\sigma}_0^2 & 78.19 \times 10^{-5}, \\ \text{lineages} & \hat{\sigma}_1^2 & 1.84 \times 10^{-5}, \\ \text{aviaries} & \hat{\sigma}_2^2 & 1.40 \times 10^{-5}. \end{array}$$

Both the lineage and aviary variance components are small relative to the between-lice variance. Despite that, there is no compelling reason to declare them null simply because they are small. The fitted slope coefficients ($\times 10^4$) for the two pigeon breeds are

| Parameter | Estimate | s.e. |
|------------|----------|------|
| Feral:time | -2.23 | 0.53 |
| Giant:time | 1.37 | 0.38 |
| Difference | 3.60 | 0.63 |

This analysis appears to provide reasonably strong evidence that lice transferred to captive feral pigeons decrease in size over time, and moderately strong evidence that lice transferred to giant runts increase in size over time. However, the analysis is based on linearity in time, which seems implausible given Fig. 5.1, and a covariance structure (5.3) that is both inadequate for the data and in conflict with randomization.

5.2.5 Conformity with Randomization

Randomization implies that the body-size measurements at $t = 0$ are exchangeable with respect to some group of permutations, here assumed to be large enough that the responses for every pair of lice have the same joint distribution regardless of whether they are assigned to the same pigeon, to different pigeons in the same aviary or to different pigeons in different aviaries. Unfortunately, randomization implies $\sigma_1 = \sigma_2 = 0$ in (5.3).

Ever since the pioneering work of Cavalli-Sforza and Edwards (1967), Brownian motion has been the standard probabilistic model for the neutral evolution of a quantitative trait (Felsenstein, 2004, chapter 23). The conflict with randomization can be fixed only by introducing non-stationary temporal processes for the lineage and aviary effects, and the most natural way to incorporate Brownian motion is as follows:

$$\text{cov}(Y_u, Y_{u'}) = \sigma_0^2 \delta_{u,u'} + \sigma_1^2 K(t, t') \delta_{l,l'} + \sigma_2^2 K(t, t') \delta_{a,a'} + \sigma_3^2 K(t, t'). \quad (5.4)$$

The Brownian covariance function $K(t, t') = \min(t, t')$ is positive semi-definite, and $K(0, 0) = 0$ ensures conformity with randomization. Environmental selective pressure exerts a genetic drift, and the mean model (5.2) contains one drift parameter for each host, so the differential drift is the treatment effect.

The rationale for (5.4) is as follows. The louse population as a whole evolves as a Brownian motion with volatility σ_3 ; each aviary evolves independently as a Brownian motion with volatility σ_2 ; and each lineage evolves independently as a Brownian motion with volatility σ_1 . For the duration of this experiment, each louse belongs to the system, a lineage and an aviary, and the value for the louse is the sum of these three processes plus white noise. All three processes are neutral or drift-free. Drifts associated with host size occur in the mean (5.2).

The REML log likelihood achieved by this Brownian modification exceeds that for (5.3) by approximately 57.9 units, and all four fitted coefficients are positive. Although these models are not nested, the difference is huge enough to leave no doubt that the authors' additive block proposal (5.3) is totally inadequate for these data.

The effect of these temporal correlations on the fitted regression coefficients is small but not negligible; their effect on standard errors is an eight-fold increase. The fitted slope coefficients ($\times 10^4$) for the two pigeon breeds are

| Parameter | Estimate | s.e. |
|------------|----------|------|
| Feral:time | -4.22 | 4.1 |
| Giant:time | 0.33 | 4.1 |
| Difference | 4.55 | 3.3 |

The conclusion from this analysis is the essence of simplicity: the data are entirely consistent with neutral evolution of louse size on both hosts. Not only is there no evidence of a differential drift in louse size for the two hosts, there is no evidence of a drift for either host.

Apart from the Brownian contribution, Table 5.2 shows that the baseline variance is substantially larger than the residual variance on subsequent occasions. This observation suggests that (5.4) is not adequate on its own, and must be supplemented by an additional diagonal matrix for baseline observations. This differential baseline variance leads to a further 64.7-unit increase in the REML criterion. However its effect on conclusions is almost negligible; for comparison, the fitted coefficients ($\times 10^4$) are as follows:

| Parameter | Estimate | s.e. |
|------------|----------|------|
| Feral:time | -4.39 | 4.3 |
| Giant:time | 0.30 | 4.1 |
| Difference | 4.69 | 3.6 |

The conclusion regarding neutrality of evolution is unaffected. The apparent evidence for a differential trend in the analysis at the end of Sect. 5.2.4 is a consequence of a demonstrably inadequate variance assumption.

Brownian motion in (5.4) does a reasonable job of describing the temporal dependence, but the fit can be improved by using a low-index fractional Brownian motion. However, this and other modifications discussed in Sect. 5.4 and Exercises 5.24–5.25 have only a small effect on drift estimates.

5.3 Critique of Published Claims

Villa et al. base their conclusions on the first principal component as a combined measure of overall louse size. Since the first principal component is essentially the standardized sum or average of the three size variables, this much is fine. The sample averages for each host are plotted in the fourth panel of Fig. 5.1, which shows that the divergence between the two mean trajectories is not appreciably greater than the temporal variability of any single trajectory. This is a disappointing conclusion for a four-year experiment, and not appealing as a headline story.

However, Villa et al. choose to emphasize the divergence over the variability by plotting the PC1 mean difference (giant runts minus controls) as a function of time in their Fig. 1C. A version of their plot is shown in Fig. 5.2, and is to be contrasted with the fourth panel of Fig. 5.1.

The plot symbol on the horizontal line in Fig. 1C or Fig. 5.2 is explicitly associated with controls. Error bars attached to zero are not mentioned in captions or in text. The visual impression of remarkable temporal stability of louse size on feral pigeons contrasts starkly with the rapid increase for lineages on giant runts. The plot title and the scale on the y-axis confirm those impressions, which are in line with the authors' conclusion *Lineages of lice transferred to different sized pigeons rapidly evolved differences in size*. In my opinion, Fig. 1C or Fig. 5.2 gives a grossly misleading impression of stability for feral pigeons contrasted with a substantial trend for giant runts. In fact, Table 5.1 shows that louse body-size changes are no more than 2% over the entire period.

Taking correlations into account, the error bars for the non-zero line in Fig. 1C or Fig. 5.2 are too small by a factor increasing from about 1.0 to 7.0, and roughly proportional to time.

Tables S2–S5 in the Appendix to their paper report regression coefficients and their standard errors for the full factorial model with (5.3) as the covariance structure. These tables are cited in the *Results and Discussion* section to support the chief claim: *Over the course of 4 y, lice on giant runts increased in size, relative to lice on feral pigeon controls (Fig. 1C and SI Appendix, Tables S2–S5)*. It is unclear which coefficients are meant to justify this claim, but the coefficient of *host:time* in the PC1 analysis is reported with a *t*-ratio of 3.15. Overlooked in this computational blizzard is the fact that both the fitted mean and the fitted covariance contradict the randomization. In addition, the covariance assumption is non-standard for an evolutionary process, and is demonstrably inadequate for the task.

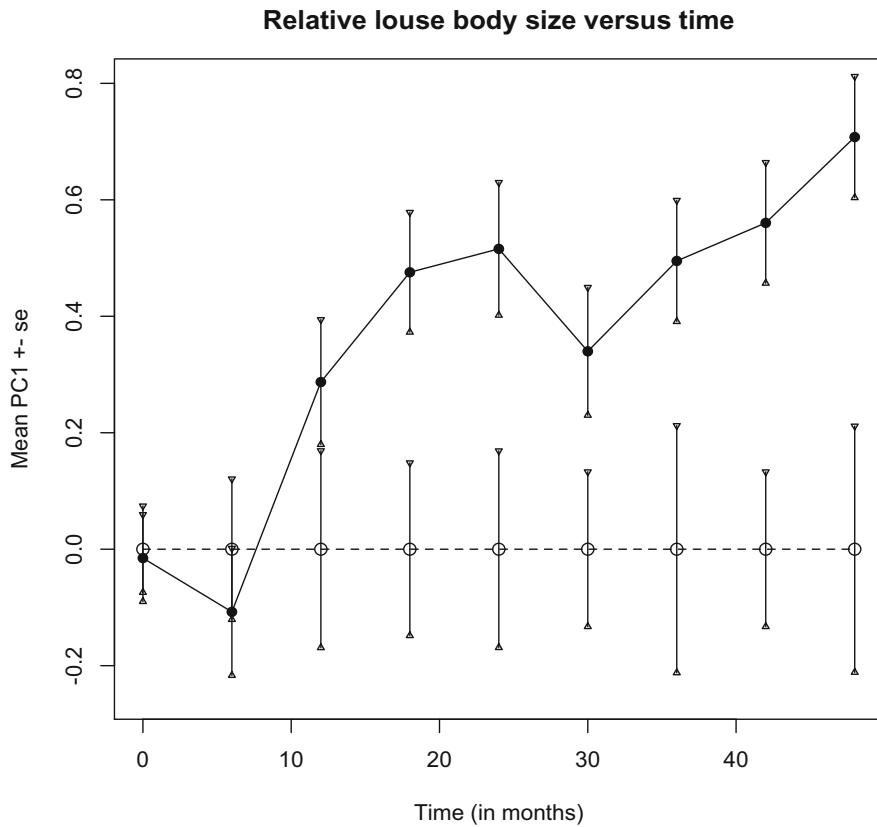


Fig. 5.2 PC1 mean difference ‘giant runt–feral’ versus time

The formal analysis of the first principal component by linear Gaussian models follows the lines of Sect. 5.2.5. Although the scale of the PC1-response is very different from that of the body length, the need for the Brownian-motion component is abundantly clear, as is the additional baseline variance. When these covariances are accommodated, the slope estimates and their standard errors are

| Parameter | Estimate | s.e. |
|------------|----------|-------|
| Feral:time | -0.0126 | 0.033 |
| Giant:time | 0.0016 | 0.033 |
| Difference | 0.0142 | 0.013 |

Nothing in this PC1 analysis points to a departure from neutral evolution of lice on either host. In conclusion, the evolutionary divergence described by Villa et al. may well exist on some time scale, but the evidence for it is not to be found in their data.

5.4 Further Remarks

5.4.1 Role of Louse Sex

The variables *host* and *lineage* are treatment factors generated immediately post-baseline by randomization, and having a known distribution. For the 800 lineage founders, louse sex is a pre-baseline variable; for the remaining lice, sex is a random variable not generated by randomization, and not recorded immediately post-baseline. One can speculate on the joint distribution, but in principle, the sex ratio for giant runts might not be the same as the sex ratio for controls. Thus, (5.1) and (5.2) are models for the conditional mean while (5.3) and (5.4) are models for the conditional covariance—given *host* and *lineage* plus the entire sex-configuration for all sampled lice.

Regardless of covariance assumptions, the interpretation in (5.2) of β_h as ‘the effect of treatment’ must be considered in the light of the fact that any additive effect possibly attributable to an effect of treatment on sex has been eliminated. Although not intermediate in the temporal sense, sex is not dissimilar mathematically to an intermediate response. It is possible that treatment could have an effect on the intermediate response, in which case the coefficients β_h in the conditional mean describe only one part of the treatment effect.

In the context of this experiment, no effect of treatment on sex is anticipated. Any effect that might be present is most likely to be a sampling artifact of little or no evolutionary interest. Nonetheless, it is not difficult to examine the sex distribution at baseline and post-baseline for both treatment groups. Table 5.3 shows the louse counts by time, host and sex.

The post-baseline total count is quite constant at 200 for giant runts, but is much more variable for captive feral pigeons. The first is presumably a design target. We are left to wonder why the control group does not have a similar target. Nevertheless, this is not a serious criticism. In both treatment groups, females account for 58% of lice at baseline, but close to 50% thereafter. As anticipated, there is little evidence of a difference in sex ratio between groups. If anything, the difference between the ratios is below expectation at nearly every time point.

The Poisson log-linear model *time:(host+sex)* is equivalent to the statement that *host* and *sex* are independent at each time point, or equivalently, that the sex ratio is the same for both pigeon breeds, but not necessarily 50:50. The residual deviance

Table 5.3 Louse counts by host, sex and time

| Host | Sex | Time in months | | | | | | | | |
|------------|-----|----------------|-----|-----|-----|-----|-----|-----|-----|-----|
| | | 0 | 6 | 12 | 18 | 24 | 30 | 36 | 42 | 48 |
| Feral | F | 231 | 67 | 39 | 57 | 38 | 56 | 23 | 55 | 19 |
| | M | 169 | 73 | 44 | 50 | 37 | 55 | 31 | 49 | 22 |
| Giant runt | F | 233 | 95 | 104 | 105 | 102 | 104 | 105 | 105 | 96 |
| | M | 167 | 102 | 95 | 92 | 98 | 97 | 91 | 95 | 104 |

of 2.8 on nine degrees of freedom falls at the lower third percentile (0.03) of the null distribution, which shows that sample log odds ratios are uniformly closer to constant than the Poisson model predicts. Certainly, there is no suggestion of a treatment effect on sex ratios. Apart from the imbalance at baseline, the subsequent ratios are close to 50:50, so we can regard the sex indicator post-baseline as a Bernoulli process independent of treatment.

5.4.2 Persistence of Initial Patterns

One unintended consequence of the Brownian covariance model (5.4) is that baseline values are independent of all subsequent values. This is a strong assumption. It is not implied by randomization, and it is not necessarily a property that we could confidently expect to be supported by detailed examination of the data. Without contradicting the randomization, it is possible to introduce temporal correlations between baseline and non-baseline values by a simple modification such as replacing the last term in (5.4) with the shifted Brownian covariance $\min(t - \tau, t' - \tau)$ for some $\tau \leq 0$. For reasons that are explained in Chap. 18, the REML criterion is independent of τ , so this particular modification has no effect on fitted values, on prediction or inference for contrasts. In fact, this covariance term could be replaced with the stationary version $-|t - t'|/2$.

The analysis of variance for baseline values already casts doubt on the fairness of the randomization with respect to aviaries, so it is natural to check for correlations between initial and subsequent values associated with the same aviary. Does the pattern of louse size differences among aviaries at baseline persist in subsequent generations? The question is concerned with persistence of aviary patterns, so fairness of the randomization is not presumed.

One way to introduce persistent initial patterns is to replace the aviary term in (5.4) with independent shifted Brownian motions, one per aviary. The covariance contribution is then

$$\sigma_2^2 \delta_{a,a'} \min(t - \tau, t' - \tau),$$

with a single temporal shift $\tau \leq 0$ to be estimated from the data using the REML criterion. One boundary point $\tau = 0$ coincides with (5.4), and the other limit $\tau \rightarrow -\infty$ implies a constant aviary effect as in (5.3). For $\tau < 0$, this modification implies positive correlations within aviaries at baseline, which is a size pattern that contradicts our understanding of randomization. The interpretation is that, by accident or by design, some aviaries start out with larger lice than others, and the initial pattern leaves an imprint on the subsequent evolution.

For the PC1 variable, the profile REML log likelihood values for τ at zero, $\hat{\tau} = -2.6$ and $-\infty$ are 0.0, 11.5 and -18.5 , showing that the constant aviary effect is decisively rejected by the data. It appears from this analysis that the initial aviary pattern for PC1 is non-zero and that it persists in the subsequent evolution. The

particular temporal offset may be pure coincidence, but $\hat{\tau} = -2.6$ months is a very close approximation to the de-lousing quarantine period during which the pigeons had to be housed somewhere.

5.4.3 Observational Units

Consider the statement near the beginning of Sect. 5.1.3: ‘since each measurement is made on one louse, it is evident that each observational unit is one louse...’. The premiss—that each measurement is made on one louse—is indisputable. Nevertheless, a conclusion that is obvious literally, is not necessarily true mathematically in the sense of the definition.

According to the definition, the observational units are the objects, or points in the domain, on which the response is defined as a random function or a stochastic process. Thus, each observational unit exists at baseline, not necessarily as a physical object, but as a non-random mathematical entity. For the models in Sect. 5.2, with louse-time pairs as observational units, there is no birth or death, and no evolving finite population—only a fixed, arbitrarily large, set of lice in each lineage. In this mathematical framework, the lice are in 1–1 correspondence with the natural numbers, they live indefinitely in the product space, and their vital statistics are random variables recorded in the state space. To each louse there corresponds a stochastic process, so the value for each louse evolves over time, but the population itself is fixed and arbitrarily large in every lineage.

It would be wrong to say that the Gaussian model is incorrect or that its flaws are fatal, but its shortcomings for this application are clear enough. If the application calls for a finite randomly-evolving lineage, a more complicated mathematical structure is required. The remarkable thing is not that this Gaussian model is exquisitely tailored to this evolutionary process, but that a generic model that is missing the defining aspects of life, namely birth and death, should have anything useful to contribute at all.

Certainly, the lice do not exist in the physical sense at baseline. But lineages are established at baseline, and it is the lineages that evolve. They evolve randomly in two senses—in their composition as a finite set of lice, and in their values or features. If both aspects are important for a given application, a more complicated model is needed in which the observational units are lineage-time pairs. The state space for one measurement on one louse is $\mathcal{S} = \{M, F\} \times \mathbb{R}^3$; the state space for one observational unit is the set of finite subsets of \mathcal{S} . One finite subset of \mathcal{S} is a complete description of the population size and the vital statistics of the residents at time t . The transitions from one finite subset to another are limited by birth, death and continuity in time.

A general process of the type described in the preceding paragraph is a complicated mathematical structure, and we make no effort to develop a general theory here. But there are simple versions that are essentially equivalent to imposing a pure birth-death process independently as a cohort restriction on the domain of a

Gaussian process. The distribution of the values thus generated coincides with the Gaussian model in Sect. 5.2, and none of the subsequent analyses are affected. For that setting, birth and death are immaterial.

The possibility that individual louse values or body-sizes might be related to the sample size or lineage size from which they come has not been considered up to this point, in part because such a dependence is not possible under the models in Sect. 5.2. The notion that a sample can be extended indefinitely from a sub-sample such that the sub-sample values remain unchanged, is usually understood in applied work as an obvious fact. Failure strikes at the heart of the most cherished notion in probability and applied statistics, which is the ‘obvious fact’ of distributional consistency for sub-samples as formulated by Kolmogorov (1933). If lice are the observational units for this process, consistency implies that the distribution for individuals is unrelated to the size of the sample from which they are taken. Fortunately, variability of sample sizes provides a weak check to test that implication.

Each of the 32×9 lineage-time pairs provides one sample, of which 15 are empty. The louse counts range from zero to 44, they are highly variable, and they tend to decrease over time. One lineage appears to go extinct at 30 months. The safest and the simplest way to test for a dependence on sample size is to include sample size as an additional ‘covariate’ in (5.2), retaining (5.4) for covariances. For both log body length and PC1, the fitted coefficient is negative and approximately one half of the standard error. This analysis offers no evidence of a sample-size dependence, which provides a little reassurance that the earlier analysis with louse as the observational unit is reasonably sound.

If some birds preened more vigorously or more thoroughly than others, and larger or older lice were preferentially removed by preening, the more assiduous preeners would then host fewer and smaller lice. Differential preening could lead to a dependence of mean louse size on lineage size or on sample size, in which case the test in the preceding paragraph is a reasonable check.

5.5 Follow-Up

5.5.1 New Design Information

Given the severity of the discrepancy between the conclusions presented above and those published by Villa et al. (2019), it seemed only appropriate to send a copy of Sects. 5.1–5.4 to the authors for comment. I contacted the lead author in early December 2020. Scott Villa, responded immediately, and later at the beginning of February 2021 offering further details about the experimental design, and challenging the conclusions on several points.

By Villa’s account, the randomization was carried out according to an elaborate protocol, which involved dislodging the CO₂-anesthetized lice over a custom-made

10×14 glass grid, generating a random grid number as the starting point for collection of specimens, and placing lice sequentially and cyclically in vials labelled 1–32 until each vial contained 25+ lice. It was designed to avoid unintentional biases, and it appeared to be adequate for the task.

The following summary of key design points that had previously been partially or totally misunderstood is taken from Villa's reply.

1. At time zero, 1600+ lice were collected from wild feral pigeons. No size measurements were made on the sub-sample of 800 founder lice that were transferred to captive birds. A second sample of 800 lice was photographed, measured, and frozen for subsequent genetic analysis.
2. The 800 founder lice were assigned to hosts at random, 25 per bird. Each founding population consisted of 13–14 females and 11–12 males with a deliberate female bias to ensure that a lineage would be established on every host.
3. The 800 lice measured at time zero did not contribute to the breeding population; their assignment to lineages was randomized, but purely virtual. The virtual sample had the same sex-ratio as the founders.
4. After baseline, the lice that were measured at 6-month intervals were frozen thereafter to use for genomic analyses of the populations over time. Throughout the experiment (months 6–48), the adult and immature lice that were removed but not photographed were immediately placed back on birds, thus ensuring stability of the lineages over time.

In light of the revised information, certain statements in the 'Materials and Methods' section of the published paper seem ambiguous or oddly phrased, for example,

We transferred 800 lice from wild caught feral pigeons to 16 giant runt pigeons and 16 feral pigeon controls (25 lice per bird). At this time (Time 0), we also randomly sampled 800 lice from the source population on wild caught feral pigeons and measured their body size.

This remark suggests, correctly as it turns out, that the measured lice and the founder lice might be disjoint subsets. But that thought was dispelled by an earlier remark

Once photographed, the live lice were returned to their respective host,

which now turns out to be incorrect.

To learn about a natural host-parasite system, the scientist must manipulate the system to some extent. But as the degree of interference increases, the more is learned about the interference and the less about the natural system. The strong approving remark in the second paragraph of Sect. 5.1.2 about the necessity of returning all lice to their host seems entirely correct as a matter of principle, if only to reduce interference and to minimize the possibility of lineage extinction. Regrettably, it seems now that photographed lice were not returned, perhaps because photography is damaging or destructive. Whether that degree of interference is acceptable or excessive is a matter of biological judgement best left to subject-

matter experts, not a matter on which statistical expertise carries weight. As always, the over-riding concern is that the experiment be reported as it was conducted.

5.5.2 *Modifications to Analyses*

At this point we accept the new design information, and ask what effect it has on the appropriateness of the analyses already performed, and what modifications are required.

Consider first the information that the association of time-zero measurements with lineages is virtual. This fact implies that the information content is unchanged if time-zero values are permuted in any manner that preserves sexes, while non-baseline values stay put. A baseline permutation that preserves sexes is one in which males are permuted with other males, females with other females, and non-baseline individuals are fixed. This set of permutations is a sub-group of size $464! \times 336!$ in the larger group of size 3105!.

Any credible analysis that accommodates the virtual randomization must be invariant with respect to this group of permutations; similar remarks apply to numerical conclusions regarding temporal trends, variance components or other effects. The authors' block-factor assumption (5.3) applies to baseline and non-baseline values, so it contradicts baseline exchangeability, virtual or otherwise. The numerical values reported in their supplementary tables S1–S5 are also not invariant.

Non-virtual baseline exchangeability as discussed in Sect. 5.2.5 implies that the marginal distribution of the initial 800 measurements is invariant with respect to sex-preserving permutation. Virtual exchangeability is a much stronger condition because it implies also that the joint distribution of all 3105 measurements is invariant. Neither condition implies independence of initial and subsequent values, but virtual exchangeability implies that the dependence must be of a trivial type, which is ignorable in practice. The Brownian model (5.4) implies $\text{cov}(Y_u, Y_{u'}) = 0$ for any pair $u \neq u'$ such that $t(u) = 0$ or $t(u') = 0$. Together with (5.2), it also satisfies the virtual exchangeability condition. By contrast, the standard random-effects model (5.3) having independent and identically distributed lineage effects that are constant in time, does not satisfy even the weaker exchangeability condition. It is also incompatible with the discussion in Sect. 5.4.2.

The Brownian-motion model is in line with the standard genetic theory for trait evolution, and is compatible with virtual randomization as described above. Thus the conclusions as stated at the end of Sect. 5.3 are confirmed. Average size differences between the two hosts shown in Table 5.1 are less than 2% and are compatible with neutral evolution in both hosts. The sex-adjusted PC1 mean differences GR – F at each non-zero time point are very similar to the unadjusted differences displayed in Fig. 5.2, but the correctly-computed standard errors in Table 5.4 tell a different story.

Both the differences and the standard errors in this table are computed from a fitted Gaussian model, in which the temporal trend, previously modelled as a zero-

Table 5.4 PC1 mean differences Giant Runt–Feral by time

| | Time in months | | | | | | | | |
|-------|----------------|--------|-------|-------|-------|-------|-------|-------|-------|
| | 0 | 6 | 12 | 18 | 24 | 30 | 36 | 42 | 48 |
| Diff | 0.000 | -0.114 | 0.111 | 0.464 | 0.496 | 0.316 | 0.251 | 0.494 | 0.750 |
| s.e. | 0.00 | 0.24 | 0.33 | 0.40 | 0.46 | 0.51 | 0.56 | 0.60 | 0.65 |
| Ratio | 0.00 | -0.48 | 0.34 | 1.16 | 1.09 | 0.62 | 0.45 | 0.82 | 1.16 |

mean random effect with covariance $\sigma_3^2(t \wedge t')$ in (5.4), is replaced with a non-random term in the mean. The moments are

$$E(Y_u) = \beta_0 + \beta_1 s(u) + \gamma_h(t), \quad (5.5)$$

$$\text{cov}(Y_u, Y_{u'}) = \sigma_0^2 \delta_{u,u'} + \sigma_1^2 \delta_{l,l'}(t \wedge t') + \sigma_2^2 \delta_{a,a'}(t \wedge t') + \sigma_3^2 \delta_{uu'} I_{t=0}. \quad (5.6)$$

The mean subspace includes an additive constant for sex, and a host-dependent temporal trend $\gamma_h(t)$. The factorial model formula

```
sex + as.factor(time) : host
```

generates a subspace of dimension $1 + 9 \times 2 = 19$, but the randomization constraint implies $\gamma_0(0) = \gamma_1(0)$, which reduces the dimension by one. The fitted differences $\hat{\gamma}_1(t) - \hat{\gamma}_0(t)$ are shown in the table, together with standard errors as estimated by REML and weighted least squares. They are automatically sex-adjusted, so they are not exactly the same as the sample differences shown in Fig. 5.2.

If the randomization constraint is ignored, the fitted difference is non-zero for $t = 0$. All estimates and standard errors throughout the table are altered, but only slightly.

At no time does the observed difference reach much above one standard error, so claims for rapid divergence are not supported by this analysis or by any modifications that include non-trivial temporal dependence: see Exercise 5.24. The same applies to the overall estimate of linear temporal trend, which is 0.0142 per month with standard error 0.013. Comparable analyses for body length and other size measurements point to similar conclusions.

It is possible to satisfy the randomization constraint by restricting the block factor terms in (5.3) to post-baseline times only. But Brownian motion implies a non-trivial temporal correlation, and is a much better fit than the restricted block factor. The implication is that the evidence for non-trivial temporal correlation is very strong (see Exercises 5.24 and 5.25). Every modified analysis that takes account of such correlations leads to very similar conclusions.

This analysis does not imply that divergent evolution does not exist on some time scale. But it is safe to say that no evidence for it exists in these data.

5.5.3 Further Remarks

According to the reply by Scott Villa, the sex ratio of lice at baseline was intentionally biased towards females, with 13–14 females and 11–12 males as founders for each lineage. Following the initial seeding, male and female lice were sampled in approximately equal numbers, so information on the evolution of the sex ratio over time is not available. In light of this information, much of the speculation in Sect. 5.4.1 is no longer relevant.

Villa also takes issue with a remark in Sect. 5.2.1 that the overall change in body size is surprisingly small, which suggests that changes of this magnitude (< 2%) cannot be biologically significant. His counter-claim is that

... body size changes on this scale are biologically relevant for this species, as the effect on mating behavior shows (Villa et al., 2019, Figs. 2–5).

The coefficient of variation of body length for female lice within aviaries is very stable at 2.4–2.6% from six months onwards; the value for males is equally stable at 2.2–2.4%. These numbers represent natural variability of body length within freely breeding populations, which is approximately 2.4% (or $\hat{\sigma}_{\text{resid}}^2 \simeq 60 \times 10^{-5}$ in Table 5.2). The mean differences between hosts are shown in Table 5.1; they are almost uniformly less than 2%.

What are the implications for mating? The root mean square size discrepancy between a random pair from the same aviary is approximately $\sqrt{2 \times 2.4^2}$, or 3.4%, so the distribution of $F - M$ -size differences is approximately $N(411, 83^2)$. A 2% increase in mean size for females implies that the distribution of size differences for mixed hosts is $N(411 + 50, 83^2)$. If size discrepancy is the chief determinant of sexual compatibility, and incompatibility is rare in each population, a mean difference of 0.6 standard deviations is not sufficient to make the incompatible fraction large in the mixed population.

The two movies provided by Villa et al. (2019) illustrate size discrepancies of 1.8 and –2.6 standard deviations, so their relevance at the 0.6σ -scale is not immediately apparent. In the absence of a detailed morphological explanation, it is difficult to accept the authors' claim that body size changes on this scale ($\sim 0.6\sigma$) are biologically important for any species.

5.6 Exercises

5.1 According to the standard definition in Sect. 11.4.2, two observational units u, u' belong to the same experimental unit if the treatment assignment probabilities given the baseline configuration satisfy $P(T_u = T_{u'}) = 1$. Section 5.1.3 makes the argument that each louse is one observational unit, and that each lineage is one experimental unit. But the author subsequently pivots to *aviary* as the experimental unit, hedging his bets by stating that ‘both seem to be relevant’. Discuss the arguments pro and con of *louse-lineage* versus *louse-aviary* versus *lineage-aviary*

as the observational-experimental units. In connection with the models in Sect. 5.2, what are the substantive implications of one choice versus another?

5.2 According to Villa *et al.*,

Pigeons combat feather lice by removing them with their beaks during regular bouts of preening. Columbicola columbae, a parasite of feral pigeons, avoids preening by hiding in spaces between adjacent feather barbs; preening selects for C. columbae small enough to fit between the barbs. Preening also exerts selection on traits critical for locomotion on the host.

In light of this information, comment on the remark in Sect. 5.1.3 ... size-biased sampling need not be a serious concern for this experiment provided that it affects all birds equally.

5.3 Download the data, compute the averages at each time point for the two pigeon breeds, and reconstruct the plots in Figs. 5.1 and 5.2.

5.4 The coefficient of variation is the standard-deviation-to-mean ratio, which is often reported as a percentage. For *body length* or other size variables, the coefficient of variation is essentially the same as the standard deviation of the log-transformed variable. Compute the coefficient of variation of *body length* separately for male and female lice on each occasion, and report this as a table of percentages. What patterns do you see in this table for males versus females or baseline versus non-baseline?

5.5 Use `anova` (...) to re-compute the mean squares in Table 5.2. Use Bartlett's statistic (Exercise 18.9) to test the hypothesis that the residual mean squares have the same expected value at all time points. What assumptions are needed to justify the null distribution?

5.6 For the model (5.3), what is the expected value of the within-lineage mean square at time t ? For the Brownian-motion model (5.4), show that the variance of Y_u increases linearly with time. What is the expected value of the within-lineage mean square?

5.7 Use `lmer` (...) to fit the variance-components model (5.3) to the log body length with (5.2) as the mean-value subspace. Report the two slopes, the slope difference, and the three standard errors.

5.8 Explain why (5.3) is in conflict with randomization.

5.9 Compute the four covariance matrices V_0, \dots, V_3 that occur in (5.4). Let Q be the ordinary least-squares projection with kernel (5.2). Compute the four quadratic forms $Y'Q'V_r QY$ and their expected values as a linear function of the four variance components. Hence or otherwise, obtain initial estimates.

5.10 Use `regress(...)` to compute the REML estimate of the variance components in (5.4). Hence obtain the estimated slopes, their difference, and the standard errors for all three.

5.11 For $n = 100$ points t_1, \dots, t_n equally spaced in the interval $(0, 48)$, compute the matrix

$$\Sigma_{ij} = \delta_{ij} + \theta(t_i \wedge t_j)$$

for small values of θ , say $0 \leq \theta \leq 0.02$. Find the maximum-likelihood estimate of β in the linear model $Y \sim N_n(\alpha + \beta t, \Sigma)$ with Σ known, and plot the variance of $\hat{\beta}$ as a function of θ . Comment on the effect of the Brownian-motion component.

5.12 Regress the 32×9 lineage-time averages (for PC1) against sample size using sample size as weights. You should find a statistically significant positive coefficient a little larger than 0.01. Explain why the conclusions from this exercise are so different from those at the end of Sect. 5.4.2.

5.13 In Table S2 of their Appendix, Villa et al. fit the eight-dimensional factorial model *host:sex:time* to the first principal component values on 3096 lice. Show that this is equivalent to fitting four separate linear regressions $E(Y_u) = \alpha + \beta t_u$, with one intercept and one slope for each of the disjoint subgroups, Fer.F, Fer.M, Gr.F, Gr.M. Feral and female are the reference levels, so $sex_u = 1$ is the indicator vector for males. Deduce that the *host:time* coefficient is equal to the slope difference $\beta_{Gr.F} - \beta_{Fer.F}$ restricted to female lice. The fitted value is 0.009. What is the fitted slope difference for male lice?

5.14 The *sex* coefficient in Table S2 is -2.437 . Which combination of the four α -values in the previous exercise does this correspond to?

5.15 The *host* coefficient in Table S2 is 0.449 with standard error 0.159. What does this imply about the average or expected baseline values for the four subgroups?

5.16 For the model with persistent aviary patterns described at the end of Sect. 5.4.2, compute and plot the REML profile log likelihood for τ in the range $0.5 \leq \tau \leq 24$. Use PC1 as the response, and (5.2) for the mean-value subspace. The covariance should be a linear combination of five matrices, one each for the identity matrix and the identity restricted to baseline, two Brownian-motion product matrices as in (5.4), and one τ -shifted B-M product matrix. Ten to twelve points equally spaced on the log scale should suffice for plotting.

5.17 Use the profile log likelihood plot in the previous exercise to obtain a nominal 95% confidence interval for τ .

5.18 Distributional invariance. Consider a simplified version of the louse model in which there are 16 feral and 16 giant runt pigeons, no sex differences between lice, and no correlations among measurements. Two lice are associated with each bird at baseline, and two at each subsequent time $t = 1, \dots, 7$ for a total of 512 observations. Each louse u is associated with a host type $h(u)$, feral or giant runt, and the joint distribution is Gaussian with moments

$$E(Y_u) = \beta_0 + \beta_{h(u)} t_u; \quad \text{cov}(Y_u, Y_{u'}) = \sigma^2 \delta_{u,u'}.$$

A baseline permutation is a 1–1 mapping $u \mapsto \tau(u)$ such that $t(u) > 0$ implies $\tau(u) = u$. Distributional invariance means that the permuted vector Y^τ with components $Y_u^\tau = Y_{\tau(u)}$ has the same distribution as Y . Show that the joint distribution is invariant with respect to baseline permutations. Note that $h(\tau(u))$ is not necessarily equal to $h(u)$.

5.19 Procedural invariance. Consider a sample of 512 observations generated according to the model in the previous exercise. The estimation procedure is invariant if $\hat{\beta}(Y) = \hat{\beta}(Y^\tau)$ and $\hat{\sigma}(Y) = \hat{\sigma}(Y^\tau)$ for every baseline permutation. Is it necessarily the case that distributional invariance implies procedural invariance? Explain why least-squares and maximum-likelihood are invariant procedures.

5.20 Consider the following statement taken from Sect. 5.5. *Any credible analysis that accommodates the virtual randomization must be invariant with respect to the same group, and similar remarks apply to numerical conclusions regarding temporal trends, variance components or other effects.* Invariance in this setting means that each distribution in the model is exchangeable, or invariant with respect to sex-preserving baseline permutations. This is a demanding standard, and it is possible that subsequent statements in that same section may not live up to it. Show that the model-formula `Host : as.factor(Time)`, which is related to Table 5.4, corresponds to a set of vectors, some of which are not group-invariant. Investigate the implications, particularly for time zero.

5.21 According to the text in Sect. 5.5, *Virtual randomization requires the time-zero average for feral hosts to be the same as that for giant runts, but the temporal trends are otherwise unconstrained.* It appears that the model matrix spanning this subspace is not constructible using factorial model formulae. Explain how to construct the desired matrix including a constant additive sex effect. What is its rank? Fit the model as described in the text following Table 5.4. Include independent Brownian motions for aviaries and lineages, plus an additional baseline error term with independent and identically distributed components.

5.22 Use the fitted model from the previous exercise to compute the linear trend coefficient

$$\frac{\sum t(\hat{y}_1(t) - \hat{y}_0(t))}{\sum t^2}$$

and its standard error. You should find both numbers in the range 0.013–0.015 per month, similar to, but not exactly the same as those reported in the text.

5.23 The model in the previous two exercises has a baseline variance that is larger than the non-baseline residual variance. What is the ratio of fitted variances?

5.24 The fact that measured lice were not returned to their hosts is an interference in the system that may reduce or eliminate temporal correlations that would otherwise be expected. One mathematically viable covariance model that is in line with virtual randomization, replaces each occurrence of $t \wedge t'$ in (5.6) with the rank-one Boolean product matrix $(t > 0)(t' > 0)$, so that the only non-zero temporal correlations are those associated with lineage and aviary as strictly post-baseline block factors. Fit this modified block-factor model to the PC1 response with (5.5) for the mean subspace. Which model fits better? Is the log likelihood difference small or large? An informal comparison suffices at this point.

5.25 Construct two versions of Table 5.4, one based on the modified block-factor model, and one based on the combined variance model that includes both. Comment on any major discrepancy or difference in conclusions based on the various models.

Chapter 6

Time Series I



6.1 A Meteorological Temperature Series

The UK Meteorological Office, maintains the longest continuous instrumental temperature record in the world. According to the MET office website,

These daily and monthly temperatures are representative of a roughly triangular area of the United Kingdom enclosed by Lancashire, London and Bristol. The monthly series, which begins in 1659, is the longest available instrumental record of temperature in the world. The daily mean-temperature series begins in 1772.

We examine here the Central England daily temperature series, from January 1, 1772 to Dec 31, 2019. The series length is 90 580 days over 248 years.

The data in tenths of a degree Celsius can be downloaded from the address

<https://www.metoffice.gov.uk/hadobs/hadcet/cetdl1772on.dat>

For each year the values are arranged in a 31×12 array, one column for each month and one row for each day in standard Gregorian format. Non-existent days are padded with the placeholder ‘value’ –999. For computational purposes, we assume that the data have been rearranged in standard data-frame format with one row for each of $n = 90\,580$ days. Each column is one variable. Apart from `temp` and `day`, it may be convenient to include the first and second-order annual harmonics

$$c(t) = \cos(2\pi t/\tau), s(t) = \sin(2\pi t/\tau); c(2t) = \cos(4\pi t/\tau), s(2t) = \sin(4\pi t/\tau),$$

where t is time measured in days counted from Jan 1, 1772, and $\tau = 365.2425$ is the mean number of days in one Gregorian year.

As is often the case with very extensive data, much can be learned from simple graphs and other summaries without resorting to formal stochastic models. We first examine the nature of the annual seasonal cycle.

6.2 Seasonal Cycles

6.2.1 Means and Variances

The average temperature for each date in the year is computed by associating with each day a calendar date, either the Gregorian calendar date or some version thereof. In the conventional Gregorian system, each date is an integer in the range 0–365, beginning with Jan. 1 coded as zero. February 28 and March 1 are coded as 58 and 60 respectively, whether these are consecutive days or not. For present purposes, it suffices to code day as sequential integers $0:(n - 1)$, where $n = 90\,580$, and to use the mathematical calendar date

```
tau <- 365.2425;      date <- trunc(day %% tau)
```

which is an integer in the range 0–365. Whatever version of the calendar date is used, the average for each date is computed as follows:

```
dailymeantemp <- tapply(temp, date, "mean")
```

Our mathematical dates do not correspond exactly with the Gregorian calendar date, mostly because the leap day is intercalated at the end of December rather than at the end of February. Thus, each calendar date 0–364 occurs 248 times, and these dates are always consecutive days, whereas the leap date occurs only 60 times, so date 0 follows 365 in leap years and 364 in non-leap years. Similar remarks apply to date number 59 (Feb. 29) in the Gregorian system. Wherever the leap date is intercalated, a minor discontinuity may be introduced, as can be seen in the volatility series in Fig. 6.1. If the Gregorian date is used, the discontinuity at Dec. 31/Jan 1 disappears, but does not reappear at Feb. 29.

Neither the mean series nor the volatility series is adequately described by a first-order harmonic function, which is a linear combination of the three basis vectors $\mathbf{1}$, $\cos(t)$, $\sin(t)$, but both are reasonably well described by second-order harmonic functions with two further basis elements $\cos(2t)$, $\sin(2t)$. The fitted harmonics shown in Fig. 6.1 were computed by ordinary least squares,

```
olsfit <- lm(dailymeantemp~c1+s1+c2+s2)
```

which is perfectly adequate for graphical purposes, but technically sub-optimal because of serial correlation. Note that all vectors at this stage, including the harmonic functions, are functions of the date, so each vector has 366 components. The leap date could be given reduced weight in the analysis, but this has not been done here. Alternatively, the leap date could be omitted, with a corresponding modification in the harmonic functions.

Apart from the discontinuity at the leap day (Dec 32), which results in a spike in volatility, there is a curious springtime anomaly of reduced temperature volatility around April 5–9. The depression in volatility is spread over several days, and is evident also in plots using Gregorian dates. Whatever its cause—social, ecclesiastical or meteorological—the volatility plots in Fig. 6.2 show that the phenomenon has persisted for over 200 years.

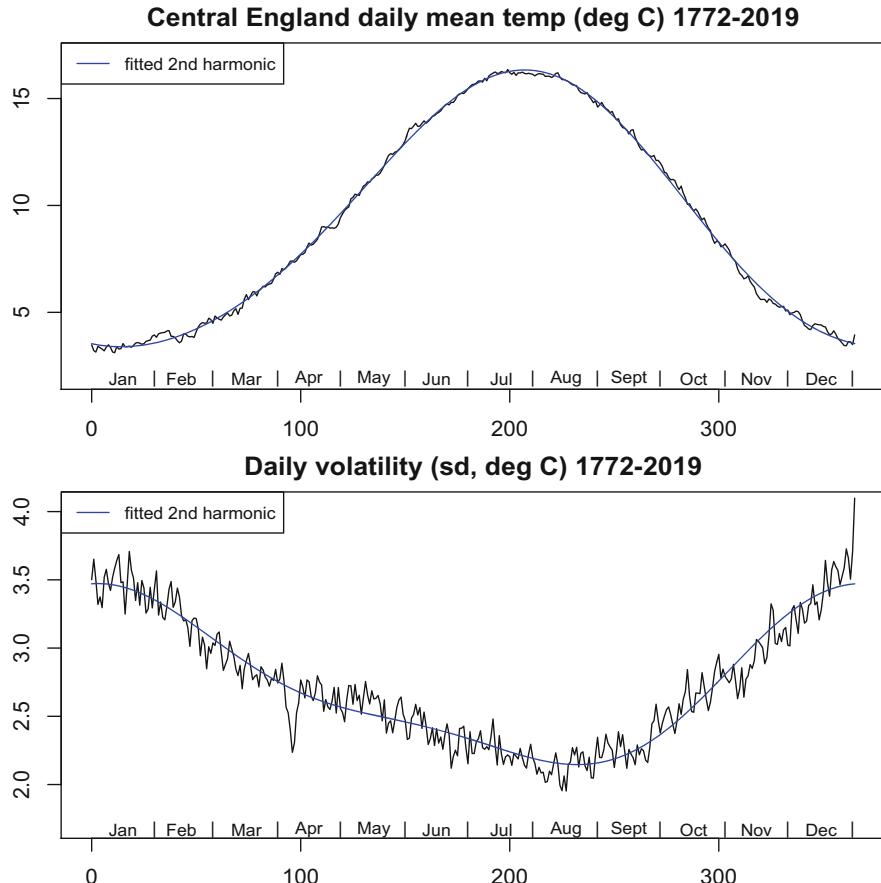


Fig. 6.1 Mean temperature and volatility by day of the year

Figure 6.2 is the same as Fig. 6.1 except that the period has been split into four non-overlapping blocks of 62 years in order that long-term trends and variations in the annual cycle might be revealed. To simplify cross-block comparisons, the plotting scales are fixed for each block, and the second-order harmonic is also fixed to serve as a historical reference.

It is evident that there has been no major shift in the seasonal cycle over this period. However, winter temperatures, particularly in January, have risen by several degrees throughout this period, and that increase began even in the nineteenth century. The low summer and autumn temperatures in the late nineteenth century are well known and are often attributed to volcanic effects such as the Krakatowa eruption in 1883. However, the lowest annual mean in this series occurs in 1879, four years before the eruption, and the expected volcanic effects are not readily apparent in the annual averages for the decade that follows: see Fig. 6.4. Other than

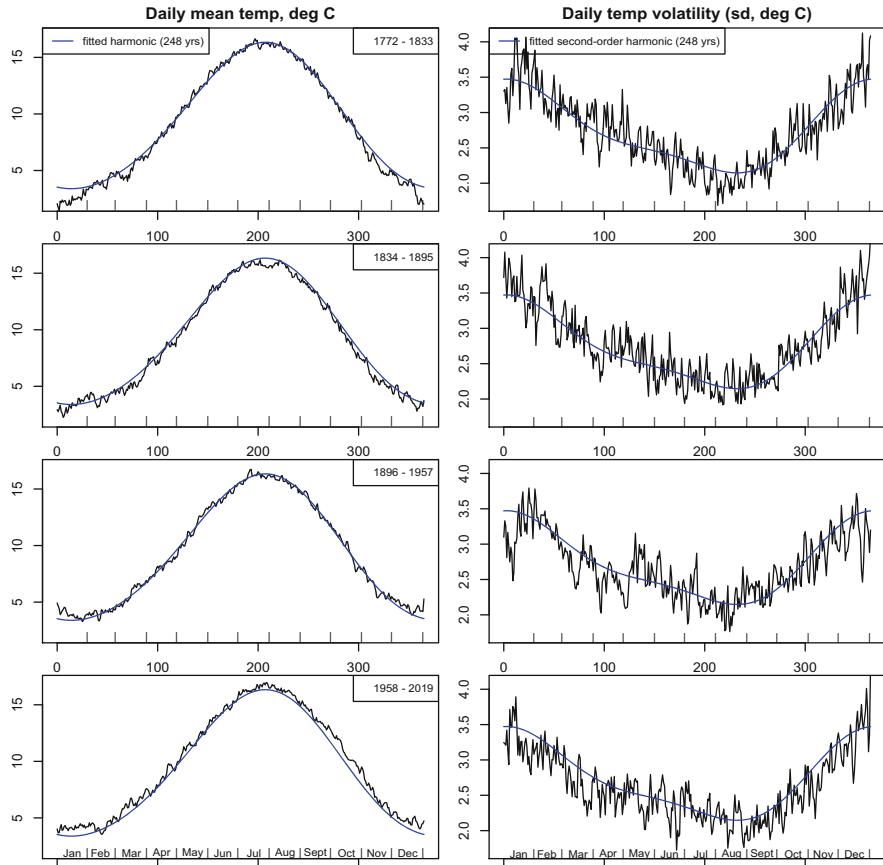


Fig. 6.2 Mean temperature and volatility by day of the year in consecutive 62-year blocks

the winter increase, the annual pattern in the early 20th century is remarkably close to that in the early 19th century. The phenomenon that stands out in Fig. 6.2 is the uniformly high temperature throughout the year in the most recent period. Only on 35 dates do the daily averages for 1958–2019 fall below the historical reference curve.

6.2.2 Skewness and Kurtosis

A k -statistic of degree r is a sequence of homogeneous polynomial symmetric functions $k_{r,n}: \mathbb{R}^n \rightarrow \mathbb{R}$ defined for each $n \geq r$. They were first defined by Fisher (1929) as a way to simplify the study of sample moments.

The first k -statistic is the sample mean. Subsequent k -statistics of order 2–4 are

$$(n - 1) k_{2,n}(x) = \sum (x_i - \bar{x}_n)^2,$$

$$(n - 1)^{\downarrow 2} k_{3,n}(x) = n \sum (x_i - \bar{x}_n)^3,$$

$$(n - 1)^{\downarrow 3} k_{4,n}(x) = n(n + 1) \sum (x_i - \bar{x}_n)^4 - 3(n - 1)^3 k_{2,n}^2(x),$$

where $n^{\downarrow r} = n(n - 1) \cdots (n - r + 1)$ is the descending factorial, and $k_{r,n}$ is defined for $n \geq r$ only. For an independent and identically distributed sample, the expected values are the population cumulants $E(k_{r,n}) = \kappa_r$, which are zero for $r \geq 3$ in Gaussian samples. The third and fourth standardized k -statistics are $k_3/k_2^{3/2}$ and k_4/k_2^2 , which are invariant with respect to affine transformation $x_i \mapsto a + bx_i$ with $b > 0$. Thus, the fact that the temperature is recorded in tenths of a degree °C rather than °F has no effect on the standardized values. These statistics are frequently used to gauge departures from normality. Here we are looking at cumulant variations as a periodic annual time series.

The standardized values are plotted by calendar date in Fig. 6.3, so each skewness and kurtosis coefficient is computed using 248 replicate temperature values for every non-leap date, or 60 for the leap date. The average skewness is close to zero, but there is a distinct sinusoidal cycle with a summer maximum, which is in phase with the mean temperature cycle. Winter temperatures are skewed negatively, summer values positively. The kurtosis values are more widely scattered with no clear pattern, but summer values are slightly larger on average than those in other months. Two thirds of the k_4 -values are negative, indicating that tails are shorter than Gaussian. The sinusoidal trend in the skewness plot is clear evidence of non-normality, but that is not an adequate reason to abandon methods of analysis based on linear decompositions.

It is worthwhile recalling the inheritance property of sample statistics $k_{r,n}$, and more general U -statistics, computed for sub-samples of various sizes. Let $[N]$ be the population and $S \subset [N]$ a sample of size $n \leq N$; let $Y[S]$ be the sample temperatures and $k_{r,n}(Y[S])$ the sample statistic. Given the population statistic $k_{r,N} \equiv k_{r,N}(Y[N])$, the average over samples of size n satisfies

$$\operatorname{ave}_{S \subset [N]} k_{r,n}(Y[S]) = k_{r,N}(Y[N]).$$

Thus, given that the variance for April 7 is low relative to April 1 or April 12 in the population of 248 years, we should expect the same to hold on average for simple random samples or simple random partitions. Although a sequential block of 62 years is not a simple random sample, it may behave as such in the absence of serial correlation, in which case the depression seen for April 5–9 variances in successive 62-year blocks in Fig. 6.2 is expected and not a surprise.

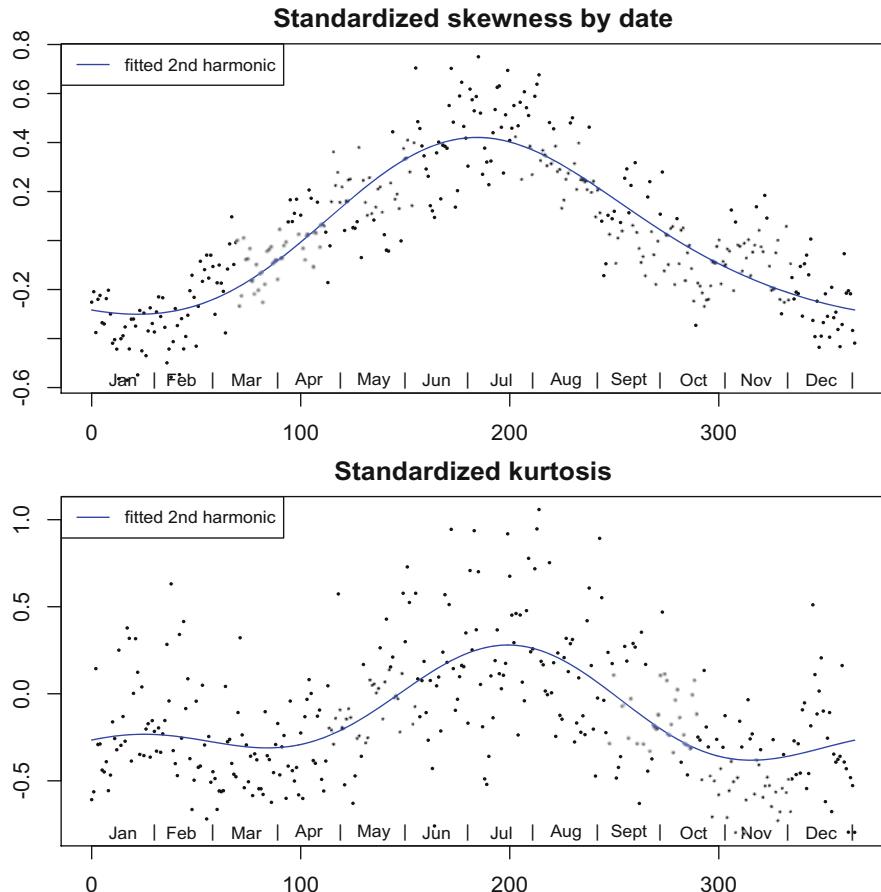


Fig. 6.3 Skewness and kurtosis of temperatures by date

This inheritance argument does not apply to the standardized skewness or standardized kurtosis, which are not U -statistics. Nevertheless, an approximate version of inheritance does hold.

6.3 Annual Statistics

6.3.1 Means and Variances

The top panel of Fig. 6.4 shows the annual average temperature for each year over the 248-year period. Post-1790 record lows and record highs are indicated: year t is a record high if $Y_t = \max\{Y_1, \dots, Y_t\}$, and a record low if $Y_t = \min\{Y_1, \dots, Y_t\}$.

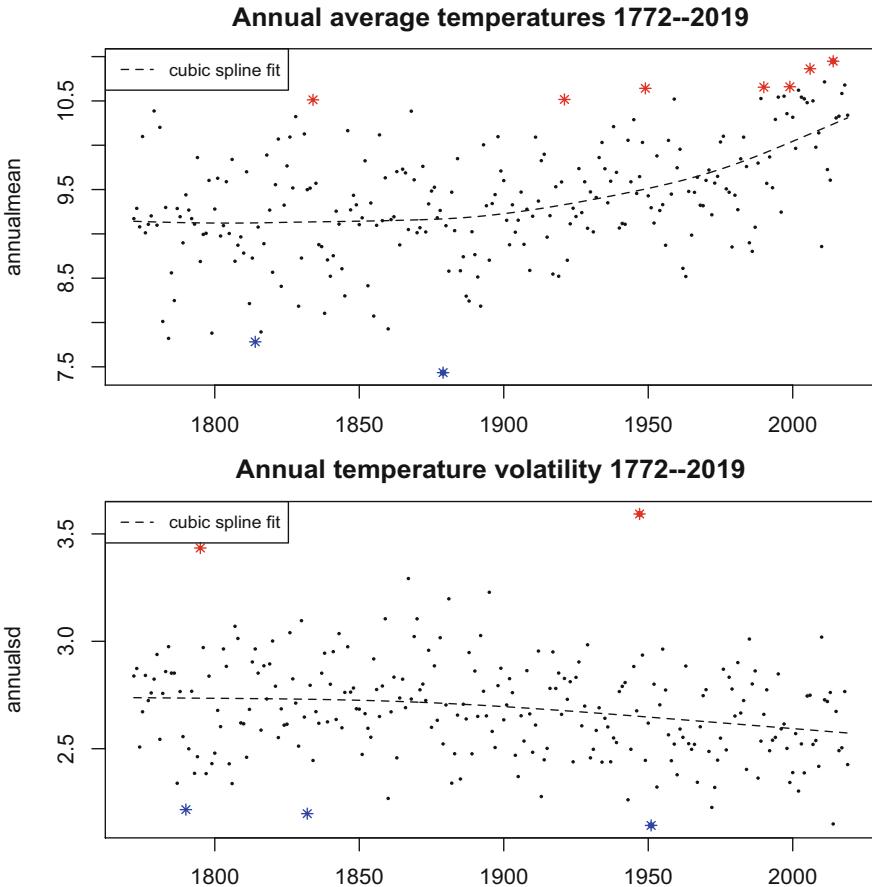


Fig. 6.4 Mean temperature and volatility over 248 consecutive years, with record highs and lows indicated

The record lows occur in 1814 and 1879, the record highs in 1834, 1921, 1949, 1990, 1999, 2006 and 2014. By visual inspection, the mean trend is constant up to about 1900, increasing slowly to about 1950, and more rapidly thereafter. The maximum-likelihood cubic-spline fit has been superimposed as a summary of the mean trend. Computational details are given in the following section.

The second panel of Fig. 6.4 is similar to the first, except that it shows the within-year standard deviation measured as the deviation from the second-order harmonic fit. The harmonic term is removed so that the effect of seasonal variation is kept to a minimum. Post 1790 record lows and highs are highlighted; the lows occur in 1790, 1832 and 1951, the highs in 1795 and 1947. The trend in volatility is downwards as indicated by the cubic spline fit, but it is not significantly non-linear over this period.

Changes in meteorological technology over the centuries must have an effect on variability of measurements, but this effect seems unlikely to be large for temper-

ature measurements. Temperatures are well calibrated relative to the freezing and boiling points of water, so the effects of technological innovation on measurements of annual average temperatures are likely to be small, if not entirely negligible.

6.3.2 Variance of Block Averages

A block of length b is an interval of the form $(t+1, t+b)$, beginning on day $t+1$ and ending on day $t+b$. The focus in this section is on the behaviour of block averages for contiguous blocks of fixed length. To eliminate seasonal variation, we restrict attention to blocks whose length is an integer number of years. In the following table, the sample averages for 5000 blocks of length b years or $365.25b$ days were obtained, and the sample variance of these block averages was computed. Blocks were sampled uniformly at random, not necessarily starting on Jan 1, so the average fractional overlap of two blocks is $b/248$. Standard theory for simple random samples tells us that, in the absence of correlation, the sample variance of the averages is proportional to $(1-f)/b$, where $f = b/248$ is the sampling fraction and $1-f$ is the finite-population correction factor. Accordingly, the second line reports the corrected variance of the block averages, with $C_b = 1/(1-f)$.

| b | Block length in years | | | | | |
|---|-----------------------|-------|-------|-------|-------|-------|
| | 4 | 8 | 16 | 32 | 64 | 128 |
| $C_b \text{ var}(\bar{Y}_b)$ | 0.213 | 0.158 | 0.133 | 0.087 | 0.058 | 0.036 |
| $b \times C_b \text{ var}(\bar{Y}_b)$ | 0.853 | 1.261 | 2.130 | 2.778 | 3.719 | 4.624 |
| $b^{1/2} \times C_b \text{ var}(\bar{Y}_b)$ | 0.427 | 0.446 | 0.533 | 0.491 | 0.465 | 0.409 |

The standard theory for uncorrelated values also applies asymptotically to block averages from a stationary processes provided that the correlations decay at a sufficiently fast rate. For a short-range dependent process the product $bC_b \text{ var}(\bar{Y}_b)$ shown in the middle line should be approximately constant in b , at least for large b . However, this product is clearly increasing as a function of the block size. The third line suggests that $b^{1/2}C_b \text{ var}(\bar{Y}_b)$ is approximately constant, and hence that the variance of block averages behaves inversely as the square root of the block size rather than $O(b^{-1})$. This phenomenon is a characteristic of long-range dependence. For a stationary process, the behaviour observed here for block averages is consistent with the assertion that the covariance function does not have a finite integral. It is incompatible with short-range dependence such as $e^{-|s|}$ or $P(s)e^{-|s|}$ for any polynomial P , or any of the finite-range Matérn models.

It is important to be aware that the variances shown above are finite-population sample variances. Since the totals for a block B of size b and its complement \bar{B} of size $n-b$ satisfy

$$b\bar{Y}_B + (n-b)\bar{Y}_{\bar{B}} = n\bar{Y} = \text{const},$$

the sample variance of fixed-length block totals is exactly equal to the sample variance of the totals on the block complements. In general, the complement of a contiguous block is not a contiguous block. However, if the sampling were done cyclically, i.e., modulo $n = 248$ years, the complement of a contiguous block is also a contiguous block.

6.3.3 Variogram at Short and Long Lags

The variogram of a stationary process at lag h is the expected value of the squared difference $|Y_t - Y_{t+h}|^2$, which is non-negative and symmetric in h . If the process has a covariance function $K(|t - t'|)$, the variogram is

$$\gamma_h = E(|Y_t - Y_{t+h}|^2) = 2K(0) - 2K(h) = 2\sigma^2(1 - \rho(h)),$$

where $\rho(h)$ is the autocorrelation at lag h , and σ^2 is the variance. The semi-variogram is one half the variogram.

For a sequence of length n , the empirical variogram is the average squared difference of sample values

$$\tilde{\gamma}_h = \frac{1}{n-h} \sum_{t=1}^{n-h} (Y_t - Y_{t+h})^2.$$

If the process has a non-constant mean, but is otherwise stationary, the residuals are used instead. The empirical variogram provides a decomposition of the total sum of squares by lags:

$$\frac{1}{n} \sum_{h=1}^n (n-h) \tilde{\gamma}_h = \frac{1}{n} \sum_{s>t} (Y_t - Y_s)^2 = \sum (Y_i - \bar{Y})^2.$$

However, it is not an orthogonal decomposition.

Figure 6.5 shows the log variogram split by short, medium and long lags. All variogram computations are based on residuals after eliminating first and second-order seasonal harmonics. The first panel shows the typical monotone increase for lags 1–30 days; the second panel is re-scaled to show a similar, but less steep, increase for lags of 30–365 days. A particular curve fitted by non-linear least-squares is superimposed on the log variogram, the same curve $-0.02 + \log(1 - K^*(h/10.9))$ in both panels. Details concerning the $SD_{1/2}$ covariance function K^* are given in Sect. 7.2.2; the behaviour for large h in excess of about 5 is $4K^*(h) \simeq 1.25h^{-3/2} - h^{-2}$. Overall, the fitted curve tracks the observed values quite well over lags $1 \leq h \leq 365$ measured in days, but it is essentially constant after

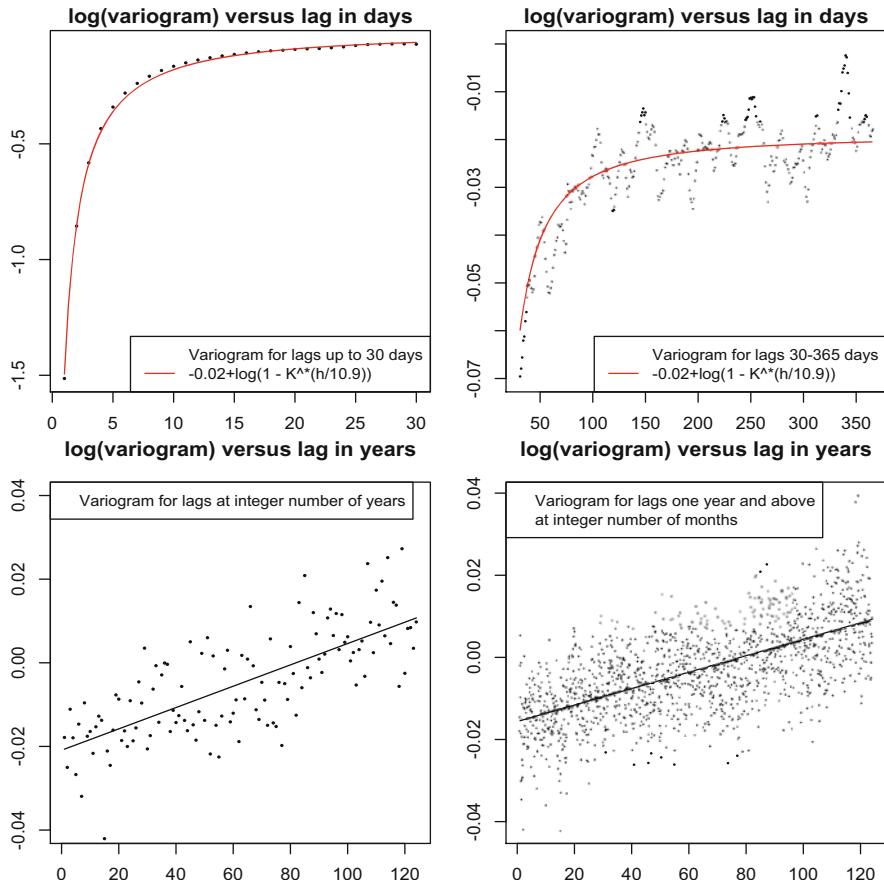


Fig. 6.5 Empirical log variogram of temperatures split into short, medium and long lags. A least-squares fitted curve for short and medium lags taken together is shown in the top two panels. For the longer lags, the least-squares straight line with slope 0.20–0.25 per millennium is shown

about 12–18 months. The standardized variogram and the autocorrelations implied by the fitted curve for lags up to one week are as follows:

| h | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|-------------------------------------|-------|-------|-------|-------|-------|-------|-------|
| $\tilde{\gamma}(h)/(2s^2)$ | 0.220 | 0.425 | 0.559 | 0.648 | 0.711 | 0.755 | 0.788 |
| $\hat{\gamma}(h)/(2\hat{\sigma}^2)$ | 0.224 | 0.441 | 0.572 | 0.658 | 0.718 | 0.761 | 0.794 |
| $\hat{\rho}_h$ | 0.776 | 0.559 | 0.428 | 0.342 | 0.282 | 0.239 | 0.206 |

For lags $2 \leq h \leq 4$, the $SD_{1/2}$ autocorrelations satisfy $\hat{\rho}_h < \hat{\rho}_1^h$, the inequality being reversed for $h > 4$.

The third and fourth panels show the variogram behaviour for very long lags in the range 1–120 years. The third panel is restricted to lags that are an integer multiple of one year, so that the sequence of values is not affected by the elimination of seasonal cycles. This graph indicates that the log variogram increases at the rate 0.25 units per millennium

$$\log \hat{\gamma}(h) \simeq \text{const} + 0.25h/1000$$

over the range $1 \leq h \leq 120$ years. The fourth panel shows lags that are integer multiples of one month over the same range. The least-squares fitted line in this case is a little flatter with slope 0.20 units per millennium. Neither scatterplot suggests a substantial deviation from linearity over the range $1 \leq h \leq 120$ years. The absence of an upper bound for the variogram is one of the hallmarks of non-stationarity.

It is striking that the $\text{SD}_{1/2}$ variogram curve $\gamma(h) = \sigma_0^2 - \sigma_1^2 K^*(h/\lambda)$, which fits the empirical variogram reasonably well for lags up to and well beyond one year, has a finite limit $\gamma(\infty) = \sigma_0^2$, and thus fails completely to capture the non-constant behaviour of the variogram at very long lags. Although the long-range trend is difficult to deny, the implied annual increase is almost imperceptible and is comparable to the width of one plotting symbol in the second panel.

6.4 Stochastic Models for the Seasonal Cycle

6.4.1 Structure of Observational Units

The observational units in a time series are the time points at which measurements are made. Usually, there are no replicate measurements at the same time. In this instance, each day is one observational unit, the observational units are completely ordered and are associated with the integers, i.e., equally spaced points on the real line. In the absence of further structure, we have at our disposal only one fundamental covariate, which is time measured in days beginning at an arbitrary point, which is taken to be zero for Jan 1, 1772. There is, however, one crucial piece of additional information, which is the length of the Gregorian year, $\tau = 365.2425$ days. From this we arrive at the first-order harmonics,

$$c(t) = \cos(2\pi t/\tau), \quad s(t) = \sin(2\pi t/\tau)$$

whose period is one calendar year. The k th-order harmonics $c(kt), s(kt)$ have a period of $1/k$ years.

One crucial property of harmonics is that the subspace spanned by each pair $c(t), s(t)$ is closed with respect to temporal translation:

$$\text{span}\{\cos(t), \sin(t)\} = \text{span}\{\cos(t + h), \sin(t + h)\}$$

for each displacement h , and the same holds for the pair $c(kt), s(kt)$. The space \mathcal{H}_k of harmonics of degree $\leq k$ is a vector space of dimension $2k + 1$, in which $\mathcal{H}_0 = \mathbf{1}$ is the space of constant functions. These are the only plausible functions that are available for use as covariates in the model for the mean temperature.

The Fourier basis vectors $c(kt), s(kt)$ are exactly orthogonal in the continuous setting as functions on $(0, 2\pi)$, and they are exactly orthogonal in certain uniformly-spaced discrete-time settings. In the present discrete setting, they are not quite orthogonal because of the leap-year complication, but this effect is very small and can be neglected.

6.4.2 Seasonal Structure

The fitted second-order harmonic shown in Fig. 6.1 is a reasonably accurate description of the seasonal pattern in mean temperature. Although the deviations from this curve are small, they are far from independent, as the following analysis demonstrates.

The additional structure on observational units comes not in the form of covariates, but in the form of relationships between pairs of observational units (t, t') . The most obvious relationship is the Euclidean metric $|t - t'|$ on the real line, but there are also at least three periodic semi-metrics that are more natural for the description of seasonal cycles:

$$\begin{aligned}\chi(t, t') &= \frac{\tau}{\pi} \sin\left(\frac{\pi|t - t'|}{\tau}\right), \\ \ell(t, t') &= \min\{|t - t'|, \tau - |t - t'|\}, \\ d(t, t') &= (t - t')(\tau - (t - t'))/\tau.\end{aligned}$$

The first two are respectively the chordal distance and the arc length on the annual circle whose perimeter is τ . In all three expressions, t, t' are understood as points in the space $\mathbb{R} \pmod{\tau}$, with addition modulo τ , so that $0 \leq t - t' < \tau$ and $t' - t = \tau - (t - t')$ are complementary arc lengths. With this understanding, it can be verified that d is a metric. For each metric, the maximum values $\tau/\pi, \tau/2$ and $\tau/4$ occur at diametrically opposite points $t - t' = \tau/2$.

For most statistical work, d and χ are essentially equivalent: see Exercises 6.11–6.14.

6.4.3 Stationary Periodic Processes

For each $\lambda > 0$, the function $K(t, t') = \exp(-\chi(t, t')/\lambda)$ is positive definite on $[0, \tau]$, and also stationary and positive semi-definite on the real line with period τ . The Gaussian random function $\eta \sim \text{GP}(0, K)$ is periodic and continuous, and is reasonably well suited as a statistical description of the temperature deviations from the seasonal harmonic in Fig. 6.1. For fixed λ , the Gaussian model in which $\mu = E(Y)$ belongs to the space of harmonics of degree two, and $\text{cov}(Y) = \sigma_0^2 I_n + \sigma_1^2 K$, is linear in the parameters. The model can be fitted using the R command

```
fit <- regress(dailymeantemp~c1+s1+c2+s2, ~K)
```

In this discrete computational setting, $\tau = 366$, or 365 if the leap day is dropped, and K is a symmetric matrix of the same order. The identity matrix, or nugget effect, is included by default, so there are two variance components and five regression coefficients to be estimated. As it happens, the nugget variance estimate is zero, or even slightly negative if not constrained. The maximized log likelihood plotted against λ has a maximum at $\hat{\lambda} \simeq 4.3$ days, and the fitted variance coefficients are $\hat{\sigma}_0^2 = 0$, $\hat{\sigma}_1^2 = 3.62$. The log likelihood is distinctly non-quadratic in λ , but it is approximately quadratic in λ^{-1} with a finite long-range limit as $\lambda \rightarrow \infty$.

The positivity constraint is enforced either through the optional argument `pos=c(1, 1)` or, in this instance, by nugget omission `identity=FALSE`. The residual log likelihood for the covariance model $\sigma_0^2 I_n + \sigma_1^2 K$ exceeds that for the independent and identically distributed sub model with $\sigma_1^2 = 0$ by 167 units, leaving no doubt about strength of the residual serial correlation.

If the arc length is substituted for chordal distance, the resulting process is essentially an autoregressive process of order one, but with a periodic constraint. The dependence is local and confined to a few days, so there is little difference between the chordal and arc-length models.

The quadratic metric is closely related to the Brownian-bridge process: see Exercises 6.15–6.16.

6.5 Estimation of Secular Trend

6.5.1 Gaussian Estimation and Prediction

Suppose that $Y = (Y_0, Y_1)$ is a pair of random vectors that are jointly Gaussian with moments

$$E(Y) = \begin{pmatrix} \mu_0 \\ \mu_1 \end{pmatrix}, \quad \text{cov}(Y) = \begin{pmatrix} \Sigma_{00} & \Sigma_{01} \\ \Sigma_{10} & \Sigma_{11} \end{pmatrix}$$

in partitioned-matrix form. Each marginal distributions is Gaussian $Y_0 \sim N(\mu_0, \Sigma_{00})$, and $Y_1 \sim N(\mu_1, \Sigma_{11})$. The conditional distribution of Y_1 given Y_0 is also Gaussian. The conditional mean is linear in Y_0 and the variance is constant. If Σ_{00} is invertible, the moments are

$$\begin{aligned} E(Y_1 | Y_0) &= \mu_1 + \Sigma_{10}\Sigma_{00}^{-1}(Y_0 - \mu_0), \\ \text{cov}(Y_1 | Y_0) &= \Sigma_{11} - \Sigma_{10}\Sigma_{00}^{-1}\Sigma_{01}. \end{aligned}$$

In statistical applications of this formula for estimation and prediction, Y_0 is the observation vector, and Y_1 is an unobserved random variable—the long-range secular trend. Usually, maximum-likelihood estimates are used for all unknown parameters as needed.

6.5.2 Application to Trend Estimation

The series of annual averages is modelled as the sum $Y(t) = \eta(t) + \varepsilon(t)$ of two independent Gaussian processes in which the components of ε are independent and identically distributed with mean zero. The secular trend is a smooth random function whose covariance function exhibits long-range dependence. Intuitively, a long-range secular trend conjures up an image of a smooth function in time, so the choice for $K \propto \text{cov}(\eta)$ must force a specific degree of smoothness on the function η . Typically, the mean of η is constant or linear $\mu(t) = \beta_0 + \beta_1 t$, but more general expressions are also possible if the circumstances require it. The statistical goal is to estimate the secular trend by computing the conditional expectation $E(\eta(\cdot) | \text{data})$, which is called the Bayes estimator. When maximum-likelihood estimates of the fitted parameters are inserted, this is known as the empirical Bayes estimator. In other types of application, the conditional expected value is sometimes called the *best linear predictor* or the Kriging estimate.

It is worth emphasizing that continuity requires $\eta(\cdot)$ to be defined at all points in \mathbb{R} , even though the process Y is observed or recorded only at a finite set of points. The conditional expected value $E(\eta(s) | \text{data})$ is linear in the observations; its behaviour as a function of s is a linear combination of covariances $K(s, t_i) = \text{cov}(\eta(s), Y(t_i))$ for observation times t_1, \dots, t_n . In practice, the degree of smoothness of K on the diagonal is crucial.

6.5.3 Matérn Models

For the present illustration, we choose the Matérn covariance function with index v

$$K(t, t') = x^v \mathcal{K}_v(x),$$

where K_v is the Bessel function of order $v > 0$, and $x = |t - t'|/\lambda$ is the standardized temporal difference. The exponential covariance function corresponds to $v = 1/2$, and in this case $\lambda \log 2$ is the range at which the serial correlation in η is reduced by half. The index range $v > 0$ guarantees continuity of $\eta(\cdot)$ as a random function, $v > 1$ guarantees continuity of first derivatives, and $v > r$ guarantees continuity of r th derivatives. For computational illustration, we set $v = 3/2$ and $\lambda = 1000$ (in units of years). The large value of λ means not only that serial correlation persists well beyond the observation period but also that the behaviour of η is governed by the behaviour of K near the diagonal.

```
nu <- 3/2; lambda <- 1000; x <- abs(outer(yr, yr, "-"))/lambda;
K <- x^nu * besselK(x, nu); diag(K) <- 2^(nu-1)*gamma(nu)
fit <- regress(annualmean~yr, ~K)
blp <- fit$fitted + fit$sigma[2]*K %*% fit$W %*% (annualmean-fit$fitted)
lines(yr, blp)
```

In the formula for the conditional expectation, `fit$fitted` is the fitted mean vector $\hat{\beta}_0 + \hat{\beta}_1 t$ with 248 components, `fit$W` is the fitted inverse covariance matrix for the observations, `fit$sigma` is the vector of fitted variance components, and `fit$sigma[2]*K` is the matrix of covariances $\text{cov}(\eta(t), Y(t'))$ for t, t' among the observation points. If we wish to make predictions beyond the range of observation times, say to 2020 or 2021, it is necessary to extend the vector of fitted means and the matrix of covariances in the obvious way.

6.5.4 Statistical Tests and Likelihood Ratios

The fitted variance components are $\hat{\sigma}^2 = (0.316, 151.6)$, and the log likelihood ratio statistic for testing the linear sub-model $\sigma_\eta^2 = 0$ against this alternative is $2(16.14 - 8.46) = 15.36$ on one degree of freedom. Note that the null hypothesis sub-model is not stationary, but the mean trend is constrained to be linear in time. Using the standard asymptotic approximation for the distribution of the likelihood-ratio statistic, the tail probability is less than 10^{-4} , so the evidence for non-linearity and/or long-range correlation is fairly strong.

If we wish to test the hypothesis of no trend versus a continuous non-linear trend, we could proceed computationally as follows:

```
nu <- 1/2; lambda <- 1000; x <- abs(outer(yr, yr, "-"))/lambda;
K <- x^nu * besselK(x, nu); diag(K) <- 2^(nu-1)*gamma(nu)
fit0 <- regress(annualmean~1)
fit1 <- regress(annualmean~1, ~K)
2*(fit1$llik - fit0$llik) # LLR=71.68
```

To be clear, the null hypothesis of no trend is interpreted here as independent and identically distributed Gaussian observations, and it is this hypothesis that is decisively rejected by the likelihood ratio statistic of 71.68. However, this interpretation of the null hypothesis is arguably unfair because short-term correlation between

consecutive annual averages seems inevitable, and no trend is not the same as no correlation.

The difficulty here is that it is unclear statistically what is implied by the null-hypothesis phrase ‘no long-term trend’. After all, the Gaussian process with constant mean and covariance $\sigma_0^2 I_n + \sigma_1^2 K$ is temporally stationary. The last snippet of code generates a test statistic that is sensitive to long-range correlation, which is arguably indistinguishable from long-term trend.

6.5.5 Rough Paths Versus Smooth Paths

To appreciate the effect of the Bessel index, it is worthwhile computing and plotting the Bayes estimate for the fitted mode shown above with $\nu = 1/2$. For $\nu = 3/2$ and large λ , the conditional mean is piecewise cubic—a cubic spline which has at least two continuous derivatives at all points. For $\nu = 1/2$ the conditional mean is piecewise linear—a linear spline with a knot at each observation. The linear spline is constrained only by continuity; the cubic spline is constrained by continuity of two derivatives, so it is less flexible and much smoother in appearance. Intermediate values such as $\nu = 1$ have Bayes estimates that are intermediate in appearance.

In the majority of applications of this sort, the likelihood function is close to constant in ν , but also slowly decreasing. In other words, the data are relatively uninformative about smoothness of η , but there is a slight preference for rougher trajectories. For visual extrapolation, however, a smooth curve provides a more compelling image and tells a more convincing story than a rough curve. Continuity of second derivatives seems about right for visual displays, and $\nu = 3/2$ is a reasonable compromise for a graphical summary.

In principle, the range parameter λ can also be estimated by maximum likelihood, but for most practical work the value is effectively infinite, in which case the limit process may be used directly. In both examples, we have used $\lambda = 1000$ for illustration, but the maximum is achieved in the long-range limit. Details of the limit process, also called the cubic spline model, are given in section 16.9. The first snippet of code shown above is satisfactory for $\nu \leq 3/2$, but it is not recommended for $\nu > 3/2$ if λ is large. The second snippet with constant mean is satisfactory for $\nu \leq 1/2$, but is not recommended for $\nu > 1/2$ if λ is large.

6.5.6 Smooth Versus Ultra-Smooth Paths

The Matérn covariance function is convenient in many ways, but it is not essential to the discussion regarding smoothness. An alternative strategy for accommodating intermediate-range dependence is to use an inverse-polynomial covariance function of the form $1/(1 + x^2)$, where $x = |t - t'|/\lambda$. Correlations decay slowly with distance. Each realization of this process is an infinitely differentiable function, so

the conditional expected value $E(\eta \mid \text{data})$ is also a C^∞ -function. Unlike the Matérn process, the long-range limit of the inverse-quadratic process is not well-behaved. Consequently, it is necessary to fix a finite range or to estimate the range, and $\hat{\lambda} \simeq 99$ years is the value suggested by the sequence of annual averages. With this choice, the conditional expected-value curve is not appreciably different in appearance from the cubic spline shown in Fig. 6.4.

The ‘Gaussian’ covariance function $\exp(-|t - t'|^2/\lambda)$ also gives rise to C^∞ trajectories. Usually this choice is not recommended for applied work because the ultra-smooth trajectories give rise to ultra-smooth, non-local, predictions whose apparent accuracy may be misleading. In addition, the long-range limit is not well-behaved as a process, so a finite range is needed for computation.

In general, the behaviour of the covariance function at the origin governs the smoothness of trajectories of the process $\eta \sim \text{GP}(0, K)$. If K is continuous at the origin, η is everywhere continuous with probability one, i.e., $\eta \in C^0$. If K has $2r$ continuous derivatives at the origin, η is r times differentiable everywhere, i.e., $\eta \in C^r$. For $r = \infty$, the situation is a little more delicate.

A covariance function that is infinitely differentiable at the origin has a series expansion whose Taylor coefficients are finite. If the radius of convergence of this series is strictly positive, we say that K is *analytic at the origin*. Otherwise, if the radius of convergence is zero, K is infinitely differentiable but not analytic. Both the inverse quadratic and the Gaussian covariance function are analytic. The trajectories are also analytic, i.e., ultra-smooth, so predictions are non-local.

The SD $_\alpha$ process has a covariance function K_α proportional to the α -stable density on the real line. The characteristic function, or spectral density, is $\exp(-|\omega|^\alpha)$, so $\alpha = 1$ corresponds to the Cauchy covariance $1/(1 + x^2)$, and $\alpha = 2$ corresponds to the Gaussian covariance. Both are analytic. The variogram shown in the top two panels of Fig. 6.5 correspond to $\alpha = 1/2$. The SD-covariance function $K_{1/2}$ is infinitely differentiable at the origin, but it is not analytic. The trajectories are also infinitely differentiable but nowhere analytic. For an illustration, see the fourth panel of Fig. 7.5.

The point of this distinction is that a C^∞ trajectory is much more flexible than an analytic trajectory. The distinction is important both for computation and prediction. Analytic covariance functions give rise to near-singular covariance matrices. In practice, it is best to avoid ultra-smooth processes entirely.

6.6 Exercises

6.1 Let $0 < n \leq N$ be positive integers, let $\varphi: [n] \rightarrow [N]$ be a 1–1 mapping from the set $[n]$ into $[N]$, and let $N^{\downarrow n}$ be the set of such functions. Show that $\#N^{\downarrow n} = N(N - 1) \cdots (N - n + 1)$ so that $\#N^{\downarrow N} = N!$.

6.2 A vector $x \in \mathbb{R}^N$ may be regarded as a function $[N] \rightarrow \mathbb{R}$, in which case the composition $x\varphi$ is a function $[n] \rightarrow \mathbb{R}$ or a vector in \mathbb{R}^n . Show that Fisher’s

k -statistics of degree $r \leq 4$ satisfy the arithmetic identity

$$\frac{1}{\#N \downarrow n} \sum_{\varphi \in N \downarrow n} k_{r,n}(x\varphi) = k_{r,N}(x)$$

for each $x \in \mathbb{R}^N$. Interpret this identity as a reverse martingale with n playing the role of time.

6.3 Given the variance components, the Bayes estimate of the secular trend is a linear combination of the fitted mean vector and the fitted residual

$$\tilde{\mu} = PY + L\Sigma^{-1}QY,$$

where PY and QY are independent Gaussian vectors. Use this representation to approximate $\text{cov}(\tilde{\mu})$.

6.4 The U.K. Met Office maintains a longer record of monthly average and annual average temperatures for Central England from 1659 onwards in the file

<https://www.metoffice.gov.uk/hadobs/hadcet/cetml1659on.dat>

Check the format, download the data, and plot the annual average temperature as a time series. For the annual mean data up to Dec. 31 of the past year, fit the Matérn model with $v = 3/2$ and range $\lambda = 1000$ as described in section 6.5.3, and plot the Bayes estimate of the secular trend. Repeat the calculation for $v = 1$ and range $\lambda = 1000$, and superimpose the two Bayes estimates. Comment briefly on the shape of the fitted curves prior to 1772.

6.5 For the cubic and quadratic models described in the preceding exercise, compute the predicted temperature for next year, i.e., the conditional distribution of the mean temperature for next year given the series of annual averages up to December 31 of the past year. The two models should give slightly different predictive distributions.

6.6 A variety of other smoothing techniques can be employed to illustrate long-term secular trends. Pick your favourite kernel density smoother, apply it to the temperature series, and compare the fitted curve with the Bayes estimates described above.

6.7 Compute the annual average temperatures for the years 1772–2019, and duplicate the first plot in Fig. 2.4. Include the Bayes estimate of the long-term secular trend up to 2025 using the Matérn covariance function with $v = 3/2$ and range $\lambda = 1000$ years. Compute the pointwise standard deviation of the Bayes estimate, and include the 95% prediction interval on your plot.

6.8 The U.K. Met Office site <https://www.metoffice.gov.uk/> keeps long-term weather records—temperature, rainfall, and so on—for a range of stations in Great Britain and Northern Ireland. Monthly rainfall totals for Oxford from 1853 onwards are available in the file

```
/pub/data/weather/uk/climate/stationdata/oxforddata.txt
```

Check the format, download the data, and plot the monthly average rainfall as a seasonal series. Take note of the units of measurement, and include this information on the graph.

6.9 This exercise is concerned with two versions of the Bayes estimate of the seasonal rainfall component, where it is required to compute $E(\eta_m | \text{data})$ for each of 12 months. As usual, χ is the chordal distance as measured on the clock whose perimeter is 12 units, and month is a factor having 12 levels. In computer notation, the code for fitting the two models is as follows, where $K = \text{const} - \chi$ is positive-definite of order $n \times n$ and rank 12:

```
fit0 <- regress(rain~1, ~month)
fit1 <- regress(rain~1, ~K)
```

Positive definiteness is not required for computation so the constant is immaterial, but K is positive definite if the constant exceeds $2\tau/\pi^2 = 24/\pi^2$. Compute the Bayes estimate of monthly means for each model and superimpose these points on the plot of monthly averages.

6.10 For the Oxford rainfall data up to Dec 2019, the first Bayes estimate in the preceding exercise is a flat 10% shrinkage of monthly averages towards the annual average; the second Bayes estimate is different. For example, the average rainfall for September is 55.6 mm, which is slightly above the overall average of 54.7, so the first Bayes estimate is 55.5 mm. The second Bayes estimate is 57.5 mm. Explain this phenomenon—why the September component, which is already above the annual average, is shifted even further from the overall average.

6.11 Let $(\varepsilon_k, \varepsilon'_k)_{k \geq 0}$ be independent and identically distributed standard Gaussian variables. For real coefficients σ_k , show that the random function

$$\eta(t) = \sum_{k=0}^{\infty} \sigma_k \varepsilon_k \cos(kt) + \sigma_k \varepsilon'_k \sin(kt)$$

is stationary with covariance

$$\text{cov}(\eta(t), \eta(t')) = \sum_{k=0}^{\infty} \sigma_k^2 \cos(k(t - t'))$$

provided that the series converges in a suitable sense.

6.12 Verify the following trigonometric integral for integer k :

$$\int_0^{2\pi} \sin(x/2) \cos(kx) dx = \frac{-4}{4k^2 - 1}.$$

Hence find the coefficients λ_k in the Fourier expansion of the function

$$2/\pi - \sin(x/2) = \sum_{k=0}^{\infty} \lambda_k \cos(kx)$$

for $0 \leq x < 2\pi$, and show that they are all positive.

6.13 In this exercise, χ is the chordal metric on the unit circle. From the results of the preceding exercise, show that $4/\pi - \chi(t, t')$ is positive definite on $[0, 2\pi]$. Use the Choleski decomposition to simulate and plot a random function having this covariance function as follows:

```
n <- 1000; t <- (1:n)*2*pi/n; chi <- 2*abs(sin(outer(t, t, "-"))/2))
eta <- t(chol(4/pi - chi)) %*% rnorm(n)
plot(t, eta, type="l")
```

6.14 Show that the quadratic function

$$\frac{2\pi^2}{3} - x(2\pi - x)$$

on $[0, 2\pi]$ has Fourier cosine coefficients $4\pi/k^2$ for $k \geq 1$. Hence or otherwise, investigate the function

$$K(t, t') = \frac{2\pi^2}{3} - |t - t'|(2\pi - |t - t'|)$$

as a candidate covariance function for a process on $[0, 2\pi]$, and by extension to a stationary periodic process on the real line. Plot a simulation of the process on $[0, 4\pi]$, and verify continuity.

6.15 Suppose that $\eta \sim \text{GP}(0, K)$, with K as defined in the preceding exercise. The tied-down process $\zeta(t) = \eta(t) - \eta(0)$ is periodic and zero at integer multiples of 2π . Find its covariance function, and investigate its connection with the classical Brownian bridge.

6.16 Simulate and plot a random function $\eta(\cdot)$ on $(0, 2\pi)$ whose covariance is $\pi/2 - \ell(t, t')$, where $\ell(\cdot)$ is the arc-length metric. This function is less well behaved than the chordal function because half of its Fourier coefficients are zero, so the simulation code must be modified to accommodate singularities.

6.17 A real-valued process $Y(\cdot)$ is called a Gaussian random affine function if the differences $Y(t) - Y(t')$ are Gaussian with covariances satisfying

$$\text{cov}(Y(x) - Y(s'), Y(t) - Y(t')) = \sigma^2(s - s')(t - t')$$

for some $\sigma \geq 0$. Show that $Z(t) = \beta_0 + \beta_1 t$ is a random affine function if and only if β_1 is Gaussian.

6.18 Let K be the covariance function of a stationary process on the real line such that K is twice differentiable on the diagonal, i.e.,

$$K(t, t') = 1 - (t - t')^2 + o(|t - t'|^2).$$

Suppose that Y is stationary with covariance $\sigma^2 K((t - t')/\lambda)$. In the long-range limit $\lambda \rightarrow \infty$ such that $\sigma \propto \lambda$, show that Y is a random affine function.

6.19 Let $K(t, t') = \sigma^2 \exp(-|t - t'|/\lambda)$ be the scaled exponential covariance, and let Y be a zero-mean Gaussian process with covariance K . Show that the long-range limit with $\sigma^2 \propto \lambda$ is such that the deviation $Y(t) - Y(0)$ is finite and equal in distribution to Brownian motion.

Chapter 7

Time Series II



7.1 Frequency-Domain Analyses

7.1.1 Fourier Transformation

A time series is, in the first instance, a function $t \mapsto Y(t)$ on time points, either $t \in \mathbb{R}$ for a continuous-time process, or $t \in \mathbb{Z}$ for a discrete-time process. Most meteorological processes exist in continuous time, but are recorded in discrete time, either as noontime values, daily totals, daily averages or daily maxima. Similar remarks apply to plant and animal growth curves, personal health as a time series, and most business series and economic series. For statistical purposes, either in modelling or analysis, it is helpful to proceed as if the process exists in continuous time but is observed discretely at a finite collection of time points. Growth curves and personal health series are typically recorded at a small collection of irregularly-spaced time points. The methods of analysis described in this section are most suitable for a long series that is observed at a large collection of equally-spaced time points.

Let $Y(t)$ be the value recorded at time $t = 1, \dots, n$, so that the recording period is an interval of length n in suitable time units. Frequency is measured in cycles per recording interval, and the focus is on Fourier frequencies, which correspond to an integer number of cycles. The discrete Fourier transformation $\omega \mapsto \hat{Y}(\omega)$ at frequency ω is a complex number

$$\hat{Y}(\omega) = \sum_{t=1}^n e^{2\pi i \omega t / n} Y_t,$$

so that $\hat{Y}(0) = \hat{Y}(n) = Y.$ is the total, which is real in most applications. For integer frequencies $0 \leq \omega \leq n,$ the real and imaginary parts are the linear combinations

$$\begin{aligned}\hat{Y}(1) &= \sum_t \cos(2\pi t/n) Y_t + i \sum_t \sin(2\pi t/n) Y_t, \\ \hat{Y}(\omega) &= \sum_t \cos(2\pi \omega t/n) Y_t + i \sum_t \sin(2\pi \omega t/n) Y_t.\end{aligned}$$

For real $Y,$ the Fourier coefficients satisfy the aliasing identity

$$\hat{Y}(n - \omega) = \sum_{t=1}^n e^{2\pi i t(n-\omega)/n} Y_t = \sum_{t=1}^n e^{-2\pi i t\omega/n} Y_t = \overline{\hat{Y}(\omega)}.$$

so that $\hat{Y}(\omega)$ and $\hat{Y}(n - \omega)$ is a complex-conjugate pair. Thus, $\hat{Y}(0) = \hat{Y}(n) = Y.$ is real, and if $n = 2m$ is even, the middle value $\hat{Y}(m)$ is also real.

7.1.2 Anova Decomposition by Frequency

If we set aside the zero-frequency component, and split the non-redundant components into real and imaginary parts, the Fourier transformation $\hat{Y} = FY$ is a linear transformation $\mathbb{R}^n \rightarrow \mathbb{R}^{n-1}.$ The cosine and sine components of the Fourier matrix for frequency ω are the real and imaginary parts of roots of unity:

$$F_{\omega,t}^{(c)} = \cos(2\pi \omega t/n); \quad F_{\omega,t}^{(s)} = \sin(2\pi \omega t/n).$$

The identity $F\mathbf{1} = 0$ defines the kernel subspace, the rows are mutually orthogonal n -vectors, $FF' = (n/2)I_{n-1}$ is the identity of order $n - 1,$ and $2F'F/n = I_n - J_n/n$ is the orthogonal projection in \mathbb{R}^n with kernel $\mathbf{1}.$ The projection matrix $2FF'/n$ can be expressed as a sum of $\lfloor (n-1)/2 \rfloor$ rank-2 projection matrices, $P_\omega,$ one for each frequency, plus an additional rank-1 matrix for frequency $n/2$ if n is even. The net result is that the total sum of squares has an analysis-of-variance decomposition by frequencies

$$\sum_t (Y_t - \bar{Y}_.)^2 = \sum_{\omega=1}^{\lfloor n/2 \rfloor} \|P_\omega Y\|^2 = \frac{1}{n} \sum_{\omega=1}^{n-1} |\hat{Y}_\omega|^2.$$

The last expression includes both conjugates $\hat{Y}_\omega, \hat{Y}_{n-\omega},$ so the sum of squares for frequency $1 \leq \omega < n/2$ is

$$\|P_\omega Y\|^2 = 2|\hat{Y}_\omega|^2/n$$

on two degrees of freedom. As a function of ω , this is called the sample power spectrum, or the power spectrum.

7.2 Temperature Spectrum

7.2.1 Spectral Plots

The Central England daily temperature series for 248 years has a length of 90580 days. The analysis and interpretation are easier if the observation period is an integer number of years, so the partial year at the end is not included in the analysis. Harmonics associated with the annual cycle are expected to have large amplitudes, so it is helpful for plotting purposes to separate out the seasonal frequencies (integer multiples of 248) from the rest.

The first panel of Fig. 7.1 is a scatterplot of $\log |\hat{Y}_\omega|^2$ against frequency, which has been re-coded in units of cycles per year rather than cycles per observation period of 248 years. Seasonal frequencies have been excluded, partly because the annual and biannual coefficients are so large. The general trend is quite clear for the mean, but the high variability and density of points tends to obscure matters. Ordinarily, we should expect $\|P_\omega Y\|^2$ to be approximately exponentially distributed, in which case $\log \|P_\omega Y\|^2$ should have constant variance, $\pi^2/6 \simeq 1.28^2$, and the distribution should be skewed to the left. The plot is reasonably consistent with those expectations.

In the middle panel, the squared Fourier components $\|P_\omega Y\|^2$ have been averaged in consecutive non-overlapping frequency blocks, and the log averages are plotted against average frequency, again coded in cycles per year. In this manner, the variability is much reduced, so the trend in mean becomes clearly delineated. Note that the goal here is to estimate the spectrum, which is $E(|\hat{Y}_\omega|^2)$ as a function of ω , so all averaging takes place on that scale, not on the log scale.

Finally, the log spectrum for seasonal frequencies is shown in the third panel with the non-seasonal cubic spline superimposed for comparison. Apart from the first and second harmonics, the variation or energy at other seasonal frequencies decreases with frequency in conformity with the decrease observed in the second panel for non-seasonal frequencies. Certainly the variation at seasonal frequencies above three per year is not greater on average than the variation at neighbouring non-seasonal frequencies. The distinction between seasonal and non-seasonal seems to matter only for the first two annual harmonics.

These spectrum plots tends to emphasize the variation at higher frequencies, on the order of 20–150 cycles per year. However, it is the behaviour of the spectrum at low frequencies and the limiting behaviour as $\omega \rightarrow 0$ that is crucial for understanding long-range behaviour of temperatures. To clarify the picture, and to give greater emphasis to lower frequencies, Fig. 7.2 consists of the same points as the middle panel in Fig. 7.1, but the values are plotted against the square root of the

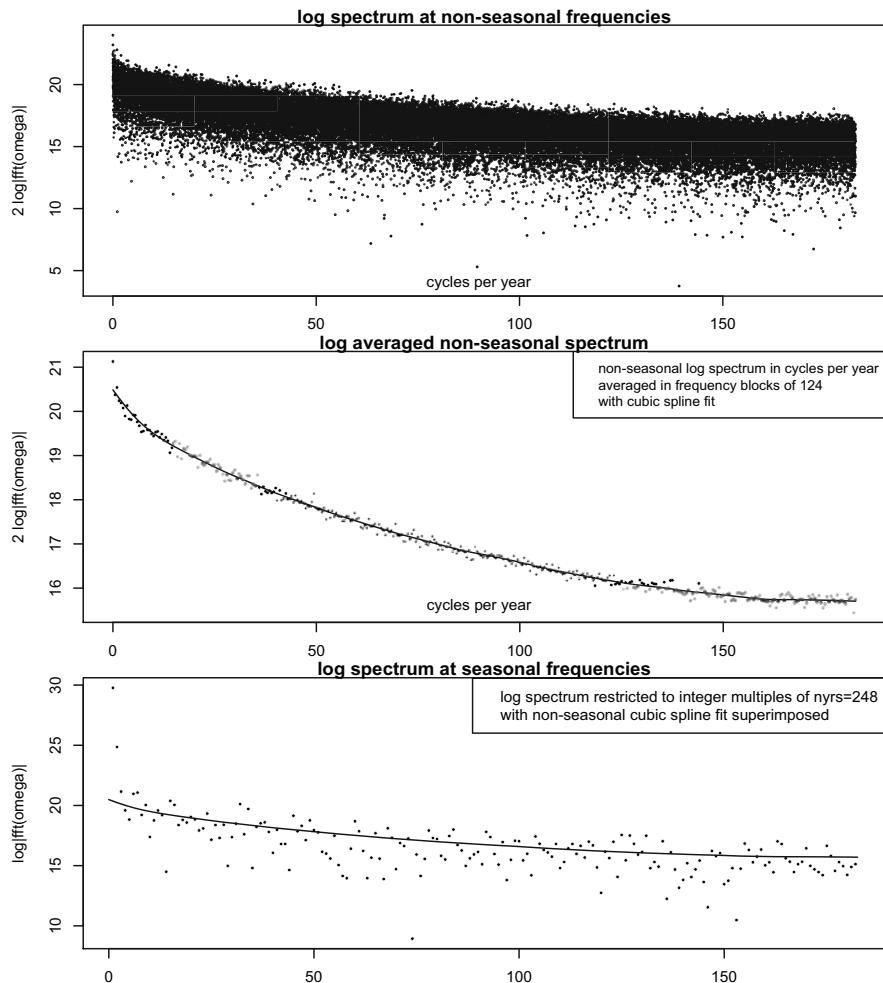


Fig. 7.1 Log power spectrum for the Central England temperature series separated by seasonal and non-seasonal frequencies

frequency. To a first order of approximation, the log spectrum is linear in $\omega^{1/2}$ over the bulk of the frequency range.

7.2.2 A Parametric Spectral Model

According to the theory discussed in the next section, the transformed coefficients $|\hat{Y}(\omega)|^2$ for non-seasonal frequencies are approximately independent exponential

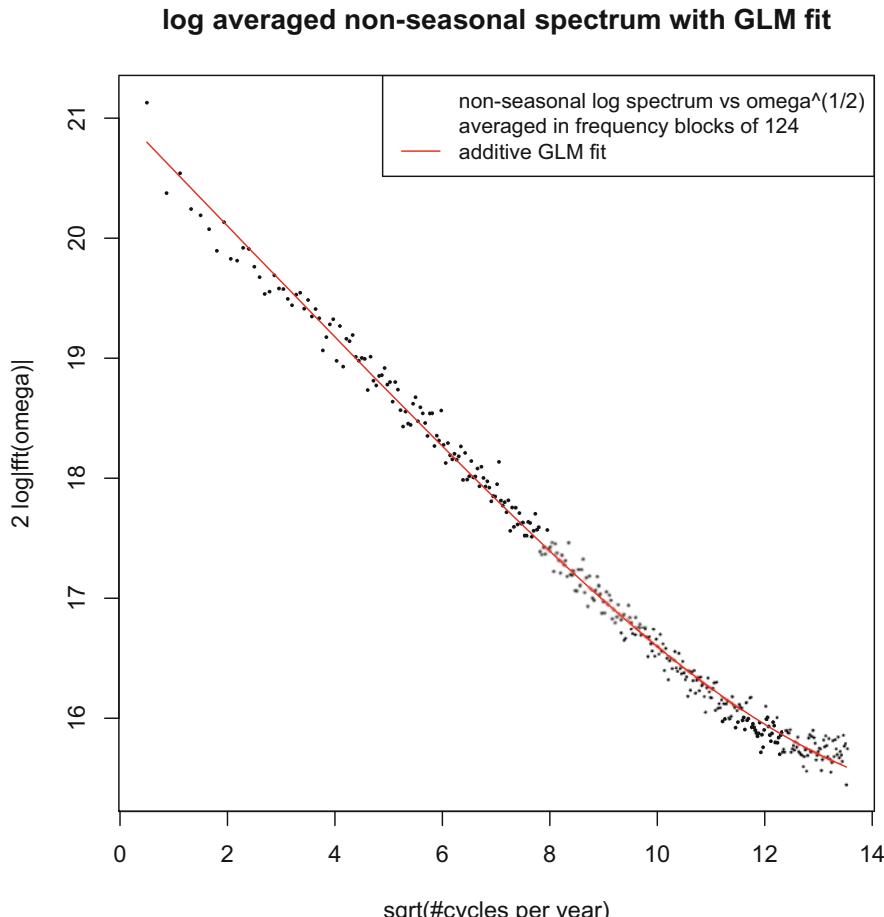


Fig. 7.2 Log power spectrum plotted against $\omega^{1/2}$. The solid line is the additive GLM spectral fit $E|\hat{Y}(\omega)|^2 \propto 1 + 391 \exp(-|0.219\omega|^{1/2})$

random variables. With this in mind, it is natural to fit a two-component additive spectral model

$$E|\hat{Y}_\omega|^2 = n\sigma_0^2 + n\sigma_1^2 \exp(-|2\pi\lambda\omega|^{1/2})$$

with three non-negative parameters $\sigma_0, \sigma_1, \lambda$ to be estimated. Aggregation by frequency blocks is helpful for plotting, but it is not needed for fitting this model, which, for fixed λ , is a gamma-type generalized linear model with unit dispersion. Additivity on the power-spectrum scale rather than on the log scale is natural if the temperature series is to be regarded as the sum $Y(t) = \sigma_0\varepsilon(t) + \sigma_1\eta(t)$ of independent processes. Typically, ε is white noise, and η is an independent serially-

correlated process whose sample paths are continuous in a suitable sense—either continuous with probability one or mean-square continuous. The spectral density proposed here decays sufficiently fast at high frequencies that η has continuous derivatives of all orders.

