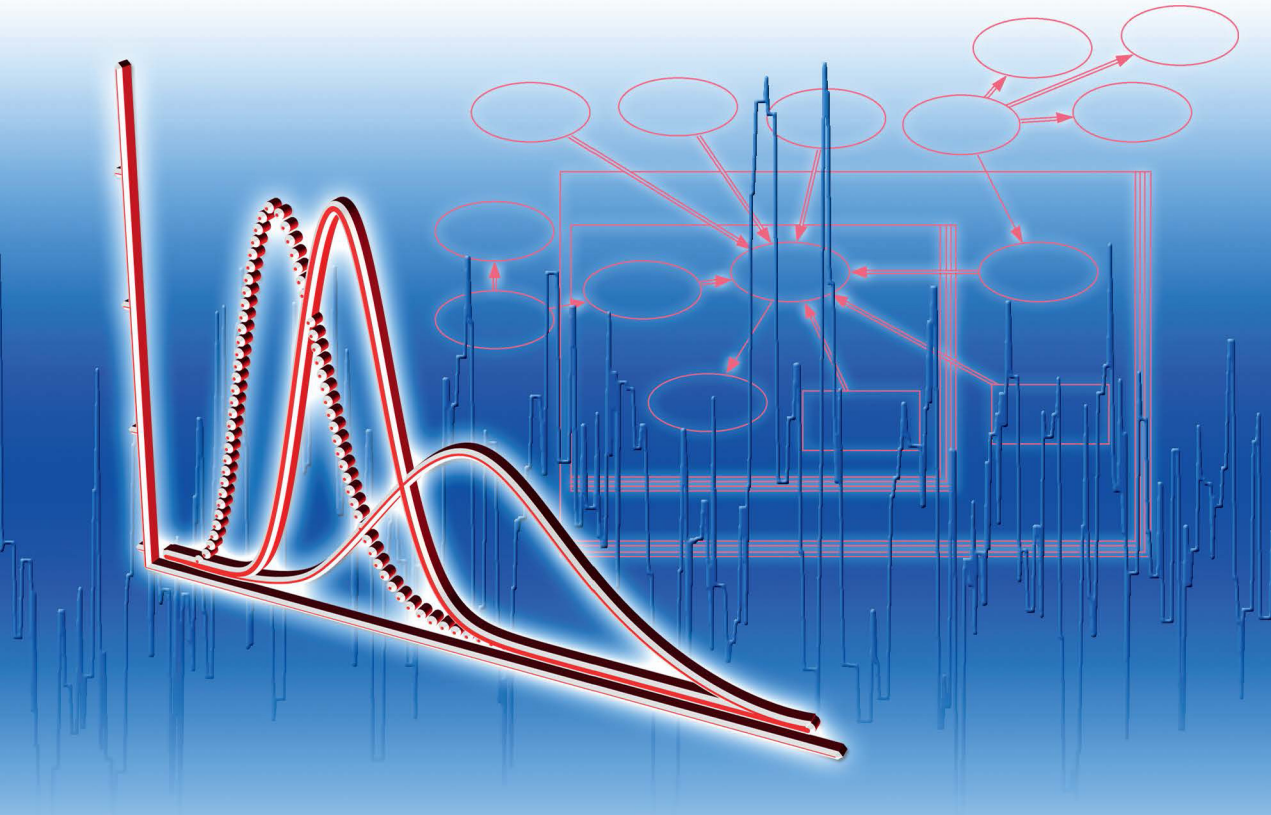
Bayesian Biostatistics



EMMANUEL LESAFFRE
and ANDREW B. LAWSON

WILEY

STATISTICS IN PRACTICE

Companion Website

Bayesian Biostatistics

Statistics in Practice

Series Advisors

Human and Biological Sciences
Stephen Senn
CRP-Santé, Luxembourg

Earth and Environmental Sciences
Marian Scott
University of Glasgow, UK

Industry, Commerce and Finance
Wolfgang Jank
University of Maryland, USA

Statistics in Practice is an important international series of texts which provide detailed coverage of statistical concepts, methods and worked case studies in specific fields of investigation and study.

With sound motivation and many worked practical examples, the books show in down-to-earth terms how to select and use an appropriate range of statistical techniques in a particular practical field within each title's special topic area.

The books provide statistical support for professionals and research workers across a range of employment fields and research environments. Subject areas covered include medicine and pharmaceutics; industry, finance and commerce; public services; the earth and environmental sciences, and so on.

The books also provide support to students studying statistical courses applied to the above areas. The demand for graduates to be equipped for the work environment has led to such courses becoming increasingly prevalent at universities and colleges.

It is our aim to present judiciously chosen and well-written workbooks to meet everyday practical needs. Feedback of views from readers will be most valuable to monitor the success of this aim.

A complete list of titles in this series can be found at
www.wiley.com/go/statisticsinpractice

Bayesian Biostatistics

Emmanuel Lesaffre

Erasmus MC, Rotterdam, The Netherlands and K.U. Leuven, Belgium

Andrew B. Lawson

Medical University of South Carolina, Charleston, USA

WILEY

A John Wiley & Sons, Ltd., Publication

This edition first published 2012

Registered office
John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, United Kingdom

Library of Congress Cataloging-in-Publication Data

Lesaffre, Emmanuel.
Bayesian biostatistics / Emmanuel Lesaffre, Andrew Lawson.
p. ; cm.
Includes bibliographical references and index.
ISBN 978-0-470-01823-1 (cloth)
1. Biometry–Methodology. 2. Bayesian statistical decision theory. I. Lawson, Andrew (Andrew B.) II. Title.
[DNLM: 1. Biostatistics–methods. 2. Bayes Theorem. QH 323.5]
QH323.5.L45 2012
570.1′5195–dc23

2012004237

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

ISBN: 978-0-470-01823-1

Typeset in 10/12pt Times by Aptara Inc., New Delhi, India

Contents

Preface **xiii**

Notation, terminology and some guidance for reading the book **xvii**

Part I BASIC CONCEPTS IN BAYESIAN METHODS

1 Modes of statistical inference 3
1.1 The frequentist approach: A critical reflection 4
1.1.1 The classical statistical approach 4
1.1.2 The *P*-value as a measure of evidence 5
1.1.3 The confidence interval as a measure of evidence 8
1.1.4 An historical note on the two frequentist paradigms* 8
1.2 Statistical inference based on the likelihood function 10
1.2.1 The likelihood function 10
1.2.2 The likelihood principles 11
1.3 The Bayesian approach: Some basic ideas 14
1.3.1 Introduction 14
1.3.2 Bayes theorem – discrete version for simple events 15
1.4 Outlook 18
Exercises 19