With ω measured in cycles per year, the fitted exponential model is

$$\hat{K}(\omega) = n\hat{\sigma}_0^2 + n\hat{\sigma}_1^2 \exp(-|2\pi\hat{\lambda}\omega|^{1/2}),$$

where $\hat{\lambda} = 0.0347$ years, or 12.67 days, $\hat{\sigma}_1/\hat{\sigma}_0 = 19.8$ for the volatility ratio, and $\hat{\sigma}_0 = 0.62$ for the nugget standard deviation in degrees Celsius. Note that the second component is formally the characteristic function of the α -stable distribution for $\alpha = 1/2$, so the associated covariance function is the density of that distribution.

The additive gamma model fits the non-seasonal power spectrum reasonably well, but it is not perfect. Small systematic deviations are apparent in Fig. 7.2 at low frequencies, and there is approximately 3% excess dispersion relative to the exponential distribution. In other words, the variance of the standardized spectral coefficients $|\hat{Y}_\omega|^2/\hat{K}_\omega$ is 1.032, while the exponential model predicts unit variance. This is a very small deviation in absolute terms, but, with 45 108 non-seasonal Fourier frequencies, a 3% deviation in variance is moderately unlikely.

In the residual plots shown in Fig. 7.3, the 3% deviation is too small to be noticed. Overall, the residual distribution seems to match the extreme-value distribution very closely. The mean of the log residuals is -0.584 , and the variance is 1.661 versus the theoretical values $-\gamma = -0.577$ (Euler's constant), and $\sigma^2 = \pi^2/6 = 1.645$.

On the negative side, the ratios of the squared Fourier coefficients to the fitted values for the ten lowest frequencies are

$$19.67, 4.44, 9.19, 4.40, 2.15, 0.73, 4.28, 0.55, 0.36, 2.99,$$

and the next largest ratio is 11.3, which occurs at one of the highest frequencies. Despite the apparent success of this parametric model for the bulk of the frequency range, these low-frequency values are not consistent with the fitted model, which predicts independent standard exponential values. The expected value of the largest of n standard exponentials is approximately $\log(n) \simeq 10.7$, and the standard deviation is approximately $\pi/\sqrt{6} \simeq 1.3$. Given that it was so selected, the second-largest ratio is entirely consistent with the fitted model, but the first 4–5 Fourier coefficients are not.

The behaviour of the low-frequency Fourier coefficients is strongly tied to the behaviour of the covariance function or variogram at the longest lags. Bearing in mind the variogram phenomenon observed in the third and fourth panels of Fig. 6.5, which is compatible with a slow random walk or an autoregressive process with semi-range λ on the order of one millennium, it is natural to look for a corresponding phenomenon in the Fourier domain. The corresponding phenomenon is an additive spectral component proportional to $1/(1 + \lambda^2\omega^2)$, which is essentially a multiple of ω^{-2} . Inclusion of the inverse-square frequency as a further covariate the spectral

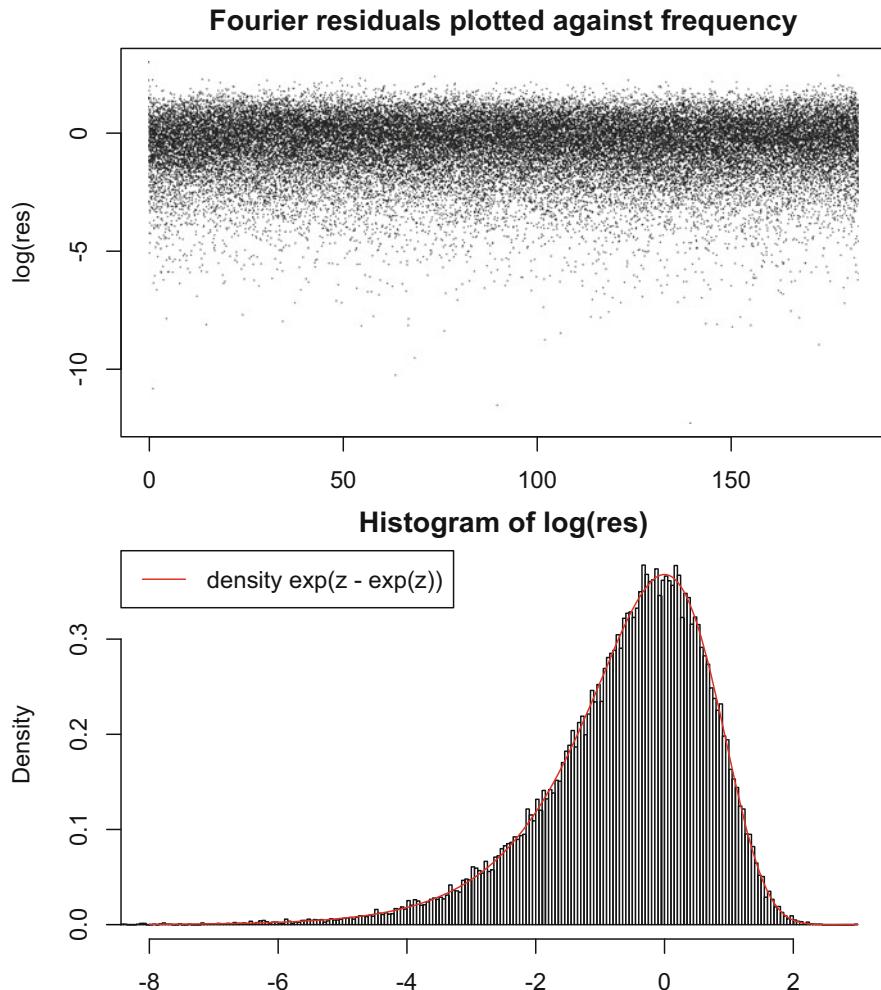


Fig. 7.3 (a) Log Fourier residuals $\log(|\hat{Y}_\omega|^2|/\text{fitted}_\omega)$ plotted against frequency; (b) Histogram of log residuals with theoretical extreme-value density superimposed

model reduces the deviance by 52+ units, which is a substantial improvement to the fit, showing conclusively that the slow linear trend seen in the variogram plots is a real phenomenon and not a statistical artifact.

7.3 Stationary Temporal Processes

7.3.1 Stationarity

This section is concerned with real-valued processes that are defined pointwise on the domain, and specifically with stationary Gaussian processes on the real line. The first part of the statement means that to each point t in the domain there corresponds a value $Y(t)$, which is a real number. First-order stationarity implies that for each pair of points $t, t + h$ in the domain the values $Y(t)$ and $Y(t + h)$ have the same distribution. For a time series, the domain is either the integers or the real line, and translation implies that the domain is a group acting on itself by addition. More generally, stationarity implies that for each ordered n -configuration $\mathbf{t} = (t_1, \dots, t_n)$ and each h -translate $\mathbf{t} + h = (t_1 + h, \dots, t_n + h)$, the values

$$Y[\mathbf{t}] = (Y(t_1), \dots, Y(t_n)) \quad \text{and} \quad Y[\mathbf{t} + h] = (Y(t_1 + h), \dots, Y(t_n + h))$$

have the same joint distribution in \mathbb{R}^n .

The focus is on Gaussian processes, which are defined by the mean function $\mu(t) = E(Y(t))$ and the covariance function $K(t, t') = \text{cov}(Y(t), Y(t'))$, which is a symmetric positive semi-definite function on the domain. Stationarity implies that the mean is constant, $\mu(t) = \mu(0)$, and that $K(t, t') = K(|t - t'|)$ is a function of the temporal separation. For example, $e^{-|t-t'|}$ is the covariance function for the standard first-order autoregressive process, and $(1 + |t - t'|)e^{-|t-t'|}$ is a related covariance function in the Matérn class with $\nu = 3/2$.

The restriction to processes defined pointwise is not vacuous because there exist temporal processes that are not defined pointwise. For example, standard white noise is a zero-mean Gaussian process defined on domain *subsets* such that $\text{cov}(Y(A), Y(B)) = \Lambda(A \cap B) < \infty$ is the Lebesgue measure of the intersection. The distribution is invariant with respect to translation, so the process is certainly stationary. However, the pointwise definition of stationarity is not satisfied because $Y(t)$ is not defined for such processes. If we attempt to define $Y(t)$ as a limit over subsets converging to $\{t\}$, then $Y(t) = 0$; if we regard $Y(A)$ as an integral of $Y(t)$ over A then $Y(t)$ cannot have finite variance. Neither of these implications is satisfactory.

The definition of stationarity given above is to be read conditionally as follows. If Y is defined pointwise, then Y is stationary if and only if, for every positive integer n , every n -configuration \mathbf{t} , and every $h \in \mathbb{R}$, the random variable $Y[\mathbf{t} + h]$ has the same distribution as $Y[\mathbf{t}]$.

For the more general definition of stationarity, the process is defined on an index set consisting of points or subsets or measures on the domain. Thus, the domain need not coincide with the index set of the process. To consider stationarity, the domain (\mathbb{Z} or \mathbb{R} or \mathbb{C} or \mathbb{R}^2) is necessarily a group, and the index set is closed under domain translation. The process is first-order stationary if, for each object A in the index set, and for each point h in the domain, the value $Y(A)$ has the same

distribution as the value $Y(A+h)$ taken on the h -translated object. Strict stationarity is defined in the same way for joint distributions. According to this definition, it is possible to make sense of the statement that $-|t - t'|$ is the covariance function for a Gaussian process or time series, sometimes called a generalized process because the index set does not coincide with the domain. Likewise for the functions $-|t - t'|^{1/2}$ and $-\log |t - t'|$. Moreover, these processes are strictly stationary. This definition paves the way to consider other group actions such as rigid motions or Euclidean congruences or similarity transformations, which are associated with isotropy and self-similarity.

7.3.2 Visualization of Trajectories

To understand what the $\text{SD}_{1/2}$ -process with spectral density $\exp(-|\omega|^{1/2})$ looks like, i.e., how a typical trajectory behaves as a function, it is helpful to compute, simulate and plot. The first step is to compute the covariance function by inversion of the spectral density. In general, this is a non-trivial computational exercise. Fortunately, this spectral density is a special case of the characteristic function of the α -stable class. The series expansion for the density (Feller, 1971, vol II, p. 582) can be simplified for $\alpha = 1/2$; for general α , see Matsui and Takemura (2006). We remark only that K is strictly positive, and monotone as a function of temporal separation. It is infinitely differentiable at all points on the real line, but it is not complex-analytic in any neighbourhood of the origin: the Taylor expansion does not have a positive radius of convergence. Accuracy to two or three significant decimal digits suffices for graphical representation of the covariance function, but at least eight-digit accuracy is needed to simulate trajectories.

Four covariance functions are shown in Fig. 7.4. At first glance, the differences among them appear to be slight: all four are continuous, symmetric and are equal at the origin and at ± 1 . The behaviour in a neighbourhood of the origin is an important characteristic, which is shown in 5x-magnified form on the right of each panel. The Matérn functions have zero, one and two derivatives at the origin, whereas the fourth has infinitely many. The first, $1 - |x| + o(x)$, is easy to see by inspection, but the others are not, even in magnified form: for $\nu = 1$, the behaviour near the origin is $1 + x^2 \log |x|/2 - 0.31x^2 + o(x^2)$, so the first derivative is zero and the second does not exist. The behaviour in the tail is the characteristic that distinguishes intermediate- and long-range dependent processes from short-range. Once again, this is easier to see in hindsight than in foresight—especially if zero has not been included for visual reference in the graph.

All four curves in Fig. 7.4 are non-negative and have finite integrals, so each is proportional to a symmetric probability distribution on the real line. The integrals are 2.0, 1.89, 1.86 and 4.58 respectively, or more generally, 2λ , $\pi\lambda$, 4λ and $\pi\lambda/2$ for the scaled versions. The first Matérn covariance is a multiple of the Laplace density; the $\text{SD}_{1/2}$ covariance is a multiple of the α -stable density for $\alpha = 1/2$.

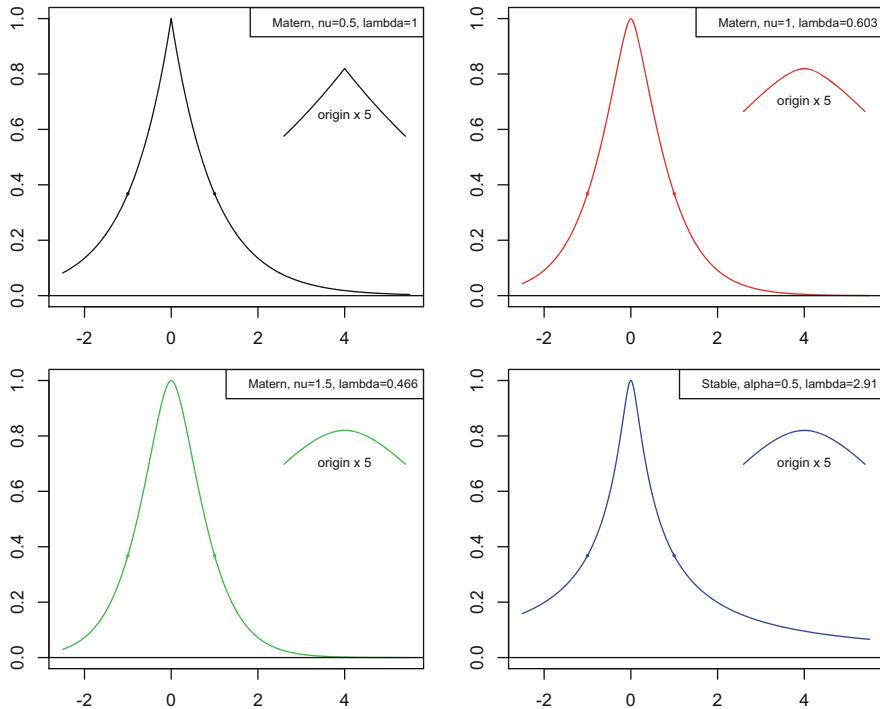


Fig. 7.4 Four covariance functions standardized to have unit variance and lag-one autocorrelation e^{-1} . The fourth is a multiple of the α -stable density function for $\alpha = 1/2$

Given computer code for the covariance function, the covariance matrix Σ for the process at 1000 points may be computed, followed by simulation of a 1000-component Gaussian variable $Y \sim N(0, \Sigma)$. These are the values of the process at the selected points in the domain. Special cases can be simulated more efficiently, but this straightforward recipe suffices for present purposes. Each of the curves in Fig. 7.5 is plotted using the values $(x, Y(x))$ at 1000 equally-spaced points in the interval $(0, 10)$.

To establish a ‘normal range’ of patterns, Fig. 7.5 shows the trajectories of three Matérn processes in the ‘typical’ index range, plus the $SD_{1/2}$ process. Each family has a variance parameter and a range parameter, both of which are strictly positive real numbers. For purposes of comparison, each covariance function is scaled to have unit variance, and the same lag-one autocorrelation $e^{-1} = 0.368$, which matches the standard order-one autoregressive process shown in the top panel. Each of the Matérn processes has a distinct character, with continuous derivatives of order zero, one and two for $\nu = 0.5$, $\nu = 1.0$ and $\nu = 1.5$ respectively. Visually speaking, the differences among these three are concerned with the degree of smoothness, which is a local property. The $SD_{1/2}$ process has its own distinct character. It has continuous derivatives of all orders so it is smoother than any Matérn process, even

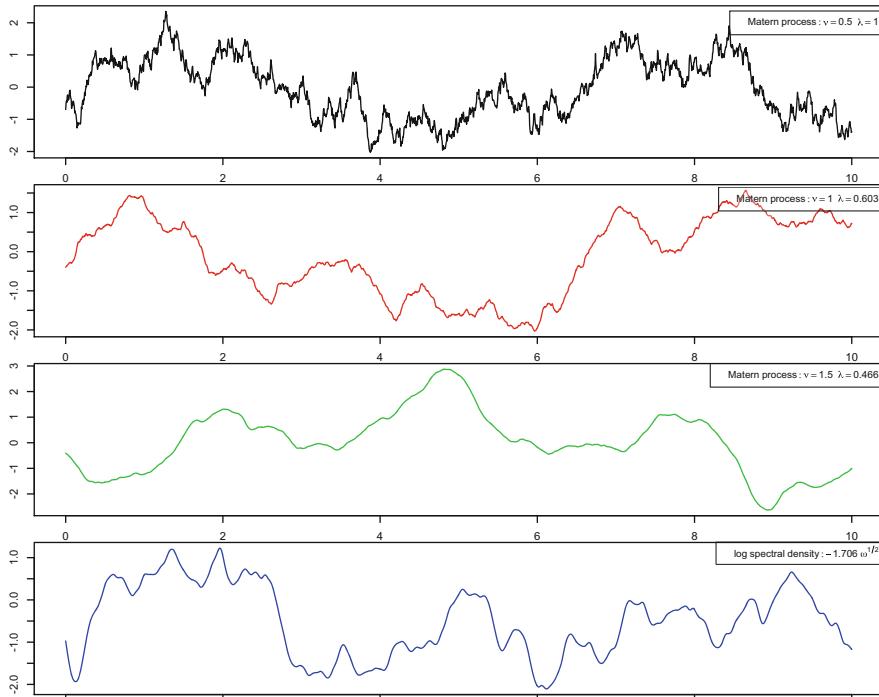


Fig. 7.5 A comparison of trajectories of four stationary continuous-time processes, three in the Matérn class, and one specified by its spectral density $e^{-|\omega|^{1/2}}$. The standard Matérn covariance function is $K_v(x) = \|x\|^v \mathcal{K}_v(\|x\|)$; special cases include $K_{1/2}(x) = e^{-\|x\|}$ and $K_{3/2}(x) = (1 + \|x\|)e^{-\|x\|}$. The spectral density is $1/(1+\omega^2)^{v+1/2}$. For visual comparison over the interval $(0, 10)$, all four processes are standardized to have unit variance and the same lag-one autocorrelation

those with $v > 3/2$. However, its medium-range oscillations are distinctly more pronounced, and somewhat similar to those of the AR1 process ($v = 1/2$).

In addition, although it is hard to point to visual consequences in the trajectories, the long-range autocorrelation of the $SD_{1/2}$ process is algebraic of order $|x-x'|^{-3/2}$, whereas the long-range Matérn correlations decay faster than any polynomial. As a result, the lag 5–10 autocorrelations are sizable 0.073–0.031 for the $SD_{1/2}$ process, but negligible for all Matérn processes and decreasing as a function of v . Long-range dependence appears to be universal for processes in nature, both for time series and for spatial processes. For such applications, we should bear in mind that each finite-range Matérn covariance has exponentially-decreasing tails whereas the $SD_{1/2}$ covariance has regularly-varying tails of order $|x - x'|^{-3/2}$. All four covariance functions have finite integrals, so all four processes are short-range dependent. On the other hand, each Matérn process has a well-behaved infinite-range limit, whereas the $SD_{1/2}$ process does not.

Algebraic, or inverse-polynomial, decay of autocorrelations is a characteristic of intermediate-range and long-range dependence. One consequence is that the sample

average over the interval $(0, t)$

$$\bar{Y}_{(0,t)} = t^{-1} \int_0^t Y(s) ds \quad \text{or} \quad \bar{Y}_{1:t} = t^{-1} \sum_{s=1}^t Y_s,$$

has a variance that tends to zero as $t \rightarrow \infty$ at a slower rate than $O(t^{-1})$. The rate is $O(t^{-1})$ for short-range dependent series, including every Matérn process, but only $O(t^{-1/2})$ for the SD_{1/2} process. Empirically, we find that the variance of the temperature average over randomly-sampled blocks of h successive years is as follows:

| h | Block length in years | | | | | |
|--------------------------------------|-----------------------|-------|-------|-------|-------|-------|
| | 4 | 8 | 16 | 32 | 64 | 128 |
| $C_h \text{ var}(\bar{Y}_h)$ | 0.213 | 0.158 | 0.133 | 0.087 | 0.058 | 0.036 |
| $h^{1/2} C_h \text{ var}(\bar{Y}_h)$ | 0.427 | 0.446 | 0.533 | 0.491 | 0.465 | 0.409 |
| $h C_h \text{ var}(\bar{Y}_h)$ | 0.853 | 1.261 | 2.130 | 2.778 | 3.719 | 4.624 |

In this table $\text{var}(\bar{Y}_h)$ is the sample variance of 5000 randomly-sampled block averages. The factor $C_h^{-1} = 1 - h/248$, which is the average fractional overlap between pairs of blocks of length h , is a finite-population bias-correction factor for sample overlap. Whole-year blocks were used to eliminate the effect of seasonal cycles. It is apparent from the table that $\text{var}(\bar{Y}_h) \propto h^{-1/2}$ is the dominant term for the variance of block averages, at least up to $h \simeq 128$ years.

7.3.3 Whittle Likelihood

When the circumstances permit it, i.e., when a series is recorded over a large number of equally-spaced points, the advantages of working in the frequency domain are considerable. For a stationary series with covariance function K , Whittle (1953) shows that the Fourier coefficients are approximately Gaussian and approximately independent for large n , with moments

$$E(\hat{Y}(\omega)\overline{\hat{Y}(\omega')}) = \begin{cases} n\hat{K}(2\pi\omega/n) + O(1) & \omega = \omega', \\ O(1) & \omega \neq \omega', \end{cases}$$

where \hat{K} is the spectral density of K . Using this approximation, we may treat the frequencies as observational units, and the Fourier coefficients as independent complex-Gaussian observations. In the standard technical sense, the squared moduli $|\hat{Y}(\omega)|^2$ are sufficient for the spectral density. To accommodate the seasonal cycle in the present application, it is necessary either to restrict attention to non-seasonal frequencies or to eliminate the first few seasonal harmonics.

The Whittle approximation overlooks the fact that, in many applications, the series is defined in continuous time, but observed in discrete time. For that reason, the expected value of the Fourier coefficient $\hat{Y}(\omega)$ is not exactly equal to the Fourier transformation of the covariance function. Sykulski et al. (2019) show how to compute the exact expectation of $\hat{Y}(\omega)$ efficiently, and to use the bias-corrected version in the Whittle likelihood.

7.4 Exercises

7.1 Let Y_1, \dots, Y_n be independent and identically distributed standard exponential variables, and let $0 \leq Y_{(1)} \leq Y_{(2)} \leq \dots \leq Y_{(n)}$ be the order statistics. Show that $Y_{(1)} \geq t$ if and only if $Y_i \geq t$ for $1 \leq i \leq n$, and deduce that $nY_{(1)}$ is exponentially distributed with unit mean. Hence or otherwise, show that the increments $(n-r)(Y_{(r+1)} - Y_{(r)})$ are independent and identically distributed. Find the mean and variance of the maximum $Y_{(n)}$, with asymptotic values for large n .

7.2 Use `fft()` to compute the Fourier coefficients for the temperature series on a whole number of years, identify and remove the frequencies that are seasonal, average the power-spectrum values in successive non-overlapping frequency blocks of a suitable size, and plot the log averages against the square root of the frequency in cycles per year.

7.3 For the non-seasonal frequencies, use `glm()` to fit the additive exponential model

$$E(|\hat{Y}(\omega)|^2) = \beta_0 + \beta_1 \exp(-|2\pi\lambda\omega|^{1/2})$$

for various values of λ in the range 7–14 days or 0.02–0.04 years. In this setting, the distributional family is Gamma, the link is `identity`, and the dispersion parameter is one. Plot the residual deviance against λ to find the maximum-likelihood estimate. Check that the fitted coefficients are non-negative. Superimpose the fitted curve on the graph of log-block-averages in the previous exercise. Plot the standardized residuals (log-ratio of observed to fitted) against frequency, and comment on any departures that are evident.

7.4 Include the inverse-square frequency as an additional covariate in the exponential model for the power spectrum. In principle, this means re-computing $\hat{\lambda}$. Compute the Wilks statistic, which is the reduction in deviance or twice the increase in log likelihood. Also compute on the Wald statistic, which is the squared ratio of the ω^{-2} -coefficient to its standard error as given by the inverse Fisher information matrix. Recall that the dispersion parameter is one, which is not the default in `summary()`. Standard asymptotic theory for large sample sizes tells us that the difference between these two statistics is $o_p(1)$, i.e., that the difference tends to zero

as $n \rightarrow \infty$, and also that the null distribution is χ_1^2 for both. In this setting the sample size is the number of non-seasonal frequencies. Comment on any discrepancy between theory and practice in this instance, and provide an explanation.

7.5 Ordinarily, Wald's likelihood ratio statistic is essentially the same as Wilks's statistic, which in one-parameter problems, is the squared ratio of the estimate to its standard error. But there are exceptional cases where a substantial discrepancy may occur, and variance-components models provide good examples. In order to understand the source of the discrepancy, simulate data with simple structure as follows:

```
set.seed(3142); n <- 1000; x <- 1:n
rx2 <- 1/x^2; beta <- c(1, 20);
X <- cbind(1, rx2); mu <- as.vector(X %*% beta)
y <- -log(runif(n)) * mu
```

The null hypothesis is that $\mu \propto \mathbf{1}$ is constant, and the alternative is that $\mu = X\beta$ for some β with non-negative components. Test this hypothesis using Wilks's likelihood ratio statistic, and also using the Wald statistic. Recall the exponential assumption, which implies that the dispersion parameter is one.

7.6 If you used the function `glm(y~rx2, family=Gamma(link=identity))` in the preceding exercise, you may have experienced a failure to converge. Write your own Newton-Raphson function with steps on the log scale, which forces the β -components to be strictly positive. As part of this exercise, you will need to compute the Fisher information matrix, $t(X/\mu)^2 \%*\% X$ for β . Report the value of I_β at the null hypothesis $\hat{\beta}_0$ and also at $\hat{\beta}$. What does this tell you about the Wald-Wilks discrepancy?

7.7 For $0 < \alpha \leq 2$, the α -stable distribution on the real line is symmetric with characteristic function $e^{-|\omega|^\alpha}$. For the sub-range $0 < \alpha < 1$, Feller (1971, eqn. 6.5) gives the series expansion for the density

$$p(t; \alpha) = \Re \frac{i}{\pi t} \sum_{k=0}^{\infty} (-1)^{k+1} \frac{\Gamma(k\alpha + 1)}{k!} t^{-k\alpha} e^{-\pi i k \alpha / 2}$$

which is convergent for $t > 0$. The goal of this exercise is to simplify the density for $\alpha = 1/2$ by splitting the sum into four parts according to $k \pmod 4$. Show that one of the four parts is zero, that the odd parts may be combined into a multiple of $t^{-3/2} \sin(1/(4t) + \pi/4)$, and that the remaining part is $O(t^{-2})$ as $t \rightarrow \infty$.

7.8 From the cosine integral $\int \cos(\omega t) e^{-|\omega|^\alpha} d\omega$, deduce that the α -stable density has a Taylor series at the origin which begins

$$\log p(t; 1/2) = \text{const} - 60t^2 + O(t^4).$$

Find the general term in this expansion and deduce the radius of convergence.

Chapter 8

Out of Africa



8.1 Linguistic Diversity

This chapter is concerned with the linguistic hypothesis and the data analysis in the paper *Phonemic diversity supports a serial founder effect model of language expansion from Africa*, published by Q.D. Atkinson in *Science* (15 April 2011). It is recommended that you read the paper and the supplementary material. The data and supplementary files can be found at . . . /ch08/

Like the genetic thesis for human migration and evolution, the ‘Out-of-Africa’ thesis for linguistic diversity holds that language evolved somewhere in Africa, and diffused from there to Asia, Europe and elsewhere as populations split and migrated. Since the genetic and linguistic diversity of a population is intrinsically related to its size, a small migrating subset carries less diversity than the population from which it originated. Accordingly, a subpopulation that splits and migrates carries less diversity than the descendants of the ancestral population that remains. Although tones and sounds are continuously gained and lost in all languages, the loss is supposedly higher for small migrating founder populations than for the ancestral population. In this way, the diversity of sounds becomes progressively reduced as the distance from the origin increases.

Atkinson’s paper is concerned with the hypothesis that human language developed in a single location and spread from there by migration. He aims to test that hypothesis by examining the relationship between the diversity of sounds in 504 extant languages and their geographic distance from a putative origin in Africa or elsewhere, taking account of speaker population size. It is not our business here to discuss the plausibility of Atkinson’s thesis. As we might expect, a wide range of mostly skeptical views on that point has already been expressed in the literature. It

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-14275-8_8.

is our business solely to examine the data carefully to assess whether the thesis is supported by the data or not.

8.2 Phoneme Inventory

The data on which Atkinson bases his analysis is a list of 504 languages from various parts of the world. The list of languages is not exhaustive, nor is it close to geographically uniform with respect to current population density. The diversity measure is not a measure of variability of sounds in the ordinary statistical sense, nor is it an inventory or list of sounds, but simply the number of distinct phonemes that the language employs. There are three distinct values for vowel inventory, three for tone inventory, five for consonant inventory, and 40 for total phoneme inventory. In principle, inventories should be all be non-negative integers, but the values have been standardized or normalized in an unspecified way. The sample means are close to zero, and the sample variances for the three phoneme constituents are close to one.

Data on phoneme inventory size were taken from the World Atlas of Language Structures, (WALS). The file ch08/S1.dat contains the main part of Atkinson's data, which is the list of 504 languages together with the following twelve variables

1. Lname Language name: text, e.g. Abkhaz, Aikan??, B?@t?©, ..., Zuni
2. WALS three character code, e.g. abk, aik, bet,...
3. Fam Language Family: text, e.g. Arawakan, Indo-European, Sino-Tibetan,...
4. Lat Latitude as a decimal number, e.g. -12.67
5. Long Longitude as a decimal number, e.g. -60.67 (meaning 60° 40'W)
6. Nvd Normalized vowel diversity based on WALS feature No. 2
7. Ncd Normalized consonant diversity based on WALS feature No. 1
8. Ntd Normalized tone diversity based on WALS feature No. 13
9. Tnpd Total normalized phoneme diversity
10. Iso ISO codes (one or more three-character codes)
11. Popn Estimated speaker population: integer 1–873 014 298
12. Dbo Distance in km. from Atkinson's best-fit origin

Regardless of its geographical range, each language is associated with a single point on the sphere, which is not necessarily the geographic centroid of the speaker domain. International languages such as English, Spanish and French are associated with their ancestral capitals. For example, English is Indo-European and is located at latitude 52.0, longitude 0.0; the speaker population 309M is dominated by parts of the former empire. Spanish is located at 40.0N, 4.0W with a population size 322M most of whom are in Latin America; Mandarin is located at 34.0N, 110.0E, with a population of 873M. The guiding principle for inclusion is not evident. Among European languages, Albanian, Basque, Catalan, Breton, Romansch and Saami are

included, but not Portuguese (178M), Italian (60M), Dutch (22M), Ukrainian (37M), Belarusian, Slovak, Slovene, Serbian or Croatian.

For whatever reason, phoneme values are rounded and normalized. In addition, the number of distinct values for each variable is very limited. For example, English, French, German and Korean have exactly the same diversity profile (1.39, 0.12, -0.77), which is shared by Turkish and 21 other languages; The values in the file are reported to seven or eight decimal digits. Donegal Irish shares its consonant-dominant diversity profile (-0.48, 1.80, 0.18) with 13 other languages including Kwakw'ala, Lezgian and Saami.

8.3 Distances

The languages were partitioned into six continental groups, Africa, Europe, Asia, Oceania, N.Amer, and S.Amer. These coincide closely with the geographic continents, but not exactly so: Malagasy, the national language of Madagascar, belongs to Oceania, not Africa.

For his analyses, Atkinson used great circle distances between points x, x' for pairs of languages belonging to the same continental group. Otherwise, for languages in different continental groups, distances were measured for the shortest path passing through certain choke points (supplementary material, Fig. S8). For example, the shortest linguistic path from Europe to N. America consists of three great-circle arcs passing through Istanbul and the Bering Strait. The great-circle distance from Aghem = (10.0, 6.67) in the Congo to Malagasy = (47.0, -20.0) in Madagascar is 5014 km, but since the latter belongs to region 4, the linguistic distance through Cairo and Phnom Penh is 18475 km.

The R-executable file ch08/OOA.R contains the following commands:

```
S1 <- read.csv(file="ch08/OOA.dat")
S4 <- read.csv(file="ch08/S4.dat", header=FALSE)
```

The same file also contains the coded list 1region of 504 linguistic groups, the geocoordinates of a small number of major cities including the choke points chokes, some code for geographic plotting, and functions for computing distances, as follows.

1. gcdist(x1, x2) great circle distance in km:

```
gcdist(Aghem, Malagasy) = 5013.585
gcdist(Paris, Chicago) = 6651.991
```

The format used here for geocoordinates is x=(long, lat) in decimal degrees.

2. chokdist(x1, x2, r1, r2) linguistic distance:

```
chokdist(Aghem, Malagasy, 1, 4) = 18474.65
chokdist(Paris, Chicago, 2, 5) = 15761.26
```

3. `vdist(Dublin, 2)` a list of 504 linguistic distances from Dublin=(-6.25, 53.33), regarded as a member of linguistic region 2. For great-circle distances, use `vdist(Dublin, 0)`. (The 504 language coordinates are assumed to be in `S1$Long, S1$Lat.`)

8.4 Maps and Scatterplots

The file `ch08/OOA.R` also contains standard R code for plotting world maps and subsets thereof, using the `ggplot2` and `rgeos` packages downloaded from the CRAN website. For illustration, Fig. 8.1 shows the location of the African and European languages used in the analysis. The African languages are heavily concentrated in equatorial Africa, roughly 10°S to 15°N. Among the eleven African countries south of 10°S, only four—Angola, Botswana, Namibia and South Africa—are represented, and Botswana has four or five out of the eight languages. Madagascar is represented by Malagasy, whose roots are non-African. Similar anomalies are evident in the European sample.

The putative linguistic origin is a geo-coordinate point x_0 for which total phoneme diversity for language i satisfies

$$E(Tnpd_i) = \beta_0 + \beta_1 \|x_i - x_0\|,$$

where $\|x_i - x_0\|$ is the linguistic distance between the origin and the geo-coordinate of language i . The least-squares criterion is thus

$$\sum_i (Tnpd_i - \beta_0 - \beta_1 \|x_i - x_0\|)^2,$$

which is to be minimized with respect to the four parameters β_0, β_1 plus the two components of x_0 . Atkinson reports the best-fitting origin at 1.25°S, 9.30°E in West Africa: see Fig. 8.4.

Support for Atkinson's thesis, that phoneme diversity—or more correctly phoneme inventory—decreases with distance from the origin is best illustrated by the aggregate scatterplot in Fig. 8.2 of total phoneme diversity against distance from the best-fitting origin. The least-squares fitted line has a definite negative slope.

What the scatterplot Fig. 8.2 fails to show is that all of the distances up to 4.5 units are in Africa, many of those in the interval 5.5–8.5 are in Europe, most of those in the interval 5.5–13.5 are in Asia, and so on. Consequently, it is natural to plot each continental group separately, which is done in Fig. 8.3. This exercise reveals that the relation between distance and phoneme inventory is negative chiefly in Africa. The negative least-squares slope for Oceania is almost entirely due to the single point (Malagasy), which has high leverage on account of its remoteness, and a low phoneme inventory. For each of the other linguistic

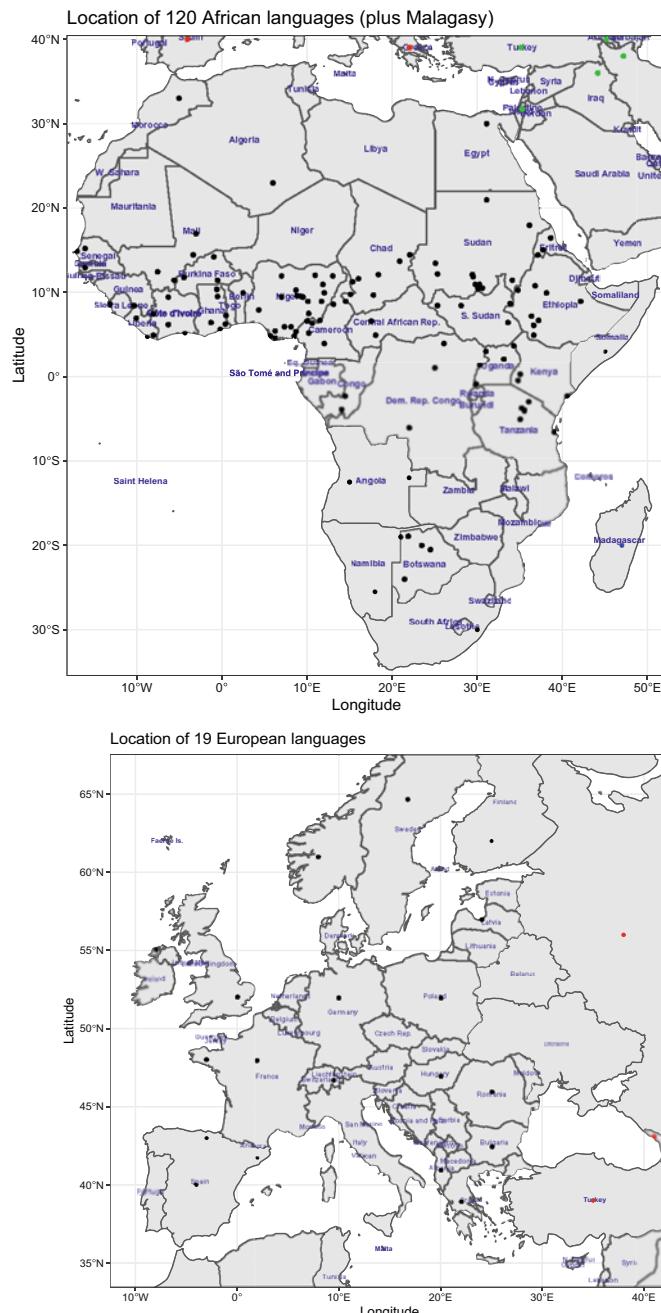


Fig. 8.1 Geographic distribution of African and European languages in Atkinson's sample

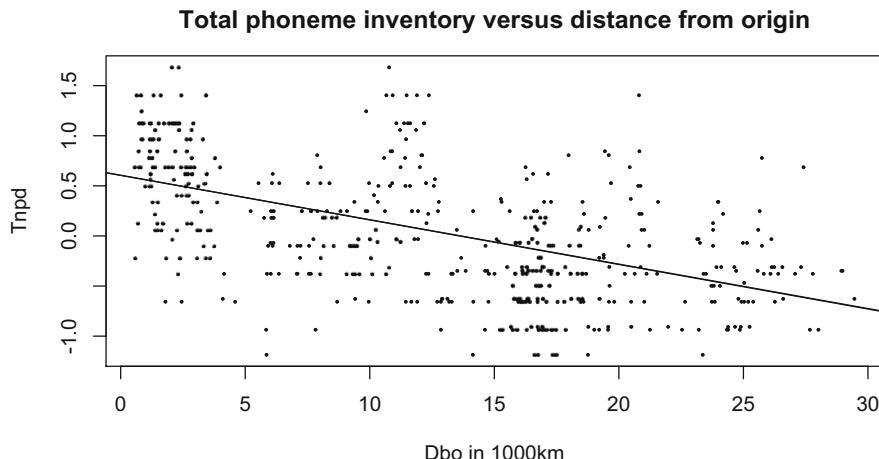


Fig. 8.2 Total phoneme diversity plotted against the distance from the best-fitting origin as reported by Atkinson at 9.5°E, 1.25°S

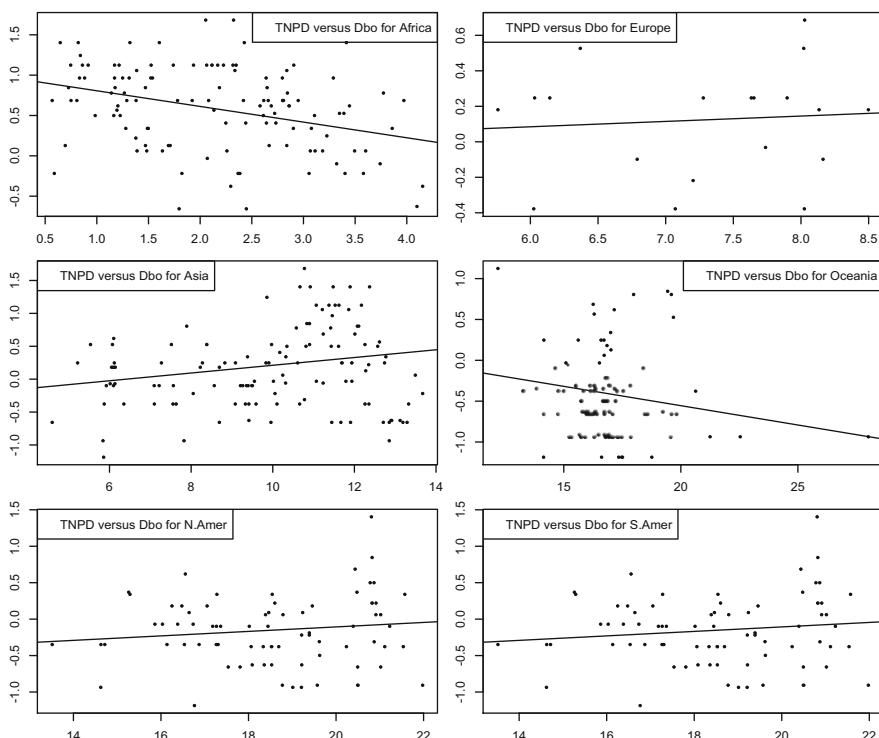


Fig. 8.3 Total phoneme diversity plotted against the distance from the best-fitting origin for each of the six linguistic groups separately

groups, the relation between inventory and distance is either negligible or positive. While the aggregate scatterplot in Fig. 8.1 seems to support Atkinson’s thesis, the disaggregated continent-by-continent plots paint a different picture. This is an instance of Simpson’s paradox for continuous data, with positive slopes on most continental subsets, but a negative slope overall.

8.5 Point Estimates and Confidence Regions

8.5.1 Simple Version

If we buy into the Out-of-Africa linguistic theory, it is natural to seek a region of plausible origins of language. In order to do this, it is necessary to know something about the statistical properties of the observations that are available. Independence of components is certainly not a reasonable assumption for this setting, but it is the easiest place to start, and it suffices to illustrate the method in its simplest form. For illustration, we assume that total phoneme inventory is related to population size and distance to the origin as follows

$$E(Tnpd_i) = \beta_0 + \beta_1 \|x_i - x_0\| + \beta_2 \log(\text{Popn}_i),$$

where x_0 is the origin and $\|x_i - x_0\|$ is the linguistic distance. Given the range of speaker populations, linearity in \log population size seems more reasonable than linearity in population size, which is in agreement with Atkinson’s principal analysis. For the moment, we assume also that the components are independent with constant variance σ^2 .

For arbitrary fixed origin, the model is linear in the remaining three regression parameters, so the least-squares estimates can be obtained in the standard way. Denote by $\text{RSS}(x)$ the residual sum of squares on $n - 3 = 501$ degrees of freedom for fixed x . The point \hat{x} that minimizes the residual sum of squares is the non-linear least-squares estimate. For these data, the minimum over the rectangular grid that covers Africa occurs at the south west corner, near 17°W, 35°S in the south Atlantic. However, the RSS function varies little throughout the bight of Africa, and is almost constant along the coast from Liberia to Cape Town. For this exercise, we restrict the parameter space to terra firma. The minimum over continental Africa occurs on the coast near the border between Angola and Namibia, roughly at 12°E, 17°S as indicated in Fig. 8.4, with $\text{RSS} = 143.02$. By the narrowest of margins, Atkinson’s fitted point on the Congo coast appears to be a local minimum with $\text{RSS} = 143.07$.

The standard recipe for the formation of a confidence set in non-linear least-squares problems uses a selected contour of the restricted residual sum of squares function $\text{RSS}(x)$ as the boundary. The residual mean square $s^2 = \text{RSS}(\hat{x})/(n - 5)$ serves as the variance estimate, which is distributed approximately as $\sigma^2 \chi_{n-5}^2$ under the stated assumptions. Moreover, s^2 is distributed approximately

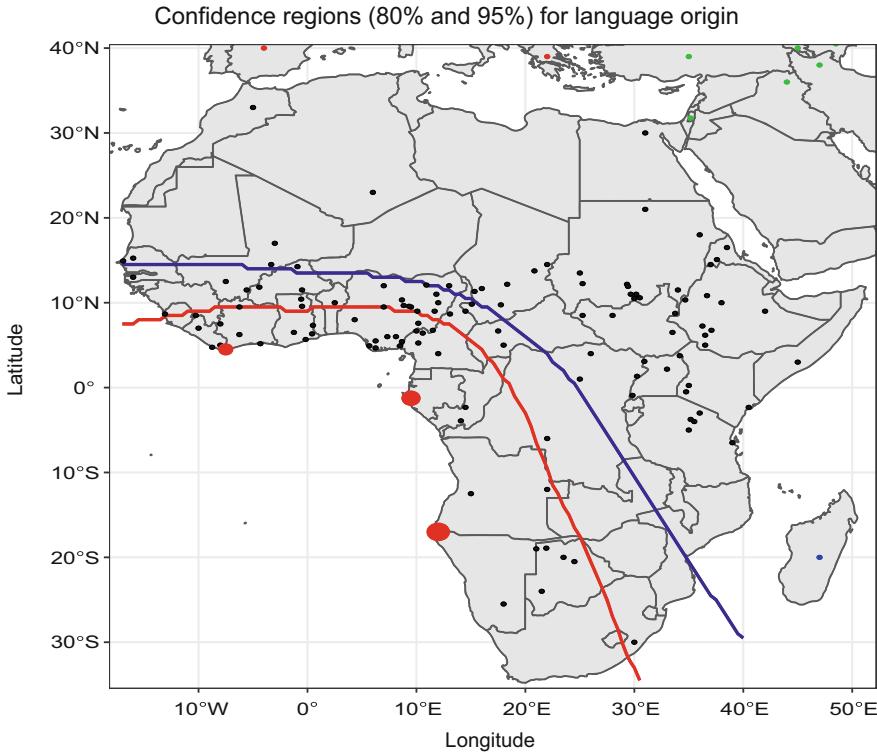


Fig. 8.4 Point estimate and confidence regions (80% and 95%) for the language origin, assuming independent observations. The RSS values for the three coastal marked points are 143.4, 143.1 and 143.0 in west-to-east order

independently of the difference $\text{RSS}(x_0) - \text{RSS}(\hat{x})$, which is distributed as $\sigma^2 \chi^2_2$ —on two degrees of freedom because the parameter space for sites is locally a two-dimensional manifold. Thus, the mean-square ratio $(\text{RSS}(x) - \text{RSS}(\hat{x}))/(\sigma^2)$ is distributed according to Fisher's $F_{2,n-5}$ distribution. Accordingly, the region

$$\left\{ x : \frac{\text{RSS}(x) - \text{RSS}(\hat{x})}{\sigma^2} \leq 2F_{2,n-5,\alpha} \right\}$$

is a $1 - \alpha$ confidence region for the linguistic origin. For $\alpha = 0.05$, the F -percentile is 3.01, so the right hand side is a little over 6.0. For reasons that are not explained, Atkinson uses a factor of four (BIC units) in place of six at this point, which gives only 86% coverage. Four BIC units is the 95% coverage factor for one degree of freedom, not for two.

Figure 8.4 shows the best-fitting origin at the Angolan-Namibian border, together with the 80% and 95% confidence regions computed according to the above formula. The 95% region includes most of western and southern Africa. If the

unrestricted maximum had been used, confidence regions at low confidence levels would cover water only, which is a difficult case for a linguist to make. However, the unrestricted 95% confidence region matches reasonably closely the restricted 80% region.

Shortcomings

The preceding analysis overlooks the fact that one of the regularity conditions fails. The restricted maximum occurs at a 1D-boundary point on the coast, so the local 2D manifold argument fails. If the linguistic hypothesis also stated that the origin must be a coastal point, the 1D argument would naturally prevail. But that is not a part of the thesis, so it seems preferable to use the more conservative 2D allowance. An alternative option is to resort to simulation—but that is neither an easy answer nor a satisfactory answer. In any event, there are more consequential effects that have so far been ignored.

8.5.2 Accommodating Correlations

In elementary statistical modelling, the difference between two means $E(Y_i) - E(Y_j)$ is associated with the difference $x_i - x_j$ between their recorded covariates: $\mu_i - \mu_j = (x_i - x_j)' \beta$. The difference between two covariances $\text{cov}(Y_i, Y_{i'})$ and $\text{cov}(Y_j, Y_{j'})$ is associated with the difference $R_{i,i'} - R_{j,j'}$ between their recorded relationships, usually but not necessarily in a linear manner: $\Sigma_{ii'} - \Sigma_{jj'} = (R_{ii'} - R_{jj'})' \tau$, where τ is a list of variance components.

For the current setting, the available covariates are the population size, continental group and the geographic location $i \mapsto x_i$. A relationship is a function on pairs of observational units, and the most obviously relevant relationships are inter-point distance $(i, j) \mapsto \|x_i - x_j\|$ and language family $(i, j) \mapsto F_{ij}$ as a Boolean matrix such that $F_{ij} = 1$ if i, j belong to the same family, and zero otherwise. These are both symmetric $n \times n$ matrices, so it is only natural that they should occur in the specification of the covariance matrix. For this project, we consider only the simplest additive model such that

$$\text{cov}(Y_i, Y_j) = \sigma_0^2 \delta_{ij} + \sigma_1^2 F_{ij} + \sigma_2^2 e^{-\|x_i - x_j\|/\lambda},$$

depending on three variance components and one range parameter λ . As it happens, linguistic distance is a little more effective than great circle distance, and for that choice the fitted range is $\hat{\lambda} \simeq 820\text{km}$. The linguistic family effect is not negligible, but the distance effect dominates. There are also linguistic sub-families, whose effects are not taken into account in this analysis.

The likelihood-based nominal 95% confidence region is the set of all candidate source points whose log likelihood is sufficiently high compared with the maximum, i.e.,

$$\{x : 2l(\hat{x}) - 2l(x) \leq \chi^2_{2,0.95}\}.$$

Here $l(x)$ denotes the profile log likelihood maximized over all other parameters. For the model suggested here, the 95% confidence region includes all of Africa except for a portion of lower Egypt; the 99% region also includes the Levant (Israel, Syria, Turkey, Jordan, Iraq) and all of Europe except for Russia and the Caucasus.

Generally speaking, failure to accommodate correlations has the effect of making the data seem more informative than they are, so the resulting confidence intervals are unrealistically narrow. Thus, it is no surprise that Fig. 8.4 is misleading in its portrayal of the strength of information in the data.

Three Points of Clarification

The code used for computing the log likelihood for one candidate point x_0 belonging to linguistic region $lregn$ has two parts:

```
ldist <- vdist(x0, lregn) # vector of distances from x0
fit <- regress(Tnpd~ldist+log(Popn), ~Fam+V, data=S1)
```

Here $S1\$Fam$ is the linguistic family coded as a factor, and V is a matrix with components $V_{ij} = \exp(-\|x_i - x_j\|/\lambda)$, which do not depend on x_0 . As it stands, this code is both computationally inefficient and technically incorrect on two counts. The efficiency can be improved substantially by including the optional argument $start=fit\$sigma$, which makes the previously-computed variance components available as the starting point for iteration.

The first technical issue is that the default likelihood function that is maximized by `regress()` is the REML likelihood for the observation Y in the space \mathbb{R}^n/\mathcal{X} of residuals modulo the subspace \mathcal{X} of mean values. Ultimately, our goal is to compute a likelihood ratio for one candidate center versus another, and the problem with the code as shown is that the mean-value subspace for one candidate point is not the same as the mean-value subspace for another candidate. The REML log likelihoods are not comparable as log likelihoods. In order for this to be done correctly, it is necessary to use the optional argument `kernel=K` to override the default kernel. While $K=0$ and $K=1$ are valid zero and one-dimensional options, the more natural choice is the two-dimensional intersection subspace `K <- model.matrix(~log(Popn), data=S1)`.

The second technical issue is that the log likelihood for a given x_0 should also be maximized over λ , which is a substantial computational overhead. For simplicity in the analysis described above, $\lambda = 820\text{km}$ has been treated as a known constant.

Shortcomings

An essential part of the Out-of-Africa thesis is that if x_0 is the linguistic origin, the regression coefficient of phoneme inventory on the linguistic distance vector $\|x_i - x_0\|$ must be negative. However, one piece of information (negativity of the coefficient) has not been used at any point in the analysis, and sign constraints have not been enforced in likelihood calculations—either by Atkinson or by me. For the rough calculations in the preceding section, the candidate points considered were restricted to existing linguistic centers in each region. In some cases, the weighted least-squares regression coefficient was positive. The fraction of negative coefficients varies considerably depending on which continental region x_0 belongs to:

| Region(x_0) | Africa | Europe | Asia | Oceania | N.Amer | S.Amer |
|-------------------|--------|--------|------|---------|--------|--------|
| Negative fraction | 1.0 | 1.0 | 1.0 | 0.87 | 0.05 | 0.00 |

Imposition of negativity constraints has no effect for candidate centers in Africa, Europe and Asia, but it must decrease the likelihood for some centers elsewhere. Given that the emphasis has been on Africa as the most plausible location, failure to impose the negativity constraint has a negligible effect on conclusions.

8.6 Matters for Further Consideration

8.6.1 Phoneme Inventory as Response

Suppose that it were possible to extract from the WALS database, the actual phoneme inventory of each language rather than the phoneme count. This statement implies a finite master list of m phonemes together with a Boolean variable $Y: [m] \rightarrow \{0, 1\}$ for each language indicating the subset of the master list that occurs in the given language. The phonemes may be labelled by type (vowel, consonant or tone), but the problem is already difficult enough without this added complication.

As it happens, the data are now available in phoneme-inventory form in the file `.../ch08/santoso.dat`. Details are given in Sect. 8.7.

Without altering notation, we may regard the phoneme inventory Y_i for language i either as a Boolean vector or as a subset $Y_i \subset [m]$ of the master list. Thus Y_i is the inventory for language i , the usual component-wise product $Y_i Y_j$ is the inventory common to a pair of languages, and the k -fold product $Y_{i_1} \cdots Y_{i_k}$ is the inventory common to a specific subset of k languages.

Setting aside the complication of phoneme type, it is mathematically natural to ask for an analysis that is invariant with respect to re-labelling of phonemes in the master list. That condition implies an analysis that depends only on phoneme inventory counts

$$i \mapsto \#Y_i, \quad (i, j) \mapsto \#(Y_i Y_j), \quad (i, j, k) \mapsto \#(Y_i Y_j Y_k)$$

and so on. The analyses presented in this chapter use only the n -vector of first-order counts. But inventory data also provide symmetric $n \times n$ matrices of second-order counts, symmetric tensors of third-order counts, and so on.

Geography and distance are essential components in the Out-of-Africa hypothesis. What bearing does the hypothesis have on data collected in inventory format? Each language i is associated with a geographical location x_i ; each pair may be associated with a pair of points $\{x_i, x_j\}$, with a line segment (x_i, x_j) or with a weighted centroid; each triple may be associated with a set of points, the convex hull of those points, or with their centroid, and so on. How is distance to be measured for singletons, pairs, triples and so on? How are the questions to be formulated statistically? Given answers to these questions, how might the analysis proceed to estimate relevant parameters and to check whether the data are consistent with the hypothesis?

8.6.2 *Vowels, Consonants and Tones*

If it is taken at face value, the *Out-of-Africa* hypothesis applies equally to vowels, to consonants and to tones. Is the evidence from these three sources consistent? This is something that can be checked by analyzing the three variables separately. We leave it as an exercise for the reader.

8.6.3 *Granularity*

In our analysis, we have ignored the fact that phoneme inventory variables are discrete, with only a few distinct values. Atkinson used three equally-spaced numbers to represent normalized tone inventory, three for vowels and five values equally spaced for normalized consonant diversity. Total normalized phoneme diversity is the sum of these three. What effect does granularity have on conclusions derived from a Gaussian model?

8.7 Follow-Up Project

8.7.1 Extended Data Frame

For her Masters project at the University of Chicago, Josephine Santoso compiled from *phoible.org* and other sources a more extensive phoneme inventory of 1277 world languages, of which 1028 have speaker populations of at least 50. This is a little more than twice as many languages as Atkinson used. The Santoso file contains all of the variables listed in Sect. 8.2 plus several others. She has kindly agreed to make her compilation available in the file `.../ch08/santoso.dat`.

Although Santoso's file lists the actual set of phonemes of each type, her analysis focuses solely on phoneme counts for languages having more than 50 speakers. Two phonemes that are distinct in one dialect are not necessarily distinct in another, so phoneme counts for any language are subjective to a certain extent, particularly those for vowels and consonants. Where Atkinson's main analysis uses total *normalized* phoneme inventory as described earlier, Santoso's main analysis uses the total phoneme count without normalization. If Y_1 , Y_2 , Y_3 are the phoneme counts for vowels, consonants and tones respectively Santoso's response variable is $Y_1 + Y_2 + Y_3$, whereas Atkinson's normalized combination is closer to

$$\frac{Y_1}{s_1} + \frac{Y_2}{s_2} + \frac{Y_3}{s_3},$$

where s_1 , s_2 , s_3 are the three standard deviations. For the new compilation, the means and standard deviations are as follows:

| Type | Vowel | Consonant | Tone |
|--------|-------|-----------|------|
| Mean | 10.81 | 24.06 | 0.74 |
| Median | 10.00 | 22.00 | 0.00 |
| SD | 6.33 | 11.35 | 1.60 |

Where Santoso's combination favours consonants as the dominant phoneme type, Atkinson's combination favours tones, which are present in a small minority of languages, mainly in central Africa and East Asia. Nonetheless, Atkinson's Out-of-Africa thesis applies non-preferentially to all types, so the total phoneme count is not unreasonable as a response.

So far as methodology is concerned, the particular selection of response is immaterial. Thus, the models and fitting methods are the same as those described earlier. There may, however, be small differences in the way that distances are computed. For example, there may be more choke points between Africa and Europe or between Europe and Asia. Despite the similarity in response and statistical technique, Santoso's conclusion is very different from Atkinson's. For total phoneme inventory, her analysis finds the best-fitting origin not in Africa, but in western Europe—in the west of Ireland, to be precise. When correlations are

taken into account, however, almost any point in Europe or Africa fits equally well as an origin.

Given the genetic and paleoanthropological evidence of migrations out of Africa as early as 200k years ago, with later migrations 70–50k years ago, the argument that traces of those migrations ought to remain in extant languages has a certain degree of plausibility. Arguably, the data contain some broad-scale evidence for it. But the linguistic evidence on its own is not at a level that enables us to point to a language origin specifically in Africa.

8.7.2 An Elementary Misconception

Let l be a language that employs c consonants, v vowels and t tones. From the perspective of a mathematician who is familiar only with European languages, it is natural to imagine that every syllable or every word is a sequence of vowels and consonants that can be enunciated in one of t tones. Each tonal enunciation is a different word and a different concept. This Cartesian-product framework does not imply that the entire set of $(v + c)t$ tonal variations occurs in the language. However, the function `table(...santoso$num_tone)` applied to Santoso's spreadsheet produces the following counts for languages having more than 50 speakers:

| # Tones | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|-------------|-----|---|----|----|----|----|---|---|---|---|----|
| # Languages | 870 | 1 | 48 | 79 | 57 | 25 | 8 | 5 | 1 | 2 | 2 |

In other words, the number of tones ranges from zero to ten. Most languages have a zero tone count, and only one has $t = 1$. If the preceding interpretation were accurate, the set of phonemic sounds available would be empty for nearly 80% of all languages. Clearly, this cannot be a correct interpretation.

It is evident that non-tonal languages are coded as $t = 0$. The singular single-tone language (Dutch) could be a coding error, but possibly not; some Franconian dialects are described as pitch-accented, which is not the same as tonal. Where a mathematician would naturally opt for the adjective 'monotonal', meaning one tone, the more common adjective 'atonal' is interpreted literally and coded as zero tones. As a mathematician, my instinctive preference is the code $t = 1$ for atonal languages, and I confess that I had taken for granted that everyone else would see it the same way. Regrettably, misconceptions of this sort are not uncommon in statistical consulting! Ultimately, the choice of one code over the other is a matter to be settled by subject-matter specialists. Since re-coding is not an additive constant, it does have a bearing on the analysis for total phonemes, normalized or otherwise.

8.8 Exercises

8.1 Use the `table()` function to extract the distinct values for vowel diversity, consonant diversity, tone diversity and total phoneme diversity. What does this tell you?

8.2 Use the function `cov(cbind(S1[...]))` to compute the sample covariance matrix of the four phoneme inventory variables. What does this tell you?

8.3 Use the function `qr(cbind(S1[...]))$rank` to deduce that total phoneme diversity is a linear combination of the three constituents. Find the coefficient vector.

8.4 The function `vdist(x0, 1)` returns a list of linguistic distances from the designated point x_0 in linguistic region 1 to each of the 504 language locations. Show that Atkinson's distance variable $S1\$Dbo$ implies that his best-fitting origin lies somewhere in the box $9\text{--}10^\circ\text{E}$, $1\text{--}2^\circ\text{S}$. Find the point and locate it on a map.

8.5 Assuming the Out-of-Africa hypothesis, total phoneme inventory necessarily depends not just on distance to the origin but also on the speaker population size. By minimizing the residual sum of squares over continental Africa, find the best-fitting origin under the linearity assumption

$$E(Tnpd_i) = \beta_0 + \beta_1 \|x_i - x_0\| + \beta_2 \log(\text{popn}_i).$$

You should not assume that the best-fitting origin lies in or near the box $9\text{--}10^\circ\text{E}$, $1\text{--}2^\circ\text{S}$.

8.6 Which language has the greatest vowel inventory in the Santoso compilation, and which has the least? Which language has the greatest consonant inventory, and which has the least?

Chapter 9

Environmental Projects



9.1 Effects of Atmospheric Warming

9.1.1 *The Experiment*

This project concerns an experiment conducted at two sites in Minnesota over the period 2009–2011 to determine the effects of climate warming on photosynthesis in juvenile trees of 11 different species. The following excerpt is taken from the data archive:

To test how climate warming and variation in soil moisture supply will jointly influence photosynthesis of southern boreal forest tree species we measured gas exchange rates of 11 species in an open-air warming experiment at two sites in northern Minnesota, USA. The experiment ran for three years and used juveniles of 11 temperate and boreal tree species under ambient and seasonally warmed (+3.4°C above- and below-ground) conditions. We measured in situ light-saturated net photosynthesis (A_{net}) and leaf diffusive conductance (gs) on numerous days across the three growing seasons. Soil and plant temperatures and soil moisture were continuously measured from sensor arrays.

Details of the experimental design, the site preparation, the species selected, the variables measured, and so on, are provided in the paper (Reich et al., 2018).

The two sites are roughly 100 miles apart, each site consisting of 12 plots arranged in three blocks of four plots. Each plot is a circular area, roughly three metres in diameter, which is adequate for 30–40 juvenile specimens. Plots in the same block are sufficiently far apart to avoid treatment interference. In other words, the separation is sufficient to ensure that the treatment applied to one plot has negligible effect on neighbouring plots.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-14275-8_9.

Several specimens of each species were planted in each plot. The heat treatment was applied to plots during the growing season only. On 50 days, roughly 15–18 days per year from mid-June till late September, measurements were made on trees in several plots. For administrative reasons, all measurements on one day occurred at the same site. On these days, soil water content was recorded for each plot together with several measurements (photosynthesis, conductance, vapour pressure gradient, temperature,...) on selected trees in each plot. Each photosynthesis and vapour-pressure measurement was made on a single leaf.

The main thrust of the paper is that atmospheric warming has two principal effects that are relevant to tree growth. One is the direct effect on leaf transpiration and photosynthesis. The other is the effect on soil moisture content. Higher temperatures at these latitudes generally means that soils are drier than they would otherwise be. The moisture deficit has an important effect on how trees respond to warmer temperatures.

9.1.2 *The Data*

The data reported in the paper are available from the Environmental Data Initiative at

<https://doi.org/10.6073/pasta/258239f68244c959de0f97c922ac313f>

For present purposes, the local file

.../ch09/borealwarming.csv

consists of a 2049×18 spreadsheet containing the data to be used for the exercises listed below. Each row consists of several measurements on one leaf on one day. A few minor discrepancies were noted and corrected, so this file differs slightly from the archived data.

Photosynthesis, conductance and vapour-pressure are measured on leaves from selected trees. Each leaf or each leaf-plot-occasion is one observational unit for all such variables, and the sample contains 2049 units of this type, one for each row in the spreadsheet. A few values are missing for some units. Responses for distinct units in the same plot may be correlated positively. Likewise for distinct units in the same block or the same site. Responses may also be correlated spatially and temporally, but spatial information is meagre, so this avenue of investigation is limited.

Soil moisture content is not measured on leaves: it is measured on plots, with at most one measurement for each plot on each visit to the site. Each plot-occasion pair is one observational unit, and the sample contains approximately 503 observational units for this variable. Two soil moisture values are missing. Responses for distinct units may be correlated spatially or temporally or in blocks.

The questions that follow are intended as a guide for analysis in an examination setting. You should first read the *Nature* paper and then answer the questions as

asked. But you are not restricted to the points mentioned below; you are free to examine the data in any way you please, so your approach might not follow the path suggested.

9.1.3 Exercises

1. Before any trees were planted, a grid of underground electrical cables was laid down at a depth of 15 cm in each plot, with sufficiently small separation that the heating effect could be deemed uniform. The cables in the treated plots were used as heating elements. What was the purpose of the cables in the control plots?
2. Ten of the variables in the file `borealwarming.csv` are as follows:

```
site, block, warming, plot_id, species,  
plant_id, year, date, Asat, soil_moisture,
```

where `Asat` is a photosynthesis measurement. According to the definitions in Sect. 11.3, one of these is a pre-baseline classification factor, two are pre-baseline quantitative variables, four are pre-baseline block factors, one is an intervention or treatment factor, and two are post-baseline responses. Indicate which is which.

3. Use the data to reproduce the authors' plots in Figs. 1 and 2 in a similar format. Show your code for Fig. 1. What, if anything, do these plots tell you about the effect of elevated temperature on deciduous versus coniferous trees?
4. Soil water content is expected to vary from day to day with the most recent weather and from plot to plot depending on the topography, for example, exposure, topsoil depth, drainage capacity of the sub-soil, and so on. You are asked to examine the effect of treatment on the soil water content taking appropriate account of such variations. You may assume initially that the treatment effect is constant over sites and over years. For purposes of this analysis, you may set aside missing response values and ignore all leaf-specific variables.

Justify your selection of terms in a suitable linear Gaussian model with soil water content as response. Explain how you fitted the model, and show the parameter estimates with standard errors. How many observational units do you have to address this question?

5. You should have found a small negative treatment effect in the preceding question. Is the treatment effect constant across sites as assumed in part 3? Is it constant across years? Explain how you might address these questions, and report your answers.
6. This question is concerned with the red-oak subset of the data, which is coded as level `queru` for the factor `species`. You are asked to analyze the relation between photosynthesis (`Asat`) and other non-leaf variables including warming treatment and soil water content. Your analysis should accommodate block, plot and temporal effects as needed. Give a brief summary of the conclusions reached on the basis of the fitted model.

9.2 The Plight of the Bumblebee

9.2.1 Introduction

Bees serve an essential function as pollinators for a very wide range of flowering plants, from fruit trees to soy beans to rapeseed plants. Over the past several decades, numerous articles have appeared in the press pointing to an alarming decline in bee survival and the dire consequences for agricultural production. In some of the more alarmist articles, the risk to honey bees is attributed to African bees or killer bees; in others, the risk is put down to habitat loss, climate change, pesticides, pathogens and parasites. All articles paint a bleak outlook, not only for bees, but for humanity as well.

Adler et al. (2020) designed and analyzed a number of experiments whose overall goal was to study the effect of a particular pathogen *Crithidia bombi* on the behaviour and survival of the bumblebee *Bombus impatiens*. For reasons unknown, it appears that some plant species are more effective than others at pathogen transmission. The following quote is taken from an introductory passage.

The role of plant species in shaping infection intensity could be influenced by bee behavior. If infected bees increase visitation to antimicrobial plant species as a form of self-medication, such plant species could play a larger role than predicted in disease dynamics. Alternatively, antimicrobial plant species may be less effective than expected if pathogens manipulate host behavior to avoid such plants. Sunflower has pollen that dramatically reduces [the pathogen] *C. bombi* in *B. impatiens* and several plant species produce nectar with secondary compounds that can reduce pathogens, although such effects are not always consistent. Only a few studies have assessed whether infection alters bee preference. In the field, infected *B. impatiens* and *Bombus vagans* had greater preference than uninfected bees for inflorescences with high nectar iridoid glycosides that can reduce pathogen infection. However, a laboratory study with *Bombus terrestris* found only weak evidence that infected bees had increased preference for nectar nicotine compared to uninfected bees. Thus, there are conflicting results across species and compounds, and very few data overall to assess whether infection changes foraging preferences.

We consider here only the first of the experiments on bumblebees by Adler et al. (2020). The authors used microcolonies of approximately 15 workers each. To confine the bees, and to force them to forage only on the flowers provided, each microcolony was housed in a tent, which contained foraging plants.

9.2.2 Risk of Infection

The first experiment was designed to study the risk of infection by the pathogen *C. bombi*, and how the risk varies with flower type. Each experimental unit was one microcolony or one tent, so the flower type was assigned to tents. The control level consisted of canola (rapeseed) alone; for the other two levels, some of the canola pots were replaced either with flowering plants that had previously been deemed to

be high-transmission or with plants that had been deemed low-transmission. Thus, the three treatment levels were labelled *canola*, *low* and *high*, and the primary goal was to compare the infection risk for one level versus another. For the part of the experiment considered here, all bees were initially infected.

The experiment was designed in five replicates or rounds, each round lasting two weeks. In each round, nine tents were available labelled in three blocks of three. Ordinarily, the chief reason for labelling units in blocks is that two units in the same block are substantially more similar than two units in different blocks. In this instance, it appears that all tents were located on one site at the Amherst Centre for Agriculture, so it is not clear that two tents in different blocks are appreciably less similar than two tents in the same block. In that case, the mean square variation between blocks would be no greater than the mean square within blocks. While we should not expect the block variance to be substantial, it costs little to retain the information provided.

The infection status of each bee was determined at the end of the period, so each observational unit is one bee. To each bee u there corresponds a round $r(u)$, a block $b(u)$, a physical tent $s(u)$, a colony $c(u)$, and a Bernoulli outcome $Y(u) \in \{0, 1\}$. There are five rounds, three physical blocks, nine tents per round, and 45 colonies. To each colony there corresponds a treatment level t_c . Thus, the set of colonies is in one-to-one correspondence with the set of round-tent pairs.

Since all bees in one colony are workers (female) and there is nothing to distinguish one bee from another, the infections were reported as totals, tent by tent in each round. Given that each colony consisted initially of approximately 15 bees, and the totals at the end of each round range from one to 18, it would appear that there must have been substantial mortality in some of the tents. The spreadsheet does not report the initial count, the paper does not discuss bee mortality rates, and mortality does not enter into the authors' analysis.

The spreadsheet contains 45 rows, one of which is missing due to a tent collapse. The first few entries are

| # Infected | # Uninfected | Round | Block | Treat |
|------------|--------------|-------|-------|---------------|
| 9 | 6 | 1 | 1 | <i>low</i> |
| 5 | 2 | 1 | 1 | <i>high</i> |
| 12 | 3 | 1 | 1 | <i>canola</i> |
| 9 | 0 | 1 | 2 | <i>canola</i> |
| 4 | 1 | 1 | 2 | <i>high</i> |
| 4 | 7 | 1 | 2 | <i>low</i> |
| : | | | | : |
| 3 | 6 | 5 | 3 | <i>canola</i> |

It is unclear whether the same nine tents were used in each round, but it is reasonable to presume so, and this has subsequently been confirmed.

To estimate the treatment effect, the authors used a generalized linear mixed model of binomial type as follows. Given the round and block effects, $\eta_{r,b}$, treated

as random variables, infections occur independently as a Bernoulli process with probabilities satisfying

$$\text{logit } \text{pr}(Y_u = 1 | \eta, \mathbf{t}) = \eta_{r,b} + \tau_{t(c)} \quad (9.1)$$

depending on the treatment $t(c)$ assigned to colony $c = c(u)$ to which u belongs. It follows that the number of infections in colony c

$$\#\{u : Y(u) = 1, c(u) = c\}$$

is a sum of conditionally independent and identically distributed Bernoulli variables, with parameter (9.1). Thus, the sum is conditionally binomial with index equal to the colony size, and logit parameter (9.1).

In the generalized linear mixed model, the colony counts are not unconditionally binomial, and the counts for different colonies are not independent. The joint distribution of counts for all 45 colonies is obtained by integration with respect to the distribution of η . The expression `round/block` in a model formula denotes a convex combination of two nested block factors, `block:round` with 15 levels, and `round` with five levels. The authors' `glmer()` model formula

```
Ytot ~ treat + (1 | round/block)
```

implies that the non-binomial random component $\eta_{r,b}$ is a sum of two independent processes, one having independent and identically distributed components for each `block:round` level, the other having independent and identically distributed components for each `round`. Equally important, η is constant within each block in each round.

Part of the `glmer()` output shows the two fitted variance components plus the estimated treatment effects with `canola` as the reference level.

```
Random effects:
 Groups      Name        Variance Std.Dev.
 block:round (Intercept) 0.0288   0.1698
 round       (Intercept) 1.3883   1.1783
 Number of obs: 44, groups: block:round, 15; round, 5

Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.7102    0.5771   1.231   0.22
treatmentHigh 0.5135    0.3144   1.634   0.10
treatmentLow   0.2518    0.3016   0.835   0.40
```

The two variance components indicate negligible between-blocks variation in each round, but more substantial variation from one round to another. Since the treatment reference level is coded as zero, the estimated infection rate for `canola` is lower than the infection rate for the other treatments. But none of the observed treatment differences comes close to statistical significance, so there is nothing here to suggest that the pathogen transmission rates differ from one plant type to another.

9.2.3 Mixed Models

The generalized linear mixed model described above is a development of recent vintage. The two-stage construction is mathematically straightforward, but maximum-likelihood estimation is computationally non-trivial. General-purpose computational packages have become available only in the past 25 years. The package `glmer()` uses the Laplace approximation for Gaussian integrals; a few others use Monte Carlo methods.

This book does not concern itself with computational tactics, but with broader strategic matters. In that respect, there is a small non-obvious error or misunderstanding in the form of the model used here. To be clear, I do not mean that the computations are in error or that the conclusion as stated is biologically incorrect. I mean only that there is a non-trivial methodological error that could, under different circumstances or with different data, give rise to misleading conclusions. It is worth revisiting the analysis in order to understand the motivation for the model and the source of the error.

The fact that the response is binary rather than a quantitative measurement is immaterial as a matter of principle. The fact that the expected value is not linear in treatment effects is also immaterial. How would the analysis have proceeded had the response $Y(u)$ on each of 389 bees been quantitative?

Standard practice recommended in every textbook on experimental design calls for a distinction between observational units and experimental units. This design has 389 observational units (bees), and 44 experimental units (colonies). Typically, the variation between observational units within the same experimental unit is less than the variation between observational units in different experimental units. In order for the variance of treatment contrasts to be estimated honestly from variation *between* experimental units rather than variation *within*, the experimental-unit factor must be included as a term in the covariance model: see chapter 1 for a simple instance. As it happens, the original `round:block` factor is not needed, so it is best to replace `(1 | round/block)` with `(1 | round) + (1 | colony)`. In effect, the 15-level block-factor `block:round` is replaced by the 44-level factor `colony`.

The output from `lmer()` after this substitution is as follows:

| Random effects: | | | | | |
|---|-------------|------------|----------|----------|--|
| Groups | Name | Variance | Std.Dev. | | |
| colony | (Intercept) | 0.2119 | 0.4603 | | |
| round | (Intercept) | 1.4924 | 1.2216 | | |
| Number of obs: 44, groups: colony, 44; round, 5 | | | | | |
| Fixed effects: | | | | | |
| | Estimate | Std. Error | z value | Pr(> z) | |
| (Intercept) | 0.6925 | 0.6100 | 1.135 | 0.26 | |
| treatmentHigh | 0.6006 | 0.3788 | 1.586 | 0.11 | |
| treatmentLow | 0.2923 | 0.3629 | 0.806 | 0.42 | |

In this instance, the between-colony variance component is not sufficiently large to alter the substantive biological conclusion that there is no evidence that the risk to initially infected bees depends on the species of foraging plant.

Subsequent to this analysis, the authors pointed out that all bees in the same block come from the same parental lineage. To that extent, pairs of bees in the same block have something in common that pairs of bees in different blocks do not. Any systematic differences associated with lineage can be accommodated by including lineage as a block factor, corresponding to either `block` if lineage is constant over rounds, or `block:round` otherwise. But there is little evidence in the data that lineage plays much role in pathogen transmission.

9.2.4 Exchangeability

The argument in the preceding section for including the experimental-unit factor `colony` is a consequence of exchangeability associated with a group that is implicit in the design. Other factors such as `round` may be included as needed, either as additive fixed effects contributing to expected values, or as block factors contributing to variances and covariances. But the inclusion of `colony` as a block factor is a matter of principle; it is not a matter to be decided based on of goodness of fit or model adequacy.

To understand the reasoning behind this attitude, it is helpful to rearrange the spreadsheet by observational units, i.e., with one row for each bee. The first 38 rows of the expanded file `ExpSS` are as follows:

| Bee | <i>Y</i> | Colony | Round | Block | Treat |
|-------|----------|--------|-------|-------|---------------|
| 1–6 | 0 | 27 | 1 | 1 | <i>low</i> |
| 7–15 | 1 | 27 | 1 | 1 | <i>low</i> |
| 16–17 | 0 | 21 | 1 | 1 | <i>high</i> |
| 18–22 | 1 | 21 | 1 | 1 | <i>high</i> |
| 23–25 | 0 | 25 | 1 | 1 | <i>canola</i> |
| 26–38 | 1 | 25 | 1 | 1 | <i>canola</i> |

It is understood that the labelling of bees within a colony is entirely arbitrary, and carries no substantive information. The labelling shown above is such that the lower-numbered bees in each colony are disease-free. Thus, the record for bee 1 is repeated for bees 1–6, the record for bee 7 is repeated for bees 7–15, and so on.

Despite the fact that the expanded spreadsheet has 389 rows while the original has only 44, it is apparent that the spreadsheets are entirely equivalent in their information content. Thus, the output of the `lmer()` expressions

```
lmer(Y~treat+(1|round/block), family=binomial, data=ExpSS)
lmer(Y~treat+(1|round)+(1|colony), family=binomial, data=ExpSS)
```

is identical in all important respects to the outputs shown in Sects. 9.2.2 and 9.2.3. In particular, all parameter estimates and standard errors are identical.

Consider now a pair of bees (u, u') with response $(Y_u, Y_{u'})$. In both of the models implied by the `lmer()` expressions shown above, the bivariate random variable (Y_1, Y_2) has the same distribution as $(Y_u, Y_{u'})$ for all pairs of distinct bees in colony 27. Likewise, (Y_{16}, Y_{17}) has the same distribution as $(Y_u, Y_{u'})$ for distinct pairs in colony 21. The responses are exchangeable within colonies. Because of different treatments applied to colonies, they are not exchangeable for different colonies either in the same round or in different rounds.

In the null versions with treatment effect omitted, both `lmer()` expressions imply that (Y_1, Y_2) also has the same distribution as the pairs (Y_{16}, Y_{17}) and (Y_{23}, Y_{24}) in colonies 21 and 25. Two pairs of units have the same joint distribution if the first pair belongs to one colony and the second pair belongs to another colony. However, the first `lmer()` expression without treatment also implies also that all observations Y_1, \dots, Y_{38} in round 1 are exchangeable. Thus $(Y_u, Y_{u'})$ has the same distribution as (Y_1, Y_2) for all distinct pairs in the same round whether they belong to the same colony or not.

Since we are dealing with an infectious disease, it is entirely possible that a part of the observed or anticipated correlation is associated with bee-to-bee transmission. In that case, there is every reason to expect that an infected bee is more likely to transmit the disease to a sister bee in the same colony than to a sister bee in a different colony with which it has no direct contact. Or, to put it the other way round, there is little reason to expect intra-colony transmissions to occur at the same pairwise rate as inter-colony transmissions in the same physical block, which is what the first version of `lmer()` implies.

It is not always easy to understand the implications of a complicated stochastic model. When it is spelled out this way, it is evident that forced block exchangeability must have been unintentional on the authors' part. Fortunately in this instance, it does not substantively alter the conclusions.

9.2.5 Role of GLMs and GLMMs

The principal purpose of a GLMM is to accommodate effects that are associated with baseline block factors such as `colony`, `block` and `round`. Treating such effects as independent random variables is most effective in situations where the number of levels is large, and exchangeability is plausible.

For the present setting, a reasonably strong case can be made for a much simpler analysis using a beta-binomial model with independent components for each colony. Alternatively a generalized linear model of binomial type may be used, provided that an over-dispersion factor is included to accommodate extra-binomial variation.

In either case, the model for the mean must include round effects in addition to treatment effects. The code used to produce the output shown below is

```
fit <- glm(Ytot~treat+round, family=binomial(), . . .)
summary(fit, dispersion=X2/resdf)
```

where Y_{tot} is the 44×2 matrix of infection counts for each colony, $X_2 = 49.56$ is Pearson's statistic, $resdf=37$ is the residual degrees of freedom, and $X_2/resdf = 1.34$ is the Lexis dispersion ratio.

| | Estimate | Std. Error | z value | $Pr(> z)$ |
|--|----------|------------|---------|------------|
| (Intercept) | 0.5506 | 0.3357 | 1.640 | 0.10 |
| treatmentHigh | 0.5331 | 0.3672 | 1.452 | 0.15 |
| treatmentLow | 0.2699 | 0.3518 | 0.767 | 0.44 |
| (Dispersion parameter for binomial family taken to be 1.339) | | | | |

Once again, this analysis shows no evidence of a treatment effect on infection rates. Pearson's statistic is 49.56 on 37 degrees of freedom, so the dispersion parameter is not especially large, and the evidence for overdispersion or correlation within colonies is not overwhelmingly strong. Nevertheless, given that colony is the experimental unit, it is essential to include the dispersion factor whether it is significantly greater than one or not.

9.3 Two Further Projects

Project I: Phenology refers to the study of cyclic and seasonal natural phenomena, especially in relation to climate and plant and animal life. The paper by Montgomery et al. (2020) is a continuation of the project discussed in Sect. 9.1. It aims to study plant phenology in response to atmospheric warming. The bud-burst dates for ten tree species over five years are available online. Is this an experiment or an observational study? If it is an experiment, what is the treatment? What do the data have to say about the change in budburst dates for each species over the five years?

Project II: Li et al. (2019) studied the effect of atmospheric warming on the amount and the blend of volatile organic compounds released by the shrub *Betula nana* in response to herbivory by moth larvae at a latitude 68°N site in the Swedish tundra. The data are available online.

What are the observational units? What are the experimental units? What role does methyl jasmonate play in the design? How do your conclusions align with those in the paper?

9.4 Exercises

9.1 Check that the authors' `lmer()` code for the bee infection experiment produces the output shown in Sect. 9.2.2. Check that the revised `lmer()` code produces the output shown in Sect. 9.2.3.

9.2 Expand the spreadsheet so that it is indexed by bees rather than by colonies. Check that the two versions of the `lmer()` code produce the same output for the expanded spreadsheet as they do for the compact format.

9.3 The analysis of bee infection rates in Sect. 9.3 is only one of many similar analyses reported by Adler et al. (2020). A subsequent analysis examines how the mean infection intensity per colony is related to treatment. The authors use a Gaussian random-effects model with `round/block` as the additive random effect. Since mean infection is a positive number with an appreciable dynamic range, check first whether a transformation might be helpful to improve additivity. On the transformed scale, check whether there is a treatment effect. Which treatment level has the lowest mean infection, and which has the highest?

9.4 McCullagh and Nelder (1989, Sect. 14.5) describe an experiment by S. Arnold and P. Verrell on the interbreeding of southern Appalachian mountain dusky salamanders. Each male has six breeding opportunities and each female also has six opportunities. The incomplete crossed design is shown in Table 14.3, and the results for three experiments in Tables 14.4–14.6 of McCullagh and Nelder (1989). Use `lmer(...)` to estimate the male and female variance components for the first salamander-mating experiment. Is there any evidence of correlation for repeat observations on the same male? Is there any evidence of correlation for repeat observations on the same female?

9.5 All of the analyses of infection rates in Sect. 9.2 are for infections among *surviving* bumblebees, with the implicit assumption that the survival distribution does not depend on infection status. Suppose that the initial colony size were available, so that the survival fraction could be obtained by subtraction. How might you use the additional information to determine the effect of treatment on survival?

Chapter 10

Fulmar Fitness



10.1 The Eynhallow Colony

10.1.1 Background

The northern fulmar *Fulmarus glacialis* is a seabird found mainly off the coasts of Iceland, the British Isles and parts of Norway. Fulmars have a life expectancy of 30–40 years, and some may live up to 60 years. Adults are monogamous, they form long-term pair bonds, and breeding pairs return to the same nest site year after year. Breeding begins in May; a single egg is laid and incubated by both parents for about 50 days. The chick is brooded for about two weeks and fully fledged after about three weeks.

A survey of the Eynhallow fulmar colony in the Orkney Islands was started in 1951 by Robert Carrick and George Dunnet; the data for this exercise concern the breeding record of 428 adult female birds for the period from 1958 to 1996. They were provided by Steven Orzack in 2006. A few of the birds were active breeders when first observed, but most were observed annually from their first attempts at breeding until 1996 or the bird's presumed death. No single bird was alive for the entire period of observation. Further details concerning the Eynhallow population can be found in Orzack et al. (2011).

According to the Wikipedia article, fulmars have a long adolescence and commence breeding at around 6–12 years of age. The records available for this project include 62 females that were born at Eynhallow and subsequently nested there at age two or later. These individuals were marked as fledglings, so their ages were known. Their nesting commenced at a wide range of ages from two to ten, with a mean of 4.65 and standard deviation 2.46.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-14275-8_10.

10.1.2 The Eynhallow Breeding Record

Every pair of birds breeding at Eynhallow during the period 1958–1996 was recorded. The record for each female in each year shows whether the bird laid an egg, and if so, whether the egg hatched, and, if hatched, whether the chick fledged, which is the event of primary interest. The data are presented in the file ...Fulmar.dat in mark-recapture format in which the first column is the bird identifier and the next 38 columns describe the reproductive event observed for that bird during each of the 38 years of the study. Each record includes the letter ‘A’, in the year that the bird was first observed, either directly if the bird fledged at Eynhallow, or indirectly if it fledged elsewhere and subsequently nested at Eynhallow. All birds born at Eynhallow are marked as fledglings. If the bird fledged at Eynhallow, ‘A’ indicates the year of fledging. However, the great majority of birds nesting at Eynhallow are blow-ins that were born elsewhere. If the bird was first observed nesting at Eynhallow, ‘A’ is the year before the first nesting. For such blow-ins, the year of birth is not known.

The code for reproductive events is:

- 0: no reproductive event observed;
- 2: egg laid;
- 3: egg hatched;
- 4: chick successfully fledged (event of primary interest).

The records for birds 116, 209, 280 and 284 are as follows:

| | |
|-----|---|
| 116 | A00444224004040000000000000000000000000000000 |
| 209 | A200433444440000000000000000000000000000000000 |
| 280 | 00000A00000032400000000000000000000000000000000 |
| 284 | 00000A4000002404043444444440440020002 |

Bird 116 was first recorded and marked in 1958 as a fledgling at Eynhallow. It was next observed as a breeding adult in 1961, when it produced a fledgling (code 4). The last sighting occurred at Eynhallow in 1971, when it produced another fledgling. Bird 209 was first observed as an adult in 1959, when it laid an egg that did not hatch. Although it was first marked in 1959, ‘A’ is inserted into the record for 1958; the year of birth is pre-1958. Likewise, bird 280 fledged at Eynhallow in 1963 and was next recorded nesting there in 1970, ’71 and ’72. Bird 284 was first recorded in 1964 as a nesting adult, so ‘A’ is inserted into the record for the previous year. This bird was not present at Eynhallow for the following six years. Given that fulmars are faithful to their nesting place, it is presumed by ornithologists who study these matters that no breeding occurred in those years either at Eynhallow or elsewhere.

A leading zero in the sequence for one bird means that no nesting event occurred in that year; the bird is classified a juvenile. A trailing zeros mean that the bird was not observed to be nesting, either because it had died or because it was no longer sexually active. An internal zero means that the bird was alive and capable of breeding, but did not nest in that year.

10.1.3 The Breeding Sequence

The breeding sequence for one bird is the subsequence beginning at the first non-zero value and ending at the last non-zero value. The ornithological rationale for dropping the leading zeros is that the bird is either not yet born or a juvenile. Since it ultimately nests at Eynhallow, and it is presumed to be faithful to its nest, the leading zeros imply that it is non-breeding in those years. In the case of trailing zeros, we know only that the bird was not seen at Eynhallow. If it is still alive, it is presumed not to be an active breeder in those years; at some point it must be presumed dead. The breeding sequence is restricted by design to the years 1958–1996.

The breeding sequences for the four birds shown above are

$$Y^{(116)} = (4, 4, 4, 2, 2, 4, 0, 0, 4, 0, 4);$$

$$Y^{(209)} = (2, 0, 0, 4, 3, 3, 4, 4, 4, 4, 4);$$

$$Y^{(280)} = (3, 2, 4)$$

$$Y^{(284)} = (4, 0, 0, 0, 0, 0, 0, 2, 4, 0, 4, 0, 4, 3, 4, 4, 4, 4, 4, 4, 4, 4, 4, 0, 4, 4, 0, 0, 2, 0, 0, 0, 2)$$

Each sequence is associated with a bird, and each sequence has a length in years. Each component in a sequence is associated with a calendar year; 1961–1971 for bird 116, 1959–1969 for bird 209, and so on. Each component in the sequence is also associated with a breeding age beginning at one for the first component. For present purposes, *b_age* is the age relative to the first egg-laying event; 1–11 for birds 116 and 209; 1–3 for bird 280, and so on.

Thirteen of the birds in this sample were born at Eynhallow but did not return as adults; their sequences are empty and do not contribute to any subsequent plots or analyses. The first and last components of each non-empty sequence are strictly positive; internal values may be zero. The non-empty sequences range in length from one to 33 years. Their combined length is 3785 bird-years.

Unless one plans to use a purpose-built computer package that is capable of digesting the file in raw mark-recapture format, it is helpful to rearrange the concatenated sequences in spreadsheet format with 3785 rows. The first few lines are as follows:

| bird_id | yr | age | b_age | rev_age | outcome |
|---------|----|-----|-------|---------|---------|
| 116 | 61 | 3 | 1 | 11 | 4 |
| 116 | 62 | 4 | 2 | 10 | 4 |

| | | | | | |
|-----|----|----|----|---|---|
| 116 | 63 | 5 | 3 | 9 | 4 |
| 116 | 64 | 6 | 4 | 8 | 2 |
| 116 | 65 | 7 | 5 | 7 | 2 |
| 116 | 66 | 8 | 6 | 6 | 4 |
| 116 | 67 | 9 | 7 | 5 | 0 |
| 116 | 68 | 10 | 8 | 4 | 0 |
| 116 | 69 | 11 | 9 | 3 | 4 |
| 116 | 70 | 12 | 10 | 2 | 0 |
| 116 | 71 | 13 | 11 | 1 | 4 |
| 193 | 59 | 1 | 1 | 2 | 4 |
| 193 | 60 | 2 | 2 | 1 | 4 |
| 194 | 59 | 1 | 1 | 9 | 3 |
| 194 | 60 | 2 | 2 | 8 | 0 |
| 194 | 61 | 3 | 3 | 7 | 0 |

The variable `age` is measured from the first recorded sighting, whereas `b_age` is breeding age measured from the first record of an egg being laid. These variables are equal for all blow-in breeders. Reverse age `rev_age` is breeding age measured backwards from the last non-zero, so that the series length is `b_age + rev_age - 1`.

10.1.4 Averages for Cohorts

The function `tapply(outcome, b_age, mean)` computes the sample mean for each breeding age. In the first panel of Fig. 10.1, the average reproductive score is plotted against breeding age. At age one, this is an average for all birds whose sequences are not empty. At age $a \geq 1$, it is an average for all birds whose breeding sequence is at least a years.

Let S_a be the subset of birds contributing to the average at age a , i.e., S_a is the set of birds whose sequence length is at least a . The sequence of subsets $S_1 \supset S_2 \supset \dots$ is an age cohort of birds, which means that the subsets are non-increasing as a function of age. The function `table(b_age)` counts the cohort size, part of which is shown below:

| | | | | | | | | | | | | | | | | |
|---------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|----|----|
| a | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 |
| # S_a | 415 | 335 | 313 | 292 | 266 | 246 | 223 | 212 | 197 | 171 | 152 | 132 | 120 | 108 | 86 | 80 |

Cohort size is encoded in the plot symbol in Fig. 10.1.

In the second panel of Fig. 10.1, the average reproductive score is plotted right-to-left against reverse age counted as one for the year in which an egg was last recorded. At reverse age one, this is an average for all birds in the sample; at reverse age a , which is coded as $-a$ in the plot, it is an average over the cohort $S_a \subset S_1$ of birds whose sequence is at least a years. Thus, the set of birds contributing to the point at age a in the first panel is the same set of birds contributing to the average at $-a$ in the second panel. Although the sequence of subsets for reverse time is the

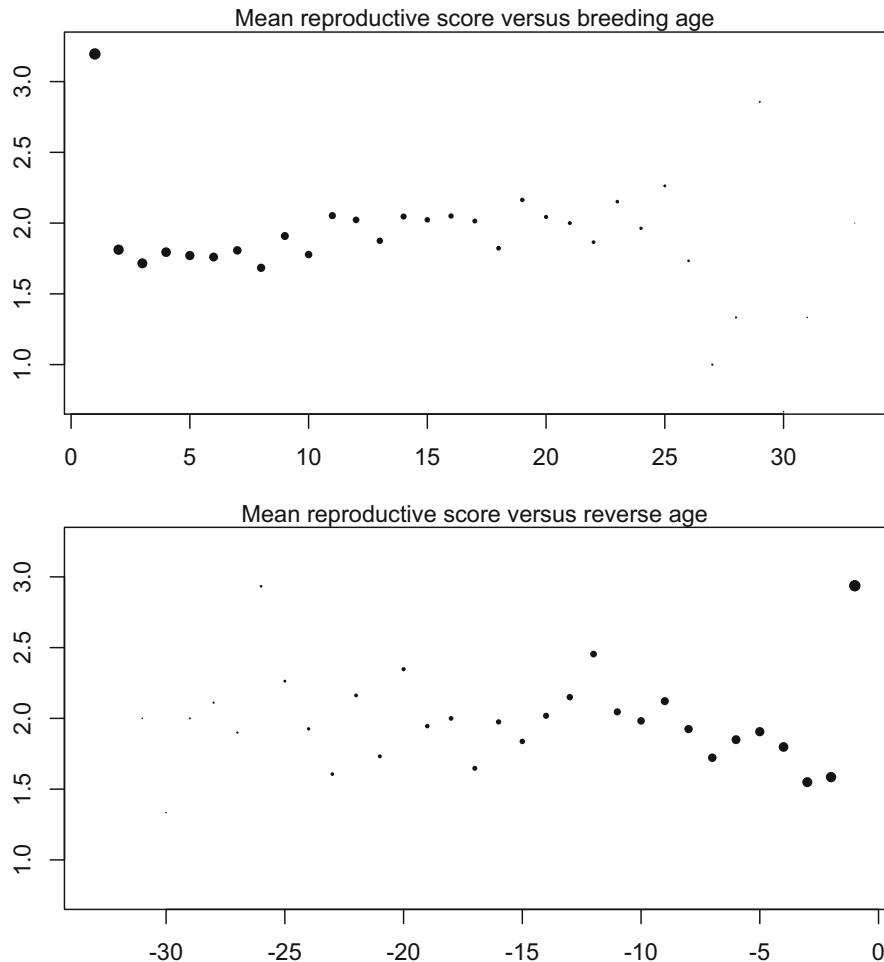


Fig. 10.1 Average reproductive score versus age in the top panel, and against reverse age coded backwards in time in the lower panel. Points are scaled to reflect cohort size

same as the original cohort, the averages are different because bird i with sequence length $n_i \geq a$ contributes $Y_{i,a}$ to the average at a in first panel, and Y_{i,n_i-a+1} to the average at $-a$ in the second.

The plot of averages against age counted forward from ‘A’ is not shown, but the pattern is essentially the same as the plot against breeding age.

Figure 10.1 shows clearly that the first and last values in each sequence are substantially larger on average than intermediate values. This fact is fairly obvious from the context because the value $Y_{i,a}$ for bird i at age $1 \leq a \leq n_i$ is a random variable taking values in $\{0, 2, 3, 4\}$. However, the event $Y_{i,a} = 0$, which has probability zero for $a = 1$ and $a = n_i$, has strictly positive probability for

$1 < a < n_i$. Figure 10.1 demonstrates clearly that the magnitude of these terminal anomalies is a dominant feature in this process.

Apart from the initial value, subsequent averages in the first panel exhibit a slow but definite increase in reproductive score as a function of breeding age, at least up to age 25. This pattern should come as a surprise because this is a breeding cohort whose average fertility, agility, ability and dexterity might be expected to decrease as a function of age. It suggests either that fertility increases with age, which is counterintuitive, or that frequent practice and parental experience are sufficient to offset any decline in fertility or foraging ability. It would be heart-warming to report this phenomenon as a triumph of avian wisdom and maternal experience over senescence and declining fertility. But any conclusion along these lines would also be a statistical misinterpretation of the facts.

In the second panel $\check{a} = n_i - a + 1$ is the breeding age in years counted backwards, while the averages are plotted against $-\check{a} = a - n_i - 1$, preserving left-to-right temporal order for all sequences of a given length. Apart from the final point, the averages in the second panel appear to decrease as a function of $-\check{a}$, i.e., to decrease as a function of the bird's age over the last 10–15 years of observation. At first sight, therefore, the two panels appear to tell contradictory stories about breeding success of females as a function of age—a slight but steady increase with age in the first panel, a moderately strong decrease with age in the second. This is a cohort paradox which is resolved in Sect. 10.1.6.

10.1.5 Averages for Disjoint Subsets

To help understand the apparent anomaly in Fig. 10.1, the birds are first partitioned into disjoint subsets according to series length, i.e., by breeding lifetime. For example, 19 birds had reproductive lifetimes of 10 years; 20 birds had reproductive lifetimes of 11 years, and so on. Figure 10.2 shows the pattern of average reproductive scores for six subsets of birds whose series lengths were 7–12 years. In all cases, the terminal values are the largest, and the first tends to be a little larger than the last. For the intermediate values, no pronounced positive or negative trend is evident. The intermediate values have the appearance of a stationary time series.

The average of the intermediate scores for series of lengths 3–15 are

$$1.24, 1.56, 1.57, 1.90, 1.55, 1.57, 1.38, 1.74, 1.48, 1.45, 1.47, 2.01, 1.72$$

Without claiming that these averages are independent or identically distributed, one can make an informal check for a trend by computing the correlation with series length. Whether we use the actual averages or their ranks, no evidence of a non-zero correlation emerges. In this respect at least, the distribution of non-terminal values appears to be unrelated to series length. Not only do the intermediate values have the appearance of a stationary series, but the distributions for different lengths seem

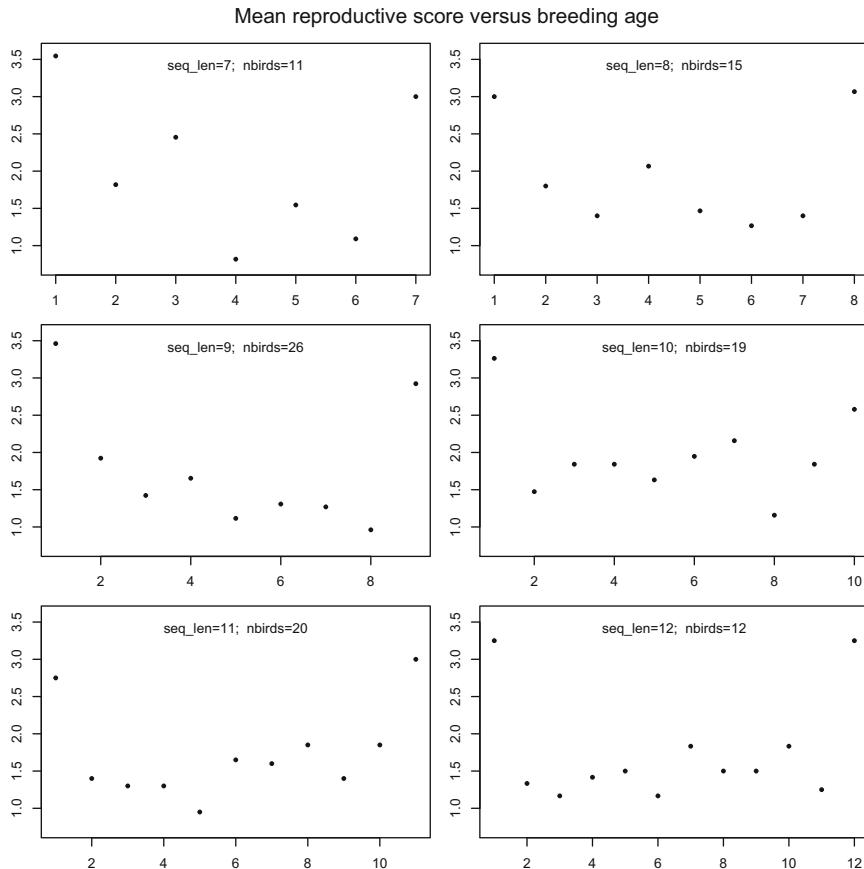


Fig. 10.2 Average reproductive score versus age for series of length 7–12 years

to have similar first moments. Visual inspection of the panels in Fig. 10.2 suggests that the second moments are also similar.