2 Bayes theorem: Computing the posterior distribution 20
2.1 Introduction 20
2.2 Bayes theorem – the binary version 20
2.3 Probability in a Bayesian context 21
2.4 Bayes theorem – the categorical version 22
2.5 Bayes theorem – the continuous version 23
2.6 The binomial case 24
2.7 The Gaussian case 30
2.8 The Poisson case 36
2.9 The prior and posterior distribution of $h(\theta)$ 40
2.10 Bayesian versus likelihood approach 40

2.11 Bayesian versus frequentist approach 41
2.12 The different modes of the Bayesian approach 41
2.13 An historical note on the Bayesian approach 42
2.14 Closing remarks 44
Exercises 44

3 Introduction to Bayesian inference 46
3.1 Introduction 46
3.2 Summarizing the posterior by probabilities 46
3.3 Posterior summary measures 47
3.3.1 Characterizing the location and variability of the posterior distribution 47
3.3.2 Posterior interval estimation 49
3.4 Predictive distributions 51
3.4.1 The frequentist approach to prediction 52
3.4.2 The Bayesian approach to prediction 53
3.4.3 Applications 54
3.5 Exchangeability 58
3.6 A normal approximation to the posterior 60
3.6.1 A Bayesian analysis based on a normal approximation to the likelihood 60
3.6.2 Asymptotic properties of the posterior distribution 62
3.7 Numerical techniques to determine the posterior 63
3.7.1 Numerical integration 63
3.7.2 Sampling from the posterior 65
3.7.3 Choice of posterior summary measures 72
3.8 Bayesian hypothesis testing 72
3.8.1 Inference based on credible intervals 72
3.8.2 The Bayes factor 74
3.8.3 Bayesian versus frequentist hypothesis testing 76
3.9 Closing remarks 78
Exercises 79

4 More than one parameter 82
4.1 Introduction 82
4.2 Joint versus marginal posterior inference 83
4.3 The normal distribution with μ and σ^2 unknown 83
4.3.1 No prior knowledge on μ and σ^2 is available 84
4.3.2 An historical study is available 86
4.3.3 Expert knowledge is available 88
4.4 Multivariate distributions 89
4.4.1 The multivariate normal and related distributions 89
4.4.2 The multinomial distribution 90
4.5 Frequentist properties of Bayesian inference 92
4.6 Sampling from the posterior distribution: The Method of Composition 93
4.7 Bayesian linear regression models 96
4.7.1 The frequentist approach to linear regression 96
4.7.2 A noninformative Bayesian linear regression model 97

4.7.3 Posterior summary measures for the linear regression model 98
4.7.4 Sampling from the posterior distribution 99
4.7.5 An informative Bayesian linear regression model 101
4.8 Bayesian generalized linear models 101
4.9 More complex regression models 102
4.10 Closing remarks 102
Exercises 102

5 Choosing the prior distribution 104
5.1 Introduction 104
5.2 The sequential use of Bayes theorem 104
5.3 Conjugate prior distributions 106
5.3.1 Univariate data distributions 106
5.3.2 Normal distribution – mean and variance unknown 109
5.3.3 Multivariate data distributions 110
5.3.4 Conditional conjugate and semiconjugate distributions 111
5.3.5 Hyperpriors 112
5.4 Noninformative prior distributions 113
5.4.1 Introduction 113
5.4.2 Expressing ignorance 114
5.4.3 General principles to choose noninformative priors 115
5.4.4 Improper prior distributions 119
5.4.5 Weak/vague priors 120
5.5 Informative prior distributions 121
5.5.1 Introduction 121
5.5.2 Data-based prior distributions 121
5.5.3 Elicitation of prior knowledge 122
5.5.4 Archetypal prior distributions 126
5.6 Prior distributions for regression models 129
5.6.1 Normal linear regression 129
5.6.2 Generalized linear models 131
5.6.3 Specification of priors in Bayesian software 134
5.7 Modeling priors 134
5.8 Other regression models 136
5.9 Closing remarks 136
Exercises 137

6 Markov chain Monte Carlo sampling 139
6.1 Introduction 139
6.2 The Gibbs sampler 140
6.2.1 The bivariate Gibbs sampler 140
6.2.2 The general Gibbs sampler 146
6.2.3 Remarks* 150
6.2.4 Review of Gibbs sampling approaches 152
6.2.5 The Slice sampler* 153
6.3 The Metropolis(–Hastings) algorithm 154
6.3.1 The Metropolis algorithm 155
6.3.2 The Metropolis–Hastings algorithm 157

6.3.3 Remarks* 159
6.3.4 Review of Metropolis(–Hastings) approaches 161
6.4 Justification of the MCMC approaches* 162
6.4.1 Properties of the MH algorithm 164
6.4.2 Properties of the Gibbs sampler 165
6.5 Choice of the sampler 165
6.6 The Reversible Jump MCMC algorithm* 168
6.7 Closing remarks 172
Exercises 173

7 Assessing and improving convergence of the Markov chain 175
7.1 Introduction 175
7.2 Assessing convergence of a Markov chain 176
7.2.1 Definition of convergence for a Markov chain 176
7.2.2 Checking convergence of the Markov chain 176
7.2.3 Graphical approaches to assess convergence 177
7.2.4 Formal diagnostic tests 180
7.2.5 Computing the Monte Carlo standard error 186
7.2.6 Practical experience with the formal diagnostic procedures 188
7.3 Accelerating convergence 189
7.3.1 Introduction 189
7.3.2 Acceleration techniques 189
7.4 Practical guidelines for assessing and accelerating convergence 194
7.5 Data augmentation 195
7.6 Closing remarks 200
Exercises 201

8 Software 202
8.1 WinBUGS and related software 202
8.1.1 A first analysis 203
8.1.2 Information on samplers 206
8.1.3 Assessing and accelerating convergence 207
8.1.4 Vector and matrix manipulations 208
8.1.5 Working in batch mode 210
8.1.6 Troubleshooting 212
8.1.7 Directed acyclic graphs 212
8.1.8 Add-on modules: GeoBUGS and PKBUGS 214
8.1.9 Related software 214
8.2 Bayesian analysis using SAS 215
8.2.1 Analysis using procedure GENMOD 215
8.2.2 Analysis using procedure MCMC 217
8.2.3 Other Bayesian programs 220
8.3 Additional Bayesian software and comparisons 221
8.3.1 Additional Bayesian software 221
8.3.2 Comparison of Bayesian software 222
8.4 Closing remarks 222
Exercises 223