10.1.6 Resolution of a Paradox

The first panel of Fig. 10.1 shows that the mean reproductive score increases with the bird's reproductive age for ages $a \geq 2$, suggesting that older birds achieve greater success on average than younger birds. Success improves with experience! However, the second panel exhibits exactly the same feature in reverse time. For the same sequence of birds, the mean reproductive score increases with reverse breeding age (≥ 2), indicating that average success also improves with lesser experience. This conclusion seems paradoxical, if not self-contradictory. It calls for a resolution.

Table 10.1 Cohort averages in forward and reverse time

| Age | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|-------------|-----|-----|-----|-----|-----|-----|-----|-----|------|------|------|-----|
| \bar{Y}_a | 3.2 | 1.8 | 1.7 | 1.8 | 1.8 | 1.8 | 1.8 | 1.7 | 1.9 | 1.8 | 2.1 | 2.0 |
| \bar{Z}_a | 2.9 | 1.6 | 1.5 | 1.8 | 1.9 | 1.8 | 1.7 | 1.9 | 2.1 | 2.0 | 2.0 | 2.5 |
| # S_a | 415 | 335 | 313 | 292 | 266 | 246 | 223 | 212 | 197 | 171 | 152 | 132 |
| h_a | 100 | 6.6 | 6.7 | 8.9 | 7.5 | 9.3 | 4.9 | 7.1 | 13.2 | 11.1 | 13.2 | 9.1 |

Table 10.1 shows the forward- and backward averages in cohorts S_a , together with the cohort size for $a \leq 12$. Each average \bar{Y}_a or \bar{Z}_a is a mean of certain reproductive scores, one score for each bird in the subset $S_a \subset S_1$. Some of these scores are initial, some are final, and the remainder intermediate; \bar{Y}_1 is the average of initial values only, while \bar{Z}_1 is the average of final values only. In a sequence of length one, the value is both initial and final, and contributes to both \bar{Y}_1 and \bar{Z}_1 . In the average over S_a for $a \geq 2$, the fraction of values that are terminal is

$$h_a = (\#S_a - \#S_{a+1})/\#S_a,$$

which is shown as a percentage in Table 10.1. The general trend in the terminal fractions is increasing as a function of $a \geq 2$. Since terminal values are appreciably larger on average than intermediate values, we should expect both \bar{Y}_a and \bar{Z}_a to increase with age for $a > 2$, precisely as observed in Fig. 10.1.

10.2 Formal Models

10.2.1 A Linear Gaussian Model

The exploratory analyses described in the preceding section suggest fitting a formal linear model to the data in concatenated sequence form. Specifically, the response for bird i is a sequence $Y_{i,1}, \dots, Y_{i,n_i}$ of length $n_i \geq 1$. We regard breeding age as a covariate and sequence length as a non-random constant, so that there is a fixed set of n_i observational units for bird i .

If the baseline is set at the initial egg-laying, then breeding age is a bona-fide covariate according to the definition in Sect. 11.3.1. Calendar year is also a baseline variable, which is treated here as a relationship. However, it is difficult to offer any comparable justification for treating sequence length as a covariate. Nonetheless, the analysis in Sect. 10.1.5 suggests that this choice is not unreasonable.

Apart from the dominant terminal effects, any formal model must aim to take account of possible trends in the mean plus non-trivial correlations associated with various baseline relationships. Some possibilities are as follows:

1. mean score as a linear function of age $E(Y_{i,a}) = \beta_0 + \beta_1 a$ with constant coefficient β_1 independent of sequence length;

2. mean score as a linear function of reverse age $E(Y_{i,a}) = \beta_0 + \beta_1 \check{a}$, with $\check{a} = n_i + 1 - a$;
3. mean score as a linear function of length-normalized breeding age $E(Y_{i,a}) = \beta_0 + \beta_1 a / (n_i + 1)$, so that age reversal $a \mapsto \check{a}$ does not alter the mean-value subspace;
4. Calendar year: covariances between $Y_{i,a}$ and $Y_{i',a'}$ for pairs of observations made in the same calendar year: $t(i, a) = t(i', a')$;
5. Bird: covariances between $Y_{i,a}$ and $Y_{i',a'}$ for observations at different ages on the same bird $i = i'$, constant as a function of $a - a'$;
6. Bird-time: covariances between $Y_{i,a}$ and $Y_{i',a'}$ for observations at ages a, a' on the same bird, non-constant as a function of temporal separation $a - a'$.

Bird-to-bird variations and year-to-year variations are expected to be substantial, so it is natural to start with a model that includes both in a reasonably simple form. For illustrative purposes, we choose option 3 as a starting point for the mean model.

$$\begin{aligned} E(Y_{i,a}) &= \beta_0 + \beta_1 a / (n_i + 1) + \beta_2 I(a = 1 | a = n_i); \\ \text{cov}(Y_{i,a}, Y_{i',a'}) &= \sigma_0^2 \delta_{i,i'} \delta_{a,a'} + \sigma_1^2 \delta_{t,t'} + \sigma_2^2 \delta_{i,i'}. \end{aligned} \quad (10.1)$$

The two covariates in the mean model are the normalized breeding age and the indicator function for initial or final values. It is possible to include two separate indicator functions, but this has not been done because a sequence of length one deserves only one terminal increment, not two. The covariance model includes calendar year as a block factor accounting for annual variations associated with weather, with no temporal correlation for successive calendar years. Apart from the block factor for birds, this model has no non-trivial temporal correlation for successive observations on the same bird.

The fitted variance components are

$$\text{Id } 2.098; \text{ year } 0.085; \text{ bird } 0.322;$$

all of which are significantly positive. The between-bird variance is nearly four times the between-year variance, showing that maternal wisdom and avian skills have greater impact on breeding success than annual weather fluctuations. The fitted standard deviations are in the ratio 5:1:2.

The fitted terminal contribution to the mean is $\hat{\beta}_2 = 1.362$ with standard error 0.064, which is in line with Fig. 10.2. The fitted temporal trend in normalized age is negative $\hat{\beta}_2 = -0.163 \pm 0.107$, but not significantly different from zero. This analysis shows no evidence of a temporal trend in normalized age or in normalized reverse age.

The block matrix for birds, $\delta_{i,i'}$ in (10.1), implies that the covariance between distinct observations on one bird is σ_2^2 , which is constant as a function of temporal separation $|a - a'|$. As it happens, the fit can be improved appreciably by allowing the

correlations to decrease with time. The block matrix $\delta_{i,i'}$ is replaced with a suitable Hadamard-product matrix such as

$$\delta_{i,i'} \times e^{-|a-a'|/\lambda}$$

with range λ to be estimated. The data suggest $\lambda \simeq 4.2$ years, meaning that the temporal correlation is reduced by half every three years. This covariance modification improves the fit by approximately 75.0 log likelihood units, which is very large, but it does not appreciably alter the principal conclusions.

10.2.2 Prediction

Whether or not the application calls for prediction in the literal sense, it is highly desirable that the fitted model have that capability. As it stands, the linear Gaussian model described in the preceding section is a family of distributions for sequences of arbitrary fixed length—padded on the right with zeros if necessary. Such a model has no capability for prediction because prediction implies that the ultimate series length is unknown. The remedy, however, is relatively straightforward.

The sequence length is regarded as a random variable N taking values in the natural numbers with probabilities $P(N = n) = p_n$, independently for each bird. The value $N = \infty$ is seldom important for prediction, but it is best included with probability $p_\infty \geq 0$. The marginal distribution of series lengths can be estimated either using the Kaplan-Meier estimator which takes account of right censoring, or it can be estimated subject to a monotone hazard constraint on the expected hazards in the last line of Table 10.1. The standard Kaplan-Meier estimator ordinarily permits immortality $\hat{p}_\infty > 0$, but smoothed versions guaranteeing finiteness are available. Conceptually, the sequence length N is generated first, and the values Y_1, \dots, Y_N are generated according to the Gaussian model (10.1) using estimated parameters as needed. The sequence may then be padded with zeros on the right. With probability $1 - p_\infty$, the sequence thus generated contains finitely many non-zero values. Given the initial sequence $Y[k] = (Y_1, \dots, Y_k)$, it is straightforward in principle to compute the conditional distribution or predictive distribution for subsequent values.

10.2.3 Model Adequacy

The model described above with either fixed or random sequence lengths is open to criticism on many fronts. The following list is by no means comprehensive.

1. Mis-match of state spaces: Each observation is a score in $\{0, 2, 3, 4\}$ while all model distributions are continuous on the entire real line.

2. Left censoring: Some of the birds were active breeders when the study commenced in 1958. In such cases, we cannot be sure that the first non-zero value for that bird is the one recorded in 1958 or later.
3. Right censoring: Some of the birds were active breeders at the end of the recording period in 1996. From the available data, we cannot be sure that the bird did not nest at Eynhallow in 1996 or later.
4. Baseline or breeding age zero for bird i is defined as the calendar year $\min\{t - 1 : Y_{i,t} > 0\}$, which is a deterministic function of the breeding process. It is difficult to regard this version of breeding age as a covariate in the spirit of the definition in Sect. 11.3.1.
5. Faith and faithfulness: The record for *some* birds shows remarkable faithfulness to the nesting site. Ornithologists assure us that *all* fulmars are equally faithful, which implies that a bird nesting once at Eynhallow never nests elsewhere. But an ounce of skepticism cures a ton of faith, and the most credulous statistician must notice that nearly 20% of birds in this study are one-time breeders. Is it plausible that such a large fraction of breeding females retire after one year? Without further evidence, it is entirely possible that some of these one-time Eynhallow breeders may be occasional or regular breeders elsewhere.

The first criticism is the most obvious, but it is also the least fundamental: all estimates and most inferences are based on first and second moments. The fact that the values are limited to the first few integers eliminates the possibility of heavy-tailed distributions, so the worst consequences of non-normality are avoided.

The second and third criticisms are intrinsic to the design of ornithological studies, which have many of the characteristics of mark-recapture designs. See Sect. 10.3.

The last point is different in a fundamental way. If it prevails, skepticism leads to a re-interpretation of all zeros in the sequence for each bird.

For this discussion, an observational unit is a slot (i, t) corresponding to bird i in calendar year t . At that time, the bird may be (i) unborn, (ii) a fledgling or juvenile, (iii) a breeder, or (iv) retired or dead. For slots in class (iii), the breeding score is a number $Y_{i,t}$ in $\{0, 2, 3, 4\}$; for all other slots, $Y_{i,t} = 0$. In the recorded sequence, each slot contains a value $X_{i,t}$, not necessarily equal to $Y_{i,t}$. Each non-zero slot (i, t) for which $X_{i,t} \neq 0$ is a breeder containing the recorded value $X_{i,t} = Y_{i,t}$. Each zero slot corresponds to a missing bird, either a non-breeder, or an unrecorded breeder. If the zero slot corresponds to a non-breeder, the breeding score is also zero $Y_{i,t} = X_{i,t} = 0$; if it corresponds to a breeder, the breeding score $Y_{i,t}$ is not necessarily zero. Although no record is available for this slot, faithfulness comes to the rescue and allows us to infer that no egg was laid either at Eynhallow or elsewhere: $Y_{i,t} = X_{i,t} = 0$. In particular, each internal zero slot in the recorded sequence is an unrecorded breeder.

In the absence of faithfulness, each breeding slot for which $X_{i,t} = 0$ corresponds to a missing bird whose score $Y_{i,t}$ is not observed. Since this bird was recorded in a previous or subsequent year, there are only two possibilities: either breeding did not occur in which case $Y_{i,t} = 0$, or breeding occurred elsewhere in which

case $Y_{i,t} > 0$ is indeterminate. More importantly, however, the sampling scheme is biased because observational units for which $Y_{i,t} = 0$ tend not to occur in a breeding colony. Without the faithfulness assumption, it is difficult to say much about the fraction of zero values among adult birds.

10.3 Mark-Recapture Designs

Mark-recapture designs are a cornerstone for sampling and experimentation in animal ecology. The basic idea is that the study begins in calendar year zero by capturing, marking and releasing a sample of individuals from the target population. Each captured animals is measured and may be treated. In subsequent years, a further sample from the same environment is captured. Previously unmarked individuals are marked and all animals are measured and released. Over a period of years, each individual accumulates an encounter history and a measurement history. The encounter history for animal i is the subset of years in which the animal was captured; the measurement history is the parallel sequence of measurements, possibly but not necessarily real-valued. In some settings, the goal is to estimate the population size or the survival distribution for animals in the wild; in other settings, the goal may be to assess the effect of some intervention on animal health.

If the goal is to determine the health of animals in the wild, the typical response is a health measurement, perhaps weight. If the goal is to study survival, the measurement $Y_{i,t}$ at each encounter is $Y_{i,t} = 1$ on the presumption that the captured animal is alive. For an animal whose encounter sequence over eight years looks like

$$E_i = (0, 0, 1, 1, 0, 0, 1, 0),$$

the measurement Y_{it} is available only at slots for which $E_{it} = 1$. Generally speaking, $E_{i,t} = 0$ implies that $Y_{i,t}$ is unmeasured and unknown. However, if the goal is to determine the survival distribution, the values for slots $t = 5, 6$ can be inferred by continuity from the encounter sequence: $Y_{i,3} = \dots = Y_{i,6} = 1$. If age is available for one slot, it can also be inferred for missing slots. Ordinarily, measurements cannot be deduced from the encounter sequence alone.

The statistical computer package MARK (Cooch & White, 2020) is designed to analyze data from mark-recapture studies. The input data are coded in mark-recapture format, one line per animal. Its goal is to accommodate statistical complications such as variable capture and survival rates, right censoring, cohort effects, and serial correlation of measurements.

The Eynhallow study has many of the features of a mark-recapture design. All birds nesting at Eynhallow are captured, measured and released. Each nesting bird generates a capture history and a measurement history. This is an observational study with no treatment. Ideally, we would like to know each bird's age, but this is available only in relation to the first encounter. However, the scheme for capturing birds favours breeding pairs, and seems to guarantee $Y_{i,t} > 0$. Without heroic

assumptions such as those in Sects. 10.1 and 10.2, it seems difficult to say much about the frequency of slots for which $Y_{i,t} = 0$.

10.4 Further References

Dunnett (1991) gives a concise historical account of the background, initiation and development of the study of fulmars at Eynhallow.

10.5 Exercises

- 10.1** Plot the average reproductive score against calendar year. Is the range of annual averages high or low in relation to the reproductive scale 0–4? Does this plot suggest serial correlation?
- 10.2** Each bird in this study has a sequence length in the range 0–xx. Compute the histogram of sequence lengths. How many sequences are empty? Report the average and the maximum length? What fraction of the birds are one-time breeders?
- 10.3** Let a be the vector of bird ages, and n the vector of sequence lengths so that $\check{a} = n + 1 - a$ is reverse age. Show that $\text{span}\{\mathbf{1}, a/(n+1)\}$ is equal to $\text{span}\{\mathbf{1}, \check{a}/(n+1)\}$. Hence justify the claim made about age reversal in the third of the list of options in Sect. 10.2.
- 10.4** Show that the model (10.1) implies exchangeability of initial values $Y_{i,1} \sim Y_{j,1}$ for every pair of birds, whether recorded in the same year or in different years.
- 10.5** Show that the model (10.1) implies exchangeability of terminal values $Y_{i,n_i} \sim Y_{j,n_j}$ for every pair of birds, regardless of whether $n_i = n_j$ and regardless of the years in which these occurred.
- 10.6** In the light of the preceding exercises, discuss the pros and cons of using normalized versus unnormalized age in the mean model (10.1).
- 10.7** Show that the extended model suggested in the last paragraph of Sect. 10.2 also has exchangeable initial values and exchangeable terminal values. Under what conditions on the parameter do initial values have the same distribution as terminal values?
- 10.8** What evidence is there in the data suggesting serial correlation in the year effects? Can the fit be improved using a model containing non-trivial serial correlation? Extend the model and report a likelihood-ratio statistic.

10.9 A breeding population is a set of birds N_t consisting of adults aged one year or more. The annual mortality rate is a constant $q = 1 - p$ from year one onwards. Each surviving adult produces one offspring per year; the survival rate for offspring is such that n individuals survive to age one. Show that the approximate size of the breeding population is $\#N_t \simeq n/q$.

What fraction of the breeding population N_t in year t are one-time breeders? The set $\cup_{1 \leq s \leq t} N_s$ consists of all birds that were adults at some year during the interval $1 \leq s \leq t$. For large t , what fraction of this set are one-time breeders?

Chapter 11

Basic Concepts



11.1 Stochastic Processes

11.1.1 Process

Probabilistic reasoning is the foundation of applied statistics. The fundamental concept that provides the basis for probabilistic reasoning is the notion of a process, and specifically a stochastic process. In the first instance, a process is a function $Y: \mathcal{U} \rightarrow \mathcal{S}$ from a domain or index set \mathcal{U} into another set \mathcal{S} called the state space or the observation space; to each point or object $u \in \mathcal{U}$, the function Y associates a point $Y(u)$ or Y_u in \mathcal{S} . A stochastic process is a probability distribution on the space of functions $\mathcal{U} \rightarrow \mathcal{S}$, i.e., a [probabilistic description of a] random function $Y: \mathcal{U} \rightarrow \mathcal{S}$.

The domain for a Markov chain or a time series is either the integers or the natural numbers; the domain for a continuous-time temporal process is the real line \mathbb{R} ; the domain for a spatial process may be the real plane or the complex plane, or possibly \mathbb{R}^d . White noise on the real line or on the plane is a little different because \mathcal{U} is not the real line or the plane but the set of lineal or planar Borel subsets. In such settings, Y is typically a random measure, i.e., a set function that is additive for disjoint subsets.

In a setting such as an agricultural field trial, the domain for the yield process is usually described loosely as the set of plots; this description is adequate for the field, but it is interpreted mathematically as the set of planar Borel subsets in any or all growing seasons. The domain for a simple clinical trial for a COVID-19 vaccine is usually described loosely as the set of patients; this is interpreted to mean all eligible patients whether or not they were recruited and observed in the AstraZeneca trial in 2021. The domain for a study of speciation or sexual compatibility of fruit flies is the set of male-female pairs—again meaning all possible pairs having the genetic characteristics of interest. The domain for a competition experiment such as a chess

or tennis tournament is the set of ordered pairs of competitors—again meaning all pairs whether or not they competed face-to-face at Wimbledon in 2021.

In each case, the state space is a set such as $\{0, 1\}$, \mathbb{R} or \mathbb{R}^2 , as a measurable space with Borel events. Depending on the setting, the response function may have context-specific properties, such as anti-symmetry $Y_{i,j} = 1 - Y_{j,i}$ in the case of a pairwise competition, or additivity $Y(A \cup B) = Y(A) + Y(B)$ for yields on disjoint plots in a field experiment.

Variability in experimental and observational data is represented stochastically using probability distributions, either simple processes with independent components, or more complicated stochastic processes exhibiting serial or spatial or other forms of dependence. The first few chapters of Davison (2003) provide a good introduction to the construction and use of stochastic models for a range of applications.

In applied statistics, \mathcal{U} is frequently called the set of observational units, and Y is called the outcome or response. A sample is a finite subset $U \subset \mathcal{U}$, and the observation is the restriction $Y[U]$ of the process to the sample.

Unless otherwise indicated, the sample is regarded as an arbitrary fixed subset. In a few applications, however, U may be a random sample. If U is independent of the process, random sampling causes no difficulty, and no adjustment is needed. If U is not independent of the process, the situation may or may not be more complicated. In a knock-out competition, for example, the occurrence of pairs in the second round depends on the outcome in round one. By design, the sample of pairs is both random and strongly dependent on the process. In a size-biased sampling scheme, physically larger units are more likely to occur in the sample than smaller units.

In the competition experiment, it is usually safe to proceed for estimation purposes as if the sample were fixed; certainly, the likelihood function is the same as if U were fixed in advance. In the second setting, the size distribution for samples is not similar to the size distribution in the population, so it is a serious error to proceed as if the sample were fixed.

11.1.2 Probability

A stochastic process, is nothing more than a probabilistic description of the function $Y: \mathcal{U} \rightarrow \mathcal{S}$ as a random variable or a collection of random variables $\{Y_u : u \in \mathcal{U}\}$. To each event $A \subset \mathcal{S}^{\mathcal{U}}$ the stochastic description associates a number $0 \leq P(A) \leq 1$, satisfying the rules of probability. Probability implies expectations, means, variances and so on, for real-valued random variables. Given a sample $U \subset \mathcal{U}$ and an observation point $y \in \mathcal{S}^U$, the process associates a conditional probability $0 \leq P(A | Y[U] = y) \leq 1$, which implies conditional expectations, conditional variances and so on.

The simplest processes have independent components. In other words, to each $u \in \mathcal{U}$ there corresponds a probability distribution P_u on the state space. Component-wise independence means that for any sample (u_1, \dots, u_n) consisting

of n distinct units, the joint distribution of Y_{u_1}, \dots, Y_{u_n} satisfies

$$P_{u_1, \dots, u_n}(A_1 \times \dots \times A_n) = P_{u_1}(A_1) \times \dots \times P_{u_n}(A_n)$$

for arbitrary events $A_r \subset \mathcal{S}$. All generalized linear models have independent components, which are usually not identically distributed because different units may have different covariate values. By general agreement in applied work, $x_u = x_{u'}$ implies $P_u = P_{u'}$. In other words, two units having the same covariate value also have the same response distribution, so the one-dimensional marginal distribution may depend *only* on the covariate value for that unit.

Gaussian processes having independent and identically distributed components are the building blocks for more general processes such as those encountered in Examples 1, 2 and 5. More general spatial and temporal processes are used throughout the examples.

11.1.3 Self-consistency

The dismissive phrase *nothing more than a probabilistic description of the function...*, which occurs at the beginning of the previous section, grossly underrates the difficulty of the assigned task. To understand the difficulty, consider a longitudinal design in which a given subject may be observed at an arbitrary finite collection of time points $\mathbf{t} \subset \mathbb{R}$ with $t_1 < t_2 < \dots < t_k$. With all covariates fixed, it is necessary to specify for each $k \geq 1$ and each \mathbf{t} , the k -dimensional joint distribution $P_{\mathbf{t}}(\cdot)$ on \mathcal{S}^k . Since the event $(Y_1, Y_4) \in A \times A'$ is the same as the event $(Y_1, Y_2, Y_4) \in A \times \mathcal{S} \times A'$, these distributional specifications are subject to logical consistency conditions such as

$$\begin{aligned} P_{1,4}(A \times A') &= P_{1,2,4}(A \times \mathcal{S} \times A') \\ &= P_{1,2,5,4}(A \times \mathcal{S} \times A') = P_{1,2,3,4}(A \times \mathcal{S}^2 \times A'). \end{aligned}$$

Without consistency, alternative ways of computing the probability of a given event might well give different answers. Kolmogorov consistency is the mathematical glue that holds it all together, and makes statistical activities such as prediction possible.

Consistent specifications are not easy to find, and a formulation that looks superficially plausible may well be self-contradictory. In a longitudinal setting where the response is real-valued and Gaussian, it may seem safe and natural to construct the joint distribution as a product of one-dimensional conditional distributions given past observations. This means specifying the conditional mean and the conditional variance given past observations—both times and values. If the joint distribution is to be Gaussian, the mean must be linear, and the variance constant, as a function of past values. However, the dependence on past observation times must also be specified, and this is not linear. It may be feasible to specify a

continuous-time process sequentially and consistently if it is Markovian; otherwise a sequential specification is most unlikely to be consistent.

Apart from Kolmogorov consistency, other forms of consistency or inconsistency sometimes arise in statistical work. Example 5 illustrates a probability model that is incompatible with randomization.

Self-consistency is an important consideration, but not necessarily a dominant part of the story. On the one hand, a consistent specification is not necessarily well-suited to a given task. On the other hand, statistical conclusions derived from an inconsistent specification are not necessarily dangerous or disastrously wrong. It all depends on the nature of the inconsistency. Nonetheless, incompatibilities and self-contradictory specifications must be strongly discouraged if not condemned outright.

11.1.4 Statistical Model

A statistical model is a non-empty set of stochastic processes $\{P_\theta : \theta \in \Theta\}$, each process having the same domain and state space. Operationally speaking, to each parameter point θ there corresponds a process P_θ , and $P_\theta(A)$ is the probability of the event A in that process. For example, $N(\mu, \sigma^2) \equiv N_{(\mu, \sigma^2)}$ denotes the normal distribution on \mathbb{R} , and, by extension, the process whose components are independent and identically distributed. Thus, $N_{(0,1)}(-1, 1) \simeq 0.683$ is the probability assigned to the interval $(-1, 1) \subset \mathbb{R}$, and $N_{(0,1)}((-1, 1)^{18}) \simeq 0.683^{18} \simeq 1/964$ is the probability assigned to the event in \mathbb{R}^{18} that the first 18 components in an independent and identically distributed sequence are all less than one in absolute value.

Every distribution on a given space can be extended automatically to an independent and identically distributed sequence on the product space. However, this extension is not always natural or relevant. Most of the processes considered in this book do not have identically distributed components, and most do not have independent components, so the extension alluded to in the previous paragraph is not one that should be taken for granted.

A word of caution is in order regarding the phrase *stochastic process*, which is used in two senses. On the one hand, a process is a sampling-consistent family of distributions, one distribution P_U for each sample $U \subset \mathcal{U}$. On the other hand, a random function $Y: \mathcal{U} \rightarrow \mathcal{S}$ with distributions $Y[U] \sim P_U$ is also called a stochastic process. Technically, the probability distributions come first, and consistency ensures the existence of the random variable Y or an ensemble of such variables having the requisite relative frequency. But, in my experience, research workers tend to put the cart before the horse.

The situation for statistical models is more complicated because there is a family of stochastic processes P_θ , one for each $\theta \in \Theta$. However, there is only one observation in the form of a function $Y: U \rightarrow \mathcal{S}$.

11.2 Samples

11.2.1 Baseline

Every experiment and every observational study has a temporal component. The baseline is the temporal origin or reference point marking the commencement of the study. Mathematically speaking, the baseline is a point at which the observational units have been assembled, together with all of the information about them that is needed to specify the probability of arbitrary outcomes. Protocols for experimentation and treatment assignment are registered at baseline. All statistical inferences are based on probabilities, and the probability model is also registered at baseline.

Generally speaking, the units available for study are not homogeneous. The baseline information records sex, age, and, in principle anything else that is available at baseline that can reasonably be deemed to have a bearing on outcome probabilities. In practice, a certain restraint or professional judgement is needed to decide what is likely to be relevant and what is not. Generally speaking, there is little point in recording distinctive information about specific units unless we have a plan for how it is to be used in the analysis, either in the immediate future with current technology or in the distant future with more advanced technology. In a field experiment, the geometric layout of the plots is ordinarily part of the baseline information, and is almost always relevant in that it affects outcome probabilities. Information about crop, treatment and yield in the previous season is sometimes available and might be judged relevant if the new plots were well-aligned with the previous plots. In a clinical trial with human patients, ethnic background might be relevant as a block factor, but the number of letters in the patient's name or the primality of his or her ID code is unlikely to be considered relevant for clinical outcomes.

For a randomized study, randomization occurs at or immediately after baseline. The randomization protocol is registered at baseline, but the randomization outcome is not. Model specification begins with randomization probabilities $p(\mathbf{t}) = \text{pr}(T = \mathbf{t})$ for each treatment assignment vector $\mathbf{t} = (t_i)_{i \in U}$, also called the treatment factor. Even if one assignment list is a permutation of the other, two assignment vectors \mathbf{t}, \mathbf{t}' may have, and usually do have, different probabilities depending on baseline information such as covariate or block structure. Most commonly, the randomization is balanced with each treatment level occurring with equal frequency in each block.

Since the probability model is registered at baseline, i.e., pre-randomization, the model specifies the joint distribution for treatment T and response Y . The joint distribution implies a marginal distribution for treatment assignments, and a conditional distribution $\mathbf{t} \mapsto P(\cdot \mid T)$, which associates with each assignment vector \mathbf{t} a conditional distribution for the response. Randomization subsequently produces a particular treatment configuration, and nearly every subsequent probability computation uses that value. In general, the conditional probability $P(A \mid T = \mathbf{t})$ of the event $Y \in A$ may depend on any and all registered baseline information.

Every variable measured post-baseline, such as T , is regarded as the outcome of a random process, and, as such, is formally a part of the response.

Baseline need not mean a fixed point in calendar time. In studies of cell development, the baseline would ordinarily be set at a key developmental stage such as fertilization, which is a point in calendar time that may vary from cell to cell. Similar remarks apply to clinical trials where the baseline is usually set at recruitment, which varies from one patient to another on the calendar scale. In a mark-recapture study such as Example 10, the baseline is set at the initial encounter.

11.2.2 Observational Unit

The *observational units* are the objects $u \in \mathcal{U}$ on which variables are defined and measurements may be made. Usually measurements are made only on a small subset of observational units (the sample), so the phrase *measurements may be made* does not imply that measurements have been made or that plans are afoot to make such measurements.

The statistical universe almost always includes infinitely many extra-sample units, notional or otherwise, for which probabilistic prediction may be required. Sometimes each unit is a physical object such as a plot, a patient, a rat, a tree, or a M-F pair of fruit flies. Sometimes the units are less tangible, such as time points or time intervals for an economic series, or spatio-temporal points or intervals for a meteorological variable such as temperature or rainfall. Very often, the set of observational units is a Cartesian product set such as

$$\{\text{mice}\} \times \{\text{front, rear}\} \times \{\text{left, right}\} \times \{\text{day0, day1, day2}\}$$

which contains 12 observational units for each mouse. As an index set, time is structured cyclically in a similar way:

$$\{\text{clock times}\} \times \{\text{7 days}\} \quad \text{or} \quad \{\text{365 calendar dates}\} \times \{\text{?? years}\}$$

The index set may be structured in other ways such as pupils within classrooms within schools, which is a nested or hierarchical structure defined by one or more relationships $R(u, u')$ on the units.

11.2.3 Population

The *population* \mathcal{U} is the set of observational units; the *sample* is a finite subset. Where necessary, the sample may be extended to include units for which observations are unavailable but response predictions are requested. In a meteorological context, the observational units are all points in the plane or sphere, or points in

the spatio-temporal product space, so the population is uncountably infinite. For a spatial process, the units may be either points in the plane, or subsets of the plane, or less tangible objects such as signed measures on the plane or planar contrasts. The sample is the finite set of points at which measurements (sample values) are planned or available or desired.

The mathematical population is the *index set* on which the response is defined as a stochastic process. As is often the case in mathematics, the mathematical index set is made sufficiently large that it encompasses every conceivable situation that might arise, and many more besides. For a clinical trial in which the experimental units are human patients, the mathematical index set need not be finite, and in fact the mathematical subset of units having a specific sex, age and body-mass-index may also be infinite. A non-mathematician might object to the fact that the mathematical index set contains more points than there are real physical or biological entities, or atoms in the universe. Such objections are not to be entertained seriously; they are on a par with rejecting the real number system for engineering or accounting purposes on the grounds that it contains infinitely many ‘useless’ values that are not needed for billing purposes.

A non-trivial stochastic theory requires the sample to be a proper finite subset of the population, but it does not require \mathcal{U} to be infinite. There are *bona fide* applications that call for a finite population, so we do not insist that all populations be infinite. However, we shall not encounter such applications in these notes.

Statistical colloquialisms. When one talks of a ‘Normal population’ or a ‘Cauchy sample’, the reference is not to the population or sample *per se*, but to the population values or sample values or their distribution, usually understood to have independent values for distinct units.

11.2.4 Biological Populations

Every biological population evolves by a process of birth and death. Tomorrow’s population is not the same as today’s population or yesterday’s population, but all three are finite. Mathematically speaking, the population is said to be locally finite in time. For most purposes, it is immaterial whether the entire population is globally finite or globally infinite. What is important is that only the current population is accessible or available for sample inclusion.

For some short-term social policy matters, voting and other political activities, the relevant population for inference is determined by democratic principles. Only the current population has a vote, so past and future generations are disenfranchised. Such populations are not constant in time, but the relevant subset is finite and constrained by democratic practice.

For medical and pharmaceutical studies, it is preferable to take a broader view, particularly if plans are afoot to use the drug or therapy for future patients. However, this broader perspective means that not all individuals in the population are accessible or available for immediate inclusion in a clinical trial.

In a clinical trial for a Covid-19 vaccine, the units available for recruitment are individuals who are alive and of a suitable age at the crucial time. It appears that the Covid-relevant population is finite. However, there are at least two reasons to reject the finiteness argument. The first is that the current population is very large. It is difficult to put a precise figure on it, say 7.5–8.0 billion, and it is even more difficult to explain why this number is biologically or mathematically relevant for the assessment of drug safety or efficacy. The second argument is that the Covid-19 relevant population is not restricted to the present, but also includes at least one future generation. Given that some units are inaccessible, it is sufficient to take \mathcal{U} to be infinite, so that the mathematical set is large enough to accommodate every conceivable demand, even beyond what is epidemiologically plausible.

11.2.5 Samples and Sub-samples

The *sample* $U \subset \mathcal{U}$ is the finite subset of observational units on which the response and other variables are recorded. Technically, U is a finite ordered list of units, ordinarily distinct, and the recorded response $Y[U]$ is the list of Y -values for $u \in U$ in the same order.

To be clear, the word ‘sample’ in these notes denotes a finite ordered subset of units. It does not imply a random sample, let alone a simple random sample. In most research settings, such as a field trial or a laboratory experiment, a random sample of any stripe is out of the question. In the case of biological populations, the sample is a subset of units that are accessible today, so the inclusion probability is necessarily zero for a great many units. Two samples consisting of the same units listed in a different order are different; their distributions are different but they are statistically equivalent for all inferential purposes.

In settings where prediction or interpolation is involved, it is necessary to consider an extended sample U' , which includes U as a sub-sample. Each $u \in U' \setminus U$ is called an extra-sample unit. Only the restriction $Y[U]$ is actually observed. Prediction refers to the conditional distribution of $Y[U']$ given the sub-sample values $Y[U]$; point prediction refers to the conditional expected value.

In a classical controlled design with a treatment factor having $k \geq 2$ levels, each unit consists of a subject i together with an assignment $i \mapsto \mathbf{t}(i)$, or $u = (i, \mathbf{t}_i)$, so $(i, 0)$ and $(i, 1)$ are distinct units having the same subject. Each sample is a finite set $I = \{i_1, \dots, i_n\}$ together with one of k^n assignments $\mathbf{t}: I \rightarrow [k]$. The sample

$$U = ((i_1, \mathbf{t}(i_1)), \dots, (i_n, \mathbf{t}(i_n))),$$

is a finite list in which $\{i_1, \dots, i_n\}$ are *distinct*. Ordinarily, I is a fixed set of subjects, and \mathbf{t} is determined by randomization.

The counterfactual setting differs from the classical setting in that each sample is an unrestricted finite collection of individual assignments, so i_1, \dots, i_n need not be distinct. Every classical sample is also a counterfactual sample,

but a generic counterfactual sample is not an assignment. For example $U = \{(i_1, 0), (i_1, 1), (i_2, 0), (i_3, 1)\}$ is not classical because it is not an assignment from the objects $\{i_1, i_2, i_3\}$ into the set of treatment levels.

11.2.6 Illustrations

In the discussion of Example 1, it was asserted that each observational unit is a site on a rat, i.e., a (rat, site) pair, and the response is a real number, i.e., the state space is the real numbers. However, one could argue that each rat is one observational unit, and the state space is \mathbb{R}^5 . At first glance, these appear to be equivalent.

What makes one choice more appropriate than the other is the nature of the five measurements on each rat. If these were five otherwise unrelated variables such as pulse rate, temperature, weight and blood pressure, each rat would be one observational unit, and the state space would be \mathbb{R}^5 . However, the observation consists of one biological variable measured at five sites. Although we do not necessarily expect the five measurements on one rat to be exchangeable or even to have the same expectation, the nature of the observation process—using the same instrument for each site—confers additional symmetry that would not otherwise be present.

For one rat, either choice leads to a response distribution on \mathbb{R}^5 . The difference is that the second version with five units per rat has more natural symmetries than the first. These symmetries arise from notionally permuting the units in various ways. For example, the model used in Example 1 has equal variances for all sites, and equal covariances for each pair of sites, which comes implicitly from assumptions about permuting sites. If we choose the rat as the observational unit, there is no possibility to permute sites, so these symmetries do not emerge as a consequence of permutation of units.

In Example 3, each observational unit was taken initially to be a mating event. But this was subsequently shown to be inappropriate for the design, and misleading for the analysis. Instead, it was deemed preferable to take one mating well as the observational unit.

For the daily temperature series, each observational unit for the analysis in Chap. 6 is a point in calendar time, consisting of a year and a date within the year. Date is a number in the range 1–365 having cyclic structure, i.e., a real number with addition modulo 365.

For the frequency analysis in Chap. 7, each observational unit is a Fourier frequency. These also come with harmonic structure such that frequency ω is associated with its harmonics $\{\omega, 2\omega, \dots\}$.

In Examples 1–9, one observational unit is (i) a site on a rat; (ii) a (log, saw) or (log, team) pair; (iii) a mating well; (iv) a (plant, date) pair; (v) a louse; (vi) a point in calendar time; (vii) a frequency; (viii) a language; (ix) a plot or a leaf. The population is some set of observational units, and there is no compelling reason in any example to restrict the population to a finite set. Many of these choices are

relatively straightforward from the definition given, but it is clear in several instances that other choices are possible. Example 10 is conceptually more complicated because the observational units are bird-year pairs, and the sampling scheme is restricted to pairs that occur in a breeding colony. For such pairs the breeding activity response is predominantly non-zero, which means that the inclusion probability is not independent of the response.

11.3 Variables

11.3.1 Ordinary Variables

An ordinary *variable* is a function on the observational units, both in-sample units and extra-sample units. The co-domain of a variable, i.e., the space in which the function takes its values, may be the set of real numbers or the set of complex numbers, the set of wheat varieties, horse breeds or hippopotami. Any set suffices for values.

Everyday examples include ‘weight in kg.’, ‘atmospheric pressure in cm. Hg.’, and ‘length in cubits’. In principle, the name of a real-valued variable includes the physical units of measurement so that the value $x_i \equiv x(i)$ of the variable x for unit i is a real number, not an expression such as ‘184.5 cm’ or ‘94.7 m.p.h.’. Mathematically speaking, weight in kilos and weight in pounds are different variables; in practice, descriptive terms such as weight, height and temperature are used flexibly in everyday speech without specified units. Flexibility is good, but ambiguity can be costly—such as the loss by NASA in 1999 of a Mars orbiter at a cost of \$125M because of a mix-up of distance units by the contractor.

Qualitative variables include *sex* taking values in $\{M, F\}$; or *occupation* taking values in a suitable set of occupations. This set of values or levels must be exhaustive, so one of the values may be the catch-all class ‘*none of the above*’.

Operations: If x, y are two variables, the ordered pair (x, y) is also a variable: the value of (x, y) for unit i is $(x, y)(i) = (x_i, y_i)$, which is a point in the Cartesian product space. Each variable is defined on the population and recorded on the sample.

Feature is a synonym for variable or attribute—a function on the units. The feature vector takes values in the feature space.

In certain settings, the response on one unit is a vector, and each feature is one component; the primary response is a class or characteristic of the unit, and the goal is to classify each unit by computing the conditional distribution over the set of classes given the features.

Quantitative Variable

A real-valued function on the observational units is called a *quantitative* variable. More generally, a quantitative variable is a function taking values in a space that permits certain arithmetic operations such as addition and scalar multiplication. Dose (of fertilizer or medication in suitable units) is a typical quantitative variable whose values are non-negative. Blood pressure (systolic, diastolic) in mm. Hg. is a quantitative variable taking values in \mathbb{R}^2 . This statement means that every realizable value of blood pressure can be found somewhere in \mathbb{R}^2 ; it does not mean that every point in \mathbb{R}^2 is realizable as a blood-pressure value for a live human subject. Values that are in conflict with hydrostatic or hydraulic or haemodynamic theory are deliberately not excluded by the definition.

Operations: If x, z are two quantitative variables taking values in the same space, so also is the linear combination $3x + 4z$. If x, z are real-valued variables, so also is the unit-wise product xz . Consequently x^2, z^3 and other monomials such as x^2z are also quantitative variables.

Qualitative Variable

A *qualitative* variable, also called a *classification factor*, is a function on the observational units taking values in a finite or countable set, called the *factor levels*. Examples include *sex*, *occupation*, *socioeconomic class*, and variables such as *genetic variant* with values ‘wild type’ and ‘mutant’. Often, one level is designated as a reference level. A qualitative variable is sometimes called an *attribute* or a feature.

Ordered pairs: If u is the qualitative variable representing COVID vaccine with four levels *Pfizer*, *Moderna*, *AstraZeneca*, *Janssen*, and v is the dose count with three integer values $\{0, 1, 2\}$, the product set contains twelve ordered pairs that are mathematically distinct. Operationally, however, the four pairs associated with zero dose are not distinguishable, so the number of physically distinguishable ordered pairs is only nine.

Response

The *response*, usually denoted by Y , is the variable of primary interest, the variable that is measured or recorded on the sample units, e.g., yield in kg. per unit area, or time to failure in a reliability study, or stage of disease, or severity of pain, or death in a 5-year period following surgery. There may be secondary or intermediate response variables such as compliance with protocol in a pharmaceutical trial, which are also part of the response. Synonyms and euphemisms include *yield*, *outcome* and *end point*.

In statistical work, the response is regarded as the realized value of a random variable, or process $u \mapsto Y_u$ taking values in the *state space* $Y_u \in \mathcal{S}$. For an

observational study, the distribution is denoted by P ; for a randomized study $P(\cdot, \mathbf{t})$ is the joint distribution of the response and treatment assignment.

To be clear, the response is not some conceptualized or notional variable that we would like to measure but are unable to measure on the sample units. By definition, the response is the variable that is actually measured on the sampled units, i.e., the value recorded by a blood pressure instrument or a treadmill task at a particular time, or by a questionnaire for a psychiatric evaluation, not some notional ‘true’ state of health. Likewise, the probability model is a probability distribution for the process corresponding to the variable measured, including the procedure or instrument used to measure it.

Many of the stochastic models considered in these notes are built from simpler processes, for example, by addition of a smooth process plus white noise, or by using a latent smooth process as the intensity for a Bernoulli process or Poisson process. Some authors are then inclined to refer to the unobserved smooth process as the ‘true value’, suggesting that the observation is the false or corrupted value. Provided that the descriptive term ‘true value’ is understood in the non-pejorative pure-mathematical sense, this terminology causes no difficulty. But it can lead to awkwardness or social embarrassment in instances where the true state of health is normal even after the patient has died.

For many examples in these notes, the response on one observational unit is a scalar. In a longitudinal study, however, an observational unit is a patient-time pair (i, t) , so the response Y_i for one patient is an entire time series whose values are recorded at a finite set of times. Similar remarks apply to mark-recapture studies in ecology: see Example 10.

Covariate

A *covariate* x is a baseline function on the observational units that is used in a probability model to permit the outcome distribution for one unit to differ from that of another unit. Ordinarily, if $x_i = x_j$, the events $Y_i \in A$ and $Y_j \in A$ are presumed to have the same probability; otherwise, if $x_i \neq x_j$, the probabilities may be different. For this to make operational sense, the covariate must be registered at baseline. Typical examples include patient age, sex of mouse, type of soil or soil pH (pre-planting).

If the set of observational units is a Cartesian product set $\mathcal{U} = \mathcal{U}_0 \times \mathcal{U}_1$, each marginal component $u \mapsto u_0$ or $u \mapsto u_1$ is available at baseline. In Example 1, each unit u is a $(\text{rat}, \text{site})$ pair, so the function $u \mapsto \text{site}(u)$ is a covariate. The function $u \mapsto \text{rat}(u)$ is also a baseline variable, but it is used as a block factor. In Example 5, each observational unit (louse) is associated with an ordered pair $(\text{aviary}_u, \text{time}_u)$, so aviary and time are baseline variables.

Operationally, a covariate is used in a randomized experiment to reduce ‘unexplained’ variation and thereby to increase the precision of treatment effect estimates. In an analysis of variance, the total sum of squares for the response is partitioned into various parts, one part associated with registered covariates and block factors, a

second part associated with treatment, the remainder being ‘unexplained’ or residual variation. The part associated with covariates and block factors, the between-blocks variation, is said to be ‘eliminated’, and the more variation eliminated the less remains to contaminate the estimates of treatment contrasts. A covariate or block factor is said to be effective for this purpose if the associated mean square is substantially larger than the mean squared residual. This means that the response variation within blocks, the intra-block mean square, should be appreciably smaller than the response variation between blocks, the inter-block mean square.

In practice, it may be acceptable to fudge matters by using as a covariate, a variable measured post-baseline before the effect of treatment has had time to develop, or an external variable whose temporal evolution is known to be independent of treatment assignment for the system under study. Louse sex in Chap. 5 is a simple, uncontroversial, example of a post-baseline variable, which is not statistically independent of the response (louse size), but whose *evolution* is ‘known to be’ independent of both treatment assignment and louse size.

At a minimum, it is necessary first to check that the variable in question is indeed unrelated to treatment assignment; otherwise its use as a covariate could be counterproductive. It is well to remember that while measurement pre-baseline is strong positive evidence that no statistical dependence on treatment assignment exists, the most that can be expected of a post-baseline measurement is absence of evidence. For a variable of dubious status, absence of evidence is considerably better than its complement, but it does not provide the same positive assurance as evidence of absence. A concomitant variable of this sort is not counted as a covariate in these notes. It is formally regarded as a component of the response whose dependence on treatment assignment is to be specified as a part of the statistical model. The dependence may be null, but that alone does not give it the status of a covariate.

As always, a probability model P allows us to compute whatever conditional distribution might be needed for inferential purposes. That includes the conditional distribution given any concomitant or intermediate outcome or the conditional distribution of health values given that the patient is alive, or the conditional distribution of the cholesterol level given that the patient has complied with the protocol, or even the probability of compliance given the cholesterol level. Whether these are the relevant distributions for the purpose at hand is an entirely different matter to be determined by the user in the given setting.

Treatment

Treatment is a function or assignment $T : \text{sample units} \rightarrow \text{levels}$ taking values in the set of treatment levels. Although the algebraic arrow points from units to treatment levels, we usually say that treatment is assigned to the units. Treatment is not a covariate because it is not a property of the observational units that is registered at baseline; it is an *intervention* that changes the status quo for the sampled units only. Usually, treatment is a random variable whose value is the outcome of a *randomization scheme*. The components of T for distinct observational units, or

even for distinct experimental units, are usually identically distributed, but seldom independent.

In computational work, the observed treatment configuration $\mathbf{t} = (T_u)_{u \in U}$ is called the treatment factor. Although T is defined only for sample units, we must bear in mind that the sample can always be extended indefinitely, at least in principle, so the restriction to U is not a major part of the distinction between a classification factor and a treatment factor. The important distinction is that a pre-baseline variable is a property of the units, whereas treatment is assigned to units at baseline.

External Variable

Any variable measured post-baseline is regarded as a random variable whose probability distribution is specified at baseline. The randomization outcome becomes available post-baseline, so the randomization outcome is a component of the response in the sense that its distribution is specified at baseline. Usually, the randomization outcome is not of scientific interest in itself, so the focus of the investigation lies elsewhere.

Apart from the randomization, there may be other post-baseline variables that are relevant and must be considered, but are not themselves of scientific interest. An external or endogenous variable is one that is usually not independent of the primary response, but whose temporal evolution is independent of the primary response. For a definition of *independent evolution*, see Sect. 11.3.3. Independent evolution is an asymmetric relation between two temporal processes, so this concept arises primarily in longitudinal designs or in time series analysis. Louse sex in Chap. 5 is a simple example of a post-baseline variable that is external in the sense of the definition.

11.3.2 Relationship

A *relationship* is a function on *pairs of units*. The occurrence of a relationship in a statistical model means that the joint distribution for one pair of units may be different from that of another pair. For this to be feasible, the values must be registered at baseline. If each unit is a point in a metric space, or is associated with such a point, the metric $d(u, u')$ is a non-negative symmetric relationship among them. Experimental units are defined by a Boolean relationship on observational units: $R(u, u') = 1$ if u, u' belong to the same experimental unit, and zero otherwise. Other examples include genetic, familial, neighbour, and adjacency relationships.

Ordinarily, the relationship is defined on the population and recorded for the sample pre-baseline. In Chap. 5, however, *Aviary* is a block factor generated by randomization, and defined on the sample. Since the randomization may have been accomplished in two waves that were not necessarily synchronous, it is difficult to

say whether this block factor is pre-baseline or post-baseline. Similar remarks might be made about the *colony* factor in Sect. 9.3.

Block Factor

A *block factor* is a Boolean function on pairs of observational units that is reflexive, symmetric and transitive—an equivalence relation registered at baseline. Each block factor (such as the experimental unit factor) partitions the set of observational units into disjoint non-empty subsets called blocks. The identity function on \mathcal{U} is a block factor whose blocks are all singletons; at the other extreme, the function J such that $J_{u,u'} = 1$ for every pair, has exactly one block.

To each variable or factor x there corresponds a block factor B defined by

$$B_{ij} = 1 \text{ if and only if } x(i) = x(j).$$

Regardless of how the information is stored in an electronic device, the chief mathematical difference between B and x is that the x -blocks are labelled by x -levels, whereas the blocks of B are unlabelled. The x -block

$$x^{-1}(x(1)) = \{j \in U : x(j) = x(1)\},$$

i.e., the subset of sample units having the same x -value as unit 1, has the label $x(1)$. Since the blocks of B are unlabelled, a block factor has no reference level or reference block.

At the risk of over-simplification, covariates typically occur in the model for the mean response; block factors and other relationships occur in the model for covariances.

In principle, there may exist relationships among triples or k -tuples of units. For example, the cross-ratio

$$\chi(z_1, z_2, z_3, z_4) = \frac{(z_1 - z_2)(z_3 - z_4)}{(z_1 - z_3)(z_2 - z_4)}$$

is a relation on real or complex 4-tuples.

11.3.3 External Variable

Let the response be a two-component temporal process, so that (Z_t, Y_t) is the value at time t . For notational simplicity, time is discrete. Independent evolution is an asymmetric relation between the two processes. We say that Z evolves independently of Y if, for each t , future Z -values are conditionally independent

of past Y -values given past Z -values. In particular, for every t ,

$$Z_t \perp\!\!\!\perp Y^{(t-1)} \mid Z^{(t-1)} \quad (11.1)$$

where $Y^{(t)} = Y[\dots, t]$ is the restriction to past values.

Independent evolution does not imply that the two processes are statistically independent, nor does it imply that Y evolves independently of Z . It is an asymmetric relationship between temporal processes, which simplifies the sequential factorization of the joint density

$$\begin{aligned} p(z_t, y_t \mid Z^{(t-1)}, Y^{(t-1)}) &= p(z_t \mid Z^{(t-1)}, Y^{(t-1)}) p(y_t \mid Z^{(t)}, Y^{(t-1)}) \\ &= p(z_t \mid Z^{(t-1)}) \times p(y_t \mid Z^{(t)}, Y^{(t-1)}). \end{aligned}$$

When the focus is on Y as the primary response, an auxiliary process satisfying (11.1) is sometimes called *external* or *exogenous*.

In circumstances where Z is exogenous, it may happen that the evolution of Y is governed by synchronous Z -values only, in which case we have

$$Y_t \perp\!\!\!\perp Z^{(t-1)} \mid Z_t, Y^{(t-1)} \quad (11.2)$$

in addition to (11.1). The joint density then factors as

$$p(z^{(T)}) \times \prod_{t=1}^T p(y_t \mid Z_t, Y^{(t-1)}),$$

where the focus is usually on the second factor.

The much stronger conditional independence condition

$$Y_t \perp\!\!\!\perp Z^{(t-1)}, Y^{(t-1)} \mid Z_t \quad (11.3)$$

severely limits the nature of the temporal dependence in Y . In this case the second factor in the joint density simplifies further to

$$p(y \mid Z) = \prod_t p(y_t \mid Z_t).$$

The occurrence of Z is similar to the occurrence of a covariate, as if Z were recorded at baseline.

Example 5 shows that pigeon lice are sexually dimorphic, so size and sex are strongly dependent. Nonetheless, louse sex is a good example of a post-baseline variable that evolves according to Mendelian laws, independently of the main response (louse size). It is also clear on general grounds that only synchronous sex-values matter, so (11.2) is satisfied. However, Brownian evolution processes do not satisfy the stronger condition (11.3), so the simpler density factorization fails.

The health of an asthmatic patient may depend on recent local weather, but the evolution of weather patterns is, to an adequate approximation, independent of the health of patients. It is obvious in this setting that only local weather patterns matter, and recent is more important than not-so-recent, but it is less obvious that only synchronous weather matters, so (11.2) is dubious. Certainly, one would not expect (11.3) to hold for values measured at moderate to high frequency. Similar remarks could be made regarding investors in the stock market.

In a statistical model with parameter $\theta \in \Theta$, a distributional factorization such as that following from (11.1) or (11.3) may hold for every θ . In that circumstance, it is usually possible to express $\Theta = \Psi \times \Phi$ as a Cartesian product, and to express the likelihood function as a product of two factors, only one of which involves the parameter of interest. Such a likelihood factorization can lead to substantial simplification for parameter estimation.

11.4 Comparative Studies

11.4.1 Randomization

The *randomization scheme* is a probabilistic protocol for the assignment of treatment levels to sample units, often uniformly at random subject to design constraints. For a completely randomized design with 12 sample units and four treatment levels, a balanced randomization scheme is a function $T: [12] \rightarrow [4]$ (from sample units to treatment levels) chosen [uniformly] at random from the set of $12!/(3!)^4 = 369600$ functions having treatment blocks $T^{-1}(1), \dots, T^{-1}(4)$ of equal size. In the randomized-blocks setting, each sample unit is an experimental unit.

The only mathematical constraint on the randomization distribution is that it be fully specified in the protocol. Full specification for a sample with covariate configuration \mathbf{x} means that the randomization protocol is declared as a function $P(\cdot; \mathbf{x})$. Most importantly, since there is only one protocol, P is parameter-free, i.e.,

$$P_\theta(T = \mathbf{t}; \mathbf{x}) = P_{\theta'}(T = \mathbf{t}; \mathbf{x}) \quad (11.4)$$

for all values \mathbf{t} and all θ, θ' in the parameter space. In this setting, the covariate configuration includes both relationships and initial values: see Sect. 13.2.

Full specification allows for considerable latitude. However, some designs are more efficient than others for treatment comparisons. For a wide range of reasons, it is best to keep the randomization protocol as simple as possible subject to efficiency considerations.

In agricultural field trials, including horticultural trials, the randomization probabilities usually depend on the block structure and covariate configuration occurring in the sample units. For a typical randomized-blocks design, the joint probability

that the pair (u, u') is assigned treatment levels (t, t') depends on whether the units belong to the same block or different blocks. More generally, the probability $\text{pr}(u \mapsto t; U)$ that treatment level t is assigned to unit u may depend not only on x_u but also on $x_{u'}$ for all other units $u' \in U$. Unless otherwise specified, we assume in these notes that the assignment probabilities $\text{pr}(u \mapsto t; U) > 0$ are strictly positive for every unit and every treatment level. Although these probabilities may be positive for every unit, they are not usually positive for pairs of units and pairs of treatments: in a crossover trial, the probability of one individual being assigned the same treatment on both occasions is usually zero.

In cases where the components of T are independent, the randomization distribution may depend on baseline covariates or classification variables such as sex. For example, a two-level treatment may be assigned in the ratio 1:2 for males and 2:1 for females. Ordinarily, a deliberately unbalanced design of this sort causes no problems in the analysis, except perhaps for a reduction in efficiency.

Randomization conventions vary greatly from one area of application to another. Ordinarily, the expectation in clinical trials with human subjects is that treatment be assigned independently of covariates and independently of initial values: see Sect. 13.2. Although the mathematics requires neither, uniformity and independence of initial values can be justified on grounds of efficiency of estimation. But the more compelling reasons for uniformity and independence are more psychological than mathematical. The essence of the matter is the need for a clear and credible account that is sufficiently compelling to convince a skeptical reader, and for that purpose, simplicity is at least as important as efficiency. A treatment assignment that is exactly or approximately 50:50 for both sexes needs no explanation; a treatment assignment that is 60:40 for men and 40:60 for women will certainly invite scrutiny and skeptical commentary from reviewers. Unless otherwise stated, randomization probabilities in clinical work are invariably assumed to be independent of covariates and initial values.

For a discussion of why and how to randomize, see Sect. 5.8 of Cox (1958), Chap. 2 of Cox and Reid (2000), or Sect. 2.2 and Chap. 14 of Bailey (2008).

11.4.2 Experimental Unit

The *experimental units* are the objects to which treatment is assigned, i.e., two distinct experimental units may be assigned different treatment levels. Or, to say the same thing in a different way, two distinct experimental units are assigned different treatment levels with strictly positive probability. Each experimental unit consists of one or more observational units, e.g., one mouse consisting of four legs, or one classroom consisting of 20–40 students in the preceding example.

Two observational units u, u' belong to the same experimental unit if the randomization scheme necessarily assigns them to the same treatment level. In mathematical terms, $R(u, u') = 1$ if and only if $T(u) = T(u')$ with probability one.

By construction, R is an equivalence relation, which partitions the sample units into disjoint blocks. Each block of R is one experimental unit.

A/B testing: This phrase, which originates in commercial internet activity, refers to a treatment having two levels A, B, which may be connected with options for on-screen presentation of internet search results. Each search is an observational unit, the response being click/no click. The experimental units may be searches or users or IP addresses, depending on the circumstances.

11.4.3 Covariate and Treatment Effects

In standard probability language, the phrase ‘ X is independent of Y ’ is not a statement about the random variables as measurable functions or the pair of outcomes (X_u, Y_u) as numerical values for a particular unit, as it is a statement about probabilities: the joint probability for each product event $(X, Y) \in A \times B$ is multiplicative. Likewise, when we talk of a statistical effect in a context such as ‘the effect of treatment on the survival of patient i ’ or ‘the effect of variety on yield’, the effect referred to is not a numerical difference of two survival times or two yields, but a difference of two probabilities or a difference between two probability distributions.

For example, if the probability model asserts that the yield in kg/Ha on plot u is distributed as $N(\mu, \sigma^2)$ for variety I and $N(\mu, 2\sigma^2)$ for variety II, the effect of variety (II versus I) is implicitly to double the yield variance. The effect of variety on [the probability of] a particular event $Y_u \in A$ is the difference $N(A; \mu, 2\sigma^2) - N(A; \mu, \sigma^2)$ between two conditional probabilities, which depends on both parameters. Similar remarks apply to the effect on linear and non-linear functionals such as means, medians or quartiles of the yield distribution.

Apart from treatment effects, there are other effects of a different nature, such as the difference in survival distributions for males versus females, or the effect of aging on mobility or cognitive function. These are covariate effects. Every treatment effect in these notes is modelled as a group action on probability distributions, which is not necessarily the case for covariate effects.

The effect of a 10-year age gap on the probability of event A is the difference between two probabilities $P_u(A)$ and $P_{u'}(A)$ for two units such that $\text{age}(u') = \text{age}(u) + 10$. The fact that $\text{age}(u) \neq \text{age}(u')$ implies $u \neq u'$. Ordinarily, such a covariate effect would be computed only for pairs that are comparable in all other respects, i.e., $x(u) = x(u')$ for all other registered covariates.

The effect of treatment on the probability of A is the difference between conditional probabilities $P_u(A | T = 1)$ and $P_{u'}(A | T = 0)$ for two units having the same covariate value, i.e., $x(u) = x(u')$ for all registered covariates. Although no unit receives more than one treatment, this difference is defined and can be evaluated for $u = u'$. However, $x(u) = x(u')$ plus exchangeability implies

$$P_u(A | T = 0) = P_{u'}(A | T = 0) \quad \text{and} \quad P_u(A | T = 1) = P_{u'}(A | T = 1),$$

so the treatment effect is the same for every pair such that $x(u) = x(u')$, whether $u = u'$ or not.

11.4.4 Additivity

Additivity refers to the combination of effects associated with classification factors, block factors and treatment factors. For a two-factor model with factors A and B , the mean response is additive if there exist real numbers $\alpha_{A(u)}$, $\beta_{B(u)}$, one for each level of each factor, such that

$$E(Y_u) = \alpha_{A(u)} + \beta_{B(u)}.$$

For non-Gaussian responses, and even for Gaussian models, it may be necessary first to apply a transformation to achieve additivity. For example, if $Y_u \sim \text{Ber}(\pi_u)$ is a Bernoulli variable, the logistic model

$$\text{logit } E(Y_u) = \alpha_{A(u)} + \beta_{B(u)}$$

exhibits additivity on the logistic scale. The coefficient vectors α, β are called effects.

Additivity usually refers to the mean model, but it can also refer to random-effects models. For example if A is a treatment factor and B is a block factor, the Gaussian model

$$Y \sim N_n(\alpha_{A(\cdot)}, \sigma_0^2 I_n + \sigma_1^2 B)$$

exhibits additive treatment and block effects.

If the effects are not additive, i.e., if the treatment effect for one level of B is different from the treatment effect at some other level, we say that interaction is present. Interaction and non-additivity are effectively synonymous terms; synergy is also used for non-additivity, particularly if the treatment effect is boosted by an increase in the level of the second variable.

11.4.5 Design

The word *design* refers to the arrangement of the sample units by blocks, by covariates, and by restrictions on treatment assignment. Nelder (1965a,b) distinguishes two aspects of the design, the *structure of the units*, meaning relationships among them, and the *treatment structure*, which is imposed on them. In a crossover design, where the same physical object occurs as a distinct experimental unit on several successive occasions, the structure of the units includes not only the temporal

sequence, but also a block factor whose blocks are the distinct physical objects. In a field experiment, the structure of the units includes the geometric shape of each plot, their physical arrangement in space, and the width of access paths or guard strips separating neighbouring plots.

11.4.6 Replication

Replication means repeating the experiment independently for different experimental units under essentially identical circumstances in order to gauge the variation in response distribution. Independence is crucial. In an animal-behaviour study, it is easy to partition a one-hour observation interval into six consecutive ten-minute intervals, and to report behaviour counts for each interval. The number of animals, or pairs of animals, is unchanged, but the number of observations is immediately increased by a factor of six. Although the experimental settings may stay the same for each sub-interval, these values, sometimes called pseudo-replicates, are not independent. For a good example of an incorrect analysis for pseudo-replicates, see the *Drosophila* courtship experiment reported in Sect. 3.5.

11.4.7 Independence

In the simplest class of statistical models, the responses on distinct observational units are assumed to be distributed independently given the treatment assignment, i.e., $Y(u_1), \dots, Y(u_n)$ are independent given t . In more complicated situations such as agricultural field experiments or crossover designs or studies involving infectious diseases, the responses on distinct observational or experimental units cannot reasonably be assumed to be conditionally independent given the treatment. For example, geographic or temporal or familial relationships may induce correlations that are detectable in the data and must be accommodated in the probability model. Most of the examples illustrated in this book exhibit non-trivial correlations.

As a general rule, lack of independence is not a serious problem provided that it is recognized, and steps are taken to make accommodations in the analysis. Ordinarily, this means that block factors and other relevant relationships are recorded at baseline and used in the model to accommodate correlations.

11.4.8 Interference

If the response Y_u for one unit is statistically independent of the treatment applied to other units, we say there is no interference, or no pairwise interference. Lack of

interference is a conditional independence assumption $Y_u \perp\!\!\!\perp \mathbf{t} \mid t_u$; it does not imply independence of components, nor does independence imply lack of interference.

Some authors use the term in a different way—as a statement about outcomes rather than distributions. In Boring et al. (2016), non-interference is *the assumption that each individual's response depends only on the treatment that individual receives and not on which treatments other individuals receive*. If the word ‘response’ were followed by ‘distribution’ or ‘conditional distribution’, this interpretation would coincide with that in the preceding paragraph. As it stands, the statement is ambiguous or meaningless from a stochastic perspective.

Unless the experiment is deliberately designed to study it, interference is best avoided by design. A typical field experiment uses guard strips to separate adjacent plots; guard strips reduce interference from root competition and fertilizer seepage, but they seldom eliminate spatial correlation.

The more general definition of no interference $Y[U'] \perp\!\!\!\perp \mathbf{t} \mid \mathbf{t}[U']$ for each $U' \subset U$ is a consistency condition that requires the distribution of each restriction $Y[U']$ to depend only on the treatment restricted to U' . In the literature on causality, lack of interference is called the *stable unit-treatment distribution assumption* (Dawid 2021, Sect. 6.2). Independence and lack of interference are not so much statements of fact or fiction as they are mathematical restrictions on probability distributions. Both have implications for model formulation and analysis.

11.4.9 State Space

In a statistical model, the response is regarded as a random variable, a function $u \mapsto Y(u)$ on the observational or experimental units taking values in the *state space* \mathcal{S} , (often the real numbers). In certain settings, particularly in observational studies where all variables are regarded as responses on an equal footing, the synonym *feature space* may be used. Usually the feature space is \mathbb{R}^k for some fixed k .

It is important that the state space contain a point for every possible response-related post-baseline event that could possibly be recorded. In a pharmaceutical trial for cholesterol reduction, individual patients give informed consent and agree to abide by the protocol. However, subsequent participation is ultimately voluntary, and not all patients comply by taking their medications on the prescribed schedule. If it is recorded, compliance or the degree of compliance is a response variable, and failure to comply is one component of the response. The probability model is a probability distribution on the state space, which specifies the compliance probability, the conditional distribution given compliance, and the probability of compliance given the cholesterol levels past and future.

In all cases, the state space is a fixed measurable set, the same set for every unit, either observational unit or experimental unit, regardless of covariates. However, this restriction may lead to mathematical contortions. Consider an animal breeding study where each experimental unit is a family, and the response is measured on individual family members (offspring only) at age six weeks. Suppose that family

size x is a covariate recorded at baseline, in which case the response Y_u for a family of size $x(u)$ is a point in $\mathbb{R}^{x(u)}$. The variation of the state space from one experimental unit to another depending on the covariate $x(u)$ appears to violate the definition of state space as a fixed set. But this violation is a mathematical illusion. We can simply re-define the state space to be the disjoint union $\mathcal{S} = \cup_{k \geq 0} \mathbb{R}^k$, and construct the probability distribution on \mathcal{S} in such a way that all of the probability mass for unit u resides in the component $x(u)$ of the state space,

$$\text{pr}(Y_u \in \mathbb{R}^k) = \begin{cases} 1 & x(u) = k \\ 0 & \text{otherwise.} \end{cases}$$

Note that x is not a random variable, so we have not written this as a conditional probability statement.

If the measurements were weights at birth rather than later at six weeks, the baseline would necessarily have to be pre-natal, implying that family size X is a part of the response, not a covariate recorded at baseline. In that setting the response Y is a random variable taking values in \mathcal{S} , and the response distribution F determines the distribution of X by $\text{pr}(X = k) = F(\mathbb{R}^k)$ (including $k = 0$). The conditional distribution given X is a function that associates with each integer $k \geq 0$ a probability distribution $F(\cdot | X = k)$ such that $F(\mathbb{R}^k | X = k) = 1$.

11.4.10 State-Space Evolution

In a study of survival times following surgery, each patient is one unit, and the response is a survival time $Y_u > 0$, which is, *prima facie* at least, a point in \mathbb{R}^+ , the positive real line. Only the most persnickety mathematician would bother to add a point at infinity to cover the remote possibility of immortality, which cannot be ruled out solely on mathematical grounds. However, the response $Y_u^{(t)}$ as it exists today or at the time of analysis, say $t = 1273$ days post-recruitment, is either a failure time in the interval $t^- = (0, t]$, or a not-yet-failure corresponding to the ‘point’ t^+ , which is required to exist as a point in the state space for today. In other words, $\mathcal{S}^{(t)} = t^- \cup \{t^+\}$, the union of a bounded interval and a topologically isolated ‘point’ exceeding each number in the interval. The limit space $\mathcal{S}^{(\infty)} = \mathbb{R}^+ \cup \{\infty\}$ differs from \mathbb{R}^+ by one isolated point that exceeds every real number.

To say the same thing in another way, the state space is a filtration that evolves as an increasing σ -field in calendar time.

To every non-negative measure Λ on the positive real numbers there corresponds a distribution F on $\mathcal{S}^{(\infty)}$ given by $F(t^+) = \exp(-\Lambda(t^-))$, where t^+ is the complement of t^- in $\mathcal{S}^{(\infty)}$. Usually, F is called the survival distribution, and Λ is the hazard measure. If the total hazard $\Lambda(\mathbb{R}^+)$ is finite, the atom of immortality $F(\{\infty\}) = \exp(-\Lambda(\mathbb{R}^+))$ is strictly positive; otherwise the atom is zero. With respect to the state of information at time t , the probability density at $y \in \mathcal{S}^{(t)}$ is

$\Lambda(dy) \exp(-\Lambda(y^-))$ for $0 < y \leq t$, and $\exp(-\Lambda(y^-))$ for $y = t^+$. In particular, if Λ is proportional to Lebesgue measure on \mathbb{R}^+ , the density is $\lambda e^{-\lambda s} ds$ for $0 < s \leq t$ with an atom $e^{-\lambda t}$ at t^+ .

Being alive at the time of analysis is one unavoidable form of censoring. In practice, some patients disappear off the radar screen at a certain point $t > 0$, and their subsequent survival beyond that time cannot be ascertained. These also are typically regarded as censored at the last time they were known to be alive.

11.4.11 Longitudinal Study

In a longitudinal study, also called a panel study, each physical unit is measured at a sequence of time points. Growth studies, of plants or of animals, are of this type, the response $Y(i, t)$ being height or weight of unit i at time t . Usually the design calls for measurements to be made at regular intervals, but in practice the intervals tend to be irregular to some degree, particularly for studies involving human subjects.

A typical longitudinal design has a large number of subjects measured on a relatively small number of occasions. The first of these measurements is made at or pre-baseline. If the experiment has a randomized treatment assignment, the first measurement is ordinarily pre-randomization before the treatment is decided, and certainly before it can have had an effect. In the modelling and analysis, it may be necessary to include a null treatment level to denote pre-randomization status; this level is in addition to the control and active post-baseline levels.

11.4.12 Cemetery State

A situation arises in geriatric and other medical studies where, beginning at recruitment, measurements on physical or mental capacity are made annually on patients—but only while they are alive. All patients ultimately die, and the number k_i of measurements on patient i is a major part of the response, which is closely connected with survival time. In this setting, each patient may be regarded as an observational unit, in which case the response $Y_i = (Y_i(0), \dots, Y_i(k_i - 1))$ is a point in the state space $\cup_{k \geq 0} \mathbb{R}^k$ implying death before time k_i . Alternatively, if each patient-time combination is regarded as one observational unit, it is necessary to add to the real numbers an absorbing state, such that $Y_i(t) = \flat$ implies that patient i is dead at time t . The state space for one observational unit is $\mathbb{R} \cup \{\flat\}$; the state space for one experimental unit (patient) is $\mathcal{S}^{(\infty)} = (\mathbb{R} \cup \{\flat\})^\infty$, each sequence \flat -padded on the right where needed.

As always, the state space at calendar time s includes only those events observed or observable up to that time; the state space is censored by the calendar, not by the death of patients.

11.5 Non-comparative Studies

11.5.1 Examples

An experiment designed to measure the speed of light *in vacuo* is not comparative; the goal is not to estimate the ratio of the speed *in vacuo* relative to that in some other medium, but to estimate the absolute speed in km/s for a particular medium. A survey whose goal is to estimate the prevalence of COVID-19 antibodies in Santa Clara County in April 2020 is not comparative; the goal is not to estimate the prevalence in Santa Clara relative to that in San Mateo, but to estimate the absolute prevalence as a percentage of the county population. An opinion poll with the aim of predicting the outcome of a plebiscite or general election is not comparative; the goal is to predict the outcome of the election on a particular day.

The avoidance of bias or systematic errors is important in all branches of science, but it is especially important in non-comparative studies. The next few sections consider the effect of response heterogeneity in a stratified population, finite or infinite.

11.5.2 Stratified Population

A function $x : \mathcal{U} \rightarrow [k]$ taking values in the finite set $[k] = \{1, \dots, k\}$ determines a partition of the units into k disjoint subsets, $\mathcal{U}_1, \dots, \mathcal{U}_k$ called strata or blocks:

$$\mathcal{U}_r = \{u : x(u) = r\}.$$

In general, \mathcal{U} may be finite or infinite; if \mathcal{U} is not finite, at least one stratum is also not finite. In practice, if \mathcal{U} is infinite, all of the strata are also infinite.

For current-population sampling applications, \mathcal{U} is finite; to a close approximation x is known from the preceding census, so the strata sizes are also known in the same sense. Every classification variable such as *sex* determines a stratification; every pair of variables such as (*sex*, *location*) determines a finer stratification, and so on. For example, *location* might have levels *rural*, *suburban*, *urban*. The classification variables that are available for survey-sampling are mostly restricted to those recorded in the census.

11.5.3 Heterogeneity

Heterogeneity means that the distribution of response values in one stratum is not the same as the distribution in another stratum, or at least similarity is not to be assumed. The implication, ironically, is that the values within each stratum can

be taken as exchangeable—infinitely exchangeable in the case of infinite strata, or finitely exchangeable otherwise. Exchangeability is either an explicit assumption, or it is forced as a consequence of random sampling.

11.5.4 Random Sample

A random sample is a finite random subset $U \subset \mathcal{U}$ taken from the population. A *simple random sample* of size n taken from a finite population of size N is a random subset selected uniformly from the set of all subsets of size n . More correctly, a simple random sample is a function φ chosen uniformly at random from the set of 1–1 functions $[n] \rightarrow [N]$. The sample $(\varphi_1, \dots, \varphi_n)$ is an ordered list consisting of n distinct units taken from the population; the sample value is the composite function $Y \circ \varphi$

$$[n] \xrightarrow{\varphi} [N] \xrightarrow{Y} \mathbb{R},$$

giving the values $(Y_{\varphi(1)}, \dots, Y_{\varphi(n)})$.

Operationally speaking, we first arrange the population units $1, \dots, N$ in uniform random order $\sigma(1), \dots, \sigma(N)$ by a uniform random permutation σ . By definition, the permuted values $(Y_{\sigma(1)}, \dots, Y_{\sigma(N)})$ are finitely exchangeable in the usual sense that the distribution is unaffected by permutation. The leading subset $\varphi = (\sigma(1), \dots, \sigma(n))$ is a simple random sample, and the sample value is $(Y_{\sigma(1)}, \dots, Y_{\sigma(n)})$. In other words, a simple random sample is a fixed sample taken from the randomly permuted population. Simple random sampling is the guarantor of exchangeability.

Ordinarily, the term *random sample* implies that the sample is not only a random subset $U \subset \mathcal{U}$, but also a subset that is independent of the process Y . A size-biased sample is one in which U is random but not independent of Y .

11.5.5 Stratified Random Sample

A stratified random sample with sizes n_1, \dots, n_k consists of k simple random samples, one independent sample from each stratum.

11.5.6 Accessibility

It is possible to select a finite random sample from an infinite population. But simple random sampling and stratified sampling are possible only for finite populations. In

practice, any form of random sampling is feasible only for the sub-population that is currently accessible. For example, a population consisting of a lineage of breeding flies that evolves in time is only partly accessible in any bounded temporal window.

11.5.7 Population Averages

The mean for stratum r

$$\mu_r = E(Y_u : x(u) = r)$$

is either a finite average if \mathcal{U}_r is finite, or a distributional mean of exchangeable random variables otherwise. In a finite or infinite population with strata fractions (π_1, \dots, π_k) adding to one, the weighted linear combination

$$\mu_\pi = \pi_1 \mu_1 + \dots + \pi_k \mu_k$$

is called the population average.

In the case of a locally finite population consisting of $N_t = \#\mathcal{U}_t$ units at time t , $N_r(t)$ is the stratum total, $\pi_r(t) = N_r(t)/N_t$ is the stratum fraction, and the democratic average $\mu_{\pi(t)} = \sum_{u \in \mathcal{U}_t} Y_u/N_t$ is the arithmetic mean in the current population.

11.5.8 Target of Estimation I

In a stratified population, the target of estimation is usually the stratum mean vector $\mu = (\mu_1, \dots, \mu_k)$. However, there are various applications, particularly related to marketing, opinion polling and voting, where the democratic average plays an outsize role. In the run-up to a crucial plebiscite such as the Brexit referendum, the democratic average of voter preferences looms so large that between-stratum variation is of little consequence.

11.5.9 Inverse Probability Weighting

Consider a stratified population consisting of 12m voters, 5m urban, 4m suburban and 3m rural. In a stratified random sample in Oct 2020, 500 voters out of 1000 declared that they would vote for candidate T; the breakdown by strata was as follows.

| | Urban | Suburban | Rural | Total |
|--------------|-------|----------|-------|-------|
| Stratum size | 5m | 4m | 3m | 12m |
| π | 5/12 | 4/12 | 3/12 | 1 |
| Sample size | 400 | 300 | 300 | 1000 |
| Candidate T | 110 | 140 | 250 | 500 |
| \bar{y} | 0.275 | 0.467 | 0.833 | 0.500 |

Note that the stratum relative proportions 5m:4m:3m are close to the sample fractions 4:3:3, but not exactly the same. The stratum averages for this sample are $\bar{y} = (0.275, 0.467, 0.833)$, and the population-weighted linear combination of stratum averages is

$$\hat{\mu}_\pi = 0.275 \times 5/12 + 0.467 \times 4/12 + 0.833 \times 3/12 = 0.4785$$

which is less than the equally-weighted poll average 500/1000.

The preceding calculation is an instance of a weighted linear combination of sample values,

$$\hat{\mu}_\pi = \sum_{i \in U} w_i Y_i \Big/ \sum_{i \in U} w_i,$$

where the weights are inversely proportional to the first-order sample inclusion probabilities (Horvitz & Thompson, 1952). Each urban voter has a sample inclusion probability 400/5m, so $w_i = 5/400$; each suburban voter has inclusion probability 300/4m, so $w_i = 4/300$; and each rural voter has inclusion probability 300/3m, so $w_i = 3/300$. The sum of these weights is 12, and the linear combination is displayed in the preceding paragraph.

11.5.10 Target of Estimation II

The calculation illustrated in the preceding section is as obvious as it is uncontroversial. It is obvious as a matter of arithmetic, and it is uncontroversial because of the political setting used for its illustration. But inverse-probability weighting is not something to be taken for granted in other settings that might appear superficially similar.

Consider the COVID-19 antibody prevalence study for Santa Clara County in April 2020. The main controversy in the Stanford study centered correctly on the false-positive rate of the antibody test, which was of a magnitude similar to the reported prevalence. See the online blog by Gelman (2020) titled *Concerns with*

that Stanford study of coronavirus prevalence. For present purposes, we set that matter aside and suppose optimistically that the false-positive rate is zero.

Suppose that a similar set of numbers—suitably scaled to represent plausible prevalences—had arisen in the COVID-19 antibody prevalence study.

| | Urban | Suburban | Rural | Total |
|----------------|-------|----------|-------|-------|
| Stratum size | 0.5m | 0.4m | 0.3m | 1.2m |
| π | 5/12 | 4/12 | 3/12 | 1 |
| Sample size | 400 | 300 | 300 | 1000 |
| Antibody cases | 4 | 8 | 13 | 25 |
| \bar{y} | 0.010 | 0.027 | 0.043 | 0.050 |

Would it be appropriate to use the same weighted procedure

$$\hat{\mu}_\pi = 0.010 \times 5/12 + 0.027 \times 4/12 + 0.043 \times 3/12 = 0.024$$

and report only the county-wide antibody prevalence at 2.4%? I should hope not!

The crucial difference is not the numbers but the setting. For the political poll, the current-population average is the natural target mandated by democratic principles and supported by the force of law. In the epidemiological setting, the democratic average or prevalence is a natural summary, but it does not carry an equivalent epidemiological or legal mandate. Nor is it necessarily the most interesting summary or the most striking feature to emerge from such a study. In the table shown above, the observed prevalence in the rural community is more than four times that in the urban community. Admittedly, the case numbers are small, so the ratio in the population might not be so extreme. But a risk ratio or prevalence ratio as large as 3–4 calls out for an explanation, and that finding could be more interesting epidemiologically than the particular value of the county-wide prevalence.

The main point is that the exclusive focus on county-wide prevalence is a distraction that has the potential to divert attention away from features that are epidemiologically more interesting. Any epidemiologist who reported only the prevalence of 2.4% would be derelict in his or her duty to draw attention to the extreme variation in rates for urban versus rural communities. To conclude, inverse-probability weighting is satisfactory as a summary statistic for a stratified population in two circumstances only: either the democratic average is mandated by law; or the degree of heterogeneity is moderate. In the latter case, the choice of stratum weights matters little.

11.6 Interpretations of Variability

11.6.1 A Tale of Two Variances

The Ewens sampling formula (Ewens, 1972), is the static probabilistic description of an exchangeable process, which can be viewed as a sequence Y_1, Y_2, \dots of species or types. In its original genetic form, $P_{n,\alpha}$ is the probability distribution of the number N of distinct alleles and the multiplicity of each type occurring in a sample of n individuals (technically haplotypes) taken from an infinite population that evolves neutrally with mutation rate α . This combinatorial stochastic process is a thing of uncommon mathematical beauty; it occurs in a surprisingly wide range of mathematical and scientific applications from linguistic studies to genetics to ecology and probabilistic number theory (Crane, 2016; Pitman, 2006; Tavaré, 2021). Only two distributional facts are relevant to the present story.

The first fact is that the number of distinct types in a sample of n objects is equal in distribution to the sum of n independent Bernoulli variables

$$N \sim X_1 + X_2 + \cdots + X_n,$$

where X_i is Bernoulli with parameter $\alpha/(i - 1 + \alpha)$. The Bernoulli parameter is the probability that the i th specimen is a new type that is different from previous specimens $1, 2, \dots, i - 1$. The sequential description leading to this conclusion is called the Chinese restaurant process: see Exercises 11.9–11. Thus, the expected value is

$$\begin{aligned} E(N) &= 1 + \frac{\alpha}{1 + \alpha} + \frac{\alpha}{2 + \alpha} + \cdots + \frac{\alpha}{n - 1 + \alpha}, \\ &= \alpha\psi(n + \alpha) - \alpha\psi(\alpha), \\ &= \alpha \log(n) + O(1), \end{aligned}$$

where ψ is the derivative of the log-gamma function. A similar calculation shows that the variance is less than the mean, but only slightly, $\text{var}(N) = \alpha \log(n) + O(1)$. In fact all cumulants of all orders differ from the mean by $O(1)$. To this order of approximation, the species count is Poisson with parameter $\alpha \log(n)$.

The second fact is that $P_{n,\alpha}$ is a one-parameter exponential family with canonical parameter $\log \alpha$, and canonical sufficient statistic N . Accordingly, the maximum-likelihood estimate is unique for $N < n$ and satisfies

$$\psi(n + \hat{\alpha}) - \psi(\hat{\alpha}) = N. \tag{11.5}$$

The Poisson approximation $N \sim \text{Po}(\alpha \log n)$ is adequate for present purposes. It implies that the maximum-likelihood estimate $\hat{\alpha} \simeq N/\log n$ is consistent with asymptotic variance

$$\text{var}(\hat{\alpha}) = \frac{\alpha}{\log n}. \quad (11.6)$$

Thirty years earlier, R.A. Fisher (1943), together with A.S. Corbet and C.B. Williams, had considered the problem of determining the distribution of the number of species to be found in a sample of n specimens. This very brief paper is now considered to be a landmark contribution to mathematical ecology. It is not exactly a joint paper in the modern sense, but rather a tripartite paper in which each author makes a separate contribution. Fisher contributed the theory; Corbet and Williams supplied the moths and butterflies.

Although Fisher's derivation is totally different from Ewens (1972), and the crucial concept of a process is absent, his gamma mixture model coincides in all essential respects with the Ewens sampling model for fixed n . In particular, Fisher's formulae contain a species-diversity parameter α which coincides with the mutation rate in genetic applications. His expression for the maximum-likelihood estimate looks nothing like (11.5), but it is numerically equivalent, at least to the present order of approximation. However, Fisher's expression for the variance is very different from (11.6). In essence, Fisher's variance formulae are equivalent to the statements

$$\text{var}(N) = \alpha \log 2; \quad \hat{\alpha} = \frac{N}{\log n}; \quad \text{var}(\hat{\alpha}) = \frac{\alpha \log 2}{(\log n)^2}, \quad (11.7)$$

the first and last being substantially at odds with (11.6).

The paragraph in which Fisher derives this formula is brief and incomprehensible. When I first read it carefully, I observed that the variance is not correct, and the conclusion (McCullagh, 2016) that he must have made a mistake somewhere in his derivation seemed unavoidable. Technical errors of this sort are exceedingly rare in Fisher's published work, so I was not entirely comfortable with this conclusion. Certainly, I could not identify a specific point where the error occurred. The source of the mysterious log 2-factor could not be identified.

It turns out that the discrepancy was noted first by F.J. Anscombe, who set out boldly to clarify the matter at a meeting with Fisher in April 1947, and in correspondence thereafter. A passage taken from Sect. 6 of Anscombe (1950) suggests that what Fisher had in mind was a different notion of variability—as measured by variability in successive samples rather than variability in repeated independent samples. According to Anscombe, Fisher's variance formula

...is appropriate to a special type of comparison, namely, between estimates of α derived from similar nearby traps, where it may be assumed that the individual species have exactly the same relative abundances, and the difference between the catches at any two traps arises solely from Poisson variation in the numbers caught of each species.

While he appeared to accept Fisher's claim at face value, Anscombe offered no proof. The phrasing of his explanation, which hews closely to passages from his correspondence with Fisher, offers no insight into Fisher's derivation.

With this in mind, we can now ask of the Ewens process, 'how much variability should be expected among a sequence of species counts N_1, \dots, N_m , or a sequence of estimates $\hat{\alpha}_1, \dots, \hat{\alpha}_m$, taken from m non-overlapping samples of n specimens all in a single trajectory?' In other words, N_1 is the number of species occurring among specimens $1, \dots, n$; N_2 is the number of species occurring among specimens $n + 1, \dots, 2n$; and so on up to sample m consisting of specimens $(m - 1)n + 1, \dots, mn$. For this scheme of samples taken in succession, Da Silva et al. (2022) show that the species counts are exchangeable Poisson variables with variances and covariances

$$\text{var}(N_i) = \alpha \log n \quad \text{and} \quad \text{cov}(N_i, N_j) = \alpha \log(n/2)$$

for $i \neq j$. Thus, the expected value of the sample variance of the species counts from successive samples is

$$\begin{aligned} E \frac{1}{m-1} \sum_{j=1}^m (N_j - \bar{N})^2 &= E \frac{1}{m(m-1)} \sum_{i < j} (N_i - N_j)^2 \\ &= \frac{1}{2} E((N_1 - N_2)^2) \\ &= \alpha \log(n) - \alpha \log(n/2) = \alpha \log 2. \end{aligned}$$

It is remarkable that this inter-sample variance is independent of the sample size.

For the sequence of parameter estimates, $\hat{\alpha}_j = N_j / \log(n)$ and $\bar{\hat{\alpha}} = \bar{N} / \log(n)$, a similar argument gives

$$E \frac{1}{m-1} \sum_{j=1}^m (\hat{\alpha}_j - \bar{\hat{\alpha}})^2 = \frac{\alpha \log 2}{(\log n)^2}$$

in agreement with (11.7). The covariances come from the species that are common to pairs of samples, each of which is distributed as $\text{Po}(\alpha \log(n/2))$, explaining the origin of the mysterious log 2-factor. It seems safe now to conclude that Fisher did not make a technical error; it looks as if his variance formula was meant to be interpreted in this unorthodox way to ensure that it would be relevant to the specific population in which the sample had been collected. Fisher gave no indication that he had any formal concept of a stochastic process in mind, so the fact that he computed and interpreted his variance in this way can only be regarded as remarkably farsighted.

11.6.2 Which Variance Is Appropriate?

Having observed a process Y on a finite sample U , the goal of statistical inference in general is to make probabilistic statements about extra-sample values that have not been observed. The goal of parametric inference is specifically to make probabilistic statements about limit values or tail behaviour. Both questions are concerned with a single trajectory of the process, extrapolating from the observed sample to new samples and ultimately to the entire population. In that narrow mathematical sense, all inference is prediction. It is now reasonably clear that Fisher had the appropriate inferential goal in mind.

In the case of the Ewens process, we first observe the species count T_n for a sample of size n . Equivalently, the observation may be recorded as an estimate $\hat{\alpha}_n = T_n / \log(n)$. The short-term goal is to predict the progress of cumulative species counts $T_n \leq T_{n+1} \leq T_{n+2} \leq \dots$ or to predict the subsequent sequence $\hat{\alpha}_{n+1}, \dots$ given T_n . These short-term activities are predictive in the literal sense. The long-term inferential goal is to predict the limit statistic

$$\hat{\alpha}_\infty = \lim_{r \rightarrow \infty} \hat{\alpha}_{n+r} = \lim_{r \rightarrow \infty} T_{n+r} / \log(n+r).$$

The existence of a deterministic limit is known in the sense $\hat{\alpha}_\infty = \alpha$ with probability one, but the limit value is not revealed. Both forms of prediction, parametric and non-parametric, refer specifically to extensions of the single trajectory that is partially observed. In other words, both inferential tasks refer to the extension that Fisher appears to have had in mind when he derived the formulae (11.7). Neither task calls for independent replication for fixed α , which is the scenario that leads to the conventional Anscombe formula (11.6), which is also the Fisher-information formula or parametric bootstrap formula.

Fisher's focus on the single-trajectory extension is entirely apposite. And his implied statement that the mean squared difference between estimates from non-overlapping samples of the same size satisfies

$$E \frac{1}{m-1} \sum_j (N_j - \bar{N})^2 = \alpha \log 2$$

is also correct—even conditionally on N_1 . Fisher's calculation for large m also implies that the observed value $\hat{\alpha}_1$ differs in mean square from the average of subsequent estimates by

$$E((\hat{\alpha}_1 - \bar{\alpha})^2 | N_1) = \frac{\alpha \log 2}{(\log n)^2}.$$

This line of argument makes it appear that Fisher's unorthodox variance formula is the correct variance for purposes of parametric inference in applications where there is a strong serial correlation between successive samples. However, there is a catch.

The combined statistic $\bar{\hat{\alpha}}$ is not the maximum-likelihood estimate based on the combined sequence of length mn . It is the average of m estimates based on non-overlapping samples of size n for which the variances and covariances are

$$\text{var}(\hat{\alpha}_i) = \frac{\alpha}{\log n}; \quad \text{cov}(\hat{\alpha}_i, \hat{\alpha}_j) = \frac{\alpha}{\log n} - \frac{\alpha \log 2}{(\log n)^2}.$$

The variance of the average is the average of variances and covariances

$$\text{var}(\bar{\hat{\alpha}}) = \frac{\alpha}{\log n} - \frac{(m-1)\alpha \log 2}{m (\log n)^2},$$

which, for fixed n , does not tend to zero as $m \rightarrow \infty$. As an estimator, the average of $\hat{\alpha}_1, \dots, \hat{\alpha}_m$ is not appreciably better than $\hat{\alpha}_1$. Although $\bar{\hat{\alpha}}$ may have a limit as $m \rightarrow \infty$, it is not a deterministic limit and it is not equal to α .

The species counts N_1, \dots, N_m for successive samples do not determine the total number T_{mn} of distinct species for the combined sample of length mn . Many species are expected to occur in several samples, but the overlaps cannot be determined from the marginal counts alone. Thus $(\hat{\alpha}_1, \dots, \hat{\alpha}_m)$ is not a sufficient statistic for the combined sample. In fact, the information in the marginal counts jointly is negligible compared with that in T_{mn} : the Fisher-information numbers are $O(\log n)$ and $O(\log n + \log m)$ respectively.

The relevant inferential calculation focuses on the difference $\hat{\alpha}_1 - \hat{\alpha}_\infty$, for which the mean square is

$$E(\hat{\alpha}_1 - \hat{\alpha}_\infty)^2 = \frac{\alpha}{\log n}.$$

Thus, Anscombe's asymptotic variance formula—computed ironically from the Fisher information formula—is the correct variance for purposes of parametric inference in this setting. Fisher's variance formula may be correct for the mean squared difference between successive samples of equal size, but it mischaracterizes the mean squared difference between the sample value $\hat{\alpha}_1$ and the limit value for a single trajectory.

11.7 Exercises

11.1 The black-footed ferret is an endangered species; it belongs to the weasel family. A ferret-breeding program has been established by various zoos throughout the United States to study the factors that affect breeding success in captive ferrets. The sample consists of 664 females, 375 males, and 1700 M-F pairings. In one season, a given female may occur in up to four pairings; a male may occur in up to

eight pairings. A given pair of ferrets may occur in two or more pairings in the same season or in different seasons.

The following variables are recorded at birth for each ferret:

```
ferret_id, sex, birth_year, zoo_id;
```

There are six participating zoos, and `zoo_id` is the zoo of birth. The following variables are recorded for each of the 1700 M-F pairings:

```
male_id, female_id, zoo_id, year, kinship, whelped;
```

The kinship coefficient is a standard measure of genetic relatedness of the breeding pair. The unit of observation is a male-female pairing, and the response is whelping success, which is binary.

In most cases, the ferrets spend their life at the zoo of birth, so `zoo_id` is the zoo of birth of both ferrets. However, inter-zoo transfers are possible, in which case the breeding zoo might be different from the zoo of birth for one or other ferret. The goal is to identify factor combinations that promote breeding success.

Discuss the role of each variable in this study.

Ferrets live about 4–6 years in captivity, so ferret age may be important for breeding success. Given that there are two ferrets in each pair, discuss how you would compute the age of each ferret, and how you would incorporate age effects into any model. Available options include linearly, symmetrically, additively, and so on.

`male_id` is an equivalence relation on pairings, as is `female_id`. What bearing does this have on modelling?

11.2 Discuss the connection between sampling consistency as described in Sect. 11.6.1 and lack of interference as described in Sect. 11.4.8.

11.3 For integer $n \geq 1$, a partition B of the set $[n] = \{1, \dots, n\}$ is a set of disjoint non-empty subsets called blocks whose union is $[n]$. A partition into k blocks, can be written as $B = \{b_1, \dots, b_k\}$, with the understanding that B is a set of subsets, not a list of subsets. For example, 12|34, 13|24, 14|23 are distinct partitions of [4] into two blocks of size two, and these are the only partitions of type 2 + 2. Equivalently, B is an equivalence relation $[n] \times [n] \rightarrow \{0, 1\}$, reflexive, symmetric and transitive. Let \mathcal{P}_n be the set of partitions of $[n]$. List the elements in \mathcal{P}_n for each $n \leq 5$, and show that $\#\mathcal{P}_1 = 1$, $\#\mathcal{P}_2 = 2$, $\#\mathcal{P}_3 = 5$, $\#\mathcal{P}_4 = 15$, $\#\mathcal{P}_5 = 52$. These are called the Bell numbers.

11.4 One of the simplest static versions of the Ewens sampling formula is stated as a probability distribution on the set of partitions of the finite set $[n]$ as follows:

$$P_{n,\alpha}(B) = \frac{\alpha^{\#B} \prod_{b \in B} (\#b - 1)!}{\alpha^{\uparrow n}},$$

where $\alpha^{\uparrow n} = \alpha(\alpha + 1) \cdots (\alpha + n - 1)$ is called the rising factorial. Show that $P_{n,\alpha}$ is an exponential-family model with canonical parameter $\theta = \log \alpha$, canonical statistic $\#B$, and cumulant function $\log((e^\theta)^{\uparrow n}) - \log(n!)$.

11.5 By direct calculation, show that the Ewens distributions satisfy the following conditions:

$$\begin{aligned} P_{2,\alpha}(12) &= P_{3,\alpha}(\{123, 12|3\}) \\ P_{2,\alpha}(1|2) &= P_{3,\alpha}(\{13|2, 1|23, 1|2|3\}) \\ P_{3,\alpha}(123) &= P_{4,\alpha}(\{1234, 123|4\}) \\ P_{3,\alpha}(1|23) &= P_{4,\alpha}(\{14|23, 1|234, 1|23|4\}) \\ P_{3,\alpha}(1|2|3) &= P_{4,\alpha}(\{14|2|3, 1|24|3, 1|2|34, 1|2|3|4\}). \end{aligned}$$

Show that $P_{4,\alpha}$ is the marginal distribution of $P_{5,\alpha}$ when the element 5 is removed from the set [5]. Hence calculate the conditional distribution $P_{5,\alpha}(x \mid B)$ for $x \in \mathcal{P}_5$ and $B = 1|3|24$ and $B = 13|24$.

11.6 The simplest sequential description of the Ewens sampling formula is called the Chinese restaurant process. The first customer arrives and is seated at a table. After n customers have been seated, the next customer is seated alone with probability $\alpha/(n + \alpha)$; otherwise, the newcomer selects one of the seated customers uniformly at random and sits at that table. Show that the configuration after n customers are seated is given by $P_{n,\alpha}$. Hence deduce that customers seven and nine are seated together with probability $1/(\alpha(\alpha + 1))$.

11.7 To the order of approximation used in Sect. 11.6, show that the maximum-likelihood estimate, $\hat{\alpha}(T_n)$, of the species-diversity parameter as a function of the cumulative species count T_n , defines a martingale.

11.8 Let $B \sim P_{n,\alpha}$ be the partition after n customers in the Chinese restaurant process with parameter α , and let $\hat{\alpha}(B)$ be the maximum-likelihood estimate. One way to approximate the variance of $\hat{\alpha}(B)$ is to generate bootstrap samples B_1^*, \dots, B_m^* by simulation, and to report the sample variance of the bootstrap estimates $\hat{\alpha}(B_1^*), \dots, \hat{\alpha}(B_m^*)$. In the parametric version, the bootstrap samples are conditionally independent and identically distributed according to the Ewens distribution $B_i^* \sim P_{n,\hat{\alpha}(B)}$. For large n , show that

$$\begin{aligned} E(\hat{\alpha}(B^*) \mid B) &= \hat{\alpha}(B) + o(1) \\ \text{var}(\hat{\alpha}(B^*) \mid B) &= \hat{\alpha}(B)/\log n + o(1/\log n) \end{aligned}$$

in agreement with the standard maximum-likelihood calculation (11.6).

11.9 In the non-parametric bootstrap, the configuration B is regarded as a list of n tables in order of occupation. Each non-parametric bootstrap sample is a sequence of n tables drawn with replacement from the empirical distribution of tables. Although $\hat{\alpha}(B^*) \leq \hat{\alpha}(B)$ for every bootstrap sample, show that

$$\begin{aligned} E(\#B - \#B^*) &= -\alpha \log(1 - e^{-1}) + o(1) \\ E(\text{var}(\#B^*) \mid B) &= -\alpha \log(1 - e^{-1}) + \alpha \log(1 - e^{-2}) + o(1). \end{aligned}$$

Hence deduce that the sample variance of bootstrap estimates satisfies

$$E(\text{var}(\hat{\alpha}(\#B^*)) \mid B) \simeq \alpha \text{ const}/\log^2 n$$

in order-of-magnitude agreement with Fisher's calculation. Show that the bootstrap constant is not the same as Fisher's constant in (11.7).

Chapter 12

Principles



12.1 Sampling Consistency

The chief guiding principle in this book is the need for a mathematical framework for thinking about measurements and how they are related to physical, chemical, biological or environmental processes. Invariably, a measurement on a sample is regarded as a random variable, so the mathematical framework is a stochastic process or family of stochastic processes.

A Gaussian model expressed in the form

$$Y \sim N_n(\mu_n(\mathbf{x}), \Sigma_n(\mathbf{x})) \quad (12.1)$$

is meant as a distributional specification that applies to an arbitrary fixed sample $U = \{i_1, \dots, i_n\}$ having covariate configuration $\mathbf{x} \equiv \mathbf{x}[U]$ and response $Y \equiv Y[U]$. For a sample of size one, $U = \{i\}$, the distribution is $Y_i \sim N(\mu_1(x_i), \sigma_1^2(x_i))$, so the mean and variance depend only on x_i —not on x_j for any unit $j \neq i$. For a sample $U = (i, j)$ of size two with $\mathbf{x} = (x_i, x_j)$, consistency implies that the mean vector is necessarily a component-wise function

$$\mu_2(\mathbf{x}) = (\mu_1(x_i), \mu_1(x_j)).$$

Consistency for covariances, implies that the diagonal values $\sigma_1^2(x_i), \sigma_1^2(x_j)$, are determined by component-wise extension of the scalar function $\sigma_1^2(\cdot)$. The off-diagonal value $\Sigma_n(x_i, x_j) = \Sigma_2(x_i, x_j)$ is a function only of covariates and pairwise relationships; it is independent of the sample size and remaining sample configuration.