Part II BAYESIAN TOOLS FOR STATISTICAL MODELING

9 Hierarchical models 227
9.1 Introduction 227
9.2 The Poisson-gamma hierarchical model 228
9.2.1 Introduction 228
9.2.2 Model specification 229
9.2.3 Posterior distributions 231
9.2.4 Estimating the parameters 232
9.2.5 Posterior predictive distributions 237
9.3 Full versus empirical Bayesian approach 238
9.4 Gaussian hierarchical models 240
9.4.1 Introduction 240
9.4.2 The Gaussian hierarchical model 240
9.4.3 Estimating the parameters 241
9.4.4 Posterior predictive distributions 243
9.4.5 Comparison of FB and EB approach 244
9.5 Mixed models 244
9.5.1 Introduction 244
9.5.2 The linear mixed model 244
9.5.3 The generalized linear mixed model 248
9.5.4 Nonlinear mixed models 253
9.5.5 Some further extensions 256
9.5.6 Estimation of the random effects and posterior predictive distributions 256
9.5.7 Choice of the level-2 variance prior 258
9.6 Propriety of the posterior 260
9.7 Assessing and accelerating convergence 261
9.8 Comparison of Bayesian and frequentist hierarchical models 263
9.8.1 Estimating the level-2 variance 263
9.8.2 ML and REML estimates compared with Bayesian estimates 264
9.9 Closing remarks 265
Exercises 265

10 Model building and assessment 267
10.1 Introduction 267
10.2 Measures for model selection 268
10.2.1 The Bayes factor 268
10.2.2 Information theoretic measures for model selection 274
10.2.3 Model selection based on predictive loss functions 286
10.3 Model checking 288
10.3.1 Introduction 288
10.3.2 Model-checking procedures 289
10.3.3 Sensitivity analysis 295
10.3.4 Posterior predictive checks 300
10.3.5 Model expansion 308
10.4 Closing remarks 316
Exercises 316

11 Variable selection 319
11.1 Introduction 319
11.2 Classical variable selection 320
11.2.1 Variable selection techniques 320
11.2.2 Frequentist regularization 322
11.3 Bayesian variable selection: Concepts and questions 325
11.4 Introduction to Bayesian variable selection 326
11.4.1 Variable selection for K small 326
11.4.2 Variable selection for K large 330
11.5 Variable selection based on Zellner's g-prior 333
11.6 Variable selection based on Reversible Jump Markov chain Monte Carlo 336
11.7 Spike and slab priors 339
11.7.1 Stochastic Search Variable Selection 340
11.7.2 Gibbs Variable Selection 343
11.7.3 Dependent variable selection using SSVS 345
11.8 Bayesian regularization 345
11.8.1 Bayesian LASSO regression 346
11.8.2 Elastic Net and further extensions of the Bayesian LASSO 350
11.9 The many regressors case 351
11.10 Bayesian model selection 355
11.11 Bayesian model averaging 357
11.12 Closing remarks 359
Exercises 360

Part III BAYESIAN METHODS IN PRACTICAL APPLICATIONS

12 Bioassay 365
12.1 Bioassay essentials 365
12.1.1 Cell assays 365
12.1.2 Animal assays 366
12.2 A generic *in vitro* example 369
12.3 Ames/Salmonella mutagenic assay 371
12.4 Mouse lymphoma assay (L5178Y TK+/−) 373
12.5 Closing remarks 374

13 Measurement error 375
13.1 Continuous measurement error 375
13.1.1 Measurement error in a variable 375
13.1.2 Two types of measurement error on the predictor in linear and nonlinear models 376
13.1.3 Accommodation of predictor measurement error 378
13.1.4 Nonadditive errors and other extensions 382
13.2 Discrete measurement error 382
13.2.1 Sources of misclassification 382
13.2.2 Misclassification in the binary predictor 383
13.2.3 Misclassification in a binary response 386
13.3 Closing remarks 389

14 Survival analysis 390
14.1 Basic terminology 390
14.1.1 Endpoint distributions 391
14.1.2 Censoring 392
14.1.3 Random effect specification 393
14.1.4 A general hazard model 393
14.1.5 Proportional hazards 394
14.1.6 The Cox model with random effects 394
14.2 The Bayesian model formulation 394
14.2.1 A Weibull survival model 395
14.2.2 A Bayesian AFT model 397
14.3 Examples 397
14.3.1 The gastric cancer study 397
14.3.2 Prostate cancer in Louisiana: A spatial AFT model 401
14.4 Closing remarks 406

15 Longitudinal analysis 407
15.1 Fixed time periods 407
15.1.1 Introduction 407
15.1.2 A classical growth-curve example 408
15.1.3 Alternate data models 414
15.2 Random event times 417
15.3 Dealing with missing data 420
15.3.1 Introduction 420
15.3.2 Response missingness 421
15.3.3 Missingness mechanisms 422
15.3.4 Bayesian considerations 424
15.3.5 Predictor missingness 424
15.4 Joint modeling of longitudinal and survival responses 424
15.4.1 Introduction 424
15.4.2 An example 425
15.5 Closing remarks 429

16 Spatial applications: Disease mapping and image analysis 430
16.1 Introduction 430
16.2 Disease mapping 430
16.2.1 Some general spatial epidemiological issues 431
16.2.2 Some spatial statistical issues 433
16.2.3 Count data models 433
16.2.4 A special application area: Disease mapping/risk estimation 434
16.2.5 A special application area: Disease clustering 438
16.2.6 A special application area: Ecological analysis 443
16.3 Image analysis 444
16.3.1 fMRI modeling 446
16.3.2 A note on software 455

17 Final chapter 456
17.1 What this book covered 456

17.2 Additional Bayesian developments 456
17.2.1 Medical decision making 456
17.2.2 Clinical trials 457
17.2.3 Bayesian networks 457
17.2.4 Bioinformatics 458
17.2.5 Missing data 458
17.2.6 Mixture models 458
17.2.7 Nonparametric Bayesian methods 459
17.3 Alternative reading 459

Appendix: Distributions 460

A.1 Introduction 460
A.2 Continuous univariate distributions 461
A.3 Discrete univariate distributions 477
A.4 Multivariate distributions 481

References 484

Index 509

Preface

The growth of biostatistics as a subject has been phenomenal in recent years and has been marked by a considerable technical innovation in methodology and computational practicality. One area that has a significant growth is the class of Bayesian methods. This growth has taken place partly because a growing number of practitioners value the Bayesian paradigm as matching that of scientific discovery. But the computational advances in the last decade that have allowed for more complex models to be fitted routinely to realistic data sets have also led to this growth.

In this book, we explore the Bayesian approach via a great variety of medical application areas. In effect, from the elementary concepts to the more advanced modeling exercises, the Bayesian tools will be exemplified using a diversity of applications taken from epidemiology, exploratory clinical studies, health promotion studies and clinical trials.