These constraints are strong and restrictive, but they are also minimal and mathematically natural. Without this form of consistency, the entire mathematical edifice underpinning all of sampling theory crumbles, and statistical inference as we have come to know it is not possible. Prediction in the sense of the conditional

distribution given the observation does not exist. Parameter estimation in the sense of probabilistic statements about infinite averages or other tail events is also impossible.

Kolmogorov consistency is not a matter of mathematical fact, nor is it a statement of physical reality; it is an assumption that can be viewed as a statement of mathematical sanity. It provides a framework for thinking about samples and sub-samples, measurements on samples, and how they are related to physical, chemical, biological or environmental processes. Without consistency, there is only mathematical chaos.

Sampling Inconsistency

To all probabilists and statisticians, the need for consistent specification of probabilities is self-evident and requires no emphasis and little explanation. However, this sentiment is not universal in applied work. A few instances taken from the literature on the spatial distribution of economic and business activities suffice as illustration.

Dong et al. (2015) discuss the problem of modelling the emerging property market in Beijing, China. For real-estate purposes, each observational unit i is an administrative region called a land parcel. There are $n = 1117$ such parcels, all disjoint subsets, which appear to cover the entirety of greater Beijing. They are partitioned into $d = 111$ districts of various sizes. For the most part, the districts and the parcels are relatively compact and simply connected. Inter-parcel distances d_{ij} are measured in km. from a representative point in each parcel; districts are related by adjacency or contiguity.

Dong et al. begin with a simultaneous autoregressive spatial formulation, which *...has been extensively studied in the spatial economics literature and is widely used in geographical research. A key characteristic ... is that it allows the observed value at a particular location to be directly dependent on the values observed at surrounding locations (or lagged y)...* Five references are cited in support of the wide usage in geographical research. So far, so good.

The following components are simplified slightly in the interests of clarity. First, W is a trace-free stochastic matrix, a non-negative function of distances whose row sums are all one; roughly speaking, $W_{ij} = \exp(-d_{ij}^2)/c_i$ for $i \neq j$. Second, M is a similar $n \times d$ contiguity matrix for districts. The simultaneous-equation rationale begins with a literal interpretation of the quoted remark in the form of a stochastic vector equation

$$Y = \rho W Y + X\beta + M\eta + \varepsilon, \quad (12.2)$$

with the usual assumptions on the distribution of η, ε . The conclusion is that $(I - \rho W)Y$ is Gaussian with mean $X\beta$ and covariance $\sigma_0^2 I_n + \sigma_1^2 MM'$, which implies that Y is Gaussian with mean and covariance

$$\mu(\mathbf{x}) = (I - \rho W)^{-1} X\beta; \quad (12.3)$$

$$\Sigma(\mathbf{x}) = (I - \rho W)^{-1} (\sigma_0^2 I_n + \sigma_1^2 MM') (I - \rho W)^{-1}. \quad (12.4)$$

It follows that $\mu_i(\mathbf{x})$ depends not only on x_i but also on x_j for $j \neq i$, in violation of the most elementary consistency condition in the preceding section.

A quick perusal shows that the use of simultaneous spatial autoregressive formulations is not at all uncommon in parts of the economic and business literature. An identical formulation was used by Cellmer et al. (2019) for land price evaluation in the city of Olsztyn in northeastern Poland. Baltagi et al. (2014) use a more general formulation of the same type with spatio-temporal weights to study the spatio-temporal pattern of house prices in England.

Why is it that the sampling inconsistencies in (12.3), which are so repugnant to many authors, are shrugged off so casually by others? It is not as if these matters have gone unnoticed. Dong et al. remark that ... *changing the value of the covariates at one location ... will affect not only its own outcomes, but also the outcomes at other locations.*... They appear to regard this property as nothing out of the ordinary—neither distasteful nor inappropriate. However, the later version used by Fingleton et al. (2018) uses the residual $Y - X\beta$ in place of Y in (12.2). This amendment has the effect of removing the inconsistency from the mean, which suggests a disapproval of (12.3). However, the inconsistency remains in the covariances. The oddities of simultaneous autoregressive covariance matrices were pointed out by Sen and Bera (2014). Those oddities are real enough, but they are not to be confused with sampling inconsistencies.

One counter-argument to consistent sampling proceeds as follows. The formulation (12.2) is not a sampling model like (12.1). It is not meant to be applied to arbitrary samples, but only to the entire sample, i.e., to the population. In other words, the population is finite, and the joint distribution is Gaussian with joint moments as specified in (12.3) and (12.4). After all, there is only one Beijing. If you want the joint distribution for any subset of units, you need only extract the relevant means and covariances.

This interpretation is mathematically honest and logically unassailable. The problem lies in the finiteness of the population. In effect, Beijing is declared to be *sui generis*—a population of one. Admittedly, the Beijing real-estate valuation space, \mathbb{R}^{111^7} , leaves plenty of room for computation. Plenty of room for an accountant, that is, but none for a probabilist or statistician.

It is natural to extend the set of units geographically to include parcels beyond the city proper, but that is not in the spirit of the authors' formulation. It is also possible to extend the domain to parcel-time pairs. The distribution (12.3), (12.4), can then be interpreted as a real-valued process in one of two ways: either Y_{it} is constant in time, or it has independent and identically distributed values for $t \neq t'$. Neither of these interpretations is appealing for real-estate valuations. But the second provides a mathematical framework within which parameters can be understood, and predictions can be interpreted.

12.2 Adequacy for the Application

Every statistical model begins with a set of observational units on which are defined covariates and relationships of various sorts. These mathematical objects must be matched carefully and appropriately with objects and relationships on the workbench or in the field. The correspondence need not be one-to-one. All parts of this activity are invariably tentative, and always require reconsideration in the light of new information. Careful compromise demands a good knowledge of what is important in the mathematics and what is important in the process under study. It is usually necessary to entertain a range of stochastic models that exhibit progressively more complicated effects.

Criticism of linear Gaussian models or generalized linear models or other stock families is a popular pastime. Glib sweeping slogans such as ‘All models are wrong!’, or ‘They don’t work!’, or ‘Data are not normal!’, or ‘Relationships are not linear!’, are certainly provocative. But they are also unhelpful and irritating—particularly so in situations where the criticism is justified. Despite the rhetoric, none of these remarks is meant as a criticism of the model *per se* as a mathematical object. For example, few would venture to claim that Brownian motion is wrong on the grounds that atoms have mass, moving atoms have momentum, and momentum implies differentiability of paths. Such a statement would expose the obvious fact that the error lies not in the behaviour of atoms nor in Brownian motion, but in the nexus that associates one with the other.

To be fair, one oft-cited slogan has commented as follows: *Since all models are wrong, the scientist must be alert to what is importantly wrong. It is inappropriate to be concerned about mice when there are tigers abroad* (Box, 1976). Apart from the lead-in phrase, Box’s sentiment is very much in line with themes in this book. In that sense, I agree with him, and I agree enthusiastically. But, of his menagerie, we must be clear that tigers are to be found neither in the model nor in the specific components of the application. Instead, both mice and tigers are interstitial species occupying exclusively the gaps in the nexus between one and the other. As a result, we cannot learn to recognize either mice or tigers by studying exclusively one domain or the other. We need to understand both.

For reasons outlined above, it is helpful to defuse the rhetoric and focus the discussion by decomposing the model and its application into four distinct aspects as follows:

- Probabilistic matters:
 - (i) observational units as the domain for a process;
 - (ii) properties such as sampling consistency, exchangeability, independence, stationarity; short, medium and long-range dependence;
 - (iii) asymptotic properties of sample means, variances,...

- Statistical matters:
 - (i) protocol, baseline, randomization,...
 - (ii) parameterization: compatibility with linear transformation,...
 - (iii) accommodation of effects;
 - (iv) compatibility with randomization;
- Match with needs of the application:
 - (i) identification of observational and experimental units;
 - (ii) compatibility with existing physical/genetic/biological theories;
 - (iii) need for response transformation or covariate transformation;
 - (iv) are baseline relationships adequately accommodated?
- Procedural aspects:
 - (i) computation for parameter estimation;
 - (ii) computation for prediction, Kriging;
 - (iii) algorithms and software.

Now that the target of criticism is more clearly identified, we can all agree to put more effort into finding a model that matches well with the demands of the application. For example, much of the effort and discussion in Examples 1–10 is directed at the critical junction, and hopefully in a constructive way. Mice are not exactly welcome, but tigers must be squeezed out.

In this scheme, procedural and computational considerations play an important, but entirely subsidiary, role. The expectation is that whatever computations are needed for estimation or prediction can be done in reasonable time. If it turns out that the required computations are more formidable than anticipated, we may need to make further compromises. For most of examples 1–10, such compromises are not needed. Thus, strategies for computational approximation do not feature in this book.

12.3 Likelihood Principle

The topic of this section is a compelling fundamental principle of parametric inference. It is debated with vigorous intensity by theoretical statisticians; it is beloved by Bayesians because it supports the cause; and it is derided or ignored by most practitioners. Why is it that this unifying principle produces such a divergent range of reactions?

The development of the LP stems from the notions of likelihood and sufficiency, which made their first appearance in the fundamental paper by Fisher (1925). Versions of the likelihood principle were discussed in a series of papers by George Barnard beginning around 1949 and culminating in Barnard et al. (1962). Credit for the formulation of the modern version goes to Birnbaum (1962), who showed

that the likelihood principle is a consequence of sufficiency and conditionality. For a thorough discussion of its history and implications, see the book by Berger and Wolpert (1988).

The statement of the likelihood principle is simple enough.

All other things being equal, two experiments yielding the same likelihood function must also lead to identical inferences concerning the parameter.

To make sense of this statement, and to understand the differences in its interpretation, we must have a clear understanding of what the likelihood function is. In these notes, $P_\theta(\cdot)$ denotes the process or probability distribution associated with the point θ in the parameter space, and $dP_\theta(\mathbf{y})$ is the density with respect to Lebesgue measure or some other fixed reference measure. Given this structure, the likelihood associates with each point \mathbf{y} in the observation space, a non-negative function $L_y: \Theta \rightarrow \mathbb{R}$ on the parameter space.

Concretely, $L_y(\theta)$ is the Radon-Nikodym derivative $dP_\theta(\mathbf{y})/d\mu(\mathbf{y})$ with respect to a dominating measure μ on the observation space. Ordinarily, the Radon-Nikodym derivative is regarded as a function of \mathbf{y} for fixed θ ; the likelihood is usually regarded as a function of θ for fixed \mathbf{y} . Since the base measure is arbitrary, the product $C_y L_y(\theta)$ is equivalent to $L_y(\theta)$. Here $C_y > 0$ is any positive function of \mathbf{y} acting as a constant multiplicative factor on Θ . It has to be admitted that zero and infinity are legitimate values for the likelihood, but that complication is ignored here.

A statistic is a function from the observation space into another ‘reduced space’. If there exists a statistic such that $T(\mathbf{y}) = T(\mathbf{y}')$ implies $L_y = L_{y'}$, we say that T is sufficient for θ , or simply that T is sufficient. The likelihood principle implies that such pairs of points \mathbf{y}, \mathbf{y}' must lead to identical inferences about the parameter.

It should be clear on general grounds that a question such as ‘are the data compatible with any distribution in the set $\{P_\theta : \theta \in \Theta\}$?’ is not addressed by the likelihood principle. Two sample-space points giving rise to the same likelihood function need not be judged equally compatible with the model. Likewise, an inferential target such as the event $Y_{n+1} \in A$ is not covered by the likelihood principle. For such a target, two points such that $L_y = L_{y'}$ may yield different predictions

$$P(Y_{n+1} \in A | Y[n] = \mathbf{y}) \neq P(Y_{n+1} \in A | Y[n] = \mathbf{y}')$$

without violating the LP. Neither of these tasks lies within the realm of parametric inference.

A predictive target event such as $\bar{Y}_\infty \in A$ is covered by LP if and only if the event has probability either zero or one for each θ . In that case, the sample-space event $\bar{Y}_\infty \in A$ can be identified with the parameter subset

$$A' = \{\theta \in \Theta : P_\theta(\bar{Y}_\infty \in A) = 1\},$$

so that $\theta \in A'$ and $\bar{Y}_\infty \in A$ are equivalent.

Some of the controversies about the application of the principle revolve around the initial clause. If the circumstances surrounding the two experiments were sufficiently different, the prior distributions might not be the same, and the posteriors would certainly be different. That much is accepted. However, if the two experiments were investigating the same phenomenon in different ways, there is only one prior (per statistician), and the conclusions must be the same. Generally speaking, Bayesian inferences obey the likelihood principle; procedures based on sample-space computations such as p -values and confidence intervals, do not.

First Illustration

Let $n \geq 1$ be a positive integer. A partition of the set $[n] = \{1, \dots, n\}$ is a set of disjoint non-empty subsets B whose union is $[n]$. The Ewens sampling formula is a family of probability distributions on set partitions

$$p_\theta(B) = \frac{\theta^{\#B} \prod_{b \in B} \Gamma(\#b)}{\theta^{\uparrow n}}, \quad (12.5)$$

where $\#B$ is the number of blocks, $\#b$ is the block size, $\theta > 0$ is the parameter, and $\theta^{\uparrow n} = (\theta(\theta + 1) \cdots (\theta + n - 1))$ is the ascending factorial. The Radon-Nikodym derivative with respect to p_1 is

$$\frac{p_\theta(B)}{p_1(B)} = \frac{\theta^{\#B} n!}{\theta^{\uparrow n}},$$

which is also the likelihood function. Although the Ewens density depends on both the number of blocks and on their sizes, the number of blocks is a sufficient statistic. In other words, the likelihood function ignores block sizes. Any inferential statement about the parameter that depends either on the block sizes or on the size of the block that contains element 1 is a violation of the likelihood principle.

Given that the observed configuration has 20 blocks, sufficiency implies that the Ewens distribution on block sizes is fixed and independent of the parameter: see Exercise 12.6. An observation consisting of 20 blocks of approximately equal size is certainly possible but it may be judged so unlikely as to cast doubt on the entire family. Questions of model adequacy and goodness-of-fit question are traditionally formulated as significance tests; such matters are not concerned with the parameter space and are not covered by the LP.

Second Illustration

The principle can be illustrated by two observations on a Bernoulli sequence $Y_i \sim \text{Ber}(\theta)$. For a sample of size $n = 20$, the density function and the likelihood function are

$$f(y; \theta) = \theta^s (1 - \theta)^{20-s},$$

where the number of successes $s = y$ is the sufficient statistic. Suppose that the two sequences are

$$y^{(1)} = (0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1)$$

$$y^{(2)} = (0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1).$$

Both sequences have $s = 10$, so the likelihood functions are equal. According to the weak version of the likelihood principle, both observations must lead to the same conclusion about θ .

From the present viewpoint, it is immaterial whether we adopt a Bayesian-type beta-binomial model or we attempt to construct a confidence interval for θ .

Third Illustration

The defining condition (11.4) for a randomization protocol means that the joint distribution of the treatment vector and response vector is

$$P_\theta(T; \mathbf{x}) \times P_\theta(Y | \mathbf{x}, T) = P(T; \mathbf{x}) \times P_\theta(Y | \mathbf{x}, T).$$

Thus, the randomization protocol does not feature in the likelihood. Or, to put it more correctly, two experiments using different assignment protocols P, P' , giving rise to the same treatment assignment and the same response must lead to identical inferences about the parameter.

12.4 Attitudes

In his assessment of the role of applied mathematics, Courant (1965) comments as follows:

Applied mathematics is not a definable scientific field but a human attitude. The attitude of the applied scientist is directed towards finding clear cut answers that can stand the test of empirical observation. To obtain the answers to theoretically often insuperably difficult problems, he must be willing to make compromises regarding rigorous mathematical completeness; he must supplement theoretical reasoning by numerical work, plausibility considerations and so on.

Principles are important both in statistical theory and in its application. While one might make a plausible argument against the likelihood principle, it is difficult to make a comparable argument against mathematical consistency. For that reason, this book puts stochastic models and mathematical consistency at the top of the hierarchy. Without consistency, there can be no guarantees of any sort, so consistency is the *sine qua non* of science in general.

It is difficult to give a complete definition that covers every aspect of consistency, but it is usually easy to spot or to confirm inconsistent formulations once the flaw is

pointed out. Kolmogorov consistency, or sampling consistency, is one facet, but it is not the only one. See Exercises 12.4–12.5 for another sort of inconsistency.

In second place, I have listed ‘adequacy of the model for the application’ as a matter of principle. Perhaps this would be better described as a goal than a principle. Certainly one must be willing to tolerate deviations whose consequences are known to be minor. To paraphrase Box (1976), it would be inappropriate to take a stand on principle solely for the avoidance of a mouse.

In third place, I have listed the likelihood principle because it is entirely dependent on, and subsidiary to, the first two. Within the realm of parametric inference, the practice of learning by updating one’s views by the application of Bayes’s theorem is self-contained and mathematically reassuring. In that sense, the likelihood principle is unassailable. Nevertheless, in my view, its implications for applied work are frequently exaggerated.

Scope for Skepticism

Suppose that $\Theta = \{0\}$, i.e., that the parameter space contains a single point, and the model consists of a single distribution P_0 . The likelihood principle is loud, clear and vacuous: there is nothing to be said about the parameter! In that case, one wonders, is there nothing to be said at all?

Despite its extreme simplicity, there are many instances in scientific work where this formulation is neither excessively naive nor excessively simplistic. Two of the *Drosophila* experiments (Sects. 3.1 and 3.5) are designed in such a way that a Bernoulli-1/2 model or a Bernoulli(\bar{p}) model is not unreasonable, if only as a place to start. Certain designs to test extra-sensory perception are set up so that the uniform distribution on permutations of a known set is natural starting point. Tests for data fabrication sometimes use uniformity of trailing decimal digits as the reference (provided that uniformity is confirmed in non-fabricated values). Most studies of Kerrich’s famous war-time coin-tossing sequence are based on the obvious Bernoulli-1/2 model.

Can anything be learned about the process from a one-point statistical model? The likelihood function and Bayes’s theorem tell us only that $\theta = 0$ with probability one. Whether or not the model is adequate, the likelihood principle is unhelpful. On the other hand, the Bernoulli model has implications that can be checked in a variety of ways, graphically or otherwise. It is a long-standing and recommended practice to examine the data in more than one way, which Kerrich does extensively in his little book (Kerrich, 1946). From the evidence thus presented, we either gain or lose confidence in the adequacy of the formulation. Following such an analysis, Kerrich (1961) remarks, *I claim that there is nothing to suggest that we are not dealing with a perfectly ordinary coin behaving in quite a usual manner*. By contrast, Fig. 3.1 leads unexpectedly to serious doubt about the Bernoulli model, and there is plenty to suggest that mating events do not behave as independent Bernoulli trials. Likewise, for whatever reason, Fig. 3.2 leads to serious doubts about independence of activities in distinct wells.

It seems to me not only that there is value in the activity just described but also that an applied statistician would be derelict in his duty if he did not engage in it.

Checking of assumptions is an essential activity, but it is also an anarchic activity that is not easily circumscribed within the likelihood principle. Figure 3.1 is a signal that leads ultimately to a fundamental reformulation with an entirely different notion of what constitutes an observational unit. The need for such a revolutionary reassessment of fundamentals is something that would be close to impossible to deduce from any likelihood function or by the application of Bayes's theorem.

Relative Importance

The second attitude addresses the relative importance of parametric inference as a component of the professional activity of the applied statistician. While one may debate the precise fractions, I venture to say that questions of parameter estimation and parametric inference represent about 20% of the effort for an applied statistician. Yet this activity accounts for possibly 70–80% of the attention in the methodological literature. Computation and parameter estimation are important, but they are not the most important activities for the applied statistician.

Lessons

The premiss of the likelihood principle is that the statistician buys into the model exactly as stated, with no probabilistic reserve in the form of an opt-out clause to cover buyer's remorse. The lesson of experience is simply to avoid being sandbagged. The likelihood principle is not rejected, but a cautious applied statistician invariably adopts the stated model provisionally, with adequate reserves to cover mistakes, misunderstandings or unanticipated events. To do otherwise would be a serious error of professional judgement.

In effect, a consulting statistician using the Bernoulli model proceeds as follows. With probability 0.65 the sequence is Bernoulli with constant parameter θ ; with probability 0.10 the sequence is Bernoulli with non-constant parameter; with probability 0.10, the sequence has some temporal dependence, possibly Markov; with probability 0.10, the design has some other feature that might lead in a different direction. The weights shown here may be varied to match the incidental information relevant to the context, but their sum is strictly less than one. After observing either sequence $y^{(1)}$ or $y^{(2)}$ the first weight component is drastically reduced. The *Drosophila* mating experiment in Chap. 3 is an instance of this sort.

12.5 Exercises

12.1 According to the discussion of Dong et al. (2015) in Sect. 12.1, the real-estate market in Beijing is divided up into 1117 land parcels, which are partitioned into 111 districts. These administrative regions remain fixed over the time period of interest. Each land parcel has a geographic position, an area, and a population density. To each district there corresponds a subset of one or more neighbouring districts. Every boundary parcel has at least one ‘foreign’ neighbour, a parcel outside of Beijing;

likewise for districts. Two Beijing districts are isolated and have no neighbours within the city.

Discuss the nature of these variables in the setting of a study of Chinese quarterly real-estate market valuations. What are the observational units? How many are there? Which variables would you classify as covariates? Which variables would you classify as relationships? At least one of these relationships is an equivalence relation. Which one? Is a district a neighbour of itself? At least one relationship is Boolean but not transitive. Which one? At least two relationships are metric. Which two? What other types of relationships might be relevant?

12.2 Each land parcel belongs to the first or inner ring, the second ring, the third ring, or beyond. To be clear, the rings are disjoint, so the phrase ‘second ring’ excludes the first. A district may straddle two or more rings. What sort of variable is *ring*?

12.3 For non-commercial sales, average sale price per square metre is recorded quarterly for each parcel. Discuss briefly how you might go about constructing a sampling-consistent Gaussian model that incorporates spatial and temporal correlations.

12.4 Consider a statistical model for a competition experiment in which each observational unit is an ordered pair (i, j) of distinct competitors (chess players). The state space consists of three outcomes *won*, *drawn*, *lost*, ordered from the viewpoint of player i , and coded for convenience as 1, 2, 3 so that $Y_{i,j} + Y_{j,i} = 4$. Consider the ordinal trinomial model

$$\text{logit } \text{pr}(Y_{i,j} \leq r) = \gamma_r + \alpha_i - \alpha_j,$$

depending on two threshold parameters $\gamma_1 \leq \gamma_2$ plus player strength parameters $\alpha_1, \dots, \alpha_n$.

In what sense is this model inconsistent? How do you modify it to resolve the inconsistency?

12.5 A different version of the preceding model uses the complementary log-log link function:

$$\log(-\log(1 - \text{pr}(Y_{i,j} \leq r))) = \gamma_r + \alpha_i - \alpha_j,$$

depending on two threshold parameters $\gamma_1 \leq \gamma_2$ plus player strength parameters $\alpha_1, \dots, \alpha_n$.

In what sense is this model inconsistent? How do you modify it to resolve the inconsistency?

12.6 For the Ewens distribution (12.5), show that the conditional distribution given $\#B = k$ is

$$p_\theta(B \mid \#B = k) = \frac{\prod_{b \in B} \Gamma(\#b)}{s_{n,k}}$$

where $s_{n,k}$ is Stirling's number of the first kind, i.e., the number of permutations $[n] \rightarrow [n]$ that have exactly k cycles.

12.7 Let $n = 12$, and let B have the Ewens distribution with parameter $\theta > 0$. Suppose B has six blocks. Which is more likely: (a) that B has all blocks of size two; (b) that B has five blocks of size one and one of size seven?

12.8 Suppose that Y_1, \dots, Y_n are independent and identically distributed with density

$$\frac{1}{2\pi}(1 + \psi \cos y)$$

on the interval $-\pi < y < \pi$. The parameter space is the interval $-1 \leq \psi \leq 1$.

Show that the log likelihood is concave. What does this imply about maximum-likelihood estimation?

12.9 Suppose that Y_1, \dots, Y_n are independent and identically distributed with density

$$\frac{1}{2\pi}(1 + \psi \cos y)(1 + \sin y/2)$$

on the interval $-\pi < y < \pi$.

Show that the vector statistic $R = \cos Y$ is sufficient and that $S = \sin Y$ is ancillary, i.e., that S is distributed independently of the parameter. Show that the conditional distribution of R given S is discrete, a Bernoulli multiple with independent components. Find the conditional likelihood, and compare it with the unconditional likelihood in Exercise 12.8.

12.10 Suppose that Y_1, \dots, Y_n are independent and identically distributed with density

$$\frac{1}{2\pi}(1 + \psi \cos y)(1 + \lambda \sin y)$$

on the interval $-\pi < y < \pi$. Show that the likelihood for (ψ, λ) factors

$$L(\psi, \lambda; y) = L_1(\psi; y) \times L_2(\lambda; y).$$

Show that the likelihood factorization is also a density factorization, i.e., for fixed ψ , that $L_1(\psi; y)$ is a probability density on $(-\pi, \pi)^n$, and likewise for L_2 . Does it follow that R and S are independent?

12.11 By definition, the randomization protocol is a known distribution on treatment assignments. In this context, ‘known’ means declared at baseline and $P_\theta(\mathbf{t}) = P_{\theta'}(\mathbf{t})$ for all pairs θ, θ' . Show that the likelihood function does not depend on the randomization protocol.

12.12 Let X_0, X_1 be given matrices of order 100×5 and 110×5 such that $X_0'X_0 = X_1'X_1 = F$, and let P_β be the Gaussian mixture model

$$P_\beta = \frac{1}{2}N_{100}(X_0\beta, I_{100}) + \frac{1}{2}N_{110}(X_1\beta, I_{110})$$

indexed by $\beta \in \mathbb{R}^5$. For any sample point, either $y \in \mathbb{R}^{100}$ or $y \in \mathbb{R}^{110}$, show that the likelihood ratio is

$$\frac{p_\beta(y)}{p_0(y)} = \exp(y'X\beta - \beta'F\beta/2).$$

Deduce that the vector $X'y$ is minimal sufficient, and that the sample size $n(y)$ is not a component of the sufficient statistic.

12.13 BIC is a sub-model selection procedure. For an observation $y \in \mathbb{R}^n$, BIC adds the penalty $-\frac{1}{2}d \log n$ to the maximized log likelihood for sub-models of dimension d . Deduce from the preceding example that this version of BIC cannot be a Bayes procedure.

Chapter 13

Initial Values



13.1 Randomization Protocols

Consider a randomized trial designed to compare two medications for hypertension. The plan calls for patients to be recruited sequentially over a six-month period. Baseline measurements recorded at presentation include date, age, sex and marital status plus blood pressure. Patients exhibiting elevated blood pressure are eligible for recruitment. This is a comparative randomized study with the goal of comparing the effectiveness of the two medications in alleviating hypertension.

Since the medications are designed to control hypertension, blood pressure at baseline is called the *initial value*. Eligible patients are randomized to one of the two treatment arms, and their blood pressure is monitored at roughly three-month intervals for one year. For purposes of discussion here, $Y_{0,i}$ is the initial value for patient i , age, sex and other baseline covariates are encoded in x_i , and $Y_{1,i}$ is the value at the one-year endpoint.

In the mathematical framework of the preceding chapter, all baseline variables are on an equal footing. Instinctively, however, it may appear that the initial value has a status that is different from other baseline variables. The goal of this section is to explore the question of whether the apparent difference is cosmetic or substantive. If the difference is substantive, what are the implications for design and analysis? Where in the mathematics does the distinction exhibit itself?

Every variable that is recorded at baseline is available for use as a covariate that modifies—or has the potential to modify—the distribution of the process being studied. That includes both the randomization scheme, which is a joint probability distribution $P(T = \mathbf{t})$, and the distribution of the ultimate response. Although the randomization distribution is fully specified by the protocol, the individual probabilities $P(T_i = 1)$ and the pairwise probabilities may depend on any and all baseline variables, relationships, and so on. Ordinarily, two individuals having the same covariate values have the same assignment distribution, but two individuals having different values may have different assignment probabilities. A complicated

randomization scheme of this sort may not be recommended for a clinical setting, but it is entirely legitimate on mathematical grounds if the scheme is fully specified in the protocol.

In many settings such as agricultural field trials, treatment assignment is deliberately balanced to take account of block factors and other baseline variables such as the geographical arrangement of plots in the field. It might well be the case that $T_i \sim T_j$ for every pair of plots, but it is rarely the case that $(T_i, T_j) \sim (T_{i'}, T_{j'})$ for every pair of distinct pairs. For example, the bivariate assignment probabilities for adjacent and non-adjacent pairs of plots are usually different. Thus, complicated randomization schemes taking account of baseline covariates and relationships are established practice in certain areas. Even though simpler randomization schemes are strongly favoured for clinical trials, the possibility of dependence of T on either \mathbf{x} or on Y_0 is left open in the discussion that follows.

This motivation for this section comes from a series of papers in the biostatistical literature, (Samuels, 1986; Liang & Zeger, 2000; Senn, 2006) in which there appears to be disagreement on the choice of statistical techniques that are appropriate for designs that focus on the change from baseline. These are sometimes called pre-post designs.

13.2 Four Gaussian Models

Four closely-related Gaussian models are described. For the most part, the formulations are entirely standard. Although the emphasis is on the treatment effect, other aspects of the joint distribution must be considered. In the third and fourth versions, it is explicitly assumed that T and Y_0 are independent. This assumption does not occur in versions I or II.

The reader should be warned that not all statements should be taken at face value. Some are debatable; others may be misleading, or inappropriate for clinical applications, or simply incorrect. See the subsequent discussion in Sect. 13.5.3.

Version I

In the simplest version of the problem, there are no baseline covariates other than the initial value. In the absence of covariates, the baseline values are exchangeable, here interpreted as independent Gaussian

$$Y_{0,i} \sim N(\mu_0, \sigma_0^2). \quad (13.1)$$

Subsequent components of the joint distribution are specified in temporal sequence. The treatment assignment distribution $p_n(T = \mathbf{t} | Y_0)$ is specified by protocol. The ultimate response Y_1 is a scalar measured at a subsequent time $t = 1$, the same time for all units. On the assumption that the treatment effect is additive with no

interference, we arrive at the standard formulation for the conditional distribution given (Y_0, T) :

$$Y_{1,i} \sim N(\mu_1 + \gamma Y_{0,i} + \tau T_i, \sigma_1^2). \quad (13.2)$$

Responses for distinct units are conditionally independent given Y_0, T . Although there are compelling reasons to expect $0 < \gamma < 1$, and perhaps $\sigma_1^2 < \sigma_0^2$, no constraints are imposed on the parameters in either (13.1) or (13.2).

Given two units with equal baseline values $Y_{0,i} = Y_{0,j}$, the expected difference in responses is

$$E(Y_{1,i} - Y_{1,j} | Y_0, T) = \tau(T_i - T_j).$$

This is the treatment effect, which is a constant independent of the initial value.

For two units whose initial values are not equal, the conditional expected response difference is a linear combination of the treatment effect plus the initial difference:

$$E(Y_{1,i} - Y_{1,j} | Y_0, T) = \tau(T_i - T_j) + \gamma(Y_{0,i} - Y_{0,j}).$$

The unconditional expected difference is

$$E(Y_{1,i} - Y_{1,j} | T) = \tau(T_i - T_j) + \gamma E(Y_{0,i} - Y_{0,j} | T),$$

which is, in general, *not* the same as the treatment effect. However, if treatment is assigned independently of initial values, exchangeability implies that the second term is zero.

The parameters in this model are the two variances plus four regression coefficients $\mu_0, \mu_1, \gamma, \tau$. Regardless of the specific model employed, the joint density has two Gaussian and one non-Gaussian factor:

$$p(y_0, T, y_1) = p_0(y_0) \times p_n(T | Y_0) \times p_1(y_1 | Y_0, T). \quad (13.3)$$

In the Gaussian model, the first factor depends on (μ_0, σ_0) ; the second factor is fully specified by protocol, i.e., constant on the parameter space; the third factor depends on $(\mu_1, \gamma, \tau, \sigma_1)$. Thus, provided that the parameters are variation independent, the density factorization is also a likelihood factorization. So far as maximum-likelihood estimation of the the treatment effect is concerned, the first two factors can be ignored, and usually are ignored. Whether T is independent of Y_0 or not, the treatment effect is estimated by ordinary least squares using (13.2) with the initial value as a covariate.

Version II

The second version admits baseline covariates \mathbf{x} whose effect on the response distribution is additive. In standard linear-model notation, μ_0 in (13.1) is replaced

by $x'_i \beta_0$ and μ_1 in (13.2) by $x'_i \beta_1$. Provided that the treatment effect is a constant independent of x , the conclusions are not appreciably different from those in version I. Whether $T \perp\!\!\!\perp Y_0$ or not, it is evident that the maximum-likelihood estimate of the treatment effect and its standard error are obtained by least squares based on the extended version of (13.2) using both \mathbf{x} and the initial value as covariates on an equal footing. It appears, therefore, that the distinction between the initial value and other baseline covariates is more cosmetic than substantive.

Version III

The third version is a minor variation on the first, but it sets the scene for the natural extension to longitudinal designs in version IV. For individuals such that $T_i = 0$, the pairs $(Y_{0,i}, Y_{1,i})$ are independent and bivariate normal

$$\begin{pmatrix} Y_{0,i} \\ Y_{1,i} \end{pmatrix} \sim N_2 \left(\begin{pmatrix} \mu_0 \\ \mu_1 \end{pmatrix}, \Sigma \right).$$

For individuals such that $T_i = 1$, the pairs are again independent and bivariate normal

$$\begin{pmatrix} Y_{0,i} \\ Y_{1,i} \end{pmatrix} \sim N_2 \left(\begin{pmatrix} \mu_0 \\ \mu_1 + \tau \end{pmatrix}, \Sigma \right).$$

This version incorporates explicitly the assumption of independence $Y_0 \perp\!\!\!\perp T$ in the form $Y_0 \sim N(\mu_0, \sigma_0^2 I_n)$ given T .

For this setting, it is natural to consider two versions of the bivariate model, one with general Σ and one with the stationarity restriction $\sigma_{00} = \sigma_{11} = \sigma^2$. Both are compatible with (13.2). However, if Σ is restricted by stationarity, the density factorization (13.3) is not a likelihood factorization because σ^2 occurs in the first and third factors. Moreover, since the initial-value coefficient is an autocorrelation, we have $-1 \leq \gamma \leq 1$ in (13.2), which is not guaranteed by least squares.

Version IV

In a longitudinal study of health, the response on each physical unit is measured at baseline and at multiple points thereafter as specified by the protocol. Apart from baseline, the observation times need not be the same for every individual. In the fourth version, the response is assumed to be a Gaussian process with covariance function $\delta_{i,j} K(t, t')$ for $t, t' \geq 0$. In other words, treatment assignment is independent of initial values, and the processes for distinct individuals $i \neq j$ are independent with the same conditional covariance function

$$\text{cov}(Y_i(t), Y_i(t') | T) = K(t, t'). \quad (13.4)$$

Apart from the covariance function, the model is determined by the conditional mean function

$$E(Y_i(t) \mid T) = \begin{cases} \mu_0(t) & (T_i = 0); \\ \mu_0(t) + \tau(t) & (T_i = 1). \end{cases} \quad (13.5)$$

In full generality, each treatment effect is a temporal function $t \mapsto \tau(t)$, presumably continuous, and subject to the randomization constraint $\tau(0) = 0$ as explained in Sect. 5.2.3. The set of treatment effects is some family of functions $\mathbb{R}^+ \rightarrow \mathbb{R}$, which is necessarily a vector space, closed under addition and scalar multiplication. For example, $\{t \mapsto \beta t/(1+t) : \beta \in \mathbb{R}\}$ is a one-dimensional vector space of bounded functions.

If it is needed, the conditional distribution of $Y_i(\cdot)$ given Y_0, T can be computed from (13.4) and (13.5), using standard formulae for the conditional mean and conditional covariance of a Gaussian process.

If other baseline covariates are present, their effect can be included in $\mu_0(t)$ if there is no covariate-treatment interaction. However, if the treatment effect for males is not the same as that for females, it is best to include one treatment effect for each sex. Clearly, version III is a special case of version IV.

13.2.1 Distribution and Likelihood

If $T \perp\!\!\!\perp Y_0$, each of the four versions is a variation on IV. The processes for distinct individuals are independent Gaussian processes with the same covariance function $K(t, t')$ for all individuals regardless of treatment status. Treatment can only have an effect post-baseline; in (13.5), the effect is additive on the mean trajectory given T , but it is not constant in time.

A treatment-assignment protocol $p_n(T \mid Y_0 = y)$ is called fixed if the distribution for each y is degenerate at $\mathbf{t}(y)$ and constant, i.e., $\mathbf{t}(y) = \mathbf{t}(0)$ for all y . Every fixed assignment is a random assignment, automatically independent of Y_0 , to which the remarks in the preceding paragraph apply. For each fixed assignment, the second factor in (13.3) is degenerate; the product of the first and third factors is the density of the Gaussian process whose mean is (13.5).

If T is not independent of Y_0 , the product of the first two factors in (13.3) is the joint density of (T, Y_0) , and the full product is the joint density of $(T, Y(\cdot))$. In general, the conditional distribution of Y_0 given T is not Gaussian, nor is the conditional distribution of $Y_i(t)$, so the conditional mean given T does not satisfy (13.5). However, the second factor in (13.3) is parameter-free, and plays no part in likelihood calculations. According to the likelihood principle, all inferential statements *concerning the parameters alone* are to be made as if the treatment assignment were fixed and independent of the initial values. In other words, if T is not independent of Y , the joint distribution given T is not Gaussian. Nonetheless,

the likelihood function is the same as if T were fixed, so the Gaussian likelihood computation is correct in that limited sense.

The likelihood principle is concerned solely with parametric inference. It has little to say about inferences that are external to the parameter space. For example, the task of predicting future values $Y_i(s)$ for patient i is a state-space task, not covered by the likelihood principle.

13.2.2 Numerical Comparison of Estimates

The following numerical example illustrates the magnitude of the differences in maximum-likelihood estimates of τ in versions I and III for a sample of 12 patients. Whether treatment assignment is independent of Y_0 or not, the likelihood is the product of the first and third factors in (13.3), which is a Gaussian likelihood.

| | | | | | | | | | | | | |
|-------|------|------|------|------|------|------|------|------|------|------|------|------|
| y_0 | 4.33 | 5.13 | 5.05 | 4.62 | 3.90 | 4.08 | 4.99 | 4.39 | 5.58 | 4.99 | 4.16 | 5.22 |
| y_1 | 5.13 | 3.86 | 4.81 | 4.88 | 4.57 | 3.13 | 6.97 | 5.86 | 5.62 | 5.67 | 4.53 | 5.56 |
| t | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 |

Maximum likelihood for the three Gaussian models described above produces the following estimates for the treatment effect. In each case, the standard REML procedure was used for the estimation of variances and covariances, followed by weighted least squares for regression coefficients.

| | $\hat{\tau}$ | s.e.($\hat{\tau}$) | s^2 |
|-------------------------------|--------------|----------------------|--------|
| (13.2): OLS | 1.1596 | 0.4843 | 0.6097 |
| III: $\sigma_0 = \sigma_1$ | 1.2147 | 0.3660 | 0.4019 |
| III: $\sigma_0 \neq \sigma_1$ | 1.1596 | 0.4277 | 0.5487 |
| Differences | 0.9350 | 0.4644 | 0.6469 |

The final column is the estimate of the conditional variance $\text{var}(Y_1 \mid Y_0)$, either computed directly from the residual mean square in (13.2), or computed indirectly as a function of the fitted 2×2 covariance matrix $\hat{\Sigma}$ in version III. The fitted matrices for the two variations of III are

$$\begin{pmatrix} 0.427 & 0.104 \\ 0.104 & 0.427 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} 0.280 & 0.110 \\ 0.110 & 0.592 \end{pmatrix}.$$

The Y_1 -values are more variable than baseline values, but not significantly so. The expression `fit2` displayed below shows why it is sometimes necessary to allow variance components to be negative.

For computational purposes, the data were coded in the response vector $Y = (Y_0, Y_1)$, the patient ID factor with 12 levels, and the treatment factor $T = (T_0, T_1)$,

which has one baseline level and two post-baseline levels. The ordinary least squares code uses the sub-vectors Y_1, Y_0 and the two-level factor T_1 . In all cases, the covariance model includes the identity I_{12} or I_{24} by default. The third part uses the additional matrix I_0 of order 24, which is the identity restricted to baseline values only.

```
fit0 <- regress(Y1~Y0+T1)
fit1 <- regress(Y~T, ~patient_id)
fit2 <- regress(Y~T ~patient_id+I0)
fit3a <- regress(Y~T+patient_id); fit3b <- regress(diff~T1)
```

The code for the fourth version differs from the second only in one respect: patient ID is used as a classification factor in the mean rather than as a block factor in the covariance. This is equivalent to assuming that the between-patient variance is infinitely large, so the same numerical value is obtained by working with the individual differences $Y_1 - Y_0$ in the code `lm(diff~T1)`. As it happens, the between-patient variance is about one third the residual variance, the regression coefficient of Y_0 in the ordinary least-squares fit is only 0.393, so differencing is not especially effective in this instance. Statistically speaking, the fourth method is strictly inferior to the first three.

13.2.3 Initial Values Versus Covariates

We have seen that no fundamental distinction can be drawn between initial values and other baseline covariates. Nevertheless, the distinction is relevant and important if only as a strategy for model construction.

If the initial value is regarded as non-random, and observations are to be made at arbitrary post-baseline points, it is necessary to specify the joint distribution of $Y(t_1), \dots, Y(t_k)$ given $Y(0)$ for arbitrary $k \geq 1$ and arbitrary configurations t_1, \dots, t_k . In other words, we need to associate with each initial value $y_0 = Y(t_0)$ a stochastic process that is devoid of symmetries such as stationarity, in such a way that the one-dimensional distributional specification for each time $t \geq 0$ is consistent with the two-dimensional specifications for times t, t' , and so on. Direct construction of conditional distributions is a very difficult exercise, and the dependence on the initial value only compounds the difficulty. Generally speaking, it is much easier and more natural to begin with a single process, which is a consistent specification of the joint distribution of $Y(t_0), \dots, Y(t_k)$ for arbitrary temporal configurations. Depending on the setting, this process might well be stationary. If it is needed—and usually it is not needed—the process itself defines the conditional distribution given $Y(t_0)$. These derived conditional distributions are automatically self-consistent. Each one is non-stationary, fixed at the temporal origin.

In conclusion, the distinction between covariates and initial values may not be fundamental, but it is strategically important. It is a strategic mistake to insist that the initial value cannot be regarded as random.

13.2.4 Initial Values in an Observational Study

All of the preceding remarks concerning initial values are made in the context of a randomized controlled experiment. However, initial values also occur naturally in every longitudinal study, even if there are only two observation times. Suppose, therefore, that $Y_i(0)$ is the initial value, x_i is a baseline classification factor such as age or sex, and that $Y_i(1)$ is the subsequent or terminal value. All patients are closely monitored, and the recommendations may be individually tailored, but the program has no randomized assignment or declaration of subsets for comparison other than subsets determined by levels of x . What sorts of analyses are possible, and how should they be conducted?

In many clinical situations, the goal is to improve symptoms, for example by alleviating pain, reducing weight or reducing blood pressure, so it is natural to focus on the difference, $Z_i = Y_i(1) - Y_i(0)$, and to examine its association with the classification factor x . Since there is no treatment and no control level for comparison, we can only look at the reduction and ask whether the program of medication appears to be effective at alleviating symptoms. In the absence of a control group, such a question can be addressed only under an assumption of stationarity, namely that any systematic difference is more naturally associated with the program than with the passage of time. Therein lies the essential weakness of an observational study.

Stationarity is not an assumption be taken lightly, particularly in studies of chronic diseases such as Lyme disease, where the persistence of symptoms is known only anecdotally and primarily from sources exhibiting the greatest pain and longest persistence. It is well to remember that volunteers are primarily or exclusively those exhibiting extreme symptoms at baseline. Even if the process is stationary and the program is entirely neutral in its effect, the regression phenomenon guarantees that average symptoms after one unit of time will be less extreme than those at baseline. The temptation to attribute such a reduction to the effectiveness of the program is undoubtedly strong for all participants, but it is also potentially misleading. For an insider's view, see Douthat (2021) and the sympathetic but more balanced review by Austin (NYT, Oct. 2021).

Given stationarity, we can investigate whether the program looks promising by examining the magnitude of the mean symptom reduction and asking whether this is appreciably better than what would be expected under a neutral program. In other words, some estimate of a neutral effect is needed for a definitive assessment. In the absence of data from a neutral group, we can ask whether the program is equally effective or ineffective for patients of all ages and both sexes. On the assumption of independence and exchangeability for distinct patients, the simplest Gaussian model takes the form

$$\begin{pmatrix} Y_{0,i} \\ Y_{1,i} \end{pmatrix} \sim N_2 \left(\begin{pmatrix} \mu_0(x_i) = \beta_{00} + \beta_{01}x_i \\ \mu_1(x_i) = \beta_{10} + \beta_{11}x_i \end{pmatrix}, \Sigma \right). \quad (13.6)$$

As always, exchangeability of responses is assumed only for patients having the same baseline covariate value, so the baseline mean is a function of x . This is to be contrasted with the randomized setting in (13.5), where treatment is a post-baseline variable, and treatment assignment occurs only in the conditional mean of the post-baseline responses.

Ordinarily, a null or neutral treatment is associated with a zero effect in treatment comparisons. In the observational setting with a binary-valued covariate such as sex, the mean difference between the two groups is β_{01} at baseline and β_{11} at $t = 1$. There is a non-zero mean difference at both time points. If we work directly with sample differences $Z_i = Y_i(1) - Y_i(0)$, the sex effect is the mean difference

$$E(Z_i - Z_j) = \beta_{11} - \beta_{01}$$

for any pair of units units $i \neq j$ such that $x_i = 1$ and $x_j = 0$. One way to estimate the sex effect is to code Y as a matrix of order $n \times 2$ and to fit (13.6) as a bivariate regression model, $Y \sim N(X\beta, \Sigma \otimes I_n)$. Maximum likelihood gives $\hat{\beta} = (X'X)^{-1}X'Y$ as a 2×2 matrix, and Σ is the 2×2 matrix of residual mean squares and products. We can then report the difference $\hat{\beta}_{11} - \hat{\beta}_{01}$, together with its standard error. Numerically, this is exactly equivalent to a simple linear regression of the differences $Z_i = Y_i(1) - Y_i(0)$ on x_i .

Given the target parameter $\beta_{11} - \beta_{01}$, it is crucial that the baseline value Y_0 *not* be included as a covariate in either regression because

$$E(Z_i | Y_0) = \mu_1(x) - \gamma \mu_0(x) + (\gamma - 1)Y_{0,i},$$

where $\gamma = \Sigma_{10}/\Sigma_{00}$. Thus, $E(Z_i - Z_j | Y_0) \neq \beta_{11} - \beta_{01}$. Apart from notation and interpretation, $\hat{\beta}_{11} - \hat{\beta}_{01}$ is what is reported in `fit3b` in Sect. 13.2.2.

The topic of this section—the conflation of a treatment factor in a randomized trial with a classification factor in an observational study—is the core of Lord’s paradox: Lord (1967). See also Bock (1975, Sect. 7.3.1) for a more detailed discussion along the lines of Exercise 13.2. As Senn has pointed out, Lord’s paradox is a feast of red herrings, the first of which is an observational study presented and analyzed as if it were a randomized experiment.

13.3 Exercises

13.1 Let Y_1, \dots, Y_n be independent and identically distributed random variables whose distribution on \mathbb{R} is atom-free, and let $r_i = \text{rank}(Y_i)/(n+1)$ be the normalized rank vector. The treatment assignment vector has conditionally independent Bernoulli components with parameter $T_i \sim \text{Ber}(r_i)$ given Y .

Deduce the following:

1. the components of T are exchangeable and $T_i \sim \text{Ber}(1/2)$;

2. for each pair $i \neq j$ the covariance is $\text{cov}(T_i, T_j) = -1/(12(n+1))$;
3. every symmetric function of (T_1, \dots, T_n) is independent of the vector Y ;
4. the sample mean is symmetrically distributed with mean one half and variance

$$\text{var}(\bar{T}_n) = \frac{n+2}{6n(n+1)}$$

as opposed to $1/(4n)$ for the independent and identically distributed Bernoulli setting.

Check these calculations for $n = 1$ and $n = 2$.

- 13.2** Let Y_1, \dots, Y_n be independent $N(\beta, \sigma^2)$ random variables with parameter (β, σ) taking values in $\mathbb{R} \times \mathbb{R}^+$, and let the components of T be conditionally independent Bernoulli variables $T_i \sim \text{Ber}(r_i)$ as described in the preceding exercise. Using results from the preceding exercise, deduce as an approximation for large n , that

$$\begin{aligned} E(Y_i | T) &= \beta + \sigma \gamma_n x_i \\ \text{cov}(Y_i, Y_j | T) &= \sigma^2(1 - \gamma_n^2)\delta_{ij} + \sigma^2 \gamma_n^2/n, \end{aligned}$$

where $x_i = 2(T_i - \bar{T}_n)$ is the normalized treatment vector, and γ_n is a sequence in $(0, 1)$ whose limit is

$$\lim_{n \rightarrow \infty} \gamma_n = 2 \int_{-\infty}^{\infty} x \Phi(x) \phi(x) dx = 0.5642.$$

- 13.3** Use the Gaussian model with second moments given in the previous exercise to compute a pseudo-log likelihood $l_0(\beta, \sigma)$ for the parameter (β, σ) . Show that the pseudo log-likelihood differs from the correct log likelihood

$$l(\beta, \sigma) = -n \log \sigma - \frac{1}{2} \sum (y - \beta)^2 / \sigma^2$$

by terms that are relatively small for large n , so that $\hat{\beta}_0 = \hat{\beta}$ and $\hat{\sigma}_0 - \hat{\sigma} = O_p(n^{-1})$. What precisely does ‘relatively small for large n ’ imply about the magnitude of the difference $l_0(\beta, \sigma) - l(\beta, \sigma)$?

- 13.4** For the pseudo log likelihood in the preceding exercise, show that the Fisher information matrix is diagonal and that it coincides with the Fisher information from the correct log likelihood.

13.5 Suppose that terminal values are conditionally independent given (Y_0, T) with conditional distribution

$$Y_{1,i} \sim N(\beta_0 + \beta_1 Y_{0,i} + \tau T_i, \sigma_1^2),$$

and that τ is estimated by ordinary linear regression of Y_1 on (Y_0, T) . Compare the asymptotic variance of $\hat{\tau}$ for three randomization schemes: (i) random permutation with balanced assignment, (ii) independent symmetric Bernoulli, and (iii) the scheme in Exercise 13.1. Hence show that the asymptotic efficiency of the exchangeable randomization scheme relative to either of the others is approximately 68%.

Chapter 14

Probability Distributions



14.1 Exchangeable Processes

14.1.1 Unconditional Exchangeability

Recall that a process with state space \mathcal{S} associates with each sample $U = (u_1, \dots, u_n)$ consisting of finitely many distinct observational units taken in a specified order, a probability distribution P_U on the observation space \mathcal{S}^U . Thus $P_U(A)$ is the probability of the event

$$(Y_{u_1}, \dots, Y_{u_n}) \in A.$$

The process is said to be *unconditionally exchangeable* if two samples consisting of the same number of units have the same joint distribution. In other words, the process is exchangeable if $\#U = \#U' = n$ implies $P_U(A) = P_{U'}(A)$ for every event $A \subset \mathcal{S}^n$. In particular, two samples consisting of the same distinct units taken in different orders have the same distribution. In that case, the n -dimensional joint distribution is usually denoted by P_n .

Unconditional exchangeability is a very demanding property that is seldom satisfied in scientific work, where experiments are almost invariably comparative. The goal is usually to study *differences* between one distribution P_u and another $P_{u'}$ that are related to covariate effects or treatment effects for pairs such that $x(u) \neq x(u')$. Nonetheless, a version of exchangeability is needed in order to make progress in situations where inhomogeneities associated with baseline covariates are anticipated.

14.1.2 Regression Processes

Let $u \mapsto x_u$ be a covariate defined as a function $\mathcal{U} \rightarrow \mathcal{X}$ for every unit in the population. It is assumed implicitly that the only baseline relations are the identity function $\delta_{u,u'}$, which tells us whether or not two units are the same, and the one-block constant function $J_{u,u'} = 1$ for all pairs.

To each sample U there corresponds a covariate configuration $\mathbf{x}[U]$, which is usually encoded as a model matrix X whose rows are indexed by observational units. The manner in which dose levels or classification factors are encoded is immaterial. The process is said to be regression-exchangeable if two samples of distinct units having the same covariate configuration automatically have the same joint distribution. In other words, $\mathbf{x}[U] = \mathbf{x}[U']$ implies $P_U = P_{U'}$.

This form of exchangeability is usually taken for granted in applied work, so much so that it is rarely judged to be worth even a brief comment. For example, all generalized linear models are regression-exchangeable in this sense. However, some spatial autoregressive formulations such as those discussed in Sect. 12.1, are not regression-exchangeable.

Planar white noise and Poisson processes are less obvious examples. In both cases, each unit is a planar subset and the covariate $x_u = \Lambda(u)$ is planar Lebesgue measure. For any collection of disjoint subsets, the white-noise values $Y(u)$ are independent zero-mean Gaussian variables with variance x_u . The Poisson-process values are independent Poisson variables with mean x_u . Disjointness of subsets is not part of either definition; it is needed here only to comply with the assumption that there are no relationships among the units other than the identity.

14.1.3 Block Exchangeability

Recall that a block factor is an equivalence relation $B : \mathcal{U}^2 \rightarrow \{0, 1\}$ on the units, which partitions the population into disjoint non-empty subsets called blocks. Each population block may be finite or infinite. The restriction of B to a finite sample U consisting of n distinct units is a symmetric binary matrix of order n . This matrix is an equivalence relation which partitions the sample into disjoint blocks. If the units are arranged in suitable order, $B[U]$ is block-diagonal.

The process is said to be block-exchangeable if two samples having the same block structure also have the same response distribution, i.e., $B[U] = B[U']$ implies $P_U = P_{U'}$. For samples of size one, $B[U] = B[U']$ is automatic, so all one-dimensional distributions are equal. For samples of size two with $u_1 \neq u_2$, either $B(u_1, u_2) = 0$ or $B(u_1, u_2) = 1$, so there are two distinct two-dimensional distributions.

A Gaussian process is block-exchangeable if and only if the mean vector is constant $\mu \in \mathbf{1}$, and the covariance matrix for a sample $U = \{u_1, \dots, u_n\}$ is a non-negative linear combination of the three matrices

$$\text{cov}(Y[U]) = \sigma_0^2 I_n + \sigma_1^2 B[U] + \sigma_2^2 J_n,$$

where $J_n(u, u') = 1$ for $u, u' \in U$. In particular, $E(Y_u) = E(Y_{u'})$ for every pair, regardless of the block sizes, and the covariances are also independent of the block sizes.

The block-exchangeability assumption that P_U depends only on $B[U]$ is most natural if all population blocks are equal in size, which usually means infinite. If some or all of the population blocks are finite we can associate with each unit u the number $x(u)$, which is the size of the block in the population to which u belongs. When this is done, x is a covariate, and the implications of exchangeability are drastically different because P_U may depend on $x[U]$ in addition to $B[U]$.

14.1.4 Stationarity

Let $\mathcal{U} = \mathbb{R}$ be the index set. No covariates are registered, and the temporal difference $R(t, t') = t' - t$ is the only registered relationship. The restriction of R to a sample is a square matrix $R[U]$ of signed temporal differences; two samples are called congruent or structurally equivalent if $R[U] = R[U']$. Congruence is an equivalence relation on samples, denoted by $U \cong U'$; in this setting, it implies $U' = U + h$ for some real number h .

A process with distributions P_U is said to be stationary, or invariant with respect to temporal translation, if $U \cong U'$ implies $P_U = P_{U'}$. In particular, stationarity implies that all singletons have the same distribution $Y_t \sim Y_{t'}$.

Any transformation of R , such as the absolute distance R^+ , is also a relationship on the units; R^+ is said to be a coarser relationship than R because the partition defined by R is a sub-partition, or a finer partition, of that defined by R^+ . In particular, $R^+(t, t') = R^+(t', t)$ is symmetric whereas R is not. If $R^+[U] = R^+[U']$ implies $P_U = P_{U'}$, the process is not only stationary but also reversible.

14.1.5 Exchangeability

The general principle of exchangeability is easy to understand and straightforward to state. Two samples are said to be congruent if they have the same structure; this is understood broadly to include not only covariates but also pairwise and higher-order relationships among units. Congruence, denoted by $U \cong U'$, is an equivalence relation among samples. It implies that the samples are of equal size, $U = (u_1, \dots, u_n)$ and $U' = (u'_1, \dots, u'_n)$; it implies that the covariate values are

equal $x(u_i) = x(u'_i)$; it implies that all pairwise relationships are equal $R(u_i, u_j) = R(u'_i, u'_j)$, and so on. In particular, $u_i = u_j$ if and only if $u'_i = u'_j$.

Exchangeability is nothing more than the statement that congruent samples are required to have the same response distribution, i.e., $U \cong U'$ implies $P_U = P_{U'}$. For singletons, $x(u) = x(u')$ implies $P_u = P_{u'}$; for pairs $(x(u_1), x(u_2)) = (x(u'_1), x(u'_2))$ and $R(u_1, u_2) = R(u'_1, u'_2)$ together imply $P_{u_1, u_2} = P_{u'_1, u'_2}$.

Exchangeability is not a statement of biological, medical or scientific fact. It is a mathematical statement of equity or equality, corresponding roughly to fairness or even-handedness, which implies that probabilistic statements are based only on facts that are registered at baseline. By supposition, all relevant facts are encoded in \mathbf{x} . Without a symmetry condition of this sort, conveniently selected alternative facts are no less compelling than recorded facts. Such a world view may be acceptable in politics and in the theatre, but it is an impediment to science.

14.1.6 Axiomatic Point

Block exchangeability is defined in Sect. 14.1.3 by the statement $B[U] = B[U']$ implies $P_U = P_{U'}$. This matrix equality $B[U] = B[U']$ makes sense only if both samples are ordered, which is the convention adopted throughout these notes although it is not the standard statistical convention. However, if $B(u, u') = 1$ and $u \neq u'$, the ordered samples $U = (u, u)$ and $U' = (u, u')$ satisfy $B[U] = B[U']$. Despite the statement, block exchangeability does not imply $(Y_u, Y_{u'}) \sim (Y_u, Y_u)$ for pairs belonging to the same block. Why not?

In the first paragraph of Sect. 14.1.3, the samples were required to consist of distinct units, so the difficulty was eliminated by this restriction. The real reason for the restriction is a more fundamental consequence of standard set theory, namely that the identity function is axiomatically a registered relationship for every set. Structural equivalence of samples implies both $I[U] = I[U']$ for the identity, and $B[U] = B[U']$ for the block factor. In the case of a regression process, structural equivalence implies $I[U] = I[U']$ and $\mathbf{x}[U] = \mathbf{x}[U']$. With this understanding the restriction to distinct units in Sects. 14.1.1–14.1.3 is not needed.

14.1.7 Block Randomization

Let B be a given partition of the finite set $[n]$ into blocks, and let \mathcal{G} be the group of permutations that preserves the partition. In other words, \mathcal{G} is the set of permutations $\sigma : [n] \rightarrow [n]$ such that $B_{\sigma(i), \sigma(j)} = B_{i,j}$.

To any vector $y = (y_1, \dots, y_n)$ in \mathcal{S}^n there corresponds a randomized vector $Y = y\sigma$ whose components $Y = (y_{\sigma(1)}, \dots, y_{\sigma(n)})$ are obtained by composing the given vector with a random permutation σ uniformly distributed over the group. Randomization defines a process with state space \mathcal{S} and finite index set $[n]$. By

definition, for each $\tau \in \mathcal{G}$, the group product $\sigma\tau$ is also uniformly distributed, so the permuted random vector $Y\tau = y\sigma\tau$ has the same distribution as Y . The distribution is invariant with respect to the natural sub-group of permutations associated with the block factor.

If all blocks of B are of equal size, the distribution of Y is block-exchangeable in the sense of Sect. 14.1.3. Otherwise, if there are blocks of different sizes, Y is not block-exchangeable. For example, if $n = 7$, and $B = 1|2|34|567$ is a partition into four blocks, the group contains $2 \times 2 \times 6 = 24$ elements. Block randomization implies $Y_1 \sim Y_2$ because the transposition $1 \leftrightarrow 2$ is a group element; it also implies $Y_3 \sim Y_4$ and $Y_5 \sim Y_6 \sim Y_7$ for similar reasons. But it does not imply $Y_1 \sim Y_3$ or $Y_3 \sim Y_5$ because there is no group element such that $\sigma(1) = 3$ or $\sigma(3) = 5$.

14.2 Families with Independent Components

14.2.1 Parametric Models

A parametric statistical model associates with each parameter point $\theta \in \Theta$ a probability distribution P_θ . In general, a distribution is a process which associates with each sample $U \subset \mathcal{U}$ a probability distribution $P_{\theta,U}$ on the observation space \mathcal{S}^U . If the process has independent components, the description can be simplified to a great extent by focusing on the one-dimensional marginal distributions.

The following examples illustrate a range of possibilities.

14.2.2 IID Model I

The parameter space is the set of probability distributions defined on the state space, say $\mathcal{S} = \mathbb{R}$ with Borel subsets. In other words, the sequence Y_u for $u \in \mathcal{U}$ has independent and identically distributed components $Y_u \sim \theta$.

Properties of a statistical model are often gauged by their behaviour under the action of a suitable group or semi-group of measurable transformations $g: \mathcal{S} \rightarrow \mathcal{S}$. If Y_1, \dots are independent and identically distributed with distribution $\theta \in \Theta$, then the transformed variables $g(Y_1), g(Y_2), \dots$ are independent and identically distributed with parameter $g\theta \in \Theta$, where

$$g\theta(A) = P_\theta(gY \in A) = P_\theta(Y \in g^{-1}A) = \theta(g^{-1}A)$$

for $A \subset \mathcal{S}$. Since Θ is the set of Borel distributions on \mathcal{S} , the transformed distribution is simply another point $\theta' = g\theta$ in the same parameter space. Provided that g is invertible, the two specifications

$$Y_1, Y_2, \dots \stackrel{\text{iid}}{\sim} \theta, \quad \text{and} \quad gY_1, gY_2, \dots \stackrel{\text{iid}}{\sim} g\theta$$

are mathematically equivalent. Equivalence, or equi-variance, implies that any inferential statement about θ after observing y must be tied to inferential statements about $g\theta$ after observing gy .

To each observation point $y = (y_1, \dots, y_n)$ there corresponds an empirical distribution function

$$\hat{\theta}(A; y) = n^{-1} \sum_{i=1}^n \delta_{y_i}(A) = n^{-1} \#\{i \in [n] : y_i \in A\}.$$

The function $y \mapsto \hat{\theta}(y)$ is equi-variant in the sense that $\hat{\theta}(gy) = g\hat{\theta}(y)$. The transformation $y \mapsto gy$ acts component-wise $\mathcal{S}^n \rightarrow \mathcal{S}^n$, while $\theta \mapsto g\theta$ is the induced transformation on distributions on \mathcal{S} . For invertible transformations, this means $\hat{\theta}(y) = g^{-1}\hat{\theta}(gy)$. The empirical distribution is sometimes called the nonparametric maximum-likelihood estimate, or the bootstrap estimate.

14.2.3 IID Model II

The parameter space is the set of Gaussian distributions on the real line. Since the Gaussian distribution is determined by its mean and variance, this statement typically means one of the following:

$$\begin{aligned} \Theta &= \mathbb{R}^2; & P_\theta &= N(\theta_1, \theta_2^2); \\ \Theta &= \mathbb{R} \times (0, \infty); & P_\theta &= N(\theta_1, \theta_2^2); \\ \Theta &= \mathbb{R}^2; & P_\theta &= N(\theta_1, e^{\theta_2}). \end{aligned}$$

Given a parameter point θ , the components are independent and identically distributed $Y_u \sim P_\theta$ on the real line.

The three versions are not mathematically equivalent. In version one, the two distinct points $(\theta_1, \pm\theta_2)$ define the same distribution, so the parameter is not identifiable. In addition, the boundary subset of Dirac distributions $N(\theta_1, 0)$ is included in the first version, but not in the other two. Versions two and three are equivalent in the sense that they contain the same set of non-degenerate distributions. Differences of this sort are sometimes important in theoretical work, for example in questions concerning the existence of a parameter point that maximizes the likelihood. But, for the most part, minor differences in parameterization are not of great importance and are usually overlooked in applied work.

All three versions are affine equi-variant in the sense that $Y_u \sim P_\theta$ implies $gY_u \sim P_{g\theta}$ for affine transformations $y \mapsto gy = g_0 + g_1y$ with $g_1 > 0$. The induced transformation on the parameter space is group composition

$$(\theta_1, \theta_2) \mapsto (g_0 + g_1\theta_1, g_2\theta_2)$$

for versions one and two, and

$$(\theta_1, \theta_2) \mapsto (g_0 + g_1 \theta_1, \theta_2 + 2 \log |g_2|)$$

in the third version where θ_2 is the log variance.

For a sample of size $n \geq 2$ and an observation $y \in \mathbb{R}^n$, the usual estimate of the parameters for version 1 or 2 is

$$\hat{\theta}_1 = \bar{y}_n; \quad \hat{\theta}_2^2 = s_n^2 = \sum (y_i - \bar{y}_n)^2 / (n - 1).$$

This estimator is affine equivariant in the sense that $\hat{\theta}(gy) = g\hat{\theta}(y)$ for affine transformations $y \mapsto gy$ acting component-wise. For this purpose, the divisor $n - 1$ could be replaced by n . Similar remarks with minor modifications apply to version 3.

The Cauchy distribution $C(\theta)$ with median θ_1 and probable error $|\theta_2|$ has a density

$$\frac{|\theta_2| dy}{\pi |y - \theta|^2},$$

where $\theta = \theta_1 + i\theta_2$ is a complex number, and $|y - \theta|^2$ is the squared modulus. Apart from the specific formulae for parameter estimates, all of the preceding remarks apply equally to the Cauchy family and every symmetric location-scale family on the real line. The complex parameterization is unimportant at this stage, but becomes relevant if reciprocal transformations are considered: $Y \sim C(\theta)$ implies $1/Y \sim C(1/\theta)$.

14.3 Non-i.d. Models

14.3.1 Classification Factor

We consider a simple model for a process in which each individual is classified as male or female. All values are independent, and they are identically distributed for individuals of the same sex. Each model is defined by a parameter space Θ , a function $\theta \mapsto P_\theta$ for males, i.e., for all u such that $x(u) = M$, and a function $\theta \mapsto Q_\theta$ for all u such that $x(u) = F$. The symbol \mathbb{R}_+ may be interpreted as either the set of non-negative numbers or the set of strictly positive numbers.

In the first model, the values are independent Bernoulli with success rates π_0 for males and π_1 for females. The parameter space may be extended to include the boundary points if so desired. There is nothing exceptional in this or in the second model, which is Gaussian with sex-dependent mean and constant variance. The third model has a similar structure with constant mean and a sex-dependent

variance, while both parameters are sex-dependent in the fourth model. In (vii), the distributions are arbitrary; $Y_u \sim \theta_0$ for males and $Y_u \sim \theta_1$ for females.

Most readers whose experience lies in applied work would blanch at the penultimate suggestion in which male values are Gaussian while female values are distributed as Cauchy. The reasons for this have nothing to do with Cauchy versus Gauss as individuals, or with male variability versus female variability, or with the suitability of this model for any specific application. Instead, they are anchored in the well-established legal principle of ‘equality under the law’, a desire to avoid overt bias related to visible factors such as race, sex and religion that are, by common agreement, incidental under law.

One mathematical statement of those principles is equi-variance under label-switching. In the present setting, the permutation σ that transposes M with F also switches P with Q . Equi-variance means that to each transposition of factor labels there corresponds a permutation of parameter components such that $P_\theta(A) = Q_{\sigma\theta}(A)$ for every event A . All of the models listed above are equi-variant except for (v) and (vi).

Equi-variance does not imply that the distribution for males is the same as the distribution for females, but it does imply that the set of distributions under consideration is the same for both. Each sex gets to pick one distribution from the same set, so there is equality of opportunity in that sense. However, the Gaussian model in the fifth row of Table 14.1 shows that equality of the sets $\{P_\theta : \theta \in \Theta\}$ and $\{Q_\theta : \theta \in \Theta\}$ is not sufficient for equi-variance.

Equi-variance is not a fundamental principle on a par with Kolmogorov consistency for a stochastic process. It is not even on a par with the principle of exchangeability for individuals having the same covariate value. Equi-variance is reasonably compelling in many circumstances and is a natural default for any factor whose levels are unordered or otherwise unstructured. For example, *occupation* is a classification factor, but the set of levels is not devoid of structure. In a survey with limited options, one level might be *employed but none of the above*. Equi-variance is a mathematical solution to the problem of accommodating distinct classes of units on an equal footing in the stochastic model.

Table 14.1 Parameterization of statistical models for a classification factor

| | θ | Θ | P_θ | Q_θ | Eqv? |
|-------|--------------------------------------|--|------------------------|------------------------|------|
| (i) | (π_0, π_1) | $(0, 1)^2$ | $Ber(\pi_0)$ | $Ber(\pi_1)$ | ✓ |
| (ii) | (μ_0, μ_1, σ) | $\mathbb{R}^2 \times \mathbb{R}_+$ | $N(\mu_0, \sigma^2)$ | $N(\mu_1, \sigma^2)$ | ✓ |
| (iii) | $(\mu, \sigma_0, \sigma_1)$ | $\mathbb{R} \times \mathbb{R}_+^2$ | $N(\mu, \sigma_0^2)$ | $N(\mu, \sigma_1^2)$ | ✓ |
| (iv) | $(\mu_0, \mu_1, \sigma_0, \sigma_1)$ | $\mathbb{R}^2 \times \mathbb{R}_+^2$ | $N(\mu_0, \sigma_0^2)$ | $N(\mu_1, \sigma_1^2)$ | ✓ |
| (v) | (μ_0, μ_1, σ) | $\mathbb{R}^2 \times \mathbb{R}_+$ | $N(\mu_0, \sigma^2)$ | $N(\mu_1, 2\sigma^2)$ | |
| (vi) | $(\mu_0, \mu_1, \sigma_0, \sigma_1)$ | $\mathbb{R}^2 \times \mathbb{R}_+^2$ | $N(\mu_0, \sigma_0^2)$ | $C(\mu_1, \sigma_1)$ | |
| (vii) | (θ_0, θ_1) | $\mathcal{P}(\mathbb{R}) \times \mathcal{P}(\mathbb{R})$ | θ_0 | θ_1 | ✓ |

14.3.2 Treatment

A treatment factor and a classification factor are accommodated in a statistical model in very different ways. The distinction is seldom emphasized and it is often not readily apparent. We begin by assuming homogeneity in the sense that no covariate is defined on the units. The first step is to specify the set of reference-level distributions $\{P_\theta : \theta \in \Theta_0\}$, each of which is interpreted as a conditional distribution given $T = 0$

$$P(Y_u \in A \mid T = 0; \theta) = P_\theta(A).$$

Treatment has an *effect*, possibly null, so the second step focuses on the set of possible treatment effects $g \in \mathcal{G}$, and on how each reference-level distribution is modulated by g . Each treatment modulation is an action on the parameter space $\theta \mapsto g\theta$ which sends P_θ to $P_{g\theta}$. The interpretation of the action by g is as follows: if the conditional distribution given $T = 0$ is P_θ , and g is the treatment effect, the conditional distribution given $T = 1$ shall be $P_{g\theta}$. To make sense of this, it is necessary that the set \mathcal{G} be a group acting on Θ_0 ; the group identity corresponds to the null treatment effect.

The overall parameter space is the product set $\Theta_0 \times \mathcal{G}$. By definition of group action, each transformation $g : \Theta_0 \rightarrow \Theta_0$ is invertible, so $g\Theta_0 = \Theta_0$. Thus, whatever the treatment effect may be, the set of conditional distributions given $T = 1$ is the same as the set of distributions given $T = 0$. Since the action is a group homomorphism, it is immaterial which level of T is used as the reference level. This condition immediately excludes the fifth and sixth models in Table 14.1 as possibilities for modelling a treatment effect.

For this setting, where there is a single treatment factor and no covariate, the Bernoulli logistic and probit models in Table 14.2 are equivalent, and both are equivalent to the Bernoulli model in Table 14.1. The distributions are in 1–1 correspondence, and the only differences are in the parameterizations.

In the first three Gaussian models, the group acts additively on the parameter, sending $(\mu, \log \sigma)$ to $(\mu + g, \log \sigma)$ in example (iii), to $(\mu, g + \log \sigma)$ in

Table 14.2 Examples of statistical models for a treatment factor

| | Θ_0 | \mathcal{G} | $g\theta$ | P_θ | $P_{g\theta}$ |
|-------|---------------------------|--------------------------|-------------------------------|--|--|
| (i) | $\theta \in \mathbb{R}$ | \mathbb{R} | $\theta + g$ | $\text{Ber}\left(\frac{e^\theta}{1+e^\theta}\right)$ | $\text{Ber}\left(\frac{e^{\theta+g}}{1+e^{\theta+g}}\right)$ |
| (ii) | $\theta \in \mathbb{R}$ | \mathbb{R} | $\theta + g$ | $\text{Ber}(\Phi(\theta))$ | $\text{Ber}(\Phi(\theta + g))$ |
| (iii) | (μ, σ) | \mathbb{R} | $(\mu + g, \sigma)$ | $N(\mu, \sigma^2)$ | $N(\mu + g, \sigma^2)$ |
| (iv) | (μ, σ) | \mathbb{R} | $(\mu, \sigma e^g)$ | $N(\mu, \sigma^2)$ | $N(\mu, \sigma^2 e^{2g})$ |
| (v) | (μ, σ) | \mathbb{R}^2 | $(\mu + g_1, \sigma e^{g_2})$ | $N(\mu, \sigma^2)$ | $N(\mu + g_1, \sigma^2 e^{2g_2})$ |
| (vi) | (μ, σ) | $\text{Aff}(\mathbb{R})$ | $(g_0 + g_1\mu, \sigma g_1)$ | $N(\mu, \sigma^2)$ | $N(g_0 + g_1\mu, \sigma^2 g_1^2)$ |
| (vii) | $\mathcal{P}(\mathbb{R})$ | Bic | $\theta \circ g^{-1}$ | θ | θg^{-1} |

example (iv), and $(\mu + g_1, g_2 + \log \sigma)$ in example (v). Each is a reparameterization of one of the Gaussian models listed in Table 14.1. Examples three and four are different reparameterizations of the same set of distributions. They are equivalent in exactly the same sense that the two Bernoulli models are equivalent.

In the fourth Gaussian model, $\theta = (\mu, \sigma)$ and (g_0, g_1) are two points in the group of affine transformations $\mathbb{R} \rightarrow \mathbb{R}$, and $g\theta$ is the group composition, which is not commutative, i.e., $g\theta \neq \theta g$. The treatment effect is equivalent in distribution to the state-space transformation $Y \mapsto gY$, so that $gY \sim N(g_0 + g_1\mu, \sigma^2 g_1^2) = P_{g\theta}$. Most of the treatment effects exhibited in Table 14.2 are not induced by an action on the state space.

In the last example, Θ_0 is the set of probability distributions on the real line, so the control distribution is an arbitrary distribution defined on Borel subsets. For this setting, the treatment effect can be modelled using any group acting on distributions, whether it is finite-dimensional or infinite-dimensional. The suggestion $\mathcal{G} = \text{Bic}(\mathbb{R})$, meaning bi-continuous transformations $\mathbb{R} \rightarrow \mathbb{R}$ having an inverse that is also continuous, is topologically natural. But there are many other possibilities such as the group of Borel-measurable transformations preserving Lebesgue measure. Regardless of the group, the product set $\Theta_0 \times \mathcal{G}$ is not in 1–1 correspondence with the product set $\mathcal{P}(\mathbb{R}) \times \mathcal{P}(\mathbb{R})$ in Table 14.1, so this pair of models is not equivalent for any group.

The group of transformations on distributions may be induced by transformations $\mathcal{S} \rightarrow \mathcal{S}$ on the state space. The set of bi-continuous transformations provides an example of that type, as does $N(\mu, \sigma^2) \mapsto N(\mu + g, \sigma^2)$, which is induced by translation. However, the Bernoulli state space is $\mathcal{S} = \{0, 1\}$ and the group $\mathcal{G} = \mathbb{R}$ does not act on \mathcal{S} , so neither Bernoulli transformation is associated with a transformation $\mathcal{S} \rightarrow \mathcal{S}$. The second and third Gaussian examples are also not associated with a state-space transformation.

14.3.3 Classification Factor Plus Treatment

Let $x : \mathcal{U} \rightarrow [k]$ be a k -level classification factor, so that x_u is the class of unit u . We assume that the levels are unordered and otherwise unstructured. For the logistic version of the Bernoulli model, the parameter space Θ_0 consists of k real numbers $\theta_1, \dots, \theta_k$, where

$$\text{logit } P_\theta(Y_u = 1 \mid T_u = 0) = \theta_{x(u)}.$$

is the conditional log odds of success for unit u , and for every unit in class $x(u)$. The treatment effect is a group action on the space $\Theta_0 = \mathbb{R}^k$, which sends θ to $g\theta$. In the absence of additional structure (such as an inner product) there are two principal options for the group and its action; either $\mathcal{G} = \mathbb{R}$ and $g\theta = (\theta_1 + g, \dots, \theta_k + g)$ or $\mathcal{G} = \mathbb{R}^k$ and $g\theta = (\theta_1 + g_1, \dots, \theta_k + g_k)$.

The first option means that

$$\text{logit } P_\theta(Y_u = 1 \mid T_u = 1) = \theta_{x(u)} + g,$$

so the conditional odds of success for unit u satisfy

$$\log\left(\frac{\text{odds}(Y_u = 1 \mid T_u = 1)}{\text{odds}(Y_u = 1 \mid T_u = 0)}\right) = g.$$

By this odds-ratio yardstick, the effect of treatment is the same number g for unit in every class. No interaction between treatment and the class means that the treatment effect on some specified scale is the same for every class, so this group action implies no interaction on the logistic scale.

The second option means that $g \in \mathbb{R}^k$ acts additively

$$\text{logit } P_\theta(Y_u = 1 \mid T_u = 1) = \theta_{x(u)} + g_{x(u)},$$

so the conditional odds of success for unit u satisfy

$$\log\left(\frac{\text{odds}(Y_u = 1 \mid T_u = 1)}{\text{odds}(Y_u = 1 \mid T_u = 0)}\right) = g_{x(u)}.$$

Unless $g \in \mathbf{1}_k \subset \mathbb{R}^k$, the effect of treatment as measured by the odds ratio is different for each class. This group action implies interaction on the logistic scale. Generally speaking, no interaction on the logistic scale implies interaction on the probit scale, and vice-versa.

14.3.4 Quantitative Covariate Plus Treatment

If a quantitative covariate $x : \mathcal{U} \rightarrow \mathcal{X}$ is defined on the units, the baseline parameter is a function $x \mapsto \theta(x)$ from \mathcal{X} into some space such as \mathbb{R} or \mathbb{R}^2 , and Θ_0 is a suitable set of such functions on which the treatment group acts. The statement that x is a quantitative covariate implies that \mathcal{X} is a vector space and that the topology is relevant, so each function $x \mapsto \theta(x)$ in Θ_0 is required to be continuous. Ordinarily, Θ_0 contains the one-dimensional space of constant functions plus the k -dimensional space of linear functionals, where $k = \dim(\mathcal{X})$.

For both Bernoulli models in Table 14.2, $\theta(x)$ is the logit or probit value, which is a real number. In the logistic version with a constant treatment effect, an individual u whose covariate value is x_u , has log odds of success either $\theta(x_u)$ if $T = 0$, or $\theta(x_u) + g$ if $T = 1$. In the probit version, $\theta(x_u)$ and $\theta(x_u) + g$ are the values on the probit scale. Other link functions such as the complementary log-log operate in the same way. In general, the treatment effect is a group action, $\theta(x) \mapsto \theta(x) + g(x)$ by addition of functions.

In a setting where Θ_0 contains the $k + 1$ -dimensional space of affine functionals $\theta(x) = \theta_0 + \theta_1(x)$, the simplest options available to accommodate treatment effects with or without interaction are the following:

$$\begin{aligned}\text{logit } P_\theta(Y_u = 1 \mid T_u = 0) &= \theta(x_u) \\ \text{logit } P_\theta(Y_u = 1 \mid T_u = 1) &= \theta(x_u) + g_0 \\ \text{logit } P_\theta(Y_u = 1 \mid T_u = 1) &= \theta(x_u) + g_1(x_u) \\ \text{logit } P_\theta(Y_u = 1 \mid T_u = 1) &= \theta(x_u) + g_0 + g_1(x_u).\end{aligned}$$

Here $g_0 \in \mathbb{R}$, and g_1 is a linear functional $\mathcal{X} \rightarrow \mathbb{R}$, so $g_1(0) = 0$. The third version implies that treatment has no effect on the set of units for which $x_u = 0$. This situation is not common, but it does occur if x represents time, and $x = 0$ is the baseline either immediately pre-treatment, or immediately post-treatment before the treatment has had time to take effect. For an illustration, see Chap. 5.

It is crucial that the space Θ_0 be closed under the group. For example, if \mathcal{X} is a vector space and every $\theta \in \Theta_0$ is a linear functional $\mathcal{X} \rightarrow \mathbb{R}$, then $x \mapsto \theta(x) + g$ sends zero to g , which is not a linear functional on \mathcal{X} . The standard choices for statistical practice are either the space $\Theta_0 = \mathbf{1}$ of constant functions on \mathcal{X} or the space of affine functions $\mathcal{X} \rightarrow \mathbb{R}$, not the space of linear functionals. Other options include the space of inhomogeneous polynomial functions of degree $\leq k$.

14.3.5 Random Coefficient Models

Consider the simple linear regression model with independent components and a quantitative covariate x . The response distribution given the parameter (η, σ) is $Y \sim N_n(\eta(x), \sigma^2 I_n)$, where the mean function $\eta(x) = \eta_0 + \eta_1 x$ is linear in x . In a random coefficients model, some or all of the regression coefficients are regarded as random variables. Suppose, therefore, that $\eta(\cdot)$ is a random linear function in which the coefficient vector (η_0, η_1) is bivariate normal with mean (β_0, β_1) , variances σ_0^2, σ_1^2 and correlation ρ . Then the marginal distribution of Y is Gaussian with moments

$$\begin{aligned}E(Y_u) &= \beta_0 + \beta_1 x_u \\ \text{cov}(Y_u, Y_{u'}) &= \sigma^2 \delta_{u,u'} + \sigma_0^2 + \sigma_1^2 x_u x_{u'} + \rho \sigma_0 \sigma_1 (x_u + x_{u'}).\end{aligned}$$

The six-parameter model is identifiable in the standard sense that distinct parameter points give rise to distinct distributions, i.e., $P_\theta = P_{\theta'}$ implies $\theta = \theta'$, at least for non-trivial designs and interior parameter points with $\sigma_0 \sigma_1 > 0$.

This version of the random-coefficients model trades a three-parameter Gaussian model for a six-parameter model. This transaction might be favourable if the larger model were capable of accommodating effects that the simpler model cannot handle. Sadly, that is not the case. The triple $(\sum Y_u, \sum x_u Y_u, \sum Y_u^2)$ is minimal sufficient

for both models, and the likelihood is maximized at the boundary point $\sigma_0 = \sigma_1 = 0$ with ρ indeterminate. Regardless of the number of observations, the variance-components components $\sigma_0^2, \sigma_1^2, \rho\sigma_0\sigma_1$ are not estimable.

The transaction is more favourable if B is a block factor, and the conditional distribution is $Y \sim N_n(\eta_{b,0} + \eta_{b,1}x, \sigma^2 I_n)$ with bivariate coefficients η_b independent and identically distributed for each block. The marginal distribution is then Gaussian with moments

$$\begin{aligned} E(Y_u) &= \beta_0 + \beta_1 x_u & (14.1) \\ \text{cov}(Y_u, Y_{u'}) &= \sigma^2 \delta_{u,u'} + B_{u,u'} (\sigma_0^2 + \sigma_1^2 x_u x_{u'} + \rho \sigma_0 \sigma_1 (x_u + x_{u'})). \end{aligned}$$

Block-exchangeability implies that the number of parameters is independent of the number of blocks.

This is a variance-components model in which the covariance matrix is a linear combination of four given matrices. It is slightly non-standard in that only three of the four coefficients are required to be positive. Nonetheless, if there are at least three blocks in the sample, the parameter is estimable, and maximum-likelihood estimation presents no serious difficulties. However, consistent estimation of between-block variance components is not possible unless the number of blocks is large.

Assumptions of randomness of coefficients are ineffective for model simplification unless they are accompanied by assumptions of distributional symmetry. Symmetry assumptions are usually most effective in situations where the number of coefficients is large and preferably infinite, and the joint distribution can reasonably be taken to be invariant with respect to the action of some large group. In the present setting, the number of coefficients is twice the number of blocks, which is usually modest in practice, and the distribution is exchangeable with respect to block permutation. The covariance model is not stationary in the sense of invariance with respect to translation $x \mapsto x + g$, but it is at least equi-variant under affine transformation: see Exercise 14.7.

My enthusiasm for finite random-coefficient models is tempered by the fact that I have yet to encounter a convincing application—one in which each block has a different mean trend and yet the trends are all exactly linear in x . In situations such as those in Chaps. 4, 5 and 8, linearity of the within-block conditional expectation $E(Y | \eta) = \eta_b(x)$ as a function of x is a mathematical possibility, but the available evidence points strongly to non-linearity—of growth curves, for example. For those projects, a more flexible model was chosen in which the processes $\eta_b(\cdot)$ are continuous random functions whose trajectories are not linear in x ; they are also exchangeable for distinct blocks. In matrix notation, the moments of the marginal distribution are

$$\begin{aligned} E(Y) &= X\beta & (14.2) \\ \text{cov}(Y) &= \sigma^2 I_n + \sigma_0^2 B + \sigma_1^2 K + \sigma_2^2 B \cdot K, \end{aligned}$$

where the matrix $K_{u,u'} = K(x_u, x_{u'})$ has full rank, and $B \cdot K$ is the Hadamard component-wise product. Once again, the model has four variance components and two regression coefficients, so the choice of one over the other is not made on the basis of parameter counting.

One crucial difference between the two versions is that (14.1) implies independence of observations in different blocks, while $\sigma_1^2 > 0$ in (14.2) implies otherwise. Note also that the sum of the last three terms in (14.1) is a Hadamard-product matrix of the form $B \cdot K$, where K has rank two with eigenvectors confined to the linear subspace $\mathcal{X} = \text{span}(\mathbf{1}, x)$.

14.4 Examples of Treatment Effects

14.4.1 Simple Gaussian Model Without Interaction

Let \mathcal{D}_n be the space of Gaussian distributions $N_n(\mu, \Sigma)$ indexed by $\mu \in \mathbb{R}^n$ and Σ in the space of positive-definite $n \times n$ matrices. The outcome of randomization is a treatment assignment vector $T \in \mathbb{R}^n$ with components in $\{0, 1\}$. The joint distribution of T is known, and specified in the protocol.

Given $T = \mathbf{t}$, the effect of treatment is an action $\mathcal{D}_n \rightarrow \mathcal{D}_n$ on distributions by some group \mathcal{G} of treatment effects. In the simplest case, $\mathcal{G} = \mathbb{R}$, and the action is additive on the mean

$$N_n(\mu, \Sigma) \xrightarrow{g} N_n(\mu + \mathbf{t}g, \Sigma), \quad (14.3)$$

keeping the covariances fixed.

Consider a standard linear model that is typical of what might be encountered in a scientific experiment, where $i \mapsto x_i$ is the covariate, and $(i, j) \mapsto V_{ij}$ is a non-identity relation that is also positive semi-definite. In the absence of treatment, i.e., if $\mathbf{t} = 0$, the response distribution is some point in the subset $\Theta_0 \subset \mathcal{P}_n$

$$N(X\beta, \sigma_0^2 I_n + \sigma_1^2 V)$$

indexed by $\beta \in \mathbb{R}^p$, with two variance components $\sigma_0^2, \sigma_1^2 > 0$. Given $T = \mathbf{t}$, the group action generates an orbit

$$\mathcal{O}(\mathbf{t}) = \{N_n(X\beta + \mathbf{t}g, \sigma_0^2 I_n + \sigma_1^2 V) : g \in \mathcal{G}\}$$

consisting of Gaussian distributions indexed by $\beta \in \mathbb{R}^p$, $g \in \mathbb{R}$, plus $\sigma_0^2, \sigma_1^2 > 0$. Provided that $\text{span}(X)$ includes the one-dimensional subspace of constant functions, the complementary treatment vectors \mathbf{t} and $\bar{\mathbf{t}} = 1 - \mathbf{t}$ generate the same orbit.

This is the standard Gaussian model for a treatment effect that is constant and additive for all units, regardless of the covariate value. In general, however, the

treatment effect need not be additive on the mean; moreover, if it is additive it need not be the same constant for every unit.

14.4.2 Additive Interaction

Loosely speaking, interaction means that the effect of treatment for one unit is not the same as the effect for another unit. In order for this to be the case, we must have $x(u) \neq x(u')$, so the treatment action depends on x . In the simplest setting, x is binary, $\mathcal{G} = \mathbb{R}^2$, and the group action (14.3) becomes

$$N_n(\mu, \Sigma) \xrightarrow{g} N_n(\mu + \mathbf{t}g_0 + \mathbf{t} \cdot x g_1, \Sigma). \quad (14.4)$$

For units at the reference level such that $x_u = 0$, the treatment effect is an additive increase in the mean by g_0 ; for units such that $x_u = 1$, the treatment effect is additive by $g_0 + g_1$. The difference g_1 is called the interaction.

In (5.2), the treatment effect is a differential drift, whose magnitude is directly proportional to time-since-baseline. That means that the action of the group element $g \in \mathbb{R}$ is an additive function of the product $g \times \text{time}$. The effect on the mean is not the same for every unit. Nonetheless, $\mathcal{G} = \mathbb{R}$, so it is not entirely clear whether this should be counted as interaction.

A similar effect can be generated artificially by restriction of (14.4) to the one-dimensional sub-group $g_0 = g_1$.

14.4.3 Survival Models

One further example may help to illustrate the options available for group action on distributions. Let Θ_0 be the set of non-negative measures on $\mathbb{R}^+ = (0, \infty)$ that are locally finite near the origin, i.e., there exists $t > 0$ such that the interval $(0, t)$ has finite measure. To each $\theta \in \Theta_0$ there corresponds a probability distribution on $\mathcal{S} = \mathbb{R}^+ \cup \{\infty\}$ defined by the survivor function

$$P_\theta(Y > t) = \exp(-\theta((0, t])),$$

which implies $P_\theta(Y > 0) = 1$ and $P_\theta(Y = \infty) = e^{-\theta(\mathbb{R}^+)}$. A distribution on \mathcal{S} is called a survival distribution; every probability distribution P on \mathbb{R}^+ is regarded as a survival distribution such that $P(\{\infty\}) = 0$. In this setting, θ is called the hazard measure.

To each survival distribution P there corresponds a hazard measure θ such that

$$\theta(0, t] = \begin{cases} -\log P(t, \infty] & P(t, \infty] > 0; \\ \infty & P(t, \infty] = 0; \end{cases}$$

for $0 < t < \infty$. Although not especially important in practice, it is worth noting that two distinct hazard measures may give rise to the same survival distribution: see Exercise 14.14.

Hazard Multiplication

Consider now the group $\mathcal{G} = \mathbb{R}^+$ of strictly positive scalars acting on Θ_0 by scalar multiplication

$$\theta(dt) \xrightarrow{g} g \times \theta(dt).$$

Each group element is a treatment effect, which is an invertible transformation $g: \Theta_0 \rightarrow \Theta_0$, or equivalently $P_\theta \mapsto P_{g\theta}$, by scalar multiplication of the hazard measure. In the absence of covariates, the parameter space is $\Theta_0 \times \mathcal{G}$.

The proportional-hazards model states that each individual has a conditional hazard given treatment, one for $T = 0$ and one for $T = 1$; if $g > 0$ is the treatment effect, the two hazard measures are $\theta_u(dt)$ and $g\theta_u(dt)$. As always, these are subject to exchangeability: $x(u) = x(u')$ implies $\theta_u = \theta_{u'}$.

The group does not act transitively on the parameter space, which means that there is more than one orbit—infinitely many in fact. Each orbit or set of orbits is a sub-model. For example, the subset $\theta(dt) \propto dt$ consisting of measures having a constant strictly positive density with respect to Lebesgue measure is an orbit corresponding to the set of exponential distributions. For each real $\alpha > 0$, the subset $\theta(dt) \propto dt t^{\alpha-1}$ is an orbit, and the union of such orbits is the family associated with the set of Weibull distributions. For $\alpha \leq 0$, the measures are not locally finite at the origin, and thus not in Θ_0 .

Temporal Dilation

Consider now the group $\mathcal{G} = \mathbb{R}^+$ of positive scalars acting on the space of hazard measures by the usual rules for the transformation of distributions by temporal dilation. For present purposes, dilation means that $(g\theta)(A) = \theta(gA)$ for $A \subset \mathbb{R}^+$. Each group element is a treatment effect, which is an invertible transformation $g: \Theta_0 \rightarrow \Theta_0$, or equivalently $P_\theta \mapsto P_{g\theta}$, by scalar dilation, either of the hazard measure or the distribution itself.

The accelerated-failure model states that each individual has a conditional hazard given treatment, one for $T = 0$ and one for $T = 1$; if $g > 0$ is the treatment effect, the two hazard densities are $\theta'_u(t)$ and $g\theta'_u(gt)$. As always, these are subject to exchangeability: $x(u) = x(u')$ implies $\theta_u = \theta_{u'}$.

Non-constant Hazard Multiplication

The group actions illustrated above are the ones most commonly encountered in survival analysis. It is evident that there are many other possibilities for group action, most of which have limited potential for applied work, either because they are implausible in one way or another, or because they lead to intractable computations. Nonetheless, it may be helpful to describe a few. In the first two examples, the group is \mathbb{R}^2 with addition, and the action on hazards is multiplicative but not constant

$$\begin{aligned}\theta(dt) &\xrightarrow{g} e^{g_1+g_2 t} \theta(dt) \\ \theta(dt) &\xrightarrow{g} e^{g_1+g_2 \log(t)} \theta(dt).\end{aligned}$$

Exponentiation is used to convert the additive group \mathbb{R} or \mathbb{R}^2 into the multiplicative group \mathbb{R}^+ or $\mathbb{R}^+ \times \mathbb{R}^+$.

There are numerous variations on this theme in which Θ_0 is replaced with some subset that is closed under the group. Ordinarily, the group action should be chosen to be compatible with temporal dilation.

Classification Factor Plus Treatment

Let x be a binary classification factor such as sex. Exchangeability plus independence implies that the values are independent and identically distributed for each sex. Suppose that the hazard measure for males in the control group is a point $\theta \in \Theta_M$. The effect of treatment on males is an action $\Theta_M \rightarrow \Theta_M$ in which the group element $g \in \mathcal{G}$ sends P_θ to $P_{g\theta}$. Thus, θ and $g\theta$ both belong to Θ_M . The first two columns of Table 14.3 illustrate this action for males, while columns 3–4 illustrate a similar action for females. In each case, $\Theta_M = \Theta_F$ is the same set of hazard measures, which is closed under scalar multiplication.

Example (v) is not a group action or group homomorphism because the group identity $g = 1$ is not associated with the identity map $\Theta \rightarrow \Theta$ on hazard measures. The set of treatment effects, i.e., the set of maps $(\theta_1, \theta_2) \mapsto (g\theta_1, 2g\theta_2)$ in (v), does not have a null element or identity map corresponding to no effect. This is strictly forbidden. Example (ii) is quirky, but it is a group action, and it is equi-variant.

Equivariance for males and females implies not only that $\Theta_M = \Theta_F$, but also that the effect of treatment is the same action either by the same group or by a second copy of the same group. If the treatment effect is an action by the same group element, we say that there is no interaction. Otherwise, the effect is sex-dependent.

In the case of the proportional-hazards model, the two hazard measures for males are θ and $g\theta$. In the absence of interaction, $\mathcal{G} = \mathbb{R}^+$ and the two hazards for females are θ' , $g\theta'$ with the same proportionality constant. If interaction is present, the group $\mathcal{G} = \mathbb{R}^+ \times \mathbb{R}^+$ consists of pairs (g, g') in which g is the hazard multiplier for males and g' is the multiplier for females.

Table 14.3 Seven examples of class-plus-treatment survival models

| | Male | | Female | | \mathcal{G} | Eqv? |
|-------|------------|---------------|------------|------------------|------------------------------------|------|
| | C | T | C | T | | |
| (i) | θ | $g\theta$ | θ | $g\theta$ | \mathbb{R}^+ | ✓ |
| (ii) | θ | $g\theta$ | θ | $g^{-1}\theta$ | \mathbb{R}^+ | |
| (iii) | θ | $g_1\theta$ | θ | $g_2\theta$ | $\mathbb{R}^+ \times \mathbb{R}^+$ | |
| (iv) | θ_1 | $g\theta_1$ | θ_2 | $g\theta_2$ | \mathbb{R}^+ | ✓ |
| (v) | θ_1 | $g\theta_1$ | θ_2 | $2g\theta_2$ | \mathbb{R}^+ | NA |
| (vi) | θ_1 | $g_1\theta_1$ | θ_2 | $g_2\theta_2$ | $\mathbb{R}^+ \times \mathbb{R}^+$ | ✓ |
| (vii) | θ_1 | $g_1\theta_1$ | θ_2 | $g_1g_2\theta_2$ | $\mathbb{R}^+ \times \mathbb{R}^+$ | ✓ |

14.5 Incomplete Processes

14.5.1 Gosset Process

Ordinarily, a real-valued process indexed by the positive integers is determined by probability distributions P_n defined on Borel subsets $\mathcal{B}(\mathbb{R}^n)$. In order for this sequence of distributions to define a process such that each consecutive n -tuple (Y_1, \dots, Y_n) is distributed as P_n , it is necessary and sufficient that each P_n be the marginal distribution of P_{n+1} under deletion of the last component, i.e., $P_{n+1}(A \times \mathbb{R}) = P_n(A)$ for each event $A \subset \mathbb{R}^n$. The process is also exchangeable if each P_n is symmetric, i.e., $P_n(\sigma A) = P_n(A)$ for each coordinate permutation σ . The following are four examples of distribution sequences, each of which determines an exchangeable Gaussian process:

$$N_n(0, I_n); \quad N_n(\mathbf{1}_n, I_n); \quad N_n(\mathbf{1}_n, I_n + J_n); \quad N_n(\mathbf{1}_n \mu, \sigma_0^2 I_n + \sigma_1^2 J_n).$$

Here, $\mathbf{1}_n = (1, 1, \dots, 1)$ is the constant vector, $J_n = \mathbf{1}_n \mathbf{1}'_n$ is the constant matrix, while μ is a real number and σ_0^2, σ_1^2 are positive.

The distributions $Q_n = N_n(\mathbf{1}_n, I_n + J_n/n)$ are finitely exchangeable for every n . But $Q_{n+1}(A \times \mathbb{R})$ is not equal to $Q_n(A)$, so the sequence is not self-consistent. Consistency is essential for inferential purposes if we wish to regard $\{Q_n\}$ as a stochastic model for a sequence of arbitrary length, or if we wish to predict Y_{n+1} or a longer-term average such as $(Y_{n+1} + \dots + Y_{n+m})/m$ given the observation $Y[n]$. An inconsistent sequence such as $N_n(\mathbf{1}_n, I_n + J_n/n)$ does not permit such activity.

The definition of a process does not require distributions to be defined on Borel subsets. It is possible to define a process that is incomplete in the sense that probabilities are defined only for a proper subset (a sub- σ -field) of Borel events $\mathcal{K}_n \subset \mathcal{B}(\mathbb{R}^n)$. In order for (P_n, \mathcal{K}_n) to be the marginal distribution of $(P_{n+1}, \mathcal{K}_{n+1})$, it is necessary that $A \in \mathcal{K}_n$ implies $A \times \mathbb{R} \in \mathcal{K}_{n+1}$. Otherwise, the consistency statement $P_n(A) = P_{n+1}(A \times \mathbb{R})$ is meaningless. For the process to be exchangeable, it must also be the case that $A \in \mathcal{K}_n$ implies $\sigma A \in \mathcal{K}_n$ for each integer n and each coordinate permutation $\sigma: \mathbb{R}^n \rightarrow \mathbb{R}^n$. Otherwise $P_n(A) = P_n(\sigma A)$ is meaningless.

A subset $A \subset \mathbb{R}^n$ is called translation-invariant if $A + \mathbf{1}_n = A$. In this setting $\mathbf{1}_n$ is the one-dimensional subspace of constant vectors. For example, the subset of \mathbb{R}^2 consisting of points (y_1, y_2) such that $-1 < y_1 - y_2 < 2$ is translation-invariant. The class of Borel events such that $A + \mathbf{1}_n = A$ is a sub- σ -field, sometimes denoted by $\mathcal{B}(\mathbb{R}^n/\mathbf{1}_n)$. Any Gaussian distribution on Borel events in \mathbb{R}^n can be restricted to $\mathcal{B}(\mathbb{R}^n/\mathbf{1}_n)$. Every Gaussian process can be restricted to the class of translation-invariant events, in which case the restricted process is called an incomplete Gaussian process.

The σ -field $\mathcal{B}(\mathbb{R}^n/\mathbf{1}_n)$ is closed under coordinate permutation in \mathbb{R}^n . If the original process on \mathbb{R}^n is exchangeable, so also is its restriction. However, the restricted process is not the same as the unrestricted process; conditional distributions for the restricted process are not the same as those for the original.

A translation-invariant subset $A \subset \mathbb{R}^n$ is called translation-scale invariant if $\lambda A = A$ for every $\lambda > 0$. For example, the subset of \mathbb{R}^3 such that $-1 < (y_1 - y_2)/(y_2 - y_3) < 2$ is translation-invariant and scale-invariant. The class of translation-scale invariant Borel events is a sub- σ -field $\mathcal{K}_n \subset \mathcal{B}(\mathbb{R}^n)$. Any Gaussian distribution on $\mathcal{B}(\mathbb{R}^n)$ can be restricted to \mathcal{K}_n . Every Gaussian process can be restricted to the class of translation-scale invariant events, in which case the restricted process is an incomplete Gaussian process. It assigns probabilities only to invariant events. If the original process is exchangeable, so also is its restriction.

The construction of Gosset's process begins fittingly with a sequence of independent random variables $\epsilon_1, \epsilon_2 \dots$ such that $\epsilon_1 = \pm 1$ with probability one half each, and $\epsilon_{n+1} \sim t_n$ is distributed as Student's t on $n \geq 1$ degrees of freedom. These are independent real-valued random variables with probability distributions on $\mathcal{B}(\mathbb{R})$. Initialization of the Gosset process commences with $Y_1 = 0$, $Y_2 = \epsilon_1$, followed thereafter by

$$Y_{n+1} = \bar{Y}_n + \alpha_n s_n \epsilon_n, \quad (14.5)$$

where \bar{Y}_n, s_n^2 are the sample mean and variance of the first n values, and $\alpha_n^2 = 1 + 1/n$. The Gosset process with parameter (μ, σ) differs only in the initialization: $Y_1 = \mu$; $Y_2 = \mu + \sigma \epsilon_1$.

As a process on Borel subsets, Gosset's process is certainly not exchangeable or Gaussian: Y_1 does not have the same distribution as Y_2 or Y_3 . Nor is it Gaussian because Y_2 has a two-point distribution, and Y_3 has infinite moments. As a process restricted to \mathcal{K} it is exchangeable: the restricted process states not only that each standardized ratio

$$\epsilon_n = \frac{Y_{n+1} - \bar{Y}_n}{s_n \sqrt{(1 + 1/n)}}$$

is distributed as t_{n-1} for $n \geq 2$, but also that the ratio is independent of $(Y[n], \mathcal{K}_n)$. In this form, it is not difficult to see that every exchangeable Gaussian process restricted to translation-scale invariant events coincides with Gosset's process on \mathcal{K} .

The lesson from the Gosset example is that a process defined on translation-scale invariant events has multiple extensions to a Borel process on \mathbb{R}^n . Every exchangeable Gaussian process is an extension of the restricted Gosset process. The sequential description (14.5) is a non-exchangeable non-Gaussian extension, an extension that is well-suited for fiduciary purposes that require parameter-free prediction. The fiducial extension implies, for example, that $\bar{Y}_\infty = \lim_{n \rightarrow \infty} \bar{Y}_n$ exists and that, its conditional distribution given $Y[n]$ is Student's t_{n-1} , centered at \bar{y}_n with scale parameter s_n/\sqrt{n} :

$$\bar{Y}_\infty \sim t_{n-1}(\bar{y}_n, s_n/\sqrt{n}).$$

There is a similar fiducial prediction for s_∞^2 , and for the limiting pair $(\bar{Y}_\infty, s_\infty^2)$.

14.5.2 Factual and Counterfactual Processes

In studies of causality and treatment effects, each unit in the population has one of k possibilities for treatment. A non-randomized design consists of a finite sample $U \subset \mathcal{U}$ together with a treatment assignment $U \xrightarrow{\mathbf{t}} [k]$, and $P_{\mathbf{t}}$ is the associated finite-dimensional response distribution. These distributional specifications—for different samples and alternative assignments—are assumed to be mutually consistent in the Kolmogorov sense, so they determine a stochastic process indexed by assignments \mathbf{t} . Mutual consistency does not imply independence for distinct units, but it does imply lack of interference, or the *stable unit-treatment distribution assumption* in Dawid (2021, Sect. 6.2). In most applications, the sample $U \subset \mathcal{U}$ is—or is regarded as—a fixed subset of the population, and treatment assignment is generated by randomization subject to design constraints. For example, all sites on one rat in Example 1 necessarily receive the same treatment.

The mathematical set-up for a counterfactual process is less complicated but more elaborate. First, the index set is extended to the Cartesian product $\mathcal{U} \times [k]$, consisting of all unit-treatment pairs, and the counterfactual process is envisaged as a function on this index set. Thus, $Y(u, r)$ is the response, or potential outcome, that would be observed if treatment r were assigned to unit u or patient u . To specify the counterfactual stochastic process, it is necessary to specify the joint distributions P_S^\dagger for each finite sample $S \subset \mathcal{U} \times [k]$ in a consistent manner. It suffices to specify the finite-dimensional distributions P_S^\dagger for finite product sets $S = U \times [k]$.

To each counterfactual sample S there corresponds a finite subset of units $U \subset \mathcal{U}$, which is obtained from S by ignoring the treatment component and eliminating duplicate units. For example,

$$S = \{(u_1, 0), (u_1, 2), (u_2, 1), (u_4, 0), (u_4, 1)\}$$

implies $U = \{u_1, u_2, u_4\}$. In general, $\#S \geq \#U$. A counterfactual sample that contains exactly one treatment level for each patient is called physical or realizable. In that case $\#S = \#U$, and treatment is a function, or assignment, $U \rightarrow [k]$. A sample that contains more than one treatment level for at least one patient is called metaphysical; it does not correspond to an assignment. The example displayed above is metaphysical. It contains two counterfactual pairs $(u_1, 0)$, $(u_1, 2)$, and $(u_4, 0)$, $(u_4, 1)$, so an observation on S contains two counterfactual replicates, $\{Y(u_1, 0), Y(u_1, 2)\}$ for patient u_1 and $\{Y(u_4, 0), Y(u_4, 1)\}$ for patient u_4 .

Although the counterfactual process is envisaged as a function on the product set $\mathcal{U} \times [k]$, it is not possible in practice to duplicate units or patients. Thus, each unit can be assigned only one treatment, so each observation on the process is confined to a realizable sample or a finite assignment. However, a counterfactual process does allow us to compute conditional or predictive distributions such as

$$P^\dagger(Y(u_1, 1) \in A \mid \text{data})$$

based on data from any sample, physical or metaphysical. For example the data might come from a physical design that includes $(u_1, 0)$ or a metaphysical design such as S that includes $(u_1, 0)$ and $(u_1, 2)$, in which case the conditioning event includes the value $Y(u_1, 0)$. In other words, although patient u was physically assigned at baseline to the control level, the counterfactual process allows us to compute the conditional distribution of the response for the same patient had he or she been assigned to the active treatment level at baseline. Mathematical duplication of patients enables us to evade the apparent contradiction.

To many authors, the flexibility afforded by the introduction of counterfactuals is embraced as a liberating experience and a cause for celebration. For example, Pearl and Mackenzie (2021) write on page 269–270,

It is impossible to overstate the importance of this development. It provided researchers with a flexible language to express almost every causal question they might wish to ask...

Indeed it does! However, that attitude is not universally embraced, and it is not regarded by this author as a conceptual advance or a liberating experience or a cause for celebration.

The situation in a nutshell is as follows. Let P^\dagger be the distribution of a counterfactual process with domain $\mathcal{U} \times [k]$, and let P be its restriction to assignments or physical designs. Equivalently, P^\dagger is a counterfactual extension of P . For any real design S corresponding to an assignment $U \xrightarrow{\mathbf{t}} [k]$, and any event $A \subset \mathbb{R}^S$, the probabilities are equal: $P^\dagger(A) = P_{\mathbf{t}}(A)$. In other words, the two processes are in agreement regarding the probability to be assigned to every observable event. Given that there is no disagreement about observables, their means, their variances, and so on, what is the explanation for the exuberance of the quote in the preceding paragraph? The answer, as Dawid (2000) argues, is that the counterfactual framework adds much to the vocabulary but brings nothing of substance to the conversation regarding observables.

The situation is entirely different when we address questions concerning counterfactuals proper. In that case, the conventional process has nothing whatever to say because the question is not mathematically legitimate: it is not within the scope of the distributions P_t . The good news is that the counterfactual extension provides an answer; the bad news is that every counterfactual extension provides a different answer. Moreover, there are infinitely many such extensions, all of which are indistinguishable on the basis of physical experiments. By definition, the extension is uncheckable.

If we restrict attention to what is, in principle, observable from experiment, counterfactual pairs such as $Y_{u,0}, Y_{u,1}$ do not arise, so a joint distribution is not needed. Nonetheless, the discussion of Dawid (2000) illustrates clearly the role of individual attitude in applied work. My own preference is to travel light and avoid unnecessary baggage. The issue is not one of fact versus fiction—whether counterfactual variables exist or do not exist—because that is settled by mathematical fiat. The issue is whether counterfactual processes in the form of a metaphysical extension to P^\dagger on $\mathcal{U} \times [k]$ can help to streamline statistical thinking, or whether they promote misleading lines of argument and false conclusions (Dawid, 2000).

Suppose that the sample consists of twelve units with treatment assignment $\{u_1 \mapsto 0, u_2 \mapsto 1, \dots, u_{12} \mapsto 1\}$, the model distribution is additive Gaussian with independent components for distinct units, and the observation vector is $y = (3.8, 7.4, \dots, 6.2)$. The question ‘What response would have been expected had unit u_1 been assigned to treatment 1 rather than treatment 0?’ is counterfactual because u_1 was actually assigned treatment 0. It can be answered only within the counterfactual framework by computing the conditional distribution given the data, which includes the value $Y(u_1, 0) = y_1$. In other words, the joint distribution of $Y(u_1, 0), Y(u_1, 1)$ is needed. Such counterfactual questions cannot be addressed within the classical framework indexed by assignments.

Some authors are content to talk the talk but not to walk the walk; they find comfort in the language of counterfactuals provided that attention is confined ultimately to realizable samples and the restricted distributions P_t . In good hands, such activity is safe and uncontroversial. It has one merit, namely that it illustrates how a treatment intervention is handled differently than a covariate in a statistical model. Beyond that, the discussion of Dawid (2000) shows no sign of consensus. My own attitude is to regard any substantive discussion of counterfactual effects such as differences $Y_{u,2} - Y_{u,1}$ or ratios $Y_{u,2}/Y_{u,1}$, whose distribution is determined by P^\dagger but not by P_t , as pointless, unnecessary, unverifiable and possibly misleading.

14.5.3 Limitations of Incomplete Processes

Mathematical experience and vocabulary gained from the manipulation of ordinary Borel distributions can be treacherous for incomplete processes. Consider, for example, the process whose distribution P_n is the restriction of $N_n(0, I_n)$ to

translation-invariant events $A \in \mathcal{B}(\mathbb{R}^n / \mathbf{1}_n)$. The process is exchangeable, so every sample of size n has the same distribution P_n . For $n = 1$, the subspace $\mathbf{1}_1 \subset \mathbb{R}$ coincides with the real numbers. It follows that $\mathcal{B}(\mathbb{R}/\mathbf{1}_1) = \{\emptyset, \mathbb{R}\}$ consists of two events only, for which the probabilities are zero and one respectively. In other words, the one-dimensional distributions are degenerate. For $n = 2$, the σ -field $\mathcal{B}(\mathbb{R}^n / \mathbf{1}_n)$ is generated by the transformation $(y_1, y_2) \mapsto y_1 - y_2$, and the distribution P_2 is fully specified by the statement $Y_2 - Y_1 \sim N_1(0, 2)$ on Borel subsets of \mathbb{R} . This means that the conditional distribution given Y_1 associates with each point y_1 a probability distribution on \mathbb{R}^2 such that $Y_2 - Y_1 \sim N_1(0, 2)$. In other words, the pair $Y[2]$ is independent of $Y[1]$ —as must be the case because of distributional degeneracy.

Fisher's fiducial version of the preceding inferential statement is ever-so-slightly different: Given $Y[1] = y_1$, the difference $Y_2 - y_1$ is distributed as $N_1(0, 2)$, which implies $Y_2 \sim N_1(y_1, 2)$. Of course, this cannot be correct because it implies that each Borel event $Y_2 \in A$ belongs to $\mathcal{B}(\mathbb{R}^2 / \mathbf{1}_n)$, which is false. Nonetheless, if $\epsilon_n \sim N(0, 1 + 1/n)$ are independent Gaussian variables, the predictive or fiducial sequence $Y_1 = \mu$ followed by

$$Y_{n+1} = \bar{Y}_n + \epsilon_n$$

is equivalent to the exchangeable Gaussian process when restricted to translation-invariant events. In other words, these are equivalent descriptions of one and the same incomplete process.

The main mathematical difficulty with fiducial statements stems from the fact that many apparently innocuous transformations such as $Y_n \mapsto Y_n^2$ or $Y[n] \mapsto \sum Y_i^2$ are not measurable in the sense that may be needed. A fiducial predictive statement such as $Y_{n+1} - \bar{Y}_n \mid Y[n] \sim N_1(0, 1 + 1/n)$ is correct conditionally as a statement about invariant events, but it has no direct implication for non-invariant events.

A similar issue arises in connection with missing values and treatment assignment. Suppose that the observation can be depicted schematically as a table indexed by n patients and k treatment levels:

| u | 1 | 2 | 3 | k |
|----------|----------|----------|----------|----------|
| u_1 | ? | 3.1 | ? | ? |
| u_2 | 2.7 | ? | ? | ? |
| u_3 | ? | ? | 4.5 | ? |
| \vdots | \vdots | \vdots | \vdots | \vdots |
| u_n | ? | ? | ? | 6.4 |

It is easy to be persuaded by the argument that one goal of inference—causal or otherwise—is to fill in the missing values. We would then know, either exactly or in distribution, what the response would have been for each patient had he or she been given a treatment other than the one actually assigned. Wouldn't that impress the medics and the folks back home! If the only tool available is the conventional

process P —with estimated parameters if needed—we must aim to compute the conditional distribution $P_{\mathbf{t}}(\cdot \mid \text{data})$ or its expected value or some such integral. However, the pattern of non-missing values tells us the realized assignment $\mathbf{t}: U \rightarrow [k]$, and $P_{\mathbf{t}}(\cdot)$ is informative only for pairs (u, r) such that $\mathbf{t}(u) = r$. The missing values are inaccessible. Too bad! This missing-value argument may be superficially persuasive, but it is, in my view, misleading and empty rhetoric.

No logical difficulty arises in using the conditional distribution $P^{\dagger}(\cdot \mid \text{data})$ associated with any counterfactual extension.

Note that it is possible to extend the sample from \mathbf{t} to \mathbf{t}' by including an extra-sample row

$$u_{n+1} \quad ? \quad ? \quad ? \quad ?$$

that is initially empty. The statement that \mathbf{t}' is an extension of \mathbf{t} means that $\mathbf{t}'(u) = \mathbf{t}(u)$ for all in-sample units. For an extension such that $\mathbf{t}'(u_{n+1}) = r$, we can compute the conditional distribution of Y_{n+1} using the conventional distribution $P_{\mathbf{t}'}$ associated with that extension. By considering various extensions, it is possible to complete all k entries in the additional row, one at a time, either by imputation or by conditional expectation. But, by definition, the so-called missing entries for in-sample patients are invisible to the conventional process.

14.6 Exercises

14.1 Let \mathbf{t} be the treatment assignment vector, and let $B_{\mathbf{t}}$ be the associated block factor, i.e., $B_{\mathbf{t}}(i, j) = 1$ if $t_i = t_j$ and zero otherwise. For $g \in \mathbb{R}$, consider the transformations

$$\Sigma \xrightarrow{g} \Sigma + g^2 B_{\mathbf{t}}$$

for Σ in the space of positive definite matrices. Discuss whether these transforms determine a group action or group homomorphism (preserving identity and composition). If not, is it a semi-group homomorphism in a suitable sense? Maybe after changing g^2 to e^g or $|g|$ to maintain positivity?

14.2 This exercise is concerned with a possible action of the additive group of real numbers on the space of positive definite matrices of order n . Let $\mathcal{X} \subset \mathbb{R}^n$ be a given subspace. To each Σ and $W = \Sigma^{-1}$ there corresponds a W -orthogonal projection P_W whose image is \mathcal{X} , and a complementary projection $Q_W = I - P_{\mathcal{X}}$. In matrix notation, $P_W = X(X'WX)^{-1}X'W$ depends on Σ . For $g \in \mathbb{R}$, show that the transformations

$$\Sigma \xrightarrow{g} Q_W \Sigma + e^g P_W \Sigma = \Sigma + (e^g - 1)X(X'\Sigma^{-1}X)^{-1}X'$$

determine a group homomorphism by linear transformations on the space of positive-definite matrices.

14.3 Let \mathbf{t} be the treatment assignment vector, and let P_W be the W -orthogonal projection onto the subspace $\text{span}(\mathbf{1}, \mathbf{t})$. Show that the transformation

$$N_n(\mu, \Sigma) \xrightarrow{g} N_n(\mu + g_0 \mathbf{t}, Q_{\mathcal{X}} \Sigma + e^{g_1} P_W \Sigma)$$

is an action of the additive group \mathbb{R}^2 on the space of Gaussian distributions. Describe the orbit of the distribution $N_n(\mathbf{1}, I_n)$.

14.4 Under what conditions does the treatment model in the preceding exercise satisfy the lack of interference condition?

14.5 Show that the log likelihood function for the simple linear regression model is

$$-n \log \sigma - \frac{1}{2} \sum (Y_u - \beta_0 - \beta_1 x_u)^2 / \sigma^2.$$

Deduce that the triple $(\sum Y_u, \sum x_u Y_u, \sum Y_u^2)$ is sufficient for the parameter. Under what conditions is this triple also minimal sufficient?

14.6 Show that the same triple is sufficient for the six-parameter random coefficient model (14.1) with one block. Deduce that the likelihood is maximized at the boundary point $\sigma_0 = \sigma_1 = 0$. Discuss the situation for two or more blocks.

14.7 Let Θ be the extended complex plane. For each $\theta = \theta_0 + i\theta_1$ let P_θ be the distribution on the extended real line with density

$$P_\theta(dy) = \frac{|\theta_1| dy}{\pi |y - \theta|^2}$$

for $\theta_1 \neq 0$, or the Dirac measure $\delta_\theta(dy)$ if $\theta_1 = 0$ or $\theta = \infty$. Each treatment effect is a non-singular 2×2 real matrix

$$g = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

acting on Θ as a fractional linear transformation

$$g\theta = \frac{a\theta + b}{c\theta + d}.$$

Show that the set of fractional linear transformations is a group, that $\theta \mapsto g\theta$ is a group action with two orbits in Θ , and that $P_{g\theta}$ is the Cauchy distribution with parameter $g\theta$.

14.8 Let $\Theta = \mathbb{R}^2$, and let P_θ be the von Mises-Fisher distribution on the unit circle with density

$$P_\theta(d\phi) = \frac{e^{\theta'y} d\phi}{I_0(|\theta|)},$$

where $y = (\cos \phi, \sin \phi)$ and $d\phi$ is arc length, and $I_0(\cdot)$ is the Bessel I function of order zero. Discuss the following groups as possible treatment effects acting on distributions: (i) the group of strictly positive numbers acting on Θ by scalar multiplication; (ii) the group of planar rotations; (iii) the group of similarity transformations generated by (i) and (ii).

14.9 Show that the random-coefficient model (14.2) is equi-variant under affine covariate transformation $x \mapsto g_0 + g_1 x$ with $g_1 \neq 0$. Show that the induced transformation on (β_0, β_1) is linear $\mathbb{R}^2 \rightarrow \mathbb{R}^2$. Show that the induced transformation on variance components is also linear

$$\begin{pmatrix} \sigma_0^2 \\ \rho\sigma_0\sigma_1 \\ \sigma_1^2 \end{pmatrix} \longmapsto \begin{pmatrix} 1 & 2g_0 & g_0^2 \\ 0 & g_1 & g_0g_1 \\ 0 & 0 & g_1^2 \end{pmatrix} \begin{pmatrix} \sigma_0^2 \\ \rho\sigma_0\sigma_1 \\ \sigma_1^2 \end{pmatrix}$$

14.10 Suppose that the observations Y_1, \dots, Y_n in a two-arm randomized design are independent bivariate Gaussian with mean vector (μ_1, μ_2) for units in the control arm, and

$$E(Y_i | \mathbf{t}) = \begin{pmatrix} \mu_1 \cos \tau - \mu_2 \sin \tau \\ \mu_1 \sin \tau + \mu_2 \cos \tau \end{pmatrix}$$

for units in the active treatment arm. The covariance in both cases is $\sigma^2 I_2$; the parameter $\mu \in \mathbb{R}^2$ is unrestricted, while $\sigma > 0$ and the treatment effect lies in $0 \leq \tau < 2\pi$. By expressing the sample averages for each treatment arm as complex numbers, show that $\hat{\tau} = \arg(\bar{Y}_0) - \arg(\bar{Y}_1)$ is the maximum-likelihood estimate of the treatment effect. Find the maximum-likelihood estimate of σ^2 .

14.11 A modification of the preceding model retains the mean vectors, while the covariance matrix is unrestricted but constant over units. Find an expression for the maximum-likelihood estimate of the treatment effect.

14.12 Show that the treatment effect in both preceding exercises is a group action on Gaussian distributions. What is the group, and how does it act? In only one case is the action on distributions induced by an action on the sample space. Explain.

14.13 In Exercise 14.10, the null hypothesis of no treatment effect $H_0: \tau = 0$ is the left endpoint of the parameter interval $\tau \in [0, 2\pi)$. Explain why this is not a boundary point in the parameter space.

14.14 Find the survival distribution P associated with the hazard measure

$$\theta(dt) = \begin{cases} dt/(1-t) & 0 < t < 1; \\ dt/t & t \geq 1. \end{cases}$$

Hence find a second hazard measure that has the same survival distribution.

14.15 Let $\Theta_0 = (0, 1)$ and let P_θ be the iid Bernoulli model $\text{Ber}(\theta)$. Let \mathcal{G} be the additive group of addition modulo one, so that $P_{g\theta} = \text{Ber}(\theta + g)$ is the Bernoulli model with parameter $\theta + g$ modulo one. Explain why $\text{Ber}(\theta) \mapsto \text{Ber}(\theta + g)$ is not a group action on distributions in the sense of Sect. 14.3.2.

Chapter 15

Gaussian Distributions



15.1 Real Gaussian Distribution

15.1.1 Density and Moments

The standard Gaussian distribution has a density

$$\Phi(dy) = \phi(y) dy = \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy$$

with respect to Lebesgue measure on the real line. It is symmetric with finite moments of all orders. The moment generating function is

$$M_0(t) = \int e^{ty} \phi(y) dy = e^{t^2/2} = \sum_{r=0}^{\infty} \mu_r t^r / r!,$$

from which the odd moments are zero, and the even moments are

$$\mu_{2r} = \frac{(2r)!}{2^r r!} = 1 \cdot 3 \cdots (2r - 1).$$

The cumulant generating function is

$$K_0(t) = \log M_0(t) = \sum \kappa_r t^r / r! = t^2 / 2,$$

so all of the cumulants are zero except for the variance $\kappa_2 = 1$.

If ε is a standard normal variable, and (μ, σ) is any pair of real numbers with $\sigma > 0$, the affine transformation $Y = \mu + \sigma \varepsilon$ is distributed according to the Gaussian

distribution with mean μ and variance σ^2 . The density function of the transformed variable at y is

$$\frac{1}{\sigma} \phi\left(\frac{y - \mu}{\sigma}\right) = \frac{1}{\sigma \sqrt{2\pi}} e^{-(y-\mu)^2/(2\sigma^2)}.$$

The moment and cumulant generating functions are

$$\begin{aligned} M_{\mu,\sigma}(t) &= E(e^{tY}) = e^{t\mu} E(e^{t\sigma\varepsilon}) = e^{t\mu + t^2\sigma^2/2} \\ K_{\mu,\sigma}(t) &= \log M_{\mu,\sigma}(t) = t\mu + t^2\sigma^2/2. \end{aligned}$$

The mean is μ , the variance is $\kappa_2 = \sigma^2$, and all other cumulants are zero.

For $x > 0$, the ratio of the right tail probability $1 - \Phi(x)$ to the density $\phi(x)$ is called Mills's ratio. The asymptotic expansion is

$$\frac{1 - \Phi(x)}{\phi(x)} = \frac{1}{x} - \frac{1}{x^3} + O(x^{-5}).$$

This stands in sharp contrast with heavy-tailed distributions for which the corresponding ratio is increasing in x ; in the case of the Cauchy distribution the ratio is asymptotically linear in x .

The approximate inverse relationship for the Gaussian quantile in terms of the right rail probability p is

$$x \simeq \sqrt{-2 \log p - \log(2\pi) - 2 \log x}.$$

For small p , this can be solved recursively starting from $x = 1$.

15.1.2 Gaussian Distribution on \mathbb{R}^n

Let $X = (X_1, \dots, X_n)$ be a random vector in \mathbb{R}^n whose components are independent and identically distributed $N(0, 1)$ variables. Independence implies that the joint density function with respect to Lebesgue measure at $x \in \mathbb{R}^n$ is the product of the marginal density functions, which is

$$\Phi_n(dx) = \phi_n(x) dx = (2\pi)^{-n/2} e^{-\|x\|^2/2} dx,$$

where $\|x\|^2 = x_1^2 + \dots + x_n^2$ is the standard Euclidean squared norm.

This is called the standard normal distribution on \mathbb{R}^n , and is denoted by $N_n(0, I_n)$. The joint moment generating function is the product of the marginal generating functions

$$M_0(t) = \int_{\mathbb{R}^n} e^{t_1 x_1 + \dots + t_n x_n} \phi_n(x) dx = e^{\|t\|^2/2},$$

and the cumulant generating function is $\|t\|^2/2$, which is quadratic and radially symmetric as a function of t . All of the joint cumulants are zero except for the variances, which are $\text{cov}(X_i, X_j) = \delta_{ij}$, i.e., one for $i = j$ and zero otherwise.

Let L be a linear transformation $\mathbb{R}^n \rightarrow \mathbb{R}^n$, so that the matrix L is of order $n \times n$. The moment generating function of the transformed variable $Y = LX$ is

$$E(e^{t'Y}) = \int_{\mathbb{R}^n} e^{t'Lx} \phi_n(x) dx = M_0(L't) = e^{\|L't\|^2/2},$$

so the cumulant generating function $\|L't\|^2/2 = t'LL't/2$ is quadratic in t but not radially symmetric. All of the cumulants are zero except for the variances and covariances, which are the components of the matrix

$$\Sigma = \text{cov}(Y) = \text{cov}(LX) = LL',$$

which is symmetric and positive semi-definite. The random variable Y has the normal distribution in \mathbb{R}^n with mean zero and covariance Σ , which is denoted by $N_n(0, \Sigma)$.

If L is invertible, the covariance matrix $\Sigma = LL'$ is also invertible with inverse $W = L'^{-1}L^{-1}$. In that case, the Jacobian of the transformation is the absolute value of the determinant of the transform matrix

$$dy = |\det(L)| dx = \det^{1/2}(\Sigma) dx.$$

The joint density of the transformed variable is

$$(2\pi)^{-n/2} |W|^{1/2} e^{-y'Wy/2} dy, \quad (15.1)$$

which is the density at y of the Gaussian distribution $N_n(0, \Sigma)$.

In general, the linear transformation L need not be invertible. In that case the subspaces $\text{Im}(L) = \text{Im}(\Sigma)$ and $\ker(L') = \ker(\Sigma)$ are complementary of dimensions $n - k$ and k respectively, and are also orthogonal with respect to the standard inner product in \mathbb{R}^n . With probability one, $Y = LX$ belongs to $\text{Im}(\Sigma)$, so the distribution $N_n(0, \Sigma)$ necessarily puts mass one on this subspace. If $k > 0$, the distribution $N_n(0, \Sigma)$ is singular and does not have a density with respect to Lebesgue measure on \mathbb{R}^n .

The translation $Y \mapsto Y + \mu$ sends the distribution $N_n(0, \Sigma)$ to $N_n(\mu, \Sigma)$, which is supported on the coset, or affine subspace, $\mu + \text{Im}(\Sigma)$. The cumulant generating

function is $t'\mu + t'\Sigma t/2$, so the mean vector is μ and the covariance matrix is Σ . If Σ is invertible, then $\text{Im}(\Sigma) = \mathbb{R}^n$, and the distribution has a density

$$(2\pi)^{-n/2} |W|^{1/2} e^{-(y-\mu)'W(y-\mu)/2} dy. \quad (15.2)$$

15.2 Complex Gaussian Distribution

15.2.1 One-Dimensional Distribution

The one-dimensional Gaussian distribution on the complex plane is nothing more than a two-dimensional Gaussian distribution on \mathbb{R}^2 that is also rotationally symmetric. The zero-mean complex Gaussian distribution with variance σ^2 has a density

$$\phi(z) = \frac{e^{-|z|^2/\sigma^2}}{\pi\sigma^2}$$

with respect to two-dimensional Lebesgue measure. The real part and the imaginary part of $Z \sim \mathbb{C}N(0, 1)$ are independent zero-mean real Gaussian variables with variance $\sigma^2/2$ each. The argument of Z is uniformly distributed on $[0, 2\pi)$, and independent of $|Z|^2$, which is exponentially distributed with mean σ^2 .

Rotational symmetry means that, for every real θ , the rotated variables $Ze^{i\theta}$ have the same distribution as Z . It follows that $Z^k e^{ki\theta} \sim Z^k$ for every integer k . Provided that the moments are finite, $\mu_k e^{ki\theta} = \mu_k$ implies that complex powers satisfy $E(Z^k) = 0 = E(\bar{Z}^k)$ for every integer $k \geq 1$. The only non-zero integer moments are $E(|Z|^{2k}) = k! \sigma^{2k}$ in which Z and \bar{Z} occur an equal number of times in the product. The k th order cumulant is $\text{cum}_k(|Z|^2) = (k-1)! \sigma^{2k}$.

15.2.2 Gaussian Distribution on \mathbb{C}^n

Let $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)$ be independent and identically distributed $\mathbb{C}N(0, 1)$ random variables, so that the joint density is

$$\pi^{-n} \prod_{r=1}^n e^{-|\varepsilon_r|^2} = \pi^{-n} e^{-\varepsilon^* \varepsilon} = \pi^{-n} e^{\|\varepsilon\|^2}$$

with respect to $2n$ -dimensional Lebesgue measure. Here ε^* is the transpose of the conjugate vector. Let $Z = L\varepsilon$, where L is a full-rank complex matrix of order n , and let $\Sigma = LL^*$ be positive-definite Hermitian. The derivative matrix of the linear

transformation $L: \mathbb{C}^n \rightarrow \mathbb{C}^n$ is L , but the Jacobian of the linear transformation $\mathbb{R}^{2n} \rightarrow \mathbb{R}^{2n}$ is $\det(LL^*) = \det(\Sigma)$. Thus, the density of the transformed vector is

$$\pi^{-n} \det(\Sigma)^{-1} e^{-z^* \Sigma^{-1} z}$$

with respect to $2n$ -dimensional Lebesgue measure at $z \in \mathbb{C}^n$. This distribution is denoted by $Z \sim \mathbb{C}N_n(0, \Sigma)$, where $E(ZZ^*) = E(L\epsilon\epsilon^*L^*) = LL^* = \Sigma$.

As a reminder, Hermitian symmetry means that the real part of Σ is symmetric, and the imaginary part is anti-symmetric or skew-symmetric. Thus the conjugate is equal to the transpose $\bar{\Sigma} = \Sigma'$, while $\bar{\Sigma}' = \Sigma^* = \Sigma$. Strict positive definiteness means that every Hermitian quadratic form $\xi^* \Sigma \xi$ in complex vectors is strictly positive unless $\xi = 0$. If Σ is strictly positive definite, so also is the complex conjugate matrix $\bar{\Sigma}$, and the real part $\Re(\Sigma)$.

The conjugate vector is distributed as $\mathbb{C}N(0, \bar{\Sigma})$, and the unit complex multiple $e^{i\theta} Z$ has the same distribution as Z . The one-dimensional marginal distribution of Z_1 is complex Gaussian with variance Σ_{11} , and the marginal distribution of Z_{i_1}, \dots, Z_{i_k} is complex Gaussian with covariance $\Sigma[\mathbf{i}, \mathbf{i}]$ restricted to rows $\mathbf{i} = \{i_1, \dots, i_k\}$, and the same columns. Note that the restriction is applied to the rows and columns of Σ , not to the rows and columns of the precision matrix Σ^{-1} .

Exercises 15.1–15.3 show that the Hermitian matrix $\Sigma = \Sigma_0 + i\Sigma_1$ can be associated with a $2n \times 2n$ real symmetric matrix in such a way that the pair of real vectors $\Re(Z), \Im(Z)$ is jointly Gaussian with covariance

$$\text{cov} \begin{pmatrix} \Re(Z) \\ \Im(Z) \end{pmatrix} = \begin{pmatrix} \Sigma_0 & \Sigma_1 \\ \Sigma_1' & \Sigma_0 \end{pmatrix} / 2,$$

where $\Sigma_1' = -\Sigma_1$. Any pair of identically distributed real Gaussian vectors X, Y defines a complex Gaussian vector $Z = X + iY$ if and only if the cross-covariances are anti-symmetric, $\text{cov}(X, Y) = -\text{cov}(Y, X)$.

15.2.3 Moments

Rotational symmetry with respect to complex unit scalar multiplication means that $E(Z_r) = E(e^{i\theta} Z_r)$ is necessarily zero. Likewise the product moment $E(Z_r Z_s) = E(e^{2i\theta} Z_r Z_s)$ is also zero. The only non-zero second-order moments are $\text{cov}(Z_r, \bar{Z}_s) = \Sigma_{rs}$. More generally, the only non-zero moments of degree $2k$ are those in which conjugated and non-conjugated components occur in equal number, such as the product $Z_{i_1} \cdots Z_{i_k} \bar{Z}_{j_1} \cdots \bar{Z}_{j_k}$.

The evaluation of Gaussian moments is a classical problem dating back to Isserlis (1918), in the case of real vectors. In the case of complex Gaussian vectors, the product moment is related to Wick's theorem (Wick, 1950), to Boson point processes McCullagh and Møller (2006), and to Feynman diagrams. The complex

case is a little simpler than the real case, and the product moment is as follows. To each permutation $\pi : [k] \rightarrow [k]$ there corresponds a 1–1 matching $i_r \mapsto j_{\pi(r)}$ of conjugated with non-conjugated components. Each matching gives rise to a product of k covariances

$$E(Z_{i_1} \cdots Z_{i_k} \bar{Z}_{j_1} \cdots \bar{Z}_{j_k}) = \sum_{\pi} \prod_{r=1}^k \Sigma_{i_r, j_{\pi(r)}} = \text{per}(\Sigma[\mathbf{i}, \mathbf{j}]),$$

which is the permanent of the $\mathbf{i} \times \mathbf{j}$ sub-matrix. Note that rows or columns may be repeated, so that $E(|Z_1|^{2k}) = \Sigma_{11}^k k!$, which are the moments of the exponential distribution. The permanent is the same as the determinant except that all $k!$ terms in the permutation expansion have coefficient +1.

Complex-valued random variables seldom occur in experimental research except in the setting of Fourier transformation for time series, as in Chap. 7. They are not used in the remainder of this chapter, but they do also arise in connection with stationary Gaussian processes, particularly space-time processes in Chap. 16.

15.3 Gaussian Hilbert Space

15.3.1 Euclidean Structure

It is often convenient to associate with the Gaussian distribution $N_n(0, \Sigma)$ or $N_n(\mu, \Sigma)$ a vector space having very specific geometric properties that match the second moments of the distribution. In doing so, the mean vector is ignored, so ‘second moments’ refers to variances and covariances. For simplicity, we assume that $\Sigma = W^{-1}$ is invertible, so the domain or support of the distribution is the entire vector space \mathbb{R}^n . The Euclidean geometric properties (length, angle, orthogonality,...) are generated by the specific inner product $\langle x, y \rangle = \sum w_{ij} x_i y_j$ matching the norm in the exponent of the density (15.1) or (15.2). This inner-product space $\mathcal{H} = (\mathbb{R}^n, \langle \cdot, \cdot \rangle)$ is called the Gaussian Hilbert space. Apart from minor modifications of notation, the algebra for complex Gaussian spaces is essentially the same as that for real vector spaces, so the notation here uses real vector spaces.

The geometry associated with \mathcal{H} is a special case of the geometry associated with the Rao-Fisher-information metric generated by a parametric model. In particular, the linear model $Y \sim N_n(X\beta, \Sigma)$ determines a subspace $\mathcal{X} \subset \mathcal{H}$, and a linear transformation $Y \mapsto \hat{\beta}$ that sends Y to the weighted least-squares coefficient vector $\hat{\beta} = (X'WX)^{-1}X'WY$ in \mathbb{R}^p . This linear transformation also sends the distribution $N_n(X\beta, \Sigma)$ on \mathbb{R}^n to $N_p(\beta, (X'WX)^{-1})$, so it is a transformation $\mathcal{H} \rightarrow \mathcal{H}_p$, where \mathcal{H}_p is p -dimensional Euclidean space with inner product matrix $X'WX$. The transformation $\mathcal{H}_p \rightarrow \mathcal{X} \subset \mathcal{H}$ that sends $\hat{\beta}$ to $\hat{\mu} = X\hat{\beta}$ is in fact a Euclidean isometry, or Hilbert-space isometry, so the geometric and distributional properties of $\hat{\beta} \in \mathcal{H}_p$ mirror exactly those of the orthogonal projection $\hat{\mu} = PY \in \mathcal{X} \subset \mathcal{H}$.

The main reasons for endowing the domain with Euclidean structure are as follows:

1. Orthogonality of subspaces is associated with independence of random variables;
2. The orthogonal projection having a given image is associated with maximum-likelihood and weighted least squares;
3. The orthogonal projection having a given kernel is associated with a number of statistically distinct operations such as least-squares residual, prediction, interpolation, smoothing and Kriging;
4. Cochran's theorem and much of the distribution-theory associated with linear regression and analysis of variance become more transparent.

15.3.2 Cautionary Remarks

From the vantage of linear algebra, it is natural to specify the inner product directly through the inner-product matrix W , which is symmetric and strictly positive definite. The inner product in the dual space of linear functionals is the matrix inverse, $\Sigma = W^{-1}$.

The order of operations in statistical work is ordinarily reversed. A Gaussian process is defined by its covariance function, which is naturally subject to restrictions such as stationarity, isotropy, or exchangeability, depending on the structure of its domain. Consequently, the matrix Σ , which is the restriction of the covariance function to the sample points, is specified first. The inverse matrix then determines the inner product in the observation space \mathcal{H} for the particular sample selected.

For a process sampled at points u_1, \dots, u_n in some domain \mathcal{U} , the matrix component $\Sigma_{ij} = \text{cov}(Y(u_i), Y(u_j))$ depends on u_i, u_j only, and is independent of the configuration of the remaining sample points. By contrast w_{ij} depends on the entire configuration of sampled points. For example, if the process is stationary on the plane, then $u_i - u_j = u_{i'} - u_{j'}$ implies $\Sigma_{ij} = \Sigma_{i'j'}$. But equal separation in the domain does not imply $w_{ij} = w_{i'j'}$.

Despite the substantial advantages listed in the preceding section, it is good to be aware of one additional limitation of associating a specific geometry with the Gaussian distribution. In statistical work it is often necessary to compare two candidate distributions on the same observation space, for example by computing the likelihood ratio. For example, the candidate distributions might be $N_n(0, \Sigma_0)$ and $N_n(0, \Sigma_1)$ for two given matrices. To compute a likelihood ratio, it is essential to compare candidate distributions on the same space, so it could be a serious mistake to associate with each distribution its own geometry.

15.3.3 Projections

Specification by Image

Let X be any matrix of order $n \times p$ whose columns span the real subspace $\mathcal{X} \subset \mathcal{H}$ of dimension p . The transformation $P: \mathcal{H} \rightarrow \mathcal{H}$ whose matrix representation is

$$P = X(X'WX)^{-1}X'W \quad (15.3)$$

has the following properties.

1. $P^2 = P$;
2. For each $x \in \mathcal{H}$, Px belongs to \mathcal{X} ;
3. For each $x \in \mathcal{X}$, $Px = x$;
4. For each $x, y \in \mathcal{H}$, $\langle x, Py \rangle = \langle Px, y \rangle$.

The first of these, called idempotence, is the definition of a projection, linear or otherwise. The second says that the image of P is a subspace of \mathcal{X} . The third says that P acts as the identity on \mathcal{X} , so $\text{Im}(P)$ contains \mathcal{X} ; the second and third together imply $\text{Im}(P) = \mathcal{X}$. The fourth is the self-adjointness condition, which implies that $\text{Im}(P)$ and $\ker(P)$ are orthogonal subspaces in \mathcal{H} . It follows that P is the orthogonal projection $\mathcal{H} \rightarrow \mathcal{H}$ whose image is \mathcal{X} , and the complementary transformation $Q = I_n - P$ is the orthogonal projection whose kernel is \mathcal{X} .

Properties 1–3 hold for any strictly positive definite matrix W whether or not it coincides with the inner product in \mathcal{H} . In particular, $P_0 = X(X'X)^{-1}X'$ is a projection with image \mathcal{X} , and $Q_0 = I_n - P_0$ is the complementary projection with kernel \mathcal{X} , but neither projection is orthogonal in \mathcal{H} unless $W \propto I_n$.

If L is $p \times p$ of full rank, then the matrices X and XL span the same space. If we replace X with XL in the definition of P , we obtain the same projection; likewise for P_0 . In other words, P and P_0 are independent of the vectors selected to span the image subspace.

These are the most familiar versions of projection matrices that arise in statistical work, where the projection is usually targeted to have a particular image. But it is occasionally convenient to specify a projection directly by a linear transformation matrix having the desired kernel.

Specification by Kernel

Let $\mathcal{K} \subset \mathcal{H}$ be a given subspace of dimension k , and let $K: \mathcal{H} \rightarrow \mathbb{R}^{n-p}$ be any matrix of order $n - k \times n$ with kernel \mathcal{K} . Then the matrix

$$Q^\dagger = \Sigma K'(K\Sigma K')^{-1}K \quad (15.4)$$

satisfies $Q^\dagger Q^\dagger = Q^\dagger$, so Q^\dagger is a projection $\mathcal{H} \rightarrow \mathcal{H}$. It is easily verified that $\ker(Q^\dagger) = \mathcal{K}$. Symmetry of WQ^\dagger implies self-adjointness, so Q^\dagger is the orthogonal projection with kernel \mathcal{K} . In particular, if we choose the matrix K so that $\mathcal{K} = \mathcal{X}$, uniqueness implies that Q^\dagger coincides with $Q = I_n - P$ as defined in (15.3). This identity is by no means obvious from the matrix algebra alone.

Self-adjointness Identity

Let P be any orthogonal projection $\mathcal{H} \rightarrow \mathcal{H}$ such as (15.3) or (15.4). Idempotence implies $P^2 = P$, and self-adjointness implies that $WP = P'W$ is a symmetric matrix. It follows that $P'WP = WP = P'W$ is symmetric and positive semi-definite.

Mixed Products

By definition, two projections such that $\text{Im}(P_0) \subseteq \text{Im}(P_1)$ satisfy $P_1 P_0 = P_0$. If both projections have the same image, then

$$P_1 P_0 = P_0; \quad P_0 P_1 = P_1;$$

i.e., the first or rightmost projection prevails in the product. Mixed products having nested kernels exhibit the opposite behaviour; $\ker(Q_0) \subseteq \ker(Q_1)$ implies $Q_1 Q_0 = Q_1$ in which the last, or leftmost, projection prevails.

In statistical work related to linear models, two linear transformations T, T' having the same kernel are statistically equivalent in the sense that there exists linear transformations L, L' such that $T = LT'$ and $T' = L'T$. One can be obtained from the other by a linear transformation; for projections, $L = T$ and $L' = T'$.

Trace and Rank

Let P be any linear projection, not necessarily an orthogonal projection $\mathcal{H} \rightarrow \mathcal{H}$. The idempotence condition $P^2 = P$ means that the eigenvalues of P satisfy $\lambda^2 = \lambda$, which implies that λ is either zero or one. Consequently, the trace of P , which is the sum of the eigenvalues, is equal to the rank of P or the dimension of the image space.

Rank Degeneracy

Suppose that $Y \sim N_n(0, \Sigma)$, where Σ has rank $n - p$. Let $K: \mathbb{R}^n \rightarrow \mathbb{R}^{n-p}$ be any linear transformation whose kernel coincides with the kernel of Σ , i.e., $\ker(K) = \ker(\Sigma) = \mathcal{X}$. This means that K is a matrix of order $n - p \times n$ and rank $n - p$.

With respect to the standard inner product in \mathbb{R}^n , $\text{Im}(K')$ is complementary and orthogonal to $\ker(K)$. Symmetry of Σ implies $\text{Im}(K') = \text{Im}(\Sigma)$.

The relevant Gaussian Hilbert space associated with $N_n(0, \Sigma)$ is either the subspace $\text{Im}(\Sigma)$, or the quotient space \mathbb{R}^n/\mathcal{X} . In either case, the dimension is $n - p$. For either representation of \mathcal{H} , the inner product is a positive semi-definite quadratic form $\langle x, y \rangle = x'Wy$ in for $x, y \in \mathbb{R}^n$, where W is the $n \times n$ symmetric matrix

$$W = K'(K\Sigma K')^{-1}K, \quad (15.5)$$

which has the same image and kernel as Σ . Evidently, $W\Sigma K' = K'$ so $W\Sigma$ is the identity on $\text{Im}(K') = \text{Im}(\Sigma)$, and $W\Sigma$ is zero on $\ker(\Sigma)$, so $W\Sigma$ is a projection. Symmetry of $W\Sigma W = W$ implies that $W\Sigma$ is self-adjoint, so $W\Sigma$ is the orthogonal projection $\mathcal{H} \rightarrow \mathcal{H}$ whose image is $\text{Im}(\Sigma)$, i.e., $W\Sigma$ is the identity $\mathcal{H} \rightarrow \mathcal{H}$.

The preceding algebra has another interpretation that is related to incomplete Gaussian distributions in which $N_n(0, \Sigma; \mathcal{X})$ is a Gaussian distribution on the quotient space \mathbb{R}^n/\mathcal{X} . In other words, $N_n(0, \Sigma; \mathcal{X})$ is the restriction of $N_n(0, \Sigma)$ to Borel subsets $A \subset \mathbb{R}^n$ such that $A + \mathcal{X} = A$. In this situation, it is necessary only that Σ be positive definite on \mathcal{X} -contrasts, which means that $K\Sigma K'$ is strictly positive definite. Two covariance matrices such that $K\Sigma_1 K' = K\Sigma_2 K'$ are equivalent on \mathcal{X} -contrasts, and determine the same distribution on \mathbb{R}^n/\mathcal{X} . In that case, the matrix W in (15.5) serves as the inner product in \mathcal{H} .

For a simple example of the latter, floating Brownian motion is a generalized Gaussian process on the real line whose covariance function is $-|u - u'|$. For any collection of points u_1, \dots, u_n in \mathbb{R} , the matrix whose components are $\Sigma_{ij} = -|u_i - u_j|$ is symmetric but clearly not positive definite or even semi-definite. However, if we take $\mathcal{X} = \mathbf{1}$, the subspace of constant functions, and $\ker(K) = \mathbf{1}$, it can be shown that $K\Sigma K'$ is positive definite. We say that $-|u - u'|$ is *positive-definite on simple contrasts*.

The Dirac difference measure $\delta_u(\cdot) - \delta_{u'}(\cdot)$ is an example of an elementary contrast, and the process takes a value $Y(\delta_u - \delta_{u'})$, conventionally written as an increment $Y(u) - Y(u')$, which is distributed as Gaussian with variance

$$(1, -1) \begin{pmatrix} 0 & -|u - u'| \\ -|u - u'| & 0 \end{pmatrix} \begin{pmatrix} 1 \\ -1 \end{pmatrix} = 2|u - u'|.$$

Despite the notation, $Y(u)$ or $Y(\delta_u)$ in isolation is not a Gaussian variable with finite variance. Provided that the phrase is understood informally as a limit, it is seldom misleading to regard $Y(u)$ as Gaussian with ‘infinite’ variance. Floating Brownian motion is not defined pointwise, but it is stationary on its domain of contrasts. Standard Brownian motion

$$B(u) = Y(u) - Y(0) \sim N(0, 2|u|)$$

is defined pointwise for $u \in \mathbb{R}$, but is not stationary. Realizations of either process are everywhere continuous but nowhere differentiable.

15.3.4 Dual Space of Linear Combinations

Let $Y = (Y_1, \dots, Y_n)$ be a random vector distributed as $N_n(0, \Sigma)$ on \mathbb{R}^n , where Σ is invertible. To each coefficient vector $\alpha = (\alpha_1, \dots, \alpha_n)$ there corresponds a linear combination

$$Y(\alpha) = \alpha_1 Y_1 + \cdots + \alpha_n Y_n.$$

Instead of indexing Y by the points $i \in [n]$, the preceding notation suggests that we use the space of linear combinations as an extended index set. Strictly speaking, this extension is unnecessary and superfluous. As a linear functional, the extension $Y(3\alpha + 4\beta) = 3Y(\alpha) + 4Y(\beta)$ is linear and additive, so all values are determined by the values on any basis.

The covariance of two linear combinations is bilinear:

$$\text{cov}(Y(\alpha), Y(\beta)) = \langle \alpha, \beta \rangle = \sum \alpha_i \beta_j \Sigma_{ij}.$$

The Hilbert space \mathcal{H}^* consisting of coefficient vectors, or linear functionals, with this inner product is the dual of \mathcal{H} . By definition, it is restricted to coefficient vectors α such that the linear combination $Y(\alpha)$ has finite variance $\|\alpha\|^2 < \infty$. The dual space arises most prominently in problems of prediction and computation of conditional distributions for spatial and temporal processes.

An observation on the process consists of a finite sample $\{x_1, \dots, x_n\}$ of sites plus the site values $Y(x_1), \dots, Y(x_n)$. The observation values serve as basis elements in the observation space \mathcal{H} , while the sample points serve as basis elements for the dual space of linear combinations $\alpha \in \mathcal{H}_0^*$. As a process, the space of samples is embedded in a larger Hilbert space $\mathcal{H}^* \supset \mathcal{H}_0^*$ associated with extended samples. For any $\beta \in \mathcal{H}^*$, the conditional distribution of $Y(\beta)$ given the sample values $\{Y(\alpha) : \alpha \in \mathcal{H}_0^*\}$ is Gaussian with moments

$$Y(\beta) \mid Y[\mathcal{H}_0^*] \sim N(Y(P\beta), \|Q\beta\|) \tag{15.6}$$

where P is the orthogonal projection $\mathcal{H}^* \rightarrow \mathcal{H}^*$ with image \mathcal{H}_0^* , and Q is the complementary projection.

15.4 Statistical Interpretations

15.4.1 Canonical Norm

In this section, \mathcal{H} is the Hilbert space associated with the distribution $N_n(0, \Sigma)$. For simplicity of exposition, $W = \Sigma^{-1}$ is invertible and $\dim(\mathcal{H}) = n$.

The squared norm of a vector $x \in \mathcal{H}$ is $\|x\|^2 = x'Wx$. For $Y \sim N(0, \Sigma)$, the distribution of the scalar random variable $\|Y\|^2$, can be obtained from its moment generating function

$$\begin{aligned} E(e^{t\|Y\|^2}) &= (2\pi)^{-n/2}|W|^{1/2} \int_{\mathcal{H}} e^{ty'Wy - y'Wy/2} dy \\ &= (2\pi)^{-n/2}|W|^{1/2} \int_{\mathcal{H}} e^{(1-2t)\|y\|^2/2} dy = (1-2t)^{-n/2} \end{aligned}$$

provided that $t < 1/2$. The moment generating function of the χ_1^2 -distribution is $(1-2t)^{-1/2}$, so $Y'WY$ is distributed as χ_n^2 , which is the distribution of the sum $Z_1^2 + \dots + Z_n^2$ of squares of n independent standard Gaussian variables.

The χ^2 density function is available in closed form, but is not especially important for either theory or applications. The cumulant generating function $-n \log(1-2t)/2$ implies that the r th cumulant is $\kappa_r = n(r-1)!2^{r-1}$. All cumulants are proportional to n , the mean and variance are n and $2n$, and the central limit theorem implies $\chi_n^2 \simeq N(n, 2n)$ for large n . For numerical work, the cumulative distribution function is available in R using the syntax `pchisq(x, df=n)`, and simulated variables are available using `rchisq(..., df=n)`.

15.4.2 Independence

Two orthogonal projections $P, Q: \mathcal{H} \rightarrow \mathcal{H}$ are said to be *mutually orthogonal* if $PQ = QP = 0$, so the projections (15.3) and (15.4) are both complementary and mutually orthogonal. Orthogonality of projections implies that the random vectors PY, QY are independent. This can be verified directly from the matrix forms (15.3) or (15.4), which satisfy

$$P\Sigma P' = P\Sigma; \quad Q\Sigma Q' = Q\Sigma; \quad \text{and} \quad P\Sigma Q' = P Q \Sigma = 0.$$

More directly, the joint moment generating function

$$E(e^{t'_1 PY + t'_2 QY}) = e^{(t'_1 P + t'_2 Q)\Sigma(P't_1 + Q't_2)/2} = e^{t'_1 P\Sigma P't_1 + t'_2 Q\Sigma Q't_2}$$

is the product of the marginal generating functions.

Cochran's Theorem

Let P_1, \dots, P_k be orthogonal projections $\mathcal{H} \rightarrow \mathcal{H}$ that are (i) mutually orthogonal in the sense $P_r P_s = 0$ for $r \neq s$, and (ii) complementary in the sense $P_1 + \dots + P_k = I_n$. Since the rank and trace are equal, complementarity implies $n_1 + \dots + n_k = n$, where $n_r = \text{tr}(P_r) = \text{rank}(P_r)$. Then, for every $Y \in \mathcal{H}$, we have the linear and Pythagorean identities

$$\begin{aligned} Y &= P_1 Y + \dots + P_k Y; \\ \|Y\|^2 &= \|P_1 Y\|^2 + \dots + \|P_k Y\|^2. \end{aligned}$$

By an obvious extension of the argument given above for $Y \sim N_n(0, \sigma^2 V)$, the projected random variables $P_r Y \sim N(0, \sigma^2 P_r V)$ are mutually independent, and $\|P_r Y\|^2 \sim \sigma^2 \chi_{n_r}^2$ are also mutually independent. In the literature on analysis of variance, this distributional decomposition is known as Cochran's theorem, or the Fisher-Cochran theorem.

For $r \neq s$, the ratio of mean squares

$$\frac{\|P_r Y\|^2/n_r}{\|P_s Y\|^2/n_s}$$

is distributed independently of σ according to Fisher's F distribution F_{n_r, n_s} .

The decomposition can be stated in an alternative way in terms of a sequence of strictly nested subspaces

$$\mathbf{0} = \mathcal{X}_0 \subset \mathcal{X}_1 \subset \dots \subset \mathcal{X}_k \subset \mathcal{X}_{k+1} = \mathcal{H}$$

of dimensions $0 < n_1 < n_2 < \dots < n_k < n$. Let P_r be the orthogonal projection onto \mathcal{X}_r so that $P_r P_s = P_{r \wedge s}$, and $Q_r Q_s = Q_{r \vee s}$ for the complementary projections. Then the increments $(\Delta P)_r = P_r - P_{r-1} = Q_{r-1} - Q_r$ are mutually orthogonal projections satisfying the conditions for Cochran's theorem. In particular, if $Y \sim N(X\beta, \sigma^2 V)$ satisfies the standard linear model assumption with non-zero mean such that $\mathcal{X}_1 = \text{span}(X)$, and $\mathcal{X}_2 = \text{span}(X, Z)$ is any proper subspace of \mathcal{H} containing \mathcal{X} as a proper subspace, then

$$\|Q_1 Y\|^2 = \|(Q_1 - Q_2)Y\|^2 + \|Q_2 Y\|^2$$

is a decomposition of the residual sum of squares into independent $\sigma^2 \chi^2$ components. Consequently, the ratio of mean squares

$$F = \frac{\|(Q_1 - Q_2)Y\|^2/(n_2 - n_1)}{\|Q_2 Y\|^2/(n - n_2)}$$

is distributed according to the F distribution.

15.4.3 Prediction and Conditional Expectation

Let $Y \sim N_n(0, \Sigma)$ on \mathbb{R}^n , and let $Z = KY$ be any linear transformation whose kernel is \mathcal{K} . Then the conditional expectation given Z is $E(Y | Z) = QY$, where Q is the orthogonal projection $\mathcal{H} \rightarrow \mathcal{H}$ whose kernel is \mathcal{K} . If K has full rank, the matrix form (15.4) makes it clear that the conditional expected value

$$QY = \Sigma K'(K\Sigma K')^{-1}KY = \Sigma K'(K\Sigma K')^{-1}Z$$

is indeed a function of the observation Z .

If we write $Y = PY + QY$ as the sum of complementary orthogonal projections, the proof is trivial because $KQ = K$ and $Z = KQY$ is independent of PY . Consequently,

$$\begin{aligned} E(Y | Z) &= E(PY + QY | KQY) = QY; \\ \text{cov}(Y | Z) &= \text{cov}(PY + QY | QY) = \text{cov}(PY) = P\Sigma. \end{aligned} \quad (15.7)$$

Thus, the conditional distribution of Y given Z is $N(QY, P\Sigma)$. These equations are dual to (15.6).

In standard probability terminology, prediction calls for the conditional distribution given the σ -field generated by the observation as a measurable transformation. By definition, the σ -field generated by a linear transformation with kernel $\mathcal{K} \subset \mathbb{R}^n$ is the Borel σ -field in \mathbb{R}^n/\mathcal{K} , i.e., all Borel subsets $A \subset \mathbb{R}^n$ such that $A + \mathcal{K} = A$. In that probabilistic sense, all linear transformations having the same kernel are equivalent.

Partitioned Matrix Representation

In applied work, it is often the case that $Y: U \rightarrow \mathbb{R}$ is a function on the units, and the observation $Z = Y[U_0]$ is the restriction of Y to a sub-sample $U_0 \subset U$ of size $n - k$. For that setting, it is convenient and computationally efficient to use partitioned-matrix notation in which $Y_0 = Y[U_0]$, and $Y_1 = Y[\bar{U}_0]$ is the restriction to the complementary subsample:

$$\begin{aligned} Y &= \begin{bmatrix} Y_0 \\ Y_1 \end{bmatrix}; & K &= [I_{n-k} : 0]; & \Sigma &= \begin{bmatrix} \Sigma_{00} & \Sigma_{01} \\ \Sigma_{10} & \Sigma_{11} \end{bmatrix}; & \Sigma^{-1} &= \begin{bmatrix} W_{00} & W_{01} \\ W_{10} & W_{11} \end{bmatrix}; \\ Q &= \begin{bmatrix} I_{n-k} & 0 \\ \Sigma_{10}\Sigma_{00}^{-1} & 0 \end{bmatrix}; & P &= \begin{bmatrix} 0 & 0 \\ -\Sigma_{10}\Sigma_{00}^{-1} & I_k \end{bmatrix}; & P\Sigma &= \begin{bmatrix} 0 & 0 \\ 0 & \Sigma_{11} - \Sigma_{10}\Sigma_{00}^{-1}\Sigma_{01} \end{bmatrix}. \end{aligned}$$

The conditional distribution of Y_1 given Y_0 is Gaussian with moments

$$\begin{aligned} E(Y_1 | Y_0) &= (QY)_1 = \Sigma_{10}\Sigma_{00}^{-1}Y_0; \\ \text{cov}(Y_1 | Y_0) &= (P\Sigma)_{11} = \Sigma_{11} - \Sigma_{10}\Sigma_{00}^{-1}\Sigma_{01} = W_{11}^{-1}. \end{aligned}$$

This component-wise version of the conditional distribution speaks directly to the goal of predicting the values for extra-sample units, and it is computationally more efficient than (15.7) because it sets aside obvious degeneracies. But it does so at the cost of obscuring a crucial aspect of the geometry, namely that Gaussian point prediction is an orthogonal projection.

In applied work, the situation is typically a little more complicated because $\mu = E(Y)$ is never zero, in which case the conditional mean is

$$E(Y_1 | Y_0) = \mu_1 + \Sigma_{10}\Sigma_{00}^{-1}(Y_0 - \mu_0).$$

In practice, unknown parameters must be estimated before this can be computed.

Example: Exchangeable Gaussian Process

A zero-mean exchangeable Gaussian process has finite-dimensional distributions

$$Y[n] \sim N(0, \Sigma_n = \sigma_0^2 I_n + \sigma_1^2 J_n),$$

where $J_n(i, j) = 1$ is the $n \times n$ matrix whose components are all one. The inverse matrix is

$$\Sigma_n^{-1} = \sigma_0^{-2} \left(I_n - \frac{\theta}{1+n\theta} J_n \right),$$

where $\theta = \sigma_1^2/\sigma_0^2$ is the variance-component ratio.

Regardless of the variance parameters, $P_n = J_n/n$ is the orthogonal projection onto the subspace $\mathbf{1}_n$ of constant functions, and $Q_n = I_n - J_n$ is the complementary orthogonal projection. Thus, the projected random vectors

$$\begin{aligned} P_n Y &\sim N(0, n^{-1} J_n \Sigma) = N(0, (\sigma_0^2/n + \sigma_1^2) J_n) \\ Q_n Y &\sim N(0, Q_n \Sigma) = N(0, \sigma_0^2 Q_n) \end{aligned}$$

are independent. To avoid confusion in statistical work where the covariance matrix is not completely known, it is best to fix the inner product in \mathcal{H} rather than having a parameter-dependent inner product: see the cautionary remarks in Sect. 15.2.2. Most

statistical work involving exchangeable or partially exchangeable processes uses the standard invariant inner product $\sum x_i y_i$. With this understanding, the squared norms

$$\|P_n Y\|^2 = n \bar{Y}_n^2 \sim (\sigma_0^2 + n\sigma_1^2) \chi_1^2,$$

$$\|Q_n Y\|^2 = (n - 1)s_n^2 \sim \sigma_0^2 \chi_{n-1}^2$$

are independent χ^2 random variables with scale factors as indicated. Much of analysis of variance for balanced designs is based on extensions of this result to partially exchangeable arrays.

For $n \geq 1$, the partitioned-matrix formulae in the preceding subsection imply that the conditional distribution of Y_{n+1}, \dots, Y_{n+m} given $Y[n]$ is exchangeable Gaussian with moments

$$E(Y[n+1:m] | Y[n]) = \frac{n\theta \bar{Y}_n}{1+n\theta} \mathbf{1}_m,$$

$$\text{cov}(Y[n+1:m] | Y[n]) = \sigma_0^2 I_m + \sigma_1^2 J_m / (1+n\theta).$$

The same partitioned-matrix formulae also imply that the conditional distribution of the average $(Y_{n+1} + \dots + Y_{n+m})/m$ given $Y[n]$ is Gaussian with moments

$$E(\bar{Y}_{n+1:m} | Y[n]) = \frac{n\theta \bar{Y}_n}{1+n\theta},$$

$$\text{var}(\bar{Y}_{n+1:m} | Y[n]) = \sigma_0^2/m + \sigma_1^2/(1+n\theta).$$

This conditional distribution has a limit as $m \rightarrow \infty$ for fixed n , implying that the infinite average is a conditionally non-degenerate random variable such that

$$\bar{Y}_\infty \sim N\left(\frac{n\theta \bar{Y}_n}{1+n\theta}, \frac{\sigma_0^2\theta}{1+n\theta}\right).$$

The limit $\theta \rightarrow \infty$ gives $\bar{Y}_\infty - \bar{Y}_n \sim N(0, \sigma_0^2/n)$. For $n \geq 2$, the internally-standardized ratio

$$\frac{\sqrt{n}(\bar{Y}_\infty - \bar{Y}_n)}{s_n} \sim t_{n-1},$$

is distributed as Student's t on $n - 1$ degrees of freedom.

15.4.4 Eddington's Formula

Scalar Signal Estimation

Given a random signal $X \sim P$ and an observation $Y = X + \varepsilon$ contaminated by independent additive Gaussian noise, how do we estimate the signal? This version of the signal estimation problem was first posed in 1926 by the Astronomer Royal, Sir Frank Watson Dyson, in connection with parallax resolution problems in astronomical observations made at Greenwich.

In the astronomical setting, there is a large number of independent signals (parallaxes), all identically distributed according to some unknown signal distribution $X_i \sim P$, and the measured parallaxes $Y_i = X_i + \varepsilon_i$ are contaminated by independent additive Gaussian error with known variance. It is feasible to estimate the marginal density by smoothing, but it is not obvious how to estimate the signal distribution or how to adjust the observation to account for measurement error. Eddington provided a simple and elegant solution.

If the signal density is $p(\cdot)$, and the noise is standard Gaussian, the joint density of (Y, X) is $p(x)\phi(y - x)$, and the marginal density is

$$m(y) = \int p(x)\phi(y - x) dx = \phi(y) \int_{\mathbb{R}} p(x)e^{xy-x^2/2} dx.$$

Thus the density ratio $m(y)/\phi(y)$ is the Laplace transform of the function $p(x)e^{-x^2/2}$. In addition, $p(x)e^{-x^2/2}\phi(0)/m(0)$ is a probability density whose cumulant generating function is $\log(m(y)/\phi(y))$. In the absence of other considerations, the goal of signal estimation is to compute the conditional expected value of the signal given the data.

$$\begin{aligned} E(e^{tX} | Y) &= \int e^{tx} p(x)\phi(y - x) dx \Big/ m(y) \\ &= \frac{\phi(y)}{m(y)} \int e^{tx+xy-x^2/2} p(x) dx \\ &= \frac{m(y+t)}{\phi(y+t)} \Big/ \frac{m(y)}{\phi(y)}; \\ E(X | Y) &= \frac{d}{dt} \log \left(\frac{m(y+t)}{\phi(y+t)} \right)_{t=0} \\ &= \frac{d}{dy} \log \left(\frac{m(y)}{\phi(y)} \right) = y + \frac{m'(y)}{m(y)}. \\ \text{var}(X | Y) &= \frac{d^2}{dy^2} \log \left(\frac{m(y)}{\phi(y)} \right). \end{aligned}$$

Eddington's solution was the additive adjustment $\sigma^2 m'(y)/m(y)$, scaled to account for the observation variance. Higher-order derivatives are the higher-order cumulants of the conditional distribution.

Dyson started out with a table or histogram of parallaxes of stars measured at Greenwich, from which he estimated the marginal density by smoothing. Knowing the observation variance from replicates, he estimated the adjustment and reported the adjusted values. He also noted that if the signal distribution happens to be normal, the correction reduces to $-y\sigma^2/(\sigma^2 + \sigma_x^2)$. Dyson's observed parallax distribution was strongly skewed in the positive direction, so his signal distribution was far from normal.

Eddington's formula is remarkable for two reasons. The most obvious is that it depends only on the marginal density of the observations. Less obvious is the fact that if the signals are restricted to an interval, say $-1 \leq x \leq 2$ or $x > 0$, then $E(X | Y)$ also lies in the interval. This aspect was crucial for Dyson's task because parallaxes are positive even if some observations are negative, and Dyson was understandably reluctant to report a negative value.

Eddington's formula was attributed by Robbins (1956) to personal correspondence with M.C.K. Tweedie. In the subsequent statistical literature, it has been called Tweedie's formula: see Efron (2011) or McCullagh and Polson (2018).

Isotropic Vector Signal Estimation

Eddington's formula applies also to vector signals $X \in \mathbb{R}^d$ contaminated by additive Gaussian noise $\varepsilon \sim N_d(0, \Sigma)$. The conditional mean given $Y = X + \varepsilon$ is then

$$E(X | Y) = \Sigma \frac{d}{dy} \log \left(\frac{m(y)}{\phi(y)} \right) = y + \Sigma m'(y)/m(y),$$

where ϕ is the density of $N_d(0, \Sigma)$, and $m'(y)$ is the gradient vector.

For the vector formula to be useful in practical work, it is usually necessary to make further simplifying assumptions. Rotational symmetry for both the signal and the noise is reasonably natural. In that case $\varepsilon \sim N(0, I_d)$, the signal density satisfies $p(\sigma x) = p(x)$ for each orthogonal transformation $\sigma: \mathbb{R}^d \rightarrow \mathbb{R}^d$, and $m(\sigma y) = m(y)$ is also rotationally symmetric. The conditional expectation $H(y) = E(X | Y = y)$ then satisfies the commutativity condition

$$H(\sigma y) = \sigma H(y).$$

Since the sub-group ± 1 commutes with every group element, the normalized vector $H(y)/\|H(y)\|$ is necessarily equal to $\pm y/\|y\|$. Examples of such transformations include scalar multiples $y \mapsto -3y$, and invariant multiples such as the Euclidean normalization $y \mapsto y/\|y\|_2$ or the inversion $y/\|y\|_2^2$. But L_1 -normalization $y \mapsto y/\|y\|_1$ does not commute with \mathcal{G} .

Ordinarily, $y \mapsto H(y)$ is a shrinkage towards the origin. But Exercise 15.10 shows that $\|H(y)\| \geq \|y\|$ is possible.

Spectral Moments for Matrix Reconstruction

Suppose that the signal X is a random matrix of order $n \times p$, and that the components of ε are independent standard normal. The first spectral moment is the conditional expected value of $\text{tr}(X'X)$, i.e., the sum of the eigenvalues, given $Y = y$. By Eddington's second-moment formula, the first spectral moment is the trace of the second partial derivatives of the conditional moment generating function

$$M(t \mid Y) = \frac{m(y + t)\phi(y)}{\phi(y + t)m(y)}$$

with respect to t at the origin. The trace of second derivatives is the standard Laplacian. The second spectral moment is the conditional expected value of the sum of squared eigenvalues, i.e., the expected value of $\text{tr}((X'X)^2)$ given $Y = y$. By Eddington's fourth-moment formula, the second spectral moment is the cyclic scalar contraction of the fourth partial derivatives of the moment generating function. Similar remarks apply to the k th-order spectral moment, which is the cyclic scalar contraction of the partial derivatives of order $2k$.

The spectral-moment formulae can be simplified if the signal distribution is rotationally symmetric with respect to left and right orthogonal transformation. In other words, $p(\sigma x \tau) = p(x)$ for all orthogonal matrices σ of order n and τ of order p . Since ε is rotationally symmetric, the convolution is also rotationally symmetric in the same sense, and the conditional expectation is equi-variant in the sense $H(\sigma y \tau) = \sigma H(y) \tau$. Equi-variance implies that H acts only on the singular values.

Although the conditional expectation $y \mapsto H(y)$ is an action on singular values, the transformation does not necessarily act component-wise, nor is it necessarily a shrinkage. Under certain sparsity assumptions, it is possible to be more specific about the nature of the transformation, which is a shrinkage towards the origin applied component-wise to the singular values: see Sect. 17.5.3.

15.4.5 Linear Regression

Let $\mathcal{X} \subset \mathbb{R}^n$ be a subspace of dimension p spanned by the columns of the given matrix X of order $n \times p$, let V be a given strictly positive definite matrix with inverse $W = V^{-1}$, and let \mathcal{H} be the Euclidean space with inner product $\langle x, y \rangle = x'Wy$.

Linear regression refers to the family of Gaussian distributions on \mathbb{R}^n

$$\{N_n(\mu, \sigma^2 V) \mid \mu = X\beta \in \mathcal{X}, \sigma > 0\}$$

indexed by $\beta \in \mathbb{R}^p$ and $\sigma > 0$. For geometrical purposes, we want to regard each of these as a distribution on the same Hilbert space, so we choose the given matrix $W = V^{-1}$ rather than Σ^{-1} in order that all operations in \mathcal{H} be computable.

We now suppose that, for some unspecified parameter point (β, σ^2) , an observation $Y \sim N_n(X\beta, \sigma^2 V)$ is generated and the value $y \in \mathcal{H}$ is observed. To estimate the parameter, we use the log likelihood function, which is the log density

$$l(\beta, \sigma^2; y) = -\frac{1}{2}\|y - X\beta\|^2/\sigma^2 - n \log \sigma + \text{const.}$$

So far as the regression parameter is concerned, maximization of the log likelihood is equivalent to minimizing the Euclidean squared distance $\|y - X\beta\|^2$ over points $\mu = X\beta$ in \mathcal{X} . Regardless of σ^2 , the minimum over \mathcal{X} occurs at the Euclidean projection

$$\hat{\mu} = X\hat{\beta} = Py = X(X'WX)^{-1}X'Wy,$$

and the minimum value achieved is the residual quadratic form $\|(I - P)y\|^2 = \|Qy\|^2$.

The projection PY and its complement QY are independent Gaussian random vectors with distributions

$$PY \sim N(X\beta, \sigma^2 PV), \quad QY \sim N(0, \sigma^2 QV).$$

It follows from Sect. 15.4.1 that $\|QY\|^2/\sigma^2$ is distributed as χ_{n-p}^2 , i.e., the weighted residual sum of squares $\|QY\|^2$ is distributed as $\sigma^2 \chi_{n-p}^2$. The conventional estimate of σ^2 is the mean-squared residual

$$s^2 = \|Qy\|^2/(n - p),$$

which is unbiased and strictly larger than the maximum-likelihood estimate $\|Qy\|^2/n$.

Statistical computer packages invariably report the least-squares coefficient vector $\hat{\beta} = (X'WX)^{-1}X'Wy$ together with the standard errors, which are the square roots of the diagonal components of the matrix

$$\text{cov}(\hat{\beta}) = s^2(X'WX)^{-1}.$$

15.4.6 Linear Regression and Prediction

Exercises 15.4–15.9 give an outline of the argument for combining linear regression with prediction. The parametric family $Y \sim N(X\beta, \sigma^2 V)$ on $\mathcal{H} = (\mathbb{R}^n, W)$ determines a parametric family on the observation space, which is the image of

a linear transformation $K: \mathbb{R}^n \rightarrow \mathbb{R}^{n-k}$. The regression parameter is estimated by weighted least squares based on the observation $Z \sim N_{n-k}(KX\beta, \sigma^2 KVK')$, or equivalently, based on $QY \sim N_n(QX\beta, \sigma^2 QV)$, where Q is the orthogonal projection having the same kernel. Identifiability requires $p = \text{rank}(KX) \leq n - k$, in which case the composite transformation $L: Y \mapsto QY \mapsto \hat{\mu}$ is a linear projection $\mathcal{H} \rightarrow \mathcal{H}$.

The conditional distribution of Y given $Z = KY$ is the same as the conditional distribution of the sum $PY + QY$ given QY . Independence of PY and QY implies that the conditional distribution is

$$N_n(QY + P\mu, \sigma^2 PV).$$

The least squares estimate is obtained by parameter substitution

$$N_n(QY + P\hat{\mu}, s^2 PV) = N(QY + L_0Y, \sigma^2 PV), \quad (15.8)$$

with σ^2 replaced by s^2 if needed.

The transformation $Y \mapsto \hat{\mu} = LY$ is a linear projection whose image is \mathcal{X} and whose kernel includes \mathcal{K} . These are arbitrary non-overlapping subspaces so L is not orthogonal in \mathcal{H} . It is the sum of two transformations $L_1 = QL$, which is the orthogonal projection with image $Q\mathcal{X}$, and $L_0 = PL$ which is nilpotent, i.e., $L_0^2 = 0$, because $LP = 0$.

Notation for Component-Wise Transformation

If the observation is a component-wise restriction of Y , it is convenient and efficient to express (15.8) in partitioned matrix notation such that $K = [I_{n-k}, 0]$, $KY = Y_0$ and $KX = X_0$, in which case

$$QY = \begin{pmatrix} Y_0 \\ V_{10}V_{00}^{-1}Y_0 \end{pmatrix}, \quad P\hat{\mu} = \begin{pmatrix} 0 \\ \hat{\mu}_1 - V_{10}V_{00}^{-1}\hat{\mu}_0 \end{pmatrix},$$

$$(PV)_{11} = V_{11} - V_{10}V_{00}^{-1}V_{01} = W_{11}^{-1}.$$

The least-squares estimate is $\hat{\beta} = (X'_0 V_{00}^{-1} X_0)^{-1} X'_0 V_{00}^{-1} Y_0$, giving the fitted mean with components $\hat{\mu}_0 = X_0 \hat{\beta}$ and $\hat{\mu}_1 = X_1 \hat{\beta}$. Given Y_0 , the predictive distribution for Y_1 has moments

$$\hat{\mu}_1 + V_{10}V_{00}^{-1}(Y_0 - \hat{\mu}_0) \quad \text{and} \quad \sigma^2 W_{11}^{-1},$$

with σ^2 replaced by s^2 where needed. The predictive mean in this setting goes by various names—best linear predictor, fiducial predictor, Kriging estimate, smoothing spline—depending on the area of application.

Fiducial Prediction

The least-squares predictive distribution (15.8) associates with each observation point $z = Ky$ in \mathbb{R}^{n-k} a probability distribution on \mathbb{R}^n . It assigns probability one to the k -dimensional coset

$$y + \mathcal{K} = Qy + \mathcal{K} = \{y \in \mathbb{R}^n \mid Ky = z\}, \quad (15.9)$$

which is the subset of points that are consistent with the observed value.

Any distribution defined on Borel subsets of \mathbb{R}^n can be restricted to a sub- σ -field if the need arises. In the linear-model setting $N_n(X\beta, \sigma^2 V)$ with $\mathcal{X} = \text{span}(X)$, an event $A \subset \mathbb{R}^n$ is said to be translation-invariant if $A + \mathcal{X} = A$. For historical reasons, the invariant events are also called fiducial events; the set of fiducial events is the Borel σ -field $\mathcal{B}(\mathbb{R}^n / \mathcal{X})$.

According to the fiducial argument, the set of distributions $N_n(X\beta, \sigma^2 V)$ indexed by $\beta \in \mathbb{R}^p$ for fixed σ is interpreted as a single distribution $N_n(0, \sigma^2 V)$ on fiducial events. For this purpose, two Gaussian distributions $N_n(\mu_0, \Sigma_0)$ and $N_n(\mu_1, \Sigma_1)$ are equivalent modulo \mathcal{X} , if, for any linear transformation T whose kernel includes \mathcal{X} , $T\mu_0 = T\mu_1$ and $T\Sigma_0 T' = T\Sigma_1 T'$. In particular, $\ker(Q_{\mathcal{X}}) = \mathcal{X}$ implies that the distributions $N_n(X\beta, V)$ and $N_n(0, Q_{\mathcal{X}} V)$ are fiducially equivalent. Thus, a single fiducial distribution has multiple covariance-matrix representations in \mathbb{R}^n .

Fiducially speaking, the response distribution is $N_n(0, \sigma^2 Q_{\mathcal{X}} V)$, the observation is a linear transformation Q^\dagger with kernel $\mathcal{X} + \mathcal{K}$, and (15.7) implies that the conditional distribution given the observation is

$$\begin{aligned} N_n(Q^\dagger Y, \sigma^2(Q_{\mathcal{X}} - Q^\dagger)Q_{\mathcal{X}} V) &= N_n(Q^\dagger Y, \sigma^2(Q_{\mathcal{X}} - Q^\dagger)V) \\ &\cong N_n(Q^\dagger Y + \hat{\mu}, \sigma^2(P^\dagger - P_{\mathcal{X}})V) \\ &\cong N_n(Q^\dagger Y + \hat{\mu}, \sigma^2 P_{\mathcal{K}} V). \end{aligned}$$

The predictive distribution is restricted to fiducial events in \mathbb{R}^n , and these various expressions are equivalent when restricted to $\mathcal{B}(\mathbb{R}^n / \mathcal{X})$. For example, the addition of $\hat{\mu}$ to the mean has no effect. The last expression is the [fiducial restriction of the] least-squares predictive distribution.

15.5 Additivity

15.5.1 IDOFNA Algorithm

One degree of freedom for non-additivity is a technique introduced by Tukey (1949) to check the adequacy of the linear model $Y \sim N(X\beta, \sigma^2 V)$ by testing for deviations from additivity and/or linearity. Tukey was particularly concerned with

additivity assumptions for factorial models in randomized-blocks and Latin-square designs, but the technique is valid more broadly for simple linear regression and multiple linear regression.

The computational procedure goes as follows. First compute the least-squares fitted vector $\hat{\mu} = X\hat{\beta} = PY$ together with the residual sum of squares $\|Y - \hat{\mu}\|^2$ on $n - p$ degrees of freedom. Second, compute the derived vector z with components $z_i = \hat{\mu}_i^2$. Third, fit the extended linear model including both X and z , i.e., $E(Y) = X\beta + z\gamma$, and compute the least-squares fitted vector $\hat{\mu}_1 = P_1Y$ by projection onto the subspace spanned by X and z .

The reduction in residual sum of squares $\|QY\|^2 - \|Q_1Y\|^2$ is the one degree of freedom for non-additivity. According to Tukey, the null distribution is exactly $\sigma^2\chi^2_1$, and the mean-square ratio

$$F = \frac{\|QY\|^2 - \|Q_1Y\|^2}{\|Q_1Y\|^2/(n - p - 1)} \quad (15.10)$$

is distributed exactly as $F_{1,n-p-1}$ if the null assumption $Y \sim N(X\beta, \sigma^2V)$ is correct. The 1DOFNA F -ratio provides an exact test of the null model, large values being interpreted as evidence of non-additivity. Alternatively, the least-squares coefficient of z can be used in the standard manner

$$T = \hat{\gamma} / \text{s.e.}(\hat{\gamma}),$$

and the value compared with the null distribution t_{n-p-1} . As always, $T^2 = F$, so the two approaches are effectively equivalent. If the F -ratio is large, the remedy suggested is to transform the response $Y \mapsto Y^\lambda$, and Tukey's suggested power transform is $\lambda = 1 - 2\hat{\gamma}\bar{Y}$.

15.5.2 1DOFNA Theory

Although this description is entirely satisfactory as a computational procedure, the notation is misleading because the transformation that sends u to P_1u is neither a projection $\mathcal{H} \rightarrow \mathcal{H}$ nor a linear transformation. It does not satisfy $P_1(u + v) = P_1u + P_1v$ for vectors $u, v \in \mathcal{H}$, nor does it satisfy $P_1^2 = P_1$. Hence, Tukey's claim is not an immediate consequence of earlier remarks such as Cochran's theorem (Sect. 15.3.2), which is concerned exclusively with linear transformations and orthogonal projections.

A correct argument proceeds as follows. First, the projected random vectors PY and QY are independent, so the conditional distribution of QY given $\hat{\mu} = PY$ is $N(0, \sigma^2 QV)$, and the conditional distribution of QY given z is also $N_n(0, \sigma^2 QV)$. It follows that $\gamma = 0$ if the null model holds. Given $\hat{\mu}$, least-squares estimate of γ

is the regression coefficient of the residuals on z —or more correctly on Qz —which is conditionally linear

$$\hat{\gamma} = (z' W Q z)^{-1} z' W Q Y,$$

and the conditional distribution given $\hat{\mu}$ is $N(0, \sigma^2(z' W Q z)^{-1})$. Given $\hat{\mu}$, the residual vector is split additively into two orthogonal parts

$$QY = Qz\hat{\gamma} + Q(Y - z\hat{\gamma}) = Qz(z' W Q z)^{-1} z' W Q Y + Q_1 Y.$$

The residual sum of squares also splits into two parts

$$\|QY\|^2 = \gamma'(z' W Q z)\gamma + \|Q_1 Y\|^2 = Y' W Q z (z' W Q z)^{-1} z' W Q Y + Y' W Q_1 Y.$$

According to Sect. 15.3.2, these are conditionally independent given $\hat{\mu}$ with distributions $\sigma^2 \chi^2_1$ and $\sigma^2 \chi^2_{n-p-1}$ respectively. Thus, Tukey's distributional assertions are upheld conditionally on $\hat{\mu}$, and therefore unconditionally.

15.5.3 Scope and Rationale

Apart from the condition $p \leq n - 2$, which is needed to ensure $Q_1 \neq 0$, one additional condition related to the algebraic properties of the subspace \mathcal{X} is needed. A subspace $\mathcal{X} \subset \mathbb{R}^n$ is said to be a *commutative ring* if, for each pair of vectors $u, v \in \mathcal{X}$, the component-wise product $uv = vu$ also belongs to \mathcal{X} . If \mathcal{X} happens to be a ring, then $\hat{\mu} \in \mathcal{X}$ implies $z = \hat{\mu}^2 \in \mathcal{X}$, so that $Qz = 0$. Thus, if \mathcal{X} is a ring under multiplication, no degree of freedom for non-additivity exists.

In the factorial-model setting with non-nested block or treatment factors A, B, C, \dots , each of the factorial subspaces

$$\mathbf{0}, \mathbf{1}, A, B, C, AB, AC, BC, ABC$$

is also a commutative ring that is closed under multiplication. (Here, AB is the subspace denoted by either $A:B$ or $A*B$ in R notation.) In order for the 1DOFNA to be non-trivial, it is necessary that \mathcal{X} not be a ring. More explicitly, the set of points $x \in \mathcal{X}$ such that $x^2 \in \mathcal{X}$ must have measure zero. Apart from degenerate designs having completely aliased factors, the 1DOFNA is non-trivial for every other factorial subspace such as $A + B$ or $AB + C$ or $AB + BC$, that includes a non-trivial $+$ operator. Note that $\mathbf{1} + A$ is the same as A and $A+B+A*B$ is the same as AB , both of which are rings.

In the case of simple linear regression with $E(Y_i) = \beta_0 + \beta_1 x_i$ for a quantitative variable x , the constructed variable $\hat{\mu}^2$ is a quadratic function of x . Provided that the design contains at least three distinct x -values, the vectors $\mathbf{1}, x, x^2$ are linearly independent. With probability one $\hat{\beta}_1 \neq 0$, in which case the vectors $\mathbf{1}, x, \hat{\mu}^2$

are also linearly independent. For that setting, the 1DOFNA is equivalent to the one degree of freedom for non-linearity, and specifically quadratic deviations from linearity.

Provided that the constructed variable is a function of $\hat{\mu}$, any component-wise non-linear transformation such as $z_i = \exp(\hat{\mu}_i)$, or any non-component-wise transformation $\mathcal{H} \rightarrow \mathcal{H}$, may be used in the algorithm. Subject to the condition $z \notin \mathcal{X}$ mentioned above, the distributional argument leading to the conclusion that the 1DOFNA F -ratio is distributed as $F_{1,n-p-1}$ is unaffected by the choice of transformation. For $\mathcal{X} \neq \mathbf{1}$, the neighbour average $z_i = \text{ave}_{j \in \text{nb}(i)} \hat{\mu}_j$ is an example of a linear non-component-wise transformation $\mathcal{H} \rightarrow \mathcal{H}$ that might arise in a spatial or graphical setting.

Note that the word transformation is used above in two distinct senses that are algebraically distinct. First, every statistical vector is a function $U \rightarrow \mathbb{R}$ on the units, and component-wise transformation $g: \mathbb{R} \rightarrow \mathbb{R}$ refers to composition $y \mapsto gy$ on the left as illustrated by the diagram $U \xrightarrow{y} \mathbb{R} \xrightarrow{g} \mathbb{R}$. Component-wise transformation exploits the fact that \mathbb{R}^U is a commutative ring. Second, every statistical vector is also a point $y \in \mathcal{H}$, and a typical linear transformation $\mathcal{H} \rightarrow \mathcal{H}$ such as $y \mapsto \hat{\mu}$ or $y \mapsto Qy$ does not act component-wise.

15.6 Exercises

15.1 Show that the set of $2n \times 2n$ real matrices of the form

$$\begin{pmatrix} A & B \\ -B & A \end{pmatrix}$$

is closed under matrix addition and multiplication. Show also that the ‘linear’ mapping into the space of complex $n \times n$ matrices

$$\begin{pmatrix} A & B \\ -B & A \end{pmatrix} \mapsto A + iB$$

is an isomorphism preserving matrix addition and multiplication.

15.2 Let $A + iB$ be a full-rank Hermitian matrix of order n . Show that the inverse matrix $C + iD$ is also Hermitian and satisfies the pair of equations

$$AD + BC = 0; \quad AC - BD = I_n.$$

Deduce that the $2n \times 2n$ real symmetric matrices

$$\begin{pmatrix} A & B \\ -B & A \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} C & D \\ -D & C \end{pmatrix}$$

are mutual inverses. What does this matrix isomorphism imply about the relation between complex Gaussian vectors and real Gaussian vectors?

15.3 By writing the complex vector z and the Hermitian matrix Γ as a linear combination of real and imaginary parts, show that the Hermitian quadratic form $z^*\Gamma z$ reduces to the following linear combination of real quadratic forms:

$$(x' - iy')\Gamma_0(x + iy) + i(x' - iy')\Gamma_1(x + iy) = x'\Gamma_0x + y'\Gamma_0y + y'\Gamma_1x - x'\Gamma_1y.$$

Hence deduce that the real and imaginary parts of $Z \sim \mathbb{C}N(0, \Sigma)$ are identically distributed Gaussian vectors $N(0, \Sigma_0)$ with covariances $\text{cov}(X, Y) = -\text{cov}(Y, X) = \Sigma_1$.

Gaussian Linear Prediction The next five exercises are concerned with estimation and prediction in the Gaussian linear model $Y \sim N_n(\mu = X\beta, \sigma^2 V)$ in which the observation is the linear transformation $Z = KY$. The matrices X of order $n \times p$, K of order $n - k \times n$, and V of order $n \times n$ are given, while β , σ^2 are parameters to be estimated. All three matrices are of full rank, the product KX has rank $p \leq n - k$, while the Hilbert space \mathcal{H} with inner-product matrix $W = V^{-1}$ determines the geometry.

15.4 Show that the maximum-likelihood estimate of β satisfies

$$[X'K'(KVK')^{-1}KX]\hat{\beta} = X'K'(KVK')^{-1}Z = X'WQY,$$

where $Q: \mathcal{H} \rightarrow \mathcal{H}$ is the orthogonal projection with kernel $\ker(K)$.

15.5 Deduce that the linear transformation $Y \mapsto LY = \hat{\mu} = X\hat{\beta}$ is a projection $\mathcal{H} \rightarrow \mathcal{H}$, but not an orthogonal projection unless $Q\mathcal{X} = \mathcal{X}$.

15.6 Deduce that the composite linear transformation $Y \mapsto L_1Y = Q\hat{\mu}$ is also a projection, and that it is the orthogonal projection whose image is the p -dimensional subspace $Q\mathcal{X}$.

15.7 For the complementary projection $P = I_n - Q$ whose image is \mathcal{K} , deduce that the composite linear transformation $Y \mapsto L_0Y = P\hat{\mu}$ is nilpotent, i.e., that $L_0^2 = 0$. What does nilpotence imply about the image and kernel of L_0 ? Construct the multiplication table for L_0, L_1 .

15.8 Show that the least-squares estimate of the conditional distribution of Y given Z is

$$N_n(QY + P\hat{\mu}, s^2PV)$$

for some scalar s^2 . Show that the least-squares estimate is singular and is supported on the k -dimensional coset $\mathcal{Q}Y + \mathcal{K}$. Explain why self-consistency requires $K\mathcal{Q} = K$.

15.9 Show that the zero-mean exchangeable Gaussian process in Sect. 15.3.3 with covariances

$$\text{cov}(Y_r, Y_s) = \sigma_0^2 \delta_{rs} + \sigma_1^2,$$

has a dynamic or sequential representation beginning with $Y_0 = 0$ followed by

$$Y_{n+1} = \frac{n\theta \bar{Y}_n}{1+n\theta} + \sigma_0 \sqrt{1+\theta/(1+n\theta)} \epsilon_{n+1}$$

for $n \geq 0$. Here $\theta = \sigma_1^2/\sigma_0^2$ is the variance ratio, and ϵ_1, \dots are independent standard normal variables.

15.10 Suppose that X is uniformly distributed on the surface of the unit sphere in \mathbb{R}^d , and that $Y \sim N(X, \sigma^2 I_d)$ is observed. Show that Eddington's formula reduces to the projection $E(X | Y) = Y/\|Y\|$.

15.11 Suppose that X is uniformly distributed on the interior of the unit sphere in \mathbb{R}^d , and that $Y \sim N(X, \sigma^2 I_d)$ is observed. Show that Eddington's formula is a radial shrinkage so that $E(X | Y)$ has norm strictly less than one.

Chapter 16

Space-Time Processes



16.1 Gaussian Processes

Let \mathcal{U} be an arbitrary index set, here identified with the domain. A Gaussian process associates with each u in the domain a random variable Z_u in such a way that for each sample $U = (u_1, \dots, u_n)$, the random variable $Z[U] = (Z_{u_1}, \dots, Z_{u_n})$ has a Gaussian distribution. It should be noted that the sample points are taken in a specific order, so U is an n -tuple of points from the domain, and the components of Z are taken in the same order. If U contains repeats, say $U = (u_1, u_1, u_2)$, then the first two components of $Z[U]$ are necessarily identical.

As a function on the index set, Z may be real-valued or complex-valued or \mathbb{R}^k -valued or \mathbb{C}^k -valued. This chapter focuses chiefly on scalar processes, either real-valued or complex-valued, so Z is a function $\mathcal{U} \rightarrow \mathbb{R}$ or a function $\mathcal{U} \rightarrow \mathbb{C}$ into the space of scalars. Since the complex numbers are in 1–1 correspondence with ordered pairs of reals, every complex-valued process $Z = X + iY$ is also a \mathbb{R}^2 -valued process (X, Y) . Any reader who has made it this far has every right to ask why, in a book that professes to be concerned with scientific applications of statistical ideas, we should concern ourselves with a complex-valued process when a \mathbb{R}^2 -valued process would serve the same purpose. However, there is a legitimate reason, which is central to the theme of this chapter. For reasons discussed below, an arbitrary \mathbb{R}^2 -valued Gaussian process (X, Y) is not a complex Gaussian process in the algebraic sense. The algebra of the complex numbers is not irrelevant in the real world.

A Gaussian process is determined by its mean function $\mu(\cdot)$ and its covariance function $K(\cdot, \cdot)$. In the case of a real-valued process, μ is a function $\mathcal{U} \rightarrow \mathbb{R}$, and K is a symmetric function $\mathcal{U} \times \mathcal{U} \rightarrow \mathbb{R}$ that is also positive definite. In the case of a complex-valued process, μ is a function $\mathcal{U} \rightarrow \mathbb{C}$, and K is a positive-definite

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-14275-8_16.

Hermitian function $\mathcal{U} \times \mathcal{U} \rightarrow \mathbb{C}$. Hermitian symmetry means that $K(u, u')$ is the complex conjugate of $K(u', u)$, so K is real and positive on the diagonal. The mean function is $\mu(u) = E(Z_u)$; the covariance function is $K(u, u') = \text{cov}(Z_u, Z_{u'})$ for a real-valued process, and $K(u, u') = \text{cov}(Z_u, \bar{Z}_{u'})$ for a complex-valued process. On the finite subset $U = (u_1, \dots, u_n)$, the covariance matrix of $Z[U]$ is the finite restriction of K to ordered pairs (u_i, u_j) . Positive definiteness means that the Hermitian form $\xi^* K \xi$ is non-negative for every complex n -vector ξ , and every finite restriction of K .

This chapter is concerned exclusively with variances and covariances, so $\mu(u) = 0$ throughout. In the case of a real-valued process the covariance function

$$K(u, u') = \text{cov}(Z(u), Z(u')) = \text{cov}(Z(u'), Z(u)) = K(u', u)$$

is necessarily real and symmetric. In the case of a complex-valued process $Z_u = X_u + iY_u$ is a pair of real-valued Gaussian processes with covariance functions K_X , K_Y , and cross-covariance K_{XY} . For each pair of points u, u' , not necessarily distinct, there are two [linearly independent] complex products and two real products whose means are as follows:

$$\begin{aligned} E(Z_u Z_{u'}) &= E(X_u X_{u'} - Y_u Y_{u'}) + iE(X_u Y_{u'} + X_{u'} Y_u) \\ &= K_X(u, u') - K_Y(u, u') + i(K_{XY}(u, u') + K_{XY}(u', u)) = 0; \\ E(Z_u \bar{Z}_{u'}) &= E(X_u X_{u'} + Y_u Y_{u'}) - iE(X_u Y_{u'} - X_{u'} Y_u) \\ &= K_X(u, u') + K_Y(u, u') - i(K_{XY}(u, u') - K_{XY}(u', u)) = K(u, u'). \end{aligned}$$

The first equation is the condition for a pair of real-valued Gaussian processes X, Y to determine a complex Gaussian process in the algebraic sense. The zero real part implies $K_X = K_Y$, so the real and imaginary parts of Z are two processes having the same distribution. The imaginary part of the first equation implies that the cross-covariances satisfy the mysterious skew-symmetry condition

$$\text{cov}(Y_{u'}, X_u) = \text{cov}(X_u, Y_{u'}) = -\text{cov}(X_{u'}, Y_u) = -\text{cov}(Y_u, X_{u'}).$$

As a consequence, all non-zero second moments of the complex-valued Gaussian process are encapsulated in the conjugated second moments $\text{cov}(Z_u, \bar{Z}_{u'}) = K(u, u')$.

It is apparent from the preceding paragraph that if X, Y are independent real Gaussian processes having the same distribution with covariance function $K/2$, then $Z = X + iY$ is a complex Gaussian process whose covariance K is real and symmetric. In that sense, the only interesting complex Gaussian process are those whose covariance function has a non-zero imaginary part.

In most instances, \mathcal{U} is a topological space such as the real line, the plane, the sphere or the torus, so the continuity or degree of smoothness of the function $u \mapsto Z(u)$ is of considerable interest. In principle, it is possible for the degree

of smoothness to vary throughout the domain, in either a random manner or in a predetermined manner. But the processes described here are all well-behaved in the sense that they have the same behaviour throughout the domain.

16.2 Stationarity and Isotropy

16.2.1 Definitions

Stationarity and isotropy are properties of a process that are associated with a group action on the domain. Stationarity is a symmetry or distributional invariance with respect to domain translation; isotropy is an invariance under rotation or orthogonal transformation. For translation to make sense, the domain is necessarily a commutative group such as a vector space or an affine space; for orthogonal transformation to make sense, the domain is necessarily a Euclidean space in which the inner product determines the geometry.

A stochastic process Z with domain \mathcal{U} is said to be stationary if the following properties hold:

1. The domain is a commutative group, usually a vector space such as \mathbb{R}^d or \mathbb{C}^d for some $d \geq 0$. Other possibilities include the integers and the integers modulo k .
2. Each $g \in \mathcal{U}$ acts on the domain by addition, sending u to $u + g$;
3. The action on the domain sends the original process to $Z^g(u) = Z(u + g)$ by composition, which is a translation by $-g$;
4. Each Z^g has the same distribution as Z .

Ordinarily, the domain is a group which acts on itself by addition. Addition acts transitively, which implies that each value $Z_u = Z(u)$ has the same distribution as Z_0 , i.e., all one-dimensional marginal distributions are equal. Since differences are invariant under translation, stationarity implies that $(Z_u, Z_{u'})$ has the same joint distribution as $(Z_v, Z_{v'})$ whenever $u - u' = v - v'$. Stationarity does not imply that the pair $(Z_u, Z_{u'})$ has the same distribution as the reverse pair $(Z_{u'}, Z_u)$.

Isotropy has a similar meaning in relation to a different group—orthogonal, special orthogonal or unitary—acting on the domain, which is necessarily a Euclidean space with an inner product.

1. The domain is Euclidean space, either \mathbb{R}^d or \mathbb{C}^d , or some subset of Euclidean space that is closed under the group;
2. The orthogonal group, or possibly the special orthogonal group with positive determinant, acts on the domain, sending u to gu ;
3. The action on the domain sends the original process to $Z^g(u) = Z(gu)$ by composition;
4. The process is isotropic if each Z^g has the same distribution as Z .

In either case, the group action by composition is described schematically by the compositional diagram such as

$$\mathcal{U} \xrightarrow{g} \mathcal{U} \xrightarrow{Z} \mathbb{R}$$

which shows the process as a function $\mathcal{U} \rightarrow \mathbb{R}$ with the group acting on the domain.

Sometimes it is necessary to ask for clarification whether the full orthogonal group, including reflections, is intended. Sometimes the domain may be a proper subset of Euclidean space on which the group acts, for example, the unit circle or the unit disk in the complex plane or the unit sphere in \mathbb{R}^3 .

Ordinarily in applied work, the domain has no natural origin, so it is better described as an affine space. In such applications isotropy is not a natural requirement on its own, but the Euclidean group of proper rigid motions (translation plus rotation) is very natural. Depending on the setting, reflections may or may not be included.

A zero-mean complex Gaussian process is stationary if and only if $K(u, u') = G(u - u')$ for some function G such that $G(-u) = \bar{G}(u)$. In the case of a real Gaussian process, G is real, and therefore symmetric. A zero-mean complex Gaussian process is stationary and isotropic if and only if $K(u, u') = G(\|u - u'\|)$ for some real-valued function G . Each function G is necessarily positive definite.

It follows that every stationary isotropic complex Gaussian process $Z = X + iY$ is a pair of independent isotropic real Gaussian processes $X \sim Y$ having the same distribution. Conversely, a pair $(X, Y) \mapsto X+iY$ of independent and identically distributed stationary isotropic real-valued Gaussian processes determines a complex Gaussian process. The situation for stationary non-isotropic processes is different.

Finally, a process is said to be reversible if $Z[-U] \sim Z[U]$ for every sample. In this setting, the group $\{\pm 1\}$ acts on the domain $u \mapsto \pm u$, and the process is invariant. Usually, the domain is $\mathcal{U} = \mathbb{R}$ representing time, in which case the process is said to be time-reversible. Every stationary real-valued Gaussian process is time-reversible; a stationary complex-valued Gaussian process is time-reversible only if the covariance function is real.

16.2.2 Stationarity on Increments

A process that is not stationary may nevertheless be stationary in a restricted sense. For application to comparative experiments, stationarity on increments is all that is really needed.

An increment is technically a two-point Dirac difference measure $\delta_u - \delta_v$, which assigns mass $+1$ to $\{u\}$ and -1 to $\{v\}$. The total mass is zero. The value on increments is defined by integration $Z(\delta_u - \delta_v) = Z(u) - Z(v)$.

Let u_1, \dots, u_n and v_1, \dots, v_n be points in the domain, and let

$$Z[\mathbf{u}] - Z[\mathbf{v}] = (Z(u_1) - Z(v_1), \dots, Z(u_n) - Z(v_n))$$

be the vector of increments. The process is *stationary on increments* if, for each integer n and each h in the domain, $Z[\mathbf{u} + h] - Z[\mathbf{v} + h]$ has the same distribution as $Z[\mathbf{u}] - Z[\mathbf{v}]$. A stationary process is automatically stationary on increments.