This book grew out of a course that the first author has taught for many years (especially) in the Master programs in (bio)statistics at the universities of Hasselt and Leuven, both in Belgium. The course material was the inspiration for two out of three parts in the book. Therefore, the intended readership of this book are Master program students in (bio)statistics, but we hope that applied researchers with a good statistical background will also find the book useful. The structure of the book allows it to be used as course material for a course in Bayesian methods at an undergraduate or early stage postgraduate level. The aim of the book is to introduce the reader smoothly into Bayesian statistical methods with chapters that gradually increase in the level of complexity. The book consists of three parts. The first two parts of this work were the chapters primarily covered by the first author, while the last five chapters were primarily covered by the second author.

In Part I, we first review the fundamental concepts of the significance test and the associated P-value and note that frequentist methods, although proved to be quite useful over many years, are not without conceptual flaws. We also note that there are other methods on the market, such as the likelihood approach, but more importantly, the Bayesian approach. In addition, we introduce (an embryonic version of) the Bayes theorem. In Chapter 2, we derive the general expression of Bayes theorem and illustrate extensively the analytical computations to arrive at the posterior distribution on the binomial, the Gaussian and the Poisson case. For this, simple textbook examples are used. In Chapter 3, the reader is introduced to various posterior summary measures and the predictive distributions. Since sampling is fundamental to contemporary Bayesian approaches, sampling algorithms are introduced and exemplified in this chapter. While these sampling procedures will not yet prove their usefulness in practice,

we believe that the early introduction of relatively simple sampling techniques will prepare the reader better for the advanced algorithms seen in later chapters. In this chapter, approaches to Bayesian hypothesis testing are treated and we introduce the Bayes factor. In Chapter 4, we extend all the concepts and computations seen in the first three chapters for univariate problems to the multivariate case, introducing also Bayesian regression models. It is then seen that, in general, no analytical methods to derive the posterior distribution are available, neither are the classical numerical approaches to integration sufficient. A new approach is then needed. Before addressing the solution to the problem, we treat in Chapter 5 the choice of the prior distribution. The prior distribution is the keystone to the Bayesian methodology. Yet, the appropriate choice of the prior distribution has been the topic of extensive discussions between non-Bayesians and Bayesians, but also among Bayesians. In this chapter, we extensively treat the various ways of specifying prior knowledge. In Chapters 6 and 7, we treat the basics of the Markov chain Monte Carlo methodologies. In Chapter 6, the Gibbs and the Metropolis–Hastings samplers are introduced again illustrated using a variety of medical examples. Chapter 7 is devoted to assessing and accelerating the convergence of the Markov chains. In addition, we cover the extension of the EM-algorithm to the Bayesian context, i.e. the data augmentation approach is exemplified. It is then time to see how Bayesian analyses can be done in practice. For this reason, we review in Chapter 8 the Bayesian software. We focus on two software packages: the most popular WinBUGS and the recently released Bayesian SAS® procedures. In both cases, a simple regression analysis serves as a guiding example helping the readers in their first analyzes with these packages. We end this chapter with a review of other Bayesian software, such as the packages OpenBUGS and JAGS, and also various R packages written to perform specific analyses.

In Part II, we develop Bayesian tools for statistical modeling. We start in Chapter 9 with reviewing hierarchical models. To fix ideas, we focus first on two simple two-level hierarchical models. The first is the Poisson-gamma model applied to a spatial data set on lip cancer cases in former East Germany. This example serves to introduce the concepts of hierarchical modeling. Then, we turn to the Gaussian hierarchical model as an introduction to the more general mixed models. A variety of mixed models are explored and amply exemplified. Also in this chapter comparisons between frequentist and Bayesian solutions aim to help the reader to see the differences between the two approaches. Model building and assessment are the topics of Chapter 10. The aim of this chapter is to see how statistical modeling could be performed entirely within the Bayesian framework. To this end, we reintroduce the Bayes factor (and its variants) to select between two statistical models. The Bayes factor is an important tool in model selection but is also fraught with serious computational difficulties. We then move to the Deviance Information Criterion (DIC). For a better understanding of this popular model selection criterion, we introduce at length the classical model selection criteria, i.e. AIC and BIC and relate them to DIC. The part on model checking describes all classical actions one would take to construct and evaluate a model, such as checking the residuals for outliers and influential observations, finding the correct scale of the response and the covariates, choosing the correct link function, etc. In this chapter, we also elaborate on the posterior predictive check as a general tool for goodness-of-fit testing. The final chapter, Chapter 11, in Part II handles Bayesian variable selection. This is a rapidly evolving topic that received a great impetus from the developments in bioinformatics. A broad overview of possible variable and model selection approaches and the associated software is given. While in the previous two chapters, the WinBUGS software and to a lesser extent the Bayesian SAS procedures were dominant, in this chapter we focus on software packages in R.

In Part III, we address particular application areas for Bayesian modeling. We include the most important areas from a practical biostatistical standpoint. In Chapter 12, we examine bioassay, where we consider preclinical testing methods: both Ames and Mouse Lymphoma *in vitro* assays and the famous Beetles LD50 toxicity assay are considered. In Chapter 13, we consider the important and pervasive problem of measurement error and also the misclassification in biostatistical studies. We discuss Berkson and classical joint models, bias such as attenuation and the problem of discrete error in the form of misclassification. In Chapter 14, we examine survival analysis from a Bayesian perspective. In this chapter, we cover basic survival time models and risk-set-based approaches and extend models to consider contextual effects within hazards. In Chapter 15, longitudinal analysis is considered in greater depth. Correlated prior distributions for parameters and also temporally correlated errors are considered. Missingness mechanisms are discussed and a nonrandom missingness example is explored. In Chapter 16, two important spatial biostatistical application areas are then considered: disease mapping and image analysis. In disease mapping, basic Poisson convolution models that include spatially structured random effects are examined for risk estimation, while in image analysis a focus on Bayesian fMRI analysis with correlated prior distributions is presented.

In Chapter 17, we end the book with a brief review of the topics that we did not cover in this book and give some key references for further reading. Finally, in the appendix, we provide an overview of the characteristics of most popular distributions used in Bayesian analyses.

Throughout the book there are numerous examples. In Parts I and II, explicit reference is made to the programs associated with the examples. These programs can be found at the website `www.wiley.com/go/bayesian_methods_biostatistics`. The programs used in Part III can also be found at this website.

Acknowledgments

When writing the early chapters, the first author benefitted from discussions with master students at Leuven, Hasselt and Leiden who pointed out various typos and ambiguities in earlier versions of the book. In addition, thanks go to the colleagues and former/current PhD students at L-Biostat at KU Leuven and at the Department of Biostatistics, Erasmus MC, Rotterdam, for illustrative discussions, critical remarks and help with software. In this respect, special thanks go to Susan Bryan, Silvia Cecere, Luwis Diya, Alejandro Jara, Arnošt Komárek, Marek Molas, Mukendi Mbuyi, Timothy Mutsvari, Veronika Rockova, Robin Van Oirbeek and Sten Willemsen. Software support was received from Sabanés Bové on the R package `glmBfp`, Fang Chen on the Bayesian SAS programs, Robert Gramacy on the R program `monomvn`, David Hastie and Peter Green on an R program for RJMCMC, David Lunn on the Jump interface in WinBUGS, Elizabeth Slate and Karen Vines. Thanks also go to those who provided data for the book or who gave permission to use their data, i.e. Steven Boonen, Elly Den Hondt, Jolanda Luime, the Signal-Tandmobiel® team, Bako Topal and Vincent van Weel. The authors also wish to thank, for interesting discussions, their colleagues in Leuven, Rotterdam and Charleston especially Dipankar Bandyopadhyay, Paul Eilers, Steffen Fieuws, Mulugeta Gebregziabher, Dimitris Rizopoulos and Elizabeth Slate. Finally, insightful conversations with George Casella, James Hodges, Helmut Küchenhoff and Paul Schmitt were much appreciated.

Last, the first author especially wishes to thank his wife Lieve Sels for her patience not only during the preparation of the ‘book’ but also during his whole professional career. To his children, Annemie and Kristof, he apologizes for the many times that their father was present but ‘absent’.

Emmanuel Lesaffre (Rotterdam and Leuven)
December 2011 Andrew B. Lawson (Charleston)

Notation, terminology and some guidance for reading the book

Notation and terminology

In this section, we review some notation used in the book. We limit ourselves to outline some general principles; for precise definitions, we refer to the text.

First, both the random variable and its realization are denoted in this book as y. The vector of covariates is most often denoted as $\boldsymbol{x}$. The distribution of a discrete y as well as the density of a continuous variable will be denoted as $p(y)$, unless otherwise stated. In the text, we make clear which of the two meanings applies. A sample of observations $y_1, \ldots, y_n$ is denoted as $\boldsymbol{y}$, but also, a d-dimensional vector will be denoted in bold, i.e. $\boldsymbol{y}$. We make it clear from the context what is implied. Further, independent, identically distributed random variables are indicated as i.i.d. A distribution (density) depending on a parameter vector $\boldsymbol{\theta}$ is denoted as $p(y \mid \boldsymbol{\theta})$. The joint distribution of y and z is denoted as $p(y, z \mid \boldsymbol{\theta})$ and the conditional distribution of y, given z will be denoted as $y \mid z, \boldsymbol{\theta} \sim p(y \mid z, \boldsymbol{\theta})$. Alternatively, we use the notation $p(y \mid z, \boldsymbol{\theta})$. The probability that an event happens will occasionally be denoted as P for reasons of clarity.

A particular distribution will be addressed in two ways. For example, $y \sim \text{Gamma}(\alpha, \beta)$ indicates that the random variable y has a gamma distribution with parameters α and β, but to indicate that the distribution is evaluated in y we will use the notation $\text{Gamma}(y \mid \alpha, \beta)$. When parameters have been given almost the same notation, say β_0, β_1, β_3, then the notation β_* is used where * is a place holder for 1, 2, 3.

In the normal case, some Bayesian textbooks use the *precision* notation while others use the *variance* notation. More specifically, if y has a normal distribution with mean μ and variance σ^2, then instead of $y \sim \text{N}(\mu, \sigma^2)$ (classical notation) the alternative notation $y \sim \text{N}(\mu, \tau^{-1})$ (or even $y \sim \text{N}(\mu, \tau)$) with the precision $\tau = \sigma^{-2}$ is used by some. This alternative notation is inspired by the fact that in Bayesian statistics some key results are better expressed in terms of the precision rather than the variance. This is also the notation used by WinBUGS. In this book, we frequently refer to classical, frequentist statistics. The use of precision would then be more confusing, rather than illuminating. In addition, when it comes to summarizing the results of a statistical analysis, the standard deviation is a better tool than the precision. For these reasons, we primarily used in this book the variance notation. But, throughout the

book (especially in the later chapters) we regularly make the transition from one notation to the other.

Finally, we use some generally accepted standard notations, such as $\boldsymbol{x}$ always denotes a column vector, $|A|$ denotes the determinant of a matrix A, tr(A) is the trace of a matrix and A^T denotes the transpose of a matrix A. The sample mean of $\{y_1, y_2, \ldots, y_n\}$ is denoted as $\overline{y}$ and their standard deviation as s, s_y or simply SD.

Guidance for reading the book

No particular guidance is needed to read this book. The flow in the book is natural: starting from elementary Bayesian concepts we gradually introduce the more complex topics. This book deals, basically, only with parametric Bayesian methods. This means that all our random variables are assumed to have a particular distribution with a finite number of parameters. In the Bayesian world, many more distributions are used than in classical statistics. So for the classical reader (whatever meaning this may have), many of the distributions that pop-up in the book will be new. A brief characterization of these distributions, together with a graphical display, can be found in the appendix of the book.

Finally, some of the sections indicated by '*' are technical and may be skipped at first reading.

Part I

BASIC CONCEPTS IN BAYESIAN METHODS

1

Modes of statistical inference

The central activity in statistics is inference. Statistical inference is a procedure or a collection of activities with the aim to extract information from (gathered) data and to generalize the observed results beyond the data at hand, say to a population or to the future. In this way, statistical inference may help the researchers in suggesting or verifying scientific hypotheses, or decision makers in improving their decisions. Inference obviously depends on the collected data and on the assumed underlying probabilistic model that generated these data, but it also depends on the approach to generalize from the known (data) to the unknown (population). We distinguish two mainstream views/paradigms to draw statistical inference, i.e. the frequentist approach and the Bayesian approach. In-between these two paradigms is the (pure) likelihood approach.

In most of the empirical research, but definitely in medical research, scientific conclusions need to be supported by a 'significant result' using the classical *P*-value. Significance testing belongs to the frequentist paradigm. However, the frequentist approach does not consist of one unifying theory but is rather the combination of two approaches, i.e. the inductive approach of Fisher who introduced the null-hypothesis, the *P*-value and the significance level and the deductive procedure of Neyman and Pearson who introduced the alternative hypothesis and the notion of power. First, we review the practice of frequentist significance testing and focus on the popular *P*-value. More specifically we look at the value of the *P*-value in practice. Second, we treat an approach that is purely based on the likelihood function not involving any classical significance testing. This approach is based on two fundamental likelihood principles that are also essential for the Bayesian philosophy. Finally, we end this chapter by introducing the principles of the Bayesian approach and we give an outlook of what the Bayesian approach can bring to the statistician. However, at least three more chapters will be needed to fully develop the Bayesian theory.