On $\mathcal{U} = \mathbb{R}^d$, the zero-mean Gaussian process with covariance

$$K(u, v) = \|u\| + \|v\| - \|u - v\|$$

is not stationary. It is, however, stationary on increments.

A process that is defined on increments automatically has a point-wise extension $u \mapsto Z(u) - Z(0)$, but stationarity on increments does not imply the existence of a stationary extension.

16.2.3 Stationary Process on $\mathbb{Z} \pmod k$

Let $k \geq 2$ be a positive integer. The domain $\mathcal{U} = \mathbb{Z} \pmod k$ means that \mathcal{U} is the finite set $\{0, 1, \dots, k-1\}$, which, with addition modulo k , is a commutative group. A zero-mean complex Gaussian process on \mathcal{U} is a random variable $Z = (Z_0, \dots, Z_{k-1})$ whose joint distribution is complex Gaussian in \mathbb{C}^k . The non-zero covariances $K_{r,s} = \text{cov}(Z_r, \bar{Z}_s)$ form a $k \times k$ matrix which is positive-definite and Hermitian, i.e., $K_{r,s} = \bar{K}_{s,r}$.

The process is stationary if the distribution is invariant with respect to translation modulo k ; in the case of a Gaussian process, this means $K_{r,s}$ is a function of $rs^{-1} = r-s$. Since the group action on pairs $(r, s) \in \mathcal{U}^2$ has k orbits, i.e., $\mathcal{O}_t = \{(r, s) : r-s=t\}$ for each $t \in \mathcal{U}$, stationarity implies one value for each orbit:

$$K_{0,0} = K_{1,1} = \dots = K_{k-1,k-1} \geq 0; \quad K_{0,r} = K_{1,r+1} = \dots = K_{k-1,k+r-1}.$$

The orbits also come in conjugate pairs, $\{\mathcal{O}_r, \mathcal{O}_{-r}\}$ with $-r \equiv k-r$ such that $K_{0,r} = \bar{K}_{0,k-r}$. The zero orbit is self-conjugate, so values on the diagonal are real (and positive). If $k \geq 2$ is even, orbit $k/2$ is also self-conjugate, so $K_{0,k/2}$ is also real.

For $k=2$, the covariance matrix of a stationary process

$$K = \begin{pmatrix} K_0 & K_1 \\ K_1 & K_0 \end{pmatrix} = K_0 \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$$

has two distinct values; both orbits are self-conjugate, so both values are real. Stationarity implies that the real and imaginary parts are independent and identically distributed processes; the reverse-time process has the same distribution as the original.

For $k = 3$, the orbits are labelled $-1, 0, 1$ with conjugate values on conjugate orbits: $K_{-1} = \bar{K}_1$. The covariance matrix

$$K = \begin{pmatrix} K_0 & K_1 & \bar{K}_1 \\ \bar{K}_1 & K_0 & K_1 \\ K_1 & \bar{K}_1 & K_0 \end{pmatrix} = K_0 \begin{pmatrix} 1 & \rho & \bar{\rho} \\ \bar{\rho} & 1 & \rho \\ \rho & \bar{\rho} & 1 \end{pmatrix}$$

has three distinct values, one real and one complex-conjugate pair. The reverse-time process is Gaussian with covariance $K' = \bar{K}$. Positive definiteness implies, $K_0 \geq 0$, $|\rho| \leq 1$ plus the determinantal condition $1 - 3|\rho|^2 + 2\Re(\rho^3) \geq 0$. These are equivalent to the conditions $K_0 \geq 0$ and ρ in the equilateral triangle with vertices at the primitive roots $\{1, e^{2\pi i/3}, e^{-2\pi i/3}\}$. See Exercise 16.1 for an explanation. The condition $|\rho| \leq 1/2$ is sufficient but not necessary for positive-definiteness.

16.3 Stationary Gaussian Time Series

16.3.1 Spectral Representation

Any process whose domain is the set of real numbers may be regarded as a time series. To underline the connection, points in the domain are denoted by t . If time is continuous but values are recorded at constant temporal increments, the recorded process is a linear transformation of a continuous-time series. The linear transformation is either a transformation by restriction to the integers $\mathbb{Z} \subset \mathbb{R}$, or a transformation by integration over unit intervals. In either case, the continuous series determines the distribution of the linear transformation. All processes discussed here are defined in continuous time.

Let $\xi(t) = e^{i\omega t}$ be the complex harmonic function with frequency ω and period, or wavelength, $2\pi/\omega$. The Hermitian function

$$K_\omega(t, t') = e^{i\omega(t-t')} = \xi(t)\bar{\xi}(t'),$$

which is the Hermitian outer product $\xi\xi^*$ of the ξ with itself, is positive definite of rank one. Accordingly, if μ is a non-negative finite measure on frequencies, the convex combination

$$K_\mu(t, t') = \int_{-\infty}^{\infty} e^{i\omega(t-t')} d\mu(\omega)$$

is positive definite Hermitian. As a function of $t - t'$, it is necessarily the covariance of a stationary Gaussian time series. Moreover, every stationary Gaussian process, real or complex, has an associated spectral measure. Finiteness of the spectral measure ensures that the covariance is finite on the diagonal: $K_\mu(t, t) < \infty$.

In the algebra that follows it is helpful to split the spectral measure into symmetric and skew-symmetric parts $\mu = \mu_{\text{sym}} + \mu_{\text{alt}}$ as follows:

$$2\mu_{\text{sym}}(A) = \mu(A) + \mu(-A) \quad (16.1)$$

$$2\mu_{\text{alt}}(A) = \mu(A) - \mu(-A). \quad (16.2)$$

The symmetric part is non-negative. The alternating part is a signed measure such that $-1 \leq (d\mu_{\text{alt}}/d\mu_{\text{sym}})(\omega) \leq 1$ for every ω . If we write t for the temporal difference, the result of this decomposition of the measure is

$$\begin{aligned} K_\mu(t) &= \int_{-\infty}^{\infty} \cos(\omega t) d\mu_{\text{sym}}(\omega) + i \int_{-\infty}^{\infty} \sin(\omega t) d\mu_{\text{alt}}(\omega) \\ &= K_\mu^{\text{sym}}(t) + i K_\mu^{\text{alt}}(t), \end{aligned}$$

which is a decomposition of K into real and imaginary parts. Every pair of symmetric and alternating measures such that $|\mu_{\text{alt}}| \leq \mu_{\text{sym}}$ gives rise to a positive definite Hermitian function, and vice-versa.

To any spectral measure μ there corresponds a family of measures $\mu(\cdot - a)$ generated by frequency translation by $a \in \mathbb{R}$. Spectral translation automatically generates a family of related covariance functions. If $K_\mu(t)$ is the covariance function associated with μ , the covariance function associated with the translated measure is

$$\begin{aligned} K_{\mu,a}(t) &= \int_{-\infty}^{\infty} e^{i\omega t} d\mu(\omega - a) \\ &= e^{iat} K_\mu(t). \end{aligned}$$

In particular, if μ is symmetric so that K_μ is real symmetric, the real part $\Re(K_{\mu,a}(t)) = K_\mu(t - t') \cos(a(t - t'))$ is also real symmetric and positive definite, while $K_\mu(t - t')e^{ia(t-t')}$ is positive-definite Hermitian. Each of the associated Gaussian processes, real or complex, is stationary; the real processes are also time-reversible, but the complex processes are not. The time-reversed complex process has the conjugate covariance $K_\mu(t - t')e^{-ia(t-t')}$ with frequency-shift parameter $-a$.

16.3.2 Matérn Class

The standard Matérn spectral measure with index v is symmetric with density

$$d\mu_{\text{sym}}(\omega) = \frac{d\omega}{(1 + \omega^2)^{v+1/2}}.$$

For $\nu > 0$, the measure is finite and proportional to the Student t distribution on 2ν degrees of freedom. The standard Matérn covariance function is the characteristic function of this distribution, which is real and symmetric. The characteristic function of the translated Student t distribution is proportional to

$$K_{\nu,a}(t) = |t|^\nu \mathcal{K}_\nu(|t|) \times (\cos(at) + i \sin(at)),$$

where \mathcal{K}_ν is the Bessel function of order ν .

The decision to generate an asymmetric spectral measure by translation of a symmetric measure is convenient but arbitrary. One other choice that pairs reasonably well with the Matérn family is

$$d\mu_{\text{alt}}(\omega) = \frac{d\omega}{(1 + \omega^2)^{\nu+1/2}} \times \frac{2b\omega}{1 + \omega^2}.$$

The condition $-1 \leq b \leq 1$ implies that the skew factor $2b\omega/(1+\omega^2)$ lies in $[-1, 1]$, so $|\mu_{\text{alt}}| \leq \mu_{\text{sym}}$. Other possibilities for the skew factor include $b\omega/(1 + \omega^2)^{1/2}$.

For the spectral measures shown above, the covariance function is proportional to

$$K_\mu(t) = |t|^\nu \mathcal{K}_\nu(|t|) (1 + ibt/(\nu + 1/2)). \quad (16.3)$$

For a derivation of the real part, see Stein (1999, section 2.10) or Exercises 16.2–16.6. In applications, t is replaced with t/ρ for some temporal range ρ .

16.4 Stationary Spatial Process

16.4.1 Spectral Decomposition

We assume in this section that the domain is \mathbb{R}^d for some $d \geq 1$. In most examples, the domain is also assumed to be Euclidean with an inner product and a norm, so that the orthogonal group may act on it. To distinguish space from time, particularly in the case $d = 1$, points in the domain are called sites and are denoted by x .

The frequency vector $\omega = (\omega_1, \dots, \omega_d)$ acts as a linear functional on the spatial domain, so that the scalar product $\omega x \equiv \omega'x$ is the value at x . The Hermitian function

$$K_\omega(x, x') = e^{i\omega'(x-x')},$$

which is the Hermitian outer product $\xi \xi^*$ of the function $\xi(x) = e^{i\omega' x}$ with itself, is positive definite of rank one. For each finite non-negative measure μ on frequencies, the convex combination

$$K_\mu(x, x') = \int_{\mathbb{R}^d} e^{i\omega'(x-x')} d\mu(\omega)$$

is positive-definite Hermitian. As a function of $x - x'$, it is necessarily the covariance of a stationary Gaussian process. Moreover, every stationary Gaussian process has an associated spectral measure.

To any spectral measure μ there corresponds a family of frequency-shifted measures $\mu(\cdot - a)$ generated by translation by $a \in \mathbb{R}^d$. If $K_\mu(x)$ is the covariance function associated with μ , the covariance function associated with the translated measure at spatial displacement x is

$$\begin{aligned} K_{\mu,a}(x) &= \int_{\mathbb{R}^d} e^{i\omega' x} d\mu(\omega - a) \\ &= e^{ia' x} K_\mu(x). \end{aligned}$$

Note that the exponent in the complex harmonic modulation factor is the inner product $a'x$, or $a'(x - x')$, of the frequency shift vector a with the spatial displacement $x - x'$ of the two sites.

If the spectral measure is rotationally symmetric, the covariance function is also rotationally symmetric, which means that $K_\mu(x)$ is a function of $\|x\|$. Both $K_\mu(x - x')$ and the real part of the frequency-shifted covariance are also real symmetric and positive definite, while $K_{\mu,a}(x - x')$ is positive-definite Hermitian.

The spectral measure may be decomposed into symmetric and alternating parts as defined in (16.1), so that

$$-1 \leq \frac{d\mu_{\text{alt}}}{d\mu_{\text{sym}}}(\omega) = -\frac{d\mu_{\text{alt}}}{d\mu_{\text{sym}}}(-\omega) \leq 1.$$

Since μ_{sym} is even and μ_{alt} is odd, the associated covariance function is

$$\begin{aligned} K_\mu(x) &= \int_{-\infty}^{\infty} \cos(\omega x) d\mu_{\text{sym}}(\omega) + i \int_{-\infty}^{\infty} \sin(\omega x) d\mu_{\text{alt}}(\omega) \\ &= K_\mu^{\text{sym}}(x) + i K_\mu^{\text{alt}}(x), \end{aligned}$$

where x is the spatial difference vector for two sites. By construction $K_\mu^{\text{sym}}(-x) = K_\mu^{\text{sym}}(x)$ is even, whereas $K_\mu^{\text{alt}}(-x) = -K_\mu^{\text{alt}}(x)$ is odd.

For a simple illustrative example, the first-order Taylor expansion about $a = 0$ of the shifted Matérn measure

$$\begin{aligned}\frac{d\omega}{(1 + \|x - a\|^2)^{d/2+\nu}} &= \frac{d\omega}{(1 + \|a\|^2 + \|\omega\|^2 - 2a'\omega)^{d/2+\nu}} \\ &= \frac{d\omega}{(1 + \|\omega\|^2)^{d/2+\nu}} + \frac{(d/2 + \nu)2a\omega d\omega}{(1 + \|\omega\|^2)^{d/2+\nu+1}} + o(\|a\|)\end{aligned}$$

is a $\mu_{\text{sym}} + \mu_{\text{alt}}$ decomposition in which μ_{sym} is also orthogonally invariant.

16.4.2 Matérn Spatial Class

For general $\nu > 0$, the Matérn spectral measure on \mathbb{R}^d is finite and radially symmetric with density

$$d\mu(\omega) = \frac{\Gamma(\nu + d/2)}{\pi^{d/2}} \frac{d\omega}{(1 + \|\omega\|^2)^{\nu+d/2}}. \quad (16.4)$$

The standard Matérn covariance function is

$$M_\nu(x - x') = \|x - x'\|^\nu \mathcal{K}_\nu(\|x - x'\|),$$

where $M_\nu(x)$ is the characteristic function of μ . This covariance function is real symmetric and rotationally invariant, so the associated zero-mean Gaussian process is isotropic in \mathbb{R}^d . If it is complex, the real and imaginary parts are independent and identically distributed Matérn processes with covariance function $M_\nu/2$.

The spectral measure is multivariate Student t , which has finite moments up to order r strictly less than 2ν . The number of moments is also the number of derivatives of M_ν at the origin, which controls the number of derivatives of the process, so larger values of ν give rise to processes with smoother trajectories. For example, the Matérn process with $\nu = 1/2$ has trajectories that, with probability one, are everywhere continuous but nowhere differentiable; the process with $\nu = 3/2$ has one continuous derivative everywhere, but no second derivative. For practical work, it is seldom advisable to consider processes having more than one or two derivatives, so $0 < \nu \leq 2$ is the range of most interest.

Spectral Convolution

Let $\mu^{*2} = \mu \star \mu$ be the two-fold convolution of the Matérn measure with itself. As a consequence of the definition, the characteristic function of a convolution is

the product of the characteristic functions. Thus, the covariance function associated with μ^{*2} is

$$M_v^2(x - x') = \|x - x'\|^{2v} \mathcal{K}_v^2(\|x - x'\|).$$

The behaviour of M_v^2 near the origin is the same as that of M_v , so the two processes are equally smooth. Convolution extends to higher-order products such as $M_v M_{v'} M_{v''}$, in which case the behaviour near the origin is governed by $\min(v, v', v'')$.

Frequency Translation

Frequency translation is a special case of spectral convolution. The characteristic function of the frequency-shifted measure $d\mu(\omega - a)$ gives rise to the Hermitian covariance

$$M_{v,a}(x - x') = M_v(x - x') e^{ia'(x - x')}.$$
 (16.5)

This defines a complex-valued process, $Z(\cdot)$, which is stationary but not isotropic in \mathbb{R}^d . The real and imaginary parts are identically distributed processes with covariance $\Re M_{v,a}$, but they are not independent unless $a = 0$. Part of the effect of anisotropy can be understood from the covariances of sums and differences at sites x, x' :

$$\begin{aligned} \text{cov}(Z(x) + Z(x'), \bar{Z}(x) + \bar{Z}(x')) &= 2M_v(0) + 2M_v(x - x') \cos(a'(x - x')); \\ \text{cov}(Z(x) - Z(x'), \bar{Z}(x) - \bar{Z}(x')) &= 2M_v(0) - 2M_v(x - x') \cos(a'(x - x')); \\ \text{cov}(Z(x) + Z(x'), \bar{Z}(x) - \bar{Z}(x')) &= -2i M_v(x - x') \sin(a'(x - x')). \end{aligned}$$

This latter is zero for sites such that $x - x'$ perpendicular to a , and a damped sinusoid for $x - x'$ parallel to a .

The real part of the frequency-shifted covariance

$$\Re M_{v,a}(x - x') = M_v(x - x') \cos(a'(x - x'))$$

is also positive definite. It defines a real-valued stationary process $Y(x) = \Re Z(x)$, which is also stationary but not isotropic. Both the real and the complex processes exhibit wave-like behaviour, with parallel waves that are perpendicular to a with spatial frequency $|a|$. However, the real process is symmetric in the sense of reversibility $Y(x) \sim Y(-x)$ and $Y[x_1, \dots, x_n] \sim Y[x_n, \dots, x_1]$.

Decomposition of Spectral Measure

To a certain extent, the relation between μ_{alt} and μ_{sym} may be decided in arbitrary fashion. Apart from frequency translation, there are a few other mathematically natural choices such as

$$d\mu_{\text{alt}}(\omega) = d\mu_{\text{sym}}(\omega) \times \frac{2a'\omega}{1 + \|\omega\|^2},$$

where $a'\omega$ is the scalar product of Euclidean vectors. The skew perturbation is the stereographic projection from the unit sphere in \mathbb{R}^{d+1} into \mathbb{R}^d of the spherical harmonic of degree one having polar vector $a \in \mathbb{R}^d$ in the equatorial plane. Unlike frequency translation in (16.5), the polar condition $\|a\| \leq 1$ is required for positivity of the density on the sphere, so $|\mu_{\text{alt}}| \leq \mu_{\text{sym}}$.

For the spectral measure $\mu_{\text{sym}} + \mu_{\text{alt}}$, the covariance function is

$$M_v(\|x - x'\|)(1 + ia'(x - x')/(v + d/2)). \quad (16.6)$$

In addition to the index and the range parameter which is not shown, this covariance also depends linearly on the polar vector. The effect of the polar asymmetry on the covariance of sums and differences is similar to (16.5) except that the sinusoids are replaced by linear functionals.

Domain Restriction

The reason for choosing the dimension-dependent power in the denominator of (16.4) is partly to guarantee integrability, but that reason is not sufficient to explain this particular choice. The real reason is that the index $v + d/2$ ensures that the measures μ_d for different spaces are mutually compatible in the sense that $\mu_{d+1}(A \times \mathbb{R}) = \mu_d(A)$ for every $d \geq 1$ and arbitrary subsets $A \subset \mathbb{R}^d$. See Exercises 16.4–16.6.

Compatibility of spectral measures is more a matter of convenience than logical necessity. It implies that if Z is a Matérn process with index v on \mathbb{R}^{d+1} , the restriction of Z to a lower-dimensional affine subspace is also a Matérn process having the same index and range parameter. This sort of inter-dimensional compatibility is convenient for discussions about differentiability, which is governed by v and not by d . Similar remarks hold for the frequency-shifted covariance (16.5).

The situation for μ_{alt} and the covariance function (16.6) is more complicated because the effect on the polar vector of the subspace restriction must also be taken into account. Since a acts as a linear functional on the domain, it is natural to associate with each subspace restriction the corresponding orthogonal projection. As they are written, the measures $\mu_{\text{alt},d}$ are not mutually compatible in that sense. Compatibility is restored if we set a finite maximum for d and replace a with

$a' = a/(v + d/2)$. Otherwise, we can regard a as an infinitesimal generator for a perturbation; see Sect. 16.6.3.

16.4.3 Illustration by Simulation

Figure 16.1 shows two independent simulations of the zero-mean complex-valued Gaussian process using the isotropic Matérn covariance function with index $v = 1$. At each point x on a 30×30 triangular grid in the square $0 \leq x_1, x_2 \leq 10$, the arrow shows the magnitude and direction of the field $Z(x)$ at that point. In fact, the plotted value is not $Z(x)$ but the deviation of $Z(x)$ from the sample average. This was done in order to reduce visual clutter and to focus attention on spatial variability.

Trajectories of the Matérn process with $v > 1$ are differentiable. Although not quite differentiable, the process with $v = 1$ is relatively smooth so that streamlines can be traced visually in Fig. 16.1. Both simulations are on a square window $0 \leq x_1, x_2 \leq 10$, with range parameter $\rho = 5$ in the first and $\rho = 1$ in the second. The large range parameter means that the most distant pairs are moderately highly correlated with correlation 0.14. In the second plot, the most distant pairs are essentially independent. The first plot can be viewed as a five-fold magnification of a part of the second process.

The simulations for $v = 0.5$ in Fig. 16.2 are in the same format. They are considerably rougher, and the streamlines more ragged. The correlation for the most distant pairs in the first plot is 0.06.

In the isotropic case, the covariance function is real, which means that the real and imaginary components of the field are independent and identically distributed. In that respect, the visual impression may be misleading.

Four anisotropic zero-mean processes are illustrated in Figs. 16.3 and 16.4. These have the frequency-shifted Matérn covariance function

$$\text{cov}(Z(x), \bar{Z}(x')) = M_v(\|x - x'\|/\rho) e^{i\theta'(x-x')/\rho}$$

with index v , range ρ , and frequency-shift vector θ . The real and imaginary parts of the process have the same distribution, but they are not independent. The frequency shift vector governs both the magnitude and the direction of the anisotropy, as is easily seen by visual inspection of the simulations.

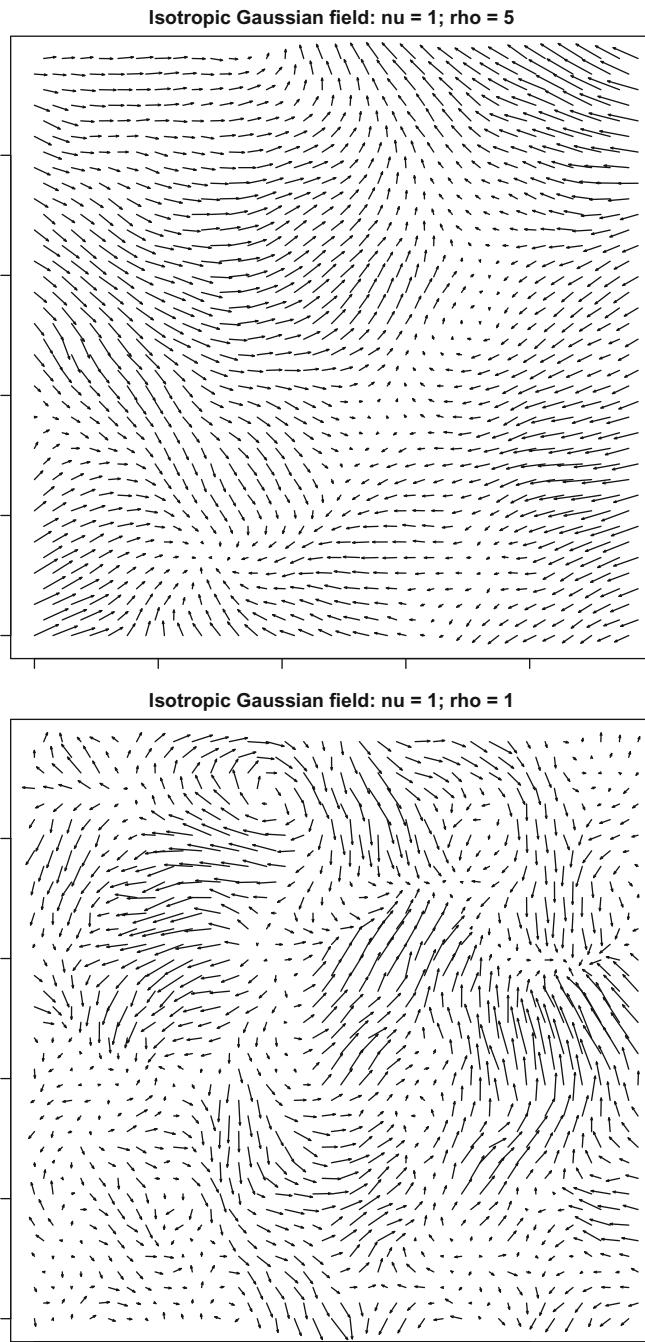


Fig. 16.1 Two simulations of an isotropic Gaussian field

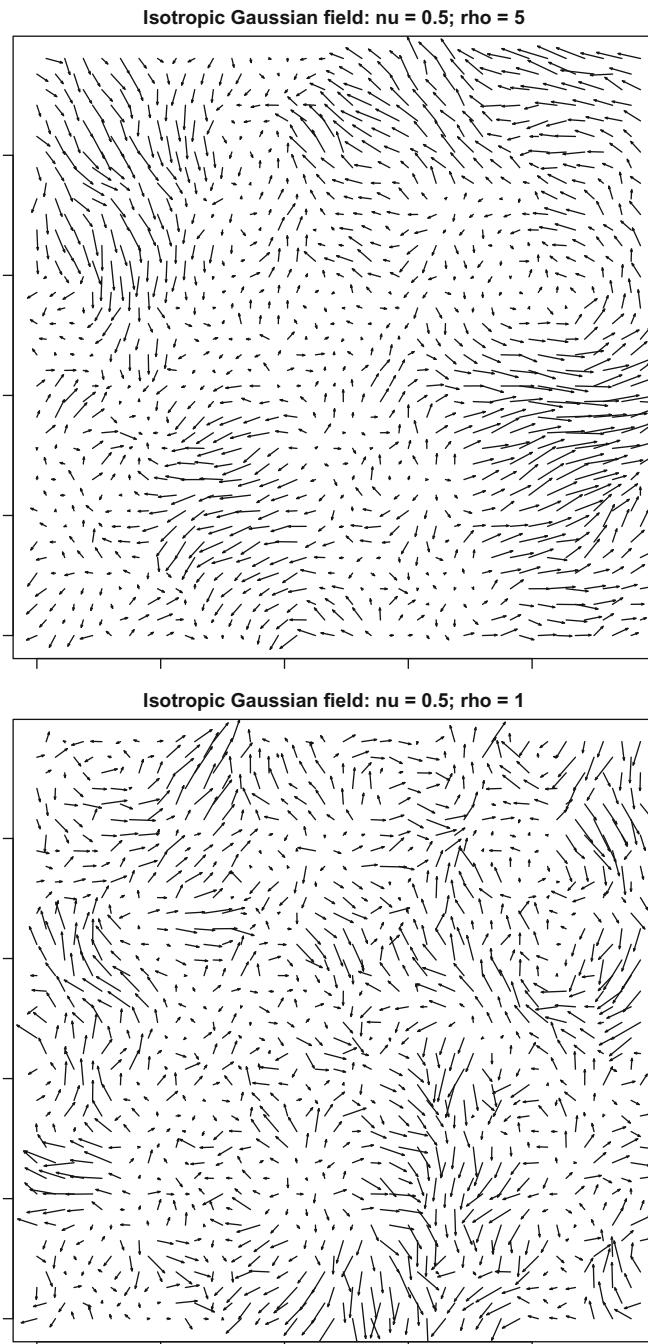


Fig. 16.2 Two simulations of an isotropic Gaussian field

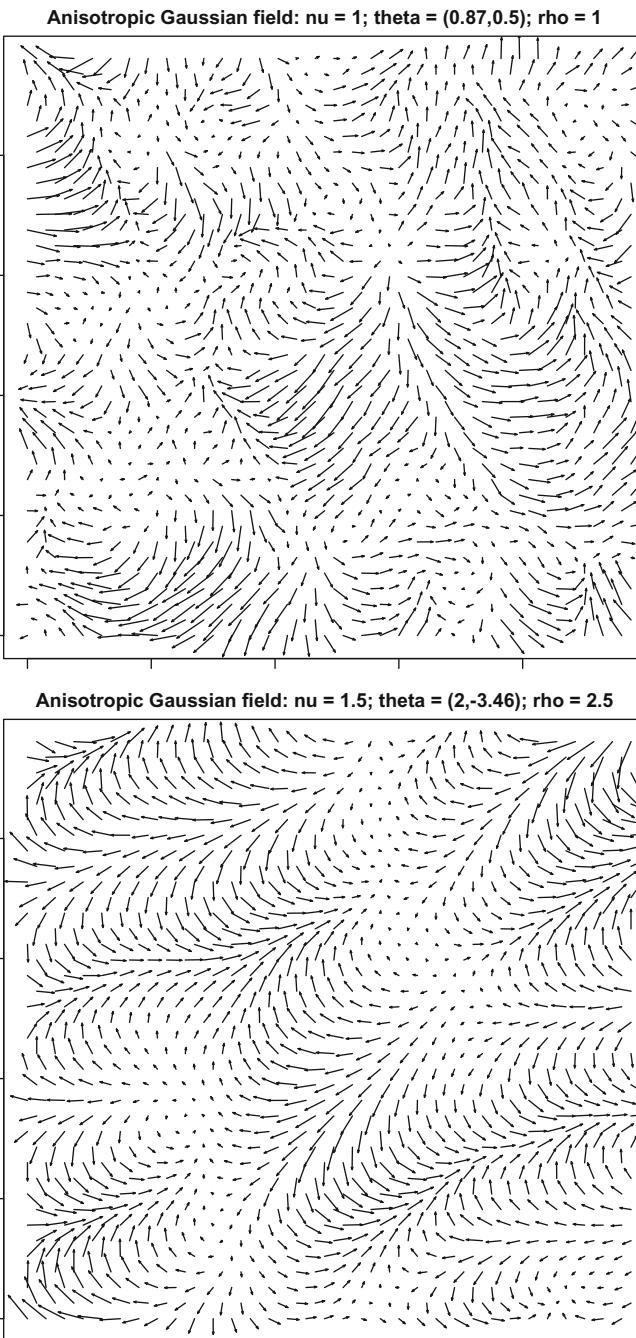


Fig. 16.3 Two anisotropic Gaussian fields with $\nu = 1$ and different anisotropy vectors

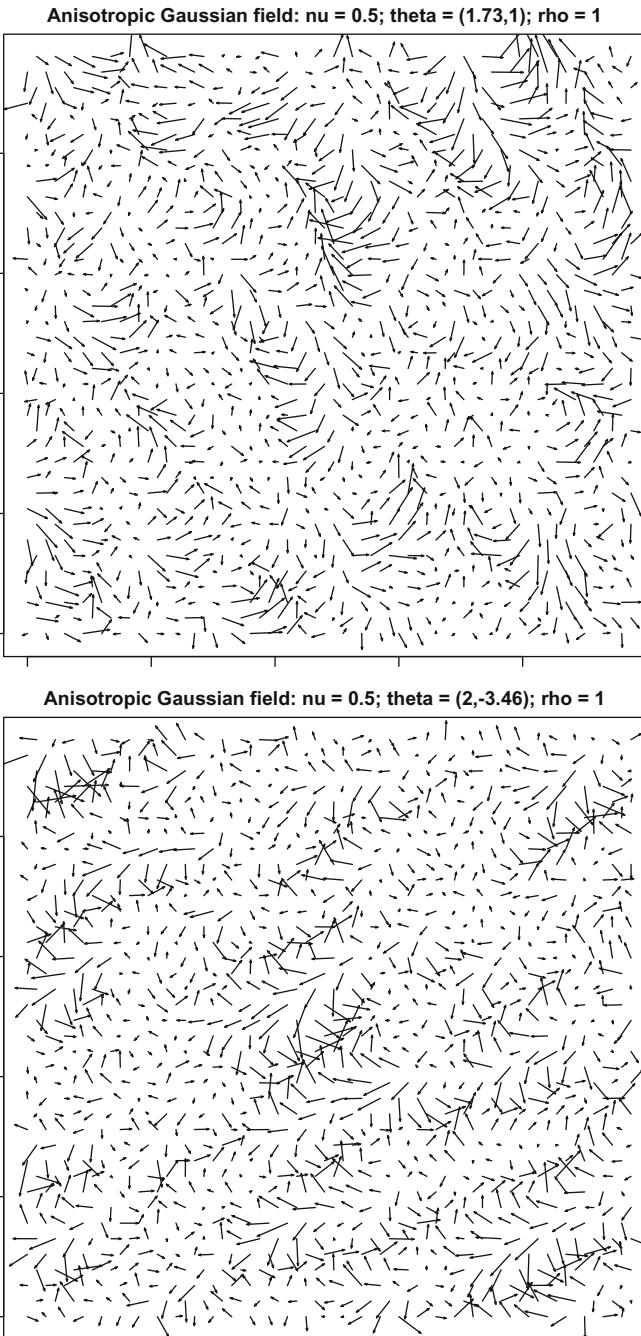


Fig. 16.4 Two anisotropic Gaussian fields with $\nu = 1/2$ and different anisotropy vectors

16.5 Covariance Products

16.5.1 Hadamard Product

Let $Z = (Z_1, \dots, Z_n)$ and $W = (W_1, \dots, W_n)$ be zero-mean independent Gaussian vectors in \mathbb{C}^n with covariance matrices K and C respectively. Then the covariances of the products are

$$\text{cov}(Z_r W_r, Z_s W_s) = 0; \quad \text{cov}(Z_r W_r, \bar{Z}_s \bar{W}_s) = K_{rs} C_{rs}.$$

The non-zero covariance is the component-wise product of the two covariance matrices. In the case of real-valued random variables, where there is no distinction between Z and \bar{Z} , the conjugated version prevails even if the distributions are non-Gaussian.

The preceding derivation is the simplest proof of Schur's product theorem, which states that the Hadamard product of positive-definite matrices is itself positive definite. The most immediate consequence for Gaussian processes is that the functional product $K(x, x')C(x, x')$ of two covariance functions *on the same space* is also a covariance function. In particular, $C = K$ implies that the squared function $K^2(x, x')$ is positive-definite Hermitian, and $C = \bar{K}$ implies that the squared modulus $|K|^2(x, x')$ is real symmetric and positive-definite.

As a first example, suppose that C and K are Matérn covariances (16.5) with index v and frequency shift vectors a and b respectively. The products CK and $C\bar{K}$ are

$$(CK)(x, x') = M_v^2(x - x') e^{i(a+b)(x-x')}, \quad (C\bar{K})(x, x') = M_v^2(x - x') e^{i(a-b)(x-x')},$$

so these are both positive definite Hermitian. In fact $M_v^2(x)$ is the characteristic function of μ^{*2} , the two-fold convolution of the Matérn spectral measure (16.4), which is spherically symmetric. and $(CK)(x)$ is the characteristic function of the convolution of the frequency-translated measures. The real part

$$\Re(CK)(x, x') = M_v^2(x - x') \times \cos((a' + b')(x - x'))$$

is a positive-definite symmetric function with bi-polar vector $\pm(a + b)$. It is the covariance function of a real Gaussian process that is stationary in \mathbb{R}^d . This process is not isotropic, but it is invariant with respect to orthogonal transformations that preserve the bi-polar vector.

For a second example, suppose that C and K are skew-Matérn covariances (16.6), with polar vectors a and b respectively. The product is

$$M_v^2(\|x - x'\|) \left(1 + \frac{ia(x - x')}{v + d/2}\right) \left(1 + \frac{ib(x - x')}{v + d/2}\right),$$

which is positive-definite Hermitian on \mathbb{R}^d . The real part of the product is symmetric and positive definite:

$$M_v^2(\|x - x'\|) \left(1 - \frac{Q(x - x')}{(v + d/2)^2}\right), \quad (16.7)$$

where $x \mapsto Q(x)$ is a quadratic form in x whose maximum absolute eigenvalue is less than one. Since a non-negative linear combination of positive-definite functions is positive definite, it follows that (16.7) is positive definite for every quadratic form Q whose operator norm is less than one. The operator norm is the largest absolute eigenvalue.

16.5.2 Separable Products and Tensor Products

Let $K(u, u')$ be a positive definite function on \mathcal{U} , and $C(v, v')$ a positive definite function on \mathcal{V} . Let the eigenvalues and eigenfunctions of K be $\{\lambda_i, \xi_i(u)\}$, so that $\int_{\mathcal{U}} K(u, u') \xi_i(u') du = \lambda_i \xi_i(u)$. Likewise, on \mathcal{V} , let the eigenvalues and vectors of C be $\{\rho_j, \zeta_j(v)\}$.

The product space $\mathcal{U} \times \mathcal{V}$ consists of ordered pairs (u, v) , and the covariance product

$$K_2((u, v), (u', v')) = K(u, u') C(v, v') \quad (16.8)$$

is a natural candidate for a covariance function on the product space. An elementary calculation shows that $\xi_i(u) \zeta_j(v)$ is an eigenfunction of the product:

$$\begin{aligned} & \int K(u, u') C(v, v') \xi_i(u') \zeta_j(v') du' dv' \\ &= \int_{\mathcal{U}} K(u, u') \xi_i(u') du' \int_{\mathcal{V}} C(v, v') \zeta_j(v') dv' \\ &= \lambda_i \rho_j \xi_i(u) \zeta_j(v). \end{aligned}$$

Thus, the eigenvalues of the product are the products of the eigenvalues. Hence the covariance product is positive definite on the product space, and the rank of the product is the product of the ranks.

One important special case occurs when the spaces \mathcal{U} and \mathcal{V} are equal. The covariance product (16.8) restricted to the diagonal of $\mathcal{U} \times \mathcal{U}$ is nothing more than the Hadamard product of two covariance functions on \mathcal{U} . Positive definiteness of the Hadamard product follows trivially by diagonal restriction.

A covariance function on the product space is said to be *separable* if it is expressible as a single product, as in (16.8). A statistical covariance model is said to be separable if each covariance function in the model is separable. For example, the

space-time covariance model consisting of the infinite set of covariance functions

$$\{\sigma^2 e^{-|t-t'|/\lambda} M_{\nu,a}(x, x'): \sigma^2, \lambda, \nu > 0; a \in \mathbb{R}\}$$

for M_μ in (16.5), is space-time separable.

A tensor product is a linear combination of pairwise products taken from two basis sets, say K_0, K_1 on \mathcal{U} and C_0, C_1, C_2 on \mathcal{V} , and the tensor product space is the set of all such combinations. Most statistical applications employ tensor products in this form as models for covariances, with constraints on the coefficients to ensure positive definiteness. A tensor product is typically not separable.

The matrix formed by restriction of the separable product to a finite product grid is the Kronecker product of the marginal restrictions, and the inverse of a Kronecker product is the Kronecker product of the inverses. This fact makes for enormous simplification of statistical calculations related to Gaussian estimation and prediction, and that computational simplicity is the chief attraction of separable covariance models.

Separable covariances have a deservedly poor reputation in applied work for several reasons including the following. Suppose that a spatio-temporal Gaussian process Z is observed at a collection of sites $\mathbf{x} = (x_1, \dots, x_n)$ at times $\mathbf{t} = \{t_1 < \dots < t_k\}$, and it is required to predict the values at the same sites at a later time t_{k+1} using the conditional distribution. Ordinarily, the conditional expected value of $Z(x_1, t_{k+1})$ given the data is a linear combination of all nk observations $Y[\mathbf{x} \times \mathbf{t}]$. However, if the covariance function is separable, the conditional expectation is a linear combination of the values at site x_1 only. See Exercises 16.11 for an outline of a proof. In other words, separability implies that earlier values observed at nearby sites are irrelevant for prediction in this regular grid-like sampling scheme. This consequence of separability is unacceptable for almost any naturally-occurring process. It is particularly egregious for a 2D spatial process to be modelled using a separable product of one-dimensional covariance functions.

16.6 Real Spatio-Temporal Process

16.6.1 Covariance Products

The simplest way to construct a positive-definite covariance on a product space is to begin with a covariance function or set of covariance functions on each space, and to use tensor products. Usually a single product is not sufficient for applied work. Elementary examples can be found in (5.4).

For purposes of illustration, we use the temporal family (16.3) and the spatial family (16.5) with the same index. Writing x and t in place of $x - x'$ and $t - t'$, the outer product is

$$M_v(\|x\|)(1 + iax/(v + d/2)) \times M_v(t)(1 + ibt/(v + 1/2)), \quad (16.9)$$

where $-1 \leq b \leq 1$ is a scalar. This product splits into four sub-products, two real and two imaginary:

$$\begin{aligned} & M_v(\|x\|) M_v(t); \\ & M_v(\|x\|) M_v(t) \times iax/(v + d/2); \\ & M_v(\|x\|) M_v(t) \times ibt/(v + 1/2); \\ & M_v(\|x\|) M_v(t) \times -ax bt/((v + 1/2)(v + d/2)). \end{aligned} \quad (16.10)$$

For a real spatio-temporal process, the two imaginary terms can be discarded. We are left with a linear combination of the two real products,

$$M_v(\|x\|) M_v(t) \left(1 - \frac{ax bt}{(v + 1/2)(v + d/2)} \right). \quad (16.11)$$

Only the first of these is positive definite. Nevertheless, the linear combination is positive definite for all $\|a\| \leq 1$.

The complex temporal process associated with (16.3) is stationary but not reversible. The complex spatial process associated with (16.5) is stationary, but the polar parameter implies a specific directional asymmetry. The real space-time process with covariance (16.11) is both spatially and temporally stationary. If $ab = 0$, the covariance is also space-time symmetric, or space-time reversible, in the sense

$$\text{cov}(Z(\text{site}_0, \text{day}_0), Z(\text{site}_1, \text{day}_1)) = \text{cov}(Z(\text{site}_0, \text{day}_1), Z(\text{site}_1, \text{day}_0)). \quad (16.12)$$

Otherwise, if $ab \neq 0$, the interaction of one temporal asymmetry with one spatial asymmetry in (16.10) or (16.5) implies that the covariance is not symmetric in the above sense. In typical real-world applications such as environmental monitoring or meteorological processes, space-time symmetry in the sense of (16.12) is highly undesirable for obvious reasons.

Since b is a real number, it can be ignored or absorbed into other factors. The product (16.10) is linear in x_1, \dots, x_k with coefficients proportional to the polar coefficients a_1, \dots, a_k . It can therefore be expressed as a linear combination of k similar terms, with coefficients to be estimated from the data.

16.6.2 Examples of Covariance Products

Patterned Covariance Matrices

A process on a finite index set $[k]$ is nothing more than a list of variables $Z = (Z_1, \dots, Z_k)$, one component Z_r for each $r \in [k]$. The covariance function is a matrix $K_{r,s}$, one value for each ordered pair $r, s \in [k]$. In the case of a complex-valued process such that $e^{i\phi}Z$ has the same distribution as Z for each complex unit scalar multiple, the product $Z_r Z_s$ has the same distribution as $e^{2i\phi} Z_r Z_s$, which implies that the expected value is zero. It is sufficient, therefore to record only the expected values of the conjugated products $K_{rs} = E(Z_r \bar{Z}_s)$; This matrix is positive definite Hermitian.

When we refer to $Z: [k] \rightarrow \mathbb{C}$ as a Gaussian process rather than a complex Gaussian vector, we usually mean that the process has certain symmetries, which are seen as patterns in the covariance matrix. Two rather simple examples suffice by way of illustration.

For $k = 2$ and real θ , let $\chi_2(\theta)$ be the skew-symmetric matrix

$$\chi_2(\theta) = \begin{pmatrix} 0 & -\theta \\ \theta & 0 \end{pmatrix}.$$

As it happens, this is the 2×2 real matrix representation of the complex number $i\theta$. Provided that $-1 \leq \theta \leq 1$, the matrix

$$K = I_2 + i \chi_2(\theta) \tag{16.13}$$

is positive-definite Hermitian; the determinant is $1 - \theta^2$. Let Z be zero-mean complex Gaussian with covariance K . Since $K_{11} = K_{22}$, the variances are equal $E(|Z_1|^2) = E(|Z_2|^2)$. Thus, the permuted vector (Z_2, Z_1) has the conjugate covariance $\tilde{K} = I_2 - i \chi(\theta)$, and $(Z_2, -Z_1)$ has the same distribution as Z .

For $k = 3$ and $\theta \in \mathbb{R}^3$, let $\chi_3(\theta)$ be the skew-symmetric matrix

$$\chi_3(\theta) = \begin{pmatrix} 0 & -\theta_3 & \theta_2 \\ \theta_3 & 0 & -\theta_1 \\ -\theta_2 & \theta_1 & 0 \end{pmatrix} = -\chi_3(-\theta). \tag{16.14}$$

Both χ_2 and χ_3 are anti-symmetric: $\chi(-\theta) = \chi'(\theta)$. Provided that $\|\theta\| \leq 1$, the matrix

$$K_\theta = I_3 + i \chi_3(\theta) \tag{16.15}$$

is positive-definite Hermitian. The determinant is $1 - \|\theta\|^2$, and the inverse satisfies

$$(1 - \|\theta\|^2) K_\theta^{-1} = I_3 - \theta \theta' - i \chi_3(\theta).$$

Let Z be zero-mean complex Gaussian with covariance K . Since $K_{11} = K_{22} = K_{33}$, the variances are equal $E(|Z_1|^2) = E(|Z_2|^2) = E(|Z_3|^2)$, so the one-dimensional distributions are equal (standard Gaussian). However, the three two-dimensional distributions are different: each pair has a joint covariance of the type (16.13), but with different parameters.

If $\sigma = (1, 2, 3)$ is the cyclic permutation $1 \mapsto 2 \mapsto 3 \mapsto 1$, the permuted vector σZ with components (Z_2, Z_3, Z_1) does not have the same distribution as Z . But it does have a distribution in the same family with parameter $\sigma\theta = (\theta_2, \theta_3, \theta_1)$. For general permutations, the permuted vector has a distribution in the same family with parameter $\sigma\theta$ if the number of cycles is odd, either one or three, and $-\sigma\theta$ if the number of cycles is even, which necessarily means two.

The family of matrices (16.15) is of interest in physical applications principally because of the following geometric property. For each real orthogonal matrix $g \in \mathcal{O}_3$, the transformed variable gZ has covariance

$$gK_\theta g' = I_3 + ig\chi_3(\theta)g' = I_3 + i \det(g)\chi_3(g\theta), \quad (16.16)$$

where $\det(g) = +1$ for a rotation and -1 for a reflection. In other words, the family of matrices (16.15) is closed under orthogonal conjugation, as is the family of inverses.

Complex Moments

On the planar domain, $\mathcal{U} = \mathbb{R}^2$ or $\mathcal{U} = \mathbb{C}$, the conjugate-product function

$$C_2(u, v) = \bar{u}v = u'v + i(u_1v_2 - u_2v_1) \quad (16.17)$$

is positive-definite Hermitian. Here $u = u_1 + iu_2$ and $v = v_1 + iv_2$ are points in the complex plane, and $u'v = u_1v_1 + u_2v_2$ is the real part of the complex product.

On the vector domain $\mathcal{U} = \mathbb{C}^d \cong \mathbb{R}^{2d}$, the inner-product function

$$C_2(u, v) = u^*v = \text{tr}(u'v) + i(u'_1v_2 - u'_2v_1) \quad (16.18)$$

is, by definition, positive-definite Hermitian. In this setting each domain point is a complex vector $u = u_1 + iu_2$, where u_1, u_2 are vectors in \mathbb{R}^d . In other words, u can be regarded as a $d \times 2$ matrix with real components, and $\text{tr}(u'v)$ is the trace of the matrix product. Although (16.17) and (16.18) are positive definite, they seldom arise in statistical work because of extreme lack of stationarity.

Complex Covariance Product

In the product space $\mathbb{C} \times [2]$, each domain point is an ordered pair (u, r) with $u \in \mathbb{C}$ and $r \in [2]$. If it is convenient, the process values $Z(u, r)$ can be taken in pairs

$W(u) = (Z(u, 1), Z(u, 2))$. In this form, W is a bivariate planar process, i.e., a function $\mathbb{C} \rightarrow \mathbb{C}^2$. If Z happens to be real-valued, W is a function $\mathbb{C} \rightarrow \mathbb{R}^2$ or a function $\mathbb{C} \rightarrow \mathbb{C}$, a planar process also taking values in the plane. By construction, the product of (16.13) on [2] and (16.17) on \mathbb{C} is positive definite Hermitian on the product space. The real part of the product is positive-definite symmetric: it associates with the ordered pair $(u, r), (v, s)$ the real number

$$K_\theta(u, r; v, s) = u' v \delta^{rs} - (u_1 v_2 - u_2 v_1) \chi_2^{rs}(\theta) = K_\theta(v, s; u, r), \quad (16.19)$$

where δ^{rs}, χ_2^{rs} are the components of I_2 and χ_2 respectively. As a covariance function for the \mathbb{R}^2 -valued process W , the matrix-valued covariances are

$$K_\theta(u, v) = \text{cov}(W(u), W(v); \theta) = u' v I_2 - (u_1 v_2 - u_2 v_1) \chi_2(\theta).$$

Although (16.19) is necessarily symmetric, the 2×2 matrix

$$K_\theta(u, v) = K'_\theta(v, u) = K_{-\theta}(v, u)$$

is not symmetric unless $\theta = 0$. In that case, the second term disappears and $K_0(u, s; v, r) = K_0(u, r; v, s)$ exhibits full symmetry, i.e., $\text{cov}(W(u), W(v)) = \text{cov}(W(v), W(u))$.

A 3D Real Process

A 3D process on the domain \mathcal{U} is a vector-valued function $Z: \mathcal{U} \rightarrow \mathbb{R}^3$, which is the same thing as a scalar function $\mathcal{U} \times [3] \rightarrow \mathbb{R}$. To construct such a vector process, let K be a positive-definite Hermitian function on \mathcal{U} :

$$K(u, v) = K_0(u, v) + i K_1(u, v).$$

By definition, K_0 is positive-definite symmetric, and the imaginary part K_1 is skew-symmetric. On the Cartesian product space $\mathcal{U} \times [3]$, the direct product with (16.15) associates with each ordered pair $(u, r), (v, s)$ the complex product

$$(u, r; v, s) \mapsto K(u, v)(\delta^{rs} + i \chi_3^{rs}(\theta)).$$

For $\|\theta\| \leq 1$, this function is positive-definite Hermitian. The real part, which is positive-definite symmetric, associates with each ordered pair $(u, r; v, s)$ the function

$$(u, r; v, s) \mapsto K_0(u, v)\delta^{rs} - K_1(u, v)\chi_3^{rs}(\theta), \quad (16.20)$$

which is the product of the real parts minus the product of the imaginary parts.

To understand what (16.20) implies, let Z be a real-valued Gaussian process on the product space $\mathcal{U} \times [3]$ with covariances (16.20). In other words, $Z(u) = (Z^1(u), Z^2(u), Z^3(u))$ is a vector-valued Gaussian process on \mathcal{U} . The covariance function for the vector process is 3×3 matrix-valued

$$\text{cov}(Z(u), Z(v)) = K_0(u, v)I_3 - K_1(u, v)\chi_3(\theta),$$

which is the transpose of $\text{cov}(Z(v), Z(u))$. For a given collection of points $\mathbf{u} = \{u_1, \dots, u_n\}$, the values $Z[\mathbf{u}]$ may be collected in an $n \times 3$ matrix with components $Z^r(u_i)$. The covariance matrix of $Z[\mathbf{u}]$ is symmetric of order $3n \times 3n$; it is most conveniently arranged as a 3×3 array of $n \times n$ matrices:

$$\begin{aligned} \text{cov}(Z[\mathbf{u}]) &= I_3 \otimes K_0[\mathbf{u}, \mathbf{u}] - \chi_3(\theta) \otimes K_1[\mathbf{u}, \mathbf{u}] \\ &= \begin{pmatrix} K_0 & \theta_3 K_1 & -\theta_2 K_1 \\ -\theta_3 K_1 & K_0 & \theta_1 K_1 \\ \theta_2 K_1 & -\theta_1 K_1 & K_0 \end{pmatrix} [\mathbf{u}, \mathbf{u}]. \end{aligned} \quad (16.21)$$

The diagonal blocks are equal and symmetric; Each off-diagonal block is proportional to $K_1[\mathbf{u}, \mathbf{u}]$, which is skew-symmetric.

On account of (16.16), the orthogonally transformed process gZ has covariance function

$$\text{cov}(gZ(u), gZ(v)) = K_0(u, v)I_3 \mp K_1(u, v)\chi_3(g\theta),$$

where the sign is $-\det(g)$. In other words, the 3D process is not rotationally symmetric, but the θ -indexed family is closed under orthogonal transformation.

16.6.3 Travelling Wave

Any continuous group acting on the spatial domain $g: x \mapsto gx$ can be made to act on the space-time product space, either in a component-wise manner $(x, t) \mapsto (gx, t)$ or in a more complicated manner. The simplest non-component-wise action is that of a wave travelling with constant velocity $g \in \mathbb{R}^d$, so that the group action $(x, t) \mapsto (x - gt, t)$ is a spatial shift proportional to time. Such a transformation on the domain induces a transformation on the process, sending $Z(x, t)$ to $W(x, t) = Z(x - gt, t)$ by composition. Positive-definiteness of the composite covariance function follows automatically from the definition.

If the original process Z has a separable covariance function $K(x, x')C(t, t')$, the transformed covariance function

$$\begin{aligned} \text{cov}(W(x, t), \bar{W}(x', t')) &= \text{cov}(Z(x - gt, t), \bar{Z}(x' - gt', t')) \\ &= K(x - gt, x' - gt') C(t, t') \end{aligned}$$

is a $K \times C$ product, but it is not a space-time separable product. At time t , the process has a random spatial profile whose covariance function is

$$\text{cov} (W(x, t), \bar{W}(x', t)) = K(x - gt, x' - gt) C(t, t).$$

Assuming that K and C are both stationary, the spatial profile is a Gaussian process with covariance $K(x - x') C(0)$, which is spatially stationary. Although the spatial distribution is constant in time, the profile itself is not static unless the spatial factor is constant $C(t) = C(0)$.

As a specific example consider a complex-valued Matérn process in \mathbb{R}^d with polar vector a and covariance function (16.5). The covariance function for the associated wave travelling at velocity $v \in \mathbb{R}^d$ is

$$M_v(\|x - x' - v(t - t')\|) \left(1 + \frac{ia(x - x' - v(t - t'))}{v + d/2} \right). \quad (16.22)$$

Using the temporal covariance (16.3) with $b = 1$, and writing x, t in place of $x - x'$ and $t - t'$, the four components of the product are obtained by replacing x with $x - vt$ in (16.11):

$$M_v(\|x - vt\|) M_v(t); \quad (16.23)$$

$$M_v(\|x - vt\|) M_v(t) \times ia(x - vt)/(v + d/2);$$

$$M_v(\|x - vt\|) M_v(t) \times it/(v + 1/2);$$

$$M_v(\|x - vt\|) M_v(t) \times \frac{-axt + avt^2}{(v + 1/2)(v + d/2)}. \quad (16.24)$$

Since we are interested in real-valued spatio-temporal processes, we focus on the two real parts whose sum is automatically positive definite. Note that ax is the scalar product of the polar vector with the spatial displacement x , and av is the scalar product with the velocity vector.

For application to fluid dynamics, other groups may be relevant, in particular the group of rigid Euclidean motions in \mathbb{R}^d , which allows the wave to rotate as it travels. It is convenient to illustrate the idea for $d = 2$ by regarding $\mathbb{R}^2 \cong \mathbb{C}$ as the spatial domain, so that the group element $g = (\theta, v)$ acts as a rigid Euclidean motion on the space-time domain by

$$g: (x, t) \mapsto (e^{i\theta t} x - vt, t).$$

The group element has two components, $\theta \in [0, 2\pi)$ or $\mathbb{R} \pmod{2\pi}$, and $v \in \mathbb{C}$. The covariance function for the transformed process is obtained by substituting $e^{i\theta t} x - vt$ for $x - vt$ in (16.22) and (16.10). If the polar vector a is also taken as a complex number, ax is the real part of the complex product $a\bar{x}$, not the product of complex numbers.

The concept of a flowing compressible fluid is of central importance in partial differential-equation models governing atmospheric physics and fluid mechanics more generally. Matter, heat and electrical particles are transferred by fluid flow, a phenomenon known as advection. It is only natural to think of v as an advection vector in (16.22). This mechanical portrayal conveys a vivid image but the picture may be seriously misleading. Although some waves do travel with the fluid, there is no essential connection between the fluid velocity and the wave velocity. As the following example demonstrates, the wave velocity may be substantially greater than the fluid velocity—and not necessarily in the same direction.

16.6.4 Perturbation Theory

One way to understand the skew-symmetric contribution in Sect. 16.4 is to view the parameter a (or $a/(v + d/2)$) as a small perturbation of the Matérn spectral measure μ_{sym} . In that case, we can approximate (16.5) by a perturbation generated by the group element e^{iax} acting on the process. As always, $x \mapsto ax$ is the scalar product of vectors in \mathbb{R}^d . In other words, if Z is an isotropic process with covariance $M_v(\|x - x'\|)$, the covariance of the non-isotropic perturbation $W(x) = e^{iax}Z(x)$ is

$$\begin{aligned}\text{cov}(W(x), \bar{W}(x')) &= M_v(\|x - x'\|)e^{ia(x-x')} \\ &= M_v(\|x - x'\|)(1 + ia(x - x') + o(\|a\|)).\end{aligned}$$

By construction, this function is positive definite for all vectors a ; it is an approximation to (16.5) only for small a .

If we also regard the temporal covariance (16.3) as a multiplicative perturbation with parameter b , we arrive at a multiplicative perturbation of the covariance product in the form

$$e^{ia(x-x')+ib(t-t')} = e^{ia(x-vt-(x'-vt'))},$$

with $v \in \mathbb{R}^d$ as velocity vector and $b = -av$ as the scalar product. The perturbation-theory covariance function

$$M_v(\|x - x'\|) M_v(t - t') e^{ia(x-vt-(x'-vt'))}$$

has some of the characteristics of a travelling wave, but it is not the same as (16.22).

16.7 Hydrodynamic Processes

16.7.1 Frame of Reference

A spatio-temporal process $Z(x, t)$ defined for $x \in \mathbb{R}^d$ and $t \in \mathbb{R}$ is called *hydrodynamic* if the state space coincides with the spatial domain. In other words, Z takes values in \mathbb{R}^d , here taken to be Euclidean space with the standard inner product. A typical example for $d = 2$ is wind velocity measured at various points in the plane. Wind speed is a scalar-valued planar process; wind velocity is a vector-valued planar process, which is hydrodynamic.

The coincidence of the state space and the spatial domain leads to symmetry considerations that do not arise otherwise. A frame of reference is needed to record points in the domain, and the same frame of reference is used to record values in the state space. Two individuals observing the same process are at liberty to use different frames of reference, each with its own origin and orthogonal axes. If they observe the process at the same points, with values and points recorded in different frames, it is essential to arrange matters so that their substantive conclusions are identical. If they observe a stationary process at two configurations that are space-time congruent, their substantive conclusions must be identical in distribution. In Newtonian relativity, the frame of reference is arbitrary up to rigid 3D Euclidean motions, and space-time congruence is built in to the mathematics by considerations of group action by composition.

Consider first the static case with $t = 0$ fixed, so that $Z(x) = Z(x, 0)$, which is assumed to be spatially stationary. To any orthogonal transformation $R: \mathbb{R}^d \rightarrow \mathbb{R}^d$ there correspond three transformed processes denoted by ZR , RZ and RZR' as follows:

$$(ZR)(x) = Z(Rx); \quad (RZ)(x) = R(Z(x)); \quad (RZR')(x) = RZ(R'x).$$

The original process associates with each point x in the domain a value $Z(x)$; we denote this diagrammatically by $x \xrightarrow{Z} Z(x)$. The domain-transformed process ZR associates with each point x a value $(ZR)(x) = Z(Rx)$ by composition on the domain:

$$x \xrightarrow{R} Rx \xrightarrow{Z} Z(Rx).$$

The state-space-transformed process RZ associates with each point x a value $(RZ)(x)$ by composition on the state:

$$x \xrightarrow{Z} Z(x) \xrightarrow{R} RZ(x).$$

The hydrodynamically transformed process RZR' associates with each point x a value $(RZR')(x)$ by composition on both the domain and the state:

$$x \xrightarrow{R'} R'x \xrightarrow{Z} Z(R'x) \xrightarrow{R} RZ(R'x).$$

The process is said to be *isotropic* if, for every rotation R , the domain-transformed process ZR has the same distribution as Z . It is *rotationally symmetric* if, for every rotation R , the state-space-transformed process RZ has the same distribution as Z . For present purposes, neither of these symmetry conditions is compelling because any modification of the frame-of-reference in the domain has an equal impact in the state space. If RZR' has the same distribution as Z , we say that the process is *hydrodynamically symmetric*; in other words, the frame of reference is immaterial. Hydrodynamic symmetry is equivalent to the condition that RZ have the same distribution as ZR . It does not imply that ZR has the same distribution as Z , so hydrodynamic symmetry does not imply isotropy or rotational symmetry. The gradient of a real-valued process on \mathbb{R}^d provides the simplest example of a hydrodynamic process: see Exercise 16.17.

Our primary focus in this section is on hydrodynamic processes in \mathbb{R}^2 and \mathbb{R}^3 . The situation for \mathbb{R}^2 is simpler and reasonably well understood, so the algebra is specific to $d = 3$. It is also specific to the proper orthogonal group \mathcal{O}_3^+ consisting of 3D rotations, excluding reflections. Typically, the domain-reflected process $Z(-x)$ has the same distribution as $-Z(x)$, which need not be the same as that of $Z(x)$.

16.7.2 Rotation and Group Action

Let $x = (x_1, x_2, x_3)$ be a non-zero point in \mathbb{R}^3 , and let $Q(x) = \|x\|^2 I_3 - xx'$, so that $Q(x)/\|x\|^2 = I_3 - xx'/\|x\|^2$ is the rank-two orthogonal projection whose kernel includes x . Define the related skew-symmetric matrix

$$\chi(x) = \begin{pmatrix} 0 & -x_3 & x_2 \\ x_3 & 0 & -x_1 \\ -x_2 & x_1 & 0 \end{pmatrix} = -\chi'(x). \quad (16.25)$$

This definition means that $x \mapsto \chi(x)$ is a linear transformation from \mathbb{R}^3 into the space of 3×3 matrices. It is a full-rank transformation: the image is the three-dimensional subspace of skew-symmetric matrices. By construction, $Q(x)x = \chi(x)x = 0$ and $\chi^2(x) = -Q(x)$.

For fixed $x \neq 0$, the normalized matrices $\hat{\chi} = \chi(x)/\|x\|$ and $\hat{Q} = Q(x)/\|x\|^2$ can be multiplied as follows:

$$\hat{Q}^2 = \hat{Q}; \quad \hat{Q}\hat{\chi} = \hat{\chi}\hat{Q} = \hat{\chi}; \quad \hat{\chi}^2 = -\hat{Q}.$$

This means that the four matrices $\{\pm \hat{Q}(x), \pm \hat{\chi}(x)\}$ form a finite group which is isomorphic with that of the four complex numbers $\{\pm 1, \pm i\}$. For each $x \neq 0$, the 2D subspace of matrices $\alpha \hat{Q} + \beta \hat{\chi}$ for $(\alpha, \beta) \in \mathbb{R}^2$ is a field that is isomorphic with the complex numbers.

The proper orthogonal group consists of 3×3 orthogonal matrices whose determinant is $+1$. Each group element $R \in \mathcal{O}_3^+$ acts as a linear transformation

$\mathbb{R}^3 \rightarrow \mathbb{R}^3$, which sends x to Rx by matrix multiplication; this mapping is a rotation. Since rotation preserves the Euclidean norm, i.e., $\|Rx\| = \|x\|$, the group action is not transitive: each sphere is a group orbit. For subspaces $V \subset \mathbb{R}^3$, however, the condition $RV \subset V$ for all $R \in \mathcal{O}_3^+$ implies either $V = 0$ or $V = \mathbb{R}^3$: there are no sub-representations. We say that the representation of the group by linear transformations $\mathbb{R}^3 \rightarrow \mathbb{R}^3$ is \mathcal{O}_3^+ -irreducible.

The group also acts on 3×3 matrices $M \in \mathbb{R}^{3 \times 3}$. Each group element determines a linear transformation $\mathbb{R}^{3 \times 3} \rightarrow \mathbb{R}^{3 \times 3}$, which sends M to RMR' by conjugation. For subspaces $V \subset \mathbb{R}^{3 \times 3}$, the condition $RV \subset V$ for all $R \in \mathcal{O}_3^+$ determines three non-overlapping sub-representations of dimensions one, three and five. The one-dimensional subspace consists of matrices $M \propto I_3$; the three-dimensional subspace consists of skew-symmetric matrices $M = -M'$; and the five-dimensional subspace consists of symmetric matrices that are also trace-free $\text{tr}(M) = 0$. These subspaces are complementary in $\mathbb{R}^{3 \times 3}$, and each sub-representation is irreducible.

The preceding remarks apply also to the full orthogonal group acting on \mathbb{R}^d and on $\mathbb{R}^{d \times d}$ for $d \geq 1$, except that the irreducible matrix representations have dimensions one, $d(d-1)/2$ and $d(d+1)/2 - 1$ respectively. Remarkably, for $d = 3$, and only for $d = 3$, the 3D vector-space representation is \mathcal{O}_3^+ -isomorphic with the skew-symmetric matrix representation, and the transformation $x \mapsto \chi(x)$ in (16.25) is an isomorphism of representations. This means not only that χ is a 1–1 linear transformation from vectors into skew-symmetric matrices, but also that the images under χ of the points x and Rx are related by

$$\chi(Rx) = R \chi(x) R' \quad (16.26)$$

for all 3D-rotation matrices R . In other words, the group relationship $x \mapsto Rx$ in \mathbb{R}^3 is carried over by χ into the group action on matrices.

If R is a reflection, we have instead

$$\chi(Rx) = -R \chi(x) R' = R \chi'(x) R',$$

so reflections in \mathbb{R}^3 are not preserved by the group action in the space of matrices. The full orthogonal group admits two three-dimensional representations, one by its action on vectors and one on matrices; both are \mathcal{O}_3 -irreducible, but they are non-isomorphic as \mathcal{O}_3 -representations.

The group property (16.26) is satisfied not only by every scalar multiple of $\chi(x)$, but also by the matrix products $\chi^2(x), \chi^3(x)$, and so on for each positive integer. Thus, every polynomial function of $\chi(x)$ and every analytic matrix function such as $\exp(\chi(x))$ also satisfies (16.26). For even powers, $\chi^{2n}(x) = (-1)^n \|x\|^{2n} \hat{Q}(x)$ is symmetric, and $\chi^{2n+1}(x) = (-1)^n \|x\|^{2n+1} \hat{\chi}(x)$ is skew-symmetric. Thus, if $x \mapsto P(x)$ is a polynomial or analytic function such that $P(0) = 0$, the matrix polynomial $P(\chi(x))$ is a linear combination of the basis elements $\hat{Q}(x)$ and $\hat{\chi}(x)$ with coefficients depending on $\|x\|$. Every even function is symmetric or real,

and every odd function is skew-symmetric or complex. In particular, the matrix exponential is

$$\exp(\chi(x)) = I_3 - \hat{Q}(x) + \hat{Q}(x) \cos\|x\| + \hat{\chi}(x) \sin\|x\|,$$

where $\hat{Q}(0) = \hat{\chi}(0) = 0$. It follows that the matrix sine function satisfies

$$\sin(\chi(Rx)) = \hat{\chi}(Rx) \sin\|x\| = R \sin(\chi(x)) R'$$

for $R \in \mathcal{O}_3^+$.

No attempt is made here to give a proof of (16.26). However, it is easy to check numerically by testing random values of R and x . It can also be checked algebraically by finding a suitable parameterization for general 3D-rotations and verifying that the two sides of (16.26) are identical for all x .

16.7.3 Action on Matrices

Suppose now that $x = (x_1, x_2, x_3)$ is a list of three $n \times n$ matrices. Then $\chi(x)$ in (16.25) is a matrix of order $3n \times 3n$, and χ is a 1–1 linear transformation $\mathbb{R}^{3n^3} \rightarrow \mathbb{R}^{9n^2}$. If x_1, x_2, x_3 are all symmetric, then $\chi(x) = -\chi'(x)$ is skew-symmetric. However, if x_1, x_2, x_3 are skew-symmetric, then $\chi(x) = \chi'(x)$ is symmetric. Since we are interested in real covariance matrices, our focus is mostly on the latter case.

Whether x is a list of scalars or a list of matrices, the rotation $R \in \mathcal{O}_3^+$ acts on x in the obvious way: $(Rx)_i = \sum_j R_{ij}x_j$, which is another list of the same type. For the same reasons outlined in the preceding section, the linear transformation satisfies $\chi(Rx) = R \chi(x) R'$.

The fact that the components of x and $\chi(x)$ are matrices rather than scalars is immaterial for linear operations such as $x \mapsto Rx$, $\chi(x) \mapsto R \chi(x)$ or $x \mapsto \chi(x)$. However, multiplicative properties such as $\chi^2(x) = -Q(x)$, which rely on commutativity, do not carry over to matrices.

16.7.4 Borrowed Products

For present purposes, it is convenient to regard each $u \in \mathbb{R}^3$ as a purely imaginary quaternion, i.e., $u = u_0 + u_1\mathbf{i} + u_2\mathbf{j} + u_3\mathbf{k}$ whose real part u_0 is zero. The only purpose of this odd choice is to borrow the quaternion scalar product

$$\langle u, v \rangle = u\bar{v} = u'v - (u_2v_3 - u_3v_2)\mathbf{i} - (u_3v_1 - u_1v_3)\mathbf{j} - (u_1v_2 - u_2v_1)\mathbf{k}; \quad (16.27)$$

see Exercises 16.24–16.25. Thus, the real part of $u\bar{v}$ is the standard scalar product $u'v = v'u$ in \mathbb{R}^3 , and the imaginary part is the negative vector product, or cross product

$$u \times v = (u_2v_3 - u_3v_2, u_3v_1 - u_1v_3, u_1v_2 - u_2v_1) = -v \times u; \quad (16.28)$$

(Jeffreys & Jeffreys, 1956, Sect. 2.071).

Equation (16.28) defines the vector product $u \times v$ in \mathbb{R}^3 as the negative imaginary part of the quaternion scalar product $u\bar{v}$. However, the quaternion algebra presumes a right-handed frame of reference in the form of the triple products $\epsilon_{123} = \mathbf{i}\mathbf{j}\mathbf{k} = -1$ and $\epsilon_{321} = \mathbf{k}\mathbf{j}\mathbf{i} = 1$ of the three basis vectors $\mathbf{i}, \mathbf{j}, \mathbf{k}$, taken in cyclic order. In general, therefore,

$$u \times v = \epsilon_{123} \Im(u\bar{v}),$$

where the sign is negative in a right-handed frame of reference and positive otherwise.

Since we work exclusively with real numbers and real matrices, it is convenient to represent the quaternion inner product as the 3×3 matrix

$$u\bar{v} = u'vI_3 + uv' - vu' = u'vI_3 - \chi(u \times v).$$

For the matrix case in which X is an n -point configuration or an $n \times 3$ matrix $X = (X_1, X_2, X_3)$, the set of 3D-vector products is a list of three $n \times n$ skew-symmetric matrices

$$X \times X = (X_2X'_3 - X_3X'_2, X_3X'_1 - X_1X'_3, X_1X'_2 - X_2X'_1).$$

The set of quaternion inner products

$$Q(X) = XX' \otimes I_3 - \chi(X \times X)$$

is a 3×3 matrix whose components are $n \times n$ matrices. The three diagonal matrices are equal and symmetric; the off-diagonal matrices are skew-symmetric. As a 3×3 matrix, it satisfies $Q(XR') = RQ(X)R'$ for $R \in \mathcal{O}_3^+$. It is also an $n \times n$ matrix whose components are 3×3 matrices, and a $3n \times 3n$ matrix whose components are real numbers. For a generic configuration with $n \geq 2$, $Q(X)$ is symmetric positive definite of rank four.

16.7.5 Hydrodynamic Symmetry

A 3D process $u \mapsto Z(u)$ taking values in \mathbb{R}^3 is hydrodynamically symmetric if the rotated process $RZ(\cdot)$ has the same distribution as the domain-rotated process $Z(R\cdot)$. To understand what this implies for a zero-mean Gaussian process,

let $X = (X_1, X_2, X_3)$ be an n -point configuration as an $n \times 3$ matrix, let $Z[X] = (Z_1[X], Z_2[X], Z_3[X])$ be the values, and, for $1 \leq r, s \leq 3$, let $K_{rs}[X] = \text{cov}(Z_r[X], Z_s[X])$ be the $n \times n$ matrix of covariances. In this section, $K[X]$ is regarded as a 3×3 array whose entries are $n \times n$ matrices satisfying $K_{sr}[X] = K'_{rs}[X]$.

Let $u \mapsto Z(u)$ be a 3D zero-mean Gaussian hydrodynamic process. Since Z takes values in \mathbb{R}^3 , the covariance function associates with each ordered pair of domain points (u, v) a 3×3 matrix with real components

$$K(u, v) = \text{cov}(Z(u), Z(v)) = E(Z(u)Z'(v)) = K'(v, u)$$

such that $K(u, v)$ is the transpose of $K(v, u)$. The process is stationary if $K(u + h, v + h) = K(u, v)$ for all $u, v, h \in \mathbb{R}^3$, which means that K is a function of the vector difference $u - v$.

The goal here is to exhibit a non-trivial hydrodynamically-symmetric process, i.e., a process that is hydrodynamically symmetric, but neither isotropic nor rotationally symmetric. One such covariance function is derived in Exercise 16.17. For another example, consider the following matrix-valued covariance functions:

$$\begin{aligned} K_0(u, v) &= M_v(\|u - v\|)V \\ K_1(u, v) &= M_v(\|u - v\|)(I_3u'v - \chi(u \times v)), \end{aligned}$$

where V is a given 3×3 matrix, and M_v is the Matérn covariance function, or any similar positive-definite function of the norm. Provided that V is symmetric and positive definite, K_0 is also positive definite, so there exists a Gaussian process with covariance K_0 . This process is stationary and isotropic. It is not, however, rotationally symmetric unless $RVR' = V$ for each $R \in \mathcal{O}_3^+$, which implies $V \propto I_3$. In that case, the three component processes are independent and identically distributed Matérn processes, and the vector process is also trivially symmetric in the hydrodynamical sense.

For reasons given in the preceding section, K_1 is also a positive-definite symmetric function, so there exists a vector-valued Gaussian process such that $\text{cov}(Z(u), Z(v)) = K_1(u, v)$. This process is neither isotropic nor rotationally invariant. However, ZR has covariance function $K_1(Ru, Rv) = RK_1(u, v)R'$, which is the covariance of RZ . Thus, Z is hydrodynamically symmetric. Finally, $K_1(0, v) = 0$ implies $Z(0) = 0$, so the process is not stationary. However, it is stationary on increments.

16.8 Summer Cloud Cover in Illinois

Figure 16.5 illustrates the fractional cloud cover on a 15×15 grid of points in central Illinois with 0.2 degrees separation in latitude and longitude, which implies

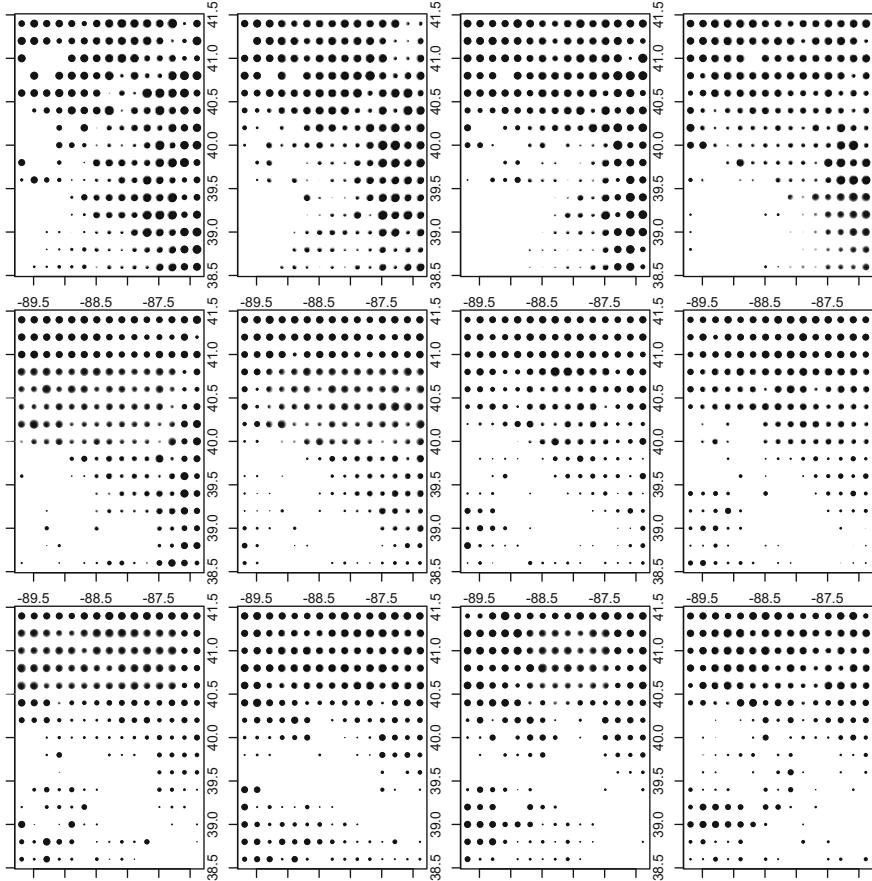


Fig. 16.5 Fractional cloud cover on a 15×15 spatial grid in central Illinois in half-hour intervals from 6.00am to 11.30am on June 9, 1998

17.0×22.2 km cells. Successive panels show the cloud cover at 30-minute intervals from 6.00am to noon on June 9, 1998.

Solar irradiance is measured by a geostationary satellite, and fractional cloud cover is the complement of the ratio of solar irradiance relative to the clear-sky maximum at that time and location. The value lies between zero and one. On this morning, the average fractional cloud cover was 35%. Cloud cover is the primary variable that limits the production of solar energy. Its evolution throughout the day is of commercial interest for short-term prediction of solar electrical generating capacity, so that alternative sources may be brought online if needed.

For this illustration, only the first 2.5 hours of data from 6.00am to 8.30am are used for parameter estimation and model fitting. This corresponds to the first six panels in Fig. 16.5. The process appears to be relatively smooth in space and in time, so we use the Matérn model (16.11) with $\nu = 1$ for both space and time. Two

range parameters, ρ_0 for time in minutes, and ρ_1 for distance in km. are also needed, so t is replaced by t/ρ_0 and x by x/ρ_1 . The isotropic sub-model has $a = 0$. The maximally anisotropic model takes $b = 1$ and advection $a = (\cos \theta, \sin \theta)$ as a unit vector in the east and north directions respectively.

For both the isotropic and the anisotropic covariance models, the mean fractional cloud cover is taken to vary linearly in both space and time. This may be adequate for short-term prediction, but it is not recommended for long-term prediction or extrapolation beyond the spatial domain. As always, a nugget term is included in the covariance model. The fitted range parameters for the isotropic model are 32.5 minutes and 27.0 km, while the variance components are 0.0219 for the identity matrix and 0.0283 for the Matérn product covariance. For the anisotropic model (16.11) using the unit vector $a = (\cos \theta, \sin \theta)$ with $\hat{\theta} = 3.01$, the fitted range parameters are 32.0 minutes and 26.0 km, while the variance components are 0.0215 and 0.0276.

The REML log likelihood for the fitted anisotropic model is 11.23 units higher than that for the isotropic model. Since the anisotropic model has two additional parameters, both $\|a\|$ and θ , the likelihood-ratio test statistic is nominally on two degrees of freedom. In fact, $\|\hat{a}\| = 1$ on the boundary, which slightly complicates the null distribution theory. Nevertheless, the observed likelihood-ratio test statistic of 22.46 leaves no doubt about the existence of space-time anisotropy for the cloud-cover process. Whether the formulation (16.11) captures adequately the full extent of anisotropy is another matter. Most likely, the polar vector could not be expected to remain constant from one day to the next.

Table 16.1 shows the fitted parameters for four spatial models, all including a nugget effect. Each of the anisotropic models is a substantial improvement over the isotropic Matérn product. The simplest travelling wave model (16.23) is the most effective; the additional polar anisotropy in (16.10) does not substantially improve the fit.

In both travelling-wave models, the estimated wave velocity is approximately 0.7 km/min, or 42 km/hr, or 26 mph from the east. However, that particular June

Table 16.1 Summary of fitted parameters for four space-time models

| Parameter | Isotropic | (16.11) | (16.23) | +(16.24) |
|-------------------------|-----------|---------|---------|----------|
| Spatial range (km) | 27.0 | 26.0 | 35.0 | 35.5 |
| Temporal range (min) | 32.5 | 32.0 | 61.0 | 62.0 |
| Nugget variance | 0.0219 | 0.0215 | 0.0240 | 0.0238 |
| Matérn variance | 0.0283 | 0.0276 | 0.028 | 0.0298 |
| Wave speed (km/min) | 0.0 | 0.0 | 0.70 | 0.70 |
| Wave direction θ | — | — | 0.033 | 0.00 |
| Polar norm $\ a\ $ | 0.0 | 1.0 | 0.0 | 0.67 |
| Polar direction ϕ | — | 3.01 | — | 0.63 |
| Log likelihood | 4.724 | 15.954 | 23.522 | 24.920 |
| RMSE 9.00am | 0.140 | 0.131 | | |

morning was calm and humid, with light and variable winds averaging four mph. In such circumstances, a wave travelling at 42 km/hr in any direction might be attributed to changes in temperature or pressure, but it cannot be attributed to atmospheric advection.

The large value of the nugget variance implies that even the best predictor has substantial variance. The nugget standard error is a lower bound for the root mean square prediction error, and the fitted values are 0.148 for the isotropic model, and 0.147 for the anisotropic model (16.11). The empirical one-step-ahead root mean square prediction error averaged over 225 sites for 9.00am are 0.140 for the isotropic model and 0.131 for the anisotropic model (16.11).

16.9 More on Gaussian Processes

16.9.1 White Noise

An alert reader may have noticed that the definition of a Gaussian process in Sect. 16.1, and the definitions of stationarity and isotropy in Sects. 16.2–16.3, are not sufficiently broad to include the simplest non-trivial Gaussian processes on the real line or on the plane. On an arbitrary domain with measure Λ , white noise is a zero-mean Gaussian process indexed by subsets such that, $\text{cov}(W(A), W(B)) = \Lambda(A \cap B)$. The process takes independent values on disjoint sets, and variances are determined by the intensity measure. If Λ is Lebesgue measure on the real line or \mathbb{R}^d , which is the standard choice for those domains, the process is both stationary and isotropic. The notation here and subsequently in this section presumes that the process is real-valued.

The earlier definition is inadequate because it assumes that the domain and the index set are one and the same set. This is sufficient for processes defined pointwise, but it is not sufficient to cover many of the generalized or intrinsic processes that occur in applied work. For planar white noise, the domain is $\mathcal{D} = \mathbb{R}^2$ or \mathbb{C} , but the index set \mathcal{U} is the set of Borel subsets in the domain. More correctly, \mathcal{U} is the proper subset consisting of Borel sets of finite Λ -measure. The definition of stationarity offered in Sect. 16.2.1 is not applicable to white noise because it presumes that $W(x)$ exists for $x \in \mathcal{D}$. In the case of standard white noise $W(\{x\})$ exists and the value is zero for all singletons.

Stationarity and isotropy refer to a group acting on the domain $x \mapsto gx$ either by translation or rotation. There is a natural induced action on the index set $A \mapsto gA$, which is a rigid Euclidean motion of subsets. With an appropriate modification to distinguish between the domain and the index set, white-noise with intensity Λ is stationary or isotropic if the measure is invariant under this action.

Every process Z that is defined pointwise and is continuous on the domain can be extended by integration to an additive process W on domain subsets

$$W(A) = \int_A Z(x) d\Lambda(x).$$

Additivity for disjoint subsets means $W(A \cup B) = W(A) + W(B)$. The covariance function $K(x, x')$ of Z is the covariance density of W

$$\text{cov}(W(A), W(B)) = \int_{A \times B} K(x, x') d\Lambda(x) d\Lambda(x').$$

White noise is not a continuous process and does not have a covariance density. However, $\text{cov}(W(A), W(B)) = \Lambda(A \cap B)$ means that there is a covariance measure, which is the Dirac-type singular measure $\Lambda(\cdot)$ concentrated on the diagonal in \mathcal{D}^2 .

The extension to subsets is a half-way house that suffices for a few purposes, but it is not adequate for mathematical work, which requires all variances to be finite, and it is not entirely adequate even for applied work. Consider, for example, the additive planar process defined for regular planar subsets as follows:

$$\text{cov}(W(A), W(B)) = \begin{cases} \Lambda_1(\partial A \cap \partial B), & \text{Int}(A) \subset \text{Int}(B) \text{ or } \text{Int}(B) \subset \text{Int}(A); \\ -\Lambda_1(\partial A \cap \partial B), & \text{Int}(A) \cap \text{Int}(B) = \emptyset. \end{cases}$$

Regular means that each planar subset A has a well-defined interior $\text{Int}(A)$, and a one-dimensional boundary ∂A of finite length $\Lambda_1(\partial A)$. Additivity implies that the variance for more general regions is the boundary length, and the covariance for two regions is the total signed length of the common boundary.

It is not clear from the preceding description how we are meant to deal with a subset having an irregular boundary or an empty interior, so this specification is not entirely satisfactory. It is also unclear whether a covariance measure exists. The more satisfactory way to study such processes is to abandon subsets and to use a suitable Hilbert space as the index set. Planar white noise is associated with the space of square-integrable functions $f : \mathbb{R}^2 \rightarrow \mathbb{R}$; a subset is nothing more than its indicator function. The process described above is associated with functions such that the norm of the derivative vector is square-integrable $\int \|f'(x)\|^2 d\Lambda(x) < \infty$.

16.9.2 Limit Processes

This section deals with questions of two types, the first related to limits of Gaussian processes, the second related to prediction and limits of conditional distributions.

Two families of processes are used to illustrate the development. Both are indexed by a single parameter $\theta > 0$, and the limit refers to $\theta \rightarrow \infty$. The first is an exchangeable Gaussian process in which the covariance function is

$$\text{cov}(Z_i, Z_j) = \delta_{ij} + \theta, \quad (16.29)$$

which means that the covariance matrix of $Z[n]$ is $I_n + \theta J_n$. The second is a Matérn-1/2 process defined pointwise on \mathbb{R}^d with covariance function

$$\text{cov}(Z(x), Z(x')) = \theta e^{-\|x-x'\|/\theta}. \quad (16.30)$$

Here, θ is the range parameter, and the limit $\theta \rightarrow \infty$ addresses long-range dependence. Inessential scalar multiples, which would almost always occur in applied work, are disregarded.

Each question gives rise to a number of subsidiary questions along the following lines:

1. Does the limit process exist? (N). If not, does any limit process exist? (Y). If a non-trivial limit exists, exhibit the index set, the covariance function, and so on.
2. For fixed n , is there an invertible normalization that produces a limit distribution? (Y).
3. What is the conditional distribution of $Z(x_0)$ given $Z(x_1), \dots, Z(x_n)$? Does the conditional distribution have a limit as $\theta \rightarrow \infty$? (Y).
4. Can the limiting conditional distribution be obtained from the limit process? (N). Can the limiting conditional distribution be obtained from the limit in part 2? (Y).
5. Can the best linear predictor (BLP) of Z_{n+1} be obtained from the limit process? (Y). Can the conditional expected value of Z_{n+1} given $Z[n]$ be computed from the limit process? (N).

The two examples are generic, so the answers indicated in parentheses apply equally to both.

Existence of a Limit Process

In the first example $Z_1 \sim N(0, 1 + \theta)$, and in the second $Z(x) \sim N(0, \theta)$. Neither sequence of distributions has a limit as $\theta \rightarrow \infty$, so the answer to the first question is negative. On the other hand, $Z_i - Z_j \sim N(0, 2)$ for $i \neq j$, while $Z(x) - Z(x')$ has variance

$$2\theta - 2\theta e^{-\|x-x'\|/\theta} = 2\|x-x'\| + O(\theta^{-1}).$$

Both limits exist and the distributions are Gaussian. More generally, for $n \geq 1$ and any coefficient vector $\alpha = (\alpha_1, \dots, \alpha_n)$ whose components add to zero, the

linear combination $\sum \alpha_j Z_j$ is Gaussian with variance $\|\alpha\|^2 = \sum \alpha_j^2$, independent of θ . In the case of the Matérn model, $\sum \alpha_j Z(x_j)$ is Gaussian with variance $\alpha' D \alpha + O(\theta^{-1})$, where $D_{ij} = -\|x_i - x_j\|$ is the $n \times n$ matrix of negative Euclidean distances. In both cases the limit exists for arbitrary contrasts.

The easiest way to characterize the preceding limit is to use the vector space of contrasts as the index set. For ease of exposition, suppose that the set of points under consideration is fixed and finite, so that a contrast $\alpha \in \mathbf{1}_n^0$ is a vector whose components add to zero. The value at contrast α is $W(\alpha) = \alpha_1 Z_1 + \dots + \alpha_n Z_n$ for (16.29), or $\sum \alpha_r Z(x_r)$ in the case of the Matérn model. The limiting covariance of two contrasts is the Hilbert-space inner product

$$\text{cov}(W(\alpha), W(\beta)) = \langle \alpha, \beta \rangle = \begin{cases} \sum_r \alpha_r \beta_r \\ - \sum_{r,s} \alpha_r \beta_s \|x_r - x_s\| \end{cases}$$

The Hilbert space \mathcal{H}_n^* has dimension $n - 1$, but it is exhibited here as the subspace $\mathbf{1}_n^0$ of contrasts in a vector space of dimension n . Consequently, the inner-product matrix is of order n , and is not unique. For the first example, we could use either the identity matrix of order n or $I_n - J_n/n$.

Since every n -contrast is also a $(n+1)$ -contrast whose last component is zero, the Hilbert-space \mathcal{H}_n^* of n -contrasts is a subspace of H_{n+1}^* . Kolmogorov consistency is automatic, but is equivalent to the statement that the restriction or insertion $\mathcal{H}_n^* \hookrightarrow H_{n+1}^*$ is an isometry. In effect, \mathcal{H}_∞^* includes all of the finite-dimensional spaces as subspaces. In fact, the restriction to contrasts determines a consistent process, not only in the limit, but for every θ .

An equivalent way of saying the same thing is that the process for finite θ is defined conventionally for finite samples as a probability distribution $P_{n,\theta}$ on the space $\mathcal{B}(\mathbb{R}^n)$ of Borel subsets in \mathbb{R}^n . For any event $A \subset \mathbb{R}^n$ such that $A + \mathbf{1}_n = A$, the value assigned by $P_{n,\theta}$ to A has a limit $P_{n,\infty}(A)$, which a Gaussian probability. These translation-invariant events $A \in \mathcal{B}(\mathbb{R}^n/\mathbf{1}_n)$ are the only events to which the limit process assigns a probability. Thus, the non-trivial limit is obtained by restriction of the σ -field.

Existence of a Limit Distribution

In all examples of the type under consideration, the covariance matrix of $Z[n]$ is

$$\text{cov}(Z[n]) = \theta J_n + \Sigma + O(\theta^{-1}) \quad (16.31)$$

where Σ is independent of θ and is also positive definite on contrasts. For the Matérn example, Σ is the matrix with components $-\|x_i - x_j\|$.

Let $P = J_n/n$ and $Q = I_n - P$ be complementary projections, so that

$$\text{cov} \begin{pmatrix} \theta^{-1/2} PZ \\ QZ \end{pmatrix} = \begin{pmatrix} J_n + P\Sigma P/\theta & P\Sigma Q/\theta^{1/2} \\ Q\Sigma P/\theta^{1/2} & Q\Sigma Q \end{pmatrix} \rightarrow \begin{pmatrix} J_n & 0 \\ 0 & Q\Sigma Q \end{pmatrix}.$$

Consequently, the vector $W = \theta^{-1/2} PZ + QZ$ with components

$$W_i = \theta^{-1/2} \bar{Z}_n + (Z_i - \bar{Z}_n)$$

has a Gaussian limit distribution with covariance matrix $J_n + Q\Sigma Q'$. For each $n \geq 1$, the transformation $Z[n] \mapsto W$ is invertible, so the answer to part 2 is affirmative.

Note that $W \in \mathbb{R}^n$ is not the restriction of the corresponding transformation in \mathbb{R}^{n+1} , so there is no W -process associated with these transformations. The existence of a limit distribution for every n does not imply the existence of a limit process.

Limit of Conditional Distributions

The conditional distribution of $Z(x_0)$ given $Z(x_1), \dots, Z(x_n)$ is Gaussian, so it is necessary only to compute the conditional mean and variance, and to observe the behaviour as $\theta \rightarrow \infty$. For the exchangeable model, the conditional mean and variance are

$$E(Z_{n+1} | Z[n]) = \frac{n\theta \bar{Z}_n}{1+n\theta}, \quad \text{var}(Z_{n+1} | Z[n]) = \frac{1+(n+1)\theta}{1+n\theta},$$

so the limit of the conditional distributions is $N(\bar{Z}_n, 1)$.

For the spatial model, the calculations for finite θ are a little more complicated, so it is necessary to take limits as the calculation progresses. Let $L = \theta^{-1/2}L_0 + L_1$ be the matrix of a linear transformation in \mathbb{R}^{n+1}

$$L_0 = \begin{pmatrix} P_n & 0 \\ 0 & 0 \end{pmatrix}, \quad L_1 = \begin{pmatrix} Q_n & 0 \\ -\mathbf{1}_n/n & 1 \end{pmatrix},$$

where P_n and $Q_n = I_n - P_n$ are the complementary projections denoted by P, Q in the preceding section. From the representation (16.31), the covariance matrix of $W = LZ$ has a limit, which is the sum of two mutually orthogonal matrices

$$\begin{aligned} \text{cov}(LZ) &= \theta L J_{n+1} L' + L \Sigma L' + O(\theta^{-1}) \\ &= \begin{pmatrix} J_n & 0 \\ 0 & 0 \end{pmatrix} + L_1 \Sigma L_1' + O(\theta^{-1}). \end{aligned}$$

For each n , it follows that $\theta^{-1/2} \bar{Z}_n$ has a standard normal limit as $\theta \rightarrow \infty$, and is asymptotically independent of every contrast $Z_i - \bar{Z}_n$, not only for $i \leq n$, but also for $i = n + 1$. Thus, the limiting conditional mean satisfies

$$E(Z_{n+1} - \bar{Z}_n | Z[n]) = \sum_{r=1}^n \beta_r Z_r, \quad (16.32)$$

$$E(Z_{n+1} | Z[n]) = \bar{Z}_n + \sum_{r=1}^n \beta_r Z_r, \quad (16.33)$$

where β is the orthogonal projection $H_{n+1}^* \rightarrow \mathcal{H}_n^*$ of the coefficient vector $(-\mathbf{1}_n/n, 1)$ associated with the contrast $Z_{n+1} - \bar{Z}_n$. The linear combination (16.33), which is not a contrast, is often called the best linear predictor (BLP). The limiting conditional variance is the reciprocal of the last diagonal component of $(L_1 \Sigma L_1')^{-1}$.

Conditional Distributions for the Limit Process

The situation regarding conditional distributions for the limit process is different in a fundamental but subtle way. The joint distribution for any set of contrasts is determined by the Hilbert-space inner product. In particular, the conditional distribution of the contrast $Z_{n+1} - \bar{Z}_n$ given the σ -field generated by Z_1, \dots, Z_n is Gaussian with mean (16.32) and variance as described above. However, the limit process is defined on contrasts only, so the σ -field generated by Z_1, \dots, Z_n is the σ -field generated by contrasts, which means that the coefficient vector β in (16.32) is a contrast in \mathcal{H}_n^* . The limit process does not admit either Z_{n+1} or \bar{Z}_n as a Gaussian variable, so the crucial statement that $Z_{n+1} - \bar{Z}_n$ is independent of \bar{Z}_n is either meaningless or mathematically trivial. In either case, the fiducial leap from (16.32) to (16.33) requires a σ -field extension, which cannot follow from the limit process alone.

Limit Process as a Markov Kernel

The limit process with probabilities defined on the σ -field generated by contrasts has a certain mathematical elegance—brutal and minimalist. But the σ -field restriction is a price too steep for any applied statistician interested in probabilistic prediction. Is there a way out, a way that retains the elegance of contrasts at a more affordable price? The answer, we hope, is yes.

A Markov kernel is a function that associates with each $\mu \in \mathbb{R}$ a Gaussian process Z such that, for each contrast $\alpha \in \mathbf{1}_n^0$, the increment or linear functional $\sum \alpha_r Z_r$ has the same distribution as that in the limit process. There is no σ -field restriction.

16.10 Exercises

16.1 Show that the 3×3 Hermitian matrix

$$\begin{pmatrix} 1 & \rho & \bar{\rho} \\ \bar{\rho} & 1 & \rho \\ \rho & \bar{\rho} & 1 \end{pmatrix}$$

has determinant $1 - 3|\rho|^2 + 2\Re(\rho^3)$. Hence or otherwise, deduce that the matrix is positive definite if and only if ρ lies in the triangle with vertices at the cube roots of unity $\{1, e^{2\pi i/3}, e^{-2\pi i/3}\}$.

16.2 By making the transformation $u = 1/(1+x^2)$ and converting to a beta-type integral on $(0, 1)$, show that

$$2 \int_0^\infty \frac{x^{d-1} dx}{(1+x^2)^{v+d/2}} = B(v, d/2) = \frac{\Gamma(v) \Gamma(d/2)}{\Gamma(v+d/2)},$$

where $B(\cdot, \cdot)$ is the beta function for strictly positive arguments.

16.3 The Matérn spectral measure on the real line is proportional to the symmetric type IV distribution in the Pearson class, which is also equivalent to the Student t family (Pearson type VII). For $v > -1/2$, show that the standardized version

$$M_1(d\omega) = \frac{\Gamma(v+1/2) d\omega}{\pi^{1/2} (1+\omega^2)^{v+1/2}}$$

has positive density, but the total mass is finite only for $v > 0$.

16.4 By transforming to spherical polar coordinates in \mathbb{R}^d , show that

$$\int_{\mathbb{R}^d} \frac{d\omega}{(1+\|\omega\|^2)^{v+d/2}} = A_{d-1} \int_0^\infty \frac{x^{d-1} dx}{(1+x^2)^{v+d/2}},$$

where $A_{d-1} = 2\pi^{d/2}/\Gamma(d/2)$ is the surface area of the unit sphere in \mathbb{R}^d . For $v > -d/2$, deduce that the Matérn measure on \mathbb{R}^d

$$M_d(d\omega) = \frac{\Gamma(v+d/2) d\omega}{\pi^{d/2} (1+\|\omega\|^2)^{v+d/2}}.$$

has finite mass if and only if $v > 0$. Show that the total mass is a constant independent of the dimension of the space.

16.5 For fixed $\nu > 0$, show that the Matérn measures are mutually consistent in the sense that $M_{d+1}(A \times \mathbb{R}) = M_d(A)$ for all $d \geq 0$ and subsets $A \subset \mathbb{R}^d$. In other words, show that M_d is the marginal distribution of M_{d+1} after integrating out the last component. For $\nu > -1$, show that the Matérn measures are mutually consistent in the sense that $M_{d+1}(A \times \mathbb{R}) = M_d(A)$ for all $d \geq 2$.

16.6 Consistency and finiteness together imply that the normalized Matérn measures define a real-valued process X_1, X_2, \dots in which $M_n / \Gamma(\nu)$ is the joint distribution of the finite sequence $X[n] = (X_1, \dots, X_n)$. This process—a special case of the Gosset process—is not only exchangeable but also orthogonally invariant for every n . Show that the conditional distribution of X_{n+1} given $X[n]$ is Student t , with a certain location parameter, scale parameter and degrees of freedom. To what extent is finiteness needed in the construction of the process?

16.7 For the Matérn process, show that the sequence of partial averages \bar{X}_n has a limit $\bar{X}_\infty = \lim_{n \rightarrow \infty} \bar{X}_n$. For $n \geq 2$, what can you say about the conditional distribution of \bar{X}_∞ given $X[n]$? Consider separately the cases $\nu = 0$ and $\nu > 0$.

16.8 One definition of the Bessel-K function is the integral

$$\int_0^\infty \frac{\cos(\omega t) d\omega}{(1 + \omega^2)^{\nu+1/2}} = \frac{\sqrt{\pi}}{2^\nu \Gamma(\nu + 1/2)} \times |t|^\nu \mathcal{K}_\nu(t).$$

Deduce that $\mathcal{K}_\nu(\cdot)$ is symmetric and that

$$\lim_{t \rightarrow 0} |t|^\nu \mathcal{K}_\nu(|t|) = 2^{\nu-1} \Gamma(\nu).$$

16.9 For any linear functional $x: \mathbb{R}^d \rightarrow \mathbb{R}$, show that

$$\begin{aligned} \int_{\mathbb{R}^d} \frac{\cos(\omega x) d\omega}{(1 + \|\omega\|^2)^{\nu+d/2}} &= I(\nu + 1/2, d - 1) \int_{\mathbb{R}} \frac{\cos(w\|x\|) dw}{(1 + w^2)^{\nu+1/2}} \\ &= \frac{\pi^{d/2}}{2^{\nu-1} \Gamma(\nu + d/2)} \times \|x\|^\nu \mathcal{K}_\nu(\|x\|), \end{aligned}$$

where $I(\nu, d) = \pi^{d/2} \Gamma(\nu) / \Gamma(\nu + d/2)$.

16.10 Use integration by parts to show that

$$\begin{aligned} \int_{-\infty}^\infty \frac{\omega \sin(t\omega) d\omega}{(1 + \omega^2)^{\nu+3/2}} &= \frac{t}{2\nu + 1} \int_{-\infty}^\infty \frac{\cos(t\omega) d\omega}{(1 + \omega^2)^{\nu+1/2}}, \\ &= \frac{\sqrt{\pi}}{2^{\nu-1} \Gamma(\nu + 1/2)} \times \frac{t}{2\nu + 2} |t|^\nu \mathcal{K}_\nu(t). \end{aligned}$$

Hence deduce that, for any pair of linear functionals $v, x : \mathbb{R}^d \rightarrow \mathbb{R}$,

$$\int_{\mathbb{R}^d} \frac{\omega v \sin(\omega x) d\omega}{(1 + \|\omega\|^2)^{v+d/2+1}} = \frac{\pi^{d/2}}{2^{v-1} \Gamma(v + d/2)} \times \frac{vx}{2v+1} \|x\|^v \mathcal{K}_v(\|x\|),$$

where vx denotes the scalar product.