Bayesian Biostatistics, First Edition. Emmanuel Lesaffre and Andrew B. Lawson.

1.1 The frequentist approach: A critical reflection

1.1.1 The classical statistical approach

It is perhaps an oversimplification to speak of a classical statistical approach. Nevertheless, we mean by this the ensemble of methods that provides statistical inference based on the classical *P*-value, the significance level, the power and the confidence interval (CI). To fix ideas, we now exemplify current statistical practice with a *randomized controlled clinical trial* (*RCT*). In fact, the RCT is the study design that, by excellence, is based on the classical statistical tool box of inferential procedures. We assume that the reader is familiar with the classical concepts in inferential statistics.

For those who have never experienced a RCT, here is a brief description. A clinical trial is an experimental study comparing two (or more) medical treatments on human subjects, most often patients. When a control group is involved, the trial is called *controlled*. For a *parallel* group design, one group of patients receives one treatment and the other group(s) receive(s) the other treatment and all groups are followed up in time to measure the effect of the treatments. In a *randomized study*, patients are assigned to the treatments in a random manner. To minimize bias in evaluating the effect of the treatments, patients and/or care givers are blinded. When only patients are blinded one speaks of a *single-blinded* study, but when both patients and care givers are blinded (and everyone involved in running the trial) one speaks of a *double-blinded* trial. Finally, when more than one center (e.g. hospital) is involved one deals with a multicenter study.

Example I.1: Toenail RCT: Evaluation of two oral treatments for toenail infection using the frequentist approach

A randomized, double-blind, parallel group, multicenter study was set up to compare the efficacy of two oral treatments for toenail infection (De Backer *et al*. 1996). In this study, two groups of 189 patients were recruited, and each received 12 weeks of treatment (Lamisil: treatment *A* and Itraconazol: treatment *B*), with 48 weeks of follow-up (FU). The significance level was set at $\alpha = 0.05$. The *primary endpoint* (upon which the sample size was based) in the original study was negative mycology, i.e. a negative microscopy and a negative culture. Here, we look at another endpoint, i.e. *unaffected nail length at week 48* on a subset of patients for whom the big toenail was the target nail. One hundred and thirty-one patients treated with *A* and 133 treated with *B* were included in this comparison. Note that we only included those patients present at the end of the study. The observed mean (SD) lengths in millimeter at week 48 were 9.07 (4.92) and 7.70 (5.33), for treatments *A* and *B*, respectively. Suppose the (population) average for treatment *A* is μ_1 while for treatment *B* it is μ_2. Therefore, the null-hypothesis is $H_0 : \Delta = \mu_1 - \mu_2 = 0$ and can be evaluated with an unpaired t-test at a two-sided significance level of $\alpha = 0.05$. Upon completion of the study, the treatment estimate was $\widehat{\Delta} = 1.38$ with an observed value of the t-statistic equal to $t_{obs} = 2.19$. This result lies in the rejection region corresponding to $\alpha = 0.05$ yielding a statistically significant result (at 0.05). Thus, according to the Neyman–Pearson (NP) approach we (can) reject that *A* and *B* are equally effective.

It is common to report also the *P*-value of the result to indicate the strength of evidence against the hypothesis of two equally effective treatments. Here, we obtained a two-sided *P*-value equal to 0.030, which is a measure of evidence against H_0 in a Fisherian sense. □

In Section 1.1.2, we reflect on what message the P-value can bring to the researcher. We will also indicate what properties the P-value does not have (but assumed to have).

1.1.2 The P-value as a measure of evidence

Fisher developed the P-value in the context of well-planned limited agricultural experiments in a time when computations had to be done by hand. Nowadays, a great variety of studies are undertaken in the medical research usually of an exploratory nature and often evaluating hundreds to thousands of P-values. The P-value is an intuitively appealing measure against the null-hypothesis, but that it is not always perceived in the correct manner. Here, we will further elaborate on the use and misuse of the P-value in practice.

The P-value is not the probability that H_0 is (not) true A common error is to interpret the P-value as a probability that H_0 is (not) true. However, the P-value only measures the extremeness of the observed result under H_0. The probability that H_0 is true is formally $p(H_0 \mid \text{data})$, which we shall call the posterior probability of the null-hypothesis in the following text, given the observed data. This probability is based on Bayes theorem and depends on the prevalence of H_0 (see also Example I.11).

The P-value depends on fictive data The P-value does not express the probability that the observed result occurred under H_0, but is rather the probability of observing this or a more extreme result under H_0. This implies that the calculation of the P-value is based not only on the observed result but also on fictive (never observed) data.

Example I.2: Toenail RCT: Meaning of P-value
The P-value is equal to the probability that the test statistic exceeds the observed value if the null-hypothesis were true. The computation of the P-value is done using probability laws, but could also be represented by a simulation experiment. For instance, in the toenail infection study the P-value is approximately equal to the proportion of studies, out of (say) 10 000 imaginary studies done under H_0 (two identical treatments), that yield a t-value more extreme than $t_{\text{obs}} = 2.19$. In Figure 1.1, the histogram of imaginary results is displayed together with the observed result. □

The P-value depends on the sample space The above simulation exercise shows that the P-value is computed as a probability using the *long-run frequency definition*, which means that a probability for an event A is defined as the ultimate proportion of experiments that generated that event to the total number of experiments. For a P-value, the event A corresponds to a t-value located in the rejection region. This makes it clear that the P-value depends on the choice of the fictive studies and, hence, also on the sample space. The particular choice can have surprising effects, as illustrated in Example I.3.

Example I.3: Accounting for interim analyses in a RCT
Suppose that a randomized controlled trial has been set up to compare two treatments and that four *interim analyses for efficacy* were planned. An interim analysis for efficacy is a statistical comparison between the treatment groups prior to the end of the study to see

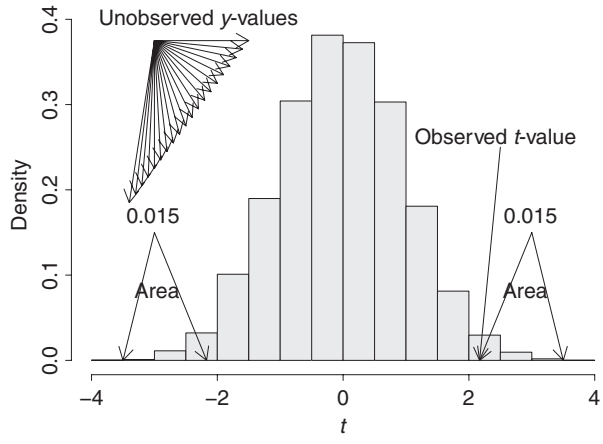


Figure 1.1 Graphical representation of the *P*-value by means of a simulation study under H_0.

whether the experimental treatment is better than the control treatment. The purpose is to stop the trial earlier if possible. When more than one comparison is planned, one needs to correct in a frequentist approach for *multiple testing*. A classical correction for multiple testing is *Bonferroni's rule* which dictates that the significance level at each comparison (*nominal significance level*) needs to be made more stringent, i.e. α/k where k is the total number of tests applied, to arrive at an overall (across all comparisons) type I error rate less or equal to α. Bonferroni's procedure is approximate; an exact control of the type I error rate is obtained with a *group sequential design* (Jennison and Turnbull 2000). With Pocock's group sequential method and 5 analyses (4 interim + 1 final analyses), a significance level of 0.016 is handled at each analysis to achieve a global significance level of 0.05. Thus, when the study ran until the end and at the last analysis a *P*-value of 0.02 was obtained, then the result cannot be claimed significant with Pocock's rule. However, if the same result had been obtained without planning interim analyses, then this trial produced a significant result! Thus, in the presence of two identical results, one cannot claim evidence against the null-hypothesis in one case, while in the other case we would conclude that the two treatments have a different effect. □

In statistical terminology, the different evidence for the treatment effect in the two RCTs of Example I.3 (with an identical *P*-value) is due to a different sample space (see below) in the two scenarios. This is further illustrated in Example I.4.

Example I.4: Kaldor' *et al.* case-control study: Illustration of sample space
In the (matched) case-control study (Kaldor *et al*. 1990) involving 149 cases (leukaemia patients) and 411 controls, the purpose was to examine the impact of chemotherapy on leukaemia in Hodgkin's survivors (Ashby *et al*. 1993). The 5-year survival of Hodgkin's disease (cancer of the lymph nodes) is about 80%, but the survivors have an excess risk of developing solid tumors, leukaemia and/or lymphomas. In Table 1.1, the cases and controls are subdivided according to exposure to chemotherapy or not.

Table 1.1 Kaldor' *et al*. case-control study (Kaldor *et al*. 1990): frequency table of cases and controls subdivided according to their exposure to chemotherapy.

Treatment	Controls	Cases
No chemo	160	11
Chemo	251	138
Total	411	149

Ignoring the matched character of the data, the analysis of the 2×2-contingency table by a Pearson chi-squared test results in $P = 7.8959 \times 10^{-13}$ with a chi-squared value of 51.3. With the Fisher's exact test, a P-value of 1.487×10^{-14} was obtained. Finally, the estimated odds ratio is equal to 7.9971 with a 95% CI of [4.19, 15.25]. □

The chi-squared test and the Fisher's exact test have a different *sample space*, which is the space of possible samples considered to calculate the null distribution of the test statistic. The sample space for the chi-squared test consists of the 2×2-contingency tables with the same total sample size (n), while for Fisher's exact test the sample space consists of the subset of 2×2-contingency tables with the same row and column marginal totals. The difference between the two sample spaces explains here partly the difference in the two test results. In Example I.3, it is the sole reason for the different evidence from the two RCTs. The conclusion of a scientific experiment, hence, not only depends on the results of that experiment but also on the results of experiments that did not and will never happen. This finding has triggered a lot of debate among statisticians (see Royall 1997).

The *P*-value is not an absolute measure A small P-value does not necessarily imply a large difference between two treatments or a strong association among variables. Indeed, as a measure of evidence the P-value does not take the size of the study into account. There have been vivid discussions on how a small P-value should be interpreted as a function of the size of the study (see Royall 1997).

The *P*-value does not take all evidence into account Let us take the following example also discussed by Ashby *et al*. (1993).

Example I.5: Merseyside registry results
Ashby *et al*. (1993) reported on data obtained from a subsequent registry study in UK (after Kaldor *et al*.'s case-control study) to check the relationship between chemotherapy and leukemia among Hodgkin's survivors. Preliminary results of the Merseyside registry were reported in Ashby *et al*. (1993) and are reproduced in Table 1.2. The P-value obtained from the chi-squared test with continuity correction equals 0.67. Thus, formally there is no reason to worry that chemotherapy may cause leukemia among Hodgkin's survivors. Of course, every epidemiologist would recognize that this study has no chance of finding a relationship between chemotherapy and leukemia because of the small study size. By simply analyzing the data of the Merseyside registry, no evidence of a relationship can be established. □

Table 1.2 Merseyside registry: frequency table of cases and controls subdivided according to their exposure to chemotherapy.

Treatment	Controls	Cases
No chemo	3	0
Chemo	3	2
Total	6	2

Is it reasonable to analyze the results of the Merseyside registry in isolation, not referring to the previous study of Kaldor *et al*. (1990)? In other words, should one forget about the historical data and assume that one cannot learn anything from the past? The answer will depend on the particular circumstances, but it is not obvious that the past should never play a role in the analysis of data.

1.1.3 The confidence interval as a measure of evidence

While the *P*-value has been criticized by many statisticians, it is more the (mis)use of the *P*-value that is under fire. Nevertheless, there is a growing preference to replace the *P*-value by the (95%) CI.

Example I.6: Toenail RCT: Illustration of 95% confidence interval
The 95% CI for Δ is equal to [0.14, 2.62]. Technically speaking we can only say that in the long run 95% of those intervals will contain the true parameter (the 95% CI is based on the long-run frequency definition of probability). But for our RCT, the 95% CI will either contain the true parameter or not (with probability 1)! In our communication to nonstatisticians, we never use the technical definition of the CI. Rather, we say that the 95% CI [0.14, 2.62] expresses that we are uncertain about the true value of Δ and that it most likely lies between 0.14 and 2.62 (with 0.95 probability). □

The 95% CI expresses our uncertainty about the parameter of interest and as such is considered to give better insight into the relevance of the obtained results than the *P*-value. However, the adjective '95%' refers to the procedure of constructing the interval and not to the interval itself. The interpretation that we give to nonstatisticians has a Bayesian flavor as will be seen in Chapter 2.

1.1.4 An historical note on the two frequentist paradigms*

In this section, we expand on the difference between the two frequentist paradigms and how they have been integrated in practice into an apparently one unifying approach. This section is not essential for the remainder of the book and can be skipped. A more in-depth treatment of this topic can be found in Hubbard and Bayarri (2003) and the papers of Goodman (1993, 1999a, 1999b) and Royall (1997).