16.11 Let Z be a real Gaussian space-time process with zero mean and full-rank separable covariance function:

$$\text{cov}(Z(x, t), Z(x', t')) = K(x, x') V(t, t').$$

Let $\mathbf{x} = \{x_1, \dots, x_n\}$ be a spatial configuration, $\mathbf{t} = \{t_1, \dots, t_k\}$ a temporal configuration, and let $Z[\mathbf{x} \times \mathbf{t}]$ be the values on the Cartesian product set. Show that the conditional expected value of $Z(x_0, t_0)$ given $Z[\mathbf{x} \times \mathbf{t}]$ satisfies

$$E(Z(x_0, t_0) | Z[\mathbf{x} \times \mathbf{t}]) = \sum_{ij} \sum_{rs} rs K(x_0, x_i) V(t_0, t_r) K[\mathbf{x}]_{ij}^{-1} V[\mathbf{t}]_{rs}^{-1} Z(x_j, t_s).$$

Show also that if the prediction site belongs to \mathbf{x} , say $x_0 = x_1$, the conditional expectation reduces to the linear combination

$$E(Z(x_0, t_0) | Z[\mathbf{x} \times \mathbf{t}]) = \sum_{rs} V(t_0, t_r) V[\mathbf{t}]_{rs}^{-1} Z(x_0, t_s)$$

depending only on the values at (x_0, \mathbf{t}) . In other words, if the model is separable, the spatial covariance is irrelevant for temporal prediction.

16.12 This exercise is concerned with stereographic projection from the unit sphere in \mathbb{R}^{d+1} onto the equatorial plane \mathbb{R}^d . Latitude on the sphere is measured by the polar angle θ , starting from zero at the north pole, through $\theta = \pi/2$ at the equator up to $\theta = \pi$ at the south pole. Every point on the sphere is a pair $z = (e \sin \theta, \cos \theta)$ where e is a unit equatorial vector. The stereographic image of z is the point

$$\omega = e \cot(\theta/2) = e \cos(\theta/2) / \sin(\theta/2),$$

so that the southern hemisphere is projected into the unit ball, and the northern hemisphere to its complement in \mathbb{R}^d . Deduce that the stereographic image of the uniform spherical distribution is

$$\frac{\Gamma(d)}{\pi^{d/2} \Gamma(d/2)} \frac{d\omega}{(1 + \|\omega\|^2)^d}$$

on the equatorial plane.

16.13 Points near the north pole are transformed stereographically to high frequencies, and points near the south pole to low frequencies. For $v > d/2$, the weighted distribution with density proportional to

$$|\sin(\theta/2)|^{2v-d}$$

reduces the mass on northern latitudes and increases that on southern latitudes, maintaining radial symmetry. Show that the stereographic image of the weighted distribution is inversely proportional to $(1 + \|\omega\|^2)^{v+d/2}$. Find the normalizing constants for both distributions.

16.14 For the special case $d = 2$, we may regard $\mathbb{R}^2 \cong \mathbb{C}$, so that ω is a complex number and e is a unit complex number. Show that the weighted spherical distribution with weight proportional to the degree k harmonic perturbation

$$1 + \Re(a\bar{e}^k) \sin^k \theta$$

is transformed to

$$\frac{\Gamma(d)}{\pi^{d/2}\Gamma(d/2)} \frac{d\omega}{(1 + \|\omega\|^2)^d} \times \left(1 + \frac{2\Re(a\bar{\omega}^k)}{(1 + |\omega|^2)^k}\right)$$

and is positive for $|a| \leq 1$.

16.15 Consider a fixed tessellation of the plane into a countable set of polygonal cells A_1, \dots , and let $0 \leq \ell_{ij} < \infty$ be the length of the common boundary $\partial A_i \cap \partial A_j$. Associate with each ordered pair of regions (i, j) a Gaussian random variable

$$\varepsilon_{ij} = -\varepsilon_{ji} \sim N(0, \ell_{ij})$$

with independent and identically distributed signs independent of $|\varepsilon|$. If all boundary lengths ℓ_i are finite, the row sums $W(A_i) = \varepsilon_i$ define a Gaussian process indexed by cells. Find the covariances $\text{cov}(W(A_i), \bar{W}(A_j))$ for $i = i$ and $i \neq j$.

16.16 In the setting of the previous exercise, let $W = L\varepsilon$, where L is a Boolean matrix. Show that W is a process defined on general planar regions and that it coincides with the process described at the end of Sect. 16.9.1.

16.17 Let Y be a stationary real-valued Gaussian process on \mathbb{R}^d with isotropic covariance function $\exp(-\|x - x'\|^2/2)$. Show that the gradient field ∂Y is an \mathbb{R}^d -valued Gaussian process with covariance function

$$K_{rs}(x, x') = \text{cov}(\partial_r Y(x), \partial_s Y(x')) = \exp(-\|x - x'\|^2/2)(\delta_{rs} - (x_r - x'_r)(x_s - x'_s)).$$

Show also that $K(Rx, Rx') = RK(x, x')R'$ for $R \in \mathcal{O}_d$, and hence that the gradient process is hydrodynamically symmetric.

16.18 Special Gaussian family on \mathbb{C}^3 : Let $\rho = (\rho_1, \rho_2, \rho_3)$ be a real vector, and let $Z = (Z_1, Z_2, Z_3)$ be a zero-mean complex Gaussian variable with covariance matrix of the form

$$\text{cov}(Z, Z^*) = \begin{pmatrix} 1 & -i\rho_3 & i\rho_2 \\ i\rho_3 & 1 & -i\rho_1 \\ -i\rho_2 & i\rho_1 & 1 \end{pmatrix} = I_3 + i\chi(\rho).$$

Show that the covariance matrix is positive definite if and only if $\|\rho\| \leq 1$. For any real 3×3 orthogonal matrix L with $\det(L) = 1$, show that LZ belongs to the same family with parameter $L\rho$, i.e., that $\chi(L\rho) = L\chi(\rho)L'$.

16.19 Under what conditions on ρ is the special Gaussian process stationary on $\mathbb{Z} \pmod{3}$?

16.20 For each $v > 0$ and $\omega \in \mathbb{R}$, the Matérn function $M_v(\|t - t'\|) e^{i\omega(t-t')}$ defines a stationary complex Gaussian process on the real line with frequency $|\omega|$. Show that the conjugate process $t \mapsto \bar{Z}(t)$ has the same distribution as the reverse-time process $t \mapsto Z(-t)$.

16.21 For each $\rho \in \mathbb{R}^3$ such that $\|\rho\| \leq 1$, deduce that the following symmetric functions are positive definite on $\mathbb{R} \times [3]$:

$$\begin{aligned} & M_v(\|t - t'\|)\delta_{rs}; \\ & M_v(\|t - t'\|) \cos(\omega(t - t'))\delta_{rs}; \\ & M_v(\|t - t'\|) \cos(\omega(t - t'))\delta_{rs} - M_v(\|t - t'\|) \sin(\omega(t - t'))\chi(\rho)_{rs}. \end{aligned}$$

16.22 For each $v > 0$ and $\omega \in \mathbb{R}^3$, the Matérn function $M_v(\|x - x'\|) e^{i\omega'(x-x')}$ defines a stationary complex Gaussian process on \mathbb{R}^3 with frequency $\|\omega\|$ and wave direction $\omega/\|\omega\|$. For each $\rho \in \mathbb{R}^3$ such that $\|\rho\| \leq 1$, deduce that the following functions are positive definite symmetric on $\mathbb{R}^3 \times [3]$:

$$\begin{aligned} & M_v(\|x - x'\|)\delta_{rs}; \\ & M_v(\|x - x'\|) \cos(\omega'(x - x'))\delta_{rs}; \\ & M_v(\|x - x'\|) \cos(\omega'(x - x'))\delta_{rs} - M_v(\|x - x'\|) \sin(\omega'(x - x'))\chi(\rho)_{rs}; \end{aligned}$$

for $x, x' \in \mathbb{R}^3$ and $r, s \in [3]$.

16.23 For each (ω, ρ) , deduce that the matrix-valued function

$$M_v(\|x - x'\|) \left(\cos(\omega'(x - x')) I_3 - \sin(\omega'(x - x')) \chi(\rho) \right)$$

is the covariance function for a stationary Gaussian process $Z: \mathbb{R}^3 \rightarrow \mathbb{R}^3$.

16.24 If $Z: \mathbb{R}^3 \rightarrow \mathbb{R}^3$ is the Gaussian process with parameter (ω, ρ) as defined in the preceding exercise, and R is a 3D rotation, show that the domain-rotated process $x \mapsto Z(R'x)$ is Gaussian with parameter $(R\omega, \rho)$. Show also that RZ is Gaussian with parameter $(\omega, R\rho)$, while RZR' is Gaussian with parameter $(R\omega, R\rho)$.

16.25 The parameters ω, ρ of the Gaussian process are two points in \mathbb{R}^3 , which determine the frequency and direction of spatial anisotropies in the given frame of reference. In the rotated frame of reference, the values are $R\omega, R\rho$. Hence justify the claim that the matrix-valued covariance in Exercise 16.20 is the covariance function of a hydrodynamic process at a fixed time.

16.26 For each $\alpha \in \mathbb{R}$ and $\omega, \rho \in \mathbb{R}^3$ with $\|\rho\| \leq 1$, deduce that the matrix-valued function

$$M_v(\|x - x'\|) M_v(t - t') \times \left(\cos(\omega'(x - x') + \alpha(t - t')) I_3 - \sin(\omega'(x - x') + \alpha(t - t')) \chi(\rho) \right)$$

is the covariance function for a Gaussian process $Z: \mathbb{R}^3 \times \mathbb{R} \rightarrow \mathbb{R}^3$ that is both temporally and spatially stationary.

16.27 Let $x = (x_0, x_1, x_2, x_3)$ be a unit vector in \mathbb{R}^4 , let $v = (x_1, x_2, x_3)$ and let $\chi(v)$ be the 3×3 matrix in (16.25). Show that image of the mapping $x \mapsto R(x)$

$$R(x) = I_3 + 2x_0 \chi(v) + 2\chi^2(v)$$

is equal to \mathcal{O}_3^+ . Find the unit vector x' such that $R(x') = R'(x)$.

16.28 A quaternion is a formal linear combination $q = q_0 + q_1\mathbf{i} + q_2\mathbf{j} + q_3\mathbf{k}$ of four basis elements $\{1, \mathbf{i}, \mathbf{j}, \mathbf{k}\}$ with real coefficients, so that the set of quaternions is a real vector space of dimension four. Unlike vectors, quaternions can also be multiplied according to Hamilton's celebrated formula

$$\mathbf{i}^2 = \mathbf{j}^2 = \mathbf{k}^2 = \mathbf{ijk} = -1.$$

Scalar multiplication, or real multiplication, is commutative. Show that Hamilton's formula implies

$$\mathbf{ij} = \mathbf{k} = -\mathbf{ji}, \quad \mathbf{jk} = \mathbf{i} = -\mathbf{kj}, \quad \mathbf{ki} = \mathbf{j} = -\mathbf{ik}.$$

16.29 The conjugate quaternion is $\bar{q} = q_0 - q_1\mathbf{i} - q_2\mathbf{j} - q_3\mathbf{k}$, so $q = \bar{q}$ means that q is real, and $q = -\bar{q}$ means that q is purely imaginary. Show that conjugate product \overline{pq} is equal to the product of the conjugates $q\bar{p}$ in reverse order, and

$$|q|^2 = q\bar{q} = \bar{q}q = q_0^2 + q_1^2 + q_2^2 + q_3^2.$$

16.30 Show that $|pq| = |p| \times |q|$, i.e., that the modulus of a product is the product of the moduli.

16.31 A quaternion of modulus one is called a unit quaternion. Show that the set of unit quaternions is a group containing the finite sub-group $\{\pm 1, \pm \mathbf{i}, \pm \mathbf{j}, \pm \mathbf{k}\}$. What is the group inverse of q ?

16.32 Show that the 4×4 matrices $e_0 = I_4$,

$$e_1 = \begin{pmatrix} 0 & -1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 \end{pmatrix}, \quad e_2 = \begin{pmatrix} 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \end{pmatrix}, \quad e_3 = \begin{pmatrix} 0 & 0 & 0 & -1 \\ 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}$$

satisfy Hamilton's multiplication formulae, i.e., $e_1e_2 = e_3$, $e_2e_3 = e_1$ and so on. Hence deduce that the quaternion-to-matrix mapping

$$q \mapsto \chi_4(q) = \begin{pmatrix} q_0 & -q_1 & -q_2 & -q_3 \\ q_1 & q_0 & -q_3 & q_2 \\ q_2 & q_3 & q_0 & -q_1 \\ q_3 & -q_2 & q_1 & q_0 \end{pmatrix}$$

is a linear representation satisfying $\chi_4(pq) = \chi_4(p)\chi_4(q)$ and $\chi_4(\bar{p}) = \chi'_4(p)$.

16.33 Let p, q be purely imaginary quaternions. Find the matrix representation $\chi_4(p\bar{q})$ of the quaternion product $p\bar{q}$, and show how this is related to $\chi(x)$ in (16.25) and $\chi(u \times v)$ in Sect. 16.7.4.

16.34 Let p be an arbitrary quaternion and let q be a unit quaternion. Show that the real part of the product $qp\bar{q}$ is equal to $\Re(p)$. Show also that the imaginary part of the product $qp\bar{q}$ is a 3D rotation of $\Im(p)$.

Chapter 17

Likelihood



17.1 Introduction

17.1.1 Non-Bayesian Model

A non-Bayesian statistical model is a set of processes or a set of probability distributions $\{P_\theta\}$ on the sample space indexed by the points θ in the parameter space Θ . According to the standard paradigm, Nature chooses a point θ^* and generates the process $\{Y(x) : x \in \mathcal{D}\}$ on some domain according to the distribution P_{θ^*} . The observer chooses a fixed design or sample $\mathbf{x} = \{x_1, \dots, x_n\}$ and observes or measures the sample values $Y[\mathbf{x}]$. The data consists of the design points \mathbf{x} , the values $Y[\mathbf{x}] \in \mathbb{R}^n$, and any other recorded baseline information.

Before the model can be used for inference, it is necessary to estimate the parameter from the data. The point estimator is a function $\hat{\theta}_n : \mathbb{R}^n \rightarrow \Theta$ from the observation space into the parameter space, defined for every adequately large design. The numerical value $\hat{\theta}_n(Y)$ determines a process $P_{\hat{\theta}}$, called the fitted process or bootstrap process, which serves as our best guess about what Nature might have been up to. If the goal is parametric inference, it is essential to quantify the estimation error, i.e., to quantify the magnitude of the difference $\hat{\theta}_n - \theta^*$ in some suitable sense. If the goal is not parametric, for example, the prediction of future values, it is necessary to compute the fitted conditional distribution

$$P_{n+m,\hat{\theta}}(A \mid Y[\mathbf{x}])$$

for events $A \subset \mathbb{R}^{n+m}$. In either case, the inferential goal requires not only a point estimate of the parameter, but also some measure of its uncertainty and the effect of uncertainty on inferences.

The estimation step is sometimes called ‘learning’ in computer-science circles. But the parameter value is never learned with the certainty that that word implies; it is only estimated with error, which might be small or large. A large error

in estimation does not necessarily lead to a large error in prediction. Generally speaking, parameters that are difficult to estimate have little effect on single-point predictions that are local in a suitable sense.

In all cases considered in this book, Θ is a smooth manifold—at least locally near most points. There may be boundary points or points of singularity. We say that the model is finite-dimensional if the dimension $\dim(\Theta) = p$ is finite. Otherwise the model is infinite-dimensional. The phrase *non-parametric model* is sometimes used in the literature as a synonym for *infinite-dimensional parametric model*. In these notes, *parametric inference* is meant literally in the sense of inferences about $\theta^* \in \Theta$ whether the dimension is finite or infinite; *nonparametric inference* refers to inferential goals that are beyond the parameter space.

Although infinite-dimensional problems occur as exercises, the focus here is on finite-dimensional models. There is an intermediate class of problems in which the space $\Theta \equiv \Theta_n$ is design-dependent or sample-size dependent, with finite dimension p_n dependent on n . Very often, dimension-dependent spaces are used as an artificial mathematical device to gauge the effect of ‘many parameters’ on the behaviour of the estimation procedure in extreme situations. Every model considered in this book is a family of processes. Although the parameter space may be infinite-dimensional, it is fixed and independent of the design.

The emphasis in this chapter is on normal behaviour of models and estimates, not on anomalies. Typically, we assume that the model is *identifiable*, which means that

$$\theta \neq \theta' \implies P_\theta \neq P_{\theta'}.$$

In other words, different parameter values correspond to distinct processes. Identifiability does not necessarily imply $P_{n,\theta} \neq P_{n,\theta'}$ for small samples, say $n = 1$ or $n = 2$. Nor does it imply that θ is estimable from the data, even for large n . Identifiability is not a strong condition, nor is it an essential condition: see the mixture problem in Exercise 17.1.

17.1.2 Bayesian Resolution

A Bayesian model has all of the ingredients listed in the preceding section—plus one other. The additional feature is a probability distribution $\pi(\cdot)$ on the parameter space. The non-Bayesian model is generally portrayed as a stochastic formulation whose appropriateness in a given application is widely agreed, whereas no broad consensus is expected regarding the choice of $\pi(\cdot)$. One is said to be objective and the other subjective. These adjectives are not only provocative and unhelpful, but also devoid of mathematical content.

The net effect of the prior is that a Bayesian model is either (i) a single distribution $\pi(d\theta)P_{\theta,n}(\cdot)$ on the product space $\Theta \times \mathbb{R}^n$; or (ii) a single mixture process $P_\pi(\cdot) = \int P_\theta(\cdot) \pi(d\theta)$. In principle, the reduction to a mixture is a huge

simplification because the estimation step is by-passed, the model comprises a single process, and the ambiguity about the choice of process for prediction is eliminated. Even for problems of parametric inference, it is usually possible in principle to by-pass the parameter space entirely by re-phrasing the target as a tail event associated with a limit statistic, for example by computing $P_\pi(\bar{Y}_\infty \in A \mid Y[\mathbf{x}])$ or $P_\pi(\hat{\theta}_\infty \in A \mid Y[\mathbf{x}])$.

In practice, two difficulties must be overcome before we can confidently take advantage of the Bayesian solution. The first is to select a suitable prior distribution and, more importantly, to convince the reader that this prior is appropriate for the problem. The Bayes resolution calls for a single prior distribution selected to represent the information available a priori. In practice, it is often better to depart from the paradigm by considering a sequence of distributions π_v for $v > 0$ such that the available information corresponds either to the limit $v \rightarrow 0$ or to the asymptote in which v is small but strictly positive.

Absence of information may be represented by a sequence of distributions such that $\pi_v(A) \rightarrow 0$ at rate $\rho_v > 0$ on bounded subsets in such a way that $\rho_v^{-1}\pi_v(dx)$ has a finite non-zero limit. The limit is a measure, sometimes termed ‘improper’ because it is not a probability distribution. However, for sufficiently large samples, the conditional distribution given the data may have a limit that is satisfactory for inference. Usually it depends on the limit measure, but is otherwise independent of the sequence. At the other end of the spectrum, strong information such as sparsity corresponds to a sequence that tends to the Dirac measure in a suitably regular way that permits limits for a certain class of integrals (McCullagh and Polson 2018).

These limit recipes are reasonably satisfactory for stylized problems in low-dimensional parameter spaces. For high-dimensional spaces, assumptions of independence for selected components are not to be taken lightly because their effect on conclusions may be substantial.

The second problem, perhaps less of an obstacle today than in the recent past, is to manage the computations. Posterior distributions can often be approximated by simulation in various ways, for example, using Markov-chain Monte Carlo. Bayesian computation is not a focus of this book, so we do not make sweeping recommendations regarding the choice of prior or how to manage the computation.

17.2 Likelihood Function

17.2.1 Definition

In the simplest setting where the response distribution has a density $P_\theta(dy) = p_\theta(y) dy$, the density $p_\theta(y)$ as a function of θ for fixed y is the likelihood function. The function $p_\theta(y)$ is the density of the probability relative to Lebesgue measure. For present purposes, there is nothing special about Lebesgue measure, so the likelihood function is the density ratio relative to an arbitrary fixed density. For

example, if zero is a point in the parameter space, we could adopt $L(\theta) = p_\theta(y)/p_0(y)$ as the likelihood function provided that $p_0(y)$ is strictly positive throughout the space. The important point to remember is that only likelihood ratios $p_\theta(y)/p_{\theta'}(y)$ are well defined. Even in that case, it is necessary to handle points of zero density and points of infinite density with care.

On the technical side, we assume that there is a dominating measure that covers all distributions in the model. Usually this is Lebesgue measure or counting measure. But in some settings such as survival models, the measure has a discrete part associated with censored values, and a continuous part associated with failure times.

The likelihood function is a fundamental object for statistical estimation and inference. For parametric Bayes tasks, the likelihood function is the ratio of the posterior distribution to the prior on Θ :

$$P_\pi(d\theta | Y) \propto L(\theta; y)\pi(d\theta).$$

Its non-Bayesian role is a little more complicated, but it is equally fundamental. Mostly it is used for point estimation and interval estimation.

17.2.2 Bartlett Identities

The likelihood is a function $L(\theta; y)$ of the parameter θ and the data y , and the same applies to the log likelihood $l(\theta; y) = \log L(\theta; y)$. Since the likelihood is defined up to an arbitrary multiplicative factor that is constant in θ , the log likelihood is defined up to an arbitrary additive term that is constant in θ .

The Bayesian goal is to compute the conditional probability of some specified inferential event given the data, and in that calculation y is regarded as a fixed constant. However, frequentist properties of estimators are connected with the statistical behaviour of log likelihood derivatives and related procedures for fixed θ as a function of the random variable whose distribution is P_θ . The Bartlett identities are fundamental for deriving large-sample asymptotic distributions in regular problems. To keep notation digestible, we pretend that θ is a scalar, so that each log likelihood derivative is also a scalar. Results for vector-valued parameters are obtained by replacing scalars with vectors or matrices as appropriate.

The first two log likelihood derivatives are

$$U_1(\theta; y) = dl(\theta; y)/d\theta; \quad U_2(\theta; y) = d^2l(\theta; y)/d\theta^2.$$

The Bartlett identities are connected with the moments of these and higher-order derivatives. The first identity follows from the constancy of the integral $\int p_\theta(y) dy$ as a function of θ .

$$\begin{aligned} 0 &= \frac{\partial}{\partial \theta} \int p_\theta(y) dy \\ &= \int \frac{\partial p_\theta(y)}{\partial \theta} dy \\ &= \int \frac{\partial \log p_\theta(y)}{\partial \theta} p_\theta(y) dy \\ &= E(U_1(\theta; Y); \theta). \end{aligned}$$

The first step in this derivation is to switch the order of differentiation with respect to θ and integration over the observation space. This step requires a regularity condition, which fails if the support of P_θ is parameter-dependent. Regularity conditions must be taught by faculty and learned by students, if only to demonstrate mastery of Fubini's theorem, but they almost never fail in practical work. In the last expression θ occurs twice, the first to indicate the differentiation point, the second to indicate that the parameter of the distribution $Y \sim P_\theta$ is the same as the point at which the derivative is computed. The random variable $U_1(\theta; Y)$ does not have zero mean under the distribution $Y \sim P_{\theta^*}$.

The second identity, which follows from the second derivative of the probability integral, establishes the role of the Fisher information matrix

$$\begin{aligned} 0 &= \frac{\partial^2}{\partial \theta^2} \int p_\theta(y) dy, \\ &= \frac{\partial}{\partial \theta} \int \frac{\partial \log p_\theta(y)}{\partial \theta} p_\theta(y) dy, \\ &= \int \left(\frac{\partial^2 \log p_\theta(y)}{\partial \theta^2} + \left(\frac{\partial \log p_\theta(y)}{\partial \theta} \right)^2 \right) p_\theta(y) dy, \\ &= E(U_2(\theta; Y); \theta) + E(U_1^2(\theta; Y); \theta). \end{aligned}$$

The final expression implies that the variance of the first derivative is the negative expected value of the second derivative, which is called the Fisher information matrix:

$$I(\theta) = -E(U_2(\theta; Y); \theta) = \text{cov}(U_1(\theta; Y); \theta).$$

It follows that $I(\theta) > 0$, and, for vector parameters, that $I(\theta)$ is positive definite.

17.2.3 Implications for Estimation

Regularity conditions for statistical work are of two types, those that can be checked or verified and those that cannot. Fubini-type conditions permitting the interchange of sample-space integration with parameter-space differentiation are verifiable. Conditions regarding the smoothness of functions or topological adequacy of the parameter space are also verifiable. Statistical models that occur in applied work are usually field-tested and are seldom in violation of verifiable conditions.

Asymptotic conditions holding in the large-sample limit are a different matter. Much of the theory of statistical estimation uses asymptotic theory as a device for distributional approximation. For simple processes having independent and identically distributed components, the only route to infinity is ‘more independent copies of the same’. For more general spatial or temporal processes, or processes involving covariates, the routes to infinity are more numerous. By their nature, asymptotic conditions are not verifiable in any finite sample because any finite design can be embedded into a sequence of larger designs in countless ways. The question to be asked is not whether the given design is part of a particular sequence but whether one conceptual design sequence provides a better distributional approximation than another.

The motivation for large-sample theory is most straightforward for independent and identically distributed sequences. Such sequences seldom occur naked in applied work, so the independent and identically distributed theory is not directly relevant. However, the crucial parts of the theory carry over with relatively minor modification to models having independent observations. Additional conditions are needed for asymptotic regularity in specialized models for genetics, time series and spatial processes. The count of individual numbers or observations or rows in a data file may be impressive, but that does not necessarily translate into an impressive quantity of information.

In a setting where the components of the response are independent, or conditionally independent given treatment, the log likelihood is a sum of n independent contributions, and the same applies to the log likelihood derivatives. In particular, the total Fisher information $I_+(\theta) = \sum I_i(\theta)$ is the sum of positive contributions coming from individual components. The first derivative at θ^* is the sum of independent random variables, U_1, \dots, U_n , having zero mean and finite variances $I_i(\theta) < \infty$. Provided that n is large and that no small subset of components dominates the contribution to the total Fisher information, the central limit theorem implies that the first derivative at θ^* is approximately normally distributed

$$U_+(\theta^*) \sim N(0, I_+(\theta^*))$$

under the distribution P_{θ^*} . Assuming that the maximum is a stationary point, Taylor approximation in a neighbourhood of θ^* gives

$$0 = U_+(\hat{\theta}) = U_+(\theta^*) - I_+(\theta^*)(\hat{\theta} - \theta^*) + O_p(1).$$

To first order in the sample size, this implies

$$\hat{\theta} - \theta^* \simeq I^{-1}(\theta^*)U(\theta^*) \sim N(0, I^{-1}(\theta^*)) \quad (17.1)$$

in which the uncomputable $I(\theta^*)$ may be replaced with the computable $I(\hat{\theta})$. Probability calculations using this asymptotic approximation have error of order $O(n^{-1/2})$ in the sample size. However, the error can often be reduced to acceptable levels by parameter transformation. More accurate approximations using bias corrections and Edgeworth series are available in the literature.

The linear approximation (17.1) is often re-packaged as a computational algorithm, which generates from a starting point $\hat{\theta}_0$ a parameter sequence satisfying

$$\hat{\theta}_{r+1} = \hat{\theta}_r + I^{-1}(\hat{\theta}_r)U(\hat{\theta}_r). \quad (17.2)$$

If this sequence converges, it converges to a stationary point of the log likelihood, which is a local maximum and usually the global maximum. Technically, the sequence (17.2) is not Newton-Raphson because it uses the Fisher information or expected second derivative at $\hat{\theta}_r$, which is not usually the same as the observed second derivative at that point.

17.2.4 Likelihood-Ratio Statistic I

Taylor expansion of the log likelihood function about θ^* including terms up to degree two in $\hat{\theta} - \theta^*$ gives

$$l(\hat{\theta}; y) - l(\theta^*; y) = U(\theta^*)(\theta^* - \hat{\theta}) - \frac{1}{2}I(\theta^*)(\theta^* - \hat{\theta})^2 + \dots.$$

For models having independent and identically distributed components, the first derivative is $O_p(n^{1/2})$, while second and higher-order derivatives are $O_p(n)$. As a result, both terms shown are formally $O_p(1)$ while the error term is $O_p(n^{-1/2})$. Under suitable asymptotic conditions, these asymptotic orders also hold more broadly for generalized linear models and many models having temporal or spatial correlation.

Using the one-step approximation (17.1) for the parameter estimate, the likelihood-ratio statistic satisfies

$$\begin{aligned} 2l(\hat{\theta}; y) - 2l(\theta^*; y) &= U(\theta^*)I^{-1}(\theta^*)U(\theta^*) + O_p(n^{-1/2}), \\ &= U(\theta^*)I^{-1}(\hat{\theta})U(\theta^*) + O_p(n^{-1/2}), \\ &= (\hat{\theta} - \theta^*)I(\hat{\theta})(\hat{\theta} - \theta^*) + O_p(n^{-1/2}). \end{aligned}$$

The first and second versions are positive definite quadratic forms in the vector of first derivatives at the true or hypothesized parameter point; the third is a quadratic form in the parameter space. The likelihood ratio statistic is invariant under smooth reparameterization, and that property is inherited by the first quadratic form shown, which is called the Rao statistic, or Fisher-Rao statistic. The third version, called the Wilks statistic, is not invariant under reparameterization. Invariance is desirable in applied work, but perhaps not absolutely essential.

The central limit approximation for the distribution of log likelihood derivatives implies that all three versions of the likelihood-ratio statistic are first-order equivalent, and that the limit distribution is χ_p^2 in all cases. They are not second-order equivalent, either in power or in distribution. A more refined analysis taking account of higher-order terms shows that the expected value of the likelihood-ratio statistic is $p(1 + b(\theta)/n)$, and that the asymptotic distribution is $(1 + b(\theta)/n)\chi_p^2$ with error $O(n^{-2})$. Division by the Bartlett adjustment factor $1 + b(\theta)/n$ greatly improves the accuracy of the χ_p^2 approximation. This adjustment holds fairly widely for regular problems with continuous distributions.

17.2.5 Profile Likelihood

Most parametric models that occur in applied work make a distinction between parameters of interest and other parameters, loosely called nuisance parameters. Despite the nomenclature, nuisance parameters are essential for satisfactory inferences.

The parameter of interest is defined by a differentiable function $T: \Theta \rightarrow \Theta'$ from the parameter space of dimension p into a manifold of dimension $q \leq p$. We suppose without loss of generality that this mapping is onto, i.e., $T\Theta = \Theta'$. To each $\tau \in \Theta'$ there corresponds a sub-manifold of dimension $p - q$

$$\Theta_\tau = \{\theta : T(\theta) = \tau\} \subset \Theta.$$

All points in Θ_τ are similar in the sense that they have the same value of the parameter of interest; differences are associated with nuisance parameters. By construction, the sub-manifolds are disjoint and exhaustive in Θ ; they form a partition or a foliation of the parameter space.

The profile likelihood for τ is the maximum achieved on Θ_τ :

$$l_p(\tau; y) = \max_{\theta \in \Theta_\tau} l(\theta; y) = l(\hat{\theta}_\tau; y).$$

To first order in the sample size, the profile likelihood behaves like an ordinary likelihood function. For example, the first derivative has mean of order $O(n^{-1})$, which is not zero but is small enough to permit the standard asymptotic argument to proceed. Likewise, the expected value of the second derivative is not exactly

the variance of the first, but the difference is small enough that it does not affect first-order asymptotic approximations under standard regularity conditions. Consequently, the subset consisting of parameter values achieving near-maximum likelihood

$$\{\tau \in \Theta' : 2l(\hat{\theta}; y) - 2l(\hat{\theta}_\tau; y) \leq \chi^2_{q, 1-\alpha}\}$$

is an approximate $1 - \alpha$ -confidence subset for the parameter of interest.

17.2.6 Two Worked Examples

Example 1: Treatment Effect Estimation

The standard Gaussian model for a completely randomized design has three parameters $\theta = (\mu_0, \mu_1, \sigma^2)$, two means and one variance $\sigma^2 > 0$, so Θ has dimension three. The log likelihood function is

$$l(\theta; y) = -\frac{n_0(\bar{y}_0 - \mu_0)^2}{2\sigma^2} - \frac{n_1(\bar{y}_1 - \mu_1)^2}{2\sigma^2} - \frac{(n-2)s^2}{2\sigma^2} - n \log \sigma$$

in standard notation for sample sizes n_0, n_1 , sample means \bar{y}_0, \bar{y}_1 , and pooled sample variance s^2 . The treatment effect is the difference $T(\theta) = \mu_1 - \mu_0$, and the focus of the analysis is primarily on that parameter. The profile log likelihood for the treatment effect is the maximum value achieved on the subset $\Theta_\tau \subset \Theta$

$$\Theta_\tau = \{\theta : \mu_1 - \mu_0 = \tau\}; \quad l(\hat{\theta}_\tau; y) = \max_{\theta \in \Theta_\tau} l(\theta; y).$$

Usually the maximum for fixed τ must be computed numerically, but it can be evaluated explicitly in this instance:

$$\begin{aligned} \hat{\mu}_0 &= (n_0\bar{y}_0 + n_1\bar{y}_1 - n_1\tau)/n; \\ \hat{\mu}_1 &= (n_0\bar{y}_0 + n_0\tau + n_1\bar{y}_1)/n = \hat{\mu}_0 + \tau; \\ n\hat{\sigma}^2 &= (n-2)s^2 + n_0(\bar{y}_0 - \hat{\mu}_0)^2 + n_1(\bar{y}_1 - \hat{\mu}_1)^2; \\ l(\hat{\theta}_\tau; y) &= \frac{n}{2} \log(\hat{\sigma}^2) + \text{const} \\ &= \frac{n}{2} \log((n-2)s^2 + n_0(\bar{y}_0 - \hat{\mu}_0)^2 + n_1(\bar{y}_1 - \hat{\mu}_1)^2) + \text{const}' \\ &= \frac{n}{2} \log((n-2)s^2 + n_0n_1(\bar{y}_0 - \bar{y}_1 + \tau)^2/n) + \text{const}'. \end{aligned}$$

The partially maximized likelihood function is called the profile likelihood for the parameter of interest. By construction, the overall maximum occurs at the ordinary maximum of the likelihood, $\hat{\tau} = \bar{y}_1 - \bar{y}_0$ in this example.

Asymptotically, the profile log likelihood has all of the essential properties of a log likelihood function. For example, an approximate level- α likelihood-based confidence interval can be obtained in the standard manner

$$\{\tau : 2l(\hat{\theta}; y) - 2l(\hat{\theta}_\tau; y) \leq \chi^2_{1,1-\alpha}\}. \quad (17.3)$$

In this example, it is possible to construct the standard exact confidence interval for τ using the ratio

$$t_\tau = \sqrt{\frac{n_0 n_1}{n}} \frac{\bar{Y}_0 - \bar{Y}_1 + \tau}{s},$$

which has the Student t distribution on $n - 2$ degrees of freedom. The exact coverage of the likelihood-based interval can be inferred from the fact that the likelihood-ratio statistic $n \log(1 + t_\tau^2/(n - 2))$ is monotone in t_τ^2 .

Example 2: Inference for the LD₉₀

Suppose that the response of unit i to dose x is a Bernoulli variable with parameter $\pi(x)$ satisfying the linear logistic model

$$\text{logit } \pi(x) = \theta_0 + \theta_1 x, \quad (17.4)$$

with independent responses for distinct units. The goal is to estimate the dose τ for which $\pi(\tau) = 0.9$, the so-called lethal dose 90%. The LD₉₀ is a non-linear function of the parameters

$$\begin{aligned} \text{logit}(0.9) &= \log(9) = \theta_0 + \theta_1 \tau; \\ \tau &= (\log(9) - \theta_0)/\theta_1, \end{aligned}$$

so we take $T(\theta) = (\log(9) - \theta_0)/\theta_1$. To compute the profile likelihood for τ , it is necessary to fit the logistic model (17.4) by maximizing over the parameter subset

$$\Theta_\tau = \{(\theta_0, \theta_1) : T(\theta) = \tau\} = \{(\log(9) - \tau\theta_1, \theta_1) : \theta_1 \in \mathbb{R}\}.$$

In other words, we aim to fit the one-parameter sub-model

$$\text{logit } \pi(x) = \log(9) - \tau\theta_1 + \theta_1 x = \log(9) + \theta_1(x - \tau)$$

for arbitrary but fixed τ . This is not a linear logistic model in the strict technical sense, but most computer packages have the option to cater for an offset, which is the constant $\log(9)$ in this setting. The likelihood-based confidence region for τ is the set of values for which the likelihood is sufficiently large in the sense of (17.3).

If we replace the linear logistic model (17.4) with a linear Gaussian model and ask for the x -value that makes the mean response zero, the goal is the abscissa or parameter ratio $\tau = -\theta_0/\theta_1$. Fieller's method is tailored for problems of this sort. However, the likelihood-ratio statistic is a function of the standardized ratio on which Fieller's method is based, so the two approaches are essentially identical.

One point to note is that a likelihood-based confidence set is not necessarily an interval. Equivariance under reparameterization makes this unavoidable. For instance, if the likelihood-based confidence set for $\tau = \theta_0/\theta_1$ is a bounded interval containing zero, the confidence set for $1/\tau$ is necessarily an ‘interval’ containing $\pm\infty$ but not zero. In both the linear logistic and Gaussian cases, the likelihood-based confidence set is either an interval or the complement of an interval, or possibly the whole space.

17.3 Generalized Linear Models

Let P_0 be a distribution on the state space \mathcal{S} , and let $T: \mathcal{S} \rightarrow \mathbb{R}$ be a real-valued random variable. The moment generating function of T is the expected value of the exponential integral

$$M_0(\theta) = \int_{\mathcal{S}} e^{\theta T(y)} P_0(dy),$$

so $M_0(0) = 1$. It is assumed that the generating function exists in the sense that the subset

$$\Theta = \{\theta : M_0(\theta) < \infty\},$$

has a non-empty interior that includes zero. This subset is a real interval, called the canonical parameter space. In that case M_0 has a Taylor series

$$\begin{aligned} M_0(\theta) &= \int \left(1 + \theta T(y) + \cdots + \frac{\theta^r T^r(y)}{r!} + \cdots \right) dP_0(y), \\ &= \sum \mu_r \theta^r / r!, \end{aligned}$$

which is convergent in a neighbourhood of the origin. The r th moment $\mu_r = E(T^r)$ is the r th derivative of M_0 at the origin.

For each $\theta \in \Theta$, the exponentially weighted distribution

$$P_\theta(dy) = \frac{e^{\theta T(y)} P_0(dy)}{M_0(\theta)}$$

is also a probability distribution on the state space. The set of distributions $\{P_\theta : \theta \in \Theta\}$ is called the exponential family associated with the pair (P_0, T) . For a random variable $Y \sim P_\theta$, the moment generating function of $T(Y)$ is

$$\begin{aligned} M_\theta(t) &= \int e^{tT(y)} dP_\theta(y) \\ &= \int_{\mathcal{S}} e^{tT(y)} e^{\theta T(y)} P_0(dy) / M_0(\theta) \\ &= M_0(\theta + t) / M_0(\theta). \end{aligned}$$

The cumulant generating function of $T(Y)$ under $Y \sim P_\theta$ is

$$K_\theta(t) = \log M_\theta(t) = K_0(\theta + t) - K_0(\theta).$$

This function also has a Taylor expansion in which the r th cumulant is the r th derivative of $K_\theta(t)$ at $t = 0$, which is the r th derivative of $K_0(\cdot)$ at θ . In particular, the mean is $\mu = K'(\theta)$, and the variance is $K''(\theta)$.

For all generalized linear models, the state space is the real line or a proper subset, and T is the identity function. The simplest example is the unit exponential distribution $P_0(dy) = e^{-y} dy$ for $y > 0$. In that case $M_0(\theta) = 1/(1-\theta)$ for $\theta < 1$, and P_θ is the exponential distribution with rate $1-\theta$ and mean $\mu = 1/(1-\theta) > 0$. The cumulant function is $K_0(\theta) = -\log(1-\theta)$. Standard examples of the same type are listed in Table 2.1 of McCullagh and Nelder (1989). They include the normal, binomial, Poisson, multinomial, hypergeometric and gamma families.

Examples in which the state space is not the real line include the Ewens distribution on set partitions (Exercises 10.4, 11.6) and the von-Mises-Fisher family on the circle or sphere (Exercise 14.8).

In every generalized linear model, the mean vector $\mu = E(Y)$ (or $\mu = E(T(Y))$) is a component-wise non-linear function of covariates $X\beta$, depending on regression parameters $\beta \in \mathbb{R}^p$. In a regular model, the derivative matrix with components $D_{ir} = \partial \mu_i / \partial \beta_r$ has full rank $p \leq n$ at every point. The gradient vector of the log likelihood with respect to β is

$$D' \Sigma^{-1} (Y - \mu)$$

where D and $\Sigma = \text{cov}(Y)$ depend on β through μ . Ordinarily, the maximum-likelihood estimate can be found by iteration

$$\hat{\beta} - \beta = (D' \Sigma^{-1} D)^{-1} D' \Sigma^{-1} (Y - \mu)$$

in which the μ , D , Σ are computed at the current value. Provided that all eigenvalues of $D'\Sigma^{-1}D$ are large, the asymptotic covariance of $\hat{\beta}$ is

$$\text{cov } \hat{\beta} \simeq (D'\Sigma^{-1}D)^{-1}.$$

17.4 Variance-Components Models

Let V_0, \dots, V_k be given $n \times n$ matrices that are linearly independent and symmetric. Usually $V_0 = I_n$ and the remaining matrices are symmetric positive semi-definite. A variance-components model is one in which the observation vector Y is a zero-mean Gaussian variable with covariance matrix satisfying the linear model

$$\Sigma = \gamma_0 V_0 + \dots + \gamma_k V_k$$

with coefficients γ_r to be estimated. The log likelihood derivative with respect to γ_r is

$$\frac{\partial l}{\partial \gamma_r} = \frac{1}{2} \text{tr} (\Sigma^{-1} V_r \Sigma^{-1} (YY' - \Sigma))$$

and the r, s component of the Fisher information matrix is

$$I_{rs} = \text{cov} \left(\frac{\partial l}{\partial \gamma_r}, \frac{\partial l}{\partial \gamma_s} \right) = \frac{1}{2} \text{tr} (\Sigma^{-1} V_r \Sigma^{-1} V_s).$$

Given an initial estimate γ , the one-step update $\hat{\gamma} - \gamma$ satisfies the linear equation

$$I(\hat{\gamma} - \gamma) = \partial l / \partial \gamma.$$

This formula works reasonably well in the vicinity of the solution, but some tempering is often needed to keep the early steps from straying into territory where Σ is not positive definite. In practice, some or all of the coefficients may also be subject to positivity conditions, so it is necessary to check for boundary components and to adjust computations for the remaining components.

17.5 Mixture Models

17.5.1 Two-Component Mixtures

Let ψ_0 and ψ_1 be the density functions of two distributions on the real line. Both densities are assumed to be strictly positive, so the density ratio $\zeta(y) =$

$\psi_1(y)/\psi_0(y)$ is finite, as is the inverse ratio. The mixture model refers to the family of distributions

$$\psi_\theta(y) = (1 - \theta)\psi_0(y) + \theta\psi_1(y),$$

which is a convex set indexed by the mixture parameter $0 \leq \theta \leq 1$.

According to the standard statistical paradigm, the observations Y_1, \dots, Y_n are independent and identically distributed as ψ_θ for some unknown parameter value. Statistically speaking, the estimation and testing problems are regular if $0 < \theta < 1$; in such circumstances, the standard asymptotic approximations hold for the distribution of $\hat{\theta}$ and for the likelihood-ratio statistic. Otherwise, if $\theta = 0$ or $\theta = 1$ on the boundary, the problem is non-regular; standard asymptotic approximations cannot be relied upon for either the distribution of $\hat{\theta}$ or of the likelihood-ratio statistic.

Given the observation $y = (y_1, \dots, y_n)$, the log likelihood function for θ is

$$l(\theta; y) = \sum \log(\psi_\theta(y_i)) = \sum \log(1 - \theta + \theta\zeta(y_i)) + \text{const}(y).$$

To understand the behaviour as a function of θ , we examine the derivatives

$$\begin{aligned} l'(\theta; y) &= \sum_i \frac{\zeta(y_i) - 1}{1 - \theta + \theta\zeta(y_i)}; \\ l''(\theta; y) &= - \sum_i \left(\frac{\zeta(y_i) - 1}{1 - \theta + \theta\zeta(y_i)} \right)^2 < 0. \end{aligned}$$

If all of the observation points satisfy $\psi_0(y_i) = \psi_1(y_i)$, then $\zeta(y_i) = 1$ for each i , and the log likelihood is constant in θ . Otherwise, the second derivative is everywhere strictly negative, implying concavity. Every stationary point is a global maximum, and there is at most one such point in $(0, 1)$.

At the left end-point $l'(0; y) = \sum \zeta(y_i) - n$; if the derivative at zero is negative, i.e., if $\sum \zeta(y_i) \leq n$, the maximum occurs at $\hat{\theta} = 0$. At the right end-point $l'(1; y) = n - \sum 1/\zeta(y_i)$. If the derivative is positive, i.e., if $\sum 1/\zeta(y_i) \leq n$, the maximum occurs at $\hat{\theta} = 1$. The likelihood function has a maximum in the interior of the interval if and only if $\sum \zeta(y_i) > n$ and $\sum 1/\zeta(y_i) > n$. In that case, the maximum can be computed by a straightforward Newton-Raphson iteration.

For $0 < \hat{\theta} < 1$, the condition $l'(\hat{\theta}; y) = 0$ implies

$$\sum \frac{\zeta(y_i)}{1 - \hat{\theta} + \hat{\theta}\zeta(y_i)} = \sum \frac{1}{1 - \hat{\theta} + \hat{\theta}\zeta(y_i)} = n,$$

which can be viewed as a self-consistency condition. If we associate with each i the class-I assignment probability

$$\hat{\theta}(y_i) = \frac{\hat{\theta}\zeta(y_i)}{1 - \hat{\theta} + \hat{\theta}\zeta(y_i)} = \text{pr}(i \mapsto \text{class I} | Y),$$

then $\hat{\theta} = n^{-1} \sum \hat{\theta}(y_i)$ is the sample mean of the assignment probabilities.

17.5.2 Likelihood-Ratio Statistic

For likelihood-ratio statistics it is convenient to take ψ_0 as the reference point. Relative to that point, the maximized likelihood ratio statistic is

$$l(\hat{\theta}; y) - l(0; y) = \sum \log(1 - \hat{\theta} + \hat{\theta}\zeta(y_i)).$$

In particular, the likelihood-ratio statistic is zero if $\hat{\theta} = 0$, i.e., if $\sum \zeta(y_i) \leq n$.

If we regard ψ_0 as the null hypothesis, it must be understood that $\theta = 0$ is a boundary point, and that the standard asymptotic theory may fail—and indeed it does fail spectacularly. For the null distribution theory, the observations are independent with distribution ψ_0 . An elementary computation shows that if $Y \sim \psi_0$, the random variable $\zeta(Y) = \psi_1(Y)/\psi_0(Y)$ is non-negative with mean one. Thus, by the law of large numbers, $n^{-1} \sum \zeta(Y_i) \rightarrow 1$. In addition, if $\zeta(Y)$ has finite variance, the central limit theorem implies asymptotic normality, so that the event $\sum \zeta(Y_i) \leq n$ occurs with limiting probability one half. In those cases, the null distribution of $\hat{\theta}$ has an atom of 1/2 at the origin, and the same goes for the likelihood-ratio statistic. This sort of behaviour is non-standard, but it is classical for boundary-point problems.

For the more usual sorts of mixtures that occur in practical applications, $\zeta(Y)$ does not have finite variance. In those cases, the convergence of the average $n^{-1} \sum \zeta(Y_i) \rightarrow 1$ does not imply that the event $n^{-1} \sum \zeta(Y_i) \leq 1$ has a limiting probability or that the limit is one half. As an example, if ψ_0 is standard normal, and ψ_1 is Cauchy, the random variable $\zeta(Y)$ has a density whose tail behaviour is $O(z^{-2} \log(z)^{-3/2})$. The mean is one, but there are no finite moments beyond the first. The limit distribution appears from simulation to be such that

$$n^{-1} \sum_{i=1}^n \zeta(Y_i) = 1 - \frac{\text{const}}{\log \log n} + \frac{\epsilon}{\log n \log \log n},$$

where ϵ is a random variable in the Landau class (stable with $\alpha = \beta = 1$). The event $\sum \zeta(Y_i) > n$ is equivalent in the limit to $\epsilon > \text{const} \times \log n$. Since the Landau density has an inverse-square right tail, the probability is $O(1/\log n)$.

Every mixture model in which ψ_1 is symmetric with inverse-square tails gives the same limit. For other distributions having sub-Gaussian tails such as $e^{-|y|}$, the same limit is approached at a possibly different rate. In all such cases, the limiting null distribution for $\hat{\theta}$ and the likelihood-ratio statistic are degenerate at zero. This is not standard asymptotic behaviour for boundary-point problems.

17.5.3 Sparse Signal Detection

Given a random signal $X \sim P$ and an observation $Y = X + \varepsilon$ contaminated by additive independent Gaussian noise, how do we estimate the signal? The non-sparse signal estimation problem was first posed by Frank Watson Dyson in 1926. Eddington's solution, which is described in Sect. 15.4.4, depends only on the marginal density $m(y)$ of the observation. The sparse version of the problem is discussed by Johnstone and Silverman (2004). For simplicity, we assume here that ε is standard normal.

A signal $X \sim P$ is said to be sparse if its distribution is symmetric and most of the mass is concentrated at or near the origin. In that case, the sparsity rate ρ is defined by the integral

$$1 - \rho = \int e^{-x^2/2} p(x) dx.$$

The statement that ρ is small is to be interpreted as a mathematical code or convention, which implies a formal limit $\rho \rightarrow 0$ even if that is not explicitly stated. The reason for focusing on the sparsity rate as opposed to the null atom $P(X = 0)$ is that the null atom may be zero; more crucially, ρ is a mixture fraction that is identifiable from observations whereas the null atom is not. Subsequent conclusions depend only on the sparsity rate, which is strictly smaller than the probability of a non-null signal.

Under regularity conditions given in McCullagh and Polson (2018), the marginal density is a Gaussian mixture

$$m(y) = \phi(y)(1 - \rho + \rho\zeta(y)) + o(\rho),$$

where the density ratio ζ is a symmetric non-negative convex function satisfying $\zeta(0) = 0$. In practice, ρ must first be estimated from the data.

The essence of the matter is that all sparse scale families having similar tail behaviour give rise to the same zeta function. The horseshoe family with density $\log(1 + y^{-2})/(2\pi)$ has the same inverse-square tail behaviour as the Cauchy family, and the zeta function for both satisfies $\zeta''(y) = \exp(y^2/2)$. This implies $\zeta(y) = \sum_{r \geq 1} \mu_{2r-2} y^{2r}/(2r)!$, where $\mu_{2r} = 1 \cdot 3 \cdots (2r-1)$ is the $2r$ th standard Gaussian moment. It is possible to make an elaborate argument for one over the other, but

such arguments are futile because the marginal distributions are indistinguishable to first order.

Eddington's signal-estimation formula reduces to

$$E(X_i | Y) = \frac{\rho \zeta'(y_i)}{1 - \rho + \rho \zeta(y_i)} + o(\rho)$$

$$E(X_i^2 | Y) = \frac{\rho \zeta''(y_i)}{1 - \rho + \rho \zeta(y_i)} + o(\rho).$$

In addition, the signal identification or conditional exceedance probability for threshold $\epsilon > 0$ is formally the same as $E(|X_i|^0 | Y)$:

$$P(|X_i| > \epsilon | Y) = \frac{\rho \zeta(y_i)}{1 - \rho + \rho \zeta(y_i)} + o(1).$$

For a given suitably low but strictly positive threshold, the exceedance probability is approximately independent of the threshold (McCullagh and Polson, 2018), and $\zeta(y)$ is interpretable as the posterior-to-prior odds ratio, also called the Bayes factor. For example, if $\rho = 0.05$ and $y = 3.5$, the inverse-square zeta value is $\zeta(y) = 55.3$ and the exceedance probability is 0.74. Provided that below-threshold signals are counted as null or false, there is a close formal connection with the concept of false discovery rate, and particularly with the local false discovery rate (Benjamini & Hochberg, 1995; Efron, 2010).

In the great majority of sparse signal identification and detection formulations the second component of the mixture $\psi_2(y) = \phi(y)\zeta(y)$ has heavy tails. The tails are governed by the signal distribution, which may be either Laplace-type $e^{-|y|}$ or Cauchy-like with regularly varying tails. In all such models, the asymptotic null distribution of $\hat{\rho}$ and of the likelihood-ratio statistic is degenerate at zero. However, a very small signal little larger than $\rho \simeq \log(n)/n$ is enough to change the calculus for signal detection, at least in the Cauchy case.

17.6 Inferential Compromises

The Dictatorial Compromise

The fundamental difficulty with the non-Bayesian model $\{P_\theta : \theta \in \Theta\}$ for inferential purposes is that it contains more than one stochastic process. Which one, if any, are we to use for prediction? The Bayesian paradigm resolves the difficulty by compromise, which—however reasonable it may be and however little its effect on conclusions may be—is ultimately dictatorial. That compromise consists of a prior distribution or mixture $\pi(d\theta)$, so that the set $\{P_\theta\}$ is replaced with the single mixture $P_\pi = \int P_\theta \pi(d\theta)$. Prediction is then straightforward in principle.

For parametric inference, the event $\theta \in A$ is identified with the set of sequences $y \in \mathbb{R}^\infty$ such that $\hat{\theta}(y) = \lim_{n \rightarrow \infty} \hat{\theta}_n(y[n])$ exists and belongs to A . In that way, the conditional probability of the event $\theta \in A$ given $Y[n]$ is computable as a tail event

$$P_\pi(\theta \in A \mid Y[n]) = P_\pi(\hat{\theta}(Y) \in A \mid Y[n]).$$

Any consistent estimator can be used in place of $\hat{\theta}_n$, so this description is not tied in any way to maximum likelihood.

The One-Time Representative

Consider a standard non-Bayesian model consisting of processes P_θ , with finite-dimensional distributions $P_{n,\theta}$ on \mathbb{R}^n . Maximum-likelihood estimation offers two ways to generate a new process that is related to the family. The first of these is the standard parametric bootstrap, or parametric simulation.

Bootstrap process: Given a observation point $y \in \mathbb{R}^m$, and the corresponding point $\hat{\theta} = \hat{\theta}_m(y)$ in Θ , the entire family $\{P_\theta\}$ is replaced with the maximum-likelihood representative $\hat{P} = P_{\hat{\theta}}$. The finite-dimensional distributions are $\hat{P}_n = P_{n,\hat{\theta}}$ on \mathbb{R}^n . In particular, if each P_θ defines a process with independent components, the bootstrap representative is also a process whose components are conditionally independent given $\hat{\theta}_n$.

The Sequential Representative

The maximum-likelihood process (MLP) operates in a different manner and exhibits fundamentally different behaviour, which is analogous to the behaviour of a Polya urn. Every process P_θ determines a Markov kernel, which associates with each $n \geq 0$ and each point $y \in \mathbb{R}^n$, a conditional distribution

$$Q_{n+1,\theta}(dy_{n+1}; y) = P_{n+1,\theta}(dy_{n+1} \mid Y[n] = y)$$

on \mathbb{R} . The finite-dimensional joint density is the product of such kernels

$$P_{n,\theta}(dy) = Q_{1,\theta}(dy_1)Q_{2,\theta}(dy_2; y[1]) \cdots Q_{n,\theta}(dy_n; y[n-1]).$$

The conditional distribution given $Y[m]$ is a truncated kernel product

$$Q_{m+1,\theta}(dy_{m+1}; y[m])Q_{m+2,\theta}(dy_{m+2}; y[m+1]) \cdots Q_{m+n}(dy_{m+n,\theta}; y[m+n-1]).$$

In the maximum-likelihood process, each transition kernel $Q_{n+1,\theta}(dy_{n+1}; y)$ is replaced with the maximum-likelihood estimate

$$\hat{Q}_{n+1}(dy_{n+1}; y) = Q_{n+1,\hat{\theta}_n(y)}(dy_{n+1}; y).$$

This makes sense only for n sufficiently large that $\hat{\theta}_n = \hat{\theta}_n(y[n])$ exists. Given an initial sequence $y[m]$, the distribution of successive values in the MLP is defined by the kernel product

$$Q_{m+1, \hat{\theta}_m}(dy_{m+1}; y[m]) \times Q_{m+2, \hat{\theta}_{m+1}}(dy_{m+2}; y[m+1]) \times \cdots$$

in which the maximum-likelihood point is updated at each stage.

17.7 Exercises

17.1 Maximum-likelihood for mixtures: Let $\psi_0(\cdot), \dots, \psi_k(\cdot)$ be given probability density functions on \mathbb{R} , and let

$$m_\theta(y) = \theta_0\psi_0(y) + \cdots + \theta_k\psi_k(y)$$

be a $k+1$ -component mixture with non-negative weights adding to one. Suppose that Y_1, \dots, Y_n are independent and identically distributed with density m_θ , assumed to be strictly positive for θ strictly positive. Under what conditions is the mixture model with independent and identically distributed observations identifiable? Show that the maximum-likelihood estimator satisfies the condition

$$\sum_{i=1}^n \frac{\psi_r(y_i)}{\hat{m}(y_i)} \leq n,$$

with equality for every r such that $\hat{\theta}_r > 0$. Discuss the ‘almost-true’ claim that \hat{m} exists and is unique for every $n \geq 1$ and every $y \in \mathbb{R}^n$, even if the model is not identifiable.

17.2 Let $\psi_0(y) = e^{-y^2/2}/\sqrt{2\pi}$ be the standard normal density. Assume that Y_1, \dots, Y_n are independent standard normal. Show that the random variables $X_i = \psi_1(Y_i)/\psi_0(Y_i)$ have unit mean, and hence, by the law of large numbers, that the sample average tends to one as $n \rightarrow \infty$.

17.3 If the claim made in the last paragraph of section 17.5.2 is to be believed, the re-scaled limit distribution of \bar{X}_n does not have a mean. Discuss this apparent contradiction.

17.4 Consider the two-component mixture with ψ_0 standard normal, and ψ_1 standard Cauchy. The null hypothesis is all Gaussian, i.e., $\theta = (1, 0)$. Show that $\hat{\theta}_1 > 0$ if and only if $\bar{X}_n > 1$. By simulation or otherwise, show that $P_0(\bar{X}_n > 1) \rightarrow 0$ as $n \rightarrow \infty$. What is the effect of changing the Cauchy scale parameter?

17.5 Show that the random variables $X_i = \psi_1(Y_i)/\psi_0(Y_i)$ in the preceding exercise have a density whose tail behaviour is $1/f(x) \sim x^2 \log(x)^{3/2}$ as $x \rightarrow \infty$.

17.6 Explain why the observation $P_0(\bar{X}_n > 1) \rightarrow 0$ as $n \rightarrow \infty$ deduced from simulations does not conflict with the law of large numbers $\bar{X}_n \rightarrow 1$.

17.7 Consider the two-component mixture with ψ_0 standard normal and ψ_1 standard Laplace, or double exponential. Investigate the behaviour of $P_0(\bar{X}_n > 1)$ as a function of n for large n . What is the effect of changing the scale parameter? What do these calculations imply about the null distribution of the likelihood-ratio statistic?

17.8 Sparse signal detection. Suppose that the observation $Y = X + \varepsilon$ is the sum of a signal X plus independent Gaussian noise $\varepsilon \sim N(0, 1)$. For any signal distribution $X \sim P_\nu$, the sparsity rate is defined by the integral

$$\rho = \int (1 - e^{-x^2/2}) P_\nu(dx).$$

Suppose that the signal is distributed according to the Dirac-Cauchy mixture $P_\nu(dx) = (1 - \nu)\delta_0(dx) + \nu C(dx)$ in which the null atom $1 - \nu$ is the null-signal rate. Find the sparsity rate corresponding to 5% non-zero signals.

17.9 For the setting of the previous exercise, show that Y is distributed according to the mixture with density

$$m(y) = (1 - \rho)\phi(y) + \rho\psi(y) + o(\rho) = \phi(y)(1 - \rho + \rho\zeta(y)) + o(\rho)$$

where $\psi(\cdot)$ is a probability density, $\zeta(y) = \psi(y)/\phi(y)$ is the density ratio, and $\zeta(0) = 0$. Fill in the details needed to express $\zeta(\cdot)$ or $\psi(\cdot)$ as a function of the family P_ν .

17.10 Suppose that Y_1, \dots, Y_n are independent and identically distributed with density $m(y)$. Ignoring the error term, show that the maximum-likelihood estimate of the mixture fraction is zero if $\sum \zeta(y_i) \leq n$, one if $\sum 1/\zeta(y_i) \leq n$, and otherwise is a point $0 < \hat{\rho} < 1$ satisfying

$$\sum \frac{\zeta(y_i) - 1}{1 - \hat{\rho} + \hat{\rho}\zeta(y_i)} = 0.$$

Hence or otherwise deduce that the maximum-likelihood estimate of the mixture satisfies the self-consistency condition

$$\sum \frac{\psi(y_i)}{\hat{m}(y_i)} = \sum \frac{\phi(y_i)}{\hat{m}(y_i)} = n.$$

In what sense does this equation imply self-consistency?

17.11 A sequence $\epsilon_v \rightarrow 0$ such that $P_v(|X| < \epsilon_v) \rightarrow 1$ as $v \rightarrow 0$ is called a signal negligibility threshold. Show that the conditional probability of a non-negligible signal is

$$P_v(|X| > \epsilon_v \mid Y) = \frac{\rho\zeta(y)}{1 - \rho + \rho\zeta(y)} + o(1),$$

which implies that the ‘true discovery rate’ is essentially independent of the threshold.

17.12 What does the preceding equation imply about the fraction of non-negligible signals among sites in the sample such that $|Y_i| \geq 3$?

17.13 Let $\kappa_0 = \rho\zeta(y)/(1 - \rho + \rho\zeta(y))$ be the exceedance probability, and let κ_r be the r th derivative of $\log(1 - \rho + \rho\zeta(y))$. For $\zeta(y) \simeq e^{y^2/2}/y^2$ for large y , show that Eddington’s formulae give

$$E(X \mid Y) \simeq \kappa_0(y^2 - 2)/|y|, \quad \text{var}(X \mid Y) \simeq \kappa_0(1 - \kappa_0)(y^2 - 3) + \kappa_0^2$$

for large y . Discuss the implications for mean shrinkage and variance inflation.

17.14 For $1 \leq i \leq k$, suppose $Y_i = \alpha_i + \epsilon_i$, where $\epsilon_1, \dots, \epsilon_k$ are independent standard normal variables, and $\alpha_1, \dots, \alpha_k$ are exchangeable and independent of ϵ . Let F be the joint distribution of α . The goal of this exercise is to find an estimator of F as a function of the observation $y \in \mathbb{R}^k$. Ideally the estimator should be the maximum-likelihood estimator or an approximation thereof within a set of distributions having some natural symmetry. In the candidate estimators listed below, λ and s are unspecified scalars, $\delta_x(\cdot)$ is the Dirac measure at x , \mathcal{M}_k is the set of functions $[k] \rightarrow [k]$, $\mathcal{S}_k \subset \mathcal{M}_k$ is the set of permutations, and τy is the composition $(\tau y)_i = y_{\tau(i)}$.

$$\hat{F}_0(\cdot) = \frac{1}{k!} \sum_{\tau \in \mathcal{S}_k} \delta_{\lambda \tau y}(\cdot);$$

$$\hat{F}_1(\cdot) = \frac{1}{k^k} \sum_{\tau \in \mathcal{M}_k} \delta_{\lambda \tau y}(\cdot);$$

$$\hat{F}_2(\cdot) = N_k(\mathbf{1}\bar{y}, s^2 I_k).$$

Show that \hat{F}_0 and \hat{F}_1 are both exchangeable with the same marginal distribution, and that \hat{F}_1 also has independent components. For $\lambda = 1$, these are called the permutation estimator and the bootstrap estimator respectively.

Acknowledgement I am grateful to Nick Polson for bringing the Eddington/Dyson paper to my attention.

Chapter 18

Residual Likelihood



18.1 Background

Residual maximum likelihood (REML) is a technique proposed by Patterson and Thompson (1971) for estimating variances and variance components in a linear Gaussian model in which the observation $y \in \mathbb{R}^n$ is regarded as a realization of the Gaussian random vector $Y \sim N_n(X\beta, \Sigma)$. The model matrix, or design matrix, X is of order $n \times p$ and known, with image subspace $\mathcal{X} = \text{span}(X)$. In the simplest version of the covariance model, the matrix is expressed as a linear combination of given symmetric non-negative definite matrices

$$\Sigma = \sigma_1^2 V_1 + \cdots + \sigma_k^2 V_k \quad (18.1)$$

with non-negative coefficients $\sigma_1^2, \dots, \sigma_k^2$ to be estimated. These coefficients are called *variance components*; the space of matrices determined by (18.1) is the convex cone spanned by the given matrices. Usually V_1 is the identity matrix of order n ; the remaining matrices are typically block factors or other known relationships among the observational units.

In other settings, the model for Σ may not be linear in all parameters, but partial linearity is fairly common, as is linearity after transformation. In a spatial or time-series setting, the variance model may be a combination such as

$$\text{cov}(Y_s, Y_t) = \sigma_0^2 \delta_{s-t} + \sigma_1^2 e^{-\lambda|s-t|}$$

which is linear for fixed λ . On the other hand, a Gaussian graphical model is an additive specification for the inverse covariance matrix. The simplest version is

$$\Sigma^{-1} = \tau_0 I_n + \tau_1 G,$$

where $G \subset [n]^2$ is the graph incidence matrix, and the coefficients are subject to positive-definiteness conditions. Usually, this means $\tau_0 > 0$ and $\tau_1 \leq 0$.

Residual likelihood differs from ordinary likelihood in that it uses only the residuals $R = LY$, where L is any linear transformation such that

$$\ker(L) = \mathcal{X} = \{X\beta : \beta \in \mathbb{R}^p\}.$$

By focusing on the residuals, the regression parameters are eliminated from the distribution

$$R \sim N_n(LX\beta, L\Sigma L') = N_n(0, L\Sigma L').$$

It is crucial that R be observable, which means that L is a fixed linear transformation independent of all parameters. Under (18.1), the residual covariance matrix $L\Sigma L'$ is a linear combination of the matrices LV_rL' , so the residual likelihood is a function of the variance components only. Ordinarily, the matrices LV_rL' are non-zero and linearly independent, which implies that the variance-components are identifiable from the residuals. After estimating the variance components by maximizing the residual likelihood, the second step is to compute

$$\hat{\Sigma} = \sum_{r=1}^k \hat{\sigma}_r^2 V_r,$$

its inverse $\hat{W} = \hat{\Sigma}^{-1}$, and the weighted least squares estimate of β

$$\hat{\beta} = (X'\hat{W}X)^{-1} X' \hat{W}Y. \quad (18.2)$$

The covariance matrix of $\hat{\beta}$ is then reported as $(X'\hat{W}X)^{-1}$.

It is important in applications that the number of variance parameters be small relative to the sample size. Otherwise, if the variance matrix is not consistently estimable, the weighted least squares estimate (18.2) may be inefficient or even inconsistent in the large-sample limit. A good rule of thumb is one decimal digit, i.e., $k \leq 9$ regardless of the sample size; $k \leq 5$ is more normal. For consistent estimation of variance components, asymptotic theory requires that the Fisher information matrix

$$I_{rs} = \frac{1}{2} \text{tr}(V_r Q \Sigma V_s Q \Sigma)$$

be well-behaved in the sense that $n^{-1} I$ has a limit whose eigenvalues are strictly positive. For the residual likelihood (18.5), $Q = I - X(X'WX)^{-1}X'W$ is the orthogonal projection whose kernel is \mathcal{X} .

18.2 Simple Linear Regression

In a simple linear regression model with a single variance component, the covariance matrix is $\Sigma = \sigma^2 V$, where V is known and strictly positive definite. It is convenient in this setting to take $W = V^{-1}$ as the inner-product matrix, so that $P_X = X(X'WX)^{-1}X'W$ and $Q = I - P$ are complementary W -orthogonal projections. Then WQ and QV are both known and symmetric. The model for the residual $QY \sim N(0, \sigma^2 QV)$ has only a single parameter. For this full exponential-family setting, the quadratic form $\|QY\|^2 = Y'WQY$ is minimal sufficient, and the REML estimate is obtained by equating the observed value $\|Qy\|^2$ to its expected value:

$$\|Qy\|^2 = E(Y'WQY; \hat{\sigma}^2) = \hat{\sigma}^2 \text{tr}(VWQ) = \hat{\sigma}^2 \text{tr}(Q).$$

Since Q is a projection with rank $\text{tr}(Q) = n - p$, the REML estimate reduces to the standard unbiased estimator that is universally recommended and used in all computer packages

$$\hat{\sigma}^2 = y'WQy/(n - p).$$

Note that the REML estimate is strictly larger than the ordinary maximum-likelihood estimator which is $y'WQy/n$. The lesson here is that REML, not ML, is the norm for variance estimation.

18.3 The REML Likelihood

18.3.1 Projections

For the development in this section, K is a matrix of order $n \times k$, whose columns span a subspace \mathcal{K} of dimension k , and T is a complementary matrix of order $n \times (n - k)$ such that $T'K = 0$. In other words, the linear transformation $T': \mathbb{R}^n \rightarrow \mathbb{R}^{n-k}$ satisfies $\ker(T') = \mathcal{K}$. For the moment, the relation between \mathcal{K} and \mathcal{X} is left unspecified, but $\mathcal{K} = 0$, $\mathcal{K} = \mathcal{X}$ and $\mathcal{K} \subset \mathcal{X}$ are the most important special cases.

The observation space \mathbb{R}^n is regarded as a real inner-product space with inner product matrix $W = \Sigma^{-1}$. For most purposes, W can be replaced with any proportional matrix, as was done in Sect. 1.2. Consider the three $n \times n$ matrices:

$$P = K(K'WK)^{-1}K'W; \quad Q = I - P; \quad A = \Sigma T(T'\Sigma T)^{-1}T'. \quad (18.3)$$

It is readily checked that $P^2 = P$, $Q^2 = Q$ and $A^2 = A$, so all three are idempotent, and thus linear projections $\mathbb{R}^n \rightarrow \mathbb{R}^n$. They are also self-adjoint, meaning that WP , WQ and WA are symmetric, which implies all three are orthogonal projections. In addition, $x \in \mathcal{K}$ implies $T'x = 0$, which implies $Ax = 0$, and hence $\mathcal{K} \subset \ker(A)$.

Finally, $Ax = 0$ implies $T'Ax = 0$, which implies $T'x = 0$, which implies $x \in \mathcal{K}$, and hence $\ker(A) \subset \mathcal{K}$. Thus, A is the orthogonal projection with kernel \mathcal{K} , and Q is also the orthogonal projection with kernel \mathcal{K} . Uniqueness of orthogonal projections implies $A = Q$ and $I - A = P$.

18.3.2 Determinants

Now consider the partitioned matrix $H = [T, K]$, which is invertible of order n , and the related matrix

$$H'\Sigma H = \begin{pmatrix} T'\Sigma T & T'\Sigma K \\ K'\Sigma T & K'\Sigma K \end{pmatrix}.$$

The condition $T'K = 0$ implies that $H'H$ is block-diagonal with determinant $\det(H'H) = \det(T'T)\det(K'K)$, and hence that

$$\det(H'\Sigma H) = \det(H'H)\det(\Sigma) = \det(T'T)\det(K'K)\det(\Sigma).$$

Using the standard formula for the determinant of a partitioned matrix, we find

$$\begin{aligned} \det(H'\Sigma H) &= \det(T'\Sigma T)\det(K'\Sigma K - K'\Sigma T(T'\Sigma T)^{-1}T'\Sigma K) \\ &= \det(T'\Sigma T)\det(K'[\Sigma - T(T'\Sigma T)^{-1}T'\Sigma]K) \\ &= \det(T'\Sigma T)\det(K'(I - A)\Sigma K) \quad \text{from (18.3)} \\ &= \det(T'\Sigma T)\det(K'K(K'WK)^{-1}K'W\Sigma K) \\ \det(T'T)\det(K'K)\det(\Sigma) &= \det(T'\Sigma T)\det^2(K'K)/\det(K'WK) \\ \frac{\det(T'T)}{\det(T'\Sigma T)} &= \frac{\det(K'K)}{\det(K'WK)\det(\Sigma)}. \end{aligned}$$

For REML applications where the kernel is specified by K , the determinantal term in the marginal likelihood is the expression on the right.

18.3.3 Marginal Likelihood with Arbitrary Kernel

For any linear transformation such as T' having kernel \mathcal{K} , the linear transformation $Y \mapsto T'Y$ is called a residual modulo \mathcal{K} . All transformations having the given kernel determine the same likelihood function. The marginal log likelihood based on the linear transformation $T'Y \sim N(T'\mu, T'\Sigma T)$ is

$$l = -\frac{1}{2}(y - \mu)'T(T'\Sigma T)^{-1}T'(y - \mu) - \frac{1}{2}\log\det(T'\Sigma T) + \text{const.}$$

In this setting, l is a function on the parameter space, and the additive constant may be any function that is constant on the parameter space. It is convenient here to take a particular constant, namely $\frac{1}{2} \log \det(T'T)$ plus any function of y . This choice ensures that, for every invertible matrix L of order $n - k$, the linear transformations T' and LT' produce identical versions of the log likelihood. With this choice, the marginal log likelihood based on the residuals modulo \mathcal{K} is one half of

$$\begin{aligned} 2l &= -(y - \mu)'WA(y - \mu) + \log \det(T'T) - \log \det(T'\Sigma T) \\ &= -(y - \mu)'WQ(y - \mu) - \log \det(\Sigma) \\ &\quad - \log \det(K'WK) + \log \det(K'K), \end{aligned} \quad (18.4)$$

where Q is the orthogonal projection with kernel \mathcal{K} , and K is any matrix whose columns span \mathcal{K} .

In applications where \mathcal{X} is the model subspace, the most common choice is $\mathcal{K} = \mathcal{X}$, but expression (18.4) is valid for all subspaces, and $\mathcal{K} \subset \mathcal{X}$ arises in the computation of likelihood-ratio statistics. The ordinary log likelihood with kernel $\mathcal{K} = 0$ is obtained by setting $K = 0$. The standard REML likelihood has $K = X$ and $\mathcal{K} = \mathcal{X}$ so that $\mu \in \mathcal{X}$ implies $Q\mu = 0$:

$$2l = -y'WQy - \log \det(\Sigma) - \log \det(X'WX) + \log \det(X'X). \quad (18.5)$$

Formulae (18.4) and (18.5) may be used directly in computer software. The constant term $\log \det(K'K)$ is included to ensure that the log likelihood depends on the kernel subspace, not on the particular choice of basis vectors.

For general-purpose computer software, these formulae are not recommended because the marginal likelihood requires only that $T'\Sigma T$ be positive definite, which is a weaker condition than positive-definiteness for Σ . Marginal likelihood modulo a suitable kernel may be used for fitting generalized Gaussian processes, sometimes called intrinsic processes, that are defined by a generalized covariance function, which is not positive definite in the normal sense, but for which $T'KT$ is positive definite. For example, if $i \mapsto z_i$ is a quantitative covariate taking values in \mathbb{R}^k , the matrix $\Sigma_{ij} = -\|z_i - z_j\|$ is positive definite in the Euclidean space $\mathbb{R}^n/\mathbf{1}$, which is the space of residuals modulo the one-dimensional subspace of constant functions. In other words, for general-purpose computer software, it is best to use a version of (18.5) that does not require Σ to be positive definite or invertible.

18.3.4 Likelihood Ratios

A likelihood ratio at E is the ratio of probabilities assigned to the event E by two probability measures:

$$LR_{\theta',\theta}(E) = \frac{P_{\theta'}(E)}{P_\theta(E)}.$$

A maximized likelihood ratio is a similar expression

$$\frac{\sup_{\theta \in \Theta_1} P_\theta(E)}{\sup_{\theta \in \Theta_0} P_\theta(E)}$$

in which the numerator and denominator are maximized over the respective parameter spaces. It is crucial that all probability measures be defined on the same σ -field and that the event in the numerator be the same as the event in the denominator; otherwise the ratio is not a fair comparison. In fact, E is always an observation or singleton event, which is best regarded as an infinitesimal event, and commonly denoted by $E = dy$. Operationally speaking, dy is the limiting ϵ -ball $B(y, \epsilon)$ centered at the observation point $y \in \mathbb{R}^n$, and the likelihood ratio is the density ratio at y .

In the case of marginal likelihood, however, the event $E \subset \mathbb{R}^n$ is necessarily an event in the σ -field generated by the linear transformation T' into the Borel space \mathbb{R}^{n-k} . The induced σ -field in \mathbb{R}^n is the class of residual events, which are the Borel subsets $E \subset \mathbb{R}^n$ such that $E + \mathcal{K} = E$. In other words, a residual is a point in the quotient space \mathbb{R}^n/\mathcal{K} , and each residual event E is a union of translates of \mathcal{K} , i.e., a union of \mathcal{K} -cosets. The residual event, $E = B(y, \epsilon) + \mathcal{K}$, is the union of \mathcal{K} -cosets that intersect the ball. This is, of course a Borel subset in the space of residuals modulo \mathcal{K} . A residual likelihood ratio statistic modulo \mathcal{K} is thus a ratio of the form

$$\frac{\sup_{\theta \in \Theta_1} P_\theta(dy + \mathcal{K})}{\sup_{\theta \in \Theta_0} P_\theta(dy + \mathcal{K})}$$

in which the limiting event $B(y, \epsilon) + \mathcal{K}$ is the observed residual. A ratio such as

$$\frac{\sup_{\theta \in \Theta_1} P_\theta(dy + \mathcal{K}_1)}{\sup_{\theta \in \Theta_0} P_\theta(dy + \mathcal{K}_0)}$$

is not a likelihood ratio unless $\mathcal{K}_0 = \mathcal{K}_1$.

18.4 Computation

18.4.1 Software Options

By default, the function `lmer(...)` estimates the variance components by maximizing the residual log likelihood (18.5). As a follow-up, it reports the weighted least squares estimate (18.2) of the regression coefficients. The square roots of the diagonal components of the inverse Fisher information $(X' \hat{W} X)^{-1}$ serve as standard errors. The optional argument `REML=FALSE` is a cop-out, which overrides the default, and reverts to ordinary maximum likelihood instead. This option produces

a valid likelihood-ratio statistic, which is not one recommended by Welham and Thompson (1997) or by this author. Maximum likelihood with $\mathcal{K} = 0$ is not recommended because the variance estimates have a multiplicative bias of order $O(p/n)$, whose effect is sometimes not negligible.

The function `regress(y~X, ~block+V, kernel=K)` has a three-part syntax, permitting greater flexibility, in which the setting for `kernel` determines the method of estimation. The first part is a standard model-formula for the mean-value subspace \mathcal{X} ; the second part, which may be empty or missing, is a simple model formula for the covariances. Each term in the second part is either a symmetric matrix or a factor; each factor is converted internally into a block matrix by `outer(fac, fac, "==")`, and $\hat{\Sigma}$ is a linear combination of these matrices. The identity matrix is included by default as the first element in the list. The set of matrices must be linearly independent as vectors in \mathbb{R}^{n^2} . For the third part, the default is $\mathcal{K} = \mathcal{X}$, i.e., REML, not $\mathcal{K} = 0$. The log likelihood value reported by `regress(...)$llik` is the maximized log likelihood (18.4), using whatever kernel is specified or implied. The zero-dimensional and one-dimensional options `kernel=0` and `kernel=1` are permitted but not recommended as defaults.

18.4.2 Likelihood-Ratios

Two probability distributions P_0 and P_1 are needed to compute a likelihood ratio. The null distribution goes in the denominator and the alternative in the numerator. The likelihood ratio is the probability ratio $P_1(E)/P_0(E)$ for the same version of the observation, typically as a limit event E in a suitable σ -field. In the present setting P_1 and P_0 are fitted distributions, one fitted under the null model and one under the alternative. It is essential that the event in the numerator be the same as the event in the denominator, i.e., that the same kernel be used in both fits.

For the comparison of mean-values $H_0: \mu \in \mathcal{X}_0$ versus $H_1: \mu \in \mathcal{X}_1 \supset \mathcal{X}_0$ as alternative, residual likelihood may be used in the following manner:

```
X0 <- model.matrix(~mf0);    X1 <- model.matrix(~mf1)
fit0 <- regress(y~mf0, ~block+V, kernel=X0);    # default kernel
fit1 <- regress(y~mf1, ~block+V, kernel=X0);    # default over-ridden
2*(fit1$llik - fit0$llik);
```

Here, `mf0` and `mf1` denote the model formulae for \mathcal{X}_0 and \mathcal{X}_1 respectively. The space of covariance matrices is fixed but arbitrary, and `block+V` is used solely for illustration.

Welham and Thompson (1997) discuss two possibilities for a likelihood-ratio statistic in this setting. The statistic denoted by A in their equation (5) is equivalent to setting the kernel equal to the null subspace, i.e., $\mathcal{K} = \mathcal{X}_0 \subset \mathcal{X}_1$, which is the computation illustrated above. Provided that \mathcal{K} is fixed, $\mathcal{X}_0 \subset \mathcal{X}_1$ and $\mu \in \mathcal{X}_0$, the log likelihood ratio is distributed approximately as χ^2 on $\dim(\mathcal{X}_1 + \mathcal{K}) - \dim(\mathcal{X}_0 + \mathcal{K})$ degrees of freedom, which simplifies for $\mathcal{K} \subseteq \mathcal{X}_0$ to $q = \dim(\mathcal{X}_1) - \dim(\mathcal{X}_0)$.

independent of the kernel. The numerical value of the likelihood-ratio statistic for $\mathcal{K} \subseteq \mathcal{X}_0$ depends on the kernel, but the first-order asymptotic approximation to the null distribution is χ_q^2 , which is independent of \mathcal{K} . The choice $\mathcal{K} = \mathcal{X}_0$ is thought to be optimal in the sense of power, and in the sense of accuracy of the χ^2 distributional approximation.

The option `kernel=0` is permitted but not encouraged; it implies ordinary maximum likelihood, and is equivalent to the `REML=FALSE` option in `lmer()`. The option `kernel = X1` is also allowed; this option produces a valid likelihood ratio statistic that is exactly zero. Why? Because the hypothesis concerns $\mu \in \mathcal{X}_1$, and the residuals modulo \mathcal{X}_1 contain no information about the parameter.

For the comparison of two nested models having the same mean-value subspace, the REML default option is recommended:

```
fit0 <- regress(y~mf0, ~block);
fit1 <- regress(y~mf0, ~block+site);
2*(fit1$llik - fit0$llik);
summary(fit1)
```

Ordinarily, the one-dimensional subspace of constant functions is a subspace of \mathcal{X} , while `block` and `site` are factors having at least two levels. Every factor that occurs in a covariance model is converted internally into a block factor or equivalence matrix

```
Vb <- outer(block, block, "==")      Vs <- outer(site, site, "=="),
```

so the first model specifies a linear combination of $V_1 = I_n$ and $V_2 = \text{block}$ as a block factor, while the second specifies a linear combination of three matrices. Positivity of coefficients is not automatically enforced. Provided that the third coefficient is not restricted to be positive, the asymptotic null distribution is χ_1^2 . To force positivity for the site variance component, the code may be modified as follows:

```
fit0 <- regress(y~mf0, ~block);
fit1 <- regress(y~mf0, ~block+site, pos=c(0,0,1), start=c(fit0$sigma, 1));
2*(fit1$llik - fit0$llik);
summary(fit1)
```

The asymptotic null distribution is a 50% mixture $0.5\delta_0 + 0.5\chi_1^2$.

18.4.3 Testing for Interaction

Consider an experimental design consisting of three physical sites, one northern one southern and one western, separated by a considerable distance that is sufficient to affect the local climate. Each site consists of four blocks of six plots, all of which are outdoors. Each plot is assigned by randomization to one of two treatment levels, which are constant in time. On 127 days over a two-year period, measurements are made on one plant in certain designated plots. By construction, there is one

treatment factor; `site` is regarded as a classification factor; `block` is a factor with 12 levels, while `plot` has 72 levels; both are naturally regarded as block factors. The levels of the remaining factor `day` have a temporal component, which may be important for certain purposes, but the temporal aspect is ignored in the initial discussion, which is concerned with treatment by site interaction. Is the treatment effect constant over sites?

The simplest null model includes additive independent and identically distributed random effects for observations in the same block, additional additive random effects for observations in the same plot, and independent additive effects for observations on the same day. The simplest models with and without interaction are specified implicitly as follows:

```
X0 <- model.matrix(~site+treat);
fit0 <- regress(y~site+treat, ~block+plot+day);
fit1a <- regress(y~site*treat, ~block+plot+day, kernel=X0);
fit1b <- regress(y~site*treat, ~block+plot+day, start=fit1a$sigma),
2*(fit1a$llik - fit0$llik); summary(fit1b)
```

The `treat`×`site` interaction space has dimension two, so the null distribution of the likelihood ratio statistic is χ^2_2 . The parameter estimates reported by `fit1a` and `fit1b` should be very similar, but not identical. If the number of observations is large, it is helpful to supply an initial value for the iteration, as illustrated for `fit1b` above. Note that `fit1b` uses the default kernel `site*treat`, so `fit1b$llik` is not comparable with `fit0$llik`.

It may happen that the response has a temporal component that is continuous in time, as opposed to the process implied by the inclusion of `day` as a block factor above, in which the daily contributions are independent and identically distributed. One simple option is to assume that the temporal component behaves like free Brownian motion, with generalized covariance function proportional to $-|t - t'|$. Free Brownian motion has independent increments on non-overlapping intervals; it is a stationary process in the sense that the distribution of increments are constant in time.

```
dayv <- as.numeric(paste(day)); BM <- -abs(outer(dayv, dayv, "-"))
X0 <- model.matrix(~site+treat);
fit0 <- regress(y~site+treat, ~block+plot+BM);
fit1a <- regress(y~site*treat, ~block+plot+ BM, kernel=X0);
fit1b <- regress(y~site*treat, ~block+plot+BM, start=fit1a$sigma),
2*(fit1a$llik - fit0$llik); summary(fit1b)
```

It is essential in this script that the kernel subspace include the constant functions. The option `kernel=1` is acceptable; `kernel=0` is not acceptable, and may produce an error message.

18.4.4 Singular Models

There are various ways in which singularities may arise in a variance-components model. Consider, for example, a design in which each subject is one experimental unit, and several response measurements of the same type are made on each individual. See project 1 in which up to five measurements are made at different sites on each rat. Then `subject` is a partition of the experimental units, which is a sub-partition of `treatment`. Ordinarily, the covariance model contains `subject` as a block factor with independent and identically distributed effects, and the model for the mean contains the treatment factor. This model is non-singular, and the computation should be straightforward. However, if the exchangeability assumption for subject effects is dropped, and the effects are included additively in the mean model, the subspace spanned by `subject` includes the subspace spanned by `treatment`. As a consequence, treatment effects are not identifiable, i.e., they are confounded with subject effects.

A different sort of singularity arises when a factor such as `block` is included both in the model for the mean and in the model for covariances. While the software may complain about singularities or lack of identifiability, it is feasible to examine this situation analytically. The model is technically identifiable in the sense that distinct parameter values give rise to distinct probability distributions. However, the likelihood function achieves its maximum on the boundary of the space at which `block` has a zero coefficient in the covariance model. In other words, the factor in the mean model trumps the block factor in the covariance model. The model can thus be fitted by dropping the block factor from the covariances.

18.5 Exercises

18.1 Welham and Thompson (1997) discuss two possibilities for a Gaussian likelihood-ratio statistic. For arbitrary mean vector μ , inner product matrix $W = \Sigma^{-1}$, and W -orthogonal projection Q with kernel $\mathcal{K} = \text{span}(K)$, W&T define the residual log likelihood as a function of the parameter $\theta = (\mu, \Sigma)$ by

$$2 \text{RL}(y, \mathcal{K}; \theta) = c(K) + \log |W| - \log |K'WK| - \|Q(y - \mu)\|^2$$

where $c(K) = \text{rank}(Q) \log(2\pi) - \log |K'K|$. Note that Q depends on Σ , and $Q\mu$ need not be zero. For nested subspaces $\mathcal{X}_0 \subset \mathcal{X}_1$, show that version A in their equation (5) is the difference

$$2 \text{RL}(y, \mathcal{X}_0; \hat{\theta}_1) - 2 \text{RL}(y; \mathcal{X}_0; \hat{\theta}_0),$$

where RL attains its maximum at $\hat{\theta}_0$ under the null, and at $\hat{\theta}_1$ under the alternative.

Discuss whether version D in their equation (6) is or is not a likelihood ratio. If it is a likelihood ratio, or a non-random multiple of a likelihood ratio, what is the event for which the probability ratio is computed?

18.2 Consider a balanced block design having m blocks each consisting of b observational units, and let B be the associated block factor as a Boolean matrix of order $n = mb$. The three-parameter Gaussian model with moments

$$\mu \in \mathbf{1}_n, \quad \Sigma = \sigma^2(I_n + \theta B),$$

is parameterized by two scalars $\mu, \sigma > 0$ and one additional parameter. For the purposes of this exercise $\theta > -1/b$ is not necessarily positive, but Σ is positive definite. In addition, the residual refers to any linear transformation, such as $Y_{ij} - \bar{Y}_{..}$, whose kernel is $\mathbf{1} \subset \mathbb{R}^n$.

Let Y_{ij} be the observation for unit j in block i . Show that the within- and between- quadratic forms

$$\text{SS}_W = \sum_{ij} (Y_{ij} - \bar{Y}_{i.})^2, \quad \text{SS}_B = b \sum_i (\bar{Y}_{i.} - \bar{Y}_{..})^2,$$

are independent with distributions $\sigma^2 \chi_{n-m}^2$ and $\sigma^2(1+b\theta) \chi_{m-1}^2$ respectively. Hence deduce that, if $\theta = 0$, the mean-square ratio

$$F = \frac{\text{SS}_B / (m-1)}{\text{SS}_W / (n-m)}$$

is distributed according to Fisher's $F_{m-1, n-m}$ distribution.

18.3 For the balanced block design in the preceding exercise, show that the implied distribution for residuals is a two-parameter full exponential-family model with canonical sufficient statistic SS_W, SS_B . Hence deduce that the residual maximum-likelihood estimate satisfies

$$\hat{\sigma}^2 = \text{SS}_W / (n-m), \quad 1 + b\hat{\theta} = F.$$

Show also that the sub-model with $\theta = 0$ is a one-parameter exponential family model with sufficient statistic $\text{SS}_B + \text{SS}_W$.

18.4 For the balanced block design, show that the log determinant is

$$\log \det(\Sigma) = n \log(\sigma^2) + m \log(1 + b\theta).$$

Show that the ML estimate satisfies $1 + b\hat{\theta} = (m-1)F/m$. Hence deduce that the ordinary log likelihood ratio statistic for testing $\theta = 0$ is

$$\log \det \hat{\Sigma}_0 - \log \det \hat{\Sigma}_1 = n \log \left(\frac{n-m+(m-1)F}{n} \right) - m \log \left(\frac{(m-1)F}{m} \right),$$

while the REML statistic is

$$(n - 1) \log \left(1 + \frac{(m - 1)(F - 1)}{n - 1} \right) - (m - 1) \log F.$$

What does this expression tell you about the null distribution of the REML likelihood-ratio statistic?

18.5 Show that the REML estimate with positivity constraint satisfies $1 + b\hat{\theta} = \max(F, 1)$. What is the REML estimate for the second component? Express the constrained REML likelihood-ratio statistic as a function of F , and compute the atom at the origin.

18.6 The following exercise is concerned with the distribution of the likelihood-ratio statistic in a ‘fixed-effects’ model for a balanced design, where $\Sigma = \sigma^2 I_n$, and either $\mu \in \mathbf{1}_n$ under the null hypothesis or $\mu \in \text{span}(B)$ under the alternative. The meaning of the term ‘residual’ is unchanged, and the F -statistic in Exercise 18.2 is also unchanged.

Show that the residual log likelihood-ratio statistic for testing $\mu \in \mathbf{1}$ versus $\mu \in \text{span}(B)$ is

$$(n - 1) \log \left(1 + \frac{(m - 1)F}{n - m} \right).$$

By simulation or otherwise, show also that the null expected value exceeds that of χ^2_{m-1} by the approximate multiplicative factor

$$1 + \frac{1}{2}(m + 1)/(n - m).$$

This is a particular instance of the Bartlett correction factor.

18.7 The null hypothesis being tested in Exercise 18.5 is the same as that in Exercise 18.3, but the alternatives are different: one implies exchangeability of block effects, the other does not. Discuss the implications of the fact that one statistic is strictly increasing as a function of F , whereas the other is strictly decreasing for $F < 1$ and strictly increasing for $F > 1$.

18.8 Positive definiteness of a function $\mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ means that, for integer $n \geq 1$ and each finite collection of points $\mathbf{x} = \{x_1, \dots, x_n\}$, the $n \times n$ matrix $K[\mathbf{x}, \mathbf{x}]$ with components

$$K_{ij} = K(x_i, x_j)$$

is positive definite and symmetric.

Let x be a point in real Euclidean space \mathbb{R}^d . For each $\lambda > 0$, the function

$$K(x, x') = e^{-\lambda \|x-x'\|}$$

is the covariance function for a stationary autoregressive process of order one if $d = 1$, also called the Ornstein-Uhlenbeck process for general d . Show that K is positive definite.

18.9 To test for equality of variances in k blocks of sizes n_1, \dots, n_k , the REML procedure goes as follows. First, blk is the k -dimensional subspace of \mathbb{R}^n spanned by the block indicators b_1, \dots, b_k . Second, I_r is the restriction of the identity matrix to block r , i.e., $I_r(i, j) = \delta_{ij} b_r(i)$. The likelihood-ratio statistic is computed by fitting null and alternative models as follows:

```
fit0 <- regress(Y~blk)
fit1 <- regress(Y~blk, ~I1+...+Ik, identity=FALSE, start=rep(1,k))
```

Show that the REML likelihood-ratio statistic $2\text{fit1\$llik} - 2\text{fit0\$llik}$ is equal to the gamma deviance statistic

$$(n. - k) \log(s_{\text{pool}}^2) - \sum (n_r - 1) \log(s_r^2),$$

where s_r^2 is the sample variance in block r , and s_{pool}^2 is the pooled variance. Bartlett's (1937) test for equality of variances is the REML statistic divided by the Bartlett correction factor

$$1 + \frac{1}{3(k-1)} \left(\sum_{r=1}^k \frac{1}{n_r - 1} - \frac{1}{n - k} \right).$$

What is the null distribution of the ratio?

Chapter 19

Response Transformation



19.1 Likelihood for Gaussian Models

In applied work, it is frequently advantageous to transform the observations prior to fitting a linear Gaussian model. Invariably, this means that the state space for each observation $Y_i \in S$ is an open real interval such as $S = (0, \infty)$ or $S = (0, 1)$ or $S = \mathbb{R}$. A transformation $g : S \rightarrow \mathbb{R}$ is identified and applied component-wise to the vector $Y \in \mathcal{S}^n$ in the hope that the transformed variable gY might be approximately normally distributed with mean $\mu \in \mathcal{X}$, and covariance Σ belonging to some family of covariance matrices such as those described in Chaps. 1–5. According to this scenario, the joint density of the observation Y at the point $y \in \mathcal{S}^n$ is

$$(2\pi)^{-n/2} |\Sigma|^{-1/2} e^{-(gy-\mu)' \Sigma^{-1} (gy-\mu)/2} \prod_{i=1}^n |g'(y_i)|$$

provided that g is 1–1 differentiable with a differentiable inverse. To specify the likelihood function, it is necessary to identify the set of transformations $g \in \mathcal{G}$ under consideration, plus the mean-value space $\mathcal{X} = \text{span}(X)$ and the space Θ of covariance matrices. To be clear, these moment spaces are moment spaces for the transformed variable gY , not for Y . As a function on $\mathcal{G} \times \mathcal{X} \times \Theta$, this density is the likelihood function.

It is helpful at this stage to insert two additional technical conditions. First, the space **1** of constant n -vectors is a subspace of \mathcal{X} ; this is not required in the theory of linear models, but it is nearly universal in applied work and it is needed at certain points in the argument that follows. Second, the space of covariance matrices is a cone, i.e., $\Sigma \in \Theta$ implies $\tau \Sigma \in \Theta$ for every scalar multiple $\tau > 0$. Both conditions are mathematically essential but relatively benign; the cone need not be convex. The

cone condition extends to Σ^{-1} and ensures that the maximum-likelihood estimate $\hat{\mu}_g$, $\hat{\Sigma}_g$ for fixed g satisfies

$$(gy - \hat{\mu}_g)' \hat{\Sigma}_g^{-1} (gy - \hat{\mu}_g) = n.$$

As a consequence, the profile log likelihood for the transformation $g \in \mathcal{G}$ is

$$l_p(g; y) = -\frac{1}{2} \log \det(\hat{\Sigma}_g) + \sum_{i=1}^n \log |g'(y_i)|. \quad (19.1)$$

Finally, for all scalars $a, b \neq 0$, the cone condition and $\mathbf{1} \subset \mathcal{X}$ imply $l_p(a+bg; y) = l_p(g; y)$, so that the profile likelihood is invariant with respect to affine composition. In other words, the transformations $y \mapsto g(y)$ and $y \mapsto a + bg(y)$ are equivalent for this comparison: $gY \sim N(\mu, \Sigma)$ implies $a + bgY \sim N(a + b\mu, b^2\Sigma)$, and vice-versa.

The preceding analysis assumes that the maximum-likelihood estimate $\hat{\mu}_g$, $\hat{\Sigma}_g$ exists. Existence and uniqueness cannot be guaranteed in general, but failure is rare in practice provided that $p < n$ and the residual space is adequate to estimate all variance components.

19.2 Box-Cox Transformation

19.2.1 Power Transformation

One very natural option is to choose a simple parametric family such as the family of power transformations (Box and Cox, 1964). Provided that $\mathbf{1} \subset \mathcal{X}$ and all observations are strictly positive, the transformation $(0, \infty) \rightarrow \mathbb{R}$ may be taken in the form $y \mapsto (y^\lambda - 1)/\lambda$ for some scalar λ , with the limit $\lambda \rightarrow 0$ corresponding to the log function. The derivative $y^{\lambda-1}$ is strictly positive, so, by (19.1), the profile log likelihood for λ is

$$l_p(\lambda; y) = -\frac{1}{2} \log \det(\hat{\Sigma}_\lambda) + (\lambda - 1) \sum_{i=1}^n \log y_i \quad (19.2)$$

provided that the maximum-likelihood estimate $\hat{\Sigma}_\lambda$ exists. It is a straightforward exercise to plot $l_p(\lambda)$ against λ to check whether there is a clear maximum in the range of interest, which is typically $-1 \leq \lambda \leq 1$. A large value of the likelihood-ratio statistic $2l_p(\hat{\lambda}) - 2l_p(1)$ indicates a need for transformation.

The profile log likelihood is meant to be used only as a rough guide. In practice, the only transformations that are ordinarily considered for linear statistical analysis are (i) the logarithm if the response scale is strictly positive with a well-

defined origin, and effects are expected to be multiplicative; (ii) the identity if treatment effects are expected to be additive on the given scale; (iii) occasionally the reciprocal, square root or cube root if there is a reasonable justification based on the physical units of measurement. For example, if the observation is a volume, an argument might be made for the cube-root; if the observation is a time or duration, conversion by reciprocals to the rate scale or frequency scale might make sense. But additivity of effects on such scales is usually dubious, so the log transformation is the preferred choice for most physical variables such as mass, volume, length, time, or ratios such as speed, density, miles per gallon, and so on. Under no circumstances should the reported analysis be done on the scale $Y^{\hat{\lambda}}$, where $\hat{\lambda}$ is the maximum-likelihood estimate from (19.2).

19.2.2 Re-scaled Power Transformation

Let $\tau > 0$ be a fixed constant, and let $y \mapsto \tau(y^\lambda - 1)/\lambda$ be the re-scaled power transformation applied component-wise on the transformed scale. The Jacobian is $\tau^n \prod y_i^{\lambda-1}$, so the log Jacobian is

$$\log J = n(\lambda - 1) \log \dot{y} + n \log \tau,$$

where \dot{y} is the geometric mean of the observations. As a numerical device, it is sometimes helpful to set the scale parameter to $\tau = \dot{y}^{1-\lambda}$ so that the Jacobian is reduced to one. With this choice, the profile likelihood reduces to the determinantal term in (19.2).

The preceding discussion refers to re-scaling $g \mapsto \tau g$ by composition on the left, i.e., by multiplication after power transformation. Composition on the right $g \mapsto g\tau$ refers to the effect of re-scaling $y \mapsto \tau y$ before power transformation:

$$\begin{aligned} \text{left : } & y \xrightarrow{g} g(y) \xrightarrow{\tau} \tau g(y) \\ \text{right : } & y \xrightarrow{\tau} \tau y \xrightarrow{g} g(\tau y). \end{aligned}$$

Composition on the right sends $g(\cdot)$ to $g(\tau \cdot)$, which is an affine transformation of $g(\cdot)$:

$$g(\tau y) = \tau^\lambda g(y) + \frac{\tau^\lambda - 1}{\lambda} = \tau^\lambda g(y) + \text{const.}$$

The assumption $\mathbf{1} \subset \mathcal{X}$, and the cone condition on covariance matrices, are sufficient to ensure that likelihood-based conclusions are unaffected by scalar composition on the right. Invariance is absolutely essential in applied work, where the choice of physical units—_inches versus centimetres or minutes versus seconds—is quite arbitrary.

The purpose of re-scaling on the left is not to modify the power transformation in a substantive way, but to simplify the computation. Nonetheless, the argument in the first paragraph could easily be misconstrued as a statement that the modified power transformation

$$y_i \mapsto \frac{y_i^\lambda}{\lambda \dot{y}^{\lambda-1}} \quad \text{or} \quad y_i \mapsto \frac{y_i^\lambda - 1}{\lambda \dot{y}^{\lambda-1}}$$

has Jacobian equal to one. As they are written above, these transformations do not act component-wise. The first transformation satisfies $g(\tau y) = \tau g(y)$, but the Jacobian $J = |\lambda|^{-1}$ is discontinuous at $\lambda = 0$. For the second transformation, the Jacobian

$$J = \frac{1}{\lambda} + \frac{\lambda - 1}{n\lambda} \sum y_i^{-\lambda}$$

is continuous at $\lambda = 0$, but there do not exist constants a, b such that $g(\tau y) = a + bg(y)$. A modified power transformation satisfying both conditions— $g(\tau y) = \tau g(y)$ and continuity in λ —is described in Exercise 19.7. None of these modifications has a parameter-independent Jacobian, so the Jacobian cannot be ignored in likelihood calculations.