The Fisherian and the NP approach are different in nature but are integrated in current statistical practice. Fisher's views on statistical inference are elaborated in two of his books: *Statistical Methods for Research Workers* (Fisher 1925) and *The Design of Experiments* (Fisher 1935). He strongly advocated the inductive reasoning to generate new hypotheses. Fisher's approach to inductive inference goes via the rejection of the *null-hypothesis*, say $H_0 : \Delta = 0$. His *significance test* constitutes of a statistical procedure based on a test statistic for which the sampling distribution, given that $\Delta = 0$ holds, is determined. He called the probability under H_0 of obtaining the observed value of that test statistic or a more extreme one, the P-value. To Fisher, the P-value was just a practical tool for inductive inference whereby the smaller the P-value implies a greater evidence against $\Delta = 0$. Further, according to Fisher the null-hypothesis should be 'rejected' when the P-value is small, say less than a prespecified threshold $\alpha = 0.05$ called the *level of significance*. Fisher's rule for rejecting H_0 is, therefore, when $P \leq 0.05$, but he recognized (Fisher 1959) that rejection may have two meanings: either that an exceptionally rare chance has occurred or the theory (according to the null-hypothesis) is not true.

In their approach to statistical testing, Neyman and Pearson (1928a, 1928b, 1933) needed an *alternative hypothesis* (H_a), say $\Delta \neq 0$. Once the data have been observed, the investigator needs to decide between two actions: reject H_0 (and accept H_a) or accept H_0 (and reject H_a). NP called their procedure an *hypothesis test*. Their approach to research has a decision theoretic flavor, i.e. decision makers can commit two errors: (1) type I error with probability P(type I error) $= \alpha$ (*type I error rate*) when H_0 is rejected while in fact true and (2) type II error with probability P(type II error) = β (*type II error rate*) when H_a is rejected while in fact true. In this respect they introduced the *power* of a test, equal to $1 - \beta$ for an alternative hypothesis $H_a : \Delta = \Delta_a$. NP argued that a statistical test must minimize the probability of making the wrong decision and demonstrated (Neyman and Pearson 1933) that the well-known *likelihood-ratio test* minimizes β for given a value for α. The NP approach is in fact deductive and reasons from the general to the particular and thereby makes claims from the particular to the general only in the long run. NP strived that one shall not be wrong too often. In other words, they rather advocated 'inductive behavior'. Fisher strongly disliked this viewpoint and both parties ended up in a never-ending debate.

Despite the strong historical disagreement between the proponents of the two approaches, nowadays the two philosophies are mixed up and presented as a unifying methodology. Hubbard and Bayarri (2003) (see also references therein) warned for the confusion this unification might imply, especially for the danger that the P-value is wrongly interpreted as a type I error rate. Indeed, the P-value was introduced by Fisher as a surprise index vis-à-vis the null-hypothesis and in the light of the data. It is an a posteriori determined probability. A problem occurs when the P-value is given the status of an a priori determined error rate. For example, suppose the significance level is $\alpha = 0.05$, chosen in advance. Thus, if upon completion of the study, we obtain $P = 0.023$ we say that H_0 is rejected at 0.05. However, we cannot say that H_0 is rejected at the 0.025 level, because in this sense α is chosen after the facts and P obtains the status of a prespecified level. In medical papers, the P-value is often given the nature of a prespecified value. For instance when significant results are indicated by asterisks, e.g. '*' for $P < 0.05$, '**' for $P < 0.01$ and '***' for $P < 0.001$, then the impression is created that for a '**' result significance at 0.01 can be claimed. Following Carl Popper (Popper 1959), Fisher claimed that one can never accept the null-hypothesis, only disprove it. In contrast, according to the NP approach subsequent to the hypothesis test there are two possible actions: one 'rejects' the null-hypothesis (and accepts the alternative hypothesis) or vice versa. This creates another clash of the two approaches. Namely, if the NP approach

is the basis for statistical inference, then there is in principle no problem in accepting the null-hypothesis. However, one of the basic principles in classical statistical practice is never to accept H_0 in case of a nonsignificant result that is in the spirit of Fisher's significance testing. Note that in clinical trials the standard approach is the NP approach, but accepting the null-hypothesis would be a major flaw.

Because of the above difficulties with the *P*-value and that classical statistical inference is claimed to be not coherent, Goodman (1993, 1999a, 1999b) and others advocated to use Bayesian inference tools such as the Bayes factor. Having said this, others still regard it as a useful tool in some circumstances. For instance, Hill (1996) writes: 'Like many others, I have come to regard the classical *P*-value as a useful diagnostic device, particularly in screening large numbers of possibly meaningful treatment comparisons.' Further, in some cases (one-sided hypothesis testing) the *P*-value and Bayesian inference come close (see Section 3.8.3). Finally, Weinberg (2001) argues that 'It is time to stop blaming the tools, and turn our attention to the investigators who misuse them.'

1.2 Statistical inference based on the likelihood function

1.2.1 The likelihood function

The concept of *likelihood* was introduced by Fisher (1922). It expresses the plausibility of the observed data as a function of the parameters of a stochastic model. As a function of the parameters the likelihood is called the *likelihood function*. Statistical inference based on the likelihood function differs fundamentally from inference based on the *P*-value, although both approaches were promoted by Fisher as a tool for inductive inference. To fix ideas, let us look at the likelihood function of a binomial sample. The following example dates back to Cornfield (1966) but is rephrased in terms of a surgery experiment.

Example I.7: A surgery experiment
Assume that a new but rather complicated surgical technique was developed in a hospital with a nonnegligible risk for failure. To evaluate the feasibility of the technique the chief surgeon decides to operate on $n = 12$ patients with this new procedure. Upon completion of the 12 operations, he reports $s = 9$ successes. Let the outcome of the ith operation be denoted as $y_i = 1$ for a success and $y_i = 0$ for a failure. The total experiment yields a sample of n independent binary observations $\{y_1, \ldots, y_n\} \equiv \boldsymbol{y}$ with s successes. Assume that the probability of success remains constant over the experiment, i.e. $p(y_i) = \theta$, $(i = 1, \ldots, n)$. Then the probability of the observed number of successes is expressed by the *binomial distribution*, i.e. the probability that s successes out of n experiments occur when the probability of success in a single experiment is equal to θ, is given by

$$f_\theta(s) = \binom{n}{s} \theta^s (1-\theta)^{n-s} \text{ with } s = \sum_{i=1}^{n} y_i, \tag{1.1}$$

where $f_\theta(s)$ is a discrete distribution (as a function of s) with the property that $\sum_{s=0}^{n} f_\theta(s) = 1$. When s is kept fixed and θ is varying, $f_\theta(s)$ becomes a continuous function of θ, called the *binomial likelihood function*. The likelihood function could be viewed as expressing the plausibility of θ in the light of the data and is therefore denoted as $L(\theta \mid s)$. The graphical representation of the binomial likelihood function is shown in Figure 1.2(a) for $s = 9$ and