19.2.3 Worked Example

In the analysis of the woodcutting efficiencies of three brands of saws in Chap. 2, the response was the time taken to complete a designated task. On the grounds that multiplicative effects were more plausible than additive effects, the log transformation was used in all analyses. We now provide an analysis justifying that choice.

Bearing in mind that the chief purpose of transformation is not so much to induce normality, but to achieve additivity of effects, two additive Gaussian models were selected as targets. In the first version, the mean of the transformed variable is additive in the four factors *species+bark+team+saw.id*, while the variances are constant, and the covariances are zero. This is a rank-14 sub-model of the standard Latin-square model, which has rank 16. The transformation model has two additional parameters, σ^2 and λ , making 16 total. In the second version, the mean is additive in the three factors *species+bark+saw.brand*, which is a subspace of dimension 6, while there are two additional variance components *team+saw.id*, making a total of ten parameters. Both profile log likelihoods for λ in Fig. 19.1 have their maxima near $\hat{\lambda} = -0.34$; both 95% confidence intervals include $\lambda = 0$, but the identity is excluded. The conclusion is that the effects on the time scale are approximately multiplicative, so taking logs is the natural remedy, as indicated by Bliss (1970, pp. 440–441). Most experienced statisticians would transform

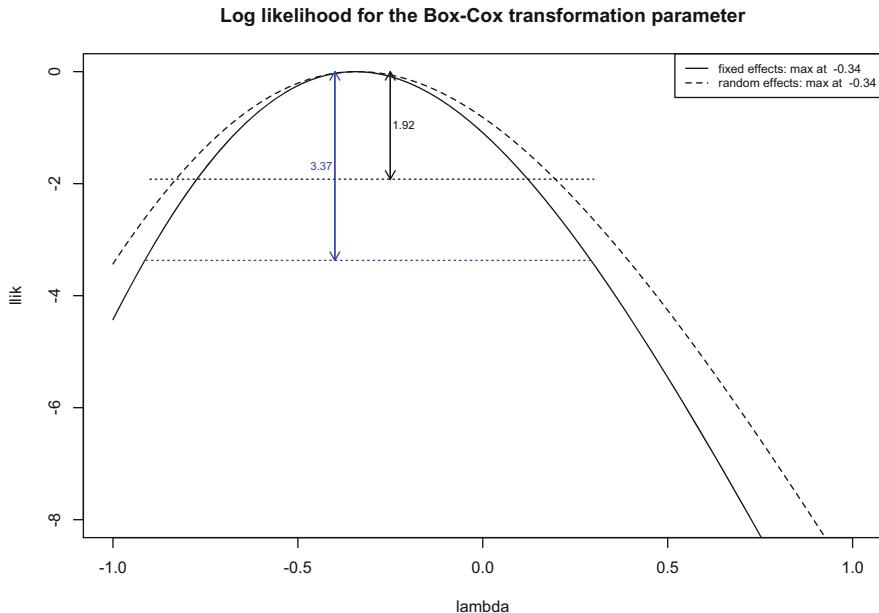


Fig. 19.1 Log likelihood for the transformation parameter λ for two linear models

instinctively to the log scale on the grounds that additive effects on the time scale are less plausible than multiplicative effects.

As it happens, the variation between duplicate saws is small, but brand three is about 15% more efficient than the others. Mean cutting times are in the ratios 1.28:1.00:0.80 for larch:spruce:pine, with an additional factor of 1.14 for bark. There is substantial variation among the teams.

A crucial point in the computation of log likelihoods for Gaussian transformation models is that REML, or residual log likelihood, must not be used under any circumstances. REML calculations are based on the distribution of residuals in \mathbb{R}^{n-p} , whereas the Jacobian is the determinant of a transformation $\mathbb{R}^n \rightarrow \mathbb{R}^n$. These are not compatible because the power transformation does not act on residuals. As a function of λ , the residual likelihood criterion is not a likelihood in the conventional sense of a density ratio in \mathbb{R}^n . If the REML criterion were used with the Jacobian as in (19.2), the plots shown in Fig. 19.1 would look substantially different. Details are discussed in the next section.

A $1 - \alpha$ confidence interval for the transformation parameter may be obtained by the likelihood-ratio formula

$$\{\lambda : l(\hat{\lambda}; y) - l(\lambda; y) \geq \frac{1}{2}\chi_{1,\alpha}^2\},$$

which is based on large-sample distribution theory. For $\alpha = 0.05$, the cutoff allowance $1.92 = 1.96^2/2$ is indicated in Fig. 19.1. In the present setting, the

effective sample size is the residual degrees of freedom, which is $36 - 14 = 22$. Given that we are interested only in whether the interval includes zero, the asymptotic approximation is reasonably adequate. However, the coverage level can be improved appreciably by replacing the $\frac{1}{2}\chi_1^2$ threshold with the threshold based on Fisher's F -ratio,

$$\frac{n}{2} \log\left(1 + \frac{F_{1,n-p-1,\alpha}}{n-p-1}\right),$$

which is the exact threshold for $(1 - \alpha)$ -coverage in the setting of nested linear models. The 95% F -threshold for $n = 36$ and $p = 14$ is 3.37, which is also shown in Fig. 19.1 for comparison. The greater allowance produces a wider interval, but does not materially alter the conclusion or subsequent analysis.

An analysis-of-variance decomposition on the log scale shows that the interactions *species.bark* and *bark.brand* are negligible. Bark removal reduces the *mean* log cutting time by an estimated 0.152 ± 0.031 units for each species and each brand, so the cutting-time distribution is reduced multiplicatively by about 14%.

19.2.4 Transformation and Residual Likelihood

If the transformation under consideration can be regarded as an invertible transformation on residuals, the residual likelihood modulo the subspace \mathcal{X} may be used in place of (19.2). A transformation $g: \mathbb{R}^n \rightarrow \mathbb{R}^n$ may be regarded as a transformation on residuals if and only if each coset $y + \mathcal{X}$ has an image that is either a coset or a subset of a coset. In that case, g induces a transformation $\mathbb{R}^n/\mathcal{X} \rightarrow \mathbb{R}^n/\mathcal{X}$ on residuals, which is assumed to be measurable with respect to the Borel subsets of \mathbb{R}^n/\mathcal{X} . For example, a linear transformation $g: \mathbb{R}^n \rightarrow \mathbb{R}^n$ induces a transformation on residuals if and only if $g\mathcal{X} \subset \mathcal{X}$. Unfortunately, a component-wise non-linear transformation does not induce a measurable transformation on residuals except for trivial cases such as $\mathcal{X} = 0$ and $\mathcal{X} = \mathbb{R}^n$. Even $\mathcal{X} = \mathbf{1}$ fails. Thus, residual likelihood is not available as an option for the comparison of response transformations.

To see why and how a naive version of REML fails, it suffices compare the standard log likelihood with a REML-style criterion $l^\dagger(\cdot)$ first proposed by Shi and Tsai (2002) and subsequently by Gurka et al. (2006). For the power-transformation model $gY \sim N(X\beta, \Sigma)$, the two criteria are

$$l = -\frac{1}{2}(gy - \mu)' \Sigma^{-1} (gy - \mu) - \frac{1}{2} \log \det \Sigma + (\lambda - 1) \sum \log(y_i),$$

$$l^\dagger = l(\mu, \Sigma, \lambda) - \frac{1}{2} \log \det(X'WX) + \frac{1}{2} \log \det(X'X),$$

where $W = \Sigma^{-1}$. Here, $gy = y^\lambda/\lambda$ for $y > 0$ is the component-wise power transformation, so that the log Jacobian is $(\lambda - 1) \sum \log(y_i)$. In either case,

maximization over the space of mean-vectors $\mu \in \mathcal{X}$ gives $\hat{\mu} = P(gy)$, the W -orthogonal projection of the transformed vector. The profile criteria are

$$l(\Sigma, \lambda; y) = -\frac{1}{2} gy' W Q gy - \frac{1}{2} \log \det \Sigma + (\lambda - 1) \sum \log(y_i),$$

$$l^\dagger(\Sigma, \lambda; y) = l(\Sigma, \lambda; y) - \frac{1}{2} \log \det(X' WX) + \log \det(X' X).$$

Suppose that two statisticians are asked to examine the same data, which is concerned with vehicle fuel economy. Statistician I analyzes the consumption rates in miles per gallon, and statistician II in kilometres per litre, so the pairs of numbers differ by a constant multiple: $y_i^{(1)} = \tau y_i^{(2)}$ with $\tau \simeq 2.8$. For each λ , the transformed values differ by a parameter-dependent factor τ^λ , the associated variance matrices satisfy $\Sigma^{(1)} = \tau^{2\lambda} \Sigma^{(2)}$, and the inverse matrices satisfy $W^{(1)} = \tau^{-2\lambda} W^{(2)}$. As we should expect, the log likelihood function is scale-invariant in the sense that, the two versions differ by an additive constant

$$l(\tau^{2\lambda} \Sigma, \lambda; \tau y) = l(\Sigma, \lambda; y) - n \log(\tau).$$

Invariance with respect to scalar multiplication means that two statisticians analyzing the same data on different scales must arrive at the same conclusion regarding the transformation parameter. By contrast, the two versions of the modified criterion differ linearly in λ :

$$l^\dagger(\tau^{2\lambda} \Sigma, \lambda; \tau y) = l^\dagger(\Sigma, \lambda; y) - n \log(\tau) + \lambda p \log(\tau),$$

where $p = \dim(\mathcal{X})$. Lack of invariance means that two statisticians using l^\dagger as the selection criterion are liable to arrive at contradictory conclusions for λ . For the continuity-modified transformation $gy = (y^\lambda - 1)/\lambda$, the analysis is slightly more complicated, but the conclusions are essentially the same provided that $\mathbf{1} \subset \mathcal{X}$.

The extreme example $X = I_n$ and $\mathcal{X} = \mathbb{R}^n$ is of little practical interest because $Q = 0$ implies that the residual is identically zero. But it suffices to show that l^\dagger is a non-trivial function of the observations, and thus not a function of residuals. With $X = I_n$ the three determinantal terms vanish, leaving

$$l^\dagger(\Sigma, \lambda; y) = (\lambda - 1) \sum \log y_i.$$

In the absence of information from residuals, constancy in Σ is correct, but linearity in λ is misleading. The slope is positive if the geometric mean observation is greater than one, and negative otherwise, so scale conversion can change the slope from positive to negative or vice-versa.

19.3 Quantile-Matching Transformation

In certain ‘big-data’ settings such as the analysis of micro-array gene-expression data, transformation to a marginal reference distribution is sometimes recommended as a way to reduce the impact of incidental or unwanted structural effects. Quantile matching is a strictly monotone transformation $h = G^{-1} \circ F$, which is defined by a domain distribution F and a target distribution G . Both cumulative distribution functions are assumed to be strictly monotone and differentiable. Quantile matching is applied component-wise to the data, and transforms $Y \sim F$ into $hY \sim G$. The empirical version, denoted by \tilde{h} , transforms the finite set $\{y_1, \dots, y_n\}$ into specific quantiles of G :

$$\begin{aligned} h : y &\longmapsto F(y) \longmapsto G^{-1}(F(y)) \\ \tilde{h} : y &\longmapsto \tilde{F}_n(y) \longmapsto G^{-1}(\tilde{F}_n(y)). \end{aligned}$$

For the specific requirements of this section, \tilde{F} is a strictly monotone continuously differentiable function satisfying $0 < \tilde{F}(t) < 1$. At each observation point $y \in \{y_1, \dots, y_n\}$, the value is the average of the left and right limits of the empirical distribution function

$$\tilde{F}_n(y) = \frac{1}{2}\hat{F}_n(y^-) + \frac{1}{2}\hat{F}_n(y^+).$$

Elsewhere in the domain, $\tilde{F}_n(t)$ is subject to differentiability and strict monotonicity, but, apart from the sample points, the values are otherwise unspecified. The numbers $\tilde{F}(y_1), \dots, \tilde{F}(y_n)$ are the uniform sample quantiles in $(0, 1)$, and the target G -quantiles are the transformed values

$$q_{i:n} = \tilde{h}(y_i) = G^{-1}(\tilde{F}(y_i)),$$

taken with multiplicity in ascending order. If there are no ties, the uniform quantiles are the numbers $(2i - 1)/2n$ for $1 \leq i \leq n$.

The Jacobian of the transformation $h : \mathbb{R}^n \rightarrow \mathbb{R}^n$ is the product of the derivatives at the domain points

$$\prod_{i=1}^n h'(y_i) = \prod_{i=1}^n \frac{F'(y_i)}{g(h(y_i))},$$

so the log Jacobian and its empirical version are

$$\sum \log F'(y_i) - \sum \log g(h(y_i)) \quad \text{and} \quad \sum \log \tilde{F}'(y_i) - \sum \log g(q_{i:n}),$$

where $g = G'$ is the target density. The last term is a quadrature sum, which is an approximation to the entropy integral

$$\tilde{J}(G) = n^{-1} \sum \log g(q_{i:n}) = \int \log g(x) dG(x) + O(n^{-1}).$$

From (19.1), the profile log likelihood function for the quantile-matching transformation \tilde{h} is

$$-\frac{1}{2} \log \det \hat{\Sigma}_h + \sum_i \log \tilde{F}'(y_i) - \sum \log g(q_{i:n}), \quad (19.3)$$

where $\hat{\Sigma}_h$ is the maximum-likelihood estimate after transformation. However, the derivatives $\tilde{F}'(y)$ are not readily available, so the log likelihood (19.3) is not computable.

Now consider two quantile-matching transformations, which are defined by their target distributions G_0, G_1 . From (19.3), the profile log likelihood ratio of G_1 to G_0 , is

$$-\frac{1}{2} \log \det(\hat{\Sigma}_1 \hat{\Sigma}_0^{-1}) - n \tilde{J}(G_1) + n \tilde{J}(G_0), \quad (19.4)$$

(McCullagh and Tresoldi, 2021). If n is sufficiently large, the quadrature sums $\tilde{J}(G_0)$ and $\tilde{J}(G_1)$ can be replaced with the corresponding integrals. The quadrature errors are typically $O(n^{-1})$ for both $J(G_0)$ and $J(G_1)$, but if the distributions are contiguous or similar, the quadrature error for the difference is $o(n^{-1})$, and thus negligible for present purposes.

19.4 Exercises

19.1 Let Y be a non-negative random variable with cumulants κ_r such that $\kappa_r/\kappa_1^r = O(\rho^{r-1})$ as $\rho \rightarrow 0$. In other words, the scale-free variable $Z = Y/\kappa_1$ has variance $\rho = \kappa_2/\kappa_1^2$, which is the squared coefficient of variation of Y , and the higher-order scale-free cumulants are $O(\rho^{r-1})$. Show that the cumulants of the power-transformed variable are

$$\begin{aligned} E(Z^\lambda) &= 1 + \frac{(\lambda - 1)\kappa_2}{2\kappa_1^2} + o(\rho); \\ \text{var}(Z^\lambda) &= \frac{\kappa_2}{\kappa_1^2} + o(\rho); \\ \text{cum}_3(Z^\lambda) &= \frac{\kappa_3}{\kappa_1^3} + 3(\lambda - 1) \frac{\kappa_2^2}{\kappa_1^2} + o(\rho^2). \end{aligned}$$

Hence deduce that the approximate symmetry-inducing power transformation is $\hat{\lambda} = 1 - \kappa_1 \kappa_3 / (3\kappa_2^2)$.

19.2 Wilson-Hilferty transformation: Show that the r th cumulant of the exponential distribution is $\kappa_r = \kappa_1(r-1)!$, and hence that $Y^{1/3}$ is approximately symmetrically distributed.

19.3 Show that the r th cumulant of the Poisson distribution is $\kappa_r = \kappa_1$, and hence that $Y^{2/3}$ is approximately symmetrically distributed.

19.4 Show that the transformation $\mathbb{R}^n \rightarrow \mathbb{R}^n$ defined by

$$\bar{u} \mapsto \bar{u} + \text{const}, \quad u_i - \bar{u} \mapsto \lambda(u_i - \bar{u})$$

is linear and invertible with Jacobian $J = |\lambda|^{n-1}$. Here, \bar{u} is the mean of the components of the vector $u \in \mathbb{R}^n$, and λ is a non-zero constant.

19.5 Consider the non-component-wise transformation

$$y_i \mapsto \frac{y_i^\lambda}{\lambda \dot{y}^{\lambda-1}}$$

where \dot{y} is the geometric mean of the components of $y \in \mathbb{R}_+^n$. Using the result of the previous exercise, show that the modified transformation is invertible $\mathbb{R}_+^n \rightarrow \mathbb{R}_+^n$ with Jacobian $J = |\lambda|^{-1}$.

19.6 As a function of λ , show that the transformation $\mathbb{R}_+^n \rightarrow \mathbb{R}^n$

$$(gy)_i = \frac{y_i^\lambda - 1}{\lambda \dot{y}^{\lambda-1}}$$

is continuous at $\lambda = 0$, and that the Jacobian is the absolute value of

$$\frac{1}{\lambda} + \frac{\lambda - 1}{n\lambda} \sum y_i^{-\lambda}.$$

Find the limits for $\lambda = 0, \pm 1$. Discuss the implications regarding invertibility? For $\tau > 0$, show that $g(\tau y)$ is not expressible in the form $a + bg(y)$ for any constants a, b depending on τ, λ .

19.7 For $\tau > 0$, show that the modified transformation $\mathbb{R}_+^n \rightarrow \mathbb{R}^n$

$$(gy)_i = \dot{y} + \frac{y_i^\lambda - \dot{y}^\lambda}{\lambda \dot{y}^{\lambda-1}}$$

is continuous at $\lambda = 0$ and satisfies $g(\tau y) = \tau g(y)$. What are the implications for statistical applications? Show that the Jacobian is

$$\frac{1}{\lambda} + \frac{\lambda - 1}{n\lambda} \dot{y}^\lambda \sum y_i^{-\lambda},$$

which is positive, and that the limits for $\lambda \rightarrow 0$ and $\lambda \rightarrow 1$ are equal.

19.8 For which values of α is the transformation

$$g_\alpha(x) = \frac{(-\log(1-x))^\alpha}{\alpha} - \frac{(-\log(x))^\alpha}{\alpha}$$

differentiable and strictly monotone $(0, 1) \rightarrow \mathbb{R}$?

19.9 Repeat the preceding exercise taking $\mathcal{X} = \mathbf{1}$, and Σ a linear combination of the block matrices I_n , `row` and `col`.

Chapter 20

Presentations and Reports



20.1 Coaching Tips I

The first nine of these tips are concerned with technical aspects of statistical reports. The last six are concerned with English usage, style and semantics.

1. Length: Reports should be no longer than necessary. A short report that makes the salient points is preferable to a long rambling philosophical essay, even if the longer essay makes the same points somewhere along the way. Above all, have compassion for the reader (and grader).
2. Graphs and plots: A plot either of the raw data or of the residuals is almost always essential at some point, if only to explain the motivation for the analysis. Some plots provide insight, some do not; only the most useful need to be shown. Although you should indicate what plots were made, it is generally not necessary to include in the report a copy of all plots made and all analyses performed. If necessary for examination purposes, extra plots and lengthy analyses can be included in an appendix.
3. Executive summary: All major conclusions should be stated at the beginning in a summary intended for a scientifically literate reader who is not a statistician. Technical terms associated with the context of the problem are unavoidable, but technical statistical terms should so far as possible be avoided. One page is the upper limit. Remember, few readers progress beyond the summary. It is up to the author to state the conclusions early in as persuasive a manner as possible if the reader is to be convinced.
4. Statistical analyses: Following the summary, the report should describe the models fitted, the tests performed, and how these support the conclusions. The relevance of the models to the context under study is important. Technical statistical terms are acceptable here only if they are essential to support the conclusions.

5. Model specification: Statistical models are used by statisticians, computer scientists, engineers, quantitative sociologists, biostatisticians and epidemiologists, and even by research workers in literature, law and the humanities. The term does not necessarily mean the same thing to all users. Some users think that a statistical model is an equation beginning with $y =$ and ending in $\dots + \epsilon$. Others are of the opinion that a statistical model is a syntactical expression such as $\sim A+x+\dots$ containing the symbol \sim , or more generally any machine-learning algorithm that is coded in R or Python. A professional statistician knows that a statistical model is a non-empty set whose elements are probability distributions on the observation space, usually \mathbb{R}^n .

A stochastic *model specification* calls for a statement indicating which distributions are included in the set and which are excluded. Parameter estimation calls for an estimation method, usually maximum likelihood, but very frequently in a modified form such as REML. Maximum likelihood calls for an algorithm, preferably one that is efficient and coded in readily accessible software. Software syntax is important, but an estimation method is not an algorithm, and an algorithm is not a model specification.

A stochastic model may be specified *indirectly* by offering a description of how a random draw $Y \sim P_\theta$ may be generated from an arbitrary distribution P_θ in the model. GLMs are usually specified in this manner by a three-step procedure:

$$\eta = X\theta; \quad \pi_i = e^{\eta_i}/(1 + e^{\eta_i}); \quad Y_i \sim \text{Ber}(\pi_i); \quad (\text{indep.})$$

A great many Gaussian models may be generated by adding several independent Gaussian processes, each associated with a different factor or interaction. For example

$$Y_{it} = \alpha_i + \eta_0(t) + \eta_i(t) + \epsilon_{it}$$

as a sum of four independent zero-mean processes, can be matched up with the direct specification of the covariance function

$$\text{cov}(Y_{it}, Y_{i',t'}) = \delta_{i,i'} + K_0(t, t') + K_1(t, t')\delta_{i,i'} + \delta_{i,i'}\delta_{t,t'}$$

provided that α has independent and identically distributed standard normal components, and the other three components are distributed as indicated.

6. Numerical precision: Adequate numerical precision is important, but ordinarily two significant digits are sufficient for standard errors. Parameter estimates should always be given with standard errors, or standard errors of differences in the case of factor levels. It is often sufficient to say that the standard error is 9–15%, or the standard error of pairwise differences is 0.35–0.45, if the range is not excessive. Parameter estimates should be accurate to 10% of a standard error. By convention, p -values are given as a percentage: rarely is there a need for more than two significant digits. The listing of excessively

many uninformative digits in estimates and standard errors betrays a lack of statistical sense, and will be penalized.

7. Computer output: While it is necessary for students to demonstrate mastery of the computer system or statistical package, tailoring the computer output to the problem at hand is always necessary if only to demonstrate that you are the computer driver not the computer slave. (i) From the computer-generated analysis-of variance table, list only the parts that are relevant to your analysis. It is your job as statistician and expert to judge what is relevant and what is not. (ii) If the estimated coefficients in a logistic regression model with several covariates are listed in tabular form with standard errors, and the coefficient of smok is 0.3365, do not repeat this number in the text, but seize the opportunity to explain in concrete terms that the estimated odds of the event (cancer) are 40% higher for smokers than for non-smokers. (iii) If the model matrix is non-standard and cannot be generated by a standard model formula, as for example the additive skew-symmetric formula $E(Y_{ij}) = \alpha_i - \alpha_j$, you need to explain what the structure of the matrix is. (iv) Do not quote a *p*-value without stating the hypothesis under test and how the value supports the stated conclusion. (v) Every parameter has a physical interpretation: do not pass up the opportunity to remind the reader what the physical interpretation of the logistic regression coefficient $\hat{\beta} = -0.684$ is in the context of the problem. (The risk of disease in the treatment group is one half that in the control group.)
 8. Physical units: Physical variables, unlike mathematical variables, always have units such as ‘length in mm,’ ‘temperature in °K,’ ‘mm Hg,’ ‘age in months,’ or ‘depth in fathoms.’ If you lose sight of the units your conclusions are liable to be ridiculous.
 9. Mathematical terminology: A set consists of points or elements, all of which are distinct; a multi-set is a set in which each element has a multiplicity, which is a positive integer. A list is a sequence of arbitrary objects such as factor levels, which are usually not distinct; a vector is a list of real or complex numbers, usually called components. An array is a doubly-indexed list of arbitrary objects; a matrix is an array of real or complex numbers. Thus, the set of eigenvalues of a square matrix is a multi-set, not a vector; the eigenvalues of a block factor are the block sizes together with their multiplicities.
- A space is a set with additional structure. Examples include groups, semi-groups, vector spaces, rings, topological spaces, measure spaces, Hilbert spaces, and so on. The structure is what is important. A homomorphism is a transformation $h: A \rightarrow B$ that preserves structure: the relation between x , x' and x'' in A is the same as the relation between $h(x)$, $h(x')$ and $h(x'')$ in B .
- In statistical work, it is necessary to be clear whether a sample is a set or a list. If it is a list, are the elements distinct? Is it necessarily finite? Is it random or non-random? In these notes, a sample is a non-random finite list of observational units; the sample values are random. Distinctness simplifies the discussion, particularly in regard to exchangeability.
10. Grammar and style: Reports should be logically organized and written in grammatical English. In particular, each sentence should have one, and only

one, main verb. Poor logical organization is a signal of a confused mind, and poor sentence structure points to a lack of attention to detail.

11. Clarity and word usage: It is good practice to cultivate an awareness of grammar and word usage. Accurate word usage is important insofar as inaccurate or careless usage sows confusion; good grammar is important insofar as poor grammar betrays faulty logic.

For example, some native English speakers who are employed as commentators at sports events seem not to understand the difference between the verbs *substitute* and *replace*. These words are also important in mathematics. Viewers are likely to be confused when a talking head recommends at the end of the first quarter that the starting quarterback be substituted! For the correct usage in the active voice, the coach may *substitute* a bench player for a starter or he may *replace* the starter with a substitute from the bench. In the passive voice, an active player may be *replaced*, in which case a bench player is *substituted*. To declare that an active player has been substituted on account of injury is to put the focus on the destination, implying that the coach's job is to ensure that the bench is well-supplied with injured players! Unintended, perhaps, but possibly accurate.

In a similar vein with relevance to genetics, the upstream region of a gene may be rich in certain motifs, meaning that those short sequences are abundant in the upstream neighbourhood. The upstream region is enriched with motifs, but the motifs themselves are neither rich nor enriched anywhere. One could say that coal is abundant in Wyoming and fruit is plentiful in Florida, but coal is not enriched in Wyoming nor is fruit richer in Florida than it is in Georgia.

use versus usage: The line *What's the use of crying?* from the song *Smile* by Nat King Cole is a rhetorical query about the utility or futility of the act. Similarly, the phrase *cocaine use* refers to the act—its utility, its benefits or its prevalence. By contrast, the title *Modern English Usage* of Fowler's celebrated book refers to the manner in which the language is spoken or written, e.g., imaginatively, in long convoluted sentences, with flair, grammatically, clichéd, and so on. In the same vein, the phrase *cocaine usage* refers to the manner of ingestion. As a statistical factors, *cocaine usage* has levels snorting, smoking, injection and other; *cocaine use* has levels never, infrequent, occasional and regular.

Verbs for computational activities: Author A writes *I created a proportional-hazards model with covariates...*; author B writes *I ran a p-h model on the data...*; author C writes *I fitted the p-h model...*; author D writes *I trained the p-h model...*; author E writes *The p-h model was trained...*; author F writes *I learned the p-h model....*

The proportional-hazards model is a set of probability distributions for survival times. Credit for its creation goes to Cox (1972), not to author A. Generally speaking, one *runs* computer code for an algorithm that is designed to pick the distribution that best fits the data. This activity is called model-fitting—or learning in CS circles. In a sense, the computer or the algorithm learns the best-fitting distribution, possibly using data from a training subsample, and

shares that wisdom with the user. Grammatically speaking, if the p-h model is trained on the data, and learns from it, it would be more accurate for author F to write *The data taught the proportional-hazards model...*, or perhaps, *I used the data to teach the proportional-hazards model...*, but the semantic anomaly would then be too evident.

12. Appropriate adjectives: Some computational tasks are easy, while other are hard; some algorithms are efficient for the task while other are inefficient. Likewise for a software implementation of an algorithm. Simulation is easy for some distributions, less so for others. Maximum-likelihood estimation for some models admits a computationally efficient algorithm; not so for other models.

A task may be easy or it may be hard, but it is neither efficient nor inefficient. A model as a set of probability distributions may be finite or infinite, finite-dimensional or infinite-dimensional; it may be suited to the task or it may not; but it is neither easy nor hard, efficient nor inefficient.

13. Verb tense: Reports are best written in the present tense. If you wish to refer to a past event, by all means use the past tense; likewise for future events. If you switch from one tense to another mid-paragraph readers will notice, and if a good reason is not apparent, the result will be confusion. It is best to keep the bulk of your report in the present tense, including references to later sections: *An open-air experiment was conducted during the period 2012–2015; the data from that experiment were analyzed and conclusions are presented in sections 4–5.*

Present tense: *Anthropogenic emissions lead to global climate warming.* Past tense: *Anthropogenic emissions led to global climate warming.* Past perfect tense: *Anthropogenic emissions have led to global climate warming.* Both versions of the past tense, but particularly the first, suggest (probably incorrectly) that anthropogenic emissions no longer have the effect that they had in the past. That incorrect implication may be deliberate if the writer is a White-House hack seeking to justify the U.S. exodus from the Paris Accord, but it is a distraction for the discerning reader. Future tense: *Anthropogenic emissions will lead to global climate warming.* The future tense suggests that emissions did not have this effect in the past.

14. Numbers in text: A sentence must not begin with a mathematical symbol or a numeral. Small integers 0–10 or 0–12 are usually spelled out when they occur in text. A 3^{4-1} design has 27 observational units indexed by four factors with three levels each. Zero is one of the dose levels. The zero subspace $\mathbf{0} = \{0\} \subset \mathbb{R}^n$, which contains one point and has dimension zero, is not to be confused with the empty subset $\emptyset \subset \mathbb{R}^n$, which contains zero points and has no dimension.
15. Quantities; number, amount, volume: *a great number of tired tourists, diving dolphins, ornery kangaroos, football supporters,...; amount of cash in low denominations, amount of food, alcohol,...; volume or tonnage of crude oil, undelivered mail, mining sludge, ripe tomatoes,...; mass of water, mass of humanity; less time, fewer people.*

20.2 Coaching Tips II

These remarks are the instructor's responses following a Statistics consulting presentation by students on 17 April, 2018. The experiment was done on mice, and the design was factorial with three factors; the observations were cell counts, all large integers.

1. Transformation: In applications of this sort where the observation is a cell count, or any large count of objects, it is much more natural for treatment effects to be multiplicative than additive. Why so? If the mean cell count for controls in the three genotypes are 1000, 2000, 3000, and the treatment effect is -0.69 , or a 50% reduction, the cell means for treated mice in the same genotype classes will be 500, 1000, 1500. So the average reduction is 1000. An additive model with an additive reduction of 1000 has treatment means 0000, 1000, 2000 for the three groups. Usually, this sort of thing—no cells at all in one group—is very implausible; negative counts are even less likely. So the conclusion is that the log scale should be the first option for analysis, the go-to choice, but not necessarily the final choice.
2. Experimental units versus observational units: In this experiment, the observational units are mice, and all responses are measured post-mortem. However, all mice in one litter have the same genotype and were given the same treatment. This is a classic distinction. It is not possible in this design for two mice in the same litter to be given different treatments. Accordingly, the mice are the observational units, and the litters are the experimental units, i.e., each litter is one experimental unit. It is one of the few universally-agreed rules of experimental design and analysis that you cannot have more degrees of freedom for the estimation of treatment-effect variances than there are experimental units available for analysis (27 in this case). One way to proceed is to reduce each litter to the litter average, and to do the standard factorial decomposition on the litter averages. My preference is to average the counts and then take the log, but you could take logs first and then average. The operations are not commutative.
3. Random effects: The use of litter averages is not ideal because litters vary in size, probably from one to six or thereabouts. A linear analysis weighted by litter size is not correct either—that weighting is too extreme. A better option is to use a random-effects model in which each litter is associated with an independent additive Gaussian variable with constant variance independent of litter size. Since each random effect is associated with the contribution of one experimental unit, the question of significance testing for a zero between-litter variance is not something that arises naturally. There is simply no reason to expect zero additional variance per experimental unit, so the litter effect must be retained whether it is statistically significant or not. Remember that it is the experimental units that govern the degrees of freedom for treatment-effect estimation: the number of observational units is entirely irrelevant, even if infinite.

4. Model selection: In this design, there are three crossed factors $3 \times 2 \times 2$, where genotype is a three-level classification factor. Ordinarily, in the analysis of a factorial design, the main effects of all three factors are retained in the ‘final model’, regardless of significance. This is sound scientific practice, and there are many reasons for it. Comparability with other studies of the same phenomenon in similar or different circumstances is paramount. For three factors, regardless of the number of levels for each, there are only 9 factorial subspaces that include all three main effects,

$$A + B + C, \quad A * B + C, \quad \dots, \quad A * B + B * C, \quad \dots, \quad A * B * C.$$

Of these, only about five are likely to be seriously contemplated: $A + B + C$ (additivity, no interactions), $A * B + C$, $A + B * C$, $B + A * C$ (one interaction only, but additivity for the other), $A * B * C$ (no additivity anywhere). The lesson: whatever you learned in class about model selection in regression is not relevant here. Subset selection is not such a big issue in most scientific work involving factorial designs, and standard covariate selection algorithms are an outright menace in this setting. However, if you are using a random-effects model, as you ought, you do need to be careful to use a proper likelihood-ratio statistic (NOT REML) for the comparison of two nested factorial models. This is one of the few instances where a technical measure-theoretic issue impinges on statistical methodology. If you are unsure about the technicality, just ask.

5. Coding of factors: Unless the factor levels are ordered or have additional structure, the fitted model should be independent of the coding. For example a factor with two levels coded ‘M’ and ‘F’ might represent sex—or it might represent parent. In one case the ‘F’ level stands for ‘Female’ in the other case it stands for ‘Father’. It is clearly unacceptable for the fitting or selection procedure to depend on the letter or character string used to represent each level. Each term in a factorial model is a vector subspace; although the labelling of the basis vectors must depend on the coding, the subspace itself is invariant with respect to coding. The coding determines the basis vectors but not the subspace. The factorial models are essentially the only subspaces that have this property, which is most naturally expressed in terms of algebraic representation theory. A model-selection procedure that is code-dependent is a plague to be avoided.
6. Graphs and tables: The purpose of a table or graph is to advance the narrative by drawing attention to the most important patterns or features in the data such as the nature and direction of various effects. A graph is helpful for presentation only if it illustrates an important effect clearly. But, without narrative support, the graph is silent; its relevance to the conclusions must be spelled out in the text. A graph of residuals may be helpful for model checking, and may be mentioned in presentation, but, unless it is explicitly requested in a homework exercise, it is seldom included as part of the report. In most factorial designs, whether balanced or otherwise, one-way and two-way tables of averages are often useful as a partial summary of conclusions.

7. Higher-order interactions: How do you present the conclusions comprehensibly if high-order interactions are present? If the additive model is a satisfactory fit, you can report estimates of main effect contrasts in the usual way—pooling higher-order interaction sums of squares to obtain an estimate of variance. There is little need to encourage lazy scientific behaviour by giving undue emphasis to p -values, but you should report the degrees of freedom of the variance estimate, particularly if it is small. If, as appears to be the case here, there is a high-order interaction, it is best to partition the units into sub-classes by genotype, and to report the treatment effects separately, but in parallel, for each genotype. Since there are four treatment combinations for each genotype, you can report the three contrasts with some reference level. Show these numbers in a 3×4 table, one row per genotype with standard errors but absolutely no p -values.
8. Rules of thumb for summary statistical tables:
 - (a) Report effect estimates and standard errors only. Ratios are OK. No asterisks please!
 - (b) Always label the effects in an informative way so that the reference level for each factor is clear.
 - (c) Always report the reference level of each factor with zero as the estimate.
 - (d) Always report the variance-component estimates.
 - (e) Estimates and regression coefficients: four significant decimal digits maximum.
 - (f) Standard errors: three decimal digits maximum.
 - (g) F -ratios: two decimal digits maximum.
 - (h) p -values: best used sparingly, but two digits maximum as a percent if needed.
 - (i) If you must report a p -value, be sure to state the null hypothesis being tested.
9. Baseline: Baseline refers to a point in time just prior to randomization and treatment assignment. Notionally, the probability model for the outcomes is registered at baseline, so all information needed to determine outcome probabilities (including the randomization outcome) must be revealed at that time. Any variable recorded at or pre-baseline is called a baseline variable. A block factor is an example of a baseline variable. Age at recruitment in a clinical trial is a baseline variable.
10. Covariate: A covariate is a function on the units that is known in advance and recorded pre-baseline for the in-sample units. Typical covariates in a clinical trial include age, sex, and medical history. In the case of a vital response variable such as blood pressure, cholesterol or blood serum level, the baseline value is usually recorded as part of the recruitment interview. As such, the initial response and other baseline variables may be used to determine eligibility for inclusion in the study, particularly if the study focuses on high-risk patients. Thus, the baseline response value is, or may be treated as, a covariate, which is regarded as fixed in the probability model.

11. Treatment assignment: Treatment assignment is determined by randomization at baseline. Usually the treatment is not assigned independently to experimental units, but is subject to design conditions such as balance and equi-replication within blocks. In general, the treatment assignment probabilities, most obviously the joint probabilities for two or more observational units, may depend on block sizes, covariates and other baseline variables. The treatment assignment vector is technically a random variable, not a covariate or baseline variable.
12. Response: Many studies have multiple responses per observational unit, for example birth weight and gestational period in a study of the effect of certain interventions in a medical setting. For such a setting, each observational/experimental unit i is a mother/baby, and the response $i \mapsto (t_i, w_i)$ is bivariate. Treatment (e.g., folic acid supplement) may have an effect on the baby's weight at birth; it may also have an effect on the probability of a premature birth. So there are at least two treatment effects to be considered. The effect of treatment on birth weight is ordinarily defined as the difference of average weights or log-weights; the effect of treatment on gestational period is defined likewise. But, in general, the full story is the effect of treatment on the joint distribution: gestational period and birth weight are strongly correlated. To estimate the effect of treatment on birth weight, it is not legitimate to include gestational period as a 'covariate' in the one-dimensional model for birth weight.

20.3 Exercises

20.1 The verb *to write* has a subject, a direct object and an indirect object. Some of the parts may be empty or missing. Identify the three parts in the following sentences
(i) *Joe wrote a letter to Anne*; (ii) *Anne sent Joe a present*; (iii) *Joe wrote Anne a long passionate letter*.

20.2 An involution is a transformation f that is self-inverse, i.e. $f(f(x)) = x$. Show that the verbs *to substitute* and *to replace* are both transformations, and that the relation between them is an involution.

20.3 Let A, B be two groups. What are the relations between x, x', x'' in A that are preserved by a group homomorphism $h: A \rightarrow B$?

20.4 Let A, B be two vector spaces, i.e., commutative groups with additional structure. What is the additional structure? What are the additional relations between x, x', x'' in A that are preserved by a vector-space homomorphism $h: A \rightarrow B$?

20.5 Discuss whether or not the mapping $x \mapsto \|x\|^2$ (or its inverse image) is a homomorphism

$$(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), N(0, I_d)) \xrightarrow{\|x\|^2} (\mathbb{R}, \mathcal{B}(\mathbb{R}), \chi_d^2)$$

of probability spaces.

Chapter 21

Q & A

21.1 Scientific Investigations

21.1.1 Observational Unit

- Q1.* Who made the world?
- A1. God made the world.
- Q2.* Was it an experiment?
- A2. We have every reason to believe so. It is the best explanation we have for the current state of pestilence and political chaos.
- Q3.* Where did He start?
- A3. If it was an experiment, He started at the baseline.
- Q4.* What is the baseline?
- A4. A point in time prior to all experience—the most recent point in time prior to the revelation of protocols and the implementation of randomization.
- Q5.* Does anything exist before the baseline?
- A5. Yes, every scientific investigation has a protocol—written or unwritten.
- Q6.* What is the protocol?
- A6. The protocol is a declaration of scientific purpose, experimental timeline, strategy, tactics, and so on. Including, of course, the stochastic model and the method of analysis to be used.
- Q7.* Tell me about the individual parts.
- A7. The purpose refers to the phenomenon to be investigated, both the response and the target population. Strategy, tactics and timeline refer to the study design, the sample, and the measurement process.
- Q8.* What is the target population?
- A8. The target population is the set of observational units.
- Q9.* Does the population exist pre-baseline?
- A9. The observational units are declared pre-baseline. If they didn't exist, they couldn't be declared.

- Q10. What does it mean for something to exist?*
- A10. Existence means occurrence as a feature or component in the mathematics as declared in the protocol.
- Q11. Does mathematical existence have any connection with reality?*
- A11. Assuming that we can agree on the meaning of reality, everything of interest that exists in reality, and every relevant event that could possibly occur in reality, must have a counterpart in the mathematics. Reality in that counterpart sense is a subset of mathematics.
- Q12. Isn't that asking a lot from mathematics?*
- A12. Yes and no. The phrase ‘everything of interest’ implies compartmentalization or restriction to objects and events that are considered relevant to the investigation.
- Q13. What objects and events are relevant to God's experiment?*
- A13. Only God can answer that. From time to time one reads of claims that His protocols have been revealed to a select few who pose as His interpreters. But the evidence presented is not especially convincing.
- Q14. Is every observational unit a physical object?*
- A14. Every observational unit is an identifiable mathematical object, which may or may not correspond to an existing physical object.
- Q15. Tell me more about that.*
- A15. The NW3 weather station near Hampstead is a physical object of sorts, but the observational units in a meteorological series are site-time pairs. A data analyst is usually content to represent the sample units by certain floating-point pairs of numbers in an electronic computer. But the mathematical system contains uncountably many units that cannot be represented in an electronic computer.
- Q16. How many observational units are there?*
- A16. Usually the number in the population is infinite. But the sample is always finite.
- Q17. So the sample is a finite random subset of the population?*
- A17. Finite, yes.
- Q18. And random?*
- A18. The status of the sample as a fixed subset or a random subset is part of what is revealed implicitly or explicitly by the protocol.
- Q19. Can you give me an example.*
- A19. The protocol identifies the baseline, the population of interest, the sample or sampling scheme, and the response variable. Suppose the baseline is Dec 31, 1899. The protocol declaration *daily noon temperature at Kew, Greenwich and Hampstead, Jan 1, 1900 to Dec 31, 2020* identifies the response and the sample points as a fixed finite set.
- Q20. And the population?*
- A20. Usually it is not necessary to be persnickety about the population, so any larger space-time domain suffices. In the absence of a compelling argument to the contrary, the entire space-time product set serves as the population.

- Q21. Didn't you say that the population must exist at baseline? Does Jan 8, 2022 exist at baseline?*
- A21. Yes, I did say that. And yes, the ordered pair (Kew, Jan 8, 2022) exists today just as it did in AD 1899. But I did not say that every unit must be accessible or observable immediately after baseline. Even a mathematician has little control over the flow of time.
- Q22. The space-time product set is uncountable in both dimensions. Isn't that excessive and unnecessarily extensive for statistical work?*
- A22. Maybe so, but imperialism is inscribed in the DNA of mathematics. Besides, if you restrict the population, you forego the opportunity to make inferences about the disenfranchised parts.
- Q23. Is that a problem?*
- A23. Only if you decide later that you want to say something about units whose existence was not declared.
- Q24. Couldn't you declare them retroactively?*
- A24. So long as the extension fits comfortably into the original scheme, yes you could and you must, and probably you won't be penalized for the oversight. But, if there are multiple extensions, a convincing argument for a particular extension may be problematic.

21.1.2 Clinical Trials

- Q1. What is the role of the protocol for a clinical trial?*
- A1. Patient eligibility is one crucial part.
- Q2. And what are the implications of eligibility criteria?*
- A2. The population consists of all eligible patients—patients who were eligible yesterday, patients who are eligible today, and most certainly those who will be eligible tomorrow. The recruitment scheme for patients is also part of the protocol. Of necessity, the sample is a subset of patients who are eligible today—while recruitment is open. Usually the sample is also restricted geographically.
- Q3. What's the point of including dead folk?*
- A3. Why not? They don't charge for service or rent.
- Q4. Why include patients who are not yet born?*
- A4. If you are interested only in the current population, so be it. That's OK for short-range planning and short-sighted politicians. If a goal is to say something about the effect of a COVID vaccine or global warming, you may wish to cast the net liberally by including future generations.
- Q5. What hope is there of saying anything useful about the effect of a vaccine on future generations?*
- A5. The purpose of casting a wide net is not to say something *useful* about future generations, but to be in a position to say anything at all.

Q6. Could you elaborate on that?

A6. The point is that the population as declared in the protocol is the universe of discourse. You, the statistician, have the obligation and the privilege of specifying that set. It may be restricted in various ways by eligibility criteria, and so on. The sample is certainly restricted geographically and temporally. If you focus solely on the sample, all you can do is arithmetic on sample values. That sort of activity is for accountants. Typical questions of scientific interest involve extra-sample units, and you cannot begin to address those unless the universe of discourse contains them.

Q7. Patients in a clinical trial are usually recruited sequentially as they present themselves at the medical centre. Is a sequentially-recruited sample fixed or random?

A7. That appears to be a philosophically complicated question, but, ... (reaching into his pocket), here is a sample of six pennies. Is it a fixed subset of all pennies or a random subset?

Q8. It's obviously a random subset.

A8. And if I say that it is a fixed subset, can anyone prove otherwise?

Q9. Perhaps not, but I would not believe you.

A9. And you might well be right to be skeptical. But the question is meaningless without mathematical context. It can be answered only as a mathematical question in a mathematical context.

Q10. So how do you formulate sequential recruitment mathematically?

A10. First, you must retain in the mathematics anything that is essential for the context. All else can and must be discarded. One obvious difficulty is that there is no master-list of eligible patients—not even a comprehensive list of patients who are eligible today. So either you pretend that there is a master list, or you figure out a way to cope without it.

Q11. How does mathematics cope without a master-list?

A11. One solution is to record eligible patients as a point process by date of presentation—with follow-up to monitor disease progress. The sample is the subset that presents at a given medical centre in a bounded temporal window.

Q12. So, is such a sample fixed or random?

A12. Well, the window is fixed and bounded, but the sample as a set of arrival times is random and finite.

Q13. And what about the patients? Can they be a random sample?

A13. That's complicated because there is no master-list that can be identified as the set of eligible patients. There is only a window and a set of presentation times, which we use to label patients in the sample.

Q14. So, what is it? Fixed or random?

A14. When it comes to patients, you need to get over your fixation with fixed versus random samples. The set of patients is neither a fixed subset nor a random subset of anything because there is no concept in the mathematics of a population of eligible patients. The recruitment window is a master-list of time-points, and the presentation times comprise a finite random subset.

Disease occurrence is a stochastic process and recruitment is a stochastic process.

- Q15. That doesn't seem to fit in with the general framework as described earlier.*
- A15. Maybe so, but the flaw is more in the language than in the mathematics. It would sound anti-social, even callous, to talk of the sample units as time points rather than patients. But that's the way it is. One patient is associated with each recruitment time, so we talk of patients rather than times. Think of it as the statistician's bedside manner: social training emphasizes the occupants rather than the beds.
- Q16. How can there be a covariate associated with recruitment times?*
- A16. The recruitment process is a marked point process that is observed in a fixed temporal window. Each patient has his own baseline, which is the time of recruitment. The marks, which include age and sex plus current and future health, are random variables. However, any marks that are revealed at recruitment are pre-baseline values. That includes the sample size or window length. A point-process sample is a random sample of time points.
- Q17. How does point-process recruitment affect the statistical analysis?*
- A17. There could be an issue about volunteers versus non-volunteers. It's possible they're not going to respond in the same way, and you're not going to find that out. But that's a different matter from sequential recruitment. In my opinion, the set of sequentially-recruited patients is best treated as a fixed subset of an infinite population.
- Q18. Only patients who have access to a qualified physician are included in your description of the population. What about those who are eligible but do not have access, either for reasons of geography or economics?*
- A18. Whether the sample is fixed or random, it can usually be guaranteed that there are units in the population that have zero probability of being included in the sample. If the outcome (the effect of COVID vaccine) were very different depending on geography or economics, we would certainly want to know. Practically speaking, it is best to recruit broadly and to record adequate baseline information.
- Q19. I want to revisit a remark that you made earlier about mathematicians being imperialistic in outlook. Could you elaborate on that?*
- A19. Far be it from me to say anything derogatory about mathematicians or statisticians, either individually or as a group. I did say that mathematics was imperialistic in outlook.
- Q20. That sounds like criticism to me. What do you mean by it?*
- A20. Oddly enough, I meant it in a positive and approving way. Mathematics has always been imperialistic, and it should be imperialistic. When Pythagoras discovered his theorem, he declared it to be a universal truth holding not only for Greek triangles but also for Egyptian and Assyrian triangles as well. Similar remarks hold for Archimedes and physics. That sort of imperialism is good. Maybe catholic ($\kappa\alpha\thetaολικος$) or universal would be a better word. But both catholicism and imperialism have unfortunate negative connotations.

- Q21. It is hard to see the relevance of catholicism or imperialism to applied statistics.*
- A21. On the contrary. Random samples and finite-population models for clinical trials are a case in point. There is nothing mathematically wrong with a finite population if that is your chosen universe of discourse. But the philosophy is democratic, short-sighted and inward-looking—all bad for science and medicine.
- Q22. How so?*
- A22. To arrange matters so that all individuals in the current population—and only those individuals—have strictly positive inclusion probability is an undeniably democratic idea. But it comes at enormous cost to subsequent generations who are not accessible today, and must be excluded. It is also contrary to the spirit of scientific catholicism, which recoils at restrictions. I would venture further to say that any medical statistician who restricts the population to the current generation is mathematically derelict in his or her duty of care to subsequent generations.
- Q23. But surely the finiteness assumption can't make much difference to procedures and conclusions.*
- A23. It absolutely makes a difference to conclusions because, if you don't admit that the next generation exists in your population, you forego the opportunity to say anything about the effect of treatment tomorrow.
- Q24. What reason is there to say that the effect of treatment today must be the same as the effect tomorrow?*
- A24. I do not claim that the effect is constant over generations. But I do insist on the opportunity to make that comparison. As do you, apparently, since you raised the question. If you were to conclude that today's data are irrelevant for tomorrow's patients, that would be fine by me. But if you don't admit the existence of tomorrow, you can't say even that.
- Q25. But medical recommendations are seldom explicitly time-constrained.*
- A25. True enough. In that case your actions imply that today's data are relevant to some degree, and that the effect is constant or nearly so.
- Q26. A great part of classical sampling theory is based on the Horvitz-Thompson estimator, which requires strictly positive inclusion probabilities. How does that fit in with your philosophy?*
- A26. Horvitz-Thompson sampling theory is a fine and mathematically elegant theory for current-population sampling. It has an important role for public policy in democratic societies. However, inverse-probability weighting is the ultimate in scientific negativism. By insisting on positive inclusion probabilities for all, it implies that tomorrow's population is beyond the scope of discussion.
- Q27. So, how does philosophy affect procedures?*
- A27. Mathematically speaking, you can't have it both ways. If you want to say something about the effect of treatment in the future, or how it might have affected past generations, you must have those generations in the target population. If you insist on a finite population with a random sample

and strictly positive sample-inclusion probabilities, future generations must be excluded, and you forego the opportunity to address certain critical questions. Philosophically, I'm a statistical catholic, so you know where I stand.

Q28. That answer is a bit strident, is it not?

A28. Strident or not, it lays out the issue as clearly as I can.

21.1.3 Agricultural Field Trials

Q1. Can you say a little about agricultural field trials?

A1. By comparison with clinical trials, field trials are very simple.

Q2. How so?

A2. Each observational unit is a plot in the field. The protocol specifies the varieties or cultivars to be tested by growing on the sample, which consists of 36 plots situated at the western end of Hoos field. That's all there is to it. No recruitment or random sampling of plots. Only random assignment of varieties to plots in the sample.

Q3. My impression was that random samples were the norm in all statistical work. Wouldn't it be better to use a random sample of plots from several fields?

A3. Try that on the farm manager! But you might make a case for a more extensive design replicated in several distant blocks differing in soil composition or weather pattern. A variety that performs well at Rothamsted might fare poorly in Rotherham or Rothesay: Student (1934) describes a good example of this phenomenon for barley varieties.

Q4. I have an image of each sample unit as a rectangular plot, all sample plots being neatly arranged by rows and columns separated by access paths. What does the population of 'all plots' look like.

A4. Your image is a bit idyllic, but the population is a family of planar subsets.

Q5. Are they all the same size and shape?

A5. Not at all. It is not necessary to include all planar subsets, but a mathematician instinctively aims to include all Borel subsets. That's a big set, big enough for most purposes, but maybe not big enough for all purposes. The units in a long-term field experiment also have a temporal component.

Q6. That seems far too big. Besides, plots cannot overlap.

A6. A catholic statistician must always think big. If the response is yield, there is no concern about overlap: yield is an additive set function.

Q7. But you cannot have different treatments on overlapping plots.

A7. That's a good reason for picking a sample of non-overlapping plots.

- Q8. If your sample of plots is a non-random subset, where does the probability come from?*
- A8. Probability comes from the mathematical framework that is implicit or explicit in the protocol. Exchangeability gives rise to probabilities. Randomization also gives rise to probabilities.
- Q9. What is the role of randomization analysis?*
- A9. Randomization is usually associated with the uniform distribution on a finite group acting on the sample units. Re-randomization enables you to generate new ‘pseudo-samples’ having the same distribution as the original. For any non-invariant statistic, you can compute its randomization distribution. This is a useful way to determine where the observed treatment effect occurs in the spectrum of treatments effects anticipated under randomization.
- Q10. So the set of units in the randomization analysis is the finite sample of plots?*
- A10. Certainly the sample is finite.
- Q11. Isn't the randomization population the same as the sample?*
- A11. In a purely arithmetical sense, yes!
- Q12. Is there any other sense?*
- A12. There must always be a wider statistical sense.
- Q13. To what end?*
- A13. Presumably you want to say something about the likely effect of treatment on other plots of a similar type in the population.
- Q14. Couldn't you just take the finite-population estimate, patch it together with the randomization distribution or bootstrap distribution, and apply that to other plots.*
- A14. If you had no principles or concerns about mathematical integrity, you could do whatever you liked.
- Q15. Isn't that what every statistician does? Are we all dishonest?*
- A15. It is true that many statisticians do exactly that—and very often it is the right thing to do if not always for the reasons stated.
- Q16. So what's the problem?*
- A16. The problem is one of honesty in mathematics. If you refuse to acknowledge extra-sample plots, the statement about treatment effect is meaningless. If you acknowledge their existence you have to establish a connection between yields on the in-sample plots and yields on extra-sample plots. That step requires an assumption such as stationarity or exchangeability with respect to extra-sample plots.
- Q17. In that case, what is the role of randomization analysis?*
- A17. Randomization analyses and bootstrap analyses are logically sound and useful statistical tools. On its own—restricted to the finite sample of plots—randomization is a basis for arithmetic and distribution-theory. It is not otherwise a basis for statistical inference in the sense of prediction for extra-sample plots.

21.1.4 Covariates

- Q1.* Apart from the observational units, what else exists before the baseline?
- A1. Covariates are recorded pre-baseline.
- Q2.* What is a covariate?
- A2. A variable recorded pre-baseline.
- Q3.* What is a variable?
- A3. A variable is a function on the observational units.
- Q4.* What types of covariate are there?
- A4. Qualitative variables or classification factors, and quantitative variables such as age or calendar date or spatial position.
- Q5.* Are there any other types of covariate?
- A5. Yes, relationships can also be recorded at baseline.
- Q6.* What is a relationship?
- A6. A relationship is a function on pairs of observational units.
- Q7.* Can you give examples.
- A7. A block factor is an equivalence relation; there are also genetic relationships, familial relationships, temporal relationships, adjacency relationships, and metric relationships.
- Q8.* What is a metric relationship?
- A8. A metric is a symmetric non-negative function on pairs that satisfies the triangle inequality.
- Q9.* Any other examples of relationships?
- A9. On any space, the identity function is a relationship on pairs; it tells you whether the two elements are the same or different. That's a rather fundamental component of elementary set theory. In a Euclidean space, the inner product is a relationship between pairs of points.
- Q10.* Are any covariates recorded post-baseline?
- A10. No. Every post-baseline variable is a random outcome subject to the rules of probability.
- Q11.* What happens at baseline?
- A11. The protocol is announced, units are assembled, treatment is assigned by randomization, and nature or Tyche takes over.
- Q12.* Who is Tyche?
- A12. Tyche is the Greek goddess of chance—Fortuna to the Romans.
- Q13.* Is treatment a covariate?
- A13. No, it is not.
- Q14.* Why not?
- A14. Treatment is the outcome of randomization as specified by protocol. It is a random variable, albeit ancillary in all settings.
- Q15.* Is treatment assigned independently to units in the sample?
- A15. No, not usually. A balanced design implies non-independent assignments.

Q16. Is the treatment assignment distribution the same for every unit?

A16. Not necessarily. In principle, the treatment assignment probability may vary from one covariate sub-group to another as specified by protocol. But this practice is not common and is not encouraged.

Q17. What is the purpose of randomized treatment assignment?

A17. Randomization is a panacea. It has many purposes.

Q18. Tell me one specific purpose.

A18. Concealment of treatment assignment promotes integrity in human trials.

Q19. Can you elaborate?

A19. Where human subjects are involved, the integrity of the experiment is at risk if the treatment assignment is revealed prematurely, either to the patient or to the physician. Concealment helps to limit the possibilities for subverting the design.

Q20. Any other purposes?

A20. To see if God is paying attention.

Q21. What has God got to do with it?

A21. Concealment means that treatment assignment is known only to the controlling statistician, who must pay attention to events as they unfold.

Q22. Any other purpose?

A22. To help convince skeptics by levelling the playing field for treatment comparisons.

Q23. Tell me about the role of exchangeability?

A23. Exchangeability is the fundamental axiom of statistical modelling.

Q24. What does exchangeability imply?

A24. It implies that two units having the same covariate value must have the same response distribution. Implicitly or explicitly, that's usually part of the protocol.

Q25. Is exchangeability a mathematical theorem?

A25. No, it is an axiom of applied statistics. You can think of it as a bill of rights or a guarantee of equality under the law. If two units are to have different response distributions, there needs to be a demonstrable reason for that difference. That's where covariates enter the story.

Q26. What is the purpose of a covariate in a randomized study?

A26. There are three inter-related purposes.

(i) to accommodate sub-group effects (sex, age,...);

(ii) to improve precision of the treatment estimate;

(iii) to check for interaction.

Q27. What does interaction mean?

A27. Interaction means that the effect of treatment on males is different from its effect on females.

Q28. So that means that you have two numbers, one for males and one for females.

A28. Yes. If the treatment effect is summarized in a single real number, you have one number for males and one for females.

- Q29. Two treatment values means two real numbers, right?*
- A29. No, not in general. You can measure the efficacy of Covid vaccine by its ability to reduce future infection. That's one effect. You can also measure efficacy by its ability to reduce mortality given a subsequent infection. That's another effect. One number does not suffice; you need at least two numbers to define the treatment effect. No interaction means that both numbers are the same for males and females.
- Q30. Is interaction always concerned with sex or gender?*
- A30. No, not at all. The effect of treatment might depend on any baseline variable. Age is a good example in the case of Covid vaccination. It appears that most vaccines have some capacity to prevent future infections in adults, but little capacity in children. On the other hand, most vaccines appear to reduce mortality and hospitalization rates regardless of age.

21.1.5 Matched Design

- Q1. What is the baseline for a matched-pairs design?*
- A1. The time when the units are assembled or declared, just pre-randomization.
- Q2. What covariates are available in the matched-pairs design?*
- A2. In the simplest setting, only the block factor indicating the pairs. In the absence of other covariates, the pairs themselves are usually regarded as exchangeable.
- Q3. Is treatment always assigned independently to units?*
- A3. That's a question about protocol. The assignments could be independent, but usually each pair is restricted to (C, T) or (T, C) . That means (T, T) and (C, C) have zero probability, so the assignments are not independent.
- Q4. Is there a difference between a matched-pairs design and a pre-post design with initial values as discussed in Sect. 13.2?*
- A4. They are fundamentally different, but the similarities are strong enough to cause confusion. The main point of similarity is that observations come in pairs that are presumed to be positively correlated. In a matched-pairs design, one unit in each pair is assigned to 'C' and the other to 'T'. In an initial-value design with randomized assignment, the first unit in each pair is at 'baseline', and the second is assigned randomly to 'C' or 'T'. If you like, you can associate with the initial value a null treatment level, distinct from 'C' and 'T'.
- Q5. I can see that there's a technical distinction. But what effect can it have on the analysis?*
- A5. It is more than a minor technical distinction: it is a fundamental difference of design. The analysis for matched-pairs would usually focus on the within-pair differences $Y_{i,T} - Y_{i,C}$, the average difference and its standard deviation. So there is one difference for every pair, and the goal is to say something about the mean difference. If you were to do something along the same lines to

accommodate initial values, you might end up comparing the improvements $Y_{i,1} - Y_{i,0}$ for treated individuals with the improvements $Y_{i,1} - Y_{i,0}$ for controls. The two treatment groups are distinct individuals, so a standard two-sample comparison suggests itself. But that argument leads to a distinctly inferior assessment of the treatment effect.

- Q6. In the discussion of initial values in Sect. 13.2, what was the baseline for the hypertension study?*
- A6. Jan 1 when the patients were first measured to determine eligibility.
- Q7. Was that pre-randomization?*
- A7. Yes. Only eligible patients are randomized, so the determination of eligibility precedes randomization.
- Q8. What covariates are available in the hypertension study?*
- A8. The initial values plus the block factor, which is just patient ID. In practice, there are always lots of others.
- Q9. Section 13.2 discusses several methods for accommodating initial values. Which is the correct one?*
- A9. There is no single universally-correct analysis. The possible dependence of treatment on covariates such as age and sex makes that impossible. But the first three suggestions are effectively equivalent and can be extended as needed.
- Q10. Should we be concerned if treatment assignment were not independent of baseline values?*
- A10. That depends very much on the area of application. Dependence of treatment on baseline factors, particularly on spatial relationships, is standard protocol for field trials. For reasons of statistical efficiency, it is best to avoid having the same variety or treatment in adjacent plots. Simpler protocols are preferred for clinical work. In principle, if the protocol is followed, there's nothing to be concerned about.
- Q11. I can understand the dependence of treatment assignment on relationships. But that's not the same as dependence on initial values, is it?*
- A11. No. The initial value is a function on units; a relationship is a function on pairs.
- Q12. So I repeat the question: Which is the correct analysis for a pre-post design?*
- A12. Regardless of whether treatment assignment is independent of initial values, analysis of covariance with linear adjustment for initial values is a good place to start.
- Q13. Isn't it perverse to make the randomization probabilities depend on initial values?*
- A13. Whether it is perverse or not, randomization theory does not exclude the possibility of dependence, so you need to examine the protocol. It is standard practice to check whether the data are in accord with the protocol, for example, to check whether the baseline distribution is the same for the two treatment groups.

21.1.6 The Effect of Treatment

- Q1.* I've read that each patient in a two-arm randomized trial has two potential outcomes or counterfactual responses, only one of which can be recorded. Is that a fair description of the way you see it?
- A1. No, not really. My inclination is to focus only on what can be observed in principle. If only one observation can be recorded per patient, there is only one outcome per patient. Counterfactuals or potential outcomes are not used in these notes.
- Q2.* Why not? What's wrong with counterfactuals?
- A2. The issue is not whether there's anything right or wrong with the concept of counterfactuals, but whether there is a need for it, either in the real world or in the mathematics. Certainly, the concept occurs in everyday language, so there's a need for it in some sense. Such issues are encountered in thorny legal matters. *X died in a work-related accident at age 32. What would X's lifetime earnings have been had he not died?* That sort of determination is important, and we need a way to address it. The question is how we address it formally in the mathematics. It is a matter of mathematical style.
- Q3.* If it is merely a matter of style, what is there to argue about?
- A3. Style is more than sufficient for an argument. In academia, the lower the stakes the more tenacious the fight.
- Q4.* What are the mathematical issues?
- A4. The counterfactual world admits duplicate copies of each patient, one copy for each treatment level, and one outcome or response for each patient-treatment combination. A counterfactual stochastic model necessarily begins with a joint distribution for all outcomes, potential or otherwise. For n patients and five treatment levels, you're talking about a probability distribution on \mathbb{R}^{5n} . The sampling scheme is restricted to one treatment level for each patient, and the counterfactual process determines the joint distribution of those particular outcomes.
- Q5.* That seems straightforward. What other options are there?
- A5. The approach taken in these notes is to construct a process indexed by assignments. A sample consists of a subset of n patients together with a treatment assignment \mathbf{t} . The stochastic model associates with each finite sample a probability distribution $P_{\mathbf{t}}$ on \mathbb{R}^n . One can extend this to a counterfactual distribution on \mathbb{R}^{5n} , but the extension is not unique and is not needed for observables. Nor is there any possibility of using data to check the extension.
- Q6.* So the two approaches are equivalent for all observables?
- A6. For observables, yes they are exactly the same. But not for counterfactuals.
- Q7.* Can you give an example where the two approaches give different answers?
- A7. Patient i was assigned medication A for hypertension, and her outcome was $Y_{i,A} = 12.3$ in suitable units. What would her outcome have been if she had been assigned medication B ? That's a counterfactual question to which

the answer is simply the conditional distribution of the B -value given the A -value for this patient. So the answer depends on the joint distribution and the counterfactual correlation.

- Q8. Couldn't the counterfactual correlation be zero?*
- A8. Yes, it could be any number between -1 and $+1$ inclusive.
- Q9. How do you address the same question without counterfactuals?*
- A9. A genuinely counterfactual question admits an answer only in the counterfactual realm. But every applied statistician knows that the initial question is merely an opening gambit: the standard reply is to re-phrase the question. If you can persuade the interrogator to accept the re-phrased version, you're in business.
- Q10. How do you re-phrase the question to avoid counterfactuals?*
- A10. The population contains extra-sample individuals who have the same covariates as the target patient, but who were assigned to medication B . These patients all have the same response distribution. For any assignment such that the target patient gets A and the other gets B , you can report the conditional distribution for the extra-sample patient given the data.
- Q11. How do you know for sure that the population contains another individual having exactly the same covariate values as the target patient?*
- A11. You can be absolutely sure of that because the statistician has the luxury of defining the population. For reasons mentioned earlier, the population contains infinitely many units for each covariate value.
- Q12. So the counterfactual question is a question about a specific individual, and the re-phrased question insists on an extra-sample individual. Is that the only difference?*
- A12. That's all there is to it. But if the two patients are one and the same individual, the answer depends on counterfactual correlations. So the counterfactual and non-counterfactual answers are numerically different in general.
- Q13. How do you address an explicit counterfactual question such as X's lifetime earnings potential?*
- A13. You could think of injury or death at age 32 as a sort of treatment assignment. You'd definitely get into trouble if you presented that to the institutional review board, so you'd have to think it quietly.
- Q14. And where does that take you?*
- A14. The counterfactual set-up has multiple versions of individual X, the real one who died at age 32, and many who survived beyond that. The literal answer to the question of future earnings is the conditional distribution of earnings for X given the counterfactual history of health and earnings up to age 32. Except for the event of death, the counterfactual history agrees with the actual history.
- Q15. That seems fair enough. So the counterfactual framework is needed to address such matters?*
- A15. No, I don't think it is needed. You have in the population infinitely many individuals who have the same covariates as X, the same employment

history up to age 32, and so on, but who did not die at that age. You can compute the distribution of lifetime earnings for one individual in that subset.

Q16. And you get the same answer?

A16. That's complicated. The answers are the same, but the questions are different. The first is an answer to a counterfactual question about a specific individual X. The second is an answer to a non-counterfactual question about individuals other than X.

Q17. Why are the two answers numerically equal in the second case but not in the first?

A17. I'll leave that up to you. But the presumption is that the post-mortem lifetime earnings of X is non-random.

Q18. Let's move back to treatment. What do you mean by the effect of treatment?

A18. The effect of treatment is to modify the response distribution by group action. The effect is to change the control distribution for each patient to the corresponding active distribution.

Q19. Why must the effect of treatment be a group action?

A19. Well, you want to be in a position of saying 'Here are the distributions under consideration for a control, and here are the distributions under consideration for a treated patient'. And the two sets had better be the same for obvious reasons; a null treatment effect means that whatever the control distribution may be, the treatment distribution is the same. It's not just a pair of distributions; it's an action that takes each control distribution to a specific treatment distribution.

Q20. What do you mean by the treatment effect?

A20. The treatment effect is a specific group element, a parameter if you like.

Q21. Is the effect of treatment the same for everyone?

A21. Yes, in the sense that it is the same group action on distributions. But no, the particular group element need not be the same for everyone. It may vary from one covariate subset to another; in (5.2), the treatment effect increases linearly as a function of time.

Q22. If the treatment effect is not the same for everyone, how should it be reported? Must we report the average treatment effect?

A22. The question makes sense only if the treatment group is a vector space, and only if the population is finite. The first is often true, but not necessarily: see Exercises 14.7 and 14.8. An average also requires a distribution on the set of sampling units, and that's not normally part of the specification unless the population is finite.

Q23. So, how do you report a variable effect?

A23. If the treatment effect for males is not the same as that for females, you must report one group element for each sex, or perhaps one distribution for each sex. Same for population strata determined by any covariate or classification factor. It would be misleading to report a single value or a single distribution if there is substantial stratum-to-stratum variation.

References

- Adler, L., Barber, N. A., Biller, O. M., & Irwin, R. E. (2020). Flowering plant composition shapes pathogen infection intensity and reproduction in bumblebee colonies. *Proceedings of the National Academy of Sciences*, 117, 11559–11565.
- Aitchison, J. (1986). *The statistical analysis of compositional data*. London: Chapman and Hall. ISBN: 0-412-28060-4.
- Andrews, D. A., & Herzberg, A. (1985). *Data*. New York: Springer.
- Armitage, P. (1955). Tests for linear trends in proportions and frequencies. *Biometrics*, 11, 375–386.
- Atkinson, Q. D. (2011). Phonemic diversity supports a serial founder effect model of language expansion from Africa. *Science*, 32, 346–349.
- Austin, S. (2021). Review of *The Deep Places: A Memoir of Illness and Discovery* by R. Doukhan. New York Times Oct. 26, 2021.
- Bailey, R. A. (2008). *Design of comparative experiments*. Cambridge: Cambridge University Press. ISBN: 978-0-521-86506-7.
- Baltagi, B. H., Fingleton, B., & Pirotte, A. (2014). Spatial lag models with nested random effects: An instrumental variable procedure with an application to English house prices. *Journal of Urban Economics*, 80, 76–86. <https://doi.org/10.1016/j.jue.2013.10.006>
- Barnard, G. A. (1949). Statistical inference (with discussion). *Journal of the Royal Statistical Society B*, 11, 115–139.
- Barnard, G. A., Jenkins, G. M., & Winsten, C. B. (1962). Likelihood inference for time series. *Journal of the Royal Statistical Society A*, 125, 321–372.
- Bartlett, M. S. (1937). Properties of sufficiency and statistical tests. *Proceedings of the Royal Society, Series A*, 160, 268–282.
- Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67, 1–48.
- Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B*, 57, 289–300.
- Berger, J., & Wolpert, R. L. (1988). *The likelihood principle. IMS Lecture Note Series* (Vol 6, 2nd ed.). Hayward, CA: Institute of Mathematical Statistics.
- Birnbaum, A. (1962). On the foundations of statistical inference (with discussion). *Journal of the American Statistical Association*, 57, 269–306.
- Bliss, C. I. (1970). *Statistics in biology, vol. II*. New York: McGraw-Hill.
- Bock, R. D. (1975). *Multivariate statistical methods in behavioral research*. New York: McGraw-Hill. ISBN: 0-07-006305-2.

- Box, G. E. P. (1976). Science and statistics. *Journal of the American Statistical Association*, 71, 791–799. <https://doi.org/10.1080/01621459.1976.10480949>
- Box, G. E. P., & Cox, D. R. (1964). An analysis of transformations (with discussion). *Journal of the Royal Statistical Society B*, 26, 211–252.
- Box, G. E. P., & Tidwell, P. W. (1962). Transformation of the independent variables. *Technometrics*, 4, 531–550.
- Brien, C. J., Bailey, R. A., Tran, T. T., & Boland, J. (2012). Quasi-Latin designs. *Electronic Journal of Statistics*, 6, 1900–1925. <https://doi.org/10.1214/12-EJS732>
- Cavalli-Sforza, L. L., & Edwards, A. W. F. (1967). Phylogenetic analysis: Models and estimation procedures. *American Journal of Human Genetics*, 19, 233–257.
- Cellmer, R., Kobylinska, K., & Belej, M. (2019). Application of hierarchical spatial autoregressive models to develop land value maps in urban areas. *International Journal of Geo-Information*, 8, 195–214. <https://doi.org/10.3390/ijgi8040195> www.mdpi.com/journal/ijgi
- Clifford, D., & McCullagh, P. (2006). The regress function. *R News*, 6, 6–10.
- Cochran, W. G. (1954). Some methods for strengthening the common χ^2 tests. *Biometrics*, 10, 417–451.
- Cooch, E. G., & White, G. (2020). *Program MARK: A gentle introduction*. Fort Collins, CO: Colorado State University.
- Courant, R. (1965). Professor Richard Courant's acceptance speech for the distinguished service award. *The American Mathematical Monthly*, 72, 377–379. <https://doi.org/10.2307/2313496>
- Cox, D. R. (1958). *Planning of experiments*. New York: J. Wiley & Sons. ISBN: 0-471-1813-8.
- Cox, D. R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society B*, 74, 187–220.
- Cox, D. R. (2006). *Principles of statistical inference*. Cambridge: Cambridge University Press. ISBN: 0-521-68567-2.
- Cox, D. R., & Reid, N. (2000). *The theory of the design of experiments*. London: Chapman and Hall. ISBN: 1-58488-195-X.
- Cox, D. R., & Snell, E. J. (1981). *Applied statistics: Principles and examples*. London: Chapman and Hall. ISBN: 0-412-16570-8.
- Crane, H. (2016). The ubiquitous Ewens sampling formula. *Statistical Science*, 3, 11–19.
- Da Silva, P. H., Jamshidpey, A., McCullagh, P., & Tavaré, S. (2022). Fisher's measure of variability in repeated samples. *Bernoulli* (to appear).
- Davison, A. C. (2003). *Statistical models*. Cambridge: Cambridge University Press. ISBN: 05217773393.
- Dawid, A. P. (2000). Causal inference without counterfactuals. *Journal of the American Statistical Association*, 95, 407–424.
- Dawid, A. P. (2021). Decision-theoretic fundations for statistical causality. *Journal of Causal Inference*, 9, 39–77. <https://doi.org/10.1515/JCI-2020-0008>
- Dawson, R. B. (1954). A simplified expression for the variance of the χ^2 function on a contingency table. *Biometrika*, 41, 280.
- Dickey, D. A. (2020). A warning about Wald tests. SAS Global Forum 2020, paper 5088.
- Diggle, P., Heagerty, P., Liang, K.-Y., & Zeger, S. L. (2013). *Analysis of longitudinal data*. Oxford: Oxford University Press. ISBN: 978-0-19-967675-0.
- Dong, G., Harris, R., Jones, K., & Yu, J. (2015). Multilevel modelling with spatial interaction effects with application to a emerging land market in Beijing, China. *PLoS One*, 10, 1–18. <https://doi.org/10.1371/journal.pone.0130761>
- Douthat, R. (2021). *The deep places: A memoir of illness and discovery*. New York: Penguin Random House. ISBN: 9780593237366.
- Dyson, F. W. (1926). A method for correcting series of parallax observations. *Monthly Notices of the Royal Astronomical Society*, 86, 686–706. <https://doi.org/10.1093/mnras/86.9.686>
- Efron, B. (2010). *Large-scale inference. IMS monograph series*. Cambridge: Cambridge University Press.
- Efron, B. (2011). Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106, 1602–1614.

- EWENS, W. J. (1972). The sampling theory of selectively neutral alleles. *Theoretical Population Biology*, 3, 87–112.
- FELLER, W. (1971). *An introduction to probability theory and its applications, vol II*. New York: Wiley.
- FELSENSTEIN, J. (1973). Maximum-likelihood estimation of evolutionary trees from continuous characters. *American Journal of Human Genetics*, 25, 471–492.
- FELSENSTEIN, J. (2004). *Inferring phylogenies*. Sunderland: Sinauer Associates.
- FINGLETON, B., LE GALLO, J., & PIROTE, A. (2018). Panel data models with spatially dependent nested random effects. *Journal of Regional Science*, 58, 63–80. <https://doi.org/10.1111/jors.12327>
- FISHER, R. A. (1925). Theory of statistical estimation. *Proceedings of the Cambridge Philosophical Society*, 22, 700–725.
- FISHER, R. A. (1929). Moments and product moments of sampling distributions. *Proceedings of the London Mathematical Society*, 30, 199–238.
- FISHER, R. A. (1943). A theoretical distribution for the apparent abundance of different species. *Journal of Animal Ecology*, 12, 54–57.
- FISHER, R. A., CORBET, A. S., & WILLIAMS, C. B. (1943). The relation between the number of species and the number of individuals in a random sample from an animal population. *Journal of Animal Ecology*, 12, 42–58.
- FOWLER, H., & WHALEN, R. E. (1961). Variation in incentive stimulus and sexual behavior in the male rat. *Journal of Comparative and Physiological Psychology*, 54, 68–71.
- GELMAN, A. (2020). Concerns with that Stanford study of coronavirus prevalence. <https://statmodeling.stat.columbia.edu/2020/04/19/fatal-flaws-in-stanford-study-of-coronavirus-prevalence>
- GURKA, M., EDWARDS, L. J., MULLER, K. E., & KUPPER, L. L. (2006). Extending the Box-Cox transformation to the linear model. *Journal of the Royal Statistical Society A*, 169, 273–288.
- HALDANE, J. B. S. (1939). The mean and variance of χ^2 when used as a test of homogeneity when expectations are small. *Biometrika*, 31, 346–365.
- HORVITZ, D. G., & THOMPSON, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47, 663–685.
- ISSERLIS, L. (1918). On a formula for the product-moment coefficient of any order of a normal frequency distribution in any number of variables. *Biometrika*, 12, 134–139.
- JEFFREYS, H., & JEFFREYS, B. S. (1956). *Methods of mathematical physics*. Cambridge: Cambridge University Press.
- JIANG, J., & NGUYEN, T. (2021). *Linear and generalized linear mixed models and their applications* (2nd ed.). New York: Springer. ISBN: 978-1-0716-1281-1. <https://doi.org/10.1007/978-1-0716-1282-8>
- JOHNSON, R. N. (1972). *Aggression in man and animals*. Philadelphia, London: Saunders.
- JOHNSTONE, I., & SILVERMAN, B. W. (2004). Needles and straw in haystacks: Empirical-Bayes estimates of possibly sparse sequences. *The Annals of Statistics*, 32, 1594–1649.
- KAISSER, J. (2021). Key cancer results fail to be reproduced. *Science*, 374, 1311.
- KERRICH, J. E. (1946). *An experimental introduction to the theory of probability*. Copenhagen: Einar Munksgaard.
- KERRICH, J. E. (1961). Random remarks. *The American Statistician*, 15, 16–20.
- KIMBLE, G. A., GARMEZY, N., & ZIGLER, E. (1980). *Principles of general psychology*. New York: J. Wiley & Sons.
- KOLMOGOROV, A. N. (1933). *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Ergebnisse der Mathematik (Vol. 2). New York: Springer.
- LI, T., HOLST, T., MICHAELSEN, A., & RINNAN, R. (2019). Amplification of plant volatile defence against insect herbivory in a warming Arctic tundra. *Nature Plants*, 5, 568–574.
- LIANG, K.-Y., & ZEGER, S. L. (2000). Longitudinal data analysis of continuous and discrete responses for pre-post designs. *Sankhya the Indian Journal of Statistics B*, 62, 134–148.
- LOD, F. M. (1967). A paradox in the interpretation of group comparisons. *Psychological Bulletin*, 68, 305–305.
- MATÉRN, B. (1986). *Spatial variation*. New York: Springer. ISBN: 0-387-96365-0.

- Matsui, M., & Takemura, A. (2006). Some improvements in numerical evaluation of symmetric stable density and its derivatives. *Communications in Statistics - Theory and Methods*, 35, 149–172. <https://doi.org/10.1080/03610920500439729>
- McCullagh, P. (2016). Two early contributions to the Ewens saga. *Statistical Science*, 31, 23–26.
- McCullagh, P. (2018) *Tensor methods in statistics* (2nd ed.). New York: Dover Publications Inc.. ISBN: 0-486823784.
- McCullagh, P., & Møller, J. (2006). The permanental process. *Advances in Applied Probability*, 38, 873–888.
- McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models* (2nd ed.) London: Chapman and Hall. ISBN: 0-412-31760-5.
- McCullagh, P., & Polson, N. (2018). Statistical sparsity. *Biometrika*, 105, 779–814.
- McCullagh, P., & Tresoldi, M. F. (2021). A likelihood analysis of quantile-matching transformations. *Biometrika*, 108, 247–251.
- Mead, R. (1988). *The design of experiments*. Cambridge: Cambridge University Press. ISBN: 0-521-28762-6.
- Montgomery, R. A., Rice, K. E., Stefanski, A., Rich, R. L., & Reich, P. B. (2020). Phenological responses of temperate and boreal trees to warming depend on ambient spring temperatures, leaf habit, and geographic range. *Proceedings of the National Academy of Sciences*, 117, 10397–10405. <https://doi.org/10.1073/pnas.1917508117>
- Nelder, J. A. (1965a). The analysis of randomised experiments with orthogonal block structure I. Block structure and the null analysis of variance. *Proceedings of the Royal Society of London. Series A*, 283, 147–162.
- Nelder, J. A. (1965b). The analysis of randomised experiments with orthogonal block structure II. Treatment structure and the general analysis of variance. *Proceedings of the Royal Society of London. Series A*, 283, 163–178.
- Orzack, S. H., Steiner, U. K., Tuljapurkar, S., & Thompson, P. (2011). Static and dynamic expression of life history traits in the northern fulmar *Fulmarus glacialis*. *Oikos*, 120, 369–380. <https://doi.org/10.1111/j.1600-0706.2010.17996.x>
- Patterson, H. D., & Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika*, 58, 545–554.
- Pearl, J., & Mackenzie, D. (2021). *The Book of Why*. Basic Books. ISBN: 978-1-5416-9896-3.
- Phillips, D. P., & Feldman, K. A. (1973). A dip in deaths before ceremonial occasions: some new relationships between social integration and mortality. *American Sociological Review*, 38, 678–696. <https://doi.org/10.2307/2094131>
- Pitman, J. W. (2006). *Combinatorial stochastic processes. Lecture Notes in Mathematics* (Vol. 1875). Berlin: Springer.
- R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. <http://www.R-project.org/>
- Reich, P. B., Sendall, K. M., Stefanski, A., Rich, R. L., Hobie, S. E., & Montgomery, R. A. (2018). Effects of climate warming on photosynthesis in boreal tree species depend on soil moisture. *Nature*, 263, 263–267.
- Robbins, H. (1956). An empirical Bayes approach to statistics. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Vol. I* (pp. 157–163). California: University of California Press.
- Samuels, M. L. (1986). The use of analysis of covariance in clinical trials: A clarification. *Controlled Clinical Trials*, 7, 325–329.
- Sen, M., & Bera, A. K. (2014). The improbable nature of the implied correlation matrix from spatial regression models. *Regional Statistics*, 4, 3–15. <https://doi.org/10.15196/RS04101>
- Senn, S. J. (2006). Change from baseline and analysis of covariance revisited. *Statistics in Medicine*, 25, 4334–4344. <https://doi.org/10.1002/sim.2682>
- Senn, S. J. (2019). Red herrings and the art of cause fishing: Lord's paradox revisited. *Error Statistics Philosophy*. <https://errorstatistics.com/2019/08/02/s-senn-red-herrings-and-the-art-of-cause-fishing-lords-paradox-revisited-guest-post/>

- Sharon, G., Segal, D., Ringo, J. M., Hefetz, A., Zilber-Rosenberg, I., & Rosenberg, E. (2010). Commensal bacteria play a role in mating preference of *Drosophila melanogaster*. *Proceedings of the National Academy of Sciences*, 107, 20051–20056. <https://doi.org/10.1073/pnas.1009906107>. Correction: (2013) *Proceedings of the National Academy of Sciences*, 110, 4853.
- Shi, P., & Tsai, C.-L. (2002). Regression model selection—A residual likelihood approach. *Journal of the Royal Statistical Society B*, 64, 237–252. Correction: 70, 1067.
- Stein, M. (1999). *Interpolation of spatial data: Some theory for kriging*. New York: Springer. ISBN: 0-387-98629-4.
- Student (1931). Agricultural field experiments. *Nature*, 127, 404–405.
- Sykulski, A. M., Olhede S. C., Guillaumin, A. P., Lilly, J. M., & Early, J. J. (2019). The debiased Whittle likelihood. *Biometrika*, 106, 251–266.
- Tan, C. K. W., Løvlie, H., Greenway, E., Goodwin, S. F., Pizzari, T., & Wigby, S. (2013). Sex-specific responses to sexual familiarity, and the role of olfaction in *Drosophila*. *Proceedings of the Royal Society B*, 280. <https://doi.org/10.1098/rspb.2013.1691>
- Tan, C. K. W., Løvlie, H., Greenway, E., Goodwin, S. F., Pizzari, T., & Wigby, S. (2013). Sex-specific responses to sexual familiarity, and the role of olfaction in *Drosophila*: A new analysis confirms original results. *Proceedings of the Royal Society B*, 281. <https://doi.org/10.1098/rspb.2014.0512>
- Tavaré, S. (2021). The magical Ewens sampling formula. *Bulletin of the London Mathematical Society*, 53, 1563–1582.
- Tukey, J. W. (1949). One degree of freedom for non-additivity. *Biometrics*, 5, 232–242.
- Villa, S. M., Altuna, J. C., Ruff, J. R., Beach, A., Mulvey, L. I., Poole, E. J., Campbell, H. E., Johnson, K. P., Shapiro, M. D., Bush, S. E., & Clayton, D. H. (2019). Rapid experimental evolution of reproductive isolation from a single natural population. *Proceedings of the National Academy of Sciences*, 116, 13440–13445.
- Wasserman, L. (2004). *All of statistics*. New York: Springer. ISBN: 0-387-40272-1.
- Welham, S. J., & Thompson, R. (1997). Likelihood ratio tests for fixed model terms using residual maximum likelihood. *Journal of the Royal Statistical Society B*, 59, 701–714.
- Whittle, P. (1953). Estimation and information in stationary time series. *Arkiv för Matematik*, 2, 423–434.
- Wilson, J. R., Kuehn, R. E., & Beach, F. A. (1963). Modification in the sexual behavior of male rats produced by changing the stimulus female. *Journal of Comparative and Physiological Psychology*, 56, 636–644.
- Wick, G. C. (1950). The evaluation of the collision matrix. *Physical Reviews*, 80, 268–272. <https://doi.org/10.1103/PhysRev.80.268>
- Yates, F. (1948). The analysis of contingency tables with groupings based on quantitative characters. *Biometrika*, 35, 176–181; corr. 35, 424.
- Zehnder, P. J., Weber, A., & Linder, A. (1951). Étude du rendement des scies par les méthodes statistiques. *Annales de l'Institut Fédéral de Recherches Forestières*, 27, 1–18.

Index

A

Accelerated-failure model, 238

Accessibility, 387

Additivity, 3, 178, 272

Adler, L., 136

Advection, 311

Age cohort, 148

Aitchison, J., 40

Alpha-stable

covariance, 97

distribution, 97, 111, 116

Altuna, J.C., 55

Analysis of variance, 104, 263, 266

Analytic sample path, 97

Ancillary statistic, 208, 393

Andrews, D.A., 38

Anscombe, F.J., 189

Armitage, P., 38

Arnold, S., 143

Assortative mating, 26, 30, 31

Atkinson, Q.D., 117

Austin, S., 218

Autocorrelation, 89

Average treatment effect, 399

B

Bailey, R.A., ix, 16, 63, 176

Baltagi, B.H., 199

Barber, N.A., 136

Barnard, G.A., 201

Bartlett correction factor, 334, 360, 361

Bartlett identities, 330

Bartlett, M.S., 76, 361

Baseline, 8, 10, 163, 385

Bates, D., ix

Bayes estimator, 94, 267

Beach, A., 55

Beach, F.A., 38

Belej, M., 199

Benjamini, Y., 343

Bera, A.K., 199

Berger, J., 202

Biased sampling, 156

BIC selection procedure, 209

Biller, O.M., 136

Birnbaum, A., 201

Bliss, C.I., 15, 366

Block

averages, 88

design, 137

exchangeability, 140, 224, 226

factor, 173

randomization, 226

BLUP, 48, 49, 94, 271

Bock, R.D., 219

Bolker, B., ix

Bootstrap, 191, 194, 195, 392

Boson process, 255

Boundary point, 249

Box, G.E.P., 47, 48, 53, 200, 205, 364

Box-Tidwell method, 47

Bradley-Terry model, 207

Brownian

bridge, 93, 100

covariance, 45, 51, 64, 73, 200

evolution, 64, 73

motion, 45, 51, 73, 101, 200, 260

Bush, S.E., 55

C

- Campbell, H.E., 55
 Carrick, R., 145
 Cauchy distribution, 229, 247
 Cavalli-Sforza, L.L., 64
 Cellmer, R., 199
 Cemetery state, 182
 Censoring, 154, 181
 Chinese restaurant process, 188, 194
 Choleski factorization, 52, 53, 100
 Chordal metric, 92
 Clayton, D.H., 55
 Clifford, D., viii
 Cochran's theorem, 257, 263
 Cochran, W.G., 38
 Coefficient of variation, 9
 Cohort plot, 148
 Commutative ring, 274
 Competition model, 207
 Compliance, 171, 180
 Compositional response, 40
 Conformity with randomization, 62, 63
 Confounding, 5
 Congruent samples, 225
 Consistency, self-, 161
 Contrast, 316
 Cooch, E.G., 156
 Coolidge effect, 38
 Corbet, A.S., 189
 Counterfactual, 166, 242, 397
 as missing value, 245
 Counterfactual sample, 166
 Covariate, 8, 16, 170, 393
 Cox, D.R., ix, 176, 364, 378
 Crane, H., 188
 Crossover design, 178
 Cumulant, 37, 85, 188
 -g.f., 194, 251, 262, 267
 Cycle of a permutation, 208

D

- Da Silva, P.A., 190
 Davison, A.C., 160
 Dawid, A.P., 242–244
 Dawson, R.B., 37
 Degrees of freedom, 3, 4, 11, 20–22, 31, 61, 69, 105, 123, 142, 273
 Denning, T., 34, 41
 Density factorization, 209
 1DOFNA, 272
 Dong, G., 198, 206
 Douthat, R., 218
 Drosophila

courtship, 27

- diet, 25
 refractory period, 27, 28
 Dunnet, G.M., 145
 Dynamic range, 3, 143
 Dyson, F.W., 267, 342

E

- Early, J.J., 115
 Eddington, A., 267, 342
 Eddington's formula, 51, 267, 277
 Edwards, A.W.F., 64
 Edwards, L.J., 368
 Effects
 covariate, 177
 treatment, 177
 Efron, B., 268, 343
 Eligibility, 387
 Entropy, 371
 Equi-variance, 228
 Ewens, W.J., 188
 sampling formula, 188, 193, 203, 208
 Exchangeability, 5, 10, 63, 73, 140, 141, 157, 177, 184, 200, 218, 223–225, 238, 394
 Exclusions, 7
 Exogenous variable, 172–174
 Experimental design, 178
 Experimental unit, 3, 10, 176
 E.U. vs. O.U., 139, 380
 Exponential family model, 188
 External variable, 172, 173
 Eynhallow, 145

F

- Factor
 block, 5, 10
 classification, 5, 6, 10
 coding, 381
 effects, 5
 subspace, 5
 treatment, 10
 Factorial model, 62, 63, 66, 74, 272
 Factorial subspace, 381
 False discovery rate, 343
 Famous Americans, 41
 Feature, 168
 Feldman, K.A., 41
 Feller, W., 111, 116
 Felsenstein, J., 64
 Feynman diagram, 255
 Fiducial process, 242, 272

Fieller confidence region, 337
 Fingleton, B., 199
 Finite population, 70, 88, 165, 184, 390
 Fisher information metric, 256
 Fisher, R.A., 84, 189, 201, 245
 Fitted value, 48
 Fowler, H., 38
 Fractional Brownian motion, 51, 52
 Fractional factorial design, 15
 Fractional linear transformation, 247
 F -ratio, 263

G

Garmez, N., 38
 Gaussian
 distribution, 254
 Hilbert space, 256
 moments, 255
 Gelman, A., 187
 Generalized linear mixed model (GLMM), 137
 Generalized linear model, viii, 107, 142, 161,
 224, 337, 376
 Goals of analysis, 17
 Goodness-of-fit, 203
 Goodwin, S.F., 39
 Gosset process, 241
 Gosset, W.S., 240, 391
 Graph, purpose of, 58, 381
 Greenway, E., 39
 Gregorian calendar, 82
 Group action, 111, 177, 231–237, 399
 Group representation, 307
 Guillaumin, A.P., 115
 Gurka, M., 368

H

Haldane-Dawson formula, 37, 41
 Haldane, J.B.S., 37
 Harris, R., 198, 206
 Hazard function, 152, 154
 Hazard measure, 181
 Hefetz, A., 25
 Hermitian symmetry, 255
 Herzberg, A., 38
 Heterogeneity, 183
 Hobbie, S.E., 133
 Hochberg, Y., 343
 Holst, L.T., 142
 Horvitz, D.G., 186
 Horvitz-Thompson estimator, 186, 390
 Hurst index, 52
 Hydrodynamic process, 305, 310, 324

Hydrodynamic symmetry, 306, 310, 323
 Hypergeometric distribution, 36

I

Idempotent matrix, 258
 Immortality, 181
 Inclusion probability, 165, 166, 168, 186, 390
 Incomplete process, 240
 Inconsistency, 198, 240
 Independence, 6, 27, 33, 160, 179, 257, 262
 Independent evolution, 173
 Inheritance for k -statistics, 85
 Interaction, 178, 233, 237, 356, 368, 382, 394,
 399
 Interference, 133, 179, 193, 242
 Intervention, 171
 Inverse polynomial, 43
 Irwin, R.E., 136
 Isomorphism, 275
 Isotropic process, 281, 282, 288, 306
 Isserlis, L., 255

J

Jamshidpey, A., 190
 Jeffreys, B.S., 309
 Jeffreys, H., 309
 Jenkins, G.M., 201
 Johnson, K.P., 55
 Johnstone, I., 342
 Jones, K., 198, 206

K

Kaplan-Meier estimator, 154
 Kerrich, J.E., 205
 Kimble, G.A., 38
 Kobylinska, K., 199
 Kolmogorov, A.N., 71
 Kolmogorov consistency, 205
 Krakatowa, 83
 Kriging, 94, 257, 271
 Kronecker symbol, 6
 k -statistic, 84
 Kuehn, R.E., 38
 Kupper, L.L., 368

L

Laplace approximation, 139
 Latin square, 15, 366
 Le Gallo, J., 199
 Lexis dispersion, 31, 142

- Liang, K.-Y., 212
 Likelihood
 factorization, 209
 ratio, 353, 355
 residual-, 351
 Lilly, J.M., 115
 Linder, A., 15
 Linear regression, 269
 Li, T., 142
 Locally finite population, 165
 Log normal distribution, 9
 Longitudinal design, 182
 Long-range dependence, 88, 91
 Lord, F.M., 219
 Lord's paradox, 219
 Løvlie, H., 39
- M**
 Mächler, M., ix
 Mackenzie, D., 243
 MARK, 156
 Mark-recapture design, 155, 164
 Matched pairs design, 395
 Matérn, B., 94
 Matérn covariance, 94, 285, 288
 Matrix exponential, 308
 Matsui, M., 111
 McCullagh, P., viii, 37, 61, 63, 143, 190, 255,
 268, 329, 338, 342, 343, 371
 Mead, R., ix
 Michelson, A., 142
 Mills's ratio, 252
 Missing values, 1, 8, 155, 245
 Mixture model, 339
 Model
 formula, 5, 11, 48, 63, 74, 138, 355, 377
 selection, 381
 specification, 376
 Møller, J., 255
 Moment generating function, 9
 Montgomery, R.A., 133, 142
 Muller, E., 368
 Multinomial model, 29
 Mulvey, L.I., 55
- N**
 Nelder, J.A., ix, 37, 63, 143, 178, 338
 Neutral evolution, 64, 188
- O**
 Observational study, 156
 Observational unit, 10, 28, 70, 134, 164, 386
 Olhede, S.C., 115
 Orzack, S.H., 145
 Out of Africa, 117
 Out of Ireland, 129
 Over-dispersion, 31, 39, 108, 142
- P**
 Patterson, H.D., 349
 Pearl, J., 243
 Pearson statistic, 31, 33, 37, 38
 Permanent of a matrix, 256
 Phillips, D.P., 41
 Pirotte, A., 199
 Pitman, J.W., 188
 Pizzari, T., 39
 Polson, N., 61, 268, 329, 342, 343, 347
 Poole, E.J., 55
 Population, 164
 average, 185, 399
 biological, 165
 locally finite, 165
 Post-baseline variable, 8
 Potential outcome, 242, 397
 Prediction, 48, 93, 166, 264
 Principal component, 66
 Probability weighting, 185
 Process, 159, 162
 counterfactual, 242
 Profile likelihood, 77
 Projection, 14, 22, 23, 76, 246, 258, 259
 Proportional-hazards model, 238
 Protocol, 10, 163, 169, 171, 175, 385
 Pseudo-replication, 179
p-value, 5, 7, 34, 203, 376, 377, 382
- Q**
 Quantile-matching transformation, 370
 Quaternion scalar product, 309, 325
- R**
 Random coefficient model, 234, 248
 Randomization, 10, 166, 171, 175, 392, 394
 Random matching, 36
 Random sample, 184
 Rao-Fisher-information, 256
 Reich, P.B., 133, 142
 Reid, N., 176
 Relationship, 10, 172, 393
 REML, 7, 12, 20, 349
 Replace *vs.* substitute, 378

- Replication, 137, 179
 Response transformation, 143
 Reversible process, 225
 Rice, K.E., 142
 Rich, R.L., 133, 142
 Ringo, J.M., 25
 Rinnan, R., 142
 Robbins, H., 268
 Rosenberg, E., 25
 Ruff, J.R., 55
- S**
- Sample
 accessibility, 184
 path, 96
 simple random-, 88, 166, 184
 stratified-, 184, 185
 Sampling consistency, 193, 194
 Sampling fraction, 88
 Samuels, M.L., 212
 Santoso, J., 128, 131
 Schur product theorem, 292
 Segal, D., 25
 Selective sampling, 160
 Self-adjoint transformation, 258, 259
 Semi-max coefficient, 43, 47
 Sendall, K.M., 133
 Sen, M., 199
 Senn, S.J., 212, 219
 Separable covariance, 297, 303, 321
 Sexual asymmetry, 27
 Sexual isolation index, 26
 Shapiro, M.D., 55
 Sharon, D., 25
 Shi, P., 368
 Significance test, 203
 Silverman, B.W., 342
 Simpson's paradox, 120
 Singular distribution, 253
 Size-biased sample, 184
 Snell, E.J., ix
 Space-time reversibility, 299
 Spatial autoregressive model, 224
 Speciation, 25, 55
 Species diversity model, 189
 Spline function, 46, 87, 96, 105, 271
 Stability over time, 58
 Standard error
 of contrasts, 18
 Stationarity, 93, 225, 281
 Statistical model, 162
 Stefanski, A., 133, 142
 Steiner, U.K., 145
- Stein, M., 286
 Stirling number, 208
 Student, 391
 Student-*t* distribution, 266
 Subset selection, 381
 Substitute vs. replace, 378
 Survival model, 237
 Survivor function, 181
 SUTDA, 180, 242
 Sykulski, A.M., 115
 Synergy, 178
- T**
- Table, purpose of, 58, 381
 Takemura, A., 111
 Tan, C.K.W., 39
 Tavaré, S., 188, 190
 Thompson, D.J., 186
 Thompson, P., 145
 Thompson, R., 20, 22, 349, 355, 358
 Tidwell, P.W., 47, 48, 53
 Time reversibility, 282
 Transformation, 3, 18, 143, 363
 Box-Cox, 364
 cube root, 372
 idempotent, 258
 image of, 258
 involution, 383
 isometric, 256
 Jacobian, 366, 370
 kernel of, 258
 nilpotent, 271
 quantile-matching, 370
 self-adjoint, 258
 Travelling wave, 303, 305
 Treatment
 assignment, 166, 171, 175, 383
 effect, 177, 231–237, 399
 factor, 16, 166, 231
 interference, 133, 179
 Trend estimation, 94
 Tresoldi, M.F., 371
 Tsai, C.-L., 368
 Tukey, J.W., 272
 Tuljapurkar, S., 145
 Tweedie's formula, 268
- U**
- Under-dispersion, 31, 35
 Use vs. usage, 378

V

- Variable
 external, 172
 qualitative, 169
 quantitative, 169
 response, 169
 types of, 135
Variogram, 89
Verrell, S., 143
Villa, S.M., 55, 75
von-Mises-Fisher distribution, 248

W

- Walker, S., ix
Weber, A., 15
Weibull distribution, 238
Welham, S.J., 20, 22, 355, 358
Whalen, R.E., 38
White, G., 156
Whittle likelihood, 114
Whittle, P., 114

Wick, G.C., 255

- Wick's theorem, 255
Wigby, S., 39
Williams, C.B., 189
Wilson-Hilferty transformation, 372
Wilson, J.R., 38
Winsten, C.B., 201
Wolpert, R.L., 202

Y

- Yates, F., 38
Yekutieli, D., 35
Yu, J., 198, 206

Z

- Zeger, S.L., 212
Zehnder, P.J., 15
Zigler, E., 38
Zilber-Rosenberg, I., 25